Tight Bounds for Universal Compression of Large Alphabets

Jayadev Acharya
ECE, UCSD
jacharya@ucsd.edu

Hirakendu Das
Yahoo!
hdas@yahoo-inc.com

Ashkan Jafarpour
ECE, UCSD
ashkan@ucsd.edu

Alon Orlitsky
ECE & CSE, UCSD
alon@ucsd.edu

Ananda Theertha Suresh
ECE, UCSD
asuresh@ucsd.edu

***Abstract*—Over the past decade, several papers, e.g., [1–7] and references therein, have considered universal compression of sources over large alphabets, often using patterns to avoid infinite redundancy. Improving on previous results, we prove tight bounds on expected- and worst-case pattern redundancy, in particular closing a decade-long gap and showing that the worst-case pattern redundancy of i.i.d. distributions is $\tilde{\mathbf{\Theta}}(\mathbf{n^{1/3}})^{\dagger}$.**

I. Introduction

Every distribution P can be compressed to its entropy $H(P)$ and no further. This optimal compression rate is achieved by assigning to each symbol x a codeword of length $\sim \log \frac{1}{P(x)}$.

Yet in most cases, the source distribution is unknown, except that it can be assumed to belong to a known distribution class $\mathcal{P}$. For example the class of all *i.i.d.*, or Markov, distributions.

Compression schemes designed for all distributions in a class are called *universal*. Their performance is measured in terms of *redundancy*, the largest number of bits beyond $\log \frac{1}{P(x)}$, or $H(P)$, that they use to compress distributions in $\mathcal{P}$. The lowest possible redundancy of any compression scheme is the *redundancy* of $\mathcal{P}$ and determines how well unknown distributions in $\mathcal{P}$ can be universally compressed.

More concretely, observe that every compression scheme for an alphabet $\mathcal{X}$ corresponds to a distribution Q over $\mathcal{X}$ where the number of bits the scheme assigns to $x \in \mathcal{X}$ is roughly $\log \frac{1}{Q(x)}$. The extra number of bits the scheme uses to encode x when the underlying distribution is P is therefore $\log \frac{P(x)}{Q(x)}$.

Let $\mathcal{P}$ be a class of distributions over $\mathcal{X}$. Two measures for the redundancy of $\mathcal{P}$ have been extensively studied [8, 9], and both play an important role in compression algorithms.

The *expected-case redundancy* of $\mathcal{P}$ is

$$\overline{R}(\mathcal{P}) \stackrel{\text{def}}{=} \min_{Q} \max_{P\in\mathcal{P}} \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$$
$$= \min_{Q} \max_{P\in\mathcal{P}} \left(\mathbb{E}_P \log \frac{1}{Q(X)} - H(P) \right),$$

the lowest possible increase, over all compression schemes Q, of the extra number of bits beyond the entropy, that Q uses to compress the worst distribution $P \in \mathcal{P}$.

Similarly, the *worst-case redundancy* of $\mathcal{P}$ is

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

the lowest possible increase, over all compression schemes Q, of the extra number of bits above $\log \frac{1}{P(x)}$ that Q uses for the worst distribution $P \in \mathcal{P}$ and the worst symbol $x \in \mathcal{X}$.

$^{\dagger} f(n) = \tilde{\Theta}(g(n))$ if the functions differ by a poly-logarithmic factor.

As evident from the definition, $\hat{R}(\mathcal{P}) \geq \overline{R}(\mathcal{P})$, hence low worst-case redundancy is a stronger performance guarantee than low expected-case redundancy. It ensures that a universal-compression scheme will be close to optimal not only on average, but for all possible outcomes.

Another interpretation, and strong motivation, for redundancy is as a measure for the quality of *prediction algorithms* [10, 11]. A prediction algorithm Q observes the output $X_1, X_2, \ldots$ of an unknown random process P, and at each time i, having observed $X^i \stackrel{\text{def}}{=} X_1, \ldots, X_i$, outputs a distribution $Q(x_{i+1}|X^i)$ over the possible values of X_{i+1}. The most common measure for the performance of prediction algorithms is their *cumulative log loss*,

$$\sum_{i=1}^{n} \log \frac{P(X_{i+1}|X^i)}{Q(X_{i+1}|X^i)}.$$

It is not difficult to see that for every n, the expected- and worst-case redundancy are exactly the expected and worst-case cumulative log-loss of the best prediction algorithm.

The most extensively studied classes of distributions are $\mathcal{I}_k^n$, the collections of all length-n *i.i.d.* distributions over an alphabet of size k. A string of works [12–18] determined the redundancy of $\mathcal{I}_k^n$ up to a diminishing additive term, in particular showing that

$$\overline{R}(\mathcal{I}_k^n) + C_1(k) = \hat{R}(\mathcal{I}_k^n) + C_2(k) = \frac{k-1}{2} \log n + o_n(1),$$

where $C_1(k)$ and $C_2(k)$ are known functions of k, independent of n. In particular this shows that while $\overline{R}(\mathcal{I}_k^n)$ and $\hat{R}(\mathcal{I}_k^n)$ grow with the block-length n, they are always within a (very small) constant from each other.

As the above equation shows, while the redundancy of $\mathcal{I}_k^n$ grows logarithmically with the length n, it grows linearly with the alphabet size k. In many practical applications, including those involving natural language processing [19, 20], the alphabet size is very large, often even larger than the block length. Hence the redundancy may be correspondingly high.

To address this fast increase in redundancy with the alphabet size, a new approach was proposed for compression and estimation over large alphabets. The *pattern* [1] of a sequence represents the relative order in which its symbols appear. For example, the pattern of *abracadabra* is 12314151231. A natural method for compressing a sequence over a large alphabet is to compress its pattern as well as the *dictionary* that maps the order to the original symbols. For example, for abracadabra, $1 \to a$, $2 \to b$, $3 \to r$, $4 \to c$, $5 \to d$.

Let $\mathcal{I}^n_{\bar\psi}$ be the class of all distributions over length-n patterns induced by all *i.i.d.* distributions, over any number of symbols. It was shown in [1, 21] that

$$\left(\frac{3}{2}\log_2 e\right)n^{1/3} \le \hat R(\mathcal{I}^n_{\bar\psi}) \le \left(\pi\sqrt{\frac{2}{3}}\right)n^{1/2}. \qquad (1)$$

It follows that patterns can be compressed with $\le \frac{1}{\sqrt{n}}$per-symbol, hence diminishing redundancy, regardless of the alphabet size.

This result also upper bounds expected redundancy. Subsequently, [22] described a proof-outline that could potentially show the following tighter upper bound on expected redundancy, and [23] proved the following lower bound, strengthening one in [6], and extensions appeared in [4],

$$1.84\left(\frac{n}{\log n}\right)^{1/3} \le \overline{R}(\mathcal{I}^n_{\bar\psi}) \le n^{0.4}.$$

Recently, [24] improved these results and determined the polynomial growth rate of expected pattern redundancy,

$$0.3\cdot n^{1/3} \le \overline{R}(\mathcal{I}^n_{\bar\psi}) \le n^{1/3}(\log n)^2.$$

II. Results

As we saw, for *i.i.d.* distributions, expected- and worst-case redundancies are very close. Yet for general distributions they may differ greatly. Example 1 shows a distribution class $\mathcal{P}$ where $\overline{R}(\mathcal{P})$ is very small while $\hat R(\mathcal{P})$ is infinite.

We therefore consider the worst-case pattern redundancy of *i.i.d.* distributions. Improving on (1), we establish a tight upper bound on the polynomial growth rate,

$$\hat R(\mathcal{I}^n_{\bar\psi}) \le n^{1/3}(\log n)^4.$$

Combined with the lower bound in (1), this determines $\hat R(\mathcal{I}^n_{\bar\psi})$ up to a poly-logarithmic factor,

$$\hat R(\mathcal{I}^n_{\bar\psi}) = \tilde\Theta(n^{1/3}).$$

We also improve on [24] and show that $\overline{R}(\mathcal{I}^n_{\bar\psi}) \le n^{1/3}(\log n)^{4/3}$. Due to lack of space, we only prove the bound on worst case pattern redundancy.

The paper is organized as follows. In Section III-A we state a few general properties of worst-case redundancy. In section III-B we define patterns and profiles, and in Section III-C we consider the method of Poisson sampling that is used to simplify the analyses. In Section IV-A we provide an overview of the proof and give the details in the final section.

III. Preliminaries

A. Worst-Case Redundancy

Let $\mathcal{P}$ be a class of distributions over $\mathcal{X}$. For $x\in\mathcal{X}$, let $\hat P(x) = \sup_{P\in\mathcal{P}} P(x)$, and let

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

be the Shtarkov sum of $\mathcal{P}$. It can be shown [25] that

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

The following example presents a class of distributions over the natural numbers, such that $\overline{R}$ is finite but $\hat R$ is infinite.

Example 1. *Let $\mathcal{X}=\mathbb{N}$ and $\mathcal{P}=\{P_1,P_2,\dots,\}$, where P_j is a distribution over $\{1,j\}$ assigning probability $1-\frac{1}{j}$ to 1 and $\frac{1}{j}$ to j. Then, $S(\mathcal{P})=\sum_{j\ge1}\frac{1}{j}=\infty$, and hence $\hat R(\mathcal{P})=\infty$. Let $Q=\frac{s}{i^2}$ be a distribution over $\mathbb{N}$, where $s=\frac{6}{\pi^2}$. Then for any j, $D(P_j||Q)=(1-\frac{1}{j})\log\frac{1-\frac{1}{j}}{s}+\frac{1}{j}\log\frac{j}{s}<\log\frac{1}{s}+1$. Hence, $\overline{R}(\mathcal{P})<1-\log s$.*

We now state a few properties of $\hat R(\mathcal{P})$. For a distribution P over $\mathcal{X}$ and a function $f:\mathcal{X}\to\mathcal{Y}$, let $f(P)$ be the distribution over $\mathcal{Y}=f(\mathcal{X})$ that assigns to $y\in\mathcal{Y}$ the probability $P(f^{-1}(y))$. For a collection $\mathcal{P}$ of distributions over $\mathcal{X}$, let $f(\mathcal{P})=\{f(P):P\in\mathcal{P}\}$.

Lemma 2 (Redundancy of functions)**.** $\hat R(f(\mathcal{P}))\le\hat R(\mathcal{P})$.

Proof:
$$S(\mathcal{P}) = \sum_{x\in\mathcal{X}}\hat P(x) = \sum_{y\in\mathcal{Y}}\sum_{x\in f^{-1}(y)}\hat P(x)$$
$$\ge \sum_{y\in\mathcal{Y}}\sup_{P\in f(\mathcal{P})} P(y) = S(f(\mathcal{P})).$$

Taking logarithm yields the result. ∎

For a class $\mathcal{P}$ consisting of product (independent) distributions over $\mathcal{X}\times\mathcal{Y}$, let $\mathcal{P}_\mathcal{X}$ and $\mathcal{P}_\mathcal{Y}$ be the class of marginals. The redundancy of $\mathcal{P}$ is at most the sum of the marginal redundancies.

Lemma 3 (Redundancy of products)**.** *For a collection $\mathcal{P}$ of product distributions over $\mathcal{X}\times\mathcal{Y}$,*

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

Proof: $\mathcal{S}(\mathcal{P}) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\sup_{(P,Q)\in\mathcal{P}} P(x)Q(y)$

$$\le \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\sup_{P\in\mathcal{P}_\mathcal{X}} P(x)\sup_{Q\in\mathcal{P}_\mathcal{Y}} Q(y)$$
$$\le \sum_{x\in\mathcal{X}}\sup_{P\in\mathcal{P}_\mathcal{X}} P(x)\sum_{y\in\mathcal{Y}}\sup_{Q\in\mathcal{P}_\mathcal{Y}} Q(y) \le \mathcal{S}(\mathcal{P}_\mathcal{X})\mathcal{S}(\mathcal{P}_\mathcal{Y}).$$ ∎

The next lemma relates the redundancy of union of classes of distributions to the individual redundancies.

Lemma 4 (Redundancy of unions)**.** *If $\mathcal{P}_1,\dots,\mathcal{P}_T$ are distribution collections, then*

$$\hat R(\bigcup_{1\le i\le k}\mathcal{P}_i) \le \max_{1\le i\le T}\hat R(\mathcal{P}_i) + \log T.$$

Proof: $\sum_x \sup_{\cup\mathcal{P}_i} P(x) \le T\times\max_{i\in[T]}\sum_x\sup_{\mathcal{P}_i}P(x)$. ∎

B. Patterns and Profiles

Recall that the *pattern* $\bar\psi$ of a sequence is the sequence of integers obtained by replacing each symbol by its order of appearance. For example, the length 4 sequence $isit$ has pattern 1213. The probability of a pattern $\bar\psi$ is the probabilities of all sequences whose pattern is $\bar\psi$,

$$P(\bar\psi) = \sum_{\overline{x}:\bar\psi(\overline{x})=\bar\psi} P(\overline{x}).$$

For example, $P(1213)=P(isit)+P(alan)+\dots$.

The *multiplicity* $\mu(x)$ of a symbol x in a sequence is the number of times it appears. The *profile* $\overline{\varphi}$ of a sequence is

the multiset of multiplicities of all symbols appearing [1, 26]. For example, the sequence $ababcde$ has multiplicities $\mu(a) = \mu(b) = 2$, $\mu(c) = \mu(d) = \mu(e) = 1$, and profile $\{1, 1, 1, 2, 2\}$. The length of a profile is the length of sequence generating it (hence the sum of all the multiplicities). Each profile of length n, is a partition of n. Let Φ_n denote all profiles of length n, and $\Phi = \cup_n \Phi_n$. Similar to patterns, the probability of a profile is the sum of probabilities of sequences with that profile.

Similar to $\mathcal{I}^n_{\bar{\psi}}$, let $\mathcal{I}^n_{\overline{\varphi}}$ be the class of all distributions over length-n profiles induced by all *i.i.d.* distributions, sampled n times. Since any *i.i.d.* distribution assigns the same probability to all patterns with the same profile, it can be shown [1] that

$$\hat{R}(\mathcal{I}^n_{\bar{\psi}}) = \hat{R}(\mathcal{I}^n_{\overline{\varphi}}). \tag{2}$$

C. Poisson Sampling and Profile Probability

When a distribution is sampled *i.i.d.* exactly n times, the multiplicities are dependent, *e.g.,* they add up to n. A standard approach [27] to overcome the dependence is to sample the distribution $\mathrm{poi}(n)$ times, where $\mathrm{poi}(n)$ is a Poisson random variable with parameter n. With high probability, the resulting sequences have their random length close to n. We let $\mathrm{poi}(\lambda, \mu) \stackrel{\mathrm{def}}{=} e^{-\lambda}\lambda^\mu/\mu!$ denote the probability that a $\mathrm{poi}(\lambda)$ random variable equals μ.

The following basic properties of Poisson sampling help simplify the analysis and relate it to fixed-length sampling.

Lemma 5. *[27] If a discrete distribution is sampled* $\mathrm{poi}(n)$ *times then: (1) the number of appearances of different symbols are independent; (2) a symbol with probability p appears* $\mathrm{poi}(np)$ *times; (3) for any fixed n_0, conditioned on the length* $\mathrm{poi}(n) \geq n_0$*, the first n_0 elements are distributed identically to sampling P exactly n_0 times.*

We now express profile probabilities and redundancy under Poisson sampling. The probability multiset of a distribution over $\mathcal{X}$ is the collection $\{P(x) : x \in \mathcal{X}\}$ of probabilities. For example, the probability multiset of $P(a) = P(c) = .4, P(b) = .2$ is $\{.4, .4, .2\}$. Since relabeling symbols in a sequence does not change its profile, distributions with the same probability multiset assign the same probability to any profile [26].

For a distribution P with multiset $\{p_1, p_2, \ldots\}$, let $\lambda_i \stackrel{\mathrm{def}}{=} np_i$, and $\Lambda \stackrel{\mathrm{def}}{=} \{\lambda_1, \lambda_2, \ldots\}$. The profile generated by $\mathrm{poi}(n)$ sampling of P is a multiset $\overline{\varphi} = \{\mu_1, \mu_2, \ldots\}$, where each μ_i is generated independently according to $\mathrm{poi}(\lambda_i)$.

Similarly, it is useful to consider profiles (set of multiplicities) generated by $\Lambda' \subseteq \Lambda$. For example, see IV-B. Henceforth, distributions are denoted by the multiset Λ instead of P.

The probability that Λ generates $\overline{\varphi}$ is [1, 28],

$$\Lambda(\overline{\varphi}) = \frac{1}{\prod_{\mu=0}^{\infty} f_\mu!} \sum_\sigma \prod_i \mathrm{poi}(\lambda_{\sigma(i)}, \mu_i). \tag{3}$$

where the summation is over permutations of the support set, and f_μ is the number of elements with multiplicity μ in $\overline{\varphi}$.

For example, for $\Lambda = \{\lambda_1, \lambda_2, \lambda_3\}$, the profile $\overline{\varphi} = \{2, 2, 3\}$ has $f_2 = 2$, and $f_3 = 1$, and and there are three possible ways to assign the multiplicities to the symbols. This is reflected by the $f_2!$ terms in the denominator, ensuring that only three out of the six permutation terms are taken into account.

Let $\mathcal{I}^{\mathrm{poi}(n)}_{\overline{\varphi}}$ be the class of all distributions induced on profiles (of all lengths, denoted Φ) by *i.i.d.* distributions under $\mathrm{poi}(n)$ sampling. Using Lemma 5 and the fact that $\mathrm{poi}(n, n) \sim \frac{1}{\sqrt{n}}$

$$\hat{R}(\mathcal{I}^{\mathrm{poi}(n)}_{\overline{\varphi}}) > \hat{R}(\mathcal{I}^n_{\overline{\varphi}}) - \frac{1}{2}\log n.$$

Hence, upper bounds on $\hat{R}(\mathcal{I}^{\mathrm{poi}(n)}_{\overline{\varphi}})$ provide bounds for $\hat{R}(\mathcal{I}^n_{\overline{\varphi}})$. Henceforth, we consider only Poisson sampling. Also, note that $\hat{R}(\mathcal{I}^{\mathrm{poi}(n)}_{\bar{\psi}}) = \hat{R}(\mathcal{I}^{\mathrm{poi}(n)}_{\overline{\varphi}})$ and hence considering profiles suffices.

IV. Proof

A. Overview

We strengthen the arguments in [24] to show the main result. Notice that a distribution in $\mathcal{I}^{\mathrm{poi}(n)}_{\overline{\varphi}}$ is a collection of positive reals that sum to n. For each distribution we divide the collection of λ's into three sub-collections, those $\leq n^{1/3}$, between $n^{1/3}$ and $n^{2/3}$, and those $> n^{2/3}$. In Lemma 6 we show that it suffices to consider distributions that have all λ's in the middle range, namely in $(n^{1/3}, n^{2/3}]$.

We then partition such distributions into T_n classes $\mathcal{I}(1), \ldots, \mathcal{I}(T_n)$. The classes are designed such that $\log(T_n) < \tilde{\mathcal{O}}(n^{1/3})$, and the redundancy of each class is at most $\tilde{\mathcal{O}}(n^{1/3})$. The result then follows by Lemma 4.

B. Details

Each distribution Λ in $\mathcal{I}^{\mathrm{poi}(n)}_{\overline{\varphi}}$ is a collection of λ's that sum to n. For any such distribution, let

$$\Lambda_{\mathrm{low}} \stackrel{\mathrm{def}}{=} \{\lambda \in \Lambda : \lambda \leq n^{1/3}\},$$
$$\Lambda_{\mathrm{med}} \stackrel{\mathrm{def}}{=} \{\lambda \in \Lambda : n^{1/3} < \lambda \leq n^{2/3}\},$$
$$\Lambda_{\mathrm{high}} \stackrel{\mathrm{def}}{=} \{\lambda \in \Lambda : \lambda > n^{2/3}\},$$

and let $\overline{\varphi}_{\mathrm{low}}, \overline{\varphi}_{\mathrm{med}}, \overline{\varphi}_{\mathrm{high}}$ denote the corresponding profile each subset generates. Let $\mathcal{I}_{\overline{\varphi}_{\mathrm{low}}}$, $\mathcal{I}_{\overline{\varphi}_{\mathrm{med}}}$ and $\mathcal{I}_{\overline{\varphi}_{\mathrm{high}}}$ be the collection of all Λ_{low}, Λ_{med}, and Λ_{high}'s respectively, for all $\Lambda \in \mathcal{I}^{\mathrm{poi}(n)}_{\overline{\varphi}}$.

By Poisson sampling, $\overline{\varphi}_{\mathrm{low}}$, $\overline{\varphi}_{\mathrm{med}}$ and $\overline{\varphi}_{\mathrm{high}}$ are independent, and $\overline{\varphi} = \overline{\varphi}_{\mathrm{low}} \cup \overline{\varphi}_{\mathrm{med}} \cup \overline{\varphi}_{\mathrm{high}}$. By Lemmas 2 and 3,

$$\hat{R}(\mathcal{I}^{\mathrm{poi}(n)}_{\overline{\varphi}}) \leq \hat{R}(\mathcal{I}_{\overline{\varphi}_{\mathrm{low}}}) + \hat{R}(\mathcal{I}_{\overline{\varphi}_{\mathrm{med}}}) + \hat{R}(\mathcal{I}_{\overline{\varphi}_{\mathrm{high}}}). \tag{4}$$

The following lemma bounds $\hat{R}(\mathcal{I}_{\overline{\varphi}_{\mathrm{low}}})$ and $\hat{R}(\mathcal{I}_{\overline{\varphi}_{\mathrm{high}}})$.

Lemma 6. *Both* $\hat{R}(\mathcal{I}_{\overline{\varphi}_{\mathrm{low}}}), \hat{R}(\mathcal{I}_{\overline{\varphi}_{\mathrm{high}}}) < 4n^{1/3}\log n$.

Proof: Any distribution in $\mathcal{I}_{\overline{\varphi}_{\mathrm{high}}}$ is a collection of λ's, each at least $n^{2/3}$ that sum to at most n. Hence, any distribution in $\mathcal{I}_{\overline{\varphi}_{\mathrm{high}}}$ consists of at most $n^{1/3}$ elements. For a profile $\overline{\varphi}_{\mathrm{high}}$, let $\mu_{\max}$ be the largest multiplicity in it. Elias coding [29] encodes every positive integer j using at most $2\log(j+1)$ bits. Hence, $\overline{\varphi}_{\mathrm{high}}$ can be encoded using $\leq n^{1/3} \times 2\log(\mu_{\max} + 1)$ bits. Since redundancy is the extra number of bits needed to encode a profile, it is upper bounded by the code-length. Hence,

$$\log\Big(\sum_{\overline{\varphi}:\mu_{max}=\mu} \hat{P}(\overline{\varphi})\Big) \le 2n^{1/3}\log(\mu+1).$$

Since every distribution in $\Lambda_{\rm high}$ has all $\lambda \le n$, probability of profiles with $\mu_{\max} > n$ falls exponentially with μ_{max}. Using this argument, it can then be shown that, $\hat{R}(\mathcal{I}_{\overline{\varphi}_{\rm high}}) \le 4n^{1/3}\log n$. $\hat{R}(\mathcal{I}_{\overline{\varphi}_{\rm low}})$ can be bounded similarly. ∎

We bound $\hat{R}(\mathcal{I}_{\overline{\varphi}_{\rm med}})$ by dividing $\mathcal{I}_{\overline{\varphi}_{\rm med}}$ into classes $\mathcal{I}(1),\ldots,\mathcal{I}(T_n)$ and bounding the redundancy of each class.

Consider any partition of $(n^{1/3}, n^{2/3}]$ into $b \stackrel{\rm def}{=} n^{1/3}$ consecutive intervals (bins) $B_1, B_2, \ldots, B_b$ of lengths $\Delta_1,\ldots,\Delta_b$. For each distribution $\Lambda \in \mathcal{I}_{\overline{\varphi}_{\rm med}}$, let $\Lambda_j \stackrel{\rm def}{=} \Lambda \cap B_j$ be the set of elements of Λ in B_j, let $m_j \stackrel{\rm def}{=} m_j(\Lambda) \stackrel{\rm def}{=} |\Lambda_j|$ be the number of elements of Λ in B_j and let $s_j \stackrel{\rm def}{=} s_j(\Lambda) \stackrel{\rm def}{=} \sum_{\lambda\in\Lambda_j}\lambda$ be the sum of all elements in Λ_j. $\overline{m}(\Lambda)$ and $\overline{s}(\Lambda)$ denote the b-tuples of m_j's and s_j's respectively.

Lemma 7. *$\mathcal{I}_{\overline{\varphi}_{\rm med}}$ can be partitioned into $T_n \le n^{100b}$ classes such that in each class, all Λ's have the same $\overline{m}(\Lambda)$ and all the $\overline{s}(\Lambda)$'s are within ℓ_1 distance $< 1/n^{99}$ from each other.*

Proof Sketch: Each $\overline{m}(\Lambda)$ is a $b-$tuple of natural numbers, each at most $n^{2/3}$. Hence, the number of possible $\overline{m}(\Lambda)$'s is at most $(n^{2/3})^b$. Similarly, quantizing $\overline{s}(\Lambda)$ in units of $\frac{1}{n^{99}}$ proves the lemma. ∎

By Lemma 4,

$$\hat{R}(\mathcal{I}_{\overline{\varphi}_{\rm med}}) \le \max_{1\le i\le T_n} \hat{R}(\mathcal{I}(i)) + 100b\log n. \tag{5}$$

We now bound the redundancy of each class. Let $\mathcal{I} \in \{\mathcal{I}(i) : 1 \le i \le T_n\}$ be any one of the classes, and let $\overline{m} = \overline{m}(\Lambda)$ for all $\Lambda \in \mathcal{I}$. Let $B_j = (\lambda_{j-1}, \lambda_j]$ be the jth interval. The following theorem, which holds when $\Delta_j \le \sqrt{\lambda_j}\log n$, bounds the redundancy as a function of $\overline{m}$, and the intervals. The proof is deferred to Section IV-C.

Theorem 8. $\hat{R}(\mathcal{I}) \le \frac{3b}{2} + 2(\log n)^2\Big(\sum_{j=1}^b \frac{m_j\Delta_j^2}{\lambda_j}\Big)$.

Using this theorem we bound $\hat{R}$ of profiles by choosing B_j's that bound the expression in the theorem. The interval size Δ_j's are chosen to be geometrically growing, namely $\Delta_j = n^{1/3}c(1+c)^{j-1}$, where c is chosen to ensure that the sum of the b intervals is $n^{2/3} - n^{1/3}$. Summing this geometric series, $n^{2/3} = n^{1/3} + \Delta_1 + \ldots + \Delta_b = n^{1/3}(1+c)^b \Rightarrow (1+c)^b = n^{1/3}$. Since $b = n^{1/3}$, the condition $b\log(1+c) = \log n/3$ yields $c < 0.4\log n/b$. By the construction of the geometric series, $\Delta_j = \lambda_{j-1}c$. Since $\lambda_j \le n^{2/3}$, $\Delta_j \le \sqrt{\lambda_j}\log n$ holds for all j, and the theorem holds. Plugging the expression for Δ_j,

$$\sum_{j=1}^b \frac{m_j\Delta_j^2}{\lambda_j} = \sum_{j=1}^b \frac{m_j c^2\lambda_j^2}{\lambda_j} = c^2\sum m_j\lambda_j < nc^2,$$

where we use the fact that n is the sum of all λ's and hence is an upper bound on $\sum m_j\lambda_j$.

Using the bound of c in Equation (5), and Lemma 6 in Equation (4), for large n,

$$\hat{R}(\mathcal{I}_{\overline{\varphi}}^{{\rm poi}(n)}) \le 108n^{1/3}\log n + 2nc^2(\log n)^2 + 1.5n^{1/3} < n^{1/3}\log^4 n.$$

C. Proof of Theorem 8

Recall that $\mathcal{I}$ was one of the classes of distributions obtained by Lemma 7. Hence, there exists $\overline{m} = (m_1,\ldots,m_b)$ and $\overline{s} = (s_1,\ldots,s_b)$, such that for all $\Lambda \in \mathcal{I}$, $m(\Lambda) = \overline{m}$ and $|\overline{s}(\Lambda) - \overline{s}| < \frac{1}{n^{99}}$. Also, for a distribution Λ, $\Lambda_j = \Lambda \cap B_j$ Let $\mathcal{B}_j \stackrel{\rm def}{=} \{\Lambda_j : \Lambda \in \mathcal{I}\}$, be the Λ_j's for all distributions in $\mathcal{I}$, and $1 \le j \le b$. Each element in $\mathcal{B}_j$ is a collection of m_j elements in B_j and their sum $\in [s_j - \frac{1}{n^{99}}, s_j + \frac{1}{n^{99}}]$.

Let $\overline{\varphi}_j$ be the profile generated by Λ_j, *i.e.,* λ's in the jth interval. Then, $\overline{\varphi}_{\rm med} = \overline{\varphi}_1 \cup \ldots \cup \overline{\varphi}_b = f(\overline{\varphi}_1,\ldots,\overline{\varphi}_b)$. By Poisson sampling, $\overline{\varphi}_j$'s are independent. By Lemmas 2 and 3,

$$\hat{R}(\mathcal{I}) \le \sum_{j=1}^b \hat{R}(\mathcal{B}_j).$$

We will prove Theorem 8 by showing that for all j,

$$\hat{R}(\mathcal{B}_j) \le \frac{3}{2} + 2(\log n)^2\frac{m_j\Delta_j^2}{\lambda_j}. \tag{6}$$

Recall that, $B_j = (\lambda_{j-1}, \lambda_j]$ and $S(\mathcal{B}_j) = \sum_{\overline{\varphi}_j\in\Phi}\sup_{\Lambda\in\mathcal{B}_j}\Lambda(\overline{\varphi}_j)$. Hence, $\Delta_j = \lambda_j - \lambda_{j-1}$.

Let $\Phi_j^{\rm near}$ be the set of all profiles with all multiplicities in $[\lambda_{j-1} - 2\sqrt{\lambda_j\log n}, \lambda_j + 2\sqrt{\lambda_j\log n}]$. In other words, $\Phi_j^{\rm near}$ consists of all $\{\mu_1,\ldots,\mu_{m_j}\}$, where each $\mu_r \in [\lambda_{j-1} - 2\sqrt{\lambda_j\log n}, \lambda_j + 2\sqrt{\lambda_j\log n}]$ By Poisson tail bounds, a ${\rm poi}(\lambda)$ random variable lies in $\lambda \pm 2\sqrt{\lambda\log n}$ with probability at least $1 - 1/n^4$, and the probability falls exponentially beyond this range. By the union bound, a profile generated by any Λ_j is in $\Phi_j^{\rm near}$ with probability $> 1 - \frac{1}{n^3}$. Using this along with the fact that $\lambda \ge n^{1/3}$, the following lemma states that it suffices to consider $\Phi_j^{\rm near}$. The proof is omitted due to lack of space.

Lemma 9. $\sum_{\overline{\varphi}_j\in\Phi}\sup_{\Lambda\in\mathcal{B}_j}\Lambda(\overline{\varphi}_j) < 2\sum_{\overline{\varphi}_j\in\Phi_j^{\rm near}}\sup_{\Lambda\in\mathcal{B}_j}\Lambda(\overline{\varphi}_j)$.

Let $\Lambda_{j0} = \{m_j \times \lambda_0 : \lambda_0 \stackrel{\rm def}{=} \frac{s_j}{m_j}\}$, be the element in $\mathcal{B}_j$ that has m_j Poisson parameters, all equal to λ_0 that add to s_j. In other words, Λ_{j0} is a uniform distribution in $\mathcal{B}_j$.

Note that the profiles in $\Phi_j^{\rm near}$ have bounded multiplicities. We show that for any $\overline{\varphi}_j \in \Phi_j^{\rm near}$, the uniform distribution Λ_{j0} does not underestimate $\sup_{\mathcal{B}_j}\Lambda_j(\overline{\varphi}_j)$. More precisely,

Theorem 10. *For any $\Theta_j > 0$, and any $\overline{\varphi}_j$ with all multiplicities in $[\lambda_0 - \frac{\Theta_j}{2}, \lambda_0 + \frac{\Theta_j}{2}]$, and any $\Lambda_j^* \in \mathcal{B}_j$,*

$$\frac{\Lambda_j^*(\overline{\varphi}_j)}{\Lambda_{j0}(\overline{\varphi}_j)} \le \sqrt{2}\exp\Big[m\Big(\frac{\Delta_j\Theta_j}{\lambda_0}\Big)^2\Big].$$

This provides a direct bound on $S(\mathcal{B}_j)$, and hence on $\hat{R}(\mathcal{B}_j)$.

Proof: Let $\overline{\varphi}_j = \{\mu_1,\ldots,\mu_{m_j}\}$ and $\Lambda_j^* = \{\lambda_1,\ldots,\lambda_{m_j}\}$. Let $s_j^* = \lambda_1 + \ldots + \lambda_{m_j}$. By Equation (3),

$$\Lambda_j^*(\overline{\varphi}_j) = N(\overline{\varphi}_j)\frac{\exp(-s_j^*)}{\prod\mu_i!}\left(\sum_{\sigma\in S_{m_j}}\prod_{l=1}^{m_i}\lambda_{j\sigma(l)}^{\mu_l}\right).$$

Taking the ratio with Λ_{j0},

$$\frac{\Lambda_j^*(\overline{\varphi}_j)}{\Lambda_{j0}(\overline{\varphi}_j)} = \frac{1}{m_j!}\exp(s_j^* - s_j)\left(\sum_{\sigma\in S_{m_j}}\prod_{l=1}^{m_j}\left(\frac{\lambda_{j\sigma(l)}}{\lambda_0}\right)^{\mu_l}\right).$$

Since, $|s_j^* - s_j| < \frac{1}{n^{99}}$, the term $\exp(s_j^* - s_j)$ is inconsequential to our calculations and is ignored. Let $\delta_i \stackrel{\text{def}}{=} \lambda_i - \lambda_0$. Let $\overline{\mu} = \frac{\sum\mu_i}{m_j}$ be the average multiplicity of $\overline{\varphi}_j$, and $\theta_j \stackrel{\text{def}}{=} \mu_j - \overline{\mu}$. Then, $\sum_{l=1}^{m_j}\theta_l = 0$, and $|\sum_{l=1}^{m_j}\delta_j| = |s_j^* - s_j| < \frac{1}{n^{99}}$. Plugging these in the equation above and using $1 + x \le \exp(x)$,

$$\begin{aligned}\frac{\Lambda_j^*(\overline{\varphi}_j)}{\Lambda_{j0}(\overline{\varphi}_j)} &= \frac{1}{m_j!}\left(\sum_{\sigma\in S_{m_j}}\prod_{l=1}^{m_j}\left(1+\frac{\delta_{\sigma(l)}}{\lambda_0}\right)^{\mu_l}\right)\\ &\le \frac{1}{m_j!}\sum_{\sigma\in S_{m_j}}\exp\left(\sum_{l=1}^{m_j}\frac{\delta_{\sigma(l)}\mu_l}{\lambda_0}\right)\\ &= \frac{1}{m_j!}\sum_{\sigma\in S_{m_j}}\exp\left(\sum_{l=1}^{m_j}\frac{\delta_{\sigma(l)}\theta_l}{\lambda_0}\right),\end{aligned}$$

where the last step used the fact that $\sum\theta_l = 0$. Since, $|\sum\delta_l| \le \frac{1}{n^{99}}$, it suffices to consider δ_l's that sum to zero. This reduces to maximizing the function

$$f(\{\delta_l\},\{\theta_l\}) = \frac{1}{m!}\sum_{\sigma}\left[\exp\left(\sum_{l=1}^{m}\frac{\delta_l\theta_{\sigma_l}}{\lambda_0}\right)\right],$$

subject to $\sum\theta_l = \sum\delta_l = 0$, and $|\theta_l| \le \Theta_j$, $|\delta_l| \le \Delta_j$. Note that for convenience we have replaced m_j with m.

f is a sum of exponentials, hence it is convex. Therefore, f is maximized when all δ's are $\pm\Delta_j$, and all θ_l's are $\pm\Theta_j$. Plugging these values and collecting equal terms,

$$\begin{aligned}&f(\{\delta_l\},\{\theta_l\})\\ &\le \frac{\left(\frac{m}{2}!\right)^2}{m!}\left(\sum_{k=0}^{m/2}\binom{\frac{m}{2}}{k}^2\exp\left((m-4k)\frac{\Delta_j\Theta_j}{\lambda_0}\right)\right)\\ &\le \frac{\left(\frac{m}{2}!\right)^2\binom{\frac{m}{2}}{\frac{m}{4}}}{m!}\exp\left(\frac{m\Delta_j\Theta_j}{\lambda_0}\right)\sum_{k=0}^{\frac{m}{2}}\binom{\frac{m}{2}}{k}\exp\left(\frac{-4k\Delta_j\Theta_j}{\lambda_0}\right)\\ &\le \frac{\sqrt{2}}{2^{m/2}}\left(1+\exp\left(-\frac{4\Delta_j\Theta_j}{\lambda_0}\right)\right)^{\frac{m}{2}}\exp\left(\frac{2\Delta_j\Theta_j}{\lambda_0}\right)^{\frac{m}{2}}\\ &= \frac{\sqrt{2}}{2^{m/2}}\left(\exp\left(-\frac{2\Delta_j\Theta_j}{\lambda_0}\right)+\exp\left(\frac{2\Delta_j\Theta_j}{\lambda_0}\right)\right)^{\frac{m}{2}}.\end{aligned}$$

By Taylor series, $\frac{e^x+e^{-x}}{2} \le e^{\frac{x^2}{2}}$, proving the theorem. ∎

If $\Delta_j \le \sqrt{\lambda_j}\log n$, profiles in Φ_j^{near} satisfy the conditions of the theorem above for $\Theta_j = \sqrt{\lambda_j}\log n + 4\sqrt{\lambda_j\log n} < \sqrt{2\lambda}\log n$. Hence $S(\mathcal{B}_j)$ can be bounded using Lemma 9,

$$S(\mathcal{B}_j) \le 2\sum_{\Phi_j^{\text{near}}}\sup_{\Lambda\in\mathcal{B}_j}\Lambda(\overline{\varphi}_j) \le 2^{\frac{3}{2}}\exp\Big(\frac{m_j\Delta_j^2\log^2 n}{\lambda_0}\Big).$$

Taking logarithms proves Equation (6), and summing over $1 \le j \le b$ proves Theorem 8.

REFERENCES

[1] A. Orlitsky, N. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE-ITT*, vol. 50, no. 7, pp. 1469– 1481, July 2004.

[2] G. M. Gemelos and T. Weissman, "On the entropy rate of pattern processes," *IEEE-ITT*, vol. 52, no. 9, pp. 3994–4007, 2006.

[3] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "A better good-turing estimator for sequence probabilities," *CoRR*, vol. abs/0704.1455, 2007.

[4] S. Boucheron, A. Garivier, and E. Gassiat, "Coding on countably infinite alphabets," *IEEE-ITT*, vol. 55, no. 1, pp. 358–373, 2009.

[5] W. Szpankowski and M. J. Weinberger, "Minimax redundancy for large alphabets," in *ISIT*, 2010, pp. 1488–1492.

[6] G. Shamir, "Universal lossless compression with unknown alphabets—the average case," *IEEE-ITT*, vol. 52, no. 11, pp. 4915–4944, Nov. 2006.

[7] B. Kelly and A. Wagner, "Near-lossless compression of large alphabet sources," in *CISS, Princeton*, 2012, pp. 1–6.

[8] B. M. Fitingof, "Optimal coding in the case of unknown and changing message statistics," *Probl. Inform. Transm.*, vol. 2, no. 2, pp. 1–7, 1966.

[9] L. Davisson, "Universal noiseless coding," *IEEE-ITT*, vol. 19, no. 6, pp. 783–795, Nov. 1973.

[10] N. Merhav and M. Feder, "Universal prediction," *IEEE-ITT*, vol. 44, no. 6, pp. 2124–2147, October 1998.

[11] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY, USA: Cambridge University Press, 2006.

[12] L. D. Davisson, R. J. McEliece, M. B. Pursley, and M. S. Wallace, "Efficient universal noiseless source codes," *IEEE-ITT*, vol. 27, no. 3, pp. 269–279, 1981.

[13] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: basic properties." *IEEE-ITT*, vol. 41, no. 3, pp. 653–664, 1995.

[14] T. Cover, "Universal portfolios," *Mathematical Finance*, vol. 1, no. 1, pp. 1–29, January 1991.

[15] J. Rissanen, "Fisher information and stochastic complexity," *IEEE-ITT*, vol. 42, no. 1, pp. 40–47, January 1996.

[16] W. Szpankowski, "On asymptotics of certain recurrences arising in universal coding," *Problems of Information Transmission*, vol. 34, no. 2, pp. 142–146, 1998.

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

[18] M. Drmota and W. Szpankowski, "Precise minimax redundancy and regret," *IEEE-ITT*, vol. 50, no. 11, pp. 2686–2707, 2004.

[19] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *Acoustics, Speech and Signal Processing, IEEE Tran. on*, vol. 35, no. 3, pp. 400 – 401, mar 1987.

[20] 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*, 1996, pp. 310–318.

[21] A. Orlitsky, N. Santhanam, and J. Zhang, "Always Good Turing: Asymptotically optimal probability estimation," *Science*, vol. 302, no. 5644, pp. 427–431, October 17 2003.

[22] G. Shamir, "A new upper bound on the redundancy of unknown alphabets," in *CISS, Princeton*, 2004.

[23] A. Garivier, "A lower-bound for the maximin redundancy in pattern coding," *Entropy*, vol. 11, no. 4, pp. 634–642, 2009.

[24] J. Acharya, H. Das, and A. Orlitsky, "Tight bounds on profile redundancy and distinguishability," in *NIPS*, 2012.

[25] Y. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, no. 3, pp. 3–17, 1987.

[26] A. Orlitsky, N. Santhanam, K. Viswanathan, and J. Zhang, "On modeling profiles instead of values," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 2004.

[27] M. Mitzenmacher and E. Upfal, *Probability and computing - randomized algorithms and probabilistic analysis*. Cambridge Univ. Press, 2005.

[28] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan, "Competitive closeness testing," in *COLT*, vol. 19, 2011.

[29] P. Elias, "Universal codeword sets and representations of integers," *IEEE-ITT*, vol. 21, no. 2, pp. 194–203, Mar. 1975.