DRDO–IISc Programme on
Advanced Research in Mathematical
Engineering

Guessing and compression subject to distortion

(TR-PME-2010-12)

by

Manjesh Kumar Hanawal and Rajesh Sundaresan
Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore

12 February 2010

Indian Institute of Science
Bangalore 560 012

Guessing and compression subject to distortion

Manjesh Kumar Hanawal and Rajesh Sundaresan

12 February 2010

ABSTRACT

The problem of guessing a random string is revisited. The relationship between guessing without distortion and compression is extended to the case when source alphabet size is countably infinite. Further, similar relationship is established for the case when distortion allowed by establishing a tight relationship between rate distortion codes and guessing strategies.

Keywords: test

1 Introduction

Let $X^n = (X_1, \cdots, X_n)$ denote n letters of a random process where each letter is drawn from a discrete set $\mathbb{X}$. The joint probability mass function (PMF) is given by $(P_n(x^n) : x^n \in \mathbb{X}^n)$. Consider a ordered list of guesses denoted by $\mathcal{G}_n := \{y^n(1), y^n(2), \cdots\}$. Our interest is in guessing the realisation of a random string by stepping through the elements of $\mathcal{G}_n$. For a given $D \geq 0$, we say $\mathcal{G}_n$ is a *D-admissible guessing strategy* if for each $x^n \in \mathbb{X}^n$ there is a $y^n(j) \in \mathcal{G}_n$ such that $d(x^n, y^n(j)) \leq nD$, where $d(\cdot, \cdot)$ denotes a given distortion measure. The ordered list $\mathcal{G}_n$ induces a *guessing function* from $\mathbb{X}^n$ onto the set of natural numbers denoted by $\mathbb{N}$, i.e., $G_n : \mathbb{X}^n \to \mathbb{N}$, and defined by

$$G_n(x^n) = \min \{j \ : \ d(x^n, y^n(j)) \leq nD\}, \ \ \forall \ x^n \in \mathbb{X}^n.$$

The above model is applicable in search problems such as approximate pattern matching and database searches where one possesses only partial information about a target [1] while executing a query. The performance criterion is the rate of growth of the expected number of guesses as the length of string being searched grows. Specifically, we wish to evaluate the optimal exponential growth rate of guessing moments, i.e.,

$$\mathcal{E}(D, \rho) = \lim_{n \to \infty} \inf_{\mathcal{G}_n} \frac{1}{n} \log \mathbb{E}[G_n(X^n)^\rho]$$

where ρ is a arbitrary positive real number and infimum is taken over all D-admissible guessing strategies. The limit above may or may not exist and an identification of classes of sources for which the limit exists is also of interest.

Arikan [2] considered the above problem for the case when $D = 0$, i.e., lossless guessing. The optimal guessing strategy in this case is the one which proceeds in the decreasing order of message probabilities [3], a strategy that is denoted G_n^*. Arikan [2] showed that the guessing exponent exists for independent and identically distributed (iid) sources taking values on a finite alphabet set. The limiting exponent is the Rényi entropy of order $1/(1+\rho)$.

Sundaresan [4] showed that for any finite alphabet source, the problem of finding the guessing exponent is the same as that of finding the exponential growth rate for moment generating function for compressed lengths, a problem proposed by Campbell [5]. This equivalence was established by associating a length function to each guessing function and vice-versa. One purpose of this report is to extend these arguments to the case when the alphabet is countably infinite.

Merhav and Arikan [1] studied the problem of guessing a random string with $D \geq 0$. They showed that the guessing exponent $\mathcal{E}(D, \rho)$ exists for an iid source, among other sources, and commented that a similar method can be used to solve the compression variant as well (exponents of moment generating function for compressed lengths). The second purpose of this report is to make this connection rigorous.

This report is organised as follows. Section 2 considers the lossless case ($D = 0$) and $\mathbb{X}$ countably infinite. Section 3 addresses $D \geq 0$.

2 Guessing without distortion

This section considers the case when $\mathbb{X}$ is countably infinite and establishes equality of exponents of guessing and moment generating functions for compressed lengths. The approach will be nearly the same as that of [4, Sec. 2] but for a minor technical point which we will resolve.

We shall denote the source by $(X^n : n \in \mathbb{N})$ with $X_i \in \mathbb{X}$, a countably infinite set. A *length function* is a mapping $L_n : \mathbb{X}^n \to \mathbb{N}$ such that the Kraft inequality

$$\sum_{x^n \in \mathbb{X}^n} 2^{-L_n(x^n)} \leq 1$$

holds. The engineering interpretation is that $L_n(x^n)$ is the compression length of the string x^n. We first associate a guessing function to each length function.

Definition 1 *Given a length function L_n, the associated guessing function G_{L_n} is the one that guesses strings in the increasing order of L_n-lengths.*

Strings with the same L_n-lengths are ordered using an arbitrary fixed rule, say the lexicographical order on $\mathbb{X}$. We also define the associated probability mass function (PMF) Q_{L_n} on $\mathbb{X}$ to be

$$Q_{L_n}(x^n) = \frac{2^{-L_n(x^n)}}{\sum_{y^n \in \mathbb{X}^n} 2^{-L_n(y^n)}}.$$

□

The following proposition is a restatement of [4, Prop. 5]. It holds verbatim even when the source alphabet size is countably infinite, and is restated here for completeness.

Proposition 1 *For a length function L_n and $B \geq 1$, the associated guessing function G_{L_n} satisfies the following:*

$$\log G_{L_n}(x^n) \leq Q_{L_n}(x^n)^{-1} \leq L_n(x^n), \tag{1}$$

$$\{x^n : G_{L_n}(x^n) \geq 2^B\} \subseteq \{x^n : L(x^n) \geq B\}. \tag{2}$$

□

We next associate a length function to every guessing function as follows.

Definition 2 *Given any guessing function G_n and $\delta > 0$, we say Q_{G_n} defined by*

$$Q_{G_n}(x^n) = c_n(\delta)^{-1} \cdot G_n(x^n)^{-1-\delta}, \ \forall x^n \in \mathbb{X}^n \tag{3}$$

is the PMF associated with G_n. The quantity $c_n(\delta)$ is the normalisation constant. We say L_{G_n} defined by

$$L_{G_n}(x^n) = \lceil -\log Q_{G_n}(x^n) \rceil, \ \forall x^n \in \mathbb{X}^n$$

is the length function associated with G_n. □

Observe that for any $\delta > 0$, we have

$$c_n(\delta) = \sum_{i=1}^{\infty} \frac{1}{i^{1+\delta}} < \infty,$$

and hence the PMF Q_{G_n} is well defined. Similar to Proposition 2 in [4] we state the relation between associated quantities in the following proposition.

Proposition 2 *To each guessing function G_n, there exists a length function L_{G_n} such that $\forall\ x^n \in \mathbb{X}^n$ and $\delta > 0$*

$$\frac{L_{G_n}(x^n) - 1 - \log c_n(\delta)}{1+\delta} \leq \log G_n(x^n) \leq \frac{L_{G_n}(x^n)}{1+\delta}.$$

□

Proof: From the definition of L_{G_n} we have

$$L_{G_n}(x^n) = \lceil -\log Q_{G_n}(x^n) \rceil \leq 1 + \log(c_n(\delta) \cdot G_n(x^n)^{1+\delta}).$$

Rearranging the above we get

$$\frac{L_{G_n}(x^n) - 1 - \log c_n(\delta)}{1+\delta} \leq \log G_n(x^n).$$

Furthermore, because of (3) and the fact that $c_n(\delta) > 1$

$$\log G_n(x^n)^{1+\delta} \leq -\log Q_{G_n}(x^n) \leq \lceil -\log Q_{G_n}(x^n) \rceil = L_{G_n}(x^n),$$

which concludes the proof. ■

The following corollary to the above proposition follows immediately.

Corollary 3 *For any given $\delta > 0$ and $B \geq 1$, a guessing function G_n and its associated length function L_{G_n} satisfy*

$$\begin{aligned} \{L_{G_n}(x^n) - 1 - \log c_n(\delta) \geq B\} &\subseteq \{(1+\delta)\log G_n(x^n) \geq B\} \\ &\subseteq \{L_{G_n}(x^n) \geq B\}. \end{aligned}$$

□

Let G_n^* denote the optimal guessing strategy. The optimal exponential growth rate of guessing moments is defined as

$$E(\rho) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[G_n^*(X^n)^\rho\right] \tag{4}$$

when the limit exists. Define the growth rate of moment generating function for compression lengths to be

$$F(\rho) := \lim_{n\to\infty} \inf_{L_n} \frac{1}{n\rho} \log \mathbb{E}\left[2^{\rho L_n(X^n)}\right] \tag{5}$$

whenever the limit exists [5].

The following proposition establishes that the two limits above are the same and hence it suffices to study one of them, say the limiting exponential rate of growth of the moment generating function for compression lengths. Sufficient conditions in the finite alphabet case for the existence of this limiting exponent can be found in our prior work [6].

Proposition 4 *Let $\rho > 0$. Suppose that the $F(\rho)$ exists and the function F is continuous at ρ; then $E(\rho)$ exists and equals to $F(\rho)$. Conversely, suppose $E(\rho)$ exists and the function E is continuous at ρ; then $F(\rho)$ exists and equals to $E(\rho)$.*

Proof: First assume that $F(\rho)$ exists and let $\rho' = \rho/(1+\delta)$. For each $\epsilon > 0$ there then exists a length function $L_n^{'}$ such that the following sequence of inequalities holds:

$$\begin{aligned}
\inf_{L_n} \log \mathbb{E}\left[2^{\rho L_n(x^n)}\right] + \epsilon &\geq \log \mathbb{E}\left[2^{\rho L_n^{'}(x^n)}\right] \\
&\geq \log \mathbb{E}\left[G_{L_n^{'}}(x^n)^{\rho}\right] && (6)\\
&\geq \log \mathbb{E}\left[G_n^*(x^n)^{\rho}\right] && (7)\\
&\geq \log \mathbb{E}\left[2^{\rho' L_{G_n^*}(x^n)}\right] - \rho'(1+\log c_n(\delta)) && (8)\\
&\geq \inf_{L_n} \log \mathbb{E}\left[2^{\rho' L_n(x^n)}\right] - \rho'(1+\log c_n(\delta)) && (9)
\end{aligned}$$

In inequality (6), $G_{L_n^{'}}$ is the guessing function associated with $L_n^{'}$ and obtained by applying Proposition 1. Inequality (7) is obtained by noting that G_n^* is the optimal guessing function. In (8), $L_{G_n^*}$ is the length function associated with G_n^* and we applied Proposition 2. Finally, inequality (9) follows after taking infimum.

After normalising both sides of (7) by n, taking limit superior on both sides, and observing that $c_n(\delta)$ is finite, we have

$$F(\rho) = \limsup_{n\to\infty} \inf_{L_n} \frac{1}{n} \log \mathbb{E}\left[2^{\rho L_n(x^n)}\right] \geq \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[G_n^*(x^n)^{\rho}\right]. \quad (10)$$

Similarly, normalising both sides of (9) by n and taking limit inferior on both sides yields

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[G_n^*(x^n)^{\rho}\right] \geq \liminf_{n\to\infty} \inf_{L_n} \frac{1}{n} \log \mathbb{E}\left[2^{\rho' L_n(x^n)}\right] = F(\rho'). \quad (11)$$

Inequalities (10) and (11) and the assumption that

$$\lim_{\delta\to 0} F(\rho') = F(\rho)$$

show that $E(\rho)$ exists and equals $F(\rho)$.

To prove the converse, assume that $E(\rho)$ exists. Consider the following

chain of inequalities:

$$\begin{aligned}
\log \mathbb{E}\left[G_n^*(x^n)^\rho\right] &\geq \log \mathbb{E}\left[2^{\rho' L_{G_n^*}(x^n)}\right] - \rho'(1+\log c_n(\delta)) && (12)\\
&\geq \inf_{L_n} \log \mathbb{E}\left[2^{\rho' L_n(x^n)}\right] - \rho'(1+\log c_n(\delta)) && (13)\\
&\geq \log \mathbb{E}\left[2^{\rho' L_n^{'}(x^n)}\right] - \epsilon - \rho'(1+\log c_n(\delta)) && (14)\\
&\geq \log \mathbb{E}\left[G_{L_n^{'}}(x^n)^{\rho'}\right] - \epsilon - \rho'(1+\log c_n(\delta)) && (15)\\
&\geq \log \mathbb{E}\left[G_n^*(x^n)^{\rho^{'}}\right] - \epsilon - \rho^{'}(1+\log c_n(\delta)) && (16)
\end{aligned}$$

In inequality (12), $L_{G_n^*}$ is the length function associated with G_n^* and we used Proposition 2. In (14), ϵ is arbitrary positive number and $L_n^{'}$ is some length function depending on this ϵ; its existence is assured by the definition of the infimum. In (15), $G_{L_n^{'}}$ is the guessing function associated with $L_n^{'}$ and we used Proposition 1. Finally, (16) is obvious from the use of the optimal guessing strategy.

Normalising both sides of (13) by n, taking limit superior on both sides, we have

$$E(\rho) = \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[G_n^*(x^n)^\rho\right] \geq \limsup_{n\to\infty} \inf_{L_n} \frac{1}{n} \log \mathbb{E}\left[2^{\rho' L_n(x^n)}\right]. \quad (17)$$

Similarly, normalising both sides of (16) by n and taking limit inferior on both sides, we get

$$\liminf_{n\to\infty} \inf_{L_n} \frac{1}{n} \log \mathbb{E}\left[2^{\rho' L_n(x^n)}\right] \geq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[G_n^*(x^n)^{\rho'}\right] = E(\rho'). \quad (18)$$

From inequalities (17) and (18) and the continuity assumption of E in ρ, i.e.,

$$\lim_{\delta\to 0} E(\rho') = E(\rho),$$

we conclude that $F(\rho)$ exists and equals $E(\rho)$. ■

3 Guessing with distortion

We now consider the case when the goal is to guess within a distortion D of the actual realisation. Let us fix a distortion metric $d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+$. Recall that

Definition 3 *For a given distortion D and distortion measure $d(\cdot,\cdot)$, an ordered list $\mathcal{G}_n = \{y^n(1), y^n(2), \cdots\}$ is a D-admissible guessing strategy if*

$$\Pr\{d(X^n, y^n(j)) \leq nD, \text{ for some } j\} = 1.$$

□

Also recall that the D-admissible guessing list $\mathcal{G}_n$ induces a guessing function

$$G_n : \mathbb{X}^n \to \mathbb{N}.$$

(If $\mathcal{G}_n$ is not D-admissible, we set $G_n(x^n) = \infty$ for all x^n with $d(x^n, y^n) > nD$ for each $y^n \in \mathcal{G}_n$.

Definition 4 *A rate distortion code (RDC) is a triple (C_n, f_n, L_n) defined as follows. C_n is a subset of $\mathbb{X}^n$. The function $f_n : \mathbb{X}^n \to C_n$ is such that for each $x^n \in \mathbb{X}^n$ there exists a $y^n \in C_n$ satisfying $d(x^n, y^n) \leq nD$. The dependence on D is implicit and understood. $L_n : C_n \to \mathbb{N}$ is a length function that satisfies Kraft's inequality; $L_n(x^n)$ denotes the length of the encoding for an element $x^n \in C_n$.*

We now associate a D-admissible guessing strategy and a guessing function to an RDC.

Definition 5 *Given an RDC $\phi = (C_n, f_n, L_n)$, let $\mathcal{G}_n$ order the elements of C_n in the increasing order of the compression L_n lengths. This shall be the associated D-admissible guessing strategy. The induced guessing function is denoted by G_{L_n}. We also define the associated probability mass function (PMF) Q_{L_n} on C_n to be*

$$Q_{L_n}(y^n) = \frac{2^{-L_n(y^n)}}{\sum_{c^n \in C^n} 2^{-L_n(c^n)}}, \quad \forall y^n \in C_n.$$

□

The following proposition is similar to Proposition 1.

Proposition 5 *For a given rate distortion code (C_n, f_n, L_n) and $B \geq 1$, the associated guessing function G_{L_n} satisfies the following:*

$$\log G_{L_n}(x^n) \leq Q_{L_n}(f(x^n))^{-1} \leq L_n(f(x^n))$$
$$\{x^n : G_{L_n}(x^n) \geq 2^B\} \subseteq \{x^n : L(f(x^n)) \geq B\}.$$

□

The proof is obviously analogous to that of Proposition 1 and is therefore omitted.

We now associate an RDC to any D-admissible guessing strategy.

Definition 6 *Given a guessing function G_n induced by a D-admissible guessing strategy $\mathcal{G}_n$, and given a $\delta > 0$, let $C_n = \mathcal{G}_n$, let f_n be the function that maps x^n to the first element y^n in the ordered list $\mathcal{G}_n$ that satisfies $d(x^n, y^n) \leq nD$. Further, define a length function L_n on C_n as in definition 2. We say (C_n, f_n, L_n) is an RDC associated with the guessing function G_n.* □

The following proposition establishes the relationship between the associated quantities defined above. The proof is very similar to that of Proposition 2 and is therefore skipped.

Proposition 6 *Let G_n be a guessing function associated with a D-admissible guessing strategy. There exists an RDC (C_n, f_n, L_n) such that $\forall\ x^n \in \mathbb{X}^n$ and $\delta > 0$*

$$\frac{L_{G_n}(f_n(x^n)) - 1 - \log c_n(\delta)}{1+\delta} \leq \log G_n(x^n) \leq \frac{L_{G_n}(f(x^n))}{1+\delta}.$$

□

We now establish a relationship between the guessing exponent and the exponent of moment generating function for compression lengths, under the distortion setting. For $D \geq 0$ and $\rho > 0$, define the optimal exponential growth rate of guessing moments to be

$$\mathcal{E}(D, \rho) := \lim_{n\to\infty} \inf_{\mathcal{G}_n} \frac{1}{n} \log \mathbb{E}[G_n(X^n)^\rho], \tag{19}$$

whenever the limit exists, where the infimum is taken over all D-admissible guessing strategies $\mathcal{G}_n$ with G_n the associated guessing function. Similar to (5), define the exponent of the moment generating function for compression lengths (with distortion) as

$$\mathcal{F}(D, \rho) := \lim_{n\to\infty} \inf_{(c_n, f_n, L_n)} \frac{1}{n\rho} \log \mathbb{E}[2^{\rho L_n(X^n)}], \tag{20}$$

whenever the limit exists, where the infimum is taken over all RDC codes with distortion within D.

Under the above definitions, we now have the following result analogous to Proposition 4.

Proposition 7 *Let $D \geq 0$ and $\rho > 0$. Suppose that the $\mathcal{F}(D, \rho)$ exists and is continuous in ρ, then $\mathcal{E}(D, \rho)$ exists and equals to $\mathcal{F}(D, \rho)$. Conversely, suppose $\mathcal{E}(D, \rho)$ exists and is continuous in ρ, then $\mathcal{F}(D, \rho)$ exists and equals to $\mathcal{E}(D, \rho)$.* □

The proof of the above proposition is easy following the proof of Proposition 4.

We have thus established that the limiting guessing exponent, subject to distortion, and the problem of identifying the exponent of the moment generating function for compressed lengths, again subject to distortion, are identical.

Acknowledgements This work was supported by the Defence Research and Development Organisation, Ministry of Defence, Government of India, under the DRDO-IISc Programme on Advanced Research in Mathematical Engineering.

References

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

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

[3] J. L. Massey, "Guessing and entropy," in *Proc. 1994 IEEE International Symposium on Information Theory*, Trondheim, Norway, Jun. 1994, p. 204.

[4] R. Sundaresan, "Guessing based on length functions," in *Proceedings of the Conference on Managing Complexity in a Distributed World, MCDES*, Bangalore, India, May 2008; *also available as* DRDO-IISc Programme in Mathematical Engineering Technical Report No. TR-PME-2007-02, Feb. 2007.
http://pal.ece.iisc.ernet.in/PAM/tech_rep07/TR-PME-2007-02.pdf.

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

[6] M. K. Hanawal and R. Sundaresan, "Guessing revisited: A large deviations approach," in *Proc. National Conference on Communications*, Guwahati, India, Jan 2009.