Minimax Optimal Bayes Mixtures for Memoryless Sources over Large Alphabets

Elias Jääsaari ELIAS.JAASAARI@AALTO.FI
Helsinki Institute for Information Technology HIIT
Department of Computer Science, Aalto University
P.O. Box 15400, FI-00076, Finland

Janne Leppä-aho JANNE.LEPPA-AHO@CS.HELSINKI.FI
Helsinki Institute for Information Technology HIIT
Department of Computer Science, University of Helsinki
P.O. Box 68, FI-00014, Finland

Tomi Silander TOMI.SILANDER@NAVERLABS.COM
NAVER LABS Europe
6 Chemin de Maupertuis, 38240 Meylan, France

Teemu Roos TEEMU.ROOS@CS.HELSINKI.FI
Helsinki Institute for Information Technology HIIT
Department of Computer Science, University of Helsinki
P.O. Box 68, FI-00014, Finland

Editors:

Abstract

The normalized maximum likelihood (NML) distribution achieves minimax log loss and coding regret for the multinomial model. In practice other nearly minimax distributions are used instead as calculating the sequential probabilities needed for coding and prediction takes exponential time with NML. The Bayes mixture obtained with the Dirichlet prior $\mathrm{Dir}(1/2,\ldots,1/2)$ and asymptotically minimax modifications of it have been widely studied in the context of large sample sizes. Recently there has also been interest in minimax optimal coding distributions for large alphabets. We investigate Dirichlet priors that achieve minimax coding regret when the alphabet size m is finite but large in comparison to the sample size n. We prove that a Bayes mixture with the Dirichlet prior $\mathrm{Dir}(1/3,\ldots,1/3)$ is optimal in this regime (in particular, when $m > \frac{5}{2}n + \frac{4}{n-2} + \frac{3}{2}$). The worst-case regret of the resulting distribution approaches the NML regret as the alphabet size grows.

Keywords: minimax regret, normalized maximum likelihood, Bayes mixture, large alphabet, universal coding, universal prediction, online learning

1. Introduction

Let $x^{n-1} = (x_1, \ldots, x_{n-1})$ be a sequence of observations from a finite set $\mathcal{X}$, and suppose we wish to predict the next observation x_n. This requires choosing the sequential probabilities for all x^{n-1}. When x_n is observed, we incur the log loss $-\log p(x_n|x^{n-1})$ (see e.g. Cesa-Bianchi and Lugosi, 2006). Predicting all n observations sequentially, our goal is then to minimize the cumulative log loss $-\sum_{k=0}^{n-1} \log p(x_{k+1}|x^k) = -\log p(x^n)$.

An analogous problem appears in source coding where we try to minimize the code-lengths for each block x^n of n symbols. For any probability distribution p, there exists a uniquely decodable code with code-lengths $-\log p(x^n)$ (ignoring integer constraints), and vice versa. The code-length also plays an important role in the minimum description length (MDL) principle (Grünwald, 2007) where it can be used as a model selection criterion.

The performance of the distribution or strategy p can be measured relative to a known class of distributions $\mathcal{P}$ such as all i.i.d. or Markov sources. For a distribution q, we define its regret relative to the sequence x^n as the excess code-length or log loss compared to the optimal distribution in $\mathcal{P}$ in hindsight:

$$\operatorname{regret}(q, x^n) = \log \frac{1}{q(x^n)} - \inf_{p \in \mathcal{P}} \log \frac{1}{p(x^n)} = \sup_{p \in \mathcal{P}} \log \frac{p(x^n)}{q(x^n)}.$$

A special role is given to the distribution that achieves minimax (pointwise) regret

$$\inf_q \sup_{x^n \in \mathcal{X}^n} \operatorname{regret}(q, x^n),$$

i.e. minimizes the regret in the worst case. Such minimax methods have been shown to be robust with respect to different data generating mechanisms where a good choice of prior is challenging (Eggeling et al., 2014; Määttä et al., 2016).

In this paper we consider the case where we have a parameterized discrete memoryless source over an alphabet $\mathcal{X}$ of size m. Each $x_i \in \mathcal{X}$ is generated independently according to a probability mass function in the parametric family $\{p(x;\theta) : \theta \in \Theta \subset \mathbb{R}^d\}$. Thus

$$p(x^n;\theta) = \prod_{i=1}^{n} p(x_i;\theta).$$

Of particular interest is the multinomial model $p(x;\theta) = \theta_x$, $\sum_{j=1}^m \theta_j = 1$, extended to sequences x^n by the i.i.d. assumption. The shortest code-length or least log loss in hindsight is achieved by the maximum likelihood model $p(x^n;\hat{\theta}(x^n))$, where

$$\hat{\theta}(x^n) = \arg\max_\theta p(x^n;\theta).$$

Shtarkov (1987) proved that for this model (and all other models for which the maximum likelihood measure is normalizable), the minimax regret is achieved by the normalized maximum likelihood (NML) distribution

$$p_{\mathrm{NML}}(x^n) = \frac{p(x^n;\hat{\theta}(x^n))}{C_n^m},$$

where $C_n^m = \sum_{x^n} p(x^n;\hat{\theta}(x^n))$ is the Shtarkov sum that goes over all $x^n \in \{1,\ldots,m\}^n$. The NML distribution has uniform regret $\log C_n^m$ for all sequences x^n.

The normalizing constant often renders using NML impractical. Even though the normalizing constant can be calculated in linear time in the multinomial case (Kontkanen and Myllymäki, 2007), obtaining the sequential probabilities needed for coding takes exponential time. Consequently, other nearly minimax distributions such as Bayes mixtures for which the sequential predictions can be obtained efficiently have been studied.

Bayes mixtures exhibit useful properties for approximation of the NML distribution. In certain exponential families Bayes mixtures are asymptotically minimax for both the worst-case and the expected regret (Takeuchi and Barron, 1998). For the multinomial model, Krichevsky and Trofimov (1981) proposed using a Bayes mixture with the $\mathrm{Dir}(1/2,\ldots,1/2)$ prior which has higher regret than the minimax level by a vanishing amount everywhere except for the boundaries of the frequency composition simplex. This distribution is also the Jeffreys prior which has a special role as the invariant (or reference) prior (Bernardo, 1979). Barron et al. (2014) showed that even exact representation of NML is possible using signed mixtures. However, this strategy requires high computational complexity when the alphabet size is large and is in practice sensitive to numerical errors.

In recent years large alphabet methods have been gaining more attention. The alphabet size can be larger than the sample size or even infinite in application areas such as natural language processing (Chen and Goodman, 1996) and Bayesian network structure learning (Silander et al., 2018). Images can also be considered as data associated with a large alphabet where each pixel can take on 2^{24} different values.

Different strategies for data compression on large alphabets have subsequently been proposed. The regret over infinite alphabets is infinite (Kieffer, 1978) since describing the symbols that appear in the sequence requires an unbounded number of bits. In particular, Orlitsky and Santhanam (2004) showed that when $n = o(m)$,

$$\log C_n^m \sim n \log \frac{m}{n},$$

and thus the regret is high when m is large. The work on minimax compression with large alphabets has subsequently focused on subclasses of i.i.d. distributions such as envelope classes (Acharya et al., 2014; Bontemps, 2011) and patterns (Orlitsky et al., 2004) that can be compressed with vanishing regret. Recently there has also been work on codes that do not assume a fixed subclass in advance but provide optimistic bounds within subclasses (Boucheron et al., 2015; Orlitsky and Suresh, 2015).

However, as these codes target a different class of distributions, their code-lengths are not interchangeable with code-lengths for i.i.d. distributions and thus they are not useful in for example model selection. A coding distribution for the i.i.d. class is still needed to calculate a target minimax distribution. Therefore such distributions have recently been proposed for large alphabets (Yang and Barron, 2017).

In this paper we study the minimax optimal Bayes mixture with a Dirichlet prior in the large (but finite) alphabet setting. Our main theorem (Theorem 6) states that the minimax optimal Dirichlet prior is $\mathrm{Dir}(1/3,\ldots,1/3)$ when $m > \frac{5}{2}n + \frac{4}{n-2} + \frac{3}{2}$. We also prove that the worst-case regret of the resulting distribution approaches that of the NML distribution when m grows and can be used as an approximation of the NML regret. The Bayes mixture also allows more efficient calculation of the marginal and conditional probabilities needed for e.g. coding and prediction than earlier proposed distributions.

The rest of this paper is structured as follows. In Section 2 we discuss related work on Bayes mixtures in the large sample size setting, and then prove the minimax optimality of the $\mathrm{Dir}(1/3,\ldots,1/3)$ prior in the large alphabet setting. In Section 3 we prove that the worst-case regret of the 1/3-mixture approaches the NML regret as the alphabet size grows and study the worst-case regret numerically and as an approximation to the NML regret. Section 4 is reserved for discussion and conclusions.

2. Main result

Given a class $\mathcal{P}$ of distributions parameterized by a parameter set Θ, the Bayes mixture is given by the weighted mixture $p_{\text{Bayes}}(x^n) = \int_\Theta p(x^n;\theta)q(\theta)\,d\theta$ for some prior distribution q on the parameter space. The corresponding conjugate prior for the multinomial model is the Dirichlet distribution. In the symmetric case where each outcome $x \in \{1,\ldots,m\}$ has equal prior probability, its density function takes the form

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

where $\alpha > 0$ is a hyperparameter and $\Gamma(\cdot)$ is the gamma function $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$. The probabilities for the sequences x^n are obtained by taking the weighted mixture

$$p_{B,\alpha}(x^n) = \int_\Theta \prod_{i=1}^n p(x_i;\theta)q(\theta;\alpha)\,d\theta = \frac{\Gamma(m\alpha)}{\Gamma(\alpha)^m}\frac{\prod_{j=1}^m \Gamma(n_j+\alpha)}{\Gamma(n+m\alpha)},$$

where n_j is the number of occurrences of symbol j in x^n.

The choice of α can be interpreted as the amount of prior mass, or "pseudo-count" given to each of the outcomes. Small differences in α can be crucial in for example Bayesian network structure learning (Silander et al., 2007; Suzuki, 2017). The Krichevsky-Trofimov estimator (Krichevsky and Trofimov, 1981) obtained with $\alpha = 1/2$ is asymptotically maximin (Xie and Barron, 1997). However, it is not asymptotically minimax as its regret is higher than the minimax regret by a nonvanishing amount on the boundaries of the frequency composition simplex where some of the symbols do not occur at all.

Xie and Barron (2000) proposed an asymptotically minimax version of the Krichevsky-Trofimov estimator that puts extra mass to the boundaries of the probability simplex with a horizon-dependent strategy. Later, Watanabe and Roos (2015) proved that a simpler Bayes procedure p_{B,α_n} with the horizon-dependent hyperparameter

$$\alpha_n = \frac{1}{2} - \frac{\log 2}{2}\frac{1}{\log n}$$

achieves asymptotic minimaxity. Here α_n converges to the asymptotic value $1/2$ at a logarithmic rate as $n \to \infty$ regardless of the alphabet size. However, in practice we deal with a finite amount of data. As the alphabet size grows, the minimax optimal hyperparameter moves further away from α_n when the sample size is finite.

In certain application areas the alphabet size can be large (but finite) in comparison to the sample size. In this case the minimax optimal hyperparameter α can be different. We now consider finding the minimax optimal Dirichlet prior for the Bayes mixture in the large alphabet setting. Specifically, if m is large compared to n, we want to find $\alpha^* = \arg\min_\alpha \max_{x^n} \text{regret}(p_{B,\alpha}, x^n)$, where

$$\text{regret}(p_{B,\alpha}, x^n) = \log\frac{p(x^n;\hat{\theta}(x^n))}{p_{B,\alpha}(x^n)} = \sum_{j=1}^m n_j \log\frac{n_j}{n} - \log p_{B,\alpha}(x^n).$$

Our main theorem (Theorem 6) states that when $m > \frac{5}{2}n + \frac{4}{n-2} + \frac{3}{2}$, the minimax optimal hyperparameter is $\alpha^* = \frac{1}{3}$.

We prove this result by showing that there is always a sequence whose regret is decreasing as a function of α when $m \geq n$ (Lemma 3) and a sequence whose regret is increasing as a function of α when $m > \frac{5}{2}n + \frac{4}{n-2} + \frac{3}{2}$ (Lemma 5). Furthermore, at the point $\alpha = 1/3$ these sequences achieve the highest regret (Lemma 2). Thus the point $\alpha = 1/3$ has to be the minimax point. We start by proving the following technical result:

Lemma 1 *For $k \geq 2$, the function*

$$d(k) = k \log k - k \log 3 - \log \Gamma(k + \frac{1}{3}) + \log \Gamma(\frac{1}{3})$$

satisfies $d(k) \leq 0$, where equality holds if and only if $k = 2$.

Proof The derivative of d is $d'(k) = \log(k/3) - \psi(k + 1/3) + 1$, where ψ is the digamma function, defined as the logarithmic derivative of the gamma function. Using the inequality $\psi'(k) > 1/k + 1/(2k^2)$ (e.g., Guo and Qi, 2013), we can show that if $k > 2/3$, we have $d''(k) = 1/k - \psi'(k + 1/3) < 0$. Moreover, since $d'(2) < 0$, the derivative d' must also be negative for all $k \geq 2$ and therefore d is decreasing. The claim holds since $d(2) = 0$. ■

Lemma 1 allows us to prove the following result which characterizes the sequences that achieve maximum regret for $p_{B,1/3}$ in the case $m \geq n$:

Lemma 2 *Let m, n be integers such that $m \geq n \geq 1$ and x^n be a sequence where all the symbols are different. Then for all sequences y^n we have $\mathrm{regret}(p_{B,\frac{1}{3}}, x^n) \geq \mathrm{regret}(p_{B,\frac{1}{3}}, y^n)$, where equality holds if and only if no symbol in y^n occurs more than twice.*

Proof Let y^n be a sequence with at least one symbol occurring more than once and j be a symbol that occurs $n_j > 1$ times in y^n. Furthermore, let z^n be the same sequence as y^n except $n_j - 1$ occurrences of j are each replaced by a different non-occurring symbol. We first note that the regret of a sequence w^n can be written as

$$\mathrm{regret}(p_{B,\alpha}, w^n) = \sum_{j=1}^{m} \{n_j \log n_j - \log \Gamma(n_j + \alpha)\} + \kappa,$$

where κ is a quantity that does not depend on the sequence w^n and $0 \log 0 = 0$. Now

$$\begin{aligned}
\mathrm{regret}(p_{B,\frac{1}{3}}, y^n) &= \mathrm{regret}(p_{B,\frac{1}{3}}, z^n) - n_j(1 \log 1 - \log \Gamma(1 + \frac{1}{3})) \\
&\quad - (n_j - 1) \log \Gamma(\frac{1}{3}) + n_j \log n_j - \log \Gamma(n_j + \frac{1}{3}) \\
&= \mathrm{regret}(p_{B,\frac{1}{3}}, z^n) + n_j \log n_j - n_j \log 3 - \log \Gamma(n_j + \frac{1}{3}) + \log \Gamma(\frac{1}{3}).
\end{aligned}$$

Hence from Lemma 1 we have

$$\mathrm{regret}(p_{B,\frac{1}{3}}, y^n) = \mathrm{regret}(p_{B,\frac{1}{3}}, z^n) + d(n_j) \leq \mathrm{regret}(p_{B,\frac{1}{3}}, z^n),$$

where equality holds if and only if $n_j = 2$. Finally, as any sequence y^n can be transformed into a sequence where all the symbols are different by repeated application of the above procedure, the claim holds. ■

We now proceed to proving that if $m \geq n$, there is always a sequence whose regret is decreasing as a function of α:

Lemma 3 *Let m, n be integers such that $m \geq n \geq 2$ and x^n be a sequence where all the symbols are different. Then the function $\alpha \mapsto \mathrm{regret}(p_{B,\alpha}, x^n)$ is decreasing.*

Proof Taking the derivative with respect to α, we obtain

$$\frac{\partial}{\partial \alpha}\mathrm{regret}(p_{B,\alpha}, x^n) = m\psi(n+m\alpha) - m\psi(m\alpha) - \frac{n}{\alpha}.$$

Repeated application of the identity $\psi(n+1) = \psi(n) + 1/n$ gives

$$m\psi(n+m\alpha) - m\psi(m\alpha) - \frac{n}{\alpha} = m\left(\sum_{k=0}^{n-1} \frac{1}{m\alpha + k}\right) - \frac{n}{\alpha} = \sum_{k=0}^{n-1}\left(\frac{1}{\alpha + \frac{k}{m}} - \frac{1}{\alpha}\right) < 0$$

for all positive α. ■

In turn, we can find a sequence whose regret is increasing as a function of α when $\alpha \geq 1/3$ and $m > \frac{5}{2}n + \frac{4}{n-2} + \frac{3}{2}$. This requires first proving the following lemma:

Lemma 4 *Let $\alpha > 0$ and m, n be integers such that $m \geq n \geq 1$. Then*

$$m\psi(n+m\alpha) - m\psi(m\alpha) > \frac{1}{2\alpha} + \frac{1}{2(\alpha+1)} + \frac{2m(n-1)}{2m\alpha + n - 1}.$$

Proof We can first write

$$m\psi(n+m\alpha) - m\psi(m\alpha) = \sum_{k=0}^{n-1} \frac{1}{\alpha + \frac{k}{m}}.$$

Applying the trapezoidal rule

$$\int_1^n f(x)\,dx < \sum_{k=1}^{n} f(k) - \frac{1}{2}(f(1) + f(n)),$$

where f is a convex function, we have

$$\begin{aligned}
\sum_{k=0}^{n-1} \frac{1}{\alpha + \frac{k}{m}} &> \frac{1}{2}\left(\frac{1}{\alpha} + \frac{1}{\alpha + \frac{n-1}{m}}\right) + \int_0^{n-1} \frac{dx}{\alpha + \frac{x}{m}} \\
&> \frac{1}{2}\left(\frac{1}{\alpha} + \frac{1}{\alpha+1}\right) + \int_0^{n-1} \frac{dx}{\alpha + \frac{x}{m}} \\
&= \frac{1}{2\alpha} + \frac{1}{2(\alpha+1)} + m\log\left(1 + \frac{n-1}{m\alpha}\right).
\end{aligned}$$

Using the inequality $\log(1+x) \geq 2x/(2+x)$ valid for $x \geq 0$ gives the result

$$\frac{1}{2\alpha} + \frac{1}{2(\alpha+1)} + m\log\left(1 + \frac{n-1}{m\alpha}\right) \geq \frac{1}{2\alpha} + \frac{1}{2(\alpha+1)} + \frac{2m(n-1)}{2m\alpha + n - 1}.$$

■

Lemma 5 *Let m, n be integers such that*

$$m > \frac{5}{2}n + \frac{4}{n-2} + \frac{3}{2}$$

if $n > 2$ or $m \geq 2$ if $n = 2$. Then the regret of a sequence where each symbol occurs twice (and one symbol occurs once if n is odd) is an increasing function of α when $\alpha \geq 1/3$.

Proof Assume first that n is even and let x^n be a sequence where $n/2$ symbols occur twice. Taking the derivative of the regret we obtain

$$\begin{aligned}\frac{\partial}{\partial\alpha}\text{regret}(p_{B,\alpha}, x^n) &= m\psi(n+m\alpha) - m\psi(m\alpha) + \frac{n}{2}(\psi(\alpha) - \psi(2+\alpha)) \\ &= m\psi(n+m\alpha) - m\psi(m\alpha) - n\left(\frac{1}{2\alpha} + \frac{1}{2(\alpha+1)}\right).\end{aligned}$$

Plugging in $n = 2$, we get $1/(\alpha+1/m) - 1/(\alpha+1)$ which is positive for $m \geq 2$. Now assume that $n > 2$ is odd and take x^n to be a sequence where $(n-1)/2$ symbols occur twice and one symbol occurs once. Taking the derivative gives

$$\begin{aligned}\frac{\partial}{\partial\alpha}\text{regret}(p_{B,\alpha}, x^n) &= m\psi(n+m\alpha) - m\psi(m\alpha) + \frac{n-1}{2}(\psi(\alpha) - \psi(2+\alpha)) - \frac{1}{\alpha} \\ &= m\psi(n+m\alpha) - m\psi(m\alpha) - (n-1)\left(\frac{1}{2\alpha} + \frac{1}{2(\alpha+1)}\right) - \frac{1}{\alpha}.\end{aligned}$$

Thus, if the derivative is positive in the odd case, it is also positive in the even case. For the derivative to be positive in the odd case, from Lemma 4 we have the inequality

$$\frac{1}{2\alpha} + \frac{1}{2(\alpha+1)} + \frac{2m(n-1)}{2m\alpha + n - 1} - (n-1)\left(\frac{1}{2\alpha} + \frac{1}{2(\alpha+1)}\right) - \frac{1}{\alpha} > 0,$$

from which we can solve

$$m > \frac{(n-1)(2\alpha(n-1)+n)}{2\alpha(n-2)}$$

for $\alpha > 0, n > 2$. Clearly this bound is decreasing as a function of α. Plugging $\alpha = 1/3$ into the bound, we have

$$m > \frac{5}{2}n + \frac{4}{n-2} + \frac{3}{2}.$$

Therefore if m satisfies the bound above, the derivative is positive for all $\alpha \geq 1/3$. ■

We are now ready to prove the main result:

Theorem 6 *Let n, m be integers such that*

$$m > \frac{5}{2}n + \frac{4}{n-2} + \frac{3}{2}$$

if $n > 2$ or $m \geq 2$ if $n = 1, 2$. Then

$$\arg\min_{\alpha} \max_{x^n} \text{regret}(p_{B,\alpha}, x^n) = \frac{1}{3}.$$

Proof We denote $r(\alpha) = \max_{z^n} \text{regret}(p_{B,\alpha}, z^n)$. From Lemma 2, we have $r(1/3) = \text{regret}(p_{B,1/3}, x^n) = \text{regret}(p_{B,1/3}, y^n)$, where x^n is a sequence where each symbol is different and y^n is a sequence where each symbol in the sequence occurs twice (and one symbol occurs once if n is odd). Using Lemma 3, we have $r(\alpha) \geq \text{regret}(p_{B,\alpha}, x^n) > r(1/3)$ for all $0 < \alpha < 1/3$. From Lemma 5, we have $r(\alpha) \geq \text{regret}(p_{B,\alpha}, y^n) > r(1/3)$ for all $\alpha > 1/3$ when $m > \frac{5}{2}n + \frac{4}{n-2} + \frac{3}{2}$ if $n > 2$ or $m \geq 2$ otherwise. Consequently, when $n \geq 2$, the function r is minimized at the point $\alpha = 1/3$. In the case $n = 1$ it is straightforward to verify that the worst-case regret is constant as a function of α. ■

The bound approaches $\frac{5}{2}n + \frac{3}{2}$, while numerical results show that the minimum m/n ratio for which the optimal hyperparameter is $a^* = \frac{1}{3}$ converges to between 2.1 and 2.2.

3. Properties of the 1/3-mixture

In this section, we study the asymptotic behavior of the $p_{B,1/3}$ mixture as the alphabet size grows and present numerical experiments for the worst-case regret in comparison to other coding distributions. Furthermore, we show how the optimal hyperparameter α^* can be found efficiently for any given n, m with ε precision.

3.1. Asymptotic behavior

The following theorem states that the worst-case regret of $p_{B,1/3}$ grows asymptotically at the same rate as the regret of the NML distribution when $n = o(m)$:

Theorem 7 *If $n = o(m)$, then*

$$\max_{x^n} \text{regret}(p_{B,\frac{1}{3}}, x^n) = n \log \frac{m}{n} + \frac{3}{2}\frac{n(n-1)}{m} + \mathcal{O}\left(\frac{n^3}{m^2}\right).$$

Proof When $n = o(m)$, by definition there is an m_0 such that $m \geq n$ for all $m \geq m_0$. Then as per Lemma 2, the worst-case regret for $p_{B,1/3}$ occurs when x^n is a sequence where all the symbols are different. Thus

$$\max_{x^n} \text{regret}(p_{B,\frac{1}{3}}, x^n) = \log \Gamma(n + \frac{m}{3}) - \log \Gamma(\frac{m}{3}) - n \log \frac{n}{3}.$$

Applying Stirling's approximation

$$\log \Gamma(x) = x \log x - x + \frac{1}{2} \log \frac{2\pi}{x} + o(1)$$

and using the identity $\log\left(\frac{m}{3} + n\right) = \log \frac{m}{3} + \log\left(1 + \frac{3n}{m}\right)$, we have

$$\max_{x^n} \text{regret}(p_{B,\frac{1}{3}}, x^n) = \left(\frac{m}{3} + n - \frac{1}{2}\right) \log\left(1 + \frac{3n}{m}\right) + n \log \frac{m}{n} - n + o(1).$$

Using the Taylor expansion $\log(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$ results in

$$\max_{x^n} \text{regret}(p_{B,\frac{1}{3}}, x^n) = n \log \frac{m}{n} + \frac{3}{2}\frac{n(n-1)}{m} + \mathcal{O}\left(\frac{n^3}{m^2}\right).$$

■

This asymptotic form is the same as for the NML regret in the case $n = o(m)$ as given by Szpankowski and Weinberger (2012). Thus, if n is fixed and $m \to \infty$, the difference between the worst-case regret of the 1/3-mixture and the NML regret approaches zero.

3.2. Numerical experiments

Even though the NML regret $\log C_n^m$, which can be used in e.g. model selection with the MDL principle, can be calculated in time $\mathcal{O}(n+m)$ (Kontkanen and Myllymäki, 2007), this may still be impractical if n or m is large. However, it directly follows from Theorem 7 that when m is large compared to n, we have

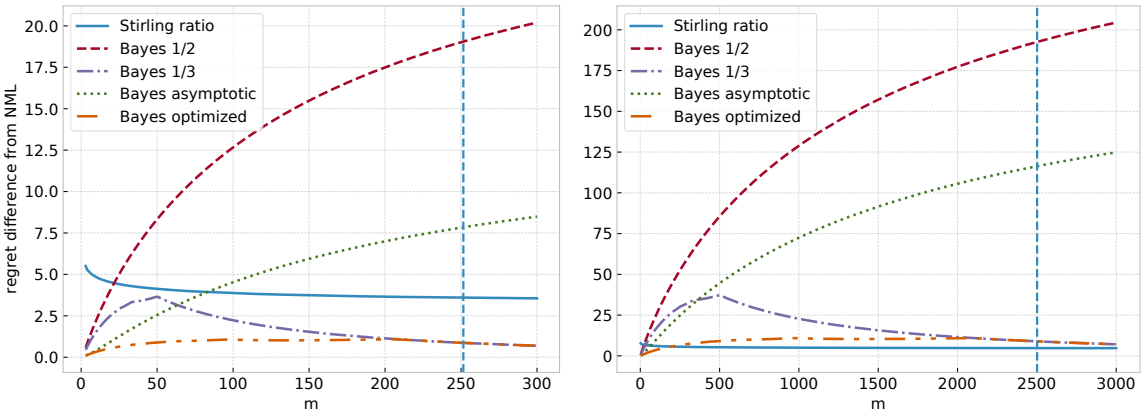
$$\log C_n^m \approx \max_{x^n} \operatorname{regret}(p_{B,\frac{1}{3}}, x^n) = \log \Gamma(n + \frac{m}{3}) - \log \Gamma(\frac{m}{3}) - n \log \frac{n}{3}.$$

That is, the NML regret $\log C_n^m$ can be approximated in constant time by the worst-case regret of the 1/3-mixture. This approximation is compared to the NML regret in Table 1 for different values of n and m. It can be seen that the approximation approaches the exact value of the NML regret as m increases.

n	m	approx	$\log C_n^m$
50	10	15.990	13.238
	100	60.555	60.004
	1000	153.292	153.276
	10000	265.282	265.281
500	100	172.606	144.029
	1000	609.691	603.928
	10000	1533.550	1533.379
	100000	2652.883	2652.881
5000	1000	1738.564	1451.782
	10000	6101.034	6043.158
	100000	15336.133	15334.406
	1000000	26528.893	26528.873

Table 1: Comparison between the approximation given by the worst-case regret of the 1/3-mixture and the NML regret for different values of n and m.

We also evaluated the actual worst-case regret of the 1/3-mixture (which differs from the approximation when $m < n$) as a function of m for $n = 100, 1000$. Furthermore, we evaluated the worst-case regrets of the Bayes procedures with $\alpha = 1/2$, the asymptotic hyperparameter $\alpha_n = 1/2 - \log 2 / \log n^2$ as given by Watanabe and Roos (2015), and the optimized hyperparameter α^*. We also include the tilted Stirling ratio distribution given by Yang and Barron (2017). These worst-case regrets are visualized in Figure 1 which shows the difference in regret from the NML regret for each distribution.



(a) $n = 100$ (b) $n = 1000$

Figure 1: Worst-case regret differences from NML regret as a function of alphabet size when (a) $n = 100$ and (b) $n = 1000$. A vertical line is shown at $\frac{5}{2}n + \frac{4}{n-2} + \frac{3}{2}$.

The worst-case regrets of the procedures given by the Jeffreys prior ($\alpha = 1/2$) and the asymptotic hyperparameter α_n stray further away from the NML regret as the alphabet size increases. The 1/3-mixture achieves lower worst-case regret already when m is of the order of $n/2$ and approaches the NML regret as m increases. The tilted Stirling ratio distribution achieves lower worst-case regret than the 1/3-mixture for larger n and moderate m, but does not approach the NML regret as $m \to \infty$ (Yang and Barron, 2017).

Similar to the Bayes procedure, the tilted Stirling ratio distribution depends on a tilting parameter which we optimized by a grid search. For the optimized Bayes procedure we used an efficient algorithm for finding the minimax optimal hyperparameter α^* for given n, m. It allows for computation of α^* with ε precision in time $\mathcal{O}(\log(\min\{n, m\}) \log(1/\varepsilon))$. The algorithm makes use of the following lemma proved by Watanabe and Roos (2015) which reduces the number of possible worst-case sequences to $\min\{n, m\}$ (when two sequences are considered the same if their count vectors $(n_1, \ldots, n_m)$ are permutations of each other and thus their regrets are equal):

Lemma 8 (Watanabe and Roos, 2015, Lemma 5) *The possible worst-case sequences in*

$$\max_{x^n} \operatorname{regret}(p_{B,\alpha}, x^n)$$

have l non-zero counts ($l = 1, 2, \ldots, m$), each of which is $\lfloor \frac{n}{l} \rfloor$ or $\lfloor \frac{n}{l} \rfloor + 1$ and all the other counts are zeros.

Since the count vector of each possible worst-case sequence contains at most two different elements, we can evaluate the regret in constant time if we know the counts of the different elements. Thus we can find the optimal α with ε precision in time $\mathcal{O}(\min\{n, m\}/\varepsilon)$ by considering all α on a grid with length ε intervals. This can be further reduced to $\mathcal{O}(\min\{n, m\} \log(1/\varepsilon))$ by using the following lemma:

Lemma 9 *If $n > 1$, the function*

$$\alpha \mapsto \max_{x^n} \operatorname{regret}(p_{B,\alpha}, x^n)$$

is unimodal on the interval $(0, \infty)$.

Proof We first consider the regret for a fixed x^n. We have

$$\frac{\partial}{\partial \alpha} \operatorname{regret}(p_{B,\alpha}, x^n) = \frac{\partial}{\partial \alpha} \log \frac{p(x^n; \hat{\theta}(x^n))}{p_{B,\alpha}(x^n)} = -\frac{\partial}{\partial \alpha} \log p_{B,\alpha}(x^n).$$

Levin and Reeds (1977) proved that when $n > 1$, the derivative of $\log p_{B,\alpha}$ with respect to α has at most one zero on the interval $(0, \infty)$ and if this happens at a finite α, the zero has to be a local maximum. Therefore the regret for any x^n as a function of α is either decreasing, increasing or decreases up to a point and increases from that point on.

All monotone functions and functions that decrease up to a point and increase from that point on are strictly quasiconvex. Since the maximum of strictly quasiconvex functions is strictly quasiconvex, the worst-case regret as a function of α is strictly quasiconvex. The claim follows from the fact that a strictly quasiconvex function is strictly unimodal. ■

Lemma 10 in Appendix A states that α^* is always on the interval $(0, 1]$. Furthermore, since the function $\max_{x^n} \operatorname{regret}(p_{B,\alpha}, x^n)$ is unimodal as a function of α, we can optimize it with an algorithm such as golden section search (Kiefer, 1953) in time $\mathcal{O}(\log(1/\varepsilon))$ on the fixed-length interval $(0, 1]$. As each evaluation of the function to be optimized by golden section search takes time $\mathcal{O}(\min\{n, m\})$, the optimal hyperparameter α^* can be found in $\mathcal{O}(\min\{n, m\} \log(1/\varepsilon))$ time with ε precision. Appendix A additionally describes how the worst-case regret can be found in $\mathcal{O}(\log(\min\{n, m\}))$ time, resulting in an $\mathcal{O}(\log(\min\{n, m\}) \log(1/\varepsilon))$ time algorithm. This algorithm can be used to find the minimax optimal α with ε precision efficiently for any practical values of n and m.

4. Discussion

The choice of α is important if we wish to achieve as low regret as possible even in the worst case, and can be critical in for example Bayesian network structure learning (Silander et al., 2007). The 1/3-mixture provides a coding distribution whose worst-case performance is almost optimal when the size of the alphabet is large. This result holds not only in the limit, but for all values of n as long as m exceeds the derived bound.

The 1/3-mixture can be useful in for example model selection with the MDL principle since its regret can be calculated in time not dependent on m, or by allowing approximation of the NML regret in constant time. The Bayes mixture also allows more efficient calculation of the marginal and conditional probabilities needed for e.g. coding and prediction than an earlier proposed tilted Stirling ratio distribution (Yang and Barron, 2017).

The minimax optimality of the $\operatorname{Dir}(1/3, \ldots, 1/3)$ prior can also serve as a theoretical justification for choosing the hyperparameters in a model with Dirichlet priors when the alphabet size is large. Application areas where large alphabets arise naturally include for example natural language processing and Bayesian networks.

There are several questions remaining for future work. One possibility is to examine the Dirichlet prior which minimizes the worst-case expected regret in the large alphabet regime. Attention should also be given to providing tight bounds between the regrets of the mixture distributions and the NML distribution. Finally, the behavior of Bayes mixtures should be studied in the large alphabet setting as building blocks in models that incorporate context.

Acknowledgments

We thank the anonymous reviewers for useful comments. E.J., J.L. and T.R. were supported by the Academy of Finland (COIN CoE and Project MachQu). J.L. was supported by the DoCS doctoral programme of the University of Helsinki.

References

J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Poissonization and universal compression of envelope classes. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT 2014)*, pages 1872–1876. IEEE, 2014.

A. R. Barron, T. Roos, and K. Watanabe. Bayesian properties of normalized maximum likelihood and its fast computation. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT 2014)*, pages 1667–1671. IEEE, 2014.

N. Batir. On some properties of digamma and polygamma functions. *Journal of Mathematical Analysis and Applications*, 328(1):452–465, 2007.

J. M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 113–147, 1979.

D. Bontemps. Universal coding on infinite alphabets: exponentially decreasing envelopes. *IEEE Transactions on Information Theory*, 57(3):1466–1478, 2011.

S. Boucheron, E. Gassiat, and M. I. Ohannessian. About adaptive coding on countable alphabets: Max-stable envelope classes. *IEEE Transactions on Information Theory*, 61 (9):4948–4967, 2015.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, UK, 2006.

S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics, 1996.

R. Eggeling, T. Roos, P. Myllymäki, and I. Grosse. Robust learning of inhomogeneous PMMs. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS 2014)*, pages 229–237, 2014.

P. Grünwald. *The Minimum Description Length Principle*. The MIT Press, Cambridge, MA, USA, 2007.

B. Guo and F. Qi. Refinements of lower bounds for polygamma functions. *Proceedings of the American Mathematical Society*, 141(3):1007–1015, 2013.

J. Kiefer. Sequential minimax search for a maximum. *Proceedings of the American mathematical society*, 4(3):502–506, 1953.

J. Kieffer. A unified approach to weak universal source coding. *IEEE Transactions on Information Theory*, 24(6):674–682, 1978.

P. Kontkanen and P. Myllymäki. A linear-time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters*, 103(6):227–233, 2007.

R. Krichevsky and V. Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.

B. Levin and J. Reeds. Compound multinomial likelihood functions are unimodal: Proof of a conjecture of I.J. Good. *The Annals of Statistics*, 5(1):79–87, 1977.

J. Määttä, D. F. Schmidt, and T. Roos. Subset selection in linear regression using sequentially normalized least squares: Asymptotic theory. *Scandinavian Journal of Statistics*, 43(2):382–395, 2016.

A. Orlitsky and N. P. Santhanam. Speaking of infinity [i.i.d. strings]. *IEEE Transactions on Information Theory*, 50(10):2215–2230, 2004.

A. Orlitsky and A. T. Suresh. Competitive distribution estimation: Why is Good-Turing good. In *Advances in Neural Information Processing Systems (NIPS 2015)*, pages 2143–2151, 2015.

A. Orlitsky, N. P. Santhanam, and J. Zhang. Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, 50(7):1469–1481, 2004.

Y. M. Shtarkov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 23(3):3–17, 1987.

T. Silander, P. Kontkanen, and P. Myllymäki. On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence (UAI 2007)*, pages 360–367. AUAI Press, 2007.

T. Silander, J. Leppä-aho, E. Jääsaari, and T. Roos. Quotient normalized maximum likelihood criterion for learning Bayesian network structures. 2018. Accepted for the *Twenty-First International Conference on Artificial Intelligence and Statistics (AISTATS 2018)*.

J. Suzuki. A theoretical analysis of the BDeu scores in Bayesian network structure learning. *Behaviormetrika*, 44(1):97–116, 2017.

W. Szpankowski and M. J. Weinberger. Minimax pointwise redundancy for memoryless models over large alphabets. *IEEE Transactions on Information Theory*, 58(7):4094–4104, 2012.

J. Takeuchi and A. R. Barron. Asymptotically minimax regret by Bayes mixtures. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT 1998)*, page 318. IEEE, 1998.

K. Watanabe and T. Roos. Achievability of asymptotic minimax regret by horizon-dependent and horizon-independent strategies. *Journal of Machine Learning Research*, 16(1):2357–2375, 2015.

Q. Xie and A. R. Barron. Minimax redundancy for the class of memoryless sources. *IEEE Transactions on Information Theory*, 43(2):646–657, 1997.

Q. Xie and A. R. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445, 2000.

X. Yang and A. R. Barron. Minimax compression and large alphabet approximation through Poissonization and tilting. *IEEE Transactions on Information Theory*, 63(5):2866–2884, 2017.

Appendix A. Fast computation of the optimal hyperparameter

Lemma 10 *The minimax optimal hyperparameter α^* is always on the interval $(0, 1]$.*

Proof We first prove that for the coding distribution $p_{B,\alpha}$ with $\alpha \geq 1$, the worst-case regret is always given by a sequence where a single symbol occurs n times. Let x^n be any sequence where two symbols u and v occur $n_u \geq 0$ and $n_v \geq 0$ times, respectively. Furthermore, let y^n be the same sequence as x^n except all occurrences of v have been replaced by u. Thus the symbol u occurs in y^n in total $N = n_u + n_v$ times. Now, assuming fixed N, the difference in regret between the sequences y^n and x^n as a function of n_v is

$$w(n_v) = \text{regret}(p_{B,\alpha}, y^n) - \text{regret}(p_{B,\alpha}, x^n) = q_\alpha(N) + q_\alpha(0) - q_\alpha(n_v) - q_\alpha(N - n_v),$$

where we denote

$$q_\alpha(k) := k \log k - \log \Gamma(k + \alpha).$$

Due to symmetry of w, we can restrict its domain to be $0 \leq n_v \leq N/2$. Taking the second derivative of w, we have

$$\frac{d^2}{dn_v^2} w(n_v) = \psi'(\alpha + N - n_v) - \frac{1}{N - n_v} + \psi'(\alpha + n_v) - \frac{1}{n_v}.$$

Using the inequality $\psi'(x) < 1/x + 1/x^2$ (e.g., Guo and Qi, 2013), it is straightforward to show that the second derivative is always negative for $\alpha \geq 1$. Furthermore, since $w'(\frac{N}{2}) = 0$, the derivative is positive for all $\alpha \geq 1$ and w is minimized at $n_v = 0$, where $w(0) = 0$. Thus w is always non-negative and $\text{regret}(p_{B,\alpha}, y^n) \geq \text{regret}(p_{B,\alpha}, x^n)$. Applying the above procedure repeatedly to any sequence x^n, we get that the sequence where a single symbol occurs n times has higher or equal regret than the arbitrary sequence x^n.

Finally, let x^n be any sequence where a single symbol occurs n times. Since

$$\begin{aligned}
\frac{\partial}{\partial \alpha} \text{regret}(p_{B,\alpha}, x^n) &= m(\psi(n + m\alpha) - \psi(m\alpha)) - (\psi(n + \alpha) - \psi(\alpha)) \\
&= m \left(\sum_{k=0}^{n-1} \frac{1}{m\alpha + k} \right) - \sum_{k=0}^{n-1} \frac{1}{\alpha + k} \\
&= \sum_{k=0}^{n-1} \frac{1}{\alpha + \frac{k}{m}} - \frac{1}{\alpha + k} > 0,
\end{aligned}$$

we have that the worst-case regret always increases as a function of α on the interval $[1, \infty)$. Thus $\max_{y^n} \text{regret}(p_{B,1}, y^n) < \max_{y^n} \text{regret}(p_{B,\alpha}, y^n)$ for all $\alpha > 1$ and the α that minimizes the worst-case regret has to belong to the interval $(0, 1]$. ■

Given n, m and α, we now describe a way to perform the search for the worst-case sequence of the Bayes mixture $p_{B,\alpha}$ in time $\mathcal{O}(\log(\min\{n, m\}))$. Combined with golden section search (see Section 3.2), this yields an $\mathcal{O}(\log(\min\{n, m\}) \log(1/\varepsilon))$ time algorithm for finding the optimal hyperparameter with ε precision. For convenience, we allow the counts n_j to be any non-negative real numbers as long as $\sum_{j=1}^{m} n_j = n$. We also denote any sequence with a count vector consisting of a symbols with count x, b symbols with count $x + 1$ and the remaining $m - b - a$ symbols with count zero as $x_{a,b}^n$.

Consider now the following function which represents the difference in regret by replacing x/y y counts in a count vector with a single x count:

$$h_\alpha(x,y) = x\log\frac{x}{y} - \log\Gamma(x+\alpha) + \frac{x}{y}\log\Gamma(y+\alpha) + (1-\frac{x}{y})\log\Gamma(\alpha).$$

This function has the following properties which are straightforward to verify:

1. $h_\alpha(x,x) = 0$
2. $h_\alpha(x,y) = -\frac{x}{y}h_\alpha(y,x)$
3. $ah_\alpha(x,y) + bh_\alpha(x+1,y) = \text{regret}(p_{B,\alpha}, x^n_{a,b}) - \text{regret}(p_{B,\alpha}, y^n_{c,0})$, where $c = n/y$.

A key observation is described in the following lemma (Watanabe and Roos, 2015):

Lemma 11 *The second derivative $\frac{\partial^2}{\partial x^2}h_\alpha(x,y)$ has at most one zero.*

Proof We have

$$\frac{\partial^2}{\partial x^2}h_\alpha(x,y) = \frac{1}{x} - \psi'(x+\alpha).$$

Using the inequality $(\psi'(x))^2 + \psi''(x) > 0$ given by Batir (2007), we can prove that the second derivative has at most one zero since the derivative

$$\frac{d}{dx}\left(x - \frac{1}{\psi'(x+\alpha)}\right) = \frac{\psi''(x+\alpha)}{\psi'(x+\alpha)^2} + 1$$

is positive, meaning that the function $x - 1/\psi'(x+\alpha)$ is increasing from $-1/\psi'(\alpha) < 0$ and thus has at most one zero coinciding with the zero of $\frac{\partial^2}{\partial x^2}h_\alpha(x,y)$. ■

Since we have $\lim_{x\to 0^+}\frac{\partial^2}{\partial x^2}h_\alpha(x,y) = \infty$, it follows from Lemma 11 that the function $x \mapsto h_\alpha(x,\cdot)$ is either convex on the whole interval $(0,\infty)$ or convex up to an inflection point c and concave from that point on. Let there now be the sequences $\ell_1{}^n_{a_1,b_1}, \ell_2{}^n_{a_2,b_2}, \ldots, \ell_t{}^n_{a_t,b_t}$, where $\ell_1,\ldots,\ell_t \in \{\lfloor \frac{n}{l} \rfloor : l = 1,2,\ldots,m\} =: \mathcal{L}$ and $\ell_1 < \ell_2 < \cdots < \ell_t$. Furthermore, we let ℓ_c be the largest $\ell_i \in \mathcal{L}$ such that $\ell_i < c$ and $\ell_d \in \mathcal{L}$ be such that $\ell_d > \ell_c$ and $h_\alpha(x,\ell_d) \le 0$ for all $c \le x \le \ell_d - 1$ and $x \ge \ell_d + 1$. In particular, this means that replacing ℓ_j/ℓ_d ℓ_d counts in a count vector with a single $\ell_j \neq \ell_d$ count would result in lower or equal regret. Consider now the following lemma:

Lemma 12 *Let c be such that the function $t \mapsto h_\alpha(t,\cdot)$ is concave for all $t \ge c$ and $x,z \in \mathbb{N}, y \in \mathbb{R}$ be such that $c \le x < x+1 \le y < y+1 \le z$. If $h_\alpha(t,z+1) \le 0$ for all $c \le t \le z$ and there exist sequences $x^n_{a,b}, y^n_{\cdot,0}, z^n_{d,e}$, then $\text{regret}(p_{B,\alpha}, x^n_{a,b}) \le \text{regret}(p_{B,\alpha}, z^n_{d,e})$. Moreover, if we have $c \le z < z+1 \le y < x$ and $h_\alpha(t,z) \le 0$ for all $t \ge z+1$, then $\text{regret}(p_{B,\alpha}, x^n_{a,b}) \le \text{regret}(p_{B,\alpha}, z^n_{d,e})$.*

Proof For the first part, we have $h_\alpha(y,z+1) \le 0$ and thus $h_\alpha(z+1,y) \ge 0$. Moreover, $h_\alpha(y,y) = 0$ and the function $t \mapsto h_\alpha(t,\cdot)$ is concave, and thus also $h_\alpha(z,y) \ge 0$. Hence

$$\text{regret}(p_{B,\alpha}, z^n_{d,e}) - \text{regret}(p_{B,\alpha}, y^n_{c,0}) = dh_\alpha(z,y) + eh_\alpha(z+1,y) \ge 0.$$

Again, by concavity we have $h_\alpha(x, y) \leq 0$ and $h_\alpha(x+1, y) \leq 0$. Therefore

$$\begin{aligned}\text{regret}(p_{B,\alpha}, x^n_{a,b}) - \text{regret}(p_{B,\alpha}, z^n_{d,e}) &\leq \text{regret}(p_{B,\alpha}, x^n_{a,b}) - \text{regret}(p_{B,\alpha}, y^n_{c,0}) \\ &= a h_\alpha(x, y) + b h_\alpha(x+1, y) \leq 0\end{aligned}$$

The other part follows from a similar argument. ■

In particular, we have $\ell_c < \ell_{c+1} < \cdots < \ell_{d-3} < \ell_{d-3} + 1 \leq \ell_{d-2} \leq y < y + 1 \leq \ell_{d-1}$, where $y \in [\ell_{d-2}, \ell_{d-2} + 1)$ and there exists a sequence $y^n_{\cdot,0}$ since $\ell_{d-2} = \lfloor n/l \rfloor$ for some l and thus $y = n/l$. Since now $h_\alpha(x, \ell_d) \leq 0$ for all $c \leq x \leq \ell_d - 1$ and $h_\alpha(\ell_d, \ell_d) = 0$, by concavity of $t \mapsto h_\alpha(t, \cdot)$, we have $h_\alpha(t, \ell_{d-1} + 1) \leq 0$ for all $c \leq t \leq \ell_{d-1}$. Lemma 12 then verifies that we only need to check the regrets of the sequences corresponding to ℓ_{d-2}, ℓ_{d-1} and ℓ_d to find the maximum regret amongst the sequences corresponding to $\ell_{c+1}, \ell_{c+2}, \ldots, \ell_{d-1}, \ell_d$.

Consider now the remaining sequences corresponding to $\ell_{d+1}, \ell_{d+2}, \ldots, \ell_t$ in the concave region. In this region we have $\ell_d < \ell_d + 1 \leq \ell_{d+1} \leq y < \ell_{d+2} < \ell_{d+3} < \cdots < \ell_t$, where $y \in [\ell_{d+1}, \ell_{d+1} + 1)$ and there exists a sequence $y^n_{\cdot,0}$. As $h_\alpha(x, \ell_d) \leq 0$ for all $x \geq \ell_d + 1$, Lemma 12 shows that the maximum regret is given by ℓ_d or ℓ_{d+1} on this interval.

Putting these together, in the concave region, the maximum regret is achieved by one of the sequences corresponding to $\ell_{d-2}, \ell_{d-1}, \ell_d$ or ℓ_{d+1}. That is, given ℓ_d, we only need to examine a constant number of cases to find the maximum regret. We now show that such ℓ_d always exists in the concave region and finding it can be done in logarithmic time. Consider first the following lemma:

Lemma 13 *Let $n \in \mathbb{N}$ and $c \in \mathbb{R}$ be such that $x \mapsto h_\alpha(x, y)$ is concave for all $c \leq x \leq n$.*

- *If there exists a smallest $z \in \mathbb{N}$ such that $c \leq z < n$ and $h_\alpha(z+1, z) \leq 0$, then $h_\alpha(x, z) \leq 0$ for all $c \leq x \leq z - 1, x \geq z + 1$ and $h_\alpha(x+1, x) \leq 0$ for all $x \geq z + 1$.*
- *If such integer z does not exist, then $h_\alpha(x, n) \leq 0$ for all $c \leq x \leq n - 1$.*

Proof Assume that there exists a smallest $z \in \mathbb{N}$ such that $c \leq z < n$ and $h_\alpha(z+1, z) \leq 0$.

- Assume $x \geq z+1$. Since $h_\alpha(z, z) = 0$ and $h_\alpha(z+1, z) \leq 0$, by concavity of $x \mapsto h_\alpha(x, y)$ we must have $h_\alpha(x, z) \leq 0$ and $h_\alpha(z, x) \geq 0$. Since also $h_\alpha(x, x) = 0$, by concavity of $x \mapsto h_\alpha(x, y)$ we must have $h_\alpha(x+1, x) \leq 0$.
- Assume $c \leq x \leq z - 1$. We must have $h_\alpha(z-1, z) \leq 0$, as otherwise $h_\alpha(z, z-1) \leq 0$ which is a contradiction since z is the smallest integer such that $h_\alpha(z+1, z) \leq 0$. Thus by concavity $h_\alpha(x, z) \leq 0$ for all $x \leq z - 1$ since $h_\alpha(z, z) = 0$.

If the smallest $z \in \mathbb{N}$ does not exist, we have $h_\alpha(n, n-1) > 0$ and thus $h_\alpha(n-1, n) < 0$. By the fact that $h_\alpha(n, n) = 0$ and concavity of the function $t \mapsto h_\alpha(t, n)$, we now have $h_\alpha(x, n) < 0$ for all $c \leq x \leq n - 1$. ■

As Lemma 13 states that if ℓ_d is the smallest integer for which $h_\alpha(\ell_d + 1, \ell_d) \leq 0$, then $h_\alpha(x+1, x) \leq 0$ for all $x \geq \ell_d + 1$, we can find ℓ_d in time $\mathcal{O}(\log n)$ by a binary search like routine (Algorithm 1). This algorithm returns the first integer z on the range `[start, end]` such that $f(z)$ is true, assuming that $f(x)$ is false for all $x < z$ and true for all $x \geq z$.

Algorithm 1: Finding smallest z such that $f(x)$ is false for $x < z$ and true for $x \geq z$

Function `BIN`$(f, \text{start}, \text{end})$
  $l, h \leftarrow \text{start}, \text{end} + 1$
  while $l \neq h$ **do**
    $m \leftarrow \lfloor (l+h)/2 \rfloor$
    if *not* $f(m)$ **then** $l \leftarrow m + 1$
    else $h \leftarrow m$
  end
  return l

Finally, we note that in the convex region, the highest regret is achieved at the boundaries of the interval, i.e. by one of the sequences corresponding to $\ell_1, \ell_2, \ell_{c-1}$ or ℓ_c. The proof is identical to the one in the concave region, except it uses the fact that if $x \mapsto h_\alpha(x, \cdot)$ is concave, then $x \mapsto -h_\alpha(x, \cdot)$ is convex. We omit the details here for brevity.

From Lemma 8, we know that the possible worst-case sequences are of the form $\ell^n_{a,b}$. The previous lemmas can then be used to formulate the final procedure which is described in Algorithm 3. Given n, m and α, the function `F` finds the worst-case regret. The algorithm first uses the `BIN` routine to find the smallest integer c such that the function $x \mapsto h_\alpha(x, \cdot)$ is concave for all $x \geq c$. On the next line, the algorithm uses `BIN` to find the smallest integer y in the concave region such that $h(y+1, y)$ is negative. The variable ℓ_d is subsequently set as the smallest $\ell \geq y$ such that there exists a sequence $\ell^n_{a,b}$ for some $a, b \in \mathbb{N}_0$.

The functions `PREV` and `NEXT` find for a given k the previous (next) x such that there exists a sequence $x^n_{a,b}$ for some $a, b \in \mathbb{N}_0$. Using these functions, the maximum regret can be found by considering all the possible cases. The regret for each case is calculated in constant time by Algorithm 2. Since there can be multiple count vectors that consist of the integers k and $k+1$, the sign of $h_\alpha(k, k+1)$ is checked by `REGRET`. It is easy to verify that if $h_\alpha(k, k+1) < 0$, the count vector with maximum amount of k counts should be preferred, and the count vector with maximum amount of $k+1$ counts otherwise.

The `MINIMAX` function uses golden section search to minimize the maximum regret `F`(n, m, α) on the fixed-length interval $(0, 1]$. Since the `BIN` routine works in time $\mathcal{O}(\log n)$ and all other operations are constant time operations, this yields an $\mathcal{O}(\log(n) \log(1/\varepsilon))$ time algorithm. However, both of the binary searches can also be performed by considering only the numbers $\lfloor n/m \rfloor, \lfloor n/(m-1) \rfloor, \ldots, n$ as possible inputs for the function f, which takes $\mathcal{O}(\log m)$ time. Thus the total time complexity is $\mathcal{O}(\log(\min\{n, m\}) \log(1/\varepsilon))$.

Algorithm 2: Calculating the regret in constant time

Function `REGRET`(n, m, k, α)
- **if** $h_\alpha(k, k+1) < 0$ **then**
  - let x^n be any sequence with a count vector consisting of k and $k+1$ with the maximum possible amount of k counts
- **end**
- **else**
  - let x^n be any sequence with a count vector consisting of k and $k+1$ with the maximum possible amount of $k+1$ counts
- **end**
- **return** regret$(p_{B,\alpha}, x^n)$

Algorithm 3: Finding the minimax optimal hyperparameter α^*

Function `F`(n, m, α)
- **Function** `PREV`(k)
  - **if** $k \leq 1$ **then**
    - **return** 0
  - **end**
  - **return** largest $x \in \mathbb{N}$ such that $x < k$ and $ax + b(x+1) = n$ for some $a, b \in \mathbb{N}_0$
- **Function** `NEXT`(k)
  - **if** $k \geq n$ **then**
    - **return** n
  - **end**
  - **return** smallest $x \in \mathbb{N}$ such that $x > k$ and $ax + b(x+1) = n$ for some $a, b \in \mathbb{N}_0$
- $c \leftarrow$ `BIN`$(f(x) := 1/x - \psi'(x+\alpha) < 0, 1, n)$
- $y \leftarrow$ `BIN`$(f(x) := h_\alpha(x+1, x) \leq 0, \max(c, \lfloor n/m \rfloor), n)$
- $\ell_d \leftarrow$ smallest $\ell \in \mathbb{N}$ such that $\ell \geq y$ and $a\ell + b(\ell+1) = n$ for some $a, b \in \mathbb{N}_0$
- maxregret $\leftarrow 0$
- **for** $k \in \{\ell_d,$ `NEXT`(ℓ_d), `PREV`(ℓ_d), `PREV`$($`PREV`$(\ell_d))$, `PREV`(c), `PREV`$($`PREV`$(c))$, $\lfloor n/m \rfloor$, `NEXT`$(\lfloor n/m \rfloor)\}$ **do**
  - **if** $\max(1, \lfloor n/m \rfloor) \leq k \leq n$ **then**
    - maxregret $\leftarrow \max($maxregret, `REGRET`$(n, m, k, \alpha))$
  - **end**
- **end**
- **return** maxregret

Function `MINIMAX`(n, m, α)
- $\alpha^* \leftarrow$ optimize $\alpha \mapsto$ `F`(n, m, α) with GSS on the range $(0, 1]$ with ε precision
- **return** α^*