Received: 9 September 2020 | Revised: 29 April 2021 | Accepted: 19 June 2021

DOI: 10.1111/sjos.12550

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Improper priors and improper posteriors

Gunnar Taraldsen | **Jarle Tufto** | **Bo H. Lindqvist**

Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Correspondence
Gunnar Taraldsen, Norwegian University of Science and Technology, Trondheim, Norway.
Email: gunnar.taraldsen@ntnu.no

Abstract

What is a good prior? Actual prior knowledge should be used, but for complex models this is often not easily available. The knowledge can be in the form of symmetry assumptions, and then the choice will typically be an improper prior. Also more generally, it is quite common to choose improper priors. Motivated by this we consider a theoretical framework for statistics that includes both improper priors and improper posteriors. Knowledge is then represented by a possibly unbounded measure with interpretation as explained by Rényi in 1955. The main mathematical result here is a constructive proof of existence of a transformation from prior to posterior knowledge. The posterior always exists and is uniquely defined by the prior, the observed data, and the statistical model. The transformation is, as it should be, an extension of conventional Bayesian inference as defined by the axioms of Kolmogorov. It is an extension since the novel construction is valid also when replacing the axioms of Kolmogorov by the axioms of Rényi for a conditional probability space. A concrete case based on Markov Chain Monte Carlo simulations and data for different species of tropical butterflies illustrate that an improper posterior may appear naturally and is

useful. The theory is also exemplified by more elementary examples.

KEYWORDS

Axioms for statistics, Bayesian inference, Bayesian problems and characterization of Bayes procedures, foundations and philosophical topics, Haldane's prior, Markov Chain Monte Carlo

1 | INTRODUCTION

The purpose of this paper is to present and exemplify recent mathematical developments (Taraldsen, 2019) that allows a theory of statistical inference that includes both improper priors and improper posteriors. It is based on a replacement of the axioms of Kolmogorov (1933, pp. 2, 14) by the axioms of Rényi (1955) as suggested by Taraldsen and Lindqvist (2016) and reviewed briefly in Appendix. This replacement was suggested already by Lindley (1965, p. xi), but the mathematics for this was not satisfactorily developed then. One recent important development is given by the notion of q-vague convergence towards improper priors as introduced by Bioche and Druilhet (2016). Another recent development linked to improper priors is given by fiducial inference as reviewed by Hannig et al. (2016). The main mathematical contribution in the presentation that follows gives existence and uniqueness of a posterior law given a model formulated in the full generality given by the axioms of Rényi.

An often voiced criticism of the use of improper priors in Bayesian inference is that such priors sometimes do not lead to a proper posterior. This can typically happen in applied settings with sparse data (Bord et al., 2018), but also in other cases as demonstrated in Section 4.2. Taraldsen and Lindqvist (2010) explain that this happens if the marginal distribution of the data is not σ-finite. The dangers of improper posteriors in Markov Chain Monte Carlo (MCMC) methods of inference are well recognized (Gelfand & Sahu, 1999; Hobert & Casella, 1996). The latter, however, suggest that a Gibbs sampler with an improper posterior may be used to obtain meaningful inference for certain model unknowns.

A particular class of problems arise from spatially varying phenomena. They are often modeled using Gaussian random fields, specified by their mean function and covariance function. The spatial correlation structure of these models is commonly specified to be of a certain form (e.g., spherical, power exponential, rational quadratic, or Matern) with a small number of unknown parameters. Berger et al. (2001) show that common choices of default prior distributions, such as the constant prior and the independent Jeffreys prior, typically result in improper posterior distributions for these models.

Berger et al. (2001) first observed this operationally while analyzing a spatial dataset. The MCMC simulations seemed to give a nice looking posterior, but a few days later the nice looking posterior had moved to a different location and had a different shape. If the posterior looks fine, but continually moves around as the MCMC runs on, then MCMC would not be trustworthy with improper posteriors. It can in practice be impossible based on simulations only to decide if the simulations have actually converged, and if the resulting posterior is proper. The result can be different, but seemingly plausible, from one day to the next given random initialization of the MCMC simulation.

Between Handcock and Stein (1993) and Berger et al. (2001) the standard prior used was the constant prior, so there were many articles written over those 8 years that had improper posteriors without being explicit on this. Improper posteriors are not uncommon, because it is difficult to determine good objective priors that avoid the problem (e.g., reference priors). It should be noted that using vague proper priors does not really solve the problem. A vague prior that approximates an improper prior will result in similar computational problems.

The structure of the remaining parts of this paper is as follows. Section 2 presents three motivating examples. The first exemplify a typical applied problem solved by MCMC methods, but it is problematic since the posterior is improper. The other two examples involves respectively the standard scale prior for a Poisson process and the Haldane prior for the binomial. All examples demonstrate directly the usefulness of allowing improper posteriors.

Section 3 presents the initial ingredients in a theory for uncertain knowledge as presented by possibly unbounded measures. The main technical result is Theorem 1 which ensures that prior knowledge P_Θ is mapped uniquely to posterior knowledge P_Θ^y given data y and a statistical model P_Y^θ.

Section 4 gives methods with examples for the actual calculation of posterior knowledge. The most elementary is a direct natural extension of the common formal manipulation with densities. It is reassuring that this follows as a consequence of the general theory from Section 3. For more complicated cases an indicated MCMC method can be used, but further developments can and should be developed.

Section 5 provides a final discussion including more comments on prior work by Kolmogorov (1933), Jeffreys (1939), Lindley (1965), Rényi (1970), Berger (1985), Schervish (1995), Robert (2007), Taraldsen and Lindqvist (2010), and Taraldsen and Lindqvist (2013). Finally, Appendix presents further measure theoretic considerations. It is intended for the more mathematically oriented reader. The main result is Theorem 2 which proves existence of a unique conditional law P^t on a Rényispace.

2 | THREE MOTIVATING EXAMPLES

Within the theory to be presented here, improper posteriors as such are well-defined mathematically, and interpretable as a representation of the state of knowledge. It is hence of interest to develop numerical methods for computing such posteriors for complex models that are used in practice. One possible method is proposed by Tufto et al. (2012, appendix S4) in the context of inference from spatial mark-recapture data. The resulting improper density for the expected life-time of certain butterflies is illustrated in Figure 1.

The key idea is to consider the family of posteriors obtained from restriction to intervals, and then glue the resulting posteriors together in a postprocessing step. The general theory presented in Section 3 implies that this simple idea represents a valid approach. Knowledge is here not represented by a probability measure, but is represented by an unbounded measure. This example is discussed in more detail in Section 4.2

As a simpler motivating example, suppose you observe a homogeneous Poisson process. Assume your state of knowledge about the rate parameter $\lambda > 0$ is appropriately represented by a scale invariant prior density (Jeffreys, 1939, p. 122)

$$\pi(\lambda) = \frac{c}{\lambda}. \tag{1}$$

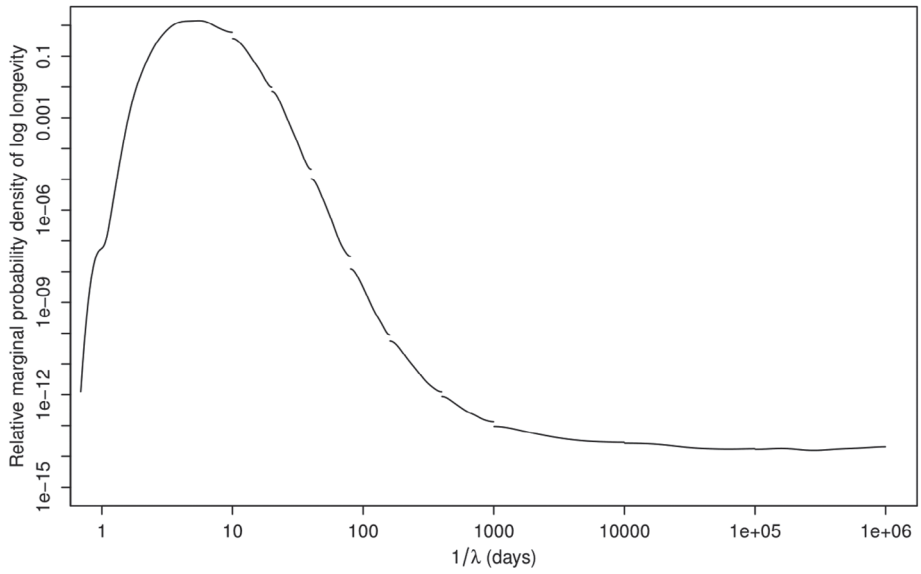

FIGURE 1 An estimate of an improper posterior density. It is obtained by alignment of kernel density estimates based on separate MCMCMarkov Chain Monte Carlo runs. Each run is restricted to different subintervals

This density is not a probability density. The positive constant c is arbitrary, and carries no information. Similar arbitrary constants will play an important role in the theory in later parts of this paper.

The density of the number X of Poisson process occurrences in the interval $(0, t]$ is

$$f(x|\lambda) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}, \quad x = 0, 1, 2, \ldots \tag{2}$$

The posterior corresponding to observing zero occurrences follows by multiplying the prior by the likelihood as usual. The result is then an improper posterior

$$\pi(\lambda|X = 0) = c\frac{e^{-\lambda t}}{\lambda}. \tag{3}$$

This posterior knowledge for λ is different from the initial prior knowledge in Equation (1). High values for λ are less reasonable given the observation $X = 0$. Further updating can be done with this posterior as a prior, and this is consistent with only one updating based on the initial prior. We claim that this is a correct way of incorporating the information given by $X = 0$. Section 3 introduces the necessary mathematics, and its interpretation. The previous argument is then a special case of the general theory.

A related example is the Beta posterior density

$$\pi(p|x) = c\, p^{x-1}(1-p)^{n-x-1}, \tag{4}$$

for the success probability p after observing x successes out of n trials in a Bernoulli process. This corresponds to the improper Haldane (1932) prior

$$\pi(p) = c\, p^{-1}(1-p)^{-1}, \tag{5}$$

discussed by Jeffreys (1939, p. 123). The Haldane prior is the invariant prior corresponding to a natural group structure. Interested readers may consult Jaynes (1968), Eaton (1989), and Terenin and Draper (2017) for further explanation. Assuming a Haldane prior is hence equivalent with a symmetry assumption similarly to assuming the scale invariant prior in Equation (1).

The observation of the number of successes x results in a corresponding updating of the uncertainty associated with p. The posterior in Equation (4) contains the information given by the binomial model, the observation x, and the prior in Equation (5). The posterior, however, is in this case improper for $x = 0$ and for $x = n$. There is nothing wrong with observing $x = 0$ or $x = n$, and a theory for inference should include these possibilities. This theory is presented next.

3 | A THEORY FOR UNCERTAIN KNOWLEDGE

3.1 | Knowledge and uncertainty

What is knowledge? Knowing the definition of the exponential function is a possible example. Another example could be that the second law of Newton gives a very precise description of certain phenomena in nature. A third example could be any of the claims made by Sigmund Freud regarding the behavior of humans. Depending on the situation at hand, many would agree that there is uncertainty involved in these examples. Both knowledge and uncertainty are concepts used in everyday life without any strict definition. The interpretation depends on the context. Usage of these terms in the context of statistics requires more precision.

A concrete example is the electron rest mass. According to Wikipedia in 2021 it equals $9.1093837015(28) \times 10^{-31}$kg. The number 28 in parenthesis gives the standard uncertainty as required and defined in the Guide to the expression of uncertainty in measurement (GUM) by the Joint Committee for Guides in Metrology (JCGM, 2008). It is important to recognize that there is an international standard for reporting uncertainty. The JCGM, chaired by the Director of the BIPM (Bureau international des poids et mesures), was formed in 1997 by the seven International Organizations that had prepared the original versions of the GUM.

The electron mass exemplifies that any quantity in physics determined by measurements, with the exception of the seven quantities defining the basic SI units (BIPM, 2019), has a corresponding uncertainty. The same is true for any estimate obtained from all kinds of data considered by statisticians. The standard uncertainty equals, in the Bayesian interpretation of the GUM, the SD of the probability distribution encoding the uncertain knowledge regarding the electron rest mass.

Uncertain knowledge is represented by a probability distribution in conventional Bayesian statistics. This is exemplified by the probability distribution for the electron rest mass. O'Hagan (2019) exemplify more generally elicitation of expert knowledge concerning an uncertain quantity. The knowledge is expressed in the form of a probability distribution. The notion of a probability distribution is defined mathematically by the axioms of Kolmogorov (1933, pp. 2, 14). Knowledge in this context is always uncertain knowledge, and the term *uncertain knowledge* is hence replaced simply by the term *knowledge*. We choose to say simply *prior knowledge* instead of *prior uncertain knowledge*. It is even customary to speak directly of the *prior* and the *posterior* as in the title of this paper.

Bayesian inference is in principle embarrassingly simple and general: Prior knowledge is transformed uniquely to posterior knowledge by the observed data and the statistical model

for the data. In the simplest cases this is proved as a consequence of Bayes theorem, but the proof in full generality is more complicated. It was given by Kolmogorov (1933, p. 53, eq. 1). With this Kolmogorov developed further the measure theoretic formulation of probability theory given by Frechet (1930) and others. Kolmogorov (1933, p. v) emphasize especially the theory of conditional probabilities and conditional expectations as important novel contribution in his book.

The main mathematical result below is a generalization of Kolmogorov's theory of conditional probabilities to include the case where knowledge is represented by a measure which can be unbounded. With this we develop further the axiomatic theory of probability formulated by Rényi (1955). The resulting Theorem 1 gives conditions such that prior knowledge is transformed uniquely to posterior knowledge by the observed data and the statistical model for the data also when knowledge is represented by a possibly unbounded measure. We consider this to be a most important and needed result given the widespread use of improper priors.

3.2 | Mathematical preliminaries

A complete understanding of the material to be presented requires familiarity with measure theory as presented by Rudin (1987). For completeness we recall some of the basic definitions and recap some less standard definitions. This is necessary to avoid confusion since some well-recognized writers use conventions that deviate from what we consider to be standard. This is exemplified by Casella and Berger (2001, p. 2) who defines an event to be any subset of a sample space, and by Halmos (1950, p. 30) who defines a measure to be a countably additive set function defined on a ring of sets.

A family $\mathcal{F}$ of subsets of a set $\mathcal{X}$ is a σ-algebra if it is closed under complements and countable unions. A set $A \subset \mathcal{X}$ is by definition measurable if $A \in \mathcal{F}$. A measurable space is a set equipped with a σ-algebra. A measure μ is a countably additive function $\mu : \mathcal{F} \to [0, \infty]$ where $\mathcal{F}$ is a σ-algebra (Rudin, 1987, p. 16). This means that $\mu(\emptyset) = 0$ and $\mu(A_1 \cup A_2 \cup \dots) = \mu A_1 + \mu A_2 + \dots$ when $A_1, A_2, \dots$ are disjoint. A measure space $\mathcal{X}$ is a measurable space equipped with a measure μ.

Definition 1 (Admissible condition). An admissible condition A in a measure space equipped with the measure μ is a measurable set A such that $0 < \mu A < \infty$.

This definition is as given by Taraldsen and Lindqvist (2016, definition 1, p. 5009) and by Rényi (1970, p. 72). A more general definition of an admissible condition is defined by Rényi (1970, p. 38) for the more general case where $\mathcal{X}$ is a conditional probability space. We discuss this in more detail in Appendix.

The measure μ is a probability measure and $\mathcal{X}$ is a probability space if $\mu\mathcal{X} = 1$. A measure μ is finite if $\mu\mathcal{X} < \infty$. More generally, the measure μ and the measure space $\mathcal{X}$ are by definition σ-finite if $\mathcal{X}$ is a countable union of admissible conditions. Even more generally, the measure μ and the measure space $\mathcal{X}$ are by definition s-finite if μ is a countable sum of finite measures (Last & Penrose, 2017).

A function $\phi : \mathcal{X} \to \mathcal{Y}$ is measurable if $B = \phi^{-1}(A) = (\phi \in A) = \{x | \phi(x) \in A\}$ is measurable whenever A is measurable. The push-forward measure $\mu_\phi = \mu \circ \phi^{-1}$ is the measure defined by $\mu_\phi(A) = \mu(\phi \in A)$. If μ is s-finite, then it follows that μ_ϕ is s-finite. If μ is σ-finite, then it does not follow that μ_ϕ is σ-finite. This motivates Definition 2.

Definition 2 (σ-Finite function). Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space and let $\mathcal{Y}$ be a measurable space. A measurable function $\phi : \mathcal{X} \to \mathcal{Y}$ is σ-finite if the push-forward measure μ_ϕ is σ-finite.

It follows as a consequence that μ is σ-finite if there exists a σ-finite ϕ. Definition 2 is as used and discussed further by Taraldsen and Lindqvist (2010), Taraldsen and Lindqvist (2016), and is a generalization of the concept of a regular random variable defined by Rényi (1970, p. 73). The push-forward measure μ_ϕ of a σ-finite measure μ by a σ-finite measurable function ϕ is a σ-finite measure. Furthermore, it follows that a conditional measure μ^y concentrated on the level sets $(\phi = y) = \{x|\phi(x) = y\}$ with the property $\mu^y(\mathcal{X}) = 1$ can be defined. This is explained in more elementary terms by Taraldsen and Lindqvist (2010). The concept of a conditional measure is discussed in more technical detail in Appendix.

The main mathematical results in the following is given by Theorem 1 and Theorem 2. These theorems prove that the conditional measure μ^y concentrated on $(\phi = y)$ can be defined also for any measurable function ϕ for, respectively, the case of a σ-finite space $\mathcal{X}$ and a conditional probability space $\mathcal{X}$. The normalization $\mu^y(\mathcal{X}) = 1$ is then not possible in general.

A statistical model is according to currently accepted theories defined as an indexed family of probability measures on the sample space (McCullagh, 2002, p. 1225). The index is the model parameter. We need an additional technical requirement. We will assume that the indexed family of probability measures is a measurable family of probability measures.

Definition 3 (Measurable family of probabilities). Let μ^y be a probability measure on a measurable space $(\mathcal{X}, \mathcal{F})$ for each y in a measurable space $\mathcal{Y}$. The family $\{\mu^y | y \in \mathcal{Y}\}$ is a measurable family of probability measures if $\{y|\mu^y(A) \le \alpha\}$ is measurable for all real α and all measurable A.

In the context of probability and statistics a sample space is by definition a measurable space. An event is a measurable set in a sample space. This corresponds to the axioms of Kolmogorov (1933, pp. 2, 14) which require that $\emptyset$ is an event, that A^c is an event when A is an event, and that $A_1 \cup A_2 \cup \dots$ is an event when $A_1, A_2, \dots$ are events. The underlying abstract space $(\Omega, \mathcal{E}, \mathrm{P})$ is assumed by Kolmogorov (1933) to be a general probability space. It is abstract in the sense of never being specified. It is simply assumed to exist and obeying the axioms. Actual existence must be proved in every concrete modelling case.

We will assume that the underlying abstract $(\Omega, \mathcal{E}, \mathrm{P})$ is allowed to be a general measure space. An admissible condition A is then from the above defined to be an event such that $0 < \mathrm{P}(A) < \infty$. The other definitions given above are similarly inherited. This is next exemplified and motivated by the uniform law on the real line. Two recipes for obtaining conditional probabilities are derived along the way. The first recipe holds for conditioning on a general σ-finite random quantity as explained by Taraldsen and Lindqvist (2010), but the second holds for a general random quantity and is a novelty here. In the latter case the result is not a single conditional probability, but a family of probabilities indexed by the admissible conditions.

3.3 | Conditional probabilities

Symmetry is important in physics, and also in the context of statistics. Knowledge can in some cases be determined by assuming symmetry. The standard prior knowledge P_Θ for a location parameter Θ with sample space $\Omega_\Theta = \mathbb{R}$ is given by letting $\mathrm{P}_\Theta(A)$ equal the length of A. The measure P_Θ is uniquely determined, up till multiplication with a positive constant, by being shift invariant. In this case $\mathrm{P}_\Theta(\Omega_\Theta) = \infty \neq 1$, and this shows that P_Θ is not a proper prior: It is not a

probability distribution. The prior P_Θ is, however, σ-finite since the sample space $\Omega_\Theta = \mathbb{R}$ is a countable union of finite intervals $A_n = [-n, n]$. A random quantity Θ is more generally said to be σ-finite if the corresponding knowledge P_Θ is σ-finite as defined by Definition 2.

Any random quantity, including Θ, is a function defined on the underlying sample space Ω equipped with a law P. It is called a random quantity since there is uncertainty associated with Θ. This uncertainty, the knowledge P_Θ of Θ, or simply the law of Θ, is defined as in the theory of Kolmogorov (1933, eq. 1, p. 21) by Taraldsen and Lindqvist (2010) to be

$$\mathrm{P}_\Theta(A) = \mathrm{P}(\Theta \in A). \tag{6}$$

A random quantity Θ is, by definition, a function $\Theta : \Omega \to \Omega_\Theta$ such that $(\Theta \in A)$ is an event for all events $A \subset \Omega_\Theta$. This ensures that P_Θ is well defined by Equation (6). The reader is hereby warned and reminded that the notation $(\Theta \in A)$ is ambiguous. It does not mean that Θ is an element in A, but it denotes the event $\{\omega|\Theta(\omega) \in A\}$ in Ω. This convention, and similar conventions for other events determined by conditions on random quantities, is used by Kolmogorov (1933, p. 22), and other researchers in probability (Doob, 1953, p. 11). We apologize for this reminder, but feel that it is necessary since there are many authors in the mathematical literature that do not use this convention.

It is assumed above, and throughout this paper, that Ω is equipped with a positive measure P defined on the family $\mathcal{E}$ of events. This is as in the theory of Kolmogorov, but the requirement $\mathrm{P}(\Omega) = 1$ is dropped. The sample space Ω is simply assumed to be a measure space (Rudin, 1987, p. 16, def. 1.18).

The above location prior assumption gives that

$$\mathrm{P}(\Omega) = \mathrm{P}(\Theta \in \Omega_\Theta) = \mathrm{P}_\Theta(\Omega_\Theta) = \infty, \tag{7}$$

so the underlying law P can not be a probability measure in this case. The law P is, however, σ-finite since Ω equals the countable union of the events $B_n = (-n \le \Theta \le n)$ and $\mathrm{P}(B_n) = 2n < \infty$. It turns out, as explained below, that assuming P to be σ-finite is sufficient for the construction of a transformation from a σ-finite prior P_Θ into a σ-finite posterior P_Θ^y given data $Y(\omega) = y$ and a statistical model P_Y^θ.

Taraldsen and Lindqvist (2010) define the conditional knowledge $\mathrm{P}^\theta(A) = \mathrm{P}(A|\Theta = \theta)$ for the case where Θ is σ-finite. It is defined as the Radon–Nikodym derivative of the measure $\mu(C) = \mathrm{P}(A \cap (\Theta \in C))$ with respect to P_Θ. This means that $\mu(d\theta) = \mathrm{P}^\theta(A)\mathrm{P}_\Theta(d\theta)$, and implies

$$\mathrm{E}(\psi(\Theta)A) = \int \psi(\theta)\mathrm{E}^\theta(A)\ \mathrm{P}_\Theta(d\theta) = \mathrm{E}[\psi(\Theta)\mathrm{E}(A|\Theta)]. \tag{8}$$

This is a generalization of the common double expectation formula used in probability theory. The case $A = \Omega$ gives as a consequence that $\mathrm{P}^\theta(\Omega) = 1$, so the conditional knowledge is normalized in this case. The underlying sample space Ω is hence equipped with a measurable family $(\mathrm{P}^\theta|\theta \in \Omega_\Theta)$ of conditional probability measures even though the measure P itself is unbounded. Taraldsen and Lindqvist (2010) discuss this in nontechnical terms with many more examples.

Consider next the random variable

$$T = (0 \le \Theta \le 1). \tag{9}$$

where again $\mathrm{P}_\Theta(A)$ is the length of $A \subset \mathbb{R}$. Note that here, and in the previous paragraph, we identify an event and the corresponding random variable given by its indicator function. This convention is as used by Finetti (1972, p. xxiii), Hartigan (1983, p. 14), and many other authors. Again, we apologize for this reminder, but feel that it is necessary since again there are other writers that do not use this convention. Many authors write χ_A or 1_A for the indicator function of an event, but we prefer to write simply A for both the event and the indicator function.

The indicator variable T takes only the values 0 and 1, and

$$\mathrm{P}(T = 0) = \infty. \tag{10}$$

The measure P_T is then not σ-finite, so the indicator variable T is not σ-finite. Another example of a non-σ-finite variable is the number X of occurrences in the interval $(0, t]$ of the homogeneous Poisson process considered in the Introduction.

The previous exemplifies that there exist many natural random quantities that are not σ-finite. The next aim is to define the conditional law P^t for these case, and more generally for any random quantity T. The definition will be a strict generalization of the above definition of P^θ for the case where Θ is σ-finite.

An event B that fulfills the condition $0 < \mathrm{P}(B) < \infty$ is by Definition 1 an admissible condition. The reason is that the conditional knowledge $\mathrm{P}(\cdot|B)$ defined by

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(AB)}{\mathrm{P}(B)}, \tag{11}$$

gives a probability measure on Ω for each admissible B. Consequently, the conditional probability

$$\mathrm{P}^t(A|B) = \mathrm{P}(A|B, T = t). \tag{12}$$

can be defined as above Equation (8), but with $\mathrm{P}(\cdot)$ replaced by $\mathrm{P}(\cdot|B)$. The resulting conditional knowledge given $T = t$ is hence represented by a family $\{\mathrm{P}^t(\cdot|B)\}$ of probability measures indexed by the admissible conditions. This is similar to how the knowledge P is represented, and interpreted, by the family $\{\mathrm{P}(\cdot|B)\}$ of probability measures as explained by Rényi (1970, pp. 33-37). The initial ingredients in the theory of Rényiare explained in Appendix.

3.4 | Posterior knowledge

We next show how a single posterior knowledge P^t is obtained in the most general case of an arbitrary random quantity T. Let Q_T be a σ-finite measure that dominates P_T. This assumption means that $Q_T(N) = 0$ implies $\mathrm{P}_T(N) = 0$. The measure Q_T is not unique, but it always exists since P is assumed to be σ-finite. A proof is given by Lemma 1 in Appendix. The conditional knowledge P^t is then defined by letting $\mathrm{P}^t(A) = \mathrm{P}(A|T = t)$ be the Radon–Nikodym derivative of the measure $\mu(C) = \mathrm{P}(A \cap (T \in C))$ with respect to Q_T. This means that $\mu(dt) = \mathrm{P}^t(A)\, Q_T(dt)$. In this case it does not follow as a consequence that $\mathrm{P}^t(\Omega) = 1$. In fact, the conditional knowledge P^t, is only unique up till multiplication by an arbitrary positive $c(t)$. This ambiguity is a consequence of the choice of Q_T. The conditional knowledge P^t is a probability measure only when T is σ-finite, and then only by the choice $Q_T = \mathrm{P}_T$.

The corresponding conditional expectation gives the important disintegration

$$\mathrm{E}(\psi(T)A) = \int \psi(t)\mathrm{E}^t(A)\ Q_T(dt), \tag{13}$$

valid for any positive random variables A and $\psi(T)$. This corresponds to the double expectation formula used in ordinary probability theory and generalizes Equation (8).

The result so far is the construction of a posterior knowledge $\mathrm{P}^t(\cdot|B)$ for any admissible condition B, and a construction of a single posterior knowledge P^t unique up till multiplication by a positive $c(t)$. Theorem 2 in Appendix shows that the two constructions are linked by the relation:

$$\mathrm{P}^t(AB) = \mathrm{P}^t(A|B)\mathrm{P}^t(B). \tag{14}$$

Equation (14) can also be used to construct P^t starting from all $\mathrm{P}^t(\cdot|B)$ as demonstrated by Taraldsen et al. (2017). The above construction using the dominating measure Q_T is more straightforward. The construction gives additionally a link between the theory of conditional probability spaces by Rényi (1970) and the theory of disintegration by pseudo-image measures as presented by Bourbaki (1959, VI.44).

A statistical model is given by a measurable family $\{\mathrm{P}_Y^\theta|\theta \in \Omega_\Theta\}$ of probability measures P_Y^θ on the data space Ω_Y indexed by the model space Ω_Θ. Measurability of the family is as in Definition 3 with $\mathcal{X} = \Omega_Y$ and $\mathcal{Y} = \Omega_\Theta$. The model Θ and the data Y are random quantities so they are measurable functions $\Theta : \Omega \to \Omega_\Theta$ and $Y : \Omega \to \Omega_Y$. The previous arguments have the following important consequence for Bayesian inference. It states that observed data y and a statistical model gives a well-defined mapping from prior P_Θ knowledge to posterior P_Θ^y knowledge.

Theorem 1. *Assume that a measurable family of probability measures is specified for the data, and that the data is given. This determines a transformation of σ-finite prior knowledge into a unique σ-finite posterior knowledge.*

Proof. The proof follows from the above arguments, but we will nonetheless summarize the main ingredients. The assumption implies that a joint law of data and model (Y, Θ) is given by $\mathrm{P}_{Y,\Theta}(dy, d\theta) = \mathrm{P}_Y^\theta(dy)\mathrm{P}_\Theta(d\theta)$. It can hence be assumed that $(Y, \Theta) : \Omega \to \Omega_Y \times \Omega_\Theta$ with the joint law determined by the underlying σ-finite law P on Ω. The prior law P_Θ is then mapped into the posterior law P_Θ^y given by $\mathrm{P}_\Theta^y(A) = \mathrm{P}^y(\Theta \in A)$. The posterior law P^y is defined by the disintegration $\mathrm{E}[\psi(Y)A] = \int \psi(y)\mathrm{P}^y(A)\ Q_Y(dy)$ where Q_Y is a σ-finite measure that dominates P_Y. Existence and uniqueness of P^y is a consequence of the Radon–Nikodym theorem. The choice of Q_Y is not unique, but different choices give equivalent posteriors. The notion of $c(y)\mathrm{P}^y$ being equivalent with P^y is motivated by the interpretation by the proper probabilities $\mathrm{P}^y(A|B) = \mathrm{P}^y(AB)/\mathrm{P}^y(B)$ for $0 < \mathrm{P}^y(B) < \infty$. This corresponds to Equation (14) which is valid more generally: If $0 < \mathrm{P}(B) < \infty$, then $\mathrm{P}^y(\cdot|B)$ can be defined from $\mathrm{P}(\cdot|B)$ and Y using the Radon–Nikodym theorem directly. This gives a family $\mathrm{P}^y(\cdot|B)$ of conditional probabilities indexed by B. Equation (14) ensures that the definition of the posterior in terms of a single posterior or as a family of conditional probabilities indexed by B's are consistent. ∎

The claimed uniqueness above does not mean that the posterior P_Θ^y is a unique σ-finite measure for each y. It does not even mean that P_Θ^y is a measure for almost all y. It can, however, be represented as a σ-finite measure if it is additionally assumed that Ω_Θ is a Borel space. This is discussed and explained in more detail in Appendix.

3.5 | Interpretation

Finally, we will explain how knowledge represented by a measure K can be interpreted. This interpretation is used for the case where K is a marginal or a conditional knowledge, or the underlying law P, and for any sample space on which K is defined. The knowledge K is interpreted by considering

$$\mathrm{K}(A|B) = \frac{\mathrm{K}(AB)}{\mathrm{K}(B)}. \tag{15}$$

This defines the family $\{\mathrm{K}(\cdot|B) \mid 0 < \mathrm{K}(B) < \infty\}$ of conditional probability measures indexed by the family of admissible conditions B. The interpretation of each conditional probability can be, depending on the situation at hand, in a frequentist sense (Kolmogorov, 1933, pp. 3-5) or in a subjective sense (Lindley, 2014, p. 19). This is explained in a plethora of introductory books on probability and statistics.

The difference now is that the single probability measure of Kolmogorov is replaced by a consistent family of probability measures. Consistency is defined and discussed in Appendix. The interpretation as given by the interpretation of all conditional probabilities in Equation (15) is explained in more detail by Rényi (1970, pp. 33-37). Additional interpretation is given by the definition of what it means to sample from an unbounded measure. This is described further down in Section 4.3.

A particular consequence of the interpretation is that we will consider the knowledge K to be equivalent with the knowledge cK for any positive c since K and cK define the same family of conditional probabilities. If K depends on some quantity q then c can also depend on q. This is exemplified in the proof of Theorem 1. Again, K and cK define the same family of conditional probabilities.

This interpretation is in particular used for the priors and posteriors for the butterfly, Poisson process, and Bernoulli process examples in Section 2. It is most important since it gives the needed interpretation of the mathematical theory in the context of statistical inference. This interpretation is in particular used for both the prior and the posterior. They are on an equal footing, and this is how uncertain knowledge is represented in a statistical model of a real world phenomena.

4 | CALCULATING POSTERIOR KNOWLEDGE

4.1 | Conditional densities and Bayesian inference

Routine Bayesian argumentation is given by specification of a prior density $\pi(\theta)$, and a family of probability densities $f(y|\theta)$ for the data y conditionally given the model θ. Combined with observed data y, this gives the posterior density $\pi(\theta|y)$. The observation and the model gives hence a transformation of the prior into the posterior. The symbols f and π are used here, and in the following, as generic symbols for densities and conditional densities. It will next be demonstrated how this can be justified also with improper priors and posteriors as a special case of the general definition of a conditional knowledge given in Section 3. This will in particular justify the inference based on sampling from the Poisson process and the Bernoulli distribution discussed in Section 2.

The above assumptions mean more precisely that the probability model for the data Y given a model $\Theta = \theta$ is given by

$$\mathrm{P}_Y^\theta(dy) = f(y|\theta)\ \mu(dy), \tag{16}$$

and the prior knowledge for the model Θ is given by

$$\mathrm{P}_\Theta(d\theta) = \pi(\theta)\ \nu(d\theta), \tag{17}$$

with σ-finite measures μ and ν. Typical examples are given by Lebesgue measure and counting measure, but the theory is not restricted to these cases. Interesting examples include measures concentrated on a manifold such as a circle, a sphere, or more exotic objects.

Equations (16) and (17) are equivalent with

$$\mathrm{P}_{Y,\Theta}(dy, d\theta) = f(y, \theta)\ \mu(dy)\nu(d\theta), \tag{18}$$

where

$$f(y, \theta) = f(y|\theta)\pi(\theta). \tag{19}$$

The assumption $\mathrm{P}_Y^\theta(\Omega_Y) = 1$ ensures in particular that the previous two equations imply $\Theta \sim \pi(\theta)\ \nu(d\theta)$ as stated in Equation (17).

From Theorem 1 in Section 3 it follows that a unique posterior $\mathrm{P}_\Theta^y(d\theta)$ is defined. Starting with a joint density as in equation (18) the posterior is given, as proved below, by $\mathrm{P}_\Theta^y(d\theta) = \pi(\theta|y)\ \nu(d\theta)$ with

$$\pi(\theta|y) = c(y)f(y, \theta). \tag{20}$$

There is no need for the arbitrary constant $c(y)$ since two proportional densities are equivalent when considered as conditional densities for the parameter θ. The $c(y)$ carries no information, but is included to show the arbitrariness of the y dependence.

The proof of Equation (20) with $c(y) = 1$ follows by observing that P_Y is dominated by $Q_T = \mu$ since

$$\mathrm{P}(Y \in C) = \int_C \left(\int f(y, \theta)\ \nu(d\theta) \right)\ \mu(dy). \tag{21}$$

The disintegration

$$\mathrm{P}[(\Theta \in A) \cap (Y \in C)] = \int_C \mathrm{P}_\Theta^y(A)\ \mu(dy) = \int_C \left(\int_A f(y, \theta)\ \nu(d\theta) \right)\ \mu(dy), \tag{22}$$

proves that $\pi(\theta|y) = f(y, \theta)$ with respect to the σ-finite measure ν as claimed. A different choice for the dominating measure Q_T will give a different $c(y)$ normalization of $\pi(\theta|y)$, but the conditional knowledge is unchanged by this. All together, this gives a unique transformation of a prior knowledge into a posterior knowledge.

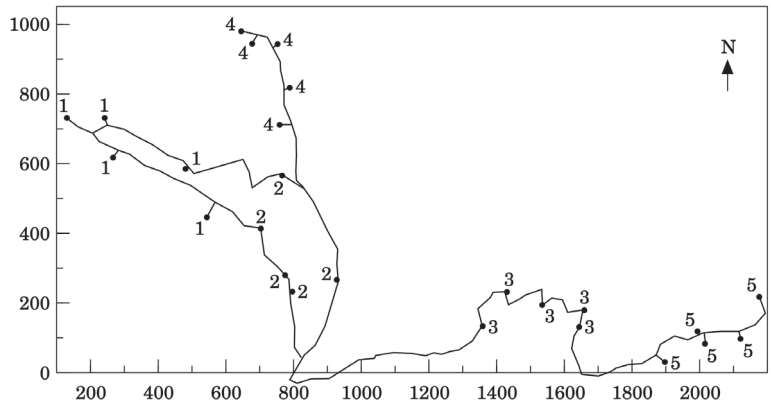

FIGURE 2 Map showing the spatial location of the traps used in the mark-recapture example (Devries & Walla, 2001, reproduced from). Solid lines represent trails. Numbers designate individual replicate trap sites in the sampling areas (1–5). Scale bars are in meters

4.2 | A mark-recapture model for butterflies

The previous subsection proved that the posterior is simply given by the product of the likelihood with the prior. This is exactly as in classical theory for cases described simply by densities, but without the need of a normalization constant. In more complicated cases the likelihood may not be available, but Theorem 1 ensures existence of a unique posterior more generally. To illustrate the application of our new theoretical framework in a realistic applied setting we consider MCMC based Bayesian inference based on spatially explicit mark-recapture data for different species of tropical butterflies (Tufto et al., 2012).

Data were collected using 25 traps located at permanent spatial location separated by distances ranging from about 40 m up to 2 km as shown in Figure 2. During sampling, the traps were baited with fermented fruit that easily attracts species of fruit feeding butterflies. Sampling was conducted approximately concurrently once every day, for five consecutive days during the first 10 days of each month over a period from 1994 to 2004. When captured, previously unmarked individuals were given unique marks before being released. For each marked individual the complete subsequent mark-recapture history was then recorded consisting of a sequence of trap identities (if recaptured) or zeros (if not recaptured) at each subsequent sampling time point. For the nine species used in the study the number of marked individuals were between 102 and 1972 and the number of recaptures between 17 and 709. Tufto et al. (2012) and Devries and Walla (2001) give more details.

It is assumed that all individuals disperse according to independent Brownian motions in two dimensions. The position of individual k at the ith sampling event, conditional on its position at the $(i-1)$th sampling event, is $\mathbf{z}_i^{(k)}|\mathbf{z}_{i-1}^{(k)} \sim N(\mathbf{z}_{i-1}^{(k)}, \sigma^2(t_i - t_{i-1})\mathbf{I}_2)$, where σ^2 is the infinitesimal variance. Furthermore, the adult life span follows an exponential distribution with mortality rate λ. The model assumes that an individual becomes trapped with probability one if its location $\mathbf{z}_i^{(k)}$ is within the trap attraction distance v of any given trap at the time t_i of a given sampling event. Individuals that are not captured are thus at a distance greater than v from all traps at the time of a given sampling event. Hence, not capturing an individual also provide some information about the model parameters.

In addition to the primary model parameters σ^2, λ and v, the statistical model also involves, for each marked individual k, a latent time of death T_k and the latent spatial locations $\mathbf{z}_i^{(k)}$ at all sampling time points up to time T_k. Conditional on the observed mark-recapture history of each individual, the primary model parameters and T_k, these latent locations have a multivariate Gaussian distribution truncated to locations inside and outside distances v from the different traps. The other model quantities similarly have simple conditional distributions facilitating inference via Gibbs sampling.

Given limited knowledge about the primary model parameters, independent, improper scale priors were used for σ, λ, and v by Tufto et al. (2012). This translates to a uniform improper prior on the log of expected adult lifespan $\log(1/\lambda)$. While diagnostic checks of the resulting Markov-chain did not in any way indicate an improper posterior, it follows that the posterior distribution for the above model must be improper with this choice of prior as explained next.

Impropriety follows from a notable feature of mark-recapture data: We can not know if a given individual is alive and not captured or dead at any given sampling event after its last capture. For a finite number of sampling time points, the probability of not recapturing an individual after its last recapture (and the total likelihood) therefore tends to a positive limiting value as the expected adult lifespan tends to infinity (or equivalently, when the adult mortality rate tends to zero) since the probability that an individual happens to be outside the attraction distance v of all traps at all sampling time points after its last recapture is strictly positive. Combined with a uniform prior on the log of expected adult life span, the resulting posterior density also tends to a limiting value for large values of $\log(1/\lambda)$, making the joint posterior distribution improper.

We have argued that there is nothing inherently incoherent with improper posteriors, but that improper posteriors is a valid outcome of Bayesian inference involving improper priors. Computing improper posteriors is therefore of interest. This is discussed more generally in the next section, but the concrete example is here explained first.

Instead of running a single Markov chain, a possible method used by Tufto et al. (2012, appendix S4) is to run several Markov chains restricted to different subintervals for expected adult longevity $\log(1/\lambda)$. For each Markov chain, an estimate of the marginal posterior density of this parameter (up to an unknown constant) can be computed using for example kernel density estimation. To account for the restricted domain of the truncated target density, the reflection method of Silverman (1986, p. 30) was used. Under the assumption that the overall marginal, improper, posterior density of the parameter is a continuous, smooth function, an estimate of this density (up to an arbitrary constant c) was obtained by alignment of the kernel density estimates computed for each subinterval.

The estimate obtained using this method is as shown in Figure 1. Although the estimate is computed for $\log_{10}(1/\lambda)$ up to 6 only, the estimate strongly indicates that the density tends to a limiting positive value for large $\log(1/\lambda)$ such that the overall density indeed is improper. It is worth noting that the density in the flat tail to the right is about 13 orders of magnitude smaller than at the mode which explains why this was undetected by traditional MCMC convergence-diagnostics.

In subsequent studies using such posteriors as prior, it would seem reasonable to estimate the density for larger parameter values in flat tails by extrapolation. It is also clear that the above method can be further improved. One shortcoming are the artifacts appearing at the boundaries between each subinterval resulting from Silverman's reflection method. How to best align the kernel density estimates for each subinterval, also accounting for the likely smoothness of the density function, is another open question. It may also be worth considering other subdivisions schemes perhaps involving overlapping intervals.

4.3 | Knowledge sampling

In simple cases the posterior knowledge is given by the product of the likelihood and the prior as proved in Section 4.1, and exemplified in the Introduction. In more complicated cases it is necessary to consider sampling based methods. This is exemplified in Section 4.2, and more generally by likelihood-free models as in fiducial inference (Taraldsen & Lindqvist, 2013) or in models treated by approximate Bayesian computation (Marin et al., 2012). The posterior density may be analytically intractable, it may be defined on an intractable manifold (Diaconis et al., 2013), or a density may simply not exist. This raises a fundamental question:

What is knowledge sampling?

The answer is well known when knowledge Kz is represented by a probability measure K, but what about the case where knowledge is represented by an unbounded measure K?

The answer presented below can also be used for interpretation purposes. This kind of interpretation parallels the interpretation of a probability as given by the law of large numbers. It will be shown that the concept of a random sample of size n from K is given by a random sample of pairs $(\gamma_1, w_1), (\gamma_2, w_2), \dots, (\gamma_n, w_n)$ from a joint probability distribution of a quantity γ and weight w. This result holds generally, but it is presented next only for the case where the knowledge is represented by a density.

Assume that knowledge for a parameter γ is represented by a density π with respect to a σ-finite measure ν. The aim of sampling can be to compute integrals of the form

$$J = \int \eta(\gamma)\pi(\gamma)\ \nu(d\gamma). \tag{23}$$

The normalization of π is arbitrary, so the computation will always be about comparing two or more integrals of this form.

The integral equals

$$J = \int \eta(\gamma)w(\gamma)\mathrm{p}(\gamma)\ \nu(d\gamma), \tag{24}$$

where p is a suitably chosen probability density and the weight $w = \pi/\mathrm{p}$. Sampling from π can then be done by sampling from p, and returning a weighted sample sequence $(\gamma_1, w_1), (\gamma_2, w_2), \dots$. The sequence can be an *iid* sequence and then $(\gamma_1, w_1), (\gamma_2, w_2), \dots, (\gamma_n, w_n)$ is by definition a random sample of size n from π. It can more generally be a Markov chain as in more modern methods. In both cases, the choice of p should be dictated by π and the family of function η under consideration, but also by implementation issues. This can require considerable skullduggery as demonstrated in a most readable way by Trotter and Tukey (1956).

The previous argument identifies knowledge sampling with weighted sampling. The interpretation explained by Rényi (1970, pp. 33-37) gives a more fundamental answer: Sampling from P is defined by sampling from $\mathrm{P}(\cdot|B)$ for all admissible conditions B. In the density case this translates into being able to compute all integrals

$$J = \int_B \eta(\gamma)w_B\mathrm{p}(\gamma)\ \nu(d\gamma), \tag{25}$$

with $w_B = \int_B \pi(\gamma)\ \nu(d\gamma)$, and $\mathrm{p}(\gamma) = \pi(\gamma|B) = \pi(\gamma)/w_B$. This is then a special case of the weighted sampling.

How and why should it be possible to sample from $\mathrm{P}(\cdot|B)$ for all admissible conditions B? It is intuitively clear that P is uniquely determined by $\mathrm{P}(\cdot|B_n)$ where $B_1 \subset B_2 \ \dots$ with $\cup_i B_i = \Omega$. A proof of this is given by Taraldsen and Lindqvist (2016, p. 5014). It follows hence that it is sufficient to determine $\mathrm{P}(\cdot|B_n)$ for appropriately chosen B_n. The general argument can be continued, but we choose instead to illustrate a general idea by the example considered in Section 4.2.

A visualization of knowledge can be given by plotting the density as in Figure 1. The abscissa is given by the expected adult life span $1/\lambda$ of the butterflies for the interval $I = (0, b)$ with $b = 1{,}000{,}000$ days. The arguments in Section 4.2 indicate that the density should approach a constant, and the choice of b is so large that this is also indicated in Figure 1. Altogether, the graph gives a complete picture of the knowledge about the expected adult life span of the butterflies.

Finally, we will explain how simulations can be used more generally to determine a posterior density on a large interval I of parameter values $\gamma = \psi(\theta)$. It is assumed that the model P_Y^θ and the prior on θ is such that $B = \{\omega|\psi(\Theta(\omega)) \in I\}$ gives that $0 < \mathrm{P}^y(B) < \infty$. This implies that $\mathrm{P}^y(\cdot|B)$ is a probability. There is then a corresponding unique posterior probability distribution for γ restricted to I. The problem has by this been reduced to the problem of sampling from a probability distribution, but it can still be problematic since I is large.

Assume that I can be divided into smaller intervals $I_1, \dots, I_m$ so that sampling can be done for each interval as from the argument in the previous paragraph. The sampling method itself can be of any of the kinds used for posterior sampling for probability distributions and may be differently adapted for each interval. The sampling for each interval is from the law for the entire interval, but normalized to be a probability on each interval.

In the case with densities this means that

$$\pi(\gamma) = w_j \pi(\gamma|I_j), \quad \gamma \in I_j, \tag{26}$$

where the weight is given by

$$w_j = \int_{I_j} \pi(\gamma)\ \nu(d\gamma). \tag{27}$$

This is a special case of the relation (14). It follows that the density π is determined by all the densities $\pi(\cdot|I_j)$ if the weights w_j can be determined. If it is assumed, as in the case illustrated in Figure 1, that the density π is continuous, then it follows that the weights w_j are determined uniquely up till multiplication by a common constant c.

In practice the previous can be implemented in a variety of ways. One approach is to use kernel density estimation of each $\pi(\gamma|I_j)$, and then glue the pieces together as explained. This gives problems at the boundary of each I_j. Another approach is to use a single kernel density estimate for the entire interval I given by putting weights on the samples in each interval. This problem, and its even more challenging versions in more dimensions, is interesting, but will not be discussed further here. Further work on this will be most important for applications.

5 | DISCUSSION

Lindley (1965, p. xi) writes in the preface of his classic book on Bayesian statistics:

> The axiomatic structure used here is not the usual one associated with the name of Kolmogorov. Instead one based on the ideas of Rényi has been used.

It can be concluded that Lindley initially supported the use of conditional probability spaces as introduced by Rényi. We have argued, essentially, that Lindley's initial intuition is correct. The theory of Rényigives a natural approach to Bayesian statistics including commonly used improper priors. Theorem 1 is a natural continuation of the theory of Rényi. It shows, in a mathematically precise way, that improper posteriors are a natural consequence of allowing improper priors.

Historically, the most influential initial work on Bayesian inference is possibly given by the book by Jeffreys (1939). Jeffreys (1939, p. 21) argues in particular that the normalization of probabilities is a rule generally adopted, but that the value ∞ is needed in certain cases. This is in line with the current usage of Bayesian arguments. It is well established that inference based on the posterior gives, indeed, a most rewarding path for obtaining useful inference procedures from both a Bayesian and a frequentist perspective (Berger, 1985; Lehmann & Romano, 2005; Robert, 2007; Schervish, 1995; Taraldsen & Lindqvist, 2013). Taraldsen and Lindqvist (2013) prove in particular that optimal frequentist decision rules are obtained from Bayesian posteriors, and also more generally from posteriors obtained by fiducial arguments.

Parts of Jeffreys arguments were mainly intuitive, and there is a lack of mathematical rigor. We suggest that a rigorous reformulation of some of the original and most important ideas of Jeffreys (1939) can be done within the mathematical theory introduced by Rényi (1970) and continued in our presentation here.

Within this framework we reach the view that improper posteriors, just as improper priors, are not "improper" but reflect the updated state of knowledge about a parameter after conditioning on the data. Returning to the introductory Poisson-process example, at time t, we have clearly learned something about λ in that our belief in large values of the Poisson intensity λ has decreased while our relative degree of belief in small values of λ has remained approximately unchanged. An improper posterior does not imply that our prior was wrong, but only that more data perhaps needs to be collected if possible. Proceeding by using the improper posterior at time t as prior in subsequent inference, say based on the number of occurrences observed in a sufficiently long subsequent interval $(t, t_2]$, we indeed eventually reach the same proper final posterior as the one reached by combining the initial scale prior and the likelihood for the data on $(0, t_2]$. We hope that the reader can appreciate that this simple argument indicates also the potential philosophical importance of representing knowledge by unbounded measures more generally.

An unbounded measure can, according to Rényi, be interpreted by the corresponding family of conditional probabilities given by conditioning on admissible conditions. These conditional probabilities are probabilities in the sense of Kolmogorov, and the interpretation depends on the application. They can, as Lindley (2014) advocates convincingly, be interpreted as personal probabilities corresponding to a range of real-life events. They can also, as needed in for instance quantum physics (Von Neumann, 1932, p. x), be interpreted as objectively true probabilities representing a knowledge for how a system behaves when observed repeatedly under idealized conditions. In classical mechanics probability statement arises from the incompleteness or our knowledge. In quantum mechanics the fundamental postulates include a probabilistic interpretation and a nonatomic probability distribution can correspond to complete knowledge. An example is given by the electron in a hydrogen atom in its ground state (Von Neumann, 1932, p. 297).

The presented theory is not a formal theory for making decisions, but it is a theory for making statistical inference. As a concrete example: If the knowledge is given by the uniform law on the real line, then it is not obvious how the best estimate can be obtained. Similar problems

can also occur in probability theory as exemplified by the uniform law on the circle. More work on the connection to decision theory should be done. A full discussion of this, starting with axioms from Savage (1954), is beyond the scope of the current presentation. The axioms of Savage imply that proper priors and posteriors are the only possible—so the axioms conflict with the theory we present. Our theory can be seen as based essentially on replacing the axioms of Kolmogorov with the axioms of Rényias explained in Appendix. This theory is not a formal theory for decision-making, but a theory for statistical inference.

Assume now that you accept a theory where the prior knowledge is given by a possibly unbounded measure. It is then natural, we claim, that you accept that a resulting posterior knowledge can also be represented by a possibly unbounded measure. Both the prior and the posterior represent knowledge of the same kind.

ACKNOWLEDGEMENTS

The comments on improper spatial statistics and problematic simulations results in the Introduction is based on personal communications with Jim Berger. His insights and comments are highly appreciated. Valuable suggestions from the review process is also acknowledged.

ORCID

Gunnar Taraldsen https://orcid.org/0000-0003-4980-7019

REFERENCES

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer.

Berger, J. O., de Oliveira, V., & Sansó, B. (2001). Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, *96*(456), 1361–1374.

Bioche, C., & Druilhet, P. (2016). Approximation of improper priors. *Bernoulli*, *22*(3), 1709–1728.

BIPM (2019). *The international system of units (SI). (Technical Report)*. Bureau International des Poids et Mesures.

Bord, S., Bioche, C., & Druilhet, P. (2018). A cautionary note on Bayesian estimation of population size by removal sampling with diffuse priors. *Biometrical Journal*, *60*(3), 450–462.

Bourbaki, N. (1959). *Elements of mathematics. Integration I* (Vol. *2004*). Springer.

Casella, G., & Berger, R. L. (2001). *Statistical inference* (2nd ed.). Cengage Learning.

Devries, P. J., & Walla, T. R. (2001). Species diversity and community structure in neotropical fruit-feeding butterflies. *Biological Journal of the Linnean Society*, *74*(1), 1–15.

Diaconis, P., S. Holmes, and M. Shahshahani (2013). Sampling from a manifold. In Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton*, Volume 10 of* IMS Collections, pp. 102–125, Institute of Mathematical Statistics

Doob, J. L. (1953). *Stochastic processes*. Wiley.

Eaton, M. L. (1989). Group invariance applications in statistics. *Paper presented at: Regional Conference Series in Probability and Statistics*, (vol. 1, pp. i–133). Institute of Mathematical Statistics and the American Statistical Association.

Finetti, B. D. (1972). *Probability, induction and statistics: The art of guessing*. Wiley.

Frechet, M. (1930). *Recherches Theoriques Modernes, Fasc. 3 Du Tome I Du Traite Des Probabilites; Par E. Borel et Divers Auteurs*.

Gelfand, A. E., & Sahu, S. K. (1999). Identifiability, improper priors, and gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, *94*(445), 247–253.

Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *28*(1), 55–61.

Halmos, P. R. (1950). *Measure theory*. Van Nostrand Reinhold.

Handcock, M. S., & Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, *35*(4), 403–410.

Hannig, J., Iyer, H., Lai, R. C. S., & Lee, T. C. M. (2016). Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, *111*(515), 1346–1361.

Hartigan, J. (1983). *Bayes theory*. Springer.
Hobert, J. P., & Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, *91*(436), 1461–1473.
Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, *4*(3), 227–241.
JCGM (2008). Evaluation of measurement data —Guide to the expression of uncertainty in measurement (GUM). *(Technical Report)*. International Organization for Standardization.
Jeffreys, H. (1939). *Theory of probability* (3rd ed.). Cambridge University Press 1966.
Kolmogorov, A. (1933). *Foundations of the theory of probability* (Vol. *1956*). Chelsea.
Last, G., & Penrose, M. (2017). *Lectures on the Poisson process Institute of Mathematical Statistics Textbooks* (). Cambridge University Press.
Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses*. Springer.
Lindley, D. V. (1965). *Introduction to probability and statistics from a Bayesian viewpoint* (Vol. *I-II*). Cambridge University Press 2008.
Lindley, D. V. (2014). *Understanding uncertainty* (Revised ed.). Wiley.
Marin, J.-M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, *22*(6), 1167–1180.
McCullagh, P. (2002). What is a statistical model? *The Annals of Statistics*, *30*(5), 1225–1310.
O'Hagan, A. (2019). Expert knowledge elicitation: Subjective but scientific. *The American Statistician*, *73*(Suppl 1), 69–81.
Rényi, A. (1955). On a new axiomatic theory of probability. *Acta Mathematica Academiae Scientiarum Hungarica*, *6*(3), 285–335.
Rényi, A. (1970). *Foundations of probability*. Holden-Day.
Robert, C. (2007). *The Bayesian choice: From decision-theoretic foundations to computational implementation* (2nd ed.). Springer-Verlag.
Royden, H. L. (1989). *Real analysis* (3rd ed.). Palgrave Macmillan.
Rudin, W. (1987). *Real and complex analysis*. McGraw-Hill.
Savage, L. J. (1954). *The foundations of statistics* (2nd Revised ed.). Dover Publications 1972.
Schervish, M. J. (1995). *Theory of statistics*. Springer.
Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall.
Taraldsen, G. (2019). Conditional probability in Renyi spaces. *arXiv:1907.11038 [math, stat]*.
Taraldsen, G., & Lindqvist, B. H. (2010). Improper priors are not improper. *The American Statistician*, *64*(2), 154–158.
Taraldsen, G., & Lindqvist, B. H. (2013). Fiducial theory and optimal inference. *Annals of Statistics*, *41*(1), 323–341.
Taraldsen, G., & Lindqvist, B. H. (2016). Conditional probability and improper priors. *Communications in Statistics - Theory and Methods*, *45*(17), 5007–5016.
Taraldsen, G., Tufto, J., & Lindqvist, B. H. (2017). Improper posteriors are not improper. *arXiv:1710.08933*.
Terenin, A., & Draper, D. (2017). A noninformative prior on a space of distribution functions. *Entropy*, *19*(8), 391.
Trotter, H. F., & Tukey, J. W. (1956). *Conditional Monte Carlo for normal samples*. In H. A. Meyer (Ed.), *Proceedings of the Symposium on Monte Carlo Methods* (pp. 64–79). Wiley.
Tufto, J., Lande, R., Ringsby, T.-H., Engen, S., Saether, B.-E., Walla, T. R., & DeVries, P. J. (2012). Estimating Brownian motion dispersal rate, longevity and population density from spatially explicit mark-recapture data on tropical butterflies. *The Journal of Animal Ecology*, *81*(4), 756–769.
Von Neumann, J. (1932). *Mathematische Grundlagen Der Quantenmechanik (Mathematical Foundations of Quantum Mechanics)*. Springer 1955.

How to cite this article: Taraldsen, G., Tufto, J., & Lindqvist, B. H. (2021). Improper priors and improper posteriors. *Scandinavian Journal of Statistics*, 1–23. https://doi.org/10.1111/sjos.12550

APPENDIX A. MEASURE THEORETIC CONSIDERATIONS

The mathematics in the previous presentation is correct, but some readers may feel that it is mathematically imprecise. The purpose of this appendix is to supply more details on measure theoretic aspects. It is assumed that the reader is well acquainted with basic measure theory as presented by Rudin (1987). The aim next is to present the initial ingredients in the theory of Rényi (1955) which replaces the theory of Kolmogorov (1933), and then formulate our main mathematical results in this context.

An event is by definition a measurable set in a measurable space Ω. A family $\mathcal{B}$ of events that obeys the following three axioms is, according to Rényi (1970, p. 38, def. 2.2.1), a *bunch*:

1. $\emptyset \notin \mathcal{B}$
2. $B, C \in \mathcal{B} \Rightarrow B \cup C \in \mathcal{B}$
3. There exist $B_1, B_2, \ldots \in \mathcal{B}$ with $\cup_i B_i = \Omega$.

A set $B \in \mathcal{B}$ is by definition an admissible condition according to Rényi (1970, p. 38). This is then a generalization of Definition 1.

A conditional probability space (Rényi, 1970, p. 38, def. 2.2.2) is a measurable space equipped with a consistent family of probability measures $(\mathrm{P}(\cdot|B)|B \in \mathcal{B})$ where $\mathcal{B}$ is a bunch. Consistency is defined by the relation

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(AB|C)}{\mathrm{P}(B|C)}, \quad \mathrm{P}(B|C) > 0, \tag{A1}$$

holding for all events A, and all $B, C \in \mathcal{B}$ with $B \subset C$. Taraldsen and Lindqvist (2016) discuss this in more detail in the context of statistics.

The structure theorem of Rényi (1970, p. 40, thm 2.2.1) gives that any conditional probability space can be represented by a σ-finite measure μ in the sense that

$$\mathrm{P}(A|B) = \frac{\mu(AB)}{\mu(B)}, \tag{A2}$$

for all events A and all $B \in \mathcal{B}$. The measure μ is determined uniquely up to a positive constant factor. Conversely, a σ-finite measure μ defines a consistent family of probability measures by Equation (A2), and by defining $\mathcal{B} = \{B|0 < \mu(B) < \infty\}$. The resulting conditional probability space is, according to Rényi (1970, p. 43), a full conditional probability space. We use the term *Rényispace* as equivalent with a full conditional probability space.

In the discussion following Equation (7) it was stated that it was sufficient to assume that P is σ-finite, but also that the corresponding knowledge is equivalently represented by $c\mathrm{P}$ where $c > 0$ is an arbitrary constant. A more precise statement is to assume directly that P equals an equivalence class $\mathrm{P} = [\mu] = \{c\mu|0 < c < \infty\}$ where μ is a σ-finite measure. Due to the Rényistructure theorem this is the same as assuming that Ω is a full conditional probability space. This is our basic assumption when the theory of Kolmogorov (1933) is replaced by the generalization given by Rényi (1970) as discussed in more detail by Taraldsen and Lindqvist (2016).

A particular consequence is that the uniform law P_Θ is not a measure, but rather equals the equivalence class given by all measures that are shift invariant, or equivalently all measures that are equivalent with Lebesgue measure on the real line Ω_Θ. In this case Θ is a regular random variable in the sense of Rényi (1970, p. 73). This is the same as assuming that Θ is a σ-finite

quantity as defined more generally by Taraldsen and Lindqvist (2010). The result is then that Ω_Θ is a full conditional probability space with a bunch $\mathcal{B}_\Theta$ given by the admissible conditions B defined by obeying $0 < \mathrm{P}_\Theta(B) < \infty$.

The law of a general random quantity Θ is defined by Equation (6), but it must be interpreted by using representatives from the equivalence classes: Let $\mu \in \mathrm{P}$. Then $\mathrm{P}_\Theta = [\mu_\Theta]$ with $\mu_\Theta(A) = \mu(\Theta \in A)$. The resulting equivalence class P_Θ does not depend on the choice of $\mu \in \mathrm{P}$. Similar interpretations in terms of representatives must be used also for the other equations and definitions presented in the previous sections. The convention of using the same symbol P for both the equivalence class and a representative measure P can be confusing, but it is similar to using the same symbol f for a function and the resulting equivalence class $f \in L^2(\mu)$.

It may seem unimportant to distinguish between a a measure μ and the corresponding equivalence class $[\mu]$. The distinction has, however, many important consequences as discussed by Taraldsen and Lindqvist (2016). An element $[\mu]$ in the quotient space is a C-measureand a C-measurespace is a measurable space equipped with a C-measure. A C-measurespace based on a σ-finite measure is equivalent with a full conditional probability space (Taraldsen & Lindqvist, 2016, proposition 3).

A particular consequence is that convergence concepts are changed by going to the quotient space defined by the equivalence relation. A particularly important example is given by vague convergence of Radon measures which is replaced by q-vague convergence. This is introduced and discussed in some detail with illustrating examples by Bioche and Druilhet (2016). It is important because improper priors are often viewed intuitively as limits of proper priors, and this intuition can then be made precise. A different convergence concept is introduced and discussed by Rényi (1970, p. 57-). Altogether, the mathematics associated with the resulting topologies is here not fully developed and more can be done.

The conditional law P^t, or equivalently $\mathrm{P}(\cdot|T = t)$, is defined as explained before Equation (8). It is, again, an equivalence class, but this time it is more complicated in several ways. Firstly, equivalence $\sim$ is defined not by a positive constant, but by a positive measurable function $c(t)$: $\mu_1^{(\cdot)} \sim \mu_2^{(\cdot)}$ if and only if there exists a positive measurable c with $\mu_1^t(A) = c(t)\mu_2^t(A)$ for almost all t for all measurable A. The exceptional set has P_T measure 0, but it may depend on A.

Furthermore, it will not be assumed that $\mu^t \in \mathrm{P}^t$ is a measure for almost all t. This is as in ordinary probability theory as explained by Halmos (1950, p. 209). It is a positive conditional measure in the sense that $\mu^t(\emptyset) = 0$, $\mu^t(A) \geq 0$, and $\mu^t(A_1 + A_2 + \dots) = \mu^t(A_1) + \mu^t(A_2) + \dots$ with almost everywhere equality again depending on the measurable sets $A, A_1, \dots$. It can be shown that there exists a version of $\mu^t \in \mathrm{P}^t$ such that it is a measure for almost all t if Ω is a Borel space using standard results from measure theory (Royden, 1989, p. 406)(Schervish, 1995, p. 618). This is not needed nor assumed here. The space Ω is only assumed to be a full conditional probability space. Integration with respect to μ^t can nonetheless be defined as demonstrated already by Kolmogorov (1933, eq. 10 on p. 54). This is what is needed for applications

In the discussion preceding Equation (8) it was assumed that there exists a σ-finite measure Q_T that dominates P_T. A more complete statement and proof is as follows.

Lemma 1. *Let Ω be a Rényispace with law* P*, let Ω_T be a measurable space, and let $T : \Omega \to \Omega_T$ be measurable. There exists then a σ-finite measure Q_T so that $Q_T(A) = 0$ implies $\mu(T \in A) = 0$ for all $\mu \in \mathrm{P}$.*

Proof. Let $Q_T(A) = \nu(T \in A)$ where ν is a probability measure with the same zero sets as $\mu \in \mathrm{P}$. The existence of a probability measure $\nu(d\omega) = w(\omega)\mu(d\omega)$ with $w > 0$ is a standard result (Rudin, 1987, lemma 6.9, p. 121). ∎

Usage of the dominating σ-finite measure Q_T for cases where P_T is not assumed to be σ-finite is the key ingredient in the construction of P^t. The idea is similar and generalizes the disintegration of a measure relative to a pseudo-image as discussed in the context of measures on topological spaces by Bourbaki (1959, VI.45). Taraldsen et al. (2017) provide an alternative route by constructing P^t from $\mathrm{P}^t(\cdot|\cdot)$ more directly.

The following Theorem is a generalization of the structure theorem of Rényi (1970, p. 40, thm 2.2.1) and includes in particular Equation (14).

Theorem 2. *Let Ω be a Rényispace with law P and bunch $\mathcal{B}$, let Ω_T be a measurable space, and let $T : \Omega \to \Omega_T$ be measurable. There exists then a unique conditional law P^t such that*

$$\mu^t(AB) = \mathrm{P}^t(A|B)\mu^t(B), \tag{A3}$$

for all $\mu^t \in \mathrm{P}^t$, all events A and all $B \in \mathcal{B}$.

Proof. The Rényistructure theorem with $\mu \in \mathrm{P}$ gives

$$\mathrm{P}_T(C|B) = \mathrm{P}(T \in C|B) = \mu((T \in C)B)/\mu(B).$$

Disintegration with respect to Q_T from Lemma 1 gives

$$\mu((T \in C)B) = \int_C \mu^t(B)Q_T(dt).$$

Combined this gives the identity

$$\mathrm{P}_T(dt|B) = (\mu^t(B)/\mu(B))\ Q_T(dt).$$

Disintegration gives also

$$\int_C \mu^t(AB)Q_T(dt)/\mu(B) = \mathrm{P}((T \in C)A|B),$$

which equals, using the previous identity,

$$\int_C \mathrm{P}^t(A|B)\mathrm{P}_T(dt|B) = \int_C \mathrm{P}^t(A|B)\mu^t(B)Q_T(dt)/\mu(B).$$

This implies

$$\int_C \mu^t(AB)\ Q_T(dt) = \int_C \mathrm{P}^t(A|B)\mu^t(B)\ Q_T(dt),$$

and Equation (A3) is proved since this holds for all events C in Ω_T. ∎

The previous Theorem 2 contains the structure theorem of Rényi (1970) as a special case by letting $T(\omega) = 1$ for all $\omega \in \Omega$. The proof and result provides a link between the disintegration theory presented by Bourbaki (1959), and the theory of conditional probability spaces introduced by Rényi (1955). We believe that this combination is an important contribution to mathematical statistics.