Exchangeability Characterizes Optimality of Sequential Normalized Maximum Likelihood and Bayesian Prediction

Fares Hedayati, Peter L. Bartlett *Member, IEEE*

Abstract

We study online learning under logarithmic loss with regular parametric models. In this setting, each strategy corresponds to a joint distribution on sequences. The minimax optimal strategy is the *normalized maximum likelihood* (*NML*) strategy. We show that the *sequential normalized maximum likelihood* (*S*NML) strategy predicts minimax optimally (i.e. as NML) if and only if the joint distribution on sequences defined by SNML is exchangeable. This property also characterizes the optimality of a Bayesian prediction strategy. In that case, the optimal prior distribution is Jeffreys prior for a broad class of 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.

Index Terms

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

I. Introduction

In online learning, 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 estimation with log loss, where the forecaster's prediction at each round takes the form of a probability density over the next outcome, and the loss suffered is the negative logarithm of the forecast density 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) [16]—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 (SNML) strategy; see Definitions 7 and 6 below. Previous work has studied the asymptotic performance of these strategies, in the limit as the number of rounds goes to infinity.

We show for a very general class of parametric models that optimality of a Bayesian strategy implies that the strategy uses Jeffreys prior. Furthermore we show that optimality of the Bayesian strategy is equivalent to optimality of sequential normalized maximum likelihood, with the exchangeability of SNML sequences being a necessary and sufficient condition for optimality of both strategies. 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.

As an important consequence of these results, we see that the exchangeability of SNML sequences characterizes when NML is independent of the horizon n. That is, optimal prediction is possible without advance knowledge of the length of the game precisely when SNML is exchangeable.

Online density estimation with log loss has been widely studied in several communities, since it is closely related to universal compression and portfolio optimization; see, for example, the textbooks [2, 6] and the review [13]. Shtarkov [17] proved that NML is the unique optimal strategy for this problem. There are few models for which NML can be efficiently computed exactly: Kontkanen and collaborators showed that it is possible for multinomial models and for finite bin histogram models [10, 11], and Rissanen showed a similar result for certain linear regression models [15]. Clarke and Barron considered Bayesian strategies with Jeffreys prior, and proved their asymptotic optimality for regular parametric families under certain constraints on the outcome sequence [3, 4]. Slightly stronger results are known for horizon-dependent modifications to the Jeffreys prior [20, 18, 19]. SNML is also known to be asymptotically optimal, again under constraints on the outcome sequence [12]. In contrast to this

F. Hedayati is with the Department of Computer Engineering, the Baha'i Institute for Higher Education (BIHE), and Upwork, San Francisco CA 94107. Part of this work has been done while F. Hedayati was at UC Berkeley

P. Bartlett is with the Computer Science Division and Department of Statistics, University of California at Berkeley, Berkeley CA 94720, and the School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia.

This paper was presented in part at AISTATS2012 and COLT2012.

previous work, which studies the asymptotic performance of these strategies in the limit as the horizon goes to infinity, we consider when these strategies are exactly optimal for all finite horizons.

Earlier versions of this work appeared in the AISTAT and COLT conferences [7, 8]. Subsequent work [1] has led to a characterization of the one-dimensional exponential family distributions for which the optimality property holds.

II. Definitions and Notation

The goal of online learning is to predict a sequence of outcomes $x_t \in X$ 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 i.i.d. 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,

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

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 a sequence x^n of length n with respect to a parametric constant model p_θ is*

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

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 6, chap. 11].

Definition 3 (Conditional Regret)**.** *Given a sequence x^{m-1}, the conditional regret of a strategy $q^{(n)}$ on a sequence x_m^n is*

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

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

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

The regret of the NML strategy achieves the minimax bound, that is, $q^{(n)} = p^{(n)}_{nml}$ minimizes $\max_{x^n} R(x^n, q^{(n)})$. This follows from the fact that NML is an *equalizer*.

Definition 5 (Equalizer)**.** *A strategy is called an equalizer if, for all n, its regrets with respect to p_θ on all sequences x^n of length n are equal.*

We will use the fact that NML is the only equalizer; we include the proof of this for completeness [see, for e.g., 6, chap. 6].

Lemma II.1. *Any equalizer is minimax optimal and is identical to NML.*

Proof. Let strategy $p^{(n)}$ be an equalizer and let $q^{(n)}$ be a strategy different from $p^{(n)}$. Then for some z^n we have $p^{(n)}(z^n) > q^{(n)}(z^n)$ which in turn makes the regret of $q^{(n)}$ for z^n larger than that of p^n. If sequence w^n maximizes the regret of $q^{(n)}$ then

$$R\left(w^n, q^{(n)}\right) \geq R\left(z^n, q^{(n)}\right) > R\left(z^n, p^{(n)}\right) = R\left(w^n, p^{(n)}\right).$$

This means that for any strategy $q^{(n)}$ different from $p^{(n)}$, the maximum regret of $q^{(n)}$ over all sequences of length n is strictly greater than the maximum regret of $p^{(n)}$, therefore $p(n)$ has the minimum value of the maximum regret, that is, it is the unique minimax optimal strategy.

NML is an equalizer, because its regret on a sequence x^n is

$$\begin{aligned} R(x^n_m, p^{(n)}_{nml}) &= \log \sup_{\theta\in\Theta} p_\theta(x^n) - \log p^{(n)}_{nml}(x^n) \\ &= \log \int_{\mathcal{X}^n} \sup_{\theta\in\Theta} p_\theta(y^n)\, d\lambda^n(y^n), \end{aligned}$$

which does not depend on x^n. Thus, NML is the unique minimax optimal strategy. □

Note that $p^{(n)}_{nml}$ might not be defined if the normalization is infinite. In many cases where this occurs, for a suitable sequence x^{m-1} and for all $n \geq m$, we can define the conditional probabilities

$$\begin{aligned} &p^{(n)}_{nml}(x^n_m | x^{m-1}) \\ &= \frac{\sup_{\theta\in\Theta}\ p_\theta(x^n)}{\int_{\mathcal{X}^{n-m+1}} \sup_{\theta\in\Theta} p_\theta(x^{m-1}, y^n_m)\, d\lambda^{n-m+1}(y^n_m)}. \end{aligned}$$

For these cases, conditional NML again attains the minimax bound, that is, $q^{(n)} = p^{(n)}_{nml}$ minimizes $\max_{x^n_m} R(x^n_m, q^{(n)} \mid x^{m-1})$. This follows from the fact that conditional NML equalizes the conditional regret. The same argument as the proof of Lemma II.1 shows that any conditional strategy with the equalizing property is optimal and is identical to conditional NML [see also 6, chap. 11].

Definition 6 (SNML)**.** *The sequential normalized maximum likelihood (SNML) strategy has*

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

Notice that this update does not depend on the sequence length. Under mild conditions, the regret of SNML is no more than a constant (independent of n) larger than the minimax regret [12]. 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 $n \geq m$, the appropriate conditional probabilities are properly defined. We restrict our attention to these cases.

Definition 7 (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 $p_\pi(x^n_m \mid x^{m-1})$.

Jeffreys prior [9] has the appealing property that it is invariant under reparameterization.

Definition 8 (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 9 (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 density of x^n is the same as the density 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 10 (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 θ_0 in the interior of Θ,*

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

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 parametric models that are appropriately *regular*; for typical regularity conditions, see for example, Theorem 3.3 in [14].

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

Definition 11 (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 d_n, if the d_n are positive and diminish to zero as n goes to infinity and there is a real number b such that for any θ, the possible maximum likelihood estimates for p_θ from n i.i.d. random variables are points in Θ that are of the form $(b + k_1 d_n, b + k_2 d_n, \cdots, b + k_d d_n)$, for some integers $k_1, k_2, \cdots, k_d$.*

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

III. Main Results

First, we show in Theorem *III*.1 that SNML and NML are equivalent if and only if p_{snml} is exchangeable. This happens only if NML is horizon-independent. We then show in Theorem *III*.2, 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 SNML. 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. In short, either both SNML and the Bayesian strategy with Jeffreys prior predict optimally or neither does. We emphasize that these results are non-asymptotic: the equivalences we consider are for all sequence lengths n.

Theorem III.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.

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

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

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

Combining the two previous equations, we get:

$$R_{snml}(x^n \mid x^{m-1}) = \log \frac{\int \sup_\theta p_\theta(x^{n-1}, x)\, dx}{p_{snml}(x_m^{n-1} \mid x^{m-1})}. \tag{1}$$

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 (Definition 5) and hence, according to Lemma II.1, 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 $p_{snml}(\cdot \mid x^{m-1})$ is exchangeable meaning that $p_{snml}\left(x_m^n \mid x^{m-1}\right)$ is permutation invariant. Consequently, for any permutation σ of x^n that leaves x^{m-1} fixed, $R_{snml}(x^n \mid x^{m-1}) = R_{snml}(\sigma(x^n) \mid x^{m-1})$. These two properties give us the following.

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

Continuing inserting y_{m+i} at the last position and swapping it with x_{m+i} we see that $R_{snml}(x^n \mid x^{m-1}) = R_{snml}(y^n \mid y^{m-1})$ (remember $y^{m-1} = x^{m-1}$). This means that SNML is an equalizer (Definition 5) and hence, according to Lemma II.1, 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,

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

then SNML is exchangeable. This is because

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

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

Theorem III.2. *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 d_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 $n \geq m$ 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 $n \geq m$ 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 $n \geq m$ 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 $n \geq m$ 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 $n \geq m$ and all x_m^n,

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

Remark III.3. *A version of this theorem, applicable to exponential families, can be proved using an extension of de Finetti's theorem due to Diaconis and Freedman [5]. For details of this result and its proof, refer to the conference version [7].*

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

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

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

Equivalence of SNML and NML implies the following is true for all n:

$$p_{snml}(x^t \mid x^m) = p^{(n)}_{nml}(x^t \mid x^m) \tag{2}$$

$$= \frac{\int \sup_\theta\ p_\theta(x^t, y^{n-t}) d\,\lambda^{n-t}(y^{n-t})}{\int \sup_\theta\ p_\theta(x^{m-1}, y^{n-m+1}) d\,\lambda^{n-m+1}(y^{n-m+1})}$$

$$= \frac{\int p_{\hat\theta_{(x^t,y^{n-t})}}(x^t, y^{n-t}) d\,\lambda^{n-t}(y^{n-t})}{\int p_{\hat\theta_{(x^m,y^{n-m})}}(x^m, y^{n-m}) d\,\lambda^{n-m+1}(y^{n-m+1})} \tag{3}$$

where $\hat\theta_{(x^t,y^{n-t})}$ is the maximum likelihood estimate upon observing the sequence (x^t, y^{n-t}) and λ is either the Lebesgue measure as in the case where observations are continuous or the counting measure for discrete observations. We will use this observation to exploit the asymptotics of the maximum likelihood estimator. We emphasize again that we will show that NML is equivalent to a Bayesian strategy with Jeffreys prior *for every* $n \geq m$. We let $c(\theta, \alpha)$ be a hypercube with center θ and sides equal to α defined in the following way, where $\theta = (\theta_1, \cdots, \theta_d)$:

$$c(\theta, \alpha) = \left[\theta_1 - \frac{\alpha}{2}, \theta_1 + \frac{\alpha}{2}\right) \times \left[\theta_2 - \frac{\alpha}{2}, \theta_2 + \frac{\alpha}{2}\right) \cdots \times \left[\theta_d - \frac{\alpha}{2}, \theta_d + \frac{\alpha}{2}\right).$$

Furthermore, in case of continuous MLE, we let $h_n = \frac{\delta_n}{\sqrt{n}}$ where δ_n is positive and diminishes to zero as n goes to infinity; for the case of lattice MLE we let $h_n = \min\left(d_n, \frac{1}{n}\right)$ and let $\delta_n = \sqrt{n} \times h_n$. Note that, in this latter case, our construction guarantees that δ_n converges to zero as n goes to infinity and that in each hypercube $c(\theta, h_n)$ there is only one MLE, namely θ, the center. Furthermore we let:

$$S^n(\theta, x^t) = \{y^{n-t} \mid \hat\theta_{(x^t,y^{n-t})} \in c(\theta, h_n)\},$$

and we let:

$$\omega^n(x^t, x^m) = \frac{\sum_{C_n(x^t)} \int_{S^n(\theta,x^t)} p_{\hat\theta_{(x^t,y^{n-t})}}(x^t, y^{n-t}) d\,\lambda^{n-t}(y^{n-t})}{\sum_{C_n(x^m)} \int_{S^n(\theta,x^m)} p_{\hat\theta_{(x^m,y^{n-m})}}(x^m, y^{n-m}) d\,\lambda^{n-m}(y^{n-m})},$$

where $C_n(x^t)$ is the largest collection of disjoint hypercubes of the form $c(\theta, h_n)$ that fit in Θ with hypercube centers from $\hat\Theta^n_{x^t} = \left\{\theta \in \Theta \mid \exists\, y^{n-t} \text{ s.t. } \hat\theta_{(x^t,y^{n-t})} = \theta\right\}$, i.e. $C_n(x^t) = \cup_{\theta\in S}\, c(\theta, h_n)$ for some $S \subseteq \hat\Theta^n_{x^t}$ with $\cup_{\theta\in S}\, c(\theta, h_n) \subseteq \Theta$ and $\cap_{\theta\in S}\, c(\theta, h_n) = \emptyset$, and $\cup_{\theta\in S}\, c(\theta, h_n)$ having maximum coverage of Θ. $C_n(x^m)$ is constructed similarly.

Note that due to δ_n converging to zero, $C_n(x^t)$ converges to the whole set Θ as n goes to infinity. Consequently $|\, \omega^n(x^t, x^m) - p^{(n)}_{nml}(x^t|x^m)\,|$ converges to zero as n goes to infinity. Therefore it would be enough to study asymptotic

behavior of $\omega^n(x^t|x^m)$. Now we construct a slightly different function than $\omega^n(x^t, x^m)$, which we call $\gamma^n(x^t, m^t)$:

$$\begin{aligned}\gamma^n(x^t, x^m) &= \frac{\sum_{C_n(x^t)} \int_{S^n(\theta,x^t)} p_\theta(x^t, y^{n-t}) d\,\lambda^{n-t}(y^{n-t})}{\sum_{C_n(x^m)} \int_{S^n(\theta,x^m)} p_\theta(x^m, y^{n-m}) d\,\lambda^{n-m}(y^{n-m})} \\ &= \frac{\sum_{C_n(x^t)} \int_{S^n(\theta,x^t)} p_\theta(x^t) p_\theta(y^{n-t}) d\,\lambda^{n-t}(y^{n-t})}{\sum_{C_n(x^m)} \int_{S^n(\theta,x^m)} p_\theta(x^m) p_\theta(y^{n-m}) d\,\lambda^{n-m}(y^{n-m})} \\ &= \frac{\sum_{C_n(x^t)} p_\theta(x^t) \int_{S^n(\theta,x^t)} p_\theta(y^{n-t}) d\,\lambda^{n-t}(y^{n-t})}{\sum_{C_n(x^m)} p_\theta(x^m) \int_{S^n(\theta,x^m)} p_\theta(y^{n-m}) d\,\lambda^{n-m}(y^{n-m})}.\end{aligned}$$

As the likelihood function $p_\theta(x^s)$ is continuous in θ for any sequence x^s and the hypercubes diminish as n goes infinity we get

$$\lim_{n\to\infty} \left|\omega^n(x^t, x^m) - \gamma^n(x^t, x^m)\right| = 0$$

This means that we only need to study asymptotic behavior of the latter function, i.e. $\gamma^n(x^t, x^m)$. Now we let $\hat{\theta}_{(x^t,Y^{n-t})}$ be the random variable of the maximum likelihood estimate of n random variables all generated by $p_\theta(\,\cdot\,)$ with the initial t observations fixed, i.e. x^t. Then

$$\begin{aligned}\int_{S^n(\theta,x^t)} p_\theta(y^{n-t}) d\,\lambda^{n-t}(y^{n-t}) &= p_\theta\left(\hat{\theta}_{(x^t,Y^{n-t})} \in c(\theta, h_n)\right) \\ &= p_\theta\left(\sqrt{n}\left(\hat{\theta}_{(x^t,Y^{n-t})} - \theta\right) \in \sqrt{n}\left(c(\theta, h_n) - \theta\right)\right) \\ &= p_\theta\left(\sqrt{n}\left(\hat{\theta}_{(x^t,Y^{n-t})} - \theta\right) \in c(0, \delta_n)\right) \\ &\equiv F_{x^t}(\theta, \delta_n)\end{aligned}$$

Therefore

$$\begin{aligned}\gamma^n(x^t, x^m) &= \frac{\sum_{C_n(x^t)} p_\theta(x^t) F_{x^t}(\theta, \delta_n)}{\sum_{C_n(x^m)} p_\theta(x^m) F_{x^m}(\theta, \delta_n)} \\ &= \frac{\sum_{C_n(x^t)} p_\theta(x^t) \frac{F_{x^t}(\theta,\delta_n)}{|c(\theta,h_n)|} \times |c(\theta, h_n)|}{\sum_{C_n(x^m)} p_\theta(x^m) \frac{F_{x^m}(\theta,\delta_n)}{|c(\theta,h_n)|} \times |c(\theta, h_n)|} \\ &= \frac{\sum_{C_n(x^t)} p_\theta(x^t) \frac{F_{x^t}(\theta,\delta_n)}{\frac{1}{\sqrt{n^d}}|c(\theta,\delta_n)|} \times |c(\theta, h_n)|}{\sum_{C_n(x^m)} p_\theta(x^m) \frac{F_{x^m}(\theta,\delta_n)}{\frac{1}{\sqrt{n^d}}|c(\theta,\delta_n)|} \times |c(\theta, h_n)|} \\ &= \frac{\sum_{C_n(x^t)} p_\theta(x^t) \frac{F_{x^t}(\theta,\delta_n)}{|c(\theta,\delta_n)|} \times |c(\theta, h_n)|}{\sum_{C_n(x^m)} p_\theta(x^m) \frac{F_{x^m}(\theta,\delta_n)}{|c(\theta,\delta_n)|} \times |c(\theta, h_n)|}\end{aligned}$$

Where $|c(\theta, h_n)|$ and $|c(\theta, \delta_n)|$ denote the volumes of $c(\theta, h_n)$ and $c(\theta, \delta_n)$ respectively. As n goes to infinity $\hat{\theta}_{(x^t,Y^{n-t})}$ becomes independent of x^t, this is because

$$\hat{\theta}_{(x^t,Y^{n-t})} = \text{argmax}_{\,\theta\in\Theta}\left(\frac{\sum_{i=1}^t \log p_\theta(x_i)}{n} + \frac{\sum_{j=t+1}^n \log p_\theta(Y_j)}{n}\right).$$

The first fraction converges to 0 as n goes to infinity. MLE's asymptotic normality tells us that $F_{x^t}(\theta, \delta_n)$ converges to the volume of a normal distribution with mean 0 and covariance matrix $I^{-1}(\theta)$ over the cube $c(\theta, \delta_n)$ as n goes to infinity. Furthermore $\frac{F_{x^t}(\theta,\delta_n)}{|c(\theta,\delta_n)|}$ converges to the density of the aforementioned normal distribution at 0 which is $K\sqrt{I(\theta)}$ for some K. Using a Riemann integral we get:

$$\lim_{n\to\infty} \gamma^n(x^t, x^m) = \frac{\int p_\theta(x^t)\sqrt{I(\theta)}\, d\theta}{\int p_\theta(x^m)\sqrt{I(\theta)}\, d\theta}$$

□

IV. Examples

Example IV.1. *Consider the parametric constant model consisting of Bernoulli distributions, with* $\mathcal{X} = \{0, 1\}$, $\Theta = (0, 1)$ *and*

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

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

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

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. The MLE is the empirical average which is asymptotically normal by the central limit theorem, hence Theorem III.2 can be applied here. This theorem tells us that SNML and NML cannot be equivalent and neither is equivalent to a Bayesian strategy.

Example IV.2. *In this example, $\mathcal{X} = \mathbb{R}$, $\Theta = \mathbb{R} \times \mathbb{R}^+$, and 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

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

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:

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

It can be shown [12] that for $n > 1$

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

This shows that the conditional SNML is an equalizer (Definition 5) and hence, according to Lemma II.1, 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 III.2 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.

Example IV.3. *In this example, $\mathcal{X} = \mathbb{R}$ and the parametric family is the class of one-dimensional asymmetric student-t distributions as defined in [21] with unknown skewness parameter $\alpha \in (0\,,1)$ and fixed left and right tail parameters $v_1 = v_2 = 1$, i.e.*

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

The maximum likelihood estimator for asymmetric student-t distributions is asymptotically normal [21]. Note that additionally for any x, $p_\alpha(x)$ is continuous in α, hence Theorem III.2 is applicable to this example. Proposition 2 in [21] 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 regret 1.1851. This means that the Bayesian strategy under Jeffreys prior is not optimal because otherwise it would have resulted in equal regrets for sequences of equal length. Furthermore Theorem III.2 shows that no prior distribution on $(0\,,1)$ can make the Bayesian strategy optimal and SNML can not be optimal either.

V. Discussion

According to Theorem III.2, the property that guarantees the optimality of SNML and Bayesian strategies is exchangeability of SNML sequences. The main question is then determining which families satisfy this optimality property. [1] showed that in one-dimensional exponential family distributions, only three classes of natural exponential family distributions, namely the Gaussian, Gamma, and the Tweedie exponential family of order 3/2, have exchangeable SNML strategies. The question remains open for multi-dimensional exponential families, for other parametric models and for non-parametric families.

Acknowledgments

We gratefully acknowledge the support of the NSF through grants CCF-1115788 and IIS-1619362 and of the Australian Research Council through an Australian Laureate Fellowship (FL110100281) and through the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS).

References

[1] Peter L. Bartlett, Peter Grünwald, Peter Harremoës, Fares Hedayati, and Wojciech Kotlowski. Horizon-independent optimal prediction with log-loss in exponential families. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 639–661. JMLR.org, 2013.

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

[3] Bertrand S Clarke and Andrew R Barron. Information-theoretic asymptotics of bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.

[4] Bertrand S Clarke and Andrew R Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical planning and Inference*, 41(1):37–60, 1994.

[5] P. Diaconis and D. A. Freedman. Cauchy's equation and de Finetti's theorem. *Scandinavian Journal of Statistics*, 17(3):235–249, 1990.

[6] Peter D. Grünwald. *The Minimum Description Length Principle*, volume 1 of *MIT Press Books*. The MIT Press, 2007.

[7] Fares Hedayati and Peter L. Bartlett. Exchangeability characterizes optimality of sequential normalized maximum likelihood and Bayesian prediction with Jeffreys prior. In Neil D. Lawrence and Mark A. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, April 21-23, 2012*, volume 22 of *JMLR Proceedings*, pages 504–510. JMLR.org, 2012.

[8] Fares Hedayati and Peter L. Bartlett. The optimality of Jeffreys prior for online density estimation and the asymptotic normality of maximum likelihood estimators. *Proceedings of the Twenty Fifth Conference on Learning Theory (COLT' 12)*, 2012.

[9] 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.

[10] Petri Kontkanen, Wray Buntine, Petri Myllymäki, Jorma Rissanen, and Henry Tirri. Efficient computation of stochastic complexity. In *Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics*, pages 233–238, 2003.

[11] Petri Kontkanen and Petri Myllymäki. A fast normalized maximum likelihood algorithm for multinomial data. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pages 1613–1615, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

[12] W. T. Kotlowski and P. D. Grünwald. Maximum likelihood versus sequential normalized maximum likelihood in online density estimation. In Sham Kakade and Ulrike von Luxburg, editors, *Proceedings of Annual Conference on Learning Theory 2011*. JMLR.org, July 2011.

[13] Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

[14] Whitney K. Newey and Daniel McFadden. Chapter 35: Large sample estimation and hypothesis testing. In Robert Engle and Dan. McFadden, editors, *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier Science, 1994.

[15] Jorma Rissanen. MDL denoising. *IEEE Transactions on Information Theory*, 46:2537–2543, 2000.

[16] Y. M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):175–186, 1987.

[17] Y. M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):175–186, 1987.

[18] Jun-ichi Takeuchi and Andrew R. Barron. Asymptotically minimax regret for exponential families. In *SITA'97*, pages 665–668, 1997.

[19] Jun-ichi Takeuchi and Andrew R. Barron. Asymptotically minimax regret by Bayes mixtures. In *Proceedings of the IEEE International Symposium on Information Theory*, page 318, 1998.

[20] Qun Xie and Andrew R. Barron. Minimax redundancy for the class of memoryless sources. *Information Theory, IEEE Transactions on*, 43(2):646–657, 1997.

[21] Dongming Zhu and John W. Galbraith. Modeling and forecasting expected shortfall with the generalized asymmetric Student-t and asymmetric exponential power distributions. *Journal of Empirical Finance*, 18(4):765–778, 2011.