Minimax Compression and Large Alphabet Approximation Through Poissonization and Tilting

Xiao Yang, *Student Member, IEEE*, and Andrew R. Barron, *Fellow, IEEE*

***Abstract*— This paper introduces a convenient strategy for coding and predicting sequences of independent, identically distributed random variables generated from a large alphabet of size m. In particular, the size of the sample is allowed to be variable. The employment of a Poisson model and tilting method simplifies the implementation and analysis through independence. The resulting strategy is optimal within the class of distributions satisfying a moment condition, and it is close to optimal for the class of all i.i.d distributions on strings of a given length. The method also can be used to code and predict strings with a condition on the tail of the ordered counts, and it can be applied to distributions in an envelope class. Moreover, we show that our model permits exact computation of the minimax optimal code, for all alphabet sizes, when conditioning on the size of the sample.**

***Index Terms*— Large alphabet, minimax regret, normalized maximum likelihood, Poisson distribution, power law, universal coding, Zipf's law.**

I. Introduction

LARGE alphabet compression and prediction problems concern understanding the probabilistic scheme of a huge number of possible outcomes. In many cases the ordered probability of individual outcomes displays a quickly falling shape, with a small number of outcomes happening most often. An example is Chinese character. A dictionary [1] containing 85568 Chinese characters in total [2] only has a few thousand that are frequently used. Here we consider an i.i.d model for this problem. Despite the possible dependence among the symbols in an alphabet like in language, it serves as a start and can be extended to models that consider dependent relationships. Some efforts to investigate alphabets with symbols having dependency with each other are included in [3].

Most source codes assume that the length of the source text is known (to the encoder and decoder) or assume that the first step in encoding is to describe the source length. Here we will work with a model that has a distribution for the source length N and show that it has desirable properties of computation and analysis both when conditioned on N=n and unconditionally. The reason is that with a suitable (Poisson) distribution for N, the counts that were dependent conditionally become independent unconditionally. Here a suitable universal distribution for independent counts is derived with a simple exact expression. The use of independent counts permits demonstration of near optimal properties for large alphabet settings. Meanwhile, with conditioning on the sample size, our model is shown to exactly match the Shtarkov conditionally minimax optimal distribution for all alphabet sizes and to provide a computationally feasible means to exactly compute the Shtarkov conditionals required for optimal coding.

Suppose a string of random variables $\underline{X} = (X_1, \ldots, X_N)$ is generated independently from a discrete alphabet $\mathcal{A}$ of size m. We allow the string length N to be variable. Thus $\underline{X}$ is a member of the set $\mathcal{X}^*$ of all finite length strings

$$\mathcal{X}^* = \bigcup_{n=0}^{\infty} \mathcal{X}^n$$

$$= \bigcup_{n=0}^{\infty} \{x^n = (x_1, \ldots, x_n) : x_i \in \mathcal{A}, i = 1, \ldots, n\}.$$

Our goal is to code/predict the string $\underline{X}$. Note that the length N is determined by the string. Our model for the data will incorporate a distribution of N, though we will also examine the case it is conditioned on a specific value.

Now suppose given N, each random variable X_i is generated independently according to a probability mass function in a parametric family $\mathcal{P}_\Theta = \{P_{\underline{\theta}}(x) : \underline{\theta} \in \Theta \subset R^m\}$ on $\mathcal{A}$. Thus

$$P_{\underline{\theta}}(X_1, \ldots, X_N | N = n) = \prod_{i=1}^{n} P_{\underline{\theta}}(X_i)$$

for $n = 1, 2, \ldots$ Of particular interest is the class of all distributions with $P_{\underline{\theta}}(j) = \theta_j$ parameterized by the simplex $\Theta = \{\underline{\theta} = (\theta_1, \ldots, \theta_m) : \theta_j \geq 0, \sum_{j=1}^{m} \theta_j = 1, j = 1, \ldots, m\}$.

As is familiar in universal coding, the normalized maximum likelihood (NML) distribution defined as $Q^*_{nml}(\underline{X}|N = n) = \max_{\underline{\theta} \in \Theta} P_{\underline{\theta}}(\underline{X}|N = n)/C^*_{m,n}$ provides the unique pointwise minimax strategy when the value $C^*_{m,n} = \sum_{\underline{X}} \max_{\underline{\theta} \in \Theta} P_{\underline{\theta}}(\underline{X}|N = n)$ is finite, and $\log C^*_{m,n}$ is the minimax regret. Coding and prediction of sequences of random variables usually involves computing conditionals of $X_{i+1}|X_1, \ldots, X_i$ as consecutive ratios of its marginals [4], [5]. This task is generally hard since the marginalization requires a sum of order m^n, which appears to take exponential time in n. A linear time algorithm (in n) for computing the NML is proposed in [6], but it is not practically useful when the alphabet size m is large. Bayes-like

Manuscript received October 1, 2013; revised March 7, 2015 and
AQ:1 August 21, 2016; accepted December 17, 2016. This paper was presented at 2013 the Sixth Workshop on Information Theoretic Methods in Science and Engineering and Center for Science of Information at Purdue University 2013.
AQ:2 The authors are with Yale University, New Haven, CT 06520 USA (e-mail: graceyx.scut@gmail.com; andrew.barron@yale.edu).
Communicated by Y. Oohama, Associate Editor for Source Coding.
Digital Object Identifier 10.1109/TIT.2017.2661990

representation of NML has been found which makes possible an easy computation of NML, but only moderate size m is computationally feasible at this point [7]. Alternatively, one can use the Krichevsky-Trofimov's method [8], which is the mixture with respect to the $Dirichlet(1/2, \ldots, 1/2)$ prior, to approximate the NML distribution. It has been shown that the Krichevsky-Trofimov probability assignment achieves regret which matches the asymptotic minimax value (to within $o(1)$) when θ lies in the interior of the parameter space and has a higher regret (by a $O(m)$ term) for boundary points [5]. As a reviewer points out, examination of Equation (2.3) in [8] shows that the regret matches $((m-1)/2)\log(n/m)$ to within a $O(m)$ error when $m = o(n)$. For $m \gg \log n$, we aim to do much better, with regret that differs from the conditional optimum by not more than $(1/2)\log n$. The distribution on the counts induced by the $Dirichlet(1/2, \ldots, 1/2)$ has the right behavior when the counts are large. But when many of the counts are small, as is the case when m is of order n or larger, we target a better level of performance, matching that of the NML distribution, but with a computationally feasible distributional set-up. We accomplish these aims by applying two tools: one is the factorization of the coding distribution of the string into a product of the distribution of the counts and the string given the counts. The distribution of the latter is uniform in accordance with the sufficiency of the counts. The other is a tilted Stirling ratio distribution which we introduce here. It simplifies the encoding of the counts as discussed later, it has suitable regret properties, and it agrees with the minimax optimal NML conditionally.

Let $\underline{N} = (N_1, \ldots, N_m)$ denote the vector of counts for symbols $1, \ldots, m$. The domain of the counts is denoted $\mathcal{N}^m = \{(N_1, \ldots, N_m) : N_i \geq 0, i = 1, \ldots, m\}$. The observed sample size N is the sum of the counts $N = \sum_{j=1}^m N_j$. Both $P_{\underline{\theta}}(\underline{X})$ and $P_{\underline{\theta}}(\underline{X}|N = n)$ have factorizations based on the distribution of the counts

$$P_{\underline{\theta}}(\underline{X}|N = n) = P(\underline{X}|\underline{N})\, P_{\underline{\theta}}(\underline{N}|N = n),$$

and

$$P_{\underline{\theta}}(\underline{X}) = P(\underline{X}|\underline{N})\, P_{\underline{\theta}}(\underline{N}).$$

The first factor of the two equations is the uniform distribution on the set of strings with given counts, which does not depend on $\underline{\theta}$. The vector of counts $\underline{N}$ forms a sufficient statistic for $\underline{\theta}$. Modeling the distribution of the counts is essential for forming codes and predictions. In the particular case of all i.i.d. distributions parameterized by the simplex, the distribution $P_{\underline{\theta}}(\underline{N}|N = n)$ is the $multinomial(n, \underline{\theta})$ distribution.

In the above, there is a need for a distribution of the total count N. Of particular interest is the case that the total count is taken to be $Poisson$, because then the resulting distribution of individual counts makes them independent [9].

Accordingly, we give particular attention to the target family $\mathcal{P}_\Lambda^m = \{P_{\underline{\lambda}}(\underline{N}) : \lambda_j \geq 0,\ j = 1, \ldots, m\}$, in which $P_{\underline{\lambda}}(\underline{N})$ is the product of $Poisson(\lambda_j)$ distribution for N_j, $j = 1, \ldots, m$. It makes the total count $N \sim Poisson(\lambda_{sum})$ with $\lambda_{sum} = \sum_{j=1}^m \lambda_j$ and yields the $multinomial(n, \underline{\theta})$ distribution by conditioning on $N = n$, where $\theta_j = \lambda_j/\lambda_{sum}$. And the induced distribution on $\underline{X}$ is

$$P_{\underline{\lambda}}(\underline{X}) = P(\underline{X}|\underline{N})P_{\underline{\lambda}}(\underline{N}).$$

The task of coding a string is equivalent to providing a probabilistic scheme. A coder Q for the string could also be a (sub)probability distribution on $\mathcal{X}^*$ which assigns a probability $Q(\underline{X})$ to each string $\underline{X}$ and produces a binary string of length $\log 1/Q(\underline{X})$ (we do not worry about the integer constraint). Ideally the true probability distribution $P_\lambda(\underline{X})$ could be used if $\underline{\lambda}$ were known, as it produces no extra bits for coding purpose. The *regret* induced by using Q instead of $P_{\underline{\lambda}}$ is

$$R(Q, P_{\underline{\lambda}}, \underline{X}) = \log \frac{1}{Q(\underline{X})} - \log \frac{1}{P_{\underline{\lambda}}(\underline{X})},$$

where log is logarithm base 2. Likewise, the *expected regret* is

$$r(Q, P_{\underline{\lambda}}) = \mathbf{E}_{P_{\underline{\lambda}}}\left(\log \frac{1}{Q(\underline{X})} - \log \frac{1}{P_{\underline{\lambda}}(\underline{X})}\right).$$

In universal coding the expected regret is also called the *redundancy*.

Here we can construct Q by choosing a probability distribution for the counts and then use the uniform distribution for the distribution of strings given the counts, written as P_{unif}. That is

$$Q(\underline{X}) = P_{unif}(\underline{X}|\underline{N})Q(\underline{N}).$$

Then the regret becomes the log ratio of the counts probability

$$\begin{aligned} R(Q, P_{\underline{\lambda}}, \underline{X}) &= \log \frac{P_{\underline{\lambda}}(\underline{N})}{Q(\underline{N})} \\ &= R(Q, P_{\underline{\lambda}}, \underline{N}). \end{aligned}$$

And the redundancy becomes

$$r(Q, P_{\underline{\lambda}}) = \mathbf{E}_{P_{\underline{\lambda}}} \log \frac{P_{\underline{\lambda}}(\underline{N})}{Q(\underline{N})}.$$

In the pointwise regret story, the set of codelengths $\log(1/P_{\underline{\lambda}}(\underline{X}))$ provides a standard with which our coder is to be compared. Given the family $\mathcal{P}_\Lambda^m$, consider the best candidate with hindsight $P_{\hat{\underline{\lambda}}}(\underline{X})$, which achieves the maximum value, $P_{\hat{\underline{\lambda}}}(\underline{X}) = \max_{\underline{\lambda} \in \Lambda}(P_{\underline{\lambda}}(\underline{X}))$ (corresponding to $\min_{\underline{\lambda} \in \Lambda} \log(1/P_{\underline{\lambda}}(\underline{X}))$), where $\hat{\underline{\lambda}}$ is the maximum likelihood estimator of $\underline{\lambda}$, and compare it to our strategy $Q(\underline{X})$. The maximization is equivalent to maximizing $\underline{\lambda}$ for the count probability, as the uniform distribution does not depend on λ, i.e.

$$\begin{aligned} \max_{\underline{\lambda} \in \Lambda}(P_{\underline{\lambda}}(\underline{X})) &= P_{unif}(\underline{X}|\underline{N}) \max_{\underline{\lambda} \in \Lambda} P_{\underline{\lambda}}(\underline{N}) \\ &= P_{unif}(\underline{X}|\underline{N})\, P_{\hat{\underline{\lambda}}}(\underline{N}). \end{aligned}$$

Moreover, the maximum likelihood estimate is $\hat{\underline{\lambda}} = \underline{N}$. Then the problem becomes: given the family $\mathcal{P}_\Lambda^m$, how to choose Q to minimize the maximized regret

$$\min_Q \max_{\underline{X}} R(Q, P_{\hat{\underline{\lambda}}}, \underline{X}) = \min_Q \max_{\underline{N}} \log \frac{P_{\hat{\underline{\lambda}}}(\underline{N})}{Q(\underline{N})},$$

or the redundancy,

$$\min_Q \max_{P_{\underline{\lambda}} \in \mathcal{P}_\Lambda^m} r(Q, P_{\underline{\lambda}}) = \min_Q \max_{P_{\underline{\lambda}} \in \mathcal{P}_\Lambda^m} \mathbf{E}_{P_{\underline{\lambda}}} \log \frac{P_{\underline{\lambda}}(\underline{N})}{Q(\underline{N})}.$$

For the regret, the maximum can be restricted to a set of counts instead of the whole space $\mathcal{N}^m$. A traditional choice being $S_{m,n} = \{(N_1, \dots, N_m) : \sum_{j=1}^m N_j = n, N_j \geq 0, j = 1, \dots, m\}$ associated with a given sample size n, in which case the minimax regret is

$$\min_Q \max_{\underline{N} \in S_{m,n}} \log \frac{P_{\hat{\lambda}}(\underline{N})}{Q(\underline{N})}.$$

The normalized maximum likelihood distribution

$$Q_{nml}(\underline{N}) = \frac{P_{\hat{\lambda}}(\underline{N})}{C(S_{m,n})} \mathbf{1}_{\{\underline{N} \in S_{m,n}\}}$$

provides the unique pointwise minimax strategy for coding and predicting the counts given $C(S_{m,n}) = \sum_{\underline{N} \in S_{m,n}} P_{\hat{\lambda}}(\underline{N})$ being finite in accordance with [4]. Again, we have $\log C(S_{m,n})$ as the minimax regret.

We introduce a coding distribution that makes the counts independent. Because it lives on the whole space $\mathcal{N}^m$, it is suboptimal on each $S_{m,n'}$. Nevertheless, we show that it is nearly optimal for every $S_{m,n'}$ with n' not too different from a target n. Moreover, our simple coding distribution may be preferable to use computationally when m is large even if the sample size n were known in advance.

To produce our desired coding distribution we make use of some basic principles. One is that the multinomial family of distributions on counts matches the conditional distribution of $N_1, \dots, N_m$ given the sum N when unconditionally the counts are independent Poisson. Another is the information theory principle [10]–[12] that the conditional distribution given a sum (or average) of a large number of independent random variables is approximately a product of distributions, each of which is the one closest in relative entropy to the unconditional distribution subject to an expectation constraint. This minimum relative entropy distribution is an exponential tilting of the unconditional distribution.

In the Poisson family with distribution $\lambda_j^{N_j} e^{-\lambda_j}/N_j!$, exponential tilting (multiplying by the factor e^{-aN_j}) preserves the Poisson family (with the parameter scaled to $\lambda_j e^{-a}$). Those distributions continue to correspond to the multinomial distribution (with parameters $\theta_j = \lambda_j/\lambda_{sum}$) when conditioning on the sum of counts N. A particular choice of $a = \ln(\lambda_{sum}/N)$ provides the product of Poisson distributions closest to the multinomial in regret. Here for universal coding, we find the tilting of individual maximized likelihood that makes the product of such closest to the Shtarkov's NML distribution. This greatly simplifies the task of approximate optimal universal compression and the analysis of its regret.

Indeed, applying the maximum likelihood step to a Poisson count k produces a maximized likelihood value of $M(k) = k^k e^{-k}/k!$. We call this maximized likelihood the Stirling ratio, as it is the quantity that Stirling's approximation shows near $(2\pi k)^{-1/2}$ for k not too small. We find that this $M(k)$ plays a distinguished role in universal large alphabet compression, even for sequences with small counts k. This measure M has a product extension to counts $N_1, N_2, \dots, N_m$,

$$M^m(\underline{N}) = M(N_1)M(N_2)\cdots M(N_m).$$

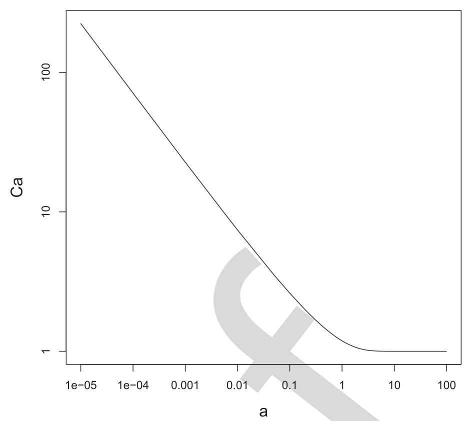

Fig. 1. Relationship between C_a and a.

Although M has an infinite sum by itself, it is normalizable when tilted for every positive a. Our model for universal coding is to arrange i.i.d. counts, where the probability distribution for the $N_1, \dots, N_m$ is given by what we call the tilted Stirling ratio distribution

$$P_a(k) = \frac{k^k e^{-k}}{k!} \frac{e^{-ak}}{C_a}, \quad (1)$$

for $k = 0, 1, 2, \dots$, with the normalizer $C_a = \sum_{k=0}^{\infty} k^k e^{-(1+a)k}/k!$. Figure 1 illustrates how C_a decreases with respect to a. For each k, the numerator (before normalizing by C_a) can be calculated by adding $k\log(1+1/k)-1-a$ to the previous one on the natural logarithm scale. The individual terms in C_a behave like $e^{-ak}/\sqrt{k}$. So the series is exponentially convergent, and accurately computed by stopping at k large compared to $1/a$.

The coding distribution we propose and analyze is simply the product of those tilted one-dimensional maximized Poisson likelihood distributions for a value of a we will specify later

$$Q_a(\underline{N}) = P_a^m(\underline{N}) = P_a(N_1)\cdots P_a(N_m).$$

By allowing description of all possible counts $N_j \geq 0$, $j = 1, \dots, m$, our codelength will be greater for some strings than codelengths designed for the case of a given sum $N = n$. Nevertheless, with N distributed $Poisson(n)$, the probability of the outcome $N = n$ is approximately $P(N = n) \approx 1/\sqrt{2\pi n}$. So the allowance of description of N (not just $N_1, \dots, N_m$ given N) adds $\log 1/P(N = n)$ which is approximately $\frac{1}{2}\log 2\pi n$ bits to the description length beyond the value which would have been ideal $\log 1/Q_a(N_1, \dots, N_m | N = n)$ if $N = n$ were known. This ideal codelength constructed from the tilted maximized Poisson, when conditioning on n, matches the Shtarkov's normalized maximum likelihood based on the multinomial. Thus, $Q_a(\underline{N})$ may also be used in construction of Shtarkov's NML distribution and its conditionals as explained in Section IV-C.

For small alphabet with $m << n$, the minimax regret is about $\frac{1}{2}\log n$ bits per free parameter (a total of $\frac{m-1}{2}\log n +$ constant); and for large alphabet when $m \sim n$ and $n = o(m)$, the minimax regret is about $O(n)$ and $n \log \frac{m}{n}$ respectively [4], [5], [13], [14]. The additional $\frac{1}{2}\log n$ bits is a small price

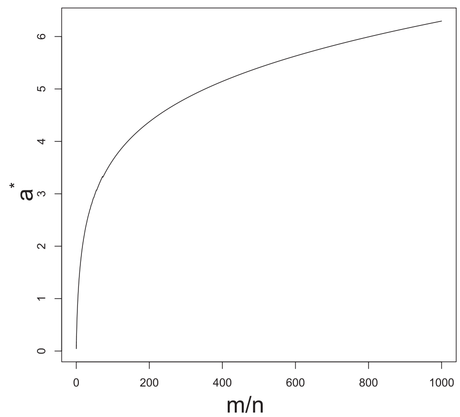

Fig. 2. Relationship between a^* and $\frac{m}{n}$.

to pay for the sake of gaining the coding simplification and additional flexibility.

If it is known that the total count is n, then the regret is a simple function of n and the normalizer C_a. The choice of the tilting parameter a^* given by the moment condition $\mathbf{E}_{Q_a} \sum_{j=1}^{m} N_j = n$ minimizes the regret over all positive a. This arises by differentiation because $\frac{\partial}{\partial a} \log C_a$ is equal to $-n/m \log e$. Moreover, a^* depends only on the ratio between the size of the alphabet and the total count m/n. Figure 2 displays a^* as a function of m/n solved numerically. These values can be stored. Given an alphabet with m symbols and a string generated of length n, one can look at the stored values and find the a^* desired according to the m/n given, and then use the a^* to encode.

If, however, the total count N is not given, then the decoder does not know the a^*. We use a mixture of a to account for the lack of advance knowledge of N, and details are discussed in Section III-D.

When a is small, the tilting of the maximized Poisson likelihood distributions does not have much effect except in the tail of the distribution. Over most of the range of count values k it follows the approximate power-law $1/k^{1/2}$ as we have indicated. Power-laws have been studied for count distributions and are shown to be related to Zipf's law for the sorted counts [15]. Our use of a distribution close to a power-law is not because a power-law is assumed to govern the data, but rather because of its near optimum regret properties within suitable set of counts, demonstrated here for the class of all Poisson count distributions, from which we obtain also its near optimality for the class of all multinomial distributions on counts.

An interesting suggestion from a reader is to simply use a count distribution that is proportional to $1/\sqrt{k}$ on $\{1 \le k \le n\}$, or equivalently proportional to $1/\sqrt{2\pi k}$ on $\{1 \le k \le n\}$, with some provision for the $k = 0$ case. This would be reasonably successful, in a part of the $m = o(n)$ regime, in those cases in which all but $o(\log n)$ of the counts are all large.

However, characteristic of large alphabet source coding is that there can be a large number of small counts. Certainly more than order $\log n$ and even up to order $\min\{m, n\}$. For small counts (e.g. $k = 0, 1, 2$), the $1/\sqrt{2\pi k}$ differs enough from the optimum $k^k e^{-k}/k!$ (which exactly reproduces NML conditional on the sum) that the use of $1/\sqrt{2\pi k}$ would be substantially sub-optimal in regret, while the $k^k e^{-k}/k!$ distribution (with suitable modification) has near optimal regret properties for all large m and exact optimal regret properties conditionally.

Shtarkov studied the universal data compression problem and identified the exact pointwise minimax strategy [4]. He showed the asymptotic minimax lower bound for the regret is $\frac{m-1}{2} \log n + O(1)$, in which the parameter set Θ is the $m-1$ dimensional simplex of all probability vectors on an alphabet of size m. However, this strategy cannot be easily implemented for prediction or compression [4], because of the computational inconvenience of computing the normalizing constant, and because of the difficulty in computing the successive conditionals required for implementation (by arithmetic coding). Let m^* be the number of different symbols that appear in a sequence. Shtarkov [16] also pointed out that when m is large, it is typical that m^* is much less than m, and the regret depends mainly on m^* rather than m. Xie and Barron [5], [17] gave an asymptotic minimax strategy for coding under both the expected and pointwise regret for fixed size alphabet, which is formulated by a modification of the mixture density using Jeffery's prior. The asymptotic value of both the redundancy and the regret are of the form $\frac{m-1}{2} \log n + C_m + o(1)$, where C_m is a constant depending on m. Orlitsky and Santhanam [18] considered the problem in a large alphabet setting. They found the main terms in the minimax regret for $m = o(n)$, $m \sim n$ and $n = o(m)$ cases take the forms $\frac{m-1}{2} \log \frac{n}{m}$, $O(m)$ and $n \log \frac{m}{n}$ respectively. Szpankowski and Weinberger [14] provided more precise asymptotics in these settings. They also calculated the minimax regret of a source model in which some symbol probabilities are fixed. Boucheron, Garivier, and Gassiat [19] focused on countably infinite alphabets with an envelope condition; they used an adapted strategy and gave upper and lower bounds for pointwise minimax regret. Later on Bontemps and Gassiat [20] worked on exponentially decreasing envelope class and provided a minimax strategy and the corresponding regret.

In this paper, we introduce a straightforward and easy to implement data model and associated method for large alphabet coding. The purpose is four-fold: first, by allowing the sample size to be variable, we are considering a larger class of distributions. This is a less restrictive assumption than presuming a particular length. But the method can also be used for fixed sample size coding and prediction. In addition to simple near optimal compression for the class of all strings of a given length, our method also provides natural extension to the conclusion of [19] and [20].

Second, it unveils an information geometry of three key distributions/measures in the problem: the unnormalized maximum Poisson likelihood measure M^m of the counts, the conditional distribution M_{cond} of M^m given the total count equals n, which matches Shtarkov's normalized maximum multinomial likelihood distribution, and a tilted distribution Q_a, with the tilting parameter a chosen to make the expected total count equal to n. This tilted distribution Q_a minimizes the relative entropy from the original measure M^m within the class $\mathcal{C}$ of distributions with the moment condition $E[N] = n$. Hence,

Q_a is the information projection of M^m onto $\mathcal{C}$. Moreover, since M_{cond} is also in $\mathcal{C}$, the Pythagorean-like equality holds [10], [21], as verified also in Appendix C.

$$D(M_{cond}||M^m) = D(M_{cond}||Q_a) + D(Q_a||M^m). \qquad (2)$$

The case of a tilted distribution (the information projection) as an approximating conditional distribution is investigated in [12] and [11]. A difference here is that our unconditional measure M^m is not normalizable.

Thirdly, the strategy designed through an independent Poisson model and tilting is much easier to analyze and compute as compared to the strategies based on multinomials. The convenience is gained through independence. To actually apply this two pass code, one could first describe the independent counts $N_1, \ldots, N_m$, for instance by arithmetic coding using $P_a(N_j)$, and then describe $X_1, \ldots, X_n$ given the counts, by arithmetic coding using the sequence of conditional distributions for X_{i+1} given both $X_1, \ldots, X_i$ and all the counts (which is the sampling without replacement distribution, proportional to the counts of what remains after step i).

Finally, the fourth purpose for our Stirling ratio model is that, as we have said, conditioning on the total count $N = n$ reproduces the Starkov normalized maximum likelihood distribution. Accordingly, as shown in Section IV-C, this method provides a computationally feasible way to exactly compute the Starkov conditionals required for minimax optimal compression.

An alternative to exponential tilting, if the source length n is given, is to use independent count distributions proportional to the Stirling ratio $k^k e^{-k}/k!\ \mathbf{1}_{\{0 \le k \le n\}}$, in which we individually condition on $N_j \le n,\ j = 1, \ldots, m$, with no need for exponential tilting. We do not examine the regret properties of this alternative here. Nevertheless, we note that it retains the independence by conditioning on a square lattice of counts rather than the simplex condition of $N_1 + N_2 + \ldots + N_m = n$, while retaining exact agreement with NML, if one does do that further conditioning on the sum. So the modification of the Stirling ratio can be either by tilting or by this individual bounding of the counts. If the source length is not known to the receiver, the individual count bounding method would require that n be first described or that there be an agreed upon upper bound.

Tilting does not force a bound on the counts to be available and works well for a range of sample sizes. Moreover, there is the allowance of mixing across choices of a as explained in Section III-D.

This paper is organized in the following way. Section II introduces the model. Section III provides results on the regret for coding with our independent counts model. Section IV gives results for exact minimax coding by conditioning on the total count. Section V gives simulated and real data examples. And details of proof are left in the appendix.

II. The Poisson Model

A Poisson model fits well into this problem. We have for each $j = 1, \ldots, m$,

$$N_j \sim Poisson(\lambda_j),$$

independently, and N also has a Poisson distribution

$$N \sim Poisson(\lambda_{sum}),$$

where $\lambda_{sum} = \sum_{j=1}^m \lambda_j$. Write $\underline{\lambda} = (\lambda_1, \ldots, \lambda_m)$, we have

$$P_{\underline{\lambda}}(\underline{X}) = P_{unif}(\underline{X}|\underline{N}) \prod_{j=1}^m P_{\lambda_j}(N_j).$$

We know that the MLE for each λ_j is $\hat{\lambda}_j = N_j$, and the first term is a uniform distribution which does not depend on $\underline{\lambda}$. So

$$P_{\hat{\underline{\lambda}}}(\underline{X}) = P_{unif}(\underline{X}|\underline{N}) \prod_{j=1}^m M(N_j).$$

where $M(k) = k^k e^{-k}/k!$, $k = 1, 2, \ldots$ (as given in the introduction) is the unnormalized maximized likelihood $M(N_j) = \max_{\lambda_j} P_{\lambda_j}(N_j)$.

If we use a distribution $Q(\underline{N})$ to code the counts, then the regret is

$$\log \frac{P_{\hat{\underline{\lambda}}}(\underline{X})}{P(\underline{X}|\underline{N})Q(\underline{N})} = \log \frac{\prod_{j=1}^m M(N_j)}{Q(\underline{N})}.$$

And the redundancy is

$$\mathbf{E}_{P_{\underline{\lambda}}} \log \frac{P(\underline{X}|\underline{\lambda})}{P(\underline{X}|\underline{N})Q(\underline{N})} = \mathbf{E}_{P_{\underline{\lambda}}} \log \frac{P(\underline{N}|\underline{\lambda})}{Q(\underline{N})}.$$

This method can also be applied to fixed total count scenario, which corresponds to the multinomial coding and prediction problem. Suppose $N = n$ is given, the Poisson model, when conditioned on $N = n$, indeed reduces to the i.i.d sampling model

$$P_{\underline{\lambda}}(X_1, \ldots, X_N|N = n) = P_{\underline{\theta}}(X_1, \ldots, X_n).$$

The right hand side is a discrete memoryless source distribution (i.i.d. $P_{\underline{\theta}}$) with probability specified by $P_{\underline{\theta}}(j) = \theta_j$, for $j = 1, \ldots, m$. Note that a sequence $X_1, \ldots, X_N$ with counts $N_1, \ldots, N_m$ of total $N = n$ satisfies

$$\begin{aligned} &P_{\underline{\lambda}}(X_1, \ldots, X_N|N = n) \\ &= \frac{P_{\underline{\lambda}}(X_1, \ldots, X_n)}{P_{\lambda_{sum}}(N = n)} \\ &= \frac{P_{unif}(X_1, \ldots, X_n|N_1, \ldots, N_m) P_{\underline{\lambda}}(N_1, \ldots, N_m)}{P_{\lambda_{sum}}(N = n)}. \end{aligned}$$

The question left is still how to model the counts. The maximized likelihood (the same target as used by Shtarkov) is thus expressible as

$$\begin{aligned} &P_{\hat{\underline{\lambda}}}(X_1, \ldots, X_N|N = n) \\ &= \frac{P_{unif}(X_1, \ldots, X_n|N_1, \ldots, N_m) \prod_{j=1}^m M(N_j)}{P_{\hat{\lambda}_{sum}}(N = n)}. \end{aligned}$$

Now again if we use $Q(N_1, \ldots, N_m)$ to code the counts, then the regret is

$$\begin{aligned}
&\log \frac{P_{\hat{\underline{\lambda}}}(X_1, \ldots, X_N | N = n)}{P_{unif}(X_1, \ldots, X_n | N_1, \ldots, N_m) Q(N_1, \ldots, N_m)} \\
&\quad = \log \frac{\prod_{j=1}^{m} M(N_j)}{P_{\hat{\lambda}_{sum}}(N = n) Q(N_1, \ldots, N_m)} \\
&\quad \approx \frac{1}{2} \log 2\pi n + \log \frac{\prod_{j=1}^{m} M(N_j)}{Q(N_1, \ldots, N_m)} \qquad (3)
\end{aligned}$$

Here $\hat{\lambda}_{sum} = n$, hence the term $\frac{1}{2} \log 2\pi n$ is Stirling's approximation of $\log 1/P_{\hat{\lambda}_{sum}}(N = n)$ with a difference bounded by $\frac{1}{12n} \log e$ by the Robbin's refinement [22] of the Stirling's approximation. The $\frac{1}{2} \log 2\pi n$ arises because here Q includes description of the total N while the more restrictive target regards it as given.

III. Regret Results Coding With Independent Counts

A. Regret

We start by looking at the performance of using independent tilted Stirling ratio distributions as a coding strategy, by examining the regret.

Let S be any set of counts, then the maximized regret of using Q as a coding strategy given a class $\mathcal{P}$ of distributions when the vector of counts is restricted to S is

$$R(Q, \mathcal{P}, S) = \max_{\underline{N} \in S} \log \frac{\max_{P \in \mathcal{P}} P(\underline{N})}{Q(\underline{N})}.$$

Theorem 1: Let P_a be the distribution specified in Equation (1) (Poisson maximized likelihood, tilted and normalized) and N denote the total count. The regret of using a product of tilted distributions $Q_a = \otimes_{j=1}^{m} P_a$ for a given vector of counts $\underline{N} = (N_1, \ldots, N_m)$ is

$$R\left(Q_a, \mathcal{P}_{\Lambda}^{m}, \underline{N}\right) = aN \log e + m \log C_a.$$

Let $S_{m,n}$ be the set of count vectors with total count n be defined as before, then

$$R\left(Q_a, \mathcal{P}_{\Lambda}^{m}, S_{m,n}\right) = an \log e + m \log C_a. \qquad (4)$$

Let a^* be the choice of a satisfying the following moment condition

$$\mathbf{E}_{P_a} \sum_{j=1}^{m} N_j = m\, \mathbf{E}_{P_a} N_1 = n. \qquad (5)$$

Then a^* is the minimizer of the regret in expression (4). Write $R_{m,n} = \min_a R(Q_a, \mathcal{P}_{\Lambda}^{m}, S_{m,n})$.

When $m = o(n)$, the $R_{m,n}$ is near $\frac{m}{2} \log \frac{ne}{m}$ in the following sense.

$$\begin{aligned}
-d_1 \frac{m}{2} \log e \le R_{m,n} &- \frac{m}{2} \log \frac{ne}{m} \\
&\le m \log(1 + \sqrt{\frac{m}{n}}), \qquad (6)
\end{aligned}$$

where $d_1 = O\left((\frac{m}{n})^{1/3}\right)$.

When $n = o(m)$, the $R_{m,n}$ is near $n \log \frac{m}{ne}$ in the following sense.

$$\begin{aligned}
m \log\left(1 + (1 - d_2)\frac{n}{m}\right) \le R_{m,n} &- n \log \frac{m}{ne} \\
&\le m \log\left(1 + \frac{n}{m} + d_3\right) \qquad (7)
\end{aligned}$$

where $d_2 = O(\frac{n}{m})$, and $d_3 = \frac{1}{2\sqrt{\pi}} \frac{n^2 e^2}{m(m-ne)}$.

When $n = bm$, the $R_{m,n} = cm$, where the constant $c = a^* b \log e + \log C_{a^*}$, and a^* is such that $\mathbf{E}_{P_a} N_1 = b$.

Proof: The expression of the regret is from the definition. The fact that a^* is the minimizer can be seen by taking partial derivative with respect to a of expression (4). The upper bounds are derived by applying Lemma 1 in the appendix. Pick $a = m/2n$ and use the first inequality, we get the upper bound for $m = o(n)$ case; pick $a = \ln(m/ne)$ and use the second inequality, we have the upper bound for $n = o(m)$. Here ln is the logarithm base e. The rest of the proof is left in Appendix B. ■

Remark 1: The regret depends only on the number of parameters m, the total counts n and the tilting parameter a. The optimal tilting parameter is given by a simple moment condition in Equation (5).

Remark 2: The regret $R_{m,n}$ is close to the minimax level in all three cases listed in Theorem 1. The main terms in the $m = o(n)$ and $n = o(m)$ cases are the same as the minimax regret given in [14] except the multiplier for $\log(ne/m)$ here is $m/2$ instead of $(m-1)/2$ for the small m scenario. For the $n = bm$ case, the $R_{m,n}$ is close to the minimax regret in [14] numerically.

Remark 3: In fact, the regret provides an upper bound for the redundancy. Recall that

$$\begin{aligned}
\mathbf{E}_{P_{\underline{\lambda}}} \log \frac{P_{\underline{\lambda}}}{Q_a} &\le \mathbf{E}_{P_{\underline{\lambda}}} \max_{\underline{\lambda}} \log \frac{P_{\underline{\lambda}}}{Q_a} \\
&= a\lambda_{sum} \log e + m \log C_a. \qquad (8)
\end{aligned}$$

Theorem 4 in Appendix D gives more detailed expression of the redundancy for using Q_a. While there is a reduction of $(m/2) \log e$ bits as compared to the pointwise case, the error depends on the λ_j's. Nevertheless, expression (8) still provides an uniform upper bound for the redundancy for all possible Poisson means $\underline{\lambda}$ with a given sum.

Corollary 1: Let $\mathcal{P}_{\Theta}^{m}$ be a family of multinomial distributions with total count n. Then the maximized regret $R(Q_a, \mathcal{P}_{\Theta}^{m}, S_{m,n})$ has an upper bound within $\frac{1}{2} \log 2\pi n + \frac{1}{12n} \log e$ above the upper bound in Theorem 1.

Proof: This can be easily seen by Equation (3). ■

B. Subset of Sequences With Partitioned Counts

One advantage of using the tilted Stirling ratio distributions is the flexibility of choosing tilting parameters. As mentioned in the introduction, the ratio m/n uniquely determines the optimal tilting parameter. In fact, different tilting parameters can be used for symbols to adjust for their relative importance in the alphabet. Here we consider a situation in which the empirical distribution has most probability captured by a small portion of the symbols. This happens when the sorted probability list is quite skewed.

The following theorem holds for strings with constraints on the sum of tail counts $\sum_{j>L} N_j = nf$. Small remainder occurs in the following regret bound when $nf/(m-L)$ and $L/(n-nf)$ are both small.

Theorem 2: Let $S_{m,n,f,L}$ be a subset of count vectors with the tail sum controlled by a value $0 \le f \le 1$, that is, $S_{m,n,f,L} = \{\underline{N} = (N_1,\ldots,N_m) : \sum_{j=1}^m N_j = n, \sum_{j>L} N_j = nf\}$. Here L is a number between 0 and m. The regret of using the tilted Stirling ratio distributions for count vectors in $S_{m,n,f,L}$ given each $L \in \{0,\ldots,m\}$ is mainly

$$\frac{L}{2}\log\frac{(n-nf)e}{L} + nf\log\frac{(m-L)}{nfe}. \tag{9}$$

The remainder is bounded below by r_1 and above by r_2, where

$$r_1 = -d_1\frac{L}{2}\log e + (m-L)\log\left(1+(1-d_2)\frac{nf}{m-L}\right),$$

and

$$r_2 = (m-L)\log\left(1+\frac{nf}{m-L}+d_3\right) + L\log\left(1+\sqrt{\frac{L}{n-nf}}\right).$$

Here d_1 is $O\left(\left(\frac{L}{n-nf}\right)^{1/3}\right)$ and d_2 is $O\left(\frac{nf}{m-L}\right)$ and $d_3 = \frac{1}{2\sqrt{\pi}}\frac{(nfe)^2}{(m-L)((m-L)-nfe)}$.

Proof: Consider the product distribution,

$$Q_{a,b}(\underline{N}) = \prod_{j=1}^m P_{a,b}(N_j) = \prod_{j=1}^m \frac{N_j^{N_j}e^{-N_j}}{N_j!}\frac{e^{-aN_j}e^{-bN_j\mathbf{1}_{\{j>L\}}}}{C_{a,b,j}},$$

where $C_{a,b,j} = C_a$ if $j \le L$, and $C_{a,b,j} = C_{a,b}$ is defined as $\sum_{k=0}^{\infty} k^k e^{-(1+a+b)k}/k!$ if $j > L$. It is in fact using an L dimensional product distribution Q_a on the first L symbols, and an $m-L$ dimensional product distribution Q_{a+b} on the rest.

The regret is the same for any $\underline{N} \in S_{m,n,f,L}$ given a and b. That is,

$$\begin{aligned} &R(Q_{a,b}, \mathcal{P}_\Lambda^m, S_{m,n,f,L}) \\ &\quad = na\log e + L\log C_a + nfb\log e + (m-L)\log C_{a,b} \\ &\quad = R(Q_a, \mathcal{P}_\Lambda^L, S_{L,n-nf}) + R(Q_{a+b}, \mathcal{P}_\Lambda^{m-L}, S_{m-L,nf}). \end{aligned}$$

Here $\mathcal{P}_\Lambda^j$ denotes the class of j independent Poisson distributions and $S_{j,k}$ is the set of j independent Poisson counts with sum equal to k. In the above case, $j = L$ or $m-L$, and $k = n-nf$ or nf.

The choice of a, b providing minimization of $R(Q_{a,b}, \mathcal{P}_\Lambda^m, S_{m,n,f,L})$ is given by the following conditions

$$\mathbf{E}_{P_{a,b}}\sum_{j=1}^m N_j = n$$

$$\mathbf{E}_{P_{a,b}}\sum_{j>L} N_j = nf.$$

This result can be derived by applying Inequality (6) and Inequality (7) in Theorem 1 to $R(Q_a, \mathcal{P}_\Lambda^L, S_{L,n-nf})$ and $R(Q_{a+b}, \mathcal{P}_\Lambda^{m-L}, S_{m-L,nf})$ respectively. ■

Remark 4: The problem here is treated as two separate coding tasks, one for a small alphabet with L symbols having a total count $n-nf$, and the other for a large alphabet with $m-L$ symbols with total count nf. The two main terms in expression (9) represent regret from coding the two subsets of symbols, with one set containing L symbols having relatively large counts, and each symbol induces $\frac{1}{2}\log\frac{n(1-f)e}{L}$ bits of regret, and the other containing the rest $m-L$ symbols with small counts and together cost $nf\log\frac{m}{nfe}$ extra bits.

Remark 5: We can add more flexibility to the code by including some extra cost. One is to adapt the choice of L between 0 and m, including $\log(m+1)$ more bits for the description of L. Next one can either work with the counts in the given order, or use an additional $\log\binom{m}{L}$ bits to describe the subset that has the L largest counts. Then one uses $\log 1/Q_{a,b}(\underline{N})$ bits to describe the counts. Rather than fixing f, one can work with the empirical tail fraction $\hat{f}(L)$, where $n\hat{f}(L)$ is the sum of the counts for the remaining $m-L$ symbols. Finally we can adapt the choices of a and b. A suggested method of doing so is described in Section III-D, in which the $Q_{a,b}$ above is replaced by a mixture over a range of choices of a and b.

C. Envelope Class

Besides a subset of strings, we can also consider subclass of distributions. Here we follow the definition of envelope class in [19]. Suppose $\mathcal{P}_{m,f}$ is a class of distributions on $1,\ldots,m$ with the symbol probability bounded above by an envelope function f, i.e.

$$\mathcal{P}_{m,f} = \{P_\theta : \theta_j \le f(j), j = 1,\ldots,m\}.$$

Given the string length n, we know the count of each symbol follows a Poisson distribution with mean $\lambda_j = n\theta_j$, $j = 1,\ldots,m$. This transfers an envelope condition from the multinomial distribution to a Poisson distribution, the mean for which is restricted to the following set

$$\Lambda_{m,f} = \{\underline{\lambda} : \lambda_j \le nf(j), j = 1,\ldots,m\}.$$

Theorem 3: The minimax regret of the Poisson class $\Lambda_{m,f}$ with envelope function f has the following upper bound

$$\begin{aligned} &R(Q_a, \Lambda_{m,f}, \underline{N}) \\ &\quad \le \min_{L\in\{1,\ldots m\}} \frac{L}{2}\log\frac{n(1-\bar{F}(L))}{L} + n\bar{F}(L)\log e + r_3, \end{aligned}$$

where $\bar{F}(L) = \sum_{j>L} f(j)$, and

$$r_3 = \frac{L}{2(1-\bar{F}(L))}\log e + L\log\left(1+\sqrt{\frac{L}{n(1-\bar{F}(L))}}\right).$$

Proof: A tilted distribution with $a = L/2n(1-\bar{F}(L))$ will give the result. Details are left in Appendix E. ■

Remark 6: Here in order for r_3 to be small, the tail sum of the envelope function $\bar{F}(L)$ needs to be small, although the upper bound holds for general envelope function f and L.

This result is of the same order as the upper bound $\inf_{L:L\le n}\left((L-1)/2\log n + n\bar{F}(L)\log e\right) + 2$ given in [19]. The first main term in the bound given in Theorem 3 also matches the minimax regret given in [5] for an alphabet with L symbols and $n(1-\bar{F}(L))$ data points by Stirling's approximation, i.e.,

$$\frac{L-1}{2}\log\frac{n(1-\bar{F}(L))}{2\pi} + \log\frac{\Gamma(1/2)^L}{\Gamma(L/2)}$$

$$\approx \frac{L-1}{2}\log\frac{n(1-\bar{F}(L))e}{L} + \frac{1}{2}\log\frac{e}{2}.$$

The extra $(1/2)\log(n(1-\bar{F}(L))e/L)$ is because the tilted distribution allows m free parameters instead of $m-1$.

Remark 7: The best choice of tilting parameters for envelope class only depends on the envelope function and the number of symbols L constituting the 'frequent' subset. Unlike the subset of strings case discussed before, neither the order of the counts nor which symbols are those with largest counts matters, all we need is an envelope function decaying fast enough when the symbol probabilities are arranged in decreasing order so that L is a small integer and $\bar{F}(L)$ is also not big.

D. Regret With Unknown Total Count

We know that a^* depends on the value of the ratio $\eta = m/n$. However, when the total count is not known, we can use a mixture of tilted distributions $Q(\underline{N})$.

$$Q(\underline{N}) = \int_0^{m/2} Q_a(\underline{N})\frac{1}{m/2}da$$

$$= \int_0^{m/2}\prod_{j=1}^{m}\frac{N_j^{N_j}e^{-N_j}}{N_j!\,C_a}e^{-aN_j}\frac{2}{m}da$$

$$\le M(\underline{N})\frac{2}{m}\int_0^{\infty}e^{-Nh(a)}da$$

where $h(a) = a + \eta\log C_a$, with $\eta = m/N$. Here the upper end of the integrated area is due to Lemma 2. We have $a^* \le m/(2n) \le m/2$.

For any realized non-negative total count $N = k$, the integrand is maximized at a^*_η with $\eta = m/k$, defined as solution to the Equation $\mathbf{E}_{P_a}N_1 = 1/\eta$. And the integral can be approximated by the Laplace method [23],

$$Q(\underline{N}) = \frac{2}{m}\left(\prod_{j=1}^{m}\frac{N_j^{N_j}e^{-N_j}}{N_j!}\right)e^{-kh(a^*_\eta)}\sqrt{\frac{2\pi}{ck}}\,(1+o(1)),$$

where $c = h''(a)|_{a=a^*_\eta}$. Note that the above approximation provides the leading term in an asymptotic expansion of $Q(\underline{N})$. Given η fixed, the leading term approaches the integral as k goes to infinity.

Hence, the regret induced by $Q(\underline{N})$ is

$$\log\frac{M(\underline{N})}{Q(\underline{N})} \approx k(a^*_\eta + \eta\log C_{a^*_\eta}) + \frac{1}{2}\log\frac{ck}{2\pi} + \log\frac{m}{2}.$$

The main part $k(a^*_\eta + \eta\log C_{a^*_\eta})$ is the answer from Theorem 1 if we had known the sample size k in advance. By definition,

$$h''(a) = \eta\frac{\partial^2}{\partial a^2}(\log C_a) = \eta Var_{P_a}(N_1),$$

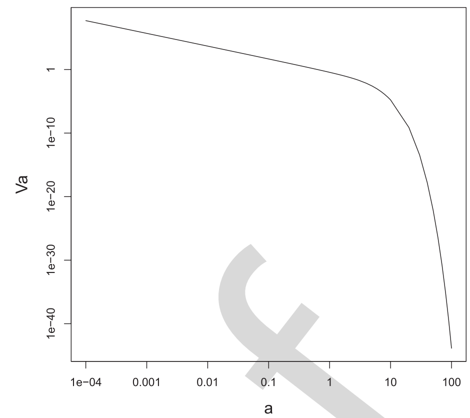

Fig. 3. Relationship between a and V_a.

since $\log C_a$ is the cumulant generating function of the tilted Stirling ratio distribution. We plot $V_a = \frac{\partial^2}{\partial a^2}(\log C_a)$ in Figure 3.

E. Prediction

A sequence of conditional distributions for X_{i+1} given the past observations $X_1, \ldots, X_i$ for $i < n$ provides a sequential prediction with cumulative log loss defined by $\sum_{i<n}\log 1/P(X_{i+1}|X_1,\ldots,X_i)$.

There are two natural ways of providing this sequence of conditionals. One is to get the conditionals from the full joint distribution P_n, which is horizon dependent as mentioned above. It produces cumulative log loss prediction regret precisely the same as the regret of using Q_a for data compression. The other is by using the sequence of distributions $P_{i+1}(X_1,\ldots,X_{i+1}), i < n$, called sequential NML [24]. The sequential prediction distribution $P_{i+1}(X_{i+1} = x|X_1,\ldots,X_i)$ is proportional to $P_{i+1}(X_1,\ldots,X_i,X_i+1 = x)$ and accordingly simplifies to

$$P(X_{i+1} = x|X_1,\ldots,X_i) = \frac{(N_x^i+1)^{N_x^i+1}/N_x^{i\,N_x^i}}{\sum_{\tilde{x}=1}^{m}(N_{\tilde{x}}^i+1)^{N_{\tilde{x}}^i+1}/N_{\tilde{x}}^{i\,N_{\tilde{x}}^i}}.$$

Note that the prediction rule does not involve a. Previous study by Shtarkov [4] shows that it is approximately proportional for large N_x to the $N_x + 1/2$ rule of the Laplace-Jeffreys $Drichlet(1/2,\ldots,1/2)$ update rule (also called the Krichevski-Trofimov rule). Yet it differs importantly from the Laplace-Jeffreys rule for small counts N_x.

However, when using two tilting parameters to adjust for relative importance of symbols within an alphabet, for example, $Q_{a,b}$ in Section III-B, the predictive distribution does depend on b, i.e.,

$$P(X_{i+1} = x|X_1,\ldots,X_i)$$

$$= \frac{e^{-\mathbf{1}_{\{x>L\}}b}(N_x^i+1)^{N_x^i+1}/N_x^{i\,N_x^i}}{\sum_{\tilde{x}=1}^{m}e^{-\mathbf{1}_{\{\tilde{x}>L\}}b}(N_{\tilde{x}}^i+1)^{N_{\tilde{x}}^i+1}/N_{\tilde{x}}^{i\,N_{\tilde{x}}^i}}.$$

Hence, all symbols beyond L are discounted by an extra fact of e^{-b} when predicted by this rule.

IV. Results Coding Conditioned on $N = n$

A. Conditioning on n and Convolutions of P_a

To account for strings of arbitrary length, our coding strategy Q_a assigns a probability distribution to all finite length strings. However, when considering strings of a known length, we are interested to see what the distribution looks like conditioning on a particular number n.

Let $\underline{N}^n$ denote any count vector in $S_{m,n}$, and N_x^n denote the x's component of $\underline{N}^n$, where $x \in \{1, \ldots, m\}$. Also, let M_{mul} be the $multinomial(n, \underline{\theta})$ maximized likelihood. We have

$$Q_a(\underline{N}^n | N = n) = \frac{Q_a(\underline{N}^n)}{Q_a(S_{m,n})} = \frac{M_{mul}(\underline{N}^n)}{M_{mul}(S_{m,n})}. \tag{10}$$

In Equation (10), the factor of difference between the independent coding distribution Q_a $(\underline{N}^n)$ and the Shtarkov NML is the factor $Q_a(S_{m.n})$. This is the probability of the event that the sum $N_1 + N_2 + \ldots + N_m$ equals n, when the individual counts are independent according to the tilted Stirling ratio distribution P_a. As such it is equal to the m-fold convolution of P_a which we also denote by $P_a^m(n)$. This is the distribution on the sample size induced by P_a.

Taking logs, we see that the difference between the unconditional and conditional codelengths is given by $\log(1/P_a^m(n))$. This is the amount by which the unconditional code differs from the Starkov minimax optimal code. One sees in Equation (10) that the relationship with the minimax optimal code holds for all $a \geq 0$. The choice of a^* to minimize the coding regret of $\log 1/Q_a(\underline{N})$ is the same as the choice maximizing $P_a^m(n)$, i.e. minimizing the difference between the unconditional codelength and the Starkov codelength.

Up to a specified n, the convolution $P_a^m(k)$, for $0 \leq k \leq n$, can be evaluated recursively in m, started with $P_a^1(k) = P_a(k)$, and iterating the evaluations

$$P_a^m(k) = \sum_{k'=0}^{k} P_a(k') P_a^{m-1}(k - k') \tag{11}$$

for $k = 0, 1, \ldots, n$. Each such update requires k multiply and adds of stored values for $k = 0, 1, \ldots, n$, which is $n(n+1)/2$ such operations. So a total of $mn(n+1)/2$ operations provide computation of $P_a^m(k)$ for $0 \leq k \leq n$.

In accordance with the relationship between our conditional distribution and Starkov's normalized maximum likelihood, this convolution provides a computationally feasible approach to evaluation of the Starkov normalizing constant $C_{m,n}^*$. Indeed it is seen that for any $a \geq 0$,

$$C_{m,n}^* = P_a^m(n) C_a^m e^{an} \frac{n!}{n^n e^n}.$$

We shall see in Subsection IV-C that evaluations of the convolutions $P_a^{m'}$ for $0 \leq m' \leq m$ also permits evaluations of the conditionals required for implementation of the minimax optimal code.

B. Two Pass Codes

The coding distribution can be implemented by a two pass code. The first pass codes the counts and then the second pass codes the string given the counts. For the coding of the counts an arithmetic code is constructed using either the tilted Stirling ratio distribution (this is the easiest to implement since this distribution makes the counts independent) or we use the distribution conditioned on the counts. Details for computation of the required conditional probabilities are in the next subsection and associated details of arithmetic coding of the counts are in Appendix G.

Then, for the second pass, use an arithmetic code again to code the string given the counts. This distribution of the string given the counts is again to code the string given the counts. The distribution of the string given the counts is uniform for all strings with the given counts. To implement arithmetic coding, one uses the conditional probability for x less than or equal to the observed X_{i+1} given its past and the counts, i.e.

$$P\,(X_{i+1} < x_{i+1} | X_1, \ldots, X_i, (N_1, \ldots, N_m)),$$

and

$$P\,(X_1, \ldots, X_i, X_{i+1} | (N_1, \ldots, N_m)),$$

for each $i = 0, \ldots, n-1$ with $n = \sum_{j=1}^m N_j$.

Indeed for $i = 1$, the $P(X_1 = x_1 | (N_1, \ldots, N_m)) = N_{x_1}/n$, and generally let $N_{j,i}$ be the count of the number of occurrence of j in $X_1, \ldots, X_i$, then the remaining counts are $N_{j,i}^{rem} = N_j - N_{j,i}$, and $P(X_{i+1} = x | X_1, \ldots, X_i, (N_1, \ldots, N_m)) = N_{j,i}^{rem}/(n - i)$. This is the consequence of the distribution of $X_1, \ldots, X_n$ given $N_1, \ldots, N_m$ being uniform on the set of strings with these counts. (It is in accordance with the theory of sampling without replacement that arises with this conditioning.)

These two pass codes make possible computationally feasible coding of exact or approximate minimax optimal codes. The simpler approximate minimax coding has desirable regret properties in the regime of $m \sim n$ and $n = o(m)$ as well as $m = o(n)$. Alternatively, the one pass Krichevsky–Trofimov [8] sequential coding rule, which is the Laplace posterior update rule with respect to the $Dirichlet(1/2, \ldots, 1/2)$ prior, can also be used for $m = o(n)$. What we propose here is a simple scheme that achieves nearly minimal regret in all situations. And its implementation is simple due to the independence of the coding distribution of the counts. Computation complexity for the codes is $O(m \log n + n \log mn)$ as explained in Appendix G. Conditioning to provide the exact minimax strategy adds an additional $(m + n) \log mn$ bits to compute the conditionals, and an additional complexity of order mn^2 to compute the convolutions of P_a. (The latter can be precomputed once off-line and stored so as to not increase the time complexity in repeated coding thereafter.) We explain more about the conditional distributions required to implement the exact minimax strategy here below in Subsection IV-C.

C. Computing Shtarkov's Distribution Using Q_a Conditionals

Exact minimax compression is regarded as challenging because of the potential difficulty with the Shtarkov joint distribution in computing either the conditional distribution of X_i given $X_1, \ldots, X_{i-1}$ for observations $i \leq n$ or the conditional distribution of the counts N_j given $N_1, \ldots, N_{j-1}$

for symbol indices $j \leq m$. Here we show how to overcome this difficulty working with the counts.

We have seen that, when n is given, the Shtarkov joint distribution $Q_{nml}(N_1, \ldots, N_m)$ of the counts is the same as the Q_a joint distribution of $N_1, \ldots, N_m$, conditioned on $N_1 + \ldots + N_m = n$. Consequently, it holds for every value of $a \geq 0$ that

$$Q_{nml}(N_j = n_j | N_1 = n_1, \ldots, N_{j-1} = n_{j-1})$$
$$= Q_a(N_j = n_j | N_1 = n_1, \ldots, N_{j-1} = n_{j-1}, \sum_{i=1}^{m} N_i = n)$$

for each $j = 1, 2, \ldots, m$. By the rules of probability this is the ratio

$$\frac{Q_a\left(N_1 = n_1, \ldots, N_j = n_j, \sum_{i=j+1}^{m} N_i = n - \sum_{i=1}^{j} n_i\right)}{Q_a\left(N_1 = n_1, \ldots, N_{j-1} = n_{j-1}, \sum_{i=j}^{m} N_i = n - \sum_{i=1}^{j-1} n_i\right)}$$

Next use that Q_a makes the N_j independent with distribution P_a and that the sums $N_{j+1} + \ldots + N_m$ have distribution P_a^{m-j} obtained by the $m - j$ fold convolution of P_a. Canceling common factors the above ratio is simply

$$\frac{P_a(n_j) P_a^{m-j}(n - (n_1 + \ldots + n_j))}{P_a^{m-j+1}(n - (n_1 + \ldots + n_{j-1}))}. \tag{12}$$

Thus computation of the Shtarkov conditionals reduces to this ratio involving the $P_a^{m'}$ for $1 \leq m' \leq m$, precomputed by convolution. Note that the dependence on n_j is only in the numerator and that the denominator is simply the sum of the numerator for n_j in the range between 0 and $n - (n_1 + \ldots + n_{j-1})$, in accordance with the rules of convolution. This identity for the Shtarkov conditionals is valid for any $a \geq 0$. Note that when $a = 0$, the numerator and denominator are not probability distributions since C_a equals infinity, but the C_a terms cancel out through conditioning and the equality still holds.

For numerical stability (to avoid ratios of very small numbers) it is advantageous to choose $a = a^*$ for which the denominator is large. This a^* may be evaluated at m/n. The choice maximizing the denominator at step j is a^* evaluated at $(m - j + 1)/(n - (n_1 + \ldots + n_{j-1}))$.

We note here that when conditioning on the count sum n, the results are unchanged if the tilted Stirling ratio distribution is restricted to the set $\{0 \leq k \leq n\}$. This is because in the convolution calculation of $P_a^m(k)$ in Equation (11), the index k is only needed for $0 \leq k \leq n$. Truncating the distribution at n would change the normalizer, though, as we have said, the normalizer cancels out in the conditional distribution.

V. Application

A. Simulation

Theorem 2 indicates we could optimize L to save coding cost when the ordered counts are skewed. We look at the performance of the tilted Stirling ratio distribution for algebraically decreasing counts with simulated data. The alphabet is partitioned into two subsets – the frequent symbols and the infrequent ones. The tilting parameter is chosen approximately according to the ratio of the number of symbols in a subset and their total count. The regret of assigning different number of symbols as ‘frequent’ (L) is shown in Fig. 4. We can see that more skewness pushes the optimizing L smaller.

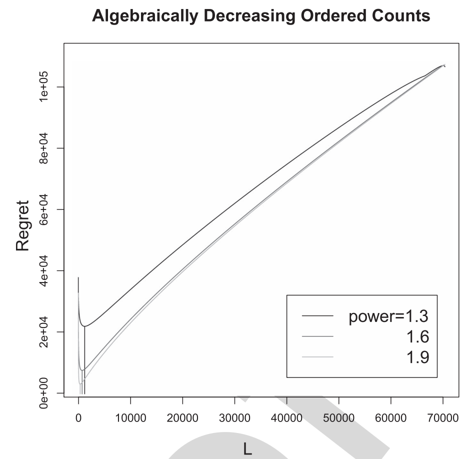

Fig. 4. Regret of using tilted Stirling ratio distribution for algebraically decreasing counts.

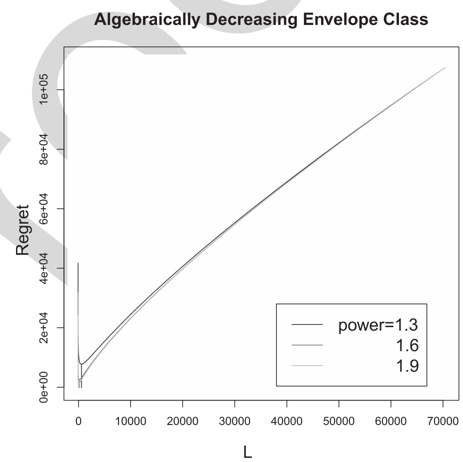

Fig. 5. Regret of using tilted Stirling ratio distribution for an algebraically decreasing envelope class.

Figure 5 shows the upper bound of the minimax regret in Theorem 3 for an algebraically decreasing envelope class.

B. Real Data

We also provide an example of using the tilted Stirling ratio distribution to code Chinese literature. The target book is an ancient collection of poems named 诗经, translated as the Classic of Poetry. It is the existing earliest collection of Chinese poetry and dates from the 10th to 7th centuries BC [25]. The book is downloaded freely from http://wenku.baidu.com/. Since many ancient words are rarely used today, the encoding is done in GB18030 [26], the largest Chinese coded character set. It contains 70244 characters, among which 2889 appear in the book with a total character count 39161. There are 792 characters appear once and 479 appear twice. The smallest regret happens at $L = 2889$ which is the total number of characters appear.

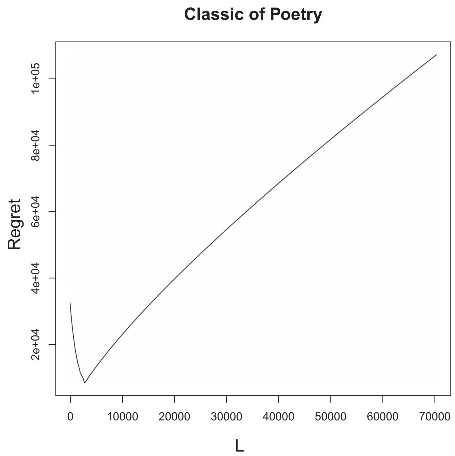

Fig. 6. Regret of $Q_{a,b}$ for L from 1 to m.

VI. Discussion

We have introduced the use of independent tilted maximized Poisson likelihood distributions (also here called tilted Stirling ratio distributions) Q_a for coding the counts of sequences of independently distributed random variables. The performance of the coding distribution is close to the minimax level. Actually, the difference between the regret and the minimax level is the probability assigned to the set with the observed total count by the tilted distribution with the optimal tilting parameter, i.e.

$$\begin{aligned} R(M_{cond}, \mathcal{P}_\Lambda^m, S_{m,n}) &= R(Q_{a^*}, \mathcal{P}_\Lambda^m, S_{m,n}) \\ &\quad + \log Q_{a^*}(S_{m,n}). \end{aligned}$$

The optimal tilting parameter a^* minimizes the difference among all possible a. Since M_{cond} reproduces the Shtarkov's NML distribution for the multinomial family of distributions on counts, it is the exact pointwise minimax strategy. As shown in this paper, our findings about the regret produced by the distribution Q_a, taken together with earlier work [4], [5], [14], [18], show that the difference is no larger than about $\log n$ in small alphabet case, and about $\frac{1}{2}\log n$ for moderate or large alphabets. The probability $Q_a(S_{m,n})$ is the probability distribution for the total count N evaluated at $N = n$ as induced by our distribution Q_a. Further analysis could be done to characterize this distribution of the total count more precisely.

Appendix A

Fact 1: For any $a > 0$,

$$\frac{1}{\sqrt{2\pi}}\int_0^1 t^{-\frac{1}{2}} e^{-at} dt < \sqrt{\frac{2}{\pi}}.$$

Proof:

$$\begin{aligned} \frac{1}{\sqrt{2\pi}}\int_0^1 t^{-\frac{1}{2}} e^{-at} dt &\overset{u=at}{=} \frac{1}{\sqrt{2\pi}}\int_0^a \left(\frac{u}{a}\right)^{-\frac{1}{2}} e^{-u}\frac{1}{a} du \\ &= \frac{1}{\sqrt{2\pi a}}\int_0^a u^{-\frac{1}{2}} e^{-u} du \end{aligned}$$

The integrand is smaller than $u^{-\frac{1}{2}}$ on $[0, a]$, so the integral is upper bounded by

$$\frac{1}{\sqrt{2\pi a}}\int_0^a u^{-\frac{1}{2}} du = \sqrt{\frac{2}{\pi}}.$$

■

Fact 2: For any $a > 0$,

$$\sum_{k=1}^\infty \frac{k^{-\frac{1}{2}}}{\sqrt{2\pi}\, e^{r_k}} e^{-ak} \ge \frac{1}{\sqrt{2\pi}}\int_1^\infty t^{-\frac{1}{2}} e^{-at} dt$$

when $\frac{1}{12k+1} \le r_k \le \frac{1}{12k}$.

Proof: It suffice to show

$$\sum_{k=1}^\infty \frac{k^{-\frac{1}{2}}}{e^{\frac{1}{12k}}} e^{-ak} \ge \int_1^\infty t^{-\frac{1}{2}} e^{-at} dt \tag{13}$$

Note that $f(t) = t^{-\frac{1}{2}} e^{-at}$ is convex in t, so we have $\int_k^{k+1} f(t)dt$ upper bounded by $(f(k) + f(k+1))/2$. Then we only need to show the latter is upper bounded by $f(k)e^{-1/12k}$. This can be done by proving the following inequality.

$$\left(1 + \left(\frac{k}{k+1}\right)^{\frac{1}{2}} e^{-a}\right) e^{\frac{1}{12k}} \le 2$$

for each $k \ge 1$ and $a > 0$. Check that the left hand side is increasing in k, its value goes up to $1 + e^{-a}$ which is not larger than the right hand side for every $a \ge 0$. Therefore, Inequality (13) follows. ■

Lemma 1 (Bounds for C_a): For any $a > 0$, the following bounds hold for C_a

$$\max\left(1, 1 - \sqrt{\frac{2}{\pi}} + \frac{1}{\sqrt{2a}}\right) < C_a < 1 + \frac{1}{\sqrt{2a}}, \tag{14}$$

and

$$1 + e^{-(a+1)} < C_a < 1 + e^{-(a+1)} + \frac{1}{2\sqrt{\pi}}\frac{e^{-2a}}{1 - e^{-a}}. \tag{15}$$

Proof: The argument to prove the upper bounds is analogous to Fact 2. Indeed,

$$C_a = \sum_{k=0}^\infty \frac{k^k e^{-k}}{k!} e^{-ak} \overset{(a)}{=} 1 + \sum_{k=1}^\infty \frac{k^{-\frac{1}{2}}}{\sqrt{2\pi}\, e^{r_k}} e^{-ak} \tag{16}$$

Here (a) is by Robbins' refinement of Stirling's approximation where $\frac{1}{12k+1} < r_k < \frac{1}{12k}$.

The sum can be bounded by a gamma integral, so

$$\begin{aligned} C_a &\le 1 + \frac{1}{\sqrt{2\pi}}\int_0^\infty t^{-\frac{1}{2}} e^{-at} dt \\ &= 1 + \frac{1}{\sqrt{2\pi}}\frac{\Gamma(\frac{1}{2})}{a^{\frac{1}{2}}} \\ &= 1 + \frac{1}{\sqrt{2a}}. \end{aligned}$$

Also, following expression (16), C_a has the following lower bound.

$$
\begin{aligned}
C_a &= 1+\sum_{k=1}^{\infty}\frac{k^{-\frac{1}{2}}}{\sqrt{2\pi}\,e^{r_k}}e^{-ak}\\
&\overset{(b)}{\geq} 1-\sqrt{\frac{2}{\pi}}+\sqrt{\frac{2}{\pi}}+\frac{1}{\sqrt{2\pi}}\int_1^{\infty}t^{-\frac{1}{2}}e^{-at}dt\\
&\overset{(c)}{>} 1-\sqrt{\frac{2}{\pi}}+\frac{1}{\sqrt{2\pi}}\int_0^{1}t^{-\frac{1}{2}}e^{-at}dt\\
&\quad+\frac{1}{\sqrt{2\pi}}\int_1^{\infty}t^{-\frac{1}{2}}e^{-at}dt\\
&= 1-\sqrt{\frac{2}{\pi}}+\frac{1}{\sqrt{2\pi}}\int_0^{\infty}t^{-\frac{1}{2}}e^{-at}dt\\
&= 1-\sqrt{\frac{2}{\pi}}+\frac{1}{\sqrt{2a}}.
\end{aligned}
$$

Here again $\frac{1}{12k+1} < r_k < \frac{1}{12k}$, and Inequality (b) is due to Fact 2 and Inequality (c) is by Fact 1.

Note that Inequality (14) is good for small a. For a moderately large a ($a > 0.2$), the following upper bound is better.

$$
\begin{aligned}
C_a &\leq 1+e^{-(a+1)}+\sum_{k=2}^{\infty}\frac{1}{\sqrt{2\pi k}}e^{-ka}\\
&< 1+e^{-(a+1)}+\frac{1}{2\sqrt{\pi}}\frac{e^{-2a}}{1-e^{-a}}.
\end{aligned}
$$

■

Lemma 2: For any $a > 0$,

$$
e^{-(a+1)} \leq \mathbf{E}_{P_a}N_1 \leq \frac{1}{2a}.
$$

Proof: Let $k^* = \text{argmin}_{k\in\mathbf{N}_+}\left|k-\frac{1}{2a}\right|$. We prove the upper bound by consider a within two different intervals. First, if $a \leq e(\sqrt{\pi}-\sqrt{2})^2$, we know

$$
\begin{aligned}
&\sum_{k=1}^{\infty}\frac{k^{k+1}e^{-k}}{k!}e^{-ak}\\
&= \sum_{k=1}^{k^*-1}\frac{k^{k+1}e^{-k}}{k!}e^{-ak}+\sum_{k=k^*+1}^{\infty}\frac{k^{k+1}e^{-k}}{k!}e^{-ak}\\
&\quad+\frac{k^{*k^*+1}e^{-k^*}}{k^*!}e^{-ak^*}\\
&\overset{(a)}{\leq} \sum_{k=1}^{k^*-1}\frac{k^{1/2}e^{-ak}}{\sqrt{2\pi}}+\sum_{k=k^*+1}^{\infty}\frac{k^{1/2}e^{-ak}}{\sqrt{2\pi}}\\
&\quad+\frac{k^{*1/2}e^{-ak^*}}{\sqrt{2\pi}}
\end{aligned}
\tag{17}
$$

where (a) is an upper bound by Stirling's approximation.

Both sums in the last expression can be upper bounded by a gamma integral, and $k^{*1/2}e^{-ak^*}$ is no larger than the maximum of the unnormalized $Gamma(3/2, 1/a)$ density, which is achieved at $1/(2a)$. Hence, we have the following upper bound for expression (17).

$$
\begin{aligned}
&\int_0^{k^*}\frac{t^{1/2}e^{-at}}{\sqrt{2\pi}}dt+\int_{k^*}^{\infty}\frac{t^{1/2}e^{-at}}{\sqrt{2\pi}}dt+\frac{(1/2a)^{1/2}e^{-1/2}}{\sqrt{2\pi}}\\
&\quad= \frac{\Gamma(3/2)}{a^{3/2}\sqrt{2\pi}}+\frac{(1/2a)^{1/2}}{\sqrt{2\pi\,e}}\\
&\quad= \frac{1}{(2a)^{3/2}}+\frac{1}{\sqrt{2\pi\,e}}\frac{1}{(2a)^{1/2}}
\end{aligned}
$$

Using this upper bound for C_a, we could prove an upper bound for the expected value.

$$
\begin{aligned}
\mathbf{E}_{P_a}N_1 &= \sum_{k=1}^{\infty}\frac{k^{k+1}e^{-k}}{k!\,C_a}e^{-ak}\\
&\overset{(b)}{\leq} \frac{\frac{1}{(2a)^{3/2}}+\frac{1}{\sqrt{2\pi e}}\frac{1}{(2a)^{1/2}}}{\frac{1}{(2a)^{1/2}}+1-\sqrt{\frac{2}{\pi}}}\\
&= \frac{1}{2a}\underbrace{\left(\frac{\frac{1}{(2a)^{1/2}}+\frac{1}{\sqrt{2\pi e}}(2a)^{1/2}}{\frac{1}{(2a)^{1/2}}+1-\sqrt{\frac{2}{\pi}}}\right)}_{(A)}
\end{aligned}
$$

The lower bound for the denominator in (b) is attributed to Lemma 1. A little algebra can show that term (A) is not larger than 1 when a is restricted to $(0, e(\sqrt{\pi}-\sqrt{2})^2]$.

If $a > e(\sqrt{\pi}-\sqrt{2})^2$, we have $\arg\max_{k\geq 1} k^{1/2}e^{-ak} = 1$. Using Stirling's approximation and split the sum into $k = 1$ and $k > 1$, we have

$$
\begin{aligned}
&\sum_{k=1}^{\infty}\frac{k^{k+1}e^{-k}}{k!}e^{-ak}\\
&\quad\leq \frac{e^{-a}}{\sqrt{2\pi}}+\sum_{k=2}^{\infty}\frac{k^{1/2}e^{-ak}}{\sqrt{2\pi}}\\
&\quad\overset{(c)}{\leq} \frac{1}{\sqrt{2\pi}}\left(\frac{1}{2}e^{-a}+\int_0^{\infty}t^{1/2}e^{-at}dt\right)\\
&\quad= \frac{1}{\sqrt{2\pi}}\left(\frac{1}{2}e^{-a}+\frac{\Gamma(3/2)}{a^{3/2}}\right)\\
&\quad= \frac{1}{2\sqrt{2\pi}}e^{-a}+\frac{1}{(2a)^{3/2}}
\end{aligned}
$$

where (c) is because the sum $\sum_{k=2}^{\infty}k^{1/2}e^{-ak}$ is bounded above by the integral $\int_1^{\infty}t^{1/2}e^{-at}dt$, and the difference between $\int_0^1 t^{1/2}e^{-at}dt$ and e^{-a} (value of $k^{1/2}e^{-ak}$ at $k = 1$) is less than $\frac{1}{2}e^{-a}$ due to the concavity of $t^{1/2}e^{-at}$ to the left of $1/2a$.

By this upper bound for the numerator and Lemma 1 again,

$$
\begin{aligned}
\mathbf{E}_{P_a}N_1 &\leq \frac{\frac{1}{(2a)^{3/2}}+\frac{1}{2\sqrt{2\pi}}e^{-a}}{\frac{1}{(2a)^{1/2}}+1-\sqrt{\frac{2}{\pi}}}\\
&= \frac{1}{2a}\underbrace{\left(\frac{\frac{1}{(2a)^{1/2}}+\frac{1}{\sqrt{2\pi}}ae^{-a}}{\frac{1}{(2a)^{1/2}}+1-\sqrt{\frac{2}{\pi}}}\right)}_{(B)}.
\end{aligned}
$$

Term (B) is not larger than 1 because $\frac{1}{\sqrt{2\pi}}ae^{-a} \leq 1-\sqrt{\frac{2}{\pi}}$ for all a.

For the lower bound,

$$\begin{aligned}
\mathbf{E}_{P_a} N_1 &= \sum_{k=1}^{\infty} \frac{k^{k+1} e^{-k}}{k!\, C_a} e^{-ak} \\
&= \frac{e^{-(a+1)} \left(\sum_{k=1}^{\infty} \frac{k^k e^{-(k-1)}}{(k-1)!} e^{-a(k-1)} \right)}{C_a} \\
&\overset{l=k-1}{=} \frac{e^{-(a+1)} \left(\sum_{l=0}^{\infty} \frac{(l+1)^{l+1} e^{-l}}{l!} e^{-al} \right)}{C_a} \\
&= e^{-(a+1)} \underbrace{\left(\frac{\sum_{l=0}^{\infty} \frac{(l+1)^{l+1} e^{-l}}{l!} e^{-al}}{\sum_{k=0}^{\infty} \frac{k^k e^{-k}}{k!} e^{-ak}} \right)}_{(C)} \\
&\overset{(d)}{\geq} e^{-(a+1)} \qquad (18)
\end{aligned}$$

Here Inequality (d) is because term (C) is above 1. Hence, the upper bound is deduced. ■

Appendix B
Proof of Theorem 1

Proof: It remains to show the two lower bounds in expression (6) and (7). In both cases we need a lower bound for $na^* \log e + m \log C_{a^*}$, and we do it by lower bounding a^* and C_{a^*}, respectively. Let $\tilde{a} = \frac{m}{2n}$.

- Bounds for a^*

We know a^* is the solution for the following equation.

$$\mathbf{E}_{P_{a^*}} N_1 = \frac{n}{m}$$

By Lemma 2, we have

$$\frac{1}{2a^*} \geq \frac{n}{m}$$

That gives

$$a^* \leq \frac{m}{2n} = \tilde{a} \qquad (19)$$

Since C_a is decreasing in a, we have

$$C_{a^*} \geq C_{\tilde{a}} > \frac{1}{\sqrt{2\tilde{a}}} = \sqrt{\frac{n}{m}}.$$

For any $j \in \{1, \ldots, m\}$, and $a > 0$, we have

$$\begin{aligned}
\mathbf{E}_{P_a} N_1 &= \sum_{k=1}^{\infty} \frac{k^{k+1} e^{-k}}{k!\, C_a} e^{-ak} \\
&\overset{(a)}{\geq} \frac{\sum_{k=1}^{\infty} \frac{k^{k+1} e^{-k}}{k!} e^{-ak}}{1 + \frac{1}{\sqrt{2a}}} \\
&\overset{(b)}{=} \frac{\sum_{k=1}^{\infty} \frac{k^{\frac{1}{2}}}{\sqrt{2\pi} e^{r_k}} e^{-ak}}{1 + \frac{1}{\sqrt{2a}}} \qquad (20)
\end{aligned}$$

Here (a) is attributed to Inequality (14), step (b) is by Stirling's approximation, and $\frac{1}{12k+1} < r_k < \frac{1}{12k}$. Pick $k_1 = a^{-1/3}$, then the numerator of expression (20) can be lower bounded by

$$\begin{aligned}
&\sum_{k=\lfloor k_1 \rfloor}^{\infty} \frac{k^{1/2}}{\sqrt{2\pi} e^{r_k}} e^{-ak} \\
&\geq \sum_{k=\lfloor k_1 \rfloor}^{\infty} \frac{k^{1/2}}{\sqrt{2\pi} e^{\frac{1}{12\lfloor k_1 \rfloor}}} e^{-ak} \\
&\geq \frac{1}{\sqrt{2\pi} e^{\frac{1}{12(k_1-1)}}} \int_{\lfloor k_1 \rfloor}^{\infty} t^{1/2} e^{-at} dt
\end{aligned}$$

Taking the integral from 0 to ∞ and subtracting the part from 0 to k_1 yields the lower bound

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi} e^{\frac{1}{12(k_1-1)}}} \left(\frac{\Gamma(3/2)}{a^{3/2}} - \int_0^{k_1} t^{1/2} e^{-at} dt \right) \\
&\geq \frac{1}{\sqrt{2\pi} e^{\frac{1}{12(k_1-1)}}} \left(\frac{\Gamma(3/2)}{a^{3/2}} - \int_0^{k_1} t^{1/2} dt \right) \\
&= \frac{1}{\sqrt{2\pi} e^{\frac{1}{12(k_1-1)}}} \left(\frac{\Gamma(3/2)}{a^{3/2}} - \frac{2}{3a^{1/2}} \right).
\end{aligned}$$

Write $r_a = \frac{1}{12(k_1-1)} = \frac{a^{1/3}}{12(1-a^{1/3})}$. By the above calculation, we have a lower bound for the expectation under the tilting distribution. For a^*,

$$\frac{\frac{1}{\sqrt{2\pi} e^{r_{a^*}}} \left(\frac{\Gamma(3/2)}{a^{*3/2}} - \frac{2}{3a^{*1/2}} \right)}{1 + \frac{1}{\sqrt{2a^*}}} \leq \mathbf{E}_{a^*} N_1 = \frac{n}{m}.$$

Arranging the terms, we have

$$\begin{aligned}
\frac{1}{2a^*} &\leq \frac{n}{m} \left(1 + \sqrt{2a^*}\right) e^{r_{a^*}} + \frac{2}{3\sqrt{\pi}} \\
&\overset{(c)}{\leq} \frac{n}{m} \left(1 + \sqrt{2\tilde{a}}\right) e^{r_{\tilde{a}}} + \frac{2}{3\sqrt{\pi}}
\end{aligned}$$

Here (c) is because $a^* \leq \tilde{a}$ by Inequality (19). So,

$$a^* \geq \frac{\tilde{a}}{\left(1 + \sqrt{2\tilde{a}}\right) e^{r_{\tilde{a}}} + \frac{4}{3\sqrt{\pi}} \tilde{a}}$$

By Taylor expansion, this is no smaller than

$$\begin{aligned}
&\frac{\tilde{a}}{\left(1 + \sqrt{2\tilde{a}}\right)\left(1 + r_{\tilde{a}} + O(r_{\tilde{a}}^2)\right) + \frac{4}{3\sqrt{\pi}} \tilde{a}} \\
&= \tilde{a} \left(1 - \frac{r_{\tilde{a}} + \sqrt{2\tilde{a}} + \sqrt{2\tilde{a}} r_{\tilde{a}} + \frac{4}{3\sqrt{\pi}} \tilde{a} + O(r_{\tilde{a}}^2)}{\left(1 + \sqrt{2\tilde{a}}\right)\left(1 + r_{\tilde{a}} + O(r_{\tilde{a}}^2)\right) + \frac{4}{3\sqrt{\pi}} \tilde{a}} \right) \\
&\geq \tilde{a} \left(1 - r_{\tilde{a}} - \sqrt{2\tilde{a}} - \sqrt{2\tilde{a}} r_{\tilde{a}} - \frac{4}{3\sqrt{\pi}} \tilde{a} - O(r_{\tilde{a}}^2) \right)
\end{aligned}$$

When $m = o(n)$, $r_{\tilde{a}}$ is the leading term, so

$$a^* \geq \tilde{a}\left(1 - O\left(r_{\tilde{a}}\right)\right) = \frac{m}{2n} \left(1 - O\left(\left(\frac{m}{n} \right)^{\frac{1}{3}} \right) \right)$$

As a result,

$$na^* \log e \geq \left(1 - O\left(\left(\frac{m}{n} \right)^{\frac{1}{3}} \right) \right) \frac{m}{2} \log e$$

Hence we get Inequality (6).

The above lower bound works when a^* is small (i.e., when m is small compared to n), yet when it is large, the following bound is better. Let $a_0 = \ln \frac{m}{ne}$.

From Lemma 2,

$$e^{-(a^*+1)} \leq \frac{n}{m}.$$

Then

$$e^{a^*} \geq \frac{m}{ne} = e^{a_0}$$
$$a^* \geq a_0 \tag{21}$$

Thus,

$$na^* \log e \geq na_0 \log e = n \log \frac{m}{ne}$$

- Bounds for C_{a^*}

Now we want to lower bound C_{a^*}. Recall Inequality (18), let term (C) be defined as

$$s_a = \frac{\sum_{l=0}^{\infty} (l+1)^{l+1} e^{-l} e^{-al}/l!}{\sum_{k=0}^{\infty} k^k e^{-k} e^{-ak}/k!}.$$

We have

$$s_{a^*} e^{-(a^*+1)} = \mathbf{E}_{P_{a^*}} N_j = \frac{n}{m} = e^{-(a_0+1)}.$$

It gives

$$e^{-(a^*+1)} = \frac{e^{-(a_0+1)}}{s_{a^*}}.$$

By definition,

$$C_{a^*} \geq 1 + e^{-(a^*+1)} = 1 + \frac{e^{-(a_0+1)}}{s_{a^*}}. \tag{22}$$

By Stirling's approxmation, the numerator of s_a is bounded above.

$$\begin{aligned}
&\sum_{l=0}^{\infty} \frac{(l+1)^{l+1} e^{-l} e^{-al}}{l!} \\
&\leq 1 + \frac{1}{\sqrt{2\pi}} \sum_{l=1}^{\infty} (1+\frac{1}{l})^l \frac{l+1}{\sqrt{l}} e^{-al} \\
&\overset{(d)}{\leq} 1 + \frac{e}{\sqrt{2\pi}} \sum_{l=1}^{\infty} \frac{l+1}{\sqrt{l}} e^{-al} \\
&\leq 1 + \frac{e}{\sqrt{2\pi}} \left(\sum_{l=1}^{\infty} l e^{-al} + \sum_{l=1}^{\infty} e^{-al} \right)
\end{aligned} \tag{23}$$

where (d) is because $(1+\frac{1}{l})^l$ is bounded above by e for each $l > 0$. We know $\sum_{l=1}^{\infty} l e^{-al}(1-e^{-a})$ is equal to the expectation of a geometric random variable with success probability $1-e^{-a}$, which equals to $1/(1-e^{-a}) - 1$. And $\sum_{l=1}^{\infty} e^{-al}(1-e^{-a}) = e^{-a}$. Hence, Equation (23) has the following upper bound

$$1 + \frac{e}{\sqrt{2\pi}} \frac{e^{-a}(2-e^{-a})}{(1-e^{-a})^2}.$$

Using the above inequality and $C_{a^*} \geq 1 + e^{-(a^*+1)}$, we have

$$\begin{aligned}
\frac{1}{s_{a^*}} &\geq \frac{1+e^{-(a^*+1)}}{1+\frac{e}{\sqrt{2\pi}}\frac{e^{-a^*}(2-e^{-a^*})}{(1-e^{-a^*})^2}} \\
&= 1 - \frac{\frac{e}{\sqrt{2\pi}}\frac{e^{-a^*}(2-e^{-a^*})}{(1-e^{-a^*})^2} - e^{-(a^*+1)}}{1+\frac{e}{\sqrt{2\pi}}\frac{e^{-a^*}(2-e^{-a^*})}{(1-e^{-a^*})^2}} \\
&= 1 - \frac{\frac{e^2}{\sqrt{2\pi}}\frac{2-e^{-a^*}}{(1-e^{-a^*})^2} - 1}{1+\frac{e}{\sqrt{2\pi}}\frac{e^{-a^*}(2-e^{-a^*})}{(1-e^{-a^*})^2}} e^{-(a^*+1)}
\end{aligned}$$

Multiply $(1-e^{-a^*})^2$ on both the numerator and denominator of the second term, we have the above expression equal to

$$\begin{aligned}
&1 - \frac{\frac{2e^2}{\sqrt{2\pi}} - 1 - (\frac{e^2}{\sqrt{2\pi}} - 2)e^{-a^*} - e^{-2a^*}}{(1-e^{-a^*})^2 + \frac{e}{\sqrt{2\pi}} e^{-a^*}(2-e^{-a^*})} e^{-(a^*+1)} \\
&= 1 - \frac{\frac{2e^2}{\sqrt{2\pi}} - 1 - (\frac{e^2}{\sqrt{2\pi}} - 2)e^{-a^*} - e^{-2a^*}}{\frac{e}{\sqrt{2\pi}} + (1-\frac{e}{\sqrt{2\pi}})(1-e^{-a^*})^2} e^{-(a^*+1)}.
\end{aligned}$$

The denominator of the second term is lower bounded by 1 since $0 < e^{-a^*} < 1$. Therefore,

$$\begin{aligned}
&\frac{1}{s_{a^*}} \\
&\geq 1 - \left(\frac{2e^2}{\sqrt{2\pi}} - 1 - (\frac{e^2}{\sqrt{2\pi}} - 2)e^{-a^*} - e^{-2a^*} \right) e^{-(a^*+1)} \\
&\geq 1 - \left(\frac{2e^2}{\sqrt{2\pi}} - 1 \right) e^{-(a^*+1)} \\
&\geq 1 - \left(\frac{2e^2}{\sqrt{2\pi}} - 1 \right) e^{-(a_0+1)}.
\end{aligned}$$

The last inequality is due to Inequality (21). Now, using Inequality (22), we have

$$C_{a^*} \geq 1 + \left(1 - c_1 e^{-(a_0+1)}\right) e^{-(a_0+1)}$$

where $c_1 = 2e^2/\sqrt{2\pi} - 1$. From this lower bound on C_a^* and using $a_0 = \log \frac{m}{ne}$, we derive that

$$m \log C_{a^*} \geq m \log \left(1 + \left(1 - O\left(\frac{n}{m}\right)\right) \frac{n}{m}\right).$$

Therefore, Inequality (7) follows. ■

Appendix C

Theorem 0: Let $M(k) = k^k e^{-k}/k!$ denote the Stirling ratio measure for $k = 0, 1, \ldots$ as defined before. Let $M^m = \otimes_{j=1}^m M$ assign a product measure to $\underline{N} = (N_1, \ldots, N_m)$. Let M_{cond} be the probability distribution on $\underline{N}$ obtained from conditioning on $\frac{1}{m}\sum_{j=1}^m N_j = \alpha$ (suppose α is a value that the average of the N_j's is possible to obtain). Define $P_a(k) = M(k)\frac{e^{-ak}}{C_a}$ for an a chosen by the condition $\mathbf{E}_{P_a} N_1 = \alpha$ (suppose such an a can be obtained). Let $\mathcal{C}_\alpha$ be a class of distributions with the expected value of the average of N_j equal to α

$$\mathcal{C}_\alpha = \{P : \mathbf{E}_P \frac{1}{m}\sum_{j=1}^m N_j = \alpha\}.$$

Then, $Q_a = \otimes_{j=1}^m P_a$ is the information projection of M on $\mathcal{C}_\alpha$ in the sense of uniquely minimizing $D(Q||M)$ among all Q in C_α. In fact,

$$D(Q||M^m) = D(Q||Q_a) + D(Q_a||M^m)$$

for all $Q \in C_\alpha$. In particular, we have

$$D(M_{cond}||M^m) = D(M_{cond}||Q_a) + D(Q_a||M^m).$$

Therefore, equality (2) stands.

This is similar to what has been shown in [10], [11], and [12]. Theorem 0 says the tilted distribution is closest to the original distribution in relative entropy among all distributions with the expected value of a function equal to α. Hence it is the redundancy minimizing distribution over the class of distributions with a given moment condition. Note that $D(Q||M^m)$ and $D(Q_a||M^m)$ could be negative since M^m is not a probability measure, but $D(Q||Q_a) \geq 0$ for all $Q \in C_\alpha$.

Proof: For any $Q \in C_\alpha$ and $m \geq 1$,

$$\begin{aligned}
&D(Q||M^m)\\
&= \sum_{N_1,\ldots,N_m} Q(N_1,\ldots,N_m) \log \frac{Q(N_1,\ldots,N_m)}{Q_a(N_1,\ldots,N_m)}\\
&\quad + \sum_{N_1,\ldots,N_m} Q(N_1,\ldots,N_m) \log \frac{Q_a(N_1,\ldots,N_m)}{M^m(N_1,\ldots,N_m)}\\
&= D(Q||Q_a) + \mathbf{E}_Q \left(\log e^{-a\sum_{j=1}^m N_j} \right)\\
&\overset{(a)}{=} D(Q||Q_a) + \mathbf{E}_{Q_a} \left(\log e^{-a\sum_{j=1}^m N_j} \right)\\
&\overset{(b)}{=} D(Q||Q_a) + D(Q_a||M^m)\\
&\geq D(Q_a||M^m).
\end{aligned}$$

Here (a) is because Q_a and Q are both in the convex set $\mathcal{C}_\alpha$, and (b) holds since $Q_a(N_j) = M(N_1,\ldots,N_m)\frac{e^{-a\sum_{j=1}^m N_j}}{C_a^m}$. ■

APPENDIX D REDUNDANCY

Theorem 4: Consider the family of distributions that makes $N_1,\ldots,N_m$ independent Poisson $\lambda_1,\ldots,\lambda_m$. Let $\lambda_{sum} = \sum_{j=1}^m \lambda_j$, and let $\mathcal{P}_{\lambda_{sum}}^m$ denote the family. The redundancy of using a tilted Stirling ratio distribution Q_a on the counts generated by any $P_{\underline{\lambda}}^m \in \mathcal{P}_{\lambda_{sum}}^m$ is mainly

$$r(Q_a, P_{\underline{\lambda}}) = \underbrace{\left((-\frac{m}{2} + a\lambda_{sum}) \log e + m \log C_a \right)}_{(A)},$$

with the error bounded by

$$\sum_{j=1}^m \left(\frac{1}{3\lambda_j^2} + \frac{5}{6\lambda_j}\right) \log e.$$

Moreover, the minimizer of the redundancy is a^*, with a^* chosen by making $\mathbf{E}_{P_a} N_1 = \lambda_{sum}/m$.

When $m = o(\lambda_{sum})$, term (A) satisfies the following inequality

$$0 \leq \left| (A) - \frac{m}{2} \log \frac{\lambda_{sum}}{m} \right| \leq m \log\left(1 + \sqrt{\frac{m}{\lambda_{sum}}}\right). \tag{24}$$

When $\lambda_{sum} = o(m)$, term (A) satisfies the following inequality

$$\begin{aligned}
&m \log\left(1 + \frac{\lambda_{sum}}{m}\right) - \lambda_{sum} \log e\\
&\leq \left| (A) - \left(\lambda_{sum} \log \frac{m}{\lambda_{sum}} - \frac{m}{2} \log e \right) \right|\\
&\leq \frac{1}{2\sqrt{\pi}} \frac{\lambda_{sum}^2 e^2}{m - \lambda_{sum} e} \log e.
\end{aligned} \tag{25}$$

Remark 8: The expression (A) for the redundancy agrees with the regret $a^*\lambda_{sum} \log e + m \log C_{a^*}$ except for the $-\frac{m}{2} \log e$. This difference is due to the difference in the numerator in which the expected $\log P_{\underline{\lambda}}(\cdot)$ is used in the redundancy, and $\log P_{\hat{\underline{\lambda}}}(\cdot)$ is used in regret. Here the expected difference $\mathbf{E} \log \frac{P_{\hat{\underline{\lambda}}}(\cdot)}{P_{\underline{\lambda}}(\cdot)}$ is shown to be near $-\frac{m}{2} \log e$. A similar phenomenon occurs in [27].

Proof: The first part of the proof follows Lemma 3 in [5], and the second part resembles the proof of Theorem 1.

$$\begin{aligned}
&\mathbf{E}_{\underline{\lambda}} \ln \frac{\prod_{j=1}^m P_{\lambda_j}(N_j)}{Q_a(\underline{N})}\\
&= \sum_{j=1}^m \left(\lambda_j \ln \lambda_j \right) - \sum_{j=1}^m \mathbf{E}_{\lambda_j} \left(N_j \ln N_j \right) + a\lambda_{sum}\\
&\quad + m \ln C_a
\end{aligned} \tag{26}$$

Following Lemma 3 in [5], by Taylor's expansion, for each j,

$$\begin{aligned}
&\mathbf{E}_{\lambda_j} \left(N_j \ln N_j \right)\\
&\geq \lambda_j \ln \lambda_j + \mathbf{E}_{\lambda_j}(N_j - \lambda_j)(1 + \ln \lambda_j)\\
&\quad + \mathbf{E}_{\lambda_j} \frac{1}{2}(N_j - \lambda_j)^2 \frac{1}{\lambda_j} + \frac{1}{6}\mathbf{E}_{\lambda_j}(N_j - \lambda_j)^3 \left(-\frac{1}{\lambda_j^2}\right)\\
&= \lambda_j \ln \lambda_j + \frac{1}{2} - \frac{1}{6\lambda_j}.
\end{aligned}$$

We also know by Jensen's inequality that

$$\mathbf{E}_{\lambda_j} \left(N_j \ln N_j \right) \geq \lambda_j \ln \lambda_j.$$

Hence,

$$\mathbf{E}_{\lambda_j} \left(N_j \ln N_j \right) \geq \lambda_j \ln \lambda_j + \frac{1}{2} + \max\left(-\frac{1}{6\lambda_j}, -\frac{1}{2}\right).$$

And by Inequality (30) in [5],

$$\begin{aligned}
&\mathbf{E}_{\lambda_j} \left(N_j \ln N_j \right)\\
&\leq \lambda_j \ln \lambda_j + (\mathbf{E}_{\lambda_j} N_j - \lambda_j)(1 + \ln \lambda_j)\\
&\quad + \frac{\mathbf{E}_{\lambda_j}(N_j - \lambda_j)^2}{2\lambda_j} - \frac{\mathbf{E}_{\lambda_j}(N_j - \lambda_j)^3}{6\lambda_j^2}\\
&\quad + \frac{\mathbf{E}_{\lambda_j}(N_j - \lambda_j)^4}{3\lambda_j^3}\\
&= \lambda_j \ln \lambda_j + \frac{1}{2} + \frac{1}{3\lambda_j^2} + \frac{5}{6\lambda_j}.
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&-\left(\sum_{j=1}^{m}\frac{1}{3\lambda_j^2}+\frac{5}{6\lambda_j}\right)\\
&\quad\le \mathbf{E}_{\underline{\lambda}}\ln\frac{\prod_{j=1}^{m}P_{\lambda_j}(N_j)}{Q_a(\underline{N})}\\
&\qquad-\left(-\frac{m}{2}+a\lambda_{sum}+m\ln C_a\right)\\
&\quad\le \min\left(\sum_{j=1}^{m}\frac{1}{6\lambda_j},\frac{m}{2}\right).
\end{aligned}
$$

The fact that a^* is the minimizer can be easily seen by taking partial derivative with respect to a for the redundancy expression (26). The two inequalities are attributed to Lemma 1, by picking $a = m/(2\lambda_{sum})$ and $a = \ln(m/\lambda_{sum}e)$ respectively. ■

APPENDIX E
PROOF OF THEOREM 3

Proof: The MLE for an envelope class is the following

$$\hat{\lambda}_j = \arg\sup_{\lambda_j \le nf(j)} P_{\lambda_j}(N_j) = N_j \wedge nf(j),$$

where $\wedge$ denotes the minimum.

We formulate a tilted distribution by multiplying the exponential tilting factor e^{-aN_j} for each $j \in \{1,\ldots,m\}$ and normalize it.

$$
P_a(N_j) = \begin{cases}\frac{N_j^{N_j}e^{-N_j}}{N_j!}\frac{e^{-aN_j}}{C_{a,j}} & \text{if } N_j \le nf(j)\\ \frac{(nf(j))^{N_j}e^{-nf(j)}}{N_j!}\frac{e^{-aN_j}}{C_{a,j}} & \text{if } N_j > nf(j)\end{cases}
$$

where $C_{a,j} = \sum_{N_j \le nf(j)}\frac{N_j^{N_j}e^{-N_j}}{N_j!}e^{-aN_j} + \sum_{N_j > nf(j)}\frac{(nf(j))^{N_j}e^{-nf(j)}}{N_j!}e^{-aN_j}$.

The regret of using independent P_a for each N_j in $\underline{N} \in S_{m,n}$ is

$$\log\prod_{j=1}^{m}\frac{P_{\hat{\lambda}_j}(N_j)}{P_a(N_j)} = na\log e + \sum_{j=1}^{m}\log C_{a,j}. \tag{27}$$

Again, a^* minimizes expression (27).

For each j and any positive a,

$$
\begin{aligned}
C_{a,j} = &\sum_{N_j \le \lfloor nf(j)\rfloor}\frac{N_j^{N_j}e^{-N_j}}{N_j!}e^{-aN_j}\\
&+\sum_{N_j > nf(j)}\frac{(nf(j))^{N_j}e^{-nf(j)}}{N_j!}e^{-aN_j}.
\end{aligned}
$$

The sum only depends on the envelope function $f(j)$ for given a and j.

Since $(nf(j))^x e^{-nf(j)} \le x^x e^{-x}$ for all $x > 0$, for any symbol j with $N_j > nf(j)$, we have

$$\frac{(nf(j))^{N_j}e^{-nf(j)}}{N_j!}e^{-aN_j} \le \frac{N_j^{N_j}e^{-N_j}}{N_j!}e^{-aN_j}.$$

Hence we have,

$$C_{a,j} \le \sum_{N_j=0}^{\infty}\frac{N_j^{N_j}e^{-N_j}}{N_j!}e^{-aN_j} \le 1+\sqrt{\frac{1}{2a}}.$$

The second inequality is due to Lemma 1.

However, if $nf(j)$ is small, the following upper bound is better. For $N_j \le \lfloor nf(j)\rfloor$,

$$
\begin{aligned}
\sum_{N_j \le \lfloor nf(j)\rfloor}\frac{N_j^{N_j}e^{-N_j}}{N_j!}e^{-aN_j} &\le \sum_{N_j \le \lfloor nf(j)\rfloor}\frac{N_j^{N_j}}{N_j!}\\
&\le \sum_{N_j \le \lfloor nf(j)\rfloor}\frac{(nf(j))^{N_j}}{N_j!}.
\end{aligned}
$$

For the second partial sum, we also have

$$
\begin{aligned}
&\sum_{N_j > nf(j)}\frac{(nf(j))^{N_j}e^{-nf(j)}}{N_j!}e^{-aN_j}\\
&\quad\le \sum_{N_j > nf(j)}\frac{(nf(j))^{N_j}}{N_j!}.
\end{aligned}
$$

Deduce,

$$C_{a,j} \le \sum_{N_j=0}^{\infty}\frac{(nf(j))^{N_j}}{N_j!} = e^{nf(j)}.$$

Hence for any given a, j and $L \in \{1,2,\ldots,m\}$, the following upper bound holds.

$$
\begin{aligned}
&na\log e + \sum_{j=1}^{m}\log C_{a,j}\\
&\quad\le na\log e\\
&\qquad+\log\left(\prod_{j=1}^{L}\left(1+\sqrt{\frac{1}{2a}}\right)\prod_{j=L+1}^{m}\left(e^{nf(j)}\right)\right)\\
&\quad= na\log e + L\log\left(1+\sqrt{\frac{1}{2a}}\right)\\
&\qquad+\left(\sum_{j=L+1}^{m}nf(j)\right)\log e.
\end{aligned}
$$

Let $a = \frac{L}{2\left(n-\sum_{j>L}nf(j)\right)}$, the result follows. ■

APPENDIX F
INCOMPATIBILITY OF P_n

$$
\begin{aligned}
&\sum_{x\in\mathcal{A}}P_{n+1}(X_1,\ldots,X_n,X_{n+1}=x)\\
&\quad=\sum_{x\in\mathcal{A}}\frac{1}{\binom{n+1}{N_1^n\ldots N_x^n+1\ldots N_m^n}}\frac{Q_a(N_1^n,\ldots,N_x^n+1,\ldots,N_m^n)}{Q_a(S_{m,n+1})}\\
&\quad=\underbrace{\frac{1}{\binom{n}{N_1^n\ldots N_x^n\ldots N_m^n}}\frac{M^m(\underline{N}^n)}{M^m(S_{m,n})}}_{(A)}\underbrace{\frac{M^m(S_{m,n})}{M^m(S_{m,n+1})}}_{(B)}\\
&\qquad\underbrace{\sum_{x\in\mathcal{A}}\left(\frac{N_x^n+1}{n+1}\frac{M(N_x^n+1)}{M(N_x^n)}\right)}_{(C)}.
\end{aligned}
$$

Term (A) equals to the distribution of the count vector $\underline{N}^n$ conditioning on its total equal to n through expression (10). Hence, it suffices to check whether the rest equals to 1. This is obviously not true, since term (C) equals

$$\frac{e^{-1}}{n+1}\sum_{x\in\mathcal{A}}\frac{(N_x^n+1)^{N_x^n+1}}{{N_x^n}^{N_x^n}}$$

which depends on the specific value of the count vector $\underline{N}^n$, while the ratio $M^m(S_{m,n})/M^m(S_{m,n+1})$ is a constant given m and n. Hence the P_n's are not compatible.

Appendix G
Computation Complexity

The computations of arithmetic coding ingredients of the two pass codes are examined. One sees that each step involves at most order $n\log m$ or order $m\log n$ bits operations. For some steps of computation log factors of computation may be possible to avoid, but we will not belabor such reductions. Moreover, we quantify the additional cost of the Shtarkov code (conditional on n) compared to the code that makes the counts i.i.d.

As a preliminary step the counts are calculated for each symbol, and we flag which symbols have positive counts. (Recall that m^* denotes the number of symbols with positive counts). The data are initially in the form of n observations $X_1,\ldots,X_n$ of symbols X_i stored in binary, $\log m$ bits each. Initializing the m counts at 0, in one pass through the data increment by one the count addressed by each of the observed X_i, for $i=1,\ldots,n$. This entails $n\log m$ binary operations (counting addressing as $\log m$).

As we have said the first pass is to code the counts either by using the tilted Stirling ratio distribution or by using the exact minimax distribution obtained by conditioning on n.

Let's examine the first pass using the tilted Stirling ratio distribution by arithmetic coding [28]–[30]. The essence of this encoding is the iterative calculation of the cumulative probabilities to the left of $N_1,\ldots,N_j$, for $j=1,\ldots,m$. As discussed the probabilities $P_a(i)$ for $i=1,\ldots,n$ have been precomputed. Each can be accessed from memory with a $\log n$ bits address. Likewise for the cumulative marginal probabilities defined by $P_{a,1}^{cum}(k)=\sum_{i=0}^{k-1}P_a(i)$ for $k=1,\ldots,n$, with $P_{a,1}^{cum}(k)$ set to 0 for $k=0$. Initialize the iterations with $P_{a,1}^{cum}(N_1)$. Then for $j\ge 1$,

$$\begin{aligned}&P_{a,j+1}^{cum}(N_1,\ldots,N_j,N_{j+1})\\&\quad=\begin{cases}P_{a,j}^{cum}(N_1,\ldots,N_j) & \text{if } N_{j+1}=0\\ P_{a,j}^{cum}(N_1,\ldots,N_j) & \text{if } N_{j+1}>0\\ +Q_a^j(N_1,\ldots,N_j)P_{a,1}^{cum}(N_{j+1}).\end{cases}\end{aligned}$$

It is only at the flagged symbols with positive counts that the cumulative probability needs to be updated. So these updates to the cumulative probabilities performs only $m^*\le\min\{n,m\}$ multiplication and addition operations, and the associated bit complexity is at most $\min\{n,m\}\log n$.

Meanwhile the joint probabilities $Q_a^j(N_1,\ldots,N_j)$ used here are products of $P_a(N_1)$ through $P_a(N_j)$ for $j=1,\ldots,m$. These can be computed by updates in which for $j=1,\ldots,m-1$ we multiply by $P_a(N_{j+1})$ for the next iteration (again accessed using $\log n$ bits operations). All of these factors, even those where the counts are 0, are needed to get the proper partial products. So this is an order $m\log n$ operation if performed this way. Here the m may be reduced to $m^*\le\min\{m,n\}$ if we only encode the flagged positive counts (this would entail computations using the conditional distribution given the set of positive counts which we do not explore here).

One sees that the core of the arithmetic coding is the use of updates based on the n stored $P_a(i)$ and their associated $P_{a,1}^{cum}(k)$.

Here we have focused on the mathematical essence. As explained in [29] and [30] practical implementation requires careful additional computation to avoid underflow. This involves computing also the cumulatives including the current $(N_1,\ldots,N_j)$, that is $P_{a,j}^{cum,+}(N_1,\ldots,N_j)$ equal to $P_{a,j}^{cum}(N_1,\ldots,N_j)+Q_a^j(N_1,\ldots,N_j)$. When their binary representations are in agreement in their leading ℓ bits (these are the initial ℓ code bits), the values may be scaled by subtracting the part in agreement and shifting left by ℓ, i.e. multiplying by 2^ℓ (noting that in this case the first ℓ bits of $Q_a^j(N_1,\ldots,N_j)$ are zeros). These rescalings are repeated whenever there is such agreement. A related matter we are not addressing here in detail is the number of bits of precision with which the $P_a(i)$ (and their products and cumulatives) are to be computed, remarking only that the final number of bits of the $P_{a,m}^{cum}$ should be of the order of the length of the code which is $\log 1/Q_a^m(N_1,\ldots,N_m)$.

The second pass is to use arithmetic coding to encode the string $X_1,\ldots,X_n$ given the counts $N_1,\ldots,N_m$. Note that being given the counts for the symbols ordered as $1,\ldots,m$ provides a sorted list of the observed symbols with repeats counted. Initialize with $P(X_1|N_1,\ldots,N_m)=N_{X_1}/n$, which is evaluated at X_1. The corresponding cumulative probability to the left of X_1 is

$$F_-(X_1|N_1,\ldots,N_m)=\frac{L_{X_1}}{n},$$

where L_{X_1} is the count of symbols to the left of X_1. For the next step, the relevant counts are for $X_2,\ldots,X_n$. Accordingly we decrease the count of N_{X_1} and decrease the cumulative counts L_x for all $x>X_1$. Then for $i\ge 1$, having decreased by 1 the counts $N_{X_i}^{rem}$ and the cumulative counts L_x^{rem} for $x>X_i$, we proceed to set the conditional probability of the next symbol given the past and the counts (as given in Subsection IV-B) to be the relative frequency of x in the remaining string

$$Prob(X_{i+1}|X_1,\ldots,X_i,(N_1,\ldots,N_m))=\frac{N_{X_{i+1}}^{rem}}{n-i}.$$

where $N_{X_{i+1}}^{rem}=N_{X_{i+1}}-N_{X_{i+1},i}$. And this associate cumulative conditional probability to the left of X_{i+1} is

$$F_-(X_{i+1}|X_1,\ldots,X_i,(N_1,\ldots,N_m))=\frac{L_{X_{i+1}}^{rem}}{n-i}.$$

Arithmetic coding requires calculation of the following probabilities

$$
\begin{aligned}
&Q^{cum}(X_1, \ldots, X_i, X_{i+1}|(N_1, \ldots, N_m)) \\
&\quad = Q^{cum}(X_1, \ldots, X_i|(N_1, \ldots, N_m)) \\
&\qquad + P_i(X_1, \ldots, X_i|(N_1, \ldots, N_m)) \\
&\qquad F_-(X_{i+1}|X_1, \ldots, X_i, (N_1, \ldots, N_m)).
\end{aligned}
$$

Note that for each i, what is needed is the value of $L^{rem}_{X_{i+1}}$ which requires the position of X_{i+1} in the sorted list of the remaining symbols. This requires $\log n$ computation time for each symbol. Therefore, the computation complexity is $O(n \log n)$. Again, these calculations are scaled at each step as in Pasco [29] or Rissanen and Langdon [30] to avoid underflow or overflow.

In a nutshell, the total computational complexity for this two pass code is $O\,(m \log n + n \log mn)$.

For implementation of Shtarkov's code, this can be computed in similar fashion, by two pass arithmetic coding using the distribution conditional on $N = n$. What is different is the first pass arithmetic code for the counts, where in place of the $P_a(i)$ the updates use the conditional probability distribution for the count for symbol $j+1$ expressed (as shown in Subsection IV-C) by

$$
\begin{aligned}
&Q_{nml}(i|N_1, \ldots, N_j, N = n) \\
&\quad = \frac{P_a(i) P_a^{m-j-1}(n - (N_1 + \ldots + N_j + i))}{P_a^{m-j}(n - (N_1 + \ldots + N_j))}. \qquad (28)
\end{aligned}
$$

Adding these on step j for $i < N_{j+1}$ produces the conditional cumulatives $Q^{cum}_{nml}(N_{j+1}|N_1, \ldots, N_j, N = n)$ which replace $P^{cum}_{a,1}(N_{j+1})$ in the code update. Likewise multiplying by this at $i = N_{j+1}$ updates the otherwise elusive joint probabilities $Q_{nml}(N_1, \ldots, N_j|N = n)$.

As before we assume the values of $P_a^{m'}(k)$ for $m' = 1, \ldots, m$ and $k = 0, \ldots, n$ have been precomputed and stored. So a main difference between the conditional and unconditional distribution codes is that in this conditional case we have a storage of size mn for these $P_a^{m'}(k)$ rather than size n for the $P_a^1(k)$. Accessing these entails $\log mn$ bit addressing. Computing the above conditional probabilities for $i = 0, \ldots, N_{j+1}$ is then $1 + N_{j+1}$ operations, which sum across j to be order $m + n$ operations on these values. So the total additional cost is only of order $(m+n)\log mn$ above the value using the independent distribution.

Acknowledgment

The authors would like to thank Prof. Jun'ichi Takeuchi, Prof. Wojciech Szpankowski and Prof. Narayana Santhanam. Some ideas of the subsequent work are based on feedback from these presentations.

References

[1] Y. Leng and Y. Wei, *Zhonghua Zihai*. Taipei, Taiwan: Zhonghua Press, 1994.

[2] *Classic of Poetry*. [Online]. Available: https://en.wikipedia.org/wiki/Chinese_characters AQ:3

[3] X. Yang and A. R. Barron, "Compression and predictive distributions for large alphabet i.i.d and Markov models," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun./Jul. 2014, pp. 2504–2508.

[4] Y. M. Shtar'kov, "Universal sequential coding of single messages," *Problemy Peredachi Informatsii*, vol. 23, no. 3, pp. 3–17, 1987.

[5] Q. Xie and A. R. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 646–657, Mar. 1997.

[6] P. Kontkanen and P. Myllymäki, "A linear-time algorithm for computing the multinomial stochastic complexity," *Inf. Process. Lett.*, vol. 103, no. 6, pp. 227–233, 2007.

[7] A. Barron, T. Roos, and K. Watanabe, "Bayesian properties of normalized maximum likelihood and its fast computation," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun./Jul. 2014, pp. 1667–1671.

[8] R. Krichevsky and V. Trofimov, "The performance of universal encoding," *IEEE Trans. Inf. Theory*, vol. 27, no. 2, pp. 199–207, Mar. 1981.

[9] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1. New York, NY, USA: Wiley, 1968.

[10] I. Csiszar, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, no. 1, pp. 146–158, 1975.

[11] I. Csiszar, "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. Probab.*, vol. 12, no. 3, pp. 768–793, 1984.

[12] J. M. van Campenhout and T. M. Cover, "Maximum entropy and conditional probability," *IEEE Trans. Inf. Theory*, vol. 27, no. 4, pp. 483–489, Jul. 1981.

[13] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Always good turing: Asymptotically optimal probability estimation," *Science*, vol. 302, no. 5644, pp. 427–431, 2003.

[14] W. Szpankowski and M. J. Weinberger, "Minimax redundancy for large alphabets," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 1488–1492.

[15] L. A. Adamic. (2000). Zipf, power-laws, and Pareto—A ranking tutorial. Xerox Palo Alto Research Center, Palo Alto, CA, USA. [Online]. Available: http://ginger.hpl.hp.com/shl/papers/ranking/ranking.html

[16] Y. M. Shtar'kov, T. J. Tjalkens, and F. M. J. Willems, "Multi-alphabet universal coding of memoryless sources," *Problemy Peredachi Informatsii*, vol. 31, no. 2, pp. 20–35, 1995.

[17] Q. Xie and A. R. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 431–445, Mar. 2000.

[18] A. Orlitsky and N. P. Santhanam, "Speaking of infinity [i.i.d. strings]," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2215–2230, 2004.

[19] S. Boucheron, A. Garivier, and E. Gassiat, "Coding on countably infinite alphabets," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 358–373, Jan. 2009.

[20] D. Bontemps, "Universal coding on infinite alphabets: Exponentially decreasing envelopes," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1466–1478, Mar. 2011.

[21] S. Kullback, *Information Theory and Statistics*. North Chelmsford, MA, USA: Courier Corporation, 1997.

[22] H. Robbins, "A remark on Stirling's formula," *Amer. Math. Monthly*, vol. 62, no. 1, pp. 26–29, 1955.

[23] N. G. De Bruijn, *The American Mathematical Monthly*, vol. 4. North Chelmsford, MA, USA: Courier Corporation, 1970.

[24] T. Roos and J. Rissanen, "On sequentially normalized maximum likelihood models," *Compare*, vol. 27, no. 31, p. 256, 2008.

[25] *Classic of Poetry*. [Online]. Available: https://en.wikipedia.org/wiki/Classic_of_Poetry

[26] *GB 18030*. [Online]. Available: https://zh.wikipedia.org/wiki/GB_18030

[27] B. S. Clarke and A. R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Statist. Planning Inference*, vol. 41, no. 1, pp. 37–60, 1994.

[28] F. Jelinek, *Probabilistic Information Theory: Discrete and Memoryless Models*. New York, NY, USA: McGraw-Hill, 1968.

[29] R. C. Pasco, "Source coding algorithms for fast data compression," Ph.D. dissertation, Stanford Univ., Stanford, CA, USA, 1976. AQ:4

[30] J. Rissanen and G. G. Langdon, "Arithmetic coding," *IBM J. Res. Develop.*, vol. 23, no. 2, pp. 149–162, Mar. 1979.

Xiao Yang Xiao Yang received her M.A. and Ph.D. in statistics from Yale University. Her research interests include statistical information theory, universal data compression, data mining and statistical learning methods, and natural language processing. She currently works as a data scientist at Apple Inc, Cupertino.

Andrew R. Barron Professor Andrew R Barron has research interests in the areas of statistical information theory, model selection, the minimum description length principle, probability limit theorems, asymptotics of Bayes procedures, estimation of functions of many variables, artificial neural networks, approximation theory, investment theory, universal data compression, and sparse regression codes for practical capacity-achieving communications. Barron is a fellow of the IEEE, Medallion Prize winner of the Institute of Mathematical Statistics and a winner along with Bertrand Clarke of the IEEE Thompson Prize (for Best Paper in all IEEE Journals for authors under 30 at time of submission). He has served as secretary of the Board of Governors and subsequently has served multiple terms as a member of the Board of Governors of the IEEE Information Theory Society. He currently chairs the Thomas M. Cover Dissertation Prize Committee. Received Ph.D., Electrical Engineering, Stanford University; M.S., Electrical Engineering, Stanford University; B.S. E.E. and Math Science, Rice University. From 1985 - 1992 Andrew was Assistant and then Associate Professor of Statistics and Electrical & Computer Engineering, University of Illinois. From 1992 to present, Andrew is a Professor of Statistics at Yale and has served terms as department chair, director of graduate studies, director of undergraduate studies in Statistics, director of undergraduate studies in Applied Mathematics, and courtesy appointment as Professor of Electrical Engineering at Yale. Xiao.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your paper. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

AQ:1 = Please confirm whether the edits made in the presentation section are OK.
AQ:2 = Please confirm whether the edits made in the current affiliation of all the authors are OK as set.
AQ:3 = Please provide the accessed date for refs. [1], [25], and [26]. Also confirm the title.
AQ:4 = Please provide the department name for ref. [29].