Testing for Priors

Josiah Lopez-Wild

Abstract

The Problem of the Priors can be stated as the following question: what norms, if any, motivate our choice of one prior over another? Subjective Bayesians argue that there is no problem: your beliefs are what they are, and there is no rational requirement beyond coherence. Nonetheless I argue that in common problems in Bayesian statistics, we are uncertain which prior best represents our beliefs. Using tools from algorithmic randomness I develop a *prior determination method* that guides agents to priors in a principled manner. I argue that the method produces priors that can all claim to represent the agent's credences. I then prove that the method is algorithmically implementable in principle, and the priors thus constructed enjoy nice merging properties.

1 Introduction

Almost all Bayesians agree on two core norms. The first is Probabilism: one's credences ought to be representable as a probability measure. The second is Conditionalization: in light of new evidence, one's credences ought to be updated according to Bayes' Rule. Probabilism tells us how our credences should be structured at any given time, while Conditionalization tells us how they should change in response to new evidence. But Probabilism only says that our credences should be represented by *some* probability measure; it does not say *which* measure. The Problem of the Priors can be stated as the following question: what norms, if any, motivate our choice of one prior over another? Historically there have been two groups of responses.

Objective Bayesians such as Jaynes ([18], [19]), Rosenkrantz ([28]), and Williamson ([35]) maintain that there are additional norms beyond the requirements of Probabilism that constrain our choice of prior. Authors in this camp argue that if Bayesian methods "are to have any relevance to science, the prior distribution must be completely 'objective' in the sense that it is

independent of the personality of the user" ([18], p. 228). Jaynes formulates this idea as the criterion: "*in two problems where we have the same prior information, we should assign the same prior probabilities*" ([18], p. 228). So, on this view, we require some additional mathematical assumption on priors so that the criterion is satisfied. For example, many objective Bayesians have appealed to the Principle of Indifference, which says, roughly, that one's prior probabilities should be uniformly distributed across the sample space. This principle has generated a large literature debating is normative status, prompting the later development of the Maximum Entropy Principle ([20], [21]). This principle states that, given a set of constraints the prior must satisfy, one should pick the prior that maximizes Shannon entropy. So supporters of maximum-entropy methods recommend a two-step process. First, list structural constraints which encode the agent's prior information about the problem at hand; any permissible prior must satisfy these constraints. These constraints define a class of permissible priors, and a single prior can then be selected: any prior that maximizes entropy.

Subjective Bayesians, such as Savage ([29]), de Finetti ([10]), Skyrms ([31]), and Jeffrey ([22]), argue that there are no additional norms beyond the requirement of probabilistic coherence. Each prior is equally as good as any other, and agents must "choose" for themselves based on their subjective beliefs. Alternatively, there is no "choice" to make: your beliefs simply are what they are, and your prior is the mathematical representation of these beliefs.

While I am broadly sympathetic with the subjective position, we must admit that in real contexts there is some difficulty in determining a prior that adequately represents one's beliefs. As an example, consider a basic problem in nonparametric Bayesian statistics ([12], pp. 36–37): estimating a probability measure on the real line $\mathbb{R}$ where the measure is of completely arbitrary form (that is, we do not assume it must be Gaussian, log-normal, β, or any other particular family of measures). The natural Bayesian approach is then to define a *random measure*, that is, a prior probability measure π on the space $\mathscr{P}(\mathbb{R})$ of all possible probability measures on $\mathbb{R}$. We think of π as averaging the predictions made by the various probability measures μ in $\mathscr{P}(\mathbb{R})$. So, for example, for any Borel set $A \subseteq \mathbb{R}$, the prior π induces a probability

$$P(A) = \int \mu(A)\, d\pi(\mu).$$

Moreover P can be updated via Bayes' Rule in the natural way. Assuming independent and identically distributed observational data drawn from some unknown μ', one can show that (under certain regularity conditions) π will, in the limit, concentrate on μ', with prior π-probability 1. So Bayesian inference works as usual in this case, with just a bit more mathematical machin-

ery.

However, $\mathscr{P}(\mathbb{R})$ is an infinite-dimensional space; in general we require an infinite number of parameters to specify any particular prior π on this space. While nonparametrics *are* increasingly useful and increasingly used, this is a mathematically subtle problem; as Hjort et al. put it,

> The mathematical complexities are more demanding, since placing well-defined probability distributions on potentially infinite-dimensional spaces is inherently harder than for Euclidean spaces. Added to this is the challenge of "understanding the prior"; the ill-defined transformation from so-called "prior knowledge" to "prior distribution" is hard enough to elicit in lower dimensions and of course becomes even more challenging in bigger spaces ([15], p. 3).

The second sentence is pivotal: it is unreasonable to expect a statistician (or anyone else) to have precise beliefs about these infinite-dimensional spaces. What can be done? We must make a choice of prior based on considerations *other* than faithfulness to our beliefs; as Ghosal says,

> Typically, subjective knowledge about the minute details of the distribution on these infinite-dimensional spaces is not available for nonparametric problems. A prior is generally chosen based on tractability, computational convenience and desirable frequentist behavior …([12], p. 35).

Statisticians can circumvent some of these issues with a family of *prior-construction methods*. This is a family of algorithms that are computationally tractable, have nice analytic properties, and provably desirable frequentist (asymptotic) behavior. These models fill the chapters of any book on Bayesian nonparametrics ([13], [14]). They partially circumvent the problem by fixing various parameters of the prior independently of the statistician's beliefs. Rather than forcing the statistician to assign infinitely many parameters, they are asked to assign a small, finite number, and the prior-construction method determines the rest. While this does result in tractable priors, we have moved quite far from the ideal of a prior that represents one's subjective beliefs—these prior construction methods produce priors that are *pragmatically* useful, but at best partially represent credences.

So, we see a gap in the story of subjective Bayesian priors. Real statistical contexts make it impractical to determine the unique prior that best represents one's credences. More generally,

one might naturally experience some uncertainty regarding how best to formalize their beliefs. Faced with such uncertainty, the subjective Bayesian stance offers no guidance.

I think subjective Bayesians can do better. I propose that we want a *prior-determination method*. Such a method should take as input some form of an agent's beliefs, and produce as output some prior(s) that can reasonably be said to *represent* the agent's beliefs. In this respect the method proposed is akin to the varieties of representation theorems from decision theory ([33], [29], [2]), which ask for some subjective input (in the form of preferences, for example) and produce some mathematical object that represents the input.

In what follows I demonstrate that such a method is possible. Using tools from algorithmic randomness, I will define a method that is algorithmic in a precisely defined sense and produces a set of priors that encode information on an agent's prior beliefs. We'll begin in §2 by discussing the intuitive idea of a *sequential test*. I will argue that these sequential tests are a natural way to capture one's credences on data even in contexts such as those considered in nonparametrics. In §3 I develop the mathematical background from the theory of alorithmic randomness required to precisely define sequential tests. In §4 I define what I call *compatible measures*, which are the "output" of the prior selection method defined by sequential tests. Finally, in §5 I argue that, in the long run, the compatible priors are all roughly equally as good by appealing to merging results. Along the way we state and prove theorems that illustrate the power of the method. Taken together these results demonstrate that the proposed method produces priors that can reasonably be said to represent an agent's beliefs.

2 Algorithmic Testing

I suggest that we can use *algorithmically specifiable tests* to help us choose priors. We can motivate the idea with the following story. Suppose you are a graduate student in the lab testing coins for their statistical behavior. As a good Bayesian you want to specify a prior over the data. How should you arrive at this prior? You could consider the class of Bernoulli(α) measures, with bias $\alpha \in [0, 1]$, and define a prior over possible biases in the unit interval. It is widely agreed that this is a good model for ideal, pristine, uniformly minted coins. Alas, these coins are not ideal—from a part of the world you don't know, made of material you don't recognize, and with irregular shapes—so you also think they might *not* be correctly modeled via Bernoulli distributions. You know that the prior should encode what *you* expect to observe. But intuitively

it is very hard to specify exactly what you expect. The possible coin flips sequences could be just about anything.

However, you don't think *just any* sequence is expected. Let a 'heads' be represented by 1, and a 'tails' be represented by 0. Then you could say, for example: "The sequence 101010101010 is unexpected; I can't rule it out for sure, but I would be surprised to see it." You could say something similar about sequences such as 11111111, 00000000, 0101010101, 111111000000, and others. Each of these sequences strike you as atypical—not impossible, but very odd. This is a subjective judgment you make about the situation at hand. This procedure can easily be extended: you can set up a procedure that lists longer and longer sequences that are *abnormal* or *atypical* in this sense.

Speaking with slightly more mathematical precision, what you've set up is a computable enumeration of increasingly long prefixes of data sequences. Specifically, your procedure lists the atypical (according to you) sequences. In the limit you'll enumerate atypical infinite data sequences. That is, you have defined *non-random sequences* with respect to your beliefs about the data-generating process. This is the basic intuition behind algorithmic randomness.

I argue that these *sequential tests* are a natural way to define which data is highly atypical. They require only one's intuitions about the possible observations they will make. They do not require anything as sophisticated as precise, real-valued credences. Rather, one need only define an iterative process that generates longer and longer sequences. Sequential tests thereby also define which data is typical; the typical data is whatever is not enumerated in the limit of the sequential test. These typical data are precisely what you *do* expect to see. One of the primary functions of a prior probability is to precisely define what the agent expects to see; so, intuitively, we have made some progress toward defining a prior.

In the next section I show how to make sequential tests mathematically precise. In particular I show that one can use the mathematics of *algorithmic randomness* to define a set of priors that encode these typicality facts.

3 Algorithmic Randomness

Algorithmic randomness is, as the name suggests, a theory that uses algorithms to describe randomness. Specifically, it mixes Turing computability with probability theory to study which outcomes are *typical* or *random* with respect to some probability measure. This field has its

origins in the work of Kolmogorov ([23]), Chaitin ([8]), Martin-Löf ([25]), Schnorr ([30]), and others; modern treatments can be found in [11] and [26]. Moreover in the philosophy of science there is now a growing literature connecting algorithmic randomness to problems in Bayesian epistemology by way of computable probability theory; for examples, see [37], [17], [36], [3], [4], and [24].

Historically, algorithmic randomness was only concerned with the randomness of infinite binary sequences. Much recent work has been done to extend the concepts to other spaces such as $\mathbb{R}, \mathbb{C}, C[0,1]$, and other important spaces in analysis (see), but in this paper we will content ourselves with binary sequences. We let $2^{<\omega}, 2^{\omega}$ denote the space of finite binary strings and the space of (countably) infinite binary sequences, respectively. Since $2^{<\omega}$ is countable we fix an effective enumeration $\{\sigma_n\}_{n\in\omega}$, e.g. in length-lexicographic order. The latter is often called "Cantor space" by logicians. Given a string σ or sequence x, we let $\sigma \upharpoonright n$ (resp. $x \upharpoonright n$) denote its restriction to the first n bits.

The canonical topology on 2^{ω} is the product topology induced by the discrete topology on the set $\{0,1\}$. This topology is generated by the set of *cones* $[\![\sigma]\!] = \{x \in 2^{\omega} : \sigma \prec x\}$, where $\sigma \in 2^{<\omega}$ and "$\sigma \prec x$" means that σ is an initial prefix of x (or equivalently x extends σ). We let $\mathscr{F}$ denote the Borel σ-algebra generated by this topology. Thus $(2^{\omega}, \mathscr{F})$ is a measurable space.

Algorithmic randomness is almost always studied with respect to *computable* probability measures. We assume the reader is familiar with basic computability notions, primarily *computable* and *computably enumerable* (or *c.e.*) sets $A \subseteq \omega$ of natural numbers and *computable functions* $f : \omega \to \omega$ on natural numbers. We define the computability of a measure in terms of its behavior on cones. A probability measure μ on $\mathscr{F}$ is computable just in case, for all finite strings σ, $\mu([\![\sigma]\!])$ is a computable real, uniformly in σ. In turn, a real $r \in \mathbb{R}$ is computable if there is a computable function $f : \omega \to \mathbb{Q}$ such that, for all n, $|f(n) - r| \leq 2^{-n}$. In other words, f computes a sequence of rationals that converge to r at a computable rate. For more details see [34].

Using tools from computability theory we can assign complexity notions to subsets of 2^{ω} as well. We define an infinite hierarchy of increasingly complex subsets, known as the *arithmetical hierarchy*; see [11] or [32] for more details.

Definition 1 (Arithmetical Hierarchy)**.** The *arithmetical complexity* of a set $A \subseteq 2^{\omega}$ is defined inductively as follows.

- Let $I \subseteq \omega$ be c.e. Then $\bigcup_{i\in I} [\![\sigma_i]\!]$ is Σ^0_1.

- Let $U \subseteq 2^\omega$ be Σ^0_n. Then $2^\omega \setminus U$ is Π^0_n.
- Let $V_i \subseteq 2^\omega$ be a computable sequence of Π^0_n sets. Then $\bigcup_i V_i$ is Σ^0_{n+1}.[1]
- $U \subseteq 2^\omega$ is Δ^0_n if and only if it is both Σ^0_n and Π^0_n.

So, for example, Σ^0_1 subsets are the *effectively open* (sometimes also called *c.e. open*) sets, Π^0_1 subsets are the *effectively closed* (sometimes also called *co-c.e. closed*) sets, and Δ^0_1 subsets are the *effectively clopen* sets. Given a computable sequence $\{U_n\}_{n\in\omega}$ of Σ^0_n (resp. Π^0_n, Δ^0_n) sets, we also call the sets *uniformly* Σ^0_n (resp. Π^0_n, Δ^0_n).

Definition 2. Let $\{U_n\}_{n\in\omega}$ be a sequence of uniformly Σ^0_1 subsets of 2^ω. We say that $\{U_n\}_{n\in\omega}$ is a *Martin-Löf test* if for all n, $\mu(U_n) \leq 2^{-n}$. The set $\bigcap_{n\in\omega} U_n$ is called a *Martin-Löf null set.*

Martin-Löf tests share many similarities with the sequential tests we saw in the previous section. The sets in a Martin-Löf test are uniformly Σ^0_1, which means that they are constructed by uniformly enumerating finite strings. Their measure decreases rapidly, at a computable rate, which means that the sequences they contain are increasingly atypical as we move through the test. In the limit they define a μ-measure zero set—a Martin-Löf null set. Intuitively, much like our sequential tests, these null sets define the maximally atypical sequences. Indeed, this is precisely the definition of Martin-Löf randomness.

Definition 3. We say that a sequence $x \in 2^\omega$ is *Martin-Löf random* if x is not a member of any μ-Martin-Löf null set.

Martin-Löf random sequences are typical with respect to the measure that defines them. It is quick to show that for any μ, the set MLR^μ of μ-Martin-Löf random sequences is μ-measure 1. Moreover Martin-Löf randoms witness theorems from classical probability theory. For example, suppose μ is a Bernoulli measure with computable bias α. By the Strong Law of Large Numbers, the set

$$\mathsf{SLLN}_\alpha = \left\{ x \in 2^\omega : (\forall k)(\exists n)(\forall m \geq n) \left| \frac{\#0(x \upharpoonright n)}{n} - \alpha \right| \leq 2^{-k} \right\}$$

has μ-measure 1. One can show that each Martin-Löf random sequence witnesses this limiting relative frequency, that is, $\mathsf{MLR}^\mu \subseteq \mathsf{SLLN}_\alpha$. For this reason Zaffora Blando says that

[1] That is, there is a computable function $f : \omega \to \omega$ such that $f(n)$ is the index of the n^{th} Π^0_n set, V_n.

> algorithmic randomness notions may be taken to encode a specific type of inductive assumptions——or commitments (either explicit or implicit)——that results from the subjective prior with respect to which randomness is defined. This is because algorithmic randomness notions embody the effective statistical laws that an agent expects to see in the data by virtue of having a certain prior ([37], p. 938).

So we can gloss random sequences as those sequences that capture all of your inductive assumptions—your expectations—about future observations. Using this intuition we can rigorously define the testing procedure that I described in the previous section.

Definition 4. We call a sequence of subsets (U_n) of 2^ω a *sequential test* if

1. it is nested and decreasing; that is, $U_i \supseteq Ui + 1$ for all $i \in \omega$);
2. the U_n are uniformly Σ^0_1 sets; and
3. for some $n \in \omega$, $U_n \neq 2^\omega$.

We can gloss these requirements as follows. The test is decreasing because as the procedure goes on we produce longer and longer data sequences that are more and more abnormal. The test is nested because the later sequences extend the earlier sequences. The test is uniformly Σ^0_1 because we have a single procedure for enumerating abnormal sequences; Finally, the test eventually decides that some sequences are nonrandom—it can't treat every sequence as abnormal. Without loss of generality we may assume that, given a sequential test (U_n), $U_1 \neq 2^\omega$, since if there is $m > 1$ such that $U_m \neq 2^\omega$ (as required by the definition of sequential tests) then the sequence (U_{n+m}) is a sequential test that defines the same null set.

For the reasons given in the previous section, sequential tests are a natural way to encode *what data you expect to see.* They are also clearly a generalization of Martin-Löf tests that do not specify a particular probability measure. We can now be more precise about the proposed procedure for prior determination. It consists of three simple steps.

1. (Specification). Define a sequential test (U_n) that captures what you do not expect to see.
2. (Formalization). Find the set of priors μ such that (U_n) is a μ-Martin-Löf test.
3. (Determination). Using the set produced by Formalization, determine a prior that best encodes your beliefs.

It remains to show that this procedure can be made precise. Specification is context-specific. We think of Specification as the "input" step, where a Bayesian presents their typicality beliefs to the procedure. As long as these beliefs can be written as a sequential test, this step of the procedure is satisfied.

Our next task is to flesh out Formalization: starting with a sequential test (U_n), what can we say about the set of priors μ such that (U_n) is a μ-Martin-Löf test? I call these priors *compatible measures,* and in the next section we will discuss what we know about them.

4 Compatible Measures

In this section we define and study compatible measures.

Definition 5. Let (U_n) be a sequential test. We say that a probability measure μ is *compatible with* (U_n) if (U_n) is a μ-Martin-Löf test.

We do not require the measure to be computable because we are interested in the arithmetical complexity of the set of all compatible measures, considered as a subset of the space $\mathscr{P}(2^\omega)$ of probability measures on 2^ω, most of whose elements are non-computable.[2] So the intuition for Definition 5 is: the sequences that (U_n) says are atypical are also nonrandom with respect to μ. This definition also leaves open that μ may have other, distinct Martin-Löf tests. If there are other tests then those tests will define distinct Martin-Löf null sets. So it's still possible that μ considers other sequences nonrandom, sequences that fall outside the null set defined by (U_n). We will address this issue in the next section. But for now, compatibility is a minimal notion—a measure is compatible if its typical sequences are *at most* those defined by (U_n), or in other words, the measure does not expect to see anything that (U_n) says is atypical.

Suppose we define a sequential test $\{U_n\}_{n\in\omega}$. We first want to know that it's possible to find a compatible probability measure. We would further like the measure to be computable, both because (i) algorithmic randomness is better behaved with respect to computable measures and

[2]Defining randomness with respect to non-computable measures is more involved than the case of computable measures; see [27] and [9] for details. Briefly, a (representation of a) non-computable measure may compute some sequences $x \in 2^\omega$ that nevertheless fall outside every μ-sequential test defined by uniformly Σ^0_1 sets. Intuitively, if the measure can exactly compute the sequence, it should not be random. Thus the more elaborate definition takes into consideration all possible representations $R_\mu \in 2^\omega$ of μ, and defines tests in terms of uniformly $\Sigma^0_1(R_\mu)$ sets. Thus the set of random points with respect to a non-computable measure is a strict subset of the points falling outside their standard Martin-Löf null sets, so they are still rightly called *compatible* in the sense of Definition 5.

because (ii) one of our desiderata was that the prior selection method should be algorithmic. Hence, we want to be able to compute a compatible measure. Our first theorem demonstrates that this is always possible.

Theorem 1. *Let (U_n) be a sequential test. There exists a computable measure μ that is compatible with (U_n).*

In order to prove this theorem we require a basic fact about Π^0_1 subsets of 2^ω. A set $T \subseteq 2^{<\omega}$ of finite binary strings is a *tree* if it is closed under prefixes; that is, for all n and all $\sigma \in 2^{<\omega}$,

$$\sigma \in T \iff \sigma \restriction n \in T.$$

Since any tree T is a set of finite strings, T is countable; thus one can straightforwardly define computability on trees (see e.g. [32]). As with any countable set, we say that a tree T is *computable* if there is a Turing machine that decides membership in T. The *body* of T, also called the *set of paths through* T, is defined

$$[T] := \{x \in 2^\omega \mid (\forall n)\ x \restriction n \in T\}$$

Finally, a *splitting node* of a tree T is a string $\sigma \in 2^{<\omega}$ such that $\sigma 0 \in T$ and $\sigma 1 \in T$. In other words, both possible extensions of σ are in T, or pictorially, the tree T "splits" at σ. Trees are particularly important in computability theory; one of their uses is illustrated by the following lemma.

Lemma 1. *A set $A \subseteq 2^\omega$ is Π^0_1 iff $A = [T]$, where T is a computable tree.*

In fact this characterization is so central to computability theory that it is often given as a *definition* of Π^0_1 sets; see e.g. Soare ([32]), Definition 3.7.1. In fact one can also show something weaker: if T is a co-c.e. tree then $[T]$ is a Π^0_1 set. This does not conflict with Lemma 1: given a co-c.e. tree T, $A = [T]$ is Π^0_1, and hence there is a (possibly distinct) computable tree T' such that $A = [T']$ also holds.

Proof of Theorem 1. Set $\mu(U_1) = 0$. We know that $2^\omega \setminus U_1$ is nonempty and Π^0_1. Thus it has the form of the set of paths $[T]$ through a computable tree $T \subseteq 2^{<\omega}$. At any node $\sigma \in T$, if σ is not a splitting node, assign $\mu[\![\sigma]\!] = \mu[\![\sigma \restriction |\sigma| - 1]\!]$, that is, let it have the same probability as the previous node. Otherwise σ splits, so assign $\mu[\![\sigma]\!] = \frac{1}{2}\mu[\![\sigma \restriction |\sigma| - 1]\!]$. Splitting in a computable tree is decidable (just check the computable conditions "$\sigma 0 \in T$" and "$\sigma 1 \in T$"), so

the probability assigned to every cone is a uniformly computable real. Thus μ is a computable probability measure. Finally $\mu(U_n) = 0$ for all n, and so (U_n) is a μ-Martin-Löf sequential test. □

Theorem 1 tells us that once we fix a sequential test (U_n), the set of measures that are compatible with (U_n) is nonempty—in particular, it contains a computable element. This is very important because it means that any sequential test we use to describe our credences admits *at least one* compatible probability measure; we are never in the position of finding that our intuitions on typicality are impossible to satisfy.

So we have at least one measure, but presumably there are many that are compatible with our sequential test. Trying to describe all of them individually is impractical, but we can instead try to describe the whole set of compatible measures. One natural question is: how difficult is this set to describe? To answer this question we can use arithmetical complexity to categorize sets of probability measures.

To see how this works we make a quick detour to consider the space $\mathscr{P}(2^\omega)$ of probability measures on Cantor space. It is a general result of functional analysis ([1]) that, whenever X is a complete separable metric space, then $\mathscr{P}(X)$ is itself a complete separable metric space when equipped with the topology of weak convergence (also known as the weak topology or the weak* topology). The weak topology characterizes so-called *weak convergence* of probability measures; we say that a sequence of probability measures (μ_n) weakly converges to μ if for all bounded continuous functions f we have $\int f \, d\mu_n \to \int f \, d\mu$. This is the most commonly studied notion of convergence for probability measures (see [5]), though of course there are others. There are a variety of suitable metrics (see [5]), but their details don't concern us here.

What does concern us is the fact that one can define the arithmetical hierarchy on subsets of $\mathscr{P}(2^\omega)$ when it is equipped with the weak topology.[3] Earlier in Definition 1 we defined arithmetical complexity inductively on subsets of 2^ω in terms of effective unions and intersections. The same trick applies in $\mathscr{P}(2^\omega)$.

Definition 6. Let $\mathscr{P}(2^\omega)$ be equipped with the weak topology. Let $U \subseteq 2^\omega$.

1. U is Σ^0_1 if there is a c.e. set $I \subseteq \mathbb{N}$ such that $U = \bigcup_{i \in I} B(s_i, q_i)$, where $s_i \in S$ and $q_i \in \mathbb{Q}$, for all i.

[3] For those familiar with computable analysis, this fact follows from the result that if X is a computable Polish space, then $\mathscr{P}(X)$ with the weak topology is a computable Polish space. See [16] for a proof.

2. U is Π^0_1 if $2^\omega \setminus U$ is Σ^0_1.

3. Let (U_n) be a sequence of uniformly Π^0_n subsets of 2^ω. Then $U = \bigcup_n U_n$ is Σ^0_{n+1}, and $2^\omega \setminus U$ is Π^0_{n+1}.

Intuitively, a set is Σ^0_1 if we can enumerate some simple points that are in the set (the s_i) together with conditions of the form "any point within distance q_i of s_i is also in the set". So while we can't necessarily enumerate all the elements of a Σ^0_1 subset of 2^ω, we *can* enumerate sufficient conditions for membership. A set is Π^0_1 if we can enumerate sufficient conditions for points to fall outside the set. In the inductive step, the uniformity requirement is crucial; if we take arbitrary unions the complexity of the resulting set can be much higher.[4]

Returning to the main line of inquiry: we want to know how difficult it is to describe the set of compatible measures. More precisely, what is its arithmetical complexity? Our next result, Theorem 2, shows that the set is about as simple as possible in terms of arithmetical complexity: it is a Π^0_1 set (a co-c.e. closed set) in the weak topology on $\mathscr{P}(2^\omega)$. To prove this result we need the following lemma on Π^0_1 sets.

Lemma 2. *Suppose $\{U_i\}_{i\in I}$ is a sequence of Π^0_1 sets indexed by a c.e. set $I \subseteq \omega$, and let T_i be the corresponding computable trees. Then $\bigcap_{i\in I} T_i$ is a co-c.e. tree, and hence $\bigcap_{i\in I} U_i = [\bigcap_{i\in I} T_i]$ is Π^0_1.*

Proof. Let $\{T_i\}_{i\in I}$ be a set of computable trees $T_i \subseteq 2^{<\omega}$, with $I \subseteq \omega$ a c.e. index set. Let $\{i_j\}_{j\in\omega}$ be an enumeration of I and let $\sigma \in \omega^{<\omega}$. If $\sigma \notin \bigcap_{i\in I} T_i$ then there must be some finite set $\{i_1, i_2, \ldots, i_n\} \subseteq I$ such that $\sigma \notin \bigcap_{j=1}^n T_{i_j}$, which can be decided by a Turing machine. On the other hand the condition $\sigma \in \bigcap_{i\in I} T_i$ cannot be decided. Since the intersection of a family of trees is itself a tree, we see that $\bigcap_{i\in I} T_i$ is a co-c.e. tree. Therefore, $[\bigcap_{i\in I} T_i]$ is the body of a co-c.e. tree and hence is Π^0_1. □

Theorem 2. *Let (U_n) be a sequential test. The set of (U_n)-compatible probability measures is Π^0_1 in the weak topology on $\mathscr{P}(2^\omega)$.*

Proof. For each n the set

$$M_{U_n} := \{\mu : \mu(U_n) \leq 2^{-n}\}$$

[4] Taking arbitrary unions of Π^0_n sets, for example, produces a $\boldsymbol{\Box^0_{n+1}}$ set, where the latter classification is in the Borel hierarchy.

is Π^0_1. Therefore (M_{U_n}) is a c.e. indexed sequence of Π^0_1 sets. Being Π^0_1 each M_{U_n} is the body of a computable tree T_n. By Lemma 2 we have that $T = \bigcap_{n\in\omega} T_n$ is a co-c.e. tree, and hence $[T]$ is itself a Π^0_1 set. That is, $M = \bigcap_n M_{U_n}$ is Π^0_1. But this is precisely the set of probability measures μ for which $\mu(U_n) \leq 2^{-n}$ for all n, i.e., the set of measures for which (U_n) is a Martin-Löf test. □

So this theorem says that the set M of compatible measures is about as computationally simple as it can be: it's an effectively closed subset of $\mathscr{P}(2^\omega)$. This is both surprising and useful. It is surprising because agreement on Martin-Löf tests is, intuitively, a condition that is cumbersome to specify. It is useful because Π^0_1 sets form a central area of study in computability theory, and their properties are well-understood; see [7] and [32] for examples. So we can expect that this result will be mathematically fruitful in future developments.

Perhaps most important of all is its philosophical relevance. One of our desiderata for a prior determination method given in §2 was: the method should be algorithmically implementable, at least in principle. One way to cash out "in-principle algorithmic implementability" is: our Formalization step should produce a computationally simple set of measures. In this case we have a set that is definable by a Π^0_1 formula of arithmetic: that is, some formula containing only universal quantifiers. Therefore it is a co-c.e. condition. That is, as we enumerate our sequential test, we are able to rule out those measures which are not compatible with our typical intuitions. In the limit, as we define the full test, we arrive at the remaining set of measures, all of which are compatible.

Ideally we would define a Δ^0_1 set of measures since this is the lowest level of the arithmetical hierarchy. This may yet be possible in $\mathscr{P}(2^\omega)$ (we leave this as an open question), but it's important to note that this is *not* possible in general; if X is some computable Polish space, it is not in general true that $\mathscr{P}(X)$ has *any* nontrivial Δ^0_1 subsets.[5] So in general we settle for either Σ^0_1 or Π^0_1 subsets as the "simplest" we can find; thus Theorem 2 tells us that M is as simple as it can be in this general setting.[6]

So we have cashed out Formalization as a satisfying, algorithmically implementable procedure. We can complete our prior determination procedure by defining the Determination step.

[5]This is because a Δ^0_1 subset of a computable Polish space X is clopen (i.e., $\Box^{\mathbf{0}}_{\mathbf{1}}$) in the topology on X, but many computable Polish spaces (for example, $\mathbb{R}$) do not have any nontrivial clopen subsets.

[6]While we don't show it here, it's quite straightforward to generalize the proof of Theorem 2 to general computable Polish spaces using techniques from computable analysis.

5 Priors and Merging

Recall our Determination step:

- (Determination). Using the set produced by Formalization, determine a prior that best encodes your beliefs.

We have shown in Theorem 1 that Formalization always returns at least one computable compatible measure, and we have also shown in Theorem 2 that the whole set of compatible measures is computationally simple to specify. Now we are faced with this set of priors. We can agree that the set has been chosen based on solid principles, but Determination is vague about how we proceed.

Advocates of imprecise probabilities can feel free to stop here. They are presented with a set of priors that, I have argued, all agree on their predictions in a principled manner—they all encode the agent's typicality beliefs encoded by the sequential test. Advocates of precise priors must still make a choice among them, however. At this point subjective and objective Bayesians must rehash their debates concerning the correctness of this choice of prior. I argue that we are in a much better position than we were before. To do this I will argue that the choice of prior does not matter too much, because in the long run the priors will come to agree anyway.

The "merging of opinions" theorem due to Blackwell and Dubins ([6]) is a cornerstone of Bayesian epistemology and statistics. It says, roughly, that if two priors are sufficiently compatible then, in the limit of increasing data, they will come to have the same credences on all propositions. More precisely, let $(X, \mathscr{F})$ be a measurable space, and let $(\mathscr{F}_n)$ be a sequence of sub-σ-algebras of $\mathscr{F}$ such that (i) $\mathscr{F}_n \subseteq \mathscr{F}_{n+1}$ for all n, and (ii) $\sigma\left(\bigcup_n \mathscr{F}_n\right) = \mathscr{F}$. We call $(\mathscr{F}_n)$ a *filtration*. Suppose μ is a probability measure on $\mathscr{F}$. Then its conditional probability $\mu(A \mid \mathscr{F}_n)$ of A conditional on $\mathscr{F}_n$ is a $\mathscr{F}_n$-measurable function such that, for all $B \in \mathscr{F}_n$,

$$\mu(A \cap B) = \int_B \mu(A \mid \mathscr{F}_n)\, d\mu.$$

Finally, given two probability measures μ, ν, we say that μ is *absolutely continuous* with ν, written $\mu \ll \nu$, if for all $A \in \mathscr{F}$, if $\mu(A) > 0$ then $\nu(A) > 0$. Then the theorem is stated as follows.

Proposition 1 ([6])**.** *Suppose μ and ν are probability measures on $(X, \mathscr{F})$, and let $\mathscr{F}_n$ be a filtration of $\mathscr{F}$. Suppose $\mu \ll \nu$. Then μ-a.s.*

$$\sup_{A \in \mathscr{F}} |\mu(A \mid \mathscr{F}_n) - \nu(A \mid \mathscr{F}_n)| \to 0$$

as $n \to \infty$. We say that ν merges with μ, written $\nu \xrightarrow{M} \mu$.

Crucially this merging happens on a set of μ-measure one, but not necessarily on a set of ν-measure 1.[7] So, $\nu \xrightarrow{M} \mu$ means that μ believes that ν will come to agree with it on all beliefs, given enough information.

How does this help with prior determination? Well, suppose you are undecided between two measures μ and ν. They distribute probability over X differently, but they agree on what is possible: anytime $\mu(A) > 0$, then $\nu(A) > 0$, and vice versa. In other words, they are mutually absolutely continuous. In that case both μ and ν believe that with probability one the other will merge to it; in the long run, both priors expect to have the same credences. Put another way, both priors expect their differences to wash out over time; their disagreements will appear only in the short-to-medium run.

Any two compatible priors in M need not be absolutely continuous with one another. So, the merging of opinions theorem does not immediately help us. But if we are undecided between multiple compatible priors then we can shift the Bayesian apparatus up a level of abstraction and put a prior on these measures, as one does in statistics. In particular, we only care about computable measures and there are only countably many of them.

Theorem 3. *Let (μ_n) be a uniformly computable sequence of probability measures on 2^ω. Let (π_n) be a uniformly computable sequence of nonnegative reals such that $\sum_{i=1}^\infty \pi_i = 1$. Let $\lambda := \sum_{i=1}^\infty \pi_i \mu_i$. Then $\lambda \xrightarrow{M} \mu_i$ whenever $\pi_i > 0$.*

Proof. Immediate from the fact that $\mu_i \ll \lambda$ for all i. □

Requiring the π_i to be uniformly computable is equivalent to defining a computable prior $\pi(\mu_i)$ on the computable elements of M. Thus we see that λ is itself a computable prior that is compatible with (U_n). And while Theorem 3 is not *mathematically* interesting, it does tell us that if we are uncertain about which prior μ_i best captures our beliefs, we can instead choose a mixture—one that each μ_i believes is just as good, in the long run.

Does λ "best encode our beliefs", as Determination demands? For a subjective Bayesian the answer is, of course, subjective. We can say something more. The Formalization step defines M in terms of the sequential test (U_n). In a precise sense these compatible measures

[7] If $\mu \ll \nu$ and also $\nu \ll \mu$ then we say that μ and ν are *mutually absolutely continuous*. So this sentence is really saying that absolute continuity does not imply mutual absolute continuity. To see this, let ν be the uniform measure on $[0, 1]$, and let μ be the uniform measure on $[0, 1/2]$. Then $\mu \ll \nu$, but $\nu \not\ll \mu$.

encode the typicality facts that are implicit in (U_n). But, as we noted before, each compatible μ_i may define other Martin-Löf tests, and hence its set of random points may be strictly smaller than the set defined by (U_n). In this sense, each μ_i may encode *some* but not *all* the typicality facts in (U_n). So a natural question is: can we find compatible measures with *bigger* sets of random sequences, measures that intuitively capture more of the typicality facts laid down during Specification? The next theorem says that this is precisely what λ does. We show that λ agrees with its component measures on Martin-Löf randoms.

Theorem 4. *Under the assumptions of Theorem 3,* $\mathsf{MLR}^{\lambda} = \bigcup_n \mathsf{MLR}^{\mu_n}$.[8]

Proof. See Appendix A. □

Let's recap. We defined a sequential test. We used our sequential test to define a nice set M of compatible priors. We then asked how to choose a single prior from this set. Appealing to the Blackwell-Dubins theorem, I argued that if the priors are absolutely continuous then, as time goes on, it will not matter much which prior we pick. However, not all priors in M are absolutely continuous. So, I suggested that we can define a computable prior over M that encodes our uncertainty about which prior to pick. Under those conditions Theorem 3 tells us that each component measure μ_i will merge with this mixture measure λ; even stronger, Theorem 4 shows us that these mixture measures encode precisely the typicality facts of all their component measures. In this sense we can construct mixture measures that represent one's beliefs, as encoded in one's typicality judgments (as (U_n)), arbitrarily well.

Where does *this* prior, π, come from? Yet again the standard disagreement between objective and subjective Bayesians comes back. But we have made progress: everyone to the party can agree that these priors in M have been selected for good reasons. So one could simply select a prior π because it captures one's subjective credences, or pick the uniform prior because it maximizes entropy. One could even define a sequential test on $\mathscr{P}(2^{\omega})$ to help define a suitable prior. However we choose to do it, we are ensured that the result is a mixture of priors that represent our expectations.

[8] Zaffora Blando ([37]) has proved a tight connection between agreement on algorithmic randomness notions, such as Martin-Löf randomness, and the merging of opinions theorem. More precisely her results show that if μ and ν are computable probability measures and if $\mathsf{MLR}^{\mu} \subseteq \mathsf{MLR}^{\nu}$ then $\mu \ll \nu$, which thereby implies the Blackwell-Dubins theorem. Hence Theorem 4 really is a strengthening of Theorem 3, as the latter is a consequence of the former.

6 Conclusion

We were motivated by two intuitions: (i) priors *should* encode your subjective expectations about data, but (ii) in many cases your expectations might not specify a unique prior. To resolve this dilemma we surveyed some prior selection methods. I proposed a method that generalizes those surveyed, isolating three desiderata: first, that the method should produce prior(s) that represent our credences; second, that the method should produce a set of priors that are all reasonably good in some precise sense; third, that the method should be algorithmically implementable, at least in principle.

I think the results given in this paper satisfy these desiderata. Sequential tests encode your expectations by effectively defining atypical sequences. The result is a Π^0_1 set of priors compatible with that test. Each of these priors expects to see some subset of the data deemed "typical" by the sequential test; thus they all agree with our expectations encoded in the sequential test. Moreover the fact that the set is Π^0_1 makes it about as computationally simple as a set of measures can be; it is algorithmically implementable in principle, as described in §5. Once we have defined our set of compatible measures a unique prior can be picked on subjective grounds, maximum entropy principles, or anything else. If we are uncertain about which prior is best, we can take mixtures; Theorem 4 guarantees that each candidate prior will merge with the mixture measure.

The work presented here also has natural avenues for further work. First, we defined measures as compatible when they agreed on a given Martin-Löf test. We did not, however, study the stronger notion of compatibility where measures agree on exactly the same set of Martin-Löf randoms. To fit this requirement into our framework it would be natural to consider a *universal* generalized sequential test, but this notion has eluded me thus far. There are also other randomness notions, such as Schnorr or Kurtz randomness, that could be studied from the perspective of generalized sequential tests. Finally, one might wonder whether the "constraints" that play a central role in maximum-entropy methods ([18]) could be studied from this perspective, since any constraint on sets of measures thereby also constrains the set of points that are random with respect to those measures.

A Proof of Theorem 4

Theorem 4. *Let (μ_n) be a uniformly computable sequence of probability measures on 2^ω. Let (π_n) be a uniformly computable sequence of positive reals such that $\sum_{i=1}^{\infty} \pi_i = 1$. Let $\lambda := \sum_{i=1}^{\infty} \pi_i \mu_i$. Then $\mathsf{MLR}^\lambda = \bigcup_n \mathsf{MLR}^{\mu_n}$.*

As is often the case, one direction of the proof is quite simple while the other is more involved. The more involved is the "forward" or "left-to-right" direction; our strategy is to take a Martin-Löf test $\{V_n^i\}_{n\in\omega}$ for each component measure μ_i and use them to construct a λ-Martin-Löf test. We do so by taking finite intersections of more and more of the component Martin-Löf tests, while simultaneously moving "down" the tests themselves. If we make sure most of λ's probability mass is concentrated on these first few component measures, then the result is a λ-Martin-Löf test that is precisely the intersection of all the component tests. See Figure 1 for an illustration of the first two stages of this construction.

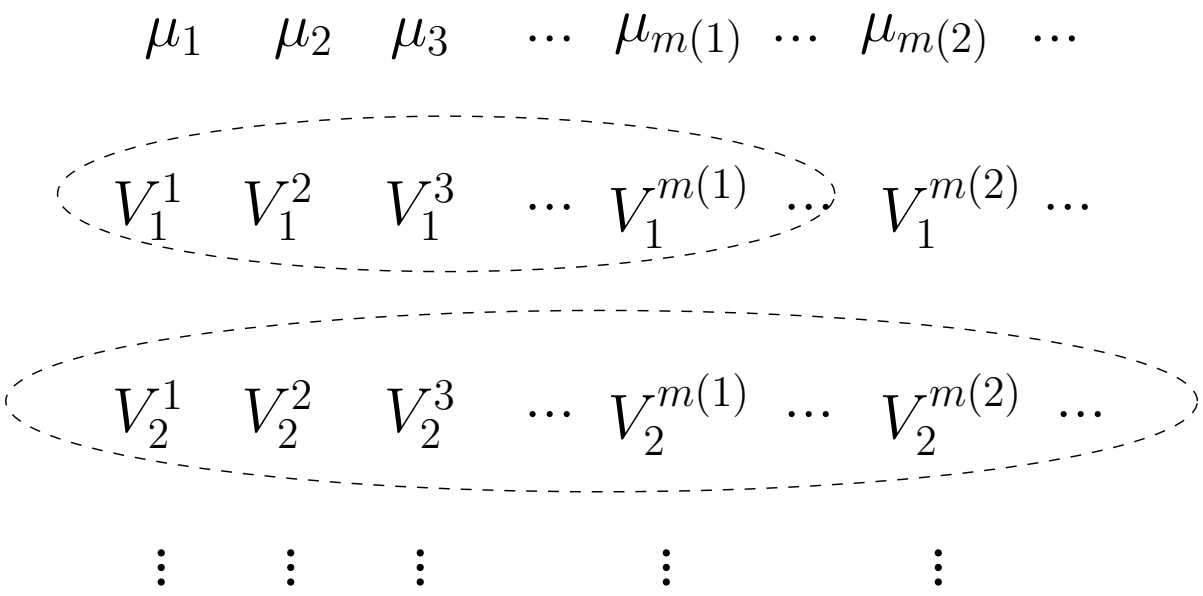

Figure 1: The construction of $\mathscr{V}$

Proof. ($\Leftarrow$). Let $x \notin \mathsf{MLR}^\lambda$, and fix a λ-Martin-Lof test $\{U_n\}_{n\in\omega}$ such that $x \in \bigcap_n U_n$. We want to show that, using this λ-Martin-Löf test, we can define a μ_i-Martin-Löf test for each i which covers x. Note that for all n,

$$2^{-n} \geq \lambda(U_n) = \sum_{i=1}^{\infty} \pi_i \mu_i(U_n),$$

so for each $i, n \in \omega$ we have

$$\mu_i(U_n) \leq \frac{2^{-n}}{\pi_i}.$$

Since π_i is independent of n we can find $N(n,i) > n$ such that $(2\pi_i)^{-N(n,i)} \leq 2^{-n}$, which means that $\mu(U_{N(n,i)}) \leq 2^{-n}$. Therefore for each μ_i there is a sequence $\{U_{N(n,i)}\}_{n\in\omega} \subseteq \{U_n\}_{n\in\omega}$ that is a μ_i-Martin-Löf test. For any i we have

$$\bigcap_{\{N(n,i)\}_{n\in\omega}} U_{N(n,i)} \supseteq \bigcap_{n\in\omega} U_n.$$

Therefore $x \notin \mathsf{MLR}^{\mu_i}$ for any i.

($\Rightarrow$). Let $x \notin \bigcup_i \mathsf{MLR}^{\mu_i}$. For each i let $\{V_n^i\}_{n\in\omega}$ be a μ_i-Martin-Löf test such that $x \in \bigcap_n V_n^i$, and assume without loss of generality that for each i, $\{V_n^i\}$ is nested. The proof strategy is similar to before: using the μ_i-Martin-Löf tests we want to construct a λ-Martin-Löf test that covers x.

First we build the test. One can effectively enumerate sufficiently large $m(n) < \infty$ such that $\sum_{i=1}^{m(n)} \pi_i > 1 - 2^{-n}$, so the set $\{m(n)\}_{n\in\omega}$ is c.e. Define

$$\mathscr{V}_n := \bigcap_{i=1}^{m(n)} V_n^i,$$

that is, at stage n we take the intersection of the n^{th} level of the Martin-Löf tests of the first $m(n)$ many measures (see Figure 1). We calculate

$$\lambda(\mathscr{V}_n) = \sum_{i=1}^{m(n)} \pi_i\mu_i\left(\bigcap_{i=1}^{m(n)} V_n^i\right) + \sum_{i>m(n)} \pi_i\mu_i\left(\bigcap_{i=1}^{m(n)} V_n^i\right); \tag{1}$$

since the μ_i are probability measures we have

$$\sum_{i=1}^{m(n)} \pi_i\mu_i\left(\bigcap_{i=1}^{m(n)} V_n^i\right) \leq 2^{-n}, \tag{2}$$

and for the second term of Equation 1 we have

$$\sum_{i>m(n)} \pi_i\mu_i\left(\bigcap_{i=1}^{m(n)} V_n^i\right) \leq \sum_{i>m(n)} \pi_i < 2^{-n}. \tag{3}$$

So we conclude that $\lambda(\mathscr{V}_n) \leq 2^{-(n-1)}$. Moreover the $\mathscr{V}_n$ are c.e.-indexed finite intersections of uniformly Σ_1^0 sets and hence are uniformly Σ_1^0 themselves. Thus $\{\mathscr{V}_{n+1}\}_{n\in\omega}$ is a λ-Martin-Löf test.

Now we show that the test covers x. Since $\pi_i > 0$ for all i, we know that for all i there is $m(n) > i$; thus, since the $\{V_n^i\}$ are nested we conclude that

$$\bigcap_n \mathscr{V}_n = \bigcap_n \bigcap_{i=1}^{m(n)} V_n^i = \bigcap_i \bigcap_n V_n^i,$$

where we can swap the order of the intersections since $m(n) \to \infty$ as $n \to \infty$. Since $x \in \bigcap_n V_n^i$ for each i, we have $x \in \bigcap_n \mathscr{V}_n$, so $x \notin \mathsf{MLR}^\lambda$. □

References

[1] Charalambos D Aliprantis and Kim C Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer, 2006.

[2] F. J. Anscombe and R. Aumann. A definition of subjective probability. *The Annals of Mathematical Statistics*, 34(1):199–205, 1963.

[3] Gordon Belot. That does not compute: David Lewis on credence and chance. *Philosophy of Science*, page 1–10, 2023.

[4] Gordon Belot. Unprincipled. *The Review of Symbolic Logic*, page 1–40, 2023.

[5] Patrick Billingsley. *Convergence of Probability Measures*. Wiley, 2nd edition, 1999.

[6] David Blackwell and Lester Dubins. Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, 33:882–886, 1962.

[7] Douglas Cenzer and Jeffrey B. Remmel. *Effectively Closed Sets*. Cambridge University Press, To appear.

[8] Gregory J Chaitin. A theory of program size formally identical to information theory. *Journal of the ACM*, 22(3):329–340, 1975.

[9] Adam Day and Joseph Miller. Randomness for non-computable measures. *Transactions of the American Mathematical Society*, 365(7):3575–3591, 2013.

[10] Bruno de Finetti. *Theory of Probability*, volume 1. Wiley, 1974.

[11] Rodney Downey and Denis Hirschfeldt. *Algorithmic Randomness and Complexity*. Springer, 2010.

[12] Subhashis Ghosal. The dirichlet process, related priors and posterior asymptotics. In *Bayesian Nonparametrics*, chapter 2, pages 35–79. Cambridge University Press, 2010.

[13] Subhashis Ghosal and Aad W Van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press, 2017.

[14] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G Walker. *Bayesian Nonparametrics*, volume 28. Cambridge University Press, 2010.

[15] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G. Walker. An invitation to bayesian nonparametrics. In *Bayesian Nonparametrics*, pages 1–21. Cambridge University Press, 2010.

[16] Mathieu Hoyrup and Cristóbal Rojas. Computability of probability measures and Martin-Löf randomness over metric spaces. *Information and Computation*, 207(7):830–847, 2009.

[17] Simon M. Huttegger, Sean Walsh, and Francesca Zaffora Blando. Bayesian convergence for computably bounded agents. *Philosophy and Phenomenological Research*, 112(1):307–324, 2026.

[18] Edwin T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):227–241, 1968.

[19] Edwin T. Jaynes. The well-posed problem. *Foundations of Physics*, 3(4):477–492, 1973.

[20] Edwin T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70:939–952, 1982.

[21] Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[22] Richard Jeffrey. *The Logic of Decision*. McGraw-Hill, 1965.

[23] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics*, 2(1-4):157–168, 1968.

[24] Josiah Lopez-Wild. A computable von Neumann-Morgenstern representation theorem. *Synthese*, 205(5):182, 2025.

[25] Per Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

[26] André Nies. *Computability and Randomness*. Oxford University Press, 2009.

[27] Jan Reimann and Theodore Slaman. Measures and their random reals. *Transactions of the American Mathematical Society*, 367(7):5081–5097, 2015.

[28] Roger D. Rosenkrantz. *Foundations and Applications of Inductive Probability*. Ridgeview, 1981.

[29] Leonard J. Savage. *The Foundations of Statistics*. Wiley & Sons, 1954.

[30] C. P. Schnorr. *Zufälligkeit und Wahrscheinlichkeit*, volume 218 of *Lecture Notes in Mathematics*. Springer, 1971.

[31] Brian Skyrms. *Pragmatics and Empiricism*. Yale University Press, 1984.

[32] Robert I. Soare. *Turing Computability: Theory and Applications*. Springer, 2016.

[33] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[34] Klaus Weihrauch. *Computable Analysis: An Introduction*. Springer, 2000.

[35] Jon Williamson. *In Defense of Objective Bayesianism*. Oxford University Press, 1981.

[36] Francesca Zaffora Blando. Pride and probability. *Philosophy of Science*, 91(3):634–660, 2024.

[37] Francesca Zaffora Blando. Bayesian merging of opinions and algorithmic randomness. *The British Journal for the Philosophy of Science*, 76(4):921–952, 2025.