Improper Priors via Expectation Measures

Peter Harremoës *

Posted Date: 2 September 2025

doi: 10.20944/preprints202509.0168.v1

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

 doi:10.20944/preprints202509.0168.v1

Article

Improper Priors via Expectation Measures

Peter Harremoës †

Niels Brock, Copenhagen Business College; harremoes@ieee.org
† Current address: Nørre Voldgade 34, Copenhagen, Denmark.

Abstract

In Bayesian statistics the prior distributions play a key role for 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 distribution is 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 they are inter-subjective 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 is described by a measure that has infinite mass so that it 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 modelling of uncertainty. Handling and interpreting improper prior distributions is a major in the Bayesian approach to statistics [1], and this will be the main focus of the present paper.

In a recent paper, expectation theory was presented as an alternative to the Kolmogorov style of probability theory [2]. The basic objects for describing uncertainty is s-finite measures rather than probability measures. 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. In [2], 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

 doi:10.20944/preprints202509.0168.v1

will provide a more detailed exposition on this problem. On the technical side we will generalize the results from [2] from discrete measures to s-finite measures.

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 [3,4]. Their result is closely related to the approach taken in this paper, but we will just briefly mention of how Haar measures are relevant for determining prior distributions. It is possible to define a monad for point processes [5]. The monad defined in [5] is also related to the observation that the Giry monad is distributive over the multiset monad as discussed in [6]. These results from category theory provide the underlying structure that allows the results presented in this paper.

In order to make this paper more self-contained there will be some slight overlap between this paper and [2], but the reader should consult [2] if the reader is interested in a more complete motivation for basing a theory of undertainty on expectation measures rather than probability measures.

1.1. Organization of the Paper

In Section 2, we provide a brief introduction to expectation theory and related topics concerning point processes. For a more detailed account, we refer the reader to [2], where the motivation for this approach is explained in detail.

In Section 3, we 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 4 contains the main contribution of this paper. We provide a probabilistic interpretation of improper priors based on point processes. The interpretation allows calculation of posterior distributions without relying on any approximation arguments.

We end the paper with a short discussion.

1.2. Terminology

A measure with a total mass of 1 is usually called a probability measure or a normalized measure. We will deviate from this terminology and use the term *unital measure* for a measure with total mass 1. The term *normalized measure* will only be used when a unital measure is the result of dividing a finite measure by the total mass of the measure. We will reserve the word *probability measure* to situations where the weights of a unital 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 are 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 [7] we use the term *proper prior* when the measure is unital, and in other cases we say that the prior distribution *improper*. Note that many statisticians only use the term improper prior when the measure has infinite total mass ([8], Chap. 8.2 Improper prior).

In standard probability theory, the probability measures lives on a space often called a sample space, but we will use the alternative term, *an outcome space*. The word *sample* will be used informally about the result of a sampling process. The result of a point process will be called *an instance* of the process and the elements of the instance will often be called *points*.

2. Preliminaries on Expectation Theory and Related Matters

Here, we will introduce the concepts and results needed in the subsequent sections. For motivation and more details, we refer to the literature. In the literature the restriction of a measure μ to a subset B is usually denoted $\mu_{|B}$, but we will use the notation $\mu_{\cap B}$ in order to avoid confusion with the notation for conditional probabilities.

 doi:10.20944/preprints202509.0168.v1

2.1. Observations and Expectations

Let $(\mathbb{B}, \mathcal{F})$ denote a measurable space. Observations are described by multisets, i.e. sets where each element has a multiplicity that is integer valued or infinte. In statistics such multisets are often given by frequency tables, but we will represent multisets by finite or countable sums of Dirac 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 the 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.

2.2. Subunital Measures and s-Finite Measures

The set of unital 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 than selves [9,10]. A measurable mapping $\mathbb{A} \to M_+^1(\mathbb{B},)$ is called a Markov kernel and a crusial property is that Markov kernels kan 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. Then the two Markov kernels can be composed by

$$(\mu \odot \nu)_a(D) = \int_{\mathbb{B}} \nu_b(D)\, \mathrm{d}\mu_a b. \tag{1}$$

From the point of view of category theory the composition is related to the fact that the functior M_+^1 is part of a monad [2,11]. A measure μ is said to be sub-unital if $\|\mu\| \leq 1$. The set of sub-unital measures will be denoted $M_+^{\leq 1}(X)$. Sub-unital kernels can be composed in just the same way as Markov kernels.

A kernel $\mu : X \to Y$ is said to be *s-finite* if there exists sub-unital kernels μ_i such that $\mu_x = \sum_{i=1}^{\infty} \mu_i$. Such s-finite kernels can be composed and the result is an s-finite kernel. 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

$$\begin{aligned} \mu \odot \nu &= \left(\sum_{i=1}^{\infty} \mu_j\right) \odot \left(\sum_{j=1}^{\infty} \nu_j\right) \\ &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \mu_i \odot \nu_j, \end{aligned} \tag{2}$$

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

2.3. Point Processes

We will define a point process with points in the measuralble space $(\mathbb{B}, \mathcal{G})$. Typically, $\mathbb{B}$ will be a d-dimensional Euclidean space, but we wil not make such 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 measure $\mu_\omega(\cdot) : \mathcal{G} \to \mathbb{R}_{0,+}$ is locally finite.
- 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.

For further details about point processes, see [12] or ([13], 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. 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.

 doi:10.20944/preprints202509.0168.v1

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

$$\mu(B) = \int_\Omega \mu_\omega(B)\, \mathrm{d}P\omega. \tag{3}$$

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

$$Po(j, \lambda) = \frac{\lambda^j}{j!} \exp(-\lambda). \tag{4}$$

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

It was proved in ([14], 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 mean value $\mu(B)$.
- If $B_1, B_2 \in \mathcal{S}$ 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 a measure μ can now be translated into results regarding the Poisson process $Po(\mu)$. This is called the Poisson interpretation of the measure.

Example 1 (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 things happen.*

Example 2 (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.*

Similarly, we may consider the spatio-temporal Poisson process $Po\left(\mu \times m_{\cap [0,\infty[}\right)$ where points continue to be created.

3. Statistical Models

In this section we will introduce statistical models, prior distributions and posterior distributions. We will provide some examples to be discussed later. Prior distributions play a major role in Bayesian statistics. A detailed discussion about how prior distributions can be determined in various cases is beyond the topic of this article. We will refer to [15] for a review of the subject including a long list of references.

3.1. 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 $\theta \in \Theta$. The goal of the statistician is to make inference on the unobserved value of θ based on an observed value $b \in \mathbb{B}$.

 doi:10.20944/preprints202509.0168.v1

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 $(\Theta \times \mathbb{B}, \sigma(\mathcal{F} \times \mathcal{G}))$ that we will denote $\mu \ltimes P_\theta$. For $A \in \mathcal{F}$ and $B \in \mathcal{G}$ the joint measure $\mu \ltimes P_\theta$ is defined by

$$(\mu \ltimes P_\theta)(A \times B) = \int_A P_\theta(B)\, \mathrm{d}\mu\theta. \tag{5}$$

Let ν denote the marginal measure of $\mu \ltimes P_\theta$ on $\mathbb{B}$, i.e. ν is the restriction of $\mu \ltimes P_\theta$ to the sub-algabra 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

$$(\mu \ltimes P_\theta)(A \times B) = \int_B Q_b(A)\, \mathrm{d}\nu b. \tag{6}$$

and we will write $\mu \ltimes P_\theta = Q_a \rtimes \nu$ 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* $\mathbb{B}$. 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 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 $\mathbb{B}^n$.

3.2. Minimax Redundancy and Jeffreys' Prior

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 [16] or [17] for a detailed exposition. The capacity of the channel is the maximal transmission rate, which is the maximal mutual information between input and output ([18], Chap. 8). According to the Gallager-Ryabko Theorem [19], the maximal transmission rate equals the minimax redundancy given by

$$\min_P \max_\theta D(P_\theta \| P) \tag{7}$$

where the Kullback-Liebler divergence is defined by

$$D(P_\theta \| P) = \int \ln\left(\frac{\mathrm{d}P_\theta}{\mathrm{d}P}\right) \mathrm{d}P_\theta, \tag{8}$$

and the minimum in Equation (7) is taken over all probability measures P on $\mathbb{B}$. Kullback-Leibler divergence quantifies *redundancy*, i.e. the mean number of bits one save if one new that the data is distributted according to P_θ rather coding as if the data were distributed according to P. If P^* is the distribution that achieves the minimum in Equation (7), then a capacity achieving input distribution is the same as a probability measure Q such that

$$P^* = \int_\Theta P_\theta \, \mathrm{d}Q\theta. \tag{9}$$

 doi:10.20944/preprints202509.0168.v1

Example 3 (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*

$$\begin{aligned} P_a(a) &= \alpha, \\ P_a(b) &= 0, \\ P_a(e) &= 1 - \alpha, \\ P_b(a) &= 0, \\ P_b(b) &= \alpha, \\ P_b(e) &= 1 - \alpha. \end{aligned} \tag{10}$$

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 ([18], Subsec. 8.1.5) for a detailed discussion of the binary erasure channel.*

Example 4 (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 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 found. 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 [20] and ([16], Sec. 5.2)).*

Kullback-Leibler divergence given by Equation (8) equals Rényi divergence of order 1. If we use Rényi divergence of order ∞ ([21], Thm. 6)

$$D_\infty(P_\theta \| P) = \ln \sup_B \frac{P_\theta(B)}{P(B)} \tag{11}$$

instead of Kullback-Leibler divergence then we get the *regret*, which tells how many bits cab 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 5 (The binomial model). *The distribution that achieves minimax regret can be calculated as the normalized maximum likelihood (NML) distribution. It has point probabilities*

$$\begin{aligned} P_{NML}(X=0) &= \frac{P_0(X=0)}{P_0(X=0) + P_{1/2}(X=1) + P_1(X=2)} = \frac{4}{9}, \\ P_{NML}(X=1) &= \frac{P_{1/2}(X=1)}{P_0(X=0) + P_{1/2}(X=1) + P_1(X=2)} = \frac{1}{9}, \\ P_{NML}(X=2) &= \frac{P_1(X=2)}{P_0(X=0) + P_{1/2}(X=1) + P_1(X=2)} = \frac{4}{9}. \end{aligned} \tag{12}$$

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

As demonstrated in Example 4 and Exampel 5 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*. 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

 doi:10.20944/preprints202509.0168.v1

twice differentiable. Note that this excludes statistical models where Θ is a discrete set. The Fisher information matrix is given by

$$[I(\theta)]_{i,j} = -E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln(f(X;\theta))|\theta\right]. \tag{13}$$

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

$$(\det(I(\theta)))^{1/2}. \tag{14}$$

Example 6 (The binomial model). *For the binomial model we have*

$$-\frac{\mathrm{d}^2}{\mathrm{d}p^2}\ln\left(\binom{n}{x}p^x(1-p)^{n-x}\right) = \frac{x}{p^2} + \frac{n-x}{(1-p)^2}. \tag{15}$$

The Fisher information equals the mean value

$$I(p) = \frac{np}{p^2} + \frac{n-np}{(1-p)^2} = \frac{n}{p(1-p)}. \tag{16}$$

Jeffreys' prior has density proportional to

$$\frac{1}{(p(1-p))^{1/2}}. \tag{17}$$

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 has been observed is a beta distribution with parameters $(x+1/2, n-x+1/2)$.

One crucial property of Jeffreys' prior is that, except for a constant factor, it does not depend on the parametrization [22,23].

Example 7 (The exponential model). *For $\lambda > 0$ the exponential distribution $Expo(\lambda)$ has density*

$$\frac{\exp\left(-\frac{x}{\lambda}\right)}{\lambda},\ x > 0. \tag{18}$$

We have

$$-\frac{2}{\lambda^2}\frac{\exp\left(-\frac{x}{\lambda}\right)}{\lambda} = \frac{2x}{\lambda^3} - \frac{1}{\lambda^2} \tag{19}$$

Hence, the Fisher information is given by

$$I(\lambda) = \lambda^{-2}, \tag{20}$$

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 statistical model as a channel has infinite capacity.

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

$$\int_0^\infty \frac{\exp(-x/\lambda)}{\lambda^2}\,\mathrm{d}\lambda = \frac{1}{x}. \tag{21}$$

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

 doi:10.20944/preprints202509.0168.v1

3.3. 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 $T\Phi_g : \Theta \to \theta$ and $\Psi_g : \mathbb{B} \to \mathbb{B}$. The group actions is said to be *covariant* if

$$\Psi_g(P_\theta) = P_{\Phi_g(\theta)}. \tag{22}$$

The notion of covariance was introduced by A. Holevo in the context of quantum information theory [24]. 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)**.** *Let $(G,\cdot)$ denote a locally compact group. Then there exist 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.

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, that is invariant under actions of the group. This measure will be the uniquely determined left invariant measure except for a multiplicative constant.

Example 8 (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 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 9 (The Binomial model)**.** *In the binomial model there is asymmetry between success and failure corresponding to the mapping $p \to 1 - p$ in the parameter space. The prior distributions in Example 4-6 are all symmetric, but the action of the symmetry group is not transitive, so symmetry alone does not determine the prior.*

Example 10 (The exponential model)**.** *The exponential model $\lambda \to Expo(\lambda)$ the group of positve numbers with multiplication $(\mathbb{R}_+,\cdot)$ has a covariant action on the stattistical 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 to any left Haar measure there exists a Poisson point process with the Haar measure as expectation measure. This gives a probabilistic interpretation that will allow a much wider use of Haar measures in probability theory.

 doi:10.20944/preprints202509.0168.v1

4. Conditioning

Many textbooks handle improper prior distributions by restricting the parameter space. In this section, we will first describe this problematic construction and then use expectation theory to give a more satisfying way of handling improper prior distributions.

4.1. Restriction of the Parameter Space

In many textbooks, improper prior distributions are handled by the selection of a "large" subset $\tilde{\Theta} \subseteq \Theta$ such that $\mu(\tilde{\Theta}) < \infty$. Then the measure $\mu_{\cap\tilde{\Theta}}$ 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 $\mu_{\cap\tilde{\Theta}_n}$ 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. This approach to handling improper prior distributions has been advocated by Akaike [25] and many others. See [26] for a more recent exposition regarding approximation of imprpoer priors by probability measures.

Such an inference is problematic for two reasons. The first reason is that the subset $\tilde{\theta}$ should be chosen before $a \in \mathbb{A}$ has been observed, 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 choosing $\tilde{\Theta}$ 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 Haar measure.

With expectation measures at our disposal, we do not need to restrict to a subset of Θ.

4.2. Normalization and Conditioning for Expectation Measures

Empirical measures can be added, one can take restrictions and one can find induced measures. 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 the same as a one-point process.

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

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

$$\mu_{\cap B} = \int \mu_{\omega \cap B} \, \mathrm{d}P\omega. \tag{23}$$

Unital measures are normally 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. *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*

$$\frac{\mu(B)}{\|\mu\|} = \int_\Omega \frac{\mu_\omega(B)}{\|\mu_\omega\|} \, \mathrm{d}P(\omega | 0 < \|\mu_\omega\| < \infty). \tag{24}$$

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 (24) does not hold can be found in ([2], Ex. 5).

 doi:10.20944/preprints202509.0168.v1

Theorem 2 states that $\mu(B)/\|\mu\|$ is the probability of observing a point in B 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 [2], and the proof of the new theorem will not rely on the proof of 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 denotes a probability measure on Ω and let $\omega \to \nu_\omega$ be a statio-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*

$$\frac{\mu(B)}{\|\mu\|} = P(b_\omega \in B). \tag{25}$$

Proof. The waiting time T_B until the first point in B is observed, has distribution $T_B \sim Expo(\mu(B)^{-1})$, and the waiting time $T_{\complement B}$ until the first observation of a point in $\complement B$ is distributed $T_{\complement B} \sim Expo\left(\mu(\complement B)^{-1}\right)$. We have

$$\begin{aligned} P(b_\omega \in B) &= P(T_B < T_{\complement B}) \\ &= \int_0^\infty \left(\int_{t_B}^\infty \exp(-t_{\complement B}\mu(\complement B))\mu(\complement B)\, \mathrm{d}t_{\complement B} \right) \exp(-t_B\mu(B))\mu(B)\, \mathrm{d}t_B \\ &= \int_0^\infty \exp(-t_B\mu(\complement B)) \exp(-t_B\mu(B))\mu(B)\, \mathrm{d}t_B \\ &= \mu(B) \int_0^\infty \exp\left(-t_B(\mu(B)+\mu(\complement B))\right) \mathrm{d}t_B \\ &= \frac{\mu(B)}{\mu(B)+\mu(\complement B)}, \end{aligned} \tag{26}$$

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

4.3. Conditioning for Imporer Prior Measures

Here we shall just look at how the results of Subsection 4.2 will allow us to give an exact interpretation of conditional probabilities with respect to an *improper prior distribution*.

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 denotes a probability measure on Ω and let $\omega \to \nu_\omega$ be a statio-temporal Poisson process with expectation measure $\mu \times m_{\cap\mathbb{R}_+}$ on*

 doi:10.20944/preprints202509.0168.v1

$\mathbb{B} \times \mathbb{R}_+$. *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*

$$\mu(B|A) = P(b_\omega \in B). \tag{27}$$

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

$$\mu(B|A) = \frac{\mu(B \cap A)}{\mu(A)} = \frac{\mu_{\cap A}(B)}{\|\mu_{\cap A}\|}. \tag{28}$$

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

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

Example 11 (The binary erasure channel)**.** *Consider the binary erasure channel discussed in Example 3. 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 12 (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 spatiotemporal 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 13 (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.

5. 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.

 doi:10.20944/preprints202509.0168.v1

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. [27].

In [28,29] it was proved that for 1-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 3-dimensional exponential families. There are still no results that relate finiteness of minimax redundancy or minimax regret with finiteness of Jeffreys' prior, and there is still a lot of open questions regarding improper prior distributions.

Funding: This research received no external funding.

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

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

References

1. Jones, A. Improper priors. Available online: https://andrewcharlesjones.github.io/journal/improper-priors.html (accessed on 2025-08-29).
2. Harremoës, P. Probability via Expectation Measures. *Entropy* **2025**, *27*, 102. https://doi.org/10.3390/e27020102.
3. Mellick, S. Point processes on locally compact groups and their cost. Phd thesis, Alfréd Rényi Institute of Mathematics, 2019.
4. Abért, M.; Mellick, S. Point processes, cost, and the growth of rank in locally compact groups. *Israel Journal of Mathematics*, *251*, 48–155. https://doi.org/https://doi.org/10.1007/s11856-022-2445-9.
5. Dash, S.; Staton, S. A Monad for Probabilistic Point Processes. In Proceedings of the ACT, 2021.
6. 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, 2021; LICS '21, pp. 1–13, [arXiv:2105.06908]. https://doi.org/10.1109/LICS52264.2021.9470678.
7. O'Hagan, A. *Kendall's Advanced Theory of Statistics*, second edition ed.; Vol. 2B, Wiley, 2010.
8. Wu, Q.; Vos, P. Chapter 6 - Inference and Prediction, 2018. https://doi.org/https://doi.org/10.1016/bs.host.2018.06.004.
9. Rényi, A. On a new axiomatic theory of probability. *Acta Mathematica Academiae Scientiarum Hungarica* **1955**, *6*, 185–335. https://doi.org/doi.org/10.1007/BF02024393.
10. Rényi, A. *Probability Theory*; North-Holland: Amsterdam, 1970.
11. Giry, M. A categorical approach to probability theory. In Proceedings of the Categorical Aspects of Topology and Analysis; Banaschewski, B., Ed., Berlin, Heidelberg, 1982; pp. 68–85.
12. Lieshout, M.V., Spatial Point Process Theory. In *Handbook of Spatial Statistics*; Handbooks of Modern Statistical Methods, Chapman and Hall, 2010; chapter 16.
13. Kallenberg, O. *Random Measures*; Springer: Schwitzerland, 2017. https://doi.org/doi:10.1007/978-3-319-41598-7.
14. Last, G.; Penrose, M. *Lectures on the Poisson Process*; Cambridge University Press, 2017.
15. Kass, R.E.; Wasserman, L.A. The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association* **1996**, *91*, 1343–1370. https://doi.org/doi.org/10.1080/01621459.1996.10477003.
16. Csiszár, I.; Shields, P. *Information Theory and Statistics: A Tutorial*; Foundations and Trends in Communications and Information Theory, Now Publishers Inc., 2004.
17. Grünwald, P. *the Minimum Description Length principle*; MIT Press, 2007.
18. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley, 1991.
19. Ryabko, B.Y. Comments on "A source matching approach to finding minimax codes". *IEEE Trans. Inform. Theory* **1981**, *27*, 780–781. Including also the ensuing Editor's Note, https://doi.org/10.1109/TIT.1981.1056409.
20. Csiszar, I. Sanov Property, Generalized *I*-Projection and a Conditional Limit Theorem. *The Annals of Probability* **1984**, *12*, 768 – 793. https://doi.org/10.1214/aop/1176993227.

 doi:10.20944/preprints202509.0168.v1

21. van Erven, T.; Harremoës, P. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Trans Inform. Theory* **2014**, *60*, 3797–3820. https://doi.org/10.1109/TIT.2014.2320500.
22. Jordan, M.I. Jeffres Prior. Available online: https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/lectures/lecture6.pdf (accessed on 2025-08-29). Lecture notes.
23. Ewing, B. What is an improper prior? Available online: https://improperprior.com/pages/what-is-an-improper-prior/index.html (accessed on 2025-08-29).
24. Holevo, A.S. *Probabilistic and Statistical Aspects of Quantum Theory*; Vol. 1, *North-Holland Series in Statistics and Probability*, North-Holland: Amsterdam, 1982.
25. Akaike, H. The interpretation of improper prior distributions as limits of data dependent proper prior distributions. *Journal of the Royal Statistical Society*, *42*, 46–52.
26. Bioche, C.; Druilhet, P. Approximation of improper priors. *Bernoulli* **2016**, *22*, 1709–1728. https://doi.org/10.3150/15-BEJ708.
27. Taraldsen, G.; Tufto, J.; Lindqvist, B.H. Improper prior and improper posterior. *Scandinavian Journal of Statistics* **2022**, *49*, 969–991. https://doi.org/10.1111/sjos.12550.
28. Grünwald, P.; Harremoës, P. Finiteness of Redundancy, Regret, Shtarkov Sums, and Jeffreys Integrals in Exponential Families. In Proceedings of the Proceedings for the International Symposium for Information Theory, Seoul, 2009. IEEE, June 2009, pp. 714–718.
29. Grünwald, P.; Harremoës, P. Regret and Jeffreys Integrals in Exp. Families. In Proceedings of the Thirtieth Symposium in Information Theory in the Benelux; Tjalkens, T.; Willems, F., Eds., Werkgemeenschap voor Informatie- en Communicatietheorie, Eindhoven, 2009; p. 143.