

DRDO–IISc Programme on
Advanced Research in Mathematical
Engineering

Guessing Under Source Uncertainty

(TR-PME-2006-11)
by
Rajesh Sundaresan

Department of Electrical Communication Engineering, Indian Institute of Science,
Bangalore

21 August 2006



Indian Institute of Science
Bangalore 560 012

Guessing Under Source Uncertainty

Rajesh Sundaresan

21 August 2006

ABSTRACT

This paper considers the problem of guessing the realization of a finite alphabet source when some side information is provided. The only knowledge the guesser has about the source and the correlated side information is that the joint source is one among a family. A notion of redundancy is first defined and a new divergence quantity that measures this redundancy is identified. This divergence quantity shares the Pythagorean property with the Kullback-Leibler divergence. Good guessing strategies that minimize the supremum redundancy (over the family) are then identified. The min-sup value measures the richness of the uncertainty class. The min-sup redundancies for two examples - the classes of discrete memoryless sources and finite-state arbitrarily varying sources - are then determined.

Keywords:

f -divergence, guessing, I -projection, mismatch, Pythagorean identity, redundancy, Rényi entropy, Rényi information divergence, side information

1 Introduction

Let X be a random variable on a finite set \mathbb{X} with probability mass function (PMF) given by $(P(x) : x \in \mathbb{X})$. Suppose that we wish to guess the realization of this random variable X by asking questions of the form “Is X equal to x ?", stepping through the elements of \mathbb{X} , until the answer is “Yes" ([1], [2]). If we know the PMF P , the best strategy is to guess in the decreasing order of P -probabilities.

The aim of this paper is to identify good guessing strategies and analyze their performance when the PMF P is not completely known. Throughout this paper, we will assume that the only information available to the guesser is that the PMF of the source is one among a class \mathbb{T} of PMFs.

By way of motivation, consider a crypto-system in which Alice wishes to send to Bob a secret message. The message is encrypted using a private key stream. Alice and Bob share this private key stream. The key stream

is generated using a random and perhaps biased source. The cipher-text is transmitted through a public channel. Eve, the eavesdropper, guesses one key stream after another until she arrives at the correct message. Eve can guess any number of times, and she knows when she has guessed right. She might know this, for example, when she obtains a meaningful message. From Alice's and Bob's points of view, how good is their key stream generating source? In particular, what is the minimum expected number of guesses that Eve would need to get to the correct realization? From Eve's point of view, what is her best guessing strategy? These questions were answered by Arikan in [2].

Taking this example a step further, suppose that Alice and Bob have access to a few sources. How can they utilize these sources to increase the expected number of guesses Eve will need to identify the realization? What is Eve's guessing strategy? We answer these questions in this paper.

When P is known, Massey [1] and Arikan [2] sought to lowerbound the minimum expected number of guesses. For a given guessing strategy G , let $G(x)$ denote the number of guesses required when $X = x$. The strategy that minimizes $E[G(X)]$, the expected number of guesses, proceeds in the decreasing order of P -probabilities. Arikan [2] showed that the exponent of the minimum value, *i.e.*, $\log [\min_G E[G(X)]]$, satisfies

$$H_{1/2}(P) - \log(1 + \ln |\mathbb{X}|) \leq \log \left[\min_G E[G(X)] \right] \leq H_{1/2}(P),$$

where $H_\alpha(P)$ is the Rényi entropy of order $\alpha > 0$.

For $\rho > 0$, Arikan [2] also considered minimization of $(E[G(X)^\rho])^{1/\rho}$ over all guessing strategies G ; the exponent of the minimum value satisfies

$$H_\alpha(P) - \log(1 + \ln |\mathbb{X}|) \leq \frac{1}{\rho} \log \left[\min_G E[G(X)^\rho] \right] \leq H_\alpha(P), \quad (1)$$

where $\alpha = 1/(1 + \rho)$.

Arikan [2] applied these results to a discrete memoryless source on \mathbb{X} with letter probabilities given by the PMF P , and obtained that the minimum guessing moment, $\min_G E[G(X^n)^\rho]$, grows exponentially with n . The minimum growth rate of this quantity (after normalization by ρ) is given by the Rényi entropy $H_\alpha(P)$. This gave an operational significance for the Rényi entropy. In particular, the minimum expected number of guesses grows exponentially with n and has a minimum growth rate of $H_{1/2}(P)$.

Suppose now that the guesser only knows that the source belongs to a set \mathbb{T} of PMFs. The uncertainty set may be finite or infinite in size. The guesser's strategy should not be tuned to any one particular PMF in \mathbb{T} , but should be designed for the entire uncertainty set. The performance of such a guessing strategy on any particular source will not be better than the optimal strategy for that source. Indeed, for any source P , the exponent of $\mathbb{E}[G(X)^\rho]$

is at least as large as that of the optimal strategy $\mathbb{E}[G_P(X)^\rho]$, where G_P is the guessing strategy matched to P that guesses in the decreasing order of P -probabilities. Thus for any given strategy, and for any source $P \in \mathbb{T}$, we can define a notion of *penalty* or *redundancy*, $R(P, G)$, given by

$$R(P, G) = \frac{1}{\rho} \log \mathbb{E}[G(X)^\rho] - \frac{1}{\rho} \log \mathbb{E}[G_P(X)^\rho],$$

which represents the increase in the exponent of the guessing moment after an appropriate normalization by ρ .

A natural means of measuring the effectiveness of a guessing strategy G on the set \mathbb{T} is to find the worst redundancy over all sources in \mathbb{T} . In this paper, we are interested in identifying the value of

$$\min_G \sup_{P \in \mathbb{T}} R(P, G),$$

and in obtaining the G that attains this min-sup value.

We first show that $R(P, G)$ is bounded on either side in terms of a divergence quantity $L_\alpha(P, Q_G)$; Q_G is a PMF that depends on G , and L_α is a measure of dissimilarity between two PMFs. The above observation enables us to transform the min-sup problem above into another one of identifying

$$\inf_Q \sup_{P \in \mathbb{T}} L_\alpha(P, Q).$$

The role of L_α in guessing is similar to the role of Kullback-Leibler divergence in mismatched source compression. The parameter α is given by $\alpha = 1/(1+\rho)$. The quantity L_α is such that the limiting value as $\alpha \rightarrow 1$ is the Kullback-Leibler divergence. Furthermore, they share the Pythagorean property with the Kullback-Leibler divergence [3]. The results of this paper thus generalize the “geometric” properties satisfied by the Kullback-Leibler divergence [3].

Consider the special case of guessing an n -string output by a discrete memoryless source (DMS) with single letter alphabet \mathbb{A} . The parameters of this DMS are unknown to the guesser. Arikan and Merhav [4] proposed a “universal” guessing strategy for the class of DMSs on \mathbb{A} . This universal guessing strategy asymptotically achieves the minimum growth exponent for all sources in the uncertainty set. Their strategy guesses in the increasing order of empirical entropy. In the language of this paper, their results imply that the normalized redundancy suffered by the aforementioned strategy is upper-bounded by a positive sequence of real numbers that vanishes as $n \rightarrow \infty$. One can interpret this fact as follows: the class of discrete memoryless sources is not “rich” enough; we have a universal guessing strategy that is asymptotically optimal.

The redundancy quantities studied in this paper also arise in the study of mismatch in Campbell’s minimum average exponential coding length problem [5], [6]. Fischer [7] addressed the same problem in the context of mismatched source coding and identified the supremum average exponential

coding length for a class of sources. In particular, he showed that the supremum value is the supremum of the Rényi entropies of the sources in the class. In contrast, our focus in this paper is on identifying the worst *redundancy* suffered by a code.

Most of the results obtained in this paper were inspired by similar results for mismatched and universal source compression ([8], [9], [10]). We now highlight some comparisons between source compression and guessing.

Suppose that the source outputs an n -string of bits. In lossless source compression, one can think of an encoding scheme as asking questions of the form, “Does $X^n \in E_i$?” where $(E_i : i = 1, 2, \dots)$ is a carefully chosen sequence of subsets of \mathbb{X}^n . More specifically, one can ask the questions “Is $X_1 = 0$?", “Is $X_2 = 0$?", and so on. The goal is to minimize the number of such questions one needs to ask (on the average) to get to the realization. The minimum expected number of questions one can hope to ask (on the average) is the Shannon entropy $H(P)$. In the context of guessing, one can only test an entire string in one attempt, *i.e.*, ask questions of the form “Is $X^n = x^n$?” The guessing moment grows exponentially with n and the minimum exponent after scaling by ρ is given by the Rényi entropy $H_\alpha(P)$.

The quantity L_α plays the same role as Kullback-Leibler divergence does in mismatched source compression. L_α shares the Pythagorean property with the Kullback-Leibler divergence [11]. Moreover, the best guessing strategy is based on a PMF that is a mixture of sources in the uncertainty class, analogous to the source compression case. The min-sup value of redundancy for the problem of compression under source uncertainty is given by the capacity of a channel [9] with inputs corresponding to the indices of the uncertainty set, and channel transition probabilities given by the various sources in the uncertainty set. We show that a similar result holds for guessing under source uncertainty. In particular, the min-sup value is the *channel capacity of order $1/\alpha$* [12] of an appropriately defined channel.

The following is an outline of the paper. In Section 2 we review known results for the problem of guessing, introduce the relevant measures that quantify redundancy, and show the relationship between this redundancy and the divergence quantity L_α . In Section 3, we see how the same quantities arise in the context of Campbell’s minimum average exponential coding length problem. In Section 4, we pose the min-sup problem of quantifying the worst-case redundancy and identify another inf-sup problem in terms L_α . In Section 5 we undertake a systematic study of the properties of L_α divergence. In particular, we show the Pythagorean property and identify the so-called *centers* and *radii* of finite as well as infinite uncertainty classes. In Section 6, we specialize our results to two examples: the class of discrete memoryless sources on finite alphabets, and the class of finite-state arbitrarily varying sources. We establish results on the asymptotic redundancies of these two uncertainty classes. In Section 7 we make some concluding remarks.

2 Inaccuracy and redundancy in guessing

In this section, we prove previously known results in guessing. Our aim is to motivate the study of quantities that measure inaccuracy in guessing. In particular, we introduce a measure of divergence, and show how it is related to the α -divergence of Csiszár [12].

Let \mathbb{X} and \mathbb{Y} be a finite alphabet sets. Consider a correlated pair of random variables (X, Y) with joint PMF P on $\mathbb{X} \times \mathbb{Y}$. Given side information $Y = y$, we would like to guess the realization of X . Formally, a guessing list G with side information is a function $G : \mathbb{X} \times \mathbb{Y} \rightarrow \{1, 2, \dots, |\mathbb{X}|\}$ such that for each $y \in \mathbb{Y}$, the function $G(\cdot, y) : \mathbb{X} \rightarrow \{1, 2, \dots, |\mathbb{X}|\}$ is a one-to-one function that denotes the order in which the elements of \mathbb{X} will be guessed when the guesser observes $Y = y$. Naturally, knowing the PMF P , the best strategy which minimizes the expected number of guesses, given $Y = y$, is to guess in the decreasing order of $P(\cdot, y)$ -probabilities. Let us denote such an order G_P . Due to lack of exact knowledge of P , suppose we guess in the decreasing order of probabilities of another PMF Q . This situation leads to mismatch. In this section, we analyze the performance of guessing strategies under mismatch.

In some of the results we will have $\rho > 0$, and in others $\rho > -1, \rho \neq 0$. The $\rho > 0$ case is of primary interest in the context of guessing. The other case is also of interest in Campbell's average exponential coding length problem where similar quantities are involved.

Following the proof in [2], we have the following simple result for guessing under mismatch.

Proposition 1 (Guessing under mismatch) *Let $\rho > 0$. Consider a source pair (X, Y) with PMF P . Let Q be another PMF with $\text{Supp}(Q) = \mathbb{X} \times \mathbb{Y}$. Let G_Q be the guessing list with side information Y obtained under the assumption that the PMF is Q , with ties broken using an arbitrary but fixed rule. Then the guessing moment for the source with PMF P under G_Q satisfies*

$$\begin{aligned} & \frac{1}{\rho} \log (\mathbb{E} [G_Q(X, Y)^\rho]) \\ & \leq \frac{1}{\rho} \log \left(\sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} P(x, y) \left[\sum_{a \in \mathbb{X}} \left(\frac{Q(a, y)}{Q(x, y)} \right)^{\frac{1}{1+\rho}} \right]^\rho \right), \end{aligned} \tag{2}$$

where the expectation \mathbb{E} is with respect to P . \square

Proof: For $\rho > 0$, for each $y \in \mathbb{Y}$, observe that

$$\begin{aligned} G_Q(x, y) &\leq \sum_{a \in \mathbb{X}} 1\{Q(a, y) \geq Q(x, y)\} \\ &\leq \sum_{a \in \mathbb{X}} \left(\frac{Q(a, y)}{Q(x, y)} \right)^{\frac{1}{1+\rho}}, \end{aligned}$$

for each $x \in \mathbb{X}$, which leads to the proposition. \blacksquare

For a source P on $\mathbb{X} \times \mathbb{Y}$, the conditional Rényi entropy of order α , with $\alpha > 0$, is given by

$$H_\alpha(P) = \frac{\alpha}{1-\alpha} \log \left(\sum_{y \in \mathbb{Y}} \left(\sum_{x \in \mathbb{X}} P(x, y)^\alpha \right)^{1/\alpha} \right).$$

For the case when $|\mathbb{Y}| = 1$, *i.e.*, when there is no side information, we may think of P as simply a PMF on \mathbb{X} . The above conditional Rényi entropy of order α is then the Rényi entropy of order α of the source P , given by

$$H_\alpha(P) = \frac{1}{1-\alpha} \log \left(\sum_{x \in \mathbb{X}} P(x)^\alpha \right).$$

It is known that

$$0 \leq H_\alpha(P) \leq \log |\mathbb{X}|. \quad (3)$$

Suppose that our guessing order is “matched” to the source, *i.e.*, we guess according to the list G_P . We then get the following corollary.

Corollary 2 (*Matched guessing, Arikan [2]*) *Under the hypotheses in Proposition 1, the guessing strategy G_P satisfies*

$$\frac{1}{\rho} \log (\mathbb{E}[G_P(X, Y)^\rho]) \leq H_\alpha(P), \quad (4)$$

where $\alpha = 1/(1+\rho)$. \square

Proof: Set $Q = P$ in Proposition 1. \blacksquare

Let us now look at the converse direction.

Proposition 3 (*Converse*) *Let $\rho > 0$. Consider a source pair (X, Y) with PMF P . Let G be an arbitrary guessing list with side information Y . Then,*

there is a PMF Q_G on $\mathbb{X} \times \mathbb{Y}$ with $\text{Supp}(Q_G) = \mathbb{X} \times \mathbb{Y}$, and

$$\begin{aligned} & \frac{1}{\rho} \log (\mathbb{E} [G(X, Y)^\rho]) \\ & \geq \frac{1}{\rho} \log \left(\sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} P(x, y) \left[\sum_{a \in \mathbb{X}} \left(\frac{Q_G(a, y)}{Q_G(x, y)} \right)^{\frac{1}{1+\rho}} \right]^\rho \right) \\ & \quad - \log(1 + \ln |\mathbb{X}|), \end{aligned} \tag{5}$$

where the expectation \mathbb{E} is with respect to P . \square

Proof: The proof is very similar to that of [2, Theorem 1]. Observe that because $\rho > 0$, for each $y \in \mathbb{Y}$, we have

$$\sum_{x \in \mathbb{X}} \left(\frac{1}{G(x, y)} \right)^{1+\rho} = \sum_{i=1}^{|\mathbb{X}|} \frac{1}{i^{1+\rho}} = c < \infty$$

regardless of the (finite) size $|\mathbb{X}|$. Define the PMF Q_G as

$$Q_G(x, y) = \frac{1}{|\mathbb{Y}|} \cdot \frac{1}{cG(x, y)^{1+\rho}}, \quad \forall (x, y) \in \mathbb{X} \times \mathbb{Y}.$$

Note that $\text{Supp}(Q_G) = \mathbb{X} \times \mathbb{Y}$. Clearly, guessing in the decreasing order of Q_G -probabilities leads to the guessing order G . By virtue of the definition of Q_G , we have

$$\begin{aligned} & \sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} P(x, y) \left[\sum_{a \in \mathbb{X}} \left(\frac{Q_G(a, y)}{Q_G(x, y)} \right)^{\frac{1}{1+\rho}} \right]^\rho \\ & = \sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} P(x, y) G(x, y)^\rho \cdot \left(\sum_{a \in \mathbb{X}} \frac{1}{G(a, y)} \right)^\rho \\ & \leq \left(\sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} P(x, y) G(x, y)^\rho \right) \cdot (1 + \ln |\mathbb{X}|)^\rho, \end{aligned} \tag{6}$$

where the last inequality follows from (as in [2])

$$\sum_{a \in \mathbb{X}} \frac{1}{G(a, y)} = \sum_{i=1}^{|\mathbb{X}|} \frac{1}{i} \leq 1 + \ln |\mathbb{X}|, \quad \forall y \in \mathbb{Y}.$$

The proposition follows from (6). \blacksquare

Observe the similarity of the terms in the right-hand sides of equations (2) and (5) in Propositions 1 and 3, respectively. The analog of this term in

mismatched source coding is $-\sum_{x \in \mathbb{X}} P(x) \log Q(x)$, which is the expected length of a codebook built using a mismatched PMF Q . The Shannon inequality (see, for example, [13]) states that

$$-\sum_{x \in \mathbb{X}} P(x) \log Q(x) \geq -\sum_{x \in \mathbb{X}} P(x) \log P(x) = H(P)$$

The next inequality is analogous to the Shannon inequality. We can interpret this as follows: if we guess according to some mismatched distribution, then the expected number of guesses can only be larger. We will let $\alpha = 1/(1 + \rho)$ and expand the range of α to $0 < \alpha < \infty$. A special case (when no side information is available) was shown by Fischer (cf. [7, Theorem 1.3]).

Proposition 4 (*Analog of Shannon inequality*) *Let $\alpha = \frac{1}{1+\rho} > 0, \alpha \neq 1$. Then*

$$\begin{aligned} \frac{\alpha}{1-\alpha} \log \left(\sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} P(x, y) \left[\sum_{a \in \mathbb{X}} \left(\frac{Q(a, y)}{Q(x, y)} \right)^\alpha \right]^{\frac{1-\alpha}{\alpha}} \right) \\ \geq H_\alpha(P), \end{aligned} \quad (7)$$

with equality if and only if $P = Q$. \square

Proof: We will prove this directly using Holder's inequality. The right side of (7) is bounded. Without loss of generality, we may assume that the left side of (7) is finite, for otherwise the inequality trivially holds and $P \neq Q$. We may therefore assume $\text{Supp}(P) \subset \text{Supp}(Q)$ under $0 < \alpha < 1$, and $\text{Supp}(P) \cap \text{Supp}(Q) \neq \emptyset$ under $1 < \alpha < \infty$ which are the conditions when the left side of (7) is finite.

With $\alpha = 1/(1 + \rho)$, (7) is equivalent to

$$\begin{aligned} & \text{sign}(\rho) \cdot \sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} P(x, y) \left[\sum_{a \in \mathbb{X}} \left(\frac{Q(a, y)}{Q(x, y)} \right)^{\frac{1}{1+\rho}} \right]^\rho \\ & \geq \text{sign}(\rho) \cdot \sum_{y \in \mathbb{Y}} \left(\sum_{x \in \mathbb{X}} P(x, y)^{\frac{1}{1+\rho}} \right)^{1+\rho}. \end{aligned}$$

The above inequality holds term by term for each $y \in \mathbb{Y}$, a fact that can be verified by using the Hölder inequality

$$\text{sign}(\lambda) \cdot \left(\sum_x u_x \right)^\lambda \cdot \left(\sum_x v_x \right)^{1-\lambda} \geq \text{sign}(\lambda) \cdot \left(\sum_x u_x^\lambda v_x^{1-\lambda} \right) \quad (8)$$

with $\lambda = \rho/(1 + \rho) = 1 - \alpha$, $u_x = Q(x, y)^{1/(1+\rho)}$,

$$v_x = P(x, y)Q(x, y)^{-\rho/(1+\rho)},$$

and raising the resulting inequality to the power $1 + \rho > 0$. From the condition for equality in (8), equality holds in (7) if and only if $P = Q$. \blacksquare

Proposition 4 motivates us to define the following quantity that will be the focus of this paper:

$$\begin{aligned} L_\alpha(P, Q) &\triangleq \\ &\frac{\alpha}{1 - \alpha} \log \left(\sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} P(x, y) \left[\sum_{a \in \mathbb{X}} \left(\frac{Q(a, y)}{Q(x, y)} \right)^\alpha \right]^{\frac{1-\alpha}{\alpha}} \right) \\ &- H_\alpha(P). \end{aligned} \quad (9)$$

Proposition 4 indicates that $L_\alpha(P, Q) \geq 0$, with equality if and only if $P = Q$.

Just as Shannon inequality can be employed to show the converse of the source coding theorem, we employ Proposition 4 to get the converse part of a guessing theorem. We thus have a slightly different proof of [2, Theorem 1(a)].

Theorem 5 (Arikan's Guessing Theorem [2]) *Let $\rho > 0$. Consider a source pair (X, Y) with PMF P . Let $\alpha = \frac{1}{1+\rho}$. Then*

$$\begin{aligned} H_\alpha(P) - \log(1 + \ln |\mathbb{X}|) &\leq \frac{1}{\rho} \log \left(\min_G \mathbb{E}[G(X, Y)^\rho] \right) \\ &\leq H_\alpha(P). \end{aligned}$$

\square

Proof: It is easy to see that the minimum is attained when the guessing list is G_P , *i.e.*, when guessing proceeds in the decreasing order of P -probabilities. Application of Proposition 3 with $G = G_P$ and Proposition 4 with $Q = Q_{G_P}$ yields the first inequality. The upper bound follows from Corollary 2. \blacksquare

Remarks: 1) Q_{G_P} may be different from P even though they lead to the same guessing order.

2) Theorem 5 gives an operational meaning to $H_\alpha(P)$; it indicates the exponent of the minimum guessing moment to within $\log(1 + \ln |\mathbb{X}|)$.

3) Loosely speaking, Proposition 4 indicates that mismatched guessing will perform worse than matched guessing. The looseness is due to the looseness of the bound in Theorem 5.

Suppose now that we use an arbitrary guessing strategy G to guess X with side information Y , when the source (X, Y) 's PMF is P . G may not necessarily be matched to the source, as would be the case when the source statistics is unknown. Let us define its *redundancy* in guessing X with side information Y when the source is P as follows:

$$R(P, G) \triangleq \frac{1}{\rho} \log (\mathbb{E} [G(X, Y)^\rho]) - \frac{1}{\rho} \log (\mathbb{E} [G_P(X, Y)^\rho]) \quad (10)$$

The dependence of $R(P, G)$ on ρ is understood and suppressed. The following proposition bounds the redundancy on either side.

Theorem 6 *Let $\rho > 0$, $\alpha = 1/(1 + \rho)$. Consider a source pair (X, Y) with PMF P . Let G be an arbitrary guessing list with side information Y and Q_G the associated PMF given by Proposition 3. Then*

$$|R(P, G) - L_\alpha(P, Q_G)| \leq \log(1 + \ln |\mathbb{X}|).$$

□

Proof: The inequality $R(P, G) \leq L_\alpha(P, Q_G) + \log(1 + \ln |\mathbb{X}|)$ follows from Proposition 1 applied with $Q = Q_G$, the first inequality of Theorem 5, and (9).

The inequality $R(P, G) \geq L_\alpha(P, Q_G) - \log(1 + \ln |\mathbb{X}|)$ follows from Proposition 3, the second inequality of Theorem 5, and (9). ■

3 Campbell's coding theorem and redundancy

Campbell in [5] and [6] gave another operational meaning to the Rényi entropy of order $\alpha > 0$. In this section we show that $L_\alpha(\cdot, \cdot)$ arises as “inaccuracy” in this problem as well, when we encode according to a mismatched source. To be consistent with the development in the previous section, we will assume that X is coded when the source coder has side information Y .

Let \mathbb{X} and \mathbb{Y} be finite alphabet sets as before. Let the true source probabilities be given by the PMF P on $\mathbb{X} \times \mathbb{Y}$. We wish to encode each realization of X using a variable-length code, given side information Y . More precisely, let the (nonnegative) integer code lengths, $l(x, y)$ satisfy the Kraft inequality,

$$\sum_{x \in \mathbb{X}} 2^{-l(x, y)} \leq 1, \quad \forall y \in \mathbb{Y}$$

The problem is then to choose l among those that satisfy the Kraft inequality so that the following is minimized:

$$\frac{1}{\rho} \log (\mathbb{E} [2^{\rho l(X, Y)}]), \quad -1 < \rho < \infty, \rho \neq 0, \quad (11)$$

where the expectation \mathbb{E} is with respect to the PMF P . As $\rho \rightarrow 0$, this quantity tends to the expected length of the code, $\mathbb{E}[l(X, Y)]$, more commonly represented as $E[l(X, Y) | Y]$.

Observe that we may assume that $\sum_{x \in \mathbb{X}} 2^{-l(x, y)} > 1/2$ for each y ; otherwise we can reduce all lengths uniformly by 1, still satisfy the Kraft inequality and get a strictly smaller value for (11). Henceforth, we focus only on length functions that satisfy

$$\frac{1}{2} < \sum_{x \in \mathbb{X}} 2^{-l(x, y)} \leq 1, \quad \forall y \in \mathbb{Y}. \quad (12)$$

Theorem 7 (Campbell's Coding Theorem, Campbell [5]) *Let $-1 < \rho < \infty$, $\rho \neq 0$. Consider a source with PMF P . Let $\alpha = \frac{1}{1+\rho}$. Then*

$$H_\alpha(P) \leq \frac{1}{\rho} \log \left(\min_l \mathbb{E} \left[2^{\rho l(X, Y)} \right] \right) \leq H_\alpha(P) + 1,$$

where the minimization is over all those length functions that satisfy (12). \square

For a PMF Q on $\mathbb{X} \times \mathbb{Y}$, let l_Q be defined by

$$l_Q(x, y) \triangleq \left\lceil -\log \left(\frac{Q(x, y)^{\frac{1}{1+\rho}}}{\sum_{a \in \mathbb{X}} Q(a, y)^{\frac{1}{1+\rho}}} \right) \right\rceil \quad (13)$$

$$= \lceil -\log(Q'(x | y)) \rceil, \quad (14)$$

where $\lceil \cdot \rceil$ refers to the ceiling function and $Q'(\cdot | y)$ is a conditional PMF on \mathbb{X} . Clearly, l_Q satisfies (12).

Analogously, for any length function satisfying (12), we can define a PMF on $\mathbb{X} \times \mathbb{Y}$ as follows:

$$Q_l(x, y) = \frac{1}{|\mathbb{Y}|} \frac{2^{-(1+\rho)l(x, y)}}{\sum_{a \in \mathbb{X}} 2^{-(1+\rho)l(a, y)}}. \quad (15)$$

We can easily check that $l_{Q_l} = l$.

Let us define the redundancy for any l satisfying (12) as

$$\begin{aligned} R_c(P, l) \\ \triangleq \frac{1}{\rho} \log \left(\mathbb{E} \left[2^{\rho l(X, Y)} \right] \right) - \frac{1}{\rho} \log \left(\min_g \mathbb{E} \left[2^{\rho g(X, Y)} \right] \right) \end{aligned}$$

Following the same sequence of steps as in the mismatched guessing problem, it is straightforward to show the following:

Theorem 8 Let $-1 < \rho < \infty$, $\rho \neq 0$, $\alpha = 1/(1 + \rho)$. Consider a source pair (X, Y) with PMF P on \mathbb{X} . Let l be an length function that denotes an encoding of X with side information Y , and Q_l the associated PMF given by (15). Then

$$|R_c(P, l) - L_\alpha(P, Q_l)| \leq 1.$$

□

The quantity $L_\alpha(P, Q_l)$ therefore gives the redundancy to within a constant. We interpret this as the penalty for mismatched coding when Q_l is not matched to P .

4 Problem statement

Let \mathbb{T} denote a set of PMFs on the finite alphabet $\mathbb{X} \times \mathbb{Y}$. \mathbb{T} may be infinite in size. Associated with \mathbb{T} is a family \mathcal{T} of measurable subsets of \mathbb{T} and thus $(\mathbb{T}, \mathcal{T})$ is a measurable space. We assume that for every $x \in \mathbb{X}$, the mapping $P \mapsto P(x)$ is \mathcal{T} -measurable.

For a fixed $\rho > 0$, we seek a good guessing strategy G that works well for all $P \in \mathbb{T}$. G can depend on knowledge of \mathbb{T} , but not on the actual source PMF. More precisely, for $P \in \mathbb{T}$ the redundancy denoted by $R(P, G)$ when the true source is P and when the guessing list is G , is given by (10). The worst redundancy under this guessing strategy is given by

$$\sup_{P \in \mathbb{T}} R(P, G)$$

Our aim is to minimize this worst redundancy over all guessing strategies, *i.e.*, find a G that attains the minimum

$$R^* = \min_G \sup_{P \in \mathbb{T}} R(P, G) \quad (16)$$

In view of Theorem 6, clearly, the following quantity is relevant for $0 < \alpha < 1$. The definition however is wider in scope.

Definition 9 For $0 < \alpha < \infty, \alpha \neq 1$,

$$C \triangleq \min_Q \sup_{P \in \mathbb{T}} L_\alpha(P, Q). \quad (17)$$

The following theorem justifies the use of “min” instead of “inf”.

Theorem 10 There exists a unique PMF Q^* such that

$$C = \sup_{P \in \mathbb{T}} L_\alpha(P, Q^*) = \inf_Q \sup_{P \in \mathbb{T}} L_\alpha(P, Q).$$

□

The proof is in Section 5.4.

Remark: 1) $C \leq \log |\mathbb{X}|$ and is therefore finite. Indeed, take Q to be uniform PMF on $\mathbb{X} \times \mathbb{Y}$. Then

$$L_\alpha(P, Q) = \log |\mathbb{X}| - H_\alpha(P) \leq \log |\mathbb{X}|, \quad \forall P \in \mathbb{T}.$$

2) The minimizing Q^* has the geometric interpretation of a *center* of the uncertainty set \mathbb{T} . Accordingly, C plays the role of *radius*; all elements in the uncertainty set \mathbb{T} are within a “squared distance” C from the center Q^* . The reason for describing $L_\alpha(P, Q)$ as “squared distance” will become clear after Proposition 14.

The following result shows how to find good guessing schemes under uncertainty.

Theorem 11 (Guessing under uncertainty) *Let \mathbb{T} be a class of PMFs. There exists a guessing list G^* for X with side information Y such that*

$$\sup_{P \in \mathbb{T}} R(P, G^*) \leq C + \log(1 + \ln |\mathbb{X}|).$$

Conversely, for any arbitrary guessing strategy G , the worst-case redundancy is at least $C - \log(1 + \ln |\mathbb{X}|)$, i.e.,

$$\sup_{P \in \mathbb{T}} R(P, G) \geq C - \log(1 + \ln |\mathbb{X}|).$$

□

Proof: Let Q^* be the PMF on $\mathbb{X} \times \mathbb{Y}$ that attains the minimum in (17), i.e.,

$$C = \sup_{P \in \mathbb{T}} L_\alpha(P, Q^*). \quad (18)$$

Let $G^* = G_{Q^*}$. Then

$$R(P, G^*) \leq L_\alpha(P, Q^*) + \log(1 + \ln |\mathbb{X}|) \quad (19)$$

follows from Proposition 1 applied with $Q = Q^*$, the first inequality of Theorem 5, and (9), as in the proof of Theorem 6. After taking supremum over all $P \in \mathbb{T}$, and after substitution of (18), we get

$$\begin{aligned} \sup_{P \in \mathbb{T}} R(P, G^*) &\leq \sup_{P \in \mathbb{T}} L_\alpha(P, Q^*) + \log(1 + \ln |\mathbb{X}|) \\ &= C + \log(1 + \ln |\mathbb{X}|), \end{aligned}$$

which proves the first statement.

For any guessing strategy G , observe that Theorem 6 implies that

$$R(P, G) \geq L_\alpha(P, Q_G) - \log(1 + \ln |\mathbb{X}|),$$

and therefore

$$\begin{aligned} \sup_{P \in \mathbb{T}} R(P, G) &\geq \sup_{P \in \mathbb{T}} L_\alpha(P, Q_G) - \log(1 + \ln |\mathbb{X}|) \\ &\geq C - \log(1 + \ln |\mathbb{X}|), \end{aligned}$$

which proves the second statement. \blacksquare

Remarks: 1) Thus one approach to obtain the minimum in (16) is to identify minimum value in (17). This minimum value will be within $\log(1 + \ln |\mathbb{X}|)$ of R^* in (16). Moreover, the corresponding minimizer Q^* can be used to generate a guessing strategy.

2) Theorem 11 can be easily restated for Campbell's coding problem. The nuisance term $\log(1 + \ln |\mathbb{X}|)$ is now replaced by the constant 1.

3) The converse part of Theorem 11 is meaningful only when $C > \log(1 + \ln |\mathbb{X}|)$. This will hold, for example, when the uncertainty class is sufficiently rich. The finite state, arbitrarily varying source is one such example. Observe that if we have $\mathbb{X} \times \mathbb{Y} = \mathbb{A}^n \times \mathbb{B}^n$, then $\log(1 + \ln |\mathbb{X}|)$ grows logarithmically with n if $|\mathbb{X}| \geq 2$. The uncertainty class will be rich enough for the converse to be meaningful if C grows with n at a faster rate.

5 Properties of L_α

Having shown how $L_\alpha(P, Q)$ arises as a penalty function for mismatched guessing and coding, let us now study its relevant properties. Throughout this section, $0 < \alpha < \infty, \alpha \neq 1$. Accordingly, $-1 < \rho < \infty, \rho \neq 0$. Let P and Q be PMFs on $\mathbb{X} \times \mathbb{Y}$.

1. As we saw before, $L_\alpha(P, Q) \geq 0$, with equality if and only if $P = Q$.
2. $L_\alpha(P, Q) = \infty$ if and only if $\text{Supp}(P) \cap \text{Supp}(Q) = \emptyset$, or $\alpha < 1$ and $\text{Supp}(P) \not\subset \text{Supp}(Q)$.
3. Given the joint PMF P , let us define the “tilted” conditional PMF on \mathbb{X} as follows:

$$P'(x | y) \stackrel{\Delta}{=} \begin{cases} P(x, y)^\alpha / \sum_{a \in \mathbb{X}} P(a, y)^\alpha, & \text{if } \sum_{a \in \mathbb{X}} P(a, y)^\alpha > 0, \\ 1/|\mathbb{X}|, & \text{otherwise.} \end{cases} \quad (20)$$

The above definition simplifies many expressions in the sequel. The dependence on α in the mapping $P \mapsto P'$ is suppressed.

4. When $|\mathbb{Y}| = 1$, we interpret that no side information is available. Then P and Q may be thought of PMFs on \mathbb{X} with no reference to \mathbb{Y} . P' and Q' given by (20) are PMFs in one-to-one correspondence with P and Q respectively.

Using the expression for Rényi entropy and (9), we have that

$$\begin{aligned} L_\alpha(P, Q) &= \frac{1}{\rho} \log \left(\sum_{x \in \mathbb{X}} P'(x)^{1+\rho} \cdot Q'(x)^{-\rho} \right) \\ &= D_{1/\alpha}(P' \parallel Q'), \end{aligned} \quad (21)$$

where $D_\beta(R \parallel S)$ is the Rényi's information divergence of order β ,

$$D_\beta(R \parallel S) = \frac{1}{\beta - 1} \log \left(\sum_{x \in \mathbb{X}} R(x)^\beta S(x)^{1-\beta} \right),$$

which is ≥ 0 and equals 0 if and only if $R = S$. For the case when $|\mathbb{Y}| = 1$ we therefore have another proof of Proposition 4.

5. The conditional Kullback-Leibler divergence is recovered as follows:

$$\lim_{\alpha \rightarrow 1} L_\alpha(P, Q) = \sum_y \sum_x P(x, y) \log \left(\frac{P(x \mid y)}{Q(x \mid y)} \right),$$

where $Q(\cdot \mid y)$ and $P(\cdot \mid y)$ are the respective conditional PMFs of X given $Y = y$.

6. In general, $L_\alpha(P, Q)$ is not a convex function of P . Moreover, it is not, in general, a convex function of Q .

7. In general, $L_\alpha(P, Q)$ does not satisfy the so-called data-processing inequality. More precisely, if \mathbb{X}' and \mathbb{Y}' are finite sets, and if $f : \mathbb{X} \times \mathbb{Y} \rightarrow \mathbb{X}' \times \mathbb{Y}'$ is a function, it is not necessarily true that $L_\alpha(P, Q) \geq L_\alpha(Pf^{-1}, Qf^{-1})$.

8. When $|\mathbb{Y}| = 1$, *i.e.*, in the no side information case, using (20) we can write $L_\alpha(P, Q)$ as follows:

$$L_\alpha(P, Q) = \frac{1}{\rho} \log [\text{sign}(\rho) \cdot I_f(P' \parallel Q')], \quad (22)$$

where $I_f(R \parallel S)$ is the f -divergence [14] given by

$$I_f(R \parallel S) = \sum_{x \in \mathbb{X}} S(x) f \left(\frac{R(x)}{S(x)} \right), \quad (23)$$

with

$$f(x) = \text{sign}(\rho) \cdot x^{1+\rho}, \quad x \geq 0. \quad (24)$$

Since f is a strictly convex function for $\rho > 0$, an application of Jensen's inequality in (23) indicates that

$$I_f(R \parallel S) \geq f(1) = \begin{cases} -1, & -1 < \rho < 0, \\ 1, & 0 < \rho < \infty. \end{cases} \quad (25)$$

Moreover, when $-1 < \rho < 0$, we have the following bounds:

$$-1 \leq I_f(R \parallel S) \leq 0. \quad (26)$$

9. Let us define

$$h(P) \triangleq \sum_{y \in \mathbb{Y}} \left(\sum_{x \in \mathbb{X}} P(x, y)^\alpha \right)^{\frac{1}{\alpha}}.$$

The dependence of h on α is understood, and suppressed for convenience. Clearly,

$$H_\alpha(P) = \frac{\alpha}{1 - \alpha} \log h(P). \quad (27)$$

Motivated by the relationship in (22), let us write L_α in the general case as follows:

$$L_\alpha(P, Q) = \frac{1}{\rho} \log [\text{sign}(\rho) \cdot I(P, Q)], \quad (28)$$

where $I(P, Q)$ is given by

$$\begin{aligned} I(P, Q) & \\ & \triangleq \frac{\text{sign}(\rho)}{h(P)} \sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} P(x, y) (Q'(x \mid y))^{-\rho}, \\ & = \frac{\text{sign}(1 - \alpha)}{h(P)} \sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} P(x, y) (Q'(x \mid y))^{\frac{\alpha-1}{\alpha}}. \end{aligned} \quad (29)$$

(30)

These expressions turn out to be useful in the sequel.

It is not difficult to show that

$$I(P, Q) = \sum_{y \in \mathbb{Y}} w(y) \cdot I_f(P'(\cdot \mid y) \parallel Q'(\cdot \mid y)),$$

where w is the PMF on \mathbb{Y} given by

$$w(y) = \frac{1}{h(P)} \cdot \left(\sum_{x \in \mathbb{X}} P(x, y)^\alpha \right)^{\frac{1}{\alpha}}.$$

Consequently, the bounds given in (25) and (26) are valid for $I(P, Q)$, under corresponding conditions on α .

10. Inequalities involving L_α result in inequalities involving I with ordering preserved. More precisely, for $r \geq 0$, if $L_\alpha(P, Q) < r$, then $I(P, Q) < t$, for $t = \text{sign}(\rho) \cdot 2^{\rho r}$.
11. From the known bounds $0 \leq H_\alpha(P) \leq \log |\mathbb{X}|$, it is easy to see the following bounds:

$$1 \leq h(P) \leq |\mathbb{X}|^{\frac{1-\alpha}{\alpha}}, \text{ for } 0 < \alpha < 1, \quad (31)$$

and

$$|\mathbb{X}|^{\frac{1-\alpha}{\alpha}} \leq h(P) \leq 1, \text{ for } 1 < \alpha < \infty. \quad (32)$$

In both cases, we see that $h(P)$ is bounded away from 0 and therefore (29) and (30) are well-defined.

The quantity $L_\alpha(P, Q)$ does not have many of the useful properties enjoyed by the Kullback-Leibler divergence, or other f -divergences, even in the case when $|\mathbb{Y}| = 1$. However, it behaves like squared distance and shares a ‘‘Pythagorean’’ property with the Kullback-Leibler divergence.

5.1 L_α -projection

We proceed along the lines of [3]. Let \mathbb{X} and \mathbb{Y} be finite alphabet sets. Let $\mathcal{P}(\mathbb{X} \times \mathbb{Y})$ denote the set of PMFs on $\mathbb{X} \times \mathbb{Y}$. Given a PMF R on $\mathbb{X} \times \mathbb{Y}$, the set

$$B(R, r) \stackrel{\Delta}{=} \{P \in \mathcal{P}(\mathbb{X} \times \mathbb{Y}) \mid L_\alpha(P, R) < r\}, \quad 0 < r \leq \infty,$$

is called an L_α -sphere (or ball) with center R and radius r . The term ‘‘sphere’’ conjures the image of a convex set. That the set is indeed convex needs a proof since $L_\alpha(P, R)$ is not convex in its arguments.

Proposition 12 $B(R, r)$ is a convex set. \square

Proof: Let $P_i \in B(R, r)$ for $i = 0, 1$. For any $\lambda \in [0, 1]$, we need to show that $P_\lambda = (1 - \lambda)P_0 + \lambda P_1 \in B(R, r)$. With $\alpha = 1/(1 + \rho)$, and $t = \text{sign}(\rho) \cdot 2^{\rho r}$, we get from (28) that

$$I(P_i, R) < t, \quad i = 0, 1. \quad (33)$$

The proof will be complete if we can show that $I(P_\lambda, R) < t$. To this end,

$$\begin{aligned}
I(P_\lambda, R) &= \frac{\text{sign}(1-\alpha)}{h(P_\lambda)} \cdot \sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} P_\lambda(x, y) (R'(x | y))^{\frac{\alpha-1}{\alpha}} \\
&= \frac{\text{sign}(1-\alpha)}{h(P_\lambda)} \cdot (1-\lambda) \sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} P_0(x, y) (R'(x | y))^{\frac{\alpha-1}{\alpha}} \\
&\quad + \frac{\text{sign}(1-\alpha)}{h(P_\lambda)} \cdot (\lambda) \sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} P_1(x, y) (R'(x | y))^{\frac{\alpha-1}{\alpha}} \\
&= \frac{(1-\lambda)h(P_0)I(P_0, R) + \lambda h(P_1)I(P_1, R)}{h(P_\lambda)} \tag{34}
\end{aligned}$$

$$< t \frac{(1-\lambda)h(P_0) + \lambda h(P_1)}{h(P_\lambda)} \tag{35}$$

$$\begin{aligned}
&= |t| \frac{(1-\lambda) \cdot \text{sign}(1-\alpha)h(P_0) + \lambda \cdot \text{sign}(1-\alpha)h(P_1)}{h(P_\lambda)} \\
&\leq |t| \frac{\text{sign}(1-\alpha)h(P_\lambda)}{h(P_\lambda)} \\
&= t;
\end{aligned} \tag{36}$$

where (34) follows from (30), (35) follows from (33), and (36) follows from the concavity of $\text{sign}(1-\alpha)h$. \blacksquare

When we talk of closed sets, we refer to the usual Euclidean metric on the $|\mathbb{X}||\mathbb{Y}|$ -dimensional Euclidean vector space. The set of PMFs on $\mathbb{X} \times \mathbb{Y}$ is closed and bounded (and therefore compact).

If \mathcal{E} is a closed and convex set of PMFs on $\mathbb{X} \times \mathbb{Y}$ intersecting $B(R, \infty)$, *i.e.* there exists a PMF P such that $L_\alpha(P, R) < \infty$, then a PMF $Q \in \mathcal{E}$ satisfying

$$L_\alpha(Q, R) = \min_{P \in \mathcal{E}} L_\alpha(P, R),$$

is called the L_α -projection of R on \mathcal{E} .

Proposition 13 (Existence of L_α -projection) *Let \mathcal{E} be a closed and convex set of PMFs on $\mathbb{X} \times \mathbb{Y}$. If $B(R, \infty) \cap \mathcal{E}$ is nonempty, then R has an L_α -projection on \mathcal{E} .*

Proof: Pick a sequence $P_n \in \mathcal{E}$ with $L_\alpha(P_n, R) < \infty$ such that $L_\alpha(P_n, R) \rightarrow \inf_{P \in \mathcal{E}} L_\alpha(P, R)$. This sequence being in the compact space \mathcal{E} has a cluster point Q and a subsequence converging to Q . We can simply focus on this subsequence and therefore assume that $P_n \rightarrow Q$ and $L_\alpha(P_n, R) \rightarrow \inf_{P \in \mathcal{E}} L_\alpha(P, R)$. \mathcal{E} is closed and hence $Q \in \mathcal{E}$. The continuity of the logarithm function, wherever it is finite, and $L_\alpha(P_n, R) < \infty$

imply that

$$\begin{aligned}
\lim_n L_\alpha(P_n, R) &= \frac{1}{\rho} \log \left(\text{sign}(\rho) \cdot \lim_n I(P_n, R) \right) \\
&= \frac{1}{\rho} \log (\text{sign}(\rho) \cdot I(Q, R)) \\
&= L_\alpha(Q, R),
\end{aligned} \tag{37}$$

where (37) follows from the observation that (30) is the ratio of a continuous linear function of P and the continuous concave function $\text{sign}(1 - \alpha)h$ that is positive, bounded, and bounded away from 0.

From the uniqueness of limits we have that $L_\alpha(Q, R) = \inf_{P \in \mathcal{E}} L_\alpha(P, R)$. Q is then an L_α -projection of R on \mathcal{E} . \blacksquare

We next state generalizations of [3, Lemma 2.1, Theorem 2.2]. Here $L_\alpha(P, Q)$ plays the role of squared Euclidean distance (analogous to the Kullback-Leibler divergence).

Proposition 14 *Let $0 < \alpha < \infty, \alpha \neq 1$.*

1. *Let $L_\alpha(Q, R)$ and $L_\alpha(P, R)$ be finite. The segment joining P and Q does not intersect the L_α -sphere $B(R, r)$ with radius $r = L_\alpha(Q, R)$, i.e.,*

$$L_\alpha(P_\lambda, R) \geq L_\alpha(Q, R)$$

for each

$$P_\lambda = \lambda P + (1 - \lambda)Q, \quad 0 \leq \lambda \leq 1,$$

if and only if

$$L_\alpha(P, R) \geq L_\alpha(P, Q) + L_\alpha(Q, R). \tag{38}$$

2. (Tangent hyperplane) *Let*

$$Q = \lambda P + (1 - \lambda)S, \quad 0 < \lambda < 1. \tag{39}$$

Let $L_\alpha(Q, R)$, $L_\alpha(P, R)$, and $L_\alpha(S, R)$ be finite. The segment joining P and S does not intersect $B(R, r)$ (with $r = L_\alpha(Q, R)$) if and only if

$$L_\alpha(P, R) = L_\alpha(P, Q) + L_\alpha(Q, R). \tag{40}$$

\square

Remarks: 1) Under the hypotheses in Proposition 14.1, we deduce that $L_\alpha(P, Q) < \infty$ as a consequence.

2) The condition (39) implies that $P \leq \lambda^{-1}Q$, and therefore $\text{supp}(P) \subset \text{supp}(Q)$. If $0 < \alpha < 1$, and $L_\alpha(Q, R) < \infty$, then we have $\text{supp}(P) \subset \text{supp}(Q) \subset \text{supp}(R)$. Thus both $L_\alpha(P, R)$ and $L_\alpha(P, Q)$ are necessarily

finite. For $\alpha \in (0, 1)$, the requirement that $L_\alpha(P, R)$ be finite can therefore be removed. The requirement is however needed for $1 < \alpha < \infty$ because even though $\text{supp}(P) \subset \text{supp}(Q)$ and $\text{supp}(Q) \cap \text{supp}(R) \neq \emptyset$, we may have $\text{supp}(P) \cap \text{supp}(R) = \emptyset$ leading to $L_\alpha(P, R) = \infty$.

3) Proposition 14.2 extends the analog of Pythagoras theorem, known to hold for the Kullback-Leibler divergence, to the family L_α parameterized by $\alpha > 0$.

4) By symmetry between P and S , (40) holds when P is replaced by S .

Proof: 1) \Rightarrow : Since $L_\alpha(P, R)$ and $L_\alpha(Q, R)$ are finite, from (29), we gather that both $\sum_y \sum_x P(x, y) R'(x | y)^{-\rho}$ and $\sum_y \sum_x Q(x, y) R'(x | y)^{-\rho}$ are finite and nonzero.

Observe that $P_0 = Q$, and $L_\alpha(P_\lambda, R) \geq L_\alpha(P_0, R)$ implies that

$$I(P_\lambda, R) \geq I(P_0, R).$$

Thus

$$\frac{I(P_\lambda, R) - I(P_0, R)}{\lambda} \geq 0 \quad (41)$$

for every $\lambda \in (0, 1]$. The limiting value as $\lambda \downarrow 0$, the derivative of $I(P_\lambda, R)$ with respect to λ evaluated at $\lambda = 0$, should be ≥ 0 . This will give us the necessary condition.

Note that the derivative evaluated at $\lambda = 0$ is a one-sided limit since $\lambda \in [0, 1]$. We will first check that this one-sided limit exists.

From (29), $I(P_\lambda, R)$ can be written as $s(\lambda)/t(\lambda)$, where $t(\lambda)$ is bounded, positive, and lower-bounded away from 0, for every λ . Let $\dot{s}(0)$ and $\dot{t}(0)$ be the derivatives of s and t evaluated at $\lambda = 0$. Clearly,

$$\begin{aligned} \dot{s}(0) &= \lim_{\lambda \downarrow 0} \frac{s(\lambda) - s(0)}{\lambda} \\ &= \text{sign}(\rho) \left(\sum_y \sum_x P(x, y) (R'(x | y))^{-\rho} \right. \\ &\quad \left. - \sum_y \sum_x Q(x, y) (R'(x | y))^{-\rho} \right). \end{aligned}$$

Similarly, it is easy to check that

$$\dot{t}(0) = \sum_y \sum_x P(x, y) (Q'(x | y))^{-\rho} - t(0),$$

with the possibility that it is $+\infty$ (only when $0 < \alpha < 1$ and $\text{supp}(P) \not\subset \text{supp}(Q)$).

Since we can write

$$\begin{aligned} &\frac{1}{\lambda} \left(\frac{s(\lambda)}{t(\lambda)} - \frac{s(0)}{t(0)} \right) \\ &= \frac{1}{t(\lambda)t(0)} \left[t(0) \frac{s(\lambda) - s(0)}{\lambda} - s(0) \frac{t(\lambda) - t(0)}{\lambda} \right], \end{aligned}$$

it follows that the derivative of $s(\lambda)/t(\lambda)$ exists at $\lambda = 0$ and is given by $(t(0)\dot{s}(0) - s(0)\dot{t}(0)) / t^2(0)$, with the possibility that it might be $+\infty$. However, (41) and $t(0) > 0$ imply that

$$\dot{s}(0) - s(0) \frac{\dot{t}(0)}{t(0)} \geq 0.$$

Consequently, $\dot{t}(0)$ is necessarily finite. In particular, when $0 < \alpha < 1$, we have ascertained that $L_\alpha(P, Q)$ is finite. After substitution of $s(0), t(0), \dot{s}(0)$, and $\dot{t}(0)$ we get

$$\begin{aligned} & \text{sign}(\rho) \cdot \sum_y \sum_x P(x, y) (R'(x | y))^{-\rho} \\ & \geq \text{sign}(\rho) \cdot \left(\sum_y \sum_x P(x, y) (Q'(x | y))^{-\rho} \right) \\ & \quad \cdot \left(\frac{\sum_y \sum_x Q(x, y) (R'(x | y))^{-\rho}}{h(Q)} \right) \end{aligned} \tag{42}$$

When $-1 < \rho < 0$, clearly, $\sum_y \sum_x P(x, y) (Q'(x | y))^{-\rho}$ cannot be zero, due to the nonzero assumptions on the other quantities in (42). This implies that $L_\alpha(P, Q)$ is finite when $1 < \alpha < \infty$ as well. An application of (28) and (29) shows that (42) and (38) are equivalent. This concludes the proof of the forward implication.

The reader will recognize that the basic idea is quite simple: evaluation of a derivative at $\lambda = 0$ and a check that it is nonnegative. The technical details above ensure that the case when the derivative of the denominator is infinite is carefully examined.

1) \Leftarrow : The hypotheses imply that $L_\alpha(P, R), L_\alpha(Q, R)$, and $L_\alpha(P, Q)$ are finite. As observed above, (42) and (38) are equivalent. Observe that both sides of (42) are linear in P . This property will be exploited in the proof. Clearly, if we set $P = Q$ in (38) and (42), we have the equalities

$$L_\alpha(Q, R) = L_\alpha(Q, Q) + L_\alpha(Q, R) \tag{43}$$

and

$$\begin{aligned} & \text{sign}(\rho) \cdot \sum_y \sum_x Q(x, y) (R'(x | y))^{-\rho} \\ & = \text{sign}(\rho) \cdot \left(\sum_y \sum_x Q(x, y) (Q'(x | y))^{-\rho} \right) \\ & \quad \cdot \left(\frac{\sum_y \sum_x Q(x, y) (R'(x | y))^{-\rho}}{h(Q)} \right) \end{aligned} \tag{44}$$

A λ -weighted linear combination of the inequalities (42) and (44) yields (42) with P replaced by P_λ . The equivalence of (38) and (42) result in

$$\begin{aligned} L_\alpha(P_\lambda, R) &\geq L_\alpha(P_\lambda, Q) + L_\alpha(Q, R) \\ &\geq L_\alpha(Q, R). \end{aligned}$$

This concludes the proof of the first part.

2) This follows easily from the first statement. For the forward implication, indeed, (42) holds for P . Moreover, (42) holds when P is replaced by S . If either of these were a strict inequality, the linear combination of these with the λ given by (39) will satisfy (44) with strict inequality, a contradiction. The reverse implication is straightforward. \blacksquare

Let us now apply Proposition 14 to the L_α -projection of a convex set. We first need the following definition.

For a convex \mathcal{E} , we call Q an algebraic inner point of \mathcal{E} if for every $P \in \mathcal{E}$, there exist λ and S satisfying (39).

Theorem 15 (Projection Theorem) *Let $0 < \alpha < \infty, \alpha \neq 1$ and \mathbb{X} a finite set. A PMF $Q \in \mathcal{E} \cap B(R, \infty)$ is the L_α -projection of R on the convex set \mathcal{E} if and only if every $P \in \mathcal{E}$ satisfies*

$$L_\alpha(P, R) \geq L_\alpha(P, Q) + L_\alpha(Q, R). \quad (45)$$

If the L_α -projection Q is an algebraic inner point of \mathcal{E} , then every $P \in \mathcal{E} \cap B(R, \infty)$ satisfies (45) with equality. \square .

Proof: This follows easily from Proposition 14. For the case when $L_\alpha(P, R) = \infty$ not covered by Proposition 14, (45) holds trivially. \blacksquare

Corollary 16 *Let $0 < \alpha < 1$, and a PMF $Q \in \mathcal{E} \cap B(R, \infty)$ be the L_α -projection of R on the convex set \mathcal{E} . If Q is an algebraic inner point of \mathcal{E} , then every $P \in \mathcal{E}$ satisfies (45) with equality.*

Proof: Clearly, for any $P \in \mathcal{E}$, we have $\text{supp}(P) \subset \text{supp}(Q) \subset \text{supp}(R)$, and therefore $\mathcal{E} \subset B(R, \infty)$. The corollary now follows from the second statement of Theorem 15. \blacksquare

While existence of L_α -projection is guaranteed for certain sets by Proposition 13, the following talks about uniqueness of the projection.

Proposition 17 (*Uniqueness of projection*) *Let $0 < \alpha < \infty, \alpha \neq 1$. If the L_α -projection of R on the convex set \mathcal{E} exists, it is unique.*

Proof: Let Q_1 and Q_2 be the projections. Then

$$\infty > L_\alpha(Q_1, R) = L_\alpha(Q_2, R) \geq L_\alpha(Q_2, Q_1) + L_\alpha(Q_1, R),$$

where the last inequality follows from Theorem 15. Thus $L_\alpha(Q_2, Q_1) = 0$, and $Q_2 = Q_1$. \blacksquare

Analogous to the Kullback-Leibler divergence case, our next result is the transitivity property.

Theorem 18 *Let \mathcal{E} and $\mathcal{E}_1 \subset \mathcal{E}$ be convex sets of PMFs on \mathbb{X} . Let R have L_α -projection Q on \mathcal{E} and Q_1 on \mathcal{E}_1 , and suppose that (45) holds with equality for every $P \in \mathcal{E}$. Then Q_1 is the L_α -projection of Q on \mathcal{E}_1 .*

Proof: The proof is the same as in [3, Theorem 2.3]. We repeat it here for completeness.

Observe that from the equality hypothesis applied to $Q_1 \in \mathcal{E}_1 \subset \mathcal{E}$, we have

$$L_\alpha(Q_1, R) = L_\alpha(Q_1, Q) + L_\alpha(Q, R). \quad (46)$$

Consequently $L_\alpha(Q_1, Q)$ is finite.

Furthermore, for a $P \in \mathcal{E}_1$, we have

$$\begin{aligned} & L_\alpha(P, R) \\ & \geq L_\alpha(P, Q_1) + L_\alpha(Q_1, R) \end{aligned} \quad (47)$$

$$= L_\alpha(P, Q_1) + L_\alpha(Q_1, Q) + L_\alpha(Q, R), \quad (48)$$

where (47) follows from Theorem 15 applied to \mathcal{E}_1 , and (48) follows from (46).

We next compare (48) with $L_\alpha(P, R) = L_\alpha(P, Q) + L_\alpha(Q, R)$ and cancel $L_\alpha(Q, R)$ to obtain

$$L_\alpha(P, Q) \geq L_\alpha(P, Q_1) + L_\alpha(Q_1, Q)$$

for every $P \in \mathcal{E}_1$. Theorem 15 guarantees that Q_1 is the L_α -projection of Q on \mathcal{E}_1 . \blacksquare

As an application of Theorem 15 let us characterize the L_α -center of a family.

Proposition 19 *If the L_α -center of a family \mathbb{T} of PMFs exists, it lies in the closure of the convex hull of the family.*

Proof: Let \mathcal{E} be the closure of the convex hull of \mathbb{T} . Let Q^* be an L_α -center of the family, and C , which is at most $\log |\mathbb{X}|$, the L_α -radius. Our first goal is to show that $Q^* \in \mathcal{E}$.

By Proposition 13, Q^* has an L_α -projection Q on \mathcal{E} , and by Proposition 17, the projection is unique on \mathcal{E} . From Theorem 15, for every $P \in \mathbb{T}$, we have

$$L_\alpha(P, Q^*) \geq L_\alpha(P, Q) + L_\alpha(Q, Q^*).$$

Thus

$$\begin{aligned} C &= \sup_{P \in \mathbb{T}} L_\alpha(P, Q^*) \\ &\geq \sup_{P \in \mathbb{T}} L_\alpha(P, Q) + L_\alpha(Q, Q^*) \\ &\geq C + L_\alpha(Q, Q^*). \end{aligned}$$

Thus $L_\alpha(Q, Q^*) = 0$, leading to $Q^* = Q \in \mathcal{E}$. \blacksquare

For the special case when $|\mathbb{T}| = m$ is finite, *i.e.*, $\mathbb{T} = \{P_1, \dots, P_m\}$, we can find a weight vector w such that $Q^* = \sum_{i=1}^m w(i)P_i$ and $\sum_{i=1}^m w(i) = 1$. This result can be obtained in a more direct way using already known results for f -divergences and Rényi information divergence of order α ; this is the subject of the next subsection.

5.2 L_α -center and radius for a finite family

In this subsection, we identify the L_α -center and radius when $|\mathbb{T}|$ is finite and when there is no side information. We will therefore use \mathbb{X} instead of the cumbersome $\mathbb{X} \times \mathbb{Y}$. Our main goals here are to verify using known results that the L_α -center exists, is unique, and lies in the closure of the convex hull of \mathbb{T} . We then briefly touch upon connections with Gallager exponents, capacity of order $1/\alpha$, and information radius of order $1/\alpha$. The development in this section will suggest an approach to prove Theorem 10 for the case when \mathbb{T} is infinite in size.

5.2.1 Proof of Theorem 10 for a finite family of PMFs

Let $\mathbb{T} = \{P_1, \dots, P_m\}$ be PMFs on \mathbb{X} . The problem of identifying the L_α -center and radius can be solved by identifying the $D_{1/\alpha}$ -center and radius of the tilted family of PMFs $\{P'_i \mid 1 \leq i \leq m\}$, where the invertible transformation from $Q \mapsto Q'$ is given by (20). Moreover, from (21) and (22), we

have

$$\inf_Q \max_{1 \leq i \leq m} L_\alpha(P_i, Q) \quad (49)$$

$$= \inf_Q \max_{1 \leq i \leq m} D_{1+\rho}(P'_i \| Q') \quad (50)$$

$$= \frac{1}{\rho} \log \left(\text{sign}(\rho) \inf_Q \max_{1 \leq i \leq m} I_f(P'_i \| Q') \right), \quad (51)$$

Csiszár considered the evaluation of (50) in [12, Proposition 1], and the evaluation of the inf-max within parenthesis in (51) in [14].

From [14, Theorem 3.2] and its Corollary (the required conditions for their application are f is strictly convex and $f(0) < \infty$; these clearly hold since $\rho \neq 0$ and $f(0) = 0$) there exists a unique PMF $(Q')^*$ on \mathbb{X} , which minimizes $\max_{1 \leq i \leq m} I_f(P'_i \| Q')$. From the bijectivity of the $Q \mapsto Q'$ mapping, the infima in (49), (50), and (51) can all be replaced by minima. From the inverse of the map $Q \mapsto Q'$, we obtain the unique minimizer Q^* for (49). This proves the existence and uniqueness result of Theorem 10 for the case when $|\mathbb{T}|$ is finite.

5.2.2 Minimizer is in the convex hull

Let \mathcal{E} be the convex hull of \mathbb{T} . That the minimizer Q^* is in the convex hull of the family, *i.e.*, $Q^* \in \mathcal{E}$, can be gleaned from the results of [14, Equation 2.25], [14, Theorem 3.2], and its Corollary. Indeed, [14, Theorem 3.2] assures that

$$\min_{Q'} \max_{1 \leq i \leq m} I_f(P'_i \| Q') \quad (52)$$

$$= \max_{\mu} \min_{Q'} \sum_{i=1}^m \mu(i) I_f(P'_i \| Q'), \quad (53)$$

where the max-min in (53) is achieved at (μ^*, Q'^*) , and Q'^* is the PMF which attains the min-max in (52). We now seek to find out the nature of Q'^* and thence Q^* .

For any arbitrary weight function μ , we have from [14, Equation 2.25] that the Q' which minimizes

$$\sum_{i=1}^m \mu(i) I_f(P'_i \| Q') \quad (54)$$

is

$$Q'(x) = c^{-1} \cdot \left(\sum_{i=1}^m \mu(i) (P'_i(x))^{1/\alpha} \right)^\alpha \quad (55)$$

$$= c^{-1} \left(\sum_{i=1}^m \frac{\mu(i)}{h(P_i)} P_i(x) \right)^{1/\alpha} \quad (56)$$

for every $x \in \mathbb{X}$, where c is the normalizing constant. From the correspondence between the primed and the unprimed PMFs, and (56), we obtain

$$Q(x) = d^{-1} \sum_{i=1}^m \frac{\mu(i)}{h(P_i)} P_i(x), \quad \forall x \in \mathbb{X} \quad (57)$$

where d is the normalizing constant

$$d = \sum_{i=1}^m \frac{\mu(i)}{h(P_i)}. \quad (58)$$

Thus, for an arbitrary μ , the Q (obtained from Q') that minimizes (54) is in the convex hull \mathcal{E} . In particular, the minimizing Q^* corresponding to the μ^* that attains the max-min objective in (53), and therefore the min-max objective in (52), is also in \mathcal{E} .

With some algebra, we can further show that

$$C = \min_Q \max_{1 \leq i \leq m} L_\alpha(P_i, Q) = \frac{\alpha}{1-\alpha} \log(d \cdot h(Q^*)), \quad (59)$$

where Q^* is given by (57) and d by (58) with $\mu = \mu^*$.

5.2.3 Necessary and sufficient conditions for finding the L_α -center and radius

From [14, Theorem 3.2], a weight vector μ maximizes (53) if and only if

$$I_f(P'_i \parallel Q') \leq K, \quad i = 1, 2, \dots, m, \quad (60)$$

where equality holds whenever $\mu(i) > 0$, and Q' is given by (55). Under this condition, clearly, the corresponding Q given by (57) is the L_α -center and $C = (1/\rho) \log(\text{sign}(\rho) \cdot K)$ is the L_α -radius.

An interesting special case occurs when $h(P_i)$ is independent of i . Then we may simplify (57) to

$$Q = \sum_{i=1}^m \mu(i) P_i, \quad (61)$$

i.e., the weights that make the optimum mixture (of PMFs) are the same as the given weights that form the objective function in (53).

5.2.4 Relationship with Gallager exponent

For the set of PMFs $\{P_i \mid 1 \leq i \leq m\}$ the tilted set $\{P'_i \mid 1 \leq i \leq m\}$ can be considered as a channel with input alphabet $\{1, 2, \dots, m\}$ and output alphabet \mathbb{X} . This channel will be represented as P' .

From the remarks in [12] on the connection between information radius of order $1/\alpha$ and the Gallager exponent of the channel P' , and from [12, Proposition 1], we have

$$\min_Q \max_{1 \leq i \leq m} L_\alpha(P_i, Q) = \max_\mu \frac{1}{\alpha - 1} E_o(\alpha - 1, \mu, P'),$$

where the right-hand side is the maximized Gallager exponent of the channel P' . ($1 < \alpha < 2$ is relevant in [15, p. 138], $1 < \alpha < \infty$ in [15, p. 157], and $0 < \alpha < 1$ in [16]).

5.2.5 The max-min problem for L_α

Thus far our focus has been on the min-max problem of finding the L_α -center. We briefly looked at identifying the max-min value of I_f in (53), but only as a means to study the min-max problem.

Suppose that our new objective is to find

$$\max_\mu \min_Q \sum_{i=1}^m \mu(i) L_\alpha(P_i, Q). \quad (62)$$

This problem is the same as identifying the “capacity of order $1/\alpha$ ” of the channel P' [12], *i.e.*,

$$\max_\mu \min_{Q'} \sum_{i=1}^m \mu(i) D_{1/\alpha}(P'_i \| Q').$$

[12, Proposition 1] solves this problem; the value is the same as the min-max value $\min_{Q'} \max_{1 \leq i \leq m} D_{1/\alpha}(P'_i \| Q')$. Consequently, the max-min value of (62) is the same as the L_α -radius of the family.

5.3 L_α -center and radius for an arbitrary family

We are now back to the case with side information and an infinite family \mathbb{T} . The development in this subsection will be analogous to Gallager’s approach [9] for the source coding problem. We first recall the technical condition indicated in Section 4. \mathbb{T} is a family of PMFs on $\mathbb{X} \times \mathbb{Y}$, $(\mathbb{T}, \mathcal{T})$ a measurable space, and for every $x \in \mathbb{X}$, the mapping $P \mapsto P(x)$ is \mathcal{T} -measurable.

Our focus will be on the following:

Definition 20 For $0 < \alpha < \infty, \alpha \neq 1$,

$$K_+ \stackrel{\Delta}{=} \min_Q \sup_{P \in \mathbb{T}} I(P, Q). \quad (63)$$

Taking Q to be the uniform PMF on $\mathbb{X} \times \mathbb{Y}$ it is easy to check that K_+ is finite; indeed $1 \leq K_+ \leq |\mathbb{X}|^\rho$ when $\rho > 0$ and $-1 \leq K_+ \leq 0$ when $-1 < \rho < 0$.

Let us define some other auxiliary quantities. Let us first define the mapping $f : \mathbb{T} \rightarrow \mathbb{R}_+^{|\mathbb{X}||\mathbb{Y}|}$, as follows:

$$f(P) \triangleq P/h(P).$$

For a probability measure μ on $(\mathbb{T}, \mathcal{T})$, let

$$F \triangleq \int_{\mathbb{T}} d\mu(P) f(P). \quad (64)$$

We define the PMF $\mu f \in \mathcal{P}(\mathbb{X})$ as the scaled version of F ,

$$\mu f \triangleq d^{-1} F \quad (65)$$

where d as in the finite case is the normalizing constant

$$d \triangleq \int_{\mathbb{T}} \frac{d\mu(P)}{h(P)} = \sum_{x \in \mathbb{X}} F(x). \quad (66)$$

These definitions are extensions of (57) and (58) to arbitrary \mathbb{T} . Moreover, let

$$J(\mu, \mathbb{T}) \triangleq \int_{\mathbb{T}} d\mu(P) I(P, \mu f). \quad (67)$$

Simple algebraic manipulations result in

$$J(\mu, \mathbb{T}) = \text{sign}(\rho) \cdot h(F) \quad (68)$$

$$= \text{sign}(\rho) \cdot d \cdot h(\mu f), \quad (69)$$

an extension of [14, Equation (2.24)] for arbitrary \mathbb{T} .

The following auxiliary problem will be useful.

Definition 21 For $0 < \alpha < \infty, \alpha \neq 1$,

$$K_- \triangleq \sup_{\mu} J(\mu, \mathbb{T}). \quad (70)$$

The following parallels will help fix ideas. The quantity μf in (65) is analogous to the PMF at the output of a channel represented by \mathbb{T} when the input measure is μ . $J(\mu, \mathbb{T})$ in (67) is the analogue of mutual information; Csiszár calls it informativity in his work on finite-sized families [14].

Proposition 22 $K_- \leq K_+$.

Proof: Fix an arbitrary PMF Q on $\mathbb{X} \times \mathbb{Y}$. It is straightforward to show that [14, Equation 2.26] holds even when $|\mathbb{T}|$ is not finite, and is given by

$$\int_{\mathbb{T}} d\mu(P) I(P, Q) = \text{sign}(\rho) \cdot J(\mu, \mathbb{T}) \cdot I(\mu f, Q).$$

Since $I(\mu f, Q) \geq \text{sign}(\rho)$, it follows that

$$\int_{\mathbb{T}} d\mu(P) \cdot I(P, Q) \geq J(\mu, \mathbb{T}).$$

Consequently

$$J(\mu, \mathbb{T}) = \min_Q \int_{\mathbb{T}} d\mu(P) I(P, Q),$$

which leads to

$$\begin{aligned} K_- &= \sup_{\mu} J(\mu, \mathbb{T}) \\ &= \sup_{\mu} \min_Q \int_{\mathbb{T}} d\mu(P) I(P, Q) \\ &\leq \min_Q \sup_{\mu} \int_{\mathbb{T}} d\mu(P) I(P, Q) \\ &= \min_Q \sup_{P \in \mathbb{T}} I(P, Q) \\ &= K^+. \end{aligned}$$

■

The following Proposition is similar to [9, Theorem A]. The proof largely runs along similar lines.

Proposition 23 *A real number R equals K_- if and only if there exist a sequence of probability measures $(\mu_n : n \in \mathbb{N})$ on $(\mathbb{T}, \mathcal{T})$ and a PMF Q^* on $\mathbb{X} \times \mathbb{Y}$ with the following properties:*

1. $\lim_n J(\mu_n, \mathbb{T}) = R$;
2. $\lim_n \mu_n f = Q^*$;
3. $I(P, Q^*) \leq R$, for every $P \in \mathbb{T}$.

Furthermore Q^* is unique, attains the minimum in (63), and $K_- = K_+$. □

Proof: \Leftarrow : Observe that on account of 1), 3), and Proposition 22 we have

$$\begin{aligned}
K_- &\geq R \\
&\geq \sup_{P \in \mathbb{T}} I(P, Q^*) \\
&\geq \min_Q \sup_{P \in \mathbb{T}} I(P, Q) \\
&= K_+ \\
&\geq K_-
\end{aligned}$$

where the first inequality follows from 1), the second from 3), and the last from Proposition 22. Consequently, all the inequalities are equalities, $R = K_- = K_+$, and the use of “min” in the definition of K_+ is justified.

\Rightarrow : Since $R = K_- \leq K_+ < \infty$, it follows from the definition of K_- that there exists a sequence $(\mu_n : n \in \mathbb{N})$ such that $\lim_n J(\mu_n, \mathbb{T}) = R$.

Now consider the sequence of $|\mathbb{X}||\mathbb{Y}|$ -dimensional vectors given by $F_n = \int_{\mathbb{T}} d\mu_n(P) f(P)$. This is a sequence of scaled PMFs given by $F_n = d_n \cdot \mu_n f$, where d_n is given by (66). The sequence resides in a compact space of scaled PMFs and therefore has a cluster point F^* which can be normalized to get the PMF Q^* . Moreover we can find a subsequence of $(F_n : n \in \mathbb{N})$ such that $\lim_k F_{n_k} = F^*$. We redefine the sequence μ_n as given by this subsequence, and properties 1) and 2) hold.

Suppose now that there is a $P_0 \in \mathbb{T}$ such that 3) is violated, *i.e.*,

$$I(P_0, Q^*) > K_-.$$

Consider the convex combinations of measures

$$\nu_{n,\lambda} = (1 - \lambda)\mu_n + (\lambda)\delta_{P_0}, \quad (71)$$

where δ_{P_0} is the atomic distribution on P_0 .

From (71), (64), and (68), we have

$$\begin{aligned}
s_n(\lambda) &\triangleq J(\nu_{n,\lambda}, \mathbb{T}) \\
&= \text{sign}(\rho) \cdot h((1 - \lambda)F_n + \lambda f(P_0)).
\end{aligned}$$

Since $\text{sign}(\rho)h(\cdot)$ is a concave and therefore continuous function of its vector-valued argument, $s_n(\lambda)$ converges point-wise to

$$s(\lambda) = \text{sign}(\rho) \cdot h((1 - \lambda)F^* + \lambda f(P_0)),$$

for $\lambda \in [0, 1]$. In particular, $s(0) = \lim_n s_n(0) = K_-$. Now, $s(\lambda)$ is a concave function of λ since $\text{sign}(\rho)h(\cdot)$ is concave and the argument is linear in λ .

Next, we can straightforwardly check that

$$\dot{s}(0) = I(P_0, Q^*) - K_- > 0,$$

with the possibility that the value (slope at $\lambda = 0$) may be $+\infty$.

We have therefore established that $s(\lambda)$ has $s(0) = K_-$, is concave and therefore continuous in $[0, 1]$, and has strictly positive slope at $\lambda = 0$. Consequently, $s(\lambda) > K_-$ for some $0 < \lambda < 1$. Since

$$J(\nu_{n,\lambda}, \mathbb{T}) = s_n(\lambda) \rightarrow s(\lambda) > K_-$$

contradicts the definition of K_- , 3) must hold.

To show uniqueness of Q^* , suppose there were another R^* and another sequence of measures $(\pi_n : n \in \mathbb{N})$ satisfying 1), 2) and 3). We can get two cluster points F^* and G^* that when normalized lead to Q^* and R^* , respectively. Then with $\nu_n = \frac{1}{2}\mu_n + \frac{1}{2}\pi_n$, we have

$$\begin{aligned} J(\nu_n, \mathbb{T}) &\rightarrow \text{sign}(\rho) \cdot h\left(\frac{1}{2}F^* + \frac{1}{2}G^*\right) \\ &> \frac{1}{2} \cdot \text{sign}(\rho) \cdot h(F^*) + \frac{1}{2} \cdot \text{sign}(\rho) \cdot h(G^*) \\ &= \frac{1}{2}K_- + \frac{1}{2}K_- \\ &= K_-, \end{aligned}$$

a contradiction. The strict inequality above is due to strict concavity of $\text{sign}(\rho)h(\cdot)$ when $\rho > -1$. ■

5.4 Proof of Theorem 10

Proof: From (28), it is clear that

$$C = \frac{1}{\rho} \log(\text{sign}(\rho) \cdot K_+).$$

Q attains the min-sup value K_+ in Definition 20 if and only if Q attains the min-sup value C in Definition 9. Proposition 23 guarantees the existence and uniqueness of such a Q . ■

6 Examples

In this section we look at two example classes of PMFs, and identify their L_α -centers and radii. We focus on guessing without side information. Throughout this section, therefore, $0 < \alpha < 1$ and $|\mathbb{Y}| = 1$. The uncertainty class will thus be PMFs in \mathbb{X} (with no reference to $|\mathbb{Y}|$).

6.1 The class of discrete memoryless sources

Let \mathbb{A} be a finite alphabet set, n a positive integer, and $\mathbb{X} = \mathbb{A}^n$. We wish to guess n -strings with letters drawn from \mathbb{A} . Let $a^n = (a_1, \dots, a_n) \in \mathbb{A}^n$. Let $\mathcal{P}(\mathbb{X})$ denote the set of all PMFs on \mathbb{X} .

Let \mathbb{T} be the class of all discrete memoryless sources (DMS) on \mathbb{A} , *i.e.*,

$$\mathbb{T} = \left\{ P_n \in \mathcal{P}(\mathbb{A}^n) \mid P_n(a^n) = \prod_{i=1}^n P(a_i), \forall a^n \in \mathbb{A}^n, \text{ and } P \in \mathcal{P}(\mathbb{A}) \right\},$$

The parameters of the source P are unknown to the guesser. Arikan and Merhav [4] provide a guessing scheme for this uncertainty class. The scheme happens to be independent of ρ . Moreover, their guessing scheme has the same asymptotic performance as the optimal guessing scheme. The guessing order was to guess in the increasing order of empirical entropies; strings with identical letters are guessed first, then strings with exactly one different letter, and so on. Within each type of sequence, the order of guessing is irrelevant. Denote this guessing list by G_n . Arikan and Merhav [4, Theorem 1] showed that for any $P_n \in \mathbb{T}$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} R(P_n, G_n) = 0.$$

The above result is couched in our notation. This indicates that \mathbb{T} , the class of all DMSs on \mathbb{A} , is not rich enough in the sense that there exists a “universal” guessing scheme. The following result makes this notion more precise.

Theorem 24 (*Class of DMS on \mathbb{A}*) *Let $m = |\mathbb{A}|$. The L_α -radius C_n of the class of discrete memoryless sources on \mathbb{A} satisfies*

$$C_n \leq \frac{m-1}{2} \log \frac{n}{2\pi} + u_m + \varepsilon_n,$$

where $u_m = \log(\Gamma(1/2)^m / \Gamma(m/2))$, a constant that depends on the alphabet size, and ε_n is a sequence in n that vanishes as $n \rightarrow \infty$. \square

Proof: Recall that $\rho > 0$. P_n is the joint PMF of the n -string with individual letter probabilities P . It is easy to verify that P'_n is the joint PMF of the n -string with individual letter probabilities P' , and therefore P'_n also belongs to \mathbb{T} . Xie and Barron [17, Theorem 2] show that there is a PMF on \mathbb{A}^n , say Q'_n , and a vanishing sequence ε_n , such that for every discrete

memoryless source P'_n , the following holds:

$$\max_{a^n \in \mathbb{A}^n} \log \frac{P'_n(a^n)}{Q'_n(a^n)} \leq \max_{a^n \in \mathbb{A}^n} \log \frac{\hat{P}'_n(a^n)}{Q'_n(a^n)} \quad (72)$$

$$\begin{aligned} &\leq r_n \\ &\triangleq \frac{m-1}{2} \log \frac{n}{2\pi} + u_m + \varepsilon_n, \end{aligned} \quad (73)$$

where \hat{P}'_n in the right-side of (72) is the DMS that maximizes $R'_n(a^n) = \prod_{i=1}^n R'(a_i)$, for a given sequence a^n .

Define the PMF Q_n as follows:

$$Q_n(\cdot) \propto (Q'_n(\cdot))^{1/\alpha},$$

the inverse of the mapping in (20). We then have the following series of inequalities:

$$\begin{aligned} L_\alpha(P_n, Q_n) \\ = \frac{1}{\rho} \log \left(\sum_{a^n \in \mathbb{A}^n} P'_n(a^n) \left(\frac{P'_n(a^n)}{Q'_n(a^n)} \right)^\rho \right) \end{aligned} \quad (74)$$

$$\leq \frac{1}{\rho} \left(\log \sum_{a^n \in \mathbb{A}^n} P'_n(a^n) \cdot \exp\{\rho r_n\} \right) \quad (75)$$

$$\begin{aligned} &= \frac{1}{\rho} \log (\exp\{\rho r_n\}) \\ &= r_n, \end{aligned}$$

where (74) follows from (21) and (75) from (73). Taking the supremum over all P_n yields the theorem. ■

Remark : Redundancy in guessing is thus upper bounded by $r_n + \log(1 + n \ln |\mathbb{A}|)$. Since the L_α -radius grows with n as $O(\log n)$, the normalized redundancy C_n/n vanishes. This implies that we can get a “universal” guessing strategy. Theorem 24 suggests the use of Q_n , which in general may depend on ρ . Arikán and Merhav’s technique of guessing in the order of increasing empirical entropy is another universal guessing technique.

Given any guessing scheme, how do we “measure” the set of DMSs which result in relatively large redundancy? The following theorem answers this question, and uses a strong version of the redundancy capacity theorem of universal coding in [18] and [19].

Theorem 25 *Let Q_n be any PMF on \mathbb{A}^n . Let μ be a probability measure on $(\mathbb{T}, \mathcal{T})$ and let $P'_{n,\mu} = \int_{\mathbb{T}} d\mu(P'_n) P'_n$. Then for any DMS P_n , we have*

$$L(P_n, Q_n) \geq D(P'_n \parallel P'_{n,\mu}) - \lambda_n$$

except on a set B of μ -probability $\mu\{B\} \leq 2^{-n\lambda_n}$.

Proof: Observe that $\rho > 0$. An application of Jensen's inequality to the concave function $\log(\cdot)$ yields

$$\begin{aligned} L(P_n, Q_n) &= \frac{1}{\rho} \log \left(\sum_{a^n \in \mathbb{A}^n} P'_n(a^n) \left(\frac{P'_n(a^n)}{Q'_n(a^n)} \right)^\rho \right) \\ &\geq \frac{1}{\rho} \sum_{a^n \in \mathbb{A}^n} P'_n(a^n) \log \left(\frac{P'_n(a^n)}{Q'_n(a^n)} \right)^\rho \\ &= D(P'_n \parallel Q'_n). \end{aligned}$$

The theorem then follows from [19, Theorem 2] which states that the source coding redundancy $D(P'_n \parallel Q'_n)$ is at least as large as $D(P'_n \parallel P'_{n,\mu}) - \lambda_n$ except on a set B of μ -probability upper bounded by $2^{-n\lambda_n}$. \blacksquare

Remark : In particular, we may do the following. We choose μ such that $D(P'_n \parallel P'_{n,\mu}) = r_n$. (This can be done since the inf-sup value of $\inf_{Q'_n} \sup_{P'_n} D(P'_n \parallel Q'_n)$ is r_n , as remarked in [17, Remark 5 after Theorem 2]. We may then choose λ_n such that $n\lambda_n \rightarrow \infty$ so that $2^{-n\lambda_n}$ vanishes with n , but λ_n is negligibly small compared to r_n . (For example, for the class of DMSs, $r_n = O(\log n)$. We may choose λ_n as $\log \log n$ or even $(\log \log n)/(\log n)$. Then, the set of sources P for which $L_\alpha(P_n, Q_n) \leq r_n - \lambda_n$ has negligible μ -probability for all sufficiently large n . Or in other words, with very high μ -probability (at least $1 - 2^{-n\lambda_n}$), the redundancy in guessing for any strategy is at least $r_n - \lambda_n - \log(1 + n \ln |\mathbb{A}|)$.

6.2 Arbitrarily varying sources

For the class of DMSs, we saw in Section 6.1 that the redundancy is upper-bounded by $O(\log n)$. In this section we look at the example of finite-state arbitrarily varying sources (FS-AVS) for which the redundancy grows linearly with n . Yet again, for exposition purposes, we assume $|\mathbb{Y}| = 1$.

As before, let $\mathbb{X} = \mathbb{A}^n$. Let \mathbb{S} be a finite set of *states*, and for each $s \in \mathbb{S}$, let $P(\cdot \mid s)$ be a PMF on the finite set \mathbb{A} . An *arbitrarily varying source* (AVS) is a sequence of \mathbb{A} -valued random variables X_1, X_2, \dots , such that X_i 's are independent and the probability of an n -string x^n is governed by an arbitrary state sequence $s^n \in \mathbb{S}^n$ as follows:

$$P_n(x^n \mid s^n) = \prod_{i=1}^n P(x_i \mid s_i).$$

Observe that for a fixed n , there are only $|\mathbb{S}|^n$ sources in the uncertainty set. Let T_{s^n} be the subset of all sequences in \mathbb{S}^n with the same letter-frequencies as s^n . T_{s^n} is also referred to as the type of the sequence s^n [20].

If the letter frequencies are given by a PMF U on \mathbb{S} , we refer to T_U as the type of sequences. Let V be a stochastic matrix given by $V(x | s)$ for $x \in \mathbb{A}$ and $s \in \mathbb{S}$. Then for a particular sequence s^n , we refer to $T_V(s^n)$, the set of sequences that are of conditional type V given s^n , as the V -shell of s^n .

Proposition 26 *Let $0 < \alpha < 1$. Let T_U be a type of sequences on \mathbb{S}^n . Let the uncertainty set \mathbb{T} be given by $\mathbb{T} = \{P_n(\cdot | s^n) \mid s^n \in T_U\}$. The L_α -radius of this family is given by*

$$R_n(T_U) \triangleq H_\alpha(Q_n^*) - \frac{1}{|T_U|} \sum_{s^n \in T_U} H_\alpha(P_n(\cdot | s^n)), \quad (76)$$

where the L_α -center Q_n^* is given by

$$Q_n^*(\cdot) = \frac{1}{|T_U|} \sum_{s^n \in T_U} P_n(\cdot | s^n). \quad (77)$$

□

Remarks : 1) It will be apparent from the proof that the quantity $H_\alpha(P_n(\cdot | s^n))$ in (76) depends on s^n only through its type, and hence the average over all sequences in the type may be replaced by the value for any specific $s^n \in T_U$.

2) All PMFs in the uncertainty set are spaced equally apart (in the sense of L_α -divergence) from the L_α -center Q_n^* .

3) Guessing in the decreasing order of Q_n^* -probabilities results in a redundancy in guessing that is upper bounded by $R_n(T_U) + \log(1 + n \ln |\mathbb{A}|)$.

4) $\text{sign}(\rho) \cdot h(P)$ is a concave function of P . It follows from (27) that $H_\alpha(P)$ is also a concave function of P for $0 < \alpha < 1$. By Jensen's inequality, $R_n(T_U) \geq 0$. (For $\alpha > 1$, $H_\alpha(P)$ is neither concave nor convex in P).

5) For any guessing strategy, there exists at least one sequence $s^n \in T_U$ for which the redundancy is lower bounded by $R_n(T_U) - \log(1 + n \ln |\mathbb{A}|)$. We will see later in Proposition 28 that if the U sequence (parameterized by n) converges as $n \rightarrow \infty$ to a PMF $U^* \in \mathcal{P}(\mathbb{S})$, then $\frac{1}{n}R_n(T_U)$ converges to a strictly positive constant. Thus $R_n(T_U)$ grows linearly with n , thereby making the converse meaningful; the nuisance term $\log(1 + n \ln |\mathbb{A}|)$ grows only logarithmically in n .

Proof: Note that given an n , the uncertainty set is finite. We will simply show that the candidate L_α -center satisfies the necessary and sufficient condition (60) given in Section 5.2.3. From (29), it is sufficient to show that

$$\begin{aligned} & I_f(P'_n(\cdot | s^n \parallel Q_n^{*\prime})) \\ &= \frac{\sum_{x^n \in \mathbb{A}^n} P_n(x^n | s^n) \left(Q_n^{*\prime}(x^n)\right)^{-\rho}}{h(P_n(\cdot | s^n))} \\ &= K, \end{aligned} \quad (78)$$

where K is some constant that depends only on n and T_U . We will show that the numerator and denominator in (78) do not depend on the actual s^n , so long as $s^n \in T_U$.

Observe that the stochastic matrix that defines the conditional PMF is given by $P(x | s)$ for $x \in \mathbb{A}$ and $s \in \mathbb{S}$. Consider $h(P_n(\cdot | s^n))$. First

$$\begin{aligned} & \sum_{x^n \in \mathbb{A}^n} (P_n(x^n | s^n))^\alpha \\ &= \sum_V |T_V(s^n)| \exp \{-n\alpha [D(V \| P | U) + H(V | U)]\} \end{aligned}$$

where the sum is over all conditional types V . All the quantities inside the summation, including $|T_V(s^n)|$, depend on s^n only through T_U , and therefore $h(P_n(\cdot | s^n))$ depends on s^n only through T_U .

Next, $Q_n^*(x^n)$ depends on x^n only through T_{x^n} . This is easily seen via a permutation argument. Given two \mathbb{A} -sequences of the same type, let π be a permutation that takes (x^n, s^n) to $((x_{\pi(1)}, \dots, x_{\pi(n)}), (s_{\pi(1)}, \dots, s_{\pi(n)}))$, where s^n and $(s_{\pi(1)}, \dots, s_{\pi(n)})$ are the two given \mathbb{A} -sequences. This permutation π leaves $P_n(x^n | s^n)$ unchanged. Moreover, the sum continues to be over

$$\begin{aligned} T_U = \{ & (s_{\pi(1)}, s_{\pi(2)}, \dots, s_{\pi(n)}) \in \mathbb{S}^n \mid \\ & s^n = (s_1, \dots, s_n) \in T_U \}. \end{aligned}$$

Thus $Q_n^*(x^n)$ and therefore $Q_n^{*\prime}(x^n)$ depend on x^n only through T_{x^n} .

Finally, given two \mathbb{A} -sequences of the same type T_U , the above permutation argument indicates that

$$\sum_{x^n \in \mathbb{A}^n} P_n(x^n | s^n) \left(Q_n^{*\prime}(x^n) \right)^{-\rho},$$

the numerator of (78), depends on s^n only through T_U .

That $R_n(T_U)$ is given by (76) follows from (57), (58), (59), the fact that $h(P_n(\cdot | s^n))$ is a constant over all $s^n \in T_U$, and (27). This concludes the proof. \blacksquare

The number of different types of sequences grows polynomially in n , in particular, this number is upper bounded by $(n+1)^{|\mathbb{S}|}$. We can use this fact to stitch together the guessing lists for the different types of sequences on \mathbb{S}^n and get one list that does only marginally worse than the list obtained by knowing the type of the state sequence.

Proposition 27 *Let $0 < \alpha < 1$. Let the uncertainty set \mathbb{T} be given by $\mathbb{T} = \{P_n(\cdot | s^n) \mid s^n \in \mathbb{S}^n\}$. There is a guessing strategy such that for every T_U , the redundancy is upper bounded by*

$$R_n(T_U) + \log(1 + n \ln |\mathbb{A}|) + |\mathbb{S}| \log(n+1).$$

whenever $s^n \in T_U$. \square

Proof: Let N be the number of types. N is upper bounded by $(n+1)^{|\mathbb{S}|}$. Fix an arbitrary order on these types. Let the k th type be T_U . Set $G_k = G_{T_U}$, where G_{T_U} is the guessing strategy that is obtained knowing that $s^n \in T_U$, via Proposition 26. It proceeds in the decreasing order of probabilities of the L_α -center of the uncertainty set indexed by T_U .

We now stitch together the guessing lists G_1, G_2, \dots, G_N to get a new guessing list G , as follows. Think of G_k as a column vector of size $|\mathbb{A}^n| \times 1$ and let H be the column vector of size $N \cdot |\mathbb{A}^n| \times 1$ obtained by reading the entries of the matrix $[G_1 \ G_2 \ \dots \ G_N]$ in raster order (one row after another). Every \mathbb{A} would have figured exactly once in the G_k list, and therefore occurs exactly N times in the H list. Next, prune the H list. For each i , if there exists an index j with $j < i$ and $H_i = H_j$, set $H_i = \delta$. This indicates that the i th string already figures in the final guessing list. Finally remove all δ 's to obtain the desired guessing list $G : \mathbb{A}^n \rightarrow \{1, 2, \dots, |\mathbb{A}|^n\}$, where $G(x^n)$ is the unique position at which x^n occurs in the pruned H list.

Clearly, for every x^n and for every k such that $1 \leq k \leq N$, we have $G(x^n) \leq NG_k(x^n)$. Indeed, x^n occurs in the position $(G_k(x^n), k)$ in the matrix constructed above. It therefore occurs in position $(G_k(x^n) - 1)N + k$ and therefore before the position $NG_k(x^n)$ in the unpruned H list. It cannot be placed any later in the pruned H list, and thus $G(x^n) \leq NG_k(x^n)$.

The above observation leads to

$$\frac{1}{\rho} \log \mathbb{E} [G(X^n)^\rho] \leq \frac{1}{\rho} \log \mathbb{E} [G(X^n)^\rho] + \log N.$$

The proposition follows from Theorem 6, Proposition 26, and the bounding $N \leq (n+1)^{|\mathbb{S}|}$. \blacksquare

We finally remark that the min-sup redundancy for the finite-state arbitrarily varying source grows linearly with n under some circumstances.

Proposition 28 *For a fixed n , let U be a PMF on \mathbb{S} and T_U the corresponding type. Let the sequence U (as a function of n) converge to a PMF $U^* \in \mathcal{P}(\mathbb{S})$ as $n \rightarrow \infty$. Then*

$$\lim_n \frac{1}{n} R_n(T_U) = R,$$

where $R \geq 0$. \square

Proof: The second term in the right-hand side of (76), after normal-

ization by n , converges to a nonnegative real number as seen below:

$$\begin{aligned}
& \frac{1}{n} H_\alpha(P_n(\cdot | s^n)) \\
&= \frac{1}{n(1-\alpha)} \log \sum_{x^n \in \mathbb{A}^n} \prod_{i=1}^n P(x_i | s_i)^\alpha \\
&= \frac{1}{n(1-\alpha)} \log \prod_{s \in \mathbb{S}} \left(\sum_{x \in \mathbb{A}} P(x | s)^\alpha \right)^{nU(s)} \\
&= \sum_{s \in \mathbb{S}} U(s) H_\alpha(P(\cdot | s)) \\
&\rightarrow \sum_{s \in \mathbb{S}} U^*(s) H_\alpha(P(\cdot | s)). \tag{79}
\end{aligned}$$

We next consider the first term on the right-hand side of (76) after normalization, *i.e.*, $H_\alpha(Q_n^*)/n$, where Q_n^* is given by (77).

Lemma 29 *For a fixed n , let U be a PMF on \mathbb{S} and T_U the corresponding type. Let the sequence U (as a function of n) converge to a PMF $U^* \in \mathcal{P}(\mathbb{S})$ as $n \rightarrow \infty$. Let V be the output PMF when the input PMF on \mathbb{S} is U and the channel is P . Furthermore, let V^* be the limiting output PMF as $n \rightarrow \infty$. Then $\lim_n \frac{1}{n} H_\alpha(Q_n^*) = H_\alpha(V^*)$.*

As a consequence of this lemma and (79), we have

$$\frac{1}{n} R_n(T_U) \rightarrow H_\alpha(V^*) - \sum_{s \in \mathbb{S}} U^*(s) H_\alpha(P(\cdot | s)) \stackrel{\Delta}{=} R.$$

By the strict concavity of $H_\alpha(\cdot)$ for $0 < \alpha < 1$, and Jensen's inequality, we have $R \geq 0$. This concludes the proof of the theorem. \blacksquare

Remarks : $R = 0$ if and only if either (i) $U(s) = 1$ for some $s \in \mathbb{S}$, or (ii) $P(\cdot | s)$ does not depend on s , *i.e.*, the state does not affect the source. Thus, for all but the trivial finite-state arbitrarily varying sources, the min-sup redundancy grows exponentially with n at a rate R . This means that the guessing strategy that achieves the min-sup redundancy has an exponential growth rate strictly bigger than that of the best strategy obtained with knowledge of the state sequence.

We now prove the rather technical Lemma 29.

Proof:

(a) We first show that $\lim_n \frac{1}{n} H_\alpha(Q_n^*) \leq H_\alpha(V^*)$.

Let U_n be the PMF on \mathbb{S}^n given by $U_n(s^n) = \prod_{i=1}^n U(s_i)$. Let $U_n\{T\}$ denote the U_n -probability of the set T . From (77), we may write

$$\begin{aligned} & \sum_{x^n \in \mathbb{A}^n} Q_n^*(x^n)^\alpha \\ &= \sum_{x^n \in \mathbb{A}^n} \left(\frac{1}{|T_U|} \sum_{s^n \in T_U} P_n(x^n | s^n) \right)^\alpha \\ &= \sum_{x^n \in \mathbb{A}^n} \left(\frac{1}{U_n\{T_U\}} \frac{U_n\{T_U\}}{|T_U|} \sum_{s^n \in T_U} P_n(x^n | s^n) \right)^\alpha \\ &= \frac{1}{U_n\{T_U\}} \sum_{x^n \in \mathbb{A}^n} \left(\sum_{s^n \in T_U} U_n(s^n) P_n(x^n | s^n) \right)^\alpha \end{aligned} \quad (80)$$

$$\leq (n+1)^{|\mathbb{S}|^\alpha} \sum_{x^n \in \mathbb{A}^n} \left(\sum_{s^n \in \mathbb{S}^n} U_n(s^n) P_n(x^n | s^n) \right)^\alpha \quad (81)$$

$$\begin{aligned} &= (n+1)^{|\mathbb{S}|^\alpha} \sum_{x^n \in \mathbb{A}^n} V_n(x^n)^\alpha \\ &= (n+1)^{|\mathbb{S}|^\alpha} \left(\sum_{x \in \mathbb{A}} V(x)^\alpha \right)^n, \end{aligned} \quad (82)$$

where (80) follows from the observation that $U_n(s^n) = U_n\{T_U\}/|T_U|$ for all $s^n \in T_U$, (81) from $U_n\{T_U\} \geq (n+1)^{-|\mathbb{S}|}$ (see proof of [20, Lemma 2.3]) and by enlarging the sum over T_U to include all of \mathbb{S}^n .

From (82) and (27), we have

$$\begin{aligned} \frac{1}{n} H_\alpha(Q_n^*) &\leq \frac{\alpha |\mathbb{S}|}{1-\alpha} \frac{\log(n+1)}{n} + H_\alpha(V) \\ &\rightarrow H_\alpha(V^*). \end{aligned}$$

(b) We now show that $\lim_n \frac{1}{n} H_\alpha(Q_n^*) \geq H_\alpha(V^*)$.

For a given PMF U on \mathbb{S} and conditional source P , let V be the induced PMF on \mathbb{X} and W the reverse conditional probability matrix, *i.e.*, $W(s | x)$ is the probability of a state s given x .

Continuing from (80), we may write

$$\begin{aligned}
& \sum_{x^n \in \mathbb{A}^n} Q_n^*(x^n)^\alpha \\
&= \frac{1}{U_n\{T_U\}} \sum_{x^n \in \mathbb{A}^n} \left(\sum_{s^n \in T_U} U_n(s^n) P_n(x^n | s^n) \right)^\alpha \\
&\geq \sum_{x^n \in T_{\bar{Q}}} \left(\sum_{s^n \in T_U} U_n(s^n) P_n(x^n | s^n) \right)^\alpha
\end{aligned} \tag{83}$$

$$\geq \sum_{x^n \in T_{\bar{Q}}} \left(\sum_{s^n \in T_{\bar{W}}(x^n) \subset T_U} V_n(x^n) W_n(s^n | x^n) \right)^\alpha \tag{84}$$

$$= \sum_{x^n \in T_{\bar{Q}}} (V_n(x^n) W_n\{T_{\bar{W}}(x^n) | x^n\})^\alpha, \tag{85}$$

where (83) follows because $U_n\{T_U\} \leq 1$ and the sum over \mathbb{A}^n is restricted to a sum over a type $T_{\bar{Q}}$ to be chosen later; (84) follows because $U_n(s^n) P_n(x^n | s^n) = V_n(x^n) W_n(s^n | x^n)$ and the sum over s^n is now restricted over a non-void \bar{W} -shell of x^n , where \bar{W} will be appropriately chosen later.

We next observe that for $x^n \in T_{\bar{Q}}$, the following hold:

$$\begin{aligned}
V_n(x^n) &= 2^{-n(H(\bar{Q})+D(\bar{Q}\|V))}, \\
W_n\{T_{\bar{W}}(x^n)\} &\geq (n+1)^{-|\mathbb{X}||\mathbb{S}|} \cdot 2^{-nD(\bar{W}\|W|\bar{Q})}, \text{ and} \\
|T_{\bar{Q}}| &\geq (n+1)^{-|\mathbb{X}|} \cdot 2^{nH(\bar{Q})}.
\end{aligned}$$

Substitution of these inequalities into (85) yields

$$\begin{aligned}
& \sum_{x^n \in \mathbb{A}^n} Q_n^*(x^n)^\alpha \\
&\geq (n+1)^{-|\mathbb{X}|(1+\alpha|\mathbb{S}|)} \\
&\quad \cdot 2^{n[(1-\alpha)H(\bar{Q})-\alpha(D(\bar{Q}\|V)+D(\bar{W}\|W|\bar{Q}))]}
\end{aligned}$$

and therefore

$$\begin{aligned}
& \frac{1}{n} H_\alpha(Q_n^*) \\
&\geq H(\bar{Q}) - \frac{1}{\rho} [D(\bar{Q} \| V) + D(\bar{W} \| W | \bar{Q})] \\
&\quad - \frac{|\mathbb{X}|(1+\alpha|\mathbb{S}|)}{1-\alpha} \frac{\log(n+1)}{n}
\end{aligned} \tag{86}$$

for any type \bar{Q} of sequences and for any \bar{W} such that $T_{\bar{W}}(x^n)$ is a non-void shell for an $x^n \in T_{\bar{Q}}$.

Clearly, the last term in (86) vanishes as $n \rightarrow \infty$.

If we can choose $\bar{Q} = V'$ and $\bar{W} = W$, we will be done since $H_\alpha(V) = H(V') - \frac{1}{\rho}D(V' \parallel V)$. We cannot do this if V' is not a type of sequences, or if W is not a conditional type given an x^n . But we will show that as $n \rightarrow \infty$, we can get close enough. The following arguments make this idea precise.

Define

$$\delta \triangleq \log \min\{W(s \mid x) \mid W(s \mid x) > 0, s \in \mathbb{S}, x \in \mathbb{X}\}$$

and consider $D(\bar{W}(\cdot \mid x) \parallel W(\cdot \mid x))$. We may restrict our choice of \bar{W} to those that are absolutely continuous with respect to W , *i.e.*, $W(\cdot \mid x) \ll W(\cdot \mid x)$ for every $x \in \mathbb{X}$. For sufficiently large n , we can choose such a \bar{W} that in addition satisfies

$$\sum_{s \in \mathbb{S}} |W(s \mid x) - \bar{W}(s \mid x)| \leq \varepsilon_n \leq \frac{1}{2}, \quad \forall x \in \mathbb{X},$$

and $\varepsilon_n \rightarrow 0$.

We then have

$$\begin{aligned} & D(\bar{W}(\cdot \mid x) \parallel W(\cdot \mid x)) \\ &= H(W(\cdot \mid x)) - H(\bar{W}(\cdot \mid x)) \\ &+ \sum_{s \in \mathbb{S}} (W(s \mid x) - \bar{W}(s \mid x)) \log W(s \mid x) \\ &\leq |H(W(\cdot \mid x)) - H(\bar{W}(\cdot \mid x))| \\ &- (\log \delta) \sum_{s \in \mathbb{S}} |W(s \mid x) - \bar{W}(s \mid x)| \\ &\leq -\varepsilon_n \log \frac{\varepsilon_n}{|\mathbb{S}|} - \varepsilon_n \delta, \end{aligned} \tag{87}$$

where (87) follows from [20, Lemma 2.7]. After averaging, we get

$$D(\bar{W} \parallel W \mid \bar{Q}) \leq -\varepsilon_n \log \frac{\varepsilon_n}{|\mathbb{S}|} - \varepsilon_n \delta \rightarrow 0.$$

A similar argument shows that

$$\begin{aligned} & H(\bar{Q}) - \frac{1}{\rho}D(\bar{Q} \parallel V) \\ &= H_\alpha(V) + [H(\bar{Q}) - H(V')] \\ &- \frac{1}{\rho} [D(\bar{Q} \parallel V) - D(V' \parallel V)] \\ &\rightarrow H_\alpha(V^*), \end{aligned}$$

where we have made use of the fact that $H_\alpha(V) = H(V') - (1/\rho)D(V' \parallel V)$. This concludes the proof of Lemma 29 \blacksquare

7 Concluding remarks

We conclude this paper by applying some of our results to guessing of binary n -strings. Let $\mathbb{X} = \{0, 1\}^n$, and P a PMF on $\{0, 1\}$. Let

$$P_n(x^n) = \prod_{i=1}^n P(X_i = x_i)$$

denote the PMF of the discrete memoryless source (DMS) where the n -string $x^n = (x_1, x_2, \dots, x_n)$. One can think of X^n as a string of outcomes of independent tosses of the same biased coin. Theorem 5 says that for $\rho = 1$, the minimum expected number of guesses grows exponentially with n ; the growth rate is given by $H_{1/2}(P)$.

If the only information that the guesser has about the source is that $P_n \in \mathbb{T}$, the guesser suffers a penalty (interchangeably called redundancy); growth rate of the minimum expected number of guesses is larger than that achievable with knowledge of P_n . The increase in growth rate is given by the normalized redundancy $R(P_n, G)/n$, where G is the guessing strategy chosen to work for all sources in \mathbb{T} . This normalized redundancy equals the normalized $L_{1/2}$ -radius of \mathbb{T} , *i.e.*, C_n/n , where C_n is given by (17).

When P_n is a DMS, and the PMF P on $\{0, 1\}$ is unknown to the guesser, Arikan and Merhav [4] have shown that guessing strings in the increasing order of their empirical entropies is a universal strategy. Their universality result implies that the normalized $L_{1/2}$ -radius of the family of DMSs satisfies $C_n/n \rightarrow 0$. The set of DMSs is thus not rich enough from the point of view of guessing. Knowledge of the PMF P is not needed; the universal strategy achieves, asymptotically, the minimum growth rate achievable with full knowledge of the source statistics.

Suppose now that two biased coins are available. To generate each X_i , one of the two coins is chosen arbitrarily, and tossed. The outcome of the toss determines X_i . This is a two-state arbitrarily varying source. We may assume $\mathbb{S} = \{a, b\}$. Let us assume that as $n \rightarrow \infty$, the fraction of time when the first coin is picked approaches a limit $U^*(a)$. Let us further assume that for each n , the receiver knows how many times the first coin was picked, *i.e.*, it knows the type of the state sequence. If the two coins are not statistically identical, the normalized $L_{1/2}$ -radius approaches a strictly positive constant as $n \rightarrow \infty$. This implies that the growth rate in the minimum expected number of guesses for a strategy without full knowledge of source statistics is strictly larger than that achievable with full knowledge of source statistics. We note that in order to maximize the expected number of guesses, the right solution may be to pick one coin, the one with the higher entropy, all the time.

The guesser's lack of knowledge of the number of times the first coin is picked results in additional redundancy. However this additional redundancy

asymptotically vanishes. The guesser “stitches” together the best guessing lists for each type of state sequences.

Acknowledgements The author gratefully acknowledges support from the following sources: Ministry of Human Resources and Development under Grant Part (2A) Tenth Plan (338/ECE), DRDO-IISc joint programme on mathematical engineering, and University Grants Commission under Grant Part (2B) UGC-CAS-(Ph.IV).

References

- [1] J. L. Massey, “Guessing and entropy,” in *Proc. 1994 IEEE International Symposium on Information Theory*, Trondheim, Norway, Jun. 1994, p. 204.
- [2] E. Arikán, “An inequality on guessing and its application to sequential decoding,” *IEEE Trans. Inform. Theory*, vol. IT-42, pp. 99–105, Jan. 1996.
- [3] I. Csiszár, “I-divergence geometry of probability distributions and minimization problems,” *The Annals of Probability*, vol. 3, pp. 146–158, 1975.
- [4] E. Arikán and N. Merhav, “Guessing subject to distortion,” *IEEE Trans. Inform. Theory*, vol. IT-44, pp. 1041–1056, May 1998.
- [5] L. L. Campbell, “A coding theorem and Rényi’s entropy,” *Information and Control*, vol. 8, pp. 423–429, 1965.
- [6] ———, “Definition of entropy by means of a coding problem,” *Z. Wahrscheinlichkeitstheorie verw. Geb.*, vol. 6, pp. 113–118, 1966.
- [7] T. R. M. Fischer, “Some remarks on the role of inaccuracy in Shannon’s theory of information transmission,” in *Trans. 8th Prague Conf. on Information Theory*, 1978, pp. 211–226.
- [8] L. D. Davisson, “Universal noiseless coding,” *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1972.
- [9] R. G. Gallager, “Source coding with side information and universal coding,” *LIDS Technical Report, LIDS-P-937*, Sept. 1979.
- [10] L. D. Davisson and A. Leon-Garcia, “A source matching approach to finding minimax codes,” *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 166–174, Mar. 1980.
- [11] R. Sundaresan, “A measure of discrimination and its geometric properties,” in *Proc. 2002 IEEE Int. Symp. on Information Theory*, Lausanne, Switzerland, Jun. 2002, p. 264.
- [12] I. Csiszár, “Generalized cutoff rates and Rényi’s information measures,” *IEEE Trans. Inform. Theory*, vol. IT-41, pp. 26–34, Jan. 1995.
- [13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, 1991.
- [14] I. Csiszár, “A class of measures of informativity of observation channels,” *Periodica Mathematica Hungarica*, vol. 2, pp. 191–213, 1972.
- [15] R. G. Gallager, *Information Theory and Reliable Communication*. New York: John Wiley & Sons, 1968.
- [16] S. Arimoto, “On the converse to the coding theorem for discrete memoryless channels,” *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 357–359, May 1973.
- [17] Q. Xie and A. R. Barron, “Asymptotic minimax regret for data compression, gambling, and prediction,” *IEEE Trans. Inform. Theory*, vol. IT-46, pp. 431–445, Mar. 2000.
- [18] N. Merhav and M. Feder, “A strong version of the redundancy-capacity theorem of universal coding,” *IEEE Trans. Inform. Theory*, vol. IT-41, pp. 714–722, May. 1995.
- [19] M. Feder and N. Merhav, “Hierarchical universal coding,” *IEEE Trans. Inform. Theory*, vol. IT-42, pp. 1354–1364, Sep. 1996.

[20] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Budapest: Academic Press, 1986.