stats

MDPI

Article

Improper Priors via Expectation Measures

Peter Harremoës

GSK Department, Niels Brock, Copenhagen Business College, Nørre Voldgade 34, 1358 Copenhagen K, Denmark; harremoes@ieee.org

Abstract

In Bayesian statistics, the prior distributions play a key role in the inference, and there are procedures for finding prior distributions. An important problem is that these procedures often lead to improper prior distributions that cannot be normalized to probability measures. Such improper prior distributions lead to technical problems, in that certain calculations are only fully justified in the literature for probability measures or perhaps for finite measures. Recently, expectation measures were introduced as an alternative to probability measures as a foundation for a theory of uncertainty. Using expectation theory and point processes, it is possible to give a probabilistic interpretation of an improper prior distribution. This will provide us with a rigid formalism for calculating posterior distributions in cases where the prior distributions are not proper without relying on approximation arguments.

Keywords: Bayesian statistics; expectation measure; improper prior distribution; expected value; point process; Poisson point process; s-finite measure; posterior distribution; statistical model; stopping time

MSC: 60A05; 60G55

1. Introduction

In Bayesian statistics, we usually use probability measures to quantify uncertainty. These probability measures are defined as measures with total mass equal to 1. Before we do any calculations, we need a prior distribution, so we need guidelines about how such prior distributions should be assigned to a specific problem. A subjective Bayesian would have consistency as the only limitation on how prior distributions are assigned. A significant problem with this approach is that it is subjective, so that more or less any conclusion can be reached by a suitable choice of prior distribution. On the contrary, an "objective" Bayesian would advocate for specific methods for determining prior distributions in particular situations. Although such methods may not be objective in any absolute sense, the aim should be that the methods are intersubjective in the sense that different scientists would get the same prior distribution if they agree that certain conditions are fulfilled.

Objective Bayesians have developed different methods for assigning prior distributions, and a significant problem is that these methods often lead to improper prior distributions, where the prior distributions are described by measures that have infinite mass so that they cannot be normalized. Although posterior distributions can often be calculated from such improper prior distributions by plugging into a formula, the formula is not well justified in the usual probabilistic models of uncertainty. Handling and interpreting improper prior distributions is a significant problem in the Bayesian approach to statistics [1], and this will be the primary focus of the present paper.

check for updates

Academic Editor: Wei Zhu

Received: 30 August 2025
Revised: 27 September 2025
Accepted: 7 October 2025
Published: 9 October 2025

Citation: Harremoës, P. Improper Priors via Expectation Measures. *Stats* **2025**, *8*, 93. https://doi.org/10.3390/stats8040093

In many textbooks, improper prior distributions are handled by the selection of a "large" subset of the parameter space. If the parameter space is Θ and the improper prior measure is μ, then one selects a subset $\tilde{\Theta} \subseteq \Theta$, such that $\mu(\tilde{\Theta}) < \infty$. Then, the measure restricted to $\tilde{\Theta}$ is normalized, so that the normalized measure can be interpreted as a probability measure. If $\tilde{\Theta}_n$ is an increasing sequence of sets, such that $\Theta = \bigcup_{n=1}^{\infty} \tilde{\Theta}_n$, then the posterior based on the normalized version of the measure μ restricted to $\tilde{\Theta}_n$ will converge to the posterior based on μ. Hence, by selecting a sufficiently large subset $\tilde{\Theta}$ of the parameter space, we get a probabilistic inference that approximately gives the right result. Akaike and many others have advocated this approach to handling improper prior distributions [2]. See [3] for a more recent exposition regarding the approximation of improper priors by probability measures.

Inference based on restriction of the parameter space is problematic for two reasons. The first reason is that the subset $\tilde{\Theta}$ should, in principle, be chosen before any observation has been made, and if μ is improper and $\mu(\tilde{\Theta}) < \infty$, there will exist observations for which the posterior based on $\tilde{\Theta}$ is very different from the posterior based on the whole parameter space Θ. The second reason is that if $\tilde{\Theta}$ is chosen with a finite measure, it will often conflict with how we justify the use of the prior measure μ. If, for instance, μ is determined as a Haar measure on a non-compact group, then the restriction of μ to a set of finite measure will, in general, not be a Haar measure.

1.1. Expectation Theory

In a recent paper, expectation theory was presented as an alternative to the Kolmogorov style of probability theory [4]. Our main result is that with expectation measures at our disposal, we can handle improper prior distributions without restricting to a subset of the parameter space. No approximation argument is required as long as we condition on an event of positive finite measure. Approximation may be relevant if we condition on an event of measure 0, but this problem is related to using continuous measures and not to the prior distribution being improper.

In [4], it was shortly mentioned that expectation theory allows us to give a probabilistic interpretation for improper prior distributions and conditioning based on such measures. Here, we will provide a more detailed exposition on this problem. Some results in [4] will be generalized from discrete measures to s-finite measures.

The basic objects for describing uncertainty in expectation theory are s-finite measures rather than probability measures that are the fundamental objects in Kolmogorov-style probability theory. These measures can be interpreted as expectation measures of specific point processes. This gives a probabilistic interpretation of expectation theory, so there is no dichotomy between probability theory and expectation theory, but the focus is slightly different in expectation theory. Expectation theory and Kolmogorov-style probability theory are two theories that both quantify uncertainty, and each of the two theories comes with a set of basic concepts, as illustrated in Table 1.

Recently, M. Albert and S. Mellick have proved that if a group is locally compact, second-countable, unimodular, non-discrete, and non-compact, then any free-probability-measure-preserving action of the group can be realized by an invariant point process [5,6]. As we will see in this paper, the idea of interpreting measures that are not probability measures via point processes can be used for any s-finite measure without reference to a group structure. Since the methods and results of M. Albert and S. Mellick are so closely related to the present work, we will briefly mention how Haar measures are relevant for determining prior distributions in Section 2.7.

Table 1. Fundamental concepts in Kolmogov-style probability theory and the corresponding fundamental concepts in expectation theory.

Probability Theory	**Expectation Theory**
Probability	Expected value
Outcome	Instance
Sample space	Multiset monad
P-value	E-Value
Probability measure	Expectation measure
Binomial distribution	Poisson distribution
Density	Intensity
Bernoulli random variable	Count variable
Empirical distribution	Empirical measure
KL-divergence	Information divergence
Uniform distribution	Poisson point process

It is possible to define a monad for point processes [7]. The monad defined in [7] is also related to the observation that the Giry monad is distributive over the multiset monad, as discussed in [8]. These results from category theory provide the underlying structure that allows for the results presented in this paper.

1.2. Terminology and Notation

In principle, there is no dichotomy between expectation theory and Kolmogorov-style probability theory, but in practice, we have to make some modifications to the terminology in order to avoid confusion.

In standard probability theory, a probability measure lives in either the sample space [9] (p. 292), [10] (Section 1.3), [11] (p. 22), the outcome space [12] (p. 10), [13] (Sec. 1.1), the space of elementary events [14] (p. 5), the sample description space [15] (p. 7), or the possibility space [16] (p. 3). In this paper, we will use the term *outcome space*, and the elements of the outcome space will be called outcomes, points, or letters. The word *sample* will be used informally about the result of sampling. Sampling can often be modeled by a point process where the result is a multiset, i.e., a set of points in the outcome space each with a weight indicating the number of observations of that point. The result of a point process will be called *an instance* of the point process, and the elements of the instance will often be called *points*. We avoid the term sample space, because it may make it less clear whether a sample leads to a point in outcome space or whether it leads to a multiset over the outcome space.

If μ is a probability measure, then the conditional probability measure given a measurable set A is denoted as $\mu(\cdot \mid A)$. In the standard approach to probability theory, the conditional probability measure may be viewed as a restriction of the original probability measure to a subset. In many expositions, the measure μ restricted to a measurable set A is denoted as $\mu \mid_A$ [17] (Sec. 3) or $\mu_{\mid A}$ [18] (p. 3). In expectation theory, the restricted measure $\mu \mid_A$ and the conditional measure $\mu(\cdot \mid A)$ are two related but distinct measures, which should not be confused. For this reason, we will denote the restricted measure as $\mu_{\cap A}$.

A measure with a total mass of 1 is usually called a probability measure. We will deviate from this terminology and use the alternative term *normalized measure* for a measure with total mass 1 [12] (p. 10). We will reserve the word *probability measure* to situations where the weights of a normalized measure are used to quantify uncertainty, and it is known that precisely one observation will be made, and one can decide which event the observation belongs to in a system of mutually exclusive events that cover the whole outcome space. Similarly, we will talk about an *expectation measure* if our interpretation

of its values is given in terms of expected values of some random variables, or if it is the expectation measure of a point process.

If a measure is used to quantify our prior knowledge about a parameter before observation, we will call it a *prior distribution*. Following [19], we use the term *proper prior* when the measure is normalized, and in other cases, we say that the prior distribution is *improper*. Note that many statisticians only use the term improper prior when the measure has infinite total mass [20] (Chap. 8.2 Improper prior).

1.3. Organization of the Paper

In order to make this paper more self-contained, there is some slight overlap between this paper and [4], but the reader should consult [4] if the reader is interested in a more complete motivation for basing a theory of uncertainty on expectation measures rather than probability measures.

In Section 2, we provide a brief introduction to expectation theory and related topics concerning point processes. We also discuss statistical models and some methods for calculating prior distributions. There are many other ways to get prior distributions, and this is not an attempt to cover this topic. We just provide enough background material to present some examples of statistical models with prior distribution.

Section 3 contains the main contribution of this paper. We provide a probabilistic interpretation of improper priors based on point processes. The interpretation allows for the calculation of posterior distributions without relying on any approximation arguments.

We end the paper with a short discussion.

2. Methods

Here, we will introduce the concepts and results needed in the subsequent sections. For motivation and more details, we refer to the literature.

2.1. Observations and Expectations

In statistics, data are often given in terms of frequency tables. To each entry, the table gives the observed frequency of that entry. An example of such a frequency table is Table 2.

Table 2. Frequencies of eyes in 68 independent throws with a six-sided die.

Number of Eyes	Frequency
One eye	10
Two eyes	15
Three eyes	15
Four eyes	7
Five eyes	8
Six eyes	13

A frequency table can be identified with a multiset, i.e., a set where each point has a multiplicity. To relate such multisets to Kolmogorov-style probability theory, we will represent them as measures. Let $(\mathbb{B}, \mathcal{F})$ denote a measurable space. Observations in $(\mathbb{B}, \mathcal{F})$ will be represented as finite or countable sums of Dirac measures.

Example 1. *The frequencies in Table 2 can be represented by the measure*

$$10\delta_1 + 15\delta_2 + 15\delta_3 + 7\delta_4 + 8\delta_5 + 13\delta_6. \tag{1}$$

A finite or countable sum of Dirac measures will be called *an observation measure*. One can define kernels with expectation measures as outcomes, and such kernels can

be composed in the same way as Markov kernels can be composed. The category of finite expectation measures was studied in [4], and this category may serve as a model of descriptive statistics.

2.2. Expectations as s-Finite Measures

Before making any observations, there will be uncertainty about what the observations will be. The uncertainty will be quantified in terms of an *expectation measure*, which is a measure μ on an outcome space $(\mathbb{B}, \mathcal{F})$, such that for $B \in \mathcal{F}$ the value $\mu(B)$ is the expected value of the number of observations in B. If we would allow all measures as expectation measures, we would get into technical problems. For instance, Tonelli's theorem does not hold for arbitrary measures, and kernels based on arbitrary measures cannot be composed. For this reason, we should look for a well-behaved category that can handle both normalized measures and observation measures.

The set of normalized measures on $(\mathbb{B}, \mathcal{F})$ will be denoted $M_+^1(\mathbb{B}, \mathcal{F})$ or $M_+^1(\mathbb{B})$ for short. Like Rényi, we are more interested in kernels than in measures [21–23]. A measurable mapping $\mathbb{A} \to M_+^1(\mathbb{B},)$ is called a Markov kernel, and an important property of Markov kernels is that they can be composed. Let $a \to \mu_a$ and $b \to \nu_b$ denote Markov kernels from $\mathbb{A}$ to $\mathbb{B}$ and from $\mathbb{B}$ to $\mathbb{D}$, respectively. The two Markov kernels can be composed by

$$(\mu \odot \nu)_a(D) = \int_{\mathbb{B}} \nu_b(D) \, \mathrm{d}\mu_a b. \tag{2}$$

With this compostion, the measurable spaces and Markov kernels form a category that Lawvere was the first to study [24]. From the point of view of category theory, the composition is related to the fact that the functor M_+^1 is part of a monad [4,25].

A kernel $a \to \mu_a$ is said to be a sub-Markov kernel if $\|\mu_a\| \leq 1$ for all $a \in \mathbb{A}$ [26] (Def.1, ii'). Sub-Markov kernels can be composed in just the same way as Markov kernels. Thus, the measurable spaces and sub-Markov kernels form a category with the category of Markov kernels as a sub-category.

A kernel $\mu : X \to Y$ is said to be *s-finite* if there exists a countable set of sub-Markov kernels μ_i, such that $\mu_x = \sum_{i=1}^{\infty} \mu_i$. Such s-finite kernels can be composed, resulting in an s-finite kernel [27]. To see that, let $\nu_x = \sum_{j=1}^{\infty} \nu_{x,j}$ be a s-finite kernel from X to Y and let $\mu_y = \sum_{i=1}^{\infty} \mu_{y,i}$ be a s-finite kernel from X to Y. Then

$$\begin{aligned} \mu \odot \nu &= \left(\sum_{i=1}^{\infty} \mu_i \right) \odot \left(\sum_{j=1}^{\infty} \nu_j \right) \\ &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \mu_i \odot \nu_j, \end{aligned} \tag{3}$$

which is clearly an s-finite kernel. With this composition, we get a category of s-finite kernels, and the category of Markov kernels is a sub-category.

Many textbooks on probability theory or general measure theory focus on σ-finite measures. The problem is that the composition of σ-finite kernels may lead to s-finite kernels that are not σ-finite. In recent years, s-finite measures have gained increasing attention among people studying denotational semantics for probabilistic programming [23,27–29].

2.3. Point Processes

We will define a point process with points in the measurable space $(\mathbb{B}, \mathcal{G})$. In the literature on point processes, $\mathbb{B}$ will be a d-dimensional Euclidean space, but we will not make such a restriction. Let $(\Omega, \mathcal{F}, P)$ denote a probability space. A transition kernel $\omega \to \mu_\omega$ from $(\Omega, \mathcal{F})$ to $M_+(\mathbb{B}, \mathcal{G})$ is called a *point process* if

- For all $\omega \in \Omega$, the function $\mu_\omega(\cdot) : \mathcal{G} \to \mathbb{R}_{0,+}$ is a s-finite measure.
- For all bounded sets $B \in \mathcal{G}$, the random variable $\omega \to \mu_\omega(B) : \Omega \to \mathbb{R}_{0,+}$ is a count variable.

In the literature, it is often assumed that that $\mu_\omega(\cdot) : \mathcal{G} \to \mathbb{R}_{0,+}$ is locally finite rather than s-finite, but we will make no such restriction. For further details about point processes, see [30] or [31] (Chapter 3).

The interpretation is that if the outcome is ω, then μ_ω is a measure that counts how many points there are in various subsets of $\mathbb{B}$, i.e., $\mu_\omega(B)$ is the number of points in the set $B \in \mathcal{G}$. Each measure μ_ω will be called *an instance* of the point process. We note that under weak topological conditions, an instance of a point process is the same as an empirical measure. In the literature on point processes, one is often interested in *simple point processes*, where $\mu_\omega(B) = 0$ when B is a singleton. However, point processes that are not simple are also crucial for the problems that will be discussed in this paper.

The definition of a point process follows the general structure of probability theory, where everything is based on a single underlying probability space. This will ensure consistency, but often this probability space has to be quite large if several point processes or many random variables are considered simultaneously.

The measure μ is called the *expectation measure* of the process $\omega \to \mu_\omega$ if for any $B \in \mathcal{S}$ we have

$$\mu(B) = \int_\Omega \mu_\omega(B) \, \mathrm{d}P\omega. \tag{4}$$

The term intensity measure is sometimes used instead of expectation measure. For simple point processes, expectation measures are often expressed in terms of the Radon–Nikodym derivative with respect to an underlying measure. In such cases, the term intensity measure is appropriate. However, we also consider point processes that are not simple, so we prefer the term expectation measure.

The expectation measure gives the mean value of the number of points in the set B. Different point processes may have the same expectation measure. *A one-point process* is a process that outputs precisely one point with probability 1. For a one-point process, the expectation measure of the process is simply a probability measure on $\mathbb{B}$. Thus, probability measures can be identified with one-point processes.

2.4. Poisson Distributions and Poisson Point Processes

For $\lambda \in [0, \infty)$, the Poisson distribution $Po(\lambda)$ is the probability distribution on $\mathbb{N}_0$ with point probabilities:

$$Po(j, \lambda) = \frac{\lambda^j}{j!} \exp(-\lambda). \tag{5}$$

For $\lambda = \infty$, we define $Po(\infty)$ as the normalized measure concentrated on ∞.

It was proven in [32] (Thm. 3.6) that for any s-finite measure on $\mathbb{B}$, there exists a point process $\omega \to \mu_\omega$, such that

- For all $B \in \mathcal{S}$, the random variable $\omega \to \mu_\omega(B)$ is Poisson distributed with a mean value $\mu(B)$.
- If B_1 and $B_2 \in \mathcal{S}$ are disjoint, then the random variables $\omega \to \mu_\omega(B_1)$ and $\omega \to \mu_\omega(B_2)$ are independent.

Such a process is called a *Poisson point process* with expectation measure μ, and we will denote it by $Po(\mu)$. All results regarding an s-finite measure μ can now be translated into results regarding the Poisson process $Po(\mu)$. We call this the *Poisson interpretation* of the measure.

Example 2 (Temporal Poisson process)**.** *Let $m_{\cap\mathbb{R}_+}$ denote the Lebesgue measure restricted to the interval $[0,\infty]$. Then, $Po\left(m_{\cap\mathbb{R}_+}\right)$ is a homogeneous Poisson process with intensity 1. This is normally considered a temporal model, where the elements in $\mathbb{R}_+$ are considered as times where certain events happen.*

Example 3 (Spatio-temporal Poisson process)**.** *If $Po(\mu)$ is a Poisson point process with points in space, then $Po\left(\mu \times m_{\cap[0,1]}\right)$ can be viewed as a spatio-temporal point process, where any points of the spatial process are created at a random time in $[0,1]$. This process has the process $Po(\mu)$ as its marginal distribution.*

Formally, one may consider the spatio-temporal Poisson process $Po\left(\mu \times m_{\cap[0,\infty]}\right)$ where points continue to be created. An instance of such a process would have infinitely many points, so it cannot be simulated. A simulation of a spatio-temporal process can be found in Figure 1.

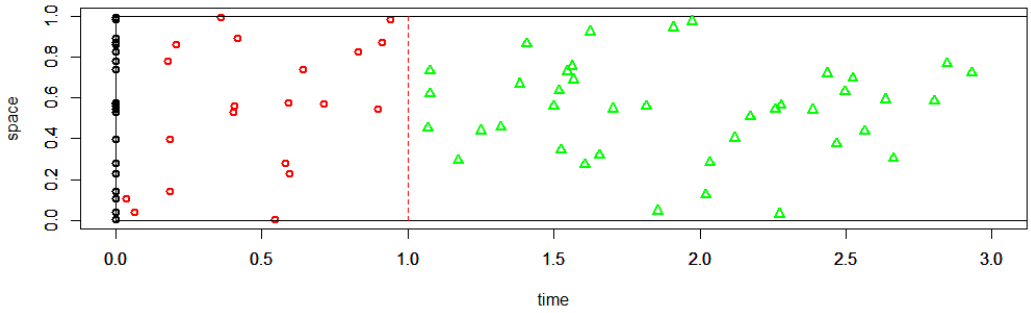

Figure 1. Simulation for some point processes. The black points are points in an instance of a Poisson point process on $[0,1]$ distributed according to a uniform distribution. The red point are point in an instance of a spatio-temporal process, where one has assigned a random time in $[0,1]$ to each of the black points. The green triangles illustrate a continuation of the same process from time 1 to time 3. One cannot simulate the whole spatio-temporal with time in $[0,\infty]$ because the expected number of points is infinite.

2.5. Measures and Kernels Associated with Statistical Models

Let $(\mathbb{B},\mathcal{G})$ be a measurable space that represents the possible outcomes. Further, let $(\Theta,\mathcal{F})$ be a measurable space that represents possible values of a parameter of a statistical model. A *statistical model* is given by a Markov kernel $\theta \to P_\theta$ that assigns a probability measure P_θ on $(\mathbb{B},\mathcal{G})$ to each parameter $\theta \in \Theta$ [33]. The goal of the statistician is to make an inference on the unobserved value of θ based on an observed value $b \in \mathbb{B}$.

Assume that our prior knowledge about the parameter θ is given by the measure μ on $(\Theta,\mathcal{F})$. This leads to a joint measure on $(\Theta \times \mathbb{B}, \sigma(\mathcal{F} \times \mathcal{G}))$. Following [26], the joint measure will denote $\mu \times P_\theta$. For $A \in \mathcal{F}$ and $B \in \mathcal{G}$, we have a measurable function from $A \to \mathbb{R}$, which is given by $\theta \to P_\theta(B)$. The joint measure $\mu \times P_\theta$ is defined by

$$(\mu \times P_\theta)(A \times B) = \int_A P_\theta(B)\, \mathrm{d}\mu\theta. \tag{6}$$

Let ν denote the marginal measure of $\mu \times P_\theta$ on $\mathbb{B}$, i.e., ν is the restriction of $\mu \times P_\theta$ to the sub-algebra of $\sigma(\mathcal{F} \times \mathcal{G})$ consisting of sets of the form $\{\theta\} \times B$. If ν is a σ-finite measure, then there exists a Markov kernel Q_b from $\mathbb{B}$ to Θ, such that

$$(\mu \times P_\theta)(A \times B) = \int_B Q_b(A)\, \mathrm{d}\nu b, \tag{7}$$

and we will write $\mu \times P_\theta = Q_a \times \nu$ for short. Remark that, at this level, the existence of the Markov kernel $a \to Q_a$ is a purely formal construction.

In information theory, a Markov kernel $(P_a)_{a\in\mathbb{A}}$ is called an *information channel* with *input alphabet* $\mathbb{A}$ and *output alphabet* $\mathbb{B}$ [34] (Chapter 8). In the branch of information theory called channel coding, the input letters are controlled by the sender (Alice), but unknown to the receiver (Bob). The goal of Bob is to make an inference about the letter $a \in \mathbb{A}$ sent by Alice based on the letter $b \in \mathbb{B}$ received by Bob.

A Markov kernel can be used to model sequences of observations in $\mathbb{B}$ in two ways. In statistics, a sequence of length n is modeled by $(\bigotimes_{i=1}^n P_\theta)_{\theta\in\Theta}$, which gives a Markov kernel from Θ to $\mathbb{B}^n$. In channel coding, a sequence of length n is modeled by $(\bigotimes_{i=1}^n P_{a_i})_{a_1^n\in\Theta^n}$. In channel coding, we get a Markov kernel from $\mathbb{A}^n$ to $\mathbb{B}^n$.

2.6. Minimax Redundancy and Jeffreys' Prior

Prior distributions play a major role in Bayesian statistics. We will provide examples of how prior distributions are calculated using ideas from information theory. The information-theoretic approach to calculating prior distributions will also lead to a motivation using Jeffreys' prior. The examples we discuss in this section will be used in subsequent sections. A detailed discussion about different methods for the calculation of prior distributions is beyond the topic of this article. We will refer to [35] for a review of the subject, including a long list of references.

One method for calculating a prior distribution for a statistical model $\theta \to P_\theta$ is to consider the model as an information channel. Here, we will only mention some of the basic ideas briefly. The reader may consult [36] or [37] for a more detailed exposition. Assume for simplicity that $\mathbb{B}$ is a finite set. If data is distributed according to P_θ, then the code that will give the shortest mean code-length uses a code-word of length proportional to $-\log(P_\theta(b))$ when the letter b is encoded. Optimal coding requires that θ is known. If θ is not known and the data is coded as if the distribution was given by the probability measure P, then the code-length of the code-word corresponding to the letter b is $-\log(P(b))$. If the letter is b, the redundancy of coding as if the distribution was given by P when it is actually P_θ is defined as the difference in code-length., i.e.,

$$\log\left(\frac{P_\theta(b)}{P(b)}\right). \tag{8}$$

The mean value of the redundancy (8) is given by the Kullback–Leibler divergence is defined by

$$D(P_\theta\|P) = \int \ln\left(\frac{\mathrm{d}P_\theta}{\mathrm{d}P}\right)\mathrm{d}P_\theta. \tag{9}$$

The Kullback–Leibler divergence quantifies *redundancy*, i.e., the mean number of bits one can save by coding according to the true distribution P_θ rather than coding as if the data were distributed according to P. The minimax redundancy is given by

$$\min_P \max_\theta D(P_\theta\|P) \tag{10}$$

where the minimum in Equation (10) takes over all probability measures P on $\mathbb{B}$. Coding according to the distribution P that minimizes the maximal redundancy is optimal, in the sense that it leads to the shortest description of data compared with what could have been achieved knowing the true distribution P_θ.

The capacity of the channel $\theta \to P_\theta$ is the maximal transmission rate, which is the maximal mutual information between input and output [34] (Chap. 8). According to the Gallager–Ryabko Theorem [38], the maximal transmission rate equals the minimax

redundancy. If P^* is the distribution that achieves the minimum in Equation (10), then a capacity-achieving input distribution is the same as a probability measure Q, such that

$$P^* = \int_\Theta P_\theta \, \mathrm{d}Q\theta. \tag{11}$$

The input distribution Q is the optimal prior distribution if we want to minimize the maximal redundancy.

Example 4 (The binary erasure channel)**.** *The binary erasure channel has an input alphabet* $\mathbb{A} = \{a, b\}$ *and an output alphabet* $\mathbb{B} = \{a, b, e\}$*. A Markov kernel* $x \to P_x$ *is given by*

$$\begin{aligned} P_a(a) &= \alpha, \\ P_a(b) &= 0, \\ P_a(e) &= 1 - \alpha, \\ P_b(a) &= 0, \\ P_b(b) &= \alpha, \\ P_b(e) &= 1 - \alpha. \end{aligned} \tag{12}$$

The output letter e represents an erasure of the input letter. The capacity achieving input distribution is the uniform distribution on the input alphabet $\mathbb{A}$*. See [34] (Subsec. 8.1.5) for a detailed discussion of the binary erasure channel.*

Example 5 (The binomial model)**.** *The binomial distributions* $p \to b(n, p)$ *form a statistical model with point probabilities* $\binom{n}{x} p^x (1-p)^{n-x}$*. In this case, there is no unique capacity-achieving distribution if the parameter space is* $\Theta = [0, 1]$*. If we restrict the parameter space to the set of possible maximum likelihood estimates* $\{0, 1/n, 2/n, \ldots, 1\}$*, there is a unique capacity-achieving distribution that can be used as a prior distribution on* Θ*. For small values of n, the exact optimal distribution can be calculated. If, for instance* $n = 2$*, the optimal distribution on* $\{0, 1/2, 1\}$ *is* $\{8/17, 1/17, 8/17\}$*. In general, no closed formula for the capacity-achieving distribution exists, but it can be approximated using an iterative algorithm (see [36] (Sec. 5.2) and [39]).*

Kullback–Leibler divergence given by Equation (9) equals the Rényi divergence of order 1. If we use the Rényi divergence of order ∞ [40] (Thm. 6)

$$D_\infty(P_\theta \| P) = \ln \sup_B \frac{P_\theta(B)}{P(B)} \tag{13}$$

instead of Kullback–Leibler divergence, then we get the *regret*, which reveals how many bits can be saved by coding with respect to P rather than coding according the model Q for the data that is least favorable without any assumption on how the data sequence is generated. From a statistical perspective, an analysis based on regret rather than redundancy is more conservative.

Example 6 (The binomial model). *The distribution that achieves minimax regret can be calculated as the normalized maximum likelihood (NML) distribution. It has point probabilities*

$$\begin{aligned} P_{NML}(X=0) &= \frac{P_0(X=0)}{P_0(X=0)+P_{1/2}(X=1)+P_1(X=2)} = \frac{4}{9}, \\ P_{NML}(X=1) &= \frac{P_{1/2}(X=1)}{P_0(X=0)+P_{1/2}(X=1)+P_1(X=2)} = \frac{1}{9}, \\ P_{NML}(X=2) &= \frac{P_1(X=2)}{P_0(X=0)+P_{1/2}(X=1)+P_1(X=2)} = \frac{4}{9}. \end{aligned} \tag{14}$$

This corresponds to the prior $(3/10, 4/10, 3/10)$ *on the parameters* $\{0, 1/2, 1\} \subseteq [0,1]$.

As demonstrated in Examples 5 and 6, finding a prior using minimax redundancy or minimax regret will, in general, lead to different results, but for long data sequences, the distributions that achieve minimax redundancy and minimax regret, respectively, can both be approximated by *Jeffreys' prior* [37] (Sec. 8.2). Thus, Jeffreys' prior can be used as an approximation of the prior that is optimal in the sense of achieving minimax redundancy or minimax regret.

Let $(P_\theta)_{\theta\in\Theta}$ denote a statistical model and assume that $\frac{dP_\theta}{dP_0}(x) = f(x,\theta)$ for some dominating measure P_0. Assume further that Θ is an open subset of $\mathbb{R}^d$, and that $\theta \to f(x,\theta)$ is twice differentiable. Note that this excludes statistical models where Θ is a discrete set. The Fisher information matrix is given by

$$[I(\theta)]_{i,j} = -E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln(f(X;\theta))|\theta\right]. \tag{15}$$

Jeffreys' prior is defined as the distribution on Θ with density

$$(\det(I(\theta)))^{1/2}. \tag{16}$$

One should note that there are other reasons for choosing Jeffreys' prior than the ones based on information theory. For instance, except for a constant factor, Jeffreys' prior does not depend on parametrization [41,42].

Example 7 (The binomial model). *For the binomial model, we have*

$$-\frac{\mathrm{d}^2}{\mathrm{d}p^2}\ln\left(\binom{n}{x}p^x(1-p)^{n-x}\right) = \frac{x}{p^2} + \frac{n-x}{(1-p)^2}. \tag{17}$$

The Fisher information equals the mean value of (17):

$$I(p) = \frac{np}{p^2} + \frac{n-np}{(1-p)^2} = \frac{n}{p(1-p)}. \tag{18}$$

Jeffreys' prior has density proportional to

$$\frac{1}{(p(1-p))^{1/2}}. \tag{19}$$

In this case, Jeffreys' prior has finite mass so that it can be normalized. The normalized Jeffreys' prior is a beta distribution with parameters $(1/2, 1/2)$. *The posterior distribution of* p, *if* x *successes and* $n-x$ *failures have been observed, is a beta distribution with parameters* $(x+1/2, n-x+1/2)$.

Example 8 (The exponential model). *For $\lambda > 0$, the exponential distribution $Expo(\lambda)$ has density:*

$$\frac{\exp\left(-\frac{x}{\lambda}\right)}{\lambda},\ x > 0. \tag{20}$$

We have

$$\begin{aligned} -\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} \ln\left(\frac{\exp\left(-\frac{x}{\lambda}\right)}{\lambda}\right) &= -\frac{2}{\lambda^2}\frac{\exp\left(-\frac{x}{\lambda}\right)}{\lambda} \\ &= \frac{2x}{\lambda^3} - \frac{1}{\lambda^2}. \end{aligned} \tag{21}$$

Hence, the Fisher information is given by

$$I(\lambda) = \lambda^{-2}, \tag{22}$$

and Jeffreys' prior has density λ^{-1}. In this case, Jeffreys' prior is improper, and it cannot be normalized. This is related to the fact that the statistical model has infinite channel capacity. Jeffreys' prior is also optimal in an information-theoretic sense without relying on any approximation argument involving long sequences [43,44].

With this prior measure, the joint measure has density $\frac{\exp(-x/\lambda)}{\lambda^2}$, $x, \lambda > 0$. The marginal measure of X is

$$\int_0^\infty \frac{\exp(-x/\lambda)}{\lambda^2}\,\mathrm{d}\lambda = \frac{1}{x}. \tag{23}$$

The conditional distribution of the parameter Λ, given $X = x$, is an inverse gamma distribution *with density $\frac{x\exp(-x/\lambda)}{\lambda^2}$, shape parameter 1, and scale parameter x.*

Example 9 (The Poisson model). *For the Poisson model $X \sim Po(\lambda)$, $\lambda \in [0, \infty]$ with $P(j) = \frac{\lambda^j}{j!}\exp(-\lambda)$, we have*

$$-\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} \ln\left(\frac{\lambda^j}{j!}\exp(-\lambda)\right) = \frac{j}{\lambda^2}. \tag{24}$$

Therefore, the Fisher information equals

$$I(\lambda) = \frac{1}{\lambda}. \tag{25}$$

Therefore, Jeffreys' prior has density $\lambda^{-1/2}$ on $[0, \infty]$, which cannot be normalized.

The marginal measure on X is $\frac{\Gamma(j+1/2)}{j!}$. The conditional distribution of the parameter as a random variable Λ given $X = j$ has the following density:

$$\frac{\lambda^{j-1/2}}{\Gamma(j+1/2)}\exp(-\lambda) \tag{26}$$

where the parameter Λ is gamma distributed with scale parameter 1 and shape parameter $j + 1/2$.

2.7. Haar Measures

Many statistical models have symmetries, and these can be useful in determining prior distributions. Let $(P_\theta)_{\theta\in\Theta}$ denote a statistical model with outcome space $\mathbb{B}$. Let G be a group that acts on both Θ and $\mathbb{B}$ via $\Phi_g : \Theta \to \theta$ and $\Psi_g : \mathbb{B} \to \mathbb{B}$. The group action is said to be *covariant* if

$$\Psi_g(P_\theta) = P_{\Phi_g(\theta)}. \tag{27}$$

The notion of covariance was introduced by A. Holevo in the context of quantum information theory [45]. Group actions on statistical models have also been discussed in the

statistical literature (see [33] (p. 1241)) and references in that paper), but the idea is less used and developed in statistics than in quantum information theory. Equation (27) can be expressed in terms of the following commutative diagram

$$\begin{array}{ccc} \Theta & \xrightarrow{\Phi_g} & \Theta \\ {\scriptstyle P}\downarrow & & {\scriptstyle P}\downarrow \\ M_+(\mathbb{B}) & \xrightarrow{M_+(\Psi_g)} & M_+(\mathbb{B}) \end{array} \tag{28}$$

If a group has a covariant action on a statistical model, then, one may argue, the prior should be invariant under the action of the group.

Theorem 1 (Existence of Haar measures [46,47])**.** *Let $(G, \cdot)$ denote a locally compact group. Then, there exists a measure μ that is invariant under left actions, i.e., for any measurable set $A \subseteq G$ and any $g \in G$ we have $\mu(g \cdot A) = \mu(A)$. The measure μ is unique except for a multiplicative constant.*

A left invariant measure is called a *left Haar measure*. The left Haar measure is finite if, and only if, the group is compact. A locally compact group also has a *right Haar measure* that may be different from the left Haar measures, but if the group acts on a set X from the left, we are mainly interested in the left Haar measures. On abelian groups, discrete groups, and compact groups, all left Haar measures are also right Haar measures. For such groups, we do not need to distinguish between left Haar measures and right Haar measures and just talk about Haar measures [48].

If a group has a left action on the parameter space, and the action is transitive, then the action induces a measure on the parameter space, which is invariant under actions of the group. This measure will be the uniquely determined left invariant measure, except for a multiplicative constant.

Example 10 (Binary erasure channel)**.** *For the binary erasure channel, there is a symmetry between the letters a and b, and this symmetry holds both for the input alphabet $\mathbb{A} = \{a, b\}$ and for the output alphabet $\mathbb{B} = \{a, b, e\}$. Measures that put equal weight on a and b are the only measures on $\mathbb{A}$ that are invariant under the symmetry. The symmetry does not depend on whether we use minimax redundancy or minimax regret as a criterion for selecting the prior, so these and many other criteria for selecting a prior all lead to the same prior except perhaps for a multiplicative constant.*

If the outcome space is discrete and the parameter space is continuous, then a covariant action of a symmetry group cannot be transitive on the parameter space.

Example 11 (The Binomial model)**.** *In the binomial model, there is a symmetry between success and failure corresponding to the mapping $p \to 1 - p$ in the parameter space. The prior distributions in Examples 5–7 are all symmetric, but the action of the symmetry group is not transitive, so symmetry alone does not determine the prior.*

Example 12 (The exponential model)**.** *For the exponential model $\lambda \to Expo(\lambda)$, the group of positive numbers with multiplication $(\mathbb{R}_+, \cdot)$ has a covariant action on the statistical model via scaling $x \to s \cdot x$. A measure with density λ^{-1} with respect to the Lebesgue measure is a Haar measure on $(\mathbb{R}_+, \cdot)$. Therefore, Jeffreys' prior must be proportional to the Haar measure.*

If a group is locally compact and σ-compact, then any left Haar measure is s-finite, and there exists a Poisson point process with the Haar measure as the expectation measure. This

gives a probabilistic interpretation that will allow for a much wider use of Haar measures in probability theory.

3. Results

Many textbooks handle improper prior distributions by restricting the parameter space. In this section, we will utilize expectation theory to provide a more satisfactory approach to handling improper prior distributions.

3.1. Normalization and Conditioning for Expectation Measures

Empirical measures can be added, restrictions can be taken, and induced measures can be found. Using the same formulas, these operations can be performed on expectation measures, but we are not only interested in the formulas but also in probabilistic interpretations.

The norm of a (positive) measure ν is defined by $\|\nu\| = \nu(\mathbb{A})$, and the *normalized measure* $\nu/\|\nu\|$ has an interpretation as a probability measure, which is equivalent to a one-point process.

The following proposition gives a probabilistic interpretation of restriction for expectation measures via the same operations applied to empirical measures. A simple calculation proves the proposition.

Proposition 1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $\omega \to \mu_\omega$ denote a point process with expectation measure μ and with points in $\mathbb{B}$. Let B be a subset of $\mathbb{B}$. Then*

$$\mu_{\cap B} = \int \mu_{\omega \cap B} \, \mathrm{d}P\omega. \tag{29}$$

Normalized measures are usually called probability measures, and the next theorem gives a probabilistic interpretation of the normalized measure $\mu/\|\mu\|$ by specifying an event that has probability equal to $\mu/\|\mu\|$.

Theorem 2 ([4] (Thm. 10))**.** *Let B be a measurable subset of $\mathbb{B}$. Let μ be a non-trivial finite measure on $\mathbb{B}$. If P denotes a probability measure on Ω and $\omega \to \mu_\omega$ is a Poisson point process with expectation measure μ, then*

$$\frac{\mu(B)}{\|\mu\|} = \int_\Omega \frac{\mu_\omega(B)}{\|\mu_\omega\|} \, \mathrm{d}P(\omega | 0 < \|\mu_\omega\| < \infty). \tag{30}$$

Proposition 1 holds for all point processes, but in Theorem 2, it is required that the point process is a Poisson point process. An example of a point process where Equation (30) does not hold can be found in [4] (Ex. 5).

Theorem 2 states that $\mu(B)/\|\mu\|$ is the probability of observing a point in B, which has an interpretation that involves two steps.

1. Observe a multiset of points as an instance of a point process.
2. Select a random point from the observed multiset.

By replacing the point process $Po(\mu)$ by a spatio-temporal point process we can replace this two-step interpretation by a one-step interpretation. The one-step interpretation will be formulated as a theorem that has a much simpler proof than the proof of Theorem 2 given in [4], and the proof of the new theorem will not rely on the proof of Theorem 2.

Consider the point process $Po(\mu)$ on $\mathbb{B}$. From this process, we construct a spatio-temporal process. To each point in an instance of the point process $Po(\mu)$, we randomly select a number in $[0, 1]$ according to a uniform distribution. The number selected for a specific point is considered as the time at which the point is created. This gives the process

$Po\left(\mu \times m_{\cap[0,1]}\right)$. Instead of choosing a random point from the instance of the original point process $Po(\mu)$, we choose the first point in the spatio-temporal point process.

For the process $Po\left(\mu \times m_{\cap[0,1]}\right)$, there is a risk that no point is created before time $\alpha = 1$. To avoid this problem, we replace the process $Po\left(\mu \times m_{\cap[0,1]}\right)$ by the process $Po\left(\mu \times m_{\cap[0,\infty]}\right)$ with points in $\mathbb{B} \times [0,\infty]$. Let T be the time at which the first point is created. Then, T is a stopping time. The distribution of the point created at time T will be $\mu/\|\mu\|$.

We can summarize this result in the following theorem:

Theorem 3. *Let B be a measurable subset of $\mathbb{B}$. Let μ be a non-trivial finite measure on $\mathbb{B}$. Let P denote a probability measure on Ω and let $\omega \to \nu_\omega$ be a spatio-temporal Poisson process with expectation measure $\mu \times m_{\cap\mathbb{R}_+}$ on $\mathbb{B} \times \mathbb{R}_+$. For an instance ν_ω of the process, let (b_ω, t_ω) denote the point (b,t) in the instance, for which t has the smallest value. Then*

$$\frac{\mu(B)}{\|\mu\|} = P(b_\omega \in B). \tag{31}$$

Proof. Let S denote the waiting time until the first point in B has been observed, and let T denote the waiting time until the first point in $\complement B$ has been observed. The S has an exponential distribution with mean $\mu(B)^{-1}$, and T has an exponential distribution with mean $\mu(\complement B)^{-1}$. We have

$$\begin{aligned} P(b_\omega \in B) &= P(S < T) \\ &= \int_0^\infty \left(\int_s^\infty \exp(-t\mu(\complement B))\mu(\complement B)\,\mathrm{d}t \right) \exp(-s\mu(B))\mu(B)\,\mathrm{d}s \\ &= \int_0^\infty \exp(-s\mu(\complement B))\exp(-s\mu(B))\mu(B)\,\mathrm{d}s \\ &= \mu(B)\int_0^\infty \exp(-s(\mu(B)+\mu(\complement B)))\,\mathrm{d}s \\ &= \frac{\mu(B)}{\mu(B)+\mu(\complement B)}, \end{aligned} \tag{32}$$

which proves the theorem because $\mu(B) + \mu(\complement B) = \|\mu\|$. □

A simulation illustrating the theorem is given in Figure 2.

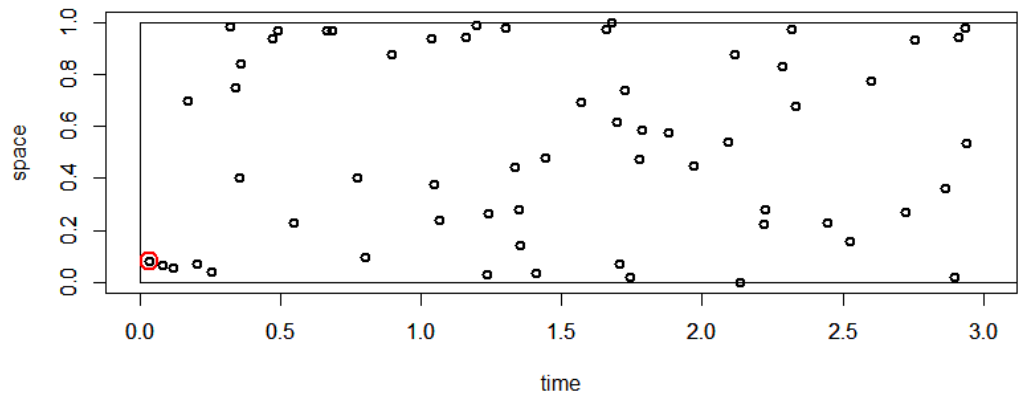

Figure 2. Simulation of a spatio-temporal based on a measure proportional to the beta distribution with $\alpha = 1/2$ and $\beta = 1/2$. A red circle marks the first point in the instance. The process is stopped at time equal to 3, but it could have been stopped at any time after the first point has been observed.

3.2. Conditioning for Improrer Prior Measures

Here, we shall look at how the results of Section 3.1 will allow us to give an exact interpretation of conditional probabilities with respect to an *improper prior distribution*. First, we note that the Poisson interpretation of normalized expectation measures carries over to conditional measures.

Theorem 4. *Let B be a measurable subset of $\mathbb{B}$. Let μ be an s-finite measure on $\mathbb{B}$. Let P denote a probability measure on Ω and let $\omega \to \nu_\omega$ be a spatio-temporal Poisson process with expectation measure $\mu \times m_{\cap \mathbb{R}_+}$ on $\mathbb{B} \times \mathbb{R}_+$. Assume that A is a measurable subset of $\mathbb{B}$ such that $0 < \mu(A) < \infty$. For an instance ν_ω of the process let (b_ω, t_ω) denote the point $(b, t) \in A$ in the instance for which t has the smallest value. Then*

$$\mu(B|A) = P(b_\omega \in B). \tag{33}$$

Proof. A conditional measure is the normalization of an expectation measure restricted to a subset.

$$\mu(B|A) = \frac{\mu(B \cap A)}{\mu(A)} = \frac{\mu_{\cap A}(B)}{\|\mu_{\cap A}\|}. \tag{34}$$

The corollary is proved by applying Theorem 2 to the measure $\mu_{\cap A}$. □

Theorem 4 is illustrated in Figure 3.

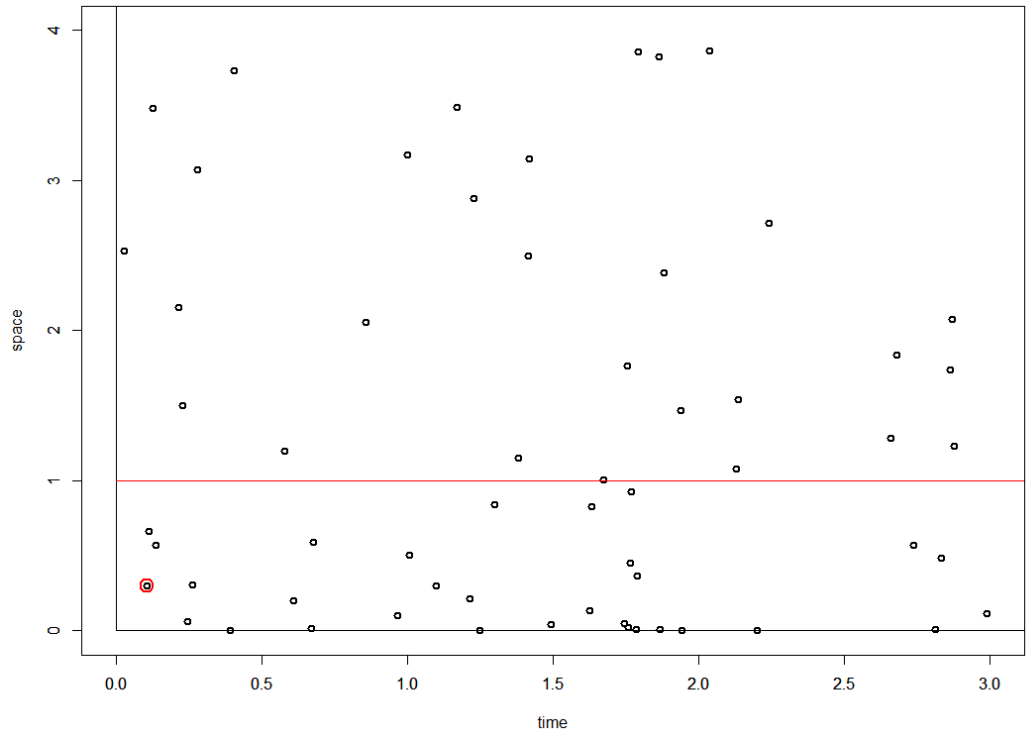

Figure 3. Simulation of a spatio-temporal process based on an un-normalized measure with density proportional to $x^{-1/2}$. We condition on $x \leq$ (indicated by the red line), which has finite measure. A red circle marks the first point in the instance below the red line. The process was stopped at time equal to 3, but it could have been stopped at any time after the marked point had been observed. Note that if we do not restrict to $x \leq 1$, then the marked point is not the first point. Only values of $x \leq 4$ were included in this simulation, but all points with $x > 1$ are irrelevant, for which point we should mark as the first point satisfying $x \leq 1$.

With this result at hand, we get an interpretation of posterior distributions calculated based on improper prior distributions.

Example 13 (The binary erasure channel)**.** *Consider the binary erasure channel discussed in Example 4. The prior measure μ gives the expected number of input letters from the alphabet $\mathbb{A} = \{a, b\}$. We run a spatio-temporal Poisson process on $\mathbb{A}$. This will give a stream of input letters at a rate of $\|\mu\|$ per time unit. Using the Markov kernel $x \to P_x$, we get a spatio-temporal process on $\mathbb{A} \times \mathbb{B}$.*

For any instance of this process, we look at the first output letter that equals e. For this first instance, we look at the corresponding input letter. The probability of the input letter a is $1/2$, and, similarly, the probability of the input letter b is $1/2$. Thus, the conditional probability distribution over input letters, given the output letter e, equals the probability that an instance with output letter e has a certain input letter.

Example 14 (The binomial model)**.** *In this example, the parameter space is the $[0, 1]$. If we fix the number n of output letters generated by a single value of the parameter and calculate the prior distribution that maximizes the transmission rate or, equivalently, minimizes the maximal redundancy, then the prior is concentrated on a finite subset of the parameter space. The prior will have a finite total mass, and it can be normalized to a probability measure. If the measure is not normalized, we will get a probabilistic interpretation by running a spatio-temporal process in exactly the same way as in the previous example.*

If we use Jeffreys' prior, which is a good approximation to the case where n is large, then it is still possible to normalize the prior measure. Normalizing the measure corresponds to selecting the first point in a point process. The posterior distribution of the parameter given the output letters equals the distribution of the parameter, given that the first point (input value of the parameter) in the spatio-temporal process leads to these output letters.

Example 15 (The Poisson model)**.** *For this model, Jeffreys' prior cannot be normalized. An instance of the point process with Jeffreys' prior as expectation measure has infinitely many points with probability 1. The same is true for the joint distribution of Λ and X. Therefore, the corresponding spatio-temporal process has no first point. This may appear as a problem, but if the joint distribution is restricted to $X = j$ then the measure is finite and an instance of the corresponding spatio-temporal process will have a first point. The distribution of this first point will be the conditional distribution of Λ given $X = j$, i.e., a gamma distribution with scale parameter 1 and shape parameter $j + 1/2$.*

In Example 15 one may object that it is not realistic to observe an instance of a point process with infinitely many points. For instance, a computer simulation will never be able to output infinitely many points. Although it is not possible to observe infinitely many points, this is irrelevant for our result because we are only interested in what happens under the condition $X = j$. What happens outside this event is irrelevant.

Example 16 (The exponential model)**.** *It is not possible to normalize Jeffreys' prior for the family of exponential distributions. Therefore, one cannot run the corresponding spatio-temporal process and take the first point because in any small time interval, there will be infinitely many points. If, instead, we have a certain interval for the output variable with finite mass, then we can take the first point in the process that lies in this interval. The conditional distribution of the parameter is a mixture of conditional distributions given the numbers in the interval weighted and normalized according to density $\frac{1}{x}$ on the interval.*

If the interval is short, then the conditional distribution given any point in the interval will be approximately constant, and conditioning on the interval will be approximately the same as conditioning on a point.

In the exponential model, one has to use some approximation argument if one has to condition with respect to the random variable having an exact value rather than being an element of an interval. This problem has nothing to do with the prior being proper or not. We will run into this problem for any continuous model, even if the parameter space is a finite set.

4. Discussion

We have applied expectation theory to give a probabilistic interpretation of improper prior distributions via the Poisson interpretation. This led to a probabilistic interpretation of conditioning with respect to improper prior distributions. With a probabilistic interpretation of improper prior measures and conditioning in place, one should go through all the arguments in favor of using specific methods for calculating prior distributions. We have briefly discussed Haar measures and Jeffreys' prior, but a careful review of all the methods is needed, which is beyond the scope of this paper.

In this paper, a statistical model was identified with a Markov kernel, as is usually done in statistics. From the point of view of expectation theory, it would be more natural to identify statistical models as s-finite kernels rather than Markov kernels. This would not make much of a difference regarding the handling of improper distributions with respect to conditioning. The idea of basing statistics on more general kernels than Markov kernels has also been promoted recently by Taraldsen et al. [49].

In [50,51], it was proven that for one-dimensional exponential families, minimax redundancy is finite if, and only if, minimax regret is finite. It was also demonstrated that a similar result does not hold for three-dimensional exponential families. There are still no results that relate the finiteness of minimax redundancy or minimax regret with the finiteness of Jeffreys' prior, and there are still a lot of open questions regarding improper prior distributions.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing does not apply to this article.

Acknowledgments: I want to thank Peter Grünwald and Tyron Lardy for stimulating discussions related to this topic.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Jones, A. Improper Priors. Available online: https://andrewcharlesjones.github.io/journal/improper-priors.html (accessed on 29 August 2025).
2. Akaike, H. The interpretation of improper prior distributions as limits of data dependent proper prior distributions. *J. R. Stat. Soc.* **1980**, *42*, 46–52. [CrossRef]
3. Bioche, C.; Druilhet, P. Approximation of improper priors. *Bernoulli* **2016**, *22*, 1709–1728. [CrossRef]
4. Harremoës, P. Probability via Expectation Measures. *Entropy* **2025**, *27*, 102. [CrossRef]
5. Mellick, S. Point Processes on Locally Compact Groups and Their Cost. Ph.D. Thesis, Alfréd Rényi Institute of Mathematics, Budapest, Hungary, 2019.
6. Abért, M.; Mellick, S. Point processes, cost, and the growth of rank in locally compact groups. *Isr. J. Math.* **2022**, *251*, 48–155. [CrossRef]
7. Dash, S.; Staton, S. A Monad for Probabilistic Point Processes. *arXiv* **2021**, arXiv:2101.10479. [CrossRef]
8. Jacobs, B. From Multisets over Distributions to Distributions over Multisets. In Proceedings of the 36th Annual ACM/IEEE Symposium on Logic in Computer Science, New York, NY, USA, 29 June–2 July 2021; LICS'21, pp. 1–13. [CrossRef]
9. Everitt, B.S. *The Cambridge Dictionary of Statistics*; Cambridge University Press: Cambridge, UK, 1998.
10. Whittle, P. *Probability via Expectation*, 3rd ed.; Springer texts in statistics; Springer: New York, NY, USA, 1992. [CrossRef]
11. Ross, S. *A first Course in Probability*, 8th ed.; Pearson Prentice Hall: Hoboken, NJ, USA, 2010.

12. Hogg, R.; Tannis, E.; Zimmerman, D. *Probability and Statistical Inference*; Pearson Education Inc.: Harlow, UK, 2013.
13. Adhikari, A.; Pitman, J. *Probability for Data Science*; Lecture notes; Berkeley, 2025. Available online: https://data140.org/textbook/content/README.html (accessed on 27 August 2025).
14. Shiryaev, A.N. *Probability*; Springer: New York, NY, USA, 1996.
15. Stark, H.; Woods, J.W. *Probability and Random Processes with Applications to Signal Processing*, 3rd ed.; Pearson: Upper Saddle River, NJ, USA, 2002.
16. Forbes, C.; Evans, M.; Hastings, N.; Peacock, B. *Statistical Distributions*, 4th ed.; Wiley: Hoboken, NJ, USA, 2011.
17. nLab Authors. Conditional Expectation. Revision 28. 2025. Available online: https://ncatlab.org/nlab/show/conditional+expectation/28 (accessed on 26 September 2025).
18. Velhinho, J. Topics of Measure Theory on Infinite Dimensional Spaces. *Mathematics* **2017**, *5*, 44. [CrossRef]
19. O'Hagan, A. *Kendall's Advanced Theory of Statistics*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2010; Volume 2B.
20. Wu, Q.; Vos, P. *Chapter 6—Inference and Prediction*; Elsevier: Amsterdam, The Netherlands, 2018. [CrossRef]
21. Rényi, A. On a new axiomatic theory of probability. *Acta Math. Acad. Sci. Hung.* **1955**, *6*, 185–335. [CrossRef]
22. Rényi, A. *Probability Theory*; North-Holland: Amsterdam, The Netherlands, 1970.
23. Vákár, M.; Ong, L. On s-Finite Measures and Kernels. Working Paper, Utricht University. 2018. Available online: http://arxiv.org/abs/arXiv:1810.01837 (accessed on 26 September 2025).
24. Lawvere. *The Category of Probabilistic Mappings*; Unpublished Lecture Notes; 1962. Available online: https://ncatlab.org/nlab/files/lawvereprobability1962.pdf (accessed on 10 October 2024).
25. Giry, M. A categorical approach to probability theory. In *Categorical Aspects of Topology and Analysis*; Banaschewski, B., Ed.; Springer: Berlin/Heidelberg, Germany, 1982; pp. 68–85.
26. Janssen, S. Markov Jump Processes. Lecture Notes, Mathematisches Institut der Universität München, 2020. Preparatory Notes for a Chapter on Spatial Birth and Death Processes in a Course on Point Processes and Gibbs Measures (LMU, Winter 2019/20). Available online: https://www.mathematik.uni-muenchen.de/~jansen/jump-processes.pdf (accessed on 26 September 2025).
27. Staton, S. Commutative Semantics for Probabilistic Programming. In *Programming Languages and Systems*; Yang, H., Ed.; Springer: Berlin/Heidelberg, Germany, 2017; pp. 855–879.
28. Affeldt, R.; Cohen, C.; Saito, A. Semantics of Probabilistic Programs using s-Finite Kernels in Coq. In Proceedings of the 12th ACM SIGPLAN International Conference on Certified Programs and Proofs, New York, NY, USA, 16–17 January 2023; CPP 2023, pp. 3–16. [CrossRef]
29. Hirata, M.; Minamide, Y. S-Finite Measure Monad on Quasi-Borel Spaces. *Arch. Form. Proofs* **2025**. Available online: https://www.isa-afp.org/entries/S_Finite_Measure_Monad.html (accessed on 24 September 2025).
30. Lieshout, M.V. Spatial Point Process Theory. In *Handbook of Spatial Statistics*; Handbooks of Modern Statistical Methods; Chapman and Hall/CRC: Boca Raton, FL, USA, 2010; Chapter 16.
31. Kallenberg, O. *Random Measures*; Springer: Cham, Switzerland, 2017. [CrossRef]
32. Last, G.; Penrose, M. *Lectures on the Poisson Process*; Cambridge University Press: Cambridge, UK, 2017.
33. McCullagh, P. What is a statistical model? *Ann. Stat.* **2002**, *30*, 1225–1310. [CrossRef]
34. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: Hoboken, NJ, USA, 1991.
35. Kass, R.E.; Wasserman, L.A. The Selection of Prior Distributions by Formal Rules. *J. Am. Stat. Assoc.* **1996**, *91*, 1343–1370. [CrossRef]
36. Csiszár, I.; Shields, P. *Information Theory and Statistics: A Tutorial*; Foundations and Trends in Communications and Information Theory; Now Publishers Inc.: Hanover, MA, USA, 2004.
37. Grünwald, P. *The Minimum Description Length Principle*; MIT Press: Cambridge, MA, USA, 2007.
38. Ryabko, B.Y. Comments on "A source matching approach to finding minimax codes". *IEEE Trans. Inform. Theory* **1981**, *27*, 780–781. [CrossRef]
39. Csiszar, I. Sanov Property, Generalized *I*-Projection and a Conditional Limit Theorem. *Ann. Probab.* **1984**, *12*, 768–793. [CrossRef]
40. van Erven, T.; Harremoës, P. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Trans Inform. Theory* **2014**, *60*, 3797–3820. [CrossRef]
41. Jordan, M.I. Jeffres Prior. Lecture Notes. Available online: https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/lectures/lecture6.pdf (accessed on 29 August 2025).
42. Ewing, B. What Is an Improper Prior? Available online: https://improperprior.com/pages/what-is-an-improper-prior/index.html (accessed on 29 August 2025).
43. Hedayati, F.; Bartlett, P. Exchangeability Characterizes Optimality of Sequential Normalized Maximum Likelihood and Bayesian Prediction with Jeffreys Prior. In Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, La Palma, Spain, 21–23 April 2012; Volume 22, pp. 504–510.

44. Bartlett, P.; Grünwald, P.; Harremoës, P.; Hedayati, F.; Kotłowski, W. Horizon-Independent Optimal Prediction with Log-Loss in Exponential Families. In Proceedings of the 26th Annual Conference on Learning Theory, Princeton, NJ, USA, 12–14 June 2013; Shalev-Shwartz, S., Steinwart, I., Eds.; Proceedings of Machine Learning Research; Volume 30, pp. 639–661. Available online: http://arxiv.org/abs/1305.4324 (accessed on 26 September 2025).
45. Holevo, A.S. *Probabilistic and Statistical Aspects of Quantum Theory*; North-Holland Series in Statistics and Probability; North-Holland: Amsterdam, The Netherlands, 1982; Volume 1.
46. Haar, A. Der Massbegriff in der Theorie der kontinuierlichen Gruppen. *Ann. Math.* **1933**, *34*, 147–169. [CrossRef]
47. Weil, A. *L'intégration Dans les Groupes Topologiques et ses Applications*; Herman: Paris, France, 1940; Volume 869.
48. Dowd, C.J. Notes on Haar Measures on Lie Groups. Lecture notes, Berkeley. 2023. Available online: https://math.berkeley.edu/~cjdowd/haar1.pdf (accessed on 26 September 2025).
49. Taraldsen, G.; Tufto, J.; Lindqvist, B.H. Improper prior and improper posterior. *Scand. J. Stat.* **2022**, *49*, 969–991. [CrossRef]
50. Grünwald, P.; Harremoës, P. Finiteness of Redundancy, Regret, Shtarkov Sums, and Jeffreys Integrals in Exponential Families. In Proceedings of the International Symposium for Information Theory, Seoul, Republic of Korea, 28 June–3 July 2009; IEEE: Seoul, Republic of Korea, 2009; pp. 714–718. [CrossRef]
51. Grünwald, P.; Harremoës, P. Regret and Jeffreys Integrals in Exp. Families. In Proceedings of the Thirtieth Symposium on Information Theory in the Benelux, Eindhoven, The Netherlands, 20–21 May 2009; Tjalkens, T., Willems, F., Eds.; Werkgemeenschap voor Informatie- en Communicatietheorie: Amsterdam, The Netherlands, 2009; p. 143.