Universal Compression of Power-Law Distributions

Moein Falahatgar
moein@ucsd.edu

Ashkan Jafarpour
ashkan@ucsd.edu

Alon Orlitsky
alon@ucsd.edu

Venkatadheeraj Pichapati
dheerajpv7@gmail.com

Ananda Theertha Suresh
asuresh@ucsd.edu

University of California, San Diego

May 5, 2015

Abstract

English words and the outputs of many other natural processes are well-known to follow a Zipf distribution. Yet this thoroughly-established property has never been shown to help compress or predict these important processes. We show that the expected redundancy of Zipf distributions of order $\alpha > 1$ is roughly the $1/\alpha$ power of the expected redundancy of unrestricted distributions. Hence for these orders, Zipf distributions can be better compressed and predicted than was previously known. Unlike the expected case, we show that worst-case redundancy is roughly the same for Zipf and for unrestricted distributions. Hence Zipf distributions have significantly different worst-case and expected redundancies, making them the first natural distribution class shown to have such a difference.

Keywords: Power-law, Zipf, Universal Compression, Distinct elements, Redundancy

1 Introduction

1.1 Definitions

The fundamental data-compression theorem states that every discrete distribution p can be compressed to its entropy $H(p) \stackrel{\text{def}}{=} \sum p(x) \log \frac{1}{p(x)}$, a compression rate approachable by assigning each symbol x a codeword of roughly $\log \frac{1}{p(x)}$ bits.

In reality, the underlying distribution is seldom known. For example, in text compression, we observe only the words, no one tells us their probabilities. In all these cases, it is not clear how to compress the distributions to their entropy.

The common approach to these cases is *universal compression*. It assumes that while the underlying distribution is unknown, it belongs to a known class of possible distributions, for example, *i.i.d.* or Markov distributions. Its goal is to derive an encoding that works well for all distributions in the class.

To move towards formalizing this notion, observe that every compression scheme for a distribution over a discrete set $\mathcal{X}$ corresponds to some distribution q over $\mathcal{X}$ where each symbol $x \in \mathcal{X}$ is

assigned a codeword of length $\log \frac{1}{q(x)}$. Hence the expected number of bits used to encode the distribution's output is $\sum p(x) \log \frac{1}{q(x)}$, and the additional number of bits over the entropy minimum is $\sum p(x) \log \frac{p(x)}{q(x)}$.

Let $\mathcal{P}$ be a collection of distributions over $\mathcal{X}$. The collection's *expected redundancy*, is the least worst-case increase in the expected number of bits over the entropy, where the worst case is taken over all distributions in $\mathcal{P}$ and the least is minimized over all possible encoders,

$$\bar{R}(\mathcal{P}) \stackrel{\text{def}}{=} \min_{q} \max_{p \in \mathcal{P}} \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

An even stricter measure of the increased encoding length due to not knowing the distribution is the collection's *worst-case redundancy* that considers the worst increase not just over all distributions, but also over all possible outcomes x,

$$\hat{R}(\mathcal{P}) \stackrel{\text{def}}{=} \min_{q} \max_{p \in \mathcal{P}} \max_{x \in \mathcal{X}} \log \frac{p(x)}{q(x)}.$$

Clearly,

$$\bar{R}(\mathcal{P}) \le \hat{R}(\mathcal{P}).$$

Interestingly, until now, except for some made-up examples, all analyzed collections had extremely close expected and worst-case redundancies. One of our contributions is to demonstrate a practical collection where these redundancies vastly differ, hence achieving different optimization goals may require different encoding schemes.

By far the most widely studied are the collections of *i.i.d.* distributions. For every distribution p, the *i.i.d.* distribution p^n assigns to a length-n string $x^n \stackrel{\text{def}}{=} (x_1, x_2, \ldots, x_n)$ probability $p(x^n) = p(x_1) \cdot \ldots \cdot p(x_n)$. For any collection $\mathcal{P}$ of distributions, the length-n *i.i.d.* collection is

$$\mathcal{P}^n \stackrel{\text{def}}{=} \{p^n : p \in \mathcal{P}\}.$$

1.2 Previous results

Let Δ_k denote the collection of all distribution over $\{1, \ldots, k\}$, where Δ was chosen to represent the simplex. For the first few decades of universal compression, researchers studied the redundancy of Δ_k^n when the alphabet size k is fixed and the block length n tends to infinity. A sequence of papers Krichevsky and Trofimov [1981], Kieffer [1978], Davisson [1973], Davisson et al. [1981], Willems et al. [1995], Xie and Barron [2000], Szpankowski and Weinberger [2010], Orlitsky and Santhanam [2004], Rissanen [1996], Cover [1991], Szpankowski [1998], Szpankowski and Weinberger [2012] showed that

$$\hat{R}(\Delta_k^n) = \frac{k-1}{2} \log \frac{n}{2\pi} + \log \frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2})} + o_k(1),$$

and that the expected redundancy is extremely close, at most $\log e$ bits lower. Note that a similar result holds for the complementary regime where n is fixed and k tends to infinity,

$$\hat{R}(\Delta_k^n) = n \log \frac{k}{n} + o(n).$$

These positive results show that redundancy grows only logarithmically with the sequence length n, therefore for long sequences, the per-symbol redundancy diminishes to zero and the underlying distribution needs not to be known to approach entropy. As is also well known, expected redundancy is exactly the same as the log loss of sequential prediction, hence these results also show that prediction can be performed with very small log loss.

However, as intuition suggests, and these equations confirm, redundancy increases sharply with the alphabet size k. In many, and possibly most, important real-life applications, the alphabet size is very large, often even larger than the block length. This is the case for example in applications involving natural language processing, population estimation, and genetics Chen and Goodman [1996]. The redundancy in these cases is therefore very large, and can be even unbounded for any sequence length n.

Over the last decade, researchers therefore considered methods that could cope with compression and prediction of distributions over large alphabets. Two main approaches were taken.

Orlitsky et al. [2004] separated compression (and similarly prediction) of large-alphabet sequences into compression of their *pattern* that indicates the order at which symbols appeared, and *dictionary* that maps the order to the symbols. For example, the pattern of "*banana*" is 123232 and its dictionary is $1 \to b$, $2 \to a$, and $3 \to n$. Letting Δ_ψ^n denote the collection of all pattern distributions, induced on sequences of length n by all *i.i.d.* distributions over any alphabet, a sequence of papers Orlitsky et al. [2004], Shamir [2006, 2004], Garivier [2009], Orlitsky and Santhanam [2004], Acharya et al. [2012, 2013] showed that although patterns carry essentially all the entropy, they can be compressed with redundancy

$$0.3 \cdot n^{1/3} \le \bar{R}(\Delta_\psi^n) \le \hat{R}(\Delta_\psi^n) \le n^{1/3} \cdot \log^4 n$$

as $n \to \infty$. Namely, pattern redundancy too is sublinear in the block length and most significantly, is uniformly upper bounded regardless of the alphabet size (which can be even infinite). It follows the per-symbol pattern redundancy and prediction loss both diminish to zero at a uniformly-bounded rate, regardless of the alphabet size. Note also, that for pattern redundancy, worst-case and expected redundancy are quite close.

However, while for many prediction applications predicting the pattern suffices, for compression one typically needs to know the dictionary as well. These results show that essentially all the redundancy lies in the dictionary compression.

The second approach restricted the class of distributions compressed. A series of works studied class of *monotone* distributions Shamir [2013], Acharya et al. [2014a]. Recently, Acharya et al. [2014a] showed that the class M_k of monotone distributions over $\{1, \ldots, k\}$ has redundancy $\hat{R}(M_k^n) \le \sqrt{20n \log k \log n}$.

More closely related to this paper are *envelope classes*. An *envelope* is a function $f : \mathbb{N}_+ \to \mathbb{R}_{\ge 0}$. For envelope function f,

$$\mathcal{E}_f \stackrel{\text{def}}{=} \{p : p_i \le f(i) \text{ for all } i \ge 1\}$$

is the collection of distributions where each p_i is at most the corresponding envelope bound $f(i)$. Some canonical examples are the power-law envelopes $f(i) = c \cdot i^{-\alpha}$, and the exponential envelopes $f(i) = c \cdot e^{-\alpha \cdot i}$. In particular, for power-law envelopes Boucheron et al. [2009, 2014] showed

$$\hat{R}(\mathcal{E}_f) \le \left(\frac{2cn}{\alpha - 1}\right)^{\frac{1}{\alpha}} (\log n)^{1 - \frac{1}{\alpha}} + \mathcal{O}(1),$$

and more recently, Acharya et al. [2014b] showed that

$$\hat{R}(\mathcal{E}_f) = \Theta(n^{1/\alpha}).$$

The restricted-distribution approach has the advantage that it considers the complete sequence redundancy, not just the pattern. Yet it has the shortcoming that it may not capture relevant distribution collections. For example, most real distributions are not monotone, words starting with 'a' are not necessarily more likely than those starting with 'b'. Similarly for say power-law envelopes, why should words in the early alphabet have higher upper bound than subsequent ones? Thus, words do not carry frequency order inherently.

1.3 Distribution model

In this paper we combine the advantages and avoid the shortfalls of both approaches to compress and predict distributions over large alphabets. As in patterns, we consider useful distribution collections, and like restricted-distributions, we address the full redundancy.

Envelope distributions are very appealing as they effectively represent our belief about the distribution. However their main drawback is that they assume that the correspondence between the probabilities and symbols is known, namely that $p_i \leq f(i)$ for the same i. We relax this requirement and assume only that an upper envelope on the sorted distribution, not the individual elements, is known. Such assumptions on the sorted distributions are believed to hold for a wide range of common distributions.

In 1935, linguist George Kingsley Zipf observed that when English words are sorted according to their probabilities, namely so that $p_1 \geq p_2 \geq \ldots$, the resulting distribution follows a power law, $p_i \sim \frac{c}{i^\alpha}$ for some constant c and power α. Long before Zipf, Pareto [1896] studied distributions in income ranking and showed it can be mathematically expressed as power-law. Since then, researchers have found a very large number of distributions such as word frequency, population ranks of cities, corporation sizes, and website users that when sorted follow this *Zipf*-, or *power*-law Zipf [1932, 1949], Adamic and Huberman [2002]. In fact, a Google Scholar search for "power-law distribution" returns around 50,000 citations.

A natural question therefore is whether the established and commonly trusted empirical observation that real distributions obey Zipf's law can be used to better predict or equivalently compress them, and if so, by how much.

In Section 2 we state our notation followed by new results in Section 3. Next, in Section 4 we bound the worst-case redundancy for power-law envelop class. In Section 5 we take a novel approach to analyze the expected redundancy. We introduce a new class of distributions which has the property that all permutations of a distribution are present in the class. Then we upper and lower bound the expected redundancy of this class based on the expected number of distinct elements. Finally, in Section 6 we show that the redundancy of power-law envelop class can be studied in this framework.

2 Preliminaries

2.1 Notation

Let $x^n \stackrel{\text{def}}{=} (x_1, x_2, .., x_n)$ denote a sequence of length n, $\mathcal{X}$ be the underlying alphabet and $k \stackrel{\text{def}}{=} |\mathcal{X}|$. The *multiplicity* μ_x of a symbol $x \in \mathcal{X}$ is the number of times x appears in x^n. Let $[k] =$

$\{1, 2, ..., k\}$ be the indices of elements in $\mathcal{X}$. The type vector of x^n over $[k] = \{1, 2, ..., k\}$, $\tau(x^n) = (\mu_1, \mu_2, \ldots, \mu_k)$ is a k-tuple of multiplicities in x^n. The *prevalence* of a multiplicity μ, denoted by φ_μ, is the number of elements appearing μ times in x^n. For example, φ_1 denotes the number of elements which appeared once in x^n. Furthermore, φ_+ denotes the number of distinct elements in x^n. The vector of prevalences for all μ's is called the profile vector.

We use $p_{(i)}$ to note the i^{th} highest probability in p. Hence, $p_{(1)} \geq p_{(2)} \geq \ldots p_{(k)}$. Moreover, we use $\text{zipf}(\alpha, k)$ to denote Zipf distribution with parameter α and support k. Hence,

$$\text{zipf}(\alpha, k)_i = \frac{i^{-\alpha}}{C_{k,\alpha}},$$

where $C_{k,\alpha}$ is the normalization factor. Note that all logarithms in this paper are in base 2 and we consider only the case $\alpha > 1$.

2.2 Problem statement

For an envelope f with support size k, let $\mathcal{E}_{(f)}$ be the class of distributions such that

$$\mathcal{E}_{(f)} = \{p : \ p_{(i)} \leq f(i) \ \forall 1 \leq i \leq k\}.$$

Note that $\mathcal{E}_f \subset \mathcal{E}_{(f)}$. We also consider the special case when f is a distribution itself, in which case we denote $\mathcal{E}_{(f)}$ by $\mathcal{P}_{(p)}$, a class that has distributions whose multi-set of probabilities is same as p. In other words, $\mathcal{P}_{(p)}$ contains all permutations of distribution p. Also we define

$$\mathcal{P}_d^n = \{p^n : \mathbb{E}_p[\varphi_+^n] \leq d\},$$

where φ_+^n is the number of distinct elements in x^n. Note that for any distribution belonging to this class, all permutations of it are also in the class.

3 Results

We first consider worst-case redundancy, lower-bound it for general unordered permutations, and apply the result to unordered power-law classes, showing that for $n \leq k^{1/\alpha}$,

$$\hat{R}(\mathcal{E}^n_{(ci^{-\alpha},k)}) \geq \hat{R}(\mathcal{P}^n_{(\text{zipf}(\alpha,k))}) \geq n \log \frac{k-n}{n^\alpha C_{k,\alpha}}.$$

This shows that the worst-case redundancy of power-law distributions behaves roughly as that of general distributions over the same alphabet.

More interestingly, we establish a general method for upper- and lower-bounding the expected redundancy of unordered envelope distributions in terms of expected number of distinct symbols. Precisely, for a class $\mathcal{P}_d^n$ we show the following upper bound

$$\bar{R}(\mathcal{P}_d^n) \leq d \log \frac{kn}{d^2} + (2 \log e + 1)d + \log(n+1).$$

Interpretation: This upper bound can be also written as

$$\log n + \log \binom{k}{d} + \log \binom{n-1}{d-1}. \tag{1}$$

This suggests a very clear intuition of the upper bound. We can give a compression scheme for any sequence that we observe. Upon observing a sequence x^n, first we declare how many distinct elements are in that sequence. For this we need $\log n$ bits. In addition to those bits, we need $\log \binom{k}{d}$ bits to specify which d distinct elements out of k elements appeared in the sequence. Finally, for the exact number of occurrences of each distinct element we should use $\log \binom{n-1}{d-1}$ bits.

We also show a lower bound which is dependent on both the expected number of distinct elements d and the distributions in the class $\mathcal{P}_d^n$. Namely, we show

$$\bar{R}(\mathcal{P}_d^n) \geq \left(\log \binom{k}{d} - d \log \frac{n}{d} - d \log \pi e\right)(1 + o_d(1)) - \sum_{np_i < 0.7} (3np_i - np_i \log np_i).$$

Using this result, we then consider expected redundancy of power-law distributions as a special case of $\mathcal{P}_d^n$ and show that it is significantly lower than that of general distributions. This shows that on average, Zipf distributions can be compressed much better than general ones. Since expected redundancy is the same as log loss, they can also be predicted more effectively. In fact we show that for $k > n$,

$$\bar{R}(\mathcal{E}^n_{(ci^{-\alpha},k)}) = \Theta(n^{\frac{1}{\alpha}} \log k).$$

Recall that general length-n *i.i.d.* distributions over alphabet of size k have redundancy roughly $n \log \frac{k}{n}$ bits. Hence, when k is not much larger than n, the expected redundancy of Zipf distributions of order $\alpha > 1$ is the $1/\alpha$ power of the expected redundancy of general distributions. For example, for $\alpha = 2$ and $k = n$, the redundancy of Zipf distributions is $\Theta(\sqrt{n} \log n)$ compared to n for general distributions. This reduction from linear to sub-linear dependence on n also implies that unordered power-law envelopes are universally compressible when $k = n$.

These results also show that worst-case redundancy is roughly the same for Zipf and general distributions. Comparing the results for worst-case and expected redundancy of Zipf distributions, it also follows that for those distributions expected- and worst-case redundancy differ greatly. This is the first natural class of distribution for which worst-case and expected redundancy have been shown to significantly diverge.

As stated in the introduction, for the power-law envelope f, Acharya et al. [2014b] showed that

$$\hat{R}(\mathcal{E}_f) = \Theta(n^{1/\alpha}).$$

Comparing this with the results in this paper reveals that if we know the envelop on the class of distributions but we do not know the true order of that, we have an extra multiplicative factor of $\log k$ in the expected redundancy, i.e.

$$\bar{R}(\mathcal{E}_{(f)}) = \Theta(n^{1/\alpha} \log k).$$

4 Worst-case redundancy

4.1 Shtarkov Sum

It is well known that the worst-case redundancy can be calculated using Shtarkov sum Shtarkov [1987], i.e. for any class $\mathcal{P}$

$$\hat{R}(\mathcal{P}) = \log S(\mathcal{P}), \tag{2}$$

where $S(\mathcal{P})$ is the Shtarkov sum and defined as

$$S(\mathcal{P}) \stackrel{\text{def}}{=} \sum_{x \in \mathcal{X}} \hat{p}(x). \tag{3}$$

For notational convenience we denote $\hat{p}(x) \stackrel{\text{def}}{=} \max_{p \in \mathcal{P}} p(x)$, to be the maximum probability any distribution in $\mathcal{P}$ assigns to x.

4.2 Small alphabet case

Recall that $\hat{R}(\Delta_k^n) \approx \frac{k-1}{2} \log n$. We now give a simple example to show that unordered distribution classes $\mathcal{P}_{(p)}$ may have much smaller redundancy. In particular we show that for a distribution p over k symbols,

$$\hat{R}(\mathcal{P}_{(p)}^n) \le \log k! \le k \log k \quad \forall n.$$

Consider the Shtarkov sum

$$\begin{aligned} S(\mathcal{P}_{(p)}^n) &= \sum_{x^n \in \mathcal{X}^n} \hat{p}(x^n) \\ &\le \sum_{x^n \in \mathcal{X}^n} \sum_{p \in \mathcal{P}_{(p)}} p(x^n) \\ &= \sum_{p \in \mathcal{P}_{(p)}} \sum_{x^n \in \mathcal{X}^n} p(x^n) \\ &= \sum_{p \in \mathcal{P}_{(p)}} 1 = |\mathcal{P}_{(p)}| = k!. \end{aligned}$$

Clearly for $n \gg k$, the above bound is smaller than $\hat{R}(\Delta_k^n)$.

4.3 Large alphabet regime

From the above result, it is clear that as $n \to \infty$, the knowledge of the underlying-distribution multi-set helps in universal compression. A natural question is to ask if the same applies for the large alphabet regime when the number of samples $n \ll k$. Recall that Acharya et al. [2014b], Boucheron et al. [2009] showed that for power-law envelopes, $f(i) = c \cdot i^{-\alpha}$, with infinite support size

$$\hat{R}(\mathcal{E}_f) = \Theta(n^{\frac{1}{\alpha}}).$$

We show that if the permutation of the distribution is not known then the worst-case redundancy is $\Omega(n) \gg \Theta(n^{\frac{1}{\alpha}})$, and thus the knowledge of the permutation is essential. In particular, we prove that even for the case when the envelope class consists of only one power-law distribution, $\hat{R}$ scales as n.

Theorem 1. *For $n \le k^{1/\alpha}$,*

$$\hat{R}(\mathcal{E}_{(ci^{-\alpha},k)}^n) \ge \hat{R}(\mathcal{P}_{(zipf(\alpha,k))}^n) \ge n \log \frac{k-n}{n^\alpha C_{k,\alpha}}.$$

Proof. Since $\mathcal{P}^n_{(\text{zipf}(\alpha,k))} \subset \mathcal{E}^n_{(ci^{-\alpha},k)}$, we have

$$\hat{R}(\mathcal{E}^n_{(ci^{-\alpha},k)}) \geq \hat{R}(\mathcal{P}^n_{(\text{zipf}(\alpha,k))}).$$

To lower bound $\hat{R}(\mathcal{P}^n_{(\text{zipf}(\alpha,k))})$, recall that

$$\begin{aligned} S(\mathcal{P}^n_{(\text{zipf}(\alpha,k))}) &= \sum_{x^n} \hat{p}(x^n) \\ &\geq \sum_{x^n:\varphi^n_+=n} \hat{p}(x^n), \end{aligned}$$

where φ^n_+ is the number of distinct symbols in x^n. Note that number of such sequences is $k(k-1)(k-2)\dots(k-n+1)$. We lower bound $\hat{p}(x^n)$ for every such sequence. Consider the distribution $q \in \mathcal{P}_{(\text{zipf}(\alpha,k))}$ given by $q(x_i) = \frac{1}{i^\alpha C_{k,\alpha}}$ $\forall 1 \leq i \leq n$. Clearly $\hat{p}(x^n) \geq q(x^n)$ and as a result we have

$$\begin{aligned} S(\mathcal{P}^n_{(\text{zipf}(\alpha,k))}) &\geq k(k-1)(k-2)\dots(k-n+1) \prod_{i=1}^{n} \frac{1}{i^\alpha C_{k,\alpha}} \\ &\geq \left(\frac{k-n}{n^\alpha C_{k,\alpha}}\right)^n. \end{aligned}$$

Taking the logarithm yields the result. ■

Thus for small values of n, independent of the underlying distribution per-symbol redundancy is $\log \frac{k}{n^\alpha}$. Since for $n \leq k$, $\hat{R}(\Delta^n_k) \approx n \log \frac{k}{n}$, we have for $n \leq k^{1/\alpha}$

$$\hat{R}(\mathcal{E}^n_{(ci^{-\alpha},k)}) \leq \hat{R}(\Delta^n_k) \leq \mathcal{O}(n \log \frac{k}{n}).$$

Therefore, together with Theorem 1, we have for $n \leq k^{1/\alpha}$

$$\Omega(n \log \frac{k}{n^\alpha}) \leq \hat{R}(\mathcal{E}^n_{(ci^{-\alpha},k)}) \leq \mathcal{O}(n \log \frac{k}{n}).$$

5 Expected redundancy based on the number of distinct elements

In order to find the redundancy of the unordered envelop classes, we follow a more systematic approach and define another structure on the underlying class of distributions. More precisely, we consider the class of all distributions in which we have an upper bound on the expected number of distinct elements we are going to observe. Lets define

$$\mathcal{P}^n_d = \{p^n : \mathbb{E}_p[\varphi^n_+] \leq d\},$$

where φ^n_+ is the number of distinct symbols in the sequence x^n. Note that for any distribution belonging to this class, all permutations of it are also in the class. We later show that envelop classes can be described in this way and the expected number of distinct elements characterizes the envelop classes; therefore we can bound the redundancy of them applying results in this section.

5.1 Upper bound

The following lemma bounds the expected redundancy of a class in terms of d.

Lemma 2. *For any class $\mathcal{P}_d^n$,*

$$\bar{R}(\mathcal{P}_d^n) \leq d \log \frac{kn}{d^2} + (2 \log e + 1)d + \log(n+1).$$

Proof. We give an explicit coding scheme that achieves the above redundancy. For a sequence x^n with multiplicities of symbols $\mu^k \stackrel{\text{def}}{=} \mu_1, \mu_2, \ldots, \mu_k$, let

$$q(x^n) = \frac{1}{N_{\varphi_+^n}} \cdot \prod_{j=1}^{k} \left(\frac{\mu_j}{n}\right)^{\mu_j}$$

be the probability our compression scheme assigns to x^n and $N_{\varphi_+^n}$ is the normalization factor given by

$$N_{\varphi_+^n} = n \cdot \binom{k}{\varphi_+^n} \cdot \binom{n-1}{\varphi_+^n - 1}.$$

Before proceeding, we show that q is a valid coding scheme by showing that $\sum_{x^n \in \mathcal{X}^n} q(x^n) \leq 1$. We divide the set of sequences as follows.

$$\sum_{x^n \in \mathcal{X}^n} = \sum_{d'=1}^{n} \sum_{S \in \mathcal{X}: |S| = d'} \sum_{\mu^k : \mu_i = 0 \text{ iff } i \notin S} \sum_{x^n : \mu(x^n) = \mu^k}$$

Now we can re-write and bound $\sum_{x^n \in \mathcal{X}^n} q(x^n)$ as the following.

$$\begin{aligned}
\sum_{d'=1}^{n} \sum_{S \in \mathcal{X}: |S| = d'} \sum_{\mu^k : \mu_i = 0 \text{ iff } i \notin S} \sum_{x^n : \mu(x^n) = \mu^k} q(x^n) &\stackrel{(a)}{\leq} \sum_{d'=1}^{n} \sum_{S \in \mathcal{X}: |S| = d'} \sum_{\mu^k : \mu_i = 0 \text{ iff } i \notin S} \frac{1}{N_{d'}} \\
&\stackrel{(b)}{=} \sum_{d'=1}^{n} \frac{\binom{k}{d'} \cdot \binom{n-1}{d'-1}}{N_{d'}} \\
&= \sum_{d'=1}^{n} \frac{1}{n} = 1.
\end{aligned}$$

where (a) holds since for a given μ^k , the maximum likelihood distribution for all sequences with same values of $\mu_1, \mu_2, \ldots \mu_k$ are same. Also (b) follows from the fact that the second summation ranges over $\binom{k}{d'}$ values and the third summation ranges over $\binom{n-1}{d'-1}$ values. Furthermore for any $p^n \in \mathcal{P}_d^n$,

$$\log \frac{p(x^n)}{q(x^n)} \leq \log N_{\varphi_+^n} + n \cdot \sum_{i=1}^{k} \frac{\mu_i}{n} \log \frac{p_i}{\mu_i / n} \leq \log N_{\varphi_+^n}.$$

Taking expectation over both sides

$$\begin{aligned}
\bar{R}(\mathcal{E}_{(f)}) &\le \mathbb{E}[\log N_{\varphi_+^n}] \\
&\le \log n + \mathbb{E}\left[\log \binom{k}{\varphi_+^n} + \log \binom{n-1}{\varphi_+^n - 1}\right] \\
&\overset{(a)}{\le} \log n + \mathbb{E}\left[\varphi_+^n \log\left(\frac{k}{\varphi_+^n} \cdot \frac{2n}{\varphi_+^n}\right) + (2\log e)\varphi_+^n\right] \\
&\overset{(b)}{\le} \log n + d\log\frac{kn}{d^2} + (2\log e + 1)d,
\end{aligned}$$

where (a) follows from the fact that $\binom{n}{d} \le \left(\frac{ne}{d}\right)^d$ and (b) follows from Jensen's inequality. ■

5.2 Lower bound

To show a lower bound on the expected redundancy of class $\mathcal{P}_d^n$, we use some helpful results introduced in previous works. First, we introduce Poisson sampling and relate the expected redundancy in two cases when we use normal sampling and Poisson sampling. Then we prove the equivalence of expected redundancy of the sequences and expected redundancy of types.

Poisson sampling: In the standard sampling method, where a distribution is sampled n times, the multiplicities are dependent, for example they add up to n. Hence, calculating redundancy under this sampling often requires various concentration inequalities, complicating the proofs. A useful approach to make them independent and hence simplify the analysis is to sample the distribution n' times, where n' is a Poisson random variable with mean n. Often called as Poisson sampling, this approach has been used in universal compression to simplify the analysis Acharya et al. [2012, 2014b], Yang and Barron [2013], Acharya et al. [2013].

Under Poisson sampling, if a distribution p is sampled *i.i.d.* $\text{poi}(n)$ times, then the number of times symbol x appears is an independent Poisson random variable with mean np_x, namely, $\Pr(\mu_x = \mu) = \frac{e^{-np_x}(np_x)^\mu}{\mu!}$ Mitzenmacher and Upfal [2005]. Henceforth, to distinguish between two cases of normal sampling and Poisson sampling we specify it with superscripts n for normal sampling and $poi(n)$ for Poisson sampling.

Next lemma lower bounds $\bar{R}(\mathcal{P}^n)$ by the redundancy in the presence of Poisson sampling. We use this lemma further in our lower-bound arguments.

Lemma 3. *For any class $\mathcal{P}$,*

$$\bar{R}(\mathcal{P}^n) \ge \frac{1}{2}\bar{R}(\mathcal{P}^{\text{poi}(n)}).$$

Proof. By the definition of $\bar{R}(\mathcal{P}^{\text{poi}(n)})$,

$$\bar{R}(\mathcal{P}^{\text{poi}(n)}) = \min_q \max_{p\in\mathcal{P}} \mathbb{E}_{\text{poi}(n)}\left[\log \frac{p_{\text{poi}(n)}(x^{n'})}{q(x^{n'})}\right], \tag{4}$$

where subscript $\text{poi}(n)$ indicates that the probabilities are calculated under Poisson sampling. Similarly, for every n',

$$\bar{R}(\mathcal{P}^{n'}) = \min_q \max_{p\in\mathcal{P}} \mathbb{E}\left[\log \frac{p(x^{n'})}{q(x^{n'})}\right].$$

Let $q_{n'}$ denote the distribution that achieves the above minimum. We upper bound the right hand side of Equation (4) by constructing an explicit q. Let

$$q(x^{n'}) = e^{-n}\frac{n^{n'}}{n'!}q_{n'}(x^{n'}).$$

Clearly q is a distribution as it adds up to 1. Furthermore, since $p_{\mathrm{poi}(n)}(x^{n'}) = e^{-n}\frac{n^{n'}}{n'!}p(x^{n'})$, we get

$$\begin{aligned}
\bar{R}(\mathcal{P}^{\mathrm{poi}(n)}) &\le \max_{p\in\mathcal{P}} \mathbb{E}_{\mathrm{poi}(n)}\left[\log\frac{p_{\mathrm{poi}(n)}(x^{n'})}{q(x^{n'})}\right] \\
&= \max_{p\in\mathcal{P}} \sum_{n'=0}^{\infty} e^{-n}\frac{n^{n'}}{n'!}\mathbb{E}\left[\log\frac{e^{-n}\frac{n^{n'}}{n'!}p(x^{n'})}{e^{-n}\frac{n^{n'}}{n'!}q_{n'}(x^{n'})}\right] \\
&\le \sum_{n'=0}^{\infty} e^{-n}\frac{n^{n'}}{n'!}\max_{p\in\mathcal{P}}\mathbb{E}\left[\log\frac{p(x^{n'})}{q_{n'}(x^{n'})}\right] \\
&= \sum_{n'=0}^{\infty} e^{-n}\frac{n^{n'}}{n'!}\bar{R}(\mathcal{P}^{n'}),
\end{aligned}$$

where the last equality follows from definition of $q_{n'}$. By monotonicity and sub-additivity of $\bar{R}(\mathcal{P}^{n'})$ (see Lemma 5 in Acharya et al. [2012]), it follows that

$$\begin{aligned}
\bar{R}(\mathcal{P}^{n'}) &\le \bar{R}(\mathcal{P}^{n\lceil\frac{n'}{n}\rceil}) \\
&\le \left\lceil\frac{n'}{n}\right\rceil \bar{R}(\mathcal{P}^{n}) \\
&\le \left(\frac{n'}{n}+1\right)\bar{R}(\mathcal{P}^{n}).
\end{aligned}$$

Substituting the above bound we get

$$\begin{aligned}
\bar{R}(\mathcal{P}^{\mathrm{poi}(n)}) &\le \sum_{n'=0}^{\infty} e^{-n}\frac{n^{n'}}{n'!}\left(\frac{n'}{n}+1\right)\bar{R}(\mathcal{P}^{n}) \\
&= 2\bar{R}(\mathcal{P}^{n}),
\end{aligned}$$

where the last equality follows from the fact that expectation of n' is n. ■

Type redundancy: In the following lemma we show that the redundancy of the sequence is same as the redundancy of the type vector. Therefore we can focus on compressing the type of the sequence and calculate the expected redundancy of that.

Lemma 4. *Lets define* $\tau(\mathcal{P}^n) = \{\tau(p^n) : p \in \mathcal{P}\}$, *then we have*

$$\bar{R}(\tau(\mathcal{P}^n)) = \bar{R}(\mathcal{P}^n).$$

Proof.

$$\begin{aligned}
\bar{R}(\mathcal{P}^n) &= \min_q \max_{p\in\mathcal{P}} \mathbb{E}\left[\log \frac{p(x^n)}{q(x^n)}\right] \\
&= \min_q \max_{p\in\mathcal{P}} \sum_{x^n\in\mathcal{X}^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} \\
&= \min_q \max_{p\in\mathcal{P}} \sum_{\tau} \sum_{x^n\in\mathcal{X}^n:\tau(x^n)=\tau} p(x^n) \log \frac{p(x^n)}{q(x^n)} \\
&\overset{(a)}{=} \min_q \max_{p\in\mathcal{P}} \sum_{\tau} \left(\sum_{x^n:\tau(x^n)=\tau} p(x^n)\right) \log \frac{\sum_{x^n:\tau(x^n)=\tau} p(x^n)}{\sum_{x^n:\tau(x^n)=\tau} q(x^n)} \\
&= \min_q \max_{p\in\mathcal{P}} \sum_{\tau} p(\tau) \log \frac{p(\tau)}{q(\tau)} \\
&= \bar{R}(\tau(\mathcal{P}^n))
\end{aligned}$$

where (a) is by convexity of KL-divergence and the fact that all sequences of a specific type have the same probability. ■

Now we reach to the main part of this section, i.e. lower bounding the expected redundancy of class $\mathcal{P}_d^n$. Based on the previous lemmas, we have

$$\bar{R}(\mathcal{P}_d^n) \geq \frac{1}{2}\bar{R}(\mathcal{P}_d^{poi(n)}) = \frac{1}{2}\bar{R}(\tau(\mathcal{P}_d^{poi(n)}))$$

and therefore it is enough to show a lower bound on $\bar{R}(\tau(\mathcal{P}_d^{poi(n)}))$. We decompose $\bar{R}(\tau(\mathcal{P}_d^{poi(n)}))$ as

$$\begin{aligned}
\bar{R}(\tau(\mathcal{P}_d^{poi(n)})) &= \min_q \max_{\tau^k\in\tau(\mathcal{P}_d^{poi(n)})} \sum_{\tau^k} p(\tau^k) \log \frac{p(\tau^k)}{q(\tau^k)} \\
&= \min_q \max_{\tau^k\in\tau(\mathcal{P}_d^{poi(n)})} \sum_{\tau^k} p(\tau^k) \log \frac{1}{q(\tau^k)} \\
&\quad - \sum_{\tau^k} p(\tau^k) \log \frac{1}{p(\tau^k)}
\end{aligned}$$

Hence it suffices to show a lower bound on $\sum_{\tau^k} p(\tau^k) \log \frac{1}{q(\tau^k)}$ and an upper bound on $\sum_{\tau^k} p(\tau^k) \log \frac{1}{p(\tau^k)}$. For the first term, we upper bound $q(\tau^k)$ based on the number of distinct elements in sequence $x^{poi(n)}$. Lemmas 5, 6, 7 prove this upper bound. Afterwards we consider the second term and it turns out that this term is nothing but the entropy of the type vectors under Poisson sampling.

The following two concentration lemmas from Gnedin et al. [2007], Ben-Hamou et al. [2014] help us to relate the expected number of distinct elements for normal and Poisson sampling. We continue by a lemma making connection between those two quantities. Denote the number of distinct elements in $x^{poi(n)}$ as $\varphi_+^{poi(n)}$, and $d^{poi(n)} = \mathbb{E}[\varphi_+^{poi(n)}]$. Similarly, φ_+^n is the number of distinct elements in x^n and $d = \mathbb{E}[\varphi_+^n]$.

Lemma 5. *(Ben-Hamou et al. [2014]) Let $v = \mathbb{E}[\varphi_1^{poi(n)}]$ be the expected number of elements which appeared once in $x^{poi(n)}$, then*

$$\Pr[\varphi_+^{poi(n)} < d^{poi(n)} - \sqrt{2vs}] \leq e^{-s}.$$

Lemma 6. *(Lemma 1 in Gnedin et al. [2007]) Let $\mathbb{E}[\varphi_2^{poi(n)}]$ be the expected number of elements which appeared twice in $x^{poi(n)}$, then*

$$|d^{poi(n)} - d| < 2\frac{\mathbb{E}[\varphi_2^{poi(n)}]}{n}.$$

Using Lemmas 5 and 6 we lower and upper bound the number of non-zero elements in $\tau(x^{poi(n)})$.

Lemma 7. *The number of non-zero elements in $\tau(x^{poi(n)})$ is more than $(1-\epsilon)d$ with probability $> 1 - e^{-\frac{d(\epsilon-2/n)^2}{2}}$. Also, the number of non-zero elements in $\tau(x^{poi(n)}) < (1+\epsilon)d$ with probability $> 1 - e^{-\frac{d(\epsilon-2/n)^2}{2}}$.*

Proof. The number of non-zero elements in τ is equal to the number of distinct elements in $x^{poi(n)}$. By Lemma 5

$$\begin{aligned}\Pr[\varphi_+^{poi(n)} < d^{poi(n)}(1-\epsilon)] &\leq e^{-\frac{(d^{poi(n)}\epsilon)^2}{2v}} \\ &\overset{(a)}{\leq} e^{-\frac{d^{poi(n)}\epsilon^2}{2}},\end{aligned}$$

where (a) is because $d^{poi(n)} > v$. Lemma 6 implies $d^{poi(n)}(1-\frac{2}{n}) < d < d^{poi(n)}(1+\frac{2}{n})$. Therefore,

$$\begin{aligned}\Pr[\varphi_+^{poi(n)} < d(1-\epsilon)] &\leq \Pr[\varphi_+^{poi(n)} < d^{poi(n)}\left(1+\frac{2}{n}\right)(1-\epsilon)] \\ &\leq e^{-\frac{d(\epsilon-\frac{2}{n})^2}{2}}.\end{aligned}$$

Proof of the other part is similar and omitted. ■

Next, we lower bound the number of bits we need to express τ^k based on the number of nonzero elements in it.

Lemma 8. *If number of non-zero elements in τ^k is more than d', then*

$$q(\tau^k) \leq \frac{1}{\binom{k}{d'}}.$$

Proof. Consider all the type vectors with the same number of non-zero elements as τ^k. It is not hard to see that q should assign same probability to all types with the same profile vector. Number of such type vectors for a given number of non-zero elements d' is at least $\binom{k}{d'}$. ■

Note that the number of non-zero elements in τ^k is same as $\varphi_+^{poi(n)}$. Based on Lemmas 7 and 8 we have

$$\begin{aligned}
\sum_{\tau^k} p(\tau^k) \log \frac{1}{q(\tau^k)} &\geq \sum_{\tau^k : \varphi_+^{poi(n)} \geq (1-\epsilon)d} p(\tau^k) \log \frac{1}{q(\tau^k)} \\
&\geq \sum_{\tau^k : \varphi_+^{poi(n)} \geq (1-\epsilon)d} p(\tau^k) \log \binom{k}{d(1-\epsilon)} \\
&\geq \left(1 - e^{-\frac{d(\epsilon - \frac{2}{n})^2}{2}}\right) \log \binom{k}{d(1-\epsilon)} \\
&= \log \binom{k}{d} (1 + o_d(1)). \qquad (5)
\end{aligned}$$

where the last line is by choosing $\epsilon = d^{-\frac{1}{3}}$. Now we focus on bounding the entropy of the type. Recall that if distribution p is sampled *i.i.d.* $poi(n)$ times, then the number of times symbol i appears, μ_i, is an independent Poisson random variable with mean $\lambda_i = np_i$. First we state a useful lemma in calculation of the entropy.

Lemma 9. *If $X \sim poi(\lambda)$ for $\lambda < 1$, then*

$$H(X) \leq \lambda[1 - \log \lambda] + e^{-\lambda} \frac{\lambda^2}{1-\lambda}.$$

Proof.

$$\begin{aligned}
H(X) &= -\sum_{i=0}^{\infty} p_i \log p_i \\
&= -\sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} \log \frac{e^{-\lambda}\lambda^i}{i!} \\
&= -\sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} \left[\log e^{-\lambda} + i \log \lambda - \log(i!)\right] \\
&= \lambda \sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} - \log \lambda \sum_{i=0}^{\infty} i e^{-\lambda} \frac{\lambda^i}{i!} + \sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} \log(i!) \\
&\overset{(a)}{=} [\lambda - \lambda \log \lambda] + e^{-\lambda} \left[\sum_{i=2}^{\infty} \frac{\lambda^i \log(i!)}{i!}\right] \\
&\leq [\lambda - \lambda \log \lambda] + e^{-\lambda} \left[\sum_{i=0}^{\infty} \lambda^i\right] \\
&\overset{(b)}{=} \lambda[1 - \log \lambda] + e^{-\lambda} \frac{\lambda^2}{1-\lambda}
\end{aligned}$$

where (a) is because the first two terms in the last summation is zero and for the rest of the terms, $\log(i!) < i!$ Also (b) follows from geometric sum for $\lambda < 1$. ■

We can write

$$
\begin{aligned}
H(\tau^k) &= \sum_{i=1}^{k} H(\mu_i) \\
&= \sum_{i=1}^{k} H(poi(\lambda_i)) \\
&= \sum_{\lambda_i<0.7} H(poi(\lambda_i)) + \sum_{\lambda_i\geq 0.7} H(poi(\lambda_i)) \\
&\overset{(a)}{=} \sum_{\lambda_i<0.7} \left(\lambda_i - \lambda_i \log \lambda_i + e^{-\lambda_i}\frac{\lambda_i^2}{1-\lambda_i}\right) + \sum_{\lambda_i\geq 0.7} H(poi(\lambda_i)) \\
&\overset{(b)}{\leq} \sum_{\lambda_i<0.7} (3\lambda_i - \lambda_i \log \lambda_i) + \sum_{\lambda_i\geq 0.7} \frac{1}{2}\log\left(2\pi e(\lambda_i + \frac{1}{12})\right) \qquad (6)
\end{aligned}
$$

where (a) is due to Lemma 9 and (b) is by using Equation (1) in Adell et al. [2010] and the fact that $e^{-x}\frac{x^2}{1-x} < 2x$ for $x < 0.7$.

In the rest of this section, we calculate an upper bound for the second term in (6).Note that the first term in the same equation, i.e. $\sum_{\lambda_i<0.7} H(poi(\lambda_i))$ is heavily dependent on the shape of distributions in the class. In other words, upper bounding this term generally, will lead us to a weak lower bound, while plugging in the exact values leads to a matching lower bound for the intended envelope class, i.e. Zipf distributions.

Let n^- be the sum of all $\lambda < 0.7$ and n^+ be the sum of all $\lambda \geq 0.7$. Similarly, we define k^- and k^+ as the number of $\lambda < 0.7$ and $\lambda \geq 0.7$ respectively. Therefore we have

$$
\begin{aligned}
k^+\left(1 - \frac{1}{e^{-0.7}}\right) &= \sum_{i:\lambda_i\geq 0.7}\left(1 - \frac{1}{e^{-0.7}}\right) \\
&\leq \sum_{i:\lambda_i\geq 0.7} 1 - exp(-\lambda_i) \\
&\leq \sum_{i} 1 - exp(-\lambda_i) \\
&\overset{(a)}{\leq} d(1+\epsilon)
\end{aligned}
$$

where (a) follows from Lemma 5 and the fact $d^{poi(n)} = \sum_i 1 - exp(-\lambda_i)$. Hence we have $k^+ \leq \frac{d(1+\frac{2}{n})}{1-\frac{1}{e^{0.7}}} \leq 2d(1+\frac{2}{n})$ and consequently $k^- \geq k - \frac{d(1+\frac{2}{n})}{1-\frac{1}{e^{0.7}}}$. For the type entropy we know

$$
\begin{aligned}
\sum_{\lambda_i\geq 0.7} \frac{1}{2}\log\left(2\pi e(\lambda_i + \frac{1}{12})\right) &\overset{(a)}{\leq} \frac{1}{2}k^+ \log\left(2\pi e(\frac{n^+}{k^+} + \frac{1}{12})\right) \\
&\overset{(b)}{\leq} d(1+\frac{2}{n})\log\left(\pi e(\frac{n}{d(1+\frac{2}{n})} + \frac{1}{6})\right) \\
&= \left(d\log(\frac{n}{d} + \frac{1}{6}) + d\log \pi e\right)(1 + o_d(1)) \qquad (7)
\end{aligned}
$$

where (a) is by concavity of logarithm and (b) is by monotonicity.

Lemma 10.

$$\bar{R}(\mathcal{P}_d^n) \geq \left(\log\binom{k}{d} - d\log\left(\frac{n}{d}+\frac{1}{6}\right) - d\log \pi e\right)(1+o_d(1)) - \sum_{\lambda_i<0.7}(3\lambda_i - \lambda_i \log \lambda_i).$$

Proof. (5), (6), and (7) leads to the theorem. ■

6 Expected redundancy of unordered power-law envelope

To use Lemmas 2 and 10 we need to bound the number of distinct elements that appear from any distribution in the envelope class $\mathcal{E}_{(f)}$ in addition to calculating the last summation in Lemma 10. For a distribution $p \in \mathcal{E}_{(f)}$ the number of distinct elements is

$$\begin{aligned}
\mathbb{E}[\varphi_+^n] &= \sum_{i=1}^k \mathbb{E}[\mathbb{I}_{\mu_i>0}] \\
&= \sum_{i=1}^k 1-(1-p_i)^n \\
&= \sum_{i=1}^k 1-(1-p_{(i)})^n \\
&\leq \sum_{i=1}^k 1-(1-f(i))^n \\
&\leq \sum_{i:f(i)\geq 1/n} 1 + \sum_{i:f(i)<1/n} 1-(1-f(i))^n \\
&\leq \sum_{i:f(i)\geq 1/n} 1 + \sum_{i:f(i)<1/n} nf(i).
\end{aligned}$$

Thus we need to bound the number of elements with envelope $\geq 1/n$ and the sum of envelopes for elements that are less than $1/n$. For $\mathcal{E}_{(ci^{-\alpha},k)}$, the first term is $\leq (n/c)^{1/\alpha}$ and the second term is

$$\begin{aligned}
\leq \sum_{i=(n/c)^{1/\alpha}}^k cni^{-\alpha} &\leq \frac{c}{\alpha-1} n(n/c)^{\frac{1-\alpha}{\alpha}} \\
&\leq \frac{c^2}{\alpha-1} n^{1/\alpha}.
\end{aligned}$$

Combining these, we get

$$d \leq \left(\frac{1}{c^{1/\alpha}} + \frac{c^2}{\alpha-1}\right) n^{1/\alpha}.$$

For $\mathcal{P}_{(\mathrm{zipf}(\alpha,k))}$, we calculate $\sum_{\lambda_i<0.7}\lambda_i$ and $\sum_{\lambda_i<0.7} -\lambda_i\log\lambda_i$ for $\alpha>1$. In the below calcula-

tions "$\approx$" means that the quantities are equal up-to a multiplicative factor of $1 + o_n(1)$.

$$\begin{aligned}
n^- = \sum_{\lambda_i < 0.7} \lambda_i = & \sum_{i=\lfloor (\frac{10n}{7C_{k,\alpha}})^{\frac{1}{\alpha}} \rfloor + 1}^{k} n \frac{i^{-\alpha}}{C_{k,\alpha}} \\
\approx & n \int_{\left(\frac{10n}{7C_{k,\alpha}}\right)^{1/\alpha}}^{k} \frac{i^{-\alpha}}{C_{k,\alpha}} di \\
\approx & \frac{n}{(\alpha-1)C_{k,\alpha}} \left(\left(\frac{10n}{7C_{k,\alpha}} \right)^{-(\alpha-1)/\alpha} - k^{-(\alpha-1)} \right) \\
\approx & \frac{n}{(\alpha-1)C_{k,\alpha}} \left(\frac{10n}{7C_{k,\alpha}} \right)^{-(\alpha-1)/\alpha} \\
= & \frac{7}{10(\alpha-1)} \left(\frac{10n}{7C_{k,\alpha}} \right)^{\frac{1}{\alpha}}
\end{aligned}$$

Now we calculate the other summation. For $\mathcal{P}_{(\text{zipf}(\alpha,k))}$,

$$\begin{aligned}
\sum_{\lambda_i < 0.7} -\lambda_i \log \lambda_i = & \sum_{i=\lfloor (\frac{10n}{7C_{k,\alpha}})^{\frac{1}{\alpha}} \rfloor + 1}^{k} n \frac{i^{-\alpha}}{C_{k,\alpha}} \log(\frac{C_{k,\alpha} i^{\alpha}}{n}) \\
= & n^- \log(\frac{C_{k,\alpha}}{n}) + \frac{n\alpha}{C_{k,\alpha}} \sum_{i=\lfloor (\frac{10n}{7C_{k,\alpha}})^{\frac{1}{\alpha}} \rfloor + 1}^{k} i^{-\alpha} \log i \\
\leq & n^- \log(\frac{C_{k,\alpha}}{n}) + \frac{n\alpha}{C_{k,\alpha}} \int_{\left(\frac{10n}{7C_{k,\alpha}}\right)^{1/\alpha} - 1}^{k} i^{-\alpha} \log i di \\
\leq & n^- \log(\frac{C_{k,\alpha}}{n}) + \frac{n\alpha}{C_{k,\alpha}} \left[\frac{x^{1-\alpha}((\alpha-1)\log x + 1)}{(\alpha-1)^2} \right]_{k}^{\left(\frac{2n}{C_{k,\alpha}}\right)^{1/\alpha}} + \frac{n\alpha}{C_{k,\alpha}} \frac{1}{\alpha} \left(\frac{10n}{7C_{k,\alpha}} \right)^{-1} \log \left(\frac{10n}{7C_{k,\alpha}} \right) \\
\leq & n^- \log(\frac{C_{k,\alpha}}{n}) + \frac{n}{C_{k,\alpha}} \frac{1}{\alpha - 1} (\frac{10n}{7C_{k,\alpha}})^{\frac{1-\alpha}{\alpha}} \log \frac{10n}{7C_{k,\alpha}} + \frac{n\alpha}{C_{k,\alpha}(\alpha-1)^2} \left(\frac{10n}{7C_{k,\alpha}} \right)^{\frac{1}{\alpha}-1} \\
& + \frac{n}{C_{k,\alpha}} \left(\frac{10n}{7C_{k,\alpha}} \right)^{-1} \log \left(\frac{10n}{7C_{k,\alpha}} \right) \\
\leq & \frac{7}{10(\alpha-1)} \left(\frac{10n}{7C_{k,\alpha}} \right)^{\frac{1}{\alpha}} \log(\frac{C_{k,\alpha}}{n}) + \frac{7}{10(\alpha-1)} (\frac{10n}{7C_{k,\alpha}})^{\frac{1}{\alpha}} \log \frac{10n}{7C_{k,\alpha}} \\
& + \frac{7\alpha}{10(\alpha-1)^2} (\frac{10n}{7C_{k,\alpha}})^{\frac{1}{\alpha}} + \frac{7}{10} \log \frac{10n}{7C_{k,\alpha}} \\
= & \frac{11.2\alpha - 4.2}{10(\alpha-1)^2} \left(\frac{10n}{7C_{k,\alpha}} \right)^{\frac{1}{\alpha}} + \frac{7}{10} \log \frac{10n}{7C_{k,\alpha}}
\end{aligned}$$

Substituting the above bounds in Lemmas 2 and 10 results in the following theorem.

Theorem 11. *For $k > n$, $c_1 = \left(\frac{1}{c^{1/\alpha}} + \frac{c^2}{\alpha - 1}\right)$, $c_1' = \left(C_{k,\alpha}^{1/\alpha} + \frac{C_{k,\alpha}^{-2}}{\alpha - 1}\right)$, and $c_2 = \frac{32.2\alpha - 25.2}{10(\alpha-1)^2}\left(\frac{10}{7C_{k,\alpha}}\right) - c_1' \cdot \log \pi e$*

$$\bar{R}(\mathcal{E}^n_{(ci^{-\alpha},k)}) \geq \bar{R}(\mathcal{P}^n_{(zipf(\alpha,k))}) \geq \left(\log \binom{k}{c_1' n^{\frac{1}{\alpha}}} - c_1'(1 - \frac{1}{\alpha}) n^{\frac{1}{\alpha}} \log \frac{n}{c_1'}\right)(1 + o_n(1)) - c_2 \cdot n^{\frac{1}{\alpha}} - \frac{7}{10} \log \frac{10n}{7C_{k,\alpha}},$$

and

$$\bar{R}(\mathcal{E}^n_{(ci^{-\alpha},k)}) \leq \log \binom{k}{c_1 n^{\frac{1}{\alpha}}} + c_1(2 - \frac{1}{\alpha} + 2\log e) \cdot n^{\frac{1}{\alpha}} \log \frac{n}{c_1} + \log(n+1).$$

We can write both of the bounds above in order notation as

$$\bar{R}(\mathcal{E}^n_{(ci^{-\alpha},k)}) = \Theta(n^{\frac{1}{\alpha}} \log k).$$

References

J. Acharya, H. Das, and A. Orlitsky. Tight bounds on profile redundancy and distinguishability. In *NIPS*, 2012. 1.2, 5.2, 5.2

J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Tight bounds for universal compression of large alphabets. In *ISIT*, pages 2875–2879, 2013. 1.2, 5.2

J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Efficient compression of monotone and m-modal distributions. In *Proceedings of IEEE Symposium on Information Theory*, 2014a. 1.2

J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Universal compression of envelope classes: Tight characterization via poisson sampling. *CoRR*, abs/1405.7460, 2014b. URL http://arxiv.org/abs/1405.7460. 1.2, 3, 4.3, 5.2

L. A. Adamic and B. A. Huberman. Zipf's law and the internet. *Glottometrics*, 3(1):143–150, 2002. 1.3

J. A. Adell, A. Lekuona, and Y. Yu. Sharp bounds on the entropy of the poisson law and related quantities. *Information Theory, IEEE Transactions on*, 56(5):2299–2306, 2010. 5.2

A. Ben-Hamou, S. Boucheron, and M. I. Ohannessian. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *CoRR*, abs/1412.8652, 2014. URL http://arxiv.org/abs/1412.8652. 5.2, 5

S. Boucheron, A. Garivier, and E. Gassiat. Coding on countably infinite alphabets. *IEEE Transactions on Information Theory*, 55(1):358–373, 2009. 1.2, 4.3

S. Boucheron, E. Gassiat, and M. Ohannessian. About adaptive coding on countable alphabets: Max-stable envelope classes. *CoRR, abs/1402.6305*, 2014. 1.2

S. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proc. of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318, 1996. 1.2

T. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, January 1991. 1.2

L. Davisson. Universal noiseless coding. *IEEE Transactions on Information Theory*, 19(6):783–795, Nov. 1973. 1.2

L. D. Davisson, R. J. McEliece, M. B. Pursley, and M. S. Wallace. Efficient universal noiseless source codes. *IEEE Transactions on Information Theory*, 27(3):269–279, 1981. 1.2

A. Garivier. A lower-bound for the maximin redundancy in pattern coding. *Entropy*, 11(4):634–642, 2009. 1.2

A. Gnedin, B. Hansen, J. Pitman, et al. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probab. Surv*, 4(146-171):88, 2007. 5.2, 6

J. Kieffer. A unified approach to weak universal source coding. *IEEE Transactions on Information Theory*, 24(6):674–682, Nov. 1978. 1.2

R. Krichevsky and V. Trofimov. The preformance of universal coding. *IEEE Transactions on Information Theory*, 27(2):199–207, Mar. 1981. 1.2

M. Mitzenmacher and E. Upfal. *Probability and computing - randomized algorithms and probabilistic analysis*. Cambridge Univ. Press, 2005. ISBN 978-0-521-83540-4. 5.2

A. Orlitsky and N. Santhanam. Speaking of infinity. *IEEE Transactions on Information Theory*, To appear, 2004. 1.2

A. Orlitsky, N. Santhanam, and J. Zhang. Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, 50(7):1469– 1481, July 2004. 1.2

V. Pareto. Cours d' économie politique, reprinted as a volume of oeuvres completes. *Droz, Geneva*, 1965, 1896. 1.3

J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, January 1996. 1.2

G. Shamir. A new upper bound on the redundancy of unknown alphabets. In *CISS, Princeton*, 2004. 1.2

G. Shamir. Universal lossless compression with unknown alphabets—the average case. *IEEE Transactions on Information Theory*, 52(11):4915–4944, Nov. 2006. 1.2

G. I. Shamir. Universal source coding for monotonic and fast decaying monotonic distributions. *IEEE Transactions on Information Theory*, 59(11):7194–7211, 2013. 1.2

Y. M. Shtarkov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 23(3):3–17, 1987. 4.1

W. Szpankowski. On asymptotics of certain recurrences arising in universal coding. *Problems of Information Transmission*, 34(2):142–146, 1998. 1.2

W. Szpankowski and M. J. Weinberger. Minimax redundancy for large alphabets. In *ISIT*, pages 1488–1492, 2010. 1.2

W. Szpankowski and M. J. Weinberger. Minimax pointwise redundancy for memoryless models over large alphabets. *Information Theory, IEEE Transactions on*, 58(7):4094–4104, 2012. 1.2

F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens. The context-tree weighting method: basic properties. *IEEE Transactions on Information Theory*, 41(3):653–664, 1995. 1.2

Q. Xie and A. R. Barron. Asymptotic minimax regret for data compression, gambling and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445, 2000. 1.2

X. Yang and A. Barron. Large alphabet coding and prediction through poissonization and tilting. In *The Sixth Workshop on Information Theoretic Methods in Science and Engineering, Tokyo*, 2013. 5.2

G. K. Zipf. Selected studies of the principle of relative frequency in language. 1932. 1.3

G. K. Zipf. Human behavior and the principle of least effort. 1949. 1.3