entropy

MDPI

Article

A Noninformative Prior on a Space of Distribution Functions

Alexander Terenin and David Draper *

Applied Mathematics and Statistics, University of California, Santa Cruz, CA 95064, USA; aterenin@ucsc.edu

* Correspondence: draper@ucsc.edu; Tel.: +1-831-459-1295

Received: 20 June 2017; Accepted: 28 July 2017; Published: 29 July 2017

Abstract: In a given problem, the Bayesian statistical paradigm requires the specification of a prior distribution that quantifies relevant information about the unknowns of main interest external to the data. In cases where little such information is available, the problem under study may possess an invariance under a transformation group that encodes a lack of information, leading to a unique prior—this idea was explored at length by E.T. Jaynes. Previous successful examples have included location-scale invariance under linear transformation, multiplicative invariance of the rate at which events in a counting process are observed, and the derivation of the Haldane prior for a Bernoulli success probability. In this paper we show that this method can be extended, by generalizing Jaynes, in two ways: (1) to yield families of approximately invariant priors; and (2) to the infinite-dimensional setting, yielding families of priors on spaces of distribution functions. Our results can be used to describe conditions under which a particular Dirichlet Process posterior arises from an optimal Bayesian analysis, in the sense that invariances in the prior and likelihood lead to one and only one posterior distribution.

Keywords: Bayesian nonparametrics; Dirichlet process; functional equations; Hyers–Ulam–Rassias stability; improper prior; invariance; optimal Bayesian analysis; transformation group

1. Introduction

Consider a statistician working on a problem P in which a vector $\boldsymbol{y} = (y_1, .., y_n)$ of real-valued outcomes is to be observed, and—prior to, i.e., without observing $\boldsymbol{y}$—the statistician's uncertainty is exchangeable, in the usual sense of being invariant under permutation of the order in which the outcomes are listed in $\boldsymbol{y}$. This situation has extremely broad real-world applicability, including (but not limited to) the analysis of a completely randomized controlled trial, in which participants—ideally, similar to elements of a population to which it is desired to generalize inferentially—are randomized. Each participant is assigned either to a control group that receives the current best treatment, or to an experimental group that receives a new treatment whose causal effect on y is of interest. This design, while extremely simple, has proven to be highly useful over the past 90 years, in fields as disparate as agriculture [1], medicine [2], and (in contemporary usage) A/B testing in data science on a massive scale [3]. We use randomized controlled trials as a motivating example below, but we emphasize that they constitute only one of many settings to which the results of this paper apply.

Focusing just on the experimental group in the randomized controlled trial, the exchangeability inherent in $\boldsymbol{y}$ implies via de Finetti's Theorem [4] that the statistician's state of information may be represented by the hierarchical model

$$y_i \mid F \overset{\text{iid}}{\sim} F \qquad\qquad F \sim \pi(F) \tag{1}$$

for $i = 1, .., n$, where F is a cumulative distribution function (CDF) on $\mathbb{R}$ and $\pi(F)$ is a prior on the space of all such CDFs, i.e., the infinite-dimensional probability simplex S_∞. Note that (1) has uniquely

specified the likelihood in a Bayesian nonparametric model for $\boldsymbol{y}$, and all that remains is specification of $\pi(F)$.

Speaking now more generally (not just in the context of a randomized controlled trial), suppose that the nature of the problem P enables the analyst to identify an alternative statistical problem $\tilde{P}$ in which

$$\tilde{P} = g(P) \qquad \text{such that} \qquad g \in G\,, \tag{2}$$

where G is a collection of transformations g from one problem to another having the property that, without having seen any data, $\tilde{P}$ and P are the *exact same problem*. Then, the prior $\tilde{\pi}$ under $\tilde{P}$ must be the same as the prior π under P! Furthermore, since this holds for any $g \in G$, the result will be, as long as G is endowed with enough structure, that there is one and only one prior π, for use in P, that respects the inherent invariance of the problem under study. Bayes' Rule then implies that there is one and only one posterior distribution under P. When this occurs—when both the likelihood function and the prior are uniquely specified, as in the example above—we say that the problem P admits an *optimal Bayesian analysis*.

The logic underlying the above argument has been used to motivate and formalize the notion of noninformative priors for decades. Indeed, in the special case where F is parametric and G is a group of transformations encoding invariance with respect to monotonically-transformed units of measurement, Jeffreys [5] derived the resulting prior distribution. As another example, Jaynes [6] derived the prior distribution for the mean number of arrivals of a Poisson process by using its characterization as a Lévy counting process to specify an appropriate transformation group. Notably, the resulting prior distribution is *not* the Jeffreys prior, because the problem's invariance and corresponding transformation group are different. See Eaton [7] for additional work on this subject.

Having studied this line of reasoning, it is natural to ponder its generality. In this paper we show that the argument can be made quite general—we prove that the argument's formal notions

(a) can be generalized to include *approximately* invariant priors in an ε–δ sense; and
(b) can be extended to infinite-dimensional priors on spaces of CDFs.

We focus on the setting described in (1) and defer more general situations to future work. In this setting we derive a number of results, ultimately showing that the Dirichlet Process [8] prior $\mathrm{DP}(\varepsilon, F_0)$ is an approximately invariant stochastic process for any CDF F_0 on $\mathbb{R}$ and sufficiently small $\varepsilon > 0$. Together with de Finetti's Theorem, this demonstrates that the posterior distribution

$$F \mid \boldsymbol{y} \sim \mathrm{DP}\left(n, \hat{F}_n\right)\,, \tag{3}$$

where $\hat{F}_n$ is the empirical CDF, corresponds in a certain sense to an optimal Bayesian analysis—see Section 3 for more on this point.

Not all approaches to noninformative priors are based on group invariance. Perhaps the earliest approach can be traced back to Laplace [9], who proposed a Principle of Indifference under which, if all that is known about a quantity θ is that $\theta \in \Theta$ (for some set Θ of possible values), then the prior should be uniform on Θ. For example, consider $\Theta = (0,1)$: the fact that $\theta \sim \mathrm{U}(0,1)$ is not consistent with $f(\theta) \sim \mathrm{U}(0,1)$ for any monotonic nonlinear f requires that the problem P under study must uniquely identify the scale on which uniformity should hold for the principle to be valid—this was a major reason for the rise of non-Bayesian theories of inference in the 19th century [10]. Bernardo [11] has proposed a notion of noninformative priors that is defined by studying their effect on posterior distributions, and choosing priors that ensure that prior impact is minimized. Jaynes [12] has proposed the Maximum Entropy Principle, which defines noninformative prior distributions via information-theoretic arguments, for use in settings in which invariance considerations do not lead to a unique prior. All of these notions are different, and applicable to problems where the corresponding notions of noninformativeness arise most naturally.

Most of the work on noninformative priors has focused on the parametric setting, in which the number of unknown quantities is finite. In contrast, Bush et al. [13] and Lee et al. [14] have derived results on noninformative priors in Dirichlet Process Mixture models. Their notion of noninformativeness is completely different from our own, as it is a posteriori, i.e., it involves examining the behavior of the posterior distribution under the priors studied. This makes their approach largely complementary to ours: in specifying priors, it is helpful to understand both the prior's effect on the posterior and the prior's behavior a priori without considering any data.

Here we study noninformative prior specification from a *strictly a priori* perspective. We do not consider the prior's effect on the posterior distribution. There is no data or discussion of computation.

Our motivation is a generalization of the following argument by Jaynes [12]. Suppose that in the randomized controlled trial described above, the outcome y of interest is binary. By de Finetti's Theorem, we know that

$$y_i \mid \theta_1 \overset{\text{iid}}{\sim} \text{Ber}(\theta_1) \tag{4}$$

is the unique likelihood for (e.g., the treatment group in) this problem. Suppose further that the statistician's state of information about θ_1 external to the data set $\boldsymbol{y}$ is what Jaynes calls "complete initial ignorance" except for the fact that $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is such that

$$\{0 \leq \theta_1 \leq 1, 0 \leq \theta_2 \leq 1, \theta_1 + \theta_2 = 1\}. \tag{5}$$

Jaynes argues that this state of information is equivalent to the statistician possessing complete initial ignorance about all possible rescaled and renormalized versions of $\boldsymbol{\theta}$, namely

$$\boldsymbol{\theta'} = \left(\frac{c_1\theta_1}{c_1\theta_1 + c_2\theta_2}, \frac{c_2\theta_2}{c_1\theta_1 + c_2\theta_2} \right) \tag{6}$$

for all positive c_1, c_2. Jaynes shows that this leads uniquely to the Haldane prior

$$\pi(\theta_1) \propto \frac{1}{\theta_1(1-\theta_1)} \qquad \text{or equivalently} \qquad \pi(\theta_1, \theta_2) \propto \frac{1}{\theta_1\theta_2}, \tag{7}$$

where $\theta_2 = 1 - \theta_1$. Combining this result with the unique Bernoulli likelihood under exchangeability, in our language Jaynes has therefore identified an instance of optimal Bayesian analysis. In what follows we (a) extend Jaynes's argument to the multinomial setting with p outcome categories for arbitrary finite p and (b) show how this generalization leads to a unique noninformative prior on S_∞.

Table 1. Notation. Bold symbols refer to vectors. Improper distributions are considered only as limits of conjugate families—we do not attempt to define DP(0) directly as a non-normalizable measure. CDF: cumulative distribution function.

Expression	Description
$\text{Dir}(\boldsymbol{\alpha})$	The Dirichlet distribution with concentration vector $\boldsymbol{\alpha}$.
$\text{Dir}(\alpha, \boldsymbol{F}_0)$	The Dirichlet distribution with concentration parameter α and mean probability vector $\boldsymbol{F}_0$.
$\text{Dir}(0)$	The improper Dirichlet distribution corresponding to $\lim_{\alpha\to 0} \text{Dir}(\alpha, \boldsymbol{F}_0)$ for $\boldsymbol{F}_0$ arbitrary.
$\text{DP}(\alpha, F_0)$	The Dirichlet Process with concentration parameter α and mean CDF F_0.
$\text{DP}(n, \hat{F}_n)$	The Dirichlet Process whose mean CDF $\hat{F}_n$ is the empirical CDF of the data set $\boldsymbol{y}$ of size n.
$\text{DP}(0)$	The improper Dirichlet Process corresponding to $\lim_{\alpha\to 0} \text{DP}(\alpha, \boldsymbol{F}_0)$ for arbitrary F_0.

The $\text{DP}(n, \hat{F}_n)$ posterior and implied $\text{DP}(0)$ prior—see Table 1 for the notational conventions used in this work—have not been subject to the same level of formal study as Dirichlet Process Mixture priors and other priors over CDFs, in part due to the simplicity and discrete nature of $\text{DP}(n, \hat{F}_n)$. On the other hand, Dirichlet and Dirichlet Process priors with small concentration parameters have been used

as low-information priors in a variety of settings (e.g., [15]), without much formal justification. In this paper we offer a mathematical foundation showing that the use of DP(0) is statistically sound.

2. Results

2.1. Preliminaries

To begin our discussion, we first introduce the notion of an invariant distribution, which describes what we mean by the term noninformative.

Definition 1. *[Invariant Distribution] A density $\pi(\boldsymbol{\theta})$ is* invariant with respect to a transformation group *G if for all $\tilde{\pi}(\boldsymbol{\theta}) = \pi[g(\boldsymbol{\theta})]$ with $g \in G$, and all measurable sets A,*

$$\int_A \pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = \int_A \pi[g(\boldsymbol{\theta})]\,\mathrm{d}g(\boldsymbol{\theta}) = \int_A \tilde{\pi}(\boldsymbol{\theta}) \left|\frac{\partial[g(\boldsymbol{\theta})]}{\partial(\boldsymbol{\theta})}\right| \mathrm{d}\boldsymbol{\theta}\,, \tag{8}$$

where $\left|\frac{\partial[g(\boldsymbol{\theta})]}{\partial(\boldsymbol{\theta})}\right|$ is the Jacobian of the transformation.

Note that in Equation (8), if we were to instead take A in the middle and right integrals to be $g^{-1}(A)$, we would exactly get the classical integration by substitution formula, which under appropriate conditions is always true. We are interested in the inverse problem: given a set of transformations in G, does there exist a unique π satisfying (8)?

In a number of practically-relevant cases, G is uniquely specified by the context of the problem being studied. If this leads to a unique prior distribution π, and when additionally a unique likelihood also arises, for example via exchangeability, an optimal Bayesian analysis is possible, as defined in Section 1. It is often the case that the prior distributions that result from this line of reasoning are limits of conjugate families, making them easy to work with—this occurs in our results below, in which the corresponding posterior distributions are Dirichlet.

The above definition is intuitive, but not sufficiently general to be applicable to spaces of functions. There are multiple technical issues:

(a) in many cases, π cannot be taken to integrate to 1;
(b) probability distributions on spaces of functions may not admit Riemann-integrable densities;
(c) G may be defined via equivalence classes of transformations, leading to singular Jacobians; and
(d) infinite-dimensional measures that are non-normalizable are not well-behaved mathematically.

As a result, the above definition needs to be extended to a measure-theoretic setting. We call a transformation group G acting on a measure space *nonsingular* if for $g \in G$ with $\tilde{\pi}(\boldsymbol{\theta}) = \pi[g(\boldsymbol{\theta})]$, we have $\pi \ll \tilde{\pi} \ll \pi$, where $\ll$ denotes absolute continuity of measures.

Definition 2. *[Invariant Measure] Let G be a nonsingular transformation group acting on a measure space. We say that a measure π is* invariant with respect to *G if for any $g \in G$ with $\tilde{\pi}(\boldsymbol{\theta}) = \pi[g(\boldsymbol{\theta})]$ and for any measurable subset A we have*

$$\int_\Omega \mathrm{I}_A\,\mathrm{d}\pi = \int_\Omega \mathrm{I}_A \frac{\mathrm{d}\tilde{\pi}}{\mathrm{d}\pi}\,\mathrm{d}\tilde{\pi}\,, \tag{9}$$

where Ω is the domain of π, I_A is the indicator function of the set A, and $\frac{\mathrm{d}\tilde{\pi}}{\mathrm{d}\pi}$ is the Radon–Nikodym derivative of $\tilde{\pi}$ with respect to π.

It can be seen by taking π to be absolutely continuous with respect to the Lebesgue measure that Equation (9) is a direct extension of Equation (8).

We would ultimately like to extend the above definition to the infinite-dimensional setting. Doing so directly is challenging, because π may be non-normalizable, in which case Kolmogorov's Consistency Theorem and other analytic tools for infinite-dimensional probability measures do not

apply. Here we sidestep this problem by instead extending the definition of invariance to allow us to define a sequence of *approximately* invariant measures, which in our setting can be taken to be probability measures. To do so, two additional definitions are needed.

Definition 3 (ε-invariant Measure)**.** *Let G be a nonsingular transformation group acting on a measure space with invariant measure $\hat{\pi}$. We say that a sequence of measures $\{\pi^{(\varepsilon)} : \varepsilon > 0\}$ is* ε-invariant with respect to G *if for any $g \in G$ with $\tilde{\pi}^{(\varepsilon)}(\boldsymbol{\theta}) = \pi^{(\varepsilon)}[g(\boldsymbol{\theta})]$ and each measurable subset A, the inequality*

$$\left| \int_{\Omega} \mathrm{I}_A \, \mathrm{d}\pi^{(\varepsilon)} - \int_{\Omega} \mathrm{I}_A \, \frac{\mathrm{d}\tilde{\pi}^{(\varepsilon)}}{\mathrm{d}\pi^{(\varepsilon)}} \, \mathrm{d}\tilde{\pi}^{(\varepsilon)} \right| < \varepsilon \tag{10}$$

implies that

$$\left| \pi^{(\varepsilon)}(A) - \hat{\pi}(A) \right| \leq \delta\mu(A)\,, \tag{11}$$

where $\mu(A)$ is a function, $\varepsilon \to 0$ implies that $\delta \to 0$, and Ω is the domain of $\pi^{(\varepsilon)}$ for all ε.

Definition 4 (ε-invariant Process)**.** *Let $\{\Pi^{(\varepsilon)} : \varepsilon > 0\}$ be a sequence of stochastic processes, and let G be a nonsingular transformation group. Let I be an arbitrary finite subset of the index set of the process, let $\pi_I^{(\varepsilon)}$ be the finite-dimensional measure of $\Pi^{(\varepsilon)}$ under I, and let G_I be a finite-dimensional homomorphism of G with invariant measure $\hat{\pi}_I$. We say that the sequence of processes $\Pi^{(\varepsilon)}$ is* ε-invariant *if, for each I, each $g_I \in G_I$ with $\tilde{\pi}_I^{(\varepsilon)}(\boldsymbol{\theta}) = \pi_I^{(\varepsilon)}[g_I(\boldsymbol{\theta})]$ and each measurable subset A, the inequality*

$$\left| \int_{\Omega_I} \mathrm{I}_A \, \mathrm{d}\pi_I^{(\varepsilon)} - \int_{\Omega_I} \mathrm{I}_A \, \frac{\mathrm{d}\tilde{\pi}_I^{(\varepsilon)}}{\mathrm{d}\pi_I^{(\varepsilon)}} \, \mathrm{d}\tilde{\pi}_I^{(\varepsilon)} \right| < \varepsilon \tag{12}$$

implies that

$$\left| \pi_I^{(\varepsilon)}(A) - \hat{\pi}_I(A) \right| \leq \delta\mu_I(A)\,, \tag{13}$$

where $\mu_I(A)$ is a function, $\varepsilon \to 0$ implies that $\delta \to 0$, Ω_I is the domain of $\pi_I^{(\varepsilon)}$ for all ε, and (ε, δ) can be taken to be identical for all I.

Definition 4 has been explicitly chosen to formalize the notion of noninformativeness on a space of functions without constructing a non-normalizable infinite-dimensional measure.

To complete our assumptions, we need to specify G. Our definitions constitute a direct generalization of the transformation group used by Jaynes to derive the Haldane prior for $p = 2$—see Section 1.

Definition 5 (Probability Function Transformation Group)**.** *Let*

$$G_\infty = \left\{ g \colon S_\infty \to S_\infty \right\} \tag{14}$$

be a nonsingular group of measurable functions under composition acting on the infinite-dimensional simplex S_∞.

Definition 6 (Probability Vector Transformation Group)**.** *For non-negative integer p and any vector $(c_1, .., c_p)$ of non-negative constants, let*

$$G_p = \left\{ g \colon (\theta_1, .., \theta_p) \to \left(\frac{c_1\theta_1}{\sum_{i=1}^{p} c_i\theta_i}, .., \frac{c_p\theta_p}{\sum_{i=1}^{p} c_i\theta_i} \right) \right\} \tag{15}$$

be a nonsingular group under composition acting on the p-dimensional simplex S_p, where each element $g \in G$ represents an equivalence class of the transformations (15).

Note that G_p is a p-dimensional homomorphism of G_∞—we use this property in our proofs below. It can also readily be seen that for any g, the constants c_i are only determined up to proportionality.

Proposition 1 (Radon–Nikodym Derivative)**.** *For each $g \in G_p$ and $\tilde{\pi}(\boldsymbol{\theta}) = \pi[g(\boldsymbol{\theta})]$, the Radon–Nikodym derivative of $\tilde{\pi}$ with respect to π is*

$$\frac{\mathrm{d}\tilde{\pi}}{\mathrm{d}\pi}(\boldsymbol{\theta}) = \frac{\prod_{i=1}^{p} c_i}{\left(\sum_{i=1}^{p} c_i \theta_i\right)^p}. \tag{16}$$

Proof. Let λ be the Lebesgue measure on the p-dimensional probability simplex, and define $\tilde{\lambda}(\boldsymbol{\theta}) = \lambda[g(\boldsymbol{\theta})]$. Note first that $\lambda \ll \tilde{\lambda} \ll \pi \ll \tilde{\pi} \ll \lambda$. Note also that

$$\frac{\mathrm{d}\pi}{\mathrm{d}\lambda} = \frac{\mathrm{d}\tilde{\pi}}{\mathrm{d}\tilde{\lambda}}\,, \tag{17}$$

because the same transformation g is used in defining $\tilde{\pi}$ and $\tilde{\lambda}$. Then, note that

$$\frac{\mathrm{d}\tilde{\pi}}{\mathrm{d}\pi} = \frac{\mathrm{d}\tilde{\pi}}{\mathrm{d}\pi}\frac{\mathrm{d}\tilde{\lambda}}{\mathrm{d}\lambda}\frac{\mathrm{d}\lambda}{\mathrm{d}\tilde{\lambda}} = \frac{\mathrm{d}\lambda}{\mathrm{d}\pi}\frac{\mathrm{d}\tilde{\lambda}}{\mathrm{d}\lambda}\frac{\mathrm{d}\tilde{\pi}}{\mathrm{d}\tilde{\lambda}} = \frac{\mathrm{d}\tilde{\lambda}}{\mathrm{d}\lambda}\,, \tag{18}$$

and hence it suffices to consider the transformation g applied to the Lebesgue measure. Consider an arbitrary hypercube B. We have

$$\lambda(B) = \lambda_1(B_1)\,..\,\lambda_p(B_p)\,, \tag{19}$$

where λ_i are 1-dimensional Lebesgue measures, for which we have that

$$\lambda_i(B_i) = b_i - a_i\,, \tag{20}$$

where $[a_i, b_i]$ is the one-dimensional projection of the hypercube B in dimension i. Consider now the transformation g. We may decompose g into d and n where

$$d : (\theta_1, .., \theta_p) \to (c_1\theta_1, .., c_p\theta_p) \qquad\qquad n : (\theta_1, .., \theta_p) \to \left(\frac{\theta_1}{\sum_{i=1}^{p}\theta_i}, .., \frac{\theta_p}{\sum_{i=1}^{p}\theta_p}\right). \tag{21}$$

Now consider the effect of d and n on λ_i. We have

$$\lambda_i[d(B_i)] = c_i(b_i - a_i) \qquad\qquad \text{and} \qquad\qquad \lambda_i[n(B_i)] = \frac{b_i - a_i}{\sum_{i=1}^{p}(b_i - a_i)}\,, \tag{22}$$

hence

$$\lambda_i[g(B_i)] = \frac{c_i(b_i - a_i)}{\sum_{j=1}^{p} c_j(b_j - a_j)}\,. \tag{23}$$

Therefore

$$\lambda[g(B)] = \prod_{i=1}^{p} \frac{c_i(b_i - a_i)}{\sum_{j=1}^{p} c_j(b_j - a_j)} \tag{24}$$

and we can compute the ratio

$$\frac{\tilde{\lambda}(B)}{\lambda(B)} = \frac{\lambda[g(B)]}{\lambda(B)} = \prod_{i=1}^{p} \frac{c_i(b_i - a_i)}{\sum_{j=1}^{p} c_j(b_j - a_j)} \left[\prod_{i=1}^{p}(b_i - a_i)\right]^{-1} = \frac{\prod_{i=1}^{p} c_i}{\left[\sum_{i=1}^{p} c_i(b_i - a_i)\right]^p}\,. \tag{25}$$

This holds for all B, hence the Radon–Nikodym derivative is just

$$\frac{\mathrm{d}\tilde{\lambda}}{\mathrm{d}\lambda}(\boldsymbol{\theta}) = \frac{\prod_{i=1}^{p} c_i}{\left(\sum_{i=1}^{p} c_i \theta_i\right)^p}, \tag{26}$$

which is the desired result. □

Since we are working with non-normalizable measures as improper priors, we cannot rigorously talk about their probability densities. In many cases, such improper priors can be shown to be limits of families of conjugate priors for which the limiting posterior distribution is well-defined, making them usable in practice. To make our discussion of improper priors rigorous, we need the following definition.

Definition 7 (Generalized Density)**.** *Let π be a measure on $\mathbb{R}^p$ (for p a positive integer) such that $\pi \ll \lambda \ll \pi$, where λ is Lebesgue measure on $\mathbb{R}^p$. Suppose that the Radon–Nikodym derivative of π with respect to λ is Riemann-integrable, and define a family of functions equal to the Radon–Nikodym derivative up to a proportionality constant. We call any function in this family a* generalized density *of π.*

2.2. Main Results

Remark 1 (Notation)**.** *In the following results, we will assume that $(\theta_1, .., \theta_p)$ is a probability vector of dimension $p \geq 2$. G_∞ and G_p will be the transformation groups identified in Definitions 5 and 6, respectively. As noted previously in Table 1, $\mathrm{Dir}(\alpha, \boldsymbol{F}_0)$ will denote the Dirichlet distribution under the alternative parametrization based on concentration parameter α and mean probability vector $\boldsymbol{F}_0$. This is equivalent to the usual parameterization in terms of concentration vector $\boldsymbol{\alpha}$ by the identity $\boldsymbol{\alpha} = \alpha \boldsymbol{F}_0$—we refer to this as the $\mathrm{Dir}(\boldsymbol{\alpha})$ distribution. Similarly, $\mathrm{DP}(\alpha, F_0)$ will refer to the Dirichlet Process with concentration parameter α and mean function F_0. We will refer to the improper priors defined via the conjugate limits as $\alpha \to 0$ of $\mathrm{Dir}(\alpha, F_0)$ and $\mathrm{DP}(\alpha, F_0)$ for arbitrary F_0 as $\mathrm{Dir}(0)$ and $\mathrm{DP}(0)$, respectively.*

We are now ready to introduce our first result. The argument below is a direct generalization of the line of reasoning in Jaynes [12]: the Haldane prior obtained is a special case of our result for $p = 2$.

Theorem 1. *Among the class of measures that admit generalized densities, the measure π with generalized density*

$$\pi\left(\theta_1, .., \theta_p\right) \propto \frac{1}{\prod_{i=1}^{p} \theta_i}, \tag{27}$$

which we call $\mathrm{Dir}(0)$, is the unique invariant measure under G_p.

Proof. An invariant measure π under G_p needs to satisfy the equation

$$\int_{S_p} \mathrm{I}_A \, \mathrm{d}\pi = \int_{S_p} \mathrm{I}_A \, \frac{\mathrm{d}\tilde{\pi}}{\mathrm{d}\pi} \, \mathrm{d}\tilde{\pi}, \tag{28}$$

where S_p is the p-dimensional simplex and $\tilde{\pi}(\boldsymbol{\theta}) = \pi[g(\boldsymbol{\theta})]$ for some $g \in G_p$. Since π is assumed to admit a generalized density, we can rewrite (28) as a Riemann integral. In addition, we substitute in the transformation and Radon–Nikodym derivative, and get

$$\int_A \pi\left(\theta_1, .., \theta_p\right) \mathrm{d}\theta_1 .. \mathrm{d}\theta_p = \int_A \pi\left(\frac{c_1\theta_1}{\sum_{i=1}^{p} c_i\theta_i}, .., \frac{c_p\theta_p}{\sum_{i=1}^{p} c_i\theta_i}\right) \frac{\prod_{i=1}^{p} c_i}{\left(\sum_{i=1}^{p} c_i\theta_i\right)^p} \mathrm{d}\theta_1 .. \mathrm{d}\theta_p \,. \tag{29}$$

This formula needs to hold for all measurable sets A, and hence the functions inside the integrals need to be equal pointwise. This yields the functional equation

$$\pi\left(\theta_1,..,\theta_p\right) = \pi\left(\frac{c_1\theta_1}{\sum_{i=1}^p c_i\theta_i},..,\frac{c_p\theta_p}{\sum_{i=1}^p c_i\theta_i}\right)\frac{\prod_{i=1}^p c_i}{\left(\sum_{i=1}^p c_i\theta_i\right)^p}, \tag{30}$$

which will be the main subject of further study. This is a multivariate functional equation that at first may appear fearsome, but is in fact solvable via elementary methods. To solve it, recognizing that (30) must hold for all probability vectors $(\theta_1,..,\theta_p)$ and all vectors $(c_1,..,c_p)$ of positive constants c_i, we set

$$\left(\theta_1,..,\theta_p\right) = \left(p^{-1},..,p^{-1}\right) \qquad \text{and} \qquad \sum_{i=1}^p c_i = 1\,, \tag{31}$$

which yields

$$\pi\left(p^{-1},..,p^{-1}\right) = \pi\left(\frac{c_1p^{-1}}{p^{-1}\sum_{i=1}^p c_i},..,\frac{c_pp^{-1}}{p^{-1}\sum_{i=1}^p c_i}\right)\frac{\prod_{i=1}^p c_i}{\left(p^{-1}\sum_{i=1}^p c_i\right)^p}\,. \tag{32}$$

Then, by swapping c_i for θ_i, (32) rearranges into

$$\pi\left(\theta_1,..,\theta_p\right) = \frac{\pi\left(p^{-1},..,p^{-1}\right)p^{-p}}{\prod_{i=1}^p\theta_i} \propto \frac{1}{\prod_{i=1}^p\theta_i}\,, \tag{33}$$

since the numerator is not a function of any θ_i, and it can easily be checked that all such generalized densities are valid solutions to the original equation. Thus (33) is the functional equation's unique solution and therefore the unique invariant measure under G_p. □

The same technique used to solve the functional equation in Theorem 1 can be used to prove a much stronger result: if the functional equation is true approximately, its solutions will approximate those of the exact equation. In the next result we make use of the definition of *stability* of a functional equation due to Hyers, Ulam and Rassias—see Jung [16] for details.

Corollary 1 (Hyers–Ulam–Rassias Stability)**.** *Suppose we have*

$$\left|\pi\left(\theta_1,..,\theta_p\right) - \pi\left(\frac{c_1\theta_1}{\sum_{i=1}^p c_i\theta_i},..,\frac{c_p\theta_p}{\sum_{i=1}^p c_i\theta_i}\right)\frac{\prod_{i=1}^p c_i}{\left(\sum_{i=1}^p c_i\theta_i\right)^p}\right| < \delta. \tag{34}$$

Then

$$\left|\pi\left(\theta_1,..,\theta_p\right) - \hat{\pi}\left(\theta_1,..,\theta_p\right)\right| < \delta\frac{e^{e^{-1}}}{\prod_{i=1}^p\theta_i}\,, \qquad \textit{where} \qquad \hat{\pi}\left(\theta_1,..,\theta_p\right) \propto \frac{1}{\prod_{i=1}^p\theta_i}. \tag{35}$$

Proof. By repeating the technique from the previous proof, we have

$$\left|\pi\left(p^{-1},..,p^{-1}\right) - \pi\left(c_1,..,c_p\right)\frac{\prod_{i=1}^p c_i}{p^{-p}}\right| < \delta\,, \tag{36}$$

which can be rewritten

$$\left|\pi\left(\theta_1,..,\theta_p\right) - \frac{\pi\left(p^{-1},..,p^{-1}\right)p^{-p}}{\prod_{i=1}^p\theta_i}\right| < \delta\frac{p^{-p}}{\prod_{i=1}^p\theta_i} < \delta\frac{e^{e^{-1}}}{\prod_{i=1}^p\theta_i}\,, \tag{37}$$

where the last inequality is strict because p is a positive integer. Letting

$$\frac{\pi\left(p^{-1},..,p^{-1}\right)p^{-p}}{\prod_{i=1}^{p}\theta_i} \propto \frac{1}{\prod_{i=1}^{p}\theta_i} \propto \hat{\pi}\left(\theta_1,..,\theta_p\right), \tag{38}$$

we get

$$\left|\pi\left(\theta_1,..,\theta_p\right) - \hat{\pi}\left(\theta_1,..,\theta_p\right)\right| < \delta\frac{e^{e^{-1}}}{\prod_{i=1}^{p}\theta_i}, \tag{39}$$

which is the stability result desired. □

This suffices to prove our result for the Dirichlet distribution.

Theorem 2. $\mathrm{Dir}(\varepsilon, \boldsymbol{F}_0)$ *is an* ε*-invariant measure under* G_p *for all* $\boldsymbol{F}_0$.

Proof. By repeating the steps of Theorem 1 and combining them with Corollary 1, we obtain that $\mathrm{Dir}(\varepsilon, \boldsymbol{F}_0)$ is ε-invariant under G_p if and only if it satisfies

$$\left|\pi^{(\varepsilon)}\left(\theta_1,..,\theta_p\right) - \hat{\pi}\left(\theta_1,..,\theta_p\right)\right| < \delta\frac{e^{e^{-1}}}{\prod_{i=1}^{p}\theta_i} \qquad \text{for some} \qquad \hat{\pi}\left(\theta_1,..,\theta_p\right) \propto \frac{1}{\prod_{i=1}^{p}\theta_i}. \tag{40}$$

Substituting in $\mathrm{Dir}(\varepsilon, \boldsymbol{F}_0)$, and choosing the constant C_ε of the generalized density $\hat{\pi}$ to be the same as for the Dirichlet, we get

$$\left|C_\varepsilon\prod_{i=1}^{p}\theta_i^{\varepsilon F_{0i}-1} - \frac{C_\varepsilon}{\prod_{i=1}^{p}\theta_i}\right| < \delta\frac{e^{e^{-1}}}{\prod_{i=1}^{p}\theta_i}, \tag{41}$$

where F_{0i} are the components of the probability vector $\boldsymbol{F}_0$, and this expression simplifies to

$$C_\varepsilon\, e^e\left|\prod_{i=1}^{p}\theta_i^{\varepsilon F_{0i}} - 1\right| < \delta. \tag{42}$$

Since $0 \leq \theta_i \leq 1$ for all i, the product is upper bounded by 1 and lower bounded by 0. Thus the inequality holds near zero if

$$C_\varepsilon < \delta \tag{43}$$

for all $(\theta_1,..,\theta_p)$, and since $C_\varepsilon \to 0$ we get that, as $\varepsilon \to 0$, we can choose δ such that $\delta \to 0$. Thus, $\mathrm{Dir}(\varepsilon, \boldsymbol{F}_0)$ is ε-invariant for all $\boldsymbol{F}_0$. □

We now extend Theorem 2 to get an analogous result for the Dirichlet Process.

Theorem 3. $\mathrm{DP}(\varepsilon, F_0)$ *is an* ε*-invariant process under* G_∞ *for all* F_0.

Proof. Consider an arbitrary finite-dimensional index I with corresponding homomorphism G_I and finite-dimensional measure $\pi_I^{(\varepsilon)}$. It follows from Theorem 2 that $\pi_I^{(\varepsilon)}$ is ε-invariant with

$$C_\varepsilon < \delta. \tag{44}$$

This inequality depends only on C_ε, so it suffices to show that this constant can be bounded by another constant that is not a function of p and approaches 0. C_ε is an instance of the inverse multivariate beta function, which is a ratio of gamma functions. It is well known that

$$\lim_{x\to 0}\left[\frac{1}{x} - \Gamma(x)\right] = \gamma, \tag{45}$$

where γ is the Euler-Mascheroni constant. Therefore, we have

$$C_\varepsilon = \frac{\Gamma(\varepsilon)}{\prod_{i=1}^{p}\Gamma(\varepsilon F_{0i})} = \frac{O(1/\varepsilon)}{\prod_{i=1}^{p} O(1/\varepsilon)} \leq \frac{O(1/\varepsilon)}{\prod_{i=1}^{2} O(1/\varepsilon)} = O(\varepsilon) \to 0 \tag{46}$$

as $\varepsilon \to 0$. Thus, for each ε, we can choose a δ to satisfy the required expressions under all finite-dimensional index sets, and $\mathrm{DP}(\varepsilon, F_0)$ is therefore an ε-invariant process. □

We conclude our theoretical investigation with a conjecture: the ε-invariance of all finite-dimensional distributions with a uniform δ should suffice for invariance with respect to the original group acting on the infinite-dimensional space.

Conjecture 4. *A stochastic process is an ε-invariant process if and only if the measure of its sample paths is an ε-invariant measure.*

One approach to attempting a proof of this conjecture would involve appropriately extending Kolmogorov's Consistency Theorem to σ-finite infinite-dimensional measures. This can be done, but the notions involved are quite technical—see Yamasaki [17] for more details.

3. Discussion

To see how our results may be applied, consider again the randomized controlled trial of Section 1, and suppose now that the outcome y_i for participant i in the experimental group is categorical with p levels. Under exchangeability, a minor extension of de Finetti's Theorem for dichotomous outcomes then yields that the likelihood can be expressed as

$$y_i \mid \boldsymbol{\theta} \overset{\text{iid}}{\sim} \mathrm{MN}(1, \boldsymbol{\theta}), \tag{47}$$

in which $\mathrm{MN}(k, \boldsymbol{\theta})$ is the multinomial distribution with parameters k and $\boldsymbol{\theta}$. Theorem 1 implies that, modulo inherent abuse of notation under improper priors,

$$(\theta_1, .., \theta_p) \sim \mathrm{Dir}(0) \tag{48}$$

is the unique prior that obeys the fundamental invariance possessed by the problem—namely, invariance with respect to all transformations of probability vectors that preserve normalization. Thus we have extended Jaynes's result for binomial outcomes to the multinomial setting, yielding another instance of optimal Bayesian analysis.

Generalizing to the setting where $\boldsymbol{y}$ is an exchangeable sequence of real-valued outcomes, de Finetti's most general representation theorem implies that

$$y_i \mid F \overset{\text{iid}}{\sim} F \tag{49}$$

is the unique likelihood. If little is known about F, and it is therefore approximately invariant under all measurable functions—i.e., under G_∞, see Definition 5—the prior given by Theorem 3 is

$$F \sim \mathrm{DP}(\varepsilon, F_0). \tag{50}$$

By the usual conjugate updating in the Dirichlet Process setting, the posterior on F given $\boldsymbol{y}$ with the prior in (50) is

$$F \mid \boldsymbol{y} \sim \mathrm{DP}\left(\varepsilon + n, \frac{\varepsilon}{\varepsilon + n} F_0 + \frac{n}{\varepsilon + n} \hat{F}_n\right), \tag{51}$$

in which $\hat{F}_n$ is the empirical CDF based on $\boldsymbol{y}$. Since ε may be taken as close to zero as one wishes, it is natural to regard

$$F \mid \boldsymbol{y} \sim \text{DP}\left(n, \hat{F}_n\right) \tag{52}$$

as an instance of approximately optimal Bayesian analysis for all ε. Conjecture 4 would strengthen this assertion—provided $\text{DP}(0)$ can be rigorously constructed as an infinite-dimensional σ-finite measure, which is beyond the scope of this work.

Though the simplicity of this analysis may at first make it seem limited, its appeal comes from its extremely general ability to characterize uncertainty. See, e.g., Terenin and Draper [18] for an example of a $\text{DP}(n, \hat{F}_n)$ analysis in two randomized controlled trials in e-commerce, one with sample sizes in the tens of millions. Furthermore, sampling from $\text{DP}(n, \hat{F}_n)$ on a discrete domain has recently been shown in a completely different setting—see Appendix B of Terenin et al. [19]—to be asymptotically equivalent to the widely-used frequentist bootstrap of Efron [20]. This also applies to the Bayesian bootstrap of Rubin [21], since it is asymptotically equivalent to the frequentist version. Our analysis provides a Bayesian nonparametric justification for this class of methods.

Bayesian analysis cannot proceed without the specification of a stochastic model—prior and sampling distribution—relating known quantities to unknown quantities: data to parameters. One of the great challenges of applied statistics is that the model is not necessarily uniquely determined by the context of the problem under study, giving rise to model uncertainty, which if not assessed and correctly propagated can cause badly calibrated and unreliable inference, prediction and decision—see, e.g., Draper [22]. Perhaps the simplest way to avoid model uncertainty is to recognize settings in which it does not exist—situations where broad and simple mathematical assumptions, rendered true by problem context, lead to unique posterior distributions. Our term for this is *optimal Bayesian analysis*. It seems worthwhile (a) to catalog situations in which optimal analysis is possible and (b) to work to extend the list of such situations—Theorems 1 and 3 are two contributions to this effort.

Acknowledgments: We are grateful to Daniele Venturi, Yuanran Zhu, and Catherine Brennan for their thoughts on differential equations, which we originally used in a much longer and more complicated proof of the solution of our functional equation. We are grateful to Dan Simpson for his thoughts on infinite-dimensional measures. We are additionally grateful to Juhee Lee for her comments on prior specification, and to Thanasis Kottas for his thoughts on Dirichlet Processes. Membership on this list does not imply agreement with the ideas expressed here, nor are any of these people responsible for any errors that may be present.

Author Contributions: Alexander Terenin and David Draper contributed to the conceptual and theoretical development of the methods in this work and co-wrote the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fisher, R.A. *Statistical Methods for Research Workers*; Oliver and Boyd: Edinburgh, UK, 1925.
2. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *Br. Med. J.* **1948**, *2*, 769–782.
3. Kohavi, R.; Longbotham, R. Online controlled experiments and AB tests. In *Encyclopedia of Machine Learning and Data Mining*; Springer: Berlin, Germany, 2015.
4. De Finetti, B. La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré* **1937**, *7*, 1–68. (In French).
5. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A Math. Phys. Eng. Sci.* **1946**, *186*, 453–461.
6. Jaynes, E.T. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, UK, 2003.
7. Eaton, M.L. *Group Invariance Applications in Statistics*; Regional Conference Series in Probability and Statistics; Institute of Mathematical Statistics: Shaker Heights, OH, USA, 1989.
8. Ferguson, T.S. A Bayesian analysis of some nonparametric problems. *Ann. Stat.* **1973**, *1*, 209–230.
9. Laplace, P.S. Mémoire sur la probabilité des causes par les événements. *Mémoires de l'Académie Royale des Sciences de Paris* **1774**, *6*, 621. (In French).
10. Hald, A. *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713–1935*; Springer: Berlin, Germany, 2007.

11. Bernardo, J.M. Reference posterior distributions for Bayesian inference. *J. R. Stat. Soc. Ser. B (Methodol.)* **1979**, *41*, 113–147.
12. Jaynes, E.T. Prior probabilities. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *4*, 227–241.
13. Bush, C.A.; Lee, J.; MacEachern, S.N. Minimally informative prior distributions for non-parametric Bayesian analysis. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2010**, *72*, 253–268.
14. Lee, J.; MacEachern, S.N.; Lu, Y.; Mills, G.B. Local-mass preserving prior distributions for nonparametric Bayesian models. *Bayesian Anal.* **2014**, *9*, 307–330.
15. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.B. *Bayesian Data Analysis*, 3rd ed.; CRC Press: Boca Raton, FL, USA, 2014.
16. Jung, S.M. *Hyers-Ulam-Rassias Stability of Functional Equations in Nonlinear Analysis*; Springer: Berlin, Germany, 2011.
17. Yamasaki, Y. *Measures on Infinite-Dimensional Spaces*; World Scientific: Singapore, 1985.
18. Terenin, A.; Draper, D. Cox's Theorem and the Jaynesian Interpretation of Probability. *arXiv* **2015**, arXiv:1507.06597.
19. Terenin, A.; Magnusson, M.; Jonsson, L.; Draper, D. Pólya Urn Latent Dirichlet Allocation: A doubly sparse massively parallel sampler. *arXiv* **2017**, arXiv:1704.03581.
20. Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **1979**, *7*, 1–26.
21. Rubin, D.B. The Bayesian Bootstrap. *Ann. Stat.* **1981**, *9*, 130–134.
22. Draper, D. Assessment and propagation of model uncertainty. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **1995**, *57*, 45–97.

© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).