How to Achieve Minimax Expected Kullback-Leibler Distance from an Unknown Finite Distribution *

Dietrich Braess[1], Jürgen Forster[2], Tomas Sauer[3], and Hans U. Simon[2]

[1] Fakultät für Mathematik, Ruhr-Universität Bochum, D-44780 Bochum
braess@num.ruhr-uni-bochum.de
[2] Fakultät für Mathematik, Ruhr-Universität Bochum, D-44780 Bochum
{forster,simon}@lmi.ruhr-uni-bochum.de
[3] Mathematisches Institut, Universität Giessen, D-35392 Giessen

Abstract. We consider a problem that is related to the "Universal Encoding Problem" from information theory. The basic goal is to find rules that map "partial information" about a distribution X over an m-letter alphabet into a guess $\widehat{X}$ for X such that the Kullback-Leibler divergence between X and $\widehat{X}$ is as small as possible. The cost associated with a rule is the maximal expected Kullback-Leibler divergence between X and $\widehat{X}$. First, we show that the cost associated with the well-known add-one rule equals $\ln(1+(m-1)/(n+1))$ thereby extending a result of Forster and Warmuth [3, 2] to $m \geq 3$. Second, we derive an absolute (as opposed to asymptotic) lower bound on the smallest possible cost. Technically, this is done by determining (almost exactly) the Bayes error of the add-one rule with a uniform prior (where the asymptotics for $n \to \infty$ was known before). Third, we hint to tools from approximation theory and support the conjecture that there exists a rule whose cost asymptotically matches the theoretical barrier from the lower bound.

1 Introduction

We consider a problem that is related to the "Universal Encoding Problem" from information theory. Loosely speaking, the basic goal is to find rules that map "partial information" about a distribution X over an m-letter alphabet into a guess $\widehat{X}$ for X such that the Kullback-Leibler divergence between X and $\widehat{X}$ is as small as possible. We mainly consider a variant of the problem (first proposed by Krichevskiy), where the partial information about X consists of a sample word drawn at random according to X^n. The cost associated with a rule is the maximal expected Kullback-Leibler divergence between X and $\widehat{X}$, where maximization ranges over all possible choices for X and expectation is taken over all words of length n drawn at random according to X^n.

* This work has been supported in part by the ESPRIT Working Group in Neural, Computational Learning II, NeuroCOLT2, No. 27150 and by the Deutsche Forschungsgemeinschaft Grant SI 498/4-1.

The problem whose roots trace back to Laplace has a long history and their is a fair amount of knowledge about it. The reader interested in a broader introduction is therefore referred to [6, 7, 9, 10] and the references given there.

The problem has been treated in considerable depth in the information-theoretic literature because its significance for code design. According to Shannon's Noiseless Coding Theorem [8], the minimal expected code-length for an alphabet with distribution X coincides (up to rounding) with the entropy $H(X)$ of X. If X is known, the optimal code can be efficiently designed [5]. A problem arises if X is not completely known and, given our current knowledge about X, approximated by $\widehat{X}$. Suppose we use a code that is optimal for outputs drawn according to $\widehat{X}$ in order to encode outputs drawn according to X. Then the Kullback-Leibler distance between X and $\widehat{X}$, denoted as $\mathrm{D}(X\|\widehat{X})$ and sometimes called relative entropy, measures how many additional bits we use compared to an optimal code for X. This motivates the general goal of finding rules mapping partial knowledge about X to a guess that exhibits a small Kullback-Leibler distance to X.

Before the problem was exposed to a worst case analysis (as in our paper), it had been investigated in a Bayesean setting, where X itself is assumed to be randomly chosen according to a (known) prior distribution. In this setting, the cost associated with a rule is the expected Kullback-Leibler distance between X and $\widehat{X}$, where expectation is taken over all X (distributed according to the prior) and all words of length n (distributed according to X^n). Cover [1] investigated the family of add-β rules, which make the estimated probability of a letter occurring k-times in the sample word proportional to $k + \beta$. He proved that the add-β rule is the Bayes rule for the Dirichlet prior with parameters $(\beta, \ldots, \beta)$. For instance, the add-one rule (sometimes named Laplace's law of succession) is the Bayes rule for the uniform prior; the add-half rule (sometimes named Jeffrey's rule) is the Bayes rule for Jeffrey's prior.

The problem with Bayesean analysis is that the choice of the prior often appears quite arbitrary. Krichevskiy was the first who exposed the problem to a worst case analysis (as we do in our paper). In [6] he showed that Jeffrey's rule is asymptotically best for many-letter prediction, where one does not only probabilistically predict a single letter (by guessing its probability) but many letters (by guessing the probability of a "long" word). [1] The results of Xie and Barron [9, 10] go in a similar direction. For single-letter prediction however, the situation is quite different. In [7], Krichevskiy showed the following results:

1. The cost of any rule is asymptotically bounded from below by $(m-1)/(2n) = 0.5(m-1)/n$ when n approaches infinity.
2. The best rule within the family of add-β rules is the add-β_0 rule for $\beta_0 = 0.50922\cdots$ whose cost asymptotically equals $\beta_0(m-1)/n$ (falling short of the theoretical barrier $0.5(m-1)/n$). The cost associated with any other rule from the add-β family is asymptotically strictly larger.

[1] Many-letter prediction can also be viewed as on-line prediction with cumulative relative entropy loss.

Thus, Jeffrey's rule is asymptotically optimal for many-letter prediction, but due to the results above it is not asymptotically optimal for single-letter prediction.

For $m = 2$, Forster and Warmuth [3, 2] have shown that the cost associated with the add-one rule equals $\ln(1 + 1/(n + 1)) < 1/(n + 1)$. Note that this result holds for all n, m (not just asymptotically when n approaches infinity).

Structure of the paper: Section 2 reviews the main problem in this paper in a more formal notation. In section 3, we generalize the result of Forster and Warmuth concerning the cost associated with the add-one rule to arbitrary values of m. In section 4, we provide an absolute lower bound on the smallest possible cost (which, of course, asymptotically matches the asymptotic lower bound provided by Krichevskiy). For this purpose, we determine sharp bounds of the Bayes error of the add-one rule. In section 5, we proceed in the framework of asymptotics as Krichevskiy [7] and hint to a connection with the approximation by Bernstein polynomials. By modifying an add-β rule we establish a rule whose cost seems to match asymptotically the theoretical barrier $(m - 1)/(2n)$.

A technical simplification of the problem: A simple symmetry argument shows that each rule can be "normalized" without increasing its cost such that, after normalization, it depends on the sample word w of length n only through the frequencies $(k_1, \ldots, k_m)$, where k_j denotes the number of occurrences of the j-th letter in w. In other words, one has only to count frequencies of letters and need not remember their ordering.

2 Notations and Reformulations of the Problem

Throughout this paper, $\mathcal{X}_m$ denotes the $(m - 1)$-dimensional "probability simplex" in $\mathbb{R}^m$:

$$\mathcal{X}_m = \left\{ x \in \mathbb{R}^m : (\forall i = 1, \ldots, m : x_i \geq 0) \text{ and } \sum_{i=1}^{m} x_i = 1 \right\} .$$

$\mathcal{K}_{m,n}$ denotes the subset of $\mathbb{Z}^m$ that contains the grid points of $n \cdot \mathcal{X}_m$:

$$\mathcal{K}_{m,n} = \left\{ k \in \mathbb{Z}^m : (\forall i = 1, \ldots, m : k_i \geq 0) \text{ and } \sum_{i=1}^{m} k_i = n \right\} .$$

A point in $\mathcal{X}_m$ represents an unknown distribution X on an m-letter alphabet and a point in $\mathcal{K}_{m,n}$ represents the frequencies that one can extract from a sample word of length n. If the sample word is generated randomly according to X^n, then the corresponding grid-point $k \in \mathcal{K}_{m,n}$ is multinomially distributed. In the sequel, we will use the shorthands with multi-indices

$$\binom{n}{k} = \frac{n!}{k_1! \cdots k_m!} \text{ and } x^k = x_1^{k_1} \cdots x_m^{k_m} .$$

In this notation, the probability assigned to k by the distribution X equals $\binom{n}{k}x^k$.

A *rule* is a function of the form $Q : \mathcal{K}_{m,n} \to \mathcal{X}_m$. It maps a vector $k \in \mathcal{K}_{m,n}$ of absolute letter-frequencies to a point in $Q(k) \in \mathcal{X}_m$, which represents the "guess" for the unknown x. In particular, the *add-β rule* is given by

$$Q_i^\beta(k) = \frac{k_i + \beta}{n + m\beta} \quad \text{for} \quad i = 1, \dots, m, \tag{1}$$

where Q_i denotes the i-th component of Q. For $\beta = 1$, we obtain the add-one rule (Laplace's law of succession), and for $\beta = 1/2$ the add-half-rule (Jeffrey's rule).

The *Kullback-Leibler divergence* between two distributions $x, y \in \mathcal{X}_m$ is defined as

$$\mathrm{D}(x\|y) = \sum_{i=1}^{m} x_i \ln \frac{x_i}{y_i} \quad .$$

The expected Kullback-Leibler divergence between x and $Q(k)$, where the expectation value is taken over the multinomially distributed random variable k, is then

$$C_{m,n}^Q(x) = \sum_{k \in \mathcal{K}_{m,n}} \binom{n}{k} x^k \mathrm{D}(x\|Q(k)) \quad . \tag{2}$$

The *cost associated with rule* Q is given by

$$C_{m,n}^Q = \sup_{x \in \mathcal{X}_m} C_{m,n}^Q(x) \quad . \tag{3}$$

As already mentioned, we aim at finding a rule whose cost is as small as possible and at determining

$$\inf_{Q:\mathcal{K}_{m,n} \to \mathcal{X}_m} C_{m,n}^Q \quad .$$

In the Bayesean approach, the supremum over all $x \in \mathcal{X}_m$ is replaced by averaging over all $x \in \mathcal{X}_m$ according to a (given) prior distribution D on $\mathcal{X}_m$. We will refer only to the uniform prior. Since $\mathcal{X}_m$ has volume $1/(m-1)!$, the resulting Bayes error is

$$\bar{C}_{m,n} = (m-1)! \inf_{Q:\mathcal{K}_{m,n} \to \mathcal{X}_m} \int_{\mathcal{X}_m} C_{m,n}^Q(x) dx \quad , \tag{4}$$

and, according to Cover's result [1], the infimum is attained at $Q = Q^1$. Since

$$\inf_{Q:\mathcal{K}_{m,n} \to \mathcal{X}_m} C_{m,n}^Q \geq \bar{C}_{m,n} \quad ,$$

the Bayesean approach can be used to derive lower bounds on $\inf_{Q:\mathcal{K}_{m,n} \to \mathcal{X}_m} C_{m,n}^Q$.

We now briefly mention some other useful reformulations of the basic equation (2).

Extraction of the entropy-term: Using the formula

$$\forall x \in \mathcal{X}_m : \sum_{k \in \mathcal{K}_{m,n}} \binom{n}{k} x^k = 1 \tag{5}$$

and the definition of the well-known entropy function

$$H(x) = -\sum_{i=1}^{m} x_i \ln x_i \ ,$$

we immediately get

$$C_{m,n}^{Q}(x) = \sum_{k \in \mathcal{K}_{m,n}} \binom{n}{k} x^k \sum_{i=1}^{m} x_i \ln \frac{1}{Q_i(k)} - H(x) \ . \tag{6}$$

An easy algebraic manipulation: Define

$$k_j^{\uparrow i} = \begin{cases} k_i + 1 \text{ if } j = i \ , \\ k_j \quad\quad \text{otherwise} \ , \end{cases} \quad \text{and} \quad k_j^{\downarrow i} = \begin{cases} k_i - 1 \text{ if } j = i \ , \\ k_j \quad\quad \text{otherwise} \ . \end{cases}$$

The inner sums in (6) are now reorganized:

$$\begin{aligned} \sum_{k \in \mathcal{K}_{m,n}} \binom{n}{k} x^k \cdot x_i \ln \frac{x_i}{Q_i(k)} &= \sum_{k \in \mathcal{K}_{m,n}} \binom{n+1}{k^{\uparrow i}} x^{k^{\uparrow i}} \cdot \frac{k_i + 1}{n+1} \ln \frac{x_i}{Q_i(k)} \\ &= \sum_{k \in \mathcal{K}_{m,n+1}} \binom{n+1}{k} x^k \cdot \frac{k_i}{n+1} \ln \frac{x_i}{Q_i(k^{\downarrow i})} \ . \end{aligned}$$

Thus, we have

$$C_{m,n}^{Q}(x) = \sum_{k \in \mathcal{K}_{m,n+1}} \binom{n+1}{k} x^k \cdot \sum_{i=1}^{m} \frac{k_i}{n+1} \ln \frac{x_i}{Q_i(k^{\downarrow i})} \ . \tag{7}$$

Clearly, this algebraic manipulation can be combined with the extraction of the entropy-term to obtain

$$C_{m,n}^{Q}(x) = \sum_{k \in \mathcal{K}_{m,n+1}} \binom{n+1}{k} x^k \cdot \sum_{i=1}^{m} \frac{k_i}{n+1} \ln \frac{1}{Q_i(k^{\downarrow i})} - H(x) \ . \tag{8}$$

3 The Cost Associated With the Add-one Rule

The cost associated with the add-one rule can be calculated in a surprisingly simple manner:

Theorem 1. $C_{m,n}^{Q^1} = \ln\left(1 + \frac{m-1}{n+1}\right)$.

Proof. We substitute $n-1$ for n and show that $C^{Q^1}_{m,n-1} = \ln(1+(m-1)/n)$. It is easy to check that $C^{Q^1}_{m,n-1}(x) = \ln(1+(m-1)/n)$ for each corner x of the probability simplex $\mathcal{X}_m$. For instance, $C^{Q^1}_{m,n-1}(1,0,\ldots,0) = \ln(1+(m-1)/n)$. Thus, $C^{Q^1}_{m,n-1} \geq \ln(1+(m-1)/n)$.
In order to show the converse direction, we first observe that any rule Q satisfies

$$C^{Q}_{m,n-1}(x) \stackrel{(7)}{=} \sum_{k\in\mathcal{K}_{m,n}} \binom{n}{k} x^k \cdot \sum_{i=1}^{m} \frac{k_i}{n} \ln \frac{x_i}{Q_i(k^{\downarrow i})} \tag{9}$$

$$\leq \sum_{k\in\mathcal{K}_{m,n}} \binom{n}{k} x^k \cdot \sum_{i=1}^{m} \frac{k_i}{n} \ln \frac{k_i/n}{Q_i(k^{\downarrow i})} \quad . \tag{10}$$

The last inequality holds because the term $\sum_{i=1}^{m} \frac{k_i}{n} \ln \frac{x_i}{Q_i(k^{\downarrow i})}$ attains its maximal value at $x_i = k_i/n$ since $x \in \mathcal{X}_m$ implies the constraint $\sum_{i=1}^{m} x_i = 1$. We proceed with the add-one rule Q^1. Recalling (1), (5), and $\sum_{i=1}^{m} k_i = n$, we obtain

$$\begin{aligned} C^{Q^1}_{m,n-1}(x) &\leq \sum_{k\in\mathcal{K}_{m,n}} \binom{n}{k} x^k \cdot \sum_{i=1}^{m} \frac{k_i}{n} \ln \frac{k_i/n}{Q^1_i(k^{\downarrow i})} \\ &= \sum_{k\in\mathcal{K}_{m,n}} \binom{n}{k} x^k \cdot \sum_{i=1}^{m} \frac{k_i}{n} \ln \frac{n+m-1}{n} \\ &= \ln\left(1+\frac{m-1}{n}\right) \quad . \end{aligned}$$

•

4 A Lower Bound on the Smallest Possible Cost

We establish a lower bound on $\inf_Q C^Q_{m,n}$ by applying the Bayesean approach for the uniform prior and calculating a sharp lower bound on $\bar{C}_{m,n}$. In a first step, we show that $\bar{C}_{m,n} = \widehat{H}_{m,n} - \bar{H}_m$ with

$$\widehat{H}_{m,n} = |\mathcal{K}_{m,n}|^{-1} \sum_{k\in\mathcal{K}_{m,n}} H\left(\frac{k+1}{n+m}\right) = \binom{n+m-1}{m-1}^{-1} \sum_{k\in\mathcal{K}_{m,n}} H\left(\frac{k+1}{n+m}\right) \quad ,$$

$$\bar{H}_m = (m-1)! \int_{\mathcal{X}_m} H(x)dx \quad ,$$

and $k/(n+m) = [k_1/(n+m), \ldots, k_m/(n+m)]$. Note that $\bar{H}_m$ is the expected entropy when we average uniformly over all $x \in \mathcal{X}_m$. Similarly, $\widehat{H}_{m,n}$ is the expected entropy when we average uniformly over all $(k+1)/(n+m)$ with $k \in \mathcal{K}_{m,n}$. The Bayes error $\bar{C}_{m,n}$ can therefore be viewed as the excess of the "discrete average entropy" (associated with the grid $\mathcal{K}_{m,n}$) over the "continuous average entropy".

In the proof of Theorem 2 below, we will refer to a well-known formula for integrals over simplices:

Lemma 1. *For each $k \in \mathbb{N}_0^m$ and $n = \sum_{j=1}^m k_j$, the following integration rule holds:*

$$\int_{\mathcal{X}_m} x^k dx = \frac{k_1! \cdots k_m!}{(n+m-1)!} \quad .$$

Theorem 2. *Let $\bar{C}_{m,n}$ be defined by (4). Then, $\bar{C}_{m,n} = \widehat{H}_{m,n} - \bar{H}_m$.*

Proof. ¿From Lemma 1 and (6), it follows that

$$\begin{aligned}
&(m-1)! \int_{\mathcal{X}_m} C_{m,n}^Q(x)dx \\
&= (m-1)! \sum_{k \in \mathcal{K}_{m,n}} \binom{n}{k} \sum_{i=1}^m \ln \frac{1}{Q_i(k)} \int_{\mathcal{X}_m} x_i x^k dx - (m-1)! \int_{\mathcal{X}_m} H(x)dx \\
&= \frac{(m-1)! \cdot n!}{(n+m-1)!} \sum_{k \in \mathcal{K}_{m,n}} \sum_{i=1}^m \frac{k_i+1}{n+m} \ln \frac{1}{Q_i(k)} - \bar{H}_m
\end{aligned}$$

The inner sums are minimal when $Q_i(k) = (k_i+1)/(n+m)$, i.e. when $Q = Q^1$; cf. [1]. Hence,

$$\begin{aligned}
\bar{C}_{m,n} &= \binom{n+m-1}{m-1}^{-1} \sum_{k \in \mathcal{K}_{m,n}} \sum_{i=1}^m \frac{k_i+1}{n+m} \ln \frac{n+m}{k_i+1} - \bar{H}_m \\
&= \widehat{H}_{m,n} - \bar{H}_m \quad .
\end{aligned}$$

•

The entropy term $\bar{H}_m$ is related to the Harmonic number $\mathbb{h}_m = \sum_{j=1}^m 1/j$ as follows:

Lemma 2. $\bar{H}_m = \sum_{j=2}^m \frac{1}{j} = \mathbb{h}_m - 1$.

The proof of Lemma 2 will be given in subsection 4.1.

The Bayes error can now be bounded from below as follows:

Theorem 3. *$\bar{C}_{m,n}$ is larger than*

$$\frac{m-1}{2(n+m)} + \frac{1}{12(n+m)^2} - \frac{(m-1)\,m}{12(n+m)(n+1)} \left(\mathbb{h}_{n+m-1} - \mathbb{h}_{m-2}\right) - \frac{1}{120(n+m)^4} \quad .$$

The proof of Theorem 3 will be given in subsection 4.2. In the full version of this paper, we will provide an upper bound on $\bar{C}_{m,n}$ which differs from the lower bound by a term of order $\ln(n)/n^2$. We conjecture that the upper bound can be strengthened such that it differs from the lower bound only by a term of order $\ln(n)/n^4$.

The reader who is not interested in the technical details of the proofs may now immediately pass to section 5 without loss of continuity.

4.1 Proof of Lemma 2

Since the domain $\mathcal{X}_m$ and the function H are symmetric in $x_1, \ldots, x_m$, we can reduce $\int_{\mathcal{X}_m} H(x)dx$ to a one-dimensional integral. Moreover, we note that the volume of the $(m-2)$-dimensional polytope $\{x \in \mathcal{X}_m : x_1 = 1-t\}$ equals $t^{m-2}/(m-2)!$:

$$\begin{aligned}
\bar{H}_m &= (m-1)! \int_{\mathcal{X}_m} H(x)dx \\
&= -m! \int_{\mathcal{X}_m} x_1 \ln(x_1) dx \\
&= -m(m-1) \int_0^1 t^{m-2}(1-t)\ln(1-t)dt \\
&= -m(m-1) \int_0^1 (t^{m-2} - t^{m-1})\ln(1-t)dt.
\end{aligned}$$

Next we perform partial integration such that boundary terms vanish:

$$\begin{aligned}
\bar{H}_m &= -m(m-1)\left[\left(\frac{t^{m-1}-1}{m-1} - \frac{t^m-1}{m}\right)\ln(1-t)\right]_0^1 \\
&\quad - m(m-1)\int_0^1 \left[\frac{t^{m-1}-1}{m-1} - \frac{t^m-1}{m}\right]\frac{1}{1-t}dt \\
&= \int_0^1 \left[\frac{m(t^{m-1}-1)}{t-1} - \frac{(m-1)(t^m-1)}{t-1}\right]dt \\
&= \int_0^1 \left[m\sum_{j=0}^{m-2} t^j - (m-1)\sum_{j=0}^{m-1} t^j\right]dt \\
&= \int_0^1 \left[\sum_{j=0}^{m-2} t^j - (m-1)t^{m-1}\right]dt \\
&= \sum_{j=0}^{m-2}\frac{1}{j+1} - \frac{m-1}{m} = \sum_{j=2}^{m}\frac{1}{j} \quad .
\end{aligned}$$

4.2 Proof of Theorem 3

For ease of reference, we first mention some useful facts that will be needed in the course of the proof:

1. For each $n \geq 1$, there exists $0 < \varepsilon < 1$ such that

$$\mathbb{h}_n = \ln n + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + \frac{\varepsilon}{120n^4} \quad , \tag{11}$$

   where $\gamma = 0.5772156649\cdots$ denotes Euler's constant.

2. The following formulas hold for binomial coefficients:

$$\sum_{i=0}^{n} \binom{i}{m} = \binom{n+1}{m+1} , \tag{12}$$

$$\binom{n+1}{i+1} = \frac{n+1}{i+1} \binom{n}{i} , \tag{13}$$

$$\sum_{j=i}^{n+1} j \binom{n+m-j-1}{m-2} = \frac{n+(m-1)i+1}{m} \binom{n+m-i}{m-1} , \tag{14}$$

$$\sum_{j=1}^{n+1} j \binom{n+m-j-1}{m-2} = \binom{n+m}{m} , \tag{15}$$

$$\binom{m+n}{m}^{-1} \sum_{j=1}^{n} \frac{1}{j} \binom{n+m-j}{m} = \mathbb{h}_{n+m} - \mathbb{h}_m . \tag{16}$$

Formulas (11), (12), and (13) are well-known as Euler's summation rule, as the rule of upper summation, and as the rule of absorption, resp., and can be found, for instance, in [4]. Formula (15) is an immediate consequence of formula (14), specialized to $i = 1$, and formula (13). The proofs for formulas (14) and (16) are found in the full version of this paper.

We are now ready to sketch the proof of Theorem 3. We first bring $\widehat{H}_{m,n}$ into another form that will be more convenient for our purpose. In the following computation, we mainly exploit (11), (13), and (15) and the fact that

$$|\{k \in \mathcal{K}_{m,n} : k_i = j\}| = \binom{n+m-j-2}{m-2}$$

for each fixed $i \in \{1, \ldots, m\}$ and each fixed $j \in \{0, \ldots, n\}$:

$$\begin{aligned}
\widehat{H}_{m,n} &= -\binom{n+m-1}{m-1}^{-1} \sum_{k \in \mathcal{K}_{m,n}} \sum_{i=1}^{m} \frac{k_i+1}{n+m} \cdot \ln \frac{n+m}{k_i+1} \\
&= -\binom{n+m-1}{m-1}^{-1} \sum_{i=1}^{m} \sum_{j=0}^{n} \binom{n+m-j-2}{m-2} \frac{j+1}{n+m} \ln \frac{n+m}{j+1} \\
&= \frac{m}{n+m} \binom{n+m-1}{m-1}^{-1} \sum_{j=0}^{n} \binom{n+m-j-2}{m-2} (j+1) \ln \frac{n+m}{j+1} \\
&\overset{(13)}{=} \binom{n+m}{m}^{-1} \sum_{j=1}^{n+1} \binom{n+m-j-1}{m-2} j \ln \frac{n+m}{j} \\
&= A - B ,
\end{aligned}$$

where

$$
\begin{aligned}
A &= \ln(n+m)\binom{n+m}{m}^{-1}\sum_{j=1}^{n+1} j\binom{n+m-j-1}{m-2} \\
&\stackrel{(15)}{=} \ln(n+m) \\
&\stackrel{(11)}{>} \mathbb{h}_{n+m} - \gamma - \frac{1}{2(n+m)} + \frac{1}{12(n+m)^2} - \frac{1}{120(n+m)^4}
\end{aligned}
$$

and

$$
\begin{aligned}
B &= \binom{n+m}{m}^{-1}\sum_{j=1}^{n+1} j\binom{n+m-j-1}{m-2}\ln j \\
&\stackrel{(11)}{<} \binom{n+m}{m}^{-1}\sum_{j=1}^{n+1} j\binom{n+m-j-1}{m-2}\left(\mathbb{h}_j - \gamma - \frac{1}{2j} + \frac{1}{12j^2}\right) \quad .
\end{aligned}
$$

Thus, $B < B_1 - B_2 - B_3 + B_4$ for

$$
\begin{aligned}
B_1 &= \binom{n+m}{m}^{-1}\sum_{j=1}^{n+1} j\binom{n+m-j-1}{m-2}\mathbb{h}_j \quad , \\
B_2 &= \gamma\binom{n+m}{m}^{-1}\sum_{j=1}^{n+1} j\binom{n+m-j-1}{m-2} \\
&\stackrel{(15)}{=} \gamma \quad , \\
B_3 &= \frac{1}{2}\binom{n+m}{m}^{-1}\sum_{j=1}^{n+1}\binom{n+m-j-1}{m-2} \\
&= \frac{1}{2}\binom{n+m}{m}^{-1}\sum_{j=0}^{n+m-2}\binom{j}{m-2} \\
&\stackrel{(12)}{=} \frac{1}{2}\binom{n+m}{m}^{-1}\binom{n+m-1}{m-1} \\
&= \frac{m}{2(n+m)} \quad , \\
B_4 &= \frac{1}{12}\binom{n+m}{m}^{-1}\sum_{j=1}^{n+1}\frac{1}{j}\binom{n+m-1-j}{m-2} \\
&= \frac{(m-1)m}{12(n+m)(n+1)}\binom{n+m-1}{m-2}\sum_{j=1}^{n+1}\frac{1}{j}\binom{n+m-1-j}{m-2} \\
&\stackrel{(16)}{=} \frac{(m-1)m}{12(n+m)(n+1)}(\mathbb{h}_{n+m-1} - \mathbb{h}_{m-2}) \quad .
\end{aligned}
$$

Here, formula (16) is applied to parameters $n-1$ in the role of n and $m-2$ in the role of m.

We move on to the computation of B_1. We start by rewriting this term as follows:

$$\begin{aligned}
B_1 &= \binom{n+m}{m}^{-1} \sum_{j=1}^{n+1} j \binom{n+m-j-1}{m-2} \sum_{i=1}^{j} \frac{1}{i} \\
&= \binom{n+m}{m}^{-1} \sum_{i=1}^{n+1} \frac{1}{i} \sum_{j=i}^{n+1} j \binom{n+m-j-1}{m-2} \\
&\overset{(14)}{=} \binom{n+m}{m}^{-1} \sum_{i=1}^{n+1} \frac{1}{i} \frac{n+(m-1)i+1}{m} \binom{n+m-i}{m-1} \\
&= B_1' + B_1'' \ ,
\end{aligned}$$

where

$$\begin{aligned}
B_1' &= \binom{n+m}{m}^{-1} \frac{m-1}{m} \sum_{i=1}^{n+1} \binom{n+m-i}{m-1} \\
&= \left(1-\frac{1}{m}\right) \binom{n+m}{m}^{-1} \sum_{i=0}^{n+m-1} \binom{i}{m-1} \\
&\overset{(12)}{=} \left(1-\frac{1}{m}\right) \binom{n+m}{m}^{-1} \binom{n+m}{m} \\
&= 1-\frac{1}{m} \ , \\
B_1'' &= \binom{n+m}{m}^{-1} \frac{n+1}{m} \sum_{i=1}^{n+1} \frac{1}{i} \binom{n+m-i}{m-1} \\
&= \binom{n+m}{m-1}^{-1} \sum_{i=1}^{n+1} \frac{1}{i} \binom{n+m-i}{m-1} \\
&\overset{(16)}{=} \mathbb{h}_{n+m} - \mathbb{h}_{m-1} \ .
\end{aligned}$$

Here, formula (16) is applied to $n+1$ in the role of n and $m-1$ in the role of m. We may now conclude that B_1 satisfies equation

$$B_1 = 1 + \mathbb{h}_{n+m} - \mathbb{h}_m \ .$$

We finally observe that

$$
\begin{aligned}
\widehat{H}_{m,n} - \bar{H}_m &= A - B - \bar{H}_m \\
&= A - B_1 + B_2 + B_3 - B_4 - \bar{H}_m \\
&> \left(\mathbb{h}_{n+m} - \gamma - \frac{1}{2(n+m)} + \frac{1}{12(n+m)^2} - \frac{1}{120(n+m)^4} \right) \\
&\quad - (1 + \mathbb{h}_{n+m} - \mathbb{h}_m) + \gamma + \frac{m}{2(n+m)} \\
&\quad - \left(\frac{(m-1)m}{12(n+m)(n+1)} (\mathbb{h}_{n+m-1} - \mathbb{h}_{m-2}) \right) - (\mathbb{h}_m - 1) \\
&= \frac{m-1}{2(n+m)} + \frac{1}{12(n+m)^2} - \frac{(m-1)m}{12(n+m)(n+1)} (\mathbb{h}_{n+m-1} - \mathbb{h}_{m-2}) \\
&\quad - \frac{1}{120(n+m)^4} \ ,
\end{aligned}
$$

which concludes the proof of Theorem 3.

5 An Asymptotically Optimal Rule ?

In this section, we will propose a rule Q^* whose cost seems to match asymptotically the lower bound $(m-1)/(2n)$. To this end, we will modify the add-0.8 rule $Q^{0.8}(k)$ at $k = 0, 1, n-1, n$ as specified in (24) below for the special case $m = 2$.

We start with some preliminary considerations. The computation of the cost associated with an add-β rule may start like the computation of the cost associated with the add-one rule in section 3 up to the following notable exception: we aim at controlling the deficit when we pass from (9) to (10). This deficit does not depend on Q. It equals the absolute difference between the entropy function H and its approximation by Bernstein polynomials. By exploiting this observation, we show in the full paper (focusing on the case $m = 2$) that

$$
C_{2,n}^{Q^\beta}(x, 1-x) \leq \frac{1}{2n} \left(1 + \frac{1}{3nx(1-x)+1} \right) \tag{17}
$$

for $1/2 \leq \beta \leq 1$. Note that this upper bound will match asymptotically the lower bound $1/(2n)$, if x is "separated away" from 0 and 1 in the sense that $x \in [\delta_n, 1-\delta_n]$ for a sequence $\delta_n = \omega(1/n)$. For instance, the upper bound (17) is already satisfactory for $x \in [1/\sqrt{n}, 1-1/\sqrt{n}]$. (Observe that the special choice of β does not matter very much, when x is separated away from 0 and 1.)

In order to control the cost of the add-β rule when x is *not* separated away from 0 and 1, we follow Krichevskiy [7] who introduced the characteristic function

$$
\Phi(\beta, z) = \beta + z \ln z - z e^{-z} \sum_{j=0}^{\infty} \frac{z^j}{j!} \ln(j+\beta) \ . \tag{18}
$$

For $m = 2$ and each fixed z, the relation between Φ and the cost-terms that we are interested in is as follows:

$$\Phi(\beta, z) = \lim_{n\to\infty} nC_{2,n}^{Q^\beta}(z/n, 1 - z/n) \ . \tag{19}$$

Krichevskiy showed that $C_{m,n}^{Q^{\beta_0}} \approx \beta_0/n$ for $\beta_0 = 0.50922\cdots$ by analyzing Φ at β_0; cf. Fig. 1.

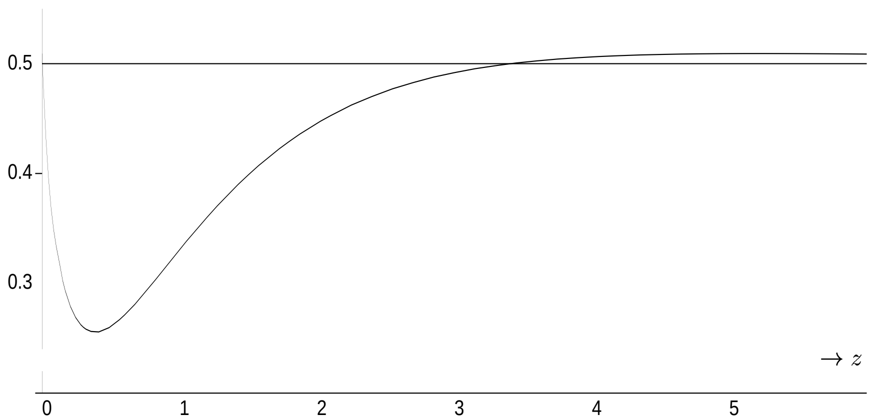

Fig. 1. Characteristic function at $\beta_0 = 0.50922\cdots$ from Krichevskiy's solution. The maximum is attained at $z \approx 5.3$.

Details of the modification for the case $m = 2$: For $m = 2$, we will use a simplified notation from now on. First, we we write C_n^Q instead of $C_{2,n}^Q$. Second, we write $Q(k)$ instead of $Q(k, n - k)$ and $C_n^Q(x)$ instead of $C_n^Q(x, 1 - x)$. The function in (2) can now be written in the form

$$C_n^Q(x) = \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} \left\{ x \ln \frac{x}{Q(k)} + (1-x) \ln \frac{1-x}{1-Q(k)} \right\} . \tag{20}$$

In the following, Q denotes a modification of Q^β that differs from Q^β only at $k = 0, 1, n - 1, n$ and satisfies

$$Q(0) < Q(1) \le 3/n \text{ and } Q(0) + Q(n) = Q(1) + Q(n - 1) = 1 \ . \tag{21}$$

Due to a symmetry argument, we restrict ourselves to $x \le 0.5$. In this part of the domain, the differences between the two rules at $k = n - 1, n$ contribute only

$O(2^{-n})$ to the difference between the associated cost-terms:

$$\begin{aligned} C_n^Q(x) - C_n^{Q^\beta}(x) &= (1-x)^n \left\{ x \ln \tfrac{Q^\beta(0)}{Q(0)} + (1-x) \ln \tfrac{1-Q^\beta(0)}{1-Q(0)} \right\} \\ &+ nx(1-x)^{n-1} \left\{ x \ln \tfrac{Q^\beta(1)}{Q(1)} + (1-x) \ln \tfrac{1-Q^\beta(1)}{1-Q(1)} \right\} + O(2^{-n}) \ . \end{aligned} \tag{22}$$

Since $Q(0)$ and $Q(1)$ are smaller than $3/n$, we get

$$\begin{aligned} \ln \tfrac{1-Q^\beta(k)}{1-Q(k)} &= \ln\{1 + \tfrac{Q(k)-Q^\beta(k)}{1-Q(k)}\} \\ &= Q(k) - Q^\beta(k) + O(n^{-2}) \quad \text{for } k = 0, 1 \ . \end{aligned}$$

Now we proceed from the Bernoulli distribution to the Poisson distribution (as done by Krichevskiy), or equivalently, setting $z = nx$ we have, for each fixed z, $(1 - z/n)^n = e^{-z} + O(1/n)$ and

$$\begin{aligned} n\{C_n^Q(z/n) - C_n^{Q^\beta}(z/n)\} &= e^{-z}\{z \ln \tfrac{Q^\beta(0)}{Q(0)} + Q(0) - Q^\beta(0)\} \\ &+ ze^{-z}\{z \ln \tfrac{Q^\beta(1)}{Q(1)} + Q(1) - Q^\beta(1)\} + O(n^{-1}). \end{aligned} \tag{23}$$

Finally, we define our candidate rule Q^* as the following modification of the add-0.8 rule:

$$Q^*(k) = \begin{cases} \frac{k+1/2}{n+1} & \text{if } k = 0, n \ , \\ \frac{k+1.05}{n+2.1} & \text{if } k = 1, n-1 \ , \\ \frac{k+0.8}{n+1.6} & \text{otherwise} \ . \end{cases} \tag{24}$$

Note that Q^* coincides with the add-half rule at $k = 0, n$, with the add-1.05 rule at $k = 1, n-1$, and with the add-0.8 rule everywhere else. Furthermore, Q^* satisfies (21). Combining (23) and a convergence condition similar[2] to (19), we show in the full paper that

$$\begin{aligned} nC_n^{Q^*}(\tfrac{z}{n}) &= \Phi(0.8, z) \\ &+ e^{-z}\left\{-0.3 + z(\ln(1.6) + 0.25) - z^2 \ln \tfrac{2.05}{1.8}\right\} + O(n^{-1/2}) \end{aligned} \tag{25}$$

holds uniformly for all $0 \leq z \leq n^{1/4}$. An interesting part of that function is shown in Fig. 2. The values do not exceed $1/2$, i.e., the number given by the lower bound (something that we have checked numerically up to $z = 20$). Moreover, the additional term $e^{-z}\left\{-0.3 + z(\ln(1.6) + 0.25) - z^2 \ln \frac{2.05}{1.8}\right\}$ on the right-hand side of (25) is negative for $z \geq 6$. We note that the cost associated with Q^* would provably asymptotically match the lower bound $1/(2n)$ if we could guarantee that the right-hand side of (25) does not exceed $1/2$ for $20 \leq z \leq n^{1/4}$. For larger z, we may appeal to (17).

[2] Actually, we need a sort of uniform convergence.

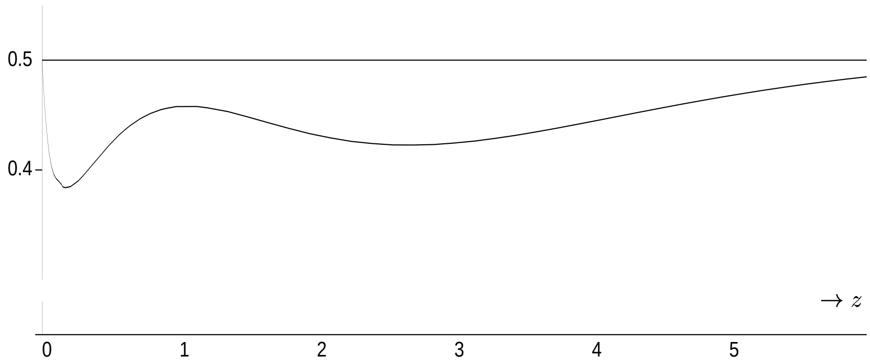

Fig. 2. A plot of the right-hand side of equation (25) for $z \in [0, 6]$, except that the vanishing term of order $n^{-1/2}$ is ignored.

The Case $m \geq 3$*:* The case $m \geq 3$ can be treated analogously except for the complication that special care must be taken to guarantee the condition $\sum_{i=1}^{m} Q_i^*(k) = 1$ for each $k \in \mathcal{K}_{m,n}$. More details can be found in the full paper.

References

1. T. M. Cover. Admissibility properties of Gilbert's encoding for unknown source probabilities. *IEEE Transactions on Information Theory*, 18:216–217, 1971.
2. Jürgen Forster and Manfred Warmuth. Relative expected instantaneous loss bounds. *Journal of Computer and System Sciences.* To appear.
3. Jürgen Forster and Manfred Warmuth. Relative expected instantaneous loss bounds. In *Proceedings of the 13th Annual Conference on Computational Learning Theory*, pages 90–99. Morgan Kaufmann, 2000.
4. Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics.* Addison Wesley, 1989.
5. D. A. Huffman. A method for the construction of minimum redundancy codes. *Proc. IRE*, 40(10):1098–1101, 1952.
6. Rafail E. Krichevskiy. *Universal Compression and Retrieval.* Kluwer Academic Publishers, 1994.
7. Rafail E. Krichevskiy. Laplace's law of succession and universal encoding. *IEEE Transactions on Information Theory*, 44(1):296–303, 1998.
8. Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423,623–656, 1948.
9. Qun Xie and Andrew R. Barron. Minimax redundancy for the class of memoryless sources. *IEEE Transactions on Information Theory*, 43(2):646–657, 1997.
10. Qun Xie and Andrew R. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445, 2000.