The Optimality of Jeffreys Prior for Online Density Estimation and the Asymptotic Normality of Maximum Likelihood Estimators

Fares Hedayati FARESHED@EECS.BERKELEY.EDU
University of California at Berkeley

Peter L. Bartlett BARTLETT@CS.BERKELEY.EDU
University of California at Berkeley,
Queensland University of Technology

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

We study online learning under logarithmic loss with regular parametric models. We show that a Bayesian strategy predicts optimally only if it uses Jeffreys prior. This result was known for canonical exponential families; we extend it to parametric models for which the maximum likelihood estimator is asymptotically normal. The optimal prediction strategy, normalized maximum likelihood, depends on the number n of rounds of the game, in general. However, when a Bayesian strategy is optimal, normalized maximum likelihood becomes independent of n. Our proof uses this to exploit the asymptotics of normalized maximum likelihood. The asymptotic normality of the maximum likelihood estimator is responsible for the necessity of Jeffreys prior.

Keywords: Online Learning, Logarithmic Loss, Bayesian Strategy, Jeffreys Prior, Asymptotic Normality of Maximum Likelihood Estimator

1. Introduction

In the online learning setup, the goal is to predict a sequence of outcomes, revealed one at a time, almost as well as a set of experts. We consider online density estimators with log loss, where the forecaster's prediction at each round takes the form of a probability distribution over the next outcome, and the loss suffered is the negative logarithm of the forecaster's probability of the outcome. The aim is to minimize the regret, which is the difference between the cumulative loss of the forecaster (that is, the sum of these negative logarithms) and that of the best expert in hindsight. The optimal strategy for sequentially assigning probability to outcomes is known to be normalized maximum likelihood (NML) (see, for e.g., Cesa-Bianchi and Lugosi, 2006; Grunwald, 2007, and see Definition 4 below). NML suffers from two major drawbacks: the horizon n of the problem needs to be known in advance, and the strategy can be computationally expensive since it involves marginalizing over subsequences. In this paper, we investigate the optimality of two alternative strategies, namely the Bayesian strategy and the sequential normalized maximum likelihood strategy; see Definitions 5 and 6 below. Bayesian prediction under Jeffreys prior has been shown to be asymptotically optimal (see, for e.g., Grunwald, 2007, chaps 7,8). Moreover the regret of SNML is within a constant of the minimax optimal (Kotlowski and Grunwald, 2011). We show that for a very general class of parametric models (Definition 1), optimality of a Bayesian strategy means that the strategy uses Jeffreys prior. Furthermore we show that optimality of the Bayesian strategy is equivalent to optimality of sequential normalized maximum likelihood. The major regularity condition for these parametric families is that the maximum likelihood estimate is asymptotically normal. This classical condition holds for a broad class of parametric models.

2. Definitions and Notation

We work in the same setup of (Hedayati and Bartlett, 2012) and use their definitions and notation. The goal is to predict a sequence of outcomes $x_t \in \mathcal{X}$, almost as well as a set of experts. We use x^t to denote $(x_1, x_2, \cdots, x_t)$, x^0 to denote the empty sequence, and x_m^n to denote $(x_m, x_{m+1}, \cdots, x_n)$. At round t, the forecaster's prediction is a conditional probability density $q_t(\cdot|x^{t-1})$, where the density is with respect to a fixed measure λ on $\mathcal{X}$. For example, if $\mathcal{X}$ is discrete, λ could be the counting measure; for $\mathcal{X} = \mathbb{R}^d$, λ could be Lebesgue measure. The loss that the forecaster suffers at that round is $-\log q_t(x_t \mid x^{t-1})$, where x_t is the outcome revealed after the forecaster's prediction. The difference between the cumulative loss of the prediction strategy and the best expert in a reference set is called the regret. The goal is to minimize the regret in the worst case over all possible data sequences. In this paper, we consider i.i.d. parametric constant experts parametrized by $\theta \in \Theta$.

Definition 1 (Parametric Constant Model) *A constant expert is an iid stochastic process, that is, a joint probability distribution p on sequences of elements of $\mathcal{X}$ such that for all $t > 0$ and for all x in $\mathcal{X}$, $p\left(x^t \left| x^{t-1}\right.\right) = p\left(x_t\right)$. A parametric constant model $(\Theta, (\mathcal{X}, \Sigma), \lambda, p_\theta)$ is a parameter set Θ, a measurable space $(\mathcal{X}, \Sigma)$, a measure λ on $\mathcal{X}$, and a parameterized function $p_\theta : \mathcal{X} \to [0, \infty)$ for which, for all $\theta \in \Theta$, p_θ is a probability density on X with respect to λ. It defines a set of constant experts via $p_\theta\left(x^t \left| x^{t-1}\right.\right) = p_\theta\left(x_t\right)$.*

For convenience, we will often refer to a parametric constant model as just p_θ.

A strategy q is any sequential probability assignment $q_t(\cdot \mid x^{t-1})$ that, given a history x^{t-1}, defines the conditional density of $x_t \in \mathcal{X}$ with respect to the measure λ. It defines a joint distribution q on sequences of elements of $\mathcal{X}$ in the obvious way,

$$q(x^n) = \prod_{t=1}^{n} q(x_t|x^{t-1}).$$

In general, a strategy depends on the sequence length n. We denote such strategies by $q^{(n)}$.

Definition 2 (Regret) *The regret of a strategy $q^{(n)}$ on sequences of length n with respect to a parametric constant model p_θ is*

$$R(x^n, q^{(n)}) = \sum_{t=1}^{n} -\log q_t^{(n)}(x_t|x^{t-1}) - \inf_{\theta\in\Theta} \sum_{t=1}^{n} -\log p_\theta(x_t|x^{t-1}) = \sup_{\theta\in\Theta} \log \frac{p_\theta(x^n)}{q^{(n)}(x^n)}$$

We consider a generalization of the regret of Definition 2. This is because some strategies are only defined conditioned on a fixed initial sequence of observations x^{m-1}. For such cases, we define the conditional regret of x^n, given a fixed initial sequence x^{m-1}, in the following way (see Grunwald, 2007, chap. 11).

Definition 3 (Conditional Regret)

$$R^{\Theta}(x_m^n, q^{(n)}|x^{m-1}) = \sum_{t=m}^{n} -\log q_t(x_t|x^{t-1}) - \inf_{\theta\in\Theta} \sum_{t=1}^{n} -\log p_\theta(x_t|x^{t-1})$$
$$= \sup_{\theta\in\Theta} \log \frac{p_\theta(x^n)}{q^{(n)}(x_m^n \mid x^{m-1})}.$$

Notice that the strategy $q^{(n)}$ defines only the conditional distribution $q^{(n)}(x_m^n \mid x^{m-1})$. We call such a strategy a conditional strategy. In what follows, where we consider a conditional strategy, we assume that x^{m-1} is such that these conditional distributions are always well defined.

Definition 4 (NML) *Given a fixed horizon n, the normalized maximum likelihood (NML) strategy is defined via the joint probability distribution*

$$p_{nml}^{(n)}(x^n) = \frac{\sup_{\theta\in\Theta}\ p_\theta(x^n)}{\int_{\mathcal{X}^n} \sup_{\theta\in\Theta} p_\theta(y^n)\, d\lambda^n(y^n)},$$

provided that the integral in the denominator exists. For $t \le n$, the conditional probability distribution is

$$p_{nml}^{(n)}(x_t \mid x^{t-1}) = \frac{p_{nml}^{(n)}(x^t)}{p_{nml}^{(n)}(x^{t-1})},$$

where $p_{nml}^{(n)}(x^t)$ and $p_{nml}^{(n)}(x^{t-1})$ are marginalized joint probability distributions of $p_{nml}^{(n)}(x^n)$:

$$p_{nml}^{(n)}(x^t) = \int_{\mathcal{X}^{n-t}} p_{nml}^{(n)}(x^n)\, d\lambda^{n-t}(x_{t+1}^n).$$

The regret of the NML strategy achieves the minimax bound, that is, $q^{(n)} = p_{nml}^{(n)}$ minimizes $\max_{x^n} R(x^n, q^{(n)})$ (see, for e.g., Grunwald, 2007, chap. 6). Note that $p_{nml}^{(n)}$ might not be defined if the normalization is infinite. In many cases, for a sequence x^{m-1} and for all $n \ge m$, we can define the conditional probabilities

$$p_{nml}^{(n)}(x_m^n|x^{m-1}) = \frac{\sup_{\theta\in\Theta}\ p_\theta(x^n)}{\int_{\mathcal{X}^{n-m+1}} \sup_{\theta\in\Theta} p_\theta(x^n)\, d\lambda^{n-m+1}(x_m^n)}.$$

For these cases the conditional NML again attains the minimax bound, that is, $q^{(n)} = p_{nml}^{(n)}$ minimizes $\max_{x_m^n} R(x_m^n, q^{(n)} \mid x^{m-1})$ (see Grunwald, 2007, chap. 11). In both cases, the nml strategy is an equalizer, meaning that the regrets of all sequences of length n are equal.

Definition 5 (SNML) *The sequential normalized maximum likelihood (SNML) strategy has*

$$p_{snml}(x_t \mid x^{t-1}) = \frac{\sup_{\theta\in\Theta}\ p_\theta(x^t)}{\int_{\mathcal{X}} \sup_{\theta\in\Theta} p_\theta(x^t)\, d\lambda(x_t)}.$$

Notice that this update does not depend on the horizon. Under mild conditions, the regret of SNML is no more than a constant (independent of n) larger than the minimax regret (Kotlowski and Grunwald, 2011). Once again, p_{snml} is not defined if the integral in the denominator is infinite. In many cases, for a sequence x^{m-1} and for all $n \ge m$, the appropriate conditional probabilities are properly defined. We restrict our attention to these cases.

Definition 6 (Bayesian) *For a prior distribution π on Θ, the Bayesian strategy with π is defined as*

$$p_\pi(x^t) \; = \int_{\theta \in \Theta} p_\theta(x^t)\, d\pi(\theta).$$

The conditional probability distribution is defined in the obvious way,

$$p_\pi(x_t \mid x^{t-1}) = \frac{p_\pi(x^t)}{p_\pi(x^{t-1})}.$$

We denote the conditional Bayesian strategy for a fixed x^{m-1} as $p_\pi(x_m^n \mid x^{m-1})$.

Jeffreys prior (Jeffreys, 1946) has the appealing property that it is invariant under reparameterization.

Definition 7 (Jeffreys prior) *For a parametric model p_θ, Jeffreys prior is the distribution over the parameter space Θ that is proportional to $\sqrt{|I(\theta)|}$, where I is the Fisher information at θ (that is, the variance of the score, $\partial/\partial\theta \ln p_\theta(X)$, where X has density p_θ).*

Our main theorem uses the notion of exchangeability of stochastic processes.

Definition 8 (Exchangeable) *A stochastic process is called* exchangeable *if the joint probability does not depend on the order of observations, that is, for any $n > 0$, any $x^n \in \mathcal{X}^n$, and any permutation σ on $\{1, \ldots, n\}$, the probability of x^n is the same as the probability of x^n permuted by σ.*

When we consider the conditional distribution $p(x_m^n \mid x^{m-1})$ defined by a conditional strategy, we are interested in exchangeability of the conditional stochastic process, that is, invariance under any permutation that leaves x^{m-1} unchanged.

The asymptotic normality of the maximum likelihood estimator is the major regularity condition of the parametric models that is required for our main result to hold.

Definition 9 (Asymptotic Normality of MLE) *Consider a parametric constant model p_θ. We say that the parametric model has an asymptotically normal MLE if, for all $\theta_0 \in \Theta$,*

$$\sqrt{n}\left(\hat{\theta}_{(x^n)} - \theta_0\right) \xrightarrow{d} N\left(0, I^{-1}\left(\theta_0\right)\right),$$

where $I(\theta)$ is the Fisher information at θ, x^n is a sample path of p_{θ_0}, and $\hat{\theta}_{(x^n)}$ is the maximum likelihood estimate of θ given x^n, that is, $\hat{\theta}_{(x^n)}$ maximizes $p_\theta(x^n)$.

Asymptotic normality holds for regular parametric models; for typical regularity conditions, see for example, Theorem 3.3 in (Newey and McFadden, 1994).

For parametric models whose maximum likelihood estimates take values in a countable set, we need the notion of a lattice MLE.

Definition 10 (Lattice MLE) *Consider a parametric model p_θ with $\theta \in \Theta \subseteq \mathbb{R}^d$. The parametric model is said to have a lattice MLE with diminishing step-size h_n, if for any θ, the possible maximum likelihood estimates of n i.i.d random variables generated by p_θ are points in Θ that are of the form $(b + k_1 h_n, b + k_2 h_n, \cdots, b + k_d h_n)$, for some integers $k_1, k_2, \cdots, k_d$ and some real numbers b and h_n. Additionally h_n is positive and diminishes to zero as n goes to infinity.*

We are now ready to state and prove our main result.

3. Main Result

We show that in parametric models with an asymptotically normal MLE, the optimality of a Bayesian strategy implies that the strategy uses Jeffreys prior. Furthermore we show that the optimality of a Bayesian strategy is equivalent to the optimality of sequential normalized maximum likelihood. This extends the result for canonical minimal exponential family distributions from (Hedayati and Bartlett, 2012) to regular parametric models. Note that NML is the unique optimal strategy, so when we say that some other strategy is equivalent to NML, that is the same as saying that strategy predicts optimally.

Theorem 11 *Suppose we have a parametric model p_θ with an asymptotically normal MLE. Assume that the MLE has a density with respect to Lebesgue measure or that the model has a lattice MLE with diminishing step-size h_n. Also assume that $I(\theta)$, the Fisher information at θ is continuous in θ, and that, for all x, $p_\theta(x)$ is continuous in θ. Also fix $m > 0$ and x^{m-1}, and assume that $p_{nml}^{(n)}(x_m^n|x^{m-1})$ and $p_\pi(x_m^n|x^{m-1})$ are well defined, where π is the Jeffreys prior. Then the following are equivalent.*

(a) NML = Bayesian:
There is a prior π on Θ such that for all n and all x_m^n,

$$p_{nml}^{(n)}(x_m^n|x^{m-1}) = p_\pi(x_m^n|x^{m-1}).$$

(b) NML = SNML:
For all n and all x_m^n,

$$p_{nml}^{(n)}(x_m^n|x^{m-1}) = p_{snml}(x_m^n|x^{m-1}).$$

(c) NML = Bayesian with Jeffreys prior:
If π denotes Jeffreys prior on Θ, for all n and all x_m^n,

$$p_{nml}^{(n)}(x_m^n|x^{m-1}) = p_\pi(x_m^n|x^{m-1}).$$

(d) $p_{snml}(\cdot|x^{m-1})$ is exchangeable.

(e) SNML = Bayesian:
There is a prior π on Θ such that for all n and all x_m^n,

$$p_{snml}(x_m^n|x^{m-1}) = p_\pi(x_m^n|x^{m-1}).$$

(f) SNML = Bayesian with Jeffreys prior:
If π denotes Jeffreys prior on Θ, for all n and all x_m^n,

$$p_{snml}(x_m^n|x^{m-1}) = p_\pi(x_m^n|x^{m-1}).$$

The proof is in the appendix.

4. Examples

Example 1 *This example is taken from (Hedayati and Bartlett, 2012). In this setting, the experts are Bernoulli distributions,*

$$p_\mu(x^n) = \mu^{\left(\sum_{i=1}^n x_i\right)}(1-\mu)^{\left(n-\sum_{i=1}^n x_i\right)},$$

with parameter space $(0,1)$. Note that this model has a lattice MLE with diminishing step-size $1/n$. Because for a fixed n the possible maximum likelihood estimates are

$$\frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \ldots, \frac{n-1}{n}.$$

The SNML is not defined for $n = 1$. However if x^{m-1} contains at least one 0 and one 1, the conditional SNML strategy is defined. Fix $x^2 = 10$. Consider $x^5 = (10011)$ and $y^5 = (10110)$. Then x^5 is a permutation of y^5 with the initial x^2 fixed. However $p_{snml}(x_3^5 \mid x^2) = p_{snml}(011 \mid 10) = 0.0930 \neq p_{snml}(110 \mid 10) = p_{snml}(y_3^5 \mid y^2) = 0.0932$. This means that $p_{snml}(\,.\mid x^2)$ is not exchangeable, hence based on our main theorem SNML and NML cannot be equivalent and neither is equivalent to a Bayesian strategy.

Example 2 *In this example the parametric family is the class of one-dimensional Gaussian distributions with unknown mean and variance μ and σ^2, i.e.*

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} + \log\sigma\right\}.$$

The MLE is

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n x_i \quad \textit{and} \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mu}_n)^2.$$

The conditional SNML satisfies

$$\begin{aligned} p_{snml}(x_n|x^{n-1}) &\propto \left(2\pi\hat{\sigma}_n^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{\sum_{i=1}^n (x_i - \hat{\mu}_n)^2}{2\hat{\sigma}_n^2}\right\} \\ &= \frac{e^{-\frac{n}{2}} n^{\frac{n}{2}}}{(2\pi(n-1))^{\frac{n}{2}}} \frac{1}{\left(\hat{\sigma}_{n-1}^2 + \frac{1}{n}(x_n - \hat{\mu}_{n-1})^2\right)^{\frac{n}{2}}}. \end{aligned}$$

Normalizing we get:

$$p_{snml}(x_n|x^{n-1}) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)} (n\hat{\sigma}_{n-1})^{-\frac{1}{2}} \left(1 + \frac{(x_n - \hat{\mu}_{n-1})^2}{n\hat{\sigma}_{n-1}^2}\right)^{-\frac{n}{2}}.$$

It can be shown (Kotlowski and Grunwald, 2011) that for $n > 1$

$$\begin{aligned} &R(x_2^n, p_{snml} \mid x_1) - R(x_2^{n-1}, p_{snml} \mid x_1) \\ &= \frac{n+1}{2}\log n - \frac{n}{2}\log(n-1) - \frac{1}{2}\log 2e + \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}. \end{aligned}$$

This shows that the conditional SNML is an equalizer and hence equivalent to the conditional NML. Moreover, asymptotic normality holds for any $\mu \in \mathbb{R}$ and any $\sigma \in \mathbb{R}^+$ and $p_{\mu,\sigma^2}(x)$ is continuous in μ and σ^2, hence Theorem 11 can be applied. This shows that conditional SNML and NML are equivalent to a conditional Bayesian strategy under Jeffreys prior. A direct computation of the Bayesian strategy with Jeffreys prior verifies this. Note that since this example is not a canonical exponential family, the results of (Hedayati and Bartlett, 2012) cannot be applied here.

Example 3 *In this example, the parametric family is the class of one-dimensional asymmetric student-t distributions as defined in (Zhu and Galbraith, 2009) with unknown skewness parameter $\alpha \in (0\,,1)$ and fixed left and right tail parameters $v_1 = v_2 = 1$, i.e.*

$$p_\alpha(x) = \begin{cases} \frac{1}{\pi}\left(1+\left(\frac{x}{2\alpha}\right)^2\right)^{-1} & \text{for } x \le 0\,, \\ \frac{1}{\pi}\left(1+\left(\frac{x}{2(1-\alpha)}\right)^2\right)^{-1} & \text{for } x > 0\,. \end{cases}$$

(Zhu and Galbraith, 2009) established asymptotic normality of maximum likelihood estimators in asymmetric student-t distributions. Note that additionally for any x, $p_\alpha(x)$ is continuous in α, hence Theorem 11 is applicable to this example. Proposition 2 in (Zhu and Galbraith, 2009) shows that the Fisher information of p_α is proportional to $\frac{1}{\alpha(1-\alpha)}$. This means that Jeffreys prior is proportional to $\frac{1}{\sqrt{\alpha(1-\alpha)}}$. After normalization we get $\frac{1}{\pi\sqrt{\alpha(1-\alpha)}}$. Calculating the regret of the Bayesian strategy under Jeffreys prior shows that for a fixed $n > 0$, the regret changes for different sequences of observations. For example, for $n = 3$, and sequence of observations $(1, 1, -1)$ the maximum likelihood estimate of α is 0.4136 and the regret of the Bayesian strategy under Jeffreys prior is 1.1472. On the other hand if we observe $(2, 2, -2)$, the maximum likelihood estimate is 0.3777 with 1.1851 for regret. This means that the Bayesian strategy under Jeffreys prior is not optimal because otherwise it should have resulted in equal regrets for sequences of equal length. Furthermore Theorem 11 shows that no prior distribution on $(0\,,1)$ can make the Bayesian strategy optimal and SNML can not be optimal either.

5. Acknowledgments

We gratefully acknowledge the support of the NSF under award CCF-1115788 and the ARC under award FL110100281.

Appendix: Proof of Theorem 11

Fix x^{m-1} so that all of the relevant conditional distributions are defined. We prove that (a), (b), and (c) are equivalent, and that (d), (e), and (f) are equivalent. The equivalence of (b) and (d) is Theorem 1 in (Hedayati and Bartlett, 2012).

(a) ⇒ (b): NML being equivalent to a Bayesian strategy means that NML is horizon-independent. Hence for any $m - 1 < t \le n$,

$$p^{(n)}_{nml}(x_t|x^{t-1}) = p_\pi(x_t|x^{t-1}) = p^{(t)}_{nml}(x_t|x^{t-1}) = p_{snml}(x_t|x^{t-1}),$$

which means that NML is equivalent to SNML.

(b) ⇒ (c): We use the asymptotic normality property to prove this below.

(c) ⇒ (a): This is immediate.
(d) ⇒ (e): We know that (d) and (b) are equivalent, and that (b) implies (a), but (b) and (a) together imply (e).
(e) ⇒ (d): Since SNML is Bayesian, $p_{snml}(x^n) = \int \prod_{i=1}^{n} p_\theta\left(x_i\right) d\,\pi(\theta)$ for some prior distribution π on Θ. As $\prod_{i=1}^{n} p_\theta\left(x_i\right)$ does not depend on the order of observations, SNML is exchangeable.
(e) ⇒ (f): (e) implies (d), which implies both (b) and (c), and together these imply (f).
(f) ⇒ (e): This is immediate.
The heart of the proof is verifying that
(b) ⇒ (c): NML being equivalent to SNML means that, for all $m-1 \le t \le n$,

$$\begin{aligned} p_{snml}(x^t \mid x^{m-1}) &= p^{(n)}_{nml}(x^t \mid x^{m-1}) \\ &= \frac{\int \sup_\theta p_\theta(x^t, y^{n-t}) d\,\lambda^{n-t}(y^{n-t})}{\int \sup_\theta p_\theta(x^{m-1}, y^{n-m+1}) d\,\lambda^{n-m+1}\left(y^{n-m+1}\right)} \\ &= \frac{\int p_{\hat\theta_{(x^t, y^{n-t})}}(x^t, y^{n-t}) d\,\lambda^{n-t}\left(y^{n-t}\right)}{\int p_{\hat\theta_{(x^{m-1}, y^{n-m+1})}}(x^{m-1}, y^{n-m+1}) d\,\lambda^{n-m+1}\left(y^{n-m+1}\right)}, \end{aligned} \tag{1}$$

where $\hat\theta_{(x^t, y^{n-t})}$ is the maximum likelihood estimate upon observing x^t, y^{n-t}. As n goes to infinity, $\hat\theta_{(x^t, y^{n-t})}$ converges to $\hat\theta_{y^{n-t}}$. This is because as n goes to infinity, $\frac{1}{n}\left[\sum_{i=1}^{t} \log p_\theta(x_i)\right]$ in the following equation goes to zero :

$$\hat\theta_{(x^t, y^{n-t})} = \arg\max_{\theta \in \Theta} \frac{1}{n}\left[\sum_{i=1}^{t} \log p_\theta(x_i) + \sum_{j=1}^{n-t} \log p_\theta(y_j)\right].$$

Now we rewrite Equation (1) in a different form. Let $C^{\theta_0}_{\Delta\theta}$ be a hypercube centered at θ_0 with all sides having length h, where $\Delta\theta = h^d$, is the volume of the hypercube. Define

$$S^n_{x^t}(\theta_0) = \left\{ z^{n-t} \,\middle|\, \hat\theta_{(x^t, z^{n-t})} \in C^{\theta_0}_{\Delta\theta/\sqrt{n^d}} \right\},$$

where $C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}$ is a hypercube that has volume $\Delta\theta/\sqrt{n^d}$ with all sides having length equal to $h/\sqrt{n}$. Let $P^\Theta_{\Delta\theta/\sqrt{n^d}}$ be the largest collection of disjoint hypercubes $C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}$ that fit in Θ. Note that as $\Delta\theta$ goes to zero $P^\Theta_{\Delta\theta/\sqrt{n^d}}$ covers the whole Θ. Define

$$g^n(x^t, x^{m-1}, \Delta\theta) = \frac{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} \int_{S^n_{x^t}(\theta_0)} p_{\theta_0}(x^t) p_{\theta_0}(y^{n-t}) d\,\lambda^{n-t}\left(y^{n-t}\right)}{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} \int_{S^n_{x^{m-1}}(\theta_0)} p_{\theta_0}(x^{m-1}) p_{\theta_0}(y^{n-m+1}) d\,\lambda^{n-m+1}\left(y^{n-m+1}\right)}.$$

First of all we show that

$$\lim_{n\to\infty} \lim_{\Delta\theta\to 0} \left| g^n(x^t, x^{m-1}, \Delta\theta) - p^{(n)}_{nml}(x^t \mid x^{m-1}) \right| = 0.$$

Since for all n, we have $p_{snml}(x^t \mid x^{m-1}) = p^{(n)}_{nml}(x^t \mid x^{m-1})$ this implies that $g^n(x^t, x^{m-1}, \Delta\theta)$ converges to $p_{snml}(x^t \mid x^{m-1})$. Then we show that the limit of $g^n(x^t, x^{m-1}, \Delta\theta)$ as n goes to infinity and $\Delta\theta$ goes to zero is a Bayesian conditional under Jeffreys prior. Now, it is easy to see the following:

$$p^{(n)}_{nml}(x^t \mid x^{m-1})$$
$$= \frac{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} \int_{S^n_{x^t}(\theta_0)} p_{\hat{\theta}_{(x^t,y^{n-t})}}(x^t) p_{\hat{\theta}_{(x^t,y^{n-t})}}(y^{n-t}) d\,\lambda^{n-t}\left(y^{n-t}\right)}{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} \int_{S^n_{x^{m-1}}(\theta_0)} p_{\hat{\theta}_{(x^{m-1},y^{n-m+1})}}(x^{m-1}) p_{\hat{\theta}_{(x^{m-1},y^{n-m+1})}}(y^{n-m+1}) d\,\lambda^{n-m+1}\left(y^{n-m+1}\right)}.$$

The only difference between this and $g^n(x^t, x^{m-1}, \Delta\theta)$ is that instead of θ_0 we have the parameter $\hat{\theta}_{(x^{m-1},y^{n-m+1})}$ for each hypercube. The distance between two points in each hypercube is at most $h\sqrt{d/n}$, hence

$$\left|\theta_0 - \hat{\theta}_{(x^t,y^{n-t})}\right| \leq h\sqrt{\frac{d}{n}}.$$

As $\Delta\theta$ and consequently h go to zero, θ_0 converges to $\hat{\theta}_{(x^t,y^{n-t})}$ for the expressions in the numerator and to $\hat{\theta}_{(x^{m-1},y^{n-m+1})}$ for those in the denominator. Due to the continuity of the likelihood for each hypercube in the numerator, we have

$$\lim_{\Delta\theta \to 0} p_{\theta_0}\left(x^t, y^{n-t}\right) = p_{\hat{\theta}_{(x^t,y^{n-t})}}\left(x^t, y^{n-t}\right).$$

Similarly, for each hypercube in the denominator we have

$$\lim_{\Delta\theta \to 0} p_{\theta_0}\left(x^{m-1}, y^{n-m+1}\right) = p_{\hat{\theta}_{(x^{m-1},y^{n-m+1})}}\left(x^{m-1}, y^{n-m+1}\right).$$

Hence $g^n(x^t, x^{m-1}, \Delta\theta)$ converges to $p^{(n)}_{nml}(x^t \mid x^{m-1})$. Furthermore as n goes to infinity the NML probability does not change, because it is equivalent to SNML and thus is horizon-independent. This means $\lim_{n\to\infty} \lim_{\Delta\theta\to 0} g^n(x^t, x^{m-1}, \Delta\theta) = p_{snml}(x^t \mid x^{m-1})$.

Next we show that the limit of $g^n(x^t, x^{m-1}, \Delta\theta)$ as n goes to infinity and $\Delta\theta$ goes to zero is a Bayesian conditional under Jeffreys prior, which completes the proof. The following is easy to see:

$$p_{\theta_0}\left(\hat{\theta}_{(x^t,Y^{n-t})} \in C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right) = \int_{S^n_{x^t}(\theta_0)} p_{\theta_0}(y^{n-t}) d\,\lambda^{n-t}\left(y^{n-t}\right).$$

Moreover, we have

$$\begin{aligned}
p_{\theta_0}\left(\hat{\theta}_{(x^t,Y^{n-t})} \in C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right) &= p_{\theta_0}\left(\hat{\theta}_{(x^t,Y^{n-t})} - \theta_0 \in C^{0}_{\Delta\theta/\sqrt{n^d}}\right) &(2)\\
&= p_{\theta_0}\left(\sqrt{n}(\hat{\theta}_{(x^t,Y^{n-t})} - \theta_0) \in \sqrt{n}C^{0}_{\Delta\theta/\sqrt{n^d}}\right) &(3)\\
&= p_{\theta_0}\left(\sqrt{n}(\hat{\theta}_{(x^t,Y^{n-t})} - \theta_0) \in C^{0}_{\Delta\theta}\right). &(4)
\end{aligned}$$

Hence

$$\int_{S^n_{x^t}(\theta_0)} p_{\theta_0}(y^{n-t}) d\,\lambda^{n-t}\left(y^{n-t}\right) = p_{\theta_0}\left(\sqrt{n}(\hat{\theta}_{(x^t,Y^{n-t})} - \theta_0) \in C^{0}_{\Delta\theta}\right).$$

Also,

$$g^n(x^t, x^{m-1}, \Delta\theta) = \frac{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} p_{\theta_0}(x^t) p_{\theta_0}\left(\sqrt{n}(\hat{\theta}_{(x^t,Y^{n-t})} - \theta_0) \in C^0_{\Delta\theta}\right)}{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} p_{\theta_0}(x^{m-1}) p_{\theta_0}\left(\sqrt{n}(\hat{\theta}_{(x^{m-1},Y^{n-m+1})} - \theta_0) \in C^0_{\Delta\theta}\right)}.$$

Let $F^n_{x^t,\theta_0}(.)$ be the cumulative distribution function of the random variable

$$\sqrt{n}(\hat{\theta}_{(x^t,Y^{n-t})} - \theta_0)$$

when the data is i.i.d. and generated by $p_{\theta_0}(\cdot)$. Define $F^n_{x^{m-1},\theta_0}(\cdot)$ similarly. With these definitions,

$$g^n(x^t, x^{m-1}, \Delta\theta) = \frac{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} p_{\theta_0}(x^t) F^n_{x^t,\theta_0}\left(C^0_{\Delta\theta}\right)}{\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} p_{\theta_0}(x^{m-1}) F^n_{x^{m-1},\theta_0}\left(C^0_{\Delta\theta}\right)}.$$

Now we find the limit as $\Delta\theta$ goes to zero. There are two possibilities: either the MLE has a density with respect to Lebesgue measure or the model has a lattice MLE with diminishing step-size h_n. In the latter case, upon constructing $P^{\Theta}_{\Delta\theta/\sqrt{n^d}}$, we choose the hypercubes so that all points of the form $(b + k_1 h_n, b + k_2 h_n, \cdots, b + k_d h_n)$ in Θ are centers of some hypercubes. Furthermore we make sure that each of these hypercubes contains at most one point of the form $(b + k_1 h_n, b + k_2 h_n, \cdots, b + k_d h_n)$, namely the center. Let $\Delta\theta_n$ be small enough to make this phenomenon hold. This construction makes many hypercubes $C^{\theta_0}_{\Delta\theta_n/\sqrt{n^d}}$ void of maximum likelihood points. Let us abbreviate $p_{\theta_0}\left(\hat{\theta}_{(x^t,Y^{n-t})} \in C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)$ in Equation (2) by $G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)$. Equation (2) shows that $G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right) = F^n_{x^t,\theta_0}\left(C^0_{\Delta\theta_n}\right)$. Many of $G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)$ are zero, namely those with θ_0 that do not correspond to a $\hat{\theta}_{(x^t,y^n-t)}$, hence:

$$\begin{aligned}
\sum_{C^{\theta_0}_{\frac{\Delta\theta_n}{\sqrt{n^d}}}} p_{\theta_0}(x^t) F^n_{x^t,\theta_0}\left(C^0_{\Delta\theta_n}\right) &= \sum_{C^{\theta_0}_{\frac{\Delta\theta_n}{\sqrt{n^d}}}} p_{\theta_0}(x^t) G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right) \\
&= \sum_{\theta_0 \in \hat{\Theta}^n_{x^t}} p_{\theta_0}(x^t) G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right),
\end{aligned}$$

where $\hat{\Theta}^n_{x^t} = \left\{\theta \in \Theta \mid \exists\, y^{n-t} \; s.t. \; \hat{\theta}_{(x^t,y^n-t)} = \theta\right\}$. Furthermore we have the following.

$$g^n(x^t, x^{m-1}, \Delta\theta_n) = \frac{\sum_{\theta_0 \in \hat{\Theta}^n_{x^t}} p_{\theta_0}(x^t) G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)}{\sum_{\theta_0 \in \hat{\Theta}^n_{x^{m-1}}} p_{\theta_0}(x^{m-1}) G^n_{x^{m-1},\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)}.$$

Note that $G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)$ is the probability that $\hat{\theta}_{(x^t,Y^n-t)}$ equals θ_0 where Y^{n-t} are $n - t$ random variables generated by p_{θ_0} in an i.i.d fashion.

As n goes to infinity, the distribution of $\hat{\theta}_{(x^t,Y^n-t)}$ becomes independent of x^t. This is because $\frac{1}{n}\sum_{i=1}^{t}\log p_\theta(x_i)$ converges to zero for all θ, and $\hat{\theta}_{(x^t,Y^n-t)}$ converges in probablity to θ_0. This along with the asymptotic normality of MLE implies that for all $\theta_0 \in \hat{\Theta}_{x^t,n}$, $G^n_{x^t,\theta_0}(\cdot)$ converges to the density of a multivariate normal distribution with mean θ_0 and covariance matrix $I^{-1}(\theta_0)$. A simple computation shows that the limit of $G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)$ as n goes to infinity is $\sqrt{n^d\,|I(\theta_0)|\,/(2\pi)^d}$. Now we construct hypercubes of sides of length h_n and centers from $\hat{\Theta}^n_{x^t}$ for the numerator and from $\hat{\Theta}^n_{x^{m-1}}$ for the denominator. Let δ_n be the volume of each of these hypercubes. It is obvious that δ_n diminishes to zero as n goes to infinity. Using Riemann integral and the continuity of Fisher information and likelihood we get:

$$\begin{aligned}\lim_{n\to\infty} g^n(x^t, x^{m-1}, \Delta\theta_n) &= \lim_{n\to\infty}\frac{\sum_{\theta_0\in\hat{\Theta}^n_{x^t}} p_{\theta_0}(x^t)G^n_{x^t,\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)\delta_n}{\sum_{\theta_0\in\hat{\Theta}^n_{x^{m-1}}} p_{\theta_0}(x^{m-1})G^n_{x^{m-1},\theta_0}\left(C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}\right)\delta_n}\\ &= \frac{\int_\Theta p_\theta(x^t)\sqrt{|I(\theta)|}\,d\theta}{\int_\Theta p_\theta(x^{m-1})\sqrt{|I(\theta)|}\,d\theta}\end{aligned}$$

which shows that the strategy is Bayesian with Jeffreys prior. On the other hand if MLE has a density with respect to Lebesgue measure then we get the following:

$$\begin{aligned}\lim_{\Delta\theta\to 0}\frac{1}{\sqrt{n^d}}\sum_{C^{\theta_0}_{\frac{\Delta\theta}{\sqrt{n^d}}}} p_{\theta_0}(x^t)F^n_{x^t,\theta_0}\left(C^0_{\Delta\theta}\right) &= \lim_{\Delta\theta\to 0}\frac{1}{\sqrt{n^d}}\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} p_{\theta_0}(x^t)\left(\frac{F^n_{x^t,\theta_0}\left(C^0_{\Delta\theta}\right)}{\Delta\theta/\sqrt{n^d}}\right)\frac{\Delta\theta}{\sqrt{n^d}}\\ &= \lim_{\Delta\theta\to 0}\sum_{C^{\theta_0}_{\Delta\theta/\sqrt{n^d}}} p_{\theta_0}(x^t)\left(\frac{F^n_{x^t,\theta_0}\left(C^0_{\Delta\theta}\right)}{\Delta\theta}\right)\frac{\Delta\theta}{\sqrt{n^d}}\\ &= \int_\Theta p_{\theta_0}(x^t)f^n_{x^t,\theta_0}(0)d\,\theta_0,\end{aligned}$$

where $f^n_{x^t,\theta_0}(\cdot)$ is the density of $F^n_{x^t,\theta_0}$. This means that

$$g^n(x^t,x^{m-1}) \equiv \lim_{\Delta\theta\to 0} g^n(x^t,x^{m-1},\Delta\theta) = \frac{\int_\Theta p_{\theta_0}(x^t)f^n_{x^t,\theta_0}(0)d\,\theta_0}{\int_\Theta p_{\theta_0}(x^{m-1})f^n_{x^{m-1},\theta_0}(0)d\,\theta_0}. \tag{5}$$

As n goes to infinity, the distribution of $\hat{\theta}_{(x^t,Y^n-t)}$ becomes independent of x^t. This is because $\frac{1}{n}\sum_{i=1}^{t}\log p_\theta(x_i)$ converges to zero for all θ, and $\hat{\theta}_{(x^t,Y^n-t)}$ converges in probablity to θ_0. This along with the asymptotic normality of MLE shows that as n goes to infinity we get the following convergence

$$\sqrt{n}\left(\hat{\theta}_{(x^t,Y^{n-t})} - \theta_0\right) \xrightarrow{d} N\left(0, I^{-1}\left(\theta_0\right)\right).$$

Let $F_{\theta_0}(\cdot)$ be the cumulative distribution function of the multivariate normal distribution with mean 0 and covariance matrix $I^{-1}(\theta_0)$. Asymptotic normality implies that

$$F^n_{x^t,\theta_0}\left(C^0_{\Delta\theta}\right) \to F_{\theta_0}(C^0_{\Delta\theta}).$$

This means that $f^n_{x^t,\theta_0}(\theta_0)$ converges to the density of a multivariate normal distribution with mean 0 and covariance matrix $I^{-1}(\theta_0)$. A simple computation shows that this value is $\sqrt{|I(\theta_0)|/(2\pi)^d}$. Now the only concern is whether we can take the limit of $n \to \infty$ inside the integral in Equation (5). We let $k^n_{x^t}(\theta) = \sqrt{(2\pi)^d} f^n_{x^t,\theta}(0)$, hence Equation (5) becomes:

$$g^n(x^t, x^{m-1}) = \frac{\int_\Theta p_\theta(x^t) k^n_{x^t}(\theta) d\theta}{\int_\Theta p_\theta(x^{m-1}) k^n_{x^{m-1}}(\theta) d\theta}.$$

As $f^n_{x^t,\theta}(\theta)$ converges to $\sqrt{I(\theta_0)/(2\pi)^d}$ when n goes to infinity, $k^n_{x^{m-1}}(\theta)$ and $k^n_{x^t}(\theta)$ converge to $\sqrt{\mid I(\theta) \mid}$ as n goes to infinity. Now we use Lebesgue's dominated convergence theorem (Weisstein, 2012b) and Fatou's lemma (Weisstein, 2012a) to show that limit and integral are interchangeable. Fatou's lemma shows that :

$$\int_\Theta p_\theta(x^{m-1}) \sqrt{\mid I(\theta) \mid} d\theta \leq \lim_{n\to\infty} \int_\Theta p_\theta(x^{m-1}) k^n_{x^{m-1}}(\theta) d\theta.$$

Let

$$h^n_{x^t}(\theta) = \frac{p_\theta(x^t) k^n_{x^t}(\theta)}{\lim_{s\to\infty} \int_\Theta p_\theta(x^{m-1}) k^s_{x^{m-1}}(\theta) d\theta}.$$

As n goes to infinity, $k^n_{x^t}(\theta)$ approaches $\sqrt{\mid I(\theta) \mid}$. Hence for $\epsilon = \sqrt{\mid I(\theta) \mid}$ there exists an n_θ such that $\mid k^n_{x^t}(\theta) - \sqrt{\mid I(\theta) \mid} \mid \leq \epsilon$ for $n > n_\theta$.
Therefore for $n > n_\theta$ we have $k^n_{x^t}(\theta) \leq 2\sqrt{\mid I(\theta) \mid}$, and

$$h^n_{x^t}(\theta) \leq \frac{2p_\theta(x^t)\sqrt{\mid I(\theta) \mid}}{\int_\Theta p_\theta(x^{m-1}) \sqrt{\mid I(\theta) \mid} d\theta}.$$

Now let $\bar{h}^n_{x^t}(\theta) = h^n_{x^t}(\theta)$ for $n > n_\theta$ and zero otherwise. For all n and $\theta \in \Theta$ we have :

$$\bar{h}^n_{x^t}(\theta) \leq \frac{2p_\theta(x^t)\sqrt{\mid I(\theta) \mid}}{\int_\Theta p_\theta(x^{m-1}) \sqrt{\mid I(\theta) \mid} d\theta}.$$

It is obvious that the limits of both are equal as n goes to infinity. Furthermore, note that $\bar{h}^n_{x^t}(\theta)$ is upper bounded by an integrable function, namely twice the conditional Bayesian density of x^t under Jeffreys prior given x^{m-1}. We know that the conditional Bayesian density of x^t under Jeffreys prior given x^{m-1} is integrable from the assumption of the theorem. Consequently Lebesgue's dominated convergence theorem is applicable here:

$$\begin{aligned}
\lim_{n\to\infty} g^n(x^t, x^{m-1}) &= \lim_{n\to\infty} \int_\Theta h^n_{x^t}(\theta) d\,\theta \\
&= \lim_{n\to\infty} \int_\Theta \bar{h}^n_{x^t}(\theta) d\,\theta \\
&= \int_\Theta \lim_{n\to\infty} \bar{h}^n_{x^t}(\theta) d\,\theta \\
&= \frac{\int_\Theta p_\theta(x^t)\sqrt{\mid I(\theta) \mid}}{\lim_{n\to\infty} \int_\Theta p_\theta(x^{m-1}) k^n_{x^{m-1}}(\theta) d\theta}.
\end{aligned}$$

Also, we have

$$\begin{aligned}\lim_{n\to\infty}\int_\Theta p_\theta(x^{m-1})k^n_{x^{m-1}}(\theta)d\theta &= \int_\Theta \lim_{n\to\infty} p_\theta(x^{m-1})k^n_{x^{m-1}}(\theta)d\theta\\ &= \int_\Theta p_\theta(x^{m-1})\sqrt{|\, I(\theta)\,|}\,d\theta,\end{aligned}$$

because otherwise $p_{snml}(x^t \mid x^{m-1}) = \lim_{n\to\infty} g^n(x^t, x^{m-1}) = \lim_{n\to\infty}\int_\Theta \bar{h}^n_{x^t}(\theta)d\,\theta$ would not be a distribution. Hence we get:

$$\lim_{n\to\infty}\lim_{\Delta\theta\to 0} g^n(x^t, x^{m-1}, \Delta\theta) = \frac{\int_\Theta p_\theta(x^t)\sqrt{I(\theta)}d\theta}{\int_\Theta p_\theta(x^{m-1})\sqrt{I(\theta)}d\theta}.$$

Notice that the proof does not use any properties of the Fisher information matrix. Thus, if the MLE is asymptotically normal with covariance $V(\theta)$, then an optimal Bayesian strategy has prior proportional to $\sqrt{|V(\theta)|}$.

References

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

P. D. Grunwald. *The minimum description length principle*. Cambridge, Mass. : MIT Press, 2007.

F. Hedayati and P. Bartlett. Exchangeability Characterizes Optimality of Sequential Normalized Maximum Likelihood and Bayesian Prediction with Jeffreys Prior. *JMLR Workshop Conference Proceedings*, 22: AISTATS 2012:504–510, 2012.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

W. Kotlowski and P. Grunwald. Maximum Likelihood vs. Sequential Normalized Maximum Likelihood in On-line Density Estimation. to appear in COLT 2011, 2011.

W. K. Newey and D. McFadden. Chapter 35: Large sample estimation and hypothesis testing. In R. Engle and D. McFadden, editors, *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier Science, 1994. ISBN 0-444-88766-0.

E. W. Weisstein. Fatou's lemma., February 2012a. URL http://mathworld.wolfram.com/FatousLemma.html.

E. W. Weisstein. Lebesgue's dominated convergence theorem., February 2012b. URL http://mathworld.wolfram.com/LebesguesDominatedConvergenceTheorem.html.

D. Zhu and J. Galbraith. A generalized asymmetric student-t distribution with application to financial econometrics. CIRANO Working Papers 2009s-13, CIRANO, Apr. 2009. URL http://ideas.repec.org/p/cir/cirwor/2009s-13.html.