Efficient Compression of Monotone and m-Modal Distributions

Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh

ECE Department, UCSD {jacharya, ashkan, alon, asuresh}@ucsd.edu

***Abstract*—We consider universal compression of n samples drawn independently according to a monotone or m-modal distribution over k elements. We show that for all these distributions, the per-sample redundancy diminishes to 0 if $k = \exp(o(n/\log n))$ and is at least a constant if $k = \exp(\Omega(n))$.**

I. Introduction

Universal compression concerns encoding the output of an *unknown* source in a *known* class of distributions. We are interested in understanding compression of *i.i.d.* samples from distributions over an underlying discrete alphabet $\mathcal{X}$ (*e.g.,* a subset of $\mathbb{N}$). Let $P^n \stackrel{\text{def}}{=} P \times P \ldots \times P$ be a product distribution over $\mathcal{X}^n$. Shannon's source coding theorem shows that $H(P) \stackrel{\text{def}}{=} \sum_{\mathcal{X}} -P(x)\log P(x)$ bits are necessary and sufficient to encode a *known* source (with distribution) P, where $H(P)$ is the entropy of P. By Kraft's Inequality, any distribution Q_n over $\mathcal{X}^n$ implies a code that uses $-\log Q_n(x_1^n)$ bits to encode $x_1^n \in \mathcal{X}^n$. The code implied by the underlying distribution P^n uses $-\log P^n(x_1^n)$ bits and the expected number of bits used is $nH(P)$. The average redundancy of Q_n with respect to P^n is

$$D(P^n||Q_n) = \mathbb{E}_{P^n}\left[\log \frac{P^n(x_1^n)}{Q_n(x_1^n)}\right],$$

the expected extra number of bits used by Q_n beyond the entropy. $D(P||Q)$ is called the Kullback-Liebler divergence between P and Q.

Let the underlying distribution be in a known class $\mathcal{P}$, and

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

be the class of all distributions of the form P^n. Then

$$\overline{R}(\mathcal{P}^n) \stackrel{\text{def}}{=} \inf_{Q_n} \sup_{P^n} D(P^n||Q_n)$$

is the *average or min-max redundancy* of $\mathcal{P}^n$, which corresponds to the best coding for the worst distribution.

While coding over unknown distributions, often some apriori knowledge about the underlying distribution is known. For example, coding the last names of people from a census data is close to coding a monotone distribution, since we have a prior knowledge about the prevalences of last names. We expect that the last name Smith is more likely than Galifianakis. In a text document we have some knowledge about word frequencies and probabilities. In such language modeling applications, Zipf distributions are common [1]. Geometric distributions over integers are useful in compressing residual signals in image compression [2].

A natural generalization is to consider distributions with at most m modes. For example, life expectancy of a population, Poisson, and Binomial distributions are unimodal. There has been a considerable interest over the past decade in approximating mixtures of distributions. Many real world phenomenon can be modeled as outcomes of mixtures of simple distributions and these have been studied since as early as the study Naples crab population by Pearson [3]. In this setup after observing measurements (say diameter over heights), the distribution that explained the data the best was a mixture of two Gaussians, predicting the presence of more than one specie. Such mixtures of m simple distributions are typically m-modal. There has been enormous work in the past decade in learning mixtures of distributions.

Before discussing monotone distributions, we state the results for the most well studied finite discrete distributions, *i.e.,* the *i.i.d.* distributions. Let $\mathcal{I}_k^n$ be the class of all *i.i.d.* distributions over k elements and block length n. Redundancy of $\mathcal{I}_k^n$ has been extensively studied [4–11]. It is now well established that

1) For $k = o(n)$

$$\overline{R}(\mathcal{I}_k^n) = \frac{k-1}{2}\log\frac{n}{k}(1+o(1)). \tag{1}$$

2) For $n = o(k)$

$$\overline{R}(\mathcal{I}_k^n) = n\log\frac{k}{n}(1+o(1)). \tag{2}$$

II. Terminology

We consider distributions over the set $\mathbb{N}$ of positive integers. Such a distribution is non-increasing, or *monotone*, if $P(i) \geq P(i+1)$ for all $i \geq 1$. Let $\mathcal{M}$ be the class of all monotone distributions over $\mathbb{N}$, and let $\mathcal{M}_k$ be the subset of $\mathcal{M}$ consisting of monotone distributions over $[k] = \{1, \ldots, k\}$. Let $\mathcal{M}_k^n$ denote the class of distributions obtained by sampling one distribution in $\mathcal{M}_k$ independently n times, *i.e.,* product distributions of the form P^n for $P \in \mathcal{M}_k$.

Generalizing monotone distributions, a consecutive set $[l, r] = \{l, \ldots, r\}$ of integers is a *mode* of a distribution P if for all $i, j \in [l, r]$ $P(i) = P(j)$, and $(P(l-1) - P(l)) \cdot (P(r+1) - P(r)) > 0$, representing a local minimum or maximum of P. P is *m-modal* if it has at most m modes. Note that a monotone distribution is 0-modal. Let $\mathcal{M}_{k,m}$ be the collection of all m-modal distributions over $[k]$, and $\mathcal{M}_{k,m}^n$, the distributions over $[k]^n$ obtained by sampling the same distribution in $\mathcal{M}_{k,m}$ independently n times.

III. Related Work

Initial work on universal compression of monotone distributions was motivated by representation of integers, hence centered on a single instance, namely $n = 1$. [12] showed that every positive integer k can be represented using roughly $\log k + 2 \log\log k$ bits. [13] derived a related lower bound, and [14] constructed a method for minimizing the expected number of bits normalized by the distribution's entropy.

Remarkably, [15] found the expected redundancy exactly,

$$\overline{R}(\mathcal{M}_k) = \log \left(1 + \sum_{i=2}^{k} \left(1 - \frac{1}{i}\right)^i \frac{1}{i-1}\right).$$

Since $(1 - \frac{1}{i})^i \to e^{-1}$ and $\sum_{i=1}^k \frac{1}{i} \sim \log k$, this shows that

$$\overline{R}(\mathcal{M}_k) = \log\log k + O(1),$$

and in particular, $\overline{R}(\mathcal{M}) = \infty$.

More recently, [16] analyzed the *min-ave redundancy* that replaces the redundancy of the worst distribution in the class by the average over all distributions. They show that monotone distributions have constant min-ave redundancy.

Recent work has concerned compression of independent identical monotone distributions, namely distributions in $\mathcal{M}^n$. The results above imply $\overline{R}(\mathcal{M}^n) = \infty$, yet [17] showed that the set of all distributions in $\mathcal{M}$ with finite entropy can be compressed with a diminishing per-sample relative redundancy. Specifically, there is a distribution Q_n over $\mathbb{N}^n$, such that for all $P \in \mathcal{M}$ with $H(P) < \infty$,

$$D(P^n || Q_n) \le nH(P) \frac{\log\log(nH(P))}{\log(nH(P))}.$$

Note that the theorem holds for all distributions with finite entropy, even those with infinite support. It therefore does not necessarily imply that the redundancy is sub-linear. For example, for the uniform distribution over $[\exp(\sqrt{n})]$, the bound is $n^{3/2}$, hence super-linear.

[18] considered the redundancy of monotone distributions as a function of the alphabet size k as well as the block length n. They show tight lower and upper bounds for $k = O(n)$ [1],

$$\overline{R}(\mathcal{M}_k^n) = \tilde{\Theta}(\min\{k, n^{1/3}\}).$$

In particular, this implies that when n grows linearly with k, the per-sample redundancy diminishes to 0. These results can be extended using methods from [19] to prove that $\mathcal{M}_k^n$ has sub-linear redundancy for $k = 2^{o(\sqrt{n})}$.

Universal compression is related to the problem of distribution estimation with Kullback-Leibler distortion. The problem of estimating monotone and few-modes distributions has been studied in statistics and theoretical computer science. [20] considered learning monotone and unimodal distributions with few samples. The sample complexity of learning m-modal distributions over an alphabet of size k was considered by [21]. Testing distributions for monotonicity has been considered in varied settings [22–26].

IV. Results

To begin, in Section V we consider approximating monotone distribution over k elements by step distributions with significantly fewer steps. While such approximation results are known for ℓ_1 distance [20], we need approximation in KL divergence, and these require more work as the KL divergence is not a metric and does not satisfy the triangle inequality. In Theorem 3 we show that for any integer $b \ge 10 \log k$, there is a partition of $[k]$ into b fixed intervals such that for any P, the step function $\bar{P}$ derived by averaging P over each of the intervals satisfies

$$D(P||\bar{P}) \le 10 \frac{\log k}{b}.$$

This result can be viewed as reducing the effective alphabet size from k to b, and in Section VI we use it to show that

$$\overline{R}(\mathcal{M}_k^n) \le 10 \frac{n \log k}{b} + \frac{(b-1)}{2} \log n.$$

Specifically, in Theorem 4 we show that for large n, and any k,

$$\overline{R}(\mathcal{M}_k^n) \le \sqrt{20 n \log k \log n}.$$

Hence in Corollary 5, we deduce that for $k = 2^{o(n/\log n)}$,

$$\overline{R}(\mathcal{M}_k^n) = o(n).$$

In Section VII we extend these results to m-modal distributions and show in Theorem 7 that for large n and any $k \ge m$

$$\overline{R}(\mathcal{M}_{k,m}^n) \le \log \binom{k}{m} + (m+1)\overline{R}(\mathcal{M}_k^n).$$

Consequently, in Corollary 8, we deduce that for any fixed m, and $k = 2^{o(n/\log n)}$,

$$\overline{R}(\mathcal{M}_{k,m}^n) = o(n).$$

Conversely, in Theorem 9, we show that for $k = 2^n$,

$$\overline{R}(\mathcal{M}_k^n) = \Omega(n).$$

By monotonicity of redundancy, the same bound holds for all $k = 2^{\Omega(n)}$. It follows that if k is subexponential in $n/\log n$ then the redundancy diminishes to 0, while if it at least exponential in n, the redundancy is at least a constant.

V. Approximating monotone distributions

We show that any monotone distribution over k elements can be approximated in KL-divergence by a step distribution with significantly fewer steps. We will use the following simple result on the average empirical variance of non-negative numbers. Its proof follows from a straight-forward expansion and is omitted for brevity.

[1] $g(n) = \tilde{\Theta}(f(n)$ if $f(n) = g(n) polylog(n)$

Lemma 1: For $0 \leq x_1 \leq x_2 \leq \ldots \leq x_n$ with mean $\bar{x}$,

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \leq n(x_n - x_1)\bar{x}. \qquad \blacksquare$$

Let $I_1^b = I_1, \ldots, I_b$ be a partition of $[k]$ into consecutive intervals. P is a *step* distribution over this partition if for every $1 \leq j \leq b$, all $i, i' \in I_j$ satisfy $P(i) = P(i')$. Let $|I|$ denote the number of integers in interval I. For a distribution P, let $\bar{P}$ be the step distribution over I_1^b whose constant value over each interval is P's average over that interval, namely for $x \in I_j$

$$\bar{P}(x) = \frac{\sum_{x \in I_j} P(x)}{|I_j|} = \frac{P(I_j)}{|I_j|}. \tag{3}$$

Let P_j^+ and P_j^- be the highest and lowest probabilities in interval I_j, and let k_j be the number of non-zero probabilities in interval I_j. The following lemma bounds the KL Divergence between P and $\bar{P}$.

Lemma 2: For any $P \in \mathcal{M}_k$,

$$D(P||\bar{P}) \leq \sum_{j=1}^{b} k_j(P_j^+ - P_j^-).$$

Proof: Using $\log x \leq x - 1$[2] for $x \geq 0$ and Lemma 1,

$$\begin{aligned}\sum_{x \in I_j} P(x) \log \frac{P(x)}{\bar{P}_j} &\leq \sum_{x \in I_j} P(x) \frac{P(x) - \bar{P}_j}{\bar{P}_j} \\ &= \sum_{x \in I_j} \frac{P^2(x) - \bar{P}_j^2}{\bar{P}_j} \\ &= \frac{1}{\bar{P}_j} \sum_{x \in I_j} (P(x) - \bar{P}_j)^2 \\ &\leq k_j(P_j^+ - P_j^-),\end{aligned}$$

and the proof follows by summing over all intervals. $\blacksquare$

We now describe a partition of $[k]$ into b intervals over which every monotone distribution P is closely approximated by the step distribution $\bar{P}$. For $\gamma = \frac{2 \log k}{b}$ define

$$|I_j| = \begin{cases} 1 & \text{for } 1 \leq j \leq \frac{b}{2}, \\ \lfloor 2(1+\gamma)^{j-b/2} \rfloor & \text{for } \frac{b}{2} < j \leq b. \end{cases}$$

Observe that

$$\begin{aligned}\sum_{j=1}^{b} |I_j| &= b/2 + \sum_{j=\frac{b}{2}+1}^{b} \lfloor 2(1+\gamma)^{j-b/2} \rfloor \\ &\geq \sum_{j=b/2+1}^{b} 2(1+\gamma)^{j-b/2} \\ &= 2\frac{1+\gamma}{\gamma}\left((1+\gamma)^{b/2} - 1\right),\end{aligned}$$

which is larger than k for $\log(1+\gamma) \geq 2 \log k/b$. Hence the intervals span $[k]$. For $j \leq \frac{b}{2}$, $|I_j| = 1$, hence $P_j^+ = P_j^-$, and for $j > b/2$, $k_j \leq |I_j| \leq 2(1+\gamma)^{j-b/2}$.

Theorem 3: Let $b \geq 10 \log k$ and the intervals I_1^b as defined above. Then for every $P \in \mathcal{M}$,

$$D(P||\bar{P}) \leq 10 \frac{\log k}{b}.$$

Proof: By Lemma 2,

$$\begin{aligned}D(P||\bar{P}) &\leq \sum_{j=1}^{b} k_j(P_j^+ - P_j^-) \\ &\leq \sum_{j=b/2+1}^{b} 2(1+\gamma)^{j-b/2}(P_j^+ - P_j^-) \\ &\leq 2(1+\gamma)P_{b/2+1}^+ + \\ &\quad \sum_{j=b/2+1}^{b} 2(P_{j+1}^+(1+\gamma)^{j+1-b/2} - P_j^-(1+\gamma)^{j-b/2}) \\ &\leq k_{b/2+1}P_{b/2+1}^+ + 2\gamma \sum_{j=b/2+1}^{b} P_j^-(1+\gamma)^{j-b/2},\end{aligned}$$

where the last step uses $P_{j+1}^+ \leq P_j^-$.

We need to consider only non-zero k_j's. If k_{j+1} is non-zero this implies that $k_j \geq (1+\gamma)^{j-b/2}$. Therefore, the summation above can be bounded as

$$\begin{aligned}D(P||\bar{P}) &\leq k_{b/2+1}P_{b/2+1}^+ + 2\gamma \sum_{j=b/2+1}^{b} P_j^- k_j \\ &\leq |I_{b/2+1}|P_{b/2+1}^+ + 2\gamma.\end{aligned}$$

The theorem follows by substituting the values of γ and $|I_{b/2+1}|$ and the inequality $P_{b/2+1}^+ \leq 2/b$. $\blacksquare$

VI. Upper bound on monotone redundancy

Theorem 4: For large n, and any k,

$$\overline{R}(\mathcal{M}_k^n) \leq \sqrt{20 n \log k \log n}.$$

Proof: Recall the definition of $\bar{P}$ in (3). As before, $\bar{P}^n$ is the *i.i.d.* distribution by sampling $\bar{P}$ n times. Then for any distribution Q_n over $[k]^n$,

$$\begin{aligned}D(P^n||Q_n) &= \sum_{x_1^n \in [k]^n} P^n(x_1^n) \log \frac{P^n(x_1^n)}{Q_n(x_1^n)} \\ &= \sum_{x_1^n \in [k]^n} P^n(x_1^n) \left[\log \frac{P^n(x_1^n)}{\bar{P}^n(x_1^n)} + \log \frac{\bar{P}^n(x_1^n)}{Q_n(x_1^n)}\right] \\ &= nD(P||\bar{P}) + \sum_{x_1^n \in [k]^n} P^n(x_1^n) \log \frac{\bar{P}^n(x_1^n)}{Q_n(x_1^n)},\end{aligned}$$

where the last step follows since the KL divergence for product distributions is the sum of KL divergence of distributions on each coordinate.

For intervals $I_1, \ldots, I_b$ satisfying Theorem 3, by the definition of redundancy

$$\overline{R}(\mathcal{M}_k^n) \leq \frac{10 n \log k}{b} + \inf_{Q_n} \sup_{P \in \mathcal{M}_k} \sum_{x_1^n \in [k]^n} P^n(x_1^n) \log \frac{\bar{P}^n(x_1^n)}{Q_n(x_1^n)}.$$

[2] We use natural logarithms throughout the paper. The results are a constant factor from base 2.

We use this inequality and Equation (1) to construct a distribution over $[k]^n$ that has a small KL divergence with respect to any distribution in $\mathcal{M}_k^n$. We do this by showing that the second expression is upper bounded by the redundancy of $n-$length *i.i.d.* sequences over an alphabet of size b.

Equation (1) states that for $b = o(n)$, there is a distribution $Q_{b,n}$ over $[b]^n$ such that any distribution P over $[b]$ satisfies

$$D(P^n||Q_{b,n}) \leq \frac{(b-1)}{2}\log n. \tag{4}$$

Using this distribution, we design a distribution over $[k]^n$ as follows. For the intervals described earlier, and any $x \in [k]$, let $f(x)$ be the index of the interval that x belongs to. Therefore, f is a map from $[k]$ to $[b]$. Then, $f(x_1^n) \overset{\text{def}}{=} f(x_1,\ldots,x_n) \overset{\text{def}}{=} f(x_1),\ldots,f(x_n)$ maps $[k]^n$ to $[b]^n$. Let $f(x_1^n) = j_1^n \overset{\text{def}}{=} j_1,\ldots,j_n$. The number of x_1^n that map to j_1^n is $|I_{j_1}|\ldots|I_{j_n}|$. Define the distribution $\bar{Q}_n$ over $[k]^n$ as

$$\bar{Q}_n(x_1^n) \overset{\text{def}}{=} \frac{Q_{b,n}(j_1^n)}{\prod_{i=1}^n |I_{j_i}|}. \tag{5}$$

It is easy to check that it sums to 1 and the distribution is well defined. Similarly by Equation (3),

$$\bar{P}^n(x_1^n) = \prod_{i=1}^n \frac{P(I_{j_i})}{|I_{j_i}|}.$$

Using these two, for any $P \in \mathcal{M}_k$

$$\sum_{x_1^n\in[k]^n} P^n(x_1^n)\log\frac{\bar{P}^n(x_1^n)}{\bar{Q}_n(x_1^n)}$$
$$= \sum_{j_1^n\in[b]^n}\left(\sum_{x_1^n:f(x_1^n)=j_1^n} P^n(x_1^n)\right)\log\frac{\prod_{i=1}^n P(I_{j_i})}{Q_{b,n}(j_1^n)}.$$

Now, for $j_1^n \in [b]^n$,

$$\sum_{x_1^n:f(x_1^n)=j_1^n} P^n(x_1^n) = \sum_{x_1^n:x_i\in I_{j_i}}\prod_{i=1}^n P(x_i) \overset{(a)}{=} \prod_{i=1}^n P(I_{j_i}),$$

where (a) exchanges the sum and product. Therefore,

$$\sum_{x_1^n\in[k]^n} P^n(x_1^n)\log\frac{\bar{P}^n(x_1^n)}{Q_n(x_1^n)} = \sum_{j_1^n\in[b]^n}\prod_{i=1}^n P(I_{j_i})\log\frac{\prod_{i=1}^n P(I_{j_i})}{Q_{b,n}(j_1^n)}$$

A distribution P induces a distribution over $I_1,\ldots,I_b$, an alphabet of size b, and this expression is the KL divergence of the product distribution over the intervals to $Q_{b,n}$. By Equation (4) it is bounded by $(b-1)\log n/2$, hence

$$\begin{aligned}\overline{R}(\mathcal{M}_k^n) &\leq \frac{10n\log k}{b} + \inf_{Q_n}\sup_{P\in\mathcal{M}_k}\sum_{x_1^n\in[k]^n} P^n(x_1^n)\log\frac{\bar{P}^n(x_1^n)}{Q_n(x_1^n)}\\ &\leq \frac{10n\log k}{b} + \sup_{P\in\mathcal{M}_k}\sum_{j_1^n\in[b]^n}\prod_{i=1}^n P(I_{j_i})\log\frac{\prod_{i=1}^n P(I_{j_i})}{Q_{b,n}(j_1^n)}\\ &\leq \frac{10n\log k}{b} + \frac{(b-1)}{2}\log n.\end{aligned}$$

Choosing $b = \sqrt{\frac{20n\log k}{\log n}}$ proves the theorem. ■

Corollary 5: For $k = 2^{o(n/\log n)}$,

$$\overline{R}(\mathcal{M}_k^n) = o(n). \qquad ■$$

VII. Upper bound on m-modal distributions

We upper bound the redundancy of m-modal distributions by decomposing them into $m+1$ monotone distributions. First note that the redundancy of a union of distribution classes is close to the highest redundancy of a class in the union.

Lemma 6 (Redundancy of unions): If $\mathcal{P}_1,\ldots,\mathcal{P}_T$ are T distribution classes over the same domain, then

$$\overline{R}\left(\bigcup_{1\leq i\leq T}\mathcal{P}_i\right) \leq \max_{1\leq i\leq T}\overline{R}(\mathcal{P}_i) + \log T.$$

Proof: Let distribution Q_i^* achieve the redundancy of $\mathcal{P}_i$, and define $Q^* = \frac{\sum_{i=1}^T Q_i^*}{T}$ to be the average of these redundancy achieving distributions. For any i and $P_i \in \mathcal{P}_i$,

$$D(P_i||Q^*) \leq D(P_i||Q_i^*) + \log T \leq \overline{R}(\mathcal{P}_i) + \log T. \qquad ■$$

We first decompose the class of m-modal distributions into $\binom{k}{m}$ classes.

Theorem 7: For large n and any $k \geq m$

$$\overline{R}(\mathcal{M}_{k,m}^n) \leq \log\binom{k}{m} + (m+1)\overline{R}(\mathcal{M}_k^n).$$

Proof: There are $\binom{k}{m}$ choices for the modes of the distributions. Divide the class of all m-modal distributions into $\binom{k}{m}$ classes such that the modes of all the distributions within one class are the same. The distributions are monotone between the modes and there are at most $m+1$ distinct regions. Each such region can be coded with redundancy at most $\overline{R}(\mathcal{M}_k^n)$ and therefore the total extra number of bits can be bounded by $(m+1)\overline{R}(\mathcal{M}_k^n)$. Combining with Lemma 6 with $T = \binom{k}{m}$ proves Theorem 7. ■

Corollary 8: For any fixed m and $k = 2^{o(n/\log n)}$,

$$\overline{R}(\mathcal{M}_{k,m}^n) = o(n). \qquad ■$$

VIII. Lower bound

Theorem 9: For $k = 2^n$,

$$\overline{R}(\mathcal{M}_k^n) = \Omega(n).$$

Proof: We show the lower bound using the redundancy capacity theorem. Our objective will be to construct a large class of *distinguishable* distributions defined below.

Definition 10: A collection $\mathcal{S} \subset \mathcal{P}^n$ of distributions over $\mathcal{X}^n$ is $(1-\epsilon)-$distinguishable if there exists a function $f: \mathcal{X}^n \to \mathcal{S}$, such that for any $P^n \in \mathcal{S}$, $Prob(f(X^n) \neq P^n) < \epsilon$.

In other words, a collection of distributions is distinguishable if given a sample generated by one of the distributions, we can identify the distribution, with error $\leq \epsilon$.

A collection of distinguishable distributions provides a lower bound on the redundancy by the following formulation of the redundancy-capacity theorem.

Lemma 11: If there is a collection $\mathcal{S} \subset \mathcal{M}_k^n$ of $(1-\epsilon)-$distinguishable distributions, then

$$\overline{R}(\mathcal{M}_k^n) \geq (1-\epsilon)\log|\mathcal{S}| - 1.$$

We now construct a class of 2^{cn} distinguishable distributions in $\mathcal{M}_k^n$ for $k = 2^n$, and a constant $c > 0$. This will give a lower bound of $cn(1-\epsilon) - 1$ on $\overline{R}(\mathcal{M}_k^n)$.

Due to lack of space, we provide an overview of the construction of distributions and a sketch the proof. We assume n is even for the ease of illustration.

Divide $[2^n]$ into n intervals as $I_1 = \{1, 2\}$, and $I_j = \{2^{j-1}+1, \ldots, 2^j\}$ for $j = 2, \ldots, n$. Our collection of distinguishable distributions will be a subset of distributions of the form P^n, where P satisfies

1) For $2 \leq j \leq n$, and $i_1, i_2 \in (2^j, 2^{j+1}]$, $P(i_1) = P(i_2)$.
2) For $n/2$ of the intervals, $P(I_j) = \frac{2}{3n}$ and for the remaining $n/2$, $P(I_j) = \frac{4}{3n}$. Furthermore, $P(I_1)$ is always $\frac{4}{3n}$.

It is then easy to verify that any P satisfying these properties is a distribution in $\mathcal{M}_k$, and furthermore, the number of such distributions is exactly $\binom{n-1}{\frac{n}{2}}$.

Consider the following bijection from the set of distributions satisfying these conditions to the set of binary strings in $\{0,1\}^n$ that have weight $n/2$ and whose first bit is 1. For $j = 1, \ldots, n$, if $P(I_j) = 4/3n$ set the jth bit to 1, and 0 otherwise, defining the bijection.

We now show the existence of a large subset of such distributions that map to strings with a large Hamming distance, using the Gilbert-Varshamov bound [27], as follows.

Lemma 12: For $\alpha < \frac{1}{2}$, there exists a class of $M \overset{\text{def}}{=} 2^{n(1-h(\alpha)-o(1))}$ distributions satisfying the properties and such that the Hamming distance between the strings that they map to satisfies,

$$|S(P_i) \cap S(P_j)| < n(1-\alpha)/2.$$

In other words for any pair of distributions, their distributions are different in at least a fraction $(1-\alpha)/2$ of the intervals.

Using this, we prove the following theorem.

Theorem 13: The class of distributions defined above is $0.9-$distinguishable and contains $\geq 2^{n/100}$ distributions.

Applying Lemma 11 to this theorem proves the lower bound.

Due to space constraints, we only sketch the function f that maps $[k]^n$ to the collection of distributions in the previous theorem and defer the complete proof to the full version of the paper.

For a distribution P in the theorem, let $L(P) \subset [n]$ be the intervals where $P(I) = 4/3n$. Note that $|L(P)| = n/2$, and we expect 2/3rd of the n symbols to appear in these intervals when P is sampled n times.

Given a sample in $[k]^n$, compute the fraction of symbols lying in $L(P)$ for each of the $2^{n/100}$ such P's. Output the P that has the highest fraction of elements in its corresponding $L(P)$. The proof of distinguishability uses Chernoff bounds and is along the lines of the proof of lower bound in [19]. ■

References

[1] G. K. Zipf, *The Psychobiology of Language*. New York, NY, USA: Houghton-Mifflin, 1935.

[2] N. Merhav, G. Seroussi, and M. J. Weinberger, "Optimal prefix codes for sources with two-sided geometric distributions," *IEEE Transactions on Information Theory*, vol. 46, no. 1, pp. 121–135, 2000.

[3] K. Pearson, "Contributions to the Mathematical Theory of Evolution," *Philosophical Transactions of the Royal Society of London. A*, vol. 185, pp. 71–110, 1894.

[4] J. Kieffer, "A unified approach to weak universal source coding," *IEEE Transactions on Information Theory*, vol. 24, no. 6, pp. 674–682, Nov. 1978.

[5] L. D. Davisson, "Universal noiseless coding," *IEEE Transactions on Information Theory*, vol. 19, no. 6, pp. 783–795, Nov. 1973.

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

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

[8] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, January 1996.

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

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

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

[12] P. Elias, "Universal codeword sets and representations of integers," *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 194–203, Mar. 1975.

[13] B. Y. Ryabko, "Coding of a source with unknown but ordered probabilities," *Problemy Peredachi Informatsii*, vol. 15, no. 2, pp. 71–77, 1979.

[14] J. Rissanen, "Minimax codes for finite alphabets," *Information Theory, IEEE Transactions on*, vol. 24, no. 3, pp. 389–392, 1978.

[15] L. D. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes." *IEEE Transactions on Information Theory*, vol. 26, no. 2, pp. 166–174, 1980.

[16] M. Khosravifard, H. Saidi, M. Esmaeili, and T. A. Gulliver, "The minimum average code for finite memoryless monotone sources." *IEEE Transactions on Information Theory*, vol. 53, no. 3, pp. 955–975, 2007.

[17] D. Foster, R. Stine, and A. Wyner, "Universal codes for finite sequences of integers drawn from a monotone distribution," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1713–1720, June 2002.

[18] G. I. Shamir, "Universal source coding for monotonic and fast decaying monotonic distributions," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7194–7211, 2013.

[19] J. Acharya, H. Das, and A. Orlitsky, "Tight bounds on profile redundancy and distinguishability," in *Neural Information Processing Systems (NIPS)*, 2012.

[20] L. Birgé, "On the risk of histograms for estimating decreasing densities," *Annals of Statistics*, vol. 15, no. 3, pp. 1013–1022, 1987.

[21] C. Daskalakis, I. Diakonikolas, and R. A. Servedio, "Learning k-modal distributions via testing," in *SODA*, 2012, pp. 1371–1385.

[22] M. Woodroofe and J. Sun, "Testing uniformity versus a monotone density," *The Annals of Statistics*, vol. 27, no. 1, pp. pp. 338–360, 1999.

[23] O. Goldreich, S. Goldwasser, E. Lehman, and D. Ron, "Testing monotonicity," *Foundations of Computer Science, IEEE Annual Symposium on*, vol. 0, p. 426, 1998.

[24] T. Batu, R. Kumar, and R. Rubinfeld, in *STOC*. ACM, pp. 381–390.

[25] R. Rubinfeld and R. A. Servedio, "Testing monotone high-dimensional distributions," *Random Struct. Algorithms*, vol. 34, no. 1, pp. 24–44, 2009.

[26] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "A competitive test for uniformity of monotone distributions," in *AISTATS*, 2013, pp. 57–65.

[27] R. M. Roth, *Introduction to coding theory*. Cambridge University Press, 2006.