THE OPTIMALITY OF JEFFREYS PRIOR FOR ONLINE DENSITY ESTIMATION AND THE ASYMPTOTIC NORMALITY OF MAXIMUM LIKELIHOOD ESTIMATORS

Fares Hedayati[1] *and Peter L. Bartlett*[2]

[1] University of California at Berkeley,
fareshed@eecs.berkeley.edu
[2] University of California at Berkeley,
Queensland University of Technology, bartlett@cs.berkeley.edu

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.

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. [1], and [2], 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. [2], chaps 7,8]. Moreover the regret of SNML is within a constant of the minimax optimal [3]. 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. The proofs and further details are in the full version of this paper [4].

2. DEFINITIONS AND NOTATION

We work in the same setup of [5] 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 $\mathcal{X} = \Re^d$, 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 condi-

tional 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,

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

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*

$$\begin{aligned} R(x^n, q^{(n)}) &= \sum_{t=1}^{n} -\log q_t^{(n)}(x_t|x^{t-1}) \\ &\quad - \inf_{\theta\in\Theta} \sum_{t=1}^{n} -\log p_\theta(x_t|x^{t-1}) \\ &= \sup_{\theta\in\Theta} \log \frac{p_\theta(x^n)}{q^{(n)}(x^n)} \end{aligned} \tag{2}$$

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 [2], chap. 11].

Definition 3 (Conditional Regret)

$$\begin{aligned} R^\Theta(x_m^n, q^{(n)}|x^{m-1}) &= \sum_{t=m}^{n} -\log q_t(x_t|x^{t-1}) \\ &\quad - \inf_{\theta\in\Theta} \sum_{t=1}^{n} -\log p_\theta(x_t|x^{t-1}) \\ &= \sup_{\theta\in\Theta} \log \frac{p_\theta(x^n)}{q^{(n)}(x_m^n \mid x^{m-1})} \end{aligned} \tag{3}$$

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*

$$p_{nml}^{(n)}(x^n) = \frac{\sup_{\theta\in\Theta}\ p_\theta(x^n)}{\int_{\mathcal{X}^n} \sup_{\theta\in\Theta} p_\theta(y^n)\, d\lambda^n(y^n)}, \tag{4}$$

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

$$p_{nml}^{(n)}(x_t \mid x^{t-1}) = \frac{p_{nml}^{(n)}(x^t)}{p_{nml}^{(n)}(x^{t-1})}, \tag{5}$$

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

$$p_{nml}^{(n)}(x^t) = \int_{\mathcal{X}^{n-t}} p_{nml}^{(n)}(x^n)\, d\lambda^{n-t}(x_{t+1}^n). \tag{6}$$

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. [2] 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

$$p_{nml}^{(n)}(x_m^n|x^{m-1}) = \frac{\sup_{\theta\in\Theta}\ p_\theta(x^n)}{\int_{\mathcal{X}^{n-m+1}} \sup_{\theta\in\Theta} p_\theta(x^n)\, d\lambda^{n-m+1}(x_m^n)} \tag{7}$$

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 [2] 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*

$$p_{snml}(x_t \mid x^{t-1}) = \frac{\sup_{\theta\in\Theta}\ p_\theta(x^t)}{\int_{\mathcal{X}} \sup_{\theta\in\Theta} p_\theta(x^t)\, d\lambda(x_t)}. \tag{8}$$

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 [3]. 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*

$$p_\pi(x^t) = \int_{\theta\in\Theta} p_\theta(x^t)\, d\pi(\theta). \tag{9}$$

The conditional probability distribution is defined in the obvious way,

$$p_\pi(x_t \mid x^{t-1}) = \frac{p_\pi(x^t)}{p_\pi(x^{t-1})}. \tag{10}$$

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

Jeffreys prior [6] 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 $\theta_0 \in \Theta$,*

$$\sqrt{n}\left(\hat{\theta}_{(x^n)} - \theta_0\right) \xrightarrow{d} N\left(0, I^{-1}\left(\theta_0\right)\right), \qquad (11)$$

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 [7].

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 $\theta \in \Theta \subseteq \Re^d$. 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 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 [5] 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 3.1 *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,

$$p_{nml}^{(n)}(x_m^n|x^{m-1}) = p_\pi(x_m^n|x^{m-1}) \qquad (12)$$

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

$$p_{nml}^{(n)}(x_m^n|x^{m-1}) = p_{snml}(x_m^n|x^{m-1}) \qquad (13)$$

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

$$p_{nml}^{(n)}(x_m^n|x^{m-1}) = p_\pi(x_m^n|x^{m-1}) \qquad (14)$$

(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,

$$p_{snml}(x_m^n|x^{m-1}) = p_\pi(x_m^n|x^{m-1}) \qquad (15)$$

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

$$p_{snml}(x_m^n|x^{m-1}) = p_\pi(x_m^n|x^{m-1}) \qquad (16)$$

4. OPEN PROBLEM

Our main result, i.e. Theorem 3.1 shows that the Bayesian strategy under Jeffreys prior, SNM and NML are all equivalent if and only if SNM is exchangeable. This equivalence holds for many exponential family distributions such as Normal, Levy, Rayleigh, Exponential. On the other hand it does not hold for some simple distributions such as Bernoulli. What properties should a distribution from an exponential family have that makes its sequential normalized maximum likelihood process exchangeable?

5. REFERENCES

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

[2] Peter D Grunwald, *The Minimum Description Length Principle*, Cambridge, Mass. : MIT Press, 2007.

[3] Wojciech Kotlowski and Peter Grünwald, "Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation," *Journal of Machine Learning Research - Proceedings Track*, vol. 19, pp. 457–476, 2011.

[4] Fares Hedayati and Peter Bartlett, "The optimality of Jeffreys prior for online density estimation and the asymptotic normality of maximum likelihood estimators," in *Proceedings of the Conference on Learning Theory (COLT2012)*, 2012, vol. 23, pp. 7.1–7.13.

[5] Fares Hedayati and Peter Bartlett, "Exchangeability Characterizes Optimality of Sequential Normalized Maximum Likelihood and Bayesian Prediction with Jeffreys Prior," *JMLR Workshop Conference Proceedings*, vol. 22: AISTATS 2012, pp. 504–510, 2012.

[6] 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*, vol. 186, no. 1007, pp. 453–461, 1946.

[7] Whitney K. Newey and Daniel McFadden, "Chapter 35: Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, Robert Engle and Dan. McFadden, Eds., vol. 4, pp. 2111–2245. Elsevier Science, 1994.