Universal Compression of Envelope Classes: Tight Characterization via Poisson Sampling[*]

Jayadev Acharya[†1], Ashkan Jafarpour[‡2], Alon Orlitksy[§2], and Ananda Theertha Suresh[¶2]

[1]Massachusetts Institute of Technology
[2]University of California, San Diego

May 30, 2014

Abstract

The Poisson-sampling technique eliminates dependencies among symbol appearances in a random sequence. It has been used to simplify the analysis and strengthen the performance guarantees of randomized algorithms. Applying this method to universal compression, we relate the redundancies of fixed-length and Poisson-sampled sequences, use the relation to derive a simple *single-letter formula* that approximates the redundancy of *any* envelope class to within an additive logarithmic term. As a first application, we consider *i.i.d.* distributions over a small alphabet as a step-envelope class, and provide a short proof that determines the redundancy of discrete distributions over a small alphabet up to the first order terms. We then show the strength of our method by applying the formula to tighten the existing bounds on the redundancy of exponential and power-law classes, in particular answering a question posed by Boucheron, Garivier and Gassiat [6].

1 Introduction

Compression concerns efficient representation of random phenomena. Let a random variable X be generated according to a distribution P over a discrete set $\mathcal{X}$. The best compression of X is achieved when P is known in advance. Roughly speaking, every $x \in \mathcal{X}$ is then represented by $\log(1/P(x))$ bits, where throughout the paper log represents logarithm to the base 2 and ln represents the natural logarithm. Hence the expected number of bits used is the distribution's *entropy*

$$H(P) \overset{\text{def}}{=} \sum_{x\in\mathcal{X}} P(x) \log \frac{1}{P(x)} = \mathop{\mathbb{E}}_{P} \log \frac{1}{P(X)}.$$

Universal compression addresses compression in the common setting where the underlying distribution P is not known, but can be assumed to belong to a known class $\mathcal{P}$ of distributions over

[*] A part of the paper will appear in IEEE International Symposium on Information Theory 2014

[†] jayadev@csail.mit.edu. Part of the work done while author was a student at UCSD.

[‡] ashkan@ucsd.edu

[§] alon@ucsd.edu

[¶] asuresh@ucsd.edu

$\mathcal{X}$, for example the class of *i.i.d.* or Markov distributions. Any compression of X can be viewed as representing $x \in \mathcal{X}$ using $\log(1/Q(x))$ bits for some distribution Q over $\mathcal{X}$ [9], and in the rest of the paper we will use the two interchangeably. The expected number of bits that Q uses to represent $X \sim P$ is the cross entropy

$$\sum_{x \in \mathcal{X}} P(x) \log \frac{1}{Q(x)} = \mathop{\mathbb{E}}_{P} \log \frac{1}{Q(X)}.$$

The expected additional number of bits, beyond the entropy, that Q uses to represent X is

$$\mathop{\mathbb{E}}_{P} \log \frac{1}{Q(X)} - \mathop{\mathbb{E}}_{P} \log \frac{1}{P(X)} = \mathop{\mathbb{E}}_{P} \log \frac{P(X)}{Q(X)}.$$

Though we will not use this fact, this difference is the KL divergence $D(P||Q)$ between P and Q.

The *expected redundancy* of a distribution collection $\mathcal{P}$ is the lowest increase in the expected number of bits achieved by any compression scheme Q,

$$\overline{R}(\mathcal{P}) \overset{\text{def}}{=} \inf_{Q} \sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{P} \log \frac{P(X)}{Q(X)}.$$

A more stringent redundancy measure, which is also the focus of this paper, is the *worst-case* redundancy (a.k.a. *minimax regret*) that represents the largest possible increase between the number of bits that Q and the best compression scheme use to represent any $x \in \mathcal{X}$,

$$\hat{R}(\mathcal{P}) \overset{\text{def}}{=} \inf_{Q} \sup_{P \in \mathcal{P}} \sup_{x \in \mathcal{X}} \log \frac{P(x)}{Q(x)}$$

Worst-case redundancy clearly exceeds expected redundancy for any distribution collection $\mathcal{P}$,

$$\hat{R}(\mathcal{P}) \geq \overline{R}(\mathcal{P}).$$

However, for many popular classes of distributions, they are almost the same.

For example, consider the collection of discrete distributions over $[k] \overset{\text{def}}{=} \{1, \ldots, k\}$,

$$\mathcal{D}_k \overset{\text{def}}{=} \left\{(p_1, \ldots, p_k) : p_i \geq 0, \sum p_i = 1\right\}.$$

It is easy to see that the uniform $Q = (1/k, \ldots, 1/k)$ achieves redundancy $\log k$, and that any other distribution will have a higher redundancy, hence

$$\hat{R}(\mathcal{D}_k) = \overline{R}(\mathcal{D}_k) = \log k.$$

The most well studied distributions are *independent identical* distributions, or *iid*, where a distribution P over $\mathcal{X}$ is sampled n times independently, and the probability of observing $x^n \in \mathcal{X}^n$ is

$$P^n(x^n) \overset{\text{def}}{=} \prod_{i=1}^{n} P(x_i).$$

Similarly, for a class $\mathcal{P}$ define the *i.i.d.* class

$$\mathcal{P}^n \overset{\text{def}}{=} \{P^n : P \in \mathcal{P}\}$$

to consist of the n independent repetitions of any single distribution $P \in \mathcal{P}$.

The most well investigated class of distributions is $\mathcal{D}_k^n$, the collection of all length-n *i.i.d.* distributions over $[k]$.

It is now well established [14, 10, 11, 23, 8, 18, 21, 24, 16, 22] that for $k = o(n)$

$$\overline{R}(\mathcal{D}_k^n) + f_1(k) = \hat{R}(\mathcal{D}_k^n) + f_2(k) = \frac{k-1}{2}\log\frac{n}{k},$$

and for $n = o(k)$

$$\overline{R}(\mathcal{D}_k^n) + g_1(n) = \hat{R}(\mathcal{D}_k^n) + g_2(n) = n\log\frac{k}{n},$$

where f_1, f_2 are independent of n and g_1, g_2 are independent of k.

The problem was traditionally studied in the setting of finite underlying alphabet size k, and large block lengths n, however in numerous applications the underlying natural alphabet best describing the data might be large. For example, a text document only contains a small fraction of all the words in the dictionary, and a natural image contains only a small portion of all possible pixel values. An extreme case of this is when $k = \infty$ and $n = 1$ in the equation above, showing that $\overline{R}(\mathcal{D}_\infty) = \infty$ [14, 10]. The class of all *i.i.d.* distributions over $\mathbb{Z}^+$ is therefore *too large* to provide meaningful compression schemes for all distributions. These impossibility results led researchers to consider natural sub-classes of all *i.i.d.* distributions and then compress sequences generated from distributions in these or to consider all *i.i.d.* distributions but decompose sequences into natural components and compress each part separately.

In recent years, three approaches have been proposed to address compression over large alphabets. [17] considered separate compression of the *dictionary* that describes the symbols appearing, and the *pattern* that specifies their order. For example, the word "paper" has pattern 12134 and dictionary 1→p, 2→a, 3→e, 4→r. They showed that while the pattern contains almost all of the sequence entropy, it can be compressed with sublinear worst-case redundancy, regardless of the alphabet size, hence is *universally compressible.*

[12] considered the subclass of monotone distributions over $\mathbb{Z}^+$. They showed that even this class is not universally compressible, but designed universal codes for all monotone distributions with finite entropy. [19] studied the class $\mathcal{M}_k^n$ of length$-n$ sequences from monotone distributions over $[k]$, and tightly characterized the redundancy for any $k = O(n)$. Recently, [3] studied $\mathcal{M}_k^n$ for much larger range of k and in particular showed that the class is universally compressible for all $k = \exp(o(n/\log n))$ and is not universally compressible for any $k = \exp(\Omega(n))$.

A third approach was proposed in [6]. They studied compression of *envelope classes*, where the probabilities are bounded by an envelope. They provide general bounds on the redundancy of envelope classes. They were motivated by the previous negative results on compressing *i.i.d.* distributions over infinite alphabets, and therefore wanted to consider classes for which it is possible to design universal codes. The upper bounds on the worst-case redundancy are obtained by bounding the Shtarkov sum. They provide bounds on the more stringent (for lower bounds) average case redundancy by employing the redundancy-capacity theorem.

In the next section we introduce and motivate envelope classes and describe some of the known results and our new results.

2 Envelope Class: Known Results

A function $f : \mathbb{Z}^+ \to \mathbb{R}^{\geq 0}$ is called an *envelope*. We abbreviate $f_i \overset{\text{def}}{=} f(i)$. Any envelope f determines an *envelope class*

$$\mathcal{E}_f \overset{\text{def}}{=} \{(p_1, p_2, \dots) \in \mathcal{D}_\infty : p_i \leq f_i\}.$$

When f is defined explicitly, we will sometimes denote $\mathcal{E}_f$ by $\mathcal{E}_{f_i}$. For example, one of our main applications is for the power envelope defined by $f_i = c \cdot i^{-\alpha}$ and the corresponding power class is denoted by $\mathcal{E}_{c \cdot i^{-\alpha}}$.

Envelope classes naturally generalize $\mathcal{D}_k$ for large and potentially infinite k. They can incorporate prior distribution knowledge that can be expressed as an upper bound on the probabilities.

[6] introduced envelope classes and proved several results about their redundancy for general, power-law, and exponential envelopes. They called an envelope function f is *summable* if $\sum_{i=1}^\infty f_i < \infty$, and used the Shtarkov sum to show that $\hat{R}(\mathcal{E}_f) < \infty$ if and only if f is summable. Letting

$$\overline{F}_u \overset{\text{def}}{=} \sum_{i=u+1}^{\infty} f_i$$

be the tail sum, they showed that

$$\hat{R}(\mathcal{E}_f^n) \leq \inf_{u \leq n} \left[n\overline{F}_u + \frac{u-1}{2} \log n \right] + 2.$$

They also used the redundancy-capacity theorem [9, Chap 13] to lower bound the expected and therefore worst-case redundancy, but their expressions are more involved and we refer the reader to their paper.

They applied these bounds to two important and natural envelope classes. For the *power-law* envelope class defined by the envelope $c \cdot i^{-\alpha}$ for $c > 0$ and $\alpha > 1$, they showed[1]

$$C_{c,\alpha} \cdot n^{\frac{1}{\alpha}} \leq \overline{R}(\mathcal{E}^n_{c \cdot i^{-\alpha}}) \leq \hat{R}(\mathcal{E}^n_{c \cdot i^{-\alpha}}) \leq \left(\frac{2cn}{\alpha - 1}\right)^{\frac{1}{\alpha}} (\log n)^{1 - \frac{1}{\alpha}} + O(1), \tag{1}$$

where the constant $C_{c,\alpha}$ depends on c and α. They also noted that the lower-bound and the upper-bound is order $(\log n)^{1-\frac{1}{\alpha}}$ apart and asked whether one of them is tight. One of our results shows that the lower bound is tight up to a constant factor.

For the *exponential-law* envelope class, defined by the envelope $f_i = c \cdot e^{-\alpha i}$ where $c, \alpha > 0$ they proved that[2]

$$\frac{\log^2 n}{8\alpha \log e}(1 + o(1)) \leq \hat{R}(\mathcal{E}^n_{c \cdot e^{\alpha i}}) \leq \frac{\log^2 n}{2\alpha \log e} + O(1).$$

[4] improved these bounds and determined the redundancy up to the first order term and showed

$$\hat{R}(\mathcal{E}^n_{c \cdot e^{\alpha i}}) = \frac{\log^2 n}{4\alpha \log e} + O(\log n \log \log n). \tag{2}$$

More recently, [5] extended the arguments of [4] to find tight universal codes for the larger class of sub-exponential envelope class whose decay is strictly faster than power-law, but slower than

[1] $g(n) = O(f(n))$ if $\exists C, N : \forall n \geq N g(n) \leq Cf(n)$.

[2] $g(n) = o(f(n))$ if $\lim_{n \to \infty} \frac{g(n)}{f(n)} = 0$.

exponential classes. However, as mentioned in these papers, a tight bound on the redundancy of heavy-tail envelopes such as power-laws remained elusive. Note that one of the important goals of [5] was to obtain sequential algorithms achieving the redundancy of these classes, not simply obtaining the bounds as we consider. [7] design efficient codes that are *adaptive* for classes of envelopes, namely they provide efficient compression schemes for families of envelopes, without the knowledge of the precise envelope being followed.

3 New Results and Techniques

Poisson sampling provides simplified analysis in various machine learning and statistical problems [15]. To the best of our knowledge, it was first used for universal compression in [2], and [1] to prove optimal bounds on pattern redundancy. Recently, [25] also apply Poisson sampling to provide simpler coding schemes for *i.i.d.* distributions over large alphabets. They observed that Poisson sampling renders the multiplicities of each symbol independent and hence can be independently coded. Using the tilting method they constructed compression schemes that are optimal upto an additional factor of $\frac{1}{2}\log\frac{n}{k}$ for $\mathcal{D}_k^n$. In a subsequent paper, they showed that it can be extended to sources with memory. The key difference between [25] and this paper is that the former focuses on finding efficient coding schemes for $\mathcal{D}_k^n$ and sources with memory, while we use the Poisson model as a proof technique to simplify the computation of the redundancy of *i.i.d.* distributions and various envelope classes.

In section 5 we describe the Poisson sampling model. We then relate the redundancy of this model to the fixed length model. In Section 6 we consider the class of simple Poisson distributions with bounded means. In Theorem 18, we give *single-letter* bounds on the worst-case redundancy of general envelope classes in terms of redundancy of single Poisson classes. The bounds are simply summations of such Poisson redundancies. The upper and lower bounds in this theorem differ by an additive factor of $O(\log n)$.

We first apply the results to the class $\mathcal{D}_k^n$ and present a short argument that bounds its redundancy tightly up to the first order term. We then apply these bounds to find the worst case redundancy of power-law class. Even though we give a single letter bound that is tight barring additive logarithmic factors, we now focus on providing *closed* form bounds that are tight to a multiplicative factor of 4. In particular, this strengthens Equation (1) answering a question posed in [6]. This also shows that in fact the lower bound of [6] on the average redundancy is within a constant factor from the worst case. We then bound the redundancy of exponential classes improving the second order term in Equation (2).

4 Preliminary Results

We first mention several simple redundancy results that will be useful later in the paper.

4.1 The Shtarkov Sum

Worst-case redundancy can be easily expressed as a sum of probabilities. Let $\mathcal{P}$ be a collection of distributions over $\mathcal{X}$. The *maximum likelihood* probability of $x \in \mathcal{X}$ is

$$\hat{P}(x) \overset{\text{def}}{=} \sup_{P \in \mathcal{P}} P(x),$$

the highest probability assigned to it by any distribution in $\mathcal{P}$. The *maximum-likelihood-*, or *Shtarkov sum* [20] is

$$S(\mathcal{P}) \overset{\text{def}}{=} \sum_{x\in\mathcal{X}} \hat{P}(x).$$

It is easy to see that

$$\hat{R}(\mathcal{P}) = \log\left(S(\mathcal{P})\right).$$

This simple formulation allows for simple proofs of several results for worst-case redundancy.

4.2 Basic Redundancy Properties

All the results mentioned in this section apply to both worst-case and expected redundancy, but for simplicity we present the proof for the worst-case.Unless otherwise mentioned, the following assumes that the distributions are over a set $\mathcal{X}$.

Lemma 1 (Subset redundancy)**.** *If $\mathcal{P}' \subseteq \mathcal{P}$, then*

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

Proof. By Shtarkov's Sum,

$$S(\mathcal{P}') = \sum_{x\in\mathcal{X}} \max_{P\in\mathcal{P}'} P(x) \le \sum_{x\in\mathcal{X}} \max_{P\in\mathcal{P}} P(x) = S(\mathcal{P}). \qquad \square$$

Lemma 2 (Union redundancy)**.** *For all distribution collections $\mathcal{P}_1,\dots,\mathcal{P}_c$,*

$$\max_{1\le i\le c} \hat{R}(\mathcal{P}_i) \le \hat{R}(\overset{c}{\underset{i=1}{\cup}} \mathcal{P}_i) \le \max_{1\le i\le c} \hat{R}(\mathcal{P}_i) + \log c.$$

Proof. The lower bound follows from subset redundancy. For the upper bound, let $\mathcal{X}_i \subseteq \mathcal{X}$ be the set of x's that are assigned the highest probability by a distribution in $\mathcal{P}_i$ and, in case of ties, not in any P_j for $j < i$. Then,

$$S(\overset{c}{\underset{i=1}{\cup}} \mathcal{P}_i) = \sum_{i=1}^{c} \sum_{x\in\mathcal{X}_i} \max_{P\in\mathcal{P}_i} P(x) \le \sum_{i=1}^{c} \sum_{x\in\mathcal{X}} \max_{P\in\mathcal{P}_i} P(x) = \sum_{i=1}^{c} S(\mathcal{P}_i) \le c\cdot \max_{1\le i\le c} S(\mathcal{P}_i). \qquad \square$$

If P is a distribution over $\mathcal{X}$ and $f:\mathcal{X}\to\mathcal{Y}$, then $f(P)$ is the distribution over $\mathcal{Y}$ defined by $[f(P)](y) = P(f^{-1}(y))$. Similarly, if $\mathcal{P}$ is a collection of distributions over $\mathcal{X}$, then $f(\mathcal{P}) \overset{\text{def}}{=} \{f(P) : P\in\mathcal{P}\}$.

Lemma 3 (Function redundancy)**.** *$\hat{R}(f(\mathcal{P})) \le \hat{R}(\mathcal{P})$ with equality iff for every $y\in\mathcal{Y}$, all $x\in f^{-1}(y)$ are assigned their highest probability by the same distribution in $\mathcal{P}$.*

Proof. Follows from the maximum-likelihood sum, with equality under the above condition,

$$S(f(\mathcal{P})) = \sum_{y\in\mathcal{Y}} \max_{P\in\mathcal{P}} \sum_{x\in f^{-1}(y)} P(x) \le \sum_{y\in\mathcal{Y}} \sum_{x\in f^{-1}(y)} \max_{P\in\mathcal{P}} P(x) = \sum_{x\in\mathcal{X}} \max_{P\in\mathcal{P}} P(x) = S(\mathcal{P}). \qquad \square$$

Corollary 4 (Bijection redundancy)**.** *If f is 1-1, then*

$$\hat{R}(f(\mathcal{P})) = \hat{R}(\mathcal{P}). \qquad \square$$

If $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ are distributions over $\mathcal{X}$ and $\mathcal{Y}$ respectively, then $P_{\mathcal{X}} \times P_{\mathcal{Y}}$ is the distribution over $\mathcal{X} \times \mathcal{Y}$ defined by $[P_{\mathcal{X}} \times P_{\mathcal{Y}}](x,y) = P_{\mathcal{X}}(x) \cdot P_{\mathcal{Y}}(y)$. Similarly, if $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{P}_{\mathcal{Y}}$ are collections of distributions over $\mathcal{X}$ and $\mathcal{Y}$ respectively, then $\mathcal{P}_{\mathcal{X}} \times \mathcal{P}_{\mathcal{Y}} \overset{\text{def}}{=} \{P_{\mathcal{X}} \times P_{\mathcal{Y}} : P_{\mathcal{X}} \in \mathcal{P}_{\mathcal{X}}, P_{\mathcal{Y}} \in \mathcal{P}_{\mathcal{Y}}\}$.

Lemma 5 (Product redundancy)**.** *For all $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{P}_{\mathcal{Y}}$,*

$$\hat{R}(\mathcal{P}_{\mathcal{X}} \times \mathcal{P}_{\mathcal{Y}}) = \hat{R}(\mathcal{P}_{\mathcal{X}}) + \hat{R}(\mathcal{P}_{\mathcal{Y}}).$$

Proof. By the Shtarkov sum,

$$S(\mathcal{P}_{\mathcal{X}} \times \mathcal{P}_{\mathcal{Y}}) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \hat{P}(x,y) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \hat{P}(x)\cdot\hat{P}(y) = \left(\sum_{x\in\mathcal{X}} \hat{P}(x)\right)\cdot\left(\sum_{y\in\mathcal{Y}} \hat{P}(y)\right) = S(\mathcal{P}_{\mathcal{X}})\cdot S(\mathcal{P}_{\mathcal{Y}}). \qquad \square$$

For class $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ consisting of product distributions with $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{P}_{\mathcal{Y}}$ being the collection of all marginal distributions over $\mathcal{X}$ and let $\mathcal{Y}$, respectively. Then, $\mathcal{P} \subseteq \mathcal{P}_{\mathcal{X}} \times \mathcal{P}_{\mathcal{Y}}$, and by combining product and subset redundancy,

Corollary 6 (Marginal redundancy)**.** *For every product distribution class $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$,*

$$\hat{R}(\mathcal{P}) \le \hat{R}(\mathcal{P}_{\mathcal{X}}) + \hat{R}(\mathcal{P}_{\mathcal{Y}}). \qquad \square$$

Remark 7. *The result can be generalized to product distributions over a countable number of coordinates $\mathcal{X}_1 \times \mathcal{X}_2 \times \ldots$.*

We now prove some results for the redundancy of *i.i.d.* distributions.

4.3 Implications to *i.i.d.* Distributions

Recall that if P is a distribution over $\mathcal{X}$, then P^n is the induced distribution over $\mathcal{X}^n$, the sequences of length n.

Lemma 8. *1.* Monotonicity: *For all n, $\hat{R}(\mathcal{P}^{n+1}) \ge \hat{R}(\mathcal{P}^n)$.*

2. Subadditivity: *For any n_1, n_2 $\hat{R}(\mathcal{P}^{n_1+n_2}) \le \hat{R}(\mathcal{P}^{n_1}) + \hat{R}(\mathcal{P}^{n_2})$.*

Proof. Monotonicity follows by marginalizing the $(n+1)$th coordinate and taking logarithm in the following.

$$\begin{aligned}
S(\mathcal{P}^{n+1}) &= \sum_{x_1^{n+1}\in\mathcal{X}^{n+1}} \sup_{P\in\mathcal{P}} P^n(x_1^{n+1}) \\
&\ge \sum_{x^n\in\mathcal{X}^n} \sup_{P\in\mathcal{P}} \left[P^n(x^n) \left(\sum_{x_{n+1}\in\mathcal{X}} P(x_{n+1}) \right) \right] \\
&= \sum_{x^n\in\mathcal{X}^n} \sup_{P\in\mathcal{P}} P^n(x^n) \\
&= S(\mathcal{P}^n).
\end{aligned}$$

For subadditivity, we give the proof of [6].

$$\begin{aligned}
S(\mathcal{P}^{n_1+n_2}) &= \sum_{x_1^{n_1+n_2}\in\mathcal{X}^{n_1+n_2}} \sup_{P\in\mathcal{P}} P(x_1^{n_1+n_2}) \\
&= \sum_{x_1^{n_1+n_2}\in\mathcal{X}^{n_1+n_2}} \sup_{P\in\mathcal{P}} \Big[P(x_1^{n_1})P(x_{n_1+1}^{n_2})\Big] \\
&\le \sum_{x_1^{n_1+n_2}\in\mathcal{X}^{n_1+n_2}} \sup_{P\in\mathcal{P}} P(x_1^{n_1}) \sup_{P\in\mathcal{P}} P(x_{n_1+1}^{n_2}) \\
&= \Big[\sum_{x_1^{n_1}\in\mathcal{X}^{n_1}} \sup_{P\in\mathcal{P}} P(x_1^{n_1})\Big] \cdot \Big[\sum_{x_1^{n_2}\in\mathcal{X}^{n_2}} \sup_{P\in\mathcal{P}} P(x_1^{n_2})\Big] \\
&= S(\mathcal{P}^{n_1}) \cdot S(\mathcal{P}^{n_2}).
\end{aligned}$$

Taking logarithm proves the second part. □

4.4 Type Redundancy

The *multiplicity* $m_{x^n}(x)$ is the number of times a symbol x appears in a sequence x^n. The *type* of a sequence x^n over $[k]$ is the k-tuple

$$\tau(x^n) \overset{\text{def}}{=} (m_{x^n}(1), m_{x^n}(2), \ldots, m_{x^n}(k))$$

of multiplicities of the k symbols in the sequence x^n, see *e.g.,* [9]. For example, for $k = 4$, $\tau(11313) = (3, 0, 2, 0)$. Note that a type is a k-tuple of non-negative integers summing to n and that for $k = \infty$, the type has infinite length.

Let $\tau(P^n)$ be the distribution induced by P^n over types. It is easy to see that $\tau(P^n)$ is the multinomial (specifically, *k-nomial*) distribution with parameters n and $(p_1, \ldots, p_k)$, namely

$$P(\tau(X^n) = (m_1, \ldots, m_k)) = \binom{n}{m_1, \ldots, m_k} \cdot \prod_{i=1}^{k} p_i^{m_i}.$$

Note that this definition applies also for infinite k as at most n multiplicities are non-zero. We let

$$\tau(\mathcal{P}^n) = \{\tau(P^n) : P \in \mathcal{P}\}$$

be the set all distributions induced by P^n over types, namely the set of all k-nomial distributions whose first parameter is n.

It is well known that for *i.i.d.* distributions, the redundancy of sequences is equal to the redundancy of types, which follows since any *i.i.d.* distribution assigns the same probability to all sequences of the same type.

Lemma 9. *For $R \in \{\overline{R}, \hat{R}\}$,*

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

Proof. We prove the result for worst case since it has simpler expressions. The average case follows similarly from the definition of average redundancy and convexity of KL divergence. Any *i.i.d.* distribution assigns the same probability to all sequences with the same type. Therefore, such sequences have the same maximum likelihood probability. We show that the Shtarkov sums of the two classes are the same and hence they have same redundancy.

$$S(\mathcal{P}^n) = \sum_{x^n\in\mathcal{X}^n} \hat{P}^n(x^n) = \sum_{\tau} \sum_{x^n:\tau(x^n)=\tau} \hat{P}^n(x^n) = \sum_{\tau} \hat{P}^n(\tau) = S(\tau(\mathcal{P}^n)).$$ □

5 The Poisson Model

Recall that if P is a distribution over $\mathcal{X}$, then P^n is the distribution over $\mathcal{X}^n$ derived by sampling according to P independently n times. A significant difficulty in analyzing such distributions is that although the samples are chosen independently, the number of appearances of different symbols are dependent. For example, they always add to n.

To overcome this difficulty, the Poisson model also considers independent samples according to $\mathcal{P}$, but eliminates the dependence between symbols by generating not a fixed, but a variable number of samples that follow a Poisson distribution.

We first recall the Poisson distribution and one of its concentration result, then define Poisson sampling and mention some of its properties, and finally relate its redundancy to fixed-length redundancy.

5.1 The Poisson Distribution

The Poisson distribution $\text{poi}(\lambda)$ with parameter λ assigns to $i \in \mathbb{N}$ probability

$$\text{poi}(\lambda, i) \overset{\text{def}}{=} e^{-\lambda} \frac{\lambda^i}{i!}.$$

The Poisson distribution concentrates exponentially around its mean, helping relate the redundancy of Poisson- and fixed-length sampling.

Lemma 10. *([15]) Let $X \sim \text{poi}(\lambda)$, then for $x \geq \lambda$,*

$$\text{poi}(\lambda)(X \geq x) \leq \exp\left(-\frac{(x-\lambda)^2}{2x}\right),$$

and for $x \leq \lambda$,

$$\text{poi}(\lambda)(X \leq x) \leq \exp\left(-\frac{(x-\lambda)^2}{2\lambda}\right).$$

□

5.2 Poisson Sampling and its Properties

Poisson-length sampling replaces a fixed number n of samples by a random number $N \sim \text{poi}(n)$ of samples, and then sampling P independently N times. Consequently, the distribution is over the set $\mathcal{X}^* \overset{\text{def}}{=} \cup_{i=0}^{\infty} \mathcal{X}^i$ of finite strings over $\mathcal{X}$, where $\mathcal{X}^0$ contains just the empty string. Therefore the probability of a sequence $x^{n'} \in \mathcal{X}^*$ is

$$P^{\text{poi}(n)}(x^{n'}) = \text{poi}(n, n') \cdot P^{n'}(x^{n'}) = \text{poi}(n, n') \cdot \prod_{i=1}^{n'} P(x_i).$$

The following lemma, which along with its elementary proof can be found in [15], states that conditioned on $N = n'$, the distribution that $P^{\text{poi}(n)}$ induces on sequences is identical to that of $P^{n'}$, and that under Poisson sampling, the multiplicity $M_i^{\text{poi}(n)}$ of symbol i is distributed Poisson, and that multiplicities of different symbols are independent.

Lemma 11. *For all k, $P \in \mathcal{D}_k$, and n, for $1 \leq i \leq k$, $M_i^{\text{poi}(n)} \sim \text{poi}(np_i)$ independently of each other.* □

Therefore, for a distribution $P = (p_1, p_1, \ldots)$ over $\mathcal{X} = \{1, 2, \ldots\}$

$$\tau(\mathcal{P}^{\text{poi}(n)}) = (\text{poi}(np_1), \text{poi}(np_2), \ldots), \tag{3}$$

where each coordinate is an independent Poisson distribution.

Lemma 12. *For all k, $P \in \mathcal{D}_k$, and n, for all n', $P^{\text{poi}(n)}(x^{n'}|N = n') = P^{n'}(x^{n'})$.* □

Similar to $\mathcal{P}^n$, let

$$\mathcal{P}^{\text{poi}(n)} \overset{\text{def}}{=} \{P^{\text{poi}(n)} : P \in \mathcal{P}\}$$

be the class of distributions over $\mathcal{X}^*$, where each distribution consists of a distribution $P \in \mathcal{P}$ sampled independently a random poi(n) times.

The next lemma relates Shtarkov sums for Poisson and fixed-length sampling.

Lemma 13. *For every distribution class $\mathcal{P}$,*

$$S\Big(\mathcal{P}^{\text{poi}(n)}\Big) = \sum_{n'=0}^{\infty} \text{poi}(n, n') \cdot S\left(\mathcal{P}^{n'}\right).$$

Proof. By Lemma 12, the Shtarkov sum conditioned on the length does not change, namely for a sequence $x^{n'}$ the same distribution attains maximum likelihood for both poi(n) sampling and sampling *i.i.d.* n' times. Therefore,

$$S\left(\mathcal{P}^{\text{poi}(n)}\right) = \sum_{n' \geq 0} \text{poi}(n, n') \sum_{x^{n'}} \sup_{P \in \mathcal{P}} P^{n'}(x^{n'}) = \sum_{n' \geq 0} \text{poi}(n, n') S\left(\mathcal{P}^{n'}\right),$$

where in the first step we sum maximum likelihoods of sequences by their lengths. □

We would like to use the independence of multiplicities to obtain simpler bounds on the redundancy of *i.i.d.* distributions. Toward this, we first show in the next result that under mild assumptions, the redundancy of a class under Poisson sampling is *close* to the redundancy under fixed length sampling. The bound we are more interested in, namely proving upper bounds on $\hat{R}(\mathcal{P}^n)$ holds even without these assumptions, namely for arbitrary classes.

Theorem 14. *For any class $\mathcal{P}$*

$$\hat{R}(\mathcal{P}^n) \leq \hat{R}(\mathcal{P}^{\text{poi}(n)}) + 1.$$

Furthermore, for $n \geq 4$ if $\hat{R}(\mathcal{P}^n) < n/16$ and $n_1 \overset{\text{def}}{=} \lfloor n - 3\sqrt{n\hat{R}(\mathcal{P}^n)} \rfloor$, then

$$\hat{R}(\mathcal{P}^{\text{poi}(n_1)}) \leq \hat{R}(\mathcal{P}^n).$$

Remark 15. *The first part of the theorem in some cases can be used to verify if the conditions for the second part hold. This is a potential method of obtaining lower bounds given good upper bounds.*

Proof of Theorem 14. By Lemma 13,

$$\begin{aligned} S(\mathcal{P}^{\text{poi}(n)}) &\overset{(a)}{=} \sum_{n' \geq 0} \text{poi}(n, n') S(\mathcal{P}^{n'}) \\ &\overset{(a)}{\geq} S(\mathcal{P}^n) \sum_{n' \geq n} \text{poi}(n, n') \\ &\overset{(b)}{\geq} \frac{1}{2} S(\mathcal{P}^n). \end{aligned}$$

where (a) from monotonicity and (b) from the fact that if the mean is an integer, then the median of a Poisson distribution is larger than its mean. Taking logarithms proves the first part.

Let $n_1 \overset{\text{def}}{=} \lfloor n - 3\sqrt{n\hat{R}(\mathcal{P}^n)} \rfloor$, then by monotonicity and the fact that median is larger than the mean,

$$S(\mathcal{P}^{\text{poi}(n_1)}) = \sum_{n'} \text{poi}(n_1, n') \cdot S(\mathcal{P}^{n'}) \leq \frac{1}{2} S(\mathcal{P}^n) + \sum_{n' \geq n} \text{poi}(n_1, n') S(\mathcal{P}^{n'}).$$

By the subadditivity and monotonicity,

$$\hat{R}(\mathcal{P}^{n'}) \leq \hat{R}(\mathcal{P}^{\lceil \frac{n'}{n} \rceil n}) \leq \left\lceil \frac{n'}{n} \right\rceil \hat{R}(\mathcal{P}^n) \leq \left(\frac{n'}{n} + 1 \right) \hat{R}(\mathcal{P}^n),$$

and hence $S(\mathcal{P}^{n'}) \leq S(\mathcal{P}^n)(S(\mathcal{P}^n))^{\frac{n'}{n}}$. Therefore,

$$\begin{aligned}
\sum_{n' \geq n} \text{poi}(n_1, n') S(\mathcal{P}^{n'}) &\leq S(\mathcal{P}^n) \sum_{n' \geq n} \text{poi}(n_1, n') (S(\mathcal{P}^n))^{\frac{n'}{n}} \\
&= S(\mathcal{P}^n) \sum_{n' \geq n} \frac{1}{n'!} e^{-n_1} n_1^{n'} (S(\mathcal{P}^n))^{\frac{n'}{n}} \\
&= S(\mathcal{P}^n) \exp(n_1 (S(\mathcal{P}^n)^{\frac{1}{n}} - 1)) \sum_{n' \geq n} \text{poi}(n_1 S(\mathcal{P}^n)^{\frac{1}{n}}, n') \\
&\overset{(a)}{\leq} S(\mathcal{P}^n) \exp\left[n_1 (S(\mathcal{P}^n)^{\frac{1}{n}} - 1) - \frac{(n_1 S(\mathcal{P}^n)^{\frac{1}{n}} - n)^2}{2n} \right] \\
&\overset{(b)}{\leq} \frac{1}{2} S(\mathcal{P}^n).
\end{aligned}$$

(a) uses Poisson tail bounds and $n_1(S(\mathcal{P}^n)^{\frac{1}{n}}) \leq n$. (b) follows from the bounds on $\hat{R}(\mathcal{P}^n)$ and n_1. □

We finally show that sequence redundancy is equal to type redundancy under Poisson sampling.

Lemma 16. *For $R \in \{\overline{R}, \hat{R}\}$,*

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

Proof. We again show only the worst case.

$$S(\mathcal{P}^{\text{poi}(n)}) \overset{(a)}{=} \sum_{n'=0}^{\infty} \text{poi}(n, n') \cdot S\left(\mathcal{P}^{n'}\right) \overset{(b)}{=} \sum_{n'=0}^{\infty} \text{poi}(n, n') \cdot S\left(\tau\left(\mathcal{P}^{n'}\right)\right) = R(\tau(\mathcal{P}^{\text{poi}(n)})),$$

where (a) uses Lemma 13 and (b) follows from Lemma 16. □

This lemma in conjunction with Theorem 14 and 18 reduces the problem of bounding the redundancy of types under Poisson sampling. Under Poisson sampling types of sequences are tuples of independent Poisson random variables, with distribution given by Equation (3).

In the next section we study the simple class of Poisson distributions with bounded means. We use this class as a primitive to bound the redundancy of envelope classes in Section 7.

6 The Redundancy of Bounded Poisson Distributions

Let

$$\mathcal{POI}_\Lambda \overset{\text{def}}{=} \{\text{poi}(\lambda) : \lambda \le \Lambda\}$$

be the class of all Poisson distributions with mean $\le \Lambda$. We approximate the class's redundancy, and will later apply it to approximate the redundancy of all envelope classes.

Lemma 17. *For* $\Lambda \le 1$,

$$\hat{R}\Big(\mathcal{POI}_\Lambda\Big) = \log\left(2 - e^{-\Lambda}\right) \le \Lambda \log e,$$

and for $\Lambda > 1$,

$$\log\sqrt{\frac{2\Lambda + 2}{\pi}} \le \hat{R}(\mathcal{POI}_\Lambda) \le \log\left(\sqrt{\frac{2\Lambda}{\pi}} + 2\right).$$

Proof. Of all Poisson distributions, the probability of $i \in \mathbb{N}$ is maximized by poi(i). Hence of all distributions in $\mathcal{POI}_\Lambda$, the probability of $i \in \mathbb{N}$ is maximized by the Poisson distribution with mean

$$\arg\max_{\lambda \le \Lambda} \text{poi}(\lambda, i) = \begin{cases} i & \text{for } i \le \Lambda, \\ \Lambda & \text{for } i \ge \Lambda. \end{cases}$$

The Shtarkov sum is therefore

$$S(\mathcal{POI}_\Lambda) = \sum_{i=0}^{\lfloor\Lambda\rfloor} e^{-i} \cdot \frac{i^i}{i!} + \sum_{i=\lfloor\Lambda\rfloor+1}^{\infty} e^{-\Lambda} \cdot \frac{\Lambda^i}{i!}.$$

For $\Lambda \le 1$,

$$S\Big(\mathcal{POI}_\Lambda\Big) = 1 + \sum_{i=1}^{\infty} e^{-\Lambda} \cdot \frac{\Lambda^i}{i!} = 1 + (1 - e^{-\Lambda}) = 2 - e^{-\Lambda} \le e^{\Lambda},$$

where the middle equality follows from $\sum_{i=0}^{\infty} e^{-\Lambda} \cdot \frac{\Lambda^i}{i!} = 1$, and the inequality from the arithmetic-geometric inequality.

For $\Lambda > 1$, first observe that Stirling's approximation, stating that for any n, there is some $\theta_n \in (\frac{1}{12n+1}, \frac{1}{12n})$ such that

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\theta_n}, \tag{4}$$

implies that for $i \ge 1$

$$e^{-i} \cdot \frac{i^i}{i!} \le e^{-i} \frac{i^i}{\sqrt{2\pi i}(\frac{i}{e})^i} \le \frac{1}{\sqrt{2\pi i}},$$

and using $e^{-x} \ge 1 - x$,

$$e^{-i} \cdot \frac{i^i}{i!} \ge e^{-i} \frac{i^i}{\sqrt{2\pi i} e^{\frac{1}{12i}} (\frac{i}{e})^i} = \frac{e^{-\frac{1}{12i}}}{\sqrt{2\pi i}} \ge \frac{1}{\sqrt{2\pi i}} - \frac{1}{12\sqrt{2\pi}} \frac{1}{i^{3/2}},$$

Hence,

$$S\Big(\mathcal{POI}_\Lambda\Big) < 2 + \sum_{i=1}^{\lfloor\Lambda\rfloor} e^{-i} \frac{i^i}{i!} \le 2 + \sum_{i=1}^{\lfloor\Lambda\rfloor} \frac{1}{\sqrt{2\pi i}} \le 2 + \sqrt{\frac{2\Lambda}{\pi}},$$

where the last inequality follows from a simple integration.

For the lower bound,

$$S\Big(\mathcal{POI}_\Lambda\Big) \geq 1 + \sum_{i=1}^{\lfloor\Lambda\rfloor} e^{-i}\frac{i^i}{i!} \geq 1 + \sum_{i=1}^{\lfloor\Lambda\rfloor} \frac{1}{\sqrt{2\pi i}} - \frac{1}{12\sqrt{2\pi}}\frac{1}{i^{3/2}}.$$

The lower bound follows by integrating this expression. □

7 The Redundancy of Envelope Classes

We use the redundancy of bounded-Poisson classes to characterize that of envelope classes. Given the similarity in redundancy of fixed-length and Poisson-sampled distributions, determined in Theorem 14, we consider envelope classes under the easier to analyze Poisson sampling.

Let f be a summable envelope. When a distribution $P = (p_1, p_2, \dots) \in \mathcal{E}_f$ is sampled poi(n) times, symbol $i \in \mathbb{Z}^+$ appears poi(np_i) times, where $np_i \leq nf_i$. Let

$$l_f \stackrel{\text{def}}{=} \min\{l : \sum_{i=l}^{\infty} f_i < 1\}$$

be the smallest integer whose tail sum is < 1, or equivalently $\sum_{i=l_f}^{\infty} nf_i < n$. Since f is summable, l_f is finite.

Our main result for envelope classes provides simple lower and upper bounds for their redundancy in terms of the redundancy of bounded Poisson classes. Note that the upper and lower bounds differ by $l_f - 1$ terms. By Lemma 17, they are tight up to an additive $l_f \cdot \log(2 + \sqrt{2n/\pi}) = O_f(\log n)$ term, where O_f implies a constant determined by f.

Theorem 18. *For any envelope f and any n,*

$$\sum_{i=l_f}^{\infty} \hat{R}(\mathcal{POI}_{nf_i}) \leq \hat{R}\left(\mathcal{E}_f^{\text{poi}(n)}\right) \leq \sum_{i=1}^{\infty} \hat{R}(\mathcal{POI}_{nf_i}).$$

Proof. By Lemma 16, it suffices to consider the redundancy of types of sequences. By Equation (3), the class of type distributions induced by $\mathcal{E}_f$ under Poisson sampling is

$$\tau\left(\mathcal{E}_f^{\text{poi}(n)}\right) = \prod_{i=1}^{\infty} \mathcal{POI}_{np_i} \subseteq \prod_{i=1}^{\infty} \mathcal{POI}_{nf_i}.$$

By Corollary 6 generalized to a countable number of dimensions

$$\hat{R}\Big(\mathcal{E}_f^{\text{poi}(n)}\Big) = \hat{R}\Big(\tau\Big(\mathcal{E}_f^{\text{poi}(n)}\Big)\Big) \leq \sum_{i=1}^{\infty} \hat{R}\Big(\mathcal{POI}_{nf_i}\Big).$$

For the lower bound, note that

$$\sum_{i \geq l_f} nf_i < n,$$

and therefore all product distributions in

$$\mathcal{POI}_{nf_{l_f}} \times \mathcal{POI}_{nf_{l_f+1}} \times \dots$$

are valid projections of a distribution in $\mathcal{E}_f^{\text{poi}(n)}$ along the coordinates $i \geq l_f$. Applying Lemma 5 proves the lower bound. □

We now apply this theorem to power law and exponential envelope classes.

8 *i.i.d.* Sequences over Small Alphabet

As the first application of Poisson sampling, we study the redundancy of $\mathcal{D}_k^n$ in the range of small k and n increasing asymptotically. [22] show that for $k = o(n)$,

$$\hat{R}(\mathcal{D}_k^n) = \frac{k-1}{2}\log n - \frac{k}{2}\log k + \frac{k}{2}\log e + o(k).$$

This is a more refined expression than the one presented in the introduction.

Using the tools developed for Poisson sampling, we derive an extremely short argument that bounds the redundancy of this class up to first order terms.

For a length−n sequence with type $\tau' = (m_1, \ldots, m_k)$, let the *abbreviated type* be

$$\tau' \overset{\text{def}}{=} (m_1, \ldots, m_{k-1}),$$

the $(k-1)$−tuple by dropping the last multiplicity. Then, for length−n sequences, there is a bijection between τ and τ'. By Corollary 4,

$$\begin{aligned}
\hat{R}(\mathcal{D}_k^n) &= \hat{R}(\tau(\mathcal{D}_k^n)) \\
&= \hat{R}(\tau'(\mathcal{D}_k^n)) \\
&\overset{(a)}{\leq} \hat{R}(\tau'(\mathcal{D}_k^{\text{poi}(n)})) + 1 \\
&\overset{(b)}{\leq} (k-1)\hat{R}(\mathcal{POI}_n) + 1 \\
&\overset{(c)}{\leq} (k-1)\log\frac{3\sqrt{n}}{\pi} + 1 \\
&\overset{(d)}{\leq} \frac{k-1}{2}\log n + k\log\frac{3}{\pi} + o(k),
\end{aligned}$$

where (a) follows similar to Theorem 14, (b) uses Lemma 5, (c) follows from Lemma 17 by using $2 + \sqrt{2n/\pi} < 3\sqrt{n}$, and (d) by a simple expansion.

We now prove a lower bound along similar lines. Let $n' = n - n^{3/4}k^{1/4}\log n$.

$$\begin{aligned}
\hat{R}(\mathcal{D}_k^n) &\overset{(a)}{\geq} \hat{R}(\mathcal{D}_k^{\text{poi}(n')}) \\
&\overset{(b)}{\geq} (k-1)\hat{R}(\mathcal{POI}_{\frac{n}{k-1}}) \\
&\overset{(c)}{\geq} \frac{(k-1)}{2}\log\frac{2n'}{\pi(k-1)} \\
&\geq \frac{(k-1)}{2}\log n - \frac{k}{2}\log k - \frac{k}{2}\log\frac{\pi}{2} + o(k),
\end{aligned}$$

where (a) uses Theorem 14, (b) follows from the equality condition of Lemma 5, (c) uses Lemma 17.

We note that the lower bound is off by $O(k)$ and the upper bound by $O(k\log k)$.

9 Power Law Envelopes

We apply Theorem 18 to the power law class and provide bounds on redundancy that are at most a factor 4 apart. The upper bound improves results of [6], and proves that their lower bound on the average redundancy is within a constant factor of the worst case redundancy. This resolves an open problem posed in [6] and is stated in the following theorem.

Theorem 19. *For large n*

$$(cn)^{1/\alpha}\Big[\frac{\alpha\log e}{2}+\frac{\log e}{2(\alpha-1)}-\frac{\log\frac{\pi}{2}}{2}\Big](1-o_n(1))\le\hat{R}(\mathcal{E}^n_{c\cdot i^{-\alpha}})\le(cn)^{1/\alpha}\Big[\frac{\alpha\log e}{2}+\frac{\log e}{\alpha-1}+\log 3\Big]+1.$$

Proof. We bound the redundancy of $\mathcal{E}^{\text{poi}(n)}_{c\cdot i^{-\alpha}}$ and then apply Theorem 14 to obtain the result for $\mathcal{E}^n_{c\cdot i^{-\alpha}}$.

9.0.1 Upper Bound

For the power-law class $\mathcal{E}_{c\cdot i^{-\alpha}}$, $np_i\le nf_i=\frac{cn}{i^\alpha}$. Let $b\overset{\text{def}}{=}(cn)^{1/\alpha}$, then

$$nf_i\ge 1 \text{ for } i\le b \text{ and } nf_i<1 \text{ for } i>b.$$

$$\begin{aligned}\hat{R}(\mathcal{E}^{\text{poi}(n)}_{c\cdot i^{-\alpha}}) &\overset{(a)}{\le}\sum_{i\le b}\hat{R}(\mathcal{POI}_{nf_i})+\sum_{i>b}\hat{R}(\mathcal{POI}_{nf_i})\\ &\overset{(b)}{\le}\sum_{i\le b}\log\left(2+\sqrt{\frac{2nf_i}{\pi}}\right)+\Big(\sum_{i>b}^{\infty}nf_i\Big)\log e,\end{aligned}$$

where (a) follows from Theorem 18 and (b) from Lemma 17.

For the first term, note that for $\Lambda\ge 1$, $2+\sqrt{2\Lambda/\pi}<3\sqrt{\Lambda}$. Therefore,

$$\sum_{i=1}^{b}\log\frac{b}{i}=\log\frac{b^b}{b!}\le\log e^b=b\log e,$$

where the inequality follows from Stirling's approximation (Equation (4)). Using $b=(cn)^{\frac{1}{\alpha}}$,

$$\begin{aligned}\sum_{i=1}^{b}\log\left(2+\sqrt{\frac{2\Lambda_i}{\pi}}\right)&<\sum_{i=1}^{b}\log\left(3\sqrt{\frac{cn}{i^\alpha}}\right)\\ &=b\log(3)+\frac{\alpha}{2}\sum_{i=1}^{b}\log\Big(\frac{b}{i}\Big)\\ &<(cn)^{1/\alpha}\Big(\log(3)+\frac{\alpha\log e}{2}\Big).\end{aligned}$$

For the second term,

$$\sum_{i=b+1}^{\infty}nf_i=cn\sum_{i=b+1}^{\infty}\frac{1}{i^\alpha}<cn\int_b^\infty\frac{dx}{x^\alpha}=\frac{c^{1/\alpha}}{\alpha-1}n^{1/\alpha}.$$

Using $\hat{R}(\mathcal{E}^n_{c\cdot i^{-\alpha}})\le\hat{R}(\mathcal{E}^{\text{poi}(n)}_{c\cdot i^{-\alpha}})+1$ from Theorem 14 and adding the terms proves the upper bound.

9.0.2 Lower Bound

The sum

$$\sum_{j=\ell+1}^{\infty} f_i = \sum_{j=\ell+1}^{\infty} \frac{c}{i^\alpha} \le \int_{\ell}^{\infty} \frac{c}{x^\alpha} dx = \frac{c\ell^{-\alpha+1}}{\alpha - 1}$$

is less than 1 for any $\ell > \ell_0 \overset{\text{def}}{=} \left(\frac{c}{\alpha-1}\right)^{\frac{1}{\alpha-1}}$.

Hence by Theorem 18,

$$\hat{R}(\mathcal{E}^{\text{poi}(n)}_{c\cdot i^{-\alpha}}) \ge \sum_{i>\ell_0} \hat{R}(\mathcal{POI}_{nf_i}) = \sum_{\ell<i\le b} \hat{R}(\mathcal{POI}_{nf_i}) + \sum_{i>b} \hat{R}(\mathcal{POI}_{nf_i}),$$

where recall that $b = (cn)^{1/\alpha}$.

For the first term, using Lemma 17,

$$\sum_{\ell_0<i\le b} \hat{R}(\mathcal{POI}_{nf_i}) \ge \sum_{\ell_0<i\le b} \log \sqrt{\frac{2nf_i + 2}{\pi}} \ge \frac{1}{2} \sum_{\ell_0<i\le b} \log \frac{2nc}{\pi i^\alpha} > \sum_{1\le i\le b} \log \frac{nc}{i^\alpha} - \ell_0 \log(nc) - \frac{b}{2} \log \frac{\pi}{2}.$$

Again by Stirling approximation, and using $B! < 2\pi B(B/e)^B$,

$$\sum_{1\le i\le b} \log \frac{nc}{i^\alpha} = \log \frac{(cn)^b}{(b!)^\alpha} \ge \frac{\alpha}{2}(cn)^{1/\alpha} \log e - \alpha \log(2\pi b)$$

For the second term, using $2 - e^{-\lambda} > 1 + \lambda/2$ for $\lambda < 1$,

$$\sum_{i>b} \hat{R}(\mathcal{POI}_{nf_i}) = \sum_{i>b} \log(2 - e^{-nf_i}) > \sum_{i>b} \frac{nf_i}{2} \log e > \frac{cn}{2} \sum_{i>b} \frac{1}{i^\alpha} > (cn)^{1/\alpha} \frac{1}{2(\alpha - 1)} - O(1).$$

Summing these two lower bounds $\hat{R}(\mathcal{E}^{\text{poi}(n)}_{c\cdot i^{-\alpha}})$. Invoking the lower bound of Theorem 14 with the upper bound from the previous part proves the lower bound. We hide the lower order and logarithmic terms in the $o(1)$ factor. □

Remark 20. *From a simple calculation, it follows that the two bounds are at most a factor of 4 apart. By obtaining tighter bounds on $\hat{R}(\mathcal{POI}_\Lambda)$, it should be possible to obtain upper and lower bounds within a multiplicative $(1 + \delta)$ factor for arbitrarily small δ.*

10 Exponential Envelopes

We provide a simple proof of Equation (2) with stronger second order terms. More precisely, we prove that

Theorem 21.

$$\hat{R}(\mathcal{E}^n_{c\cdot e^{\alpha i}}) = \frac{\log^2 n}{4\alpha} + O(\log c \log n).$$

Proof. For the exponential class, $f_i = ce^{-\alpha i}$. Therefore,

$$i \le \frac{\ln(cn)}{\alpha} \quad \Leftrightarrow \quad nf_i \ge 1.$$

Similar to the argument for power-law, let $b \overset{\text{def}}{=} \frac{\ln(cn)}{\alpha}$, and by Theorem 18 and Lemma 17,

$$\hat{R}(\mathcal{E}^{\text{poi}(n)}_{c\cdot e^{\alpha i}}) \le \sum_{i\le b}\hat{R}(\mathcal{POI}_{nf_i}) + \sum_{i>b}\hat{R}(\mathcal{POI}_{nf_i}) \le \sum_{1\le i\le b}\log\left(2+\sqrt{\frac{2nf_i}{\pi}}\right) + (\sum_{i>b} nf_i)\log e$$

Using $e^{b\alpha} = cn$, the second term can be bounded using

$$\sum_{i>b} nf_i = \sum_{i>b} cne^{-\alpha i} < cne^{-\alpha b}\frac{1}{1-e^{-\alpha}} = \frac{1}{1-e^{-\alpha}}.$$

Following the same steps as power-law and using $e^{b\alpha} = n$,

$$\begin{aligned}
\sum_{i=1}^{b}\log\left(2+\sqrt{\frac{2nf_i}{\pi}}\right) &\le \sum_{i=1}^{b}\log\left(3\sqrt{nf_i}\right)\\
&\le b\log 3 + \frac{1}{2}\sum_{i=1}^{b}\log[n\cdot ce^{-\alpha i}]\\
&\le b\log 3 + \frac{1}{2}\log[(cn)^b\cdot(cn)^{-(b+1)/2}]\\
&= b\log 3 + \frac{(b-1)}{4}\log(cn)\\
&< \frac{b}{4}\log(81c) + \frac{b}{4}\log n.
\end{aligned}$$

Substituting $b = \ln(cn)/\alpha$,

$$\hat{R}(\mathcal{E}^{n}_{c\cdot e^{\alpha i}}) \le \frac{\log^2 n}{4\alpha\log e} + \frac{\log cn\log(81c)}{4\alpha\log e} + \frac{\log n\log c}{4\alpha\log e} + \frac{\log e}{1-e^{-\alpha}} + 1.$$

We now prove the lower bound by a similar argument. Clearly, for $\ell \ge \ell_0 \overset{\text{def}}{=} \frac{1}{\alpha}\log(\frac{c}{1-e^{-\alpha}})$,

$$\sum_{i=\ell}^{\infty} f_i \le 1.$$

Recall that $b = \ln(cn)/\alpha$, then by Lemma 17,

$$\begin{aligned}
\hat{R}(\mathcal{E}^{\text{poi}(n)}_{c\cdot e^{\alpha i}}) &\ge \sum_{i=l_0}^{b}\log\left(\sqrt{\frac{2nf_i+2}{\pi}}\right)\\
&\ge \frac{1}{2}\Big[\sum_{i=1}^{b}\log\frac{2nf_i}{\pi} - \sum_{i=1}^{\ell_0}\log\frac{2nf_i}{\pi}\Big]\\
&\ge \frac{1}{2}\Big[b\log\frac{2}{\pi} + \log[(cn)^b\cdot(cn)^{-(b+1)/2}] - \sum_{i=1}^{\ell_0}\log\frac{2nf_i}{\pi}\Big]\\
&\ge \frac{1}{2}\Big[b\log\frac{2}{\pi} + \frac{b-1}{2}\log(cn) - \sum_{i=1}^{\ell_0}\log\frac{2nf_i}{\pi}\Big]\\
&\ge \frac{b-1}{4}\log\frac{4cn}{\pi^2} - \frac{\ell_0}{4}\log\frac{4cn}{\pi^2}\\
&\ge \frac{\log^2 n}{4\alpha\log e} - O(\log c\log n).
\end{aligned}$$

To obtain a bound on fixed length sampling, we apply Theorem 14 with the upper bound proved earlier. □

11 A note on expected redundancy

Since the worst case redundancy is always larger than the average redundancy, a class with infinite average redundancy has infinite worst case redundancy. However, there exist classes of distributions with a finite average redundancy and infinite worst case redundancy (See Example 1 in [1]).

However, in Lemma 22 we show that an envelope class with infinite worst case redundancy also has infinite average case redundancy. On a related note, [6] showed that for any distribution class with finite worst case redundancy, the worst case redundancy grows sub-linearly with n, namely if $\hat{R}(\mathcal{P}) < \infty$, then $\hat{R}(\mathcal{P}) = o(n)$. Recently, [13] showed that there exists a class $\mathcal{P}$ such that $\overline{R}(\mathcal{P}) < \infty$, and yet, $\overline{R}(\mathcal{P}^n) = \Omega(n)$, *i.e.,* the redundancy grows linearly with the block length.

Lemma 22. *For any envelope function f,*

$$\hat{R}(\mathcal{E}_f) < \infty \Leftrightarrow \overline{R}(\mathcal{E}_f) < \infty.$$

Our proof of this lemma uses the following result.

Lemma 23. *If $\mathcal{P}$ contains M distributions over mutually disjoint supports, then $\overline{R}(\mathcal{P}) \geq \ln M$.*

Proof. Let $P_1, \ldots, P_M$ be M distributions over disjoint sets $\mathcal{A}_1, \ldots, \mathcal{A}_M$. For any distribution Q, there is a j, such that $Q(\mathcal{A}_j) \leq 1/M$. For that distribution,

$$D(P_j\|Q) = \sum_{x\in\mathcal{A}_j} P_j(x) \ln \frac{P(x)}{Q(x)} \geq -\ln\left(\sum_{x\in\mathcal{A}_j} Q(x)\right) \geq \ln M,$$

where the inequality is from the convexity of logarithms. Since this holds for any distribution Q, it holds for the infimum and plugging in the definition of $\overline{R}$ proves the result. □

Proof of Lemma 22. The forward direction is trivial since $\overline{R} \leq \hat{R}$. For the other direction, we show that if $\hat{R}(\mathcal{P}) = \infty$, there exist an infinite number of distributions $P_1, P_2, \ldots$ that all have disjoint supports, and applying the previous lemma to this proves the result. Now, $\hat{R}(\mathcal{P}) = \infty$ means $\sum f_i = \infty$. Using this we construct an infinite sequence of distributions as follows. After constructing distribution P_i consider the first integer not in its support, and construct a distribution over the smallest possible integers starting from that location. This process can be repeated infinitely many times giving an infinite number of distributions with disjoint supports. □

References

[1] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Tight bounds for universal compression of large alphabets. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 2875–2879, 2013.

[2] Jayadev Acharya, Hirakendu Das, and Alon Orlitsky. Tight bounds on profile redundancy and distinguishability. In *Neural Information Processing Systems (NIPS)*, 2012.

[3] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Efficient compression of monotone and m-modal distributions. In *To appear in Proceedings of IEEE Symposium on Information Theory*, 2014.

[4] Dominique Bontemps. Universal coding on infinite alphabets: Exponentially decreasing envelopes. *IEEE Transactions on Information Theory*, 57(3):1466–1478, 2011.

[5] Dominique Bontemps, Stéphane Boucheron, and Elisabeth Gassiat. About adaptive coding on countable alphabets. *IEEE Transactions on Information Theory*, 60(2):808–821, 2014.

[6] Stéphane Boucheron, Aurelien Garivier, and Elisabeth Gassiat. Coding on countably infinite alphabets. *IEEE Transactions on Information Theory*, 55(1):358–373, 2009.

[7] Stephane Boucheron, Elisabeth Gassiat, and Mesrob I. Ohannessian. About adaptive coding on countable alphabets: Max-stable envelope classes. *CoRR*, abs/1402.6305, 2014.

[8] Thomas M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, January 1991.

[9] Thomas M. Cover and Joy Thomas. *Elements of Information Theory, 2nd Ed.* Wiley Interscience, 2006.

[10] Lee D. Davisson. Universal noiseless coding. *IEEE Transactions on Information Theory*, 19(6):783–795, Nov. 1973.

[11] Lee D. Davisson, Robert J. McEliece, Michael B. Pursley, and Mark S. Wallace. Efficient universal noiseless source codes. *IEEE Transactions on Information Theory*, 27(3):269–279, 1981.

[12] Dean P. Foster, Robert A. Stine, and Abraham J. Wyner. Universal codes for finite sequences of integers drawn from a monotone distribution. *IEEE Transactions on Information Theory*, 48(6):1713–1720, June 2002.

[13] Maryam Hosseini and Narayana Santhanam. On redundancy of memoryless sources over countable alphabets. *CoRR*, abs/1404.0062, 2014.

[14] John C. Kieffer. A unified approach to weak universal source coding. *IEEE Transactions on Information Theory*, 24(6):674–682, Nov. 1978.

[15] Michael Mitzenmacher and Eli Upfal. *Probability and computing - randomized algorithms and probabilistic analysis*. Cambridge Univ. Press, 2005.

[16] Alon Orlitsky and Narayana Prasad Santhanam. Speaking of infinity [i.i.d. strings]. *IEEE Transactions on Information Theory*, 50(10):2215 – 2230, oct 2004.

[17] Alon Orlitsky, Narayana Prasad Santhanam, and Junan Zhang. Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, 50(7):1469– 1481, July 2004.

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

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

[20] Yuri M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3–17, 1987.

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

[22] Wojciech Szpankowski and Marcelo J. Weinberger. Minimax redundancy for large alphabets. In *ISIT*, pages 1488–1492, 2010.

[23] Frans M. J. Willems, Yuri M. Shtarkov, and Tjalling J. Tjalkens. The context-tree weighting method: basic properties. *IEEE Transactions on Information Theory*, 41(3):653–664, 1995.

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

[25] Xiao Yang and Andrew R. Barron. Large alphabet coding and prediction through poissonization and tilting. In *Workshop on Information Theoretic Methods in Science and Engineering*, 2013.