Bayesian Analysis (2014) **9**, Number 2, pp. 331–370

On Divergence Measures Leading to Jeffreys and Other Reference Priors

Ruitao Liu * †, Arijit Chakrabarti ‡, Tapas Samanta §, Jayanta K. Ghosh ¶ and Malay Ghosh ‖

Abstract. The paper presents new measures of divergence between prior and posterior which are maximized by the Jeffreys prior. We provide two methods for proving this, one of which provides an easy to verify sufficient condition. We use such divergences to measure information in a prior and also obtain new objective priors outside the class of Bernardo's reference priors.

Keywords: α-divergences, Jeffreys prior, Reference prior

1 Introduction

Reference priors, originally introduced by Bernardo (1979), and further developed by Berger and Bernardo (1989, 1992a, 1992b), are quite a popular choice for objective priors. In the construction of reference priors, the parameters are arranged in increasing order of importance, and a step by step algorithm is used to construct conditional priors given the parameters appearing earlier in the order. It is perhaps not unfair to say that these priors are descendents of the Jeffreys prior with some of the standard objections to that prior well taken care of. Moreover, when all the parameters are given equal importance, the reference priors turn out to be the Jeffreys prior in regular problems. For more details on the above, see, for example, the articles referred to earlier in this paragraph. An important recent contribution in this area is Berger, Bernardo and Sun (2009).

A major interest in the Bayesian literature is in finding objective priors which are primarily algorithmic in nature, depending only on the model or equivalently the likelihood function. Bernardo (1979) proposed such an algorithmic method based on maximizing a particular divergence measure between prior and posterior (see (2) below), supported on a given compact parameter space. Bernardo appealed to Shannon's notion of missing information in a channel (Shannon 1948), to justify the divergence and the resulting priors, but a more direct justification for maximizing any divergence between prior and posterior may be given as follows. If the prior is a point mass, the most informative case, the posterior will also be a point mass no matter what the data is. So the divergence is zero. The more informative the prior is the less we expect a given data to change it

*Department of Statistics and Actuarial Science, University of Iowa, United States
†ACT, Inc, United States, ruitao.liu@act.org
‡Applied Statistics Unit, Indian Statistical Institute, India, arc@isical.ac.in
§Applied Statistics Unit, Indian Statistical Institute, India, tapas@isical.ac.in
¶Department of Statistics, Purdue University, United States, ghosh@stat.purdue.edu
‖Department of Statistics, University of Florida, United States, ghoshm@stat.ufl.edu

 DOI:10.1214/14-BA862

in the course of forming the posterior, leading to a smaller divergence. This is another way of saying the bigger the divergence, the lower the information in the prior or the smaller its influence on the posterior. A measure of the influence of the prior is taken as a measure of information in the prior. For example, it is often routine to examine prior-posterior plots to see if the prior is too informative. We want to investigate what kinds of priors are obtained by maximizing divergences other than the one considered by Bernardo (1979).

Bernardo (1979) considers maximization of a measure based on the Kullback-Leibler divergence (Kullback and Leibler 1951) and Jeffreys prior comes out as the maximizer. This was observed earlier in Ibragimov and Hasminskii (1973) in a different context. Two other principles of constructing objective priors, namely, probability matching and weak limits of a sort of discrete uniforms on a finite set approximating the compact parameter space in the Hellinger metric, also lead to the Jeffreys prior, (Ghosh and Ramamoorthi 2003, Ch.8, Ghosh, Delampady and Samanta 2006, Ch. 5; see also Zhang, 1994, in this context).

We believe that the Jeffreys prior is a basic objective prior and therefore, it is of natural interest to find divergence measures maximized by the Jeffreys prior. One of the main goals of this paper is to come as close as possible to characterizing such divergences. In Section 2 of this paper, we pursue this goal. We fall short of a characterization but provide a sufficient condition for a divergence measure to be maximized by Jeffreys prior.

We consider observations $\boldsymbol{X} = (X_1, X_2, \ldots, X_n) \sim p(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ has prior density $p(\boldsymbol{\theta})$. Let $p(\boldsymbol{\theta}|\boldsymbol{x})$ and $m(\boldsymbol{x})$ denote respectively the posterior density of $\boldsymbol{\theta}$ given $\boldsymbol{x}$ and the marginal density of $\boldsymbol{X}$. In Section 2 we consider a class of divergence measures $D(p(\cdot), p(\cdot|\boldsymbol{x}))$ between the prior and the posterior and the corresponding average divergence

$$J(p) = \int D(p(\cdot), p(\cdot|\boldsymbol{x}))m(\boldsymbol{x})d\boldsymbol{x}.$$

Bernardo (1979) considered the Kullback-Leibler divergence

$$D(p(\cdot), p(\cdot|\boldsymbol{x})) = \int \log\left\{\frac{p(\boldsymbol{\theta}|\boldsymbol{x})}{p(\boldsymbol{\theta})}\right\} p(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}.$$

It is observed that the functional J is expected to be a nice function of the prior only asymptotically (see, e.g., Berger and Bernardo 1989) and this suggests evaluating the divergence using the limiting posterior rather than the finite sample posterior. Thus instead of J we may also consider the functional

$$\hat{J}(p) = \int D(p(\cdot), \hat{p}(\cdot|\boldsymbol{x}))m(\boldsymbol{x})d\boldsymbol{x} \tag{1}$$

where $\hat{p}(\cdot|\boldsymbol{x})$ is the limiting posterior density of $\boldsymbol{\theta}$. This may be taken either as an approximation to J or as an idealized version of J obtained through a limit (avoiding unacceptable discrete maximizers as shown in Berger and Bernardo 1989). From a technical point of view we need to work with $\hat{J}$ to achieve one of our major goals,

namely, understanding in a general way the form of divergences which will lead to the Jeffreys prior upon maximization. In Section 2.1 we describe briefly how the asymptotic maximization is done for the Kullback-Leibler divergence. Section 2.2 deals with a general class of divergence measures. We note that if the divergence measure is invariant under smooth one-to-one transformations of the parameter then the maximizing prior is also invariant. This leads us to consider a general class of divergence measures (see (7)) which are invariant. We find easy to verify sufficient conditions on the divergence under which the average divergence measure $\hat{J}(p)$ is maximized by the Jeffreys prior. It follows as an immediate consequence that for a class of divergence measures, known as α-divergence measures (Amari 1982, 1985, Cressie and Read 1984), the Jeffreys prior is obtained as a maximizer of $\hat{J}(p)$ for $-1 < \alpha < 0$ and $0 < \alpha < 1$. Note that the class of α-divergence measures includes the squared-Hellinger distance ($\alpha = 1/2$) and the Kullback-Leibler divergence in a limiting sense ($\alpha \to 0$). We also prove our results separately for the L_1 and Hellinger distances and obtain Jeffreys prior. The proof for Hellinger distance requires some necessary modification on the argument of our general theorem (Theorem 2.1) for the case of α-divergence with $\alpha = 1/2$. The case for the L_1 distance requires a substantially different treatment and we consider this as a stand-alone case outside the purview of our general theorem that deals with sufficiently "smooth" divergences. Although we maximize $\hat{J}(p)$ instead of $J(p)$, it is expected that the Jeffreys prior will also asymptotically maximize $J(p)$ for all these divergences. For instance, we have verified this for the L_1 distance. A more precise statement of this result is presented in the Appendix.

In Section 3 we present an alternative approach based on the so called "shrinkage argument", due to one of us, which deals with $J(p)$ directly for the α-divergence measures. It is shown that the Jeffreys prior maximizes $J(p)$ for the α-divergence measures for $-1 < \alpha < 0$ and $0 < \alpha < 1$. Section 3.3 deals with an α-divergence measure with $\alpha = -1$, known as chi-square divergence. The resulting prior turns out to be different from Jeffreys prior. This is illustrated through a couple of examples. We explain at the end of Section 2 why we require two methods of proof in Sections 2 and 3 to obtain our results.

After obtaining new divergence measures (other than Kullback-Leibler divergence) that give rise to Jeffreys prior, an immediate natural question is whether we get a satisfactory reference prior if we use such a new divergence measure and proceed to construct the reference prior following the step by step algorithm of Bernardo. It may be recalled that Bernardo (1979) and Berger and Bernardo (1992a, 1992b) proposed a stepwise procedure for finding reference priors when the parameters can be arranged according to their order of importance. In particular, the situation where some of the parameters are nuisance parameters can also be handled using this approach. In Section 4 we consider selection of priors in the presence of nuisance parameters using the class of α-divergence measures considered in Section 2. We briefly outline a derivation of the result without going into the details and give one example. We present this result very briefly, as it is a problem, closely related to the main theme of the paper, which is to be studied more extensively in the future. We would like to mention in this context that an innovative use of the Berger-Bernardo strategy for constructing reference priors to

avoid the marginalization paradox can be found in Fraser et al. (2010). A relatively complete treatment of reference priors in the presence of nuisance parameters can be found in Yuan and Clarke (2004).

Finally, we would like to mention another approach to measuring information in a prior (without going into details), which seems to be related in a sense to the approach described in this article. Bernardo's concern was with estimation problems. An approach to measuring information in a prior for testing problems by associating with it a conceptual equivalent (expected) sample size has been developed by Clarke (1996) and Lin, Pittman and Clarke (2007). A similar approach in the context of clinical trials can be found in Morito, Thall and Mueller (2008, 2010).

2 Jeffreys Prior as a Maximizer of Divergence between Prior and Posterior

Jeffreys prior is one of the widely used objective priors which is defined as

$$\pi(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2},$$

where $I(\boldsymbol{\theta})$ denotes the Fisher information matrix and $|I(\boldsymbol{\theta})|$ denotes its determinant. An important property of this prior is that it is invariant under one-to-one transformations of the parameter $\boldsymbol{\theta}$ and thus does not lead to inconsistencies if applied to different parametrizations of the same problem. This means Jeffreys priors $\pi(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ and $\pi^*(\boldsymbol{\eta})$ for any smooth one-to-one function $\boldsymbol{\eta}(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ are related by the usual Jacobian formula

$$\pi(\boldsymbol{\theta}) = \pi^*(\boldsymbol{\eta}(\boldsymbol{\theta})) \left| \frac{d\boldsymbol{\eta}}{d\boldsymbol{\theta}} \right|.$$

In this section we show how Jeffreys prior can be obtained as a maximizer of divergence between prior and posterior for a class of divergence measures. Section 2.1 considers the Kullback-Leibler divergence. Section 2.2 presents a class of divergence measures which are maximized by the Jeffreys prior. It also presents specific examples of divergences falling within this class.

2.1 Jeffreys Prior as a Maximizer of Kullback-Leibler Divergence between Prior and Posterior

The method presented in this subsection is well known, but we describe it briefly to help later exposition.

Let $\boldsymbol{X} = (X_1, \ldots, X_n) \sim p(\boldsymbol{x}|\boldsymbol{\theta})$, $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}, \boldsymbol{\theta} \in \Theta \subset \mathcal{R}^k$, where $\boldsymbol{\theta}$ has prior density $p(\boldsymbol{\theta})$. Let $p(\boldsymbol{\theta}|\boldsymbol{x})$ and $m(\boldsymbol{x})$ denote respectively the posterior density of $\boldsymbol{\theta}$ given $\boldsymbol{x}$ and the marginal density of $\boldsymbol{X}$. Following Lindley (1956), Bernardo (1979) used the

average Kullback-Leibler divergence between prior and posterior, namely,

$$\begin{aligned} J(p) &= \int_{\mathcal{X}} \left[\int_{\Theta} \log \left\{ \frac{p(\boldsymbol{\theta}|\boldsymbol{x})}{p(\boldsymbol{\theta})} \right\} p(\boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{\theta} \right] m(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int_{\Theta} \left\{ \int_{\mathcal{X}} \left[\int_{\Theta} \log \left\{ \frac{p(\boldsymbol{\theta'}|\boldsymbol{x})}{p(\boldsymbol{\theta'})} \right\} p(\boldsymbol{\theta'}|\boldsymbol{x}) d\boldsymbol{\theta'} \right] p(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x} \right\} p(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \end{aligned} \tag{2}$$

as a measure of information about θ provided by the experiment. This measure was derived from the work of Shannon (1948). Maximizing $J(p)$ with respect to p will mean minimizing information in a prior. We briefly describe below how the (asymptotic) maximization is done.

Fix an increasing sequence of compact rectangles K_i whose union is the whole parameter space Θ. Fix i and consider only priors p_i supported on K_i and let $n \to \infty$.

We assume $X_1, \ldots, X_n$ are i.i.d. and also assume conditions under which the posterior distribution is asymptotically normal in an appropriate sense (see Clarke and Barron 1990, 1994). Then the functional J can be suitably approximated by

$$\begin{aligned} \hat{J}(p_i) &= \int_{K_i} \left\{ \int_{\mathcal{X}} \left[\int_{\mathcal{R}^d} \log \hat{p}_i(\boldsymbol{\theta'}|\boldsymbol{x}) \hat{p}_i(\boldsymbol{\theta'}|\boldsymbol{x})\, d\boldsymbol{\theta'} \right] p(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x} \right\} p_i(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &\quad - \int_{K_i} (\log p_i(\boldsymbol{\theta})) p_i(\boldsymbol{\theta})\, d\boldsymbol{\theta} \end{aligned}$$

where $\hat{p}_i$ is the approximating $N_k(\hat{\boldsymbol{\theta}}_n, I^{-1}(\boldsymbol{\theta})/n)$ density, $\hat{\boldsymbol{\theta}}_n$ being the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$. We now use a well-known fact about the exponent of a multivariate normal density (see, e.g., Ghosh et al. 2006, p.127) and get

$$\hat{J}(p_i) = \frac{k}{2} \log n - \{ \frac{k}{2} \log(2\pi) + \frac{k}{2} \} + \int_{K_i} \log \left\{ \frac{(\det(I(\boldsymbol{\theta})))^{\frac{1}{2}}}{p_i(\boldsymbol{\theta})} \right\} p_i(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{3}$$

The prior that maximizes the above is given by

$$p_i(\boldsymbol{\theta}) = \begin{cases} c_i [\det(I(\boldsymbol{\theta}))]^{1/2} & \text{on } K_i; \\ 0 & \text{elsewhere} \end{cases} \tag{4}$$

where c_i is a normalizing constant such that (4) is a probability density on K_i.

2.2 Divergence Measures Leading to Jeffreys Prior

The measure $J(p)$, considered in Section 2.1, is based on the average Kullback-Leibler divergence between prior and posterior. The question that we try to address in this section is for which kind of divergence measures do the corresponding functionals $J(p)$, when maximized with respect to the prior p, lead to Jeffreys prior. In the next few paragraphs we try to motivate the choice of a general class of divergence measures within which we will restrict our attention for this investigation.

An important property of the (average) Kullback-Leibler divergence $J(p)$, as given in (2), is that it is invariant under smooth one-to-one transformations of the parameter. This is because for a smooth one-to-one function $\boldsymbol{\eta}(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$,

$$\frac{p(\boldsymbol{\theta}|\boldsymbol{x})}{p(\boldsymbol{\theta})} = \frac{p^{\boldsymbol{\eta}}(\boldsymbol{\eta}|\boldsymbol{x})}{p^{\boldsymbol{\eta}}(\boldsymbol{\eta})}$$

and therefore,

$$\begin{aligned} J(p) &= \int_{\mathcal{X}} \int \log \frac{p(\boldsymbol{\theta}|\boldsymbol{x})}{p(\boldsymbol{\theta})} p(\boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{\theta}\, m(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int_{\mathcal{X}} \int \log \frac{p^{\boldsymbol{\eta}}(\boldsymbol{\eta}|\boldsymbol{x})}{p^{\boldsymbol{\eta}}(\boldsymbol{\eta})} p^{\boldsymbol{\eta}}(\boldsymbol{\eta}|\boldsymbol{x}) d\boldsymbol{\eta}\, m(\boldsymbol{x}) d\boldsymbol{x}. \end{aligned} \tag{5}$$

Here $p^{\boldsymbol{\eta}}(\boldsymbol{\eta})$ and $p^{\boldsymbol{\eta}}(\boldsymbol{\eta}|\boldsymbol{x})$ denote the prior and posterior densities of $\boldsymbol{\eta}$. This notion of invariance is the same as Shannon's (Shannon 1948) requirement of invariance under change of dominating measure (see, e.g., Ghosh et al. 2006, p. 124). That $J(p)$ is invariant implies that the maximizing prior, if unique, is also invariant under smooth one-to-one transformations. For uniqueness we need a concavity assumption for the functional J. We simply use the fact that the maximum of $J(p)$ over all $p(\boldsymbol{\theta})$ and the maximum of $J_{\boldsymbol{\eta}}(p^{\boldsymbol{\eta}})$ over all $p^{\boldsymbol{\eta}}$ are the same and the corresponding unique maximizers $p_*(\boldsymbol{\theta})$ and $p_*^{\boldsymbol{\eta}}(\boldsymbol{\eta})$ are related by the usual Jacobian formula

$$p_*(\boldsymbol{\theta}) = p_*^{\boldsymbol{\eta}}(\boldsymbol{\eta}(\boldsymbol{\theta})) \left| \frac{d\boldsymbol{\eta}}{d\boldsymbol{\theta}} \right|.$$

We now consider a divergence measure $D_p(\boldsymbol{x}) = D(p(\cdot), p(\cdot|\boldsymbol{x}))$ between the prior and the posterior and the corresponding functional

$$\begin{aligned} J(p) &= \int_{\mathcal{X}} D_p(\boldsymbol{x}) m(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int \int D_p(\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \end{aligned} \tag{6}$$

Examples of such divergence measures include the Hellinger distance with

$$D_p(\boldsymbol{x}) = \left\{ \int |p^{\frac{1}{2}}(\boldsymbol{\theta}) - p^{\frac{1}{2}}(\boldsymbol{\theta}|\boldsymbol{x})|^2 d\boldsymbol{\theta} \right\}^{\frac{1}{2}}$$

and the L_r-distance, $r > 0$, with

$$D_p(\boldsymbol{x}) = \left\{ \int |p(\boldsymbol{\theta}) - p(\boldsymbol{\theta}|\boldsymbol{x})|^r d\boldsymbol{\theta} \right\}^{\frac{1}{r}}.$$

Our aim is to see for which kind of divergence measures the corresponding functionals in (6) are maximized by the Jeffreys prior, a prior invariant under smoooth one-to-one transformations. Thus we need to restrict ourselves within the class of divergence

measures for which the maximizers of (6) remain invariant under such transformations. Keeping this in mind, we consider a class of divergence measures of the form

$$D_p(\boldsymbol{x}) = \int_\Theta d\left(\frac{p(\boldsymbol{\theta}')}{p(\boldsymbol{\theta}'|\boldsymbol{x})}\right) p(\boldsymbol{\theta}'|\boldsymbol{x}) d\boldsymbol{\theta}' \tag{7}$$

for some function $d(\cdot)$ defined on $(0, \infty)$. The divergence measure (7) with convex $d(\cdot)$ is indeed the same as what was proposed as a measure of divergence by Ali and Silvey (1966); see also Csiszár (1963) and Morimoto (1963). Following the same arguments as used above for the Kullback-Leibler divergence, one can see that the maximizer of the functional (6) using the divergence measure (7) will be invariant under smooth one-to-one transformations. Particular examples are the Kullback-Leibler divergence ($d(u) = -\log(u)$), L_1-distance ($d(u) = |u-1|$) and α-divergences (see (27)) which include the squared Hellinger distance ($\alpha = 1/2$, $d(u) = (\sqrt{u}-1)^2$). It is to be noted that the functional (6) corresponding to the L_r-distance with $r \neq 1$ is not invariant.

We now restrict attention to the class of divergence measures $D_p(x)$ of the form (7). We consider the multiparameter case and the set up of Section 2.1.

As in Section 2.1 we consider a sequence of increasing compact rectangles whose union is the whole of Θ. We fix such a compact rectangle K and consider priors $p(\cdot)$ supported on K. As mentioned in the Introduction (see (1) and the paragraph following it), we replace the posterior $p(\cdot|\boldsymbol{x})$ by its limit $N_k(\hat{\boldsymbol{\theta}}_n, (nI(\boldsymbol{\theta}))^{-1})$ in (7) and (6) to obtain the functional

$$\hat{J}(p) = \int\int \hat{D}_p(\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x}\, p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{8}$$

where

$$\hat{D}_p(\boldsymbol{x}) = \int_K d\left(\frac{p(\boldsymbol{\theta}')}{\hat{p}(\boldsymbol{\theta}'|\boldsymbol{x})}\right) \hat{p}(\boldsymbol{\theta}'|\boldsymbol{x}) d\boldsymbol{\theta}' \tag{9}$$

and $\hat{p}(\cdot|\boldsymbol{x})$ is the $N_k(\hat{\boldsymbol{\theta}}_n, (nI(\boldsymbol{\theta}))^{-1})$ density. We present below a set of sufficient conditions on the function $d(\cdot)$ under which the functional $\hat{J}(p)$ is asymptotically maximized by the Jeffreys prior restricted to K.

We assume that $d(\cdot)$ is of the form

$$d(u) = A + B d_0(u) \tag{10}$$

where A and B are constants and either

$$\text{(i) } d_0(uv) = d_0(u) d_0(v) \quad \text{for all } u, v > 0 \tag{11}$$

or

$$\text{(ii) } d_0(uv) = d_0(u) + d_0(v) \quad \text{for all } u, v > 0. \tag{12}$$

We also make the following assumptions

A(i) For some constant $B_0 > 0$ and for some $u_0 > 0$ and $\alpha < 1$,

$$|d_0(u)| \leq B_0 u^{\alpha}$$

for all $u \geq u_0$.

A(ii) The function $d_0(\cdot)$ has a continuous derivative $d_0'(\cdot)$.

A(iii) The function defined by $d^*(u) = Bd_0(1/u),\ u > 0$, is a concave function of u.

We now have the following result.

Theorem 2.1. *Consider priors $p(\boldsymbol{\theta})$, positive and differentiable on K with continuous partial derivatives. Assume that $I(\boldsymbol{\theta})$ is positive and continuous on K. Then under assumptions A(i) and A(ii), for any $d(\cdot)$ of the form given in (10)-(12), we have*

$$\hat{J}(p) = A + Bd_0(n^{-k/2}) \left[C_0 \int_K d_0 \left(\frac{p(\boldsymbol{\theta})}{|I(\boldsymbol{\theta})|^{1/2}} \right) p(\boldsymbol{\theta}) d\boldsymbol{\theta} + o(1) \right] \tag{13}$$

for some constant $C_0 > 0$ for Case (i), and

$$\hat{J}(p) = A + B_n + B \int_K d_0 \left(\frac{p(\boldsymbol{\theta})}{|I(\boldsymbol{\theta})|^{1/2}} \right) p(\boldsymbol{\theta}) d\boldsymbol{\theta} + o(1) \tag{14}$$

for some constant B_n depending on n for Case (ii).

Further, under assumption A(iii), for both the cases, $\hat{J}(p)$ is asymptotically maximized by the Jeffreys prior restricted to K (unless $d_0(u) \equiv 1/u$).

Proof. Making a change of variable $\boldsymbol{t} = \sqrt{n} I^{1/2}(\boldsymbol{\theta})(\boldsymbol{\theta}' - \hat{\boldsymbol{\theta}}_n)$ in (9), we have

$$\hat{D}_p(\boldsymbol{x}) = A + B \int_{K_n} d_0 \left(\frac{p\left(\hat{\boldsymbol{\theta}}_n + I^{-1/2}(\boldsymbol{\theta})\boldsymbol{t}/\sqrt{n}\right)}{n^{k/2}|I(\boldsymbol{\theta})|^{1/2}\phi_k(\boldsymbol{t})} \right) \phi_k(\boldsymbol{t}) d\boldsymbol{t}, \tag{15}$$

where $K_n = \sqrt{n} I^{1/2}(\boldsymbol{\theta})(K - \hat{\boldsymbol{\theta}}_n)$ and $\phi_k(\boldsymbol{t}) = (2\pi)^{-k/2} \exp(-\frac{1}{2}\sum_{i=1}^k t_i^2)$ is the standard k-dimensional normal density.

We first prove for Case (i). Let

$$\int_{K_n} d_0 \left(\frac{p\left(\hat{\boldsymbol{\theta}}_n + I^{-1/2}(\boldsymbol{\theta})\boldsymbol{t}/\sqrt{n}\right)}{n^{k/2}|I(\boldsymbol{\theta})|^{1/2}\phi_k(\boldsymbol{t})} \right) \phi_k(\boldsymbol{t}) d\boldsymbol{t} = I_1 + I_2 \tag{16}$$

where I_1 and I_2 are the integrals over $S_{1n} = \{\|\boldsymbol{t}\| \leq \sqrt{c \log n}\} \cap K_n$ and $S_{2n} = \{\|\boldsymbol{t}\| > \sqrt{c \log n}\} \cap K_n$ respectively for a constant $c > 0$. Below, many of the statements made are valid only for sufficiently large n but it will not be mentioned explicitly.

Since $I(\boldsymbol{\theta})$ is positive and continuous on K, $p(\theta)$ is continuous on K and $d_0(u)$ is continuous, we have

$$\begin{aligned} |I_2| &\leq |d_0(n^{-k/2})|\cdot|d_0(|I(\boldsymbol{\theta})|^{-1/2})| \\ &\quad \times \int_{S_{2n}} \left|d_0\left(p\left(\hat{\boldsymbol{\theta}}_n + I^{-1/2}(\boldsymbol{\theta})\boldsymbol{t}/\sqrt{n}\right)\right)\right| \cdot |d_0(\phi_k^{-1}(\boldsymbol{t}))|\phi_k(\boldsymbol{t})d\boldsymbol{t} \\ &\leq C_1|d_0(n^{-k/2})| \int_{\{\|\boldsymbol{t}\|>\sqrt{c\log n}\}} |d_0(\phi_k^{-1}(\boldsymbol{t}))|\phi_k(\boldsymbol{t})d\boldsymbol{t} \end{aligned}$$

for some constant $C_1 > 0$. By Assumption A(i), $\|\boldsymbol{t}\| > \sqrt{c\log n}$ implies $|d_0(\phi_k^{-1}(\boldsymbol{t}))| \leq B_0(\phi_k^{-1}(\boldsymbol{t}))^\alpha$ and therefore,

$$\begin{aligned} |I_2| &\leq B_0C_1|d_0(n^{-k/2})| \int_{\{\|\boldsymbol{t}\|>\sqrt{c\log n}\}} (\phi_k(\boldsymbol{t}))^{1-\alpha}d\boldsymbol{t} \\ &\leq C_2|d_0(n^{-k/2})|n^{-(1-\alpha)c/4} \end{aligned} \tag{17}$$

for some constant $C_2 > 0$. The same arguments also lead to

$$\begin{aligned} &\left|\int_{S_{2n}} d_0\left(p(\hat{\boldsymbol{\theta}}_n)n^{-k/2}|I(\boldsymbol{\theta})|^{-1/2}\phi_k^{-1}(\boldsymbol{t})\right)\phi_k^{-1}(\boldsymbol{t})d\boldsymbol{t}\right| \\ \leq\ & C_2|d_0(n^{-k/2})|n^{-(1-\alpha)c/4}. \end{aligned} \tag{18}$$

Now, by the Mean-value Theorem,

$$d_0\left(p\left(\hat{\boldsymbol{\theta}}_n + I^{-1/2}(\boldsymbol{\theta})\frac{\boldsymbol{t}}{\sqrt{n}}\right)\right) = d_0(p(\hat{\boldsymbol{\theta}}_n)) + d_0'(p(\boldsymbol{\theta}_n^*))p'(\boldsymbol{\theta}_n^*)I^{-1/2}(\boldsymbol{\theta})\frac{\boldsymbol{t}}{\sqrt{n}} \tag{19}$$

where $p'(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial\theta_1}p(\boldsymbol{\theta}), \cdots, \frac{\partial}{\partial\theta_k}p(\boldsymbol{\theta})\right)$ is the vector of partial derivatives of $p(\boldsymbol{\theta})$ and $\boldsymbol{\theta}_n^*$ lies between $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\theta}}_n + I^{-1/2}(\boldsymbol{\theta})\frac{\boldsymbol{t}}{\sqrt{n}}$. Therefore,

$$\begin{aligned} I_1 &= \int_{S_{1n}} d_0\left(p(\hat{\boldsymbol{\theta}}_n)n^{-k/2}|I(\boldsymbol{\theta})|^{-1/2}\phi_k^{-1}(\boldsymbol{t})\right)\phi_k^{-1}(\boldsymbol{t})d\boldsymbol{t} \\ &\quad + d_0(n^{-k/2}|I(\boldsymbol{\theta})|^{-1/2}) \int_{S_{1n}} d_0'(p(\boldsymbol{\theta}_n^*))p'(\boldsymbol{\theta}_n^*)I^{-1/2}(\boldsymbol{\theta})\frac{\boldsymbol{t}}{\sqrt{n}}d_0(\phi_k^{-1}(\boldsymbol{t}))\phi_k(\boldsymbol{t})d\boldsymbol{t}. \end{aligned} \tag{20}$$

Since $d_0'(\cdot)$ and $p'(\cdot)$ are continuous and $I(\boldsymbol{\theta})$ is positive and continuous on K, the second term on the right hand side of (20) is bounded above in absolute value by

$$C_3|d_0(n^{-k/2})|\left(\frac{\log n}{n}\right)^{1/2} \int_{\mathcal{R}^k} |d_0(\phi_k^{-1}(\boldsymbol{t}))|\phi_k(\boldsymbol{t})d\boldsymbol{t} \tag{21}$$

for some constant $C_3 > 0$. By Assumption A(i),

$$\int |d_0(\phi_k^{-1}(\boldsymbol{t}))|\phi_k(\boldsymbol{t})d\boldsymbol{t} < \infty$$

and therefore, (21) is bounded above by $C_4\sqrt{\frac{\log n}{n}}$ for some constant $C_4 > 0$. Note that all the constants C_1, C_2, C_3 and C_4 are free of $\boldsymbol{\theta}$ and $\boldsymbol{x}$. Thus we have from (16)-(18) and (20),

$$\begin{aligned} & \int_{K_n} d_0\left(\frac{p\left(\hat{\boldsymbol{\theta}}_n + I^{-1/2}(\boldsymbol{\theta})\boldsymbol{t}/\sqrt{n}\right)}{n^{k/2}|I(\boldsymbol{\theta})|^{1/2}\phi_k(\boldsymbol{t})}\right)\phi_k(\boldsymbol{t})d\boldsymbol{t} \qquad (22) \\ = \; & d_0(n^{-k/2})\left[\left(\int_{K_n} d_0(\phi_k^{-1}(\boldsymbol{t}))\phi_k(\boldsymbol{t})d\boldsymbol{t}\right) d_0\left(p(\hat{\boldsymbol{\theta}}_n)/\sqrt{|I(\boldsymbol{\theta})|}\right) + \zeta_n(\boldsymbol{x},\boldsymbol{\theta})\right] \end{aligned}$$

where

$$|\zeta_n(\boldsymbol{x},\boldsymbol{\theta})| \leq C_5\left[\sqrt{\frac{\log n}{n}} + n^{-(1-\alpha)c/4}\right]$$

for some constant $C_5 > 0$ not depending on $\boldsymbol{x}$ and $\boldsymbol{\theta}$. From (8), (15) and (22), by application of the dominated convergence theorem twice, one can now show that

$$\hat{J}(p) = A + Bd_0(n^{-k/2})\left[C_0\int_K d_0\left(p(\boldsymbol{\theta}/\sqrt{|I(\boldsymbol{\theta})|}\right)p(\boldsymbol{\theta})d\boldsymbol{\theta} + o(1)\right]$$

where $C_0 = \int_{\mathcal{R}^k} d_0(\phi_k^{-1}(\boldsymbol{t}))\phi_k(\boldsymbol{t})d\boldsymbol{t}$. We here use the facts that under any fixed $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}$ almost surely, $\int|d_0(\phi_k^{-1}(\boldsymbol{t}))|\phi_k(\boldsymbol{t})d\boldsymbol{t} < \infty$, $I(\boldsymbol{\theta})$ is positive and continuous on K and continuity of $d(\cdot)$ and $p(\cdot)$. By Assumption A(iii) and an application of Jensen's Inequality, the result is proved.

We now prove for Case (ii). From (12) and (19), we have

$$\begin{aligned} & \int_{K_n} d_0\left(\frac{p\left(\hat{\boldsymbol{\theta}}_n + I^{-1/2}(\boldsymbol{\theta})\boldsymbol{t}/\sqrt{n}\right)}{n^{k/2}|I(\boldsymbol{\theta})|^{1/2}\phi_k(\boldsymbol{t})}\right)\phi_k(\boldsymbol{t})d\boldsymbol{t} \\ = \; & \int_{K_n} d_0\left(\frac{p(\hat{\boldsymbol{\theta}}_n)}{n^{k/2}|I(\boldsymbol{\theta})|^{1/2}\phi_k(\boldsymbol{t})}\right)\phi_k(\boldsymbol{t})d\boldsymbol{t} \\ & + \int_{K_n} d_0'(p(\boldsymbol{\theta}_n^*))p'(\boldsymbol{\theta}_n^*)I^{-1/2}(\boldsymbol{\theta})\frac{\boldsymbol{t}}{\sqrt{n}}\phi_k(\boldsymbol{t})d\boldsymbol{t}. \qquad (23) \end{aligned}$$

The second term in the right hand side of (23) is bounded above in absolute value by

$$C_6 \frac{1}{\sqrt{n}} \int_{\mathcal{R}^k} \|\boldsymbol{t}\| \phi_k(\boldsymbol{t}) d\boldsymbol{t}$$

which is bounded above by $C_7 \frac{1}{\sqrt{n}}$. Here C_6 and C_7 are positive constants not depending on $\boldsymbol{x}$ and $\boldsymbol{\theta}$. Therefore, from (15) and (23), we have

$$\begin{aligned} \hat{D}_p(\boldsymbol{x}) &= A + Bd_0(n^{-k/2}) \int_{K_n} \phi_k(\boldsymbol{t}) d\boldsymbol{t} \\ &+ B \int_{K_n} d_0(\phi_k^{-1}(\boldsymbol{t})) \phi_k(\boldsymbol{t}) d\boldsymbol{t} + Bd_0 \left(\frac{p(\hat{\boldsymbol{\theta}}_n)}{\sqrt{|I(\boldsymbol{\theta})|}} \right) \int_{K_n} \phi_k(\boldsymbol{t}) d\boldsymbol{t} + \zeta_n(\boldsymbol{x}, \boldsymbol{\theta}), \end{aligned}$$

where $|\zeta_n(\boldsymbol{x}, \boldsymbol{\theta})| \leq C_8 \frac{1}{\sqrt{n}}$ for some constant $C_8 > 0$, not depending on $\boldsymbol{x}$ and $\boldsymbol{\theta}$. Therefore, as in Case (i), using the dominated convergence theorem we have

$$\hat{J}(p) = A + B_n + B \int_K d_0 \left(\frac{p(\boldsymbol{\theta})}{\sqrt{|I(\boldsymbol{\theta})|}} \right) p(\boldsymbol{\theta}) d\boldsymbol{\theta} + o(1)$$

where B_n is a constant depending on n. Under assumption A(iii), the result follows from an application of Jensen's Inequalty. Q.E.D.

Remark 2.1. We would like to make some comments about the sufficient conditions presented in Theorem 2.1 above. Consider the functional $\hat{J}(p)$ given by (8) and (9). By a change of variable, we have

$$\hat{D}_p(\boldsymbol{x}) = \int d \left(\frac{p\left(\hat{\boldsymbol{\theta}}_n + I^{-1/2}(\boldsymbol{\theta})\boldsymbol{t}/\sqrt{n}\right)}{n^{k/2}|I(\boldsymbol{\theta})|^{1/2}\phi_k(\boldsymbol{t})} \right) \phi_k(\boldsymbol{t}) d\boldsymbol{t}, \tag{24}$$

$\phi_k(\boldsymbol{t})$ being the standard k-dimensional normal density. Since the integration is with respect to a normal density, for sufficiently well behaved $d(\cdot)$, the contribution to the integral for $\boldsymbol{t}$ outside a set of the form $\{\|\boldsymbol{t}\| \leq \sqrt{c \log n}\}$ will be negligible for appropriate c. Indeed, assuming a condition on the growth of $d(\cdot)$, (24) can be approximated as

$$\hat{D}_p(\boldsymbol{x}) \approx \int d \left(\frac{p(\hat{\boldsymbol{\theta}}_n)}{n^{k/2}|I(\boldsymbol{\theta})|^{1/2}\phi_k(\boldsymbol{t})} \right) \phi_k(\boldsymbol{t}) d\boldsymbol{t} \tag{25}$$

in an appropriate sense, so that the functional $\hat{J}(p)$ may be obtained as

$$\hat{J}(p) = J_n^*(p) + \zeta_n \tag{26}$$

where

$$J_n^*(p) = \int\int d\left(\frac{p(\boldsymbol{\theta})}{n^{k/2}|I(\boldsymbol{\theta})|^{1/2}\phi_k(\boldsymbol{t})}\right)\phi_k(\boldsymbol{t})d\boldsymbol{t}p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

and ζ_n is negligible with respect to $J_n^*(p)$. For example, if we assume that $|d'(u)| \leq u^\lambda$ for some $\lambda < 0$, it also implies Condition A(i) of Theorem 2.1 and may be used to prove (26).

We note that if the function $d^*(u) = d(\frac{1}{u})$ is a concave function of u, by Jensen's Inequality, $J_n^*(p)$ is maximized (with respect to $p(\cdot)$) by the Jeffreys prior restricted to K. But this does not ensure that $\hat{J}(p)$ is also asymptotically maximized by the Jeffreys prior. For example, $J_n^*(p)$ may split into two parts as $J_n^*(p) = J_{1n} + J_{2n}(p)$ where J_{1n} is a term free of $p(\cdot)$ (as in the case of Kullback-Leibler divergence). In general we cannot exclude the possibility that the part of $J_{2n}(p)$, that involves $p(\cdot)$, is of smaller order than ζ_n in which case ζ_n cannot be neglected for the purpose of asymptotic maximization. A natural way to ensure that the asymptotic maximizers of $J_n^*(p)$ and $\hat{J}(p)$ are the same is to assume a form of $d(\cdot)$ such that the integral (24) can be simplified to a form where a term only involving n (such as $d_0(n^{-k/2})$ in the proof of Theorem 2.1) comes out as a multiplicative or an additive factor to an integral involving only $p(\hat{\boldsymbol{\theta}}_n)$, $\sqrt{|I(\boldsymbol{\theta}|}$ and $\phi_k(\boldsymbol{t})$. Our conditions (10)-(12) ensure that this is indeed the case.

Remark 2.2. We have derived Jeffreys prior as an objective prior. Typically the Jeffreys prior would be improper and hence there is a possibility of an improper posterior. For any improper prior with a low-dimensional parameter, one typically requires a sufficient amount of data to have a proper posterior. A rule of thumb that is often valid is that one needs k units of data for k unknown parameters. Moreover, the Jeffreys prior has a proper posterior in many examples. In fact there is a general belief that Jeffreys prior always has a proper posterior under the condition mentioned above except for mixture models. For mixture models the likelihood is rather complicated and even the Jeffreys prior will have an improper posterior if the data do not have samples from all components of the mixture.

We consider below examples of $d(\cdot)$ for which Theorm 2.1 holds.

Example 2.1 (Kullback-Leibler divergence). For the Kullback-Leibler divergence $d(u) = d_0(u) = -\log u$, $u > 0$, which satisfies (12). The other conditions are also satisfied and therefore,

$$\hat{J}(p) = C_n + \int_K \log\left(|I(\boldsymbol{\theta})|^{1/2}/p(\boldsymbol{\theta})\right)p(\boldsymbol{\theta})d\boldsymbol{\theta} + o(1)$$

which is the same as what is obtained in (3).

Example 2.2 (α-divergence). We now consider a class of divergence measures, known as α-divergences (Amari 1982, 1985, Cressie and Read 1984), which are defined by (7)

with

$$d(u) = (\alpha u + (1-\alpha) - u^{\alpha})/(\alpha(1-\alpha)). \tag{27}$$

Particular choices of α correspond to standard divergence measures. The α-divergence smoothly connects the squared Hellinger distance ($\alpha = 1/2$), Kullback-Leibler divergence ($\alpha \to 0$), and Chi-square divergence ($\alpha = -1$). It has the following basic properties:

- Nonnegativity: For any real valued α, the α-divergence is nonnegative, and equal to zero if and only if the prior is identical to the posterior.
- Convexity: The α-divergence is convex with respect to both prior and posterior.
- Continuity: The α-divergence is a continuous function of the real variable α.

A very nice review of the α-divergence can be found in Cichocki and Amari (2010).

For the α-divergence measure $d(\cdot)$ can also be taken to be

$$d(u) = \frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha(1-\alpha)} u^{\alpha},\ u > 0,$$

which satisfies (10) and (11) and also satisfies A(i) if $\alpha < 1$. We consider the cases $-1 < \alpha < 0$ and $0 < \alpha < 1$. In both these cases $d^*(u) = -\frac{1}{\alpha(1-\alpha)} u^{-\alpha}$ is a concave function and by Theorem 2.1

$$\hat{J}(p) = \frac{1}{\alpha(1-\alpha)} - \frac{n^{-\alpha/2}}{\alpha(1-\alpha)} \left[C_0 \int_K \left(\frac{|I(\boldsymbol{\theta})|^{1/2}}{p(\boldsymbol{\theta})} \right)^{-\alpha} p(\boldsymbol{\theta}) d\boldsymbol{\theta} + o(1) \right] \tag{28}$$

which is asymptotically maximized by the Jeffreys prior restricted to K.

For the case $\alpha = -1$, the conditions of the theorem are satisfied but $d_0(u) = \frac{1}{u}$ implies that the term that we maximize with respect to p, viz., $\int d_0\left(p(\boldsymbol{\theta})/|I(\boldsymbol{\theta})|^{1/2}\right) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$, is free of $p(\cdot)$. Thus the approximation of $\hat{J}(p)$ obtained in Theorem 2.1 cannot be used and we need to look at smaller order terms in the expansion of $\hat{J}(p)$ which can be obtained using an asymptotic expansion of posterior (see, e.g., Ghosh 1994, p. 47).

What we have obtained in Theorem 2.1 is based on the leading term $\hat{p}(\boldsymbol{\theta}'|\boldsymbol{x})$ of the expansion of the posterior. We could use such an expansion to get smaller order terms of $\hat{J}(p)$ which could be maximized with respect to $p(\cdot)$ but the calculations would be very messy. We present in Section 3.3 an alternative approach based on the so called "shrinkage argument" due to one of the authors, which makes the derivation much simpler. Interestingly, for the case of $\alpha = -1$, the maximizer of $J(p)$ turns out to be different from the Jeffreys prior.

Example 2.3 (Hellinger distance). Recall that the squared Hellinger distance is given by (7) with $d(u) = (\sqrt{u}-1)^2$ or $d(u) = 2-2\sqrt{u}$ for which Theorem 2.1 is applicable as seen in Example 2.2 taking $\alpha = 1/2$. For the Hellinger distance, Theorem 2.1 is not directly applicable since $J(p)$ corresponding to Hellinger distance is not expressible in

the form (6) where $D_p(\boldsymbol{x})$ is as in (7) for some $d(\cdot)$. Instead, we make a slight change in the definition of $D_p(\boldsymbol{x})$ (and $\hat{D}_p(\boldsymbol{x})$) as given in (7), by taking the square root of the right hand side of (7) with $d(u) = 2 - 2\sqrt{u}$. Again, as in (15), making a change of variable $\boldsymbol{t} = \sqrt{n} I^{1/2}(\boldsymbol{\theta})(\boldsymbol{\theta}' - \hat{\boldsymbol{\theta}}_n)$, we have

$$\hat{D}_p(\boldsymbol{x}) = \left[2 - 2 \int_{K_n} d_0 \left(\frac{p\left(\hat{\boldsymbol{\theta}}_n + I^{-1/2}(\boldsymbol{\theta})\boldsymbol{t}/\sqrt{n}\right)}{n^{k/2}|I(\boldsymbol{\theta})|^{1/2}\phi_k(\boldsymbol{t})} \right) \phi_k(\boldsymbol{t}) d\boldsymbol{t} \right]^{1/2}$$

where $d_0(u) = u^{1/2},\ u > 0$, and K_n and $\phi_k(\boldsymbol{t})$ are as in (15). Then as shown in the proof of Theorem 2.1, Case (i), for any constant $c > 0$,

$$\begin{aligned} & \int_{K_n} d_0 \left(\frac{p\left(\hat{\boldsymbol{\theta}}_n + I^{-1/2}(\boldsymbol{\theta})\boldsymbol{t}/\sqrt{n}\right)}{n^{k/2}|I(\boldsymbol{\theta})|^{1/2}\phi_k(\boldsymbol{t})} \right) \phi_k(\boldsymbol{t}) d\boldsymbol{t} \\ = \ & d_0(n^{-k/2}) \left[\left(\int_{K_n} d_0(\phi_k^{-1}(\boldsymbol{t}))\phi_k(\boldsymbol{t}) d\boldsymbol{t} \right) d_0(p(\hat{\boldsymbol{\theta}}_n)/|I(\boldsymbol{\theta})|^{1/2}) + \xi_n(\boldsymbol{x}, \boldsymbol{\theta}) \right] \end{aligned}$$

where $|\xi_n(\boldsymbol{x}, \boldsymbol{\theta})| \leq C_1 \left[n^{-1/2} (\log n)^{1/2} + n^{-(1-\alpha)c/4} \right]$ for some constant $C_1 > 0$ not depending on $\boldsymbol{x}$ and $\boldsymbol{\theta}$. Therefore,

$$\begin{aligned} \hat{D}_p(\boldsymbol{x}) &= \left[2 - 2n^{-k/4} B_n (p(\hat{\boldsymbol{\theta}}_n)/|I(\boldsymbol{\theta})|^{1/2})^{1/2} + O\left(\frac{(\log n)^{1/2}}{n^{1/2} n^{k/4}} \right) \right]^{1/2} \\ &= \sqrt{2} \left[1 - \frac{1}{2} n^{-k/4} B_n (p(\hat{\boldsymbol{\theta}}_n)/|I(\boldsymbol{\theta})|^{1/2})^{1/2} + O\left(\frac{(\log n)^{1/2}}{n^{1/2} n^{k/4}} \right) + O(n^{-k/2}) \right] \end{aligned}$$

where $B_n = \int_{K_n} d_0(\phi_k^{-1}(\boldsymbol{t}))\phi_k(\boldsymbol{t}) d\boldsymbol{t}$. As $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}$ almost surely, we then have

$$\hat{J}(p) = \sqrt{2} + B n^{-k/4} \int_K g\left(\frac{\sqrt{|I(\boldsymbol{\theta})|}}{p(\boldsymbol{\theta})} \right) p(\boldsymbol{\theta}) d\theta + o(n^{-k/4}), \tag{29}$$

with $g(u) = -u^{-1/2},\ u > 0$ (which is concave), and $B = \frac{1}{2} \int_{\mathcal{R}^k} d_0(\phi_k^{-1}(\boldsymbol{t}))\phi_k(\boldsymbol{t}) d\boldsymbol{t}$. Thus the (asymptotic) maximizer of $\hat{J}(p)$ is the Jeffreys prior restricted to K.

Remark 2.3. We have proved Theorem 2.1 for sufficiently smooth functions $d(\cdot)$. The function $d(u) = |u - 1|$ that corresponds to L_1-distance is not such a smooth function, so the general arguments for a smooth $d(\cdot)$ does not work for L_1-distance. For the L_1-distance, we take a somewhat different route from what is described above for a smooth $d(\cdot)$. We show below that for the L_1-distance, an expansion of $\hat{J}(p)$ similar to that obtained in Theorem 2.1 can be obtained and Jeffreys prior asymptotically maximizes $\hat{J}(p)$.

Proof for L_1***-distance***. For simplicity we consider the case with $k = 1$. We assume that $I(\theta)$ is positive and continuous and consider only positive, continuously differentiable

priors $p(\theta)$ on a fixed compact subset $[a, b]$ of Θ. For the L_1-distance $d(u) = |u-1|$ and therefore, from (9), making a change of variable $t = \sqrt{nI(\theta)}(\theta' - \hat{\theta}_n)$, we have

$$\hat{D}_p(\boldsymbol{x}) = \int \left| \frac{p\left(\hat{\theta}_n + t/\sqrt{nI(\theta)}\right)}{\sqrt{nI(\theta)}} - \phi(t) \right| dt$$

under a fixed θ, where $\phi(t)$ is the standard normal density. We then note that

$$\begin{aligned} & \int \left| \frac{p\left(\hat{\theta}_n + t/\sqrt{nI(\theta)}\right)}{\sqrt{nI(\theta)}} - \phi(t) \right| dt \\ = \; & 2\int_{t\in A} \left[\phi(t) - \frac{p\left(\hat{\theta}_n + t/\sqrt{nI(\theta)}\right)}{\sqrt{nI(\theta)}} \right] dt \\ = \; & 2 - 2\int_{t\in A^c} \phi(t)dt - 2\int_{t\in A} \frac{p\left(\hat{\theta}_n + t/\sqrt{nI(\theta)}\right)}{\sqrt{nI(\theta)}} dt \end{aligned} \tag{30}$$

where

$$\begin{aligned} \{t \in A\} &= \left\{ t : \phi(t) > p\left(\hat{\theta}_n + t/\sqrt{nI(\theta)}\right)/\sqrt{nI(\theta)} \right\} \\ &= \left\{ t : |t| < \left(\log n + \log(I(\theta)/2\pi) - 2\log p\left(\hat{\theta}_n + t/\sqrt{nI(\theta)}\right) \right)^{\frac{1}{2}} \right\}. \end{aligned}$$

Using the Mean-value Theorem we have

$$\int_{t\in A} \frac{p\left(\hat{\theta}_n + t/\sqrt{nI(\theta)}\right)}{\sqrt{nI(\theta)}} dt$$

$$= \int_{t\in A\cap K_n} \frac{p(\hat{\theta}_n)}{\sqrt{nI(\theta)}} dt + \int_{t\in A\cap K_n} \frac{\left(t/\sqrt{nI(\theta)}\right) p'(\theta_n^*)}{\sqrt{nI(\theta)}} dt \tag{31}$$

where $K_n = [\sqrt{nI(\theta)}(a-\hat{\theta}_n), \sqrt{nI(\theta)}(b-\hat{\theta}_n)]$ and θ_n^* lies between $\hat{\theta}_n$ and $\hat{\theta}_n + t/\sqrt{nI(\theta)}$. By our assumptions on $I(\cdot)$ and $p(\cdot)$, and using tail properties of the standard normal distribution, we have

$$\int_{t\in A^c} \phi(t)dt \le C_1(n\log n)^{-1/2} \tag{32}$$

and by the same assumptions,

$$\int_{t\in A\cap K_n} \frac{\left|\left(t/\sqrt{nI(\theta)}\right) p'(\theta_n^*)\right|}{\sqrt{nI(\theta)}} dt \le C_2 \frac{\log n}{n} \tag{33}$$

where C_1 and C_2 are constants not depending on x, θ, n. As $\hat{\theta}_n \to \theta$ almost surely under any fixed θ, using the dominated convergence theorem twice, we have from (30)-(33),

$$\hat{J}(p) = 2 + 4\left(\frac{\log n}{n}\right)^{1/2} \int_a^b g\left(\frac{\sqrt{I(\theta)}}{p(\theta)}\right) p(\theta)d\theta + o\left(\left(\frac{\log n}{n}\right)^{1/2}\right) \tag{34}$$

with $g(x) = -\frac{1}{x}$, $x > 0$, which is concave. Now by an application of Jensen's Inequality, (34) is asymptotically maximized by Jeffreys prior restricted to $[a, b]$. Q.E.D.

The method used in the present section helps us make significant steps towards our goal of characterizing the divergence measures leading to Jeffreys prior, i.e., to find conditions on a divergence measure under which it is asymptotically maximized by the Jeffreys prior. Here we work with the measure $\hat{J}(p)$, an idealized version of $J(p)$ obtained through a limit, as it is easier to deal with for the purpose of characterization. Above we have obtained Jeffreys prior as the asymptotic maximizer of $\hat{J}(p)$. The prior maximizing $\hat{J}(p)$ also (asymptotically) maximizes $J(p)$ if $J(p)$ can be approximated by $\hat{J}(p)$ up to the correct order in the sense that the error in approximating $J(p)$ by $\hat{J}(p)$ is of the same order as the error term in the expansion of $\hat{J}(p)$. For instance, we have proved this for the L_1-distance (see Appendix for a formal statement). In the proof we need to establish a certain asymptotic property of the L_1-distance between $p(\cdot|\boldsymbol{x})$ and $\hat{p}(\cdot|\boldsymbol{x})$. This requires technical calculations and uses results already available in the literature (namely, Ghosh et al. (1982)). For the case of α-divergence, we will need a similar

result about the asymptotic property of the L_1-distance between $p^\alpha(\cdot)p^{(1-\alpha)}(\cdot|\boldsymbol{x})$ and $p^\alpha(\cdot)\hat{p}^{(1-\alpha)}(\cdot|\boldsymbol{x})$. Proving such results will need more technical work compared to that needed for the L_1-distance. We didn't try such calculations as we have an alternative approach for the α-divergences based on the "Shrinkage Argument" presented in Section 3 that deals with $J(p)$ directly. In the particular case of $\alpha = -1$, we need to use an asymptotic expansion of the posterior and we use the shrinkage argument to obtain our results. The approach using the "Shrinkage Argument" directly expresses J through an integral of the expectation of a power of the posterior density. An expansion of the posterior density already available in the literature is employed while approximating the said expectation, and this in turn produces an expansion for J. For the case of $\alpha \neq -1$, only the leading term of the expansion of J is needed, while for the case of $\alpha = -1$, higher order terms in the expansion are used since the first order term does not involve the prior. If one wants to prove the result for the case of $\alpha = -1$ in the approach of Section 2, one would need two new major results. First, a result like Theorem 2.1 needs to be proved where in the definition of $\hat{J}$, the posterior normal approximation is replaced by a higher order expansion of the posterior. Proving such a result with increasingly complex higher order terms will be messy. Once such a result is proved, one has to then show that J is appropriately close to this new $\hat{J}$. For this one would need a result on asymptotic property of the L_1 distance between $p^\alpha(\cdot)p^{(1-\alpha)}(\cdot|\boldsymbol{x})$ and $p^\alpha(\cdot)\hat{p}_1^{(1-\alpha)}(\cdot|\boldsymbol{x})$ where $\hat{p}_1(\cdot|\boldsymbol{x})$ would be a higher order expansion of the posterior density as just mentioned. The "Shrinkage Argument" deals with such higher order terms in a simpler way.

3 An Alternative Approach to α-divergences – The Shrinkage Argument

What we present in this section is a multiparameter generalization of previous work of Ghosh, Mergel and Liu (2011). This result was briefly mentioned in Ghosh (2011). Here we give a detailed description. We consider selection of priors for the regular multiparameter family of distributions based on maximizing the α-divergence measures (see (35) below). We first provide a general expression for the expected α-divergence measure between the prior and the posterior and then use it to obtain a maximizing prior.

3.1 The Expected α-divergence Between the Prior and the Posterior

Consider the set up of Section 2.1 and consider a prior $p(\boldsymbol{\theta})$ which puts all its mass on a compact set in $\mathcal{R}^k$. The general expected α-divergence between the prior and the posterior is given by

$$J^\alpha(p) = \frac{1 - \int \left[\int p^\alpha(\boldsymbol{\theta}) p^{1-\alpha}(\boldsymbol{\theta}|\,\boldsymbol{x})\, d\boldsymbol{\theta}\right] m(\boldsymbol{x}) d\boldsymbol{x}}{\alpha(1-\alpha)} \tag{35}$$

(see (7) and (27)). For $\alpha = 0$ or 1, we need to interpret $J^\alpha(p)$ as its limiting value (when it exists). In particular,

$$J^0(p) = \int\int \left\{ \log \frac{p(\boldsymbol{\theta} \mid \boldsymbol{x})}{p(\boldsymbol{\theta})} \right\} p(\boldsymbol{\theta} \mid \boldsymbol{x}) m(\boldsymbol{x})\, d\theta d\boldsymbol{x}, \tag{36}$$

which is the KL divergence between the prior and the posterior considered for example in Lindley (1956), Bernardo (1979), Clarke and Barron (1990, 1994), and Ghosh and Mukerjee (1992).

From the relation $p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \boldsymbol{x})m(\boldsymbol{x})$, one can reexpress $J^\alpha(p)$ given in (35) as

$$\begin{aligned} J^\alpha(p) &= \frac{1 - \int\int p^{\alpha+1}(\boldsymbol{\theta}) p^{-\alpha}(\boldsymbol{\theta} \mid \boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x}\, d\boldsymbol{\theta}}{\alpha(1-\alpha)} \\ &= \frac{1 - \int p^{\alpha+1}(\boldsymbol{\theta}) E\left(p^{-\alpha}(\boldsymbol{\theta} \mid \boldsymbol{X}) | \boldsymbol{\theta}\right)\, d\boldsymbol{\theta}}{\alpha(1-\alpha)}, \end{aligned} \tag{37}$$

where $E(\cdot|\boldsymbol{\theta})$ denotes the conditional expectation given $\boldsymbol{\theta}$.

Let $l_n(\boldsymbol{\theta}) = n^{-1} \log p(\boldsymbol{x}|\boldsymbol{\theta})$ and let $I(\boldsymbol{\theta}) = \Big(I_{jr}(\boldsymbol{\theta})\Big)_{k\times k}$ denote the per observation Fisher information matrix. We write $I^{-1}(\boldsymbol{\theta}) = \Big(I^{jr}(\boldsymbol{\theta})\Big)_{k\times k}$. Let $p(x|\boldsymbol{\theta})$ denote the density of a single observation.

Before stating the main theorem of this section, we need a few more notations. Let $\hat{\boldsymbol{\theta}}_n = (\hat{\theta}_{n1}, \cdots, \hat{\theta}_{nk})^T$ denote the MLE of $\boldsymbol{\theta}$. Also, let $\nabla l_n(\boldsymbol{\theta})$ and $\nabla^2 l_n(\boldsymbol{\theta})$ denote the gradient and the Hessian of $l_n(\boldsymbol{\theta})$ and let $\hat{I}_n = -\nabla^2 l_n(\hat{\boldsymbol{\theta}}_n)$. Further, let

$$\begin{gathered} a_{jrs} = \left[\partial^3 l_n(\boldsymbol{\theta})/\partial\theta_j\partial\theta_r\partial\theta_s\right]|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n}, \quad a_{jrsu} = \left[\partial^4 l_n(\boldsymbol{\theta})/\partial\theta_j\partial\theta_r\partial\theta_s\partial\theta_u\right]|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n}, \\ A_{jrs} = E\left[\partial^3 l_n(\boldsymbol{\theta})/\partial\theta_j\partial\theta_r\partial\theta_s|\boldsymbol{\theta}\right], \quad A_{jrsu} = E\left[\partial^4 l_n(\boldsymbol{\theta})/\partial\theta_j\partial\theta_r\partial\theta_s\partial\theta_u|\boldsymbol{\theta}\right], \\ p_j(\boldsymbol{\theta}) = \partial p/\partial\theta_j, \ p_{jr}(\boldsymbol{\theta}) = \partial^2 p/(\partial\theta_j\partial\theta_r), \ 1 \le j, r, s, u \le k. \end{gathered}$$

The model assumptions for the following theorem are the same as the conditions (AI)-(AV) of Ghosh et al.(1982, pp.416-418). These assumptions imply the strong consistency of the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_n$ and the existence of the uniformly asymptotic normal expansion of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$. (Although the conditions in their paper are written in the form of one-dimensional parameter space, they can be easily extended to the multi-dimensional parameter space.)

The following is the main theorem of this section. This theorem is closely related to the results in Clarke and Sun (1999). They considered asymptotic expansions for the expected value of the posterior and squared posterior, which are, respectively, corresponding to $\alpha = -1$ and $\alpha = -2$ in our theorem.

Theorem 3.1. *Let $p(\boldsymbol{\theta})$ be a density function which is positive and three times continuously differentiable on the compact parameter space. Assume conditions (AI)-(AV) of Ghosh et al. (1982). Then*

$$
\begin{aligned}
& E\left[p^{-\alpha}(\boldsymbol{\theta}|\boldsymbol{X})|\boldsymbol{\theta}\right] = (2\pi)^{\frac{k\alpha}{2}} n^{-\frac{k\alpha}{2}} |I(\boldsymbol{\theta})|^{-\frac{1}{2}\alpha} (1-\alpha)^{-\frac{k}{2}} \qquad (38)\\
\times \quad & \left[1 + n^{-1}\left\{\alpha(1-\alpha)^{-1} \sum_{1\le j,r\le k} \left[\frac{1}{2}\alpha|I(\boldsymbol{\theta})|\frac{\partial|I^{-1}(\boldsymbol{\theta})|}{\partial\theta_r} I^{jr}(\boldsymbol{\theta}) + \frac{\partial I^{jr}(\boldsymbol{\theta})}{\partial\theta_r}\right]\frac{p_j(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}\right.\right.\\
+ \quad & \frac{2\alpha-\alpha^2}{2(1-\alpha)} \sum_{1\le j,r\le k} p_{jr}(\boldsymbol{\theta}) I^{jr}(\boldsymbol{\theta})/p(\boldsymbol{\theta}) - \frac{1}{2}\alpha \sum_{1\le j,r\le k} \frac{p_j(\boldsymbol{\theta})p_r(\boldsymbol{\theta})}{p^2(\boldsymbol{\theta})} I^{jr}(\boldsymbol{\theta})\\
- \quad & \left.\left. \frac{\alpha^2}{2(1-\alpha)} \sum_{1\le j,r,s,u\le k} A_{jrs} \frac{p_u(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} I^{jr}(\boldsymbol{\theta}) I^{su}(\boldsymbol{\theta}) + k(\boldsymbol{\theta})\right\} + o(n^{-1})\right],
\end{aligned}
$$

where $k(\boldsymbol{\theta})$ does not involve $p(\boldsymbol{\theta})$ or its derivatives and $\alpha < 1$.

In the proof, we adopt the shrinkage argument due to Ghosh (1994, Chapter 9). The shrinkage argument is a Bayesian approach for frequentist computations. Suppose that $\boldsymbol{X}$ is a random vector with density function $f(\boldsymbol{x};\theta)$ where the parameter θ belongs to an open subset of $\mathcal{R}^p$. In general, for any given value of θ, the shrinkage argument is used to find an asymptotic expansion for $E[q(\boldsymbol{X},\theta)|\theta]$ where q is a measurable function, and the expectation is known to exist. The use of this method can greatly simplify the computation of the higher order asymptotic expansions in many applications. For example, in the process of developing different kinds of probability matching priors, people extensively use the shrinkage argument to evaluate the asymptotic frequentist coverage probability of a posterior credible set. The shrinkage argument consists of three steps, and a detailed description of this method can be found in Datta and Mukerjee (2004, p.3). Also, we want to make it clear that neither the James-Stein shrinkage estimator nor the lasso-type shrinkage estimators in regularized regression models have connections to the shrinkage argument described here.

Proof. Step one. We consider any arbitrary thrice differentiable prior $\bar{p}$ vanishing outside a compact set and obtain $M(\boldsymbol{X}) = \int p^{-\alpha}(\boldsymbol{\theta}|\boldsymbol{X})\bar{p}(\boldsymbol{\theta}|\boldsymbol{X})d\boldsymbol{\theta}$. The expression of $M(\boldsymbol{X})$ is given in the following lemma.

Lemma 1. *Let $p(\boldsymbol{\theta})$ and $\bar{p}(\boldsymbol{\theta})$ be positive and three times continuously differentiable on the compact parameter space. Assume conditions (AI)-(AV) of Ghosh et al. (1982), the*

asymptotic expansion of $M(\boldsymbol{X})$ *is:*

$$
\begin{aligned}
& M(\boldsymbol{X}) = \int p^{-\alpha}(\boldsymbol{\theta}|\boldsymbol{X})\bar{p}(\boldsymbol{\theta}|\boldsymbol{X})d\boldsymbol{\theta} \qquad (39)\\
= \;& (2\pi)^{\frac{k\alpha}{2}} n^{-\frac{k\alpha}{2}} |\hat{I}_n|^{-\frac{\alpha}{2}} (1-\alpha)^{-\frac{k}{2}}\\
\times \;& \Bigg[1 - \frac{\alpha}{n(1-\alpha)} \sum_{1\le j,r,\le k} \frac{p_j(\hat{\boldsymbol{\theta}}_n)}{p(\hat{\boldsymbol{\theta}}_n)} \frac{\bar{p}_r(\hat{\boldsymbol{\theta}}_n)}{\bar{p}(\hat{\boldsymbol{\theta}}_n)} I^{jr}(\hat{\boldsymbol{\theta}}_n)\\
+ \;& \frac{\alpha(\alpha+1)}{2n} \sum_{1\le j,r\le k} \frac{p_j(\hat{\boldsymbol{\theta}}_n)p_r(\hat{\boldsymbol{\theta}}_n)}{p^2(\hat{\boldsymbol{\theta}}_n)} I^{jr}(\hat{\boldsymbol{\theta}}_n)\\
- \;& \frac{\alpha^2}{2n(1-\alpha)} \sum_{1\le j,r\le k} \Big\{ p_{jr}(\hat{\boldsymbol{\theta}}_n)/p(\hat{\boldsymbol{\theta}}_n) \Big\} I^{jr}(\hat{\boldsymbol{\theta}}_n)\\
- \;& \frac{\alpha^2}{2n(1-\alpha)} \sum_{1\le j,r,s,u\le k} a_{jrs} \Big\{ p_u(\hat{\boldsymbol{\theta}}_n)/p(\hat{\boldsymbol{\theta}}_n) \Big\} I^{jr}(\hat{\boldsymbol{\theta}}_n) I^{su}(\hat{\boldsymbol{\theta}}_n)\\
+ \;& \frac{k(\hat{\boldsymbol{\theta}}_n)}{n} + n^{-\frac{3}{2}} k_1(\hat{\boldsymbol{\theta}}_n, p) \Bigg].
\end{aligned}
$$

The main difference between $k(\hat{\boldsymbol{\theta}}_n)$ and $k_1(\hat{\boldsymbol{\theta}}_n, p)$ is that $k(\hat{\boldsymbol{\theta}}_n)$ does not involve $p(\boldsymbol{\theta})$ and its derivatives. The coefficient in front of $k_1(\hat{\boldsymbol{\theta}}_n, p)$ is $n^{-\frac{3}{2}}$. Therefore, this term can be treated as the approximation error term, and later will be labeled as $o(n^{-1})$.

The proof of this lemma is given in the appendix. The main idea of the proof is based on an asymptotic expansion for the posterior $p(\boldsymbol{\theta}|\boldsymbol{X})$ in Datta and Mukerjee, (2004, p.13). As they pointed out, the expansion for the posterior is valid for sample points in a set S. The $P_{\boldsymbol{\theta}}-$probability of S is $1 - o(n^{-1})$ uniformly on the compact parameters space. The set S can be constructed in the line of Bickel and Ghosh (1990, Section 3).

Step two. For any $\boldsymbol{\theta}$ in the interior of the support of $\bar{p}$, we find $\lambda_n(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} M(\boldsymbol{X})$ which is the expectation of $M(\boldsymbol{X})$ over the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{\theta}$. Noting that $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} = O_p(n^{-\frac{1}{2}})$ $(P_{\boldsymbol{\theta}})$ and using the argument applied in Datta and Mukerjee

(2004, p.7), we get

$$\begin{aligned}\lambda_n(\boldsymbol{\theta}) &= (2\pi)^{\frac{k\alpha}{2}} n^{-\frac{k\alpha}{2}} |I_n(\boldsymbol{\theta})|^{-\frac{\alpha}{2}} (1-\alpha)^{-\frac{k}{2}} \\ &\times \Bigg[1 - \frac{\alpha}{n(1-\alpha)} \sum_{1\le j,r,\le k} \frac{p_j(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \frac{\bar{p}_r(\boldsymbol{\theta})}{\bar{p}(\boldsymbol{\theta})} I^{jr}(\boldsymbol{\theta}) \\ &+ \frac{\alpha(\alpha+1)}{2n} \sum_{1\le j,r\le k} \frac{p_j(\boldsymbol{\theta}) p_r(\boldsymbol{\theta})}{p^2(\boldsymbol{\theta})} I^{jr}(\boldsymbol{\theta}) \\ &- \frac{\alpha^2}{2n(1-\alpha)} \sum_{1\le j,r\le k} \Big\{ p_{jr}(\boldsymbol{\theta})/p(\boldsymbol{\theta}) \Big\} I^{jr}(\boldsymbol{\theta}) \\ &- \frac{\alpha^2}{2n(1-\alpha)} \sum_{1\le j,r,s,u\le k} A_{jrs} \Big\{ p_u(\boldsymbol{\theta})/p(\boldsymbol{\theta}) \Big\} I^{jr}(\boldsymbol{\theta}) I^{su}(\boldsymbol{\theta}) \\ &+ \frac{k(\boldsymbol{\theta})}{n} + n^{-\frac{3}{2}} k_1(\boldsymbol{\theta}, p) \Bigg], \end{aligned} \tag{40}$$

where $k(\boldsymbol{\theta})$ does not involve p or its derivatives.

Based on conditions (AI)-(AV) of Ghosh et al. (1982) and the uniformity property about the $P_{\boldsymbol{\theta}}-$probability of the set S, using the method in Ghosh et al. (2011, 53-54), one can get that the last term in (40), $n^{-\frac{3}{2}} k_1(\boldsymbol{\theta}, p)$, is $o(n^{-1})$ uniformly over compact sets in the interior of the support of $\bar{p}(\boldsymbol{\theta})$.

Step three. Step 3 of the shrinkage argument involves integrating $\lambda_n(\boldsymbol{\theta})$ with respect to $\bar{p}(\boldsymbol{\theta})$ and then making $\bar{p}(\boldsymbol{\theta})$ degenerate at $\boldsymbol{\theta}$. In the present context, we need evaluation of

$$\sum_{1\le j,r\le k} \int \frac{p_j(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \bar{p}_r(\boldsymbol{\theta}) \frac{I^{jr}(\boldsymbol{\theta})}{|I(\boldsymbol{\theta})|^{\alpha/2}} d\boldsymbol{\theta}.$$

Integration by parts gives

$$\int \frac{p_j(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \bar{p}_r(\boldsymbol{\theta}) \frac{I^{jr}(\boldsymbol{\theta})}{|I(\boldsymbol{\theta})|^{\alpha/2}} d\theta_r = -\int \frac{d}{d\theta_r} \left\{ \frac{p_j(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \frac{I^{jr}(\boldsymbol{\theta})}{|I(\boldsymbol{\theta})|^{\alpha/2}} \right\} \bar{p}(\boldsymbol{\theta}) d\theta_r.$$

Hence,

$$\begin{aligned} & \sum_{1\le j,r\le k}\int \frac{p_j(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}\bar{p}_r(\boldsymbol{\theta})\frac{I^{jr}(\boldsymbol{\theta})}{|I(\boldsymbol{\theta})|^{\alpha/2}}d\boldsymbol{\theta} \qquad (41)\\ = & -\sum_{1\le j,r\le k}\int\left\{\frac{p_{jr}(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}-\frac{p_j(\boldsymbol{\theta})p_r(\boldsymbol{\theta})}{p^2(\boldsymbol{\theta})}\right\}\frac{I^{jr}(\boldsymbol{\theta})}{|I(\boldsymbol{\theta})|^{\alpha/2}}\bar{p}(\boldsymbol{\theta})d\boldsymbol{\theta}\\ - & \sum_{1\le j,r\le k}\int\frac{p_j(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}\left\{-\frac{\alpha}{2}|I(\boldsymbol{\theta})|^{-\frac{\alpha}{2}-1}\frac{d}{d\theta_r}|I(\boldsymbol{\theta})|\right.\\ + & \left.|I(\boldsymbol{\theta})|^{-\frac{\alpha}{2}}\frac{d}{d\theta_r}I^{jr}(\boldsymbol{\theta})\right\}\bar{p}(\boldsymbol{\theta})d\boldsymbol{\theta}. \end{aligned}$$

Notice that $\bar{p}(\boldsymbol{\theta})$ eventually will converge weakly to the degenerate prior at the true $\boldsymbol{\theta}$ which is an interior point of the support of $\bar{p}(\boldsymbol{\theta})$. Using the fact noted after (40), and combining (40), (41), one gets after some simplification,

$$\begin{aligned} & \int\lambda_n(\boldsymbol{\theta})\bar{p}(\boldsymbol{\theta})d\boldsymbol{\theta}=(2\pi)^{\frac{k\alpha}{2}}n^{-\frac{k\alpha}{2}}(1-\alpha)^{-\frac{k}{2}}\\ \times & \left[\int|I(\boldsymbol{\theta})|^{-\frac{\alpha}{2}}\bar{p}(\boldsymbol{\theta})d\boldsymbol{\theta}-\frac{\alpha^2}{2n(1-\alpha)}\sum_{1\le j,r\le k}\int\frac{p_j(\boldsymbol{\theta})\frac{d}{d\theta_r}|I(\boldsymbol{\theta})|}{p(\boldsymbol{\theta})|I(\boldsymbol{\theta})|^{\frac{\alpha}{2}+1}}\bar{p}(\boldsymbol{\theta})d\boldsymbol{\theta}\right.\\ + & \frac{\alpha}{n(1-\alpha)}\sum_{1\le j,r\le k}\int\frac{p_j(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}\frac{\frac{d}{d\theta_r}I^{jr}(\boldsymbol{\theta})}{|I(\boldsymbol{\theta})|^{\frac{\alpha}{2}}}\bar{p}(\boldsymbol{\theta})d\boldsymbol{\theta}\\ - & \frac{\alpha}{2n}\sum_{1\le j,r\le k}\int\frac{p_j(\boldsymbol{\theta})p_r(\boldsymbol{\theta})}{p^2(\boldsymbol{\theta})}\frac{I^{jr}(\boldsymbol{\theta})}{|I(\boldsymbol{\theta})|^{\frac{\alpha}{2}}}\bar{p}(\boldsymbol{\theta})d\boldsymbol{\theta}\\ + & \frac{2\alpha-\alpha^2}{2n(1-\alpha)}\sum_{1\le j,r\le k}\int\frac{p_{jr}(\boldsymbol{\theta})}{p(\boldsymbol{\theta})|I(\boldsymbol{\theta})|^{\frac{\alpha}{2}}}I^{jr}(\boldsymbol{\theta})\bar{p}(\boldsymbol{\theta})d\boldsymbol{\theta}\\ - & \frac{\alpha^2}{2n(1-\alpha)}\sum_{1\le j,r,s,u\le k}\int A_{jrs}\frac{p_u(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}I^{jr}(\boldsymbol{\theta})I^{su}(\boldsymbol{\theta})\bar{p}(\boldsymbol{\theta})d\boldsymbol{\theta}\\ + & \left.n^{-1}\int\frac{k(\boldsymbol{\theta})}{|I(\boldsymbol{\theta})|^{\frac{\alpha}{2}}}\bar{p}(\boldsymbol{\theta})d\boldsymbol{\theta}+o(n^{-1})\right]. \end{aligned}$$

By making $\bar{p}(\boldsymbol{\theta})$ degenerate at $\boldsymbol{\theta}$, and noting $\frac{d}{d\theta_r}|I(\boldsymbol{\theta})| = -|I(\boldsymbol{\theta})|^2 \frac{d}{d\theta_r}|I^{-1}(\boldsymbol{\theta})|$, we have

$$\begin{aligned}
\lambda_n(\boldsymbol{\theta}) &= (2\pi)^{\frac{k\alpha}{2}} n^{-\frac{k\alpha}{2}} (1-\alpha)^{-\frac{k}{2}} |I(\boldsymbol{\theta})|^{-\frac{\alpha}{2}} \\
&\times \Bigg[1 + \frac{\alpha^2}{2n(1-\alpha)} |I(\boldsymbol{\theta})| \sum_{j,r} \frac{\partial}{\partial \theta_r} |I^{-1}(\boldsymbol{\theta})| I^{jr}(\boldsymbol{\theta}) \frac{p_j(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \\
&+ \frac{\alpha}{n(1-\alpha)} \sum_{j,r} \frac{\partial}{\partial \theta_r} I^{jr}(\boldsymbol{\theta}) \frac{p_j(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} - \frac{\alpha}{2n} \sum_{j,r} \frac{p_j(\boldsymbol{\theta}) p_r(\boldsymbol{\theta})}{p^2(\boldsymbol{\theta})} I^{jr}(\boldsymbol{\theta}) \\
&+ \frac{2\alpha - \alpha^2}{2n(1-\alpha)} \sum_{j,r} \frac{p_{jr}(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} I^{jr}(\boldsymbol{\theta}) - \frac{\alpha^2}{2n(1-\alpha)} \sum_{j,r,s,u} A_{jrsu} \frac{p_u(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} I^{jr}(\boldsymbol{\theta}) I^{su}(\boldsymbol{\theta}) \\
&+ \frac{k(\boldsymbol{\theta})}{n} + o(n^{-1}) \Bigg].
\end{aligned}$$

This proves the theorem.

Sections 3.2 and 3.3 will focus on derivation of optimal priors under the given divergence loss.

3.2 Jeffreys Prior as Maximizer of Expected α-divergences

Here, we assume that (38) holds uniformly in $\boldsymbol{\theta}$ on the compact parameter space. This assumption is typically made in the reference prior literature. Formal justification of this assumption is difficult and may need more conditions. But as pointed out by Clark and Sun (1997), "Egoroff's theorem guarantees the existence of a set with arbitrarily large probability on which the convergence can be taken as uniform". Here, we follow this idea. In (37), we integrate with respect to $\boldsymbol{\theta}$ over the compact set to get an asymptotic expansion of $J^\alpha(p)$.

In view of (37) and Theorem 3.1, neglecting the $O_p(n^{-1})$ term, the selection of a prior p amounts to minimization of

$$\frac{1}{\alpha(1-\alpha)} \int p^{1+\alpha}(\boldsymbol{\theta}) |I(\boldsymbol{\theta})|^{-\frac{\alpha}{2}} d\boldsymbol{\theta} = \frac{1}{\alpha(1-\alpha)} \int \left(\frac{|I(\boldsymbol{\theta})|^{1/2}}{p(\boldsymbol{\theta})} \right)^{-\alpha} p(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

with respect to a prior p. Note that this is similar to maximization of (28) of Section 2.2. As argued in Section 2.2, one thus obtains the Jeffreys prior as the (asymptotic) maximizer of $J^\alpha(p)$ for $-1 < \alpha < 0$ and $0 < \alpha < 1$.

When $\alpha \longrightarrow 0$, using either the shrinkage argument, or alternatively from Clarke and

Barron (1990, 1994), one gets

$$J^0(p) = \frac{p}{2}\log\left(\frac{n}{2\pi e}\right) - \int p(\boldsymbol{\theta})\log\frac{p(\boldsymbol{\theta})}{|I(\boldsymbol{\theta})|^{1/2}}d\boldsymbol{\theta} + o(1),$$

which is maximized up to first order of approximation by $p(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}$.

Remark. When $\alpha < -1$, for the one parameter case, Ghosh et al. (2011) showed that Jeffreys prior is the minimizer of $J^\alpha(p)$, and there is no maximizer in this case. Their result can be extended to the multiparameter case without difficulty.

3.3 Maximizing Prior for α-divergence with $\alpha = -1$

In 3.2 we considered α-divergence measures for $\alpha < 1$ and $\alpha \neq -1$. For $-1 < \alpha < 1$, Jeffreys prior maximizes the $\alpha-$divergence. When $\alpha < -1$, Jeffreys prior turns out to be the minimizer. Then what is the desired prior when $\alpha = -1$? Intuitively, Jeffreys prior may not be the choice, because $\alpha = -1$ is the dividing point for choosing Jeffreys prior or not choosing Jeffreys prior. Also, by looking at the expression of $\alpha-$divergence, we can confirm that $\alpha = -1$ is a crucial point. The key component of the $\alpha-$divergence is $\int \left[\frac{p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{x})}\right]^\alpha p(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}$. When $-1 < \alpha < 0$, the importance of those values of $\boldsymbol{\theta}$ at which the posterior is greater than the prior is lessened by the factor $\left[\frac{p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{x})}\right]^\alpha$. On the contrary, when $\alpha < -1$, the importance of those values is increased by the same factor.

When $\alpha = -1$, the corresponding divergence is called the chi-square divergence which was considered in Clarke and Sun (1997) for the one parameter exponential family and also in Ghosh, Mergel and Liu (2011) for the general one-parameter family of distributions. The chi-square divergence is motivated by the chi-square goodness-of-fit statistic. Clarke and Sun (1997) gave a nice discussion on this divergence.

In the following theorem, we give a complete result about the reference prior under the chi-square divergence.

Theorem 3.2. *Under the assumptions of Theorem 3.1, for the chi-square divergence, the desired reference prior $p(\boldsymbol{\theta})$ is the solution of the following partial differential equations:*

$$\frac{\partial \log p(\boldsymbol{\theta})}{\partial\theta_i} = -\frac{1}{4}\sum_{j=1}^{k}\sum_{r=1}^{k} A_{j,r,i}I^{jr}(\boldsymbol{\theta}) + \frac{1}{2}|I(\boldsymbol{\theta})|^{-1}\frac{\partial|I(\boldsymbol{\theta})|}{\partial\theta_i}, \quad i = 1,\cdots,k,$$

where

$$A_{j,r,i} = E\left(\frac{\partial \log p(\boldsymbol{X}|\boldsymbol{\theta})}{\partial\theta_j}\frac{\partial \log p(\boldsymbol{X}|\boldsymbol{\theta})}{\partial\theta_r}\frac{\partial \log p(\boldsymbol{X}|\boldsymbol{\theta})}{\partial\theta_i}\right).$$

Proof. Here $p^{\alpha+1}(\boldsymbol{\theta}) = 1$ so that the first order term appearing in Theorem 3.1 will not suffice in finding the prior p and the coefficient of n^{-1} is needed in finding the optimal p. To this end, since $\alpha = -1$ so that $\alpha(1-\alpha) = -2$, using (37), (38), and neglecting all

terms which do not involve p or its derivatives, it suffices to maximize up to the second order approximation,

$$\int |I(\boldsymbol{\theta})|^{1/2} \left[\frac{|I(\boldsymbol{\theta})|}{4} \sum_{j,r} \left\{ \frac{\partial}{\partial \theta_r} |I^{-1}(\boldsymbol{\theta})| I^{jr}(\boldsymbol{\theta}) - 2 \frac{\partial I^{jr}(\boldsymbol{\theta})}{\partial \theta_r} \right\} \frac{p_j(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \right. \tag{42}$$
$$- \frac{3}{4} \sum_{j,r} \frac{p_{jr}(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} I^{jr}(\boldsymbol{\theta}) + \frac{1}{2} \sum_{j,r} \frac{p_j(\boldsymbol{\theta}) p_r(\boldsymbol{\theta})}{p^2(\boldsymbol{\theta})} I^{jr}(\boldsymbol{\theta})$$
$$\left. - \frac{1}{4} \sum_{j,r,s} A_{jrs} \frac{p_u(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} I^{jr}(\boldsymbol{\theta}) I^{su}(\boldsymbol{\theta}) \right] d\boldsymbol{\theta}.$$

Writing $p_{jr}(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_r} \left(\frac{p_j(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \right) + \frac{p_j(\boldsymbol{\theta}) p_r(\boldsymbol{\theta})}{p^2(\boldsymbol{\theta})}$, (42) simplifies to

$$\int |I(\boldsymbol{\theta})|^{1/2} \left[\frac{|I(\boldsymbol{\theta})|}{4} \sum_{j,r} \left\{ \frac{\partial}{\partial \theta_r} |I^{-1}(\boldsymbol{\theta})| I^{jr}(\boldsymbol{\theta}) - 2 \frac{\partial I^{jr}(\boldsymbol{\theta})}{\partial \theta_r} \right\} \frac{p_j(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \right. \tag{43}$$
$$- \frac{1}{4} \sum_{j,r} \frac{p_j(\boldsymbol{\theta}) p_r(\boldsymbol{\theta})}{p^2(\boldsymbol{\theta})} I^{jr}(\boldsymbol{\theta}) - \frac{3}{4} \sum_{j,r} \frac{\partial}{\partial \theta_r} \left(\frac{p_j(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \right)$$
$$\left. - \frac{1}{4} \sum_{j,r,s} A_{jrs} \frac{p_u(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} I^{jr}(\boldsymbol{\theta}) I^{su}(\boldsymbol{\theta}) \right] d\boldsymbol{\theta}.$$

Let

$$\boldsymbol{y}(\boldsymbol{\theta}) = (\boldsymbol{y}_1(\boldsymbol{\theta}), \cdots, \boldsymbol{y}_k(\boldsymbol{\theta})) = \left(\frac{p_1(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}, \cdots, \frac{p_k(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \right)$$
$$\nabla \boldsymbol{y}(\boldsymbol{\theta}) = \left(\frac{\partial \boldsymbol{y}_1(\boldsymbol{\theta})}{\partial \theta_1}, \cdots, \frac{\partial \boldsymbol{y}_1(\boldsymbol{\theta})}{\partial \theta_k}, \cdots, \frac{\partial \boldsymbol{y}_k(\boldsymbol{\theta})}{\partial \theta_1}, \cdots, \frac{\partial \boldsymbol{y}_k(\boldsymbol{\theta})}{\partial \theta_k} \right).$$

Note that (43) can be expressed as

$$\int F(\boldsymbol{\theta}, \boldsymbol{y}(\boldsymbol{\theta}), \nabla \boldsymbol{y}(\boldsymbol{\theta})) d\boldsymbol{\theta}, \tag{44}$$

so we need to find $\boldsymbol{y}(\boldsymbol{\theta})$ to maximize the above integral. From Giaquinta (1983, p.8), the maximizer should satisfy the Euler-Lagrange equations:

$$\frac{\partial F}{\partial \boldsymbol{y}_i(\boldsymbol{\theta})} - \sum_{j=1}^{k} \frac{\partial}{\partial \theta_j} \left(\frac{\partial F}{\partial (\partial \boldsymbol{y}_i / \partial \theta_j)} \right) = 0, \quad i = 1, \cdots, k. \tag{45}$$

Equivalently, the Euler-Lagrange equations are

$$\begin{aligned} \sum_{j=1}^{k} \frac{p_j}{p} I^{ij}(\boldsymbol{\theta}) &= -\frac{1}{2} \sum_{j=1}^{k} \sum_{r=1}^{k} \sum_{s=1}^{k} A_{jrs} I^{jr}(\boldsymbol{\theta}) I^{si}(\boldsymbol{\theta}) \\ &+ \frac{1}{2} \sum_{j=1}^{k} \left[-\frac{1}{2} |I(\boldsymbol{\theta})| \frac{\partial |I^{-1}(\boldsymbol{\theta})|}{\partial \theta_j} I^{ij}(\boldsymbol{\theta}) + \frac{\partial I^{ij}(\boldsymbol{\theta})}{\partial \theta_j} \right], \quad i = 1, \cdots, k. \end{aligned} \tag{46}$$

In matrix notations, (46) is:

$$I^{-1}(\boldsymbol{\theta}) \begin{pmatrix} \frac{p_1}{p} \\ \vdots \\ \frac{p_k}{p} \end{pmatrix} = -\frac{1}{2} I^{-1}(\boldsymbol{\theta}) \begin{pmatrix} A_1 \\ \vdots \\ A_k \end{pmatrix} + \frac{1}{4} I^{-1}(\boldsymbol{\theta}) \begin{pmatrix} B_1 \\ \vdots \\ B_k \end{pmatrix} + \frac{1}{2} I^{-1}(\boldsymbol{\theta}) \begin{pmatrix} C_1 \\ \vdots \\ C_k \end{pmatrix}, \tag{47}$$

where

$$\begin{aligned} A_i &= \sum_{j=1}^{k} \sum_{r=1}^{k} A_{jri} I^{jr}(\boldsymbol{\theta}), \quad i = 1, \cdots, k, \\ B_i &= |I^{-1}(\boldsymbol{\theta})| \frac{\partial |I(\boldsymbol{\theta})|}{\partial \theta_i} = -\sum_{j=1}^{k} \sum_{l=1}^{k} I_{jl}(\boldsymbol{\theta}) \frac{\partial I^{lj}(\boldsymbol{\theta})}{\partial \theta_i}, \quad i = 1, \cdots, k, \\ C_i &= \sum_{l=1}^{k} I_{il}(\boldsymbol{\theta}) \left(\sum_{j=1}^{k} \frac{\partial I^{lj}(\boldsymbol{\theta})}{\partial \theta_j} \right) = \sum_{j=1}^{k} \sum_{l=1}^{k} I_{il}(\boldsymbol{\theta}) \frac{\partial I^{lj}(\boldsymbol{\theta})}{\partial \theta_j}, \quad i = 1, \cdots, k. \end{aligned}$$

Next by the Bartlett identity (Bartlett (1953)),

$$A_{jri} = \frac{1}{2} \left[A_{j,r,i} - \frac{\partial I_{ri}(\boldsymbol{\theta})}{\partial \theta_j} - \frac{\partial I_{ji}(\boldsymbol{\theta})}{\partial \theta_r} - \frac{\partial I_{jr}(\boldsymbol{\theta})}{\partial \theta_i} \right],$$

where

$$A_{j,r,i} = E \left(\frac{\partial \log p(\boldsymbol{X}|\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \log p(\boldsymbol{X}|\boldsymbol{\theta})}{\partial \theta_r} \frac{\partial \log p(\boldsymbol{X}|\boldsymbol{\theta})}{\partial \theta_i} \right).$$

One can then simplify the Euler-Lagrange equations to

$$\frac{\partial \log p(\boldsymbol{\theta})}{\partial \theta_i} = -\frac{1}{4}\sum_{j=1}^{k}\sum_{r=1}^{k} A_{j,r,i} I^{jr}(\boldsymbol{\theta}) + \frac{1}{2}|I(\boldsymbol{\theta})|^{-1}\frac{\partial |I(\boldsymbol{\theta})|}{\partial \theta_i}, \quad i = 1, \cdots, k. \tag{48}$$

Then, the desired divergence prior $p(\boldsymbol{\theta})$ is obtained by solving 48. We will denote such a prior p as p_{chi}. Q.E.D.

Below we illustrate through a couple of examples how optimal priors can be obtained. The optimal priors are different from Jeffreys prior but, by the observation made in the paragraph following (5) of Section 2.2, are invariant under one-to-one differentiable transformations.

Example 1. When $p(x|\boldsymbol{\theta})$ belongs to exponential family ($\boldsymbol{\theta}$ is canonical parameter vector), one can check that

$$A_i = C_i = -B_i; \quad i = 1, \cdots, k,$$

since $A_{jri} = -\frac{\partial I_{jr}(\boldsymbol{\theta})}{\partial \theta_i}$ and $\frac{\partial I_{ik}(\boldsymbol{\theta})}{\partial \theta_j} = \frac{\partial I_{jk}(\boldsymbol{\theta})}{\partial \theta_i}$. Based on above relations, the Euler-Lagrange equations are

$$\begin{pmatrix} \frac{p_1}{p} \\ \vdots \\ \frac{p_k}{p} \end{pmatrix} = \frac{1}{4}|I^{-1}(\boldsymbol{\theta})| \begin{pmatrix} \frac{\partial |I(\boldsymbol{\theta})|}{\partial \theta_1} \\ \vdots \\ \frac{\partial |I(\boldsymbol{\theta})|}{\partial \theta_k} \end{pmatrix}.$$

Hence, $p(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{\frac{1}{4}}$.

Remark. In general, it is hard to study the propriety of the posterior when using p_{chi}. The above example shows that $p_{\text{chi}} \propto |I(\boldsymbol{\theta})|^{\frac{1}{4}}$, when $p(x|\boldsymbol{\theta})$ belongs to the exponential family, and $\boldsymbol{\theta}$ is the canonical parameter vector. In this case, p_{chi} may result in an improper posterior for some models. But because p_{chi} is the square root of Jeffreys prior, the chance of getting an improper posterior when using p_{chi} is smaller than the chance of getting an improper posterior when using Jeffreys prior.

Outside the multiparameter exponential family, when $\alpha = -1$, an optimal prior can even be different from Jeffreys prior and the above prior. Here is one example:

Example 2. Consider the Inverse Gaussian distribution with probability density function

$$f(x|\mu, \lambda) = \frac{\lambda^{1/2}}{\sqrt{2\pi x^3}} \exp\left[-\frac{\lambda}{2x}\left(\frac{x}{\mu} - 1\right)^2\right] = \frac{\lambda^{1/2}}{\sqrt{2\pi x^3}} \exp\left[-\frac{\lambda}{2}\left(\frac{x}{\mu^2} - \frac{2}{\mu} + \frac{1}{x}\right)\right].$$

One gets

$$E\left[-\frac{\partial^2 \log f}{\partial \mu^2}\middle| \mu, \lambda\right] = \frac{\lambda}{\mu^3}, \quad E\left[-\frac{\partial^2 \log f}{\partial \mu \partial \lambda}\middle| \mu, \lambda\right] = 0,$$

$$E\left[-\frac{\partial^2 \log f}{\partial \lambda^2}\middle|\mu,\lambda\right]=\frac{1}{2\lambda^2}$$

$$E\left[\frac{\partial^3 \log f}{\partial \mu^3}\middle|\mu,\lambda\right]=\frac{6\lambda}{\mu^4},\quad E\left[\frac{\partial^3 \log f}{\partial \mu\partial \lambda^2}\middle|\mu,\lambda\right]=0,$$

$$E\left[\frac{\partial^3 \log f}{\partial \mu^2\partial \lambda}\middle|\mu,\lambda\right]=-\frac{1}{\mu^3},\quad E\left[\frac{\partial^3 \log f}{\partial \lambda^3}\middle|\mu,\lambda\right]=\frac{1}{\lambda^3},$$

and

$$A_1=\frac{6}{\mu},\ \ A_2=\frac{1}{\lambda},\ \ B_1=-\frac{3}{\mu},\ \ B_2=-\frac{1}{\lambda},\ \ C_1=\frac{3}{\mu},\ \ C_2=\frac{2}{\lambda}.$$

Hence, according to (48), the prior $p(\mu,\lambda)$ should satisfy these equations:

$$\frac{\partial \log p(\mu,\lambda)}{\partial \mu}=-\frac{9}{4\mu},$$

$$\frac{\partial \log p(\mu,\lambda)}{\partial \lambda}=\frac{1}{4\lambda}.$$

Therefore, the optimal prior $p(\mu,\lambda)$ is proportional to $\mu^{-\frac{9}{4}}\lambda^{\frac{1}{4}}$.

4 Reference Priors in the Presence of a Nuisance Parameter

For the multiparameter case, if we consider all the parameters to be equally important, we maximize the average divergence

$$J(p)=\int D_p(\boldsymbol{x})m(\boldsymbol{x})d\boldsymbol{x}$$

where $D_p(\boldsymbol{x})=D(p(\cdot),p(\cdot|\boldsymbol{x}))$ is a divergence measure between the prior and the posterior. It is known that for the Kullback-Leibler divergence, this leads to the Jeffreys prior. In Section 2 above, we have also seen that $J(p)$ is asymptotically maximized by the Jeffreys prior for a class of divergence measures.

We now assume that there is an ordering of the parameters according to their importance. We consider the case with two parameters $\boldsymbol{\theta}=(\theta_1,\theta_2)$ where θ_1 is assumed to be more important than θ_2, e.g., θ_2 is a nuisance parameter. We are interested in the reference prior in the sense of Berger and Bernardo (1989, 1992a, 1992b). In this approach, $p(\theta_2|\theta_1)$ is chosen as a "reasonable" objective prior while $p(\theta_1)$ is obtained as the prior maximizing the functional $J^*(p)$, given by

$$J^*(p)=\int D_p(\boldsymbol{x})m(\boldsymbol{x})d\boldsymbol{x},$$

where $D_p(\boldsymbol{x})=D(p(.),p(.|\boldsymbol{x}))$. Here $p(.)$ and $p(.|\boldsymbol{x})$ are the prior of θ_1 and posterior of θ_1 respectively and $m(\boldsymbol{x})$ is the marginal density of the data with respect to the joint

prior $p(\theta_1, \theta_2) = p(\theta_2|\theta_1)p(\theta_1)$. Thus $J^*(p)$ is the average divergence between $p(\theta_1)$ and $p(\theta_1|\boldsymbol{x})$ with respect to the marginal density of the data. Usually $p(\theta_2|\theta_1)$ is chosen as the conditional Jeffreys prior which is proportional to $\sqrt{I_{22}(\boldsymbol{\theta})}$. For the Kullback-Leibler divergence, the prior $p(\theta_1)$ maximizing $J^*(p)$ is obtained as the geometric mean of $(I^{11}(\boldsymbol{\theta}))^{-1/2}$ with respect to $p(\theta_2|\theta_1)$ where $I^{11}(\boldsymbol{\theta})$ denotes the $(1,1)$-th element of $I^{-1}(\boldsymbol{\theta})$ (see, e.g., Ghosh et al. 2006, Sec. 5.1.10). Instead of the Kullback-Leibler, one may consider a divergence measure obtained in Section 2 leading to the Jeffreys prior. It is our interest to see if we get a satisfactory reference prior when the algorithm of Berger and Bernardo is applied on such divergence measures. We consider below the average α-divergence $J^*(p)$ between $p(\theta_1)$ and $p(\theta_1|\boldsymbol{x})$ with

$$\begin{aligned} D_p(\boldsymbol{x}) &= \left\{ \frac{1}{\alpha(1-\alpha)} \int [\alpha p(\theta_1') + (1-\alpha)p(\theta_1'|\boldsymbol{x}) - p^\alpha(\theta_1')p^{1-\alpha}(\theta_1'|\boldsymbol{x})] d\theta_1' \right\} \\ &= \frac{1}{1-\alpha} + \frac{1}{\alpha} - \frac{1}{\alpha(1-\alpha)} \int \left(\frac{p(\theta_1')}{p(\theta_1'|\boldsymbol{x})} \right)^\alpha p(\theta_1'|\boldsymbol{x}) d\theta_1'. \end{aligned}$$

We now briefly outline a derivation of the maximizer $p(\theta_1)$ of $J^*(p)$ without going into the rigorous arguments. We take a compact rectangle $K_1 \times K_2$. We fix $p(\theta_2|\theta_1)$ as a conditional prior for θ_2 on K_2 and consider priors $p(\theta_1)$ supported on K_1. Using normal approximation of the posterior $p(\theta_1'|\boldsymbol{x})$ and proceeding as in the one-parameter case we can approximate $J^*(p)$ as

$$\begin{aligned} J^*(p(.)) &\approx \text{Constant} - \frac{\text{Constant } n^{-\alpha/2}}{\alpha(1-\alpha)} \int_{K_1 \times K_2} \left(\sqrt{I^{11}(\boldsymbol{\theta})} p(\theta_1) \right)^\alpha p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \text{Constant} + \frac{\text{Constant } n^{-\alpha/2}}{\alpha(1-\alpha)} \int_{K_1} \psi(\theta_1) p^\alpha(\theta_1) p(\theta_1) d\theta_1 \end{aligned}$$

where $\psi(\theta_1) = \int_{K_2} (I^{11}(\boldsymbol{\theta}))^{\alpha/2} p(\theta_2|\theta_1) d\theta_2$. The last expression above can be rewritten as

$$\text{Constant} + \frac{\text{Constant } n^{-\alpha/2}}{\alpha(1-\alpha)} \int_{K_1} g\left(\frac{\psi^{-\frac{1}{\alpha}}(\theta_1)}{p(\theta_1)} \right) p(\theta_1) d\theta_1,$$

where $g(x) = -x^{-\alpha}$. This is maximized if $p(\theta_1) \propto \psi^{-\frac{1}{\alpha}}(\theta_1)$ for $0 < \alpha < 1$ and $-1 < \alpha < 0$ (see the result for the one-parameter case obtained in Section 2.2).

We now consider an example to see if the reference prior obtained above using the α-divergence measure is a reasonable one. At least in the following example this is indeed the case. Hopefully other examples will be found later.

Example 4.1. Consider the $N(\theta_2, \theta_1)$ model for the data, where the variance θ_1 is considered more important. We take $p(\theta_2|\theta_1)$ as the conditional Jeffreys prior, proportional to $\sqrt{I_{22}(\boldsymbol{\theta})}$. Here

$$I(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{2\theta_1^2} & 0 \\ 0 & \frac{1}{\theta_1} \end{pmatrix}, \; I_{22}(\boldsymbol{\theta}) = \frac{1}{\theta_1}, \; I^{11}(\boldsymbol{\theta}) = 2\theta_1^2.$$

Thus $p(\theta_2|\theta_1) = c$ on K_2 where $c^{-1} =$ length of K_2, and therefore,

$$\psi(\theta_1) = \int (2\theta_1^2)^{\alpha/2} c.d\theta_2 \ \propto \theta_1^\alpha$$

and

$$p(\theta_1) \propto \frac{1}{\psi^{\frac{1}{\alpha}}(\theta_1)} = \frac{1}{(\theta_1^\alpha)^{1/\alpha}} = \frac{1}{\theta_1}.$$

5 Concluding Remarks

In this paper we have provided a sufficient condition for a divergence measure (between prior and posterior) to be maximized by the Jeffreys prior. The Kullback-Leibler divergence used by Bernardo, Hellinger distance, L_1-distance and the α-divergences of Amari with $-1 < \alpha < 0,\ 0 < \alpha < 1$, satisfy this condition. We believe that among the divergence measures commonly seen in the literature, the divergences maximized by the Jeffreys prior are only those which satisfy the sufficeint conditions obtained in this paper. Although we have not verified this, it seems possible that for any particular candidate divergence, some suitable numerical approach may settle the issue for that particular divergence but developing such an approach would need further work. For example, for a particular divergence, one may consider a simple model $f(x|\theta)$, such as a location parameter family (or, more specifically, $N(\theta, 1)$ model) so that the Jeffreys prior is constant, and depending on the divergence, try to find a prior p with a higher value of $J(p)$ than the Jeffreys prior. The calculation of $J(p)$ may be done numerically by simulating from the prior $p(\theta)$ and the density $f(x|\theta)$.

For all the divergences mentioned above, our proof is based on first order asymptotic approximation of the average divergence measure $J(p)$. For the α-divergence measure with $\alpha = -1$ (known as chi-square divergence) our sufficient condition is satisfied but the first order approximation of $J(p)$ is free of the prior p. Therefore we need to consider the next smaller order term in the asymptotic expansion of $J(p)$ which is maximized by a prior different from the Jeffreys prior.

The new divergences that lead to Jeffreys prior strengthen the foundation of Jeffreys prior. From these new divergence measures one can construct new reference priors, enlarging the class of objective priors. They might provide in the future new insights about a general notion of information in a prior and the concept of nonsubjective, low information priors. The study of all such priors will help us settle whether inference based on objective priors is robust with respect to choice of objective priors, at least with respect to the reference priors arising as indicated above.

Such divergences also provide quantitative assessment of whether a given prior has too much influence on the posterior for given data. Judging this visually has been a standard practice among Bayesians. Our divergences provide a quantitative evaluation of the closeness of prior to posterior for given data. Small values of the divergence will

alert the analyst that the prior needs to be changed. This can help quantify the visual comparisons of priors as done in Gelman (2006).

6 Appendix

We first present a proof of lemma (1) and then discuss on approximation of $J(p)$ by $\hat{J}(p)$.

6.1 Proof of Lemma 1

To prove this lemma, we begin with an asymptotic expansion of the posterior $p(\boldsymbol{\theta}|\boldsymbol{X})$ (Datta and Mukerjee, 2004, p.13)

$$p(\boldsymbol{\theta}|\boldsymbol{X}) = (2\pi)^{-\frac{k}{2}} n^{k/2} |\hat{I}_n|^{1/2} \exp\left[-\frac{n}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)^T \hat{I}_n (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)\right] \tag{49}$$

$$\times \Bigg[1 + n^{-\frac{1}{2}} \Bigg\{ \sum_{j=1}^{k} \sqrt{n}(\theta_j - \hat{\theta}_{nj}) \frac{p_j(\hat{\boldsymbol{\theta}}_n)}{p(\hat{\boldsymbol{\theta}}_n)} + \frac{1}{6} \sum_{1\le j,r,s\le k} n^{\frac{3}{2}} (\theta_j - \hat{\theta}_{nj})(\theta_r - \hat{\theta}_{nr})(\theta_s - \hat{\theta}_{ns}) a_{jrs} \Bigg\}$$

$$+ n^{-1} \Bigg\{ \frac{1}{2} \sum_{1\le j,r\le k} \Big(n(\theta_j - \hat{\theta}_{nj})(\theta_r - \hat{\theta}_{nr}) - I^{jr}(\hat{\boldsymbol{\theta}}_n) \Big) \frac{p_{jr}(\hat{\boldsymbol{\theta}}_n)}{p(\hat{\boldsymbol{\theta}}_n)}$$

$$+ \frac{1}{6} \sum_{1\le j,r,s,u\le k} \Big(n^2 (\theta_j - \hat{\theta}_{nj})(\theta_r - \hat{\theta}_{nr})(\theta_s - \hat{\theta}_{ns})(\theta_u - \hat{\theta}_{nu})$$

$$- I^{jr}(\hat{\boldsymbol{\theta}}_n) I^{su}(\hat{\boldsymbol{\theta}}_n) - I^{js}(\hat{\boldsymbol{\theta}}_n) I^{ru}(\hat{\boldsymbol{\theta}}_n) - I^{ju}(\hat{\boldsymbol{\theta}}_n) I^{rs}(\hat{\boldsymbol{\theta}}_n) \Big) a_{jrs} \frac{p_u(\hat{\boldsymbol{\theta}}_n)}{p(\hat{\boldsymbol{\theta}}_n)}$$

$$+ k_*(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n) \Bigg\} + O_p(n^{-\frac{3}{2}}) \Bigg],$$

where k_* involves functions of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_n$ but not p or its derivatives.

The expansion for the posterior is valid for sample points in a set S. The $P_{\boldsymbol{\theta}}-$probability of S is $1 - o(n^{-1})$ uniformly on the compact parameters space. The set S can be constructed in the line of Bickel and Ghosh (1990, Section 3).

Using this expansion, one gets

$$p^{-\alpha}(\boldsymbol{\theta}|\boldsymbol{X}_n) = (2\pi)^{\frac{k\alpha}{2}} n^{-\frac{k\alpha}{2}} |\hat{I}_n|^{1/2} \exp\left[\frac{n\alpha}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)^T \hat{I}_n(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)\right] \tag{50}$$

$$\times \left[1 - \frac{\alpha}{n^{1/2}} \left\{ \sum_{j=1}^{k} \sqrt{n}(\theta_j - \hat{\theta}_{nj}) \frac{p_j(\hat{\boldsymbol{\theta}}_n)}{p(\hat{\boldsymbol{\theta}}_n)} \right.\right.$$

$$\left. + \frac{1}{6} \sum_{1\le j,r,s\le k} n^{\frac{3}{2}} (\theta_j - \hat{\theta}_{nj})(\theta_r - \hat{\theta}_{nr})(\theta_s - \hat{\theta}_{ns}) a_{jrs} \right\}$$

$$+ \frac{\alpha(\alpha+1)}{2n} \left\{ \sum_{1\le j,r\le k} n(\theta_j - \hat{\theta}_{nj})(\theta_r - \hat{\theta}_{nr}) p_j(\hat{\boldsymbol{\theta}}_n) p_r(\hat{\boldsymbol{\theta}}_n)/p^2(\hat{\boldsymbol{\theta}}_n) \right.$$

$$\left. + \frac{1}{3} \sum_{1\le j,r,s,u\le k} n^2 (\theta_j - \hat{\theta}_{nj})(\theta_r - \hat{\theta}_{nr})(\theta_s - \hat{\theta}_{ns})(\theta_u - \hat{\theta}_{nu}) a_{jrs} p_u(\hat{\boldsymbol{\theta}}_n)/p(\hat{\boldsymbol{\theta}}_n) \right\}$$

$$- \frac{\alpha}{2n} \sum_{1\le j,r,\le k} \left[n(\theta_j - \hat{\theta}_{nj})(\theta_r - \hat{\theta}_{nr}) - I^{jr}(\hat{\boldsymbol{\theta}}_n) \right] p_{jr}(\hat{\boldsymbol{\theta}}_n)/p(\hat{\boldsymbol{\theta}}_n)$$

$$- \frac{\alpha}{6n} \sum_{1\le j,r,s,u\le k} \left\{ n^2 (\theta_j - \hat{\theta}_{nj})(\theta_r - \hat{\theta}_{nr})(\theta_s - \hat{\theta}_{ns})(\theta_u - \hat{\theta}_{nu}) \right.$$

$$\left. - I^{jr}(\hat{\boldsymbol{\theta}}_n) I^{su}(\hat{\boldsymbol{\theta}}_n) - I^{js}(\hat{\boldsymbol{\theta}}_n) I^{ru}(\hat{\boldsymbol{\theta}}_n) - I^{ju}(\hat{\boldsymbol{\theta}}_n) I^{rs}(\hat{\boldsymbol{\theta}}_n) \right\} a_{jrs} \frac{p_u(\hat{\boldsymbol{\theta}}_n)}{p(\hat{\boldsymbol{\theta}}_n)}$$

$$\left. + k_{**}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_n) + O_p(n^{-\frac{3}{2}}) \right],$$

where once again k_{**} does not involve the prior p or its derivatives. Now in step 1 of the shrinkage argument, for any arbitrary thrice differentiable prior $\bar{p}$ vanishing outside a compact set, we have

$$\bar{p}(\boldsymbol{\theta}|\boldsymbol{X}_n) = (2\pi)^{-\frac{k}{2}} n^{k/2} |\hat{I}_n|^{1/2} \exp\left[-\frac{n}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)^T \hat{I}_n(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)\right] \tag{51}$$

$$\times \left[1 + n^{-\frac{1}{2}} \left\{ \sum_{j=1}^{k} \sqrt{n}(\theta_j - \hat{\theta}_{nj}) \frac{\bar{p}_j(\hat{\boldsymbol{\theta}}_n)}{\bar{p}(\hat{\boldsymbol{\theta}}_n)} + \frac{1}{6} \sum_{1 \le j,r,s \le k} n^{\frac{3}{2}} (\theta_j - \hat{\theta}_{nj})(\theta_r - \hat{\theta}_{nr})(\theta_s - \hat{\theta}_{ns}) a_{jrs} \right\} \right.$$

$$\left. + O_p(n^{-1}) \right],$$

where $\bar{p}_j(\boldsymbol{\theta})$ is the first derivative of $\bar{p}(\boldsymbol{\theta})$ with respect to θ_j and the $O_p(n^{-1})$ terms involve second derivatives of $\bar{p}$.

Based on the properties of the multivariate normal distribution and noting that a_{jrs} is symmetric in its arguments, one can get the following result from (50) and (51)(omitting terms which integrate out to zero.)

$$\begin{aligned}
M(\boldsymbol{X}) &= \int p^{-\alpha}(\boldsymbol{\theta}|\boldsymbol{X}) \bar{p}(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta} \qquad (52) \\
&= (2\pi)^{\frac{k\alpha}{2}} n^{-\frac{k\alpha}{2}} |\hat{I}_n|^{-\frac{\alpha}{2}} (1-\alpha)^{-\frac{k}{2}} \\
&\times \left[1 - \frac{\alpha}{n(1-\alpha)} \sum_{1 \le j,r, \le k} \frac{p_j(\hat{\boldsymbol{\theta}}_n)}{p(\hat{\boldsymbol{\theta}}_n)} \frac{\bar{p}_r(\hat{\boldsymbol{\theta}}_n)}{\bar{p}(\hat{\boldsymbol{\theta}}_n)} I^{jr}(\hat{\boldsymbol{\theta}}_n) \right. \\
&+ \frac{\alpha(\alpha+1)}{2n} \sum_{1 \le j,r \le k} \frac{p_j(\hat{\boldsymbol{\theta}}_n) p_r(\hat{\boldsymbol{\theta}}_n)}{p^2(\hat{\boldsymbol{\theta}}_n)} I^{jr}(\hat{\boldsymbol{\theta}}_n) \\
&- \frac{\alpha^2}{2n(1-\alpha)} \sum_{1 \le j,r \le k} \left\{ p_{jr}(\hat{\boldsymbol{\theta}}_n)/p(\hat{\boldsymbol{\theta}}_n) \right\} I^{jr}(\hat{\boldsymbol{\theta}}_n) \\
&- \frac{\alpha^2}{2n(1-\alpha)} \sum_{1 \le j,r,s,u \le k} a_{jrs} \left\{ p_u(\hat{\boldsymbol{\theta}}_n)/p(\hat{\boldsymbol{\theta}}_n) \right\} I^{jr}(\hat{\boldsymbol{\theta}}_n) I^{su}(\hat{\boldsymbol{\theta}}_n) \\
&+ \left. \frac{k(\hat{\boldsymbol{\theta}}_n)}{n} + n^{-\frac{3}{2}} k_1(\hat{\boldsymbol{\theta}}, p) \right],
\end{aligned}$$

where $k(\hat{\boldsymbol{\theta}}_n)$ does not involve p or its derivatives but $k_1(\hat{\boldsymbol{\theta}}, p)$ depends on the prior p.

This completes the proof.

6.2 Approximation of $J(p)$ by $\hat{J}(p)$

We consider the functionals $J(p)$ and $\hat{J}(p)$ with the L_1-distance for priors p supported on a compact subset $[a, b]$ of the parameter space Θ. In Section 2.2 an approximation to $\hat{J}(p)$ is obtained (see (15)). The same approximation can be shown to hold for $J(p)$ if we can show that

$$\sqrt{\frac{n}{\log n}}|J(p) - \hat{J}(p)| \text{ is asymptotically negligible.} \tag{53}$$

For proving (53), it is enough to show that

$$\sqrt{\frac{n}{\log n}} \int_{[a,b]} g_n(\theta)p(\theta)d\theta \text{ is negligible} \tag{54}$$

where

$$g_n(\theta) = \int \|p(\cdot|\boldsymbol{x}) - \hat{p}(\cdot|\boldsymbol{x})\|_1 p(\boldsymbol{x}|\theta)d\boldsymbol{x},$$

$\hat{p}(\cdot|\boldsymbol{x})$ is the approximating $N(\hat{\theta}, (nI(\theta))^{-1})$ density and $\|p(\cdot|\boldsymbol{x}) - \hat{p}(\cdot|\boldsymbol{x})\|_1$ is the L_1-distance between $p(\cdot|\boldsymbol{x})$ and $\hat{p}(\cdot|\boldsymbol{x})$.

To prove (54) we show the following:

(A) For some $[a_n, b_n] \subset [a, b]$, $a < a_n < b_n < b$ (a_n close to a, b_n close to b),

$$\sqrt{\frac{n}{\log n}} g_n(\theta) \to 0 \text{ uniformly in } \theta \in [a_n, b_n]$$

and

(B) $\sqrt{\frac{n}{\log n}} \int_a^{a_n} g_n(\theta)p(\theta)d\theta$ and $\sqrt{\frac{n}{\log n}} \int_{b_n}^b g_n(\theta)p(\theta)d\theta$ are negligible.

(A) implies

$$\int_{a_n}^{b_n} \frac{\sqrt{n}}{\sqrt{\log n}} g_n(\theta)p(\theta)d\theta \to 0 \text{ as } n \to \infty.$$

Thus (A) and (B) will imply (54) and hence will imply (53).

In order to prove (A), we need the result stated in Proposition 1 below. We do not present the proof of Proposition 1 here. A proof is given in the supplemental materials. For proving Proposition 1, we assume that Conditions (AI)-(AV) of Ghosh et al. (1982, pp. 416-418) on the densities $p(x|\theta)$ hold for some intervals $[c, d]$ and $[a_0, b_0]$ containing $[a, b]$ in their interiors. We also assume that the prior density $p(\cdot)$ is positive and continuous on $[a, b]$ with a bounded derivative on (a, b).

Proposition 1. Let $X_1, X_2, \ldots, X_n$ be i.i.d. observations with a common density $p(x|\theta)$, $\theta \in \Theta = \mathcal{R}$. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ and $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and let $p(\boldsymbol{x}|\theta)$ denote the density of $\boldsymbol{X}$ under θ. Under the above assumptions on the densities $p(x|\theta)$ and the prior $p(\theta)$, there exist constants $M > 0$, $c > 0$ and $r \geq 1/2$ such that, with $a_n = a + c\sqrt{\frac{\log n}{n}}$ and $b_n = b - c\sqrt{\frac{\log n}{n}}$,

$$P_\theta[\|p(\cdot|\boldsymbol{X}) - \hat{p}(\cdot|\boldsymbol{X})\|_1 \leq Mn^{-1/2}] = 1 - O(n^{-r})$$

uniformly in $\theta \in [a_n, b_n]$ as $n \to \infty$.

We now prove (A) using Proposition 1. Splitting the integral

$$\int \|p(\cdot|\boldsymbol{x}) - \hat{p}(\cdot|\boldsymbol{x})\|_1 p(\boldsymbol{x}|\theta) d\boldsymbol{x}$$

into two parts – one on the set $\{\boldsymbol{x} : \|p(\cdot|\boldsymbol{x}) - \hat{p}(\cdot|\boldsymbol{x})\|_1 > Mn^{-1/2}\}$ and the other on its complement, one can show that

$$\frac{\sqrt{n}}{\sqrt{\log n}} \int \|p(\cdot|\boldsymbol{x}) - \hat{p}(\cdot|\boldsymbol{x})\|_1 p(\boldsymbol{x}|\theta) d\boldsymbol{x} \leq \frac{M}{\sqrt{\log n}} + \frac{2\sqrt{n}}{\sqrt{\log n}} O(n^{-r})$$

uniformly in $\theta \in [a_n, b_n]$. This implies (A) with $a_n = a + c\sqrt{\frac{\log n}{n}}$ and $b_n = b - c\sqrt{\frac{\log n}{n}}$ for some constant $c > 0$.

We now prove part (B). Let p^* be a prior on $\Theta = \mathcal{R}$ and p be its restriction on $[a, b]$. Let I_1 and I_2 denote respectively the first and second integral in part (B), i.e.,

$$I_1 = \sqrt{\frac{n}{\log n}} \int_a^{a_n} g_n(\theta) p(\theta) d\theta \text{ and } I_2 = \sqrt{\frac{n}{\log n}} \int_{b_n}^{b} g_n(\theta) p(\theta) d\theta.$$

We show that for sufficiently small a and sufficiently large b, I_1 and I_2 are negligible for large n. The idea is as follows. We first choose a sufficiently large compact set $[a, b]$ so that I_1 and I_2 are negligible for sufficiently large n and then for this $[a, b]$, apply the result of part (A). We thus show that for large compact sets $[a, b]$, for any prior p^* satisfying some conditions, $J(p)$ with $p =$ Jeffreys prior restricted to $[a, b]$ is greater than or equal to $J(p)$ with $p = p^*$ restricted to $[a, b]$ for sufficiently large n.

We note that, as $g_n(\theta) \leq 2$,

$$I_1 \leq 2c \sup\{p(\theta),\ \theta \in [a, a_n]\} \text{ and } I_2 \leq 2c \sup\{p(\theta),\ \theta \in [b_n, b]\}.$$

If we use these bounds, in order to show that I_1 and I_2 are negligible, we need to assume the following:

(B1) Given any $\epsilon > 0$, there exists a_0, b_0, δ_0 such that for $a \leq a_0, b \geq b_0$,

$$\sup\{p(\theta), \theta \in [a, a + \delta_0]\} < \epsilon \text{ and } \sup\{p(\theta), \theta \in [b - \delta_0, b]\} < \epsilon.$$

The above condition (Condition (B1)) is satisfied if, for example, any one of the following two conditions (B1a) and (B1b) holds.

(B1a) $p^*(\cdot)$ is a proper prior such that $p^*(\theta)$ decreases to zero as $|\theta|$ increases to ∞.

(B1b) $p^*(\cdot)$ is an improper prior such that for some a_1 and b_1,

$$\sup_{\theta \leq a_1} p^*(\theta) < \infty \quad \text{and} \quad \sup_{\theta \geq b_1} p^*(\theta) < \infty.$$

Condition (B1b) holds, e.g., for the improper uniform prior over $\mathcal{R}$ which is the Jeffreys prior for the location parameter case and for which the restriction on $[a, b]$ is uniform density over $[a, b]$.

We note that scale parameter models like $N(0, \sigma^2)$, where the parameter space is $(0, \infty)$, can also be treated as above if we consider the reparameterization $\theta = \log \sigma$. A method that should work more generally is as follows. We truncate the parameter space $(0, \infty)$ on the left, say at $\theta = a > 0$ and extend the parameter space sufficiently to the right, i.e., we consider compact subsets $[a, b]$ of the parameter space for sufficiently large b so that the condition on the prior for I_1 and I_2 to be negligible holds.

Summarizing the above discussion we have the following result about approximating $J(p)$ by $\hat{J}(p)$:

Suppose $X_1, \ldots, X_n$ are iid observations with a common density $p(x|\theta)$, $\theta \in \mathcal{R}$ and $p(x|\theta)$ satisfies the conditions (AI)-(AV) of Ghosh et. al. (1982, pp. 416-418) for each interval $[a_0, b_0]$ and each interval $[c, d]$ in $\mathcal{R}$. Suppose further that θ has a prior density p^* which is positive and continuously differentiable on $\mathcal{R}$ such that condition (B1) above is satisfied by its restrictions p on compact sets. Then, given any $\epsilon > 0$, there exist l and u such that for any interval $[a, b]$ with $a \leq l$ and $b \geq u$, $\sqrt{\frac{n}{\log n}}|J(p) - \hat{J}(p)| < \epsilon$ for all sufficiently large n, where p is the restriction of p^* on $[a, b]$.

This result, in conjunction with the result proved in Section 2.2, gives the following. Consider iid observations with density satisfying the above conditions. Suppose p^* is a prior satisfying the above conditions. Then for large compact sets $[a, b]$, $J(p)$ with $p =$ Jeffreys prior restricted to $[a, b]$ is greater than or equal to $J(p)$ with $p = p^*$ restricted to $[a, b]$ for sufficiently large n.

References

Ali, S. M. and Silvey, S. D. (1966). "A general class of coefficients of divergence of one distribution from another." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 28: 131–142.

Amari, S. (1982). "Differential geometry of curved exponential families - curvatures and information loss." *The Annals of Statistics*, 10: 357–387.

— (1985). *Differential-geometrical Methods in Statistics*. Springer-Verlag.

Bartlett, M. S. (1953). "Approximate Confidence Intervals. II. More than one Unknown Parameter." *Biometrika*, 40: 306–317.

Berger, J. and Bernardo, J. (1989). "Estimating a product of means: Bayesian analysis with reference priors." *Journal of the American Statistical Association*, 84: 200–207.

— (1992a). "On the development of the reference priors." In Bernardo, J., Berger, J., Dawid, A., and Smith, A. (eds.), *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, 35–60. Oxford University Press.

— (1992b). "Ordered group reference priors with application to the multinomial problem." *Biometrika*, 79: 25–37.

Berger, J., Bernardo, J., and Sun, D. (2009). "The formal definition of reference priors." *The Annals of Statistics*, 37: 905–938.

Bernardo, J. (1979). "Reference posterior distributions for Bayesian inference." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 41: 113–147.

Bickel, P. and Ghosh, J. (1990). "decomposition for the likelihood ratio statistic and the Bartlett correction - a Bayesian argument." *The Annals of Statistics*, 18: 1070–1090.

Cichocki, A. and Amari, S. (2010). "Families of Alpha- Beta- and Gamma-Divergences: Flexible and Robust Measures of Similarities." *Entropy*, 12: 1532–1568.

Clarke, B. (1996). "Implications of reference priors for prior information and for sample size." *Journal of the American Statistical Association*, 91: 173–184.

Clarke, B. and Barron, A. (1990). "Information-theoretic asymptotics of Bayes methods." *IEEE Transactions on Information Theory*, 36: 453–471.

— (1994). "Jeffreys' prior is asymptotically least favorable under entropy risk." *Journal of Statistical Planning and Inference*, 41: 37–60.

Clarke, B. and Ghosal, S. (2010). "Reference priors for exponential families with increasing dimension." *Electronic Journal of Statistics*, 4: 737–780.

Clarke, B. and Sun, D. (1997). "Reference priors under the chi-square distance." *Sankhya, Series. A*, 59: 215–231.

— (1999). "Asymptotics of the expected posterior." *Annals of the Institute of Statistical Mathematics*, 51: 163–185.

Clarke, B. and Yuan, A. (2004). "Partial information reference priors: derivation and interpretations." *Journal of Statistical Planning and Inference*, 123: 313–345.

Cressie, N. and Read, T. (1984). "Multinomial Goodness-of-Fit Tests." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 46: 440–464.

Csiszár, I. (1963). "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitt von Markoffschen Ketten." *Magyar Tudományos Akadémia Matematikai Kutatintzet Intézet Kozl*, 8: 85–108.

Datta, G. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*. New York: Springer.

Fraser, D., Reid, N., Marras, E., and Yi, G. (2010). "Default priors for Bayesian and frequentist inference." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72: 631–654.

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1: 515–533.

Ghosal, S., Ghosh, J., and Ramamoorthi, R. (1997). "Noninformative priors via sieves and packing numbers." In Panchapakesan, S. and Balakrishnan, N. (eds.), *Advances in Statistical Decision Theory and Applications*, 119–132. Birkhäuser.

Ghosh, J. (1994). *Higher Order Asymptotics*. Hayward, California: Institute of Mathematical Statistics and American Statistical Association.

Ghosh, J., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis – Theory and Methods*. New York: Springer-Verlag.

Ghosh, J. and Mukerjee, R. (1992). "Non-informative Priors (with discussion)." In Bernardo, J., Berger, J., Dawid, A., and Smith, A. (eds.), *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, 195–210. Oxford University Press.

Ghosh, J. and Ramamoorthi, R. (2003). *Bayesian nonparametrics*. New York: Springer-Verlag.

Ghosh, J., Sinha, B., and Joshi, S. (1982). "Expansion for posterior probability and integrated Bayes risk." In Gupta, S. and Berger, J. (eds.), *Statistical Decision Theory and Related Topics, III(1)*, 403–456.

Ghosh, M. (2011). "Objective Priors: an Introduction for Frequentists." *Statistical Science*, 26: 187–211.

Ghosh, M., Mergel, V., and Liu, R. (2011). "A General Divergence Criterion For Prior Selection." *Annals of the Institute of Statistical Mathematics*, 63(1): 43–58.

Giaquinta, M. (1983). *Multiple integrals in the calculus of variations and nonlinear elliptic systems*. Princeton, New Jersey: Princeton University Press.

Ibragimov, I. and Has'minskii, R. (1981). *Statistical Estimation - Asymptotic Theory*. New York: Springer-Verlag.

Kullback, S. and Leibler, R. (1951). "On Information and Sufficiency." *Annals of Mathematical Statistics*, 22: 79–86.

Lin, X., Pittman, J., and Clarke, B. (2007). "Information conversion, effective samples, and parameter size." *IEEE Transactions on Information Theory*, 53: 4438–4456.

Lindley, D. (1956). "On the measure of the information provided by an experiment." *Annals of Mathematical Statistics*, 27: 986–1005.

Morimoto, T. (1963). "Markov processes and the H-theorem." *Journal of the Physical Society of Japan*, 18: 328–331.

Morito, S., Thall, P., and Mueller, P. (2008). "Determining the effective sample size of a parametric prior." *Biometrics*, 64: 595–602.

— (2010). "Evaluating the impact of prior assumptions in Bayesian Biostatistics." *Statistics in Biosciences*, 2: 1–17.

Shannon, C. (1948). "A mathematical theory of communication." *Bell System Technical Journal*, 27: 379–423 and 623–656.

Weiss, L. and Wolfowitz, J. (1974). *Maximum probability estimators and related topics*. Berlin: Springer-Verlag.

Ye, K. (1993). "Reference priors when the stopping rule depends on the parameter of interest." *Journal of the American Statistical Association*, 88: 360–363.

Yuan, A. and Clarke, B. (2010). "Reference priors for empirical likelihoods." In Chen, M., Muller, P., Sun, D., Ye, K., and Dey, D. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, 56–68.

Zhang, Z. (1994). *Discrete noninformative priors (Ph.D. Thesis)*. Department of Statistics, Yale University.

Acknowledgments

We are thankful to the referees for their very useful comments and constructive suggestions that helped us improve the paper. One of us, J.K. Ghosh, was visiting SAMSI, the NSF funded Statistical and Applied Mathematical Sciences Institute, while the paper was being written. The visit threw light on the meaning of information in a prior.