Guessing Under Source Uncertainty

Rajesh Sundaresan, *Senior Member, IEEE*

***Abstract*— We study the problem of guessing the realization of a finite alphabet source, when the only knowledge available about the source is that it belongs to a (finite or infinite) family. We identify good guessing strategies that minimize the supremum *redundancy* (over the family) due to mismatch, and identify this min-sup value.**

***Index Terms*— f-divergence, guessing, mismatch, redundancy, Rényi information divergence**

I. Inaccuracy and redundancy in guessing

Let $\mathbb{X}$ be a finite alphabet set. Let P be the PMF of a random variable X taking values in $\mathbb{X}$. We need to guess the realization of X. Formally, a guessing list G is a one-to-one function $G : \mathbb{X} \to \{1, 2, \cdots, |\mathbb{X}|\}$ that indicates the order in which the guesses are made. Naturally, knowing the PMF P, the best strategy that minimizes the expected number of guesses goes in the decreasing order of P-probabilities. Let us call such a guessing list G_P. Arikan [1] showed the following general result that gave an operational meaning to the Rényi entropy $H_\alpha(P)$ of order α.

Theorem 1: (Arikan's Guessing Theorem) Let $\rho > 0$ and $\mathbb{X}$ a finite alphabet set. Consider a source on $\mathbb{X}$ with PMF P. Let $\alpha = 1/(1+\rho)$. Then

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

□

Suppose that due to lack of exact knowledge of P we guess in the decreasing order of probabilities of another PMF Q, *i.e.*, we guess in the order given by G_Q. This situation leads to *mismatched* guessing. The penalty suffered in the guessing moment, as a result of the mismatch, is given by [2]

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

to within $\log(1 + \ln|\mathbb{X}|)$.

A universal guessing strategy that guesses in the increasing order of empirical entropy was proposed by Arikan and Merhav in [3]. Their strategy is universal inasmuch as it is asymptotically optimal (within $O((\log n)/n)$) for all finite-alphabet, memoryless sources. In our work, we are interested in understanding this universality and why the maximum penalty is within $O((\log n)/n)$. Moreover, we would like to study and identify good guessing strategies that work well over richer classes of sources.

The quantity $L_\alpha(P, Q)$ also arises in the context of redundancy for Campbell's average exponential coding length problem [4], [5]. In that case, the values that the parameter ρ takes are expanded to $-1 < \rho < 0$ (resp. $1 < \alpha < \infty$) and $0 < \rho < \infty$ (resp. $0 < \alpha < 1$). Our results below for $L_\alpha(P, Q)$ are valid for all these α's.

$L_\alpha(P, Q)$ can be written in terms of some well-understood divergence quantities. Indeed,

$$\begin{aligned} L_\alpha(P, Q) &= D_\beta(P' \parallel Q') & (2) \\ &= \frac{1}{\rho} \log\left(\operatorname{sign}(\rho) \cdot I_f(P' \parallel Q')\right), & (3) \end{aligned}$$

where $\beta = 1/\alpha = 1 + \rho$, $D_\beta(R \parallel S)$ is the Rényi divergence of order β (see for example [6]), I_f is Csiszár's f-divergence (see [7])

$$I_f(R \parallel S) = \sum_{x \in \mathbb{X}} S(x) f\left(\frac{R(x)}{S(x)} \right), \tag{4}$$

$f(x) = \operatorname{sign}(\rho) \cdot x^{1+\rho}$, and P' is the tilted PMF obtained from P and given by

$$P'(x) = \frac{P(x)^\alpha}{\sum_{a \in \mathbb{X}} P(a)^\alpha}.$$

It is known that $\lim_{\alpha \to 1} L_\alpha(P, Q) = D(P \parallel Q)$, the Kullback-Leibler divergence.

II. Problem statement

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

For a fixed $\rho > 0$, we seek a guessing strategy G that works well for all $P \in \mathbb{T}$. More precisely, the redundancy denoted by $R(P, G)$ when the true source is P and when the guessing list is G, is given by

$$R(P, G) \triangleq \frac{1}{\rho} \log\left(\mathbb{E}\left[G(X)^\rho\right]\right) - \frac{1}{\rho} \log\left(\mathbb{E}\left[G_P(X)^\rho\right]\right), \tag{5}$$

where G_P is the optimal guessing scheme when the source PMF is P, and the expectation is with respect to P. The worst redundancy under this guessing strategy is given by

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

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

This work was supported in part by the National Science Foundation under Grant NCR 9523805, and in part by the Ministry of Human Resources and Development (MHRD, India) under Grant Part(2A) Tenth Plan (338/ECE).

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

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

As indicated earlier, the redundancy under mismatch is roughly given by $L_\alpha(P,Q)$. The guessing strategy G can be associated with a certain PMF Q_G. If Q_G is not the same as P, there is mismatch, and the quantity $L_\alpha(P,Q_G)$ is an indication of the suffered penalty. So the following definition is of relevance for the subset of α satisfying $0<\alpha<1$.

Definition 2: For $0<\alpha<\infty, \alpha\neq 1$,

$$C \triangleq \min_Q \sup_{P\in\mathbb{T}} L_\alpha(P,Q). \tag{6}$$

One of the main contributions of this paper is to demonstrate the existence of a minimizing Q^*. For the case when $|\mathbb{T}|$ is finite, previously known results for f-divergences and Rényi divergences can be used to show the existence and characterization of Q^*. In this paper, we address the problem for the case of an infinite uncertainty family $\mathbb{T}$.

Theorem 3: There exists a unique PMF Q^* such that

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

□

Proof: See Section III ■

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" is because it shares the Pythagorean property with Euclidean squared distance and with Kullback-Leibler divergence. This was discussed briefly in [8].

The following result shows that guessing in the decreasing order of Q^*-probabilities, where Q^* attains the min-sup in Definition 2, results in min-sup redundancy, to within $\log(1+\ln|\mathbb{X}|)$.

Theorem 4: (*Guessing under uncertainty*) Let $\mathbb{T}$ be a family of PMFs on $\mathbb{X}$. For a fixed $\rho>0$, there exists a guessing list G^* such that

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

Indeed, we may take $G^* = Q_{G^*}$.

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

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

□

The converse part of Theorem 4 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 $\mathbb{X}=\mathbb{A}^n$, an n-fold cartesian product of a finite alphabet set $\mathbb{A}$, then $\log(1+\ln|\mathbb{X}|)$ grows logarithmically with n. The uncertainty set is therefore rich enough for the converse to be meaningful if C grows with n at a faster rate.

III. Proof of Theorem 3

From Theorem 4, it is clear that the problem at hand is one of identifying the minimizing PMF Q^* and identifying the minimum in (6). Guessing in the decreasing order of Q^*-probabilities leads to a guessing order that works well for all PMFs in the family. The min-sup redundancy is given by C to within $\log(1+\ln|\mathbb{X}|)$. In this section, we show that minimizer Q^* exists, and is unique, for the case when $|\mathbb{T}|$ is not necessarily finite.

The proof outline is as follows. We first solve the related problem

$$\sup_\mu \inf_Q \int_{\mathbb{T}} d\mu(P)\cdot I_f(P' \parallel Q'),$$

where the sup is over measures μ on $(\mathbb{T},\mathcal{T})$. We show that the inf can be replaced by min, and that the sup-min can be interchanged. These are extensions of Csiszár's results to the infinite $|\mathbb{T}|$ case. The proof technique is analogous to a technique used by Gallager in [9, Theorem A]. We then make a connection to the min-sup problem in Definition 2 via (3).

A. L_α-center and radius for an arbitrary family

The development in this subsection is analogous to Gallager's approach [9] for the source coding problem. We first recall the technical conditions put forth in Section II. $\mathbb{T}$ is a family of PMFs on $\mathbb{X}$. $(\mathbb{T},\mathcal{T})$ is a measurable space, and for every $x\in\mathbb{X}$, the mapping $P\mapsto P(x)$ is $\mathcal{T}$-measurable.

We next define a few auxiliary quantities. For a given P, let the α-norm of the PMF P be given by

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

The dependence of h on α is understood, and suppressed for convenience. The function $\text{sign}(\rho)\cdot h$ is a concave function of its argument. The Rényi entropy is given by

$$H_\alpha(P) = \frac{\alpha}{1-\alpha}\log h(P). \tag{7}$$

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

$$1 \le h(P) \le |\mathbb{X}|^{\frac{1-\alpha}{\alpha}},\ \text{for } 0<\alpha<1, \tag{8}$$

and

$$|\mathbb{X}|^{\frac{1-\alpha}{\alpha}} \le h(P) \le 1,\ \text{for } 1<\alpha<\infty. \tag{9}$$

In both cases, we see that $h(P)$ is bounded away from 0.

For the sake of clarity, let us define

$$I(P,Q) \triangleq I_f(P' \parallel Q'). \tag{10}$$

Then from (3) we can write

$$L_\alpha(P,Q) = \frac{1}{\rho}\log\left(\text{sign}(\rho)\cdot I(P,Q)\right). \tag{11}$$

Our focus will be on the following:

Definition 5: For $0 < \alpha < \infty, \alpha \neq 1$,

$$K_+ \;\triangleq\; \min_Q \sup_{P \in \mathbb{T}} I(P,Q). \tag{12}$$

Taking Q to be the uniform PMF on $\mathbb{X}$ 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 $-1 < \rho < 0$.

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

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

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

$$F \triangleq \int_{\mathbb{T}} d\mu(P) \cdot f(P). \tag{13}$$

Define the PMF μf on $\mathbb{X}$ as the scaled version of F,

$$\mu f \triangleq d^{-1} F \tag{14}$$

where d is the normalizing constant

$$d \triangleq \int_{\mathbb{T}} \frac{d\mu(P)}{h(P)} = \sum_{x \in \mathbb{X}} F(x). \tag{15}$$

Moreover, let

$$J(\mu, \mathbb{T}) \triangleq \int_{\mathbb{T}} d\mu(P) \cdot I(P, \mu f). \tag{16}$$

Simple algebraic manipulations result in

$$J(\mu,\mathbb{T}) \;=\; \operatorname{sign}(\rho) \cdot h(F) \tag{17}$$

$$\;=\; \operatorname{sign}(\rho) \cdot d \cdot h(\mu f); \tag{18}$$

these are extensions of [7, Equation (2.24)] for arbitrary $\mathbb{T}$.

The following definition will be useful in the proof.

Definition 6: For $0 < \alpha < \infty, \alpha \neq 1$,

$$K_- \;\triangleq\; \sup_\mu J(\mu, \mathbb{T}). \tag{19}$$

To help fix ideas, we now describe some parallels with classical information theoretic quantities. $\mathbb{T}$ represents a channel where the input alphabet is any index set for the PMFs in $\mathbb{T}$. The output alphabet is $\mathbb{X}$. The quantity μf in (14) is analogous to the PMF at the output of the channel $\mathbb{T}$ when the input measure is μ. $J(\mu, \mathbb{T})$ in (16) is the analogue of mutual information; Csiszár calls it informativity in his work on finite families [7].

Proposition 7: $K_- \leq K_+$.

Proof: Fix an arbitrary PMF Q on $\mathbb{X}$. It is straightforward to show that [7, Equation 2.26] holds for the arbitrary $\mathbb{T}$ case as well, and is given by

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

From the convexity of f, we have $I(\mu f, Q) \geq \operatorname{sign}(\rho)$. It follows that

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

Consequently

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

which leads to

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

■

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

Proposition 8: 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 $\mathbb{X}$ 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, and $K_- = K_+$. □

Proof: $\Leftarrow$: Observe that on account of 1), 3), and Proposition 7, we have

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

where the first inequality follows from 1) and the definition of K_-, the second from 3), and the last from Proposition 7. Consequently, all the inequalities are equalities, $R = K_- = K_+$, and the use of min in the definition of K_+ is justified.

$\Rightarrow$: Since $R = K_- \leq K_+$ and is therefore finite, by definition of K_-, there exists a sequence $(\mu_n : n \in \mathbb{N})$ such that $\lim_n J(\mu_n, \mathbb{T}) = R$.

Now consider the sequence of $|\mathbb{X}|$-dimensional vectors given by $F_n = \int_{\mathbb{T}} d\mu_n(P) \cdot f(P)$. This is a sequence of scaled PMFs given by $F_n = d_n \cdot \mu_n f$, where d_n is given by (15). This is clearly a bounded quantity. The sequence therefore 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 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

$$\nu_{n,\lambda} = (1-\lambda)\mu_n + (\lambda)\delta_{P_0}, \tag{20}$$

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

From (20), (13), and (17), we have

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

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

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

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

Next, we can straightforwardly check that the one-sided derivative at $\lambda = 0$, denoted by $\dot{s}(0)$, is given by

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

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

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

we have a contradiction. So 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

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

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

B. Proof of Theorem 3

Proof: From (11), it is clear that

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

Q^* attains the min-sup value K_+ in Definition 5 if and only if Q^* attains the min-sup value C in Definition 2. Proposition 8 guarantees the existence and uniqueness of such a Q^*. ■

IV. Discussion

In this section, we will specialize our results to binary n-strings. Let $\mathbb{X} = \{0,1\}^n$, and P a PMF on $\{0,1\}$. Let

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

denote the PMF of the discrete memoryless source (DMS) with $x^n = (x_1, x_2, \cdots, x_n)$. Theorem 1 says that for $\rho = 1$, the minimum expected number of guesses grows exponentially with n; the growth rate is given by $H_{1/2}(P)$. Such results find application in the analysis of private-key cryptosystems where users have access to a fixed random source to generate the key-string. The higher the growth rate of the minimum expected number of guesses, the better the security of the system.

When all that the guesser knows is that the source $P^n \in \mathbb{T}$, the guesser suffers a penalty (also 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 for $\mathbb{T}$. This normalized redundancy equals the normalized $L_{1/2}$-radius of $\mathbb{T}$, *i.e.*, C_n/n, where C_n is given by (6).

When P^n is a DMS, but the PMF P on $\{0,1\}$ is unknown to the guesser, Arikan and Merhav [3] have shown that there is a universal guessing strategy that works well for all DMSs. Their universal guessing strategy, as indicated earlier, guesses strings in the increasing order of their empirical entropies. Their universality result implies that the normalized $L_{1/2}$-radius of the family of DMSs satisfies $C_n/n \to 0$. The set of DMSs is thus not rich enough from the point of view of guessing. Knowledge of the PMF P is not needed; the universal strategy achieves, asymptotically, the minimum growth rate achievable with full knowledge of the source statistics.

It is known that the family of finite-state, arbitrarily varying sources has normalized $L_{1/2}$-radius approaching a strictly positive constant as $n \to \infty$ under some circumstances [2]. Such a family is rich in the following sense; the growth rate of the minimum expected number of guesses without knowledge of source statistics is strictly larger than that achievable with full knowledge of source statistics.

References

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

[2] R. Sundaresan, “Guessing under source uncertainty,” *Manuscript under preparation*, Feb. 2006.

[3] E. Arikan and N. Merhav, “Guessing subject to distortion,” *IEEE Trans. Inform. Theory*, vol. IT-44, pp. 1041–1056, May 1998.

[4] L. L. Campbell, “A coding theorem and Rényi’s entropy,” *Information and Control*, vol. 8, pp. 423–429, 1965.

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

[6] I. Csiszár, “Generalized cutoff rates and rényi’s information measures,” *IEEE Trans. Inform. Theory*, vol. IT-41, pp. 26–34, Jan. 1995.

[7] ——, “A class of measures of informativity of observation channels,” *Periodica Mathematica Hungarica*, vol. 2, pp. 191–213, 1972.

[8] R. Sundaresan, “A measure of discrimination and its geometric properties,” in *Proc. 2002 IEEE Int. Symp. on Information Theory*, Lausanne, Switzerland, Jun. 2002, p. 264.

[9] R. G. Gallager, “Source coding with side information and universal coding,” *LIDS Technical Report, LIDS-P-937*, Sept. 1979.