ACHIEVABILITY OF ASYMPTOTIC MINIMAX OPTIMALITY IN ONLINE AND BATCH CODING

Kazuho Watanabe[1], *Teemu Roos*[2] *and Petri Myllymäki*[2]

[1]Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5, Takayama-cho, Ikoma-shi, Nara, 630-0192, Japan, wkazuho@is.naist.jp
[2] Department of Computer Science, University of Helsinki,
Helsinki Institute for Information Technology HIIT, PO Box 68, FI-00014, Finland,
teemu.roos@cs.helsinki.fi, petri.myllymaki@cs.helsinki.fi

ABSTRACT

The normalized maximum likelihood model achieves the minimax coding (log-loss) regret for data of fixed sample size n. However, it is a batch strategy, i.e., it requires that n be known in advance. Furthermore, it is computationally infeasible for most statistical models, and several computationally feasible alternative strategies have been devised. We characterize the achievability of asymptotic minimaxity by batch strategies (i.e., strategies that depend on n) as well as online strategies (i.e., strategies independent of n). On one hand, we conjecture that for a large class of models, no online strategy can be asymptotically minimax. We prove that this holds under a slightly stronger definition of asymptotic minimaxity. We also show that in the multinomial model, a Bayes mixture defined by the conjugate Dirichlet prior with a simple dependency on n achieves asymptotic minimaxity for all sequences, thus providing a simpler asymptotic minimax strategy compared to earlier work by Xie and Barron. The numerical results also demonstrate superior finite-sample behavior by a number of novel batch algorithms.

1. INTRODUCTION

The normalized maximum likelihood (NML) distribution is derived as the optimal solution to the minimax problem which minimizes the worst-case regret in code-length (log-loss) of data with fixed sample size n. Although a direct evaluation of the NML distribution involves the computation of a sum over all possible data sets, taking exponential time, linear-time algorithms have been developed for certain models such as multinomials [1, 2]. However, the computation of the NML distribution is still intractable for most models.

Approximating the minimax solution by other easily implementable strategies has been studied. Asymptotic minimaxity is a key feature of such strategies, where the worst-case code-length converges to that of the NML as the sample size tends to infinity. For the multinomial model, Xie and Barron [3] showed that a Bayes procedure defined by a modified Jeffreys prior, where additional mass is assigned to the boundaries of the parameter space, can achieve asymptotic minimax optimality. An alternative technique to this procedure was studied for a more general model class [4].

In the context of online prediction of individual sequences, the focus has been on prediction strategies which can be computed without knowing the sequence length n in advance. We call such strategies *online*, while strategies that take advantage of the knowledge of the sample size n are called *batch*. For online strategies, regret bounds of the form $k \ln n + O(1)$, where k is a constant, have been obtained [5, 6, 7]. Furthermore, it was proved for the Bernoulli model and the exponential families with a constrained parameter space that the minimax optimal regret is achieved, up to the $O(1)$ term, by the Bayesian strategy using the Jeffreys prior and the last-step minimax strategy (a.k.a. the sequential normalized maximum likelihood) [8, 9]. That is, if the regret of the NML is asymptotically expanded as $k^* \ln n + c^* + o(1)$ with constants k^* and c^*, $k = k^*$ holds for these strategies. The asymptotic minimax optimality examines if the optimal constant c^* is also achieved and the maximum regret matches that of the NML up to the $o(1)$ term.

In this extended abstract, we investigate achievability of asymptotic minimaxity by batch and online strategies. We consider a slightly stronger asymptotic minimax property and prove that under a generic condition on the model class, it cannot be achieved by any online strategy (Thm. 1). We conjecture that a similar result also holds for the standard asymptotic minimax notion. We also show that for the multinomial model, a sample-size-dependent Bayes procedure defined by a simpler prior than the modified Jeffreys prior in [3] achieves asymptotic minimaxity under the standard definition, as well as approximately in our stronger sense (Thm. 2). Through numerical experiments (Sect. 4), we demonstrate the achievability of asymptotic minimaxity for batch strategies.

2. NORMALIZED MAXIMUM LIKELIHOOD AND ASYMPTOTIC MINIMAXITY

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,

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

is given by

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

where $C_n = \sum_{x^n} p(x^n|\hat{\theta}(x^n))$ and is called the normalized maximum likelihood (NML) distribution [10]. The minimax regret is given by $\ln C_n$ for all x^n. We mention that in addition to coding and prediction, the code length $-\ln p^{(n)}_{\mathrm{NML}}(x^n)$ has been used as a model-selection criterion [11]; see also [12, 2] and references therein.

Since the normalizing constant C_n is computationally intractable in most models, we consider approximating the minimax optimal NML model by another model $g(x^n)$ and focus on *asymptotic* minimax optimality of g, which is defined by

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

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

Under the following assumption, we can show (Thm. 1 below) that the model g must be dependent on the sample size n to achieve the asymptotic minimax optimality in a slightly stronger sense, as characterized in the theorem.

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*

$$\ln \frac{p^{(\tilde{n})}_{\mathrm{NML}}(x^{\tilde{n}})}{\sum_{x^n_{\tilde{n}+1}} p^{(n)}_{\mathrm{NML}}(x^n)} \to M \quad (n \to \infty), \qquad (2)$$

where $\sum_{x^n_{\tilde{n}+1}} = \sum_{x_{\tilde{n}+1}} \cdots \sum_{x_n}$ denotes the marginalization over $x_{\tilde{n}+1}, \cdots, x_n$.

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

Theorem 1 *If the model g is independent of the sample size n and satisfies $\sum_{x^n_{\tilde{n}+1}} g(x^n) = g(x^{\tilde{n}})$, then it never satisfies*

$$\ln C_n - \underline{M} + o(1) \leq \ln \frac{p(x^n|\hat{\theta}(x^n))}{g(x^n)} \leq \ln C_n + o(1), \quad (3)$$

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 $n \to \infty$.

Note that the condition in Eq. (3) is stronger than the usual asymptotic minimax optimality in Eq. (1), where only the right inequality in Eq. (3) is required. Intuitively, our stronger notion of asymptotic minimaxity requires not only that for all sequences, the regret of the model 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 since reducing the code length for one sequence (compared to the NML model), necessarily increases the code length for at least one other sequence.

When we take g as a Bayes mixture,

$$g(x^n) = \int p(x^n|\theta) q(\theta) d\theta,$$

$\sum_{x^n_{\tilde{n}+1}} g(x^n) = g(x^{\tilde{n}})$ holds if the prior distribution $q(\theta)$ does not depend on n. On the contrary, if $q(\theta)$ depends on n, it is possible that the Bayes mixture achieves Eq. (3) for all x^n. In fact, for the multinomial model (with m categories), the Dirichlet prior $\mathrm{Dir}(\hat{\alpha}_n, \cdots, \hat{\alpha}_n)$ with $\hat{\alpha}_n = \frac{1}{2} - \frac{\ln 2}{2}\frac{1}{\ln n}$ provides an example of such a case as will be proven in Sect. 3.2. Section 3.1 demonstrates that the sequence of all 1s (or all 2s, 3s, etc.) gives $M = \frac{m-1}{2} \ln 2$ in the multinomial model.

3. ASYMPTOTIC MINIMAXITY IN MULTINOMIAL MODEL

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

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

Although a linear-time (in n) algorithm has been obtained for computing the NML distribution of this model [1], we examine asymptotic minimaxity of other strategies for this model whose theoretical properties have been studied in depth [3].

For the multinomial model, the Dirichlet distribution is a conjugate prior, taking the form

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

where $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ is the gamma function and $\alpha > 0$ is a hyperparameter. The Bayes mixture is obtained as follows,

$$\begin{aligned} p_{B,\alpha}(x^n) &= \int \prod_{i=1}^{n} p(x_i|\theta) q(\theta) d\theta \\ &= \frac{\Gamma(m\alpha)}{\Gamma(\alpha)^m} \frac{\prod_{j=1}^{m} \Gamma(n_j + \alpha)}{\Gamma(n + m\alpha)}, \end{aligned} \qquad (4)$$

where n_j is the number of js in x^n. The minimax regret is asymptotically given by [3]

$$\ln C_n = \frac{m-1}{2}\ln\frac{n}{2\pi} + \ln\frac{\Gamma(1/2)^m}{\Gamma(m/2)} + o(1). \quad (5)$$

In the following two subsections, we evaluate the constant M of Assumption 1 and derive an asymptotically minimax optimal hyperparameter α. We use the following lemma in this section.

Lemma 1 *Let*

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

Then for $x \geq 0$*,*

$$0 \leq f(x) < \frac{\ln 2}{2} \quad (6)$$

and $\lim_{x\to\infty} f(x) = \frac{\ln 2}{2}$.

3.1. Change of NML Model

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

$$\begin{aligned} &\ln\frac{p^{(\tilde{n})}_{\mathrm{NML}}(x^{\tilde{n}})}{\sum_{x^n_{\tilde{n}+1}} p^{(n)}_{\mathrm{NML}}(x^n)} \\ = &\ln\frac{\prod_{j=1}^m \left(\frac{l_j}{\tilde{n}}\right)^{l_j}}{\sum_{n_j\geq l_j}\binom{n-\tilde{n}}{n_j-l_j}\prod_{j=1}^m\left(\frac{n_j}{n}\right)^{n_j}} + \ln\frac{C_n}{C_{\tilde{n}}}, \quad (7)\end{aligned}$$

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 $j = 1,2,\cdots,m$. The following lemma 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 Eq. (7). The proof is omitted here.[1]

Lemma 2 $C_{n|x^{\tilde{n}}}$ *is asymptotically evaluated as*

$$\ln C_{n|x^{\tilde{n}}} = \frac{m-1}{2}\ln\frac{n}{2\pi} + \ln\tilde{C}_{\frac{1}{2}} + o(1), \quad (8)$$

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

$$\tilde{C}_\alpha = \frac{\prod_{j=1}^m\Gamma(l_j+\alpha)}{\Gamma(\tilde{n}+m\alpha)}. \quad (9)$$

Substituting Eq. (8) and Eq. (5) into Eq. (7), we have

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

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

$$\sum_{j=1}^m\left\{l_j\ln l_j - \ln\Gamma\left(l_j+\frac{1}{2}\right) - l_j + \frac{1}{2}\ln 2\pi\right\} + o(1).$$

Taking $l_1 = \tilde{n}$, $l_j = 0$ for $j = 2,\cdots,m$, from Lemma 1, we have $\ln\frac{p^{(\tilde{n})}_{\mathrm{NML}}(x^{\tilde{n}})}{p^{(n)}_{\mathrm{NML}}(x^{\tilde{n}})} = \frac{m-1}{2}\ln 2 + o(1)$, that is, Assumption 1 holds with $M = (m-1)\ln 2/2$.

3.2. Optimal Hyperparameter and its Asymptotic Minimaxity

We examine the asymptotic minimaxity of the Bayes mixture in Eq. (4). More specifically, we investigate the minimax optimal hyperparameter

$$\operatorname*{argmin}_\alpha \max_{x^n}\ln\frac{p(x^n|\hat{\theta}(x^n))}{p_{B,\alpha}(x^n)} \quad (10)$$

and show that it is asymptotically approximated by

$$\hat{\alpha}_n = \frac{1}{2} - \frac{\ln 2}{2}\frac{1}{\ln n}. \quad (11)$$

We assume that the maximum regret is attained by both x^n consisting of a single symbol repeated n times as well as x^n with a uniform number n/m of each symbol j. Let the regrets of these two cases be equal,

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

Taking logarithms, using Stirling's formula and ignoring diminishing terms, we have

$$\begin{aligned}&(m-1)\left(\alpha-\frac{1}{2}\right)\ln n - (m-1)\ln\Gamma(\alpha) \\ &-m\left(\alpha-\frac{1}{2}\right)\ln m + (m-1)\frac{\ln 2\pi}{2} = 0. \quad (12)\end{aligned}$$

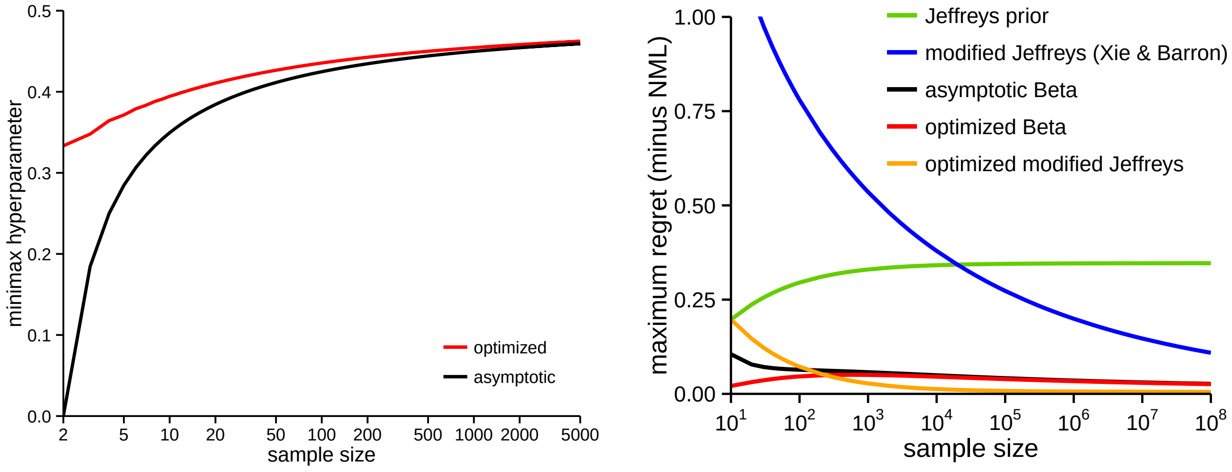
This implies that the optimal α is asymptotically given by

$$\hat{\alpha}_n \simeq \frac{1}{2} - \frac{a}{\ln n}, \quad (13)$$

for some constant a. Substituting this back into Eq. (12) and solving it for a, we obtain Eq. (11).

We numerically calculated the optimal hyperparameter defined by Eq. (10) for the binomial model ($m = 2$). Figure 1(a) shows the optimal α obtained numerically and its asymptotic approximation in Eq. (11). We see that the optimal hyperparameter is well approximated by $\hat{\alpha}_n$

[1] 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 [12, Eq. (11.47), p.323] is valid for the multinomial model with the full parameter set rather than the restricted parameter set discussed in [12].



(a)

(b)

Figure 1. (a) Minimax optimal hyperparameter α for sample size n. (b) Maximum regret for sample size n. The regret of the NML model, $\ln C_n$, is subtracted from the maximum regret of each strategy. The first two algorithms are from earlier work, while the remaining ones are novel.

in Eq. (11) 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 $\hat{\alpha}_n$. It also shows that the lower bound in Eq. (3) is almost attainable for the multinomial model. We will examine the regret of $\hat{\alpha}_n$ numerically in Sect. 4.

Theorem 2 *The Bayes mixture defined by the Dirichlet prior* $\mathrm{Dir}(\hat{\alpha}_n, \cdots, \hat{\alpha}_n)$ *is asymptotically minimax and satisfies*

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

for all x^n*, where* $M = (m-1)\ln 2/2$.

The proof is omitted.

4. NUMERICAL RESULTS

In this section, we numerically calculate the maximum regrets of several methods in the binomial model ($m = 2$).

We calculated the maximum regrets of the Bayes mixtures in Eq. (4) with the hyperparameter optimized by the golden section search and with its asymptotic approximation in Eq. (11). We also investigated the maximum regrets of Xie and Barron's modified Jeffreys prior which is proved to be asymptotically minimax [3]. The modified Jeffreys prior is defined by

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

where δ is the Dirac's delta function and $b_{1/2}(\theta)$ is the density function of the beta distribution with hyperparameters $1/2$, $\mathrm{Beta}(1/2, 1/2)$, which is the Jeffreys prior for the Bernoulli model. We set $\epsilon_n = n^{-1/8}$ as proposed in [3] 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 1(b) shows the maximum regrets of these Bayes mixtures: asymptotic and optimized Beta refer to mixtures with Beta priors (Sect. 3.2), 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 [13], which is known not to be asymptotically minimax. To better show the differences, the regret of the NML model, $\ln C_n$, is subtracted from the maximum regret of each model.

We see that the maximum regrets of these models, except the one based on Jeffreys prior, decrease toward zero 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 while that with the unoptimized weight performs much worse. Note here that we have the explicit form of the asymptotically minimax hyperparameter in Eq. (11) whereas the optimal weight for the modified Jeffreys prior is not known analytically. Note also that unlike the NML, Bayes mixtures can be computed in a sequential manner with respect to $(x_1, \cdots, x_n)$ even if the prior depends on n.

The differences in the maximum regrets under the binomial model in Fig. 1(b) are small (less than 1 nat). However, they may be important even from a practical point of view. For instance, it has been empirically observed that the slightest differences in the Dirichlet hyperparameter can be significant in Bayesian network structure learning [14]. Furthermore, the differences are likely to be greater under multinomial ($m > 2$) and other kinds of models.

5. DISCUSSION & CONCLUSION

In this extended abstract, we proved that the knowledge of the sample size n is required for a strategy to be asymptotically minimax in the sense of Eq. (3). Bartlett et al. proved, as a corollary to their main result, that NML is sample-size dependent in the general exponential family [15]. We have not observed any asymptotically minimax strategy independent of n. This suggests that the lower bound in Eq. (3) may be removed from the condition and the asymptotic minimaxity in the usual sense may be characterized by the dependency on n; in other words, no online strategy can be asymptotically minimax.

For the multinomial model, Thm. 2 shows that a simple dependency on n is sufficient to provide an accurate approximation. In practice, our numerical experiments suggest the superiority of a number of novel algorithms, whose performance is very near that of the NML (Fig. 1(b)).

Future directions include verifying our conjecture about non-achievability of minimax regret by online strategies, developing approximation schemes of the NML distribution for more complex models, and their applications in prediction, data compression, and model selection.

6. ACKNOWLEDGMENTS

This work was supported in part by JSPS KAKENHI Grant Number 23700175 and by the Academy of Finland (project Prime and the COIN Center of Excellence) and by the Finnish Funding Agency for Technology and Innovation (project D2I). Part of this work was carried out while the first author was visiting HIIT.

7. REFERENCES

[1] P. Kontkanen and P. Myllymäki, “A linear-time algorithm for computing the multinomial stochastic complexity,” *Information Processing Letters*, vol. 103, no. 6, pp. 227–233, 2007.

[2] T. Silander, T. Roos, and P. Myllymäki, “Learning locally minimax optimal Bayesian networks,” *International Journal of Approximate Reasoning*, vol. 51, no. 5, pp. 544–557, 2010.

[3] Q. Xie and A. R. Barron, “Asymptotic minimax regret for data compression, gambling, and prediction,” *IEEE Trans. Information Theory*, vol. 46, no. 2, pp. 431–445, 2000.

[4] J. Takeuchi and A. R. Barron, “Asymptotically minimax regret for exponential families,” in *Proc. of the 20th Symposium on Information Theory and its Applications (SITA’97)*, 1997, pp. 665–668.

[5] K. S. Azoury and M. K. Warmuth, “Relative loss bounds for on-line density estimation with the exponential family of distributions,” *Machine Learning*, vol. 43, no. 3, pp. 211–246, 2001.

[6] N. Cesa-Bianchi and G. Lugosi, “Worst-case bounds for the logarithmic loss of predictors,” *Machine Learning*, vol. 43, no. 3, pp. 247–264, 2001.

[7] Y. Freund, “Predicting a binary sequence almost as well as the optimal biased coin,” in *Proc. of Computational Learning Theory (COLT’ 96)*, 1996, pp. 89–98.

[8] E. Takimoto and M. K. Warmuth, “The last-step minimax algorithm,” in *Algorithmic Learning Theory, Lecture Notes in Computer Science*, 2000, vol. 1968, pp. 279–290.

[9] 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*, 2011, vol. 19, pp. 457–476.

[10] Y. M. Shtarkov, “Universal sequential coding of single messages,” *Problems of Information Transmission*, vol. 23, no. 3, pp. 175–186, 1987.

[11] J. Rissanen, “Fisher information and stochastic complexity,” *IEEE Trans. Information Theory*, vol. IT-42, no. 1, pp. 40–47, 1996.

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

[13] R. E. Krichevsky and V. K. Trofimov, “The performance of universal encoding,” *IEEE Trans. Information Theory*, vol. IT-27, no. 2, pp. 199–207, 1981.

[14] Tomi Silander, Petri Kontkanen, and Petri Myllymäki, “On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter,” in *UAI*, 2007, pp. 360–367.

[15] 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*, 2013, vol. 30, pp. 639–661.