INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Minimax Coding and Prediction

A Dissertation

Presented to the Faculty of the Graduate School

of

Yale University

in Candidacy for the Degree of

Doctor of Philosophy

by

Qun Xie

Dissertation Director: Andrew R. Barron

May 1997

Acknowledgment

I owe deeply to my advisor, Professor Andrew R. Barron, who shines my research path with his insightful guidance. He generously spends incredibly much time and energy throughout this research, without which this work could not be carried out. I have learned from him not only the effective research methodology but also a positive attitude toward life in general and a considerate thinking of people around us.

I thank Professor David Pollard, who introduces me to the research of statistics. He noticed a topic related to this dissertation. He greatly encouraged me during my time of difficulty.

I must also express my appreciation to Professor John Hartigan. He inspired me in doing statistics theoretically and practically.

Professors Joe Chang sets a good example in expressing a mathematical idea clearly and pursuing thorough understanding of science relentlessly. Professor Nicholas Hengartnar kindly agreed to read my dissertation.

I wish to thank my friend Yuewu Xu for his helpful comments during my research.

Finally, I express my deepest gratitude to my parents and brother who have encouraged and supported me throughout my student life.

Minimax Coding and Prediction

Qun Xie

Yale University

1997

ABSTRACT

A technique for statistical prediction and coding is developed using asymptotic minimax criteria under some probabilistic and non-probabilistic assumptions. The motivation is to identify the asymptotic minimax distance between a parametric family of discrete distributions and arbitrary distributions, provide implementable algorithms incurring a minimum loss, and apply the results in prediction, coding and related areas. Relationships between coding and prediction are explored.

Target levels of loss are based on the best performance achieved by competitors using a parametric family of distributions. For each sequence $x_1, \ldots, x_n$, there exists a best competitor in that family who suffers the lowest cumulative loss. To achieve this ideal performance level, in principle one would need the hindsight of an empirically optimal parameter value. Our prediction algorithm provides a distribution of x_{k+1} based on the previous observations $x_0, \ldots, x_k$, for $k = 1, \ldots, n$. The aim of our strategy is to achieve without hindsight almost as good a performance as the ideal target level.

It is discovered that Jeffreys' prior plays a major role in determining the asymptotic minimax regret, deriving online prediction procedures and providing asymptotically minimax coding strategies. We study the limiting behavior of procedures based on the Jeffreys' prior, particularly when the parameters or relative frequencies are on or around the boundaries. We manage to modify this prior to generate a sequence of asymptotic minimax strategies useful for prediction and coding. We also show that surprisingly the very same algorithm based on the modifications of Jeffreys' prior work in both the expected regret and worst-case regret cases.

Our results find applications including probability density estimation, universal source coding, categorical data prediction with side information, gambling, and a comparison between frequentists and Bayesians in hypotheses testing.

Contents

Chapter 1

Introduction

1.1 Overview

Statistical inference concerns itself with data summarization and prediction. People propose various (mostly parametric) probability models to understand these random events. While estimation of parameters is of interest, we sometimes need estimation or prediction of the probability functions of the random variables as in contexts of coding and gambling that we shall describe.

Let $X_1, \ldots, X_n$ be a sequence of letters from a finite alphabet $\mathcal{X}$. We are interested in finding a probability mass function $q(x^n)$ such that it is useful for prediction and universal coding while suffering a minimum loss. We approach this problem under two assumptions, and each approach has its own interpretations.

First we assume that these $(X_1, ..., X_n)$ follow some distribution with probability mass function $p(x_1, ..., x_n|\theta)$, where $\theta \in \Theta \subset R^d$. For example, we could assume that given θ, the X_i's are independent and identically distributed. We desire to code such data with nearly minimal expected codelength, when we have no information about the generating parameter θ other than it belongs to the set Θ. This is universal coding, first systematically treated by Davisson [13].

It is known that the expected codelength is lower bounded by the entropy of the distribution. When the true θ is known, this bound can be achieved within one bit. When θ is unknown, and if we use a (sub) probability mass function q_n on $\mathcal{X}^n$ and $-\log q_n(x^n)$ bits to code data string x^n, then it induces a redundancy in the expected length of $D(p_\theta^n||q_n)$, where p_θ^n is the joint distribution of $X^n = (X_1, X_2, ..., X_n)$, 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 θ from Θ and a statistician chooses a distribution q_n on $\mathcal{X}^n$ as his best guess of p_θ^n. The loss is measured by the total relative entropy $D(p_\theta^n||q_n)$. Then for finite n and prior $W(d\theta)$ on Θ the best strategy q_n to minimize the average risk $\int D(p_\theta^n||q_n)W(d\theta)$ is the mixture density $m_n^W(x^n) = \int p_\theta^n(x^n)W(d\theta)$ (called the Bayes procedure), and the resulting average risk

is the Shannon mutual information $I(\Theta; X^n)$ (see [13], [11]). Suppose Θ is compact and that $p_\theta(x)$ depends continuously on $\theta \in \Theta$ for every $x \in \mathcal{X}$. Then the minimax value $\min_{q_n} \max_{\theta \in \Theta} D(p_\theta^n || q_n)$ is equal to the maximin value $\max_W \int D(p_\theta^n || m_n^W) W(d\theta)$, 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 [14] using [20], and is attributed there to Gallager [22]; see [24] 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 $m_n(x^n) = \int p_\theta^n(x^n) W_n^*(d\theta)$ is both maximin and minimax (see the discussion following Lemma 5 in the appendix). An interesting property of this least favorable prior is that it is usually discrete [45]. The problem of choosing a prior to maximize $I(\Theta; X^n)$ arises in Bayesian statistics as the reference prior method (Bernardo [5]).

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 $\sum_{n'=0}^{n-1} D(p_\theta || \widehat{p}_{n'})$, where the probability function p_θ is estimated using a sequence $\widehat{p}_{n'}$ based on $X^{n'}$ for $n' = 0, \ldots, n-1$ (see [11], [12]). Consequences of this prediction interpretation are developed in [25], [27] and [3].

In this dissertation we study the behavior of the minimax redundancy $\min_{q_n} \max_{\theta \in \Theta} D(p_\theta^n || q_n)$ as $n \to \infty$. In the case that $\{p_\theta : \theta \in \Theta\}$ is the whole simplex of probabilities

on the finite alphabet $\mathcal{X}$. We determine the exact constant in the asymptotic value, and we identify asymptotically maximin and minimax procedures. We find that Jeffreys' prior plays an important role in this characterization.

The second approach to the problem is to consider the loss at each individual sequence $x_1, ..., x_n$. No probability distribution is assumed to govern the sequence. Nevertheless, probability mass functions arise operationally in the choice of data compression, gambling, or prediction strategies. Instead of a stochastic analysis of performance, our focus is the worst-case behavior of the difference between the loss incurred and a target level of loss.

We are to choose a probability mass function $q(x_1, ..., x_n)$ on $\mathcal{X}^n$ such that its conditionals $q(x_i|x_1, ..., x_{i-1})$ provide a strategy for coding, gambling and prediction of a sequence x_i, $i = 1, 2, ..., n$. We desire large values of $q(x_1, ..., x_n)$ or equivalently small values of $\log 1/q(x_1, ..., x_n) = \sum_{i=1}^{n} \log 1/q(x_i|x_1, ..., x_{i-1})$ relative to the value achieved by a target family of strategies. Specifically let $\{p(x_1, ..., x_n|\theta), \theta \in \Theta\}$ be a family of probability mass functions on $\mathcal{X}^n$. One may think of θ as indexing a family of players that achieve value $\log 1/p(x_1, ..., x_n|\theta)$ for a sequence $x_1, ..., x_n$. With hindsight the best of these values is $\log 1/p(x_1, ..., x_n|\hat{\theta})$ where $\hat{\theta} = \hat{\theta}(x_1, ..., x_n)$ achieves the maximum of $p(x_1, ..., x_n|\theta)$. The game-theoretic problem is this: choose q to minimize the maximum regret

$$\max_{x_1, ..., x_n} \left(\log 1/q(x_1, ..., x_n) - \log 1/p(x_1, ..., x_n|\hat{\theta}) \right),$$

evaluate the minimax value of the regret, identify the minimax and maximin solutions, and determine computationally feasible approximate solutions. Build-

ing on past work by Shtarkov [36] and others, in this dissertation we accomplish these goals in an asymptotic framework including exact constants, in the case of the target family of all memoryless probability mass functions on a finite alphabet of size m.

The asymptotic minimax value takes the form $\frac{m-1}{2}\log\frac{n}{2\pi}+C_m+o(1)$, where the constant C_m is identified. The choice of $q(x_1,....,x_n)$ that is a mixture with respect to Jeffreys' prior (the Dirichlet(1/2,...,1/2) in this case) is shown to be asymptotically maximin. A modification in which lower-dimensional Dirichlet components are added near the faces of the probability simplex is shown to be asymptotically minimax. We also study other forms of modifications. All these strategies are relatively easy to implement using variants of Laplace's rule of succession. Moreover, these asymptotically optimal strategies are also asymptotically optimal for the corresponding expectation version of the problem.

The above game has interpretations in data compression, gambling and prediction as we discuss in Chapter 4. The choice of $q(x_1,....,x_n)$ determines the codelength $l(x_1,....,x_n)=\log_2(1/q(x_1,....,x_n))$ (rounded up to an integer) of a uniquely decodable binary code; it results in a cumulative wealth $S_n(x_1,....,x_n)=q(x_1,....,x_n)O(x_1,...,x_n)$ after sequentially gambling according to proportions $q(x_{k+1}|x_1,....,x_k)$ on outcome x_{k+1} with odds $O(x_{k+1}|x_1,....,x_k)$ for $k=0,....,n-1$; and for prediction a strategy based on $q(x_1,....,x_n)$ incurs a cumulative logarithmic loss $\log((1/q(x_1,....,x_n))=\sum_{k=0}^{n-1}\log 1/q(x_{k+1}|x_1,....,x_k)$. Likewise for each $p(x_1,...,x_n|\theta)$ there is a corresponding codelength $\log_2 1/p(x_1,....,x_n|\theta)$.

wealth $p(x_1, ..., x_n|\theta)O(x_1, ..., x_n)$ and cumulative log loss $\sum_{i=0}^{n-1} \log(1/p(x_i|\theta))$. The target value corresponds to the maximum likelihood. The regret measures the difference in codelengths, the log wealth ratio and the difference in total prediction loss between $q(x_1, ... x_n)$ and the target level in the parametric family. This regret is

$$\log \frac{1}{q(x_1, ..., x_n)} - \log \frac{1}{p(x_1, ... x_n|\hat{\theta})}.$$

To differentiate the two measurement of difference in losses, we use *redundancy* for the relative entropy distance $D(p_\theta^n||q_n)$ (the expectation version), and *regret* for $\log\left(P(x^n|\widehat{\theta})/Q(x^n)\right)$, the logarithm of probability ratio between the best of the family $p(x^n|\hat{\theta})$ and our choice $q(x^n)$.

1.2 Layout of Thesis

As outlined in the Introduction section, we basically study two versions of asymptotic minimax distances between discrete probability distributions: the expectation version and the individual sequence version. The first version assumes probability distributions on the sequence $X_1, ... X_n$ while for the second version considers competing with the best from a family of distributions.

In Chapter 2 we study some minimax and maximin properties using this quantity. Then we give our theorem which identifies the asymptotic minimax redundancy. Moreover we show that Jeffreys' prior is asymptotically maximin but not asymptotically minimax. We also modify this prior so that the mixture

is both maximin and minimax. The proof of theorem is first carried out for alphabet size $k = 2$ case, which gives an intuitive picture in the proof. We then generalize the proof to arbitrary case. The results of Chapter 2 have appeared in our paper [44].

In Chapter 3 we study the regret $\log\left(P(x^n|\widehat{\theta})/Q(x^n)\right)$. Our competitors act according to a model in which $x_1, \ldots, x_n$ are independent with joint distribution of the form $p(x^n|\theta) = \prod p(x_i|\theta)$ for some $\theta \in \Theta$. We show that the same strategy identified in Chapter 2 also asymptotically minimizes the worst regret. Shtarkov identifies the unique minimax strategy of problem, and comments on the difficulty of its implementation to prediction. We modify Jeffreys' prior to generate a mixture which is asymptotically minimax and we also give the limiting behavior of this minimax regret. Moreover, this modified mixture is easy to calculate by simple recursive computation, thus may be used for prediction. We discovering that in essence the regret is the same for individual sequence as for the expected version of the problem. In this way the minimax regret solution of Chapter 3 strengthens the conclusions of Chapter 2.

In Chapter 4 we apply our result of Chapter 3 in data compression, gambling and prediction (with and without side information). In Chapter 5, we extend the iid case to the (first-order) Markov case. This setting is of more practical importance. Consider weather, for example, where the sequence $x_1, \ldots, x_n$ indicates rain or shine on consecutive days. You would not expect these outcomes to i.i.d. but rather to have some dependence which might well fit in a Markov

model. The parameters are the transition probabilities. Jeffreys' prior in this case is more complicated, however the Laplace integration method does work here for a certain interior set of sequences x^n. When relative frequencies based on x^n are near the boundary, we use some lemmas developed for the iid case and successfully solve the boundary problem in determining the asymptotic minimax regret.

Chapter 2

Minimax Redundancy for the Class of Memoryless Sources

2.1 Literature Review and Statement of Result

As we have outlined in Chapter 1, we assume a sequence of independent observations $X_1, \ldots, X_n$ from the same distribution $p(\cdot|\theta)$ for some θ. Lacking knowledge of this θ, we use q_n as a guess of the joint distribution of $x^n = (x_1, \ldots, x_n)$. We are interested in the Kullback-Leibler divergence between the "true" and our

guess joint distributions

$$D(p_\theta^n||q_n) = \int p(x^n|\theta) \log \frac{p(x^n|\theta)}{q_n(x^n)} d\theta.$$

where p_θ^n is the joint density of $X^n = (X_1, X_2, ..., X_n)$. In particular We are interested to know the behavior of the minimax redundancy $\min_{q_n} \max_{\theta \in \Theta} D(p_\theta^n||q_n)$ as $n \to \infty$.

Krichevsky and Trofimov [29] and Davisson et al. [15] 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 [32] shows that for any code, $(d/2)\log n - o(\log n)$ is an asymptotic lower bound on the redundancy for almost all θ in the family, and [31] 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 [11] determine the constant in the redundancy $(d/2)\log n + c_\theta + O(1)$ for codes based on mixtures. When regularity conditions are satisfied, including the finiteness of the determinant of Fisher information $I(\theta)$ and the restriction of θ to a compact subset C of the interior of Θ, Clarke and Barron [12] 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 $\log \int_C \sqrt{\det I(\theta)} d\theta$. However, their 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 chapter we take the underlying distribution p_θ to be any proba-

bility on a finite alphabet $\mathcal{X} = \{a_1, ..., a_k\}$. We assume that p_θ puts mass θ_i on letter $\{a_i\}$, for $i = 1, ..., k$. The parameter space Θ is the simplex $S_{k-1} = \left\{\theta = (\theta_1, ..., \theta_{k-1}) : \sum_{i=1}^{k-1} \theta_i \leq 1, \text{ all } \theta_i \geq 0\right\}$, or equivalently, $S'_k = \left\{\theta = (\theta_1, ..., \theta_k) : \sum_{i=1}^{k} \theta_i = 1, \text{ all } \theta_i \geq 0\right\}$, where $\theta_k = 1 - (\theta_1 + ... + \theta_{k-1})$. The Fisher information determinant is $1/(\theta_1 \cdot \theta_2 \cdot ... \cdot \theta_k)$, which is infinite when any θ_i equals 0. The Dirichlet$(\lambda_1, ..., \lambda_k)$ distribution has density proportional to $\theta_1^{\lambda_1 - 1} \cdot ... \cdot \theta_k^{\lambda_k - 1}$ on Θ for $\lambda_1, ..., \lambda_k$ 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$(1/2, ..., 1/2)$ density.

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

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

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 $\int D(p_\theta^n || m_n^{W_n}) W_n(d\theta) - ((k-1)/2) \log(n/(2\pi e))$ 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 $\max_\theta D(p_\theta^n || q_n) - ((k-1)/2) \log(n/(2\pi e))$ converges to V.

Our main result is the following.

Theorem 2.1. The asymptotic minimax and maximin redundancy satisfy

$$\lim_{n \to \infty} \left(\min_{q_n} \max_{\theta \in \Theta} D(p_\theta^n || q_n) - \frac{k-1}{2} \log \frac{n}{2\pi e} \right)$$

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

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 $\theta \in \Theta$, to $\log\int_\Theta \sqrt{\det I(\theta)}d\theta = \log(\Gamma(1/2)^k/\Gamma(k/2))$, as what we would expect from Clarke and Barron [12].

Remark 3. Previously the best upper and lower bounds on the asymptotic minimax value were based on the values achieved using the Dirichlet$(1/2, \ldots, 1/2)$ prior, see [29], [15] and more recently [37]. 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 rest of the chapter is as follows. Section 2.2 contains some notations and definitions, mostly for the Bernoulli family case ($k = 2$), and the proof for the this case is presented in Section 2.3. 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 2.4. In the Appendix of dissertation we include some propositions and lemmas used in the main analysis.

2.2 Notations and definitions

For the Bernoulli distribution $\{p_\theta(x) = \theta^x(1-\theta)^{1-x} : x \in \{0,1\}, \theta \in [0,1]\}$, 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 $X^n = (X_1, X_2, ..., X_n)$, 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 $m_n^*(x^n) = \int_0^1 p_\theta^n(x^n) w^*(\theta) d\theta = \pi^{-1} \int_0^1 \theta^{\Sigma x_i - 1/2}(1-\theta)^{n-\Sigma x_i - 1/2} d\theta$ 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 $\underline{V} = \varliminf_{n\to\infty} \underline{V}_n$ the asymptotic lower value.

Similarly the upper value (the minimax value) is

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

and the asymptotic upper value is $\overline{V} = \varlimsup_{n\to\infty} \overline{V}_n$. We remind the reader that $\overline{V}_n = \underline{V}_n$. 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.

2.3 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 1.1).

Lemma 2.1. 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 $|D(p_\theta^n||m_n^*) - (1/2)\log n/(2\pi e) - \log\pi| \leq 5/(n\min(\theta, 1-\theta))$. Similar inequalities with error $O(1/(n\delta))$ for $\delta \leq \theta \leq 1 - \delta$ have recently been obtained by Suzuki [37].

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 [12]) 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 2.1 of this Chapter 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 ei-

ther 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 can't identify this value yet.

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

2.3.1 *Lower value* $\underline{V} \geq \log \pi$

Proof. By definition, we need to show that

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

It suffices to prove that $\int_{c/n}^{1-c/n} D(p_\theta^n||m_n^*)w^*(\theta)d\theta - (1/2)\log(n/(2\pi e)) \geq \log \pi - o_n(1)$ for any $c > 0$, where $w^*(\theta) = \theta^{-1/2}(1-\theta)^{-1/2}/\pi$ is Jeffreys' prior on $[0, 1]$. In fact, from Lemma 2.1 of this Chapter, 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

$$\begin{aligned}\int_{c/n}^{1-c/n} \left(D(p_\theta^n||m_n^*) - \frac{1}{2}\log\frac{n}{2\pi e}\right) w^*(\theta)d\theta &\geq \int_{c/n}^{1-c/n} (\log \pi - \varepsilon)\, w^*(\theta)d\theta \\ &\geq (\log \pi - \varepsilon)\left(1 - \frac{4}{\pi}\sqrt{\frac{c}{n-c}}\right) \quad (2.1)\end{aligned}$$

where the last inequality is from

$$\int_0^{c/n} \theta^{-1/2}(1-\theta)^{-1/2}d\theta \leq (1-c/n)^{-1/2}\int_0^{c/n} \theta^{-1/2}d\theta = \left(1 - \frac{c}{n}\right)^{-1/2} 2\left(\frac{c}{n}\right)^{1/2}.$$

The same bound holds for the integral from $1 - c/n$ to 1. Therefore we have that the limit 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 2.3.3 below).

Remark 5. An alternative demonstration that $\underline{V} \geq \log \pi$ follows from the weaker result of [12]. In particular if we restrict $\theta \in [\delta, 1-\delta]$, then $D(p_\theta^n||m_{n,\delta}^*) - (1/2)\log n/(2\pi e) \to \int_\delta^{1-\delta} \theta^{-1/2}(1-\theta)^{-1/2}d\theta$ 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 [12] so the proof in the present thesis is self-contained.

2.3.2 *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 2.1 of this Chapter, 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

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

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

$$\begin{aligned}\overline{V}_n &\leq \max_{\theta\in[0,1]} D(p_\theta^n||m_n^\varepsilon) - \frac{1}{2}\log\frac{n}{2\pi e} \\ &= \max\left\{\max_{[0,\frac{c}{n}]} D(p_\theta^n||m_n^\varepsilon), \max_{[\frac{c}{n},1-\frac{c}{n}]} D(p_\theta^n||m_n^\varepsilon), \max_{[1-\frac{c}{n},1]} D(p_\theta^n||m_n^\varepsilon)\right\} - \frac{1}{2}\log\frac{n}{2\pi e}.\end{aligned} \tag{2.2}$$

We next show that for large n, an upperbound M_n for the supremum over $[c/n, 1-c/n]$ also upperbounds 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 $\theta \in [0, c/n]$,

$$\begin{aligned}D(p_\theta^n||m_n^\varepsilon) &= E_\theta \log \frac{p_\theta^n(X^n)}{\eta p_{c/n}^n(X^n) + \eta p_{1-c/n}^n(X^n) + (1-2\eta)\int_0^1 p_s^n(X^n)W^*(ds)} \\ &\leq E_\theta \log \frac{p_\theta^n(X^n)}{\eta p_{c/n}^n(X^n)} \\ &= \log\frac{1}{\eta} + nD(p_\theta||p_{c/n}) \\ &\leq \log\frac{1}{\eta} + nD(p_0||p_{c/n})\end{aligned} \tag{2.3}$$

$$= \log\frac{1}{\eta} + n\log\frac{1}{1-c/n}$$
$$\leq \log\frac{1}{\eta} + 2c, \tag{2.4}$$

where inequality (2.3) 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 2.1 of this Chapter,

$$\begin{aligned} D(p_\theta^n||m_n^\varepsilon) &\leq E_\theta \log \frac{p_\theta^n(X^n)}{(1-2\eta)\int_0^1 p_s(X^n)w^*(s)ds} \\ &= \log\frac{1}{1-2\eta} + D(p_\theta^n||m_n^*) \\ &\leq \log\frac{1}{1-2\eta} + \log\pi + \frac{1}{2}\log\frac{n}{2\pi c} + \varepsilon. \end{aligned} \tag{2.5}$$

for all $n \geq 2c$.

Now it's seen that (2.5) eventually will exceed (2.4) when n increases, as we intended to show. From (2.2), $\overline{V}_n \leq \log 1/(1-2\eta) + \log\pi + \varepsilon$, for all large n and hence $\overline{V} \leq \log(1/(1-2\eta)) + \log\pi + \varepsilon$. Therefore upon taking the infimum over $0 < \eta < 1/2$ and $\varepsilon > 0$, we obtain that $\overline{V} \leq \log\pi$.

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 [12], 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, $\eta_n \geq (2c/(n\pi))^{1/4}$ and $\varepsilon_n \geq 10/\log(n\pi/(2c))$.

2.3.3 *Jeffreys' prior is asymptotically least favorable*

Since $\underline{V} = \log \pi$, to prove that Jeffreys' prior w^* is asymptotically least favorable, we need $\underline{\lim}_n \left[\int_0^1 D(p_\theta^n||m_n^*)w^*(\theta)d\theta - (1/2)\log(n/(2\pi e))\right] \geq \log \pi$, which is already shown in Section 2.3.1. Moreover, a choice of $\varepsilon_n = 1/\sqrt{n}$ in Lemma 2.1 of this chapter 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 1.4 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 $1/\sqrt{n}$.

2.3.4 *Jeffreys' prior is not asymptotically minimax*

To see that Jeffreys' prior is not asymptotically minimax we use the fact, recently studied in Suzuki [37], 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

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

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 [37], we have

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

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 $\theta = 1$. It's the same scenario on the other boundary point $\theta = 0$. This completes the proof of the theorem.

Remark 6. Davisson et al. [15, 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 [37, Thm.3] points out that this bound is achieved at the end point 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 We have also 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 Chapter 3 which addresses minimal worst case redundancy over all sequences x^n.

2.4 Extension to $k \geq 3$ cases

For the case of an alphabet of size k we recall from Section 3.1 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, ... 1/2) density $w^*(\theta) = \theta_1^{-1/2} \cdot ... \cdot \theta_k^{-1/2} / D_k(1/2, ..., 1/2)$. Here $D_k(\lambda_1, ..., \lambda_k) = \int_\Theta \theta_1^{\lambda_1 - 1} \cdot ... \cdot \theta_k^{\lambda_k - 1} d\theta_1 ... d\theta_{k-1}$ is the Dirichlet integral. In terms of Gamma functions the Dirichlet function may be expressed as

$$D_k(\lambda_1, ..., \lambda_k) = \frac{\Gamma(\lambda_1) \cdot ... \cdot \Gamma(\lambda_k)}{\Gamma(\sum_{i=1}^k \lambda_i)}. \tag{2.6}$$

It follows that $\int_\Theta \sqrt{\det(I(\theta))} d\theta = D_k(1/2, ..., 1/2) = \Gamma(1/2)^k / \Gamma(k/2)$. We will first show that $\underline{V}(k) \geq \log(\Gamma(1/2)^k / \Gamma(k/2))$ using Jeffreys' prior in Part 1. then $\overline{V}(k) \leq \log(\Gamma(1/2)^k / \Gamma(k/2))$ 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 $D(p_\theta^n || m_n^*)$ at the boundary of Θ is demonstrated in Part 4.

Part 1. *Asymptotic lower value* $\underline{V}(k) \geq \log(\Gamma(1/2)^k / \Gamma(k/2))$.

This is parallel to part (A) of the $k = 2$ case, except that θ is replaced by $\boldsymbol{\theta}$, Lemma 2.1 of this chapter is replaced by Proposition 1.1 of the appendix, and inequality (2.1) is replaced by the following argument. With the Dirichlet$(1/2, \ldots, 1/2)$ prior the marginal distribution of θ_i is Beta$(1/2, (k-1)/2)$, thus the contribution of $\{\theta_i \leq c/n\}$ to the integral of $w^*(\boldsymbol{\theta})$ is bounded by

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

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* $\overline{V}(k) \leq \log(\Gamma(1/2)^k/\Gamma(k/2))$.

Proof. For any $\varepsilon > 0$, let L_i be the intersection of $\{\boldsymbol{\theta} : \theta_i = c/n\}$ with the probability simplex Θ, for $i = 1, \ldots, k$, where $c = c(\varepsilon)$ is chosen as in Proposition 1.1 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}),

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

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

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

For this prior, the Bayes procedure to minimize $\int D(p_\theta^n||q_n)W_n^\varepsilon(d\theta)$ uses

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

where $T_i(X^n) = \sum_{j=1}^n 1_{\{X_j=\{a_i\}\}}$ and

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

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

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

$$\sup_{\theta\in\Theta} D(p_\theta^n||q_n) = \max\left\{\sup_{R_1} D(p_\theta^n||q_n),\ldots\sup_{R_k} D(p_\theta^n||q_n),\sup_R D(p_\theta^n||q_n)\right\} \quad (2.7)$$

We will find an upperbound for $\sup_{\theta\in\Theta} D(p_\theta^n||q_n)$ by showing that it upperbounds all the supremums over $R_1,\ldots R_k, R$.

For $\theta\in R$, we have

$$\begin{aligned} D(p_\theta^n||q_n) &\le E_\theta \log\frac{\theta_1^{T_1}\cdots\theta_k^{T_k}}{(1-\varepsilon)m_n^*(X^n)} \\ &= \log\frac{1}{1-\varepsilon} + D(p_\theta^n||m_n^*) \end{aligned}$$

$$\leq \log\frac{1}{1-\varepsilon} + \log\frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2})} + \frac{k-1}{2}\log\frac{n}{2\pi e} + \varepsilon\log e. \quad (2.8)$$

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

For $\theta \in R_i$, say $i = 1$, that is, $0 \leq \theta_1 \leq c/n$.

$$\begin{aligned} D(p_\theta^n||q_n) &\leq E_\theta \log \frac{p_\theta^n(X^n)}{\frac{\varepsilon}{k}m_1(x^n)} \\ &= \log\frac{k}{\varepsilon} + n\theta_1\log\theta_1 + \sum_{j=2}^{k} n\theta_j\log\theta_j + \log D_{k-1}(\frac{1}{2},\ldots,\frac{1}{2}) - \\ &\quad - E_\theta \log D_{k-1}(T_2+\frac{1}{2},\ldots,T_k+\frac{1}{2}) \end{aligned} \quad (2.9)$$

We now construct a set of multinomial variables $(T_2',\ldots,T_k')$ with parameters $(n,\theta_2/(1-\theta_1),\ldots,\theta_k/(1-\theta_1))$ from $(T_1,\ldots,T_k) \sim$Multinomial$(n,\theta_1,\ldots,\theta_k)$, by randomly reassigning the T_1 occurrences of the outcome $\{a_1\}$ to $\{a_2\},\ldots,\{a_k\}$ with probabilities $\theta' = \theta_2/(1-\theta_1),\ldots,\theta_k/(1-\theta_1)$, respectively. That is, given T_1, we obtain new counts $T_j' = T_j + \xi_j$ for $j = 2,\ldots,k$, where $(\xi_2,\ldots,\xi_k) \sim$ Multinomial(T_1,θ'). Hence $(T_2',\ldots,T_k') \sim$ Multinomial(n,θ'), 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

$$E_\theta \log D_{k-1}(T_2+\frac{1}{2},\ldots,T_k+\frac{1}{2}) \geq E_{\theta'}\log D_{k-1}(T_2'+\frac{1}{2},\ldots,T_k'+\frac{1}{2}). \quad (2.10)$$

Also observe that

$$\begin{aligned} \sum_{j=2}^{k} n\theta_j\log\theta_j &= \left[\sum_{j=2}^{k}\left(n\frac{\theta_j}{1-\theta_1}\log\frac{\theta_j}{1-\theta_1} + n\frac{\theta_j}{1-\theta_1}\log(1-\theta_1)\right)\right](1-\theta_1) \\ &\leq \sum_{j=2}^{k} n\frac{\theta_j}{1-\theta_1}\log\frac{\theta_j}{1-\theta_1}. \end{aligned} \quad (2.11)$$

Applying (2.10) and (2.11) to (2.9), we obtain

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

where m_n^{**} is the procedure based on Jeffreys' prior on the reduced $(k-2)$-dimensional probability simplex S'_{k-1} and $\theta' \in S'_{k-1}$. Now a course upper bound on $D(p_{\theta'}^{n}||m_{n}^{**})$ is sufficient for this lower dimensional piece. Lemma 1.4 gives

$$D(p_{\theta'}^{n}||m_{n}^{**}) \leq \frac{k-2}{2}\log\frac{n}{2\pi e} + C_{k-1}. \tag{2.12}$$

for all $\theta' \in \Theta$ and some constant C_{k-1}. Observe that $(k-2)/2$ in (2.12) provides a smaller multiplier of the $\log n$ factor than achieved in the middle region R (see term (2.8)). Consequently, for all large n,

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

uniformly in $\theta \in \Theta$. 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 following appendix, for any θ with $\theta_i > 0$ for $i = 1, \ldots, k$, we have

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

For a vertex point such as $\mathbf{e} = (1, 0, \ldots, 0)$, as shown by Suzuki [37],

$$\begin{aligned} D(p_\theta^n || m_n^*) &= \log \frac{\Gamma(\frac{1}{2})^k / \Gamma(\frac{k}{2})}{\int \theta_1^{n-1/2} \theta_2^{-1/2} \cdots \theta_k^{-1/2} d\theta_1 \ldots d\theta_{k-1}} \\ &= \log \frac{\Gamma(n + \frac{k}{2})}{\Gamma(n + \frac{1}{2}) \Gamma(\frac{1}{2}) \ldots \Gamma(\frac{1}{2})} + \log \frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2})} \\ &\approx \left(\frac{k-1}{2} \log \frac{n}{2\pi e} + \log \frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2})} \right) + \frac{k-1}{2} \log 2e. \end{aligned} \tag{2.13}$$

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 $\theta = (\theta_1, \ldots, \theta_L, 0, \ldots, 0)$, where $1 \le L \le k-1$ and $\theta_j > 0$ for $j = 1, \ldots, L$, we have

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

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

$$+\log \frac{D_k(\frac{1}{2},\ldots,\frac{1}{2})\Gamma(n+\frac{k}{2})}{D_L(\frac{1}{2},\ldots,\frac{1}{2})\Gamma(n+\frac{L}{2})\Gamma(\frac{1}{2})^{k-L}} \tag{2.14}$$

$$= D(p^n_{\theta^L}||m^{**}_n) + \log \frac{\Gamma(n+\frac{k}{2})\Gamma(\frac{L}{2})}{\Gamma(n+\frac{L}{2})\Gamma(\frac{k}{2})}. \tag{2.15}$$

where $\theta^L = (\theta_1,\ldots,\theta_L)$ and m^{**}_n is the mixture density with Jeffreys' prior on the L-dimensional simplex. Stirling's formula yields the following approximation

$$\log \frac{\Gamma(n+\frac{k}{2})}{\Gamma(n+\frac{L}{2})} = \frac{k-L}{2}\log n + o(1). \tag{2.16}$$

From (2.15) and (2.16), and expanding $D(p^n_{\theta^L}||m^{**}_n)$ using Proposition 1 of the appendix, we have

$$D(p^n_{\theta^L}||m^{**}_n) = \left(\frac{L-1}{2}\log\frac{n}{2\pi e} + \log\frac{\Gamma(\frac{1}{2})^L}{\Gamma(\frac{L}{2})}\right) + \left(\frac{k-L}{2}\log n + \log\frac{\Gamma(\frac{L}{2})}{\Gamma(\frac{k}{2})}\right) + o(1)$$

$$= \left(\frac{k-1}{2}\log\frac{n}{2\pi e} + \log\frac{(\Gamma(\frac{1}{2}))^k}{\Gamma(\frac{k}{2})}\right) + \frac{k-L}{2}\log(2e) + o(1). \tag{2.17}$$

Comparing (2.17) with (2.13), we see that the asymptotic redundancy at a θ 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 $L = k$ non-zero 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. [15, inequality (61)] and Suzuki [37, Thm.3] proves that for each n, the value of $D(p^n_\theta||m^*_n)$ is maximized at the vertices. Here we have determined the asymptotic gap between vertex, face and interior points.

Chapter 3

Asymptotic Minimax Regret for Individual Sequences

3.1 Introduction and main result

We are interested in problems of data compression, gambling, and prediction of arbitrary sequences $x_1, x_2, ..., x_n$ of symbols from a finite alphabet $\mathcal{X}$. No probability distribution is assumed to govern the sequence. Nevertheless, probability mass functions arise operationally in the choice of data compression, gambling, or prediction strategies. Instead of a stochastic analysis of performance, our

focus is the worst-case behavior of the difference between the loss incurred and a target level of loss.

The following game-theoretic problem arises in the applications we discuss. We are to choose a probability mass function $q(x_1, \ldots, x_n)$ on $\mathcal{X}^n$ such that its conditionals $q(x_i|x_1, \ldots, x_{i-1})$ provide a strategy for coding, gambling and prediction of a sequence x_i, $i = 1, 2, \ldots, n$. We desire large values of $q(x_1, \ldots, x_n)$ or equivalently small values of $\log 1/q(x_1, \ldots, x_n) = \sum_{i=1}^{n} \log 1/q(x_i|x_1, \ldots, x_{i-1})$ relative to the value achieved by a target family of strategies. Specifically let $\{p(x_1, \ldots, x_n|\theta), \theta \in \Theta\}$ be a family of probability mass functions on $\mathcal{X}^n$. One may think of θ as indexing a family of players that achieve value $\log 1/p(x_1, \ldots, x_n|\theta)$ for a sequence $x_1, \ldots, x_n$. With hindsight the best of these values is $\log 1/p(x_1, \ldots, x_n|\hat{\theta})$ where $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ achieves the maximum of $p(x_1, \ldots, x_n|\theta)$. The game-theoretic problem is this: choose q to minimize the maximum regret

$$\max_{x_1, \ldots, x_n} \left(\log 1/q(x_1, \ldots, x_n) - \log 1/p(x_1, \ldots, x_n|\hat{\theta}) \right)$$

, evaluate the minimax value of the regret, identify the minimax and maximin solutions, and determine computationally feasible approximate solutions. Building on past work by Shtarkov [36] and others, we accomplish these goals in an asymptotic framework including exact constants, in the case of the target family of all memoryless probability mass functions on a finite alphabet of size m.

The asymptotic minimax value takes the form $\frac{m-1}{2} \log \frac{n}{2\pi} + C_m + o(1)$, where the constant C_m is identified. The choice of $q(x_1, \ldots, x_n)$ that is a mixture with respect to Jeffreys' prior (the Dirichlet(1/2,...,1/2) in this case) is shown to be

asymptotically maximin. A modification in which lower-dimensional Dirichlet components are added near the faces of the probability simplex is shown to be asymptotically minimax. We also study other forms of modifications. All these strategies are relatively easy to implement using variants of Laplace's rule of succession. Moreover, these asymptotically optimal strategies are the same as the strategies shown in Xie and Barron [44] to be asymptotically optimal for the corresponding expectation version of the problem.

Recent literature has examined the regret for individual sequences in the context of coding, prediction and gambling, in some cases building on past work on expected regret. Shtarkov [36] introduced and studied the minimax regret problem for universal data compression and gave asymptotic bounds of the form $(d/2)\log n + O(1)$ for discrete memoryless and Markov sources where d is the number of parameters. Extensions of that work to tree sources is in Willems, Shtarkov and Tjalkens [43], see also [40] and [41]. Rissanen [34] related the stochastic complex criterion for model selection to Shtarkov's regret and showed that the minimax regret takes the form $\frac{d}{2}\log n$ plus a constant he identified under certain conditions (and shows that it is related to the constant that arises in the expectation version in [12]). Feder, Merhav and Guttman [16], Haussler and Barron [25], Foster [18], Haussler, Kivinen and Warmuth [26], Vovk [39] and Freund [19] studied prediction problems for individual sequences. Cover and Ordentlich ([8], [30]) presented a sequential investment algorithm and related it to universal data compression.

Other related work considers expected regret. Davisson [13] systematically studied universal noiseless coding and the problem of minimax expected regret (redundancy). Davisson, McEliece, Pursely and Wallace [15] and Krichevsky and Trofimov [29] identified the minimax redundancy to the first order. Other work giving bounds on expected redundancy includes Davisson and Leon-Garcia [14], Rissanen [31][32], Clarke and Barron [11][12], Suzuki [37] and Haussler and Opper [27].

The minimax expected regret with smooth target families is of order $\frac{d}{2}\log n + C + o(1)$. The constant C and asymptotically minimax and maximin strategies are identified in Clarke and Barron [12] (for the minimax value over any compact region internal to the parameter space) and in Chapter 2 of this thesis published in [44] (for the minimax value over the whole finite alphabet probability simplex).

In the present chapter we show that the same strategy identified in Chapter 2 also asymptotically minimizes the worst case regret.

Before specializing to a particular target family we state some general definitions and results. We occasionally abbreviate $(x_1, \ldots, x_n)$ to x^n and omit the subscript n from probability functions p_n and q_n. Let the regret for using strategy $q_n(x^n)$ be defined by

$$r_n(q_n, x_1, \ldots, x_n) = \log \frac{p(x_1, \ldots, x_n|\hat{\theta})}{q_n(x_1, \ldots, x_n)}.$$

The minimax regret is

$$\overline{r}_n = \min_{q_n} \max_{x^n} r_n(q_n, x_1, \ldots, x_n).$$

A strategy q_n is said to be minimax if

$$\max_{x_1,\ldots,x_n} r_n(q_n, x_1, \ldots, x_n) = \overline{r}_n.$$

and it is said to be an equalizer (constant regret) strategy if $r_n(q_n, x_1, \ldots, x_n) = \overline{r}_n$ for all $x_1, \ldots, x_n \in \mathcal{X}^n$. The maximin value of the regret is defined to be $\underline{r}_n = \max_{p_n} \min_{q_n} \sum_{x^n} p_n(x^n) r_n(q_n, x_1, \ldots, x_n)$, where the maximum is over all distributions on $\mathcal{X}^n$. A strategy q_n is average case optimal with respect to a distribution p_n if it minimizes $\sum_{x^n} p_n(x^n) r_n(q_n, x^n)$ over choices of q_n. It is known from Shannon that the unique average case optimal strategy is $q_n(x^n) = p_n(x^n)$. A choice $q_n = p_n^*$ is said to be a maximin (or least favorable) strategy if $\sum_{x^n} r(p_n^*, x^n) p_n^*(x^n) = \underline{r}_n$. The following is basically due to Shtarkov [36] in the coding context.

Theorem 3.0 *Let $c_n = \sum_{x^n} p(x^n|\hat{\theta})$ where $\hat{\theta} = \hat{\theta}(x^n)$ is the maximum likelihood estimator. The minimax regret equals the maximin regret and equals*

$$\overline{r}_n = \underline{r}_n = \log c_n.$$

Moreover, $q_n^*(x^n) = p(x^n|\hat{\theta})/c_n$ is the unique minimax strategy, it is an equalizer rule achieving regret $\log p(x^n|\hat{\theta})/q_n^*(x^n) = \log c_n$ for all x^n, and it is the unique least favorable (maximin) distribution. The average regret for any other $p_n(x^n)$ equals $\sum_{x^n} p_n(x^n) \log(p(x^n|\hat{\theta})/p_n(x^n)) = \log c_n - D(p_n||q_n^*)$.

Proof. Note that $\sum_{x^n} q_n^*(x^n) = 1$ and that $r_n(q_n^*, x^n) = \log c_n$. thus q_n^* is an equalizer rule. For any other $q(x^n)$ with $\sum_{x^n} q(x^n) = 1$. we must have $q(x^n) < q_n^*(x^n)$ for some x^n and hence $r_n(q_n, x^n) > r_n(q_n^*, x^n) = \log c_n$ for that x^n. Thus q_n^* is minimax and $\bar{r}_n = \log c_n$. Now the last statement in the theorem holds by the definition of relative entropy and hence the maximin value $\underline{r}_n = \max_{p_n} \sum r(p_n, x^n) p_n(x^n) = \max_{p_n} \sum p_n(x^n) \log \frac{p(x^n|\hat{\theta})}{p_n(x^n)} = \max_{p_n} (\log c_n - D(p_n||q_n^*))$ where $D(p_n||q_n^*)$ is the relative entropy (Kullback-Leibler divergence). It is uniquely optimized at $p_n = q_n^*$, and therefore $\underline{r}_n = \log c_n$. ■

Thus the normalized maximized likelihood $q_n^*(x^n) = p(x^n|\hat{\theta})/c_n$ is minimax. However, it is not easily implementable for online prediction or gambling which requires the conditionals, nor for arithmetic coding which also requires the marginals for $x_1, \ldots, x_k$, $k = 1, \ldots, n$. The marginals obtained by summing out $x_{k+1}, \ldots, x_n$ is not the same as $p(x^t|\hat{\theta}(x^t))/c_t$. See Shtarkov [36] for his comment on the difficulty of implementing q_n^* in the universal coding context. It is natural to inquire whether there is an asymptotically maximin strategy $q_n(x^n) = \int p(x_1, \ldots, x_n|\theta) W(d\theta)$ for some fixed prior W distribution.

The choice of Jeffreys' prior density $w(\theta) \propto |I(\theta)|^{1/2}$ is asymptotically maximin for expected regret and slight modifications of it are asymptotically minimax as shown in a general setting in [12] (but with a restriction that the minimax value is taken over a compact set interior to the parameter space). For probabilities taken over the whole simplex, a modification of Jeffreys' prior is identified in [44] that is asymptotically maximin and minimax in the expected

regret setting. It would be convenient and natural for the same strategies to be maximin and minimax in the present setting.

Here we focus on the case that $p(x_1, \ldots, x_n|\theta) = \prod_{k=1}^n p(x_k|\theta)$ where $p(x = i|\theta) = \theta_i$, $i = 1, 2, \ldots, m$, is the model of conditionally independent outcomes from $\theta = (\theta_1, \ldots, \theta_m)$ on the probability simplex $S_m = \{(\theta_1, \ldots, \theta_m) : \theta_i \geq 0$ and $\sum_{i=1}^m \theta_i = 1\}$. The alphabet is taken to be $\mathcal{X} = \{1, 2, \ldots, m\}$. Jeffreys' prior in this case is the Dirichlet$(1/2, \ldots, 1/2)$ distribution. Previously Shtarkov [36] showed hat the mixture with this prior achieves maximal regret that differs from the minimax regret asymptotically by not more than a constant.

We say that a procedure $q_n(x^n)$ is asymptotically minimax if $\max_{x_1, \ldots, x_n} r_n(q_n, x_1, \ldots, x_n) = \overline{r}_n + o(1)$. It is an asymptotically constant regret strategy if $r_n(q_n, x_1, \ldots, x_n) = \overline{r}_n + o(1)$ for all x^n. A sequence $p_n(x^n)$ is asymptotically maximin if $\min_{q_n} \sum p_n(x^n) r_n(q_n, x_1, \ldots, x_n) = \underline{r}_n + o(1)$. We denote the minimax = maximin value by $r_n = \overline{r}_n = \underline{r}_n = \log c_n$.

Theorem 3.1 *The minimax regret satisfies*

$$r_n = \frac{d}{2} \log \frac{n}{2\pi} + C_m + o(1)$$

where $d = m - 1$ *and* $C_m = \log((\Gamma(1/2))^m / \Gamma(m/2))$. *The choice* $q(x^n) = m_J(x^n) = \int p(x^n|\theta) w_J(\theta) d\theta$ *with* $w_J(\theta)$ *being the Dirichlet*$_m(1/2, \ldots, 1/2)$ *prior (Jeffreys' prior in the present context) is asymptotically maximin. It has asymptotically constant regret for sequences with relative frequency composition internal to the simplex. But it is not asymptotically minimax. The maximum regret on the boundary of the simplex is* $r_n + \frac{d}{2} \log 2 + o(1)$, *which is higher than the asymptotic minimax value. Finally we give a modification of the*

Dirichlet$(1/2, \ldots, 1/2)$ *prior that provides strategies of the form* $\widetilde{q}_n(x^n) = \int p(x^n|\theta)\widetilde{W}_n(d\theta)$ *that are both asymptotically minimax and maximin, where* $\widetilde{W}_n = (1-\varepsilon_n)W_J + \varepsilon_n V$ *is a mixture of Jeffreys' prior* W_J *on* $(\theta_1, \ldots, \theta_m)$ *and a small contribution from a prior* $V = \frac{1}{m}\sum_{i=1}^{m} J_i$ *with* J_i *on the lower dimension spaces* $\{(\theta_1, \ldots, \theta_{i-1}, 1/n, \theta_{i+1}, \ldots, \theta_m) : \sum_{i' \neq i} \theta_{i'} = 1 - 1/n\}$, *where* $J_i = J_{i,n}$ *makes* $(\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_m)/(1-\frac{1}{n})$ *have the Dirichlet*$_{m-1}(1/2, \ldots, 1/2)$ *distribution and* $\theta_i = 1/n$. *Here* $\varepsilon_n = n^{-1/8}$.

Corollary *The Kullback-Leibler distance between Jeffreys' mixture and the normalized maximum likelihood probability function* $D(m_J||q_n^*)$ *converges to zero. Similarly,* $D(\widetilde{q}_n||q_n^*)$ *and* $D(m_J||\widetilde{q}_n)$ *converge to zero as* $n \to \infty$.

Remark 1 The above strategies $m_J(x^n)$ and $\widetilde{q}_n(x^n)$ based on Jeffreys' prior and its modification here shown to be asymptotically maximin and minimax for regret are the same as shown to be asymptotically maximin and minimax for the expected regret in Chapter 2. Other satisfactory modifications of Jeffreys' prior are given in Section 3.3.

Remark 2 By asymptotic minimaxity the difference between the worst case regret of the strategy and the asymptotic value $(d/2)\log(n/2\pi) + C_m$ converges to zero with n (i.e. this difference is $o(1)$). We do not seek here to determine the optimal rate at which this difference converges to zero. Nevertheless, some bounds for it are given in Section 3.3.

Remark 3 The joint probability $m_J(x^n) = \int p(x^n|\theta)w_J(\theta)d\theta$ can be expressed

directly in terms of Gamma functions as $m_J(x^n) = D_m(T_{1,n} + 1/2, \ldots, T_{m,n} + 1/2)$, where $T_{i,n} = T_i(x^n)$ is the number of occurrences of the symbol i in $(x_1, \ldots, x_n)$, for $i = 1, 2, \ldots, m$, and $D_m(\lambda_1, \ldots, \lambda_m) = \prod_{i=1}^m \Gamma(\lambda_i)/\Gamma(\sum_{i=1}^m \lambda_i)$ is the Dirichlet function. It can be more easily computed by the usual variant of Laplace's rule for conditionals. The conditionals $m_J(x_{i+1}|x_1, \ldots, x_i)$ are computed by

$$m_J(x_{i+1} = k + 1|x_1, \ldots, x_i) = \frac{T_{i,k} + \frac{1}{2}}{i + \frac{m}{2}}$$

where $T_{i,k}$ is the number of occurrences of the symbol i in the sequence $(x_1, \ldots, x_k)$, and then $m_J(x_1, \ldots, x_n) = \prod_{k=0}^n m_J(x_{k+1}|x_1, \ldots, x_k)$. Similarly the asymptotically minimax (and maximin) strategy uses

$$\tilde{q}_n(x^n) = (1 - \varepsilon_n) m_J(x^n) + \frac{\varepsilon_n}{m} \sum_{i=1}^m m_{i,n}(x^n)$$

where $m_J(x^n)$ is the Dirichlet mixture and $m_{i,n}(x^n) = \int p(x^n|\theta) J_{i,n}(d\theta)$ is Jeffreys' mixture with the prior component $J_{i,n}$ in which $\theta_i = 1/n$ is fixed. Here $m_{i,n}(x^n)$ can be expressed directly as

$$\frac{D_{m-1}(T_1 + \frac{1}{2}, \ldots, T_{i-1} + \frac{1}{2}, T_{i+1} + \frac{1}{2}, \ldots, T_m + \frac{1}{2})}{D_{m-1}(\frac{1}{2}, \ldots, \frac{1}{2})} \cdot \left(\frac{1}{n}\right)^{T_i} \left(1 - \frac{1}{n}\right)^{n - T_i}.$$

This strategy $\tilde{q}_n$ can be more easily computed by updating marginals according to

$$\tilde{q}_n(x^{k+1}) = \tilde{q}_n(x_{k+1}|x^k) \tilde{q}_n(x^k).$$

where the conditional probability is

$$\tilde{q}_n(x_{k+1}|x^k) = \frac{(1 - \varepsilon_n) m_J(x^{k+1}) + \varepsilon_n \frac{1}{m} \sum_{i=1}^m m_{i,n}(x^{k+1})}{(1 - \varepsilon_n) m_J(x^k) + \varepsilon_n \frac{1}{m} \sum_{i=1}^m m_{i,n}(x^k)}. \qquad (3.1)$$

and $m_J(x^k)$, $m_{i,n}(x^k)$ are updated according to

$$m_J(x^{k+1}) = m_J(x_{k+1}|x^k)m_J(x^k)$$

and

$$m_{i,n}(x^{k+1}) = m_{i,n}(x_{k+1}|x^k)m_{i,n}(x^k).$$

where

$$m_{i,n}(x_{k+1} = j|x_1, \ldots, x_k) = \begin{cases} \frac{T_{j,k}+1/2}{k-T_{i,k}+(m-1)/2}(1-\frac{1}{n}), & \text{for } j \neq i \\ \frac{1}{n}, & \text{for } j = i. \end{cases}$$

Therefore simple recursive computations suffice. The total computation time is not more than the order of nm^2. Note however that our strategy requires knowledge of the time horizon n when evaluating the conditionals for x_{k+1} given $x_1, \ldots, x_k$ for $k = 0, 1, \ldots, n-1$.

Remark 4 The answer $\frac{d}{2}\log\frac{n}{2\pi}+C_m$ is in agreement with the answer $\frac{d}{2}\log\frac{n}{2\pi}+\log\int_S\sqrt{|I(\theta)|}d\theta$ that we would expect to hold more generally for smooth d-dimensional families with Fisher information $I(\theta)$, and parameter θ restricted to a set S, in accordance with Rissanen [34]. It also corresponds to the answer for expected regret from Clarke and Barron [12]. However, the present case of the family of all distributions on the simplex does not satisfy the conditions of [12] or [34].

Remark 5 Comparing r_n with the minimax value using expected loss in [44] and [12], $\frac{m-1}{2}\log\frac{n}{2\pi e}+\log\frac{\Gamma(\frac{1}{2})^m}{\Gamma(\frac{m}{2})}+o(1)$, we see that there is a difference of

$\frac{m-1}{2}\log e$. The difference is due to the use in the expected loss formulation of a target value of $E_\theta \log 1/p(X^n|\theta)$ rather than $E_\theta \log 1/p(X^n|\hat{\theta})$, which differ by $E_\theta \log(p(X^n|\hat{\theta})/p(X^n|\theta)$, which is approximately one-half the expectation of a chi-square random variable with $m-1$ degrees of freedom. It may be surprising that there is no difference asymptotically between the answers for minimax regret for individual sequences $\min_q \max_{x^n} \log p(x^n|\hat{\theta})/q(x^n)$ and minimax expected regret $\min_q \max_\theta E_\theta \log p(x^n|\hat{\theta})/q(x^n)$.

Remark 6 The constant in the asymptotic minimax regret $C_m = \log((\Gamma(1/2))^m/\Gamma(m/2))$ is also identified in Ordentlich and Cover [30] in a stock market setup and by Freund [19] for the $m = 2$ case using Riemann integration to analyze the Shtarkov value $c_n = \sum_{k=0}^n \binom{n}{k}(k/n)^k(1-k/n)^{n-k}$, see Section 3.4 Also for $m = 2$, detailed asymptotics for c_n can be identified using the results of [28] and [38] that arise in other information theory contexts (as pointed out to us by Ordentlich). This constant $\log((\Gamma(1/2))^m/\Gamma(m/2))$ can also be obtained by inspection of inequality (15) in Shtarkov [36]. Here the determination of the constant is a by-product of our principal aim of identifying natural and easily implementable asymptotically maximin and minimax procedures.

Remark 7 Since $\Gamma(1/2) = \sqrt{\pi}$ and $\log\Gamma(m/2) = \log(\sqrt{2\pi}(\frac{m}{2})^{\frac{m-1}{2}}e^{-\frac{m}{2}}) + rem_m$ by Stirling's approximation to the Gamma function, see [42, pp. 253], an alternative expression for the asymptotic minimax regret from Theorem 1 is

$$r_n = \frac{m-1}{2}\log\frac{n}{m} + \frac{m}{2}\log e - \frac{1}{2}\log 2 - rem_m + o(1).$$

where $o(1) \to 0$ as $n \to \infty$ and the remainder rem_m in Stirling's approximation is between 0 and $\frac{1}{6m} \log e$. Thus with the remainder terms ignored, the minimax regret equals

$$\frac{m-1}{2} \log \frac{ne}{m}$$

plus a universal constant $\frac{1}{2} \log \frac{c}{2}$.

3.2 Proof of the main theorem

The statements of the theorem and the corollary are based on the following inequalities which we will prove.

$$\begin{aligned}
\frac{m-1}{2} \log \frac{n}{2\pi} + C_m &\leq \sum_{x^n} m_J(x^n) \log \frac{p(x^n|\hat{\theta})}{m_J(x^n)} && (3.2)\\
&\leq \max_W \sum_{x^n} m_W(x^n) \log \frac{p(x^n|\hat{\theta})}{m_W(x^n)} \\
&\leq \min_q \max_{x^n} \log \frac{p(x^n|\hat{\theta})}{q(x^n)} \\
&\leq \max_{x^n} \log \frac{p(x^n|\hat{\theta})}{\tilde{q}(x^n)} \\
&\leq \frac{m-1}{2} \log \frac{n}{2\pi} + C_m + o(1). && (3.3)
\end{aligned}$$

where $C_m = \log(\Gamma(1/2)^m/\Gamma(m/2))$. Since both ends in the above are asymptotically equal, it follows that

$$\begin{aligned}
\frac{m-1}{2} \log \frac{n}{2\pi} + C_m + o(1) &= \sum_{x^n} m_J(x^n) \log \frac{p(x^n|\hat{\theta})}{m_J(x^n)} && (3.4)\\
&= \log c_n = \underline{r}_n = \overline{r}_n \\
&= \max_{x^n} \log \frac{p(x^n|\hat{\theta})}{\tilde{q}(x^n)} + o(1)
\end{aligned}$$

$$= \frac{m-1}{2}\log\frac{n}{2\pi} + C_m + o(1).$$

and therefore, $C_m = \log(\Gamma(1/2)^m/\Gamma(m/2))$ is the asymptotic constant in the minimax regret, Jeffreys' mixture m_J is asymptotically maximin (least favorable), and the modified Jeffreys' mixture $\tilde{q}_n$ is asymptotically minimax.

As a corollary, we claim that $D(m_J||q_n^*) \to 0$. Indeed, $\sum_{x^n} m_J(x^n)\log(p(x^n|\hat{\theta})/m_J(x^n)) = \log c_n - D(m_J||q_n^*)$. Both $\sum_{x^n} m_J(x^n)\log(p(x^n|\hat{\theta})/m_J(x^n))$ and $\log c_n$ equal $\frac{m-1}{2}\log\frac{n}{2\pi} + C_m + o(1)$ asymptotically, by (3.4), thus the desired convergence of $D(m_{J,n}||q_n^*)$ follows. In the same way, $D(p_n||q_n^*) \to 0$ for any asymptotically maximin procedure p_n. Next we show that $D(m_{J,n}||\tilde{q}_n)$ converges to zero. Indeed more generally $D(p_n||q_n) \to 0$ for any asymptotically maximin p_n and asymptotically minimax q_n since $D(p_n||q_n) = D(p_n||q_n^*) + \sum p_n(x^n)\log(q_n^*(x^n)/q_n(x^n))$ and $\max_{x^n}\log(q_n^*(x^n)/q_n(x^n))$ tends to zero by asymptotic minimaxity of q_n.

We consider the regret using Jeffreys mixture $m_J(x^n)$. From Lemma 2.1 of the appendix, this regret is asymptotically constant (independent of x^n) for sequences with relative frequency composition internal to the simplex, that is, when $\min(T_1, \ldots, T_m) \to \infty$.

Lemma 2.3 exhibits a constant higher regret on vertex points when using Jeffreys' mixture, thus Jeffreys' mixture is not asymptotically minimax on the whole simplex of relative frequencies.

Now we verify inequalities (3.2) and (3.3). The three inequalities between them follow from the definitions and from maximin $\leq$ minimax.

The proof for line (3.2) follows directly from Lemma 2.2, which is actually

a stronger result. An alternative interpretation of this bound follows from the decomposition

$$\sum_{x^n} m_J(x^n) \log \frac{p(x^n|\hat{\theta})}{m_J(x^n)} = \int w_J(\theta) E_\theta \log \frac{p(X^n|\hat{\theta})}{p(X^n|\theta)} d\theta + \int w_J(\theta) E_\theta \log \frac{p(X^n|\theta)}{m_J(X^n)} d\theta. \quad (3.5)$$

where E_θ denotes expectation with respect to $p(x^n|\theta)$, and w_J is Jeffreys' prior. The first integral converges to $\frac{m-1}{2} \log e$ in agreement with the asymptotic χ^2_{m-1} distribution for $2 \log p(X^n|\hat{\theta})/p(X^n|\theta)$ for θ in the interior. The second integral in (3.5) is studied in [44] and [12], where it is shown to equal $\frac{m-1}{2} \log \frac{n}{2\pi e} + C_m$ asymptotically where $C_m = \log(\Gamma(1/2)^m/\Gamma(m/2))$ is the same constant as here.

The proof for line (3.3) follows. We denote the count of symbol i in a sequence x^n by $T_i = T_{i,n}$. Let $\tau_n \geq 1$ be a sequence with $\tau_n \to \infty$. Observe that for x^n in the region of $\mathcal{X}^n$ where $T_i \geq \tau_n$ for all $i = 1, \ldots, m$, using the upper bound from Lemma 2.1 in the appendix, we have

$$\begin{aligned} \log \frac{p(x^n|\hat{\theta})}{\tilde{q}_n(x^n)} &< \log \frac{p(x^n|\hat{\theta})}{(1-\varepsilon_n) m_J(x^n)} \\ &\leq \left(\log \frac{\Gamma(\frac{1}{2})^m}{\Gamma(\frac{m}{2})} + \frac{m-1}{2} \log \frac{n}{2\pi} \right) \\ &\quad + \left(\left(\frac{m}{4\tau_n + 2} + \frac{m^2}{4n} \right) \log e + \log \frac{1}{1-\varepsilon_n} \right) \quad (3.6) \\ &= \left(\log \frac{\Gamma(\frac{1}{2})^m}{\Gamma(\frac{m}{2})} + \frac{m-1}{2} \log \frac{n}{2\pi} \right) + o(1). \quad (3.7) \end{aligned}$$

where the remainder term in (3.6) tends to zero uniformly (for sequences with $T_i \geq \tau_n$) as $n \to \infty$.

Now we consider the region of $\mathcal{X}^n$ where $T_i < \tau_n$ for some i. Here we take

$\tau_n \log \tau_n \leq \frac{1}{8} \log n$ (a choice of $\tau_n \sim \frac{1}{8} \log n / \log \log n$ suffices). This region is the union of the regions where $T_i < \tau_n$ for $i = 1, \ldots, m$. For the ith such region we use $\varepsilon_n m_i(x^n)$ to lower bound $\tilde{q}_n(x^n)$. For notational simplicity take $i = 1$. Then

$$\begin{aligned} m_1(x^n) &= \int_{\theta_2+\ldots+\theta_m=1-\frac{1}{n}} \left(\frac{1}{n}\right)^{T_1-\frac{1}{2}} \theta_2^{T_2-\frac{1}{2}} \ldots \theta_m^{T_m-\frac{1}{2}} d\theta_2 \ldots d\theta_{m-1} / \\ &\quad / \int_{\theta_2+\ldots+\theta_m=1-\frac{1}{n}} \left(\frac{1}{n}\right)^{-\frac{1}{2}} \theta_2^{-\frac{1}{2}} \ldots \theta_m^{-\frac{1}{2}} d\theta_2 \ldots d\theta_{m-1} \\ &= \frac{D_{m-1}(T_2+\frac{1}{2}, \ldots T_m+\frac{1}{2}) \cdot \left(\frac{1}{n}\right)^{T_1} \left(1-\frac{1}{n}\right)^{n-T_1}}{D_{m-1}(\frac{1}{2}, \ldots \frac{1}{2})}. \end{aligned}$$

and it follows that

$$\begin{aligned} \log \frac{p(x^n|\hat{\theta})}{\tilde{q}_n(x^n)} &\leq \log \frac{p(x^n|\hat{\theta})}{\varepsilon_n m_1(x^n)} \\ &= \log \frac{\left(\frac{T_1}{n}\right)^{T_1} \left(\frac{n-T_1}{n}\right)^{n-T_1} \prod_{i=2}^{m} \left(\frac{T_i}{n-T_1}\right)^{T_i}}{\varepsilon_n \left(\frac{1}{n}\right)^{T_1} \left(1-\frac{1}{n}\right)^{n-T_1} \frac{D_{m-1}(T_2+\frac{1}{2}, \ldots, T_m+\frac{1}{2})}{D_{m-1}(\frac{1}{2}, \ldots, \frac{1}{2})}} \\ &\leq \log \frac{1}{\varepsilon_n} + \log \frac{\prod_{i=2}^{m} \left(\frac{T_i}{n-T_1}\right)^{T_i}}{\frac{D_{m-1}(T_2+\frac{1}{2}, \ldots, T_m+\frac{1}{2})}{D_{m-1}(\frac{1}{2}, \ldots, \frac{1}{2})}} + T_1 \log T_1 \qquad (3.8) \\ &\leq \log \frac{1}{\varepsilon_n} + \frac{m-2}{2} \log \frac{n-T_1}{2\pi} + T_1 \log T_1 \\ &\quad + \left(\frac{m^2}{4n} + \frac{m}{2}\right) \log e + \log \frac{(\Gamma(\frac{1}{2}))^{m-1}}{\Gamma(\frac{m-1}{2})}. \qquad (3.9) \end{aligned}$$

where in the last inequality we used the conclusion of Lemma 2.1 in the appendix for the lower dimension Jeffreys' mixture. Now if we let ε_n be such that $\log \varepsilon_n^{-1} \leq \frac{1}{8} \log n$, then uniformly for $T_1 \leq \tau_n$ (i.e. $T_1 \log T_1 \leq \frac{1}{8} \log n$) we have

$$\log \frac{p(x^n|\hat{\theta})}{\tilde{q}_n(x^n)} \leq \frac{m-3/2}{2} \log n + \left(\frac{m^2}{4n} + \frac{m}{2}\right) \log e. \qquad (3.10)$$

Comparison of (3.7) and (3.10) shows that with the strategy $\tilde{q}_n$ the contribution from the boundary regions produces regret asymptotically less than the asymptotic regret in the interior $\frac{m-1}{2}\log\frac{n}{2\pi}+C_m$. As an aside we note more generally that for the bound from (3.9) to be less than the desired expression (3.7), ε_n and τ_n should be chosen such that

$$\log\frac{1}{\varepsilon_n}+\tau_n\log\tau_n\le\frac{1}{2}\log\frac{n}{2}-\log\frac{\Gamma(\frac{m}{2})}{\Gamma(\frac{m-1}{2})}-\left(\frac{m^2}{4n}+\frac{m}{2}\right)\log e.$$

The right side is not greater than $(1/2)\log(n/2)-(m/2)\log e$. Thus to obtain the desired bound uniformly over $\mathcal{X}^n$ it is sufficient to set a value of $\log(1/\varepsilon_n)=\tau_n\log\tau_n$ to be not larger than $(1/4)\log(n/2)-(m/4)\log e$.

Since the value of the asymptotic constant is the same for the upper and lower bounds the inequalities in (3.2) through (3.4) collapse into asymptotic equalities and the conclusions follow.

Finally we show that the modification to produce an asymptotically minimax procedure $\tilde{q}_n$ retains the asymptotic least favorable (maximin) property of Jeffreys' mixture. That is, $\sum_{x^n} r(\tilde{q}_n,x^n)\tilde{q}_n(x^n)=\log c_n-D(\tilde{q}_n||q_n^*)=\log c_n+o(1)$ or equivalently $D(\tilde{q}_n||q_n^*)\to 0$. Indeed, we have $D(\tilde{q}_n||q_n^*)=D((1-\varepsilon_n)m_{J,n}+\varepsilon_n m_{V,n}||q_n^*)$ which by convexity is not greater than $(1-\varepsilon_n)D(m_{J,n}||q_n^*)+\varepsilon_n D(m_{V,n}||q_n^*)$. We already showed the first term goes to zero. The second term also converges to zero since $D(m_{V,n}||q_n^*)\le\log c_n$ and $\varepsilon_n\to 0$ faster than logarithmically. Thus $D(\tilde{q}_n||q_n^*)\to 0$ as $n\to\infty$. ■

3.3 Other modifications of Jeffreys' prior

In this section we explore other possibilities of modifying Jeffreys' mixture, and we also discuss the achievable rates of convergence of the modifications proposed.

In Section 3.2 we added some point mass to the Jeffreys' prior near the boundary of the simplex to pull down the regret incurred by sequences with relative frequencies close to or on the boundary. It produced maximal regret that exceeds the asymptotic minimax value by not more than order $\log\log n/\log n$ as determined by the choice of τ_n; see (3.6). For that procedure, we may modify the prior using components J_t with $\theta_t \sim (\log n)/n$ rather than $1/n$ and with probability $\varepsilon_n \sim 1/\log n$ instead of $n^{-1/8}$ to permit a slight improvement in the rate for the remainder in the maximum regret from $\log\log n/\log n$ to order $1/\log n$. In this section we show that a modification based on Dirichlet priors with parameters less that $1/2$ provides a convenient algorithm and a faster convergence rate.

The modified Jeffreys' prior we study here is

$$W_n^{(2)} = (1-\varepsilon_n)\text{Dirichlet}_m(\frac{1}{2},\ldots,\frac{1}{2}) + \varepsilon_n \text{Dirichlet}_m(\alpha,\ldots,\alpha),$$

where $0 < \alpha < 1/2$ and ε_n will be specified. The above prior $W_n^{(2)}$ yields a mixture probability mass function

$$q_n^{(2)}(x^n) = (1-\varepsilon_n)\frac{D_m(T_1+\frac{1}{2},\ldots,T_m+\frac{1}{2})}{D_m(\frac{1}{2},\ldots,\frac{1}{2})} + \varepsilon_n \frac{D_m(T_1+\alpha,\ldots,T_m+\alpha)}{D_m(\alpha,\ldots,\alpha)},$$

and we are to show that $q_n^{(2)}$ is also asymptotically minimax. The proof follows.

For x^n in the region of $\mathcal{X}^n$ where $T_i \geq \tau_n = n^p$ for some $p > 0$ and all $i = 1, \ldots, m$, we have, from Lemma 2.1, that

$$\log \frac{p(x^n|\hat{\theta})}{q_n^{(2)}(x^n)} \leq \left(\log \frac{\Gamma(\frac{1}{2})^m}{\Gamma(\frac{m}{2})} + \frac{m-1}{2} \log \frac{n}{2\pi}\right) + \left(\left(\frac{m^2}{4n} + \frac{m}{4\tau_n + 2}\right) \log e + \log \frac{1}{1-\varepsilon_n}\right). \quad (3.11)$$

For x^n in the region of $\mathcal{X}^n$ where $T_i < \tau_n = n^p$ for some i, we use Lemma 2.4 of the appendix to get that

$$\log \frac{p(x^n|\hat{\theta})}{q_n^{(2)}(x^n)} \leq \left(\frac{m-1}{2} - (1/2 - \alpha)(1-p)\right) \log n + \left(K_m \log \frac{1}{\alpha} + \log \frac{1}{\varepsilon_n}\right),$$

where K_m is a constant depending only on m. Let $\varepsilon_n = n^{-s}$ for some $s > 0$. Then as long as $(1/2 - \alpha)(1-p) > s$, for large enough n, we have

$$\begin{aligned} & \left(\frac{m-1}{2} - (\frac{1}{2} - \alpha)(1-p) + s\right) \log n + K_m \log \frac{1}{\alpha} \\ \leq \; & \frac{m-1}{2} \log \frac{n}{2\pi} + \log \frac{\Gamma(\frac{1}{2})^m}{\Gamma(\frac{m}{2})} + o(1). \end{aligned} \quad (3.12)$$

Combining (3.11) and (3.12), we conclude that for certain choice of p, s and α, the regret $r(q_n^{(2)}, x^n)$ is asymptotically upperbounded by $\log \frac{\Gamma(1/2)^m}{\Gamma(m/2)} + \frac{m-1}{2} \log \frac{n}{2\pi} + o(1)$, uniformly for all x^n. For example a choice of $p = s = 1/4$ and $\alpha = 1/8$ would satisfy (3.12). Consequently $q_n^{(2)}$ is asymptotically minimax.

Let's take $p = s = 1/4$ and $\alpha = 1/8$, then

$$q_n^{(2)}(x^n) = (1 - n^{-\frac{1}{4}}) \frac{D_m(T_1 + \frac{1}{2}, \ldots, T_m + \frac{1}{2})}{D_m(\frac{1}{2}, \ldots, \frac{1}{2})} + n^{-\frac{1}{4}} \frac{D_m(T_1 + \frac{1}{8}, \ldots, T_m + \frac{1}{8})}{D_m(\frac{1}{8}, \ldots, \frac{1}{8})}.$$

The predictive density is

$$q_n^{(2)}(x_{k+1} = j | x^k) = \frac{(1 - n^{-\frac{1}{4}}) m_j(x^{k+1}) + n^{-\frac{1}{4}} m_{1/8}(x^{k+1})}{(1 - n^{-\frac{1}{4}}) m_j(x^k) + n^{-\frac{1}{4}} m_{1/8}(x^k)}$$

with

$$m_J(x^{k+1}) = m_J(x^k)m_J(x_{k+1}|x^k) \text{ and } m_J(x_{k+1} = j|x^k) = \frac{T_{k,j} + \frac{1}{2}}{k + \frac{m}{2}}$$

and

$$m_{\frac{1}{8}}(x^{k+1}) = m_{\frac{1}{8}}(x^k)m_{1/8}(x_{k+1}|x^k) \text{ and } m_{\frac{1}{8}}(x_{k+1} = j|x^k) = \frac{T_{k,j} + \frac{1}{8}}{k + \frac{m}{8}}.$$

The total (recursive) computation time is of order nm.

We study how fast the corresponding regrets approach the asymptotic minimax value for each $\alpha < 1/2$. From (3.11) we have that for $n \geq m$,

$$\log \frac{p(x^n|\hat{\theta})}{q_n^{(2)}(x^n)} - \left(\frac{m-1}{2}\log\frac{n}{2\pi} + C_m\right) \leq \frac{m\log e}{\tau_n} + 2\varepsilon_n \log e. \tag{3.13}$$

To balance the rates at which $1/\tau_n$ and ε_n tend to zero in this upper bound, we set $p = s$. Then condition (3.12) reduces to

$$(s - (1/2 - \alpha)(1 - s))\log n + K_m \log\frac{1}{\alpha} \leq K'_m$$

for large enough n, where the constant $K'_m = \frac{m-1}{2}\log(1/2\pi) + \log\Gamma(\frac{1}{2})^m/\Gamma(\frac{m}{2})$. That is,

$$s \leq \frac{(\frac{1}{2} - \alpha)\log n + (K' + K_m \log\alpha)/\log n}{(\frac{3}{2} - \alpha)\log n}. \tag{3.14}$$

We can achieve a value of $s_n = 1/3 - O(\log\log n/(\log n)^2)$ by setting $\alpha = K_m/(\log n)^2$ to maximize the numerator of (3.14). Recall from (3.13) that the difference between our regret and the asymptotic minimax values is bounded by $(m+2)n^{-s}\log e$. Plugging $s = s_n$ in we obtain a bound of order $n^{-1/3}$.

Now we compare the three priors: Jeffreys' and the two modifications. Jeffreys' mixture achieves the asymptotic minimax value for sequences internal to

the relative frequency simplex. The two modifications are asymptotically minimax for the set of all sequences x^n. We studied some upperbounds on the rate at which the regret converges to the asymptotic minimax value. The first modification approaches the asymptotic minimax value at least at a rate of $1/\log n$, and the second modification at a rate of $n^{-1/3}$. From the asymptotics of c_n in [38] we know it differs from the asymptotic value by order $1/\sqrt{n}$ when $m = 2$.

3.4 An alternative method for determination of the asymptotic minimax value

The minimax value is $\log c_n = \log \sum_{x^n} p(x^n|\hat{\theta})$, by Theorem 0. This is asymptotically equal to $\frac{d}{2}\log\frac{n}{2\pi} + \log(\Gamma(1/2)^m/\Gamma(m/2)) + o(1)$ as proven above. Here we give an alternative direct proof using Stirling's formula and extended Riemann integration. This is the method of Cover and Ordentlich [8] and of Freund [19] in handling the $m = 2$ case. Also Szpankowski [38] gave an expansion of c_n accurate to arbitrary order for $m = 2$ case, however that method does not apply when $m \geq 3$. Here we give it for arbitrary m.

For the lower bound of c_n, recalling that $c_n = \sum_{x^n} p(x^n|\hat{\theta})$, we may rewrite c_n as

$$c_n = \sum_{T_1+\ldots+T_m=n} \frac{n!}{T_1!\cdot\ldots\cdot T_k!}\prod_{i=1}^{m}(\frac{T_i}{n})^{T_i}.$$

Now we apply Stirling's formula for $n!$ (see, e.g., [17, pp. 53])

$$\sqrt{2\pi n}n^n e^{-n} \leq n! \leq \sqrt{2\pi n}n^n e^{-n+1/n} \tag{3.15}$$

to get that

$$\begin{aligned} & \frac{n!}{T_1! \cdot \ldots \cdot T_k!} \\ \geq & \frac{\sqrt{2\pi n} n! e^{-n}}{\prod \left(\sqrt{2\pi T_i} T_i^{T_i} e^{-T_i + 1/T_i} \right)} \\ = & (2\pi)^{-\frac{m-1}{2}} \left(\frac{n}{T_1 \cdot \ldots \cdot T_m} \right)^{\frac{1}{2}} \left(\prod_{i=1}^{m} \left(\frac{T_i}{n} \right)^{-T_i} \right) e^{-\sum_{i=1}^{m} \frac{1}{T_i}}. \end{aligned}$$

Hence

$$c_n \geq \left(\frac{2\pi}{n} \right)^{-\frac{m-1}{2}} \sum_{\frac{T_1}{n} + \ldots + \frac{T_m}{n} = 1} \left(\frac{T_1}{n} \right)^{-\frac{1}{2}} \cdot \ldots \cdot \left(\frac{T_m}{n} \right)^{-\frac{1}{2}} e^{-\sum_{i=1}^{m} \frac{1}{T_i}}. \quad (3.16)$$

The factor $e^{-\sum_{i=1}^{m} 1/T_i}$ is near one for $T_1/n, \ldots, T_m/n$ sufficiently internal to the simplex. Thus we have for any $\mathcal{K} > 0$,

$$c_n \geq \left(\frac{n}{2\pi} \right)^{(m-1)/2} e^{-m\mathcal{K}/n} \cdot I_{n,\mathcal{K}}$$

where

$$I_{n,\mathcal{K}} = \sum_{\substack{\frac{T_1}{n} + \ldots + \frac{T_m}{n} = 1 \\ \text{all } T_i/n > 1/\mathcal{K}}} \left(\frac{T_1}{n} \right)^{-\frac{1}{2}} \cdot \ldots \cdot \left(\frac{T_m}{n} \right)^{-\frac{1}{2}} n^{-(m-1)}. \quad (3.17)$$

This sum reminds us of a Riemann integral. Let $S = \{(t_1, \ldots, t_{m-1}) : \text{all } t_i > 0$ and $\sum_{i=1}^{m-1} t_i < 1\}$ be the simplex and let $S_{\mathcal{K}}$ be the subset of S in which each $t_i \geq 1/\mathcal{K}$, $i = 1, \ldots, m$ where $t_m = 1 - \sum_{i=1}^{m-1} t_i$. Intersecting the set of cubes with corners of the form $T_1/n, \ldots, T_m/n$ with $S_{\mathcal{K}}$ provides a partition of $S_{\mathcal{K}}$ into sets of volume not larger than $n^{-(m-1)}$. Thus the sum in (3.17) is an upperbound on a Riemann approximation to the integral of $(t_1 \cdot \ldots \cdot t_m)^{-1/2}$ on $S_{\mathcal{K}}$. This integrand is continuous so by Riemann integration $\liminf_n I_{n,\mathcal{K}} \geq$

$\int_{0<t_1+..+t_{m-1}<1,\ \text{all } t_i\geq 1/\mathcal{K}} (t_1\cdot\ldots\cdot t_m)^{-1/2}\,dt_1\cdot\ldots\cdot dt_{k-1}$ where $t_m = 1-\sum_{i=1}^{m-1} t_i$.

Therefore it follows that asymptotically as $n\to\infty$,

$$c_n \geq \left(\frac{n}{2\pi}\right)^{(m-1)/2}\int_{S_{\mathcal{K}}}\frac{1}{\sqrt{t_1\cdot\ldots\cdot t_m}}dt_1\cdot\ldots\cdot dt_{m-1}\cdot(1-o(1)),$$

that is,

$$\liminf_{n\to\infty}\left(\log c_n - \frac{m-1}{2}\log\frac{n}{2\pi}\right)\geq \log\int_{S_{\mathcal{K}}}\frac{1}{\sqrt{t_1\cdot\ldots\cdot t_m}}dt_1\ldots dt_{m-1}. \qquad (3.18)$$

Furthermore $(t_1\cdot\ldots\cdot t_m)^{-1/2}$ is known to be Lebesgue integrable on S, thus letting $\mathcal{K}\to\infty$ in (3.18) we find that

$$\liminf_{n\to\infty}\left(\log c_n - \frac{m-1}{2}\log\frac{n}{2\pi}\right)\geq \log\int_{S}\frac{1}{\sqrt{t_1\cdot\ldots\cdot t_m}}dt_1\ldots dt_{m-1}.$$

The integral equals $D_m(1/2,\ldots,1/2) = \Gamma(\frac{1}{2})^m/\Gamma(\frac{m}{2})$, thus

$$\log c_n \geq \frac{m-1}{2}\log\frac{n}{2\pi} + \log\frac{\Gamma(\frac{1}{2})^m}{\Gamma(\frac{m}{2})} + o(1).$$

The upperbound can be established similarly using (3.15). Yet another demonstration of the asymptotic upper bound is by examination of an inequality in Shtarkov [36, Ineq. (15)]. In our notation, his bound is

$$c_n < \sum_{i=1}^{m}\binom{m}{i}\frac{\sqrt{\pi}}{\Gamma(i/2)}\left(\frac{n}{2}\right)^{(i-1)/2} \qquad (3.19)$$

The dominant term on the right side of (3.19) is for $i=m$. Thus we get an asymptotic upperbound

$$\begin{aligned} c_n &\leq \sum_{i=1}^{m}\binom{m}{i}\frac{\sqrt{\pi}}{\Gamma(i/2)}\left(\frac{n}{2}\right)^{(i-1)/2} \\ &\leq \frac{\sqrt{\pi}}{\Gamma(m/2)}\left(\frac{n}{2}\right)^{(m-1)/2}\left(1+\sum_{i=1}^{m-1}\binom{k}{i}\frac{\Gamma(m/2)}{\Gamma(i/2)}\left(\frac{n}{2}\right)^{-(m-i)/2}\right) \\ &= \frac{\sqrt{\pi}}{\Gamma(m/2)}\left(\frac{n}{2}\right)^{(m-1)/2}(1+o(1)). \end{aligned}$$

Thus $c_n \leq \frac{\Gamma(1/2)^m}{\Gamma(m/2)} \left(\frac{n}{2\pi}\right)^{(m-1)/2} (1 + o(1))$, or equivalently,

$$\log c_n \leq \frac{m-1}{2} \log \frac{n}{2\pi} + \log \frac{\Gamma(1/2)^m}{\Gamma(m/2)} + o(1).$$

Chapter 4

Applications in Prediction, Data Compression and Gambling

4.1 Application in gambling

Suppose in a horse race we index the horses by $1, \ldots, m$, and we are going to bet on n races. For race k, let the odds be $O_k(x|x_1, \ldots, x_{k-1})$ to 1 for horse x to win. We bet our fortune according to some proportion $q_n(x_k|x_1, \ldots, x_{k-1})$ at game k. Let $X^n = (X_1, \ldots, X_n)$ be the indices of the winning horses. Then the

asset at time n would be

$$\begin{aligned} S(X^n, q_n) &= \prod_{k=1}^{n} (q_n(X_k|X_1, \ldots, X_{k-1}) O_k(X_k|X_1, \ldots, X_{k-1})) \\ &= q_n(X_1, \ldots, X_n) O(X_1, \ldots, X_n), \end{aligned}$$

where $O(X_1, \ldots, X_n) = \prod_{k=1}^{n} O_k(X_k|X_1, \ldots, X_{k-1})$. If the horse races were random, with outcomes $X_1, \ldots, X_n$, if the win probabilities for each race were $(\theta_1, \ldots, \theta_m)$, and if we knew the parameter θ, we would bet with proportion $q_n(i) = \theta_i$ on horse i (see Cover and Thomas [9], Chapter 6). Whether or not the races are random, the wealth at time n with such a constant betting strategy θ is

$$\begin{aligned} S(X^n, p_\theta^n) &= \prod_{k=1}^{n} (p(X_k|\theta) O_k(X_k|X_1, \ldots, X_{k-1})) \\ &= p(X_1, \ldots, X_n|\theta) O(X_1, \ldots, X_n). \end{aligned}$$

where $p(x_1, \ldots, x_n|\theta) = \theta_1^{T_1} \cdot \ldots \cdot \theta_m^{T_m}$ and T_i is the number of wins for horse i. With hindsight the best of these values is at the maximum likelihood. Hence the ratio of current wealth to the ideal wealth is

$$\begin{aligned} R(X^n, q_n) &= \frac{S(X^n, q_n)}{S(X^n, p_{\hat{\theta}}^n)} \\ &= \frac{q(X_1, \ldots, X_n) O(X_1, \ldots, X_n)}{p(X_1, \ldots, X_n|\hat{\theta}) O(X_1, \ldots, X_n)} \\ &= \frac{q_n(X^n)}{p(X^n|\hat{\theta})}. \end{aligned}$$

We want to choose a $q_n(x^n)$ to optimize this ratio, in the worst case. That is, we pick a q_n to achieve

$$\min_{q_n} \max_{\theta, X^n} \log \frac{p(X^n|\theta)}{q_n(X^n)} = \min_{q_n} \max_{X^n} \log \frac{p(X^n|\hat{\theta})}{q_n(X^n)}.$$

This is the quantity our paper has analyzed, and we have provided an asymptotic minimax q_n. We achieve

$$\frac{q_n(X^n)}{p(X^n|\hat{\theta})} \geq C'_m \cdot n^{-\frac{m-1}{2}}(1+o(1)) \tag{4.1}$$

uniformly for all horse race outcomes X^n, where $C'_m = 2^{(m-1)/2}\Gamma(m/2)/\sqrt{\pi}$ is the best such constant. Here $n^{-\frac{m-1}{2}}$ expresses the cost (as a factor of wealth) of the lack of foreknowledge of $\hat{\theta}$. A gambling procedure that achieves (4.1) is to bet proportion $\tilde{q}(x_{k+1}|x^k)$ of our wealth on the possible outcomes of successive races using the modified Jeffreys' mixture as in equation (3.1).

There is an extension of this gambling problem to the stock market with m stocks. In this case

$$S(X^n, q_n) = \prod_{k=1}^{n}\left(\sum_{i=1}^{m} q_n(i|X_1, \ldots, X_{k-1})X_{ki}\right)$$

where X_{ki} is the wealth factor (price ratio) for stock i during investment period (day) k and $q(i|x_1, \ldots, x_{k-1})$ is the proportion of wealth invested in stock i at the beginning of day k. Recent work of Cover and Ordentlich [8], [30] shows that for all sequences $x_1, \ldots, x_n$, the minimax log wealth ratio for stocks is the same as the minimax log wealth ratio for horse racing with m horses:

$$\min_{q_n}\max_{\theta, x^n}\frac{S(x^n, p^n_\theta)}{S(x^n, q_n)} = \min_{q_n}\max_{x^n}\frac{p(x^n|\hat{\theta})}{q_n(x^n)}$$

where on the left side the maximum is over all $x_1, \ldots, x_n$ with each stock vector x_i in R^n_+ and on the right side the maximum is over all $x_1, \ldots, x_n$ with each x_i in $\{1, \ldots, m\}$. Thus from our analysis of the latter problem we have for the stock

market that the asymptotic minimax wealth ratio is $\min_{q_n} \max_{\theta, x^n} S(x^n, p_\theta^n)/S(x^n, q_n) = n^{\frac{m-1}{2}}/C'_m \cdot (1 + o(1))$ in agreement with Cover and Ordentlich [30]. However it remains an open problem whether there is an asymptotically minimax strategy that can be evaluated in polynomial time in n and m for the stock market. The best available algorithms in Cover and Ordentlich [30] runs in time of order n^{m-1} compared to time nm^2 obtained here for the horse race case.

4.2 Application in prediction

Suppose we have observed a sequence $x^k = (x_1, ..., x_k)$. We want to give a predictive probability function for the next x_{k+1} based on the past i observations, and we denote it by $\hat{p}_k(x|x^k) = q(x|x_1, ..., x_k)$ for all $x \in \mathcal{X}$. When x_{k+1} occurs we measure the loss by $\log 1/\hat{p}_k(x_{k+1}|x^k)$. Thus the loss is greater than or equal to zero (and equals zero iff the symbol x_{k+1} is predicted with $\hat{p}_k(x_{k+1}|x^k) = 1$). We initiate with a choice $\hat{p}_0(x) = q(x)$ of an arbitrary probability. We denote by

$$q(x_1, ..., x_n) = \prod_{k=0}^{n-1} q(x_{k+1}|x_1, ..., x_k).$$

the probability mass function obtained as the product of the predictive probabilities. The total cumulative log-loss is

$$\sum_{k=0}^{n-1} \log 1/q(x_{k+1}|x^k) = \log 1/q(x_1, ..., x_n). \tag{4.2}$$

A class $p(x_1, ..., x_n|\theta) = \prod_{k=1}^{n} p(x_k|\theta), \theta \in \Theta$ of memoryless predictors incurs cumulative log-loss $\sum_{k=0}^{n-1} \log 1/p(x_k|\theta) = \log 1/p(x_1, ..., x_n|\theta)$ for each θ and

with hindsight the best such predictor corresponds to the maximum likelihood. (Using this target class the aim of prediction is not to capture dependence between the $x_1, \ldots, x_n$ but rather to overcome the lack of advance knowledge of $\hat{\theta}$). The log-loss for prediction is chosen for the mathematical convenience of the chain rule (4.2). Direct evaluation of regret bounds is easier for such a loss than for other loss function. Moreover, log-loss regret provides bounds for minimax regret for certain other natural cumulative loss functions including 0-1 loss and squared error loss, see [26], [39] and [25]. The minimax cumulative regret is

$$\min_q \max_{\theta, x_1, \ldots, x_n} \sum_{k=0}^{n-1} \log \frac{p(x_{k+1}|\theta)}{q(x_{k+1}|x^k)} = \min_q \max_{x_1, \ldots, x_n} \frac{p(x_1, \ldots, x_n|\hat{\theta})}{q(x_1, \ldots, x_n)}$$

for which we have identified the asymptotics.

The Laplace–Jeffreys update rule is asymptotically maximin and its modification (as given in Theorem 1) is asymptotically minimax for online prediction.

4.3 Application in data compression

Shannon's noiseless source coding theory states that for each source distribution $p(x^n|\theta)$, the optimal code length of x^n is $\log 1/p(x^n|\theta)$, ignoring the integer rounding problem (if we do round it up to integer, the extra codelength is within one bit of optimum), where in Shannon's theory optimality is defined by minimum expected codelength. Kraft's inequality requires that the code length function $l(x^n)$ of a uniquely decodable code must satisfy $l(x^n) = \log 1/q(x^n)$ for some subprobability $q(x^n)$. When θ is unknown, we use a probability mass

function $q(x^n)$ such that for all θ and all x^n, the codelength using q is (to the extent possible) close to the smallest of the values $\log 1/p(x^n|\theta)$ over $\theta \in \Theta$. That is, we want to q to achieve

$$\min_q \max_{\theta, x_1, \ldots, x_n} (\log 1/q(x^n) - \log 1/p(x^n|\theta)) = \min_q \max_{\theta, x_1, \ldots, x_n} \frac{p(x^n|\hat{\theta})}{q(x^n)}.$$

The choice $q(x^n) = p(x^n|\hat{\theta}(x^n))$ is not available because Kraft's inequality is violated. Shtarkov showed that the minimax optimal choice is the normalized maximum likelihood $q(x^n) = p(x^n|\hat{\theta})/\sum_{x^n} p(x^n|\hat{\theta})$. Implementation of such codes for long block-length n would require computation of the marginals and conditionals associated with such a $q(x_1, \ldots, x_n)$. For the normalized maximum likelihood these conditionals (as required for arithmetic coding) are not easily computed. Instead we recommend the use of $q(x^n) = m_J(x^n)$ equal to Jeffreys' mixture or its modification, for which the conditionals are more easily calculated (see Remark 3). The arithmetic code for x^n is $\overline{F}(x^n) = \sum_{a^n < x^n} q(a^n) + \frac{1}{2} q(x^n)$ expressed in binary to an accuracy of $\lceil \log \frac{1}{q(x^n)} \rceil + 1$ bits. We can recursively update both $F(x^k)$ and $q_n(x^k)$ using the conditionals $q_n(x_k|x_1, \ldots, x_{k-1})$ in the course of the algorithm. For details see [9, pp. 104-107]. We remark here that the second modification from Chapter 4 also provides a straightforward algorithm for this arithmetic coding.

4.4 Categorical data prediction

We now look at some applications of our Theorem 1 in Xie and Barron (1996) in categorical data prediction.

Suppose a sequence of data $(x_j, y_j)_{j=1}^n$ are observed, where $y_j \in \{1, \ldots, m\}$ and $x_j \in \{1, \ldots, k\}$. We call y_j the response variable and x_j the explanatory variable. We wish to provide a choice of conditional distribution $q(y_1, \ldots, y_n | x_1, \ldots, x_n) = \prod_{j=1}^n q(y_j | y^{j-1}, x^j)$ for prediction, gambling, and data compression that perform well compared to a target family of competitors, uniformly over all sequences. The target family of procedures act according to an assumption that $y_1, \ldots, y_n$ are conditionally independent given $x_1, \ldots, x_n$, with the following conditional probability distribution

$$p(y_j = y | x_j = x) = \theta_{x,y}$$

for $k = 1, \ldots, n$, $y = 1, \ldots, m$ and $x = 1, \ldots, s$. These $\theta_{x,y}$'s are called parameters of the model. Denote the collection of these parameters by θ, that is, $\theta = (\theta_1, \ldots, \theta_k)$ with $\theta_s = (\theta_{s,1}, \ldots, \theta_{s,m})$ for $s = 1, \ldots, m$. (These parameters may be organized into a matrix.) Then the joint conditional probability under the competitor's model can be written as

$$\begin{aligned} p(y_1, \ldots, y_n | x_1, \ldots, x_n) &= \prod_{j=1}^n p(y_j | x_j) = \prod_{s=1}^k \prod_{j: x_j = s} p(y_j | s, \theta_s) \\ &= \prod_{s=1}^k p(y^{n \cdot} | \theta_s), \end{aligned}$$

where y^{n_s} is subsequence for which $x_j = s$. Thus

$$p(y^{n_s}|\theta_s) = \prod_{j:x_j=s} p(y_j|s,\theta_s)$$

treats the observations in this subsequence as if they were independent and identically distributed. The maximum likelihood estimator is

$$\hat{\theta}_s = \left(\frac{n_{s,1}}{\sum_{i=1}^{m} n_{s,i}}, \ldots, \frac{n_{s,m}}{\sum_{i=1}^{m} n_{s,i}} \right)$$

for $s = 1, \ldots, k$, where

$$n_{s,i} = \sum_{j=1}^{n} 1_{\{x_j=s, y_j=i\}}$$

is the number of observations for which the response is i when the explanatory variable is s.

We define the regret $r(x^n, y^n, q)$ for using a conditional probability function $q(y^n|x^n)$ as the log ratio between the best of the competitors probability $p(y^n|x^n, \hat{\theta})$ to our choice $q(y^n|x^n)$ at data points (x^n, y^n), that is,

$$r(x^n, y^n, q) = \log \frac{p(y^n|x^n, \hat{\theta})}{q(y^n|x^n)}.$$

We are interested to know the asymptotic minimax value $\overline{r}_n = \min_{q(\cdot|\cdot)} \max_{x^n, y^n} r(x^n, y^n, q)$, and a probability $q(y^n|x^n)$ that asymptotically achieves this minimax value. Moreover, we desire a "causal" q that is independent of future x_i's in the implementation process.

An asymptotic upperbound for the minimax value is derived from the following argument. Observe that

$$\overline{r}_n = \min_{q(\cdot|\cdot)} \max_{x^n, y^n} \log \frac{p(y^n|x^n, \hat{\theta})}{q(y^n|x^n)}$$

$$= \max_{x^n} \min_{q(\cdot|x^n)} \max_{y^n} \log \frac{p(y^n|x^n, \widehat{\theta})}{q(y^n|x^n)}. \tag{4.3}$$

Let $n_s(x^n) = \{j : x_j = s\}$ be the set of indices corresponding to the subsample of observations for which the explanatory variable takes value s. With slight abuse of notation we also use n_s to denote the size of this subsample, i.e., the cardinality of $n_s(x^n)$. We obtain an upperbound in (4.3) by restricting q to have property that

$$q(y^n|x^n) = \prod_{s=1}^{k} q(y^{n_s}|s). \tag{4.4}$$

where $y^{n_s} = (y_j : j \in n_s)$. Focus attention on the subsequence n_s. From [29] we have that mixture with respect to modified Dirichlet priors achieve asymptotically minimax regret for the target class of memoryless distribution on the m simplex. Motivated by that work we take $q(y^{n_s}|s)$ to be such a modified Dirichlet mixture of $p(y^{n_s}|s)$ for observations in the subsequence $n_s(x^n)$. Then from (4.4) and [29] the regret in (4.3) is upper bounded by

$$\begin{aligned}
\max_{x^n} \max_{y^n} \sum_{s=1}^{k} \log \frac{p(y^{n_s}|s, \widehat{\theta}_s)}{q(y^{n_s}|s)} &\leq \max_{x^n} \sum_{s=1}^{k} \max_{y^n} \log \frac{p(y^{n_s}|s, \widehat{\theta}_s)}{q(y^{n_s}|s)} \\
&\leq \max_{x^n} \sum_{s=1}^{k} \left(\frac{m-1}{2} \log \frac{n_s}{2\pi} + C_m + o(1/\log\log n_s) \right) \\
&= \frac{k(m-1)}{2} \log \frac{n}{2k\pi} + kC_m + o(1). \qquad (4.6)
\end{aligned}$$

where $n_{s,\min} = \min(n_{s,1}, \ldots, n_{s,m})$. See [29, Eq. (6)] for the validity of (4.5). Inequality (4.6) is obtained by letting all $n_s = n/k$ which maximizes the summation quantity in (4.5).

For a lower bound of $\overline{r}_n$ we use minimax $\geq$ maximin (in fact $\overline{r}_n = \underline{r}_n$ as

Theorem 0 shows). The maximin value is

$$\begin{aligned} \underline{r}_n &= \max_{x^n} \max_{p(y^n|x^n)} \min_{q(\cdot|x^n)} \sum_{y^n} p(y^n|x^n) \log \frac{p(y^n|x^n, \hat{\theta})}{q(y^n|x^n)} \\ &= \max_{x^n} \max_{p(y^n|x^n)} \sum_{y^n} p(y^n|x^n) \log \frac{p(y^n|x^n, \hat{\theta})}{p(y^n|x^n)}. \end{aligned} \tag{4.7}$$

We obtain a lower bound in (4.7) by choosing for each x^n

$$p^*(y^n|x^n) = \prod_{s=1}^{k} p^*(y^{n_s}|s),$$

where $p^*(y^{n_s}|s)$ is the mixture of $p(y^{n_s}|\theta_s)$ with respect to the Dirichlet$(1/2, \ldots, 1/2)$ prior. Then from Lemma 2 of [29], we know that

$$\begin{aligned} \log \frac{p(y^n|x^n, \hat{\theta})}{p^*(y^n|x^n)} &= \sum_{s=1}^{k} \log \frac{p(y^{n_s}|s, \hat{\theta}_s)}{p^*_{n_s}(y^{n_s})} \\ &\geq \sum_{s=1}^{k} \left(\frac{m-1}{2} \log \frac{n_s}{2\pi} + C_m \right). \end{aligned}$$

Hence continuing from (4.7), we have

$$\begin{aligned} \underline{r}_n &\geq \max_{x^n} \sum_{s=1}^{k} \left(\frac{m-1}{2} \log \frac{n_s}{2\pi} + C_m \right) \\ &= \frac{k(m-1)}{2} \log \frac{n}{2k\pi} + kC_m. \end{aligned}$$

Thus we have shown that the asymptotic minimax regret is

$$r_n = \frac{k(m-1)}{2} \log \frac{n}{2k\pi} + kC_m + o(1).$$

Furthermore, recalling the choice of q in (4.4), we have found a causal $q(y^n|x^n)$ that is asymptotically minimax. By causality we mean that q satisfies

$$q(y^n|x^n) = \prod_{j=1}^{n} q(y_j|x^j, y^{j-1}).$$

Here it is not necessary to condition on function x values as in the general decomposition $q(y^n|x^n) = \prod_{j=1}^{n} q(y_j|x^n, y^{j-1})$. Moreover the conditional distribution of y_j given x^j and y^{j-1} depends only on the subsample of past y_i of which $x_i = s$ when $x_j = s$. The advantage of using such a q is that, we can give an "online" prediction as data are revealed to us.

Chapter 5

Asymptotic Minimax Regret for the Class of Markov Sources

Suppose X^n is a Markov chain with stationary transition probabilities, with initial state X_0 already in a stationary status. Let $p(1|0) = \Pr(X_{m+1} = 1|X_m = 0) = \alpha$, $p(0|1) = \Pr(X_{m+1} = 0|X_m = 1) = \beta$. The stationary probabilities are $\pi_0 = \Pr(X_t = 0) = \frac{\beta}{\alpha+\beta}$ and $\pi_1 = \Pr(X_t = 1) = \frac{\alpha}{\alpha+\beta}$, at any time t.

The probability of a sample $x_0x_1...x_n$ is the product of $p(x_0)p(x_1|x_0) \cdot ... \cdot p(x_n|x_{n-1})$, by Markov property. It equals $p_{00}^{n_{00}} p_{01}^{n_{01}} p_{10}^{n_{10}} p_{11}^{n_{11}} \cdot p(x_0)$, where $p_{ij} = p_{ij}(\alpha, \beta) = \Pr(X_{t+1} = j|X_t = i)$ and n_{ij} is number of occurrences of $\{ij\}$ in

$x_0 x_1 ... x_n$. We also let n_0 and n_1 denote the number of zeros and ones in the sequence $x_1 ... x_n$. Therefore the "conditional Fisher information" I given X_0 is

$$I(\alpha, \beta | X_0) = -E \frac{\partial^2}{\partial(\alpha, \beta)^2} \log p_{00}^{n_{00}} p_{01}^{n_{01}} p_{10}^{n_{10}} p_{11}^{n_{11}}. \tag{5.1}$$

From the definition of α and β, we have $p_{00} = 1 - \alpha$, $p_{01} = \alpha$, $p_{11} = 1 - \beta$ and $p_{10} = \beta$. Also we observe that

$$\begin{aligned} En_{00} &= E \sum_{i=0}^{n-1} 1_{\{X_i = 0, X_{i+1} = 0\}} \\ &= \sum_{i=0}^{n-1} P\{X_i = 0, X_{i+1} = 0\} \\ &= \sum_{i=0}^{n-1} P\{X_{i+1} = 0 | X_i = 0\} P\{X_i = 0\} \\ &= n\pi_0(1 - \alpha). \end{aligned}$$

Similarly, $En_{01} = n\pi_0\alpha$, $En_{10} = n\pi_1\beta$ and $En_{11} = n\pi_1(1 - \beta)$. Thus

$$\begin{aligned} I(\alpha, \beta | X_0) &= \begin{vmatrix} n\pi_0/(\alpha(1-\alpha)) & 0 \\ 0 & n\pi_1/(\beta(1-\beta)) \end{vmatrix} \\ &= n^2\pi_0\pi_1/(\alpha(1-\alpha)\beta(1-\beta)). \end{aligned}$$

"Jeffreys' prior" $w^*(\alpha, \beta)$ is thus proportional to $J(\alpha, \beta) = (\alpha + \beta)^{-1}((1 - \alpha))^{-1/2}((1 - \beta))^{-1/2}$. This is a proper prior, since

$$\begin{aligned} C_J \stackrel{\text{def}}{=} \int_{[0,1]\times[0,1]} \frac{1}{(\alpha+\beta)\sqrt{(1-\alpha)(1-\beta)}} d\alpha d\beta &\leq \int_{[0,1]\times[0,1]} \frac{1}{(\alpha+\beta)} d\alpha d\beta \\ &= \int_0^1 (\ln(1+\beta) - \ln\beta)\, d\beta \\ &= 2\ln 2 < \infty. \end{aligned}$$

Let $m^{\heartsuit}(x^n)$ be the mixture of probability function p with $J(\alpha,\beta)$, that is,

$$m^{\heartsuit}(x^n) = \int (\alpha+\beta)^{-1}((1-\alpha))^{-1/2}((1-\beta))^{-1/2}\cdot\alpha^{n_{01}}(1-\alpha)^{n_{00}}\beta^{n_{10}}(1-\beta)^{n_{11}}d\alpha d\beta.$$

Note that $m^{\heartsuit}$ is not a probability mass function since J has not been scaled to a probability measure. This is for the computing convenience purpose, and the scaling can be carried out later.

We now study the regret

$$r(x^n) = \log \frac{p(x^n|\hat{\alpha},\hat{\beta})}{m^{\heartsuit}(x^n)}$$

for individual sequences x^n, where $\hat{\alpha} = n_{00}/n_0$ and $\hat{\beta} = n_{01}/n_1$ are the maximum likelihood estimators.

Theorem 5.1. The minimax regret for Markov class satisfies

$$\min_q \max_{x^n} \log \frac{p(x^n|\hat{\alpha},\hat{\beta})}{q(x^n)} \leq \log \frac{n}{2\pi} + \log C_J.$$

Let $I(\alpha) = 1/(\alpha(1-\alpha))$, and we restrict $\hat{\alpha},\hat{\beta}$ with $0 < d \leq \hat{\alpha},\hat{\beta} \leq 1-d < 1$. Let δ be any number such that $0 < \delta < d$ hence a circle of center $\hat{\alpha}$ or $\hat{\beta}$ with radius δ fall inside the square $[0,1] \times [0,1]$. For convenience we will study the inverse of regret, $1/r(x^n)$, i.e.,

$$\frac{1}{r(x^n)} = \frac{m^{\heartsuit}(x^n)}{p(x^n|\hat{\alpha},\hat{\beta})} = \int \exp\left[\log J(\alpha,\beta) - \log \frac{p(x^n|\hat{\alpha},\hat{\beta})}{p(x^n|\alpha,\beta)}\right] d\alpha d\beta. \qquad (5.2)$$

The second logarithm term in (5.2) is

$$\begin{aligned}
\log \frac{p(x^n|\hat{\alpha},\hat{\beta})}{p(x^n|\alpha,\beta)} &= \log \frac{\hat{\alpha}^{\hat{\alpha} n_0}(1-\hat{\alpha})^{(1-\hat{\alpha})n_0}}{\alpha^{\hat{\alpha} n_0}(1-\alpha)^{(1-\hat{\alpha})n_0}} + \log \frac{\hat{\beta}^{\hat{\beta} n_0}(1-\hat{\beta})^{(1-\hat{\beta})n_0}}{\beta^{\hat{\beta} n_0}(1-\beta)^{(1-\hat{\beta})n_0}} \\
&= n_0(\hat{\alpha}\log\hat{\alpha} + (1-\hat{\alpha})\log(1-\hat{\alpha}) - \hat{\alpha}\log\alpha - (1-\hat{\alpha})\log(1-\alpha)) + \\
&\quad n_1(\hat{\beta}\log\hat{\beta} + (1-\hat{\beta})\log(1-\hat{\beta}) - \hat{\beta}\log\beta - (1-\hat{\beta})\log(1-\beta)).
\end{aligned}$$

Now we use Taylor's expansion to get that

$$\hat{\alpha}\log\hat{\alpha} + (1-\hat{\alpha})\log(1-\hat{\alpha}) - \hat{\alpha}\log\alpha - (1-\hat{\alpha})\log(1-\alpha) = \frac{1}{2}I(\hat{\alpha},\tilde{\alpha})(\hat{\alpha}-\alpha)^2$$

for some $\tilde{\alpha}$ between $\hat{\alpha}$ and α, where I is the second derivative of the left side of the above equation.

$$I(\hat{\alpha},\alpha) = \frac{\hat{\alpha}}{\alpha^2} + \frac{1-\hat{\alpha}}{(1-\alpha)^2}. \tag{5.3}$$

Since $\tilde{\alpha}$ is between $\hat{\alpha}$ and α, we also have

$$\begin{aligned} I(\hat{\alpha},\tilde{\alpha}) &\leq \frac{d}{d-\delta}\cdot\frac{1}{\hat{\alpha}(1-\hat{\alpha})} \\ &= u\cdot I(\hat{\alpha}) \text{ (notation! } u \text{ for unit because } u\to 0) \end{aligned}$$

Returning to (5.2), we observe that when $|\alpha-\hat{\alpha}|\leq\delta, |\beta-\hat{\beta}|\leq\delta$.

$$\begin{aligned} \log J(\alpha,\beta) - \log J(\hat{\alpha},\hat{\beta}) &= \log(\hat{\alpha}+\hat{\beta}) - \log(\alpha+\beta) + \frac{1}{2}\log(1-\alpha) - \frac{1}{2}\log(1-\hat{\alpha}) + \\ &\quad + \frac{1}{2}\log(1-\beta) - \frac{1}{2}\log(1-\hat{\beta}). \end{aligned} \tag{5.4}$$

but

$$\begin{aligned} \log\frac{\alpha+\beta}{\hat{\alpha}+\hat{\beta}} &\leq \log\frac{d+d}{(d-\delta)+(d-\delta)} \\ &\leq \frac{\delta}{d-\delta}. \end{aligned}$$

and similarly

$$\begin{aligned} \frac{1}{2}\log(1-\hat{\alpha}) - \frac{1}{2}\log(1-\alpha) &= \frac{1}{2}\log\frac{1-\alpha}{1-\hat{\alpha}} \\ &\leq \frac{1}{2}\frac{\delta}{d-\delta}. \end{aligned}$$

hence it follows that

$$\log J(\alpha,\beta) \geq \log J(\hat{\alpha},\hat{\beta}) - \varepsilon.$$

where

$$\varepsilon \stackrel{\text{def}}{=} \frac{2\delta}{d-\delta}$$

Having controlled the two logarithm terms in (5.2), we now give a lower bound of the integral. Let

$$M_\delta(\hat{\alpha},\hat{\beta}) = \left\{(\alpha,\beta) : |\alpha-\hat{\alpha}| \le \delta, |\beta-\hat{\beta}| \le \delta\right\}.$$

Using the bounds we have obtained, we have

$$\begin{aligned}
&\frac{m^{\heartsuit}(x^n)}{p(x^n|\hat{\alpha},\hat{\beta})} \\
\ge\ & J(\hat{\alpha},\hat{\beta})\exp(-\varepsilon)\int_{M_\delta}\exp\left[-\frac{n_0 uI(\hat{\alpha})}{2}(\alpha-\hat{\alpha})^2-\frac{n_1 uI(\hat{\beta})}{2}(\beta-\hat{\beta})^2\right]d\alpha d\beta \\
=\ & J(\hat{\alpha},\hat{\beta})\exp(-\varepsilon)\cdot\int_{|\alpha-\hat{\alpha}|\le\delta}-\frac{n_0 uI(\hat{\alpha})}{2}(\alpha-\hat{\alpha})^2 d\alpha\cdot\int_{|\beta-\hat{\beta}|\le\delta}-\frac{n_1 uI(\hat{\beta})}{2}(\beta-\hat{\beta})^2 d\beta.
\end{aligned} \tag{5.5}$$

But

$$\begin{aligned}
\int_{|\alpha-\hat{\alpha}|\le\delta}\exp-\frac{n_0 uI(\hat{\alpha})}{2}(\alpha-\hat{\alpha})^2 d\alpha &= \int_R \exp-\frac{n_0 uI(\hat{\alpha})}{2}(\alpha-\hat{\alpha})^2 d\alpha - \\
&\quad -\int_{|\alpha-\hat{\alpha}|>\delta}\exp-\frac{n_0 uI(\hat{\alpha})}{2}(\alpha-\hat{\alpha})^2 d\alpha
\end{aligned} \tag{5.6}$$

and we study the two integrals in the right side of (5.6) separately.

The first one is relatively easy, since

$$\begin{aligned}
\int_R \exp-\frac{n_0 uI(\hat{\alpha})}{2}(\alpha-\hat{\alpha})^2 d\alpha &= \sqrt{\frac{2\pi}{n_0 uI(\hat{\alpha})}}\int_R\sqrt{\frac{n_0 uI(\hat{\alpha})}{2\pi}}\exp-\frac{n_0 uI(\hat{\alpha})}{2}(\alpha-\hat{\alpha})^2 d\alpha \\
&= \sqrt{\frac{2\pi}{n_0 uI(\hat{\alpha})}}.
\end{aligned}$$

The second integral is upperbounded as follows.

$$\begin{aligned}\int_{|\alpha-\hat{\alpha}|>\delta} \exp -\frac{n_0 uI(\hat{\alpha})}{2}(\alpha-\hat{\alpha})^2 d\alpha &= 2\int_{\alpha>\delta} \exp -\frac{n_0 uI(\hat{\alpha})}{2}\alpha^2 d\alpha \\ &\leq 2\sqrt{\frac{1}{n_0 I(\hat{\alpha})}} \exp\left(-\frac{n_0 uI(\hat{\alpha})}{2}\delta^2\right).\end{aligned}$$

Thus for (5.6) we have

$$\begin{aligned}\int_{|\alpha-\hat{\alpha}|\leq\delta} \exp -\frac{n_0 uI(\hat{\alpha})}{2}(\alpha-\hat{\alpha})^2 d\alpha &\geq \sqrt{\frac{2\pi}{n_0 uI(\hat{\alpha})}}\left(1-\sqrt{\frac{2}{\pi}}\exp\left(-\frac{n_0 uI(\hat{\alpha})}{2}\delta^2\right)\right) \\ &\geq \sqrt{\frac{2\pi}{n_0 uI(\hat{\alpha})}}\left(1-\exp\left(-\frac{n_0 uI(\hat{\alpha})}{2}\delta^2\right)\right) \\ &\geq \sqrt{\frac{2\pi}{n_0 uI(\hat{\alpha})}}\exp\left(-\frac{2}{n_0 uI(\hat{\alpha})\delta^2}\right). \qquad (5.7)\end{aligned}$$

where (5.7) is from

$$1-\exp(-x) \geq \exp(-1/x) \text{ for } x \geq 0 \qquad \text{(to show!)}$$

Plug (5.6) into (5.5), we get

$$\frac{m^{\heartsuit}(x^n)}{p(x^n|\hat{\alpha},\hat{\beta})} \geq J(\hat{\alpha},\hat{\beta})\exp(-\varepsilon)\cdot\frac{2\pi}{u\sqrt{n_0 n_1 I(\hat{\alpha})I(\hat{\beta})}}\exp\left(-\frac{2}{n_0 uI(\hat{\alpha})\delta^2}-\frac{2}{n_1 uI(\hat{\beta})\delta^2}\right).$$

But

$$\begin{aligned}n_0 I(\hat{\alpha}) &= n\frac{n_0}{n}I(\hat{\alpha}) \\ &= n\frac{\hat{\beta}}{\hat{\alpha}+\hat{\beta}}\frac{1}{\hat{\alpha}(1-\hat{\alpha})}\end{aligned}$$

and similarly

$$n_1 I(\hat{\beta}) = n\frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}}\frac{1}{\hat{\beta}(1-\hat{\beta})}.$$

Thus

$$\begin{aligned}\frac{m^{\triangledown}(x^n)}{p(x^n|\hat{\alpha},\hat{\beta})} &\geq J(\hat{\alpha},\hat{\beta})\exp(-\varepsilon)\cdot\frac{2\pi}{J(\hat{\alpha},\hat{\beta})nu}\exp\left(-\frac{2}{n_0uI(\hat{\alpha})\delta^2}-\frac{2}{n_1uI(\hat{\beta})\delta^2}\right)\\ &= \frac{2\pi}{n}\exp\left(-\varepsilon-\frac{2}{n_0uI(\hat{\alpha})\delta^2}-\frac{2}{n_1uI(\hat{\beta})\delta^2}-\log u\right).\end{aligned}$$

Thus we see that as long as that uniformly in x^n.

- $\varepsilon=\frac{2\delta}{d-\delta}\to 0$;
- $u\to 1$
- $n_0I(\hat{\alpha})\delta^2\to\infty$;
- $n_1I(\hat{\beta})\delta^2\to\infty$.

then we have that uniformly for x^n with $(\hat{\alpha},\hat{\beta})\in[d,1-d]\times[d,1-d]$.

$$\log\frac{p(x^n|\hat{\alpha},\hat{\beta})}{m^{\triangledown}(x^n)}\leq\log\frac{n}{2\pi}+o(1).$$

Such a choice of δ,d could be, for example

$$d_n=\frac{\log n}{n^{1/2}},\delta_n=\frac{1}{n^{1/2}}. \tag{5.8}$$

In fact, we need only to show $n_0I(\hat{\alpha})\delta^2\to\infty$. In fact,

$$\begin{aligned}n_0I(\hat{\alpha})\delta^2 &= n\frac{\hat{\beta}}{\hat{\alpha}+\hat{\beta}}\frac{1}{\hat{\alpha}(1-\hat{\alpha})}\delta_n^2\\ &\geq n\frac{\log n}{2n^{1/2}}\frac{1}{1/4}\frac{1}{n^{1/2}}\to\infty.\end{aligned}$$

When $\hat{\alpha},\hat{\beta}$ does not fall into the region $[d,1-d]\times[d,1-d]$, we use different measures to approach $p(x^n|\hat{\alpha},\hat{\beta})$. Specifically, when $\hat{\alpha}\in[0,1]$ and $\hat{\beta}\leq d$, let

$q'(x^n) = \beta(n_{00}+1/2, n_{01}+1/2)/\beta(1/2,1/2)\cdot\beta(n_{10}+1/4, n_{11}+1/4)/\beta(1/4,1/4) \triangleq$ $\beta_{1/2}(n_{00}, n_{01})\cdot\beta_{1/4}(n_{10}, n_{11})$, where $\beta(\cdot|\cdot)$ is a beta function. Then

$$\begin{aligned}\log\frac{p(x^n|\hat{\alpha},\hat{\beta})}{q'(x^n)} &= \log\frac{\hat{\alpha}^{n_{01}}(1-\hat{\alpha})^{n_{00}}\cdot\hat{\beta}^{n_{10}}(1-\hat{\beta})^{n_{11}}}{\beta_{1/2}(n_{00},n_{01})\cdot\beta_{1/4}(n_{10},n_{11})}\\ &= \log\frac{\hat{\alpha}^{n_{01}}(1-\hat{\alpha})^{n_{00}}}{\beta_{1/2}(n_{00},n_{01})}+\log\frac{\hat{\beta}^{n_{10}}(1-\hat{\beta})^{n_{11}}}{\beta_{1/4}(n_{10},n_{11})}.\end{aligned}$$

For the first integral in (??), from Lemma 2.1 of the appendix, we have

$$\log\frac{\hat{\alpha}^{n_{01}}(1-\hat{\alpha})^{n_{00}}}{\int\alpha^{n_{01}}(1-\alpha)^{n_{00}}d\alpha} \le \frac{1}{2}\log\frac{n_0}{2\pi}+\frac{1}{\min(n_{01},n_{00})+2}.$$

We control the second term citing the Lemma 2.4: observe that $n_{10}/n_1 < (\log n)/\sqrt{n}$, implying that $n_{10} < n_1^{3/4}$, hence that lemma gives

$$\begin{aligned}\log\frac{\hat{\beta}^{n_{10}}(1-\hat{\beta})^{n_{11}}}{\beta_{1/4}(n_{10},n_{11})} &\le \left(\frac{1}{2}-\frac{1}{4}\frac{1}{4}\right)\log n_1+Const.\\ &= \frac{7}{16}\log n_1+Const.\end{aligned}$$

Together we have

$$\begin{aligned}\log\frac{p(x^n|\hat{\alpha},\hat{\beta})}{q_1(x^n)} &\le \frac{1}{2}\log n_0+\frac{7}{16}\log n_1+C''\\ &< \log n+C''\end{aligned}$$

where C',C'' are constants. That is, we have provided a q_1 that incurs a smaller regret for x^n near the boundary $[0,1]\times\{0\}$. Similarly for other boundaries. Similar convex combinations of q_i's and Jeffreys' mixture lead to the Theorem 5.1.

Chapter 6

Appendix

Proposition 1.1 *(Pointwise asymptotic behavior of $D(p_\theta^n||m_n^*)$):* For an interior point θ of the simplex S_k', i.e., $\theta_i > 0$ for $i = 1, \ldots, k$, the following holds.

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

In particular, for any $\varepsilon > 0$, if we take $c = 2k/\varepsilon$, then for $n > kc$ and $n\theta_i \geq c$ for $i = 1, \ldots, k$, 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 $T_j = \sum_1^n 1_{\{X_i = \{a_j\}\}}$ for $j = 1, \ldots, k$, we have

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

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

$$= \sum_{i=1}^{k} n\theta_i \ln \theta_i - E_{\boldsymbol{\theta}} \ln D_k(T_1 + \frac{1}{2}, \ldots, T_k + \frac{1}{2}) + \ln D_k(\frac{1}{2}, \ldots, \frac{1}{2}). \tag{6.2}$$

Now applying the relationship between Dirichlet integrals and Gamma functions (2.6) and Stirling's approximation refined by Robbins [?].

$$\Gamma(x) = \sqrt{2\pi} x^{x-\frac{1}{2}} e^{-x}(1 + r) \quad \text{with} \quad |r| \leq e^{\frac{1}{12x}} - 1. \tag{6.3}$$

we may rewrite the middle term of (6.2):

$$\begin{aligned} & E_{\boldsymbol{\theta}} \ln D_k(T_1 + \frac{1}{2}, \ldots, T_k + \frac{1}{2}) \\ = \; & E_{\boldsymbol{\theta}} \ln \frac{\prod_1^k (\sqrt{2\pi}(T_i + \frac{1}{2})^{T_i})}{\sqrt{2\pi}(n + \frac{k}{2})^{n + \frac{k-1}{2}}} + E_{\boldsymbol{\theta}} \ln \frac{\prod_1^k (1 + r_i)}{1 + r_0} \\ = \; & \overbrace{\frac{k-1}{2} \ln 2\pi + \sum_1^k E_{\theta_i} T_i \ln(T_i + \frac{1}{2})}^{(A)} - \\ & \overbrace{- (n + \frac{k-1}{2}) \ln(n + \frac{k}{2})}^{(B)} + \overbrace{E_{\boldsymbol{\theta}} \ln \frac{\prod_1^k (1 + r_i)}{1 + r_0}}^{(C)} \end{aligned} \tag{6.4}$$

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 (6.4) separately.

For the deterministic term (B), we have

$$\left| \left(n + \frac{k-1}{2}\right) \ln \left(n + \frac{k}{2}\right) - \left(n \ln n + \frac{k-1}{2} \ln n + \frac{k}{2}\right) \right| \leq \frac{k(k-1)}{4n}. \tag{6.5}$$

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

$$-\sum_{i=1}^{k}\frac{1}{6n\theta_i}-\frac{1}{6n}\le E_{\boldsymbol{\theta}}\ln\frac{\prod_1^k(1+r_i)}{1+r_0}\le\sum_{i=1}^{k}\frac{1}{12n\theta_i}+\frac{1}{6n}. \tag{6.6}$$

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

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

$$E_{\theta_i}T_i\ln(T_i+\frac{1}{2})=\overbrace{E_{\theta_i}T_i\ln T_i}^{(A_1)}+\overbrace{E_{\theta_i}T_i\ln(1+\frac{1}{2T_i})}^{(A_2)}. \tag{6.7}$$

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

$$-\frac{1}{48n\theta_i}\le E_{\boldsymbol{\theta}}(T_i\ln T_i)\ -n\theta_i\ln n\theta_i-\frac{1-\theta_i}{2}\le\frac{1}{n\theta_i}. \tag{6.8}$$

Now we lower bound the (A_2) term in (6.7):

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

where the first inequality holds because $x\log(1+1/(2x))\ge 1/2-1/(2x+2)$ for $x\ge 0$, and the second one holds because $E_\theta(1/(T+1))\le 1/(n\theta)$, a useful lemma (Lemma 2) in [2] which is also used in the proof of Lemma 2. Now observe that $1/2$ upperbounds term (A_2), since $x\log(1+1/(2x))\le 1/2$ for $x\ge 0$.

$$\frac{1}{2}-\frac{1}{2n\theta_i}\le E_{\theta_i}T_i\ln(1+\frac{1}{2T_i})\le\frac{1}{2}. \tag{6.9}$$

Combining (6.8) and (6.9) then summing the result over i yields a bound for term (A)

$$-\frac{25}{48n}\sum_{i=1}^{k}\frac{1}{\theta_i}\le\sum_{i=1}^{k}E_{\theta_i}T_i\ln(T_i+\frac{1}{2})-\sum_{i=1}^{k}n\theta_i\ln n\theta_i-(k-\frac{1}{2})\le\frac{1}{n}\sum_{i=1}^{k}\frac{1}{\theta_i}. \tag{6.10}$$

Now we incorporate (6.5), (6.6) and (6.10) into a bound for $D(p_\theta^n||m_n^*)$:

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

In particular, if we take $c = 2k/\varepsilon$, then for $n \geq kc$ and $n\theta_i \geq c$ for $i = 1, \ldots, k$, 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 $c(\varepsilon) = (10/3)\varepsilon^{-1}\log_2 e \leq 5/\varepsilon$ 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 1.2 *(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 $\theta \notin \{0, 1\}$.

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

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

$$-\frac{1}{6n\theta}\log e \leq E_\theta \log(1+r) \leq \frac{1}{12n\theta}\log e.$$

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 (6.3) with $x = T + 1/2$, the residual r satisfies

$$|r| \leq \exp(\frac{1}{12T+6}) - 1. \tag{6.11}$$

Thus

$$\begin{aligned}\ln(1+r) &\geq \ln(2-\exp\frac{1}{12T+6}) \\ &\geq \ln(\exp(-\frac{2}{12T+6})) \\ &= -\frac{1}{6T+3}\end{aligned}$$

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

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

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

The upper bound is more direct. Again using (6.11), we have

$$\begin{aligned}\ln(1+r) &\leq \ln(\exp\frac{1}{12T+6}) \\ &= \frac{1}{12T+6}\end{aligned}$$

Thus we have completed the proof of Lemma 2.

Lemma 1.3 *(Local property of $E_\theta(T\log T)$)*: Let $T \sim \text{Binomial}(n,\theta)$. For any $\theta \notin \{0,1\}$ and $n\theta > 2$,

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

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,

$$\begin{aligned}y\ln y &= z\ln z + (y-z)(1+\ln z) + \frac{1}{2}(y-z)^2\frac{1}{z} + \frac{1}{6}(y-z)^3(-\frac{1}{z^2}) + \frac{1}{24}(y-z)^4\frac{2}{y^3} \\ &\geq z\ln z + (y-z)(1+\ln z) + \frac{1}{2}(y-z)^2\frac{1}{z} + \frac{1}{6}(y-z)^3(-\frac{1}{z^2})\end{aligned}$$

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

$$\begin{aligned} & E_\theta(T \ln T) \\ \geq\ & n\theta \ln n\theta + \frac{1}{2}\mathrm{Var}_\theta(T) \cdot \frac{1}{n\theta} + \frac{1}{6}E_\theta(T - n\theta)^3 \cdot (-\frac{1}{(n\theta)^2}) \\ =\ & n\theta \ln n\theta + \frac{1-\theta}{2} + \frac{1}{6}E_\theta(T - n\theta)^3 \cdot \frac{-1}{(n\theta)^2} \\ \geq\ & n\theta \ln n\theta + \frac{1-\theta}{2} - \frac{1}{48n\theta}. \end{aligned}$$

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 $y \geq 0$, $z > 0$.

$$y \ln y \leq z \ln z + (y - z)(1 + \ln z) + \frac{(y-z)^2}{2z} - \frac{(y-z)^3}{6z^2} + \frac{(y-z)^4}{3z^3}. \quad (6.12)$$

To prove (6.12), we substitute y with $(t+1)z$, then it reduces to show that for all $t \geq -1$.

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

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

Now replace y with $T \sim$Binomial(n, θ) and z with $n\theta$ in (6.12) and take expectation to get

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

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$(1/2, \ldots, 1/2)$ prior. See [29], [15] and [35]. (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 1.4 *(A uniform upper bound for $D(p_\theta^n||m_n^*)$):* There is a constant C_k such that for all $\theta \in S_k'$, $n \geq 1$, we have

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

Moreover, for all sequences X^n,

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

Proof. We still use e as the logarithm base in the proof. Let $\hat{\theta}$ be the maximum likelihood estimator of θ, that is, $\theta_i = T_i/n$ for $i = 1, \ldots, k$ where $T_i = \sum 1_{\{X_i = \{a_i\}\}}$, then

$$\begin{aligned}
\ln \frac{p_\theta^n(X^n)}{m_n^*(X^n)} &\leq \ln \frac{p_{\hat{\theta}}^n(X^n)}{m_n^*(X^n)} \\
&= \ln \frac{\prod_{i=1}^k \left(\frac{T_i}{n}\right)^{T_i}}{D_k(T_1 + \frac{1}{2}, \ldots, T_k + \frac{1}{2}) \,/\, D_k(\frac{1}{2}, \ldots, \frac{1}{2})} \\
&= \sum_{i=1}^n T_i \ln T_i - n\ln n - \ln \frac{\prod_{i=1}^k \Gamma(T_i + \frac{1}{2})}{\Gamma(n + \frac{k}{2})} + \ln \frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2})} \quad (6.13)
\end{aligned}$$

By Stirling's formula,

$$\ln \frac{\prod_{i=1}^k \Gamma(T_i + \frac{1}{2})}{\Gamma(n + \frac{k}{2})} = \frac{k-1}{2}\ln 2\pi + \sum_{i=1}^k T_i \ln(T_i + \frac{1}{2}) +$$

$$-(n+\frac{k-1}{2})\ln(n+\frac{k}{2}) - \sum_{i=1}^{k} \ln \frac{1+r_i}{1+r_0}$$

$$\geq \sum_{i=1}^{k} T_i \ln T_i + -(n+\frac{k-1}{2})\ln n - \text{Constant}(k)$$

Incorporation of the above inequality in (6.13) yields

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

The following Lemma is verified by standard decision theory.

Lemma 1.5 *(Maximin procedure is minimax):* Under relative entropy loss, if the game has a value, if there is a minimax procedure and if there is 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 [11, 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

$$\int D(P_\theta||Q^*)W^*(d\theta) = \inf_Q \int D(P_\theta||Q)W^*(d\theta). \tag{6.14}$$

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. 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 [24]. 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 of 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. [10, p.310], [21, p.96], [23, p.96] or [4]. That is, for any prior W (even a continuous prior) there exist $\theta_1, ..., \theta_J$ and $(w_1, ..., w_J) \in S_J$ 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 $T_1, ..., T_k$ 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 $\sum_{i=1}^{J} w_i D(p_{\theta_i}^n || \sum_{j=1}^{J} w_j p_{\theta_j}^n)$ is a continuous function of $(\theta_1, ..., \theta_J, w_1, ..., w_J)$ in the compact set $\Theta^J \times S_J$ and hence there exists a point $(\theta_1^*, ..., \theta_J^*, w_1^*, ..., w_J^*)$ 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 $\mathcal{X}$ 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.

Lemma 2.1 *(A uniform bound for log-ratio of maximum likelihood and Jeffreys' mixture) Suppose* $p(x^n|\theta_1, ..., \theta_m) = \theta_1^{T_1} \cdot ... \cdot \theta_m^{T_m}$, *where* T_i*'s are the counts of the* i*th symbol in alphabet, and* $m_J(x^n)$ *is Jeffreys' mixture, i.e.,* $m_J(x^n) = \int_S p(x^n|\theta_1, ..., \theta_m) \cdot \theta_1^{-1/2} \cdot ... \cdot \theta_m^{-1/2} d\theta_1 ... d\theta_{m-1}$, *where* $S = \left\{ (\theta_1, ..., \theta_{m-1}) : \theta_i \geq 0, \sum_{i=1}^{m-1} \theta_i \leq 1 \right\}$. *Then for all* x^n, *we have*

$$\log \frac{p(x^n|\hat{\theta})}{m_J(x^n)} = \frac{m-1}{2} \log \frac{n}{2\pi} + C_m + R_n, \tag{6.15}$$

where

$$C_m = \log \frac{\Gamma(\frac{1}{2})^m}{\Gamma(\frac{m}{2})}$$

and

$$0 < R_n \leq \frac{m^2}{4n} + \frac{m}{4 \min(T_1, ..., T_m) + 2}. \tag{6.16}$$

In particular,

$$\log \frac{p(x^n|\hat{\theta})}{m_J(x^n)} \leq \frac{m-1}{2} \log \frac{n}{2\pi} + C_m + \frac{m^2}{4n} + \frac{m}{2}. \tag{6.17}$$

Note: Equation (6.16) shows that we have an accurate characterization of regret in the interior of the relative frequency simplex. On the full simplex the bound in (6.17) is a somewhat larger (as it must be since the regret at each vertex of the relative frequency simplex, corresponding to a constant sequence, is higher in than interior, see Lemma 3). Similar bounds for Jeffreys' mixture in the $m = 2$ case are in Freund [19]. We use inequality (6.17) with a modification of Jeffreys' prior on a reduced dimension simplex in the proof of the main theorem.

Proof. We leave the lower bound proof to Lemma 2 and only prove the upper bound here.

By Stirling's formula for real-valued $x > 0$ (see [42, pp. 253])

$$\Gamma(x) = x^{x-1/2}e^{-x}\sqrt{2\pi}e^{s/(12x)}, \tag{6.18}$$

where the remainder $s = s(x)$ satisfies $0 < s < 1/(12x)$. Thus Jeffreys' mixture $m_J(x^n)$ can be approximated as the following.

$$\begin{aligned} m_J(x^n) &= D_k(T_1 + \frac{1}{2}, \ldots, T_k + \frac{1}{2})/D_k(\frac{1}{2}, \ldots, \frac{1}{2}) \\ &= \frac{\prod_{i=1}^k \Gamma(T_i + \frac{1}{2})}{\Gamma(n + \frac{k}{2})} / \frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2})} \\ &= \frac{\prod_{i=1}^k (\sqrt{2\pi}(T_i + \frac{1}{2})^{T_i})}{\sqrt{2\pi}(n + \frac{k}{2})^{n+(k-1)/2}} \frac{\prod_{i=1}^k \exp(s_i)}{\exp(s_n)} / \frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2})} \end{aligned}$$

where the remainder $s_i = s(T_i + 1/2)$ and $s_n = s(n + 1/2)$ are bounded by $1/(12T_i + 6)$ and $1/(12n + 6)$, respectively. Hence

$$\log \frac{p(x^n|\hat{\theta})}{m_J(x^n)} \tag{6.19}$$

$$= \log \frac{\Gamma(\frac{1}{2})^m}{\Gamma(\frac{m}{2})} + \frac{m-1}{2} \log \frac{n}{2\pi} + \left(s_n - \sum_{i=1}^{m} s_i \right) \log e$$
$$+ \left(\frac{m-1}{2} \log(1 + \frac{m}{2n}) + \log(1 + \frac{m}{2n}) - \sum_{i=1}^{m} T_i \log(1 + \frac{1}{2T_i}) \right). \quad (6.20)$$

where collectively the remainder term from the Stirling's approximation satisfies

$$s_n - \sum_{i=1}^{m} s_i < \frac{1}{12n+6}. \quad (6.21)$$

Now we handle the additional remainder term in (6.20). We use the following inequality

$$\frac{1}{2} - \frac{1}{4(x+1/2)} \leq x \log \left(1 + \frac{1}{2x} \right) \leq \frac{1}{2}, \text{ for } x \geq 0 \quad (6.22)$$

to get that

$$\frac{m-1}{2} \log(1 + \frac{m}{2n}) + n \log \left(1 + \frac{m}{2n} \right) - \sum_{i=1}^{m} T_i \log \left(1 + \frac{1}{2T_i} \right)$$
$$\leq \frac{m(m-1)}{4n} + \frac{m}{2} - \sum_{i=1}^{m} T_{\min} \log \left(1 + \frac{1}{2T_{\min}} \right)$$
$$\leq \frac{m^2}{4n} + \frac{m}{4T_{\min} + 2}. \quad (6.23)$$

where $T_{\min} = \min(T_1, \ldots, T_m)$. Summation of (6.21) and (6.23) yields the upperbound in (6.16). Thus continuing from (6.19) and (6.20) we obtain that

$$\log \frac{p(x^n|\hat{\theta})}{m_J(x^n)} = \log \frac{\Gamma(\frac{1}{2})^m}{\Gamma(\frac{m}{2})} + \frac{m-1}{2} \log \frac{n}{2\pi} + R_n$$

with R_n satisfying the upper bound in (6.16) (the lower bound $R_n > 0$ is shown in Lemma 2). Inequality (6.17) follows using $T_{\min} \geq 0$. ■

Lemma 2.2 *(A uniform lower bound for log-ratio of maximum likelihood and Jeffreys' mixture) Using the same notation as in Lemma 1, we have* $R_n > 0$.

Moreover $\log p(x^n|\hat{\theta})/m_J(x^n) - \frac{m-1}{2}\log\frac{n}{2\pi}$ *is a decreasing function of the counts* $T_1, \ldots, T_m$.

Proof. Define

$$f(T_1, \ldots, T_m) = \frac{p(x^n|\hat{\theta})}{m_J(x^n) n^{(m-1)/2}},$$

where $n = \sum_{i=1}^m T_i$. Once we show that f is decreasing in each variable, it will follow that

$$\begin{aligned} f(T_1, \ldots, T_m) &> f(T_{\max}, \ldots, T_{\max}) \\ &\geq \lim_{L\to\infty} f(L, \ldots, L) \\ &= \frac{\Gamma(\frac{1}{2})^m}{\Gamma(\frac{m}{2})}/(2\pi)^{\frac{m-1}{2}}. \end{aligned} \tag{6.24}$$

where $T_{\max} = \max(T_1, \ldots, T_m)$, from which it follows that $R_n > 0$.

Now we show that $f(T_1 + 1, T_2, \ldots, T_m) < f(T_1, T_2, \ldots, T_m)$. We have

$$\begin{aligned} f(T_1, T_2, \ldots, T_m) &= \frac{\{\Gamma(\frac{1}{2})^m/\Gamma(m/2)\} \cdot (\prod_{i=1}^m T_i^{T_i})/n^n}{\{(\prod_{i=1}^m \Gamma(T_i + \frac{1}{2}))/\Gamma(n + \frac{m}{2})\} \cdot n^{\frac{m-1}{2}}} \\ &= f(T_1 + 1, T_2 \ldots, T_m) \frac{(T_1 + \frac{1}{2})T_1^{T_1}}{(1 + T_1)^{1+T_1}} \frac{(n+1)^{n+1+\frac{m-1}{2}}}{(n + \frac{m}{2}) \cdot n^{n+\frac{m-1}{2}}}. \end{aligned} \tag{6.25}$$

The factor $(T_1 + \frac{1}{2})T_1^{T_1}/(1 + T_1)^{1+T_1}$ is decreasing in T_1 as seen by examining its logarithm. Indeed $g(t) = \log(t + \frac{1}{2}) + t\log t - (t+1)\log(t+1)$ has derivative $g'(t) = (t + \frac{1}{2})^{-1} + \log(t/(t+1))$, which (upon setting $t + \frac{1}{2} = \frac{1}{2u}$) equals $2u + \log\frac{1-u}{1+u}$, which is negative by examination of the Taylor expansion of $\log(1+u)$. Consequently, replacing T_1 with n in this factor, we obtain

$$\frac{(T_1 + \frac{1}{2})T_1^{T_1}}{(1 + T_1)^{1+T_1}} \frac{(n+1)^{n+1+\frac{m-1}{2}}}{(n + \frac{m}{2}) \cdot n^{n+\frac{m-1}{2}}} \geq \frac{(n + \frac{1}{2})n^n}{(1+n)^{1+n}} \frac{(n+1)^{n+1+\frac{m-1}{2}}}{(n + \frac{m}{2}) \cdot n^{n+\frac{m-1}{2}}}$$

$$= \frac{n+\frac{1}{2}}{n+\frac{m}{2}}(1+\frac{1}{n})^{\frac{m-1}{2}}$$

$$> 1, \qquad (6.26)$$

where (6.26) is equivalent to $(n+\frac{1}{2})^2(1+\frac{1}{n})^{m-1} > (n+\frac{m}{2})^2$, which is verified using the binomial expansion of $(1+\frac{1}{n})^{m-1}$. Recalling (6.25), we have shown that $f(T_1, T_2, ..., T_m) > f(T_1+1, T_2, ..., T_m)$, so it is decreasing in T_1. The same arguments show that f is decreasing in each of the counts.

Finally the limit of $f(L, ..., L)$ as $L \to \infty$ is obtained from

$$f(L, ..., L) = \frac{(1/m)^{mL}}{\{\Gamma(L+\frac{1}{2})/\Gamma(mL+\frac{m}{2})\}\{\Gamma(\frac{m}{2})/\Gamma(\frac{1}{2})^m\}(mL)^{\frac{m-1}{2}}}$$

and then using Stirling's approximation. ■

Note: A similar monotonicity argument is given [43] for the $m = 2$ case.

Lemma 2.3 *(Asymptotic regret on vertex points) At the vertices of the frequency composition simplex (such as $T_1 = n$, and $T_i = 0$ for $i = 2, ..., m$), the regret of the Jeffreys' mixture is higher than the asymptotic regret in the interior.*

Proof. On the vertex $(n, 0, ..., 0)$ we have

$$\begin{aligned}
\log \frac{p(x^n|\hat{\theta}_1)}{m_J(x^n)} &= \log \frac{1}{D_k(n+\frac{1}{2}, \frac{1}{2}, ..., \frac{1}{2})/D_k(\frac{1}{2}, ..., \frac{1}{2})} \\
&= \log \frac{\Gamma(\frac{1}{2})^m}{\Gamma(\frac{m}{2})} - \log \frac{\Gamma(n+\frac{1}{2})\Gamma(\frac{1}{2})^{m-1}}{\Gamma(n+\frac{m}{2})} \\
&= \log \frac{\Gamma(\frac{1}{2})^m}{\Gamma(\frac{m}{2})} + \frac{m-1}{2}\log\frac{n}{\pi} + o(1),
\end{aligned}$$

see also Suzuki [37] and Freund [19]. The asymptotic regret for interior point is

$$\log \frac{\Gamma(\frac{1}{2})^m}{\Gamma(\frac{m}{2})} + \frac{m-1}{2}\log\frac{n}{2\pi} + o(1),$$

(in agreement with $r_n = \log c_n$). Thus the regret on the vertex is larger by the amount $\frac{m-1}{2}\log 2$, asymptotically. ■

Lemma 2.4 *(Regret incurred by other Dirichlet mixtures) Suppose that* $\alpha < 1/2$ *and let* $m_\alpha(x^n) = D_m(T_1+\alpha,\ldots,T_m+\alpha)/D_m(\alpha,\ldots,\alpha)$. *Suppose* $n \geq n$. *If* $T_i < n^p$ *for some* $i \leq m$ *and some* $p < 1$, *then*

$$\log \frac{p(x^n|\hat{\theta})}{m_\alpha(x^n)} \leq \left(\frac{m-1}{2} - (\frac{1}{2}-\alpha)(1-p)\right)\log n + K_m \log\frac{1}{\alpha},$$

where K_m is a constant depending only on m.

Proof. Without loss of generality we assume that $T_1 < n^p$. Stirling's formula gives the following expansion

$$m_\alpha(x^n) = \frac{\prod_{i=1}^m(\sqrt{2\pi}(T_i+\alpha)^{T_i+\alpha-1/2})}{\sqrt{2\pi}(n+m\alpha)^{n+m\alpha-1/2}\cdot D_m(\alpha,\ldots,\alpha)}\cdot e^R,$$

where $R = \sum_{i=1}^m s(T_i+\alpha) - s(n+m\alpha)$ is the residual from the Stirling approximation and thus satisfies

$$\begin{aligned} R &\geq -\frac{1}{12(n+m\alpha)} \\ &\geq -\frac{1}{12n}. \end{aligned} \tag{6.27}$$

Take the logarithm to get

$$\begin{aligned} &\log\frac{p(x^n|\hat{\theta})}{m_\alpha(x^n)} \\ \leq\ & -\frac{m-1}{2}\log(2\pi) - \sum T_i\log(1+\frac{\alpha}{T_i}) + (\frac{1}{2}-\alpha)\sum_{i=1}^m \log(T_i+\alpha) \qquad (6.28) \\ & + n\log(1+\frac{m\alpha}{n}) + (m\alpha-\frac{1}{2})\log(n+m\alpha) + \log D_m(\alpha,\ldots,\alpha) - R\log e. \end{aligned}$$

(6.29)

In this bound we use

$$\begin{aligned}\sum \log(T_i+\alpha) &= \log(T_1+\alpha)+\sum_{i=2}^{m}\log(T_i+\alpha)\\ &\le \log(T_1+\alpha)+(m-1)\log\left(\frac{n-T_1}{m-1}+\alpha\right)\\ &\le p\log n+\alpha+(m-1)\log n+\frac{(m-1)^2}{2n}.\end{aligned}$$

Furthermore, we use $\sum T_i \log(1+\alpha/T_i) > 0$ and $\log(1+x) \le x$ to simplify some terms in (6.29). Collectively these yield an upperbound for $\log p(x^n|\hat{\theta})/m_\alpha(x^n)$.

$$\log\frac{p(x^n|\hat{\theta})}{m_\alpha(x^n)} \le \left(\frac{m-1}{2}-(\frac{1}{2}-\alpha)(1-p)\right)\log n+b. \tag{6.30}$$

where the constant b satisfies

$$b \le \left(\frac{(m-1)^2}{4n}+\frac{1}{12n}+\frac{m(m+1)}{4}\right)\log e+\log D_m(\alpha,\ldots,\alpha).$$

By Stirling's approximation,

$$\begin{aligned}D_m(\alpha,\ldots,\alpha) &= \frac{\Gamma(\alpha)^m}{\Gamma(m\alpha)}\\ &\le (2\pi)^{(m-1)/2}\alpha^{1/2-m/2}m^{-m\alpha+1/2}.\end{aligned}$$

hence there exists some K_m such that

$$b \le K_m \log\frac{1}{\alpha}.$$

This completes the proof. ■

Bibliography

[1] A. R. Barron, "Logically smooth density estimation," Ph.D. dissertation, Stanford, 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] J. Berger, J. M. Bernardo and M. Mendoza, "On priors that maximize expected information," in *Recent Developments in Statistics and their applications*, Freedom Academy Publishing, Seoul, 1989.

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

[6] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, pp. 453–471, May 1990.

[7] B. S. Clarke and A. R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk." *J. Statistic. Planning and Inference*, vol. 41, pp.37-60, Aug. 1994.

[8] T. M. Cover and E. Ordentlich, "Universal portfolios with side information," *IEEE Trans. Inform. Theory*, vol. 42, pp.348-363, March 1996.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc. 1991.

[10] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*, Academic Press: Orlando, Florida, 1986.

[11] 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.

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

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

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

[15] L. D. Davisson, R. J. McEliece, M. B. Pursley and M. S. Wallace, "Efficient Universal Noiseless Source Codes," *IEEE Trans. Inform. Theory*, vol. 27, pp.269-279, 1981.

[16] M. Feder, N. Merhav and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1258-1268, July 1992.

[17] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I, third edition, John Wiley & Sons, New York, 1968.

[18] D. P. Foster, "Prediction in the worst case," *The Annals of Statistics*, vol. 19, no. 2, pp. 1084-1090, 1991.

[19] Y. Freund, "Predicting a binary sequence almost as well as the optimal biased coin," *Proceedings of the 9th Annual Workshop on Computational Learning Theory*, pp. 89-98, Morgan Kaufmann, 1996.

[20] T. S. Ferguson, *Mathematical Statistics: A Decision-theoretic Approach*, Academic Press: New York, 1967.

[21] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley and sons, 1968.

[22] R. G. Gallager, "Supplementary Notes #3, Notes on Universal Coding," MIT course 6.441 notes, March 1974.

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

[24] D. Haussler, "A general minimax result for relative entropy," preprint, 1995.

[25] D. Haussler and A. R. Barron, "How well do Bayes methods work for on-line prediction of $\{\pm 1\}$ values," *Proceedings 1992 NEC Symposium on Computation and Cognition*, Chapter 4.

[26] D. Haussler, J. Kivinen and M. K. Warmuth, "Tight worst-case loss bounds for prediction with expert advice," to appear in *IEEE Trans. Inform. Theory*.

[27] D. Haussler and M. Opper, "General bounds on the mutual information between a parameter and n conditionally independent observation," preprint, 1995.

[28] T. Kløve, "Bounds on the worst case probability of undetected error," *IEEE Trans. Inform. Theory*, vol. 41, pp. 298-300, Jan. 1995.

[29] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. 27, pp. 199-207, March 1981.

[30] E. Ordentlich and T. M. Cover, "The cost of achieving the best portfolio in hindsight," preprint.

[31] J. Rissanen, "Universal coding, information, prediction and estimation," *IEEE Trans. Inform. Theory*, vol. 30, pp. 629-636, 1984.

[32] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, pp. 1080-1100, Sep. 1986.

[33] J. Rissanen, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inform. Theory*, vol. 32, pp. 526-532, July 1986.

[34] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 40, pp. 40-47, Jan. 1996.

[35] Yu. M. Shtar'kov, "Coding of Discrete Sources with Unknown Statistics," *Topics in Information Theory*, (ed. I. Csiszár and P. Elias.) Coll. Math. Soc. J. Boyai, no. 16, North Holland, Amsterdam: pp. 559-574, 1977.

[36] Yu. M. Shtar'kov, "Universal sequential coding of single messages," *Problems of Information Transmission*, Vol. 23, pp. 3-17, July 1988.

[37] J. Suzuki, "Some notes on universal noiseless coding," *IEICE Trans. Fundamentals*, vol. E78-A, No. 12, December 1995.

[38] W. Szpankowski, "On asymptotics of certain sums arising in coding theory," *IEEE Trans. Inform. Theory*, vol. 41, No. 6, pp. 2087-2090, November 1995.

[39] V. Vovk, "Aggregating strategies," *Proc. 3rd Annual Workshop on Computer Learning Theory*, Morgan Kaufmann, pp. 371-383, 1990.

[40] M. J. Weinberger, N. Merhav and M. Feder, "Optimal sequential probability assignment for individual sequences," *IEEE Trans. Inform. Theory*, vol. 40, No. 2, pp. 384-396, March 1994.

[41] M. J. Weinberger, N. Rissanen and M. Feder, "A universal finite memory sources," *IEEE Trans. Inform. Theory*, vol. 41, No. 3, pp. 643-652, May 1995.

[42] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis* (4th edition). Cambridge University Press, 1963.

[43] F. M. Willems, Y. M. Shtarkov and T. J. Tjalkens, "The context tree weighting method: Basic properties," *IEEE Trans. Inform. Theory*, vol. 41, No. 3, pp. 653-664, May 1995.

[44] Q. Xie and A. R. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Trans. Inform. Theory*, vol. 43, NO. 2, March 1997.

[45] Z. Zhang, "Discrete Noninformative Priors", Ph.D. dissertation, Yale University, December 1994.