Sequential Universal Modeling for Non-Binary Sequences with Constrained Distributions

Michael Drmota
Discrete Mathematics and Geometry
TU Wien
A-1040 Wien, Austria
Email: michael.drmota@tuwien.ac.at

Gil I. Shamir
Google
6425 Penn Ave. Suite 700
Pittsburgh, PA 15206-4037, USA
Email: gshamir@google.com

Wojciech Szpankowski
Department of Computer Science
Purdue University
West Lafayette, IN, USA
Email: spa@cs.purdue.edu

***Abstract*—Sequential probability assignment and universal compression go hand in hand. We propose sequential probability assignment for non-binary (and large alphabet) sequences with empirical distributions whose parameters are known to be bounded within a limited interval. Sequential probability assignment algorithms are essential in many applications that require fast and accurate estimation of the maximizing sequence probability. These applications include learning, regression, channel estimation and decoding, prediction, and universal compression. On the other hand, constrained distributions introduce interesting theoretical twists that must be overcome in order to present efficient sequential algorithms. Here, we focus on universal compression for memoryless sources, and present precise analysis for the maximal minimax and the average minimax for constrained distributions. We show that our sequential algorithm based on modified Krichevsky-Trofimov (KT) estimator is asymptotically optimal up to $O(1)$ for both maximal and average redundancies. This paper follows and addresses the challenge presented in [10] that suggested "results for the binary case lay the foundation to studying larger alphabets".**

I. Introduction

Universal coding and universal modeling (probability assignments) are two driving forces of information theory, model selection, and statistical inference. In universal coding one is to construct a code for data sequences generated by an unknown source from a known family such that, as the length of the sequence increases, the average code length approaches the entropy of whatever processes in the family has generated the data. In seminal works of Davisson [2], Rissanen [6], Krichevsky and Trofimov [4], and Shtarkov [7] it was shown how to construct such codes for finite alphabet sources. Universal codes are often characterized by the average *minimax* redundancy which is the excess over the entropy of the *best* code from a class of decodable codes for the worst process in the family.

As pointed out by Rissanen [6], over years universal coding evolved into *universal modeling* where the purpose is no longer restricted to just coding but rather to learn optimal models [6]. The central question of interest in universal modeling seems to be in universal codes achievable for *individual* sequences. The burning question is how to measure it. The *worst case* minimax redundancy became handy since it measures the worst case excess of the best code maximized over the processes in the family. Unfortunately, low-complexity universal codes that are optimal for the worst case minimax are not easily implementable. Therefore, we design a sequential algorithm based on the KT-estimator that is asymptotically optimal on average (i.e., for the average minimax redundancy), and show that both redundancies differ by a small constant.

In this paper we focus on universal compression and probability assignment/learning for a class of memoryless sources with constrained distributions. Let us start with some definitions and notation. We define a code $C_n : \mathcal{A}^n \to \{0,1\}^*$ as a mapping from the set $\mathcal{A}^n$ of all sequences $x^n = (x_1, \ldots, x_n)$ of length n over the finite alphabet $\mathcal{A} = \{1, \ldots, m\}$ of size m to the set $\{0,1\}^*$ of all binary sequences. Given a probabilistic source model, we let $P(x^n)$ be the probability of the message x^n; given a code C_n, we let $L(C_n, x^n)$ be the code length for x^n. However, in practice the probability distribution (i.e., source) P is unknown, and one looks for *universal codes* for which the redundancy is $o(n)$ for all $P \in \mathcal{S}$ where $\mathcal{S}$ is a class of source models (distributions). It is convenient to ignore the integer nature of the code length and replace it by its best distributional guess, say $Q(x^n)$. In other words, we just write $L(C_n, X_1^n) = -\log Q(x^n)$ and use it throughout the paper. The question is how well Q approximates P within the class $\mathcal{S}$. Minimax redundancy enters. Usually, we consider two types of minimax redundancy, namely *average* and *maximal* or *worst case* defined, respectively, as

$$\overline{R}_n(\mathcal{S}) = \min_Q \sup_{P \in \mathcal{S}} \mathbf{E}[\log P(X^n)/Q(x^n)], \tag{1}$$

$$R_n^*(\mathcal{S}) = \min_Q \sup_{P \in \mathcal{S}} \mathbf{max}_{x^n}[\log P(X^n)/Q(x^n)]. \tag{2}$$

In this paper we analyze precisely both redundancies for *memoryless sources* over m-ary alphabet $\mathcal{A} = \{1, \ldots, m\}$ with *restricted symbol probability* θ_i, that is, we assume that $\boldsymbol{\theta} \in \mathcal{S}$, where $\mathcal{S}$ is a proper subset of

$$\Theta = \{\boldsymbol{\theta} : \theta_i \geq 0 \ (1 \leq i \leq m), \ \theta_1 + \cdots + \theta_m = 1\}.$$

We will assume that $\mathcal{S}$ is a convex polytope. As a special case we have the *interval* restriction $0 \leq a_i \leq \theta_i \leq b_i \leq 1$ for $i = 1, \ldots m-1$, where $\sum_i b_i \leq 1$. Here, we present a sequential algorithm that estimates asymptotically optimal probability $P(x^n)$ for all x^n. It turns out that restricting the set of parameters is important from practical point of view and

at the same time introduces new interesting theoretical twists that we explore in this paper. We first prove in Theorem 1 that (for fixed m that can still be large)

$$\overline{R}_n(\mathcal{S}) = R_n^*(\mathcal{S}) + O(1) = \frac{m-1}{2}\log(n) + O(1)$$

where the constant implied by the O-term depends on m and on the constrains. Second we provide in Theorem 2 precise asymptotics for $\overline{R}_n(\Theta)$ and $R_n^*(\Theta)$ if $m = o(n)$. While these results are not new [5], [8], [9], [12], we derive *precise* asymptotics up to $O(m/\sqrt{n})$ term in a uniform manner that can be used to extend our analysis to the constrained case in this regime. Finally, we present in Theorem 3 a sequential add-1/2 KT-like estimator to compute $P(x_{n+1}|x^n)$ for the constrained distributions that is asymptotically optimal up to a constant for both the maximal and average redundancy.

II. Main Results

In this section we present our main results including asymptotically optimal probability estimation for the class $\mathcal{S} \subset \Theta$ of memoryless sources with constrained distributions.

We start with the *worst case redundancy* defined in (2). We recall that the empirical distribution is of the following form

$$P(x^n) = \prod_{i=1}^m \theta_i^{k_i}, \qquad \theta_i \geq 0, \ \sum_{i=1}^m \theta_i = 1,$$

where k_i is the number of symbol $i \in \mathcal{A}$ in the sequence x^n. The probabilities θ_i are unknown to us except that we restrict them to the subset $\mathcal{S} \subseteq \Theta$. Following Shtarkov [7] and [3] we can re-write the worst case redundancy for $\mathcal{S}$, by noting that $\max$ and $\sup$ commute, as

$$\begin{aligned} R_n^*(\mathcal{S}) &= \min_Q \sup_{P\in\mathcal{S}} \max_{x_1^n} (-\log Q(x_1^n) + \log P(x_1^n)) \\ &= \min_Q \max_{x_1^n} [-\log Q(x_1^n) + \sup_{P\in\mathcal{S}} \log P(x_1^n)] \\ &= \min_Q \max_{x_1^n} [\log Q^{-1}(x_1^n) + \log P^*(x_1^n) + \log \sum_{z_1^n} \sup_P P(z_1^n)] \\ &= \log \sum_{x_1^n} \sup_P P(x_1^n) =: \log S_n \end{aligned}$$

where

$$P^*(x_1^n) := \frac{\sup_{P\in\mathcal{S}} P(x_1^n)}{\sum_{z_1^n} \sup_P P(z_1^n)} \tag{3}$$

is the *maximum-likelihood distribution* and we choose $Q(x_1^n) = P^*(x_1^n)$.

To estimate the sum $S_n = \sum_{x_1^n} \sup_P P(x_1^n)$ we need first to find $\sup \prod_{i=1}^m \theta_i^{k_i}$ when $\boldsymbol{\theta} \in \mathcal{S}$. For the unrestricted case ($\mathcal{S} = \Theta$) we know that the optimal $\theta_i = k_i/n$. The situation is more complicated in the constrained case. For example, if we assume an interval restriction $a_i \leq \theta_i \leq b_i$, $i = 1, \ldots, m-1$, then for $k_i < na_i$ or $k_i > nb_i$ the optimal θ_i may be a_i or b_i, respectively. Fortunately, we are able to prove that the main contribution to S_n comes from those $\mathbf{k} = (k_1, \ldots, k_m)$ for which $\mathbf{k}/n \in \mathcal{S}$. So we are led to analyze the following sum

$$S_n^{(\mathcal{S})} = \sum_{\mathbf{k}\in n\mathcal{S}} \binom{n}{k_1, \cdots k_m} \prod_{i=1}^m \left(\frac{k_i}{n}\right)^{k_i}$$

which is of order $n^{\frac{m-1}{2}}$. The contribution of the remaining terms only is of order $O(n^{\frac{m-2}{2}})$.

We need to introduce one more notation element. Let us define the Dirichlet density as

$$\mathrm{Dir}(x_1, \ldots, x_m; \alpha_1, \ldots \alpha_m) = \frac{1}{B(\alpha_1, \ldots, \alpha_m)} \prod_{i=1}^m x_i^{\alpha_i - 1},$$

where $\sum_{i=1}^m x_i = 1$ and

$$B(\alpha_1, \ldots, \alpha_m) = \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_m)}{\Gamma(\alpha_1 + \cdots \alpha_m)}$$

is the beta function. We shall write $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_m)$ and $\mathbf{x} = (x_1, \ldots x_m)$ with $\sum_{i=1}^m x_i = 1$. Finally, we set for $\mathcal{S} \subset \Theta$

$$\mathrm{Dir}(\mathcal{S}; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \int_{\mathcal{S}} \mathbf{x}^{\boldsymbol{\alpha}-1}\, d\mathbf{x}.$$

It is our goal to present a sequential low-complexity algorithm for the probability assignment, that is, an iterative procedure to compute $P(x_{n+1}|x^n)$. Unfortunately, the maximum-likelihood distribution (3) is not well suited for it. To find one, we switch to the *average* minimax redundancy (1) and we re-cast in the Bayesian framework.

Let $\mathcal{S} \subseteq \Theta$. Then the average minimax problem is then

$$\overline{R}_n(\mathcal{S}) = \inf_Q \sup_{\theta\in\mathcal{S}} D_n(P^\theta \| Q)$$

where $D(p\|Q)$ is the Kullback-Leibler divergence. In the Bayesian framework, one assumes that the parameter θ is generated by the density $w(\theta)$ and the mixture $M_n^w(x^n)$ is

$$M_n^w(x^n) = \int P^\theta(x_1^n) w(d\theta).$$

Observe now

$$\begin{aligned} \inf_Q \mathbf{E}_w[D_n(P^\theta\|Q)] &= \inf_Q \int_{\mathcal{S}} D_n(P^\theta\|Q) dw(\theta) \\ &= \int_{\mathcal{S}} D_n(P^\theta\|M_n^w) dw(\theta) \end{aligned}$$

where we use the fact that $\min_Q \sum_i P_i \log 1/Q_i = \sum_i P_i \log 1/P_i$. As pointed out by Gallager, and Davisson the minimax theorem of game theory entitles us to conclude that

$$\overline{R}_n(\mathcal{S}) = \inf_Q \sup_{\theta\in\mathcal{S}} D_n(P^\theta\|Q) = \sup_w \inf_Q \mathbf{E}_w[D_n(P^\theta\|Q)]$$

leading to our final formula

$$\overline{R}_n(\mathcal{S}) = \int_{\mathcal{S}} D(P^\theta\|M_n^w) dw^*(\theta) \tag{4}$$

where $w^*(\theta)$ is the maximizing prior distribution. For the unconstrained case (i.e., $\mathcal{S} = \Theta$) Bernardo [1] proved that asymptotically the maximizing density is $\mathrm{Dir}(\boldsymbol{\theta}; \mathbf{1/2})$. For the constrained case we modified it to

$$w^*(\boldsymbol{\theta}) = \frac{1}{C(\mathcal{S}) \cdot B(\mathbf{1/2})} \frac{1}{\sqrt{\theta_1 \cdots \theta_m}} \tag{5}$$

where $C(\mathcal{S})$ defined as

$$C(\mathcal{S}) = \mathrm{Dir}(\mathcal{S}; \mathbf{1/2}) = \frac{1}{B(\mathbf{1/2})} \int_{\mathcal{S}} \frac{d\mathbf{x}}{\sqrt{x_1 \cdots x_m}} \tag{6}$$

is the probability that the Dirichlet distribution with $\alpha_i = 1/2$ falls into the subset $\mathcal{S}$.

Now, the mixture distribution $M^{w^*}(x^n)$ can be calculated as follows

$$\begin{aligned} M^{w^*}(x^n) &= \frac{1}{C(\mathcal{S}) \cdot B(\mathbf{1/2})} \int_{\mathcal{S}} \prod_{i=1}^{m} \theta_i^{k_i - 1/2} \\ &= \frac{1}{C(\mathcal{S}) \cdot B(\mathbf{1/2})} B(k_1 + 1/2, \cdots, k_m + 1/2) \\ &\cdot \frac{1}{B(k_1 + 1/2, \cdots, k_m + 1/2)} \int_{\mathcal{S}} \prod_{i=1}^{m} \theta_i^{k_i - 1/2} \\ &= \frac{1}{C(\mathcal{S}) \cdot B(\mathbf{1/2})} B(k_1 + 1/2, \cdots, k_m + 1/2) \\ &\cdot \mathrm{Dir}([\mathcal{S} : \mathbf{k} + \mathbf{1/2}). \end{aligned} \tag{7}$$

Observe that for the unconstrained case $\mathrm{Dir}(\Theta : \mathbf{k} + 1/2) = 1$.

In summary

$$D(P^\theta \| M^{w^*}) = \log\left(C(\mathcal{S}) B(\mathbf{1/2})\right) + \tag{8}$$

$$+ \sum_{\mathbf{k}} \binom{n}{\mathbf{k}} \prod_{i=1}^{m} \theta_i^{k_i} \log \frac{\prod_{i=1}^{m} \theta_i^{k_i}}{B(\mathbf{k} + \mathbf{1/2}) \mathrm{Dir}(\mathcal{S} : \mathbf{k} + \mathbf{1/2})}$$

and

$$\overline{R}_n(S) = \int_{\mathcal{S}} D(p^\theta \| M^{w^*}) dw^*(\theta).$$

We are now ready to formulate our first main result that reads as follows.

Theorem 1. *Consider a memoryless constrained source $\mathcal{S} \subset \Theta$ with fixed but arbitrarily large $m \geq 2$ where $\mathcal{S}$ is a convex polytope. Then the worst case redundancy for $\mathcal{S}$ is*

$$\begin{aligned} R_n^*(\mathcal{S}) &= \frac{m-1}{2} \log(n/2) - \log \Gamma(m/2) + \log C(\mathcal{S}) \\ &+ \frac{1}{2} \log \pi + O(1/\sqrt{n}) \end{aligned} \tag{9}$$

and the corresponding average redundancy is

$$\begin{aligned} \overline{R}_n(\mathcal{S}) &= \frac{m-1}{2} \log(n/2e) - \log \Gamma(m/2) + \log C(\mathcal{S}) \\ &+ \frac{1}{2} \log \pi + O(1/\sqrt{n}) \end{aligned} \tag{10}$$

where $C(\mathcal{S})$ is defined above in (6).

In Theorem 1 we assumed that m is fixed to avoid complications with constrains $\mathcal{S}_m$ that may depend of m. We handle it in the forthcoming paper, but here we present our results for large $m = o(n)$. While they are not completely new (see [5], [8], [9], [12]), our novel derivations will allow us to consider more general constrained sources.

Theorem 2. *Consider a memoryless unconstrained source Θ with $m = o(n)$. Then the unconstrained maximal redundancy is*

$$\begin{aligned} R_n^*(\Theta) &= \frac{m-1}{2} \log\left(\frac{e\,n}{m}\right) + \frac{1}{2}(1 - \log 2) \\ &+ O(1/m) + O(m/\sqrt{n}). \end{aligned} \tag{11}$$

and the unconstrained average redundancy becomes

$$\overline{R}_n(\Theta) = \frac{m-1}{2} \log\left(\frac{n}{m}\right) + \frac{1}{2}(1 - \log 2) \tag{12}$$

$$+ O(1/m) + O(m/\sqrt{n}). \tag{13}$$

We observe that $R_n^*(\Theta) - \overline{R}_n(\Theta) = O(m)$. This fact should be compared with a general results of [3] (see Theorem 6) where it was proved that for a large class sources

$$\widetilde{R}_n(\mathcal{S}) - \widetilde{R}_n^*(\mathcal{S}) = O(c_n(\mathcal{S}))$$

where

$$c_n(\mathcal{S}) = \sup_{P \in \mathcal{S}} \sum_{x_1^n} P(x_1^n) \lg \frac{\sup_{P \in \mathcal{S}} P(x_1^n)}{P(x_1^n)}.$$

Actually, for binary memoryless sources $c_n(\mathcal{S}) = O(1)$ and $c_n(\mathcal{S}) = O(m)$ for m-ary memoryless sources.

Now, we are ready to present our probability assignment algorithm. We start with formula (7) on the mixture $M(x^n)$. Then we observe that $M(x_{n+1}|x^n) = M(x^{n+1})/M(x^n)$. For example, if assume that x_{n+1} symbol is $i \in \mathcal{A}$. Thus

$$\begin{aligned} M(x^{n+1}) &= \frac{B(k_1 + 1/2, \cdots, k_i + 3/2, \cdots, k_m + 1/2)}{C(\mathcal{S}) \cdot B(\mathbf{1/2})} \\ &\cdot \mathrm{Dir}(\mathcal{S} : k_1 + 1/2, \cdots, k_i + 3/2, \cdots, k_m + 1/2). \end{aligned}$$

Using the functional equation of the gamma function, namely $\Gamma(x+1) = x\Gamma(x)$ allows us to write a simple sequential update algorithm that we present next.

Theorem 3. *Suppose that m is fixed and that $\mathcal{S} \subseteq \Theta$ is a convex polytope. Let $N_i(x^n)$ be the number of symbol i in x^n. Then*

$$M(x_{n+1}|x^n) = \frac{N_{x_{n+1}}(x^n) + 1/2}{n + m/2} \cdot \tag{14}$$

$$\cdot \frac{\mathrm{Dir}(\mathcal{S}; N_i(x^n) + 1/2 + 1(x_{n+1} = i),\ i = 1 \cdots m)}{\mathrm{Dir}(\mathcal{S}; N_i(x^n) + 1/2,\ i = 1 \cdots m)}.$$

Observe that for the unconstrained case $\mathrm{Dir}(\Theta; N_i(x^n) + 1/2 + 1(x_{n+1} = i),\ i = 1 \cdots m) = \mathrm{Dir}(\Theta; N_i(x^n) + 1/2,\ i = 1 \cdots m) = 1$, and then our estimation algorithm reduces to the KT-estimator, that is,

$$M(x_{n+1}|X^n) = \frac{N_{x_{n+1}}(x^n) + 1/2}{n + m/2}. \tag{15}$$

III. Analysis and Proofs

In this section we sketch proofs of our main results. We start with Theorem 2.

A. Proof of Theorem 2

By definition he have $R_n^*(\Theta) = \log S_n$, where

$$\begin{aligned} S_n &= \sum_{\mathbf{k}} \binom{n}{\mathbf{k}} \prod_{i=1}^{m} \left(\frac{k_i}{n}\right)^{k_i} \\ &= (2\pi)^{-\frac{m-1}{2}} \sum_{\mathbf{k}} \frac{\sqrt{n}}{\sqrt{k_1 \cdots k_m}} \left(1 + O\left(\sum_{i=1}^{m} \frac{1}{k_i + 1}\right)\right). \end{aligned}$$

By a standard but tedious analysis we have

$$\sum_{\mathbf{k}} \frac{1}{\sqrt{k_1 \cdots k_m}} = n^{\frac{m}{2}-1} B(\mathbf{1/2}) + O\left(m\, n^{\frac{m}{2}-\frac{3}{2}}\right) \quad (16)$$

and

$$\sum_{\mathbf{k}} \frac{1}{\sqrt{k_1 \cdots k_m}} \frac{1}{k_i+1} = O\left(B(\mathbf{1/2}) n^{\frac{m}{2}-\frac{3}{2}}\right). \quad (17)$$

Then (16) and (17) imply

$$S_n = (2\pi)^{-\frac{m-1}{2}} n^{\frac{m-1}{2}} B(\mathbf{1/2}) \left(1 + O\left(\frac{m}{\sqrt{n}}\right)\right).$$

Since

$$\log B(\mathbf{1/2}) = m \log \Gamma(1/2) - \log \Gamma(m/2)$$

we directly obtain the proposed representation (11) for $R_n^*(\Theta) = \log S_n$.

Our starting point for the analysis of $\overline{R}_n(\Theta)$ is

$$\overline{R}_n(\Theta) = \frac{1}{B(\mathbf{1/2})} \int_\Theta \sum_{\mathbf{k}} \binom{n}{\mathbf{k}} \boldsymbol{\theta}^{\mathbf{k-1/2}} \log\left(\frac{\boldsymbol{\theta}^{\mathbf{k}} B(\mathbf{1/2})}{B(\mathbf{k}+\mathbf{1/2})}\right) \quad (18)$$

where we write $\boldsymbol{\theta}^{\mathbf{k-1/2}} := \prod_i \theta_i^{k_1-1/2}$. We need to estimate different parts of the above sum. We first observe that

$$\sum_{\mathbf{k}} \binom{n}{\mathbf{k}} B(\mathbf{k}+\mathbf{1/2}) = \int_\Theta \sum_{\mathbf{k}} \binom{n}{\mathbf{k}} \boldsymbol{\theta}^{\mathbf{k-1/2}} d\boldsymbol{\theta}$$
$$= \int_\Theta \boldsymbol{\theta}^{\mathbf{-1/2}} d\boldsymbol{\theta} = B(\mathbf{1/2}).$$

More importantly we notice that

$$\int_\Theta \sum_{\mathbf{k}} \binom{n}{\mathbf{k}} \boldsymbol{\theta}^{\mathbf{k-1/2}} \log \boldsymbol{\theta}^{\mathbf{k}} d\boldsymbol{\theta} =$$

$$= \sum_{\mathbf{k}} \binom{n}{\mathbf{k}} \sum_{i=1}^m k_i \frac{\partial}{\partial k_i} B(\mathbf{k}+\mathbf{1/2})$$

and

$$\int_\Theta \sum_{\mathbf{k}} \binom{n}{\mathbf{k}} \boldsymbol{\theta}^{\mathbf{k-1/2}} \log B(\mathbf{k}+\mathbf{1/2}) =$$

$$= \sum_{\mathbf{k}} \binom{n}{\mathbf{k}} B(\mathbf{k}+\mathbf{1/2}) \log B(\mathbf{k}+1/2).$$

Thus

$$\overline{R}_n(\Theta) = \log B(\mathbf{1/2}) + \frac{1}{B(\mathbf{1/2})} \sum_{\mathbf{k}} \binom{n}{\mathbf{k}} B(\mathbf{k}+\mathbf{1/2})$$

$$\cdot \left(\sum_{i=1}^m k_i \frac{\partial}{\partial k_i} B(\mathbf{k}+\mathbf{1/2}) - \log B(\mathbf{k}+\mathbf{1/2})\right). \quad (19)$$

To deal with such sums we use the relation between the beta function, the gamma function, and the psi function [11]. For example

$$\frac{\partial}{\partial k_i} B(\mathbf{k}+\mathbf{1/2}) = \Psi(k_i+1/2) - \Psi(n+m/2)$$

where $\Psi(x) = \Gamma'(x)/\Gamma(x)$. Using these and Stirling's formula

$$\log \Gamma(x+1/2) = x \log x - x + \log \sqrt{2\pi} - \frac{1}{24x} + O(1/x^2),$$

$$\log \Gamma(x+m/2) = x \log x - x + \frac{m-1}{2} \log(x+m/2)$$
$$+ \log \sqrt{2\pi} + \left(\frac{1}{12} - \frac{m^2}{8}\right) \frac{1}{x} + O(m^3/x^2)$$

and we find

$$\sum_{i=1}^m k_i \frac{\partial}{\partial k_i} B(\mathbf{k}+\mathbf{1/2}) - \log B(\mathbf{k}+\mathbf{1/2}) = O(m^2/n)$$

$$\frac{m-1}{2} \left(\log(n+m/2) - 1 - \log(2\pi)\right) + O\left(\sum_i k_i^{-1}\right).$$

Now we obtain similarly to (17)

$$\sum_{i=1}^m \binom{n}{\mathbf{k}} \frac{B(\mathbf{k}+\mathbf{1/2})}{k_i+1} = O\left(\frac{B(\mathbf{1/2})}{\sqrt{n}}\right).$$

Summing up we arrive at

$$\overline{R}_n(\Theta) = \frac{m-1}{2} \log(n/2\pi e) + \log \frac{\Gamma^m(1/2)}{\Gamma(m/2)} + O(m/\sqrt{n}).$$

We now use Stirling's formula for $\Gamma(m/2)$ when m is large and $o(n)$.

B. Proof of Theorem 1

As above our starting point is $R_n^*(\mathcal{S}) = \log S_n$ where

$$S_n = \sum_{\mathbf{k}} \binom{n}{\mathbf{k}} \sup_{\theta \in \mathcal{S}} \prod_{i=1}^m \theta_i^{k_i}.$$

The problem is now that we have to distinguish between the case, where $\mathbf{k}/n \in \mathcal{S}$ and the case, where $\mathbf{k}/n \notin \mathcal{S}$. If $\mathbf{k}/n \in \mathcal{S}$ then we have

$$\sup_{\theta \in \mathcal{S}} \prod_{i=1}^m \theta_i^{k_i} = \prod_{i=1}^m \left(\frac{k_i}{n}\right)^{k_i}$$

as in the unconstrained case. If $\mathbf{k}/n \notin \mathcal{S}$ then we have

$$\sup_{\theta \in \mathcal{S}} \prod_{i=1}^m \theta_i^{k_i} = \prod_{i=1}^m \theta_{i,\mathrm{opt}}^{k_i}$$

where $(\theta_{i,\mathrm{opt}})$ is on the boundary of $\mathcal{S}$. For example, for $m = 2$ and $\mathcal{S} = \{(\theta, 1-\theta) : \theta \in [a,b]\}$ we have for $0 \le k_1 < an$

$$\sup_{0 \le \theta \le 1} \theta^{k_1} (1-\theta)^{n-k_1} = a^{k_1} (1-a)^{n-k_1}$$

and similar for $bn < k_1 \le n$. Hence, as above we obtain

$$S_n^{(\mathcal{S})} = \sum_{\mathbf{k}/n \in \mathcal{S}} \binom{n}{\mathbf{k}} \prod_{i=1}^m \left(\frac{k_i}{n}\right)^{k_i}$$
$$= \left(\frac{n}{2\pi}\right)^{\frac{m-1}{2}} C(\mathcal{S}) B(\mathbf{1/2}) \left(1 + O(1/\sqrt{n})\right).$$

The sum over $\mathbf{k}$ for which $\mathbf{k}/n \notin \mathcal{S}$ is more difficult to handle. But if $\mathcal{S}$ is a convex polytope we obtain after some (involved) algebra

$$S_n - S_n^{(\mathcal{S})} = O\left(n^{\frac{m}{2}-1}\right).$$

For example, if $m = 2$ and $\mathcal{S} = \{(\theta, 1-\theta) : \theta \in [a, b]\}$ then

$$\begin{aligned} S_n - S_n^{(\mathcal{S})} &= \sum_{0 \le k_1 < an} \binom{n}{k_1} a^{k_1} (1-a)^{n-k_1} \\ &\quad + \sum_{bn < k_1 \le n} \binom{n}{k_1} b^{k_1} (1-b)^{n-k_1} \\ &= 1 + O(1/\sqrt{n}) = O(1) \end{aligned}$$

Finally by using $B(\mathbf{1/2}) = \Gamma(1/2)^m/\Gamma(m/2)$ and $\Gamma(1/2) = \sqrt{\pi}$ we directly arrive at (9).

Our starting point for the average redundancy is (8), however, we rewrite it in terms of $\mathcal{S} \subseteq \Theta$ as follows

$$\overline{R}_n(\mathcal{S}) = \frac{1}{B_{\mathcal{S}}(\mathbf{1/2})} \int_{\mathcal{S}} \sum_{\mathbf{k}} \binom{n}{\mathbf{k}} \boldsymbol{\theta}^{\mathbf{k}-\mathbf{1/2}} \log \left(\frac{\boldsymbol{\theta}^{\mathbf{k}} B_{\mathcal{S}}(\mathbf{1/2})}{B_{\mathcal{S}}(\mathbf{k}+\mathbf{1/2})} \right) \tag{20}$$

where we use the short hand notation

$$B_{\mathcal{S}}(\boldsymbol{\alpha}) = \int_{\mathcal{S}} \mathbf{x}^{\boldsymbol{\alpha}-1}\, d\mathbf{x} = \mathrm{Dir}(\mathcal{S}; \boldsymbol{\alpha}) B(\boldsymbol{\alpha}).$$

As above we obtain

$$\begin{aligned} \overline{R}_n(\mathcal{S}) &= \log B_{\mathcal{S}}(\mathbf{1/2}) + \sum_{\mathbf{k}} \binom{n}{\mathbf{k}} \frac{B_{\mathcal{S}}(\mathbf{k}+\mathbf{1/2})}{B_{\mathcal{S}}(\mathbf{1/2})} \\ &\quad \cdot \left(\sum_{i=1}^m k_i \frac{\partial}{\partial k_i} B_{\mathcal{S}}(\mathbf{k}+\mathbf{1/2}) - \log B_{\mathcal{S}}(\mathbf{k}+\mathbf{1/2}) \right). \end{aligned}$$

Again we split the summation over $\mathbf{k}$ into several parts. If $\mathbf{k}/n \in \mathcal{S}^-$, where $\mathcal{S}^-$ denotes all points in the interior of $\mathcal{S}$ with distance $\ge n^{-1/2+\varepsilon}$ to the boundary (for some $\varepsilon > 0$), then the saddle point $\theta_i = k_i/n$ of the integrand $\boldsymbol{\theta}^{\mathbf{k}}$ of the integral of $B_{\mathcal{S}}(\mathbf{k}+\mathbf{1/2})$ or $\frac{\partial}{\partial k_i} B(\mathbf{k}+\mathbf{1/2})$, respectively, is contained in $\mathcal{S}^-$. Consequently we find for any $L > 0$

$$\begin{aligned} B_{\mathcal{S}}(\mathbf{k}+\mathbf{1/2}) &= B_{\mathcal{S}}(\mathbf{k}+\mathbf{1/2}) + O(n^{-L}), \\ \frac{\partial}{\partial k_i} B_{\mathcal{S}}(\mathbf{k}+\mathbf{1/2}) &= \frac{\partial}{\partial k_i} B(\mathbf{k}+\mathbf{1/2}) + O(n^{-L}). \end{aligned}$$

Hence

$$\begin{aligned} &\sum_{\mathbf{k}/n \in \mathcal{S}^-} \binom{n}{\mathbf{k}} B_{\mathcal{S}}(\mathbf{k}+\mathbf{1/2}) \\ &\cdot \left(\sum_{i=1}^m k_i \frac{\partial}{\partial k_i} B_{\mathcal{S}}(\mathbf{k}+\mathbf{1/2}) - \log B_{\mathcal{S}}(\mathbf{k}+\mathbf{1/2}) \right) \\ &= \sum_{\mathbf{k}/n \in \mathcal{S}^-} \binom{n}{\mathbf{k}} B(\mathbf{k}+\mathbf{1/2}) \\ &\cdot \left(\sum_{i=1}^m k_i \frac{\partial}{\partial k_i} B(\mathbf{k}+\mathbf{1/2}) - \log B(\mathbf{k}+\mathbf{1/2}) \right) \\ &= \sum_{\mathbf{k}/n \in \mathcal{S}^-} \binom{n}{\mathbf{k}} B(\mathbf{k}+\mathbf{1/2}) \cdot \\ &\left(\frac{m-1}{2} \log \frac{n}{2\pi e} + O\left(\sum_{i=1}^m 1/(k_i+1) \right) \right) \\ &= \left(\frac{m-1}{2} \log \frac{n}{2\pi e} + O\left(1/\sqrt{n}\right) \right) B_{\mathcal{S}}(\mathbf{1/2}). \end{aligned}$$

The other parts of the summation over $\mathbf{k}$ are more difficult to handle. As an example we indicate the difficulties that appear for $m = 2$ and $\mathcal{S} = \{(\theta, 1-\theta) : \theta \in [a, b]\}$. Suppose that $|k_1 - nb| \le n^{1/2+\varepsilon}$, that is $(k_1/n, 1-k_1/n)$ is at distance $\le n^{-1/2+\varepsilon}$ from the boundary of $\mathcal{S}$. Here we have

$$\begin{aligned} B_{\mathcal{S}}(k_1+1/2, n-k_1+1/2) &= \sqrt{\frac{2\pi}{n}} \left(\frac{k_1}{n} \right)^{k_1} \left(\frac{n-k_1}{n} \right)^{n-k_1} \\ &\quad \cdot \left(\Phi\left(\frac{nb-k_1}{\sqrt{nb(1-b)}} \right) + O(1/\sqrt{n}) \right), \end{aligned}$$

where $\Phi(u)$ denotes the normal distribution function. A similar representation holds for the derivatives $\frac{\partial}{\partial k_i} B_{\mathcal{S}}(\mathbf{k}+\mathbf{1/2})$. After some algebra it follows that

$$\begin{aligned} &\sum_{|k_1 - nb| \le n^{1/2+\varepsilon}} \binom{n}{\mathbf{k}} B_{\mathcal{S}}(\mathbf{k}+\mathbf{1/2}) \cdot \\ &\left(\sum_{i=1}^m k_i \frac{\partial}{\partial k_i} B_{\mathcal{S}}(\mathbf{k}+\mathbf{1/2}) - \log B_{\mathcal{S}}(\mathbf{k}+\mathbf{1/2}) \right) = \\ &= O(1/\sqrt{n}). \end{aligned}$$

The summation for $nb + n^{1/2+\varepsilon} < k_1 \le n$ is much easier to handle, so we skip it. This completes the proof of Theorem 1.

ACKNOWLEDGMENT

M. Drmota was supported in part by the the grant FWF Grant SFB F50-02. W. Szpankowski was supported in part by NSF Grants CCF-0939370 and NSF Grant CCF-1524312.

REFERENCES

[1] J. Bernardo, Reference Posterior Distributions for Bayesian Inference, *J. Roy. Stat. Soc. B.*, 41, 113–147, 1979.
[2] L. Davisson, Universal Noiseless Coding, *IEEE Trans. Inform. Theory*, 19, 783–795, 1973.
[3] M. Drmota and W. Szpankowski. Precise minimax redundancy and regrets. *IEEE Trans. Information Theory*, 50:2686–2707, 2004.
[4] R. Krichevsky and V. Trofimov, The Performance of Universal Coding, *IEEE Trans. Information Theory*, 27, 199-207, 1983.
[5] A. Orlitsky and N. P. Santhanam, Speaking of infinity, *IEEE Transactions on Information Theory*, 50, 2215-2230, 2004
[6] J. Rissanen, Fisher Information and Stochastic Complexity, *IEEE Trans. Information Theory*, 42, 40–47, 1996.
[7] Y. Shtarkov, Universal Sequential Coding of Single Messages, *Problems of Information Transmission*, 23, 175–186, 1987.
[8] G. Shamir, Universal lossless compression with unknown alphabets - the average case *IEEE Transactions on Information*, 52, 4915-4944, 2006.
[9] G. Shamir, On the MDL Principle for i.i.d. Sources With Large Alphabets, *IEEE Transactions on Information*, 52, 1939-1955, 2006.
[10] G. Shamir, T. Tjalkens and F. Willems, Low-Complexity Sequential Probability Estimation and Universal Compression for Binary Sequences with Constrained Distributions, *ITST*, 2008.
[11] W. Szpankowski, Average Case Analysis of Algorithms on Sequences, Wiley, New York, 2001.
[12] W. Szpankowski and M. Weinberger, Minimax Pointwise Redundancy for Memoryless Models over Large Alphabets, *IEEE Trans. Information Theory*, 58, 4094-4104, 2012.
[13] Q. Xie, A. Barron, Minimax Redundancy for the Class of Memoryless Sources, *IEEE Trans. Information Theory*, 43, 647-657, 1997.