Guessing Based On Length Functions

Rajesh Sundaresan
ECE Department
Indian Institute of Science
Bangalore 560012, India
E-mail: rajeshs@ece.iisc.ernet.in

***Abstract*— Close relationships between guessing functions and length functions are established. Good length functions lead to good guessing functions. In particular, guessing in the increasing order of Lempel-Ziv lengths has certain universality properties for finite-state sources. As an application, these results show that hiding the parameters of the key-stream generating source in a private key crypto-system may not enhance the privacy of the system, the privacy level being measured by the difficulty in brute-force guessing of the key stream.**

I. INTRODUCTION

We consider the problem of guessing the realization of a random variable and relate the required number of guesses to a lossless code's length function. Specifically, we sandwich the number of guesses on either side by a suitable length function. This directly establishes Arikan's result [1] that the best value of the guessing exponent is close to the average exponential coding length for Campbell's coding problem which is given by Rényi entropy of appropriate order. Our approach also shows that guessing based on lossless universal compressors leads to good universal guessing strategies. Indeed, guessing in the increasing order of Lempel-Ziv lengths for finite-state sources and increasing description lengths for unifilar sources achieves optimality in a sense made precise in the sequel.

In Section II we establish the relationship between guessing and compression. In Section III we show that guessing based on Lempel-Ziv lengths is universal. We end with an application in Section IV where we show that hiding parameters of a key-stream generating source, even if the source comes from a fairly large uncertainty set, does not enhance the privacy of the crypto-system.

Detailed proofs and extensions to guessing with key rate constraints can be found in a more recent work [2].

II. GUESSING AND SOURCE COMPRESSION

Let $\mathbb{X}$ be a finite alphabet set. A guessing function

$$G : \mathbb{X} \to \{1, 2, \cdots, |\mathbb{X}|\}$$

is a bijection that denotes the order in which the elements of $\mathbb{X}$ are guessed. If $G(x) = i$, then the ith guess is x. A length function

$$L : \mathbb{X} \to \mathbb{N}$$

is one that satisfies Kraft's inequality

$$\sum_{x \in \mathbb{X}} \exp\{-L(x)\} \leq 1. \tag{1}$$

To each guessing function G, we associate a probability mass function (PMF) Q_G on $\mathbb{X}$ and a length function L_G as follows.

Definition 1: Given a guessing function G, we say

$$Q_G(x) = c^{-1} \cdot G(x)^{-1}, \ \forall x \in \mathbb{X}, \tag{2}$$

is the PMF on $\mathbb{X}$ associated with G. The quantity c in (2) is the normalization constant. We say L_G defined by

$$L_G(x) = \lceil -\log Q_G(x) \rceil, \ \forall x \in \mathbb{X}, \tag{3}$$

is the length function associated with G. □

Observe that

$$c = \sum_{a \in \mathbb{X}} G(a)^{-1} = \sum_{i=1}^{|\mathbb{X}|} \frac{1}{i} \leq 1 + \ln |\mathbb{X}|, \tag{4}$$

and therefore the PMF in (2) is well-defined. We record the intimate relationship between these associated quantities in the following theorem.

Theorem 2: Given a guessing function G, the associated quantities satisfy

$$c^{-1} \cdot Q_G(x)^{-1} = G(x) \leq Q_G(x)^{-1}, \tag{5}$$

$$L_G(x) - 1 - \log c \leq \log G(x) \leq L_G(x). \tag{6}$$

□

Proof: The first equality in (5) follows from the definition in (2), and the second inequality from the fact that $c \geq 1$.

The upper bound in (6) follows from the upper bound in (5) and from (3). The lower bound in (6) follows from

$$\begin{aligned}
\log G(x) &= \log\left(c^{-1} \cdot Q_G(x)^{-1}\right) \\
&= -\log Q_G(x) - \log c \\
&\geq (\lceil -\log Q_G(x) \rceil - 1) - \log c \\
&= L_G(x) - 1 - \log c.
\end{aligned}$$

■

We now associate a guessing function G_L to each length function L.

Definition 3: Given a length function L, the associated guessing function G_L guesses in the increasing order of L-lengths. Sequences with the same L-length are ordered using an arbitrary fixed rule, say the lexicographic order. We also define the associated PMF Q_L on $\mathbb{X}$ to be

$$Q_L(x) = \frac{\exp\{-L(x)\}}{\sum_{a \in \mathbb{X}} \exp\{-L(a)\}}. \tag{7}$$

□

Theorem 4: For a length function L, the associated PMF and the guessing function satisfy the following:

1) G_L proceeds in the decreasing order of Q_L-probabilities;
2)
$$\log G_L(x) \le \log Q_L(x)^{-1} \le L(x). \tag{8}$$

□

Proof: The first statement is clear from the definition of G_L and from (7).

Letting $1\{E\}$ denote the indicator function of an event E, we have as a consequence of statement 1) that

$$\begin{aligned} G_L(x) &\le \sum_{a\in\mathbb{X}} 1\left\{Q_L(a) \ge Q_L(x)\right\} \\ &\le \sum_{a\in\mathbb{X}} \frac{Q_L(a)}{Q_L(x)} \\ &= Q_L(x)^{-1}, \end{aligned} \tag{9}$$

which proves the left inequality in (8). This inequality was known to Wyner [3].

The last inequality in (8) follows from (7) and Kraft's inequality (1) as follows:

$$Q_L(x)^{-1} = \exp\{L(x)\} \cdot \sum_{a\in\mathbb{X}} \exp\{-L(a)\} \le \exp\{L(x)\}.$$

■

The inequalities between the associates in (6) and (8) indicate the direct relationship between guessing moments and Campbell's coding problem [4] and that the Rényi entropies are the optimal growth exponents for guessing moments. They also establish that the minimum expected value of the logarithm of the number of guesses is close to the Shannon entropy.

We now demonstrate other relationships between guessing moments and average exponential coding lengths which will be useful in establishing universality properties.

Theorem 5: Let L be any length function on $\mathbb{X}$, G_L the guessing function associated with L, P a PMF on $\mathbb{X}$, $\rho \in (0,\infty)$, L^* the length function that minimizes $\mathbb{E}\left[\exp\{\rho L^*(X)\}\right]$, where the expectation is with respect to P, G^* the guessing function that proceeds in the decreasing order of P-probabilities and therefore the one that minimizes $\mathbb{E}\left[G^*(X)^\rho\right]$, and c as in (4). Then

$$\frac{\mathbb{E}\left[G_L(X)^\rho\right]}{\mathbb{E}\left[G^*(X)^\rho\right]} \le \frac{\mathbb{E}\left[\exp\{\rho L(X)\}\right]}{\mathbb{E}\left[\exp\{\rho L^*(X)\}\right]} \cdot \exp\{\rho(1+\log c)\}. \tag{10}$$

Analogously, let G be any guessing function, and L_G its associated length function. Then

$$\frac{\mathbb{E}\left[G(X)^\rho\right]}{\mathbb{E}\left[G^*(X)^\rho\right]} \ge \frac{\mathbb{E}\left[\exp\{\rho L_G(X)\}\right]}{\mathbb{E}\left[\exp\{\rho L^*(X)\}\right]} \cdot \exp\{-\rho(1+\log c)\}. \tag{11}$$

□

Proof: Observe that

$$\begin{aligned} &\mathbb{E}\left[\exp\{\rho L(X)\}\right] \\ &\quad \ge \mathbb{E}\left[G_L(X)^\rho\right] &(12)\\ &\quad \ge \mathbb{E}\left[G^*(X)^\rho\right] \\ &\quad \ge \mathbb{E}\left[\exp\{\rho L_{G^*}(X)\}\right]\exp\{-\rho(1+\log c)\} &(13)\\ &\quad \ge \mathbb{E}\left[\exp\{\rho L^*(X)\}\right]\exp\{-\rho(1+\log c)\}, \end{aligned}$$

where (12) follows from (8), and (13) from the left inequality in (6). The result in (10) immediately follows. A similar argument shows (11). ■

We end this section by recording the following rather obvious corollary to Theorems 2 and 4. We use the short form $\{L(x) \ge B\}$ to denote the set $\{x \in \mathbb{X} \mid L(x) \ge B\}$.

Corollary 6: For a given G, its associated length function L_G, and any $B \ge 1$, we have

$$\begin{aligned} &\{L_G(x) \ge B + 1 + \log c\} \\ &\qquad \subseteq \{G(x) \ge \exp\{B\}\} \\ &\qquad \subseteq \{L_G(x) \ge B\}. \end{aligned} \tag{14}$$

Analogously, for a given L, its associated guessing function G_L, and any positive $B \ge 1$, we have

$$\{G_L(x) \ge \exp\{B\}\} \subseteq \{L(x) \ge B\}. \tag{15}$$

□

III. Universal Guessing

In this section, we give an application of the above inclusions to conclude a universality property of guessing in the increasing order of Lempel-Ziv lengths [5]. We also show that universality for Campbell's coding problem implies universality for guessing.

Let $x^n = (x_1, \cdots, x_n)$ be a string taking values in $\mathbb{X}^n$, where $|\mathbb{X}| < \infty$. The string x^n needs to be guessed. Let $s^n = (s_1, \cdots, s_n)$ be another sequence taking values in $\mathbb{S}^n$ where $|\mathbb{S}| < \infty$. Let $s_0 \in \mathbb{S}$ be a fixed initial state. A probabilistic source P_n is finite-state with $|\mathbb{S}|$ states [6] if the probability of observing the sequence pair (x^n, s^n) is given by

$$P_n(x^n, s^n \mid s_0) = \prod_{i=1}^{n} P(x_i, s_i \mid s_{i-1}),$$

where $P(x_i, s_i \mid s_{i-1})$ is the joint probability of letter x_i and state s_i given the previous state s_{i-1}. Typically, the letter sequence x^n is observable and the state sequence s^n is not. We will let H denote the entropy-rate of a finite-state source, i.e.,

$$H \triangleq -\lim_{n\to\infty} n^{-1} \sum_{x^n\in\mathbb{X}^n} P_n(x^n \mid s_0) \log P_n(x^n \mid s_0).$$

Let $U_{LZ} : \mathbb{X}^n \to \mathbb{N}$ be the length function for the Lempel-Ziv code [5]. The following theorem due to Merhav [6] indicates that the Lempel-Ziv algorithm is asymptotically optimal in achieving the minimum probability of buffer overflow.

Theorem 7 (Merhav [6]): For any length function L_n, every finite-state source P_n, every $B_n \in (nH, n\log|\mathbb{X}|)$ where

H is the entropy-rate of the source P_n, and all sufficiently large n,

$$P_n\{U_{LZ}(X^n) \geq B_n + n\varepsilon(n)\} \\ \leq (1+\delta(n)) \cdot P_n\{L_n(X^n) \geq B_n\} \quad (16)$$

where $\varepsilon(n) = \Theta(1/\sqrt{\log n})$ is a positive sequence that depends on $|\mathbb{X}|$ and $|\mathbb{S}|$, and $\delta(n) = n^2 \exp\{-n\varepsilon(n)\}$. □

Theorem 7 is a variant of [6, Th. 1]. Merhav states [6, Th. 1] for $B_n = nB$ for a constant $B \in (H, \log|\mathbb{X}|)$, but his proof is valid for any sequence $B_n \in (nH, n\log|\mathbb{X}|)$.

Let G_{LZ} be the short-hand notation for the more cumbersome $G_{U_{LZ}}$, the guessing function associated with U_{LZ}. We show that G_{LZ} has the following asymptotic optimality property for large deviations performance. Let c_n be as given in (4) with $\mathbb{X}^n$ replacing $\mathbb{X}$.

Theorem 8: For any guessing function G_n, every finite-state source P_n, every $B \in (H, \log|\mathbb{X}|)$ where H is the entropy-rate of the source P_n, and all sufficiently large n,

$$P_n\left\{n^{-1}\log G_{LZ}(X^n) \geq B + \varepsilon(n) + \gamma(n)\right\} \\ \leq (1+\delta(n)) \cdot P_n\left\{n^{-1}\log G_n(X^n) \geq B\right\} \quad (17)$$

where $\varepsilon(n)$ and $\delta(n)$ are the sequences in (16), and $\gamma(n) = (1+\log c_n)/n = \Theta(n^{-1}\log n)$. □

Proof: Observe that

$$\begin{aligned}(1+\delta(n))&P_n\{G_n(X^n) \geq \exp\{nB\}\} \\ &\geq (1+\delta(n))P_n\{L_{G_n}(X^n) \geq nB + 1 + \log c_n\} \quad (18)\\ &\geq P_n\{U_{LZ}(X^n) \geq nB + 1 + \log c_n + n\varepsilon(n)\} \quad (19)\\ &\geq P_n\{G_{LZ}(X^n) \geq \exp\{n(B+\varepsilon(n)+\gamma(n))\}\}, \quad (20)\end{aligned}$$

where (18) follows from the first inclusion in (14), and (19) from (16) with $B_n = nB+1+\log c_n$. The last inequality (20) follows from (15). This proves the theorem. ■

Observe that $\varepsilon(n) + \gamma(n) = \Theta(1/\sqrt{\log n})$. For unifilar sources (where s_i is a deterministic function of (x_i, s_{i-1}), and given given s_{i-1} this function is a bijection), a stronger statement can be made when the number of states $|\mathbb{S}|$ is known. In particular, $\varepsilon(n) + \gamma(n) = \Theta(n^{-1}\log n)$, and guessing for this class of sources proceeds in the order of increasing description lengths.

We now demonstrate a competitive optimality property for G_{LZ}. From [6, eqn. (28)] extended to finite-state sources, we have for any competing code L_n

$$P_n\{U_{LZ}(X^n) > L_n(X^n) + n\varepsilon(n)\} \\ \leq \quad P_n\{U_{LZ}(X^n) < L_n(X^n) + n\varepsilon(n)\} \quad (21)$$

where $\varepsilon(n) = \Theta((\log\log n)/(\log n))$. From (8), we get

$$U_{LZ}(x^n) \geq \log G_{LZ}(x^n),$$

and from (6), we obtain

$$\log G(x^n) \geq L_G(x^n) - 1 - \log c_n.$$

We therefore conclude that

$$\{\log G_{LZ}(x^n) > \log G(x^n) + n(\varepsilon(n)+\gamma(n))\} \\ \subseteq \quad \{U_{LZ}(x^n) > L_G(x^n) + n\varepsilon(n)\}$$

and that

$$\{U_{LZ}(x^n) < L_G(x^n) + n\varepsilon(n)\} \\ \subseteq \quad \{\log G_{LZ}(x^n) < \log G(x^n) + n(\varepsilon(n)+\gamma(n))\}.$$

From these two inclusions and (21), we easily deduce the following result.

Theorem 9: For any finite-state source and any competing guessing function G, we have

$$P_n\{\log G_{LZ}(X^n) > \log G(X^n) + n\varepsilon'(n)\} \\ \leq \quad P_n\{\log G_{LZ}(X^n) < \log G(X^n) + n\varepsilon'(n)\}$$

where $\varepsilon'(n) = \varepsilon(n) + \gamma(n)$. □

Yet again, for unifilar sources, the above sequence of arguments for minimum description length coding and [6, eqn. (28)] imply that we may take $\varepsilon'(n) = \Theta(n^{-1}\log n)$.

We now show that universality in the average exponential coding rate sense implies the existence of a universal guessing strategy that achieves the optimal exponent for guessing moments.

Consider a class of sources. For each source in the class, let P_n be its restriction to strings of length n and let L_n^* denote an optimal length function that attains the minimum value $\mathbb{E}\left[\exp\left\{\rho L_n(X^n)\right\}\right]$ among all length functions, the expectation being with respect to P_n. On the other hand, let L_n be a sequence of length functions for the class of sources that does not depend on the actual source within the class. Suppose further that the length sequence L_n is asymptotically optimal, i.e.,

$$\lim_{n\to\infty}\frac{1}{n\rho}\log\mathbb{E}\left[\exp\{\rho L_n(X^n)\}\right] \\ = \lim_{n\to\infty}\frac{1}{n\rho}\log\mathbb{E}\left[\exp\{\rho L_n^*(X^n)\}\right].$$

for every source belonging to the class. L_n is thus "universal" for (i.e., asymptotically optimal for all sources in) the class. An application of (10) and the fact $(1+\log c_n)/n \to 0$ indicate that the sequence of guessing strategies G_{L_n} is asymptotically optimal for the class, i.e.,

$$\lim_{n\to\infty}\frac{1}{n\rho}\log\mathbb{E}\left[G_{L_n}(X^n)^\rho\right] \\ = \lim_{n\to\infty}\frac{1}{n\rho}\log\mathbb{E}\left[G^*(X^n)^\rho\right].$$

Arikan and Merhav [7] provide a universal guessing strategy for the class of discrete memoryless sources (DMS). For the class of unifilar sources with a known number of states, the minimum description length encoding is asymptotically optimal for Campbell's coding length problem (see Merhav [6]). It follows as a consequence of the above argument that guessing in the increasing order of description lengths is asymptotically optimal for this class. The left side of (10) is the extra factor in the expected number of guesses (relative to the optimal value) due to lack of knowledge of the specific source in class. Our prior work [8] characterizes this loss as a function of the uncertainty class.

IV. An Application

Consider a crypto-system using 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.

The expected number of Eve's guesses grows exponentially with an optimal growth exponent given by the Rényi entropy of the source of order $1/2$ [1]. From Alice's and Bob's points of view, this growth exponent is a measure of goodness of their key stream generating source. Merhav and Arikan [9] have generalized this result to systems with specified key rates. Recently, in [8] we looked at the scenario where Alice and Bob have a *bag of sources* from which the source that generates the key stream is chosen. While the sources in the bag are known to Eve, she is unaware of the exact source chosen.

For example, Alice and Bob may realize that they have a bad source, but have no other option than to use this source to generate their private key stream. The following question then arises naturally: does hiding the parameters of this source enhance privacy.

We showed in [8] that even if the source is any discrete memoryless source whose parameters are unknown to Eve, she has a guessing strategy that is asymptotically optimal. In other words, Eve's number of guesses to arrive at the correct key stream grows exponentially with the length of the key stream with an exponential growth rate asymptotically the same as that obtainable with knowledge of source statistics. Eve's lack of knowledge of the chosen memoryless source's parameters makes the crypto-system marginally more secure, but the extra work Eve has to do to guess right is asymptotically negligible. In the sense of the growth exponent of the expected number of guesses,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[G_n(X^n)\right], \tag{22}$$

the system is not any more secure due to Eve's lack of knowledge of the source parameters. This negative result does not change if the measure of security is modified to moments of guessing of order $\rho \in (0, \infty)$ as measured by their exponent ([8], [7])

$$\liminf_{n\to\infty} \frac{1}{n\rho} \log \mathbb{E}\left[G_n(X^n)^\rho\right], \tag{23}$$

or even when the measure of security is modified to large deviations performance [7] as indicated by the rate at which the tail probability vanishes:

$$F(B; G) = -\liminf_{n\to\infty} \frac{1}{n} \log \Pr\left\{G_n(X^n) \geq \exp\{nB\}\right\}, \tag{24}$$

where $G = \{G_n : n \geq 1\}$, and G_n is a guessing strategy on strings of length n.

Our universality results of this paper imply that the crypto-system *cannot* be made more secure, based on the above quantitative measures, even if the key stream generating source comes from a wider class of sources. Specifically, we showed that the attacker has an asymptotically optimal guessing strategy when the key stream generating source is chosen from unifilar sources of given order in the senses of (23) and (24). A similar negative result holds in the sense of (24) for finite-state sources of arbitrary order.

As shown in [8] there does exist an uncertainty set for which hiding the information enhances privacy in the sense of (22). The set of arbitrarily varying sources is one example. Sources from this uncertainty set are not finite-state sources and hence results from this paper are not applicable to such sources.

Acknowledgment

This work was supported by the Defence Research & Development Organisation (DRDO), Ministry of Defence, Government of India, under the DRDO-IISc Programme on Advanced Research in Mathematical Engineering, and by the University Grants Commission under Grant Part (2B) UGC-CAS-(Ph.IV)

The author thanks an anonymous referee for pointing out reference [3] and for suggestions that simplified the proof of Theorem 2.

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 based on length functions," *submitted to the IEEE Trans. Inform. Theory*, http://arxiv.org/abs/cs.IT/0702115, Feb. 2007.

[3] A.D.Wyner, "An upper bound on the entropy series," *Information and Control*, vol. 20(2), pp. 176–181, Mar. 1972.

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

[5] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. 24, no. 5, pp. 530–536, Sept. 1978.

[6] N. Merhav, "Universal coding with minimum probability of codeword length overflow," *IEEE Trans. Inform. Theory*, vol. 37, no. 3, pp. 556 – 563, May 1991.

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

[8] R. Sundaresan, "Guessing under source uncertainty," *IEEE Trans. Inform. Theory*, vol. 53, no. 1, pp. 269–287, Jan. 2007.

[9] N. Merhav and E. Arikan, "The Shannon cipher system with a guessing wiretapper," *IEEE Trans. Inform. Theory*, vol. 45, no. 6, pp. 1860–1866, Sep. 1999.