Exchangeability Characterizes Optimality of Sequential Normalized Maximum Likelihood and Bayesian Prediction with Jeffreys Prior

Fares Hedayati
Computer Science Division,
University of California at Berkeley

Peter L. Bartlett
Computer Science Division and Department of Statistics,
University of California at Berkeley, and
Mathematical Sciences, Queensland University of Technology

Abstract

We study online prediction of individual sequences under logarithmic loss with parametric constant experts. The optimal strategy, normalized maximum likelihood (NML), is computationally demanding and requires the length of the game to be known. We consider two simpler strategies: sequential normalized maximum likelihood (SNML), which computes the NML forecasts at each round as if it were the last round, and Bayesian prediction. Under appropriate conditions, both are known to achieve near-optimal regret. In this paper, we investigate when these strategies are optimal. We show that SNML is optimal iff the joint distribution on sequences defined by SNML is exchangeable. This property also characterizes the optimality of a Bayesian prediction strategy for an exponential family. The optimal prior distribution is Jeffreys prior.

1 Introduction

The aim of online learning under logarithmic loss is to predict a sequence of outcomes $x_i \in \chi$, revealed one at a time, almost as well as a set of experts. At round t, the forecaster's prediction takes the form of a conditional probability density $q_t(\cdot|x^{t-1})$, where $x^{t-1} \equiv (x_1, x_2, \cdots, x_{t-1})$ and the density is with respect to a fixed measure λ on χ. For example, if χ is discrete, λ could be the counting measure; for $\chi = \Re^d$, λ could be Lebesgue measure. The loss that the forecaster suffers at that around is $-\log q_t(x_t|x^{t-1})$, where x_t is the outcome revealed after the forecaster's prediction. The performance of the prediction strategy is measured relative to the best in a reference set of experts. The difference between the accumulated loss of the prediction strategy and the best expert in the 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 only consider i.i.d canonical exponential families, parametrized by $\theta \in \Theta$, which is a subset of the class of parametric constant experts. A *parametric constant expert* is a parameterized probability density p_θ such that for all $t > 0$ and for all $x \in \chi$, $p_\theta\left(x|x^{t-1}\right) = p_\theta\left(x\right)$.

Let $x^n \equiv (x_1, x_2, \cdots, x_n)$, $x_m^n \equiv (x_m, x_{m+1}, \cdots, x_n)$ and $x^0 \equiv ()$. We call any *sequential probability assignment* of the form $q_t(\cdot|x^{t-1})$, a *strategy*. The regret of a strategy on sequence x^n with respect to a class of parametric constant experts indexed by Θ, is defined as follows.

Definition 1 (Regret)

$$R^\Theta(x^n, q^{(n)}) = \sum_{t=1}^{n} -\log q_t(x_t|x^{t-1})$$

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

Note that any sequential probability assignment of length n defines a joint distribution on the n outcomes and vice versa [see Cesa-Bianchi and Lugosi, 2006, pg. 248]. In our definition of regret, $q^{(n)}$ denotes the joint probability defined by the product of the n sequential probability assignments $q_t(x_t|x^{t-1})$. Note that $q_1(x_1|x^0) = q_1(x_1)$.

The optimal strategy for this problem is known to be normalized maximum likelihood (NML) [see Grunwald, 2007, chap. 7] and see Definition 3 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

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume 22 of JMLR: W&CP 22. Copyright 2012 by the authors.

marginalizing over subsequences. In this paper, we consider the optimality of two approaches that address these difficulties: Bayesian strategies, and sequential normalized maximum likelihood (SNML): For what classes is SNML optimal; for what classes does there exist a prior for which the Bayesian strategy is optimal; and, in those cases, what is the optimal prior? For certain parametric classes of experts, Bayesian prediction with a particular choice of prior (Jeffreys prior) has been shown to be asymptotically optimal [see Grunwald, 2007, chaps 7,8]. SNML is within a constant of the minimax regret [Kotlowski and Grunwald, 2011]. We give characterizations of the optimality of these strategies in terms of an elementary property of the joint distribution defined by the SNML strategy. We show that SNML is optimal precisely when its joint distribution is exchangeable. In the case of canonical exponential family distributions on $\Re^d$, that is,

$$p_\theta(x) = h(x) \exp\left(x^\intercal \theta - A(\theta)\right),$$

where $\theta, x \in \Re^d$, h is a reference measure, and the log normalization A ensures that p_θ is a probability distribution, we show that the optimal strategy is a Bayesian strategy iff SNML is exchangeable and in this case the optimal prior is Jeffreys prior.

2 Definitions and Notations

We consider a generalization of the regret of Definition 1. To motivate it consider the setting where $\Theta = R$ and the experts take the form of a normal distribution of mean $\theta \in \Theta$ and variance one. The regret on a sequence of length $n = 1$ is

$$R(x_1, q^{(1)}) = \frac{1}{2}\log 2\pi - \log q_1(x_1).$$

Furthermore, as x goes to ∞, $q_1(x)$ should go to zero (since it is a probability distribution), so the regret can be arbitrarily large. 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 2 (Conditional Regret)

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

Notice that the strategy $q^{(n)}$ defines only the conditional distribution $q^{(n)}(x_m^n|x^{m-1})$. 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 3 (NML) *Given a fixed horizon n, the normalized maximum likelihood (NML) strategy is defined via the joint probability distribution $p_{nml}^{(n)}$, defined as*

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

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

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

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

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

The regret of the NML strategy achieves the minimax bound, that is, $q^{(n)} = p_{nml}^{(n)}$ minimizes $\max_{x^n} R^\Theta(x^n, q^{(n)})$. Furthermore, this strategy is an equalizer, meaning that the regrets of all sequences of observations of length n are equal. Note that $p_{nml}^{(n)}$ might not be defined if the normalization is infinite. In some cases, there exits an $m > 0$, such that for all $n \ge m$, we can define the conditional probabilities

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

For these cases the conditional NML again attains the minimax bound, that is, $q^{(n)} = p_{nml}^{(n)}$ minimizes $\max_{x_m^n} R^\Theta(x_m^n, q^{(n)}|x^{m-1})$ [see Grunwald, 2007, chap. 11].

Definition 4 (SNML) *In the sequential normalized maximum likelihood (SNML) update, the conditional probability distribution is defined in the following way.*

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

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 some cases, there exists an $m > 0$, such that for all $n \ge m$, the appropriate conditional probabilities are properly defined.

Definition 5 (Bayesian) *In a Bayesian strategy, the joint probability for t observations x^t, is defined in the following way:*

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

And the conditional probability distribution is:

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

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

We shall focus on Bayesian strategies for canonical exponential family distributions. Under mild conditions, the regret of this strategy is no more than a constant (independent of n) larger than the minimax regret, and for Jeffreys prior, the regret asymptotically approaches the minimax regret [see Grunwald, 2007, chaps. 7,8].

3 Main Results

First, we show in Theorem 3.1 that SNML and NML are equivalent if and only if p_{snml} is exchangeable. This implies that NML is horizon-independent. Then, we show in Theorem 3.2 that exchangeability of p_{snml} further implies the equivalence of NML, the Bayesian strategy with Jeffreys prior, and SNML. This theorem shows that the SNML and the Bayesian strategy with Jeffreys prior are optimal.

A stochastic process is called *exchangeable* if the joint probability does not depend on the order of observations. In other words, for any $n > 0$ and any permutation σ, the joint probability of the first n observations is equal to the joint probability of the same n observations permuted under σ. When we consider the conditional distribution $p(x_m^n|x^{m-1})$ 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. Now we are ready to state and prove our main results. The first result applies to any class (countable or uncountable) for which the conditional strategies SNML and NML are defined.

Theorem 3.1 *SNML is equivalent to NML and hence is minimax optimal if and only if p_{snml} is exchangeable.*

Proof Fix the x^{m-1}. Write the conditional regret under SNML in the following way.

$$\begin{aligned}
&R^\Theta_{snml}(x^n|x^{m-1}) \equiv R^\Theta(x_m^n, p_{snml}|x^{m-1}) = \\
&\log \sup_{\theta\in\Theta} p_\theta(x^n) \ - \ \log p_{snml}(x_m^n|x^{m-1}) = \\
&\log \ \frac{p_{\hat\theta}(x^n)}{p_{snml}(x_m^n|x^{m-1})},
\end{aligned}$$

where $\hat\theta$ is the maximum likelihood estimate of x^n. Now we show that the regret of SNML is independent of x_n:

$$\begin{aligned}
&p_{snml}(x_m^n|x^{m-1}) = p_{snml}(x_n|x^{n-1}) p_{snml}(x_m^{n-1}|x^{m-1}) \\
&= \frac{p_{\hat\theta}(x^n)}{\int \sup_\theta \ p_\theta(x^{n-1}, x)\, dx} p_{snml}(x_m^{n-1}|x^{m-1}).
\end{aligned}$$

Combining the two previous equations, we get:

$$R^\Theta_{snml}(x^n|x^{m-1}) = \log \ \frac{\int \sup_\theta \ p_\theta(x^{n-1}, x)\, dx}{p_{snml}(x_m^{n-1}|x^{m-1})}. \quad (1)$$

Therefore the regret is independent of the last observation. Now, we show that if p_{snml} is exchangeable, then the regret becomes independent of other observations, which implies that it is an equalizer and hence equivalent to NML. Let $y^n = x^{m-1} z_m^n$ be a sequence of observations where z_m^n is different from x_m^n. We show that the regret of y^n is equal to that of x^n. Under any permutation of x_m^n, $\sup_{\theta\in\Theta} p_\theta(x^n)$ does not change due to the fact that $p_\theta(x^n) = \prod_{i=1}^n p_\theta(x_i)$. On the other hand $p_{snml}(\cdot|x^{m-1})$ is exchangeable meaning that $p_{snml}\left(x_m^n|x^{m-1}\right)$ is permutation invariant. Consequently, for any permutation σ of x^n that leaves x^{m-1} fixed, $R^\Theta_{snml}(x^n|x^{m-1}) = R^\Theta_{snml}(\sigma(x^n)|x^{m-1})$. These two properties give us the following.

$$\begin{aligned}
&R^\Theta_{snml}(x^{m-1}, x_m^n|x^{m-1}) = \\
&R^\Theta_{snml}(x^{m-1}, x_m, \ldots, x_{n-1}, y_m|x^{m-1}) = \\
&R^\Theta_{snml}(x^{m-1}, y_m, x_{m+1}, \ldots, x_{n-1}, x_m|x^{m-1}) = \\
&R^\Theta_{snml}(x^{m-1}, y_m, x_{m+1}, \ldots, x_{n-1}, y_{m+1}|x^{m-1}) = \\
&R^\Theta_{snml}(x^{m-1}, y_m, y_{m+1}, x_{m+2}, \ldots, x_{n-1}, x_{m+1}|x^{m-1}).
\end{aligned}$$

Continuing inserting y_{m+i} at the last position and swapping it with x_{m+i} we see that $R^\Theta_{snml}(x^n|x^{m-1}) = R^\Theta_{snml}(y^n|y^{m-1})$ (remember $y^{m-1} = x^{m-1}$). This means that SNML is an equalizer and hence it is equivalent to conditional normalized maximum likelihood. Now, we prove the other direction. If SNML is equivalent to NML, meaning that for any $n \geq m$ and any x_m^n,

$$p_{snml}(x_m^n|x^{m-1}) = p^{(n)}_{nml}(x_m^n|x^{m-1}) = \frac{p^{(n)}_{nml}(x^n)}{p^{(n)}_{nml}(x^{m-1})}$$

then SNML is exchangeable. This is because

$$p^{(n)}_{nml}(x^n) \propto \sup_\theta \prod_{i=1}^n p_\theta(x_i)$$

which makes the probability permutation invariant and hence exchangeable. That is for any n and x_m^n the conditional probability $p_{snml}(x_m^n|x^{m-1})$ is invariant over permutations of x_m^n.

The next theorem shows that some Bayesian strategy is optimal for a canonical exponential family iff SNML is exchangeable. In that case, the optimal prior is Jeffreys prior.

Theorem 3.2 *Suppose the class of parametric constant experts is a canonical maximal exponential family as defined in Lemma 3.3 below, and p_{snml} satisfies Equation* (3)*. Then the following are equivalent.*

(a) SNML is exchangeable

(b) SNML = NML

(c) SNML = Bayesian

(d) SNML = Bayesian with Jeffreys prior

(e) NML = Bayesian

(f) NML = Bayesian with Jeffreys prior

Proof See the appendix.

For the proof of this theorem we need a different notion of exchangeability called Q-exchangeability. De Finetti's theorem says that a binary stochastic process is exchangeable if and only if it is a mixture of Bernoulli distribution, i.e. for any $n > 0$

$$p(x^n) = \int_{\theta\in[0,1]} \theta^{\left(\sum_{i=1}^n x_i\right)}(1-\theta)^{\left(n-\sum_{i=1}^n x_i\right)}\pi(\theta)\,d\theta$$

and the prior in this equation is unique. Freedman and Diaconis extended this to exponential families [Diaconis and Freedman, 1990], as follows.

Lemma 3.3 *A general stochastic process p is a mixture of a canonical maximal exponential family $p_\theta(x) = h(x)e^{x^\intercal\theta - A(\theta)}$ over $\Theta = \{\theta \in \Re^d | A(\theta) < \infty\}$ where h is positive, finite, and locally integrable Borel function on $\Re^d$, if and only if $\forall n > 0$*

$$p\left(x_1,\cdots,x_n|\sum_{i=1}^n x_i = s\right) = \frac{\prod_{i=1}^n h(x_i)}{h^{(n)}(s)} \qquad (2)$$

and

$$p\left(h^{(n)}\left(\sum_{i=1}^n x_i\right) < \infty\right) = 1 \qquad (3)$$

where $h^{(n)}$ is the nth convolution of h, i.e.

$$h^{(n)}(s) = \int_{\sum_{i=1}^n x_i = s} \prod_{i=1}^n h(x_i)d\,x_1\cdots d\,x_n \qquad (4)$$

A p satisfying (2) *and* (3) *is called Q-exchangeable.*

4 Examples

Bernoulli Distribution 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 $(0,1)$. Converting this to the canonical form we get $p_\theta = \exp\left(\sum_{i=1}^n x_i\theta - \log\left(e^\theta+1\right)\right)$ with $\Theta = \Re$, where we use the transformation $\theta = \ln\frac{\mu}{1-\mu}$. 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 $p_{snml}(x_3^5|x^2) = p_{snml}(011|01) = 0.0930 \neq p_{snml}(110|01) = p_{snml}(y_3^5|y^2) = 0.0932$. This means that $p_{snml}(\,.\,|x^2)$ is not exchangeable, hence SNML and NML cannot be equivalent and neither is equivalent to a Bayesian strategy. It turns out that the regret of SNML in this case is better than Bayesian with Jeffreys prior but worse than NML [Azoury and Warmuth, 2001].

Exponential Distribution The distributions are of the form $p_\theta(x) = \frac{1}{\theta}e^{-x}/\theta$ with $\Theta = (0,\infty)$. It is easy to check that for $n = 1$, $p_{snml}(x) \propto \frac{1}{x}e^{-x}/x = \frac{1}{x}$ which does not normalize. Jeffreys prior is proportional to $1/\theta$ which does not normalize either. However for x_1, subsequent conditionals for Bayesian with Jeffreys prior and SNML will be properly defined. For $n > 1$ we have

$$\begin{aligned} p_{snml}(x_n|x^{n-1}) &\propto \sup_\theta p_\theta(x^n) \\ &= \frac{1}{\frac{\sum_{i=1}^n x_i}{n}} e^{\left(-\frac{\sum_{i=1}^n x_i}{\frac{\sum_{i=1}^n x_i}{n}}\right)} \propto \frac{1}{\left(\sum_{i=1}^n x_i\right)^n} \end{aligned}$$

Normalizing this we get

$$p_{snml}(x_n|x^{n-1}) = \frac{(n-1)^n\left(\frac{\sum_{i=1}^{n-1} x_i}{n-1}\right)^{n-1}}{\left(\sum_{i=1}^n x_i\right)^n}$$

This shows that conditioned on x_1, the joint probability depends on the sum of the observations only. This in turns implies exchangeability which in turn implies that SNML and NML are equivalent. On the other hand, since this is an instance of an exponential family distribution satisfying the conditions of Theorem 3.2, we can conclude that SNML and NML are also equivalent to the Bayesian strategy with Jeffreys prior, conditioned on the first observation. It is straightforward to verify this.

Appendix

Proof of Theorem 3.2 Fix an appropriate x^{m-1} as before.

$(a) \iff (b)$: We showed this in Theorem 3.1.

$(b) \Rightarrow (c)$: $p_{snml}(x_m^n|x^{m-1}) = p_{nml}^{(n)}(x_m^n|x^{m-1})$
For ease of notation we let $q(x_m^n) \equiv p_{snml}(x_m^n|x^{m-1}) = p_{nml}^{(n)}(x_m^n|x^{m-1})$. Let $\sum_{i=1}^{m-1} x_i = t$, and let $\sum_{i=m}^{n} x_i = s$. The maximum likelihood estimate is then $\hat{\theta} = (\triangledown A)^{-1}\left(\frac{s+t}{n}\right)$. We have

$$
\begin{aligned}
q(x_m^n|\sum_{i=m}^{n} x_i = s) &= \frac{q(x_m^n)}{\int_{\sum_{i=m}^{n}\bar{x}_i=s} q(\bar{x}_m^n)\, d\bar{x}_m \cdots d\bar{x}_n} \\
&= \frac{p_{nml}^{(n)}(x_m^n|x^{m-1})}{\int_{\sum_{i=m}^{n}\bar{x}_i=s} p_{nml}^{(n)}(\bar{x}_m^n|x^{m-1}) d\,\bar{x}_m \cdots, d\,\bar{x}_n} \\
&= \frac{p_{nml}^{(n)}(x^n)/p_{nml}^{(n)}(x^{m-1})}{\int_{\sum_{i=m}^{n}\bar{x}_i=s} p_{nml}^{(n)}(x^{m-1},\bar{x}_m^n) d\,\bar{x}_m \cdots, d\,\bar{x}_n/p_{nml}^{(n)}(x^{m-1})} \\
&= \frac{\prod_{i=m}^{n} h(x_i) e^{(s+t)^\intercal \hat{\theta} - nA(\hat{\theta})}}{\int_{\sum_{i=m}^{n}\bar{x}_i=s} \prod_{i=m}^{n} h(\bar{x}_i) e^{(s+t)^\intercal \hat{\theta} - nA(\hat{\theta})} d\bar{x}_m \cdots d\bar{x}_n} \\
&= \frac{\prod_{i=m}^{n} h(x_i)}{h^{(n-m+1)}(s)}.
\end{aligned}
$$

Furthermore, $p_{snml}(h^{(n-m+1)}(\sum_{i=m}^{n} x_i) < \infty) = 1$, and therefore $q(\,.\,) \equiv p_{snml}(\,.\,|x^{m-1})$ is Q-exchangeable and hence a mixture of $h(x)e^{x^\intercal\theta - A(\theta)}$.

$$
q(x_m^n) \equiv p_{snml}(x_m^n|x^{m-1}) = \int p_\theta(x_m^n)\pi(\theta)\, d\theta. \quad (5)
$$

Now we let

$$
\pi_1(\theta) = K \times \frac{\pi(\theta)}{p_\theta(x^{m-1})} \quad (6)
$$

for a $K > 0$, so that π_1 is a density. Substituting this into Equation (5) we get:

$$
p_{snml}(x_m^n|x^{m-1}) = \frac{\int_\Theta p_\theta(x^n)\pi_1(\theta)\, d\theta}{\int_\Theta p_\theta(x^{m-1})\pi_1(\theta)\, d\theta}
$$

Hence, there exists a prior that makes the process Bayesian.

$(c) \Rightarrow (d)$: We showed in the proof of the previous statement that

$$
\begin{aligned}
p_{snml}(x_m^n|x^{m-1}) &= \int_\theta p_\theta(x_m^n)\pi(\theta) d\theta \\
&= \frac{\int_\theta p_\theta(x^n)\pi_1(\theta) d\theta}{\int_\theta p_\theta(x^{m-1})\pi_1(\theta) d\theta}
\end{aligned}
$$

Now, we consider the regret of $p_{snml}(x_m^{n-1}|x^{m-1})$. If the maximum likelihood estimate $\hat{\theta}$ lies in a fixed, bounded, closed subset of Θ which is bounded away from the boundary of Θ, then the regret of a Bayesian strategy with prior w is [see Grunwald, 2007, chap. 8]:

$$
\frac{d}{2}\log\frac{n}{2\pi} - \log w(\hat{\theta}) + \log\sqrt{\det I(\hat{\theta})} + o(1),
$$

We apply this theorem to $z^{n-m+1} \equiv x_m^n$ and π. Note that $\hat{\theta}_{x_m^n}$, is the maximum likelihood estimate of x_m^n. The reason we can apply Grunwald's theorem here is twofold. First, the maximum likelihood estimate always exists because the family is full rank and A invertible. Second, the parameter space Θ is open and for any maximum likelihood estimate there should exist a bounded subset that contains the maximum likelihood estimate and is bounded away from the boundary of the parameter space. Let's denote the regret of a Bayesian strategy with prior π on a sequence z^p by $R_\pi^\Theta(z^p)$ and the regret of SNML on z^p by $R_{snml}^\Theta(z^p)$. Then

$$
\begin{aligned}
R_\pi^\Theta(z^{n-m+1}) = R_{snml}^\Theta(x_m^n) &= \frac{d}{2}\log\frac{n-m+1}{2\pi} \\
&- \log\pi(\hat{\theta}_{x_m^n}) + \log\sqrt{\det I(\hat{\theta}_{x_m^n})} + o(1)
\end{aligned}
$$

However, here we are calculating the conditional regret. It is easy to verify the following relationship:

$$
R^\Theta(x_m^n) = R^\Theta(x_m^n|x^{m-1}) - \log\sup_\theta p_\theta(x^n) + \log\sup_\theta p_\theta(x_m^n)
$$

Hence for conditional SNML we get the following, where $n_1 = n - m + 1$:

$$
\begin{aligned}
&R_{snml}^\Theta(x_m^n|x^{m-1}) = R_{snml}^\Theta(x_m^n) + \\
&\log\sup_\theta p_\theta(x^n) - \log\sup_\theta p_\theta(x_m^n) \\
&= \frac{d}{2}\log\frac{n_1}{2\pi} - \log\pi(\hat{\theta}_{x_m^n}) + \log\sqrt{\det I(\hat{\theta}_{x_m^n})} \\
&+ o(1) + \log\frac{p_{\hat{\theta}_{x^n}}(x^n)}{p_{\hat{\theta}_{x_m^n}}(x_m^n)} \quad (7)
\end{aligned}
$$

If conditional SNML is Bayesian then it is exchangeable, and hence because $(a) \Rightarrow (b)$, conditional SNML is also equivalent to conditional NML and hence has equal regret for all x_m^n. Hence the conditional regret in (7) should not vary for fixed n and different x_m^n. We denote the value of this regret as $c_{n_1}(x^{m-1})$, emphasizing the fact that it depends on n_1 and x^{m-1} only. Simplifying (7) we get

$$
\begin{aligned}
\pi(\hat{\theta}_{x_m^n}) &= \left(\frac{n_1}{2\pi}\right)^{d/2} \times \sqrt{\det I(\hat{\theta}_{x_m^n})} \\
&\times \frac{e^{o(1)}}{c_{n_1}(x^{m-1})} \times \frac{p_{\hat{\theta}_{x^n}}(x^n)}{p_{\hat{\theta}_{x_m^n}}(x_m^n)} \quad (8)
\end{aligned}
$$

Fix $\theta_0 = \hat{\theta}_{x_m^n}$. We let $N = kn_1$ (k is a positive integer). There exists a sequence y^N whose maximum likelihood estimate is θ_0. This sequence is nothing but k copies of x_m^n, concatenated. The family is of full rank, therefore A is strictly convex and its gradient invertible. This

means $\hat{\theta}_{Y^N}$, the maximum likelihood of Y^N, is

$$\hat{\theta}_{Y^N} = (\nabla A)^{-1}\left(\frac{\sum_{i=1}^{N} y_i}{N}\right) = (\nabla A)^{-1}\left(\frac{k \times \sum_{i=m}^{n-1} x_i}{n_1 k}\right) = (\nabla A)^{-1}\left(\frac{\sum_{i=m}^{n} x_i}{n_1}\right) = \hat{\theta}_{x_m^n} = \theta_0.$$

As N grows to infinity then $\hat{\theta}_{(x^m Y^N)} \to \hat{\theta}_{Y^N} = \theta_0$. This means that $\frac{p_{\hat{\theta}_{x^n}}(x^n)}{p_{\hat{\theta}_{x_m^n}}(x_m^n)}$ in Equation (8) converges to $p_{\theta_0}(x^{m-1})$ as $N \to \infty$. Using this and Equation (8) we get:

$$\lim_{N\to\infty} \pi(\hat{\theta}_{Y^N}) = \pi(\theta_0) = \sqrt{\det I(\theta_0)} p_{\theta_0}(x^{m-1}) \times \lim_{N\to\infty}\left(\frac{N}{2\pi}\right)^{d/2} \frac{1}{c_N(x^{m-1})}$$

Since $c_N(x^{m-1})$ does not depend on θ_0, $\pi(\theta_0) = c(x^{m-1}) p_{\theta_0}(x^{m-1})\sqrt{\det I(\theta_0)}$, for some function c. Hence $\pi(\theta) \propto p_\theta(x^{m-1})\sqrt{\det I(\theta)}$, which in turn by Equation (6) means $\pi_1(\theta) \propto \sqrt{\det I(\theta)}$.

$(d) \Rightarrow (e)$: This is because, SNML being Bayesian implies exchangeability of SNML and hence SNML is equal to NML (by $(a) \Rightarrow (b)$) which makes NML Bayesian too.

$(e) \Rightarrow (b)$: NML being Bayesian means that there exists a prior π, such that for any $n > m$ and x_m^n we have

$$p_{nml}^{(n)}(x_m^n|x^{m-1}) = \frac{\int p_\theta(x^n)\pi(\theta)d\theta}{\int p_\theta(x^{m-1})\pi(\theta)d\theta}$$

Let $A(n) = \int \sup_\theta p_\theta(x^{m-1}, z^{n-m+1})\, dz^{n-m+1}$.

$$p_{nml}^{(n-1)}(x_m^{n-1}|x^{m-1}) = \frac{\sup_\theta p_\theta(x^{n-1})}{A(n-1)}$$

We can also get $p_{nml}^{(n-1)}$ by marginalizing $p_{nml}^{(n)}$ (remember NML is horizon independent because it is Bayesian):

$$p_{nml}^{(n-1)}(x_m^{n-1}|x^{m-1}) = \int_x p_{nml}^{(n)}(x_m^{n-1}, x|x^{m-1})dx = \int_x \sup_\theta \frac{p_\theta(x^{n-1}, x)}{A(n)}\, dx$$

Therefore

$$\frac{\sup_\theta p_\theta(x^{n-1})}{A(n-1)} = \int_x \sup_\theta \frac{p_\theta(x^{n-1}, x)}{A(n)}\, dx$$

Hence

$$\int_x \sup_\theta p_\theta(x^{n-1}, x)\, dx = \frac{A(n)}{A(n-1)} \sup_\theta p_\theta(x^{n-1}) \quad (9)$$

We know from Equation (1) that the conditional regret of x^n under SNML is

$$R^{\Theta}_{snml}(x^n|x^{m-1}) = \log\left(\frac{\int \sup_\theta p_\theta(x^{n-1}, x)\, dx}{p_{snml}(x_m^{n-1}|x^{m-1})}\right)$$

using Equation (9) we get

$$R^{\Theta}_{snml}(x^n|x^{m-1}) = \log[\frac{A(n)}{A(n-1)} \times \frac{\sup_\theta p_\theta(x^{n-1})}{p_{snml}(x_m^{n-1}|x^{m-1})}] = R^{\Theta}_{snml}(x^{n-1}|x^{m-1}) + \log\frac{A(n)}{A(n-1)}$$

Continuing this we get

$$R^{\Theta}_{snml}(x^n|x^{m-1}) = R^{\Theta}_{snml}(x^{m-1}|x^{m-1}) + \sum_{i=m}^{n} log\frac{A(i)}{A(i-1)} = \log\sup_\theta p_\theta(x^{m-1}) + \log\frac{A(n)}{A(m-1)} = \log A(n)$$

Note that it is easy to verify that $\sup_\theta p_\theta(x^{m-1}) = A(m-1)$. This shows that the conditional regret is fixed for a fixed x^{m-1} and hence the conditional $SNML$ is an equalizer and equivalent to conditional NML.

$(e) \Rightarrow (f)$: If NML is Bayesian then it is equal to SNML and therefore SNML is Bayesian with Jeffreys prior and hence so is NML. This is by $(e) \Rightarrow (b) \Rightarrow (c) \Rightarrow (d)$.

$(f) \Rightarrow (e)$: This is trivial because Bayesian with Jeffreys prior is a special case of being Bayesian.
Note that $(e) \Rightarrow (b)$ was proved in Theorem 5 in [Kotlowski and Grunwald, 2011].

References

Katy S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.*, 43:211–246, June 2001. ISSN 0885-6125. doi: 10.1023/A:1010896012157. URL `http://portal.acm.org/citation.cfm?id=599611.599643`.

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

P. Diaconis and D. A. Freedman. Cauchy's equation and de finetti's theorem. *Scandinavian Journal of Statistics*, 17(3):pp. 235–249, 1990. ISSN 03036898. URL `http://www.jstor.org/stable/4616171`.

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

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