# Shannon optimal priors on iid statistical experiments converge weakly to Jeffreys' prior

Holger R. Scholl

Department of Mathematics, University of Kaiserslautern

Postfach 3049, 67653 Kaiserslautern, Germany

## SUMMARY

In 1979, J.M. Bernardo argued heuristically that in the case of regular product experiments his information theoretic reference prior is equal to Jeffreys' prior. In this context, B.S. Clarke and A.R. Barron showed in 1994, that in the same class of experiments Jeffreys' prior is asymptotically optimal in the sense of Shannon, or, in Bayesian terms, Jeffreys' prior is asymptotically least favorable under Kullback Leibler risk. In the present paper, we prove, based on Clarke and Barron's results, that every sequence of Shannon optimal priors on a sequence of regular iid product experiments converges weakly to Jeffreys' prior. This means that for increasing sample size Kullback Leibler least favorable priors tend to Jeffreys' prior.

*Keywords*: Jeffreys' prior, Shannon optimal priors, Kullback Leibler distance, noninformative prior, reference prior, minimax risk, Bayes risk, channel capacity

## 1. INTRODUCTION

In 1956 D.V. Lindley adapted the concept of the information theoretic transmission or information rate to the theory of statistical experiments, see Lindley (1956). An experiment is considered as a Shannon information transmitting channel, the parameter being the unknown character sent, and the data being the characters observed after some trials. A prior's transmission or information rate measures the expected information gain of the experiment given a prior.

Lindley's concept fits nicely into the theory of Bayesian inference. There, the choice of appropriate priors is of considerable importance, and since the fundamental paper of J.M. Bernardo (1979), there has been much research on procedures how to select nonsubjectivistic priors, see e.g. Berger and Bernardo (1992) and Ghosh and Mukerjee (1992). The objective is to find priors that contain little information about the quantity of interest (the parameter) relative to the information contained in the data. Among these so called noninformative or conventional priors, Bernardo's information theoretic reference priors play a major role. Based on Lindley's ideas they are closely connected with those priors achieving the maximum transmission rate, or, in Lindley's context, to (Shannon) optimal priors. These priors maximize the expected amount of information (about the parameter) provided by the experiment. From a statistical point of view we can characterize optimal priors in terms of a Bayes strategy: If we choose the Kullback Leibler distance as the risk function, then optimal priors are those priors realizing both the maximin (Bayes) risk and the minimax risk.

In the important case of regular, continuous parameter experiments with no nuisance parameter present, Bernardo informally identified his asymptotically defined reference prior to be Jeffreys' prior. This special prior, which has a Lebesgue density proportional to the square root of the Fisher Information, was proposed by H. Jeffreys in 1946 (see Jeffreys (1961/1983)), and it turned out to have many desired properties. Extending their earlier results, Clarke and Barron showed in their 1994 paper that Jeffreys' prior is the unique asymptotically optimal prior in the class of positive priors with a Lebesgue density , cf. Clarke and Barron (1990 and 1994). The proof of their result is based on a uniform asymptotic for the risk of the Kullback Leibler Bayes strategy. As the corresponding Bayes risk given a prior is the transmission rate of the prior, the maximum transmission rate (this is called the capacity of the experiment) is the maximin risk as well as the minimax risk. This is shown in Krob (1992) (see Krob and Scholl (1997)) and asymptotically confirmed by Clarke and Barron. See also Haussler (1995) for a corresponding minimax result. Further, Clarke and Barron give the asymptotic value and the rate of convergence for this quantity.

Clarke and Barron concentrate their analysis on fixed priors that are either absolutely continuous (with respect to Lebesgue measure) or are discrete. In contrast, we investigate a sequence of optimal priors in this article, so that we have a different prior for each repetition of the experiment, which is in accord to Bernardo's reference prior method. Thus, we interpret his conjecture in the following sense: Optimal priors asymptotically resemble Jeffreys' prior. Examples show that in many typical cases optimal priors are discrete for finite sample size. Especially if the sample space is finite there is always a discrete optimal prior. See also Berger, Bernardo and Mendoza (1989) or Zhang (1994) in this context. Therefore, we do not assume any additional properties of the priors besides optimality.

This article is structured as follows: In the next section, we first restate some useful information theoretic facts, including a minimax characterization of the channel capacity and optimal priors. Then, in section 3, we demonstrate, in the simpler case of strict orthogonality, how the information theoretic comparison of experiment and sub-experiment leads to a relation of the proportions of optimal priors and the experiment's capacity. In the fourth section, we will use Clarke and Barron's asymptotics to perform this comparison in the general situation, which leads to our main result. Finally, in the last section, we will discuss the main practical implications of our result.

## 2. INFORMATION THEORETIC PRELIMINARIES

To begin with, let us first introduce some notation. With Prob$X$ or Prob$\Theta$ we will respectively denote the space of all probability measures over the (measurable) spaces $X$, and $\Theta$.

**Definition 1:** *A* **statistical experiment** *is a triple* $((X,\mathcal{X}),\mathcal{P},\Theta)$ *consisting of a measurable space* $(X,\mathcal{X})$ *which we will additionally assume to be Polish, and a family* $\mathcal{P} := (P_\theta)_{\theta\in\Theta}$ *of probability distributions over* $(X,\mathcal{X})$ *with parameter space* $\Theta \subseteq \mathbb{R}^d, d \in \mathbb{N}$.

*If additionally* $\mathcal{P}$ *is compact in the topology of the variational distance, and* $\Theta$ *is compact in the* *usual topology on $\mathbb{R}^d$, then the experiment is called* **compact**.

*If all $P_\theta, \theta \in \Theta$, are dominated by a $\sigma$–finite measure $\mu \in \mathrm{Prob}X$ then there exist $\mu$–densities $f_\theta$ for all $\theta \in \Theta$. In this situation the experiment is called* **dominated**. *Let $f_\varphi := \int_\Theta f_\theta \varphi(d\theta)$ denote the density of the mixture distribution with respect to the prior $\varphi$. If the family*

$$(f_\lambda \log f_\lambda)_{\lambda \in \mathrm{Prob}\Theta}$$

*is uniformly $\mu$–integrable then we will say that the experiment is* **uniformly integrable**.

In most cases, we will identify $\mathcal{P}$ with the experiment, when the parameter set and the observation space $(X, \mathcal{X})$ are fixed. The following example includes some relevant probability families fitting into our definition.

**Example 1:** *Let $\mathcal{P}$ be any $d$–dimensional standard exponential family, like the Normal family, the Gamma family, the Binomial family or the Poisson family for suitable $d \in \mathbb{N}$. If $\Theta$ is a compact subset of the natural parameter space, then $\mathcal{P}$ is a dominated, compact and uniformly integrable experiment in the sense of definition 1.*

The quantity

$$\mathcal{I}(\varphi) := -\int_{\mathrm{X}} p_\varphi(x) \log p_\varphi(x) \mu(dx) + \int_\Theta \int_{\mathrm{X}} f_\theta(x) \log f_\theta(x) \mu(dx) \varphi(d\theta)$$

gives us the information rate of the prior distribution $\varphi \in \mathrm{Prob}\Theta$. The assumption of uniform integrability guarantees the existence of all integrals involved. Further, we are then allowed to interchange limits and integrals which makes sure that $\mathcal{I}$ is a continuous function.

There are some rather straight forward criteria to check whether a given experiment is uniformly integrable: E.g. if $\mu$ is a finite measure and there is a constant $K < \infty$, such that we have for $\mu$–almost all $x \in X$ that

$$f_\theta(x) \leq K \quad \text{for all } \theta,$$

then $\mathcal{P}$ is a uniformly integrable family in our sense, i.e. $(f_\lambda \log f_\lambda)_{\lambda \in \mathrm{Prob}\Theta}$ is uniformly integrable. This holds especially for experiments with a finite sample space. Further, experiments with

$(f_\theta \log f_\theta)_{\theta\in\Theta}$ uniformly integrable are also uniformly integrable in the sense of definition 1. This can be seen, for instance, with the help of the Theorem of de la Vallée Poussin (see e.g. Meyer (1966)).

If our uniformly integrable experiment is also compact, then the continuous mapping $\mathcal{I}$ has maximum points. Thus we can find priors maximizing the average information gain by the experiment. These priors achieving maximal information rate are called (Shannon) optimal, and their information rate is defined to be the capacity of the experiment:

$$\mathcal{C} := \max_{\varphi\in\mathrm{Prob}\Theta} \mathcal{I}(\varphi).$$

Similar to Bernardo (1979) and Clarke and Barron (1994) we need the compactness of $\mathcal{P}$ and, corresponding, of $\Theta$, to guarantee the existence of this maximum. As we will assume a continuous one–to–one parametrization, the compactness of $\mathcal{P}$ and $\Theta$ is essentially the same. Compactness seems to be crucial here, but as we can choose any (large) compact subset of the natural parameter space, the assumption may not be too narrow for practical examples.

Information rate and Kullback Leibler distance are related by the following statements that are well known in their classical version.

**Theorem 1:** *Let $\mathcal{P}$ be a compact, dominated and uniformly integrable experiment. Then the following holds for any $\varphi \in \mathrm{Prob}\Theta$:*

1. $\mathcal{I}(\varphi) = \int_\Theta I\!\!K(P_\theta, P_\varphi)\varphi(d\theta)$.

2. $\mathcal{I}(\varphi) = \int_\Theta I\!\!K(P_\theta, P_\lambda)\varphi(d\theta) - I\!\!K(P_\varphi, P_\lambda)$ *for any* $\lambda \in \mathrm{Prob}\Theta$.

3. *Let* $\varphi^{(1)}, \ldots, \varphi^{(N)} \in \mathrm{Prob}\Theta$ *with* $N \in I\!\!N$ *and let* $s := (s_1, \ldots, s_N) \in S_{N-1}$ *be a probability vector in the* $N-1$*-dimensional unit simplex. Let* $\varphi^{(0)} := \sum_{k=1}^N s_k\varphi^{(k)}$. *Then for every* $\lambda \in \mathrm{Prob}\Theta$*:*

$$\sum_{k=1}^N s_k\mathcal{I}(\varphi^{(k)}) + \sum_{k=1}^N s_k I\!\!K(P_{\varphi^{(k)}}, P_\lambda) = \mathcal{I}(\varphi^{(0)}) + I\!\!K(P_{\varphi^{(0)}}, P_\lambda).$$

**Proof:** See Krob and Scholl (1997). ■

Because the Kullback Leibler distance is a non–negative function vanishing if both arguments coincide, we see by point (2), that

$$\mathcal{I}(\varphi) = \inf_{\lambda \in \text{Prob}\Theta} \int_\Theta I\!\!K(P_\theta, P_\lambda)\varphi(d\theta),$$

which means that

$$\mathcal{C} = \max_{\varphi \in \text{Prob}\Theta} \inf_{\lambda \in \text{Prob}\Theta} \int_\Theta I\!\!K(P_\theta, P_\lambda)\varphi(d\theta).$$

Thus, an experiment's capacity is the maximin risk for the Kullback Leibler distance, achievable by optimal priors.

The next theorem is often called the Main Theorem for Computing the Channel Capacity. It goes back in parts to Shannon (1948), and to Eisenberg and Gallager, see Gallager (1968). The theorem gives a sufficient and necessary condition for a prior to be optimal.

**Theorem 2:** *Let $\mathcal{P}$ be a compact, dominated and uniformly integrable experiment with capacity $\mathcal{C}$. Then the following holds:*

*If $\varphi \in \text{Prob}\Theta$ is a prior with corresponding mixture distribution $P_\varphi \in \text{Prob}X$ then the following condition is sufficient and necessary for $\varphi$ being optimal:*

*There is a constant $C < \infty$ so that*

1. *$I\!\!K(P_\theta, P_\varphi) = C$ for $\varphi$–almost all $\theta \in \Theta$,*

2. *$I\!\!K(P_\theta, P_\varphi) \leq C$ for all $\theta \in \Theta$.*

*If the condition holds then $C = \mathcal{I}(\varphi) = \mathcal{C}$.*

**Proof:** See Krob and Scholl (1997). ■

The theorem tells us, that $\mathcal{C}$ is the largest occuring Kullback Leibler distance of a $P_\theta$ to $P_\varphi$. Combined with Theorem 1 it leads to the following minimax equals maximin result, which is due to J. Krob (1992).

**Theorem 3:** *Let $\mathcal{P}$ be a compact, dominated and uniformly integrable experiment.*

1. *Each Shannon optimal prior $\varphi \in \mathrm{Prob}\Theta$ maximizes the minimal Bayes risk. The set of all optimal $\varphi \in \mathrm{Prob}\Theta$ is a non empty convex subset of* $\mathrm{Prob}\Theta$.
2. *There is a unique distribution $Q \in \mathrm{convexhull}(\mathcal{P}) \subseteq \mathrm{Prob}X$, which minimizes the maximal value of the Bayes risk function. If $\varphi$ is any optimal prior, then $Q = P_\varphi$.*
3. *If $\mathcal{C}$ is the capacity of the experiment, then*

$$\begin{aligned}\mathcal{C} &= \min_{\varphi\in\mathrm{Prob}\Theta}\ \max_{\theta\in\Theta} \boldsymbol{K}(P_\theta, P_\varphi)\\ &= \max_{\varphi\in\mathrm{Prob}\Theta}\ \min_{\lambda\in\mathrm{Prob}\Theta} \int_\Theta \boldsymbol{K}(P_\theta, P_\lambda)\ \varphi(d\theta).\end{aligned}$$

**Proof:** See Krob and Scholl (1997). ■

This means that $\mathcal{C}$ is both minimax– and maximin–risk value for a Bayes strategy with $\boldsymbol{K}$ as risk function, optimal priors thus being least favorable under Kullback Leibler risk. For a related minimax result, see also Haussler (1995).

## 3. OPTIMAL PRIORS ON ORTHOGONAL EXPERIMENTS

In order to demonstrate our argumentation for the main result, let us see, how an orthogonal structure of the experiment reflects in the structure of optimal priors.

**Theorem 4:** *Consider a compact, dominated and uniformly integrable experiment $((X,\mathcal{X}),\mathcal{P},\Theta)$ of the following structure:*

$$\mathcal{P} := \begin{pmatrix} \mathcal{P}_1 & \mathcal{O}_1 \\ \mathcal{O}_2 & \mathcal{P}_2 \end{pmatrix},$$

*where $\mathcal{P}_1, \mathcal{P}_2$ are compact, dominated and also uniformly integrable experiments on the disjoint spaces $(X_1,\mathcal{X}_1)$, respectively $(X_2,\mathcal{X}_2)$, with disjoint parameter spaces $\Theta_1$ and $\Theta_2$. $\mathcal{O}_1$ and $\mathcal{O}_2$ expand $\mathcal{P}_1$ and $\mathcal{P}_2$ to $(X,\mathcal{X})$ by zeros, so that $(\mathcal{P}_i,\mathcal{O}_i)$ resp. $(\mathcal{O}_i,\mathcal{P}_i)$ are (sub–)experiments on $(X,\mathcal{X})$ with parameter set $\Theta_i$. Assume the capacities of the involved experiments to be $\mathcal{C},\mathcal{C}_1$ and $\mathcal{C}_2$. Then the following holds:*

1. *Let $\varphi_1, \varphi_2$ be priors optimal for the sub-experiments $\mathcal{P}_1$ resp. $\mathcal{P}_2$. Then the prior*

$$\varphi := \frac{e^{\mathcal{C}_1}}{e^{\mathcal{C}_1} + e^{\mathcal{C}_2}}\overline{\varphi}_1 + \frac{e^{\mathcal{C}_2}}{e^{\mathcal{C}_1} + e^{\mathcal{C}_2}}\overline{\varphi}_2,$$

   *is optimal for $\mathcal{P}$, with $\overline{\varphi}_i \in \mathrm{Prob}\Theta, i = 1, 2$, being the respective expansion of $\varphi_i \in \mathrm{Prob}\Theta_i$ to the full parameter set $\Theta$ by zero.*

2. *If $\varphi$ is an optimal prior for $\mathcal{P}$ then the priors*

$$\varphi_i := \frac{\varphi|_{\Theta_i}}{\varphi(\Theta_i)}, \quad i = 1, 2,$$

   *are optimal for $\mathcal{P}_i, i = 1, 2$.*

3. *The capacity of the experiment is given by $\mathcal{C} = \log(e^{\mathcal{C}_1} + e^{\mathcal{C}_2})$.*

**Proof:** First, let $\varphi_1$ and $\varphi_2$ be optimal for the sub-experiments. By setting $\alpha := \frac{e^{\mathcal{C}_1}}{e^{\mathcal{C}_1}+e^{\mathcal{C}_2}}$ we have $\varphi = \alpha\overline{\varphi}_1 + (1-\alpha)\overline{\varphi}_2$. We write $\overline{f}_\theta$ for the densities of the whole experiment and $f_\theta$ for the densities on the sub-experiments. Thus

$$\begin{aligned} p_\varphi(x) &= \alpha p_{\overline{\varphi}_1}(x) + (1-\alpha)p_{\overline{\varphi}_2}(x) \\ &= \begin{cases} \alpha p_{\overline{\varphi}_1}(x) & \text{if} \quad x \in X_1 \\ (1-\alpha)p_{\overline{\varphi}_2}(x) & \text{if} \quad x \in X_2 \end{cases} . \end{aligned}$$

Therefore we can split $\mathbb{K}$ for $\theta \in \Theta_1$ into

$$\begin{aligned} \mathbb{K}(\overline{P}_\theta, P_\varphi) &= \int_{X_1 \cup X_2} \overline{f}_\theta(x) \log \frac{\overline{f}_\theta(x)}{p_\varphi(x)} \mu(dx) \\ &= \int_{X_1} f_\theta(x) \log \frac{f_\theta(x)}{p_\varphi(x)} \mu(dx) \\ &\quad \text{as } \overline{f}_\theta(x) = 0 \text{ for all } x \in X_2 \text{ if } \theta \in \Theta_1, \\ &= \int_{X_1} f_\theta(x) \log \frac{f_\theta(x)}{\alpha p_{\overline{\varphi}_1}(x)} \mu(dx) \\ &= \log(\frac{1}{\alpha}) \cdot \int_{X_1} f_\theta(x)\mu(dx) + \int_{X_1} f_\theta(x) \log \frac{f_\theta(x)}{p_{\varphi_1}(x)} \mu(dx) \\ &= \log(\frac{1}{\alpha}) + \mathbb{K}(P_\theta, P_{\varphi_1}) \\ &\leq \log(\frac{1}{\alpha}) + \mathcal{C}_1, \end{aligned}$$

with equality $\varphi$–almost surely. For $\theta \in \Theta_2$ a similar calculation shows that

$$\mathbb{K}(\overline{P}_\theta, P_\varphi) \leq \log(\frac{1}{1-\alpha}) + \mathcal{C}_2,$$

also with equality $\varphi$–almost surely. But as

$$\log(\frac{1}{1-\alpha}) + \mathcal{C}_2 = \log(\mathrm{e}^{\mathcal{C}_1} + \mathrm{e}^{\mathcal{C}_2}) = \log(\frac{1}{\alpha}) + \mathcal{C}_1,$$

Theorem 2 implies that $\varphi$ is optimal for $\mathcal{P}$. It also implies (3). Now, we want to show that those priors defined in (2) fulfil the condition of Theorem 2 on the level of the sub-experiments. For $\theta \in \Theta_1$ we have by the same calculation as above that

$$\mathbb{K}(\overline{P}_\theta, P_\varphi) = \log(\frac{1}{\alpha}) + \mathbb{K}(P_\theta, P_{\varphi_1}).$$

But because $\mathbb{K}(\overline{P}_\theta, P_\varphi) \leq \mathcal{C}$ and $\mathbb{K}(\overline{P}_\theta, P_\varphi) = \mathcal{C}$ $\varphi$–almost surely, we have that

$$\mathbb{K}(P_\theta, P_{\varphi_1}) \leq \mathcal{C} - \log(\frac{1}{\alpha})$$

with equality $\varphi$– and therefore also $\varphi_1$–almost surely. A similar argumentation holds for the second sub-experiment, which concludes the proof. ■

We can summarize the theorem by saying that in an experiment with orthogonal sub-experiments, a prior is globally optimal if and only if it is locally optimal on these sub-experiments. It gives us also a link of the probability of such a sub-experiment with its capacity:

$$\varphi(\Theta_i) = \frac{\mathrm{e}^{\mathcal{C}_i}}{\mathrm{e}^{\mathcal{C}_1} + \mathrm{e}^{\mathcal{C}_2}} = \frac{\mathrm{e}^{\mathcal{C}_i}}{\mathrm{e}^{\mathcal{C}}} \quad \text{or in other words} \quad \log\varphi(\Theta_i) = \mathcal{C}_i - \mathcal{C}.$$

These considerations lead to a heuristic argument for our main result of the next section: It is well known that under some regularity assumptions the probability measures in a product sequence of experiments become more and more mutually singular. So we have by any splitting of the experiment into sub-experiments asymptotically an orthogonal structure as in the above theorem. Thus we should have a tendency of optimal priors to reveal the described structure. Clarke and Barron showed in in their 1994 paper that the asymptotic capacity of any suitable experiment is given by its weight under Jeffreys' prior, and consequently, we have asymptotically that any sequence of optimal priors must assign the same weights to sub-experiments as Jeffreys' prior does.

## 4. CONVERGENCE TO JEFFREYS' PRIOR

Before we state and prove our result rigorously, let us fix some notation. We will denote the product experiment by $\mathcal{P}^n := (P^n_\theta)_{\theta\in\Theta}$ with $P^n_\theta \in \mathrm{Prob}(\mathrm{X}^n, \mathcal{X}^n)$. It is generated by independent and identical repetition of $\mathcal{P}$. Each prior $\varphi \in \mathrm{Prob}\Theta$ induces a mixture $P_\varphi^n := \int_\Theta P_\theta^n \varphi(d\theta)$. As $\mathcal{P}^n$ is dominated by $\mu^n$ we have densities $f_\theta^n$ and $p_\varphi^n$, respectively. $\mathcal{I}_n$ and $\mathcal{C}_n$ denote the information rate over and the capacity of $\mathcal{P}^n$. Further the notation $d_n := \frac{d}{2}\log\frac{n}{2\pi e}$ will be used for the rate term of Clarke and Barron's asymptotic results. We will use the same symbol (e.g. $\lambda$) for a prior with a Lebesgue density and for its density to reduce the inflation of symbols.

In order to use the results of Clarke and Barron assume that our compact, dominated and uniformly integrable experiment $\mathcal{P}$ fulfils the following additional conditions:

**Conditions 1:**

1. *The compact parameter space* $\Theta$ *has non–void interior:* $\mathrm{int}\Theta \neq \emptyset$.

2. *The density* $f_\theta : \mathrm{X} \longrightarrow \mathbb{R}$ *is twice continuously differentiable in* $\theta$ *for* $P_\theta$*–almost all* $x \in \mathrm{X}$*, and there is a* $\delta = \delta(\theta)$ *so that for each* $j,k = 1,\ldots,d$ *the expectation*

$$\int_{\mathrm{X}} \sup_{\{\theta' : \|\theta'-\theta\|<\delta\}} \left| \frac{\partial^2}{\partial\theta_j'\partial\theta_k'} \log f_{\theta'}(x) \right|^2 P_\theta(dx)$$

*is finite and continuous as a function of* $\theta$*, and for each* $j = 1,\ldots,d$ *the expectation*

$$\int_{\mathrm{X}} \left| \frac{\partial}{\partial\theta_j} \log f_\theta(x) \right|^{2+\xi} P_\theta(dx)$$

*is finite and a continuous function of* $\theta \in \Theta$ *for a* $\xi > 0$.

3. *The Fisher Information matrix* $I$ *is positive definite and coincides with the second derivative matrix* $J$ *of the* $\mathbb{K}$*–distance, i.e.*

$$\begin{aligned} I(\theta) &:= \left( \int_{\mathrm{X}} \frac{\partial}{\partial\theta_j} \log f_\theta(x) \frac{\partial}{\partial\theta_k} \log f_\theta(x) P_\theta(dx) \right)_{j,k=1,\ldots,d} \\ &= \left( \frac{\partial^2}{\partial\theta_j'\partial\theta_k'} \mathbb{K}(P_\theta, P_{\theta'})|_{\theta'=\theta} \right)_{j,k=1,\ldots,d} \\ &=: J(\theta). \end{aligned}$$

4. *The parametrization* $\Pi : \Theta \longrightarrow \mathcal{P}, \theta \mapsto P_\theta$ *is one–to–one.*

These conditions in particular imply that $I : \Theta \to \boldsymbol{R}^{d \times d}$ is positive definite and continuous and thus bounded on $\Theta$, so that the integral over the square root of the Fisher information always exists. Thus Jeffreys' prior is well defined in our situation. Further, Condition 3 guarantees that differentiation and integration may be interchanged in this special case. It also implies that the parametrization $\Pi$ is continuous, which, together with the bijectivity, means that $\Pi$ is a homeomorphism, due to the compactness. Condition 3 together with the characteristics of $\Pi$ imply the consistency of the posterior distribution $\varphi(\cdot|X^n)$. This means that $\varphi(\cdot|X^n)$ concentrates at the true parameter at a fast enough rate. For more details see Clarke and Barron (1990 and 1994).

Finally, denote Jeffreys' prior by $w^*$. It is defined by

$$w^*(A) := \frac{\int_A \sqrt{\det I(\theta)}\, d\theta}{\int_\Theta \sqrt{\det I(\theta)}\, d\theta}$$

for all Lebesgue measurable $A \subseteq \Theta$. Referring to Clarke and Barron (1994), Theorem 1, Jeffreys' prior has the following asymptotics for its transmission rate:

$$\lim_{n \to \infty} \left| \mathcal{I}_n(w^*) - d_n \; - \; \log \int_\Theta \sqrt{\det I(\theta)} \; d\theta \right| = 0, \tag{1}$$

which implies its asymptotic optimality, because for the capacity $\mathcal{C}_n$ we have

$$\lim_{n \to \infty} \left| \mathcal{C}_n - d_n \; - \; \log \int_\Theta \sqrt{\det I(\theta)} \; d\theta \right| = 0. \tag{2}$$

We can now formalize the intuitive idea given in the previous section and state our main result.

**Theorem 5:** *Let $\mathcal{P}$ be a compact, $\mu$–dominated and uniformly integrable experiment which fulfils conditions 1. Let $\mathcal{P}^n$ denote the nth–fold product experiment generated by identical and independent repetition of $\mathcal{P}$. Then if $(\varphi_n)_{n \in \boldsymbol{N}} \subset \mathrm{Prob}\Theta$ is a sequence of priors with $\varphi_n$ being optimal for $\mathcal{P}^n$, i.e. $\mathcal{I}_n(\varphi_n) = \mathcal{C}_n$, then $(\varphi_n)_{n \in \boldsymbol{N}}$ converges weakly to Jeffreys' prior $w^* \in \mathrm{Prob}\Theta$:*

$$\varphi_n \overset{n \to \infty}{\Longrightarrow} w^*.$$

All experiments included in Example 1 fulfil the assumptions and can also serve as (theoretic) examples here. But as the exact analytical derivation of Shannon optimal priors is very difficult, even in the case of exponential family experiments, we can only give some numerical examples. They are presented below, in the last section.

Before we begin the proof of this theorem we make some preparatory remarks. Prohorov's Theorem (e.g. see Dudley (1989), Theorem 11.5.4]) guarantees the existence of a weakly convergent subsequence $(\varphi_{n_k})_{k \in N}$ as Prob$\Theta$ is compact. It is enough to show that $w^*$ and an arbitrary limit point $\varphi^*$ coincide on a $\pi$–system[1] which generates the Borel $\sigma$–algebra, c.f. Williams (1991), Lemma 1.6. Suitable to the problem of weak convergence is the $\pi$–system of closed $\varphi^*$–continuity sets. These sets $B$ have a topological boundary with $\varphi^*(\partial B) = 0$.

**Proof of the Theorem:** Let $(\varphi_n)_{n \in N}$ be a sequence of $n$–optimal priors, and let $\varphi^* := \lim_{k \to \infty} \varphi_{n_k}$ be any weak limit point of it. Let $\Theta_1 \subseteq \Theta$ be a closed $\varphi^*$–continuity set and write $\Theta_2 := \Theta \backslash \Theta_1$ for the complement, which is also a $\varphi^*$–continuity set.

Let $\mathcal{C} := \log \int_\Theta \sqrt{I(\theta)}\, d\theta$ denote the asymptotic capacity of Equation (2). Analogously, we write $\mathcal{C}_1 := \log \int_{\Theta_1} \sqrt{I(\theta)} d\theta > 0$ and $\mathcal{C}_2 := \log \int_{\Theta_2} \sqrt{I(\theta)} d\theta > 0$. Further, we write $\mathcal{C}_{n_k}$ for the capacity of the $n_k$-fold product experiment, and $\mathcal{C}_1^{n_k}$, respectively $\mathcal{C}_2^{n_k}$ for the capacity of the sub-experiments with parameter space $\Theta_1$ and $\Theta_2$.

We can rewrite the priors by

$$\varphi_{n_k} = \varphi_{n_k}(\Theta_1)\overline{\varphi}_{n_k}^{(1)} + \varphi_{n_k}(\Theta_2)\overline{\varphi}_{n_k}^{(2)},$$

with

$$\varphi_{n_k}^{(1)} := \frac{\varphi_{n_k}|_{\Theta_1}}{\varphi_{n_k}(\Theta_1)} \quad \text{and} \quad \varphi_{n_k}^{(2)} := \frac{\varphi_{n_k}|_{\Theta_2}}{\varphi_{n_k}(\Theta_2)}.$$

Again, we write $\overline{\varphi}_{n_k}^{(1)}$ and $\overline{\varphi}_{n_k}^{(2)}$ for the priors extended by zero to the full parameter set $\Theta$.

If $x_{n_k} \in X^{n_k}$ then

$$\begin{aligned} p_{\varphi_{n_k}}^{n_k}(x_{n_k}) &= \varphi_{n_k}(\Theta_1) p_{\overline{\varphi}_{n_k}^{(1)}}^{n_k}(x_{n_k}) + \underbrace{\varphi_{n_k}(\Theta_2) p_{\overline{\varphi}_{n_k}^{(2)}}^{n_k}(x_{n_k})}_{\geq 0} \\ &\geq \varphi_{n_k}(\Theta_1) p_{\overline{\varphi}_{n_k}^{(1)}}^{n_k}(x_{n_k}). \end{aligned}$$

[1] i.e. a system of sets stable under finite intersection.

Therefore we have

$$\begin{aligned}
\mathbb{K}(P_\theta^{n_k}, P_{\varphi_{n_k}}^{n_k}) &\leq \int_{X^{n_k}} f_\theta^{n_k}(x_{n_k}) \log \frac{f_\theta^{n_k}(x_{n_k})}{\varphi_{n_k}(\Theta_1) p^{n_k}_{\overline{\varphi}^{(1)}_{n_k}}(x_{n_k})} \mu^{n_k}(dx_{n_k}) \\
&= \int_{X^{n_k}} f_\theta^{n_k}(x_{n_k}) \log \frac{1}{\varphi_{n_k}(\Theta_1)} \mu^{n_k}(dx_{n_k}) + \int_{X^{n_k}} f_\theta^{n_k}(x_{n_k}) \log \frac{f_\theta^{n_k}(x_{n_k})}{p^{n_k}_{\overline{\varphi}^{(1)}_{n_k}}(x_{n_k})} \mu^{n_k}(dx_{n_k}) \\
&= \log \frac{1}{\varphi_{n_k}(\Theta_1)} + \mathbb{K}(P_\theta^{n_k}, P^{n_k}_{\overline{\varphi}^{(1)}_{n_k}}).
\end{aligned}$$

This inequality leads to the following:

$$\begin{aligned}
&\int_{\Theta_1} \mathbb{K}(P_\theta^{n_k}, P_{\varphi_{n_k}}^{n_k}) \varphi_{n_k}(d\theta) - \varphi_{n_k}(\Theta_1) d_{n_k} \\
\leq\ &\int_{\Theta_1} \mathbb{K}(P_\theta^{n_k}, P^{n_k}_{\overline{\varphi}^{(1)}_{n_k}}) \varphi_{n_k}(d\theta) - \varphi_{n_k}(\Theta_1) \log \varphi_{n_k}(\Theta_1) - \varphi_{n_k}(\Theta_1) d_{n_k}. \qquad (3)
\end{aligned}$$

We have $P_{\varphi^{(1)}_{n_k}}(A) = P_{\overline{\varphi}^{(1)}_{n_k}}(A)$ for any measurable set $A \subseteq \mathcal{X}^{n_k}$, and therefore Theorem 1 implies that

$$\begin{aligned}
\int_{\Theta_1} \mathbb{K}(P_\theta^{n_k}, P^{n_k}_{\overline{\varphi}^{(1)}_{n_k}}) \varphi_{n_k}(d\theta) &= \varphi_{n_k}(\Theta_1) \int_{\Theta_1} \mathbb{K}(P_\theta^{n_k}, P^{n_k}_{\varphi^{(1)}_{n_k}}) \varphi^{(1)}_{n_k}(d\theta) \\
&\overset{\text{by } 1}{=} \varphi_{n_k}(\Theta_1)\, \mathcal{I}_1^{n_k}(\varphi^{(1)}_{n_k}) \\
&\leq \varphi_{n_k}(\Theta_1)\, \mathcal{C}_1^{n_k}. \qquad (4)
\end{aligned}$$

So we have combining both steps (3) and (4)

$$\begin{aligned}
&\int_{\Theta_1} \mathbb{K}(P_\theta^{n_k}, P_{\varphi_{n_k}}^{n_k}) \varphi_{n_k}(d\theta) - \varphi_{n_k}(\Theta_1) d_{n_k} \\
\leq\ &\varphi_{n_k}(\Theta_1) \big(\mathcal{C}_1^{n_k} - d_{n_k}\big) - \varphi_{n_k}(\Theta_1) \log \varphi_{n_k}(\Theta_1) \overset{k\to\infty}{\longrightarrow} \varphi^*(\Theta_1)\big(\mathcal{C}_1 - \log \varphi^*(\Theta_1)\big), \qquad (5)
\end{aligned}$$

by Equation (2). Together with Theorem 2 this equation also implies that

$$\int_{\Theta_1} \mathbb{K}(P_\theta^{n_k}, P_{\varphi_{n_k}}^{n_k}) \varphi_{n_k}(d\theta) - \varphi_{n_k}(\Theta_1) d_{n_k} \overset{k\to\infty}{\longrightarrow} \varphi^*(\Theta_1)\mathcal{C}.$$

These calculations lead to the following estimation

$$\varphi^*(\Theta_1)\mathcal{C} \leq \varphi^*(\Theta_1)\big(\mathcal{C}_1 - \log \varphi^*(\Theta_1)\big). \qquad (6)$$

We now see that $\varphi^*$ cannot be a Dirac measure: If we have $\varphi^* = \delta_{\theta_0}$ for a $\theta_0 \in \Theta$, and if we take a closed $w^*$–continuous $\Theta_1$ such that $0 < w^*(\Theta_1) < 1$ and $\theta_0 \in \mathrm{int}\Theta_1$, then Inequality (6) will reduce to

$$\mathcal{C} \leq \mathcal{C}_1.$$

This implies $\mathcal{C} = \mathcal{C}_1$ and $\mathcal{C}_2 = \mathcal{C}(\Theta_2) = \mathcal{C}(\Theta_2 \cup \partial\Theta_2) = 0$, taking into account that $\partial\Theta_2$ does not contribute to the asymptotic capacity. This is, of course, a contradiction, as we assumed $w^*(\Theta_2) > 0$, so that by Clarke and Barron (1994), Theorem 1, (Equations (1) and (2)) we must have $\mathcal{C}_2 = \mathcal{C}(\Theta_2 \cup \partial\Theta_2) > 0$.

Therefore, we may assume $\Theta_1$ to be a closed $\varphi^*$–continuity set with $0 < \varphi^*(\Theta_1) < 1$ and with $0 < w^*(\Theta_1) < 1$ in the sequel. That is why we have from Equation (6)

$$\log \varphi^*(\Theta_1) \leq \mathcal{C}_1 - \mathcal{C} = \log \frac{\int_{\Theta_1} \sqrt{\det I(\theta)}\, d\theta}{\int_{\Theta} \sqrt{\det I(\theta)}\, d\theta} = \log w^*(\Theta_1) \tag{7}$$

for all closed $\varphi^*$–continuity sets $\Theta_1 \subset \Theta$ with $0 < \varphi^*(\Theta_1), w^*(\Theta_1) < 1$. This exactly means

$$\varphi^*(\Theta_1) \leq \mathrm{e}^{\mathcal{C}_1 - \mathcal{C}} = w^*(\Theta_1) \tag{8}$$

for all closed $\varphi^*$–continuity sets $\Theta_1 \subset \Theta$ with $0 < \varphi^*(\Theta_1), w^*(\Theta_1) < 1$. For the complementary set $\Theta_2$ we have

$$\varphi^*(\Theta_2) \geq w^*(\Theta_2).$$

Let $(A_i)_{i \in \mathbb{N}}$ be a sequence of closed $\varphi^*$–continuity sets with $A_i \subset A_{i+1} \subset \Theta_2$ for all $i \in \mathbb{N}$ and $\bigcup_{i \in \mathbb{N}} A_i = \Theta_2$. Then we have for each $i \in \mathbb{N}$, as $i \to \infty$, that

$$\begin{tikzcd}
\varphi^*(\Theta\backslash A_i) \arrow[r, phantom, "\geq"] \arrow[d] & w^*(\Theta\backslash A_i) \arrow[d] \\
\varphi^*(\Theta_1) \arrow[r, phantom, "\geq"] & w^*(\Theta_1).
\end{tikzcd}$$

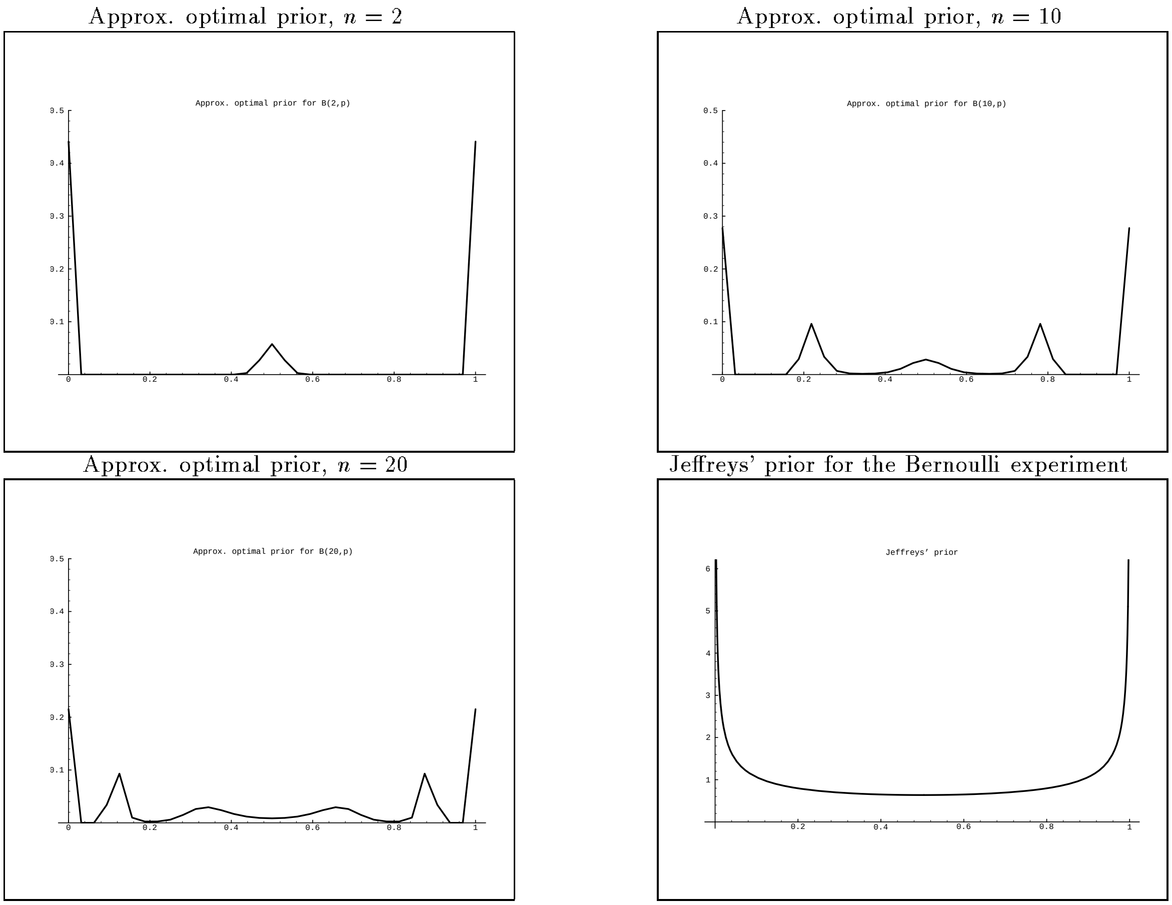
This proves $\varphi^*(\Theta_1) = w^*(\Theta_1)$ for all closed $\varphi^*$–continuity sets $\Theta_1 \subset \Theta$ with $0 < \varphi^*(\Theta_1), w^*(\Theta_1) < 1$. The equality of $\varphi^*$ and $w^*$ in the remaining cases of zero– respectively one–sets follows analogously by the continuity properties of measures. This completes the proof, because $\varphi^*$ and $w^*$ coincide on the $\pi$–system of all closed $\varphi^*$–continuity sets. As $\varphi^*$ is an arbitrary limit point of the sequence $(\varphi_n)_{n \in \mathbb{N}}$, the assertion of the theorem follows:

$$\lim_{n \to \infty} \varphi_n = \lim_{k \to \infty} \varphi_{n_k} = w^*.$$

■

## 5. CONCLUSION

The figure below approximately shows the effect of Theorem 5 for the Binomial family. With $n$ being the possible outcomes of the experiment, the first three diagrams display (approximative) optimal priors (discrete points joined by lines) calculated by the algorithm of Arimoto and Blahut for a fixed discretization (33 points) of the parameter interval. The last diagram shows the density of Jeffreys' prior (for the complete parameter interval) for the Bernoulli experiment. It should be kept in mind that the convergence is in distribution and not pointwise.



Approx. optimal prior, $n = 2$

Approx. optimal prior, $n = 10$

Approx. optimal prior, $n = 20$

Jeffreys' prior for the Bernoulli experiment

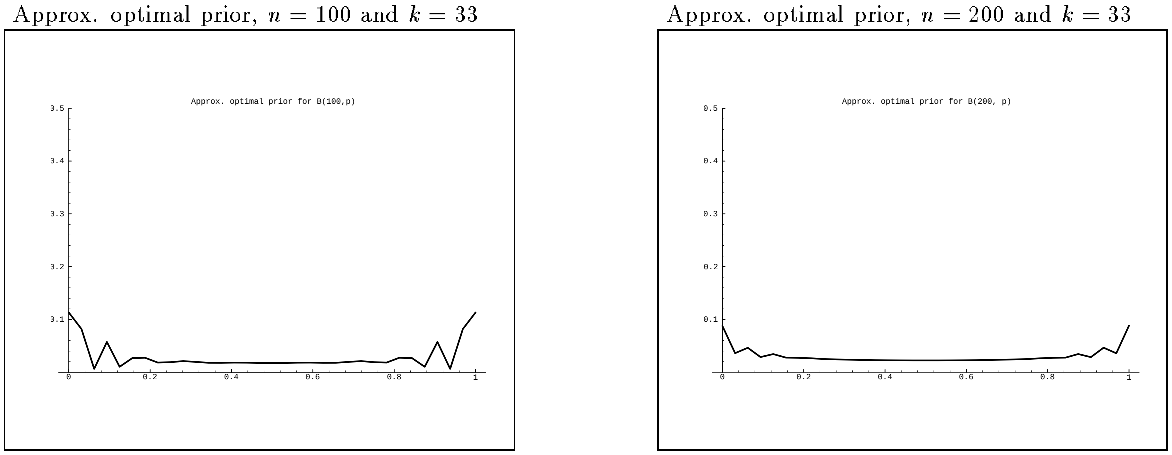
Figure 1: Approximative weak convergence of optimal priors to Jeffreys' prior for the Binomial family

Least favourable priors are often of discrete support, as they are in the present example. In general, for a finite sample space $X$, those optimal priors $\varphi$ with $|\text{supp } \varphi| = |X|$ are the extreme points of the convex set of all optimal priors. So, all least favourable priors are mixtures of discrete optimal priors in this situation. This property is very unappealing to the Bayesian statistician, see e.g.

Bernardo (1994), and it is a major advantage of Bernardo's reference priors that they overcome these difficulties. Theorem 5 links both pieces of theory together: For large sample sizes the possibly discrete optimal priors reveal more and more of the reference prior's structure, giving both parts an additional theoretic justification.

In combination with Clarke and Barron's result (1994), Theorem 5 substantiates Bernardo's practically well approved reference prior approach to the problem of objectivistic least favourable priors in the given context. But it has more than this theoretic implication. Finding exact least informative priors for a given statistical experiment is an in general intractable problem in finite samples, since it involves optimization over infinite–dimensional spaces. The often discrete structure might be somtimes of some help, but normally only numerical solutions will be available in praxis. See e.g. Spall and Hill (1990), or, Arimoto (1972) and Blahut (1972) for numerical approaches. Unfortunately, even in the case of an exponential family experiment these numerics are very expensive, especially for large sample sizes. Theorem 5 leads to an approximative bypass of these problems. For sample sizes that are large enough, the ready at hand Jeffreys' prior is a good approximation, avoiding costly calculations.

The following problem with the numerical calculation of optimal priors must be noted: All reasonable numerical algorithms (known to the author) can only deal with a discretized parameter space (i.e. a finite subexperiment). But for large sample sizes optimal priors on finite parameter sets reveal a different asymptotic behavior than priors on a 'continuous' parameter set. This can be proved, for instance, with the methods used for Theorem 5. Keeping in mind that the 'asymptotic capacity' of an experiment with parameter space $\Theta = \{\theta_1, \dots, \theta_k\}$, with $k \in I\!N$, is $\log k$, it follows that any sequence of optimal priors converges to the uniform distribution on $\Theta$. The following two diagrams together with the previous figure suggest that for a finite subfamily optimal priors tend to Jeffreys' prior (in the sense of mass distribution) as long as $k > n$, then showing their real asymptotic behavior in the long run, for $n \gg k$.


Approx. optimal prior, $n = 100$ and $k = 33$

Approx. optimal prior, $n = 200$ and $k = 33$

Figure 2: Convergence of optimal priors on a finite parameter set to the uniform distribution

## ACKNOWLEDGEMENTS

At this point, the author would like to express his thanks to Prof. H. v. Weizsäcker and his group for many valuable discussions on the subject. I would also like to thank the referees and editors for pointing out related literature and other constructive comments.

# REFERENCES

Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Information Theory* **18**, 14–20.

Blahut, R.E. (1972). Computation of channel capacity and rate distortion functions. *IEEE Trans. Information Theory* **18**, 460–473.

Berger, J. and Bernardo, J.M. (1992). On the development of the reference prior method. *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith, eds.). Oxford: University Press, 35–60 (with discussion).

Berger, J., Bernardo, J.M. and Mendoza, M. (1989). On priors that maximize expected information. *Recent Developments in Statistics and Their Applications* (J.P. Klein and J.C. Lee, eds.). Seoul: Freedom Academy Publishing. 1–20.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113–147 (with discussion).

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian theory.* Chichester: Wiley and Sons.

Clarke, B.S. and Barron. A.R. (1990). Information–theoretic asymptotics of Bayes methods. *IEEE Information Theory* **36**, 453–471.

Clarke, B.S. and Barron, A.R. (1994). Jeffreys prior is asymptotically least favorable under entropy risk. *J. of Stat. Planning and Inference* **41**, North–Holland, 37–60.

Gosh, J.K. and Mukerjee, R. (1992). Non–informative priors. *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith, eds.). Oxford: University Press, 195–210 (with discussion).

Dudley, R.M. (1989). *Real analysis and probability.* Pacific Grove, California: Wadsworth and Brooks/Cole.

Gallager, R.G. (1968). *Information theory and reliable communication.* New York: Wiley and Sons.

Haussler, D. (1995). *A general minimax result for relative entropy.* Preprint. Santa Cruz: Department of Statistics, University of California.

Jeffreys, H. (1961/1983). *Theory of probability* (3rd. ed.). Oxford: Clarendon Press.

Krob, J. (1992). *Kapazität statistischer Experimente.* Ph.D. Dissertation. Kaiserslautern: Department of Mathematics, University of Kaiserslautern.

Krob, J. and Scholl, H. (1997). A general minimax result for the Kullback Leibler Bayes risk. *Economic quality control*, to appear.

Lindley, D.V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Stat.* **27**. 986–1005.

Meyer, P.–A. (1966). *Probabilités et potential.* Paris: Maison d'edition Hermann.

Shannon, C.E. and Weaver, W. (1949). *The mathematical theory of communication.* Urbana: University of Illinois Press.

Spall, J.C. and Hill, S.D. (1990). Least informative Bayesian prior distributions for finite samples based on information theory. *IEEE Trans. Automatic Control* **35**, 580–583.

Topsøe, F. (1974). *Informationstheorie.* Stuttgart: B.G. Teubner Verlag.

Williams, D. (1991). *Probability with martingales.* Cambridge: University Press.

Zhang, Z. (1994). *Discrete noninformative priors.* Ph.D. Dissertation. New Haven: Department of Statistics, Yale University.