Journal of Machine Learning Research 16 (2015) 1-48 Submitted 7/14; Revised 12/14; Published 8/15

Achievability of Asymptotic Minimax Regret by Horizon-Dependent and Horizon-Independent Strategies

Kazuho Watanabe WKAZUHO@CS.TUT.AC.JP
Department of Computer Science and Engineering
Toyohashi University of Technology
1-1, Hibarigaoka, Tempaku-cho, Toyohashi, 441-8580, Japan

Teemu Roos TEEMU.ROOS@CS.HELSINKI.FI
Helsinki Institute for Information Technology HIIT
Department of Computer Science, University of Helsinki
PO Box 68, FI-00014, Finland

Editor: Manfred Warmuth

Abstract

The normalized maximum likelihood distribution achieves minimax coding (log-loss) regret given a fixed sample size, or horizon, n. It generally requires that n be known in advance. Furthermore, extracting the sequential predictions from the normalized maximum likelihood distribution is computationally infeasible for most statistical models. Several computationally feasible alternative strategies have been devised. We characterize the achievability of asymptotic minimaxity by horizon-dependent and horizon-independent strategies. We prove that no horizon-independent strategy can be asymptotically minimax in the multinomial case. A weaker result is given in the general case subject to a condition on the horizon-dependence of the normalized maximum likelihood. Motivated by these negative results, we demonstrate that an easily implementable Bayes mixture based on a conjugate Dirichlet prior with a simple dependency on n achieves asymptotic minimaxity for all sequences, simplifying earlier similar proposals. Our numerical experiments for the Bernoulli model demonstrate improved finite-sample performance by a number of novel horizon-dependent and horizon-independent algorithms.

Keywords: on-line learning, prediction of individual sequences, normalized maximum likelihood, asymptotic minimax regret, Bayes mixture

1. Introduction

The normalized maximum likelihood (NML) distribution is derived as the optimal solution to the minimax problem that seeks to minimize the worst-case coding (log-loss) regret with fixed sample size n (Shtarkov, 1987). In this problem, any probability distribution can be converted into a sequential prediction strategy for predicting each symbol given an observed initial sequence, and *vice versa*. A minimax solution yields predictions that have the least possible regret, i.e., excess loss compared to the best model within a model class.

The important multinomial model, where each symbol takes one of $m > 1$ possible values, has a long history in the extensive literature on universal prediction of individual sequences especially in the Bernoulli case, $m = 2$ (see e.g. Laplace, 1795/1951; Krichevsky and Trofimov, 1981; Freund, 1996; Krichevsky, 1998; Merhav and Feder, 1998; Cesa-Bianchi and Lugosi, 2001). A linear time algorithm for computing the NML probability of any individual sequence of full length n was given by Kontkanen and Myllymäki (2007). However, this still leaves two practical problems. First, given a distribution over sequences of length n, obtaining the marginal and conditional probabilities needed for predicting symbols before the last one requires evaluation of exponentially many terms. Second, the total length of the sequence, or the *horizon*, is not necessarily known in advance in so called online scenarios (see e.g. Freund, 1996; Azoury and Warmuth, 2001; Cesa-Bianchi and Lugosi, 2001). The predictions of the first $\tilde{n}$ symbols under the NML distribution depend on the horizon n in many models, including the multinomial. In fact, Bartlett et al. (2013) showed that NML is horizon-dependent in this sense in all one-dimensional exponential families with three exceptions (Gaussian, Gamma, and Tweedy). When this is the case, NML cannot be applied, and consequently, minimax optimality cannot be achieved without horizon-dependence. Similarly, in a somewhat different adversarial setting, Luo and Schapire (2014) show a negative result that applies to loss functions bounded within the interval $[0, 1]$.

Several easily implementable nearly minimax optimal strategies have been proposed (see Shtarkov, 1987; Xie and Barron, 2000; Takeuchi and Barron, 1997; Takimoto and Warmuth, 2000; Kotłowski and Grünwald, 2011; Grünwald, 2007, and references therein). For asymptotic minimax strategies, the worst-case total log-loss converges to that of the NML distribution as the sample size tends to infinity. This is not equivalent to the weaker condition that the average regret per symbol converges to zero. It is known, for instance, that neither the Laplace plus-one-rule that assigns probability $(k+1)/(n+m)$ to a symbol that has appeared k times in the first n observations, nor the Krichevsky-Trofimov plus-one-half-rule, $(k+1/2)/(n+m/2)$, which is also the Bayes procedure under the Jeffreys prior, are asymptotically minimax optimal over the full range of possible sequences (see Xie and Barron, 2000). Xie and Barron (2000) showed that a Bayes procedure defined by a modified Jeffreys prior, wherein additional mass is assigned to the boundaries of the parameter space, achieves asymptotic minimax optimality. Takeuchi and Barron (1997) studied an alternative technique for a more general model class. Both these strategies are horizon-dependent. An important open problem has been to determine whether a horizon-independent asymptotic minimax strategy for the multinomial case exists.

We investigate achievability of asymptotic minimaxity by horizon-dependent and horizon-independent strategies. Our main theorem (Theorem 2) answers the above open problem in the negative: no horizon-independent strategy can be asymptotic minimax for multinomial models. We give a weaker result that applies more generally under a condition on the horizon-dependence of NML. On the other hand, we show that an easily implementable horizon-dependent Bayes procedure defined by a simpler prior than the modified Jeffreys prior by Xie and Barron (2000) achieves asymptotic minimaxity. The proposed procedure assigns probability $(k+\alpha_n)/(n+m\alpha_n)$ to any outcome that has appeared k times in a sequence of length n, where m is the alphabet size and $\alpha_n = 1/2 - \ln 2/(2\ln n)$ is a prior mass assigned to each outcome. We also investigate the behavior of a generalization of the last-step minimax algorithm, which we call the k-last-step minimax algorithm and which is horizon-independent. Our numerical experiments (Section 5) demonstrate superior finite-sample performance by the proposed horizon-dependent and horizon-independent algorithms compared to existing approximate minimax algorithms.

2. Preliminaries

Consider a sequence $x^n = (x_1, \cdots, x_n)$ and a parametric model

$$p(x^n|\theta) = \prod_{i=1}^{n} p(x_i|\theta),$$

where $\theta = (\theta_1, \cdots, \theta_d)$ is a d-dimensional parameter. We focus on the case where each x_i is one of a finite alphabet of symbols and the maximum likelihood estimator

$$\hat{\theta}(x^n) = \operatorname*{argmax}_{\theta} \ln p(x^n|\theta)$$

can be computed.

The optimal solution to the minimax problem,

$$\min_{\overline{p}} \max_{x^n} \ln \frac{p(x^n|\hat{\theta}(x^n))}{\overline{p}(x^n)},$$

assuming that the solution exists, is given by

$$p_{\mathrm{NML}}^{(n)}(x^n) = \frac{p(x^n|\hat{\theta}(x^n))}{C_n}, \tag{1}$$

where $C_n = \sum_{x^n} p(x^n|\hat{\theta}(x^n))$ and is called the normalized maximum likelihood (NML) distribution (Shtarkov, 1987). For model classes where the above problem has no solution and the normalizing term C_n diverges, it may be possible to reach a solution by conditioning on some number of initial observations (see Liang and Barron, 2004; Grünwald, 2007). The regret of the NML distribution is equal to the minimax value $\ln C_n$ for all x^n. We mention that in addition to coding and prediction, the code length $-\ln p_{\mathrm{NML}}^{(n)}(x^n)$ can be used as a model selection criterion according to the minimum description length (MDL) principle (Rissanen, 1996); (see also Grünwald, 2007; Silander et al., 2010, and references therein).

In cases where the minimax optimal NML distribution cannot be applied (for reasons mentioned above), it can be approximated by another strategy, i.e., a sequence of distributions $(g^{(n)})_{n\in\mathbb{N}}$. A strategy is said to be *horizon-independent* if for all $1 \leq \tilde{n} < n$, the distribution $g^{(\tilde{n})}$ matches with the marginal distribution of $x^{\tilde{n}}$ obtained from $g^{(n)}$ by summing over all length n sequences that are obtained by concatenating $x^{\tilde{n}}$ with a continuation $x_{\tilde{n}+1}^n = (x_{\tilde{n}+1}, \cdots, x_n)$:

$$g^{(\tilde{n})}(x^{\tilde{n}}) = \sum_{x_{\tilde{n}+1}^n} g^{(n)}(x^n). \tag{2}$$

For horizon-independent strategies, we omit the horizon n in the notation and write $g(x^n) = g^{(n)}(x^n)$. This also implies that the ratio $g(x_{\tilde{n}+1}^n|x^{\tilde{n}}) = g(x^n)/g(x^{\tilde{n}})$ is a valid conditional probability distribution over the continuations $x_{\tilde{n}+1}^n$ assuming that $g(x^{\tilde{n}}) > 0$.[1]

[1] Note that even if a strategy is based on *assuming* a fixed horizon (or an increasing sequence or horizons like in the so called doubling-trick, see Cesa-Bianchi et al., 1997), as long as the assumed horizon is independent of the true horizon, the strategy is horizon-independent.

A property of interest is *asymptotic* minimax optimality of g, which is defined by

$$\max_{x^n} \ln \frac{p(x^n|\hat{\theta}(x^n))}{g(x^n)} \leq \ln C_n + o(1), \tag{3}$$

where $o(1)$ is a term converging to zero as $n \to \infty$.

Hereafter, we focus mainly on the multinomial model with $x \in \{1, 2, \cdots, m\}$,

$$p(x|\theta) = \theta_x, \quad \sum_{j=1}^{m} \theta_j = 1, \tag{4}$$

extended to sequences by the i.i.d. assumption. The corresponding conjugate prior is the Dirichlet distribution. In the symmetric case where each outcome $x \in \{1, \dots, m\}$ is treated equally, it takes the form

$$q(\theta|\alpha) = \frac{\Gamma(m\alpha)}{\Gamma(\alpha)^m} \prod_{j=1}^{m} \theta_j^{\alpha-1},$$

where $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ is the gamma function and $\alpha > 0$ is a hyperparameter. Probabilities of sequences under Bayes mixtures with Dirichlet priors can be obtained from

$$p_{B,\alpha}(x^n) = \int \prod_{i=1}^{n} p(x_i|\theta) q(\theta|\alpha) d\theta = \frac{\Gamma(m\alpha)}{\Gamma(\alpha)^m} \frac{\prod_{j=1}^{m} \Gamma(n_j + \alpha)}{\Gamma(n + m\alpha)}, \tag{5}$$

where n_j is the number of js in x^n. The Bayes mixture is horizon-dependent if α depends on n and horizon-independent otherwise.

The minimax regret is asymptotically given by Xie and Barron (2000),

$$\ln C_n = \frac{m-1}{2} \ln \frac{n}{2\pi} + \ln \frac{\Gamma(1/2)^m}{\Gamma(m/2)} + o(1). \tag{6}$$

3. (Un)achievability of Asymptotic Minimax Regret

We now give our main result, Theorem 2, showing that no horizon-independent asymptotic minimax strategy for the multinomial case exists. In the proof, we use the following lemma. The proof of the lemma is given in Appendix A.

Lemma 1 *Let*

$$f(x) = \ln \Gamma\left(x + \frac{1}{2}\right) - x \ln x + x - \frac{1}{2} \ln 2\pi,$$

for $x > 0$ and $f(0) = -\frac{\ln 2}{2}$. Then for $x \geq 0$,

$$-\frac{\ln 2}{2} \leq f(x) < 0 \tag{7}$$

and $\lim_{x\to\infty} f(x) = 0$.

Theorem 2 *For the multinomial model in (4), no horizon-independent strategy is asymptotic minimax.*

Proof Let g be an arbitrary horizon-independent strategy satisfying (2). First, by the properties of the Gamma function, we have $\ln\Gamma(n+\frac{m}{2}) = \ln\Gamma(n+\frac{1}{2}) + \frac{(m-1)}{2}\ln n + o(1)$. Applying this to (5) in the case of the Jeffreys mixture $p_{B,1/2}$ yields

$$\ln p_{B,1/2}(x^n) = \ln\frac{\Gamma(m/2)}{\Gamma(1/2)^m} + \sum_{j=1}^m \Big\{\ln\Gamma(n_j+1/2)\Big\} - \ln\Gamma(n+1/2) - \frac{m-1}{2}\ln n + o(1). \quad (8)$$

We thus have

$$\begin{aligned}\ln\frac{p^{(n)}_{\mathrm{NML}}(x^n)}{p_{B,1/2}(x^n)} &= \sum_{j=1}^m \left\{-\ln\Gamma\left(n_j+1/2\right) + n_j\ln n_j - n_j + \frac{1}{2}\ln 2\pi\right\} \\ &\quad + \ln\Gamma(n+1/2) - n\ln n + n - \frac{1}{2}\ln 2\pi + o(1) \\ &= -\sum_{j=1}^m f(n_j) + f(n) + o(1). \end{aligned} \quad (9)$$

By Lemma 1, for the sequence of all js (for any $j\in\{1,2,\cdots,m\}$),

$$\ln\frac{p^{(n)}_{\mathrm{NML}}(x^n)}{p_{B,1/2}(x^n)} \to \frac{m-1}{2}\ln 2 \quad (n\to\infty),$$

which means that the Jeffreys mixture is not asymptotically minimax. Hence, we can assume that g is not the Jeffreys mixture and pick $\tilde{n}$ and $x^{\tilde{n}}$ such that for some positive constant ε,

$$\ln\frac{p_{B,1/2}(x^{\tilde{n}})}{g(x^{\tilde{n}})} \geq \varepsilon. \quad (10)$$

By (9) and Lemma 1, we can find n_0 such that for all $n > n_0$ and all sequences x^n,

$$\ln\frac{p^{(n)}_{\mathrm{NML}}(x^n)}{p_{B,1/2}(x^n)} \geq -\frac{\varepsilon}{2}. \quad (11)$$

Then for all $n > \max\{\tilde{n}, n_0\}$, there exists a sequence x^n which is a continuation of the $x^{\tilde{n}}$ in (10), such that

$$\begin{aligned}\ln\frac{p^{(n)}_{\mathrm{NML}}(x^n)}{g(x^n)} &= \ln\frac{p^{(n)}_{\mathrm{NML}}(x^n)}{p_{B,1/2}(x^n)} + \ln\frac{p_{B,1/2}(x^n)}{g(x^n)} \\ &= \ln\frac{p^{(n)}_{\mathrm{NML}}(x^n)}{p_{B,1/2}(x^n)} + \ln\frac{p_{B,1/2}(x^n_{\tilde{n}+1}|x^{\tilde{n}})}{g(x^n_{\tilde{n}+1}|x^{\tilde{n}})} + \ln\frac{p_{B,1/2}(x^{\tilde{n}})}{g(x^{\tilde{n}})} \\ &\geq -\frac{\varepsilon}{2} + \varepsilon = \frac{\varepsilon}{2}, \end{aligned} \quad (12)$$

where the identity $\ln g(x^n) = \ln g(x^n_{\tilde{n}+1}|x^{\tilde{n}}) + \ln g(x^{\tilde{n}})$ implied by horizon-independence is used on the second row. The last inequality follows from (10), (11) and the fact that $g(x^n_{\tilde{n}+1}|x^{\tilde{n}})$ is a conditional probability distribution of $x^n_{\tilde{n}+1}$. Note that since (11) holds for all continuations of $x^{\tilde{n}}$, it is sufficient that there exists one continuation for which $p_{B,1/2}(x^n_{\tilde{n}+1}|x^{\tilde{n}})/g(x^n_{\tilde{n}+1}|x^{\tilde{n}}) \geq 1$ holds on the second row of (12). ■

It will be interesting to study whether similar results as above can be obtained for other models than the multinomial. For models where the NML is horizon-dependent and the Jeffreys mixture satisfies the convergence to NML in the sense of (11), we can use the same proof technique to prove the non-achievability by horizon-independent strategies. Here we provide an alternative approach that leads to a weaker result, Theorem 3, showing that a slightly stronger notion of asymptotic minimaxity is unachievable under the following condition on the horizon-dependence of the NML distribution.

Assumption 1 *Suppose that for $\tilde{n}$ satisfying $\tilde{n} \to \infty$ and $\frac{\tilde{n}}{n} \to 0$ as $n \to \infty$ (e.g. $\tilde{n} = \sqrt{n}$), there exist a sequence $x^{\tilde{n}}$ and a unique constant $M > 0$ such that*

$$\ln \frac{p^{(\tilde{n})}_{\mathrm{NML}}(x^{\tilde{n}})}{\sum_{x^n_{\tilde{n}+1}} p^{(n)}_{\mathrm{NML}}(x^n)} \to M \quad (n \to \infty). \tag{13}$$

Assumption 1 means that the NML distribution changes over the sample size n by an amount that is characterized by M. The following theorem proves that under this assumption, a stronger notion of asymptotic minimaxity is never achieved simultaneously for the sample sizes $\tilde{n}$ and n by a strategy g that is independent of n.

Theorem 3 *Under Assumption 1, if a distribution g is horizon-independent, then it never satisfies*

$$\ln C_n - \underline{M} + o(1) \leq \ln \frac{p(x^n|\hat{\theta}(x^n))}{g(x^n)} \leq \ln C_n + o(1), \tag{14}$$

for all x^n and any $\underline{M} < M$, where M is the constant appearing in Assumption 1 and $o(1)$ is a term converging to zero uniformly on x^n as $n \to \infty$.

The proof is given in Appendix B.

The condition in (14) is stronger than the usual asymptotic minimax optimality in (3), where only the second inequality in (14) is required. Intuitively, this stronger notion of asymptotic minimaxity requires not only that for all sequences, the regret of the distribution g is asymptotically at most the minimax value, but also that for no sequence, the regret is asymptotically *less* than the minimax value by a margin characterized by $\underline{M}$. Note that non-asymptotically (without the $o(1)$ terms), the corresponding strong and weak minimax notions are equivalent.

The following additional result provides a way to assess the amount by which the NML distribution depends on the horizon in the multinomial model. At the same time, it evaluates the *conditional regret* of the NML distributions as studied by Rissanen and Roos (2007), Grünwald (2007), and Hedayati and Bartlett (2012).

Let l_j be the number of js in $x^{\tilde{n}}$ ($0 \leq l_j \leq \tilde{n}$, $\sum_{j=1}^m l_j = \tilde{n}$). It follows that

$$\ln \frac{p^{(\tilde{n})}_{\mathrm{NML}}(x^{\tilde{n}})}{\sum_{x^n_{\tilde{n}+1}} p^{(n)}_{\mathrm{NML}}(x^n)} = \ln \frac{\prod_{j=1}^m \left(\frac{l_j}{\tilde{n}}\right)^{l_j}}{\sum_{n_j \geq l_j} \binom{n-\tilde{n}}{n_j - l_j} \prod_{j=1}^m \left(\frac{n_j}{n}\right)^{n_j}} + \ln \frac{C_n}{C_{\tilde{n}}}, \tag{15}$$

where $\binom{n-\tilde{n}}{n_j-l_j} \equiv \binom{n-\tilde{n}}{n_1-l_1,\cdots,n_m-l_m}$ is the multinomial coefficient and $\sum_{n_j\geq l_j}$ denotes the summation over n_js satisfying $n_1+\cdots+n_m=n$ and $n_j\geq l_j$ for $j=1,2,\cdots,m$. Lemma 4 evaluates

$$C_{n|x^{\tilde{n}}} \equiv \sum_{n_j\geq l_j} \binom{n-\tilde{n}}{n_j-l_j} \prod_{j=1}^{m} \left(\frac{n_j}{n}\right)^{n_j}$$

in (15). The proof is in Appendix C.[2]

Lemma 4 *$C_{n|x^{\tilde{n}}}$ is asymptotically evaluated as*

$$\ln C_{n|x^{\tilde{n}}} = \frac{m-1}{2}\ln\frac{n}{2\pi} + \ln\tilde{C}_{\frac{1}{2}} + o(1), \tag{16}$$

where $\tilde{C}_\alpha$ is defined for $\alpha>0$ and $\{l_j\}_{j=1}^m$ as

$$\tilde{C}_\alpha = \frac{\prod_{j=1}^m \Gamma(l_j+\alpha)}{\Gamma(\tilde{n}+m\alpha)}. \tag{17}$$

Substituting (16) and (6) into (15), we have

$$\ln\frac{p^{(\tilde{n})}_{\mathrm{NML}}(x^{\tilde{n}})}{p^{(n)}_{\mathrm{NML}}(x^{\tilde{n}})} = -\frac{m-1}{2}\ln\frac{\tilde{n}}{2\pi} + \sum_{j=1}^m l_j\ln\frac{l_j}{\tilde{n}} - \ln\frac{\prod_{j=1}^m\Gamma(l_j+1/2)}{\Gamma(\tilde{n}+m/2)} + o(1),$$

where $p^{(n)}_{\mathrm{NML}}(x^{\tilde{n}}) = \sum_{x^n_{\tilde{n}+1}} p^{(n)}_{\mathrm{NML}}(x^n)$. Applying Stirling's formula to $\ln\Gamma(\tilde{n}+m/2)$ expresses the right hand side as

$$-\sum_{j=1}^m f(l_j) + o(1),$$

where f is the function defined in Lemma 1.

To illustrate the degree to which the NML distribution depends on the horizon, take $l_1=\tilde{n}$, $l_j=0$ for $j=2,\cdots,m$. By Lemma 1, we then have $\ln p^{(\tilde{n})}_{\mathrm{NML}}(x^{\tilde{n}}) - \ln p^{(n)}_{\mathrm{NML}}(x^{\tilde{n}}) = \frac{1}{2}(m-1)\ln 2 + o(1)$.

4. Asymptotic Minimax via Simpler Horizon-Dependence

We examine the asymptotic minimaxity of the Bayes mixture in (5). More specifically, we investigate the minimax optimal hyperparameter

$$\operatorname*{argmin}_\alpha \max_{x^n} \ln\frac{p(x^n|\hat{\theta}(x^n))}{p_{B,\alpha}(x^n)} \tag{18}$$

[2] For the Fisher information matrix $I(\theta)$ whose ijth element is given by $(I(\theta))_{ij} = -\sum_x p(x|\theta)\frac{\partial^2 \ln p(x|\theta)}{\partial\theta_i\partial\theta_j} = \delta_{i,j}/\theta_j$, the constant $\tilde{C}_{1/2}$ coincides with $\int\sqrt{|I(\theta)|}\prod_{j=1}^m\theta^{l_j}d\theta$. This proves that the asymptotic expression of the regret of the conditional NML (Grünwald, 2007, Equation (11.47), p.323) is valid for the multinomial model with the full parameter set rather than the restricted parameter set discussed by Grünwald (2007).

and show that it is asymptotically approximated by

$$\alpha_n = \frac{1}{2} - \frac{\ln 2}{2}\frac{1}{\ln n}. \tag{19}$$

As a function of $(n_1, \cdots, n_{m-1})$, the regret of $p_{B,\alpha}$ is

$$\ln \frac{p(x^n|\hat{\theta}(x^n))}{p_{B,\alpha}(x^n)} = \sum_{j=1}^{m} \{n_j \ln n_j - \ln \Gamma(n_j + \alpha)\} + \kappa \tag{20}$$

where $n_m = n - \sum_{j=1}^{m-1} n_j$ and κ denotes a constant that does not depend on $(n_1, \cdots, n_{m-1})$. We first prove the following lemma (Appendix D).

Lemma 5 *The possible worst-case sequences in (18) have l nonzero counts ($l = 1, 2, \cdots, m$), each of which is $\lfloor \frac{n}{l} \rfloor$ or $\lfloor \frac{n}{l} \rfloor + 1$ with all the other counts are zeros. Here $\lfloor \cdot \rfloor$ is the floor function, the largest integer not exceeding the argument.*

From this lemma, we focus on the regrets of the two extreme cases of x^n consisting of a single symbol repeated n times and x^n with a uniform number n/m of each symbol j. Let the regrets of these two cases be equal,

$$\Gamma(\alpha)^{m-1}\Gamma(n+\alpha) = \Gamma(n/m+\alpha)^m m^n. \tag{21}$$

Equating the regrets of these two cases also equates the regrets of $(n/l, \cdots, n/l, 0, \cdots, 0)$ for $1 \le l \le m$ up to $o(1)$ terms, which is verified by directly calculating the regrets. Note that equating the regrets of the m possible worst-case sequences leads to the least maximum regret. This is because the regrets at the m possible worst-case sequences are not equal, we can improve by reducing the regret at the actual worst-case sequence until it becomes equal to the other cases.

Taking logarithms, using Stirling's formula and ignoring diminishing terms in (21), we have

$$(m-1)\left(\alpha - \frac{1}{2}\right)\ln n - (m-1)\ln\Gamma(\alpha) - m\left(\alpha - \frac{1}{2}\right)\ln m + (m-1)\frac{\ln 2\pi}{2} = 0. \tag{22}$$

This implies that the optimal α is asymptotically given by

$$\alpha_n \simeq \frac{1}{2} - \frac{a}{\ln n}, \tag{23}$$

for some constant a. Substituting this back into (22) and solving it for a, we obtain (19).

We numerically calculated the optimal hyperparameter defined by (18) for the Bernoulli model ($m = 2$). Figure 1 shows the optimal α obtained numerically and its asymptotic approximation in (19). We see that the optimal hyperparameter is well approximated by α_n in (19) for large n. Note here the slow convergence speed, $O(1/\ln n)$ to the asymptotic value, 1/2.

The next theorem shows the asymptotic minimaxity of α_n (the second inequality in (24)). We will examine the regret of α_n numerically in Section 5.1.

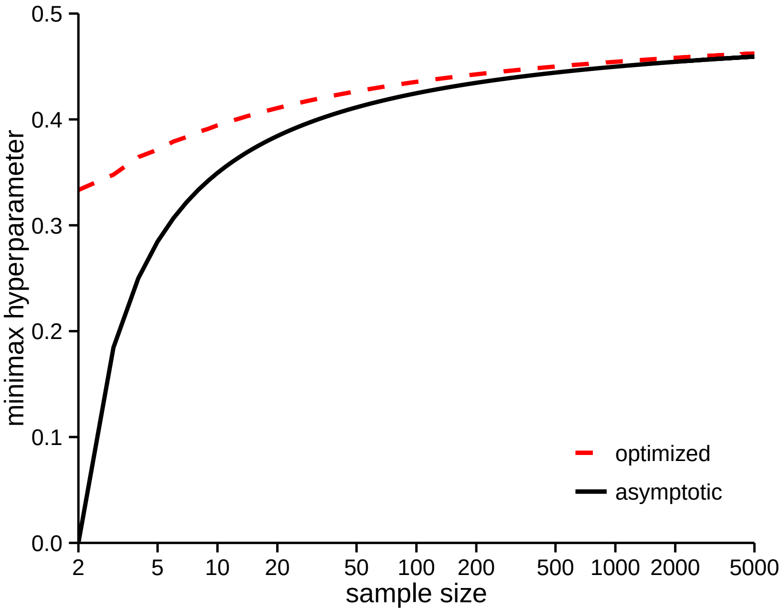


Figure 1: Minimax optimal hyperparameter α for sample size n

Theorem 6 *For the multinomial model in (4), the Bayes mixture defined by the prior* $\mathrm{Dir}(\alpha_n, \cdots, \alpha_n)$ *is asymptotic minimax and satisfies*

$$\ln C_n - M + o(1) \leq \ln \frac{p(x^n|\hat{\theta}(x^n))}{p_{B,\alpha_n}(x^n)} \leq \ln C_n + o(1), \tag{24}$$

for all x^n *where* $M = (m-1)\ln 2/2$*, and* $\ln C_n$ *is the minimax regret evaluated asymptotically in (6).*

The proof is given in Appendix E.

5. Numerical Results

In this section, we numerically calculate the maximum regrets of several methods in the Bernoulli model ($m = 2$). The following two subsections respectively examine horizon-dependent algorithms based on Bayes mixtures with prior distributions depending on n and last-step minimax algorithms, which are horizon-independent.

5.1 Optimal Conjugate Prior and Modified Jeffreys Prior

We calculated the maximum regrets of the Bayes mixtures in (5) with the hyperparameter optimized by the golden section search and with its asymptotic approximation in (19). We also investigated the maximum regrets of Xie and Barron's modified Jeffreys prior which is proved to be asymptotic minimax (Xie and Barron, 2000). The modified Jeffreys prior is defined by

$$q_{\mathrm{MJ}}^{(n)}(\theta) = \frac{\epsilon_n}{2}\left\{\delta\left(\theta - \frac{1}{n}\right) + \delta\left(\theta - 1 + \frac{1}{n}\right)\right\} + (1-\epsilon_n)b_{1/2}(\theta),$$

where δ is the Dirac's delta function and $b_{1/2}(\theta)$ is the density function of the beta distribution with hyperparameters 1/2, Beta$(1/2, 1/2)$, which is the Jeffreys prior for the Bernoulli model. We set $\epsilon_n = n^{-1/8}$ as proposed by Xie and Barron (2000) and also optimized ϵ_n by the golden section search so that the maximum regret

$$\max_{x^n} \ln \frac{p(x^n|\hat{\theta}(x^n))}{\int p(x^n|\theta)q_{\mathrm{MJ}}^{(n)}(\theta)d\theta}$$

is minimized.

Figure 2(a) shows the maximum regrets of these Bayes mixtures: asymptotic and optimized Beta refer to mixtures with Beta priors (Section 4), and modified Jeffreys methods refer to mixtures with a modified Jeffreys prior as discussed above. Also included for comparison is the maximum regret of the Jeffreys mixture (Krichevsky and Trofimov, 1981), which is not asymptotic minimax. To better show the differences, the regret of the NML distribution, $\ln C_n$, is subtracted from the maximum regret of each distribution.

We see that the maximum regrets of these distributions, except the one based on Jeffreys prior, decrease toward the regret of NML as n grows as implied by their asymptotic minimaxity. The modified Jeffreys prior with the optimized weight performs best of these strategies for this range of the sample size. For moderate and large sample sizes ($n > 100$), the asymptotic minimax hyperparameter, which can be easily evaluated by (19), performs almost as well as the optimized strategies which are not known analytically. Note that unlike the NML, Bayes mixtures provide the conditional probabilities $p(x_{\tilde{n}} \mid x_1, \ldots, x_{\tilde{n}-1})$ even if the prior depends on n. The time complexity for online prediction will be discussed in Section 5.3.

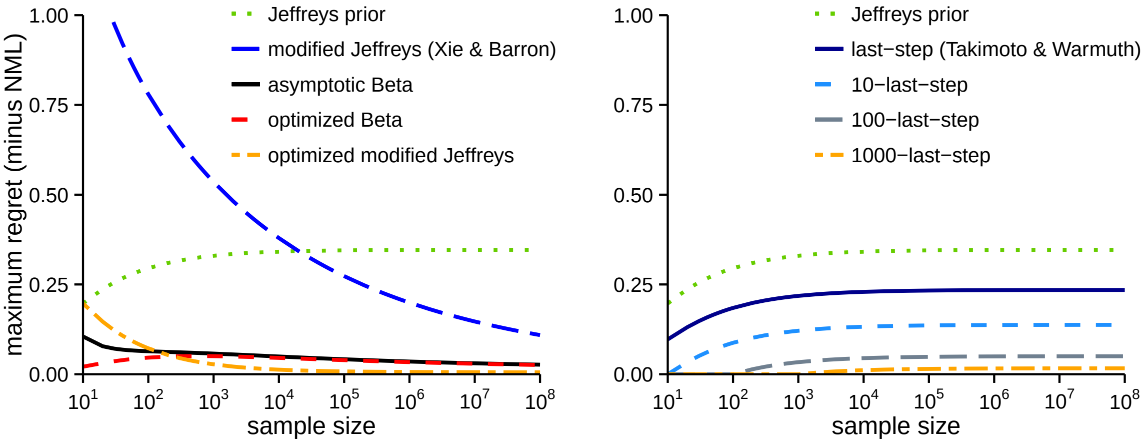
5.2 Last-Step Minimax Algorithms

The last-step minimax algorithm is an online prediction algorithm that is equivalent to the so called sequential normalized maximum likelihood method in the case of the multinomial model (Rissanen and Roos, 2007; Takimoto and Warmuth, 2000). A straightforward generalization, which we call the k-last-step minimax algorithm, normalizes $p(x^t|\hat{\theta}(x^t))$ over the last $k \geq 1$ steps to calculate the conditional distribution of $x_{t-k+1}^t = \{x_{t-k+1}, \cdots, x_t\}$,

$$p_{\mathrm{kLS}}(x_{t-k+1}^t | x^{t-k}) = \frac{p(x^t|\hat{\theta}(x^t))}{L_{t,k}},$$

where $L_{t,k} = \sum_{x_{t-k+1}^t} p(x^t|\hat{\theta}(x^t))$. Although this generalization was mentioned by Takimoto and Warmuth (2000), it was left as an open problem to examine how k affects the regret of the algorithm.

Our main result (Theorem 2) tells that k-last-step minimax algorithm with k independent of n is not asymptotic minimax. We numerically calculated the regret of the k-last-step minimax algorithm with $k = 1$, 10, 100 and 1000 for the sequence $x^n = 1010101010\cdots$ since



(a) Horizon-dependent algorithms

(b) Horizon-independent algorithms (lower bounds)

Figure 2: Maximum regret for sample size n. The regret of the NML distribution, $\ln C_n$, is subtracted from the maximum regret of each strategy. The first two algorithms (from the top) in each panel are from earlier work, while the remaining ones are novel.

it is infeasible to evaluate the maximum regret for large n. The regret for this particular sequence provides a lower bound for the maximum regret. Figure 2(b) shows the regret as a function of n together with the maximum regret of the Jeffreys mixture. The theoretical asymptotic regret for the Jeffreys mixture is $\frac{\ln 2}{2} \approx 0.34$ (Krichevsky and Trofimov, 1981), and the asymptotic bound for the 1-last-step minimax algorithm is slightly better, $\frac{1}{2}\left(1 - \ln \frac{\pi}{2}\right) \approx 0.27$ (Takimoto and Warmuth, 2000). We can see that although the regret decreases as k grows, it still increases as n grows and does not converge to that of the NML (zero in the figure).

5.3 Computational Complexity

As mentioned above, in the multinomial model, the NML probability of individual sequences of length n can be evaluated in linear time (Kontkanen and Myllymäki, 2007). However, for prediction purposes in online scenarios, we need to compute the predictive probabilities $p^{(n)}_{\mathrm{NML}}(x_t|x^{t-1})$ by summing over all continuations of x^t. Computing all the predictive probabilities up to n by this method takes the time complexity of $O(m^n)$. For all the other algorithms except NML, the complexity is $O(n)$ when m is considered fixed. More specifically, for Bayes mixtures, the complexity is $O(mn)$ and for k-laststep minimax algorithms, the complexity is $O(m^k n)$.

We mention that it was recently proposed that the computational complexity of the prediction strategy based on NML may be significantly reduced by representing the NML distribution as a Bayes-like mixture with a horizon-dependent prior (Barron et al., 2014). The authors show that for a parametric family with a finite-valued sufficient statistic, the exact NML is achievable by a Bayes mixture with a signed discrete prior designed depending on the horizon n. The resulting prediction strategy may, however, require updating as many as $n/2+1$ weights on each prediction step even in the Bernoulli case, which leads to total time complexity of order n^2.

6. Conclusions

We characterized the achievability of asymptotic minimax coding regret in terms of horizon-dependency. The results have implications on probabilistic prediction, data compression, and model selection based on the MDL principle, all of which depend on predictive models or codes that achieve low logarithmic losses or short code-lengths. For multinomial models, which have been very extensively studied, our main result states that no horizon-independent strategy can be asymptotic minimax. A weaker result involving a stronger minimax notion is given for more general models. Future work can focus on obtaining precise results for different model classes where achievability of asymptotic minimaxity is presently unknown.

Our numerical experiments show that several easily implementable Bayes and other strategies are nearly optimal. In particular, a novel predictor based on a simple asymptotically optimal horizon-dependent Beta (or Dirichlet) prior, for which a closed form expression is readily available, offers a good trade-off between computational cost and worst-case regret. Overall, differences in the maximum regrets of many of the strategies under the Bernoulli model (Figure 2) are small (less than 1 nat). Such small differences may nevertheless be important from a practical point of view. For instance, it has been empirically observed that slight differences in the Dirichlet hyperparameter, leading to relatively small changes in the marginal probabilities, can be significant in Bayesian network structure learning (Silander et al., 2007). Furthermore, the differences are likely to be greater under multinomial ($m > 2$) and other models, which is another direction for future work.

Acknowledgments

The authors thank Andrew Barron and Nicolò Cesa-Bianchi for valuable comments they provided at the WITMSE workshop in Tokyo in 2013. We also thank the anonymous reviewers. This work was supported in part by the Academy of Finland (via the Center-of-Excellence COIN) and by JSPS grants 23700175 and 25120014. Part of this work was carried out when KW was visiting HIIT in Helsinki.

Appendix A. Proof of Lemma 1

Proof The function f is non-decreasing since $f'(x) = \psi(x+1/2) - \ln x \geq 0$ where $\psi(x) = (\ln \Gamma(x))'$ is the digamma function (Merkle, 1998). $\lim_{x\to\infty} f(x) = 0$ is derived from Stirling's formula,

$$\ln \Gamma(x) = \left(x - \frac{1}{2}\right) \ln x - x + \frac{1}{2}\ln(2\pi) + O\left(\frac{1}{x}\right).$$

It immediately follows from $f(0) = -\frac{\ln 2}{2}$ and this limit that $-\frac{\ln 2}{2} \leq f(x) < 0$ for $x \geq 0$. ■

Appendix B. Proof of Theorem 3

Proof Under Assumption 1, we suppose (14) holds for all sufficiently large n and derive contradiction. The inequalities in (14) are equivalent to

$$-\underline{M} + o(1) \leq \ln \frac{p_{\mathrm{NML}}^{(n)}(x^n)}{g(x^n)} \leq o(1).$$

For a horizon-independent strategy g we can expand the marginal probability $g(x^{\tilde{n}})$ in terms of the following sum and apply the above lower bound to obtain

$$\begin{aligned} g(x^{\tilde{n}}) &= \sum_{x_{\tilde{n}+1}^n} g(x^n) = \sum_{x_{\tilde{n}+1}^n} p_{\mathrm{NML}}^{(n)}(x^n) e^{-\ln \frac{p_{\mathrm{NML}}^{(n)}(x^n)}{g(x^n)}} \\ &\leq e^{\underline{M}+o(1)} \sum_{x_{\tilde{n}+1}^n} p_{\mathrm{NML}}^{(n)}(x^n) \qquad (25) \end{aligned}$$

for all $x^{\tilde{n}}$. Then we have

$$\begin{aligned} \max_{x^{\tilde{n}}} \ln \frac{p_{\mathrm{NML}}^{(\tilde{n})}(x^{\tilde{n}})}{g(x^{\tilde{n}})} &= \max_{x^{\tilde{n}}} \left\{ \ln \frac{p_{\mathrm{NML}}^{(\tilde{n})}(x^{\tilde{n}})}{\sum_{x_{\tilde{n}+1}^n} p_{\mathrm{NML}}^{(n)}(x^n)} + \ln \frac{\sum_{x_{\tilde{n}+1}^n} p_{\mathrm{NML}}^{(n)}(x^n)}{g(x^{\tilde{n}})} \right\} \\ &\geq \max_{x^{\tilde{n}}} \left\{ \ln \frac{p_{\mathrm{NML}}^{(\tilde{n})}(x^{\tilde{n}})}{\sum_{x_{\tilde{n}+1}^n} p_{\mathrm{NML}}^{(n)}(x^n)} \right\} - \underline{M} + o(1) \\ &\geq \epsilon + o(1), \end{aligned}$$

where $\epsilon = M - \underline{M} > 0$. The first inequality follows from (25) and the second inequality follows from Assumption 1, which implies $\max_{x^{\tilde{n}}} \ln \frac{p_{\mathrm{NML}}^{(\tilde{n})}(x^{\tilde{n}})}{\sum_{x_{\tilde{n}+1}^n} p_{\mathrm{NML}}^{(n)}(x^n)} \geq M + o(1)$. The above inequality contradicts the asymptotic minimax optimality in (14) with n replaced by $\tilde{n}$. ■

Appendix C. Proof of Lemma 4

Proof In order to prove Lemma 4, we modify and extend the proof in Xie and Barron (2000) for the asymptotic evaluation of $\ln C_n = \ln \sum_{x^n} p(x^n|\hat{\theta}(x^n))$ given by (6) to that of $\ln C_{n|x^{\tilde{n}}} = \ln \sum_{x_{\tilde{n}+1}^n} p(x^n|\hat{\theta}(x^n))$, which is conditioned on the first $\tilde{n}$ samples, $x^{\tilde{n}}$. More specifically, we will prove the following inequalities. Here, $p_{B,w}$ denotes the Bayes mixture defined by the prior $w(\theta)$, $p_{B,1/2}$ and p_{B,α_n} are those with the Dirichlet priors, $\mathrm{Dir}(1/2, \cdots, 1/2)$ (Jeffreys mixture) and $\mathrm{Dir}(\alpha_n, \cdots, \alpha_n)$ where $\alpha_n = \frac{1}{2} - \frac{\ln 2}{2} \frac{1}{\ln n}$ respectively.

$$\frac{m-1}{2} \ln \frac{n}{2\pi} + \tilde{C}_{\frac{1}{2}} + o(1) \leq \sum_{x_{\tilde{n}+1}^n} p_{B,1/2}(x_{\tilde{n}+1}^n | x^{\tilde{n}}) \ln \frac{p(x^n|\hat{\theta}(x^n))}{p_{B,1/2}(x_{\tilde{n}+1}^n|x^{\tilde{n}})} \qquad (26)$$

$$
\begin{aligned}
&\leq \max_{w} \sum_{x_{\tilde{n}+1}^{n}} p_{B,w}(x_{\tilde{n}+1}^{n}|x^{\tilde{n}}) \ln \frac{p(x^{n}|\hat{\theta}(x^{n}))}{p_{B,w}(x_{\tilde{n}+1}^{n}|x^{\tilde{n}})} \\
&= \max_{w} \min_{\overline{p}} \sum_{x_{\tilde{n}+1}^{n}} p_{B,w}(x_{\tilde{n}+1}^{n}|x^{\tilde{n}}) \ln \frac{p(x^{n}|\hat{\theta}(x^{n}))}{\overline{p}(x_{\tilde{n}+1}^{n}|x^{\tilde{n}})} \\
&\leq \min_{\overline{p}} \max_{x_{\tilde{n}+1}^{n}} \ln \frac{p(x^{n}|\hat{\theta}(x^{n}))}{\overline{p}(x_{\tilde{n}+1}^{n}|x^{\tilde{n}})} \\
&= \ln \sum_{x_{\tilde{n}+1}^{n}} p(x^{n}|\hat{\theta}(x^{n})) = \ln C_{n|x^{\tilde{n}}} \\
&\leq \max_{x_{\tilde{n}+1}^{n}} \ln \frac{p(x^{n}|\hat{\theta}(x^{n}))}{p_{B,\alpha_{n}}(x_{\tilde{n}+1}^{n}|x^{\tilde{n}})} \\
&\leq \frac{m-1}{2} \ln \frac{n}{2\pi} + \tilde{C}_{\frac{1}{2}} + o(1), \qquad (27)
\end{aligned}
$$

where the first equality follows from Gibbs' inequality, and the second equality as well as the second to last inequality follow from the minimax optimality of NML (Shtarkov, 1987). Let us move on to the proof of inequalities (26) and (27). The rest of the inequalities follow from the definitions and from the fact that maximin is no greater than minimax. To derive both inequalities, we evaluate $\ln \frac{p(x^{n}|\hat{\theta}(x^{n}))}{p_{B,\alpha}(x_{\tilde{n}+1}^{n}|x^{\tilde{n}})}$ for the Bayes mixture with the prior $\mathrm{Dir}(\alpha, \cdots, \alpha)$ asymptotically. It follows that

$$
\begin{aligned}
\ln \frac{p(x^{n}|\hat{\theta}(x^{n}))}{p_{B,\alpha}(x_{\tilde{n}+1}^{n}|x^{\tilde{n}})} &= \ln \frac{\prod_{j=1}^{m} \left(\frac{n_j}{n}\right)^{n_j}}{\frac{\Gamma(\tilde{n}+m\alpha)}{\Gamma(n+m\alpha)} \prod_{j=1}^{m} \frac{\Gamma(n_j+\alpha)}{\Gamma(l_j+\alpha)}} \\
&= \sum_{j=1}^{m} n_j \ln n_j - n \ln n - \sum_{j=1}^{m} \ln \Gamma(n_j+\alpha) + \ln \Gamma(n+m\alpha) + \ln \tilde{C}_{\alpha} \\
&= \sum_{j=1}^{m} \left\{ n_j \ln n_j - n_j - \ln \Gamma(n_j+\alpha) + \frac{1}{2}\ln(2\pi) \right\} \\
&\quad + \left(m\alpha - \frac{1}{2} \right) \ln n - (m-1)\frac{1}{2}\ln(2\pi) + \ln \tilde{C}_{\alpha} + o(1), \qquad (28)
\end{aligned}
$$

where $\tilde{C}_{\alpha}$ is defined in (17) and we applied Stirling's formula to $\ln \Gamma(n+m\alpha)$.

Substituting $\alpha = 1/2$ into (28), we have

$$
\ln \frac{p(x^{n}|\hat{\theta}(x^{n}))}{p_{B,1/2}(x_{\tilde{n}+1}^{n}|x^{\tilde{n}})} = \sum_{j=1}^{m} \left(c_{n_j} + \frac{\ln 2}{2} \right) + \frac{m-1}{2} \ln \frac{n}{2\pi} + \ln \tilde{C}_{1/2} + o(1),
$$

where

$$
c_k = k \ln k - k - \ln \Gamma(k+1/2) + \frac{1}{2} \ln \pi, \qquad (29)
$$

for $k \geq 0$. Since from Lemma 1, $-\frac{\ln 2}{2} < c_k$,

$$
\ln \frac{p(x^{n}|\hat{\theta}(x^{n}))}{p_{B,1/2}(x_{\tilde{n}+1}^{n}|x^{\tilde{n}})} > \frac{m-1}{2} \ln \frac{n}{2\pi} + \ln \tilde{C}_{1/2} + o(1),
$$

holds for all x^n, which proves the inequality (26).

Substituting $\alpha = \alpha_n = \frac{1}{2} - \frac{\ln 2}{2}\frac{1}{\ln n}$ into (28), we have

$$\begin{aligned}\ln \frac{p(x^n|\hat{\theta}(x^n))}{p_{B,\alpha_n}(x^n_{\tilde{n}+1}|x^{\tilde{n}})} &= \sum_{j=1}^{m}\left\{n_j \ln n_j - n_j - \ln\Gamma(n_j+\alpha_n) + \frac{1}{2}\ln\pi\right\} \\ &\quad + \frac{m-1}{2}\ln\frac{n}{2\pi} + \ln\tilde{C}_{1/2} + o(1).\end{aligned}$$

Assuming that the first l n_js ($j = 1, \cdots, l$) are finite and the rest are large (tend to infinity as $n \to \infty$) and applying Stirling's formula to $\ln\Gamma(n_j+\alpha_n)$ ($j = l+1, \cdots, m$), we have

$$\ln \frac{p(x^n|\hat{\theta}(x^n))}{p_{B,\alpha_n}(x^n_{\tilde{n}+1}|x^{\tilde{n}})} = \sum_{j=1}^{l} c_{n_j} + \sum_{j=l+1}^{m} d_{n_j} + \frac{m-1}{2}\ln\frac{n}{2\pi} + \ln\tilde{C}_{1/2} + o(1), \tag{30}$$

where c_k is defined in (29) and

$$d_k = \frac{\ln 2}{2}\left(\frac{\ln k}{\ln n} - 1\right)$$

for $1 < k \le n$. Since $c_k \le 0$ follows from Lemma 1 and $d_k \le 0$, we obtain

$$\ln \frac{p(x^n|\hat{\theta}(x^n))}{p_{B,\alpha_n}(x^n_{\tilde{n}+1}|x^{\tilde{n}})} \le \frac{m-1}{2}\ln\frac{n}{2\pi} + \ln\tilde{C}_{1/2} + o(1), \tag{31}$$

for all x^n, which proves the inequality (27). ■

Appendix D. Proof of Lemma 5

Proof The summation in (20) is decomposed into three parts,

$$\begin{aligned}&\{n_1 \ln n_1 - \ln\Gamma(n_1+\alpha)\} + \{(n'-n_1)\ln(n'-n_1) - \ln\Gamma(n'-n_1+\alpha)\} \\ &+ \sum_{j=2}^{m-1}\{n_j \ln n_j - \ln\Gamma(n_j+\alpha)\},\end{aligned}$$

where $n' = n - \sum_{j=2}^{m-1} n_j$. We analyze the regret of the multinomial case by reducing it to the binomial case since the summation in the above expression is constant with respect to n_1. Hence, we focus on the regret of the binomial case with sample size n',

$$R(z) = z\ln z - \ln\Gamma(z+\alpha) + (n'-z)\ln(n'-z) - \ln\Gamma(n'-z+\alpha),$$

as a function of $0 \le z \le \frac{n'}{2}$ because of the symmetry. We prove that the maximum of R is attained at the boundary ($z = 0$) or at the middle $z = \frac{n'}{2}$. We will use the following inequalities for $z \ge 0$,

$$\left(\Psi'(z)\right)^2 + \Psi^{(2)}(z) > 0, \tag{32}$$

and

$$2\left(-\Psi^{(2)}(z)\right)^{3/2} - \Psi^{(3)}(z) > 0, \tag{33}$$

which are directly obtained from Theorem 2.2 of Batir (2007).

The derivative of R is

$$R'(z) = h(z) - h(n' - z),$$

where

$$h(z) = \ln z - \Psi(z + \alpha).$$

We can prove that $h'(z) = \frac{1}{z} - \Psi'(z + \alpha)$ has at most one zero since (32) shows that the derivative of the function $z - \frac{1}{\Psi'(z+\alpha)}$ is positive, which implies that it is monotonically increasing from $-1/\Psi'(\alpha) < 0$ and hence has at most one zero coinciding with the zero of h'. Noting also that $\lim_{z\to 0} h(z) = -\infty$ and $\lim_{z\to\infty} h(z) = 0$, we see that there are the following two cases: (a) $h(z)$ is monotonically increasing in the interval $(0,\ n')$, and (b) $h(z)$ is unimodal with a unique maximum in $(0,\ n')$. In the case of (a), R' has no zero in the interval $(0,\ n'/2)$, which means that R is V-shaped, takes global minimum at $z = \frac{n'}{2}$, and has the maxima at the boundaries. In the case of (b), $R'(z) = 0$ has at most one solution in the interval $(0,\ n'/2)$, which is proved as follows.

The higher order derivatives of R are

$$\begin{aligned} R^{(2)}(z) &= h'(z) + h'(n' - z), \\ R^{(3)}(z) &= h^{(2)}(z) - h^{(2)}(n' - z), \end{aligned}$$

where $h^{(2)}(z) = -\frac{1}{z^2} - \Psi^{(2)}(z + \alpha)$. Let the unique zero of $h'(z)$ be z^* (if there is no zero, let $z^* = \infty$). If $z^* < \frac{n'}{2}$, since for $z^* \le z < n'/2$, $h'(z) \le 0$ and $h'(n' - z) \le 0$, we have $R^{(2)}(z) \le 0$, which means that $R'(z)$ is monotonically decreasing to $R'\left(\frac{n'}{2}\right) = 0$. That is, $R'(z) > 0$ for $z^* \le z < \frac{n'}{2}$. Hence, we focus on $z \le z^*$ and prove that $R'(z)$ is concave for $z \le z^*$, which, from $\lim_{z\to 0} R'(z) = -\infty$, means that $R'(z)$ has one zero if $R^{(2)}\left(\frac{n'}{2}\right) = 2h'\left(\frac{n'}{2}\right) < 0$, and $R'(z)$ has no zero otherwise.[3]

For $z \le z^*$, since $\frac{1}{z} > \Psi'(z + \alpha)$ holds, we have

$$h^{(2)}(z) = -\frac{1}{z^2} - \Psi^{(2)}(z + \alpha) < -\Psi'(z + \alpha)^2 - \Psi^{(2)}(z + \alpha) < 0, \tag{34}$$

from (32). Define $\tilde{h}(z) = z - \frac{1}{\sqrt{-\Psi^{(2)}(z+\alpha)}}$, for which $\tilde{h}(z) = 0$ is equivalent to $h^{(2)}(z) = 0$. Then $\tilde{h}(0) < 0$ and it follows from (33) that

$$\tilde{h}'(z) = 1 - \frac{\Psi^{(3)}(z + \alpha)}{2\left(-\Psi^{(2)}(z + \alpha)\right)^{3/2}} > 0,$$

which implies that $\tilde{h}(z)$ is monotonically increasing, and hence that $h^{(2)}(z) = 0$ has at most one solution. Let z^{**} be the unique zero of $h^{(2)}(z)$ (if there is no zero, let $z^{**} = \infty$). Noting

[3] In case (b) where $h(z)$ is unimodal with a maximum in $(0,\ n')$, the condition that $h'\left(\frac{n'}{2}\right) \ge 0$ is equivalent to $z^* \ge \frac{n'}{2}$.

that $\lim_{z\to 0} h^{(2)}(z) = -\infty$, we see that $h^{(2)}(z) < 0$ for $z < z^{**}$ and $h^{(2)}(z) > 0$ for $z > z^{**}$. From (34), $z^* < z^{**}$ holds. For $z < z^{**}$, since $h^{(2)}(z) < 0$ implies that $-\frac{1}{z^2} < \Psi^{(2)}(z+\alpha)$, and hence $\frac{1}{z} > \sqrt{-\Psi^{(2)}(z+\alpha)}$ holds, it follows from (33) that

$$h^{(3)}(z) = \frac{2}{z^3} - \Psi^{(3)}(z+\alpha) > 2\left(-\Psi^{(2)}(z+\alpha)\right)^{3/2} - \Psi^{(3)}(z+\alpha) > 0.$$

This means that $h^{(2)}(z)$ is monotonically increasing for $z < z^{**}$. Therefore, $h^{(2)}(z)$ is negative and monotonically increasing for $z < z^{**}$, implying that $R^{(3)}(z)$ has no zero for $z \le z^{**}$ since $h^{(2)}(z) < h^{(2)}(n'-z)$, that is, $R^{(3)}(z) < 0$ holds. Thus $R'(z)$ is concave for $z \le z^* < z^{**}$, and hence $R'(z)$ has at most one zero between 0 and z^*.

Note that $\lim_{z\to 0} R'(z) = -\infty$ and $R'(n'/2) = 0$. If $R'(z) = 0$ has no solution in $(0,\ n'/2)$, that is, if $h'\left(\frac{n'}{2}\right) = \frac{2}{n'} - \Psi'\left(\frac{n'}{2}+\alpha\right) \ge 0$, the regret function looks similarly to the case of (a), and the maxima are attained at boundaries. If $R'(z) = 0$ has a solution in $(0,\ n'/2)$, that is, if $\frac{2}{n'} - \Psi'\left(\frac{n'}{2}+\alpha\right) < 0$, R' changes its sign around the solution from negative to positive as z grows. In this case, R is W-shaped with possible maximum at the boundaries or at the middle.

We see that in any case, the maximum is always at the boundary or at the middle. Therefore, as a function of the count n_1, $R(n_1)$ is maximized at $n_1 = 0$ or at $n_1 = \lfloor \frac{n'}{2} \rfloor$ (or $n_1 = \lfloor \frac{n'}{2} \rfloor + 1$ if n' is odd). The same argument applies to optimizing n_j $(j = 2, 3, \cdots, m-1)$. Thus, if the counts are such that for any two indices, i and j, $n_i > n_j + 1 > 1$, then we can increase the regret either by replacing one of them by the sum, $n_i + n_j$ and the other one by zero or by replacing them by new values n'_i and n'_j such that $|n_i - n_j| \le 1$. This completes the proof of the lemma. ■

Appendix E. Proof of Theorem 6

Proof The proof of Lemma 4 itself applies to the case where $\tilde{n} = 0$ and $l_j = 0$ for $j = 1, \cdots, m$ as well. Since, in this case, $\tilde{C}_{1/2} = \ln \frac{\Gamma(1/2)^m}{\Gamma(m/2)}$, the inequality (31) in the proof gives the right inequality in (24).

Furthermore, in (30), we have

$$\sum_{j=1}^{l} c_{n_j} + \sum_{j=l+1}^{m} d_{n_j} > -(m-1)\frac{\ln 2}{2} + o(1). \tag{35}$$

This is because, from Lemma 1 and definition, $c_{n_j}, d_{n_j} > -\frac{\ln 2}{2}$ and for at least one of j, n_j is in the order of n since $\sum_{j=1}^{n} n_j = n$, which means that $d_{n_j} = o(1)$ for some j. Substituting (35) into (30), we obtain the left inequality in (24) with $M = \frac{1}{2}(m-1)\ln 2$. ■

References

K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

A. R. Barron, T. Roos, and K. Watanabe. Bayesian properties of normalized maximum likelihood and its fast computation. In *Proc. 2014 IEEE International Symposium on Information Theory*, pages 1667–1671, 2014.

P. Bartlett, P. Grünwald, P. Harremoës, F. Hedayati, and W. Kotłowski. Horizon-independent optimal prediction with log-loss in exponential families. In *JMLR: Workshop and Conference Proceedings: 26th Annual Conference on Learning Theory*, volume 30, pages 639–661, 2013.

N. Batir. On some properties of digamma and polygamma functions. *Journal of Mathematical Analysis and Applications*, 328(1):452–465, 2007.

N. Cesa-Bianchi and G. Lugosi. Worst-case bounds for the logarithmic loss of predictors. *Machine Learning*, 43(3):247–264, 2001.

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

Y. Freund. Predicting a binary sequence almost as well as the optimal biased coin. In *Proc. 9th Annual Conference on Computational Learning Theory*, pages 89–98, 1996.

P. D. Grünwald. *The Minimum Description Length Principle*. The MIT Press, 2007.

F. Hedayati and P. L. Bartlett. Exchangeability characterizes optimality of sequential normalized maximum likelihood and Bayesian prediction with Jeffreys prior. In *JMLR: Workshop and Conference Proceedings: 15th International Conference on Artificial Intelligence and Statistics*, volume 22, pages 504–510, 2012.

P. Kontkanen and P. Myllymäki. A linear-time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters*, 103(6):227–233, 2007.

W. Kotłowski and P. D. Grünwald. Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. In *JMLR: Workshop and Conference Proceedings: 24th Annual Conference on Learning Theory*, volume 19, pages 457–476, 2011.

R. E. Krichevsky. Laplace's law of succession and universal encoding. *IEEE Trans. Information Theory*, 44(1):296–303, 1998.

R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *IEEE Trans. Information Theory*, IT-27(2):199–207, 1981.

P. S. Laplace. *A Philosophical Essay on Probabilities*. Dover, New York, 1795/1951.

F. Liang and A. Barron. Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Trans. Informaton Theory*, 50:2708–2726, 2004.

H. Luo and R. Schapire. Towards minimax online learning with unknown time horizon. In *JMLR: Workshop and Conference Proceedings: 31st International Conference on Machine Learning*, volume 32, pages 226–234, 2014.

N. Merhav and M. Feder. Universal prediction. *IEEE Trans. Information Theory*, 44: 2124–2147, 1998.

M. Merkle. Conditions for convexity of a derivative and some applications to the Gamma function. *Aequationes Mathematicae*, 55:273–280, 1998.

J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Information Theory*, IT-42(1):40–47, 1996.

J. Rissanen and T. Roos. Conditional NML universal models. In *Proc. 2007 Information Theory and Applications Workshop*, pages 337–341. IEEE Press, 2007.

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

T. Silander, P. Kontkanen, and P. Myllymäki. On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In *Proc. 27th Conference on Uncertainty in Artificial Intelligence*, pages 360–367, 2007.

T. Silander, T. Roos, and P. Myllymäki. Learning locally minimax optimal Bayesian networks. *International Journal of Approximate Reasoning*, 51(5):544–557, 2010.

J. Takeuchi and A. R. Barron. Asymptotically minimax regret for exponential families. In *Proc. 20th Symposium on Information Theory and its Applications*, pages 665–668, 1997.

E. Takimoto and M. K. Warmuth. The last-step minimax algorithm. In *Algorithmic Learning Theory, Lecture Notes in Computer Science*, volume 1968, pages 279–290, 2000.

Q. Xie and A. R. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Trans. Information Theory*, 46(2):431–445, 2000.