# LARGE ALPHABET CODING AND PREDICTION THROUGH POISSONIZATION AND TILTING

*Xiao Yang and Andrew R. Barron*

Department of Statistics, Yale University
24 Hillhouse Ave, New Haven, CT, USA
Correspondence: xiao.yang@yale.edu, andrew.barron@yale.edu

## ABSTRACT

This paper introduces a convenient strategy for compression and prediction of 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 is close to optimal for a smaller class – the class of distributions with an analogous condition on the counts. Moreover, the method can be used to code and predict sequences in a subset with the tail counts satisfying a given condition, and it can also be applied to envelope classes.

## 1. INTRODUCTION

Large alphabet coding and prediction problems concern understanding the probabilistic scheme of a huge number of possible outcomes. In many cases the ordered probabilities of individual outcomes display a quickly falling shape. An example is language. The set of frequent words that cover our everyday use is only a small portion of the whole vocabulary. Here we consider an i.i.d model for coding and predicting such alphabets. Despite the possible dependence among the outcomes in the alphabet as in text coding/prediction problem, it serves as a starting point and can be extended to models which take dependence into account.

Theoretical analysis usually assumes the length of a message is known in advance when it is coded. This is not always true in practice. Serialization writers do not know how many words a novel contains before he finishes the last sentence. Nevertheless, given a limited time/space, one could possibly guess how many words on average can be accommodated.

Suppose a string of random variables $\underline{X} = (X_1, ..., X_N)$ is generated independently from a discrete alphabet $\mathcal{A}$ of size $m$. We allow the sequence length $N$ to be variable. A special case is when $N$ is given as a fixed number, or it can be random. In either case, $\underline{X}$ is a member of the set $\mathcal{X}^*$ of all finite length sequences

$$\begin{aligned}\mathcal{X}^* &= \bigcup_{n=0}^{\infty} \mathcal{X}^n \\ &= \bigcup_{n=0}^{\infty} \{x^n = (x_1, ..., x_n) : x_i \in \mathcal{A}, i = 1, ..., n\}.\end{aligned}$$

Our goal is to code/predict the sequence $\underline{X}$. Note that the length $N$ is determined by the string. There will be an agreed upon distribution of the string length $N$, perhaps Poisson or deterministic.

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

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

for $n = 1, 2, ...$ Here $\theta$ only parameterizes the distribution of the sequence conditional on the length. Of particular interest is the class of all distributions on the given alphabets with $P_\theta(j) = \theta_j$ parameterized by the simplex $\Theta = \{\underline{\theta} : \theta_j \geq 0, \sum_{j=1}^m \theta_j = 1\}$.

Let $\underline{N} = (N_1, ..., N_m)$ denote the counts of symbols $1, ..., m$ that occur in the sequence $\underline{X}$. The observed sample size $N$ is the sum of the counts $N = \sum_{j=1}^m N_j$. Both $P_\theta(\underline{X})$ and $P_{\underline{\theta}}(\underline{X}|N = n)$ have factorizations based on the distribution of the counts.

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

and

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

The first factor of the two equations is the uniform distribution on the set of strings with given counts, which does not depend on $\theta$. Hence, the vector of counts $\underline{N}$ forms a sufficient statistic for $\theta$. Modeling the distribution of the counts is essential for forming codes and predictions.

The independent $Poisson(\lambda_j)$, $j = 1, ..., m$ family of distributions on counts is sufficiently rich via conditioning on the total counts to account for all i.i.d. distributions on

the finite alphabet, with the relationship $\theta_j = \frac{\lambda_j}{\lambda_{sum}}$ where $\lambda_{sum} = \sum_{j=1}^{m} \lambda_j$.

The task of coding a string is equivalent to providing a probabilistic scheme. A coder $Q$ for the string is also a (sub)probability distribution on $\mathcal{X}^*$ which assigns probabilities $Q(X_1, ..., X_N)$ to strings $X_1, ..., X_N$ and produces a binary string of length $\log 1/Q(\underline{X})$ (we do not worry about the integral constraint). Ideally the true probability distribution $P_\theta(X_1, ..., X_N)$ could be used if $\theta$ were known. So the regret induced by using $Q$ instead of $P_\theta$ is

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

where $\log$ is logarithm base 2.

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.

$$R(Q, P_\theta, \underline{X}) = \log \frac{P_\theta(\underline{N})}{Q(\underline{N})}.$$

Given the family $\mathcal{P}_\Theta$, consider the best candidate with hindsight $P_{\hat{\theta}}(\underline{X})$ which achieves the maximum value at $P_{\hat{\theta}}(\underline{X}) = \max_{\theta\in\Theta} P_\theta(\underline{X})$ (also corresponding to $\min_{\theta\in\Theta} \log(1/P_\theta(\underline{X}))$), where $\hat{\theta}$ is the maximum likelihood estimator of $\theta$. The maximization is equivalent to maximizing $\theta$ for the count probability, as the uniform distribution dose not depend on $\theta$, i.e.

$$\begin{aligned} \max_{\theta\in\Theta}(P_\theta(\underline{X})) &= P(\underline{X}|\underline{N}) \max_{\theta\in\Theta} P_\theta(\underline{N}) \\ &= P(\underline{X}|\underline{N})P_{\hat{\theta}}(\underline{N}). \end{aligned}$$

Then the problem becomes: given a family of distributions $\mathcal{P}_\theta$, how to choose $Q$ to minimize the maximized pointwise redundancy,

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

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

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

As is familiar in universal coding [1], [2] the normalized maximum likelihood (NML) distribution

$$Q^*(\underline{N}) = \frac{P_{\hat{\theta}}(\underline{N})}{C(S_{m,n})},$$

is the unique pointwise minimax strategy when $C(S_{m,n}) = \sum_{\underline{N}\in S_{m,n}} P_{\hat{\theta}}(\underline{N})$ is finite, and $\log C(S_{m,n})$ is the minimax value. When $m$ is large, the NML distribution can be unwieldy to compute for compression or prediction. Instead we will introduce a slightly suboptimal coding distribution that makes the counts independent and show that it is nearly optimal for every $S_{m,n'}$ with $n'$ not too different from a target $n$. Indeed we advocate that our simple coding distribution is 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 two basic principles. One is that the multinomial family of distributions of counts matches the conditional distribution of $N_1, ..., N_m$ given the sum $N$ when unconditionally the counts are independent Poisson. Another is the information theory principle [3][4][5] 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.

Apply the maximum likelihood step to the independent counts. This produces a maximized likelihood value of $M(N_j) = N_j^{N_j} e^{-N_j}/N_j!$, for each $j \in \{1, ..., m\}$. Although this measure has an infinite sum by itself, it is normalizable when tilted for every positive $a$. The tilted distribution is

$$P_a(N_j) = \frac{N_j^{N_j} e^{-N_j}}{N_j!} \frac{e^{-aN_j}}{C_a},$$

with the normalizer $C_a = \sum_{N_j=0}^{\infty} N_j^{N_j} e^{-(1+a)N_j}/N_j!$.

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,

$$Q_a(N_1, ..., N_m) = P_a(N_1)...P_a(N_m).$$

By allowing description of all possible counts $N_j \geq 0$, $j = 1, ..., m$, our code length will be greater for some sequences than code lengths 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 $N_1, ..., N_m$ given $N$) adds $\log 1/P(N = n)$ which is approximately $\frac{1}{2} \log 2\pi n$ bits to the description length beyond that which would have been ideal $\log 1/Q_a(N_1, ..., N_m|N = n)$ if $N = n$ were known. This ideal code length constructed from the tilted maximized Poisson, when conditioning on $n$, matches the Shtarkov's normalized maximum likelihood based on the multinomial.

For small alphabet with $m << n$, the minimax redundancy is about $\frac{1}{2} \log n$ bits per free parameter (i.e. a total of $\frac{m-1}{2} \log n$ + constant); and for large alphabet when $m \sim n$ and $n = o(m)$, the minimax redundancy is about $O(n)$ and $n \log \frac{m}{n}$ respectively [1][2][6][7]. The additional $\frac{1}{2} \log n$ bits is a small price to pay for the sake of

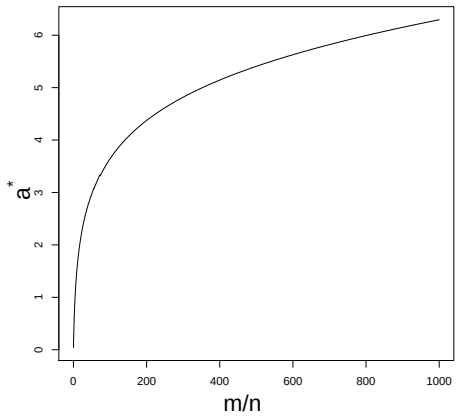


Figure 1. Relationship between $a^*$ and $m/n$.

gaining the coding simplification and additional flexibility.

If it is known that the total count is $n$, then the pointwise redundancy is a simple function of $n$ and the normalizer $C_a$. The choice of the tilting parameter $a^*$ given by the moment condition $E_{Q_a} \sum_{j=1}^{m} N_j = n$ minimizes the pointwise redundancy over all positive $a$. This arises by differentiation because $\frac{\partial}{\partial a} \log C_a$ is equal to the given moment. Moreover, $a^*$ depends only on the ratio between the size of the alphabet and the total counts $m/n$. Figure 1 displays $a^*$ as a function of $m/n$ solved numerically. Given an alphabet with $m$ symbols and a string generated from it of length $n$, one can check the plot and find the $a^*$ desired according to the $m/n$ given, and then use the $a^*$ to code or predict the data.

If, however, the total count $N$ is not given, then the pointwise redundancy depends on $N$. We use a mixture of $a$ to account for the lack of knowledge in advance, and details are contained in section 3.5.

Shtarkov studied the universal data compression problem and identified the exact pointwise minimax strategy [1]. He showed the asymptotic minimax lower bound for the regret is $\frac{m-1}{2} \log n + O(1)$, in which the parameter set $\Theta$ 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 [1], because the computational inconvenience of computing the normalizing constant. Shtarkov [8] also pointed out that the typical sequence usually contains $M << m$ different symbols, and the regret depends mainly on $M$ instead of $m$. Xie and Barron [2][9] 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 expected regret and the pointwise 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 [10] considered the problem in a large alphabet setting in which the number of symbols $m$ is much larger than the sequence length $n$ or even infinite. They found the main terms in the pointwise minimax redundancy 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 [7] provided more precise asymptotics in these settings. They also calculated the minimax redundancy of a source model in which some symbol probabilities are fixed. Boucheron, Garivier and Gassiat [11] 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 [12] worked on exponentially decreasing envelope classes and provided a minimax strategy and the corresponding regret.

In this paper, we introduce a straightforward and easy to implement method for large alphabet coding and prediction. The purpose is three-fold: first, by allowing the sample size to be variable, we are considering a larger class of distributions. This is a more realistic and less restrictive assumption than presuming a particular length. But the method can also be used for fixed sample size coding and prediction.

Second, it unveils an information geometry of three key distributions or measures in the problem: the unnormalized maximum Poisson likelihood measure $M$ of the counts, the conditional distribution $M_{cond}$ of $M$ given the total count equal to $n$, which matches Shtarkov's normalized maximum multinomial likelihood distribution, and a tilted distribution $P_a$, with tilting parameter $a$. The tilted distribution $P_a$ is closest to the original distribution $M$ in Kullback-Leibler distance within the class $\mathcal{C}$ of distributions with the moment equal to the observed value. Hence $P_a$ is the information projection of $M$ onto $\mathcal{C}$. Moreover, since $M_{cond}$ is also in $\mathcal{C}$, the Pythagorean-like equality holds [13][3], i.e.

$$D(M_{cond}||M) = D(M_{cond}||P_a) + D(P_a||M).$$

The case of a tilted distribution (the information projection) as an approximating conditional distribution is investigated in [5] and [4]. A difference here is that our unconditional measure $M$ is not normalizable.

Thirdly, the strategy designed through an independent Poisson model and tilting method is much easier to calculate and implement as compared to the strategies based on multinomial models. The convenience is gained through independence.

The paper is organized in the following way. Section II introduces the model, Section III provides general results and outlines the proof, and Section IV extends the result to subsets of sequences satisfying a tail condition and an envelope class. Details of proof are left in a journal version of this paper.

## 2. THE POISSON MODEL

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

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

independently, and $N$ also has a Poisson distribution

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

Write $\underline{\lambda} = (\lambda_1, ..., \lambda_m)$. 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 the MLE for each $\lambda_j$ is $\hat{\lambda}_j = N_j$, and the first term is a uniform distribution which does not depend on $\lambda$. So

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

where $M(k) = k^k e^{-k}/k!$, $k = 1, 2, ...$(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 pointwise redundancy is

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

This method can also be applied to fixed total counts scenario, which reduces 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

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

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 $j = 1, ..., m$. Note that a sequence $X_1, ..., X_N$ with counts $N_1, ..., N_m$ of total $N = n$ satisfies

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

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

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

Now again if we use $Q(N_1, ..., N_m)$ to code the counts, then the pointwise redundancy is

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

Here $\hat{\lambda}_{sum} = n$, hence the term $\frac{1}{2} \log 2\pi n$ is the Sterling's approximation of $\log(1/P_{\hat{\lambda}_{sum}}(N = n))$. The $\frac{1}{2} \log 2\pi n$ arises because here $Q$ includes description of the total $N$ while the more restrictive target regarded it as given.

## 3. RESULTS

### 3.1. Pointwise redundancy for a general string set

Let $S$ be any set of strings. The maximized redundancy of using $Q$ as a coding strategy given a class $\mathcal{P}$ of distributions when the string is restricted to $S$ is

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

**Theorem 1.** *Let $P_a(k) = k^k e^{-(1+a)k}/k!\, C_a$ be the distribution for $N_j$ specified in the introduction (Poisson maximized likelihood, tilted and normalized). Let $\mathcal{P}_m^{Poi}$ be the class of $m$ independent Poisson distributions. The pointwise redundancy of using a product of tilted distributions $Q_a = \otimes_{j=1}^{m} P_a$ for a given $\underline{N} = (N_1, ..., N_m)$ is*

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

*Let $S_{m,n}$ be the set of counts with total count $n$ be defined as before, then for each $\underline{N} \in S_{m,n}$,*

$$R\left(Q_a, \mathcal{P}_m^{Poi}, S_{m,n}\right) = an \log e + m \log C_a. \quad (2)$$

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

$$E_{P_a} \sum_{j=1}^{m} N_j = m\, E_{P_a} N_1 = n. \quad (3)$$

*Then $a^*$ is the minimizer of the pointwise redundancy in expression (2), written as $R_{m,n} = \min_a R(Q_a, \mathcal{P}_m^{Poi}, S_{m,n})$.*

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

$$\begin{aligned} -d_1 \frac{m}{2} \log e \;\; &\leq \;\; R_{m,n} - \frac{m}{2} \log \frac{ne}{m} \\ &\leq \;\; m \log(1 + \sqrt{\frac{m}{n}}), \end{aligned} \quad (4)$$

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

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

$$\begin{aligned} m \log\left(1 + (1 - d_2)\frac{n}{m}\right) \;\; &\leq \;\; R_{m,n} - n \log \frac{m}{ne} \\ &\leq \;\; m \log\left(1 + \frac{n}{m} + d_3\right) \end{aligned} \quad (5)$$

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

*Outline of proof.* The expression of the pointwise redundancy is from the definition. The fact that $a^*$ is the minimizer can be seen by taking partial derivative with respect to $a$ of expression (2). The bounds is derived by approximating $a^*$ by $m/2n$ and $\log(m/ne)$ respectively $\square$

**Remark 1**: This theorem shows the product of titled distributions with the tilting parameter given by the moment condition minimizes pointwise redundancy over all tilted distributions. The redundancy is close to the minimax level in either small or large alphabets. The main terms in the last two approximations are the same as what is given in [7] except the multiplier for $\log(n/m)$ here is $m/2$ instead of $(m-1)/2$ for the small $m$ scenario.

**Corollary 1.** *Let $\mathcal{P}_{m,n}^{mul}$ be a family of multinomial distributions with total counts $n$. Then the maximized redundancy $R(Q_a, \mathcal{P}_{m,n}^{mul}, S_{m,n})$ has an upper bound within $\frac{1}{2}\log 2\pi n$ above the upper bounds in Theorem 1.*

*Proof.* This can be easily seen by equation (1). □

### 3.2. Subset of sequences with partitioned counts

One advantage of using independent tilted 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 is highly skewed.

The following theorem holds for constraints on a tail sum of counts $\sum_{j>L} N_j = nf$. Small remainder occur in the following redundancy bound when $\frac{nf}{m-L}$ and $\frac{L}{n-nf}$ are small.

**Theorem 2.** *Let $S_{m,n,f,L}$ be a subset of sequences with tail counts controlled by a given function $0 \le f \le 1$, i.e., $S_{m,n,f,L} = \{\underline{N} = (N_1, ..., N_m) : \sum_{j=1}^m N_j = n, \sum_{j>L} N_j = nf\}$. Here $L$ is a number between $1$ and $m$. The pointwise redundancy of using independent tilted distributions for sequences in $S_{m,n,f,L}$ given each $L \in \{1, ..., m\}$ is mainly*

$$\frac{L}{2}\log\frac{(n-nf)e}{L} + nf\log\frac{(m-L)}{nfe}. \qquad (6)$$

*The difference between the exact value and the main terms are bounded below by $r_1$ and above by $r_2$, where*

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

*and*

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

*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 $d_3 = \frac{1}{2\sqrt{\pi}}\frac{(nfe)^2}{(m-L)((m-L)-nfe)}$.*

*Outline of proof.* Consider the product distribution,

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

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 $j > L$. The result can be derived by applying Theorem 1 to $R(Q_a, \mathcal{P}_L, S_{L,n-nf})$ and $R(Q_{a+b}, \mathcal{P}_{m-L}, S_{m-L,nf})$ respectively, where $\mathcal{P}_j$ denotes the class of $j$ independent Poisson distributions and $S_{j,k}$ is the set of $j$ independent Poisson counts with sum equal to $k$. □

**Remark 3**: 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 one for a large alphabet with $m-L$ symbols with total count $nf$. The two main terms in expression (6) represent redundancy from coding the two subsets of symbols, with one set containing $L$ symbols having relatively large probabilities, and each symbol induces $\frac{1}{2}\log\frac{n(1-f)e}{L}$ bits redundancy, and the other containing the rest $m-L$ symbols with small probabilities and together requires $nf\log\frac{m}{nfe}$ extra bits.

### 3.3. Envelope class

Here we follow the definition of envelope class in [11], suppose $\mathcal{P}_{m,f}$ is a class of distributions on $1, ..., m$ with the symbol probability bounded above by an envelope function $f$, i.e.

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

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

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

**Theorem 3.** *The pointwise minimax redundancy of the Poisson class $\Lambda_{m,f}$ with envelope function $f$ has the following upper bound*

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

*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.* using a tilted distribution with $a = L/2n(1-\bar{F}(L))$ will give the result. □

**Remark 5**: Here in order for $r_3$ to be small, the tail sum $\bar{F}(L)$ of the envelope function 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 $\inf_{L:L\leq n}\left((L-1)/2\log n + n\bar{F}(L)\log e\right)+2$ given in [11].

**Remark 6**: The best choice of tilting parameters for envelope class only dependents on the envelope function and the number of symbols $L$ constituting the 'frequent' subset. unlike the subset of sequences case discussed before, neither of the order of counts or which symbols are those with largest counts matter, all we need is an envelope function decays fast enough when the symbols are arranged in decreasing order so that $L$ and $\bar{F}(L)$ is small compared to $n(1-\bar{F}(L))$.

### 3.4. Arbitrary frequent subset

Section 3.2 and 3.3 discuss using independent tilted distributions to code skewed empirical distributions and envelope classes, both of which assume a structure in which the symbols are partitioned in a way according to its counts or probabilities of occurrence. This assumption is restrictive unless prior knowledge of an 'important' subset of symbols exists. An easy fix is a mixture of the product tilted distributions across all subsets containing $L$ symbols, for example, for a set $S^{arb}_{m,n,f,L}$ with any subset of $m-L$ symbols having a total count equal to $nf$,

$$\{\underline{N}=(N_1,...,N_m):\sum_{j=1}^{m}N_j=n,\sum_{j>L}N_{(j)}=nf\}.$$

Here $N_{(j)}$ is the $j$th largest count in $\underline{N}$. It is easy to see $S^{arb}_{m,n,f,L}=\cup_k S^k_{m,n,f,L}$ where each $S^k_{m,n,f,L}$ is a permutations of $S_{m,n,f,L}$.

For each given $S^k_{m,n,f,L}$, an optimal tilted distribution $Q_{S^k_{m,n,f,L}}$ can be used by choosing tilting parameters according to the symbol and total count ratios. So given $L$ and $f$, a mixture of tilted distributions is

$$Q^*_L(\underline{N}) = \frac{1}{\binom{m}{L}}\sum_j Q_{S^k_{m,n,f,L}}$$

The pointwise redundancy of using the mixture distribution $Q^*_L$ for any string in $S^{arb}_{m,n,f,L}$ is no larger than $\log\binom{m}{L}$ above the minimum redundancy if the 'important' subset were known, i.e., for each $\underline{N}\in S^{arb}_{m,n,f,L}$

$$\begin{aligned}&R(Q^*_L,\mathcal{P}^{Poi}_m,\underline{N})\\ \leq\ &\min_{k:\underline{N}\in S^k_{m,n,f,L}} R(Q_{S^k_{m,n,f,L}},\mathcal{P}^{Poi}_m,S^k_{m,n,f,L})\\ &+\log\binom{m}{L}\end{aligned}$$

This extra term is still acceptable as long as $\log\binom{m}{L}$ is small compared to $n$. And in fact this term also exists in the pointwise minimax redundancy, because any $\underline{N}$ in $S^{arb}_{m,n,f,L}$ is a permuted version of an $\underline{N}$ in $S_{m,n,f,L}$, with the same ordered statistic. Remember that the log of the Shtarkov's sum for $m$ independent Poisson random variables with counts $\underline{N}$ in $S^{arb}_{m,n,f,L}$ is

$$C(S^{arb}_{m,n,f,L})=\log\binom{m}{L}+C(S_{m,n,f,L})$$

Therefore the simple mixture of tilted distributions does not add any extra redundancy.

### 3.5. Pointwise redundancy with total count unknown

When the total count is not known, we can use a mixture of tilted distributions $Q(\underline{N})$ to code the strings, where

$$\begin{aligned}Q(\underline{N}) &= \int_0^{m/2} P_a(\underline{N})\frac{2}{m}da\\ &= M(\underline{N})\frac{2}{m}\int_0^{m/2} e^{-aN}C_a^{-m}da.\end{aligned}$$

For any $k\in\mathbf{N}$, the integrand is maximized at $a^*_N$, which is a solution to $\mathbf{E}_{P_a}N=k$. And the integral can be approximated by the Laplace method,

$$Q(\underline{N}) \approx M(\underline{N})\frac{2}{m}e^{-a^*_N N}C^{-m}_{a^*_N}\sqrt{\frac{2\pi}{c}},$$

where $c=-\frac{\partial^2}{\partial a}\ln\left(e^{-Na}C_a^{-m}\right)|_{a=a^*_N}$.

Hence the pointwise redundancy of $Q(\underline{N})$ is

$$\begin{aligned}&\log\frac{M(\underline{N})}{Q(\underline{N})}\\ \approx\ &\log e^{a^*_N N}C^m_{a^*_N}\sqrt{\frac{c}{2\pi}}+\log\frac{m}{2}\\ \leq\ &\log e^{a^*_N N}C^m_{a^*_N}+\frac{3}{2}\log\frac{m}{(2\pi)^{1/3}}+\log C_{a^*_N}\end{aligned}$$

The redundancy of $Q(\underline{N})$ above the optimal level is approximately bounded by $\frac{3}{2}\log\frac{m}{(2\pi)^{1/3}}+\log C_{a^*_N}$.

## 4. APPLICATION

Here we give an examples of using tilted distributions in Section 3.2 to code Chinese literature. It is the existing earliest collection of Chinese poetry dating from the 10th to 7th centuries BC [14] translated as the Classic of Poetry. The book is downloaded freely from `http://wenku.baidu.com/`. Since many ancient words are rarely used now, the encoding is in GB18030 [15], 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 alphabet is partitioned into two subsets – the frequent ones and the infrequent ones. The tilting parameter is chosen approximately according to the ratio of the number of symbols in a group and their total counts. The redundancy of assigning different number of symbols as 'frequent' ($L$) is shown in Figure 2. The smallest redundancy happens at $L=2889$ which is the total number of characters that appear.

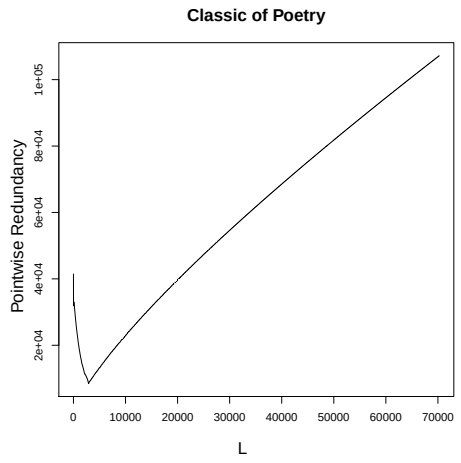


Figure 2. Pointwise redundancy of $Q_{a,b}$ for $L$ from 1 to $m$.

## 5. DISCUSSION

We have considered using independent tilted distributions as a coding or prediction strategy for independent random variables generated from a memoryless source with a sample size also being random. The performance of the strategy is close to the minimax level as shown above. Actually, the difference between the redundancy induced by using independent tilted distributions and the minimax redundancy is the probability assigned to the set with the observed condition by the tilted distribution with the parameter given by the moment condition, i.e.

$$R(Q_{a^*}, \mathcal{P}_m, S_{m,n}) = C(S_{m,n}) + \log 1/Q_{a^*}(S_{m,n}).$$

The choice of $a = a^*$ minimizes the difference $\log Q_{a^*}(S_{m,n})$ among all possible choices of $a$. Our initial finding is this term decreases with $n$ even if the tilting parameter is adjusted for $n$, however, we do not have a precise evaluation yet. Further exploration could be done to characterize this term and understand the relationship between the tilted distribution and the exact minimax distribution.

## 6. REFERENCES


[1] Y. M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmissions*, vol. 23, pp. 3–17, July 1988.

[2] Q. Xie and A. R. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Transactions on Information Theory*, vol. 43, pp. 646–657, May 1997.

[3] I. Csiszar, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, vol. 3, no. 1, pp. 146–158, Feb 1975.

[4] ——, "Sanov property, generalized I-projection and a conditional limit theorem," *The Annals of Probability*, vol. 12, no. 3, pp. 768–793, Jan 1984.

[5] J. V. Campenhout and T. Cover, "Maximum entropy and conditional probability," *IEEE Transactions on Information Theory*, vol. 27, no. 4, July 1981.

[6] A. Orlistsky, N. P. Santhanam, and J. Zhang, "Always good turing: Asymptotically optimal probability estimation," *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Sciene*, 2003.

[7] W. Szpankowski and M. J. Weinberger, "Minimax redundancy for large alphabets," *Information Theory Proceedings*, June 2010.

[8] Y. M. Shtarkov, T. J. Tjalkens, and F. M. J. Willems, "Multi-alphabet universal coding of memoryless sources," *Problems of Information Transmissions*, vol. 31, no. 2, pp. 114–127, 1995.

[9] Q. Xie and A. R. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 431–445, March 2000.

[10] A. Orlistsky and N. P. Santhanam, "Speaking of infinity," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2215–2230, October 2004.

[11] A. G. S. Boucheron and E. Gassiat, "Coding on countably infinite alphabets," *IEEE Transactions on Information Theory*, vol. 55, no. 1, Jan 2009.

[12] D. Bontemps, "Universal coding on infinite alphabets: exponentially decreasing envelopes," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1466–1478, 2011.

[13] S. Kullback, *Information Theory and Statistics*. Wiley, New York, 1959.

[14] [Online]. Available: https://en.wikipedia.org/wiki/Classic_of_Poetry

[15] [Online]. Available: http://zh.wikipedia.org/wiki/GB_18030