Minimax Redundancy for the Class of Memoryless Sources

Qun Xie and Andrew R. Barron, *Member, IEEE*

***Abstract*—Let $X^n = (X_1, \cdots, X_n)$ be a memoryless source with unknown distribution on a finite alphabet of size k. We identify the asymptotic minimax coding redundancy for this class of sources, and provide a sequence of asymptotically minimax codes. Equivalently, we determine the limiting behavior of the minimax relative entropy $\min_{Q_{X^n}} \max_{P_{X^n}} D(P_{X^n}\|Q_{X^n})$, where the maximum is over all independent and identically distributed (i.i.d.) source distributions and the minimum is over all joint distributions. We show in this paper that the minimax redundancy minus $((k-1)/2)\log(n/(2\pi e))$ converges to $\log \int \sqrt{\det I(\boldsymbol{\theta})}\, d\boldsymbol{\theta} = \log(\Gamma(1/2)^k/\Gamma(k/2))$, where $I(\boldsymbol{\theta})$ is the Fisher information and the integral is over the whole probability simplex. The Bayes strategy using Jeffreys' prior is shown to be asymptotically maximin but not asymptotically minimax in our setting. The boundary risk using Jeffreys' prior is higher than that of interior points. We provide a sequence of modifications of Jeffreys' prior that put some prior mass near the boundaries of the probability simplex to pull down that risk to the asymptotic minimax level in the limit.**

***Index Terms*—Universal noiseless coding, minimax redundancy, minimax total relative entropy risk, Jeffreys' prior, asymptotic least favorable prior.**

I. Introduction

We start with a general discussion. Suppose we have a parameterized discrete memoryless source. That is, we have a parametric family of probability mass functions $\{p_{\boldsymbol{\theta}}(x)\colon \boldsymbol{\theta} \in \Theta \subset R^d\}$ on a discrete finite set $\mathfrak{X}$, which generate independent identically distributed (i.i.d.) random variables $X_1, X_2, \cdots, X_n$. Our goal is to code such data with nearly minimal expected codelength, in a minimax sense to be defined later, when we have no information about the generating parameter $\boldsymbol{\theta}$ other than it belongs to the set Θ. This is universal coding, first systematically treated by Davisson [10]. Of particular interest is the case that the family consists of all (i.i.d.) distributions on the alphabet $\mathfrak{X}$.

It is known that the expected codelength is lower-bounded by the entropy of the distribution. When the true $\boldsymbol{\theta}$ is known, this bound can be achieved within one bit. When $\boldsymbol{\theta}$ is unknown, and if we use a mass function q_n on $\mathfrak{X}^n$ and $-\log q_n(x^n)$ bits to code data string x^n, then it induces a redundancy in the expected length of $D(p_{\boldsymbol{\theta}}^n\|q_n)$, where $p_{\boldsymbol{\theta}}^n$ is the joint density of $X^n = (X_1, X_2, \cdots, X_n)$, and $D(\cdot\|\cdot)$ is the Kullback divergence (relative entropy). (Here we ignore the rounding of $-\log q_n(x^n)$ up to an integer required for the coding interpretations, which changes the redundancy by at most one bit from what is identified here.)

Moreover, we may link the above setup with game theory and statistics. Suppose nature picks a $\boldsymbol{\theta}$ from Θ and a statistician chooses a distribution q_n on $\mathfrak{X}^n$ as his best guess of $p_{\boldsymbol{\theta}}^n$. The loss is measured by the total relative entropy $D(p_{\boldsymbol{\theta}}^n\|q_n)$. Then for finite n and prior $W(d\boldsymbol{\theta})$ on Θ the best strategy q_n to minimize the average risk

$$\int D(p_{\boldsymbol{\theta}}^n\|q_n)W(d\boldsymbol{\theta})$$

is the mixture density

$$m_n^W(x^n) = \int p_{\boldsymbol{\theta}}^n(x^n)W(d\boldsymbol{\theta})$$

(called the Bayes procedure), and the resulting average risk is the Shannon mutual information $I(\Theta; X^n)$ (see [8], [10]). Suppose Θ is compact and that $p_\theta(x)$ depends continuously on $\theta \in \Theta$ for every $x \in \mathfrak{X}$. Then the minimax value

$$\min_{q_n} \max_{\theta\in\Theta} D(p_{\boldsymbol{\theta}}^n\|q_n)$$

is equal to the maximin value

$$\max_W \int D(p_{\boldsymbol{\theta}}^n\|m_n^W)W(d\boldsymbol{\theta})$$

which is the capacity of the channel $\Theta \to X^n$. This equality of the minimax and maximin values can be found in Davisson and Leon–Garcia [11] using [13], and is attributed there to Gallager [15]; see [17] for a recent generalization. Moreover, there is a unique minimax procedure and it is realized by a Bayes procedure. Indeed, there exists a least favorable prior W_n^* (also called a capacity achieving prior), for which the corresponding procedure

$$m_n(x^n) = \int p_{\boldsymbol{\theta}}^n(x^n)W_n^*(d\boldsymbol{\theta})$$

is both maximin and minimax (see the discussion following Lemma 5 in the Appendix). The problem of choosing a prior to maximize $I(\Theta; X^n)$ arises in Bayesian statistics as the reference prior method (Bernardo [6]).

Another interpretation of this game is prediction with a cumulative relative entropy loss. Indeed the minimax problem for the total relative entropy is the same as the minimax estimation problem with cumulative relative entropy loss

$$\sum_{n'=0}^{n-1} D(p_\theta\|\hat{p}_{n'})$$

Manuscript received January 1, 1996; revised August 10, 1996. This work was supported in part by NSF under Grant ECS-9410760.

The authors are with the Department of Statistics, Yale University, New Haven, CT 06520 USA.

Publisher Item Identifier S 0018-9448(97)01281-9.

0018–9448/97$10.00 © 1997 IEEE

where the probability function p_θ is estimated using a sequence $\hat{p}_{n'}$ based on $X^{n'}$ for $n' = 0, \cdots, n-1$ (see [8], [9]). Consequences of this prediction interpretation are developed in [3], [18], and [19].

We are interested to know the behavior of the minimax redundancy

$$\min_{q_n} \max_{\boldsymbol{\theta} \in \Theta} D(p_{\boldsymbol{\theta}}^n || q_n) \quad \text{as } n \to \infty.$$

Krichevsky and Trofimov [20] and Davisson *et al.* [12] show that it is $((k-1)/2) \log n + O(1)$ for the family of all distributions on an alphabet of size k (dimension $d = k-1$), and they also provide bounds on the $O(1)$ term. In a more general parametric setting, Rissanen [22] shows that for any code, $(d/2) \log n - o(\log n)$ is an asymptotic lower bound on the redundancy for almost all $\boldsymbol{\theta}$ in the family, and [21] gives a redundancy of $(d/2) \log n + O(1)$ for particular codes based on the minimum description length principle. Barron [1] and Clarke and Barron [8] determine the constant in the redundancy $(d/2) \log n + c_{\boldsymbol{\theta}} + o(1)$ for codes based on mixtures. When regularity conditions are satisfied, including the finiteness of the determinant of Fisher information $I(\boldsymbol{\theta})$, and the restriction of $\boldsymbol{\theta}$ to a compact subset C of the interior of Θ, Clarke and Barron [9] show that the code based on the mixture with respect to Jeffreys' prior is asymptotically maximin and that the maximin and the minimax redundancy minus $(d/2) \log (n/(2\pi e))$ both converge to $\log \int_C \sqrt{\det I(\boldsymbol{\theta})}\, d\boldsymbol{\theta}$. However, the restriction to sets interior to Θ left open the question of the constant in the case of the whole simplex of probabilities on a finite-alphabet case.

In this paper, we allow the distribution $p_{\boldsymbol{\theta}}$ to be any probability on a finite alphabet $\mathfrak{X} = \{a_1, \cdots, a_k\}$. We assume that $p_{\boldsymbol{\theta}}$ puts mass θ_i on letter $\{a_i\}$, for $i = 1, \cdots, k$. The parameter space Θ is the simplex

$$S_{k-1} = \{\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{k-1}) \colon \sum_{i=1}^{k-1} \theta_i \le 1, \text{ all } \theta_i \ge 0\}$$

or equivalently,

$$S_k' = \{\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k) \colon \sum_{i=1}^{k} \theta_i = 1, \text{ all } \theta_i \ge 0\}$$

where

$$\theta_k = 1 - (\theta_1 + \cdots + \theta_{k-1}).$$

The Fisher information determinant is $1/(\theta_1 \cdot \theta_2 \cdot \cdots \cdot \theta_k)$, which is infinite when any θ_i equals 0. The Dirichlet $(\lambda_1, \cdots, \lambda_k)$ distribution has density proportional to $\theta_1^{\lambda_1 - 1} \cdot \cdots \cdot \theta_k^{\lambda_k - 1}$ on Θ for $\lambda_1, \cdots, \lambda_k$ positive. Jeffreys' prior is the one proportional to the square root of the determinant of the Fisher information matrix. In the present context, it coincides with Dirichlet $(1/2, \cdots, 1/2)$ density.

Let the minimax value $V_n = V_n(k)$ for sample size n and alphabet size k be defined by

$$V_n = \min_{q_n} \max_{\theta} D(p_{\boldsymbol{\theta}}^n || q_n) - \frac{k-1}{2} \log \frac{n}{2\pi e}.$$

As we shall see V_n has a limit $V = V(k)$. A sequence of priors W_n is said to be asymptotically least favorable (or capacity achieving) if

$$\int D(p_{\boldsymbol{\theta}}^n || m_n^{W_n}) W_n(d\boldsymbol{\theta}) - ((k-1)/2) \log (n/(2\pi e))$$

converges to V, and the corresponding procedures (based on $m_n^{W_n}$) are said to be asymptotically maximin. A sequence of procedures q_n is said to be asymptotically minimax if

$$\max_{\theta} D(p_{\boldsymbol{\theta}}^n || q_n) - ((k-1)/2) \log (n/(2\pi e))$$

converges to V.

Our main result is the following.

Theorem: The asymptotic minimax and maximin redundancy satisfy

$$\begin{aligned} &\lim_{n \to \infty} \left(\min_{q_n} \max_{\boldsymbol{\theta} \in \Theta} D(p_{\boldsymbol{\theta}}^n || q_n) - \frac{k-1}{2} \log \frac{n}{2\pi e} \right) \\ &\quad = \lim_{n \to \infty} \left(\max_{W \text{on} \Theta} \int_\Theta D(p_{\boldsymbol{\theta}}^n || q_n) W(d\boldsymbol{\theta}) - \frac{k-1}{2} \log \frac{n}{2\pi e} \right) \\ &\quad = \log \frac{\Gamma(1/2)^k}{\Gamma(k/2)}. \end{aligned}$$

Moreover, Jeffreys' prior is asymptotically least favorable (capacity achieving). The corresponding procedure is asymptotically maximin but not asymptotically minimax. A sequence of Bayes procedures using modifications of Jeffreys' prior is exhibited to be asymptotically maximin and asymptotically minimax.

Remark 1: The first equality is free, since minimax equals maximin for each n. The novel part is the identification of the limit and specification of sequences of minimax and maximin procedures.

Remark 2: For finite n, the maximin procedure W_n is also minimax, on the other hand, the asymptotically maximin Jeffreys' procedure is not asymptotically minimax on Θ. The boundary risk using Bayes strategy m_n with Jeffreys' prior is higher than that of interior points, asymptotically. However, after modifying Jeffreys' prior, we find an asymptotically minimax sequence. The redundancy minus $(d/2) \log (n/(2\pi e))$ converges, uniformly for $\boldsymbol{\theta} \in \Theta$, to

$$\log \int_\Theta \sqrt{\det I(\boldsymbol{\theta})}\, d\boldsymbol{\theta} = \log \left(\Gamma(1/2)^k / \Gamma(k/2)\right)$$

as what we would expect from Clarke and Barron [9].

Remark 3: Previously, the best upper and lower bounds on the asymptotic minimax value were based on the values achieved using the Dirichlet $(1/2, \cdots, 1/2)$ prior, see [12], [20], and more recently [25]. Now that we know that this prior is not asymptotically minimax on the whole simplex, we see that the gap between the lower and upper values previously obtained can be closed only by modifying the sequence of procedures.

The outline for the remainder of the paper is as follows. Section II contains some notations and definitions, mostly for the Bernoulli family case $(k = 2)$, and the proof for the this case is presented in Section III. It begins by studying the asymptotic behavior of the redundancy using Jeffreys' prior,

which in turn implies that the asymptotic lower value is at least $\log \pi$. Then we proceed to show that the asymptotic upper value is not greater than $\log \pi$ by providing a sequence of modifications of Jeffreys' prior. From these two results we conclude that the asymptotic value is $\log \pi$ and furthermore Jeffreys' prior is asymptotically least favorable. However, it is not asymptotically minimax because the redundancy at the boundary is higher than $\log \pi$. The extension to higher dimensions is straightforward, as we will show in Section IV. In the Appendix, we include some propositions and lemmas used in the main analysis.

II. Notations and Definitions

For the family of Bernoulli distributions

$$\{p_\theta(x) = \theta^x(1-\theta)^{1-x}: x \in \{0, 1\}, \theta \in [0, 1]\}$$

the Fisher information is $I(\theta) = (\theta(1-\theta))^{-1}$ and Jeffreys' prior density function $w^*(\theta)$ is calculated to be $\theta^{-1/2}(1-\theta)^{-1/2}/\pi$, the Beta $(1/2, 1/2)$ density. Denote $X^n = (X_1, X_2, \cdots, X_n)$, where all X_i's are independent with the Bernoulli (θ) distribution. Let

$$p_\theta^n(x^n) = \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$$

be the joint probability mass of X^n given θ, let

$$\begin{aligned} m_n^*(x^n) &= \int_0^1 p_\theta^n(x^n) w^*(\theta)\, d\theta \\ &= \pi^{-1} \int_0^1 \theta^{\Sigma x_i - 1/2}(1-\theta)^{n-\Sigma x_i - 1/2}\, d\theta \end{aligned}$$

be the mixture with Jeffreys' prior, and let $q_n(x^n)$ be any joint probability mass function on $\{0, 1\}^n$. We use base 2 when writing log.

For $n \geq 1$, define the lower value (the maximin value) as

$$\begin{aligned} \underline{V}_n &= \max_W \min_{q_n} \int_0^1 D(p_\theta^n || q_n) W(d\theta) - \frac{1}{2} \log \frac{n}{2\pi e} \\ &= \max_W \int_0^1 D(p_\theta^n || m_n^W) W(d\theta) - \frac{1}{2} \log \frac{n}{2\pi e} \end{aligned}$$

where the maximum is taken over all probability measures W on $[0, 1]$, and

$$m_n^W(x^n) = \int_0^1 p_s^n(x^n) W(ds)$$

is the mixture density of $p_\theta^n(x^n)$ with prior $W(d\theta)$. We call $\underline{V} = \underline{\lim}_{n\to\infty} \underline{V}_n$ the asymptotic lower value.

Similarly, the upper value (the minimax value) is defined as

$$\overline{V}_n = \min_{q_n} \max_\theta D(p_\theta^n || q_n) - \frac{1}{2} \log \frac{n}{2\pi e}$$

and the asymptotic upper value is $\overline{V} = \overline{\lim}_{n\to\infty} \overline{V}_n$. We remind the reader that $\underline{V}_n = \overline{V}_n$. We maintain the distinction in the notation to focus attention in the proof on obtaining lower and upper bounds, respectively (which will coincide asymptotically as we will see).

For the $k > 2$ case the maximin and minimax values $\underline{V}_n(k)$ and $\overline{V}_n(k)$ and their limits are defined similarly.

III. Proof of the Main Theorem for $k = 2$

Before we go to the formal proof of the main theorem, we give a lemma on the pointwise-asymptotic behavior of $D(p_\theta^n || m_n^*)$ in the Bernoulli case. It is useful in the main proof and may also be of interest itself. The proof for the following lemma may be found in the Appendix (at the end of the proof of Proposition 1).

Lemma 1: For any $\varepsilon > 0$, there exists a $c(\varepsilon)$ such that for $n > 2c$ the following holds uniformly over $\theta \in [c/n, 1-c/n]$:

$$\left| D(p_\theta^n || m_n^*) - \frac{1}{2} \log \frac{n}{2\pi e} - \log \pi \right| \leq \varepsilon.$$

Remark 4: The analysis we give shows that the bound holds with $c(\varepsilon) = 5/\varepsilon$, corresponding to the bound

$$\begin{aligned} |D(p_\theta^n || m_n^*) - (1/2)\log(n/(2\pi e)) - \log \pi| \\ \leq 5/(n \min(\theta, 1-\theta)). \end{aligned}$$

Similar inequalities with error $O(1/(n\delta))$ for $\delta \leq \theta \leq 1-\delta$ have recently been obtained by Suzuki [25].

This lemma extends the range of θ where the pointwise asymptotics is demonstrated from the case of intervals $[\delta, 1-\delta]$, with δ fixed (from [9]) to the case of intervals $[5/(n\varepsilon), 1-5/(n\varepsilon)]$. For instance, with $\varepsilon = 1/\sqrt{n}$ we find that the difference between $D(p_\theta^n || m_n^*)$ and $(1/2)\log n/(2\pi e) + \log \pi$ is bounded by $1/\sqrt{n}$ uniformly in $[5/\sqrt{n}, 1-5/\sqrt{n}]$. As we shall see the asymptotics do not hold uniformly on $[0, 1]$. In essence, Lemma 1 holds because the posterior distribution of θ given X^n is asymptotically normal when θ is bounded away from 0 and 1, or when θ moves at some certain rate to either of these points. But if the rate is too fast, it will destroy the posterior normality. We will show later that when θ is on the boundary, the limiting value is higher than that of any fixed interior point. For $\theta = c_0/n$ with c_0 fixed, $D(p_\theta^n || m_n^*) - (1/2)\log n/(2\pi e)$ may have a limiting value between those achieved at the boundary and at interior points, though we cannot identify this value yet.

We now proceed to the proof of the main theorem for the $k = 2$ case.

A. Lower Value $\underline{V} \geq \log \pi$

Proof: By definition, we need to show that

$$\begin{aligned} \underline{\lim}_n \sup_W \int_0^1 [D(p_\theta^n || m_n^W) - (1/2)\log(n/(2\pi e))] W(d\theta) \\ \geq \log \pi. \end{aligned}$$

It suffices to prove that

$$\begin{aligned} \int_{c/n}^{1-c/n} D(p_\theta^n || m_n^*) w^*(\theta)\, d\theta - (1/2)\log(n/(2\pi e)) \\ \geq \log \pi - o_n(1) \end{aligned}$$

for some $c > 0$ where

$$w^*(\theta) = \theta^{-1/2}(1-\theta)^{-1/2}/\pi$$

is Jeffreys' prior on $[0, 1]$. In fact, from Lemma 1, given any $\varepsilon > 0$, there exists a $c(\varepsilon)$ such that for $n \geq 2c$ and

$\theta \in [c/n, 1-c/n]$

$$D(p_\theta^n||m_n^*) \geq \log \pi + \frac{1}{2}\log\frac{n}{2\pi e} - \varepsilon.$$

Hence

$$\begin{aligned}\int_{c/n}^{1-c/n} \left(D(p_\theta^n||m_n^*) - \frac{1}{2}\log\frac{n}{2\pi e}\right) w^*(\theta)\,d\theta &\geq \int_{c/n}^{1-c/n} (\log\pi - \varepsilon) w^*(\theta)\,d\theta \\ &\geq (\log\pi - \varepsilon)\left(1 - \frac{4}{\pi}\sqrt{\frac{c}{n-c}}\right) \qquad (1)\end{aligned}$$

where the last inequality is from

$$\begin{aligned}\int_0^{c/n} \theta^{-1/2}(1-\theta)^{-1/2}\,d\theta &\leq (1-c/n)^{-1/2}\int_0^{c/n}\theta^{-1/2}\,d\theta \\ &= \left(1-\frac{c}{n}\right)^{-1/2}\cdot 2\left(\frac{c}{n}\right)^{1/2}.\end{aligned}$$

The same bound holds for the integral from $1-c/n$ to 1. Therefore, we have that the liminf of

$$\int_0^1 [D(p_\theta^n||m_n^*) - (1/2)\log(n/(2\pi e))] w^*(\theta)\,d\theta$$

is at least $\log\pi - \varepsilon$. But ε is arbitrary, thus $\underline{V} \geq \log\pi$.

What we have demonstrated will show that Jeffreys' prior is asymptotically least favorable once we have confirmed that $\underline{V}$ cannot exceed $\log\pi$ (see Section III-C below).

Remark 5: An alternative demonstration that $\underline{V} \geq \log\pi$ follows from the weaker result of [9]. In particular, if we restrict $\theta \in [\delta, 1-\delta]$, then

$$\begin{aligned}D(p_\theta^n||m_{n,\delta}^*) - (1/2)\log n/(2\pi e) &\\ \to \int_\delta^{1-\delta} \theta^{-1/2}(1-\theta)^{-1/2}\,d\theta &\end{aligned}$$

uniformly in $\theta \in [\delta, 1-\delta]$, where $m_{n,\delta}^*$ is the mixture with Jeffreys' prior on $[\delta, 1-\delta]$. Letting $\delta \to 0$ establishes $\underline{V} \geq \log\pi$. However, that reasoning uses a sequence of priors depending on δ and does not identify a fixed prior that is asymptotically least favorable on $[0,1]$. The proof we have given above permits identification of an asymptotically least favorable prior. It does not require use of [9] so the proof in the present paper is self-contained.

B. Upper Value $\overline{V} \leq \log\pi$

We show that $\overline{V}_n \leq \log\pi + o_n(1)$ by upper-bounding the risk achieved in the limit by certain procedures. For any given $\varepsilon > 0$, define a prior (which is a modification of Jeffreys' prior) on $[0,1]$ by

$$W_n^\varepsilon(ds) = \eta\delta_{c/n}(ds) + \eta\delta_{1-c/n}(ds) + (1-2\eta)w^*(s)\,ds$$

where δ_a is the distribution that puts unit mass at the point a, the quantity $c = c(\varepsilon)$ is as in Lemma 1, the mass η satisfies $0 < \eta < 1/2$, and $w^*(s)$ is Jeffreys' prior. We also require $n \geq 2c$. The Bayes procedure with respect to the prior W_n^ε uses

$$\begin{aligned}m_n^\varepsilon(x^n) = &\eta p_{c/n}^n(x^n) + \eta p_{1-c/n}^n(x^n) \\ &+ (1-2\eta)\int_0^1 p_s^n(x^n)w^*(s)\,ds.\end{aligned}$$

By definition

$$\overline{V}_n = \min_{q_n}\max_{\theta\in[0,1]} D(p_\theta^n||q_n) - \frac{1}{2}\log\frac{n}{2\pi e}.$$

Use the procedure m_n^ε and partition $[0,1]$ into three intervals to get

$$\begin{aligned}\overline{V}_n &\leq \max_{\theta\in[0,1]} D(p_\theta^n||m_n^\varepsilon) - \frac{1}{2}\log\frac{n}{2\pi e} \\ &= \max\left\{\max_{[0,c/n]} D(p_\theta^n||m_n^\varepsilon), \max_{[c/n,1-c/n]} D(p_\theta^n||m_n^\varepsilon), \right. \\ &\qquad \left. \cdot \max_{[1-c/n,1]} D(p_\theta^n||m_n^\varepsilon)\right\} - \frac{1}{2}\log\frac{n}{2\pi e}. \qquad (2)\end{aligned}$$

We next show that for large n, an upperbound M_n for the supremum over $[c/n, 1-c/n]$ also upper-bounds that over $[0, c/n]$ and $[1-c/n, 1]$, hence $\overline{\lim}_n \overline{V}_n$ is not larger than $\overline{\lim}_n M_n$.

When $\theta \in [0, c/n]$

$$\begin{aligned}D(p_\theta^n||m_n^\varepsilon) &= E_\theta \log\frac{p_\theta^n(X^n)}{m_n^\varepsilon(X^n)} \\ &\leq E_\theta \log\frac{p_\theta^n(X^n)}{\eta p_{c/n}^n(X^n)} \\ &= \log\frac{1}{\eta} + nD(p_\theta||p_{c/n}) \\ &\leq \log\frac{1}{\eta} + nD(p_0||p_{c/n}) \qquad (3) \\ &= \log\frac{1}{\eta} + n\log\frac{1}{1-c/n} \\ &\leq \log\frac{1}{\eta} + 2c \qquad (4)\end{aligned}$$

where inequality (3) holds since $D(p_\theta||p_{c/n})$ is decreasing in θ when $\theta \in [0, c/n]$.

When $\theta \in [1-c/n, 1]$, the same inequality holds.

When $\theta \in [c/n, 1-c/n]$, from Lemma 1

$$\begin{aligned}D(p_\theta^n||m_n^\varepsilon) &\leq E_\theta \log\frac{p_\theta^n(X^n)}{(1-2\eta)\displaystyle\int_0^1 p_s(X^n)w^*(s)\,ds} \\ &= \log\frac{1}{1-2\eta} + D(p_\theta^n||m_n^*) \\ &\leq \log\frac{1}{1-2\eta} + \log\pi + \frac{1}{2}\log\frac{n}{2\pi e} + \varepsilon \qquad (5)\end{aligned}$$

for all $n \geq 2c$.

Now it is seen that (5) eventually will exceed (4) when n increases, as we intended to show. From (2),

$$\overline{V}_n \leq \log 1/(1-2\eta) + \log\pi + \varepsilon$$

for all large n and hence

$$\overline{V} \le \log(1/(1-2\eta)) + \log \pi + \varepsilon.$$

Therefore, upon taking the infimum over $0 < \eta < 1/2$ and $\varepsilon > 0$, we obtain that $\overline{V} \le \log \pi$.

Hence, we have proved that for $\theta \in [0,1]$, the game has a limiting minimax value in agreement with the value $\log \int \sqrt{I(\theta)}\, d\theta$ as in [9], despite the violation of conditions they require. The limiting minimax value is achieved asymptotically by a sequence of modifications of Jeffreys' prior, indexed by η_n and ε_n. Checking the steps in the above proof, we see that the above modification works with $\eta_n \to 0$, $\varepsilon_n \to 0$, and, say, $\eta_n \ge (2e/(n\pi))^{1/4}$, and $\varepsilon_n \ge 10/\log(n\pi/(2e))$.

C. Jeffreys' Prior is Asymptotically Least Favorable

Since $\underline{V} = \log \pi$, to prove that Jeffreys' prior w^* is asymptotically least favorable, we need

$$\underline{\lim}_n \left[\int_0^1 D(p_\theta^n || m_n^*) w^*(\theta)\, d\theta - (1/2)\log(n/(2\pi e)) \right] \ge \log \pi$$

which is already shown in Section III-A. Moreover, a choice of $\varepsilon_n = 1/\sqrt{n}$ in Lemma 1 together with the fact that $|D(p_\theta^n || m_n^*) - 1/2 \log n|$ is bounded by a constant over $\theta \in [0,1]$ (see Lemma 4 in the Appendix) shows that

$$\int_0^1 D(p_\theta^n || m_n^*) w^*(\theta)\, d\theta - (1/2)\log(n/(2\pi e))$$

converges to the asymptotic maximin value at rate $1/\sqrt{n}$.

D. Jeffreys' Prior is Not Asymptotically Minimax

To see that Jeffreys' prior is not asymptotically minimax we use the fact, recently studied in Suzuki [25], that the value of $D(p_\theta^n || m_n^*)$ is largest at the boundary and remains asymptotically larger at the boundary than in the interior.

Indeed, at any interior point θ in $(0,1)$, the asymptotic value of $D(p_\theta^n || m_n^*)$ satisfies

$$\left| D(p_\theta^n || m_n^*) - \frac{1}{2}\log \frac{n}{2\pi e} - \log \pi \right| \le \frac{5}{n\theta(1-\theta)}$$

due to Proposition 1 in the Appendix. Hence

$$D(p_\theta^n || m_n^*) - \frac{1}{2}\log \frac{n}{2\pi e} - \log \pi \to 0$$

as $n \to \infty$, for any interior point θ.

When θ is on the boundary of $[0,1]$, take $\theta = 1$ for example, then using the mixture m_n^* based on Jeffreys' prior, as in Suzuki [25], we have

$$\begin{aligned} D(p_1^n || m_n^*) &= E_1 \log \frac{1}{\displaystyle\int s^n \cdot \frac{1}{\pi} s^{-1/2}(1-s)^{-1/2}\, ds} \\ &= -\log \frac{\Gamma(n+\frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(n+1)\pi} \\ &\approx -\log \frac{(n+\frac{1}{2})^n \cdot e^{-n-1/2}}{(n+1)^{n+1/2} \cdot e^{-n-1}} \frac{1}{\sqrt{\pi}} \\ &\approx \frac{1}{2}\log \frac{n}{2\pi e} + \log \pi + \frac{1}{2}\log(2e) \end{aligned}$$

where we omit the proof of the negligibility of the residual errors from Stirling's approximations.

Therefore $D(p_1^n || m_n^*) - (1/2)\log(n/(2\pi e)) - \log \pi$ converges to $(1/2)\log(2e)$ instead of 0. The limit has a higher value at boundary $\theta = 1$. The scenario is the same on the other boundary point $\theta = 0$. This completes the proof of the theorem.

Remark 6: Davisson *et al.* [12, inequality (61)] obtained $-\log(\Gamma(n+1/2)\Gamma(1/2)/(\Gamma(n+1)\pi))$ as an upper bound on the redundancy for all θ in $[0,1]$. Suzuki [25, Theorem 3] points out that this bound is achieved at the endpoint using Jeffreys' prior. Our analysis shows the perhaps surprising conclusion that it is the lower value of risk achieved by Jeffreys' prior in the interior that matches the asymptotic minimax value.

Remark 7: After the submission of this paper, we have developed other modifications of Jeffreys' prior that are asymptotically minimax. For instance, in place of the small mass points put near the boundary, one can also use a small Beta(α, α) component with $\alpha < 1/2$ mixed with the main Beta$(1/2, 1/2)$ component. Further developments on these priors are in the manuscript [4] which addresses minimal worst case redundancy over all sequences x^n.

IV. Extension to $k \ge 3$ Cases

For the case of an alphabet of size k we recall from Section I that the parameter space is the $(k-1)$-dimensional simplex $\Theta = S_{k-1}$ and that Jeffreys' prior density is given by the Dirichlet $(1/2, \cdots, 1/2)$ density

$$w^*(\boldsymbol{\theta}) = \theta_1^{-1/2} \cdot \cdots \cdot \theta_k^{-1/2} / D_k(1/2, \cdots, 1/2).$$

Here

$$D_k(\lambda_1, \cdots, \lambda_k) = \int_\Theta \theta_1^{\lambda_1 - 1} \cdot \cdots \cdot \theta_k^{\lambda_k - 1} d\theta_1 \cdots d\theta_{k-1}$$

is the Dirichlet integral. In terms of Gamma functions the Dirichlet function may be expressed as

$$D_k(\lambda_1, \cdots, \lambda_k) = \frac{\Gamma(\lambda_1) \cdot \cdots \cdot \Gamma(\lambda_k)}{\Gamma\left(\displaystyle\sum_{i=1}^k \lambda_i\right)}. \tag{6}$$

It follows that

$$\int_\Theta \sqrt{\det(I(\boldsymbol{\theta}))}\, d\boldsymbol{\theta} = D_k(1/2, \cdots, 1/2) = \Gamma(1/2)^k / \Gamma(k/2).$$

We will first show that $\underline{V}(k) \ge \log(\Gamma(1/2)^k/\Gamma(k/2))$ using Jeffreys' prior, in Part 1, then $\overline{V}(k) \le \log(\Gamma(1/2)^k/\Gamma(k/2))$ using modifications of Jeffreys' prior, in Part 2. Consequently, $V(k) = \log(\Gamma(1/2)^k/\Gamma(k/2))$ and Jeffreys' prior is asymptotically least favorable (Part 3). The higher asymptotic value of $D(p_{\boldsymbol{\theta}}^n || m_n^*)$ at the boundary of Θ is demonstrated in Part 4.

Part 1. Asymptotic Lower Value
$\underline{V}(k) \ge \log(\Gamma(1/2)^k/\Gamma(k/2))$

This is parallel to Section III-A of the $k = 2$ case, except that θ is replaced by $\boldsymbol{\theta}$, Lemma 1 is replaced by Proposition

1 of the Appendix, and inequality (1) is replaced by the following argument. With the Dirichlet $(1/2, \cdots, 1/2)$ prior the marginal distribution of θ_i is Beta$(1/2, (k-1)/2)$, thus the contribution of $\{\theta_i \le c/n\}$ to the integral of $w^*(\theta)$ is bounded by

$$\frac{\int_0^{c/n} \theta_i^{-1/2}(1-\theta_i)^{(k-3)/2}\,d\theta_i}{D_2\left(\frac{1}{2}, \frac{k-1}{2}\right)} \le \frac{\int_0^{c/n} \theta_i^{-1/2}\,d\theta_i}{D_2\left(\frac{1}{2}, \frac{k-1}{2}\right)} \le \frac{2(c/n)^{1/2}}{D_2\left(\frac{1}{2}, \frac{k-1}{2}\right)}.$$

Thus as in the previous case, the interior region in which all $\theta_i > c/n$ provides the desired bound and the Bayes risk does not drop below the target level $\log(\Gamma(1/2)^k/\Gamma(k/2))$ by more than order $1/\sqrt{n}$.

Part 2. Asymptotic Upper Value
$\overline{V}(k) \le \log(\Gamma(1/2)^k/\Gamma(k/2))$

Proof: For any $\varepsilon > 0$, let L_i be the intersection of $\{\boldsymbol{\theta}: \theta_i = c/n\}$ with the probability simplex Θ, for $i = 1, \cdots, k$, where $c = c(\varepsilon)$ is chosen as in Proposition 1 in the Appendix. We first define a probability measure μ_i concentrated on L_i with density function (with respect to $d_i\boldsymbol{\theta} = d\theta_1 \cdots d\theta_{i-1} \cdot d\theta_{i+1} \cdots d\theta_{k-1}$, the Lebesgue measure on R^{k-2})

$$\mu_i(\boldsymbol{\theta}) = \frac{\theta_1^{-1/2} \cdots \theta_{i-1}^{-1/2}\theta_{i+1}^{-1/2} \cdots \theta_k^{-1/2} 1_{L_i}(\theta)}{\int_{L_i} (\theta_1^{-1/2} \cdots \theta_{i-1}^{-1/2}\theta_{i+1}^{-1/2} \cdots \theta_k^{-1/2})\,d_i\boldsymbol{\theta}}.$$

Then we define a prior on Θ (which is a modification of the original Jeffreys' prior) as

$$W_n^{\varepsilon}(d\boldsymbol{\theta}) = \frac{\varepsilon}{k}\sum_{i=1}^{k} \mu_i(\boldsymbol{\theta})\,d_i\boldsymbol{\theta} + (1-\varepsilon)w^*(\boldsymbol{\theta})\,d\boldsymbol{\theta}.$$

For this prior, the Bayes procedure to minimize

$$\int D(p_{\boldsymbol{\theta}}^n||q_n) W_n^{\varepsilon}(d\boldsymbol{\theta})$$

uses

$$\begin{aligned} q_n(x^n) &= \int_{\Theta} p_{\boldsymbol{\theta}}^n(x^n) W_n^{\varepsilon}(d\boldsymbol{\theta}) \\ &= \frac{\varepsilon}{k}\sum_{i=1}^{k}\int_{L_i} p_{\boldsymbol{\theta}}^n(x^n)\mu_i(\boldsymbol{\theta})d_i\boldsymbol{\theta} \\ &\quad + (1-\varepsilon)\int_{\Theta} p_{\boldsymbol{\theta}}^n(x^n)w^*(\boldsymbol{\theta})\,d\boldsymbol{\theta} \\ &= \frac{\varepsilon}{k}\sum_{i=1}^{k} m_i(x^n) + (1-\varepsilon)\frac{\prod_{i=1}^{k}\Gamma(T_i+\frac{1}{2})}{\Gamma\left(n+\frac{k}{2}\right)} \end{aligned}$$

where

$$T_i = \sum_{j=1}^{n} 1_{\{X_j = a_i\}}$$

is the number of occurances of the symbol a_i in the sequence x^n, and

$$\begin{aligned} &m_i(x^n) \\ &\quad = \int_{L_i} p_{\boldsymbol{\theta}}^n(x^n)\mu_i(\boldsymbol{\theta})\,d_i\boldsymbol{\theta} \\ &\quad = \frac{\int_{L_i} p_{\boldsymbol{\theta}}^n(x^n)(\theta_1^{-1/2} \cdots \theta_{i-1}^{-1/2}\theta_{i+1}^{-1/2} \cdots \theta_k^{-1/2})\,d_i\boldsymbol{\theta}}{\int_{L_i} (\theta_1^{-1/2} \cdots \theta_{i-1}^{-1/2}\theta_{i+1}^{-1/2} \cdots \theta_k^{-1/2})\,d_i\boldsymbol{\theta}} \\ &\quad = \frac{D_{k-1}(T_1+\frac{1}{2}, \cdots, T_{i-1}+\frac{1}{2}, T_{i+1}+\frac{1}{2}, \cdots, T_k+\frac{1}{2})}{D_{k-1}(\frac{1}{2}, \cdots, \frac{1}{2})} \end{aligned}$$

where the last equality is by the substitution $\theta_j = \theta_j'(1-c/n)$ (for $j \ne i$, $j < k$), $\theta_k = \sum_{j\ne i, j<k}\theta_j$.

Define $R_i = \{\boldsymbol{\theta}: n\theta_i \le c\}$ (for $i = 1, \cdots, k$) and $R = \Theta - \cup R_i$. Now observe that

$$\sup_{\boldsymbol{\theta}\in\Theta} D(p_{\boldsymbol{\theta}}^n||q_n) = \max\Big\{\sup_{R_1} D(p_{\boldsymbol{\theta}}^n||q_n), \cdots, \sup_{R_k} D(p_{\boldsymbol{\theta}}^n||q_n), \sup_{R} D(p_{\boldsymbol{\theta}}^n||q_n)\Big\}. \quad (7)$$

We will find an upper bound for $\sup_{\boldsymbol{\theta}\in\Theta} D(p_{\boldsymbol{\theta}}^n||q_n)$ by showing that it upper-bounds all the suprema over $R_1, \cdots, R_k, R$.

For $\boldsymbol{\theta} \in R$, we have

$$\begin{aligned} D(p_{\boldsymbol{\theta}}^n||q_n) &\le E_{\boldsymbol{\theta}} \log \frac{\theta_1^{T_1} \cdots \theta_k^{Tk}}{(1-\varepsilon)m_n^*(X^n)} \\ &= \log\frac{1}{1-\varepsilon} + D(p_{\boldsymbol{\theta}}^n||m_n^*) \\ &\le \log\frac{1}{1-\varepsilon} + \log\frac{\Gamma(\frac{1}{2})^k}{\Gamma\left(\frac{k}{2}\right)} + \frac{k-1}{2\pi e} + \varepsilon\log e \end{aligned} \quad (8)$$

where the last inequality is by Proposition 1 of the Appendix.

For $\boldsymbol{\theta} \in R_i$, say $i = 1$, that is, $0 \le \theta_1 \le c/n$

$$\begin{aligned} D(p_{\boldsymbol{\theta}}^n||q_n) &\le E_{\boldsymbol{\theta}} \log \frac{p_{\boldsymbol{\theta}}^n(X^n)}{\frac{\varepsilon}{k} m_1(x^n)} \\ &= \log\frac{k}{\varepsilon} + n\theta_1 \log\theta_1 \\ &\quad + \sum_{j=2}^{k} n\theta_j \log\theta_j + \log D_{k-1}(\tfrac{1}{2}, \cdots, \tfrac{1}{2}) \\ &\quad - E_{\boldsymbol{\theta}} \log D_{k-1}(T_2+\tfrac{1}{2}, \cdots, T_k+\tfrac{1}{2}). \end{aligned} \quad (9)$$

We now construct a set of multinomial variables $(T_2', \cdots, T_k')$ with parameters $(n, \theta_2/(1-\theta_1), \cdots, \theta_k/(1-\theta_1))$ from $(T_1, \cdots, T_k) \sim$Multinomial $(n, \theta_1, \cdots, \theta_k)$, by randomly reassigning the T_1 occurrences of the outcome $\{a_1\}$ to $\{a_2\}, \cdots, \{a_k\}$ with probabilities

$$\boldsymbol{\theta}' = \theta_2/(1-\theta_1), \cdots, \theta_k/(1-\theta_1)$$

respectively. That is, given T_1, we obtain new counts $T_j' = T_j + \xi_j$ for $j = 2, \cdots, k$, where $(\xi_2, \cdots, \xi_k) \sim$Multinomial $(T_1, \boldsymbol{\theta}')$. Hence $(T_2', \cdots, T_k') \sim$Multinomial $(n, \boldsymbol{\theta}')$, conditionally for each value of T_1 and hence unconditionally. Now

since $T_j' \geq T_j$ and by the property of the Dirichlet integral that it decreases in any parameter, we have

$$E_{\boldsymbol{\theta}} \log D_{k-1}(T_2+\tfrac{1}{2}, \cdots, T_k+\tfrac{1}{2}) \geq E_{\boldsymbol{\theta}'} \log D_{k-1}(T_2'+\tfrac{1}{2}, \cdots, T_k'+\tfrac{1}{2}). \quad (10)$$

Also observe that

$$\sum_{j=2}^{k} n\theta_j \log \theta_j = \left[\sum_{j=2}^{k}\left(n\frac{\theta_j}{1-\theta_1}\log\frac{\theta_j}{1-\theta_1} + n\frac{\theta_j}{1-\theta_1}\log(1-\theta_1)\right)\right](1-\theta_1) \leq \sum_{j=2}^{k} n\frac{\theta_j}{1-\theta_1}\log\frac{\theta_j}{1-\theta_1}. \quad (11)$$

Applying (10) and (11) to (9), we obtain

$$\begin{aligned} D(p_{\boldsymbol{\theta}}^n||q_n) &\leq \log\frac{k}{\varepsilon} + \sum_{j=2}^{k} n\frac{\theta_j}{1-\theta_1}\log\frac{\theta_j}{1-\theta_1} \\ &\quad + \log D_{k-1}(\tfrac{1}{2}, \cdots, \tfrac{1}{2}) \\ &\quad - E_{\boldsymbol{\theta}'} \log D_{k-1}(T_2'+\tfrac{1}{2}, \cdots, T_k'+\tfrac{1}{2}) \\ &= \log\frac{k}{\varepsilon} + \sum_{j=2}^{k} n\frac{\theta_j}{1-\theta_1}\log\frac{\theta_j}{1-\theta_1} \\ &\quad - E_{\boldsymbol{\theta}'} \log\frac{D_{k-1}(T_2'+\frac{1}{2}, \cdots, T_k'+\frac{1}{2})}{D_{k-1}(\frac{1}{2}, \cdots, \frac{1}{2})} \\ &= \log\frac{k}{\varepsilon} + D(p_{\boldsymbol{\theta}'}^n||m_n^{**}) \end{aligned}$$

where m_n^{**} is the procedure based on Jeffreys' prior on the reduced $(k-2)$-dimensional probability simplex S_{k-1}' and $\boldsymbol{\theta}' \in S_{k-1}'$. Now a course upper bound on $D(p_{\boldsymbol{\theta}'}^n||m_n^{**})$ is sufficient for this lower-dimensional piece. Lemma 4 gives

$$D(p_{\boldsymbol{\theta}'}^n||m_n^{**}) \leq \frac{k-2}{2}\log\frac{n}{2\pi e} + C_{k-1} \quad (12)$$

for all $\boldsymbol{\theta}' \in \Theta$ and some constant C_{k-1}. Observe that $(k-2)/2$ in (12) provides a smaller multiplier of the $\log n$ factor than achieved in the middle region R (see term (8)). Consequently, for all large n

$$D(p_{\boldsymbol{\theta}}^n||q_n) - \frac{k-1}{2}\log\frac{n}{2\pi e} \leq \log\frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2})} + \log\frac{1}{1-\varepsilon}$$

uniformly in $\boldsymbol{\theta} \in \Theta$. Let n go to ∞ and then ε go to 0. The proof is completed.

Part 3. Jeffreys' Prior is Asymptotically Least Favorable

As shown in Part 1, the Bayes average risk using Jeffreys' prior converges to the value V, now identified to be the asymptotically maximin value. Thus Jeffreys' prior is asymptotically least favorable.

Part 4. Jeffreys' Prior is not Asymptotically Minimax

On the k-dimensional simplex, the asymptotic maximum redundancy of the procedure based on Jeffreys' prior is achieved at vertex points, and it is higher asymptotically than in the interior or on any face of the simplex. Here we quantify the asymptotic redundancy within each dimensional face.

From Proposition 1 of the Appendix, for any $\boldsymbol{\theta}$ with $\theta_i > 0$ for $i = 1, \cdots, k$, we have

$$D(p_{\boldsymbol{\theta}}^n||m_n^*) - \frac{k-1}{2}\log\frac{n}{2\pi e} - \log\frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2})} \to 0 \quad \text{as } n \to \infty.$$

For a vertex point such as $\boldsymbol{e} = (1, 0, \cdots, 0)$, as shown by Suzuki [25]

$$\begin{aligned} D(p_{\boldsymbol{\theta}}^n||m_n^*) &= \log\frac{\Gamma(\frac{1}{2})^k \Big/ \Gamma(\frac{k}{2})}{\int \theta_1^{n-1/2}\theta_2^{-1/2}\cdots\theta_k^{-1/2}\,d\theta_1\cdots d\theta_{k-1}} \\ &= \log\frac{\Gamma(n+\frac{k}{2})}{\Gamma(n+\frac{1}{2})\Gamma(\frac{1}{2})\cdots\Gamma(\frac{1}{2})} + \log\frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2})} \\ &\approx \left(\frac{k-1}{2}\log\frac{n}{2\pi e} + \log\frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2})}\right) + \frac{k-1}{2}\log 2e \end{aligned} \quad (13)$$

which is asymptotically larger than in the interior by the amount of $((k-1)/2)\log 2e$.

More generally, for a face point such as $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_L, 0, \cdots, 0)$, where $1 \leq L \leq k-1$ and $\theta_j > 0$ for $j = 1, \cdots, L$, we have ((14) and (15) at the top of the following page) where $\boldsymbol{\theta}^L = (\theta_1, \cdots, \theta_L)$ and m_n^{**} is the mixture density with Jeffreys' prior on the L-dimensional simplex. Stirling's formula yields the following approximation:

$$\log\frac{\Gamma(n+\frac{k}{2})}{\Gamma(n+\frac{L}{2})} = \frac{k-L}{2}\log n + o(1). \quad (16)$$

From (15) and (16), and expanding $D(p_{\boldsymbol{\theta}^L}^n||m_n^{**})$ using Proposition 1 of the Appendix, we have

$$\begin{aligned} &D(p_{\boldsymbol{\theta}^L}^n||m_n^{**}) \\ &= \left(\frac{L-1}{2}\log\frac{n}{2\pi e} + \log\frac{\Gamma(\frac{1}{2})^L}{\Gamma(\frac{L}{2})}\right) + \left(\frac{k-L}{2}\log n + \log\frac{\Gamma(\frac{L}{2})}{\Gamma(\frac{k}{2})}\right) + o(1) \quad (17) \\ &= \left(\frac{k-1}{2}\log\frac{n}{2\pi e} + \log\frac{(\Gamma(\frac{1}{2}))^k}{\Gamma(\frac{k}{2})}\right) + \frac{k-L}{2}\log(2e) + o(1). \quad (18) \end{aligned}$$

Comparing (18) with (13), we see that the asymptotic redundancy at a $\boldsymbol{\theta}$ on a face (i.e., $1 < L < k$) of the simplex is less than the risk at vertex points (i.e., $L = 1$) by the amount of $((L-1)/2)\log(2e)$. In the interior we have $L = k$ nonzero

$$D(p_{\boldsymbol{\theta}}^n||m_n^*) = E_{\boldsymbol{\theta}} \log \frac{\theta_1^{T_1} \cdots \theta_L^{T_L}}{\int \theta_1^{T_1-1/2} \cdots \theta_L^{T_L-1/2} \cdot \theta_{L+1}^{-1/2} \cdots \theta_k^{-1/2} / D_k(\frac{1}{2}, \cdots, \frac{1}{2})\, d\theta_1 \cdots d\theta_{k-1}}$$

$$= E_{\boldsymbol{\theta}} \log \frac{\theta_1^{T_1} \cdots \theta_L^{T_L} \cdot D_k(\frac{1}{2}, \cdots, \frac{1}{2})}{\Gamma(T_1+\frac{1}{2}) \cdots \Gamma(T_L+\frac{1}{2}) \cdot \Gamma(\frac{1}{2})^{k-L} \Big/ \Gamma(n+\frac{k}{2})}$$

$$= E_{(\theta_1, \cdots, \theta_L)} \log \frac{\theta_1^{T_1} \cdots \theta_L^{T_L}}{\Big(\Gamma(T_1+\frac{1}{2}) \cdots \Gamma(T_L+\frac{1}{2}) \Big/ \Gamma(n+\frac{L}{2})\Big) \Big/ D_L(\frac{1}{2}, \cdots, \frac{1}{2})}$$

$$+ \log \frac{D_k(\frac{1}{2}, \cdots, \frac{1}{2}) \Gamma(n+\frac{k}{2})}{D_L(\frac{1}{2}, \cdots, \frac{1}{2}) \Gamma(n+\frac{L}{2}) \Gamma(\frac{1}{2})^{k-L}} \tag{14}$$

$$= D(p_{\boldsymbol{\theta}^L}^n||m_n^{**}) + \log \frac{\Gamma(n+\frac{k}{2}) \Gamma(\frac{L}{2})}{\Gamma(n+\frac{L}{2}) \Gamma(\frac{k}{2})} \tag{15}$$

coordinates, and the asymptotic value is less than at a vertex by the amount $((k-1)/2) \log (2e)$, as we have seen.

Remark 8: Using Davisson *et al.* [12, inequality (61)], Suzuki [25, Theorem 3] proves that for each n, the value of $D(p_{\boldsymbol{\theta}}^n||m_n^*)$ is maximized at the vertices. Here we have determined the asymptotic gap between vertex, face, and interior points.

Appendix

Proposition 1. Pointwise Asymptotic Behavior of $D(p_{\boldsymbol{\theta}}^n||m_n^*)$: For an interior point $\boldsymbol{\theta}$ of the simplex S_k', i.e., $\theta_i > 0$ for $i = 1, \cdots, k$, the following holds.

$$\left| D(p_{\boldsymbol{\theta}}^n||m_n^*) - \frac{k-1}{2} \log \frac{n}{2\pi e} - \log \frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2})} \right| \le \left(\frac{k(k-1)}{3n} + \frac{3}{2n} \sum_{i=1}^k \frac{1}{\theta_i} \right) \log e. \tag{19}$$

In particular, for any $\varepsilon > 0$, if we take $c = 2k/\varepsilon$, then for $n > kc$ and $n\theta_i \ge c$ for $i = 1, \cdots, k$, the last quantity is less than $\varepsilon \log e$. For $k = 2$, when $c = 10/(3\varepsilon)$, the above quantity is less than $\varepsilon \log e$.

Proof: The bound is invariant to the choice of base of the logarithm. It suffices to prove the bound with the choice of the natural logarithm. By definition, and letting

$$T_j = \sum_1^n 1_{\{X_i = \{a_j\}\}} \quad \text{for } j = 1, \cdots, k$$

we have

$$D(p_{\boldsymbol{\theta}}^n||m_n^*) = E_{\boldsymbol{\theta}} \ln \frac{p_{\boldsymbol{\theta}}^n(X^n)}{m_n^*(X^n)}$$

$$= \sum_{i=1}^k n\theta_i \ln \theta_i - E_{\boldsymbol{\theta}} \ln \frac{\int_{S_{k-1}} \theta_1^{T_1-1/2} \cdots \theta_k^{T_k-1/2}\, d\boldsymbol{\theta}}{\int_{S_{k-1}} \theta_1^{-1/2} \cdots \theta_k^{-1/2}\, d\boldsymbol{\theta}}$$

$$= \sum_{i=1}^k n\theta_i \ln \theta_i - E_{\boldsymbol{\theta}} \ln D_k(T_1+\tfrac{1}{2}, \cdots, T_k+\tfrac{1}{2}) + \ln D_k(\tfrac{1}{2}, \cdots, \tfrac{1}{2}). \tag{20}$$

Now applying the relationship between Dirichlet integrals and Gamma functions (6) and Stirling's approximation refined by Robbins [23], and shown to be valid for real $x \ge 0$ in Whittaker and Watson [26, p. 253],

$$\Gamma(x) = \sqrt{2\pi} x^{x-1/2} e^{-x} (1+r)$$

with

$$|r| \le e^{1/12x} - 1 \tag{21}$$

we may rewrite the middle term of (20)

$$E_{\boldsymbol{\theta}} \ln D_k(T_1+\tfrac{1}{2}, \cdots, T_k+\tfrac{1}{2}) = E_{\boldsymbol{\theta}} \ln \frac{\prod_1^k \left(\sqrt{2\pi}(T_i+\frac{1}{2})^{T_i}\right)}{\sqrt{2\pi}(n+\frac{k}{2})^{n+(k-1)/2}} + E_{\boldsymbol{\theta}} \ln \frac{\prod_1^k (1+r_i)}{1+r_0}$$

$$= \frac{k-1}{2} \ln 2\pi + \overbrace{\sum_1^k E_{\theta_i} T_i \ln (T_i+\tfrac{1}{2})}^{\text{(A)}} - \overbrace{\left(n+\frac{k-1}{2}\right) \ln \left(n+\frac{k}{2}\right)}^{\text{(B)}} + \overbrace{E_{\boldsymbol{\theta}} \ln \frac{\prod_1^k (1+r_i)}{1+r_0}}^{\text{(C)}} \tag{22}$$

where r_i and r_0 are residuals from Stirling's approximations to $\Gamma(T_i+1/2)$ and $\Gamma(n+k/2)$, respectively.

We now upper- and lower-bound terms (A), (B), and (C) in (22) separately.

For the deterministic term (B), we have

$$\left|\left(n+\frac{k-1}{2}\right)\ln\left(n+\frac{k}{2}\right)-\left(n\ln n+\frac{k-1}{2}\ln n+\frac{k}{2}\right)\right| \le \frac{k(k-1)}{4n}. \quad (23)$$

For term (C), we apply Lemma 2 of this Appendix to get

$$-\sum_{i=1}^{k}\frac{1}{3n\theta_i}-\frac{1}{6n} \le E_{\boldsymbol{\theta}}\ln\frac{\prod_1^k(1+r_i)}{1+r_0} \le \sum_{i=1}^{k}\frac{1}{6n\theta_i}+\frac{1}{6n} \quad (24)$$

where $1/(6n)$ is a bound for $\log(1+r_0)$.

For term (A), we first rewrite each summand in (A),

$$E_{\theta_i}T_i\ln(T_i+\tfrac{1}{2}) = \overbrace{E_{\theta_i}T_i\ln T_i}^{(A_1)}+\overbrace{E_{\theta_i}T_i\ln\left(1+\frac{1}{2T_i}\right)}^{(A_2)}. \quad (25)$$

Term (A_1) is well-controlled: from Lemma 3 of this Appendix, we have

$$-\frac{1}{48n\theta_i} \le E_{\boldsymbol{\theta}}(T_i\ln T_i)-n\theta_i\ln n\theta_i-\frac{1-\theta_i}{2} \le \frac{1}{n\theta_i}. \quad (26)$$

Now we lower-bound the (A_2) term in (25)

$$E_{\theta_i}T_i\ln\left(1+\frac{1}{2T_i}\right) \ge \frac{1}{2}-E_{\theta_i}\left[\frac{1}{2(T_i+1)}\right] \ge \frac{1}{2}-\frac{1}{2n\theta_i}$$

where the first inequality holds because

$$x\log(1+1/(2x)) \ge 1/2-1/(2x+2) \quad \text{for } x\ge 0$$

and the second one holds because $E_\theta(1/(T+1)) \le 1/(n\theta)$, a useful lemma ([2, Lemma 2]) which is also used in the proof of Lemma 2. Now observe that term (A_2) is upper-bounded by $1/2$ since

$$x\log(1+1/(2x)) \le 1/2 \quad \text{for } x\ge 0.$$

Consequently,

$$\frac{1}{2}-\frac{1}{2n\theta_i} \le E_{\theta_i}T_i\ln\left(1+\frac{1}{2T_i}\right) \le \frac{1}{2}. \quad (27)$$

Combining (26) and (27) then summing the result over i yields a bound for term (A)

$$-\frac{25}{48n}\sum_{i=1}^{k}\frac{1}{\theta_i} \le \sum_{i=1}^{k}E_{\theta_i}T_i\ln\left(T_i+\frac{1}{2}\right) -\sum_{i=1}^{k}n\theta_i\ln n\theta_i-(k-\tfrac{1}{2}) \le \frac{1}{n}\sum_{i=1}^{k}\frac{1}{\theta_i}. \quad (28)$$

Now we incorporate (23), (24), and (28) into a bound for $D(p_{\boldsymbol{\theta}}^n||m_n^*)$

$$\left|D(p_{\boldsymbol{\theta}}^n||m_n^*)-\frac{k-1}{2}\ln\frac{n}{2\pi e}-\ln\frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2})}\right| \le \frac{k(k-1)}{3n}+\frac{3}{2n}\sum_{i=1}^{k}\frac{1}{\theta_i}.$$

In particular, if we take $c = 2k/\varepsilon$, then for $n \ge kc$ and $n\theta_i \ge c$ for $i = 1, \cdots, k$, the last quantity is less than ε. This completes the proof of Proposition 1.

When $k = 2$, we may take $c = 10/(3\varepsilon)$. In fact, Lemma 1 follows from the proposition by setting

$$c(\varepsilon) = (10/3)\varepsilon^{-1}\log_2 e \le 5/\varepsilon$$

to get an error bound of ε uniformly over $[c(\varepsilon)/n,\ 1-c(\varepsilon)/n]$. (Recall that we used base 2 for the logarithm in Lemma 1.)

Lemma 2. Negligibility of Residuals: Let r be the residual from Stirling's approximation to $\Gamma(T+1/2)$, where $T \sim$ Binomial (n, θ). Then for any $\varepsilon > 0$, when $\theta \notin \{0, 1\}$

$$-\frac{1}{6T+3}\log e \le \log(1+r) \le \frac{1}{12T+6}\log e.$$

Consequently, using that $E_\theta(1/(T+1)) \le 1/(n\theta)$, we have

$$-\frac{1}{3n\theta}\log e \le E_\theta\log(1+r) \le \frac{1}{6n\theta}\log e.$$

Proof: As before, assume e as the base of the logarithm in the proof. We first prove the lower bound part. From Stirling's approximation (21) with $x = T+1/2$, the residual r satisfies

$$|r| \le \exp\left(\frac{1}{12T+6}\right)-1. \quad (29)$$

Thus

$$\ln(1+r) \ge \ln\left(2-\exp\frac{1}{12T+6}\right) \ge \ln\left(\exp\left(-\frac{2}{12T+6}\right)\right) = -\frac{1}{6T+3}$$

where the second inequality is from a simple inequality verified by calculus

$$2-e^{s/2} \ge e^{-s}$$

for $0 \le s \le 1/3$. Here we have plugged in $s = 1/(6T+3)$.

The upper bound is more direct. Again using (29), we have

$$\ln(1+r) \le \ln\left(\exp\frac{1}{12T+6}\right) = \frac{1}{12T+6}.$$

Thus we have completed the proof of Lemma 2.

Lemma 3. Local Property of $E_\theta(T \log T)$: Let $T \sim$ Binomial (n, θ). For $n\theta > 2$

$$-\frac{1}{48n\theta} \log e \le E_\theta(T \log T) - n\theta \log n\theta - \frac{1-\theta}{2} \log e \le \frac{1}{n\theta} \log e.$$

Proof: Base e for the logarithm is still assumed in the proof. We begin with the lower bound part. By Taylor's expansion of $y \ln y$ around z

$$\begin{aligned} y \ln y &= z \ln z + (y-z)(1+\ln z) \\ &\quad + \frac{1}{2}(y-z)^2 \frac{1}{z} + \frac{1}{6}(y-z)^3\left(-\frac{1}{z^2}\right) \\ &\quad + \frac{1}{24}(y-z)^4 \frac{2}{y_*^3} \\ &\ge z \ln z + (y-z)(1+\ln z) \\ &\quad + \frac{1}{2}(y-z)^2 \frac{1}{z} + \frac{1}{6}(y-z)^3\left(-\frac{1}{z^2}\right) \end{aligned}$$

where y_* is between y and z. Replace y with T and z with $n\theta$, then take expectation with respect to E_θ to get

$$\begin{aligned} E_\theta(T \ln T) &\ge n\theta \ln n\theta + \frac{1}{2}\operatorname{Var}_\theta(T) \cdot \frac{1}{n\theta} \\ &\quad + \frac{1}{6} E_\theta(T-n\theta)^3 \cdot \left(-\frac{1}{(n\theta)^2}\right) \\ &= n\theta \ln n\theta + \frac{1-\theta}{2} + \frac{1}{6} E_\theta(T-n\theta)^3 \cdot \frac{-1}{(n\theta)^2} \\ &\ge n\theta \ln n\theta + \frac{1-\theta}{2} - \frac{1}{48n\theta} \end{aligned}$$

where for the last inequality we used

$$E_\theta(T-n\theta)^3 = -n\theta(1-3\theta+2\theta^2).$$

For the upper bound part, we need the following inequality: for $y \ge 0$, $z > 0$,

$$y \ln y \le z \ln z + (y-z)(1+\ln z) + \frac{(y-z)^2}{2z} - \frac{(y-z)^3}{6z^2} + \frac{(y-z)^4}{3z^3}. \quad (30)$$

To prove (30), we substitute y with $(t+1)z$, then it reduces to show that for all $t \ge -1$

$$(t+1)\ln(t+1) \le t + \frac{t^2}{2} - \frac{t^3}{6} + \frac{t^4}{3}$$

and this simplified inequality is readily verifiable by using $\log(t+1) \le t - t^2/2 + t^3/3$.

Now replace y with $T \sim$ Binomial (n, θ) and z with $n\theta$ in (30) and take expectation to get

$$\begin{aligned} E_\theta T \ln T &\le n\theta \ln(n\theta) + \frac{1-\theta}{2} - \frac{1-3\theta+2\theta^2}{6n\theta} \\ &\quad + \frac{1+3n\theta(1-\theta)}{3(n\theta)^2} \\ &\le n\theta \ln(n\theta) + \frac{1-\theta}{2} - \frac{1-3\theta}{6n\theta} \\ &\quad + \frac{1}{6n\theta} + \frac{1-\theta}{n\theta} \\ &\le n\theta \ln(n\theta) + \frac{1-\theta}{2} + \frac{1}{n\theta} \end{aligned}$$

when $n\theta > 2$. Thus we have proved Lemma 3.

We recall in the next lemma a bound of the form $((k-1)/2)\log n + O(1)$ on the redundancy of the code based on the Dirichlet $(1/2, \cdots, 1/2)$ prior; see [12], [20], and [24]. (Such a bound without precise determination of the constant plays a role in our analysis of the minimax asymptotics with the modified Jeffreys' prior in the vicinity of lower-dimensional faces of the simplex.)

Lemma 4. A Uniform Upper Bound for $D(p_{\boldsymbol\theta}^n || m_n^*)$: There is a constant C_k such that for all $\boldsymbol\theta \in S_k'$, $n \ge 1$, we have

$$D(p_{\boldsymbol\theta}^n || m_n^*) \le \frac{k-1}{2} \log n + C_k.$$

Moreover, for all sequences X^n

$$\log \frac{p_{\boldsymbol\theta}^n(X^n)}{m_n^*(X^n)} \le \frac{k-1}{2} \log n + C_k.$$

Proof: We still use e as the logarithm base in the proof. Let $\hat{\boldsymbol\theta}$ be the maximum-likelihood estimator of $\boldsymbol\theta$, that is, $\hat\theta_i = T_i/n$ for $i = 1, \cdots, k$. Then

$$\begin{aligned} \ln \frac{p_{\boldsymbol\theta}^n(X^n)}{m_n^*(X^n)} &\le \ln \frac{p_{\hat{\boldsymbol\theta}}^n(X^n)}{m_n^*(X^n)} \\ &= \ln \frac{\prod_{i=1}^k \left(\frac{T_i}{n}\right)^{T_i}}{D_k(T_1+\frac{1}{2}, \cdots, T_k+\frac{1}{2}) / D_k(\frac{1}{2}, \cdots, \frac{1}{2})} \\ &= \sum_{i=1}^n T_i \ln T_i - n \ln n \\ &\quad - \ln \frac{\prod_{i=1}^k \Gamma(T_i+\frac{1}{2})}{\Gamma(n+\frac{k}{2})} + \ln \frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2})}. \end{aligned} \quad (31)$$

By Stirling's formula,

$$\begin{aligned} \ln \frac{\prod_{i=1}^k \Gamma(T_i+\frac{1}{2})}{\Gamma(n+\frac{k}{2})} &= \frac{k-1}{2} \ln 2\pi + \sum_{i=1}^k T_i \ln\left(T_i+\frac{1}{2}\right) + \\ &\quad -\left(n+\frac{k-1}{2}\right) \ln\left(n+\frac{k}{2}\right) - \sum_{i=1}^k \ln \frac{1+r_i}{1+r_0} \\ &\ge \sum_{i=1}^k T_i \ln T_i + -\left(n+\frac{k-1}{2}\right) \ln n \\ &\quad - \text{Constant } (k) \end{aligned}$$

Incorporation of the above inequality in (31) yields

$$\ln \frac{p_{\boldsymbol\theta}^n(X^n)}{m_n^*(X^n)} \le \frac{k-1}{2} \ln n + C_k.$$

This completes the proof of Lemma 4.

The following Lemma is verified by standard decision theory.

Lemma 5. Maximin Procedure is Minimax: Under relative entropy loss, if the game has a value, if there is a minimax procedure, and if there is a least favorable prior, then the minimax procedure is unique, and the procedure corresponding to any least favorable prior is minimax.

Proof: Suppose that $\{p_\theta: \theta \in \Theta\}$ is a parametric family, W^* is any least favorable prior, and Q^* is any minimax procedure. By [8, Proposition 3.A]

$$m^{W^*} = \int p_\theta W^*(d\theta)$$

is the unique Bayes procedure with respect to the prior W^*. To prove the lemma, it suffices to show that $Q^* = m^{W^*}$, that is, Q^* is Bayes with respect to the prior W^*. Thus the desired equation is

$$\int D(P_\theta||Q^*)W^*(d\theta) = \inf_Q \int D(P_\theta||Q)W^*(d\theta). \quad (32)$$

Let the minimax value be $\overline{V}$ and maximin value be $\underline{V}$. Since W^* is a least favorable prior, we have

$$\inf_Q \int D(P_\theta||Q^*)W^*(d\theta) = \underline{V}.$$

Also since Q^* is minimax, we have

$$\sup_\theta D(P_\theta||Q^*) = \overline{V}.$$

Now observe that

$$\int D(P_\theta||Q^*)W^*(d\theta) \geq \inf_Q \int D(P_\theta||Q)W^*(d\theta) = \underline{V}$$

and that

$$\int D(P_\theta||Q^*)W^*(d\theta) \leq \sup_\theta D(P_\theta||Q^*) = \overline{V}.$$

Finally, since $\overline{V} = \underline{V}$, we obtain the desired conclusion. This completes the proof of Lemma 5.

Note that the conclusion holds for any loss for which the Bayes procedure given a prior is unique.

Remark 9: The conditions of this lemma are satisfied in our context. Indeed, it is known that with relative entropy loss, the game has value and there exists a minimax procedure, see, e.g., Haussler [17]. Next since $\mathfrak{X}$ is finite, one may view $p_\theta(x^n)$, $x^n \in \mathfrak{X}^n$ as a point in a bounded set of dimension $|\mathfrak{X}|^n - 1$ (contained within the probability simplex) and view a Bayes mixture $m_n(x^n)$, $x^n \in \mathfrak{X}^n$ as a point in the closure of the convex hull of this set, so from convex set theory any such mixture may be represented as a convex combination using not more than $|\mathfrak{X}|^n$ points θ. Imposing one more convex combination constraint we may at the same time represent the Bayes risk value

$$\int D(p_\theta^n||m_n)w(d\theta)$$

as a finite convex combination of the values $D(p_\theta^n||m_n)$, using not more than $|\mathfrak{X}|^n + 1$ points θ to represent both m_n and the Bayes risk; see, e.g., [7, p. 310], [14, p. 96], [16, p. 96], or [5]. That is, for any prior W (even a continuous prior) there exist $\theta_1, \cdots, \theta_J$ and $(w_1, \cdots, w_J) \in S_J$ with $J \leq |\mathfrak{X}|^n + 1$ such that

$$m^{W^*}(x^n) = \int p_\theta(x^n)W(d\theta) = \sum_{j=1}^J w_j p_{\theta_j}(x^n)$$

and

$$\int D(p_\theta^n||m_n)W(d\theta) = \sum_{i=1}^J w_i D(p_\theta^n||m_n)$$

(using the counts $T_1, \cdots, T_k$ as sufficient statistics reduces the cardinality bound to $J \leq \binom{n+k-1}{k-1} + 2$). If also Θ is compact and $p_\theta(x)$ is continuous in θ for each x, then

$$\sum_{i=1}^J w_i D\left(p_{\theta_i}^n||\sum_{j=1}^J w_j p_{\theta_j}^n\right)$$

is a continuous function of $(\theta_1, \cdots, \theta_J, w_1, \cdots, w_J)$ in the compact set $\Theta^J \times S_J$ and hence there exists a point $(\theta_1^*, \cdots, \theta_J^*, w_1^*, \cdots, w_J^*)$ that achieves the maximum Bayes risk. That is, there exists a least favorable prior. This confirms the conditions of Lemma 5 under the continuity and compactness conditions of the family p_θ when $\mathfrak{X}$ is discrete, and justifies the claim that there exist least favorable priors yielding a unique maximin and minimax procedure. Since these exact considerations are not essential to our asymptotics, we have relegated Lemma 5 and this discussion to the Appendix.

REFERENCES

[1] A. R. Barron, "Logically smooth density estimation," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1985.

[2] A. R. Barron, E. C. Van der Meulen, and L. O. Györfy, "Distribution estimation consistent in total variation and in two types of information divergence," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1437–1454, 1992.

[3] A. R. Barron and T. M. Cover, "A bound on the financial value of information," *IEEE Trans. Inform. Theory*, vol. 34, pp. 1097–1100, 1988.

[4] A. R. Barron and Q. Xie, "Asymptotic minimax regret for data compression, gambling and prediction," preprint, May 1996.

[5] J. Berger, J. M. Bernardo, and M. Mendoza, "On priors that maximize expected information," in *Recent Developments in Statistics and Their Applications.* Seoul, Korea: Freedom Academy Pub., 1989.

[6] J. M. Bernardo, "Reference posterior distributions for Bayesian inference," *J. Roy. Statist. Soc. Ser. B*, vol. 41, pp. 113–147, 1979.

[7] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Orlando, FL: Academic, 1986.

[8] B. S. Clarke and A. R. Barron, "Asymptotic of Bayes information—Theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, pp. 453–471, 1990.

[9] ———, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Stat. Planning and Inference*, vol. 41, pp. 37–60, 1994.

[10] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, 1973.

[11] L. D. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 166–174, 1980.

[12] L. D. Davisson, R. J. McEliece, M. B. Pursley, and M. S. Wallace, "Efficient universal noiseless source codes," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 269–279, 1981.

[13] T. S. Ferguson, *Mathematical Statistics: A Decision-Theoretic Approach.* New York: Academic, 1967.

[14] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[15] ———, "Supplementary notes #3, notes on universal coding," MIT course 6.441 notes, Mar. 1974.

[16] J. A. Hartigan, *Bayes Theory.* New York: Springer-Verlag, 1983.

[17] D. Haussler, "A general minimax result for relative entropy," preprint, 1995.
[18] D. Haussler and A. R. Barron, "How well do Bayes methods work for on-line prediction of $\{\pm 1\}$ values," in *Proc. 1992 NEC Symp. on Computational Cognition*, ch. 4.
[19] D. Haussler and M. Opper, "General bounds on the mutual information between a parameter and n conditionally independent observation," preprint, 1995.
[20] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199–207, 1981.
[21] J. Rissanen, "Universal coding, information, prediction and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, 1984.
[22] _____, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, pp. 1080–1100, Sept. 1986.
[23] H. E. Robbins, "A remark on Stirling's formula," *Ameri. Math. Monthly*, vol. 62, pp. 26–29, 1955.
[24] J. Shtarkov, "Coding of discrete sources with unknown statistics," in *Topics in Information Theory* (Coll. Math. Soc. J. Boyai, no. 16), I. Csiszár and P. Elias, Eds. Amsterdam, The Netherlands: North Holland, 1977, pp. 559–574.
[25] J. Suzuki, "Some notes on universal noiseless coding," *IEICE Trans. Fundamentals*, vol. E78-A, no. 12, Dec. 1995.
[26] E. T. Whittaker amd G. N. Watson, *A Course of Modern Analysis*, 4th ed. Cambridge, U.K.: Cambridge Univ. Press, 1963.