ISIT'2013 1569710695

Extendable MDL

Peter Harremoës Copenhagen Business College Copenhagen, Denmark E-mail: harremoes@ieee.org

***Abstract*—In this paper we show that a combination of the minimum description length principle and an exchange-ability condition leads directly to the use of Jeffreys prior. This approach works in most cases even when Jeffreys prior cannot be normalized. Kraft's inequality links codes and distributions but a closer look at this inequality demonstrates that this link only makes sense when sequences are considered as prefixes of potential longer sequences. For technical reasons only results for exponential families are stated. Results on when Jeffreys prior can be normalized after conditioning on a initializing string are given. An exotic case where no initial string allow Jeffreys prior to be normalized is given and some way of handling such exotic cases are discussed.**

I. Introduction

A major problem in Bayesian statistics is to assign prior distributions and to justify the choice of prior. The minimum description length (MDL) approach to statistics is often able to overcome this problem, but although MDL may look quite similar to Bayesian statistics the inference is different. One of the main results in MDL is that Jeffreys prior is asymptotically minimax optimal with respect to both redundancy and regret. Despite this positive result there are two serious technical complications that we will address in this paper.

The first complication is that in MDL the use of a code based on Jeffreys prior is normally considered as suboptimal to the use of the normalized maximum likelihood distribution (NML). Jeffreys prior turn out to be optimal if we make a more sequential approach to online prediction and coding. The key idea is to consider *extended sequences*.

The second complication is that in many important applications, Jeffreys prior cannot be normalized. When Jeffreys prior cannot be normalized it is often (but not always) the case that the Shtarkov integral is infinite so that the NML distribution does not exist. This problem is often handled by conditioning by a short sequence of initial data. In Bayesian statistics this has lead to a widespread use of *improper prior distributions* and in MDL it has lead to the definition of the SNML predictor. Our sequential approach will justify the use of improper Jeffreys priors and describe in which sense the use of improper Jeffreys distributions is normally preferable to the SNML predictor.

In the classical frequential approach to statistics a finite sequence is considered as a sub-sequence of an infinite sequence. I Bayesian statistics a finite sequence is normally considered without reference to longer sequences. In this paper we will take a standpoint in between. We will think of a finite sequence as a prefix of *potentially longer finite sequences*. Only in this way we can justify the equivalence between codes and distributions via Kraft's inequality. In this short paper we shall restrict our attention to exponential families to avoid technical complications related to measurablity etc. Despite this restriction our results cover many important applications and the model is still sufficiently flexible to illustrate ideas that can be generalized to a more abstract setting.

The rest of this paper is organized as follows. In Section II notation is fixed and some well-known basic results are stated in the way that we are going to use them. In Section III we will see that the use of Kraft's inequality is relevant if we consider short sequences as sub-sequences of longer sequences. In Section IV we define exponential prediction systems and we will see how such systems are given by prior measures on the parameter space and for which sequences conditional distributions exists. In Section V the optimality of Jeffreys prior is described and some results on when conditional distributions exists are stated. These sections are given in the logical order of reasoning but they can be read quite independently. In the proceeding version of this paper most proofs have been left out. A longer version of this paper with an appendix that contains proofs of all theorems, can be found on arXiv.org . The paper ends with a short discussion.

II. Preliminaries

A. Definitions for exponential families

The exponential family $\{P_\beta \mid \beta \in \Gamma^{\mathrm{can}}\}$ based on the probability measure P_0 is given in a canonical parametrization,

$$\frac{\mathrm{d}P_\beta}{\mathrm{d}P_0} = \frac{\exp(\beta x)}{Z(\beta)}, \beta \in \Gamma^{\mathrm{can}} \tag{1}$$

where Z is the partition function $Z(\beta) = \int \exp(\beta x)\,\mathrm{d}P_0 x$, and $\Gamma^{\mathrm{can}} := \{\beta \mid Z(\beta) < \infty\}$ is the *canonical parameter space*. Note that we allow the measure P_0 to have both discrete and continuous components. The trivial case where Γ^{can} has no interior points is excluded from the analysis. In Equation 1 βx will denote the product of real numbers when the exponential family is 1-dimensional and βx will denote a scalar product when the exponential family has dimension $d > 1$ so that β and x are vectors in $\mathbb{R}^d$. See [1] for more details on exponential families.

For our problem it is natural to work with *extended exponential families* as defined in [2]. For a probability distribution Q on $\mathbb{R}^k$ the convex support $cs(Q)$ is the intersection of all convex *closed* sets that have Q-probability 1. The convex core $cc(Q)$ is the intersection of all convex *measurable* sets with Q-probability 1, [3]. We have $cc(Q) \subseteq cs(Q)$. An extreme point x in $cs(Q)$ belongs to $cc(Q)$ if and only if $Q(x) > 0$. In its mean value parametrization the exponential family based on a measure with bounded support has a natural extension to $cc(Q)$. In particular δ_x belongs to the extended exponential family if Q has a point mass in x and x is an extreme point of $cs\left(Q\right)$.

The elements of the exponential family are also parametrized by their mean value μ. We write μ_β for the mean value corresponding to the canonical parameter β and $\hat{\beta}\left(\mu\right)$ for the canonical parameter corresponding to the mean value μ. Note that we allow infinite values of the mean. The element in the exponential family with mean μ is denoted P^μ. The mean value range M of the exponential family is the range of $\beta \to \mu_\beta$ and is a subset of the convex core. For 1-dimensional families we write $\mu_{\sup} = \sup M$, and $\mu_{\inf} = \inf M$. If P_0 has a point mass at $\mu_{\inf} > -\infty$ and the support of P_0 is a subset of $[\mu_{\inf}, \infty[$, then the exponential family is extended by the element $P_{-\infty} = P^{\mu_{\inf}} = \delta_{\mu_{\inf}}$, and likewise the exponential family is extended if Q has a point mass in $\mu_{\sup} < \infty$ and the support of Q is a subset of $]\text{-}\infty, \mu_{\sup}]$. For any x the distribution $P_{\hat{\beta}(x)} = P^x$ is the maximum likelihood distribution.

The covariance function V is the function that maps $\mu \in M$ into the covariance of P^μ. If M has interior points then the exponential family is uniquely determined by its covariance function. The Fisher information of an exponential family in its canonical parametrization is $I_\beta = V\left(\mu_\beta\right)$ and the Fisher information of the exponential family in its mean value parametrization is $I^\mu = V\left(\mu\right)^{\text{-}1}$.

For elements of an exponential family we introduce *information divergence* as

$$D\left(x\| y\right) := D\left(P^x\| P^y\right) = \int \ln\left(\frac{\mathrm{d}P^x}{\mathrm{d}P^y}\right) \mathrm{d}P^x.$$

This defines a *Bregman divergence* on the mean value range and under some regularity conditions this Bregman divergence uniquely characterizes the exponential family [4].

B. Posterior distributions

If the mean value parameter has prior distribution ν and x has been observed then the posterior distribution has density

$$\frac{\mathrm{d}\nu\left(\cdot|x\right)}{\mathrm{d}\nu}\left(y\right) \sim \exp\left(-D\left(x\| y\right)\right).$$

Notation We use x^m to denote $\left(x_1, x_2, \ldots, x_m\right)$ and x_m^n to denote $\left(x_m, x_{m+1}, \ldots, x_n\right)$. We use τ as short for 2π and $\sim$ to denote that two functions or measures are proportional.

If a sequence $x_1, x_2, \ldots, x_m$ has been observed then the posterior distribution has density

$$\begin{aligned}\frac{\mathrm{d}\nu\left(\cdot|x^m\right)}{\mathrm{d}\nu}\left(y\right) &\sim \prod_{i=1}^m \exp\left(\text{-}D\left(x_i\| y\right)\right)\\ &= \left(\prod_{i=1}^m \exp\left(-D\left(x_i\| \bar{x}\right)\right)\right)\cdot \exp\left(\text{-}nD\left(\bar{x}\| y\right)\right)\end{aligned}$$

where $\bar{x}$ denotes the average of the sequence x^m, where we have an equality that is of general validity for Bregman divergences. Since the first factor does not depend on y we have

$$\frac{d\nu\left(\cdot\,|x^m\right)}{d\nu}\left(y\right) \sim \exp\left(-mD\left(\bar{x}\| y\right)\right).$$

C. MDL in exponential families

For some exponential families the *minimax regret* C_∞ is finite. See [5] for details about how this quantity is defined. If C_∞ is finite the minimax regret is assumed if we code according to the *NML distribution*. In general the optimal code for X_1 will depend on whether the sample size is $n = 1$ or whether X_1 is considered as a sub-sequence of X^n. In cases where C_∞ is infinite one may use a conditional versions instead such as *sequential NML* (SNML).

Of central importance for our approach are result developed by Barron, Rissanen et al. that if the parameter space of an exponential family is restricted to a non-empty compact subset of the interior of the convex core, then the minimax regret is finite and equal to

$$C_\infty = \frac{d}{2}\ln\frac{n}{\tau} + \ln J + o(1), \tag{2}$$

where J denotes the *Jeffreys integral*

$$J = \int_{\Gamma^{\text{can}}} \left(\det I_\beta\right)^{1/2}\,\mathrm{d}\beta = \int_M \left(\det V\left(x\right)\right)^{\text{-}1/2}\,\mathrm{d}x. \tag{3}$$

where I_β denotes the *Fisher information matrix* [5]. Moreover, the same asymptotic regret (2) is achieved by the Bayesian marginal distribution equipped with Jeffreys prior. In MDL this result is often used as the most important reason for using *Jeffreys prior* with density $w(\mu) = \left(\det V\left(\mu\right)\right)^{\text{-}1/2}/J$. The use of the NML predictor requires knowledge of the sample size and the performance of the SNML predictor will depend on the order of the observations except if it corresponds to the use of Jeffreys prior [6].

If the parameter space is restricted to a non-empty compact subset of the interior of the convex core (called an *ineccsi* set in [5]) the Jefftreys integral is automatically finite but typically there is no natural way of restricting the parameter space in applications and in most cases the Jeffreys integral is infinite. It thus becomes quite relevant to investigate what happens if the parameter space is *not* restricted to an ineccsi set. To answer this question, one needs to know when the Jeffreys integral is finite, and how to handle situations where Jeffreys integral is not finite.

D. Exchangeability, sufficiency, and consistency

Prediction in exponential families satisfy the exchangeability condition that the probability of sequence does not depend on the order of the elements. We may also say the predictor is invariant under permutations of the elements in a sequence. The importance of this exchangablity condition in MDL was emphasized in [6]. A related but more important type of exchangablity is that the probability of a sequence given a sub sequence x^j does not depend on the order of the observations in the sub-sequence. A stronger requirement is that the predicted probability of a sequence given a sub-sequence x^j only depends the average $\bar{x}$ of the sub-sequence, i.e. the sample average is a *sufficient statistic*. We are also interested in consistency of the system of predictors. Note that $P\left(x_{\ell+1}^n \middle| x^\ell\right) = P\left(x^n \middle| x^\ell\right)$. A system of predictors is *consistent* if the prediction $P\left(x^n \middle| x^\ell\right) = P\left(x^n \middle| x^m\right)\cdot P\left(x^m \middle| x^\ell\right)$.

A consistent system of predictors is generated from predictions of the next symbol given by the past symbols.

III. MDL AND KRAFT'S INEQUALITY

We recall that a code is uniquely decodable if any finite sequence of input symbols give a unique sequence of output symbols. It is well-known that a uniquely decodable code satisfies Kraft's inequality

$$\sum_{a\in\mathbb{A}}\beta^{-\ell(a)}\leq 1 \tag{4}$$

where $\ell(a)$ denotes the length of the codeword corresponding to the input symbol $a\in\mathbb{A}$ and β denotes the size of the output alphabet. The length of a codeword is an integer. Normally the use of non-integer valued code length functions is justified by reference to the noiseless coding theorem which require some interpretation of the notion of probability distributions and their mean values. To emphasize our sequential point of view we formulate a version of Kraft's inequality that allow the code length function to be non-integer valued.

Theorem 1. *Let $\ell:\mathbb{A}\to\mathbb{R}$ be a function. Then the function ℓ satisfies Kraft's inequality (4) if and only if for all $\varepsilon>0$ there exists an integer n and a uniquely decodable fixed-to-variable length block code $\kappa:\mathbb{A}^n\to\mathbb{B}^*$ such that*

$$\left|\bar{\ell}_\kappa(a^n)-\frac{1}{n}\sum_{i=1}^{n}\ell(a_i)\right|\leq\varepsilon$$

where $\bar{\ell}_\kappa(a^n)$ denotes the length $\ell_\kappa(a^n)$ divided by n. The uniquely decodable block code can be chosen to be prefix free.

It is only possible to obtain a unique correspondence between code length functions and (discrete) probability measures by considering codewords as prefixes of potentially longer codewords. If we restrict our attention to code words of some finite fixed length then Kraft's inequality does not give a necessary and sufficient condition of decodability. Like in Bayesian statistics we focus on finite sequences. Contrary to Bayesian statistics we should always consider a finite sequence as a prefix of *longer finite* sequences. Contrary to frequential statistics we do not have to consider a finite sequence as a prefix of an *infinite* sequence.

If the set of input symbols is not discrete one has to introduce some type of distortion measure, but we will abstain from discussing this complication in this short note.

IV. IMPROPER PRIORS

In this section we will talk about a prior measure even when it cannot be normalized and we will call it a *proper prior* when it can be normalized to a probability measure.

A. Finiteness structure

If a sequence of length m with average $\bar{x}$ is observed then the prior integral is either finite or infinite. Let F_m denote the subset of average values in the convex core such that the prior integral is finite for samples of size m.

Theorem 2. *The sets F_m form an increasing sequence of convex subsets of the convex core, i.e. $F_1\subseteq F_2\subseteq F_3\subseteq\dots cc$.*

Example 3. Consider the Gaussian location family. For this family $D(y\|x)=\frac{(x-y)^2}{2}$. If the prior has density $\exp(\alpha x^2)$, then the prior can be normalized to a posterior distribution when

$$\int_{-\infty}^{\infty}\exp\left(\alpha x^2\right)\exp\left(-m\frac{(x-y)^2}{2}\right)\mathrm{d}x$$

so the integral is finite when $m>2\alpha$.

If the prior has density $\exp\left(x^4\right)$ then there exists no m for which the prior can be normalized.

Theorem 4. *Assume that $x_1\in F_m$ and μ_0 is in the convex core. Then $\left(1-\frac{m}{n}\right)x_0+\frac{m}{n}\mu_1\in F_n$.*

An important special case is when the convex core equals $\mathbb{R}^d$. In this case we have that if $F_n\neq\emptyset$ then $F_{n+1}=\mathbb{R}^k$.

The next example shows that Theorem 4 is 'tight'.

Example 5. The family of exponential distributions has $D(\lambda\|\mu)=\frac{\lambda}{\mu}-1-\ln\frac{\lambda}{\mu}$. Consider the prior density $\exp\left(x^{-1}\right)\cdot x^{-2}$. The conditional integral is

$$\int_0^\infty\exp\left(x^{-1}\right)x^{-2}\cdot\exp\left(-m\left(\frac{\bar{x}}{x}-1-\ln\frac{\bar{x}}{x}\right)\right)\mathrm{d}x.$$

The integral $\int_1^\infty\exp\left(x^{-1}\right)\cdot x^{-2}\,\mathrm{d}x$ is finite so we only have to consider the integral

$$\int_0^1\exp\left(x^{-1}\right)x^{-2}\cdot\exp\left(-m\left(\frac{\bar{x}}{x}-1-\ln\frac{\bar{x}}{x}\right)\right)\mathrm{d}x$$
$$=\bar{x}^n\exp(n)\int_0^1\exp\left((1-m\bar{x})\,x^{-1}\right)\cdot x^n x^{-2}\,\mathrm{d}x\,.$$

The substitution $y=x^{-1}$ gives

$$\int_0^1\exp\left((1-m\bar{x})\,x^{-1}\right)\cdot x^n x^{-2}\,\mathrm{d}x$$
$$=\int_1^\infty\exp\left((1-m\bar{x})\,y\right)\cdot y^{-n}\,\mathrm{d}y\,.$$

We see that for $n>1$ the integral is finite if and only if $\bar{x}\geq 1/m$, which implies that $F_m=[1/m,\infty[$.

B. Existence of a prior

We will now define an *exponential prediction system*. We consider a sequence of variables $X_1,X_2,\dots$ with values in $\mathbb{R}^d$. For some sequences of outcomes x^m a probability measure $P(\cdot|x^m)$ on $\mathbb{R}^d$ is given and the interpretation of this probability measure is that it gives the probability or prediction of the next variable X_{m+1} given the values of the previous variables. Equivalently we may think of $P(\cdot|x^m)$ as an instruction about how the next variable should be coded given the values of the previous variables. Further we will assume that if $P(\cdot|x^m)$ is defined then $P(\cdot|x^n)$ is also defined for any sequence x^n with x^m as prefix. Further we will assume that the sum is sufficient for prediction, i.e. $P(\cdot|x^m)$ only depends on the value of the sum $x_1+x_2+\dots+x_m$.

An exponential prediction system as described above can be extended to a consistent prediction system for sequences and we note that the sum is still sufficient for predicting sequences. Conversely, a consistent prediction system for sequences can be reconstructed from its restriction to predictions of the next symbol.

Assume that $P(\cdot|x^m)$ exists. Then we have a consistent system of probability measures on the variables $X_{m+1}, X_{m+2}, \ldots$ for which the sums of the previous variables are sufficient statistics for the following variables. According to results of S. Lauritzen any such system is a mixture of elements in an exponential family when the predictor is defined even for initial sequences of length $m = 0$ [7]. Therefore there exists a measure P_0 and a probability measure ν_{x^n} over the convex core such that

$$\frac{\mathrm{d}P(\cdot|x^m)}{\mathrm{d}P_0}(x) = \int_{cc} \frac{\exp\left(x \cdot \hat{\beta}(y)\right)}{Z\left(\hat{\beta}(y)\right)} \mathrm{d}\nu_{x^m} y.$$

These 'prior distributions' ν_{x^n} are updated to 'posterior distributions' in the usual fashion

$$\frac{\mathrm{d}\nu_{x^{m+1}}}{\mathrm{d}\nu_{x^m}}(x) \sim \exp\left(\text{-}D\left(x_{m+1}\| x\right)\right).$$

The following theorem extends results of S. Lauritzen to cases where $m > 0$.

Theorem 6. *For an exponential prediction system there exists an exponential family based on a probability measure P_0 and a prior measure η over the mean value range M of the exponential family such that*

$$\frac{\mathrm{d}P(\cdot|x^m)}{\mathrm{d}P_0}(x) = \int_M \frac{\exp\left(x \cdot \hat{\beta}(z)\right)}{Z\left(\hat{\beta}(z)\right)} \frac{\exp\left(\text{-}mD\left(\bar{x}\| z\right)\right)}{\int_M \exp\left(\text{-}mD\left(\bar{x}\| z\right)\right) \mathrm{d}\eta z} \mathrm{d}\eta z.$$

V. Jeffreys prior

A. Conditional regret

We will use conditional regret to evaluate the quality of a predictor. For a conditional setup Peter Grünwald has defined three different notions of conditional regret [5, subsection 11.4.2]. First we assume that the sample space is finite. We let P^t denote a distribution in the exponential family and we compare it with a predictor $Q(\cdot|\cdot)$. If a sequence x^n is observed then the optimal code based on an element in the exponential family would provide code-length $-\ln P^t(x^n)$. In order to code the same sequence using a predictor $Q(\cdot|\cdot)$ when the initial string x^m has been observed, the code length for the rest of the sequence is $\text{-}\ln Q(x^n|\, x^m)$. The *regret-2* is defined as the difference

$$\text{-}\ln Q(x^n|\, x^m) - \left(\text{-}\ln P^t(x^n)\right).$$

If the optimal distribution from the exponential family is used the regret of the predictor with respect to the sequence is

$$\begin{aligned} REG_Q(x^n|\, x^m) &= -\ln Q(x^n|\, x^m) - \left(-\ln P^{\bar{x}}(x^n)\right). \\ &= \ln \frac{P^t(x^n)}{Q(x^n|\, x^m)}. \end{aligned}$$

If the sample space is not finite then we replace probabilities with densities with respect to a fixed measure P_0 in the exponential family.

Kraft's inequality implies that one code based on a probability measure cannot have shorter codewords than another code for all outcomes. The following theorem states that something similar holds for consistent predictors.

Theorem 7. *Let Q_1 and Q_2 denote two different exponential prediction systems for the same exponential family. Then there exist a sequence $x_1, x_2, \ldots$ and a number m such that*

$$\lim_{n\to\infty} \inf \left(REG_{Q_2}(x^n|\, x^m) - REG_{Q_1}(x^n|\, x^m)\right) > 0.$$

B. Optimality of Jeffreys prior

We are now able to combine our sequential approach with existing results on optimality of Jeffreys prior.

Theorem 8. *Assume that (P^x) is a exponential family based on the probability measure P_0 and that $Q(\cdot|\cdot)$ denotes an exponential prediction system based on the probability measure Q_0 with prior measure ν on the mean value range M.*

If $Q_0 = P_0$ and the support of the prior measure ν equals the closure of the mean value range of the exponential family, then for any P^x in the extended exponential family with x in the convex core and any sequence $x_1, x_2, \ldots$ satisfying

$$\lim\inf D\left(P^{\bar{x}}\middle\| P^x\right) > 0$$

then the conditional regret-2 of the exponential prediction system $Q(\cdot|\cdot)$ is eventually less than the conditional regret of P^x with respect to the sequence $x_1, x_2, \ldots$

Exponential prediction systems based on P_0 and with dense prior are the only exponential prediction systems satisfying this property.

Further conditions are needed in order to single out the Jeffreys prior. The conditional Jeffreys integral is defined as

$$J \mid x^m = \int \frac{\exp\left(\text{-}mD\left(P^{\bar{x}}\| P^x\right)\right)}{\left(\det V(x)\right)^{1/2}} \mathrm{d}x$$

where $\bar{x}$ is the sample average of x^m. The following theorem states that an exponential prediction system is asymptotically optimal with respect to minimax regret if and only if it is based on Jeffreys prior. A proof of essentially the same theorem can be found in [5].

Theorem 9. *If an exponential prediction system Q is based on Jeffreys prior and an element P^x in the exponential family corresponding to an interior point x in the convex core and $x_1 x_2 \ldots$ is a sequence such that $x_n \to x$ then*

$$\lim_{n\to\infty}\left(REG_Q(x^n|\, x^m) - \frac{k}{2}\ln\frac{n}{\tau}\right) = \ln\left(J\,|x^m\right).$$

Since Jeffreys prior has regret that is asymptotically constant and since according to Theorem 7 one prediction system cannot be uniformly better than another we see that an exponential prediction system based on Jeffreys prior is optimal with respect to regret in the following sense.

Corollary 10. *For any exponential prediction system there exists an element P^x in the exponential family corresponding to an interior point x in the convex core and a sequence $x_1 x_2 \ldots$ such that $x_n \to x$ such that the regret of the exponential prediction system satisfies*

$$\lim_{n\to\infty} \inf \left(REG\left(x^n \middle| x^m\right) - \frac{k}{2} \ln \frac{n}{\tau} \right) \geq \ln\left(J \left| x^m\right.\right).$$

This theorem has important consequences. For instance it becomes much easier to prove the recent result that the SNML predictor is exchangeable if and only if it is equivalent to the use for Jeffreys prior [6].

C. When is conditional Jeffreys Finite?

After having identified Jeffreys prior as optimal it is of interest to see how long sequences are needed before the conditional Jeffreys integral becomes finite. Most exponential families used in applications have finite conditional Jeffreys integral after just one sample point. For a one dimensional exponential family one can divide the parameter interval into a left part and a right part and treat these independently. The following results seem to cover all cases relevant for applications.

Theorem 11. *Let Q be a measure for which the convex core is lower bounded. Assume that a is the left end point of M. If Q has density $f(x) = (x-a)^{\gamma-1} g(x)$ in an interval just to the right of a where g is an analytic function and $g(a) > 0$ then the conditional Jeffreys integral of the right truncated exponential family is finite.*

Grünwald and Harremoës have previously shown that under the conditions of the previous theorem if there is a point mass in a then the unconditional Jeffreys integral is also finite [8].

Theorem 12. *Let (Γ_0^{can}, Q) represent a left-truncated exponential family that is light tailed in the sense that there exists a Gamma exponential family such that the variance function V of (Γ_0^{can}, Q) satisfy*

$$\lim_{x\to\infty} \inf \frac{V(x)}{V_\gamma(x)} > 0$$

then the conditional Jeffreys integral is finite where $V_\gamma(x)$ denotes the variance function of the gamma exponential family.

The following theorem extends a theorem from [8].

Theorem 13. *Let (Γ_0^{can}, Q) represent a left-truncated exponential family such that $\beta_{\sup} = 0$ and Q admits a density q either with respect to Lebesgue measure or counting measure. If q is heavy tailed the Jeffreys integral is finite, if and only if all the conditional Jeffreys integrals are finite. If $q(x) = O(x^{-1-\alpha})$ for some $\alpha > 0$, then Jeffreys integral $\int_M V(x)^{-1/2}\,\mathrm{d}x$ is finite.*

Most exponential families with finite minimax regret also have finite Jeffreys integral but there are counter examples and they give exponential families for which the Jeffreys integral is always infinite.

Example 14. If Y is a Cauchy distributed random variable then $X = \exp(Y)$ has a very heavy tailed distribution that we will call a exponentiated Cauchy distribution. A probability measure Q is defined as a $1/2$ and $1/2$ mixture of a point mass in 0 and an exponentiated Cauchy distribution. As shown by Grünwald and Harremoës [8] the exponential family based on Q has finite minimax regret, but infinite Jeffreys integral. Since the minimax regret is finite the divergence is bounded and the conditional Jeffreys integrals are all infinite for any initial sequence x^m of any length.

VI. Discussion

The notion of sufficiency has been generalized by S. Lauritzen [7] and generalizations of his results to the setting presented here is highly relevant but cannot be covered in this short note. In cases where the Jeffreys integral is infinite and the minimax regret is finite one cannot find an optimal exponential prediction system, so exchangeability cannot be achieved. In such cases the usual NML predictor or the SNML predictor may be good alternatives. Much of what has been said here about regret will also hold for mean redundancy [9] or for any capacity of order α as defined in [10].

Acknowledgment

The author thank Peter Grünwald, Fares Hedayati, Wojciech Kotłowski and Peter Bartlett for stimulating discussions. Peter Grünwald and Wojciech Kotłowski have also provided useful comments to a previous version of this manuscript.

References

[1] O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory*. New York: John Wiley, 1978.

[2] I. Csiszár and F. Matús, "Information projections revisited," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1474–1490, June 2003.

[3] I. Csiszár and F. Matús, "Convex cores of measures on $\mathbb{R}^d$," *Studia Sci. Math. Hungar.*, vol. 38, pp. 177–190, 2001.

[4] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[5] P. Grünwald, *the Minimum Description Length principle*. MIT Press, 2007.

[6] F. Hedayati and P. Bartlett, "The optimality of jeffreys prior for online density estimation and asymptotic normality of manximum likelihood estimators.," in *Proccedings of the Twenty Fifth Conference on learning Theory (COLT' 04)*, 2012.

[7] S. L. Lauritzen, *Statistical Models as Extremal Families*. Aalborg: Aalborg Univ. Press, 1982.

[8] P. Grünwald and P. Harremoës, "Finiteness of redundancy, regret, Shtarkov sums, and Jeffreys integrals in exponential families," in *Proceedings for the International Symposium for Information Theory, Seoul, 2009*, pp. 714–718, IEEE, June 2009.

[9] F. Liang and A. R. Barron, "Exact minimax strategies for predictive density estimation, data compression, and model selection," *IEEE Tranns. Inform. Theory*, vol. 50, pp. 2708–2726, 2004.

[10] T. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence." Submitted, June 2012.