Minimax Regret on Patterns Using Kullback-Leibler Divergence Covering

Jennifer Tang JSTANG@MIT.EDU
MIT, Cambridge, MA USA

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

This paper considers the problem of finding a tighter upper bound on the minimax regret of patterns, a class used to study large-alphabet distributions which avoids infinite asymptotic regret and redundancy. Our method for finding upper bounds for minimax regret uses cover numbers with Kullback-Leibler (KL) divergence as the distance. Compared to existing results by Acharya et al. (2013), we are able to improve the power of the exponent on the logarithmic term, giving a minimax regret bound which matches the best known minimax redundancy bound on patterns.

Keywords: Patterns, profiles, minimax regret, universal prediction, Kullback-Leibler divergence

1. Introduction

How well can an estimator predict the next symbol in a sequence when the alphabet size is large? While there are many variations and nuances of this general question, we will focus on the particular case when the loss function is log-loss, a function which connects learning problems to ideas of data compression in information theory. Our main objective is to evaluate minimax regret, which represents how well an estimator can predict a sequence compared to an oracle who has some advance knowledge of the sequence (we give the specific details in Section 1.3). A related quantity to regret is redundancy (an average case version of regret).

A long line of work has been dedicated to studying redundancy and regret for the class of iid distributions. While it is known that for a fixed alphabet size k that the per-symbol redundancy can be driven to zero as the length n goes to infinity, this does not hold when k is as large as n (see Section 1.4 for more discussion). To deal with this, Orlitsky and Santhanam (2004) developed the notion of patterns, a class which captures large-alphabet distributions (so that we can let k be as large as n) but still has per-symbol redundancy which goes to zero.

We now present our main result. Formal definitions of all quantities involved are given in Sections 1.1 to 1.3. Denote the minimax redundancy over a class $\mathcal{I}$ as $\bar{R}(\mathcal{I})$ and the minimax regret as $\hat{R}(\mathcal{I})$. Let the class of patterns on length n sequences be $\mathcal{I}_\Psi^n$.

The tightest results in the literature for minimax redundancy and minimax regret on the class of patterns for large n, given by Acharya et al. (2013), are

$$0.3 \cdot n^{1/3} \leq \bar{R}(\mathcal{I}_\Psi^n) \leq n^{1/3}(\log n)^{4/3} \tag{1}$$

$$\left(\frac{3}{2\log 2}\right) n^{1/3} \leq \hat{R}(\mathcal{I}_\Psi^n) \leq n^{1/3}(\log n)^4 \tag{2}$$

where the bounds hold for large n.

In this work, our main result is that we can improve the upper bound for minimax regret so that it matches the known upper bound for minimax redundancy (up to a multiplicative constant).

Theorem 1 *For some constant c,*

$$\hat{R}(\mathcal{I}_\Psi^n) \leq cn^{1/3}(\log n)^{4/3}\,. \tag{3}$$

This result is not asymptotic. Setting $c = 38$ gives a bound for all n. Compared to (2), our result improves the exponent on the logarithmic term.

Our main technique for improving the bound on regret is to use the idea of covering under Kullback-Leibler (KL) divergence. Generally KL divergence covering bounds are not typically used since it is difficult to get precise covering numbers (due to probabilities near the boundary). Also, KL divergence covering, known to be able to compute redundancy bounds, was not previously known to be a viable approach for computing regret with logarithmic loss, where results are usually computed by summing over many probability values (Shtarkov's sum in Orlitsky and Santhanam (2004)) which are similar to ideas used in Shtarkov et al. (1995). In this work, we are able to find close-enough covering number upper bounds on the class of probabilities needed to get improved regret bounds. We are inspired from some of the ideas in Acharya et al. (2013) but we use a very different approach.

In the next sections, we give the necessary background for understanding our result, specifically on two concepts: patterns and regret.

1.1. The Class of Patterns and Profiles

Let $x^n = x_1, \ldots, x_n$ be a length n sequence which takes symbols from a set $\mathcal{X}$. Since we are interested in sequences over large alphabets, the set $\mathcal{X}$ can be arbitrarily large and possibly even infinite. The number of possible sequences x^n becomes arbitrarily large if the size of $\mathcal{X}$ is arbitrarily large, making it difficult (or nearly impossible) to store or enumerate all possible sequences. Instead, it is more tractable to enumerate the *patterns* associated with sequences. In this section, we will discuss patterns and a related quantity, profiles, which will be important for studying patterns.

Patterns are discussed in Jevtic et al. (2002); Orlitsky and Santhanam (2003, 2004); Orlitsky et al. (2004); Acharya et al. (2012); Acharya et al. (2013). Given a sequence x^n, its pattern is when each symbol is relabeled with a number based on when the symbol first appears relative to the other unique symbols. Orlitsky and Santhanam (2004) formally define patterns as the following: Let $\mathcal{A}(x^n)$ be the set of symbols appearing in x^n. Let the notation x_1^j mean the subsequence $x_1, \ldots, x_j$. The *index* of $x \in \mathcal{A}(x^n)$ is

$$i_{x^n}(x) \stackrel{\triangle}{=} \min\{|\mathcal{A}(x_1^j)| : 1 \leq j \leq n, x_j = x\}\,. \tag{4}$$

The pattern of x^n is given by

$$\mathsf{patt}(x^n) \stackrel{\triangle}{=} i_{x^n}(x_1) i_{x^n}(x_2) \ldots i_{x^n}(x_n)\,. \tag{5}$$

For example, $\mathsf{patt}(banana) = 123232$. The letter b is the first to appear, so $i_{banana}(b) = 1$. The letter a is second to appear, so $i_{banana}(a) = 2$. All a's in $banana$ are replaced with 2's.

When we also disregard the order of relabeled symbols in a pattern, we get a *profile*. A profile is the multiset of multiplicities of symbol counts for a given sequence. (Profiles will be the important quantity we will work with to determine regret of patterns, which we will discuss in Section 1.4.)

To define a profile, we will first define a type. A *type* is a vector which indicates the number of times each symbol appears in a sequence or a pattern. Suppose the alphabet size is k. For a sequence x^n let $\#\{i : x_i = j\}$ be the number of indices where $x_i = j$; then the type is a length k vector where

$$\mathsf{type}(x^n) \triangleq (\#\{i : x_i = 1\}, \#\{i : x_i = 2\}, \dots, \#\{i : x_i = k\}) \ . \tag{6}$$

As an example, if the alphabet size is $k = 6$, then $\mathsf{type}(123232) = (1, 3, 2, 0, 0, 0)$. This is because the symbol 1 appears once, the symbol 2 appears three times, the symbol 3 appears twice, and the symbols $4, 5, 6$ do not appear at all. (For more information on types, see Cover and Thomas (2006).) If we take the type of the pattern of x^n, since the length of x^n is n, the pattern of x^n has at most n different symbols. By default, we will assume that when we take a type on a pattern for any sequence x^n, that the alphabet size is n.

A profile of a sequence x^n is the multiset of values in $\mathsf{type}(\mathsf{patt}(x^n))$. A multiset does not specify any order on the values, however for the purposes of our notation, we will order the values in the multiset from greatest to least. This allows us to define a profile for a sequence x^n as

$$\mathsf{prof}(x^n) \triangleq \mathtt{SORT}(\mathsf{type}(\mathsf{patt}(x^n))) \ . \tag{7}$$

When we denote a profile, we will ignore the trailing zeros. For example, we will write that

$$\begin{aligned} \mathsf{prof}(banana) &= \mathtt{SORT}(\mathsf{type}(123232)) && (8)\\ &= \mathtt{SORT}((1, 3, 2, 0, 0, 0)) && (9)\\ &= \{3, 2, 1\} \ . && (10) \end{aligned}$$

In (9), we express the type of 123232 over an alphabet of size 6 since the pattern has length 6. Notice that in (10), as stated, we removed the zeros from the sorted multiset.

Let Φ^n be the set of all profiles for length n sequences. As an example, for length 4 sequences, the complete set of possible profiles are

$$\Phi^4 = \{\{4\}, \{3, 1\}, \{2, 2\}, \{2, 1, 1\}, \{1, 1, 1, 1\}\} \ . \tag{11}$$

Because the number of profiles for each length n is very limited, enumerating profiles is a much more feasible task than enumerating all the sequences. In the above example, for $n = 4$, there are 5 possible profiles, much less than the number of all the possible sequences, which is equal to the alphabet size (assuming it is finite, which it may not be) raised to the fourth power. In general, if the length of the sequence is n and alphabet size is k, the total number of sequences is k^n. The total number of types is $\binom{n+k-1}{k-1}$. The number of profiles is even smaller. One of the most important properties of profiles (and patterns) is that they do not depend on k. Thus, when studying large-alphabet distributions, using profiles can drastically reduce the complexity.

1.2. Induced Probabilities of Profiles

Suppose we have a distribution P_{X^n} on sequences x^n. We can then ask, if sequences x^n are drawn randomly according to P_{X^n}, what is the probability $\mathsf{prof}(x^n)$ is particular profile $\varphi \in \Phi^n$? This is the induced probability of a profile φ under P_{X^n}. Let this be denoted as $\mu_{P_{X^n}}(\varphi)$, where

$$\mu_{P_{X^n}}(\varphi) \triangleq \sum_{x^n : \mathsf{prof}(x^n) = \varphi} P_{X^n}(x^n) \ . \tag{12}$$

We will mostly be concerned with the case when P_{X^n} is an iid distribution. Let

$$P = (p(1), \dots, p(k)) \tag{13}$$

be a discrete distribution over an alphabet size of k (P is a probability on $[k] \triangleq \{1, \dots, k\}$), where $p(i)$ is the probability of symbol i occurring. For each P, we use $P^{\otimes n}$ to be the n-fold product distribution (iid distribution) on sequence x^n, e.g.

$$P^{\otimes n}(x^n) \triangleq \prod_{t=1}^{n} p(x_t)\,. \tag{14}$$

For later parts of this work, we will need to define classes of probability distributions. Our first class will be for profiles induced by iid distributions:

Definition 2 *Let the class of induced iid probabilities for length-n profiles be*

$$\mathcal{I}_\Psi^n = \{\mu_{P^{\otimes n}}(\cdot) : P \textit{ is a probability on } [n]\ \}\,. \tag{15}$$

Unfortunately, finding the induced distribution over profiles can be very complicated.

1.3. Regret for Online Learning

In this section, we discuss regret and redundancy.

In order to define regret in the context of learning problems, we first define log-loss (short for *logarithmic loss*, also called the *self-information loss* Merhav and Feder (1998)). When evaluating with log-loss, the task of the estimator is to pick a probability distribution Q on the possible outcome symbols in some set $\mathcal{X}$. Given the true symbol $x \in \mathcal{X}$, the log-loss of the estimator's prediction Q is

$$L(Q, x) \triangleq \log \frac{1}{Q(x)}\,. \tag{16}$$

While the above is the loss the estimator incurs, the log-loss alone does not characterize whether the estimator's prediction Q was 'good' or 'bad'. To measure how well the estimator's prediction is, one method is to compare the log-loss of the estimator's prediction Q to the true distribution P of x. This quantity is called *redundancy*. Redundancy for an estimator's prediction Q on random variable X, where $X \sim P$, is

$$\bar{R}(P, Q) \triangleq \mathbb{E}_{X \sim P}\left[L(Q, X) - L(P, X)\right] = \mathbb{E}_{X \sim P}\left[\log \frac{1}{Q(X)} - \log \frac{1}{P(X)}\right]\,. \tag{17}$$

We can also express $\bar{R}(P, Q) = D(P \| Q)$ where

$$D(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \tag{18}$$

is the Kullback-Leibler (KL) divergence between P and Q. To get the lowest possible redundancy, the estimator should give Q which represents her belief of the true distribution of x.

Let $\mathcal{I}$ be the class of possible P which can be the true distribution of x. Given an estimator's prediction Q, we let the *adversary* (or nature) choose the $P \in \mathcal{I}$ with the goal of maximizing the redundancy. The estimator meanwhile is trying to minimize the redundancy. This gives the *minimax redundancy*, first defined by Davisson (1973):

$$\bar{R}(\mathcal{I}) \stackrel{\triangle}{=} \inf_{Q} \sup_{P \in \mathcal{I}} \bar{R}(P, Q)\,. \tag{19}$$

Redundancy and minimax redundancy both rely on the fact that there is some true distribution P generating x which we are comparing to. The realizations of x which we are evaluating log-loss is averaged with respect to this P. Regret is analogous to redundancy, but removes the need to average over any distribution P; it is evaluated on an individual realization x. For this to work, the log-loss of the estimator is compared to the log-loss of the best choice of P in class $\mathcal{I}$ which minimizes $L(P, x)$. To define *minimax regret*, the adversary chooses the worst value of x to maximize the regret.

Definition 3 (Regret and Minimax Regret) *The regret of prediction Q, class $\mathcal{I}$ and outcome x is*

$$\hat{R}(Q, x, \mathcal{I}) = L(Q, x) - \sup_{P \in \mathcal{I}} L(P, x) = \sup_{P \in \mathcal{I}} \log \frac{P(x)}{Q(x)}\,. \tag{20}$$

The minimax regret for class $\mathcal{I}$ is

$$\hat{R}(\mathcal{I}) = \inf_{Q} \sup_{x} \hat{R}(Q, x, \mathcal{I}) = \inf_{Q} \sup_{x} \sup_{P \in \mathcal{I}} \log \frac{P(x)}{Q(x)}\,. \tag{21}$$

In some other works in the literature, redundancy as we have defined it is called *average-case redundancy* (since it averages over some distribution P) and regret is called *worst-case redundancy* (since it is evaluated for the worst realization).

Redundancy and regret are important concepts for universal compression in information theory Davisson (1973); Merhav and Feder (1998); Rissanen (1984). When the true distribution P is known, the entropy $H(P)$ is the expected number of bits needed to compress a random x generated from P. However, in many practical applications P is not known. The redundancy $\bar{R}(P, Q)$ represents the extra expected number of bits above $H(P)$ needed to represent x when the compressor uses Q instead of P to model the distribution of x. In this context, the minimax redundancy gives a worst-case bound on the expected excess code length independent of P. Minimax regret, which uses the worst-case x instead of averaging, is a more stringent value than redundancy. It upper bounds minimax redundancy and represents how many more bits are necessary to compress the worst possible value. Redundancy and regret also have connections to minimum description length Barron et al. (1998); Grünwald and Rissanen (2007), gambling Kelly (1956); Feder (1991); Cover and Ordentlich (1996); Xie and Barron (2000), and sequential prediction Merhav and Feder (1998); Cesa-Bianchi and Lugosi (2006). Our work is relevant to sequential prediction.

1.4. Regret and Redundancy in Sequential Prediction

In the setting of sequential prediction, the redundancy and regret measure the cumulative performance of an estimator in an online game. The estimator is tasked with predicting the probability of x^n one symbol at a time. At each time step t in the online game, the estimator must first give his/her best estimate of the probability distribution of symbol x_t, after which x_t is revealed. The estimator's predictions can depend on symbols revealed in the past. The *cumulative* loss of the estimator is the sum of the losses at each step. This cumulative loss can be expressed as the log loss on a joint distribution of the whole sequence of outcomes:

$$\log \frac{1}{Q_{X^n}(x^n)} \,. \tag{22}$$

Let $\mathcal{I}^n$ be (any arbitrary) class of probability distributions on sequences of length n. The cumulative loss (22) is compared against the log-loss of probability $P_{X^n} \in \mathcal{I}^n$. For redundancy, P_{X^n} is the true distribution and we average all possible sequences x^n over P_{X^n}. For regret, we use $P_{X^n} \in \mathcal{I}^n$ which gives the largest probability for the outcome x^n.

An important class studied extensively is the class of iid distributions on sequences, which we denote as $\mathcal{I}_k^n$, where k specifies the alphabet size and n specifies the sequence length. This is a class on probabilities over sequences x^n.

$$\mathcal{I}_k^n \triangleq \{P^{\otimes n} : P \text{ is a probability on } [k]\} \tag{23}$$

A long line of work which includes Krichevsky and Trofimov (1981); Shtarkov (1987); Clarke and Barron (1990, 1994); Shtarkov et al. (1995); Cover and Ordentlich (1996); Xie and Barron (2000) determined that

$$\bar{R}(\mathcal{I}_k^n) = \frac{k-1}{2} \log \frac{n}{2\pi e} + \log \frac{\Gamma^k\left(\frac{1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} + o_k(1) \tag{24}$$

$$\hat{R}(\mathcal{I}_k^n) = \frac{k-1}{2} \log \frac{n}{2\pi} + \log \frac{\Gamma^k\left(\frac{1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} + o_k(1) \,. \tag{25}$$

The formulas (24) and (25) are for when the alphabet size k is fixed and the sequence length n goes to infinity. They do not give satisfying enough solutions for when the alphabet size is large compared to the sequence length, which is the case of large-alphabet.

For the large-alphabet case, Davisson (1973) observed that while for finite alphabet, per-symbol minimax redundancy (this is the minimax redundancy divided by the sequence length n) can go to zero, for this to occur with infinite alphabet, more conditions need to be shown. Following this, Kieffer (1978) clarified what these conditions are and showed that for an iid source with an infinite alphabet, per-symbol redundancy does not go to zero. Orlitsky and Santhanam (2004) determined asymptotic values of minimax regret for iid distributions with various alphabet sizes k relative to n. They showed that $\hat{R}(\mathcal{I}_k^n) = \Theta(n)$ when $k = \Theta(n)$,

Because there is no diminishing per-symbol minimax redundancy in the large-alphabet case, an alternative direction in the study of redundancy and regret for large-alphabets is to consider patterns (or profiles) of infinite alphabets sequences.

Since iid distributions assign the same probability to patterns with the same profile, Orlitsky et al. (2004) showed that the minimax regret of the class of patterns is the same as the minimax regret of the class of profiles. Hence, we can work directly with profiles to find the minimax regret of patterns.

In order to find its redundancy and regret, Acharya et al. (2012); Acharya et al. (2013) work with a Poissonized process on profiles. Prior to their result, Orlitsky et al. (2004) determined an upper bound using a result of Hardy and Ramanujan (1918) on the number of integer partitions. This previous upper bound was that $\hat{R}(\mathcal{I}_\Psi^n) \le \pi\sqrt{2/3}n^{1/2}$.

2. Overview of Proof

The rest of this work is dedicated to proving Theorem 1. We split the proof into a few sections. In Section 3 we discuss KL divergence coverings, a key technique in our proof method. In Section 4, we show how the problem of minimax regret for the class of profiles is related to finding a specific divergence covering number. Some key ideas used in this part include the log-sum inequality to replace working with the complicated distributions over profiles, to working with simpler distributions. This turns our regret problem over profiles, to a regret problem over a special class of monotonic probability distributions. Another key idea is to use the Grenander estimator, which gives the maximum likelihood estimator when the estimators are restricted to be in this special monotonic class. In Section 5, we show the necessary divergence covering number result for our monotonic distributions class. This completes the proof.

3. Upper Bounds Using Divergence Covering

Let $\triangle_{k-1}$ be the probability simplex over size k alphabet. Our procedure for finding our minimax regret result focuses on using a covering on a subset of $\triangle_{k-1}$ under KL divergence (18). A *covering* is a set of points in a space (we will call them centers) for which all other points in the space are within a certain distance ε to. Traditionally, the distance is defined with as the Euclidean distance, but for our purposes, we need the distance to be KL divergence. Note that KL divergence is not a metric.

Definition 4 (KL Divergence Covering Number) *Let k be the alphabet size, $\varepsilon > 0$, and $\mathcal{B} \in \triangle_{k-1}$. Define*

$$M(k, \varepsilon, \mathcal{B}) = \inf\{m : \exists\{Q_{(1)}, ..., Q_{(m)} \in \triangle_{k-1}\} \; s.t \; \max_{P \in \mathcal{B}} \min_{Q_{(i)}} D(P||Q_{(i)}) \leq \varepsilon\} \,. \tag{26}$$

We call ε the *radius*. The elements $Q_{(1)}, ..., Q_{(m)}$ are the centers of the covering. We use $\mathcal{Q}$ to mean the set of centers. KL divergence covering is discussed in more detail in Tang (2022).

With a KL divergence covering, we can use a technique of Yang and Barron (1999) to get a bound on mutual information. Under an information theory interpretation, the minimax redundancy is equivalent to a channel capacity between the parameter space and observed sequence (first discovered by Gallager (1974)) which can be expressed also as a mutual information. This gives us a method to bound minimax redundancy with results from divergence covering bounds. To use divergence covering to get a bound on minimax regret, we need an additional step. This step is non-trivial since the class of induced iid distributions on profiles is complicated to work with. We show how to do this in the next section.

Covering numbers, using a different distance than KL divergence, are also used to bound minimax regret in Cesa-Bianchi and Lugosi (1999). Covering numbers for KL divergence are used to determine the capacity of noisy permutation channels in Tang and Polyanskiy (2022).

4. Regret on Monotonic Probabilities

Recall that the induced distribution on profiles sums over all sequences which map to a particularly profile (12). Since profiles disregard identities of symbols, if P_1 and P_2 are both probabilities on $[n]$ which are equivalent up to a permutation of the symbols, then $\mu_{P_1^{\otimes n}}$ and $\mu_{P_2^{\otimes n}}$ have the exact same value on every profile. Thus, to specify all the induced iid distributions on profiles of length n, it is enough to define the following class of *monotonic* probabilities:

$$\mathcal{P}^n_{\searrow} = \{\text{Probability } P = (p(1), \ldots, p(n)) \text{ on } [n] : p(1) \geq p(2) \geq \cdots \geq p(n)\} \; . \tag{27}$$

Each probability on profiles in $\mathcal{I}^n_\Psi$ (Definition 2) is induced by some probability in $\mathcal{P}^n_{\searrow}$.

Minimax regret on profiles is defined as the following, where the distribution μ can be any distribution on profiles.

$$\begin{aligned}\hat{R}(\mathcal{I}^n_\Psi) &= \inf_{\mu} \sup_{\varphi} \sup_{\mu_{P^{\otimes n}} \in \mathcal{I}^n_\Psi} \log \frac{\mu_{P^{\otimes n}}(\varphi)}{\mu(\varphi)} && (28)\\ &= \inf_{\mu} \sup_{\varphi} \sup_{\mu_{P^{\otimes n}} \in \mathcal{I}^n_\Psi} \frac{1}{\mu_{P^{\otimes n}}(\varphi)} \mu_{P^{\otimes n}}(\varphi) \log \frac{\mu_{P^{\otimes n}}(\varphi)}{\mu(\varphi)} \; . && (29)\end{aligned}$$

For an upper bound, we can restrict μ to distributions $\mu_{Q_{X^n}}$, those induced by some distribution Q_{X^n} on sequences. We chose this so that we can apply log-sum inequality next.

$$\begin{aligned}\hat{R}(\mathcal{I}^n_\Psi) &\leq \inf_{\mu_{Q_{X^n}}} \sup_{\varphi} \sup_{\mu_{P^{\otimes n}} \in \mathcal{I}^n_\Psi} \frac{1}{\mu_{P^{\otimes n}}(\varphi)} \mu_{P^{\otimes n}}(\varphi) \log \frac{\mu_{P^{\otimes n}}(\varphi)}{\mu_{Q_{X^n}}(\varphi)} && (30)\\ &= \inf_{Q_{X^n}} \sup_{\varphi} \sup_{P \in \mathcal{P}^n_{\searrow}} \frac{1}{\mu_{P^{\otimes n}}(\varphi)} \left(\sum_{x^n : \mathsf{prof}(x^n) = \varphi} P^{\otimes n}(x^n) \right) \log \frac{\sum_{x^n : \mathsf{prof}(x^n) = \varphi} P^{\otimes n}(x^n)}{\sum_{x^n : \mathsf{prof}(x^n) = \varphi} Q_{X^n}(x^n)} && (31)\\ &\leq \inf_{Q_{X^n}} \sup_{\varphi} \sup_{P \in \mathcal{P}^n_{\searrow}} \frac{1}{\mu_{P^{\otimes n}}(\varphi)} \sum_{x^n : \mathsf{prof}(x^n) = \varphi} P^{\otimes n}(x^n) \log \frac{P^{\otimes n}(x^n)}{Q_{X^n}(x^n)} && (32)\\ &\leq \inf_{Q_{X^n}} \sup_{\varphi} \sup_{P \in \mathcal{P}^n_{\searrow}} \frac{1}{\mu_{P^{\otimes n}}(\varphi)} \sum_{x^n : \mathsf{prof}(x^n) = \varphi} P^{\otimes n}(x^n) \left(\sup_{x^n} \log \frac{P^{\otimes n}(x^n)}{Q_{X^n}(x^n)} \right) && (33)\\ &= \inf_{Q_{X^n}} \sup_{P \in \mathcal{P}^n_{\searrow}} \sup_{x^n} \log \frac{P^{\otimes n}(x^n)}{Q_{X^n}(x^n)} \; . && (34)\end{aligned}$$

In order to evaluate the expression above, we need to determine which $P \in \mathcal{P}^n_{\searrow}$ gives the largest $P^{\otimes n}(x^n)$ for any x^n. For this we need the following:

Let $\boldsymbol{v} = (v(1), \ldots, v(n)) = \mathsf{type}(x^n)$. If $v(1) \geq v(2) \geq \cdots \geq v(n)$, then we know that the $P \in \mathcal{P}^n_{\searrow}$ which maximizes $P^{\otimes n}(x^n)$ is where $P = \boldsymbol{v}/n$. However, when $\boldsymbol{v}$ does not have this monotonic form, we need to determine which $P \in \mathcal{P}^n_{\searrow}$ is the maximum likelihood estimator for $\boldsymbol{v}$.

Define

$$P^{\searrow}_{ML}(x^n) = P^{\searrow}_{ML}(\boldsymbol{v}) = \max_{P = (p(1), \ldots, p(n)) \in \mathcal{P}^n_{\searrow}} \prod_{i=1}^{n} p(i)^{v(i)} \; . \tag{35}$$

Finding $P^{\searrow}_{ML}(x^n)$ is a well known problem. The solution is given by the Grenander estimator developed in Grenander (1956). The Grenander estimator gives us the following lemma:

Lemma 5 *Fix $\boldsymbol{v}$ to be the type of x^n (e.g. $\sum_{i=1}^n v(i) = n$). There exists a vector $\boldsymbol{u} = (u(1), \ldots, u(n))$ where*

- $u(1) \geq u(2) \geq ... \geq u(n)$
- $\sum_{i=1}^{n} u(i) = n$
- *Let* $\mathcal{J}(u_0)$ *be all indices* i *where* $u(i) = u_0$ *for some value* u_0*. Then*

$$u_0 = \frac{\sum_{i \in \mathcal{J}(u_0)} v(i)}{|\mathcal{J}(u_0)|} \tag{36}$$

so that

$$P^{\searrow}_{ML}(\boldsymbol{v}) = P^{\searrow}_{ML}(\boldsymbol{u})\,. \tag{37}$$

Note that (37) gives the likelihood of x^n where $\mathsf{type}(x^n) = \boldsymbol{v}$ under the maximum likelihood estimator. The maximum likelihood estimator itself is given by $\boldsymbol{u}$.

Expression (36) in Lemma 5 gives the local average property of the Grenander estimator. Because of the local average property we know that

$$u(i) = \frac{z}{r} \tag{38}$$

for $z \in \mathbb{Z}_{\geq 0}$ and $r \in [n-1]$. This is because $u(i)$ is an average of integers and in fact it is an average of less than n integers. If it were an average of all n integers, then $u(i) = 1$, so a denominator of n is not necessary. The proof of Lemma 5, more details on the Grenander estimator and the local average property is given in Appendix A.

There are two important points we need from Lemma 5. First, is the form of $\boldsymbol{u}$. For any x^n, let $\boldsymbol{v} = \mathsf{type}(x^n)$. Let $\boldsymbol{u}$ be the vector given by Lemma 5 for $\boldsymbol{v}$.

$$\sup_{P \in \mathcal{P}^n_{\searrow}} \log \frac{P^{\otimes n}(x^n)}{Q_{X^n}(x^n)} = \log \frac{P^{\searrow}_{ML}(\boldsymbol{v})}{Q_{X^n}(x^n)} = \log \frac{P^{\searrow}_{ML}(\boldsymbol{u})}{Q_{X^n}(x^n)} = \log \frac{\prod_{i=1}^{n} \left(\frac{u(i)}{n}\right)^{u(i)}}{Q_{X^n}(x^n)}\,. \tag{39}$$

Now suppose that there is a set $\mathcal{Q}$ which is a divergence covering of $\mathcal{P}^n_{\searrow}$ with radius ε. For any $\boldsymbol{u}$, we let $\widetilde{Q} \in \mathcal{Q}$ be the probability which covers $\boldsymbol{u}/n$, i.e.

$$\widetilde{Q} = \arg\min_{Q \in \mathcal{Q}} D(\boldsymbol{u}/n \| Q) \leq \varepsilon\,. \tag{40}$$

We will add an additional constraint on $\mathcal{Q}$: For $\widetilde{Q} = (\widetilde{q}(1), \ldots, \widetilde{q}(n)) \in \mathcal{Q}$ which covers $P = (p(1), \ldots, p(n))$, for indices i where $p(i) = p(i+1)$, we require that $\widetilde{q}(i) = \widetilde{q}(i+1)$.

Let $\bar{Q}$ be such that

$$\bar{Q}_{X^n}(x^n) = \frac{1}{|\mathcal{Q}|} \sum_{Q \in \mathcal{Q}} Q^{\otimes n}(x^n)\,. \tag{41}$$

While we do not know the value of Q_{X^n} which achieves the infimum in (34), we can instead use $\bar{Q}_{X^n}(x^n)$ and get an upper bound. Then (34) becomes

$$\inf_{Q_{X^n}} \sup_{P \in \mathcal{P}^n_{\searrow}} \sup_{x^n} \log \frac{P^{\otimes n}(x^n)}{Q_{X^n}(x^n)} \leq \sup_{x^n} \log \frac{\prod_{i=1}^n \left(\frac{u(i)}{n}\right)^{u(i)}}{\bar{Q}_{X^n}(x^n)} \tag{42}$$

$$= \sup_{x^n} \log \frac{\prod_{i=1}^n \left(\frac{u(i)}{n}\right)^{u(i)}}{\frac{1}{|\mathcal{Q}|} \sum_{Q \in \mathcal{Q}} Q^{\otimes n}(x^n)} . \tag{43}$$

Since $\widetilde{Q}^{\otimes n}(x^n) \leq \sum_{Q \in \mathcal{Q}} Q^{\otimes n}(x^n)$, we have

$$\log \frac{\prod_{i=1}^n \left(\frac{u(i)}{n}\right)^{u(i)}}{\frac{1}{|\mathcal{Q}|} \sum_{Q \in \mathcal{Q}} Q^{\otimes n}(x^n)} \leq \log |\mathcal{Q}| + \log \frac{\prod_{i=1}^n \left(\frac{u(i)}{n}\right)^{u(i)}}{\widetilde{Q}^{\otimes n}(x^n)} \tag{44}$$

$$= \log |\mathcal{Q}| + \log \frac{\prod_{i=1}^n \left(\frac{u(i)}{n}\right)^{u(i)}}{\prod_{i=1}^n \widetilde{q}(i)^{v(i)}} \tag{45}$$

$$= \log |\mathcal{Q}| + \log \frac{\prod_{i=1}^n \left(\frac{u(i)}{n}\right)^{u(i)}}{\prod_{i=1}^n \widetilde{q}(i)^{u(i)}} . \tag{46}$$

The inequality in (44) is the key inequality used in Yang and Barron (1999). In the last equality (46), we chose $\widetilde{Q}$ to cover $\boldsymbol{u}/n$. This means when $u(i) = u(j)$, we have that $\widetilde{q}(i) = \widetilde{q}(j)$. For the set of indices $\mathcal{J}$ where $u(i)$ are the same, we have that $\sum_{i \in \mathcal{J}} u(i) = \sum_{i \in \mathcal{J}} v(i)$. Thus, $\prod_{i \in \mathcal{J}} \widetilde{q}(i)^{v(i)} = \prod_{i \in \mathcal{J}} \widetilde{q}(i)^{u(i)}$. Continuing with (46),

$$\log |\mathcal{Q}| + \log \frac{\prod_{i=1}^n \left(\frac{u(i)}{n}\right)^{u(i)}}{\prod_{i=1}^n \widetilde{q}(i)^{u(i)}} = \log |\mathcal{Q}| + \sum_{i=1}^n u(i) \log \frac{\left(\frac{u(i)}{n}\right)}{\widetilde{q}(i)} \tag{47}$$

$$= \log |\mathcal{Q}| + n \sum_{i=1}^n \frac{u(i)}{n} \log \frac{\left(\frac{u(i)}{n}\right)}{\widetilde{q}(i)} \tag{48}$$

$$= \log |\mathcal{Q}| + nD(\boldsymbol{u}/n \| \widetilde{Q}) \tag{49}$$

$$= \log |\mathcal{Q}| + n\varepsilon \tag{50}$$

and thus

$$\hat{R}(I^n_{\Psi}) \leq \sup_{P \in \mathcal{P}^n_{\searrow}} \sup_{x^n} \log \frac{P^{\otimes n}(x^n)}{\bar{Q}_{X^n}(x^n)} \tag{51}$$

$$\leq \sup_{x^n} (\log |\mathcal{Q}| + n\varepsilon) \tag{52}$$

$$= \log |\mathcal{Q}| + n\varepsilon . \tag{53}$$

To get an upper bound on minimax regret for profiles (and patterns), it remains to find a covering of the space $\mathcal{P}^n_{\searrow}$ with the additional constraint that if $\widetilde{Q}$ covers P, $\widetilde{Q}$ is equal on indices for each

P has equal values on. The second fact we use from Lemma 5 is that the minimum non-zero value of $u(i)/n$ is some positive integer divided by $n(n-1)$. If $p(i)$ is not zero, then $p(i) > 1/n^2$. Our covering only needs to look at P of this type. We formally define this class of distributions which is a subset of the class $\mathcal{P}^n_{\searrow}$ but with a lower bound on the minimum non-zero value:

Definition 6 (Monotonic Class with Minimum) *Let*

$$\begin{aligned}\mathcal{P}^n_{\searrow}\{\alpha\} =\{&\textit{Probability } P = (p(1), \dots, p(n)) \textit{ on } [n] : p(1) \geq p(2) \geq \cdots \geq p(n) \\ &\textit{and } \forall i, p(i) \geq \alpha \textit{ or } p(i) = 0\}\,. \end{aligned} \tag{54}$$

5. Divergence Covering of Monotonic Distributions with a Minimum

For the next lemma, we will look at covering the subset $\mathcal{P}^n_{\searrow}\{1/n^2\}$ at the radius necessary for our result.

Lemma 7 *For the set of probabilities $\mathcal{P}^n_{\searrow}\{1/n^2\}$, we have that*

$$M\left(n, c_0 \frac{\log^{2/3} n}{n^{2/3}}, \mathcal{P}^n_{\searrow}\{1/n^2\}\right) = n^{c_1 n^{1/3} \log^{1/3} n} \tag{55}$$

for some absolute constants c_0 and c_1. This covering also satisfies the condition that if $P = (p(1), \dots, p(n))$ is covered by some $Q = (q(1), \dots, q(n))$ and $p(i) = p(i+1)$, then $q(i) = q(i+1)$.

This proof is inspired by some of the ideas Acharya et al. (2013) used to prove (2).

Proof

We want to choose a set $\mathcal{Q}$ of centers for a divergence covering for some appropriate radius ε on $\mathcal{P}^n_{\searrow}\{1/n^2\}$. We generate the centers for our covering by grouping 'similar' probability distributions together and then providing a center for each group. Each probability distribution P's values $p(x)$ will be divided among $t+1$ tiers $T_1, \dots, T_t$ and T_{zero} (we will define t later) according to their magnitude; this division defines the tier structure of distribution P, and distributions with the same tier structure (the same number of values in each tier) are 'similar' and hence placed in the same group. We define the tiers as follows:

For $i \in [t]$,

$$x \in T_i \text{ if } p(x) \in \left(\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^i, \left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^{i-1}\right] \tag{56}$$

and

$$x \in T_{\text{zero}} \text{ if } p(x) = 0\,. \tag{57}$$

To ensure that all x belong to some tier (recall either $p(x) \geq 1/n^2$ or $p(x) = 0$), it is sufficient that the last tier T_t is such that

$$\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^t \leq \frac{1}{n^2}\,. \tag{58}$$

If we let (ignoring integer constants, since at most it affects the result by a constant),

$$t = \frac{n^{1/3}\log(n^2)}{\log^{2/3} n} = 2n^{1/3}\log^{1/3} n \tag{59}$$

then

$$\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^{\frac{n^{1/3}\log(n^2)}{\log^{2/3} n}} \leq (e^{-1})^{\log(n^2)} \leq \frac{1}{n^2} \tag{60}$$

For each $P \in \mathcal{P}^n_{\searrow}\{1/n^2\}$, the tier structure of P is defined as $T(P) = (a_1, \ldots, a_t, a_{\text{zero}})$ where $a_j = |\{x : x \in T_j\}|$. Because P is sorted by value, knowing $T(P)$ specifies exactly which x belongs to each tier. If $T(P_1) = T(P_2)$, then P_1 and P_2 must have the same values of x in each tier (because $P_1(x)$ and $P_2(x)$ are sorted). The tier structure defines an equivalence class on the probabilities and we place P_1 and P_2 in the same group if $T(P_1) = T(P_2)$.

Let $\mathcal{G}$ be the set of all groups. We will assign one $Q^G \in \mathcal{Q}$ to each group $G \in \mathcal{G}$. We want to have that $\forall P \in G$, the Q^G assigned is such that

$$D(P||Q^G) \leq 30\frac{\log^{4/3} n}{n^{2/3}}\,. \tag{61}$$

and thus set $\varepsilon = 30\frac{\log^{4/3} n}{n^{2/3}}$.

To define the probability $Q^G = (q^G(1), \ldots, q^G(n))$ for each group G, we pick any (arbitrary) $P^G = (p^G(1), \ldots, p^G(n)) \in G$ and let

$$q^G(x) = \frac{1}{|T_j|}\sum_{y \in T_j} p^G(y) \text{ if } x \in T_j \tag{62}$$

for all $x \in [n]$. Since for every $P \in G$, the same symbols x are in the same tiers, if $x \in T_i$, then for all $P \in G$ (including P^G)

$$p(x) \in \left(\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^i, \left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^{i-1}\right]\,. \tag{63}$$

and thus

$$q^G(x) \in \left(\left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^i, \left(1 - \frac{\log^{2/3} n}{n^{1/3}}\right)^{i-1}\right]\,. \tag{64}$$

(So $p(x)$ and $q^G(x)$ should be pretty close.) For $x \in T_{\text{zero}}$, $q^G(x) = p(x) = 0$ for every $p \in G$. For every $p \in G$, if $p(x) = p(y)$, then x and y will be in the same tier. If x and y are in the same tier, then $q^G(x) = q^G(y)$.

For each Q^G, we want compute the KL divergence to any $P \in G$. We need to define some helpful quantities first. Fix some P. For each tier T_i, for $i \in \{[t] \cup \text{zero}\}$, define

$$\beta_i(P) = \sum_{x \in T_i} p(x)\,. \tag{65}$$

Naturally $\sum_{i\in\{[t]\cup\text{zero}\}}\beta_i(P)=1$.

Let $m_i(P)$ be the number of symbols in T_i. Then for $i\in[t]$, we can bound

$$m_i(P)\le\frac{\beta_i(P)}{\left(1-\frac{\log^{2/3}n}{n^{1/3}}\right)^i}\,. \tag{66}$$

This uses the fact that the smallest possible value of $p(x)$ is where $x\in T_i$ is given by $\left(1-\frac{\log^{2/3}n}{n^{1/3}}\right)^i$. Now we can compute the KL divergence.

$$\begin{aligned} D(P\|Q^G) &\le \sum_x \frac{(p(x)-q^G(x))^2}{q^G(x)} && (67)\\ &\le \sum_{i\in[t]} m_i(P)\frac{\left(\left(1-\frac{\log^{2/3}n}{n^{1/3}}\right)^i-\left(1-\frac{\log^{2/3}n}{n^{1/3}}\right)^{i-1}\right)^2}{\left(1-\frac{\log^{2/3}n}{n^{1/3}}\right)^i}\,. && (68)\end{aligned}$$

In (67) we use an upper bound on KL divergence given in Csiszar and Talata (2006). Next we use (66).

$$\begin{aligned} D(P\|Q^G) &\le \sum_{i\in[t]}\frac{\beta_i(P)}{\left(1-\frac{\log^{2/3}n}{n^{1/3}}\right)^i}\frac{\left(1-\frac{\log^{2/3}n}{n^{1/3}}\right)^{2i-2}\left(1-\frac{\log^{2/3}n}{n^{1/3}}-1\right)^2}{\left(1-\frac{\log^{2/3}n}{n^{1/3}}\right)^i} && (69)\\ &=\sum_{i\in[t]}\beta_i(P)\frac{\left(-\frac{\log^{2/3}n}{n^{1/3}}\right)^2}{\left(1-\frac{\log^{2/3}n}{n^{1/3}}\right)^2} && (70)\\ &=\left(\sum_{i\in[t]}\beta_i(P)\right)\left(\frac{\left(-\frac{\log^{2/3}n}{n^{1/3}}\right)^2}{\left(1-\frac{\log^{2/3}n}{n^{1/3}}\right)^2}\right) && (71)\\ &=\frac{\log^{4/3}n}{n^{2/3}\left(1-\frac{\log^{2/3}n}{n^{1/3}}\right)^2}\,. && (72)\end{aligned}$$

The quantity $1/\left(1-\frac{\log^{2/3}n}{n^{1/3}}\right)^2$ in (72) has a maximum occurring at $n=7$, where

$$\frac{1}{\left(1-\frac{\log^{2/3}n}{n^{1/3}}\right)^2}\le 29.1542\,. \tag{73}$$

This shows (61).

Next, we want to count the total number of groups, i.e. determine $|\mathcal{G}|$. We do this by counting the number of possible values of $T(P)$. There are $t+1$ different tier levels. While it is true that tiers corresponding to larger values of $p(x)$ cannot possibly contain too many x's, for the purposes of getting an upper bound, we assume the values a_i in each $T(P)$ can be of any value between 0 and n. Then

$$
\begin{aligned}
|\mathcal{G}| &= \binom{n+1+t+1}{t+1} && (74)\\
&\leq \left(\frac{e(n+1+t+1)}{t+1}\right)^{t+1} && (75)\\
&\leq \left(\frac{e(n+2n^{1/3}\log^{1/3} n+2)}{2n^{1/3}\log^{1/3} n+1}\right)^{2n^{1/3}\log^{1/3} n+1} && (76)\\
&\leq (e2n)^{2n^{1/3}\log^{1/3} n+1} && (77)\\
&\leq n^{8n^{1/3}\log^{1/3} n} && (78)
\end{aligned}
$$

Finally, $|\mathcal{Q}| = |\mathcal{G}|$, because we have one center for every group.

■

To prove Theorem 1:

Using (53) and Lemma 7,

$$
\begin{aligned}
\hat{R}(\mathcal{I}_\Psi^n) &\leq \log|\mathcal{Q}| + n \min_{Q\in\mathcal{Q}} \max_{P\in\mathcal{P}_{\searrow}^n\{1/n^2\}} D(P||Q) && (79)\\
&= \log\left(n^{8n^{1/3}\log^{1/3} n}\right) + n\cdot 30\frac{\log^{2/3} n}{n^{2/3}} && (80)\\
&= 8n^{1/3}\log^{4/3} n + 30n^{1/3}\log^{2/3} n && (81)\\
&\leq 38n^{1/3}\log^{4/3} n\,. && (82)
\end{aligned}
$$

Remark 8 *The constant term in (82) is generous since it is an upper bound for all n. In the limit as $n\to\infty$, we can replace the constant with 2.*

Acknowledgments

We thank Yury Polyanskiy for bringing this problem to our attention. We also thank the anonymous reviewers for their very helpful comments and suggestions, many of which we implemented. This work was supported in part by the NSF grant CCF-2131115.

References

J. Acharya, H. Das, and A. Orlitsky. Tight bounds on profile redundancy and distinguishability. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Tight bounds for universal compression of large alphabets. In *2013 IEEE International Symposium on Information Theory*, pages 2875–2879, 2013.

A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.

N. Cesa-Bianchi and G. Lugosi. Minimax regret under log loss for general classes of experts. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, COLT '99, pages 12–18, New York, NY, USA, 1999. Association for Computing Machinery.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, USA, 2006.

B.S. Clarke and A.R. Barron. Information-theoretic asymptotics of bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.

B.S. Clarke and A.R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1):37–60, 1994.

T.M. Cover and E. Ordentlich. Universal portfolios with side information. *IEEE Transactions on Information Theory*, 42(2):348–363, 1996.

T.M. Cover and J.A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

I. Csiszar and Z. Talata. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Transactions on Information Theory*, 52(3):1007–1016, March 2006.

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

M. Feder. Gambling using a finite state machine. *IEEE Transactions on Information Theory*, 37(5): 1459–1465, 1991.

R.G. Gallager. Source coding with side information and universal coding. Unpublished manuscript; also presented at the International Symposium on Information Theory (ISIT) Oct. 1974, 1974.

U. Grenander. On the theory of mortality measurement. *Scandinavian Actuarial Journal*, 1956(2): 125–153, 1956.

P.D. Grünwald and J. Rissanen. *The Minimum Description Length Principle*. Adaptive computation and machine learning. MIT Press, 2007.

G. H. Hardy and S. Ramanujan. Asymptotic formulae in combinatory analysis. *Proceedings of the London Mathematical Society*, s2-17(1):75–115, 1918.

H.K. Jankowski and J.A. Wellner. Estimation of a discrete monotone distribution, 2009.

N. Jevtic, A. Orlitsky, and N. Santhanam. Universal compression of unknown alphabets. In *Proceedings IEEE International Symposium on Information Theory,*, pages 320–, 2002.

J.L. Kelly. A new interpretation of information rate. *The Bell System Technical Journal*, 35(4): 917–926, 1956.

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

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

N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6): 2124–2147, 1998.

A. Orlitsky and N.P. Santhanam. Performance of universal codes over infinite alphabets. In *Data Compression Conference, 2003. Proceedings. DCC 2003*, pages 402–410, 2003.

A. Orlitsky and N.P. Santhanam. Speaking of infinity [i.i.d. strings]. *IEEE Transactions on Information Theory*, 50(10):2215–2230, 2004.

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

J. Rissanen. Universal coding, information, prediction, and estimation. *Information Theory, IEEE Transactions on*, 30:629 – 636, 08 1984.

T. Robertson. *Order restricted statistical inference*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, Chichester, 1988.

Y. Shtarkov. Universal sequential coding of single messages. *Probl. Peredachi Inf.*, 23:3–17, 1987.

Y. Shtarkov, T. Tjalkens, and F. Willems. Multialphabet universal coding of memoryless sources. *Problems of Information Transmission*, 31:114–127, 1995.

J. Tang. *Divergence Covering*. PhD thesis, Massachusetts Institute of Technology, 2022.

J. Tang and Y. Polyanskiy. Capacity of noisy permutation channels. In *(to appear) 2022 IEEE International Symposium on Information Theory (ISIT)*, 2022.

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

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

Appendix A. Grenander Estimator

Suppose we draw n samples from a discrete set $\mathcal{X} = [k]$. In these draws, symbol i occurs $v(i)$ times out of the n draws. The maximum likelihood estimator (MLE) for the probability of the n draws is defined as

$$P_{\text{MLE}} \triangleq \arg\max_P \prod_{i=1}^{k} p(i)^{v(i)} . \tag{83}$$

A basic fact is that the empirical distribution of the samples gives the MLE, that is, if $P_{\text{MLE}} = (p_{\text{MLE}}(1), \ldots, p_{\text{MLE}}(n))$, then for each $i \in \mathcal{X}$ we have $p_{\text{MLE}}(i) = v(i)/n$.

But suppose that we want the MLE but with an additional *shape constraint*. Let this constraint be the *monotone* constraint given by $\mathcal{P}^k_{\searrow}$ (see (27)), where we must have

$$p_{\text{MLE}}(1) \geq p_{\text{MLE}}(2) \geq \cdots \geq p_{\text{MLE}}(k)\,. \tag{84}$$

The P_{MLE} meeting this constraint for these n discrete samples turns out to be the Grenander estimator, developed in Grenander (1956). The Grenander estimator is formally described as the left derivative of the least concave majorant (LCM) of the cumulative distribution of the data points. (This property applies to both continuous and discrete data, though we are only concerned with the discrete case.) For our discrete data problem, to define the Grenander estimator, we use the notation of Robertson (1988) and first define the function

$$W(i) = \sum_{j=1}^{i} \frac{v(j)}{n}\,. \tag{85}$$

Define $W(0) = 0$.

Let $W^*(i)$ be the infimum, at each point i (integer or non-integer), of all concave functions which lie entirely above W (this is equivalent to W^* being the the upper convex hull of W). The function W^* is the LCM. Since W^* is concave, there are well-defined left-derivatives at each point j which we denote as $w^*(i)$.

We make the following observations about w^* at integer points i:

- If $W(i) < W^*(i)$, then $w^*(i) = w^*(i+1)$. If $W(i) = W^*(i)$, due to concavity, $w^*(i) \geq w^*(i+1)$. Thus, we must have that $w^*(i) \geq w^*(i+1)$ for all i.
- $W(k) = W^*(k)$ and $W(0) = W^*(0)$.
- We can compute $w^*(i) = W^*(i) - W^*(i-1)$. This gives that $\sum_{j=1}^{k} w^*(j) = \sum_{i=1}^{k} W^*(i) - W^*(i-1) = W^*(k) = 1$. Hence, $\boldsymbol{w}^* = (w^*(1), \ldots, w^*(k))$ is a probability distribution on $[k]$.

The Grenander estimator is $\boldsymbol{w}^*$.

Proposition 9 *The MLE P_{MLE} with constraint (84) is equal to $\boldsymbol{w}^*$.*

Proposition 9 was shown in Grenander (1956) for continuous data points (where Grenander used it to study estimating the laws of mortality.) Proposition 9 is given in Jankowski and Wellner (2009) specifically for discrete data. The Grenander estimator on discrete data is equivalent to the isotonic regression given in Robertson (1988). Modification of the proof in Robertson (1988) can be used to prove Proposition 9.

The form of $\boldsymbol{w}^*$ naturally gives it the *local average property*. This property is described below:

Corollary 10 *If the set $\mathcal{J}$ is a maximal set of adjacent symbols i that have the same value of $w^*(i)$, then*

$$\sum_{i \in \mathcal{J}} w^*(i) = \sum_{i \in \mathcal{J}} \frac{v(i)}{n} \tag{86}$$

$$\implies w^*(i) = \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} \frac{v(i)}{n} \tag{87}$$

The value of $w^*(i)$ is the average of the frequency, $v(i)/n$, over an adjacent set of values, hence why it is a local average.

Proof For a set $\mathcal{J}$, let i_a be the first (smallest) index in the set and i_b be the last (largest) index in the set. All points in $\mathcal{J}$ have the same left-derivative and therefore W^* between $i_a - 1$ and i_b is linear. At $i_a - 1$ and i_b, the slope of W^* changes. Since W^* is an upper convex hull, it must be that $W(i_a - 1) = W^*(i_a - 1)$ and $W(i_b) = W^*(i_b)$. The slope between $i_a - 1$ and i_b is then

$$\frac{W^*(i_b) - W^*(i_a)}{i_b - i_a + 1} = \frac{W(i_b) - W(i_a - 1)}{i_b - i_a + 1} = \frac{\sum_{i=i_a}^{i_b} \frac{v(i)}{n}}{|\mathcal{J}|} \tag{88}$$

which is the value of $w^*(i)$ for all $i \in \mathcal{J}$. ■

Proof of Lemma 5

Let $\boldsymbol{w}^*$ be the Grenander estimator with $k = n$ and let $\boldsymbol{u} = n\boldsymbol{w}^*$. Based on the properties of $\boldsymbol{w}^*$, we can directly get that $u(1) \geq \cdots \geq u(n)$ and $\sum_{i=1}^{n} u(i) = n$. We get (36) from Corollary 10.

Since $\boldsymbol{w}^* = \boldsymbol{u}/n$ is the MLE under the shape constraint as given by Proposition 9,

$$P^{\searrow}_{ML}(\boldsymbol{v}) = \prod_{i=1}^{n} \left(\frac{u(i)}{n}\right)^{v(i)} = \prod_{u}^{n} \left(\frac{u}{n}\right)^{\sum_{i\in\mathcal{J}(u)} v(i)} \tag{89}$$

Next we apply (36):

$$\prod_{u}^{n} \left(\frac{u}{n}\right)^{\sum_{i\in\mathcal{J}(u)} v(i)} = \prod_{u}^{n} \left(\frac{u}{n}\right)^{\sum_{i\in\mathcal{J}(u)} u} = \prod_{i=1}^{n} \left(\frac{u(i)}{n}\right)^{u(i)} = P^{\searrow}_{ML}(\boldsymbol{u})\,. \tag{90}$$

This shows (37). ■