Guessing Under Source Uncertainty

Rajesh Sundaresan, *Senior Member, IEEE*

***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 set. The min-sup redundancies for two examples—the families of discrete memoryless sources and finite-state arbitrarily varying sources—are then determined.**

***Index Terms*—f-divergence, I-projection, guessing, information geometry, mismatch, Pythagorean identity, Rényi entropy, Rényi information divergence, redundancy, side information.**

I. 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 family $\mathbb{T}$ of PMF's.

By way of motivation, consider a crypto-system in which Alice wishes to send a secret message to Bob. 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] and generalized to systems with specified key rate by Merhav and Arikan in [3].

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? What is Eve's guessing strategy? We answer these questions in this paper.

When P is known, Massey [1] and Arikan [2] sought to lower bound 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 $\mathbb{E}[G(X)]$, 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., $\log[\min_G \mathbb{E}[G(X)]]$, satisfies

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

where $H_\alpha(P)$ is the Rényi entropy of order $\alpha > 0$. Boztaş [4] obtains a tighter upper bound.

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

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

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, $\min_G \mathbb{E}[G(X^n)^\rho]$, 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 $H_{1/2}(P)$. The study of $\mathbb{E}[G(X)^\rho]$, as a function of ρ, is motivated by the fact that it is the convex conjugate (Legendre–Fenchel transformation) of a function that characterizes the large deviations behavior of the number of guesses. See [3] for more details.

Suppose now that the guesser only knows that the source belongs to a family $\mathbb{T}$ 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

Manuscript received March 16, 2006; revised September 8, 2006. This work was supported in part by the Ministry of Human Resources and Development under Grant Part (2A) Tenth Plan (338/ECE), by the DRDO-IISc joint programme on mathematical engineering, and the University Grants Commission under Grant Part (2B) UGC-CAS-(Ph.IV). The material in this paper was presented in part at the IEEE International Symposium on Information Theory (ISIT 2002) held in Lausanne, Switzerland, July 2002, at the National Conference on Communication (NCC 2006), New Delhi, India, Jan. 2006, at the Conference on Information Sciences and Systems (CISS 2006), Princeton, NJ, March 2006, and at the IEEE International Symposium on Information Theory (ISIT 2006)Seattle, WA, July 2006.

The author is with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India.

Communicated by Y. Steinberg, Associate Editor for Shannon Theory.
Digital Object Identifier 10.1109/TIT.2006.887466

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 normalized by ρ.

A natural means of measuring the effectiveness of a guessing strategy G on the family $\mathbb{T}$ 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 $\alpha \to 1$ is the Kullback–Leibler divergence. Furthermore, L_α shares the Pythagorean property with the Kullback–Leibler divergence [5]. The results of this paper thus generalize the "geometric" properties satisfied by the Kullback–Leibler divergence [5].

Consider the special case of guessing an n-string put out 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 [6] proposed a "universal" guessing strategy for the family 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 $n \to \infty$. One can interpret this fact as follows: the family 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. Campbell ([7] and [8]) identified a code that depended on knowledge of the source PMF. The code has redundancy within a constant of the optimal value and is analogous to the Shannon code for source compression. Blumer and McEliece [9] studied a modified Huffman algorithm for this problem and tightened the bounds on the redundancy. Fischer [10] addressed the problem in the context of mismatched source compression and identified the supremum average exponential coding length for a family of sources. In particular, he showed that the supremum value is the supremum of the Rényi entropies of the sources in the family. In contrast to Fischer's work, 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 ([11]–[13]). 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 $(E_i : i = 1, 2, \ldots)$ 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 $H_\alpha(P)$.

As indicated earlier, 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 [14]. Moreover, the best guessing strategy is based on a PMF that is a mixture of sources in the uncertainty set, 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 [12] 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 $1/\alpha$ [15] of an appropriately defined channel.

The following is an outline of the paper. In Section II 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 III, we see how the same quantities arise in the context of Campbell's minimum average exponential coding length problem. In Section IV, we pose the min-sup problem of quantifying the worst-case redundancy and identify another inf-sup problem in terms L_α. In Section V we study the relations between L_α and other known divergence measures. In Section VI we identify the so-called *center* and *radius* of an uncertainty set. In Section VII, we specialize our results to two examples: the family of discrete memoryless sources on finite alphabets, and the family of finite-state arbitrarily varying sources. We establish results on the asymptotic redundancies of these two uncertainty sets. We further refine the redundancy upper bound for the family of binary memoryless sources. In Section VIII we conduct a further study of L_α divergence and show that it satisfies the Pythagorean property. Section IX closes the paper with some concluding remarks.

II. Inaccuracy and Redundancy in Guessing

In this section, we prove previously known results on 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 [15].

Let $\mathbb{X}$ and $\mathbb{Y}$ be 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 $G : \mathbb{X} \times \mathbb{Y} \to \{1, 2, \ldots, |\mathbb{X}|\}$ such that for each $y \in \mathbb{Y}$, the function $G(\cdot, y) : \mathbb{X} \to \{1, 2, \ldots, |\mathbb{X}|\}$ 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 $\mathrm{Supp}(Q) = \mathbb{X} \times \mathbb{Y}$. 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

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

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

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. ■

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

$$H_\alpha(P) = \frac{\alpha}{1-\alpha} \log \left(\sum_{y \in \mathbb{Y}} \left(\sum_{x \in \mathbb{X}} P(x,y)^\alpha \right)^{1/\alpha} \right). \quad (3)$$

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

$$H_\alpha(P) = \frac{1}{1-\alpha} \log \left(\sum_{x \in \mathbb{X}} P(x)^\alpha \right). \quad (4)$$

Note that the left-hand side of (3) is written as a functional of P instead of the more common $H_\alpha(X \mid Y)$. We do not use the latter because the dependence on the PMF needs to be made explicit in many places in the sequel. Also note that both (3) and (4) define $H_\alpha(P)$—(3) for a pair of random variables and (4) for a single random variable. The actual definition being referred to will be clear from the context. It is well known that

$$0 \leq H_\alpha(P) \leq \log |\mathbb{X}|. \quad (5)$$

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

$$\frac{1}{\rho} \log \left(\mathbb{E}\left[G_P(X,Y)^\rho \right] \right) \leq H_\alpha(P) \quad (6)$$

where $\alpha = 1/(1+\rho)$. □

Proof: Set $Q = P$ in Proposition 1. ■

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 $\mathrm{Supp}(Q_G) = \mathbb{X} \times \mathbb{Y}$, and

$$\begin{aligned} &\frac{1}{\rho} \log \left(\mathbb{E}\left[G(X,Y)^\rho \right] \right) \\ &\quad \geq \frac{1}{\rho} \log \left(\sum_{y \in \mathbb{Y}} \sum_{x \in \mathbb{X}} P(x,y) \left[\sum_{a \in \mathbb{X}} \left(\frac{Q_G(a,y)}{Q_G(x,y)} \right)^{\frac{1}{1+\rho}} \right]^\rho \right) \\ &\quad\quad - \log(1 + \ln |\mathbb{X}|) \end{aligned} \quad (7)$$

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

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

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

Define the PMF Q_G as

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

Note that $\mathrm{Supp}(Q_G) = \mathbb{X} \times \mathbb{Y}$. 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

$$\begin{aligned}\sum_{y\in\mathbb{Y}}\sum_{x\in\mathbb{X}} P(x,y)\left[\sum_{a\in\mathbb{X}}\left(\frac{Q_G(a,y)}{Q_G(x,y)}\right)^{\frac{1}{1+\rho}}\right]^{\rho} &= \sum_{y\in\mathbb{Y}}\sum_{x\in\mathbb{X}} P(x,y)G(x,y)^{\rho}\cdot\left(\sum_{a\in\mathbb{X}}\frac{1}{G(a,y)}\right)^{\rho}\\ &\le \left(\sum_{y\in\mathbb{Y}}\sum_{x\in\mathbb{X}} P(x,y)G(x,y)^{\rho}\right)\cdot(1+\ln|\mathbb{X}|)^{\rho} \qquad (8)\end{aligned}$$

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

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

The proposition follows from (8). ■

Observe the similarity of the terms in the right-hand sides of (2) and (7) in Propositions 1 and 3, respectively. The analog of this term in mismatched source compression 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, [16]) states that

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

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. [10, Th. 1.3]).

Proposition 4: (Analog of Shannon Inequality): Let $\alpha = \frac{1}{1+\rho} > 0$, $\alpha \ne 1$. Then

$$\frac{\alpha}{1-\alpha}\log\left(\sum_{y\in\mathbb{Y}}\sum_{x\in\mathbb{X}} P(x,y)\left[\sum_{a\in\mathbb{X}}\left(\frac{Q(a,y)}{Q(x,y)}\right)^{\alpha}\right]^{\frac{1-\alpha}{\alpha}}\right) \ge H_\alpha(P), \qquad (9)$$

with equality if and only if $P = Q$. □

Proof: We will prove this directly using Holder's inequality. The right side of (9) is bounded. Without loss of generality, we may assume that the left side of (9) is finite, for otherwise the inequality trivially holds and $P \ne Q$. We may therefore assume $\mathrm{Supp}(P) \subset \mathrm{Supp}(Q)$ under $0 < \alpha < 1$, and $\mathrm{Supp}(P) \cap \mathrm{Supp}(Q) \ne \emptyset$ under $1 < \alpha < \infty$ which are the conditions when the left side of (9) is finite.

With $\alpha = 1/(1+\rho)$, (9) is equivalent to

$$\mathrm{sign}(\rho)\cdot\sum_{y\in\mathbb{Y}}\sum_{x\in\mathbb{X}} P(x,y)\left[\sum_{a\in\mathbb{X}}\left(\frac{Q(a,y)}{Q(x,y)}\right)^{\frac{1}{1+\rho}}\right]^{\rho} \ge \mathrm{sign}(\rho)\cdot\sum_{y\in\mathbb{Y}}\left(\sum_{x\in\mathbb{X}} P(x,y)^{\frac{1}{1+\rho}}\right)^{1+\rho}.$$

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

$$\mathrm{sign}(\lambda)\cdot\left(\sum_x u_x\right)^{\lambda}\cdot\left(\sum_x v_x\right)^{1-\lambda} \ge \mathrm{sign}(\lambda)\cdot\left(\sum_x u_x^{\lambda}v_x^{1-\lambda}\right) \qquad (10)$$

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 (10), equality holds in (9) if and only if $P = Q$. ■

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

$$\begin{aligned}L_\alpha(P,Q) \triangleq{}& \frac{\alpha}{1-\alpha}\log\left(\sum_{y\in\mathbb{Y}}\sum_{x\in\mathbb{X}} P(x,y)\left[\sum_{a\in\mathbb{X}}\left(\frac{Q(a,y)}{Q(x,y)}\right)^{\alpha}\right]^{\frac{1-\alpha}{\alpha}}\right)\\ &- H_\alpha(P). \qquad (11)\end{aligned}$$

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

Just as Shannon inequality can be employed to show the converse part 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, Th. 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

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

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. ■

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:

$$R(P,G) \triangleq \frac{1}{\rho}\log\left(\mathbb{E}\left[G(X,Y)^{\rho}\right]\right) - \frac{1}{\rho}\log\left(\mathbb{E}\left[G_P(X,Y)^{\rho}\right]\right). \tag{12}$$

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

$$|R(P,G) - L_\alpha(P,Q_G)| \le \log(1+\ln|\mathbb{X}|). \tag{13}$$

□

Proof: The inequality $R(P,G) \le L_\alpha(P,Q_G) + \log(1+\ln|\mathbb{X}|)$ follows from Proposition 1 applied with $Q = Q_G$, the first inequality of Theorem 5, and (11).

The inequality $R(P,G) \ge L_\alpha(P,Q_G) - \log(1+\ln|\mathbb{X}|)$ follows from Proposition 3, the second inequality of Theorem 5, and (11). ■

Remark: It is possible that two different PMFs P and Q lead to the same guessing order, i.e., $G_P = G_Q$. Thus $R(P,G_P) = R(P,G_Q) = 0$. Yet, it is possible that $L_\alpha(P,Q)$ and $L_\alpha(P,Q_{G_Q})$ are nonzero. This remains consistent with Theorem 6 since (13) only provides bounds for $R(P,G_Q)$ on either side to within $\log(1+\ln|\mathbb{X}|)$, and is not an entirely accurate measure of $R(P,G_Q)$. One can only conclude that

$$L_\alpha(P,Q_{G_Q}) \le \log(1+\ln|\mathbb{X}|).$$

This is unlike the case in source compression with mismatch where the "nuisance" term is not $\log(1+\ln|\mathbb{X}|)$ but the constant 1. Yet, in the examples in Section VII on guessing we see how to make good use of these bounds. See also the discussion following Theorem 8 at the end of the next section.

III. Campbell's Coding Theorem and Redundancy

Campbell in [7] and [8] gave another operational meaning to the Rényi entropy of order $\alpha > 0$. In this section, we show that L_α 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

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

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

$$\frac{1}{\rho}\log\left(\mathbb{E}\left[2^{\rho l(X,Y)}\right]\right), \quad -1 < \rho < \infty, \rho \ne 0 \tag{14}$$

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

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 (14). Henceforth, we focus only on length functions that satisfy

$$\frac{1}{2} < \sum_{x\in\mathbb{X}} 2^{-l(x,y)} \le 1, \ \ \forall y \in \mathbb{Y}. \tag{15}$$

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

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

where the minimization is over all those length functions that satisfy (15). □

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

$$l_Q(x,y) \triangleq \left\lceil -\log\left(\frac{Q(x,y)^{\frac{1}{1+\rho}}}{\sum_{a\in\mathbb{X}} Q(a,y)^{\frac{1}{1+\rho}}}\right)\right\rceil \tag{16}$$
$$= \lceil -\log\left(Q'(x \mid y)\right)\rceil \tag{17}$$

where $\lceil\cdot\rceil$ refers to the ceiling function and $Q'(\cdot \mid y)$ is a conditional PMF on $\mathbb{X}$. Clearly, l_Q satisfies (15).

Analogously, for any length function satisfying (15), we can define a PMF on $\mathbb{X}\times\mathbb{Y}$ as follows:

$$Q_l(x,y) = \frac{1}{|\mathbb{Y}|}\frac{2^{-(1+\rho)l(x,y)}}{\sum_{a\in\mathbb{X}} 2^{-(1+\rho)l(a,y)}}. \tag{18}$$

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

Let us define the redundancy for any l satisfying (15) as

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

analogous to the definition without side information in [9]. Following the same sequence of steps as in the mismatched guessing problem, it is straightforward to show the following.

Theorem 8: Let $-1 < \rho < \infty$, $\rho \ne 0$, $\alpha = 1/(1+\rho)$. Consider a source pair (X,Y) with PMF P on $\mathbb{X}$. Let l be a length function that denotes an encoding of X with side information Y, and Q_l the associated PMF given by (18). Then

$$|R_c(P,l) - L_\alpha(P,Q_l)| \le 1. \tag{19}$$

□

We interpret $L_\alpha(P, Q_l)$ as the penalty for mismatched coding when Q_l is not matched to P. $L_\alpha(P, Q_l)$ is indicative of the redundancy to within a constant, as the Kullback–Leibler divergence is in mismatched source compression. By comparing (19) with (13), we see that the nuisance term in this problem is a constant that does not depend on the size of the source alphabet; $L_\alpha(P, Q_l)$ is therefore a more faithful representation of $R_c(P, l)$ than $L_\alpha(P, Q_G)$ is of $R(P, G)$.

IV. 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 $(x, y) \in \mathbb{X} \times \mathbb{Y}$, the mapping $P \mapsto P(x, y)$ 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 (12). 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

$$R^* = \min_G \sup_{P \in \mathbb{T}} R(P, G). \tag{20}$$

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 $0 < \alpha < \infty$, $\alpha \neq 1$,

$$C \triangleq \min_Q \sup_{P \in \mathbb{T}} L_\alpha(P, Q). \tag{21}$$

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 VI-C.

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

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

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 24.

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

Theorem 11: (Guessing Under Uncertainty): Let $\mathbb{T}$ be a set 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 (21), i.e.

$$C = \sup_{P \in \mathbb{T}} L_\alpha(P, Q^*). \tag{22}$$

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

$$R(P, G^*) \leq L_\alpha(P, Q^*) + \log(1 + \ln |\mathbb{X}|) \tag{23}$$

follows from Proposition 1 applied with $Q = Q^*$, the first inequality of Theorem 5, and (11), as in the proof of Theorem 6. After taking supremum over all $P \in \mathbb{T}$, and after substitution of (22), 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. ■

Remarks:

1) Thus one approach to obtain the minimum in (20) is to identify minimum value in (21). This minimum value will be within $\log(1+\ln |\mathbb{X}|)$ of R^* in (20). 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 set 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 set will be rich enough for the converse to be meaningful if C grows with n at a faster rate.

V. Relations Between L_α and Other Divergence Quantities

Having shown how $L_\alpha(P,Q)$ arises as a penalty function for mismatched guessing and coding, we now study it in greater detail and relate it to other divergence quantities. The relationships we discover here will be useful in the sequel. 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 $\mathrm{Supp}(P) \cap \mathrm{Supp}(Q) = \emptyset$, or $\alpha < 1$ and $\mathrm{Supp}(P) \not\subset \mathrm{Supp}(Q)$.
3) Given the joint PMF P, let us define the "tilted" conditional PMF on $\mathbb{X}$ as in (24) shown at the bottom of the page. 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 (24) are PMFs in one-to-one correspondence with P and Q respectively. Using the expression for Rényi entropy and (11), we have that

$$\begin{aligned} L_\alpha(P,Q) &= \frac{1}{\rho} \log \left(\sum_{x \in \mathbb{X}} P'(x)^{1+\rho} \cdot Q'(x)^{-\rho} \right) \\ &= D_{1/\alpha}(P' \parallel Q') \end{aligned} \tag{25}$$

where $D_\beta(R \parallel S)$ is the Rényi 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:

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

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 $f : \mathbb{X} \times \mathbb{Y} \to \mathbb{X}' \times \mathbb{Y}'$ 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 (24) we can write $L_\alpha(P,Q)$ as follows:

$$L_\alpha(P,Q) = \frac{1}{\rho} \log \left[\mathrm{sign}(\rho) \cdot I_f(P' \parallel Q') \right] \tag{26}$$

where $I_f(R \parallel S)$ is the f-divergence [17] given by

$$I_f(R \parallel S) = \sum_{x \in \mathbb{X}} S(x) f\left(\frac{R(x)}{S(x)} \right) \tag{27}$$

with

$$f(x) = \mathrm{sign}(\rho) \cdot x^{1+\rho}, \quad x \geq 0. \tag{28}$$

Since f is a strictly convex function for $\rho \neq 0$, an application of Jensen's inequality in (27) indicates that

$$I_f(R \parallel S) \geq f(1) = \begin{cases} -1, & -1 < \rho < 0 \\ 1, & 0 < \rho < \infty. \end{cases} \tag{29}$$

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

$$-1 \leq I_f(R \parallel S) \leq 0. \tag{30}$$

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

$$H_\alpha(P) = \frac{\alpha}{1-\alpha} \log h(P). \tag{31}$$

Motivated by the relationship in (26), let us write L_α in the general case as follows:

$$L_\alpha(P,Q) = \frac{1}{\rho} \log \left[\mathrm{sign}(\rho) \cdot I(P,Q) \right] \tag{32}$$

$$P'(x \mid y) \triangleq \begin{cases} P(x,y)^\alpha \Big/ \sum_{a \in \mathbb{X}} P(a,y)^\alpha, & \text{if } \sum_{a \in \mathbb{X}} P(a,y)^\alpha > 0 \\ 1/|\mathbb{X}|, & \text{otherwise.} \end{cases} \tag{24}$$

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

$$I(P,Q) \triangleq \frac{\mathrm{sign}(\rho)}{h(P)} \sum_{y\in\mathbb{Y}}\sum_{x\in\mathbb{X}} P(x,y)\,(Q'(x\mid y))^{-\rho} \tag{33}$$

$$= \frac{\mathrm{sign}(1-\alpha)}{h(P)} \sum_{y\in\mathbb{Y}}\sum_{x\in\mathbb{X}} P(x,y)\,(Q'(x\mid y))^{\frac{\alpha-1}{\alpha}}. \tag{34}$$

These expressions turn out to be useful in the sequel. It is not difficult to show that

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

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 (29) and (30) 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 $t = \mathrm{sign}(\rho)\cdot 2^{\rho r}$.

11) From the known bounds $0 \leq H_\alpha(P) \leq \log|\mathbb{X}|$, it is easy to see the following bounds:

$$1 \leq h(P) \leq |\mathbb{X}|^{\frac{1-\alpha}{\alpha}}, \text{ for } 0<\alpha<1 \tag{35}$$

and

$$|\mathbb{X}|^{\frac{1-\alpha}{\alpha}} \leq h(P) \leq 1, \text{ for } 1<\alpha<\infty. \tag{36}$$

In both cases, we see that $h(P)$ is bounded away from 0 and therefore (33) and (34) 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 $|\mathbb{Y}| = 1$. See, for example, comments 6 and 7 made earlier in this section. However, it behaves like squared distance and shares a "Pythagorean" property with the Kullback–Leibler divergence. This is explored in Section VIII.

VI. L_α-Center and Radius of a Family

In this section we identify the L_α-center and radius of a family. We first begin with a finite family and subsequently study an arbitrary family (that satisfies some measurability conditions). We finally conclude the section with a proof of Theorem 10.

A. L_α-Center and Radius of a Finite Family

Let $|\mathbb{T}|$ be finite. For simplicity, assume that no side information is available. 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 is 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.

Proof of Theorem 10 for a Finite Family of PMFs: Let $\mathbb{T} = \{P_1,\ldots,P_m\}$ 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 (24). Moreover, from (25) and (26), we have

$$\inf_Q \max_{1\leq i\leq m} L_\alpha(P_i,Q) \tag{37}$$

$$= \inf_Q \max_{1\leq i\leq m} D_{1+\rho}(P'_i\,\|\,Q') \tag{38}$$

$$= \frac{1}{\rho}\log\left(\mathrm{sign}(\rho)\inf_Q \max_{1\leq i\leq m} I_f(P'_i\,\|\,Q')\right) \tag{39}$$

Csiszár considered the evaluation of (38) in [15, Prop. 1], and the evaluation of the inf-max within parenthesis in (39) in [17].

From [17, Th. 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 (37)–(39) can all be replaced by minima. From the inverse of the map $Q\mapsto Q'$, we obtain the unique minimizer Q^* for (37). This proves the existence and uniqueness result of Theorem 10 when $|\mathbb{T}|$ is finite.

1) 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 [17, eq. 2.25], [17, Th. 3.2], and its Corollary. Indeed, [17, Th. 3.2] assures that

$$\min_{Q'}\max_{1\leq i\leq m} I_f(P'_i\,\|\,Q') \tag{40}$$

$$= \max_\mu \min_{Q'} \sum_{i=1}^m \mu(i) I_f(P'_i\,\|\,Q') \tag{41}$$

where the max-min in (41) is achieved at (μ^*, Q'^*), and Q'^* is the PMF which attains the min-max in (40). We now seek to find out the nature of Q'^* and thence Q^*.

For any arbitrary weight function μ, we have from [17, eq. 2.25] that the Q' which minimizes

$$\sum_{i=1}^m \mu(i) I_f(P'_i\,\|\,Q') \tag{42}$$

is

$$Q'(x) = c^{-1}\cdot\left(\sum_{i=1}^m \mu(i)(P'_i(x))^{1/\alpha}\right)^\alpha \tag{43}$$

$$= c^{-1}\left(\sum_{i=1}^m \frac{\mu(i)}{h(P_i)} P_i(x)\right)^{1/\alpha} \tag{44}$$

for every $x \in \mathbb{X}$, where c is the normalizing constant. From the correspondence between the primed and the unprimed PMFs, and (44), we obtain

$$Q(x) = d^{-1} \sum_{i=1}^{m} \frac{\mu(i)}{h(P_i)} P_i(x), \ \ \forall x \in \mathbb{X} \tag{45}$$

where d is the normalizing constant

$$d = \sum_{i=1}^{m} \frac{\mu(i)}{h(P_i)}. \tag{46}$$

Thus, for an arbitrary μ, the Q (obtained from Q') that minimizes (42) is in the convex hull $\mathcal{E}$. In particular, the minimizing Q^* corresponding to the μ^* that attains the max-min objective in (41), and therefore the min-max objective in (40), is also in $\mathcal{E}$. This result will be proved in wider generality in Section VIII.

With some algebra, we can further show that

$$C = \min_{Q} \max_{1 \leq i \leq m} L_\alpha(P_i, Q) = \frac{\alpha}{1-\alpha} \log(d \cdot h(Q^*)) \tag{47}$$

where Q^* is given by (45) and d by (46) with $\mu = \mu^*$.

2) Necessary and Sufficient Conditions for Finding the L_α-Center and Radius: From [17, Th. 3.2], a weight vector μ maximizes (41) if and only if

$$I_f(P_i' \parallel Q') \leq K, \ \ i = 1, 2, \ldots, m \tag{48}$$

where equality holds whenever $\mu(i) > 0$, and Q' is given by (43). Under this condition, clearly, the corresponding Q given by (45) is the L_α-center and $C = (1/\rho) \log(\mathrm{sign}(\rho) \cdot K)$ is the L_α-radius.

An interesting special case occurs when $h(P_i)$ is independent of i. Then we may simplify (45) to

$$Q = \sum_{i=1}^{m} \mu(i)\, P_i \tag{49}$$

i.e., the weights that make the optimum mixture (of PMFs) are the same as the given weights that form the objective function in (41).

3) 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 $\{1, 2, \ldots, m\}$ and output alphabet $\mathbb{X}$. This channel will be represented as P'.

From the remarks in [15] on the connection between information radius of order $1/\alpha$ and the Gallager exponent of the channel P', and from [15, Prop. 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 [18, p. 138], $1 < \alpha < \infty$ in [18, p. 157], and $0 < \alpha < 1$ in [19]).

4) 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 (41), but only as a means to study the min-max problem. We now make some remarks about the max-min problem for the finite family case. Its extension to arbitrary uncertainty sets is not considered in this paper.

Suppose that our new objective is to find

$$\max_{\mu} \min_{Q} \sum_{i=1}^{m} \mu(i) L_\alpha(P_i, Q). \tag{50}$$

This problem is the same as identifying the "capacity of order $1/\alpha$ " of the channel P' [15], i.e.

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

[15, Prop. 1] solves this problem; the value is the same as the min-max value

$$\min_{Q'} \max_{1 \leq i \leq m} D_{1/\alpha}(P_i' \parallel Q').$$

Consequently, the max-min value of (50) is the same as the L_α-radius of the family.

B. 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 [12] for source compression. We first recall the technical condition indicated in Section IV. $\mathbb{T}$ is a family of PMFs on $\mathbb{X} \times \mathbb{Y}$, $(\mathbb{T}, \mathcal{T})$ a measurable space, and for every $(x, y) \in \mathbb{X} \times \mathbb{Y}$, the mapping $P \mapsto P(x, y)$ is $\mathcal{T}$-measurable.

Our focus will be on the following.

Definition 12: For $0 < \alpha < \infty$, $\alpha \neq 1$

$$K_+ \triangleq \min_{Q} \sup_{P \in \mathbb{T}} I(P, Q). \tag{51}$$

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. Define the mapping $f : \mathbb{T} \to \mathbb{R}_+^{|\mathbb{X}||\mathbb{Y}|}$ as follows:

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

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

$$F \triangleq \int_{\mathbb{T}} d\mu(P) f(P). \tag{52}$$

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

$$\mu f \triangleq d^{-1} F \tag{53}$$

where d as in the finite case is the normalizing constant

$$d \triangleq \int_{\mathbb{T}} \frac{d\mu(P)}{h(P)} = \sum_{x\in\mathbb{X}} F(x). \tag{54}$$

These definitions are extensions of (45) and (46) to arbitrary $\mathbb{T}$. Moreover, let

$$J(\mu, \mathbb{T}) \triangleq \int_{\mathbb{T}} d\mu(P)\, I(P, \mu f). \tag{55}$$

Simple algebraic manipulations result in

$$\begin{aligned} J(\mu, \mathbb{T}) &= \mathrm{sign}(\rho) \cdot h(F) && (56)\\ &= \mathrm{sign}(\rho) \cdot d \cdot h(\mu f) && (57)\end{aligned}$$

an extension of [17, eq. (2.24)] for arbitrary $\mathbb{T}$.

The following auxiliary quantity will be useful.

Definition 13: For $0 < \alpha < \infty, \alpha \neq 1$

$$K_- \triangleq \sup_{\mu} J(\mu, \mathbb{T}). \tag{58}$$

The quantity μf in (53) 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 (55) is the analogue of mutual information; Csiszár calls it informativity in his work on finite-sized families [17].

Proposition 14: $K_- \leq K_+$.

Proof: Fix an arbitrary PMF Q on $\mathbb{X} \times \mathbb{Y}$. It is straightforward to show that [17, eq. 2.26] holds even when $|\mathbb{T}|$ is not finite, and is given by

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

Since $I(\mu f, Q) \geq \mathrm{sign}(\rho)$, 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 [12, Th. A]. The proof largely runs along similar lines.

Proposition 15: 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 (51), and $K_- = K_+$. □

Proof:

$\Leftarrow$: Observe that on account of 1), 3), and Proposition 14, 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 14. 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 vectors in $\mathbb{R}^{|\mathbb{X}||\mathbb{Y}|}$ 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 (54). 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

$$\nu_{n,\lambda} = (1 - \lambda)\mu_n + (\lambda)\delta_{P_0} \tag{59}$$

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

From (59), (52), and (56), we have

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

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

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

for $\lambda \in [0, 1]$. In particular, $s(0) = \lim_n s_n(0) = K_-$. $s(\lambda)$ is a concave function of λ since $\mathrm{sign}(\rho)h(\cdot)$ is concave and the argument is linear in λ. Let $\dot{s}(0)$ be the one-sided derivative of $s(\lambda)$ evaluated at $\lambda = 0$ (i.e., limit as $\lambda \downarrow 0$). 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

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

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

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

a contradiction. The strict inequality above is due to strict concavity of $\mathrm{sign}(\rho)h(\cdot)$ when $\rho > -1$ and $\rho \neq 0$. ■

C. *Proof of Theorem 10*

Proof: From (32), it is clear that

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

Q attains the min-sup value K_+ in Definition 12 if and only if Q attains the min-sup value C in Definition 9. Proposition 15 guarantees the existence and uniqueness of such a Q. ■

VII. Examples

In this section, we look at two example families of PMFs, and identify their L_α-centers and radii. We focus on guessing without side information. We also take a closer look at the binary memoryless channel and obtain tighter upper bounds on redundancy than those obtained via Theorem 11. Throughout this section, therefore, $0 < \alpha < 1$ and $|\mathbb{Y}| = 1$. The uncertainty set will thus be PMFs in $\mathbb{X}$ (with no reference to $|\mathbb{Y}|$).

A. *The Family 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 set of all discrete memoryless sources (DMS) on $\mathbb{A}$, i.e.

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

The parameters of the source P are unknown to the guesser. Arikan and Merhav [6] provide a guessing scheme for this uncertainty set. The scheme happens to be independent of ρ. Moreover, their guessing scheme has the same asymptotic performance as the optimal guessing scheme. Their guessing order proceeds 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 inconsequential. Denote this guessing list by G_n. Arikan and Merhav [6, Th. 1] showed that for any $P_n \in \mathbb{T}$,

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

The above result is couched in our notation. This indicates that $\mathbb{T}$, the family 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 16 (Family of DMSs on $\mathbb{A}$): Let $m = |\mathbb{A}|$. The L_α-radius C_n of the family of discrete memoryless sources on $\mathbb{A}$ satisfies

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

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

Proof: Recall that $\rho > 0$. P_n is the joint PMF of the n-string with individual letter probabilities P. Let $P_n \mapsto P'_n$ according to the mapping given in (24). It is easy to verify that P'_n is the joint PMF of the n-string with individual letter probabilities P', where $P \mapsto P'$ according to the mapping (24), and therefore P'_n also belongs to $\mathbb{T}$. Furthermore, for a fixed $a^n \in \mathbb{A}^n$, let $\hat{S}_{a^n}$ be the PMF of letter frequencies in a^n, and define

$$\hat{S}_{a^n,n}(x^n) \triangleq \prod_{i=1}^n \hat{S}_{a^n}(x_i)$$

for every $x^n \in \mathbb{A}^n$. Note that $\hat{S}_{a^n,n}$ is not necessarily a PMF. Xie and Barron [20, Th. 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:

$$\begin{aligned} \max_{a^n\in\mathbb{A}^n} \log\frac{P'_n(a^n)}{Q'_n(a^n)} &\le \max_{a^n\in\mathbb{A}^n} \log\frac{\hat{S}_{a^n,n}(a^n)}{Q'_n(a^n)} \qquad (60) \\ &\le r_n \\ &\triangleq \frac{m-1}{2}\log\frac{n}{2\pi} + u_m + \varepsilon_n. \qquad (61) \end{aligned}$$

Define the PMF Q_n as follows:

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

the inverse of the mapping in (24). We then have the following series of inequalities:

$$\begin{aligned} &L_\alpha(P_n, Q_n) \\ &= \frac{1}{\rho}\log\left(\sum_{a^n\in\mathbb{A}^n} P'_n(a^n)\left(\frac{P'_n(a^n)}{Q'_n(a^n)}\right)^\rho\right) \qquad (62) \end{aligned}$$

$$\leq \frac{1}{\rho}\left(\log \sum_{a^n \in \mathbb{A}^n} P'_n(a^n)\cdot \exp\{\rho r_n\}\right)$$
$$= \frac{1}{\rho}\log\left(\exp\{\rho r_n\}\right)$$
$$= r_n \tag{63}$$

where (62) follows from (25) and (63) from (61). 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 16 suggests the use of Q_n, which in general may depend on ρ. Arikan 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 [21] and [22].

Theorem 17 : 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

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

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

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

The theorem then follows from [22, Th. 2] which states that the redundancy in source compression $D(P'_n \parallel Q'_n)$ 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 $2^{-n\lambda_n}$. ■

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 [20, Remark 5 after Theorem 2]. We may then choose λ_n such that $n\lambda_n \to \infty$ so that $2^{-n\lambda_n}$ vanishes with n, but λ_n is negligibly small compared to r_n. (For example, for the family of DMSs, $r_n = O(\log n)$ and therefore we may set $\lambda_n = (\log\log n)/(\log n)$). 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. Equivalently, with high μ-probability (at least $1 - 2^{-n\lambda_n}$), $L_\alpha(P_n, Q_n) > r_n - \lambda_n$.

Since L_α quantifies the redundancy in Campbell's coding problem to within unity, the above remark leads us to conclude that the redundancy in that problem is tightly bounded as $\frac{m-1}{2}(\log n)$ (up to a constant).

In the guessing context, since the nuisance term $\log(1 + n\ln m)$ grows as $\log n + \log\ln m$ for large n, we deduce that with high μ-probability (at least $1 - 2^{-n\lambda_n}$), the guessing redundancy of any strategy is at least $r_n - \lambda_n - \log(1+n\ln m)$, which for large n is

$$\frac{m-3}{2}\log n + u_m + \frac{m-1}{2}\log(2\pi) - \log\ln m + \varepsilon_n - \lambda_n. \tag{64}$$

This fact and Theorem 16 immediately lead us to conclude that for $m \geq 4$, the redundancy is between $\frac{m-3}{2}\log n$ and $\frac{m+1}{2}\log n$ for large n (ignoring constants and smaller order terms). For $m = 2$ and $m = 3$, the lower bound in (64) is useless, and the upper bound $\frac{m+1}{2}\log n$ may not be tight. The case of $m = 2$ is addressed in the next subsection. Tighter upper bounds for $m = 3$ remain to be found.

B. Guessing an Unknown Binary Memoryless Source

The L_α-based bounding technique suggested by Theorem 11 provides good bounds on guessing redundancy for large n when the DMSs alphabet size $m \geq 4$. In this subsection, we identify tighter upper bounds on the guessing redundancy of a binary memoryless source using a more direct approach.

Let $\mathbb{A} = \{0, 1\}$. There is only one unknown parameter, i.e., $p = P(1)$. The probability of any n-string is given by

$$P_n(x^n) = p^{N(x^n)}(1-p)^{n-N(x^n)} = (1-p)^n\left(\frac{p}{1-p}\right)^{N(x^n)}$$

where $N(x^n)$ is the number of 1s in the string x^n. Since $P_n(x^n)$ is monotonic in $N(x^n)$, it immediately follows that when $p > 1/2$, the optimal guessing order is to guess the string of all 1's, followed by all strings with exactly one 0, followed by all strings with exactly two 0's, and so on, *viz.*, in the decreasing order of number of 1's in the string. Note that the optimal guessing sequence is the same for all sources whose $p > 1/2$. Exactly the opposite is true when $p < 1/2$—the guessing proceeds in the increasing order of number of 1's, the first guess being the string of all 0's.

Thus there are only two optimal guessing lists for the binary memoryless source. By guessing one element from each list, skipping those already guessed, we obtain a guessing list that requires at most twice the optimal number of guesses, i.e., $G(x^n) \leq 2G_{P_n}(x^n)$ for every $x^n \in \mathbb{A}^n$. This guessing list is one of those that proceed in the increasing order of empirical entropy. Clearly then, the redundancy is upper bounded by the constant $\log 2$, a bound tighter than Theorem 16. C_n/n therefore vanishes as $(\log 2)/n$. It is not known if this is the tightest upper bound.

C. Arbitrarily Varying Sources

For the family of DMSs, we saw in Section VII-A 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 $X_1, X_2, \ldots$, 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 [23]. 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 $V(x \mid s)$ 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 18: 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 $\mathbb{T} = \{P_n(\cdot \mid s^n) \mid s^n \in T_U\}$. The L_α-radius of this family is given by

$$R_n(T_U) \triangleq H_\alpha(Q_n^*) - \frac{1}{|T_U|} \sum_{s^n \in T_U} H_\alpha(P_n(\cdot \mid s^n)) \tag{65}$$

where the L_α-center Q_n^* is given by

$$Q_n^*(\cdot) = \frac{1}{|T_U|} \sum_{s^n \in T_U} P_n(\cdot \mid s^n). \tag{66}$$

□

Remarks:

1) It will be apparent from the proof that the quantity $H_\alpha(P_n(\cdot \mid s^n))$ in (65) 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) $\operatorname{sign}(\rho) \cdot h(P)$ is a concave function of P. It follows from (31) that $H_\alpha(P)$ is also a concave function of P for $0 < \alpha < 1$. By Jensen's inequality, $R_n(T_U) \ge 0$. (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 20 that if the U sequence (parameterized by n) converges as $n \to \infty$ 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 (48) given in Section VI-A2. From (33), it is sufficient to show that

$$\begin{aligned} I_f(P_n'(\cdot \mid s^n \parallel Q_n^{*'}) &= \frac{\sum_{x^n \in \mathbb{A}^n} P_n(x^n \mid s^n) \left(Q_n^{*'}(x^n)\right)^{-\rho}}{h(P_n(\cdot \mid s^n))} \\ &= K \end{aligned} \tag{67}$$

where K is some constant that depends only on n and T_U. We will show that the numerator and denominator in (67) 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 $P(x \mid s)$ for $x \in \mathbb{A}$ and $s \in \mathbb{S}$. Consider $h(P_n(\cdot \mid s^n))$. First

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

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 $h(P_n(\cdot \mid s^n))$ 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 $((x_{\pi(1)}, \ldots, x_{\pi(n)}), (s_{\pi(1)}, \ldots, s_{\pi(n)}))$, where s^n and $(s_{\pi(1)}, \ldots, s_{\pi(n)})$ are the two given $\mathbb{A}$-sequences. This permutation π leaves $P_n(x^n \mid s^n)$ unchanged. Moreover, the sum continues to be over

$$T_U = \left\{(s_{\pi(1)}, s_{\pi(2)}, \ldots, s_{\pi(n)}) \in \mathbb{S}^n \; s^n = (s_1, \ldots, s_n) \in T_U\right\}.$$

Thus $Q_n^*(x^n)$ and therefore $Q_n^{*'}(x^n)$ 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 the numerator of (67)

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

depends on s^n only through T_U.

That $R_n(T_U)$ is given by (65) follows from (45)–(47), the fact that $h(P_n(\cdot \mid s^n))$ is a constant over all $s^n \in T_U$, and (31). This concludes the proof. ■

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 19: Let $0 < \alpha < 1$. Let the uncertainty set $\mathbb{T}$ be given by $\mathbb{T} = \{P_n(\cdot \mid s^n) \mid s^n \in \mathbb{S}^n\}$. 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 $s^n \in T_U$. □

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 kth 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 18. 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 $G_1, G_2, \ldots, G_N$ 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 $[G_1\ G_2\ \cdots\ G_N]$ 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 ith string already figures in the final guessing list. Finally remove all δ's to obtain the desired guessing list $G : \mathbb{A}^n \to \{1, 2, \ldots, |\mathbb{A}|^n\}$, 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 $1 \le k \le N$, we have $G(x^n) \le NG_k(x^n)$. 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 $G(x^n) \le NG_k(x^n)$.

The above observation leads to

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

The proposition follows from Theorem 6, Proposition 18, and the bounding $N \le (n+1)^{|\mathbb{S}|}$. ■

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

Proposition 20: 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 $n \to \infty$. Then

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

where $R \ge 0$. □

Proof: The second term in the right-hand side of (65), after normalization by n, converges to a nonnegative real number as seen below:

$$\begin{aligned}
&\frac{1}{n}H_\alpha(P_n(\cdot \mid s^n)) \\
&\quad= \frac{1}{n(1-\alpha)}\log \sum_{x^n \in \mathbb{A}^n}\prod_{i=1}^{n} P(x_i \mid s_i)^\alpha \\
&\quad= \frac{1}{n(1-\alpha)}\log \prod_{s\in\mathbb{S}}\left(\sum_{x\in\mathbb{A}} P(x \mid s)^\alpha\right)^{nU(s)} \\
&\quad= \sum_{s\in\mathbb{S}} U(s) H_\alpha(P(\cdot \mid s)) \\
&\quad\to \sum_{s\in\mathbb{S}} U^*(s) H_\alpha(P(\cdot \mid s)).
\end{aligned} \tag{68}$$

We next consider the first term on the right-hand side of (65) after normalization, i.e., $H_\alpha(Q_n^*)/n$, where Q_n^* is given by (66).

Lemma 21: 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 $n \to \infty$. 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 $n \to \infty$. Then $\lim_n \frac{1}{n}H_\alpha(Q_n^*) = H_\alpha(V^*)$. □

As a consequence of this lemma and (68), we have

$$\frac{1}{n}R_n(T_U) \to H_\alpha(V^*) - \sum_{s\in\mathbb{S}} U^*(s) H_\alpha(P(\cdot \mid s)) \triangleq R.$$

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

Remarks: $R = 0$ if and only if either (*i*) $U(s) = 1$ for some $s \in \mathbb{S}$, or (*ii*) $P(\cdot \mid s)$ 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 linearly 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 21.

Proof:

a) We first show that $\lim_n \frac{1}{n}H_\alpha(Q_n^*) \le H_\alpha(V^*)$. 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 (66), we may write

$$\begin{aligned}
&\sum_{x^n\in\mathbb{A}^n} Q_n^*(x^n)^\alpha \\
&\quad= \sum_{x^n\in\mathbb{A}^n}\left(\frac{1}{|T_U|}\sum_{s^n\in T_U} P_n(x^n \mid s^n)\right)^\alpha \\
&\quad= \sum_{x^n\in\mathbb{A}^n}\left(\frac{1}{U_n\{T_U\}}\frac{U_n\{T_U\}}{|T_U|}\sum_{s^n\in T_U} P_n(x^n \mid s^n)\right)^\alpha \\
&\quad= \frac{1}{U_n\{T_U\}^\alpha}\sum_{x^n\in\mathbb{A}^n}\left(\sum_{s^n\in T_U} U_n(s^n)P_n(x^n \mid s^n)\right)^\alpha &(69)\\
&\quad\le (n+1)^{|\mathbb{S}|\alpha}\sum_{x^n\in\mathbb{A}^n}\left(\sum_{s^n\in \mathbb{S}^n} U_n(s^n)P_n(x^n \mid s^n)\right)^\alpha &(70)\\
&\quad= (n+1)^{|\mathbb{S}|\alpha}\sum_{x^n\in\mathbb{A}^n} V_n(x^n)^\alpha \\
&\quad= (n+1)^{|\mathbb{S}|\alpha}\left(\sum_{x\in\mathbb{A}} V(x)^\alpha\right)^n, &(71)
\end{aligned}$$

where (69) follows from the observation that $U_n(s^n) = U_n\{T_U\}/|T_U|$ for all $s^n \in T_U$, (70) from $U_n\{T_U\} \ge (n+1)^{-|\mathbb{S}|}$ (see proof of [23, Lemma 2.3]) and by enlarging the sum over T_U to all of $\mathbb{S}^n$.

From (71) and (31), we have

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

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

Continuing from (69), we may write

$$\begin{aligned}\sum_{x^n \in \mathbb{A}^n} & Q_n^*(x^n)^\alpha \\ &= \frac{1}{U_n\{T_U\}^\alpha} \sum_{x^n \in \mathbb{A}^n} \left(\sum_{s^n \in T_U} U_n(s^n) P_n(x^n \mid s^n) \right)^\alpha \\ &\geq \sum_{x^n \in T_{\overline{Q}}} \left(\sum_{s^n \in T_U} U_n(s^n) P_n(x^n \mid s^n) \right)^\alpha \qquad (72) \\ &\geq \sum_{x^n \in T_{\overline{Q}}} \left(\sum_{s^n \in T_{\overline{W}}(x^n) \subset T_U} V_n(x^n) W_n(s^n \mid x^n) \right)^\alpha \qquad (73) \\ &= \sum_{x^n \in T_{\overline{Q}}} \left(V_n(x^n) W_n \left\{ T_{\overline{W}}(x^n) \mid x^n \right\} \right)^\alpha \qquad (74)\end{aligned}$$

where (72) follows because $U_n\{T_U\}^\alpha \leq 1$ and the sum over $\mathbb{A}^n$ is restricted to a sum over a type $T_{\overline{Q}}$ to be chosen later; (73) follows because $U_n(s^n) P_n(x^n \mid s^n) = V_n(x^n) W_n(s^n \mid x^n)$ and the sum over s^n is now restricted over a nonvoid $\overline{W}$-shell of x^n, where $\overline{W}$ will be appropriately chosen later.

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

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

Substitution of these inequalities into (74) yields

$$\sum_{x^n \in \mathbb{A}^n} Q_n^*(x^n)^\alpha \geq (n+1)^{-|\mathbb{X}|(1+\alpha|\mathbb{S}|)} \cdot 2^{n\left[(1-\alpha)H(\overline{Q}) - \alpha\left(D(\overline{Q}\|V) + D\left(\overline{W}\|W|\overline{Q}\right)\right)\right]}$$

and therefore

$$\begin{aligned}\frac{1}{n} & H_\alpha(Q_n^*) \\ &\geq H(\overline{Q}) - \frac{1}{\rho}\left[D\left(\overline{Q} \parallel V\right) + D\left(\overline{W} \parallel W \mid \overline{Q}\right)\right] \\ &\quad - \frac{|\mathbb{X}|(1+\alpha|\mathbb{S}|)}{1-\alpha} \frac{\log(n+1)}{n} \qquad (75)\end{aligned}$$

for any type $\overline{Q}$ of sequences and for any $\overline{W}$ such that $T_{\overline{W}}(x^n) \subset T_U$ is a nonvoid shell for an $x^n \in T_{\overline{Q}}$.

Clearly, the last term in (75) vanishes as $n \to \infty$.

If we can choose $\overline{Q} = V'$ and $\overline{W} = W$, 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 $n \to \infty$, we can get close enough. The following arguments make this idea precise.

Define

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

and consider $D(\overline{W}(\cdot \mid x) \parallel W(\cdot \mid x))$. We may restrict our choice of $\overline{W}$ 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 $\overline{W}$ that in addition satisfies

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

and $\varepsilon_n \to 0$.

We then have

$$\begin{aligned} D(\overline{W}&(\cdot \mid x) \parallel W(\cdot \mid x)) \\ &= H(W(\cdot \mid x)) - H(\overline{W}(\cdot \mid x)) \\ &\quad + \sum_{s \in \mathbb{S}} \left(W(s \mid x) - \overline{W}(s \mid x) \right) \log W(s \mid x) \\ &\leq \left| H(W(\cdot \mid x)) - H(\overline{W}(\cdot \mid x)) \right| \\ &\quad - (\log \delta) \sum_{s \in \mathbb{S}} \left| W(s \mid x) - \overline{W}(s \mid x) \right| \\ &\leq -\varepsilon_n \log \frac{\varepsilon_n}{|\mathbb{S}|} - \varepsilon_n \log \delta \qquad (76)\end{aligned}$$

where (76) follows from [23, Lemma 2.7]. After averaging, we get

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

A similar argument shows that

$$\begin{aligned} H(\overline{Q}) - \frac{1}{\rho} D\left(\overline{Q} \parallel V\right) &= H_\alpha(V) + \left[H(\overline{Q}) - H(V') \right] \\ &\quad - \frac{1}{\rho}\left[D\left(\overline{Q} \parallel V\right) - D\left(V' \parallel V\right) \right] \\ &\to H_\alpha(V^*) \end{aligned}$$

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 21 ■

VIII. L_α-Projection

In this section we look at an interesting geometric property of L_α divergence that makes it behave like squared Euclidean distance, analogous to the Kullback–Leibler divergence. Throughout this section, we assume $\alpha > -1$ and $\alpha \neq 0$.

We proceed along the lines of [5]. 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

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

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 22: $B(R,r)$ is a convex set. □

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 $t = \mathrm{sign}(\rho)\cdot 2^{\rho r}$, we get from (32) that

$$I(P_i,R) < t, \quad i=0,1. \tag{77}$$

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

$$\begin{aligned}
&I(P_\lambda,R)\\
&= \frac{\mathrm{sign}(1-\alpha)}{h(P_\lambda)}\cdot\sum_{y\in\mathbb{Y}}\sum_{x\in\mathbb{X}} P_\lambda(x,y)\,(R'(x\mid y))^{\frac{\alpha-1}{\alpha}}\\
&= \frac{\mathrm{sign}(1-\alpha)}{h(P_\lambda)}\cdot(1-\lambda)\sum_{y\in\mathbb{Y}}\sum_{x\in\mathbb{X}} P_0(x,y)\,(R'(x\mid y))^{\frac{\alpha-1}{\alpha}}\\
&\quad + \frac{\mathrm{sign}(1-\alpha)}{h(P_\lambda)}\cdot(\lambda)\sum_{y\in\mathbb{Y}}\sum_{x\in\mathbb{X}} P_1(x,y)\,(R'(x\mid y))^{\frac{\alpha-1}{\alpha}}\\
&= \frac{(1-\lambda)h(P_0)I(P_0,R)+\lambda h(P_1)I(P_1,R)}{h(P_\lambda)} &(78)\\
&< t\frac{(1-\lambda)h(P_0)+\lambda h(P_1)}{h(P_\lambda)} &(79)\\
&= |t|\frac{(1-\lambda)\cdot\mathrm{sign}(1-\alpha)h(P_0)+\lambda\cdot\mathrm{sign}(1-\alpha)h(P_1)}{h(P_\lambda)}\\
&\le |t|\frac{\mathrm{sign}(1-\alpha)h(P_\lambda)}{h(P_\lambda)}\\
&= t &(80)
\end{aligned}$$

where (78) follows from (34), (79) from (77), and (80) from the concavity of $\mathrm{sign}(1-\alpha)h$. ■

Proposition 22 shows that $L_\alpha(P,R)$ is a *quasi-convex* function of P, its first argument.

When we talk of closed sets, we refer to the usual Euclidean metric on $\mathbb{R}^{|\mathbb{X}||\mathbb{Y}|}$. 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 23: (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 $L_\alpha(P_n,R)\to\inf_{P\in\mathcal{E}}L_\alpha(P,R)$. 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 $P_n\to Q$ and $L_\alpha(P_n,R)\to\inf_{P\in\mathcal{E}}L_\alpha(P,R)$. $\mathcal{E}$ is closed and hence $Q\in\mathcal{E}$. The continuity of the logarithm function, wherever it is finite, and the condition $L_\alpha(P_n,R)<\infty$ imply that

$$\begin{aligned}
\lim_n L_\alpha(P_n,R) &= \frac{1}{\rho}\log\left(\mathrm{sign}(\rho)\cdot\lim_n I(P_n,R)\right)\\
&= \frac{1}{\rho}\log\left(\mathrm{sign}(\rho)\cdot I(Q,R)\right)\\
&= L_\alpha(Q,R) &(81)
\end{aligned}$$

where (81) follows from the observation that (34) is the ratio of a continuous linear function of P and the continuous concave function $\mathrm{sign}(1-\alpha)h$ that is bounded, and moreover 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 $\mathcal{E}$. ■

We next state generalizations of [5, Lemma 2.1, Th. 2.2] which show that $L_\alpha(P,Q)$ plays the role of squared Euclidean distance (analogous to the Kullback–Leibler divergence).

Proposition 24: 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.

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

for each

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

if and only if

$$L_\alpha(P,R)\ge L_\alpha(P,Q)+L_\alpha(Q,R). \tag{82}$$

2) *(Tangent hyperplane)* Let

$$Q = \lambda P + (1-\lambda)S, \quad 0<\lambda<1. \tag{83}$$

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

$$L_\alpha(P,R) = L_\alpha(P,Q)+L_\alpha(Q,R). \tag{84}$$

□

Remarks:

1) Under the hypotheses in Proposition 24.1, we deduce that $L_\alpha(P,Q)<\infty$ as a consequence.
2) The condition (83) implies that $P \le \lambda^{-1}Q$ (i.e., every component satisfies the inequality), and therefore $\mathrm{supp}(P)\subset\mathrm{supp}(Q)$. If $0<\alpha<1$, and $L_\alpha(Q,R)<\infty$, then we have $\mathrm{supp}(P)\subset\mathrm{supp}(Q)\subset\mathrm{supp}(R)$. 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 $\mathrm{supp}(P) \subset \mathrm{supp}(Q)$ and $\mathrm{supp}(Q) \cap \mathrm{supp}(R) \neq \emptyset$, we may have $\mathrm{supp}(P) \cap \mathrm{supp}(R) = \emptyset$ leading to $L_\alpha(P,R) = \infty$.

3) Proposition 24.2 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, (84) holds when P is replaced by S.

Proof: 1) $\Rightarrow$. Since $L_\alpha(P,R)$ and $L_\alpha(Q,R)$ are finite, from (33), we gather that both $\sum_y \sum_x P(x,y) R'(x \mid y)^{-\rho}$ and $\sum_y \sum_x Q(x,y) R'(x \mid y)^{-\rho}$ 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

$$\frac{I(P_\lambda, R) - I(P_0, R)}{\lambda} \geq 0 \tag{85}$$

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 (33), $I(P_\lambda, R)$ 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

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

Similarly, it is easy to check that

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

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

Since we can write

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

it follows that the derivative of $s(\lambda)/t(\lambda)$ exists at $\lambda = 0$ and is given by $\left(t(0)\dot{s}(0) - s(0)\dot{t}(0) \right) / t^2(0)$, with the possibility that it might be $+\infty$. However, (85) 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

$$\begin{aligned}&\mathrm{sign}(\rho) \cdot \sum_y \sum_x P(x,y) (R'(x \mid y))^{-\rho} \\ &\quad \geq \mathrm{sign}(\rho) \cdot \left(\sum_y \sum_x P(x,y) \left(Q'(x \mid y) \right)^{-\rho} \right) \\ &\qquad \cdot \left(\frac{\sum_y \sum_x Q(x,y) \left(R'(x \mid y) \right)^{-\rho}}{h(Q)} \right).\end{aligned} \tag{86}$$

When $-1 < \rho < 0$, clearly

$$\sum_y \sum_x P(x,y) \left(Q'(x \mid y) \right)^{-\rho}$$

cannot be zero, due to the nonzero assumptions on the other quantities in (86). This implies that $L_\alpha(P,Q)$ is finite when $1 < \alpha < \infty$ as well. An application of (32) and (33) shows that (86) and (82) 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, (86) and (82) are equivalent. Observe that both sides of (86) are linear in P. This property will be exploited in the proof. Clearly, if we set $P = Q$ in (82) and (86), we have the equalities

$$L_\alpha(Q,R) = L_\alpha(Q,Q) + L_\alpha(Q,R) \tag{87}$$

and

$$\begin{aligned}&\mathrm{sign}(\rho) \cdot \sum_y \sum_x Q(x,y) (R'(x \mid y))^{-\rho} \\ &\quad = \mathrm{sign}(\rho) \cdot \left(\sum_y \sum_x Q(x,y) \left(Q'(x \mid y) \right)^{-\rho} \right) \\ &\qquad \cdot \left(\frac{\sum_y \sum_x Q(x,y) \left(R'(x \mid y) \right)^{-\rho}}{h(Q)} \right).\end{aligned} \tag{88}$$

A λ-weighted linear combination of the inequalities (86) and (88) yields (86) with P replaced by P_λ. The equivalence of (82) and (86) 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, (86) holds for P. Moreover, (86) 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 (83) will satisfy (88) with strict inequality replacing the equality, a contradiction. The reverse implication is straightforward. ■

Let us now apply Proposition 24 to the L_α-projection of a convex set. For a convex $\mathcal{E}$, we call Q an algebraic inner point of $\mathcal{E}$ if for every $P \in \mathcal{E}$, there exist $S \in \mathcal{E}$ and λ satisfying (83).

Theorem 25 (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

$$L_\alpha(P,R) \geq L_\alpha(P,Q) + L_\alpha(Q,R). \tag{89}$$

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 (89) with equality. □.

Proof: This follows easily from Proposition 24. For the case when $L_\alpha(P,R) = \infty$ not covered by Proposition 24, (89) holds trivially. ■

Corollary 26: 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 (89) with equality.

Proof: Clearly, for any $P \in \mathcal{E}$, we have $\mathrm{supp}(P) \subset \mathrm{supp}(Q) \subset \mathrm{supp}(R)$, and therefore $\mathcal{E} \subset B(R, \infty)$. The corollary now follows from the second statement of Theorem 25. ■

While existence of L_α-projection is guaranteed for certain sets by Proposition 23, the following talks about uniqueness of the projection.

Proposition 27: (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 25. Thus $L_\alpha(Q_2,Q_1) = 0$, and $Q_2 = Q_1$. ■

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

Theorem 28: 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 (89) 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 [5, Th. 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

$$L_\alpha(Q_1,R) = L_\alpha(Q_1,Q) + L_\alpha(Q,R). \tag{90}$$

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

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

$$\begin{aligned} &L_\alpha(P,R) \\ &\quad \geq L_\alpha(P,Q_1) + L_\alpha(Q_1,R) && (91)\\ &\quad = L_\alpha(P,Q_1) + L_\alpha(Q_1,Q) + L_\alpha(Q,R) && (92)\end{aligned}$$

where (91) follows from Theorem 25 applied to $\mathcal{E}_1$, and (92) follows from (90).

We next compare (92) 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 25 guarantees that Q_1 is the L_α-projection of Q on $\mathcal{E}_1$. ■

As an application of Theorem 25 let us characterize the L_α-center of a family.

Proposition 29: 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 23, Q^* has an L_α-projection Q on $\mathcal{E}$, and by Proposition 27, the projection is unique on $\mathcal{E}$. From Theorem 25, 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 $Q^* = Q \in \mathcal{E}$. ■

For the special case when $|\mathbb{T}| = m$ is finite, i.e., $\mathbb{T} = \{P_1, \ldots, P_m\}$, we found the weight vector w such that $Q^* = \sum_{i=1}^m w(i)P_i$ and $\sum_{i=1}^m w(\theta) = 1$. This was done in an explicit fashion in Section VI-A2 using results on f-divergences.

IX. Concluding Remarks

We conclude this paper by applying some of our results to guessing of strings of length n with letters in $\mathbb{A}$. Let $\mathbb{X} = \mathbb{A}^n$, $m = |\mathbb{A}|$, and P a PMF on $\mathbb{A}$. 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 $x^n = (x_1, x_2, \ldots, x_n)$. 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 (21), to within $\log(1 + n \ln m)$.

When P_n is a DMS, and the PMF P on $\mathbb{A}$ is unknown to the guesser, Arikan and Merhav [6] have shown that guessing strings in the increasing order of their empirical entropies is a universal strategy. Their universality result is implied by the fact that the normalized $L_{1/2}$-radius of the family of DMSs satisfies $C_n/n \to 0$. The family 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 $\mathbb{A} = \{0, 1\}$; we may think of an n-string as the outcome of independent coin tosses. Suppose further 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 $n \to \infty$, 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 $n \to \infty$. 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.

REFERENCES

[1] J. L. Massey, "Guessing and entropy," in *Proc. 1994 IEEE Int. Symp. Inf. Theory*, Trondheim, Norway, Jun. 1994, p. 204.

[2] E. Arikan, "An inequality on guessing and its application to sequential decoding," *IEEE Trans. Inf. Theory*, vol. IT-42, pp. 99–105, Jan. 1996.

[3] N. Merhav and E. Arikan, "The Shannon cipher system with a guessing wiretapper," *IEEE Trans. Inf. Theory*, vol. 45, pp. 1860–1866, Sep. 1999.

[4] S. Boztaş, "Comments on "An inequality on guessing and its application to sequential decoding"," *IEEE Trans. Inf. Theory*, vol. 43, pp. 2062–2063, Nov. 1997.

[5] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Ann. Prob.*, vol. 3, pp. 146–158, 1975.

[6] E. Arikan and N. Merhav, "Guessing subject to distortion," *IEEE Trans. Inf. Theory*, vol. IT-44, pp. 1041–1056, May 1998.

[7] L. L. Campbell, "A coding theorem and Rényi's entropy," *Inf. Contr.*, vol. 8, pp. 423–429, 1965.

[8] ——, "Definition of entropy by means of a coding problem," *Z. Wahrscheinlichkeitstheorie verw. Geb.*, vol. 6, pp. 113–118, 1966.

[9] A. C. Blumer and R. J. McEliece, "The Rényi redundancy of generalized Huffman codes," *IEEE Trans. Inf. Theory*, vol. 34, pp. 1242–1249, Sep. 1988.

[10] T. R. M. Fischer, "Some remarks on the role of inaccuracy in Shannon's theory of information transmission," in *Proc. Trans. 8th Prague Conf. Inf. Theory*, 1978, pp. 211–226.

[11] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inf. Theory*, vol. IT-19, pp. 783–795, Nov. 1972.

[12] R. G. Gallager, "Source Coding with Side Information and Universal Coding," Technical Report LIDS, LIDS-P-937, 1979.

[13] L. D. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inf. Theory*, vol. IT-26, pp. 166–174, Mar. 1980.

[14] R. Sundaresan, "A measure of discrimination and its geometric properties," in *Proc. 2002 IEEE Int. Symp. Inf. Theory*, Lausanne, Switzerland, Jun. 2002, p. 264.

[15] I. Csiszár, "Generalized cutoff rates and Rényi's information measures," *IEEE Trans. Inf. Theory*, vol. IT-41, pp. 26–34, Jan. 1995.

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[17] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathematica Hungarica*, vol. 2, pp. 191–213, 1972.

[18] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[19] S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. IT-19, pp. 357–359, May 1973.

[20] Q. Xie and A. R. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Trans. Inf. Theory*, vol. IT-46, pp. 431–445, Mar. 2000.

[21] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inf. Theory*, vol. IT-41, pp. 714–722, May 1995.

[22] M. Feder and N. Merhav, "Hierarchical universal coding," *IEEE Trans. Inf. Theory*, vol. IT-42, pp. 1354–1364, Sep. 1996.

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