# Guessing Revisited: A Large Deviations Approach

Manjesh Kumar and Rajesh Sundaresan
Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore 560012, India
Email: {manjesh, rajeshs}@ece.iisc.ernet.in

***Abstract*— The problem of guessing a random string is revisited and some prior results on guessing exponents are re-derived using the theory of large deviations. It is shown that if the sequence of distributions of the information spectrum satisfies the large deviation property with a certain rate function, then the limiting guessing exponent exists and is a scalar multiple of the Legendre-Fenchel dual of the rate function. Example applications re-deriving prior results are also given.**

## I. Introduction

Let $X^n = (X_1, \cdots, X_n)$ denote $n$ letters of a process where each letter is drawn from a finite set $\mathbb{X}$ with joint probability mass function (pmf) $(P_n(x^n) : x^n \in \mathbb{X}^n)$. Let $x^n$ be a realisation and suppose that we wish to guess this realisation by asking questions of the form "Is $X^n = x^n$?", stepping through the elements of $\mathbb{X}^n$ until the answer is "Yes". We wish to do this using the minimum number of expected guesses. There are several applications that motivate this problem. Consider cipher systems employed in digital television or DVDs to block unauthorised access to special features. The ciphers used are amenable to such exhaustive guessing attacks and it is of interest to quantify the effort needed by an attacker.

Massey [1] observed that the expected number of guesses is minimised by guessing in the decreasing order of $P_n$-probabilities. Define the *guessing function* $G_n^* : \mathbb{X}^n \to \{1, 2, \cdots, |\mathbb{X}|^n\}$ to be one such optimal guessing order [1]. $G_n^*(x^n) = g$ implies that $x^n$ is the $g$th guess. Massey's question was to characterise $\mathbb{E}[G_n^*(X^n)]$. Arikan [2] considered the more general problem of identifying the growth of $\mathbb{E}[G_n^*(X^n)^\rho]$ as a function of $n$ for an independent and identically distributed (iid) source with marginal pmf $P_1$ and $\rho > 0$. He showed that the growth is exponential in $n$; limiting exponent

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

exists and equals $\rho H_\alpha(P_1)$ with $\alpha = 1/(1+\rho)$, where $H_\alpha(P_n)$ is the Rényi entropy of order $\alpha$ for the pmf $P_n$, given by

$$\frac{1}{1-\alpha} \log \left( \sum_{x^n \in \mathbb{X}^n} P_n(x^n)^\alpha \right), \ \alpha \neq 1. \tag{2}$$

Malone and Sullivan [3] showed that the limiting exponent $E(\rho)$ of an irreducible Markov chain exists and equals the logarithm of the *Perron-Frobenius eigenvalue* of a matrix formed by raising each element of the transition probability matrix to the power $\alpha$. From their proof, one obtains the more general result that the limiting exponent exists for any source if the Rényi entropy *rate* of order $\alpha$,

$$\lim_{n\to\infty} n^{-1} H_\alpha(P_n), \tag{3}$$

exists for $\alpha = 1/(1+\rho)$. Pfister and Sullivan [4] showed the existence of (1) for a class of stationary probability measures where the probability of finite-length strings are approximately determined by letter combinations. For such a class, they showed that the guessing exponent has a variational characterisation (see (4) later). For unifilar sources Sundaresan [5] obtained a simplification of this variational characterisation using a direct approach and the method of types.

In this paper, we give a different proof of Malone & Sullivan's implicit result in [3] that the limiting exponent exists if and only if the limiting Rényi entropy rate exists. Our proof exploits a connection between guessing and compression highlighted by Sundaresan [5]. A simple argument then leads to the following useful result: if the sequence of distributions of the *information spectrum* $(1/n)\log(1/P_n(X^n))$ (see Han [6]) satisfies the *large deviation property*, then the limiting exponent exists. This is useful because several existing large deviations results can be readily applied. Our approach generalises all prior results on guessing (without side information and key-rate constraints).

## II. Main results

We begin with some words on notation. Let $\mathcal{M}(\mathbb{X}^n)$ denote the set of pmfs on $\mathbb{X}^n$. The Shannon entropy for a $P_n \in \mathcal{M}(\mathbb{X}^n)$ is

$$H(P_n) = -\sum_{x^n \in \mathbb{X}^n} P_n(x^n) \log P_n(x^n)$$

and the Rényi entropy of order $\alpha \neq 1$ is (2). The Kullback-Leibler divergence or relative entropy between two pmfs $Q_n$ and $P_n$ is

$$D(Q_n \parallel P_n) = \begin{cases} \displaystyle\sum_{x^n \in \mathbb{X}^n} Q_n(x^n) \log \frac{Q_n(x^n)}{P_n(x^n)}, & \text{if } Q_n \ll P_n, \\ \infty, & \text{otherwise,} \end{cases}$$

where $Q_n \ll P_n$ means $Q_n$ is absolutely continuous with respect to $P_n$. By a source, we mean a sequence of pmfs $(P_n : n \in \mathbb{N})$ where $P_n \in \mathcal{M}(\mathbb{X}^n)$ and $\mathbb{N}$ is the set of natural numbers. Recall the definitions of limiting guessing exponent

[1] If there are several sequences with the same probability of occurrence, they may be guessed in any order without affecting the expected number of guesses.

in (1) and Rényi entropy rate in (3) when the limits exist. $G_n^*$ is an optimal guessing function for a pmf $P_n \in \mathcal{M}(\mathbb{X}^n)$. Our proof route will use results from source compression. We therefore define a length function $L_n : \mathbb{X}^n \to \mathbb{N}$ to be one which satisfies Kraft's inequality

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

Given a length function, it is well-known that there exists a source code such that the compression length of any string $x^n$ is $L_n(x^n)$. For a $\rho > 0$, we define $\alpha = 1/(1+\rho)$ and $\beta = \rho/(1+\rho)$, and use these consistently throughout this paper.

Our first contribution is a proof of the following implicit result of Malone & Sullivan [3]. The proof is given in Section IV-A.

*Proposition 1:* Let $\rho > 0$. For a source $(P_n : n \in \mathbb{N})$, $E(\rho)$ exists if and only if the Rényi entropy rate exists. Furthermore, $E(\rho)/\rho$ equals the Rényi entropy rate. □

The question now boils down to the existence of the limit in the definition of Rényi entropy rate. The theory of large deviations immediately yields a sufficient condition. We begin with a definition.

*Definition 1 (Large deviation property):* A sequence $(\nu_n : n \in \mathbb{N})$ of probability measures on $\mathbb{R}$ satisfies the *large deviation property (LDP)* with rate function $I : \mathbb{R} \to [0, \infty]$ if the following conditions hold:

- $I$ is lower semicontinuous on $\mathbb{R}$;
- $I$ has compact level sets;
- $\limsup_{n\to\infty} n^{-1} \log \nu_n\{K\} \leq -\inf_{t\in K} I(t)$ for each closed subset $K$ of $\mathbb{R}$;
- $\liminf_{n\to\infty} n^{-1} \log \nu_n\{G\} \geq -\inf_{t\in G} I(t)$ for each open set $G$ of $\mathbb{R}$.

□

Several commonly encountered sources satisfy the LDP with known and well-studied rate functions. We describe some of these in the examples treated subsequently.

Let $\nu_n$ denote the distribution of the information spectrum given by the real-valued random variable $-n^{-1} \log P_n(X^n)$.

*Proposition 2:* Let the sequence of distributions $(\nu_n : n \in \mathbb{N})$ of the information spectrum satisfy the LDP with rate function $I$. Then the limiting Rényi entropy rate of order $1/(1+\rho)$ exists for all $\rho > 0$ and equals

$$\beta^{-1} \sup_{t\in\mathbb{R}} \{\beta t - I(t)\},$$

where $\beta = \rho/(1+\rho)$. Consequently, the limiting guessing exponent exists and equals

$$(1+\rho) \sup_{t\in\mathbb{R}} \{\beta t - I(t)\}.$$

□

The function $I^*(\beta) := \sup_{t\in\mathbb{R}} \{\beta t - I(t)\}$ is the Legendre-Fenchel dual of the rate function $I$. As we will see in Section IV, the proofs of the aforementioned results provide a ready connection between guessing, compression, and large deviations. Before giving the proofs, we show how known prior results can be obtained using the large deviations approach in the following examples.

## III. Examples

*Example 1 (An iid source):* This example was first studied by Arikan [2]. Recall that an iid source is one for which $P_n(x^n) = \prod_{i=1}^n P_1(x_i)$, where $P_1$ is the marginal of $X_1$. It is then clear that the information spectrum can be written as a sample mean of iid random variables

$$-n^{-1} \log P_n(X^n) = -n^{-1} \sum_{i=1}^n \log P_1(X_i).$$

The sequence $(\nu_n : n \in \mathbb{N})$ of distributions of information spectrum therefore satisfies the LDP with rate function given by the Legendre-Fenchel dual of the cumulant of the random variable $-\log P_1(X_1)$ (see for example [7, Thm. II.4.1] or [6, eqn. (1.9.66-67)]):

$$\begin{aligned} \log \mathbb{E}\left[\exp\left\{\beta(-\log P_1(X_1))\right\}\right] &= \log\left(\sum_{x\in\mathbb{X}} P_1(x)^\alpha\right) \\ &= (1-\alpha) H_\alpha(P_1). \end{aligned}$$

The Legendre-Fenchel dual of the rate function is therefore the cumulant itself ([7, Thm. VI.4.1.e]). An application of Proposition 2 yields that $(1+\rho)$ times this cumulant, given by $\rho H_\alpha(P_1)$, is the guessing exponent. We thus recover Arikan's result [2].

The rate function $I$ can also be obtained using the so-called *contraction principle* [8, Th. 6.12] as follows. Consider a mapping that takes $x^n$ to its empirical pmf in $\mathcal{M}(\mathbb{X})$. Empirical pmf is then a random variable. The distribution of $X^n$ induces a distribution on $\mathcal{M}(\mathbb{X})$. The sequence of these distributions of empirical pmfs, indexed by $n$, satisfies the *level-2* LDP[2] with rate function $I_{P_1}^{(2)}(\cdot) = D(\cdot \parallel P_1)$. See for example [7, Thm II.4.3]. The contraction principle provides a formula for $I$ in terms of $D(\cdot \parallel P_1)$ as follows [7, Thm II.5.1]. Let

$$\theta(t) := \left\{Q \in \mathcal{M}(\mathbb{X}) : H(Q) + D(Q \parallel P_1) = t\right\}.$$

Then

$$I(t) = \inf\{I_{P_1}^{(2)}(Q) : Q \in \theta(t)\}.$$

Using this, we can write

$$\begin{aligned} I^*(\beta) &= \sup_{t\in\mathbb{R}} \left\{\beta t - \inf_{Q\in\theta(t)} D(Q \parallel P_1)\right\} \\ &= \sup_{t\in\mathbb{R}} \sup_{Q\in\theta(t)} \left\{\beta t - D(Q \parallel P_1)\right\} \\ &= \sup_{Q\in\mathcal{M}(\mathbb{X})} \left\{\beta(H(Q) + D(Q \parallel P_1)) - D(Q \parallel P_1)\right\} \\ &= (1+\rho)^{-1} \sup_{Q\in\mathcal{M}(\mathbb{X})} \left\{\rho H(Q) - D(Q \parallel P_1)\right\}, \end{aligned}$$

[2]Level-1 refers to sequence of distributions (indexed by $n$) of the sample means, level-2 refers to sample histograms, and level-3 to sample paths.

thus yielding

$$E(\rho) = \sup_{Q \in \mathcal{M}(\mathbb{X})} \Big\{ \rho H(Q) - D(Q \parallel P_1) \Big\}. \tag{4}$$

This formula extends to more general sources, as is seen in the next few examples. □

*Example 2 (Markov source):* This example was studied by Malone & Sullivan [3]. Consider an irreducible Markov chain taking values on $\mathbb{X}$ with transition probability matrix $\pi$. Our goal is to calculate $E(\rho)$ defined by (1) for this source.

Let $\mathcal{M}_s(\mathbb{X}^2)$ denote the set of *stationary* pmfs defined by

$$\mathcal{M}_s\left(\mathbb{X}^2\right) = \Big\{ Q \in \mathcal{M}\left(\mathbb{X}^2\right) : \sum_{x_1 \in \mathbb{X}} Q(x_1, x) = \sum_{x_2 \in \mathbb{X}} Q(x, x_2) \forall x \in \mathbb{X} \Big\}.$$

Denote the common marginal by $q$ and let

$$\eta(\cdot \mid x_1) := \begin{cases} Q(x_1, \cdot)/q(x_1), & \text{if } q(x_1) \neq 0, \\ 1/|\mathbb{X}|, & \text{otherwise.} \end{cases}$$

We may then denote $Q = q \times \eta$ where $q$ is the distribution of $X_1$ and $\eta$ the conditional distribution of $X_2$ given $X_1$. Then, the empirical pmf random variable satisfies the level-2 LDP with rate function $I_\pi^{(2)}(Q)$, given by [9]

$$\begin{aligned} I_\pi^{(2)}(Q) &= D(\eta \parallel \pi \mid q) \\ &:= \sum_{x_1 \in \mathbb{X}} q(x_1) D(\eta(\cdot \mid x_1) \parallel \pi(\cdot \mid x_1)). \end{aligned}$$

The contraction principle then yields that the sequence of distributions of information spectrum satisfies the LDP with rate function $I$ given by

$$I(t) = \inf\{ I_\pi^{(2)}(Q) : Q \in \theta(t) \}.$$

where $\theta(t) \subset \mathcal{M}_s(\mathbb{X}^2)$ is defined by

$$\theta(t) = \left\{ Q \in \mathcal{M}_s(\mathbb{X}^2) : \sum_{x_1, x_2} Q(x_1, x_2) \log \frac{1}{\pi(x_2 | x_1)} = t \right\}.$$

By Proposition 1 the limiting guessing exponent exists. Perron-Frobenius theory (Seneta [10, Ch. 1], see also [11, pp.60-61]) yields the cumulant directly as $\log \lambda(\beta)$ where $\lambda(\beta)$ is unique largest eigenvalue (Perron-Frobenius eigenvalue) of a matrix formed by raising each element of matrix $\pi$ to the power $\alpha$. (Recall that $\alpha = 1/(1+\rho)$ and $\beta = \rho/(1+\rho)$). Thus $E(\rho) = (1+\rho) \log \lambda(\beta)$, and we recover the result of Malone & Sullivan [3]. It is useful to note that the steps that led to (4) hold in the Markov case (with appropriate changes to entropy and divergence terms) and we may write

$$E(\rho) = \sup_{Q \in \mathcal{M}_s(\mathbb{X}^2)} \Big\{ \rho H(\eta \mid q) - D(\eta \parallel \pi \mid q) \Big\}, \tag{5}$$

where $H(\eta \mid q)$ is the conditional entropy of $X_2$ given $X_1$ under the joint distribution $Q$, i.e.,

$$H(\eta \mid q) := -\sum_{x \in \mathbb{X}} q(x) H(\eta(\cdot \mid x)).$$

□

*Example 3 (Unifilar source):* This example was studied by Sundaresan in [5]. A unifilar source is a generalisation of the Markov case in Example 2. Let $\mathbb{X}$ denote the alphabet set as before. In addition, let $\mathbb{S}$ denote a set of finite states. Fix an initial state $s_0$ and let the joint probability of observing $(x^n, s^n)$ be

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

where $\pi(x_i, s_i \mid s_{i-1})$ is the joint probability of $(x_i, s_i)$ given the previous state $s_{i-1}$. The dependence of $P_n$ on $s_0$ is understood. Furthermore, assume that $\pi(x_i, s_i \mid s_{i-1})$ is such that $s_i = \phi(s_{i-1}, x_i)$, a deterministic function. Such a source is called a unifilar source.

$P_{S,X}(s_{i-1}, x_i)$ and $\phi$ completely specify the process: the initial state $S_0$ is random with distribution the marginal of $S$ in $P_{S,X}$, the rest being specified by $P_{X|S}(x_i \mid s_{i-1})$ and $\phi$. Example 2 is a unifilar source with $\mathbb{S} = \mathbb{X}$, $\phi(s_{i-1}, x_i) = x_i$, and $P_{S,X} = q \times \pi$ where $q$ is the stationary distribution of the Markov chain.

Let $\mathcal{M}_s(\mathbb{S} \times \mathbb{X})$ denote the set of joint measures on the indicated space so that the resulting process $(S_n : n \geq 0)$ is a stationary and irreducible Markov chain. Let a $Q \in \mathcal{M}_s(\mathbb{S} \times \mathbb{X})$ be written as $Q = q \times \eta$. For a $t \in \mathbb{R}$ let

$$\theta(t) := \left\{ Q \in \mathcal{M}_s(\mathbb{S} \times \mathbb{X}) : \sum_{(s,x)} Q(s,x) \log \frac{1}{\pi(x \mid s)} = t \right\}.$$

Then the sequence of distributions of information spectrum $-n^{-1} \log P_n(X^n)$ satisfies the LDP ([6, eqn. (1.9.30)]) with rate function given (once again via contraction principle) by

$$I(t) = \inf\{ D(\eta \parallel \pi \mid q) : Q \in \theta(t) \}.$$

The limiting exponent therefore exists. Following the same procedure that led to (4) in the iid case and (5) for a Markov chain, we get

$$E(\rho) = \sup_{Q \in \mathcal{M}_s(\mathbb{S} \times \mathbb{X})} \Big\{ \rho H(\eta \mid q) - D(\eta \parallel \pi \mid q) \Big\}, \tag{6}$$

where $H(\eta \mid q)$ and $D(\eta \parallel \pi \mid q)$ are analogously defined, and the result of Sundaresan [5] is recovered. □

*Example 4 (A class of stationary sources):* Pfister & Sullivan [4] consider a class of stationary sources with distribution $P \in \mathcal{M}\left(\mathbb{X}^{\mathbb{N}}\right)$ that satisfy two hypotheses (H1 and H2 of [4, Sec. II-B]). They prove that $E(\rho)$ exists, and provide a variational characterisation analogous to (6), i.e.,

$$E(\rho) = \sup_{Q \in \mathcal{M}_s^P} \Big\{ \rho \overline{H}(Q) - \overline{D}(Q \parallel P) \Big\}, \tag{7}$$

where $\overline{H}(Q)$ is the Shannon entropy rate and with $P_n$ and $Q_n$ restrictions of $P$ and $Q$ to $n$ letters

$$\overline{D}(Q \parallel P) = \lim_{n \to \infty} n^{-1} \sum_{x^n} Q_n(x^n) \log \frac{Q_n(x^n)}{P_n(x^n)}.$$

$\mathcal{M}_s^P$ is the set of stationary sources that satisfy $Q_n \ll P_n$ for all $n$.

En route to their result they show that the sequence of distributions of the empirical process satisfies the level-3 LDP with rate function $I_P^{(3)}(Q) = \overline{D}(Q \parallel P)$ given above. In order to prove (7) using our recipe, the contraction principle is first applied to argue that an approximation to the information spectrum sequence satisfies the level-1 LDP with the contracted rate function. A result [8, Th. 6.14] is then be used to show that the information spectrum sequence too satisfies the LDP with the same rate function. Proposition 2 immediately yields that the limit $E(\rho)$ exists. Finally, the Legendre-Fenchel dual of the rate function is computed similar to the technique used to obtain (4), (5), and (6), thus yielding (7). The details are technical, but quite straightforward, and therefore omitted. □

*Example 5 (Mixed source):* Consider a mixture of two iid sources with letters from $\mathbb{X}$. We may write

$$P_n(x^n) = \lambda \prod_{i=1}^{n} R(x_i) + (1-\lambda) \prod_{i=1}^{n} S(x_i)$$

where $\lambda \in (0,1)$ with $R, S \in \mathcal{M}(\mathbb{X})$ the two marginal pmfs that define the iid components of the mixture. It is easy to see directly that the guessing exponent is the maximum of the guessing exponent for the two component sources. We next verify this using Proposition 2.

The sequence of distributions of the information spectrum satisfies the LDP with rate function given as follows (see Han [6, eqn. (1.9.41)]). Define

$$\begin{aligned}\theta_1 &= \Big\{Q \in \mathcal{M}(\mathbb{X}) : D(Q \parallel S) - D(Q \parallel R) \geq 0\Big\},\\ \theta_2 &= \Big\{Q \in \mathcal{M}(\mathbb{X}) : D(Q \parallel S) - D(Q \parallel R) \leq 0\Big\},\end{aligned}$$

and for $t \in \mathbb{R}$

$$\begin{aligned}A_t &= \theta_1 \cap \Big\{Q \in \mathcal{M}(\mathbb{X}) : H(Q) + D(Q \parallel R) = t\Big\}\\ B_t &= \theta_2 \cap \Big\{Q \in \mathcal{M}(\mathbb{X}) : H(Q) + D(Q \parallel S) = t\Big\}.\end{aligned}$$

The rate function (via the contraction principle) is given by

$$I(t) = \min\left\{\inf_{Q \in A_t} D(Q \parallel R), \inf_{Q \in B_t} D(Q \parallel S)\right\}.$$

From Proposition 2 we conclude that the limiting guessing exponent exists. $I^*(\beta)$ is then

$$\begin{aligned}&\sup_{t \in \mathbb{R}} \left\{\beta t - \min\left\{\inf_{Q \in A_t} D(Q \parallel R), \inf_{Q \in B_t} D(Q \parallel S)\right\}\right\}\\ &= \max\left\{\sup_{t \in \mathbb{R}} \sup_{Q \in A_t} \Big\{\beta t - D(Q \parallel R)\Big\},\right.\\ &\qquad \left.\sup_{t \in \mathbb{R}} \sup_{Q \in B_t} \Big\{\beta t - D(Q \parallel S)\Big\}\right\}\\ &= \max\left\{\sup_{Q \in \theta_1} \Big\{\beta H(Q) - (1-\beta) D(Q \parallel R)\Big\},\right.\\ &\qquad \left.\sup_{Q \in \theta_2} \Big\{\beta H(Q) - (1-\beta) D(Q \parallel S)\Big\}\right\}\\ &= (1+\rho)^{-1} \max\left\{\sup_{Q} \Big\{\rho H(Q) - D(Q \parallel R)\Big\},\right.\\ &\qquad \left.\sup_{Q} \Big\{\rho H(Q) - D(Q \parallel S)\Big\}\right\}\\ &= (1+\rho)^{-1} \max\Big\{\rho H_\alpha(R), \rho H_\alpha(S)\Big\},\end{aligned}$$

yielding

$$E(\rho) = \max\Big\{\rho H_\alpha(R), \rho H_\alpha(S)\Big\}.$$

□

## IV. Proofs

We now provide proofs of Propositions 1 and 2. The approach taken is via compression, but with exponentiated costs.

### A. Proof of Proposition 1

We first transform the guessing problem into a compression problem with an exponentially weighted cost structure. A result from Sundaresan [5, Prop. 6] implies that the limit in (4) exists if and only if the following problem originally considered by Campbell [12] has a limit:

$$\lim_{n \to \infty} \inf_{L_n} \frac{1}{n} \log \mathbb{E}\Big[\exp\{\rho L_n(X^n)\}\Big], \tag{8}$$

where the infimum is taken over all length functions $L_n : \mathbb{X}^n \to \mathbb{N}$. Moreover, the two limits are equal. This result arises from [5, Prop. 6] because the difference between two quantities as a function of $n$ decays as $O((\log n)/n)$. It is therefore sufficient to show that the limit in (8) for Campbell's coding problem exists if and only if the Rényi entropy rate exists, with the former $\rho$ times the latter.

Fix $n$. In the rest of the proof, we use the notation $\mathbb{E}_{P_n}[\cdot]$ for expectation with respect to distribution $P_n$. The length function can be thought of as a bounded (continuous) function from $\mathbb{X}^n$ to $\mathbb{R}$ and therefore our interest is in the logarithm of its moment generating function of $\rho$, the cumulant. The cumulant associated with a bounded continuous function (here $L_n$) has a variational characterisation [13, Prop. 1.4.2] as the following Legendre-Fenchel dual of the Kullback-Leibler divergence, i.e.,

$$\begin{aligned}&\log \mathbb{E}_{P_n}\Big[\exp\{\rho L_n(X^n)\}\Big]\\ &\quad = \sup_{Q_n \in \mathcal{M}(\mathbb{X}^n)} \Big\{\rho \mathbb{E}_{Q_n}[L_n(X^n)] - D(Q_n \parallel P_n)\Big\}. \quad (9)\end{aligned}$$

Taking infimum on both sides over all length functions, we arrive at the following chain of inequalities:

$$\begin{aligned}&\inf_{L_n} \log \mathbb{E}_{P_n}\Big[\exp\{\rho L_n(X^n)\}\Big] \quad (10)\\ &\quad = \inf_{L_n} \sup_{Q_n \in \mathcal{M}(\mathbb{X}^n)} \Big\{\mathbb{E}_{Q_n}[\rho L_n(X^n)] - D(Q_n \parallel P_n)\Big\}\\ &\quad = \sup_{Q_n \in \mathcal{M}(\mathbb{X}^n)} \inf_{L_n} \Big\{\mathbb{E}_{Q_n}[\rho L_n(X^n)] - D(Q_n \parallel P_n)\Big\} + \Theta(1) \quad (11)\\ &\quad = \sup_{Q_n \in \mathcal{M}(\mathbb{X}^n)} \Big\{\rho H_n(Q_n) - D(Q_n \parallel P_n)\Big\} + \Theta(1) \quad (12)\\ &\quad = \rho H_{\frac{1}{1+\rho}}(P_n) + \Theta(1). \quad (13)\end{aligned}$$

Equation (11) follows because (i) the mapping

$$(L_n, Q_n) \mapsto \mathbb{E}_{Q_n}[\rho L_n(X^n)] - D(Q_n \parallel P_n)$$

is a concave function of $Q_n$, (ii) for fixed $Q_n$ and for any two length functions $L_n^1$ and $L_n^2$, for any $\lambda \in [0,1]$, the function $L_n = \lceil \lambda L_n^1 + (1-\lambda) L_n^2 \rceil$ is also a length function and

$$\mathbb{E}_{Q_n}[L_n] = \lambda \mathbb{E}_{Q_n}[L_n^1] + (1-\lambda)\mathbb{E}_{Q_n}[L_n^2] + \Theta(1).$$

(iii) $\mathcal{M}(\mathbb{X}^n)$ is compact and convex, and therefore the infimum and supremum may be interchanged upon an application of a version of Ky Fan's minimax result [14]. This yields a compression problem, the infimum over $L_n$ of expected lengths with respect to a distribution $Q_n$. The answer is the well-known Shannon entropy $H(Q_n)$ to within 1 bit, and (12) follows. Lastly, (13) is a well-known identity which may also be obtained directly by writing the supremum term in (12) as

$$(1+\rho) \sup_{Q_n \in \mathcal{M}(\mathbb{X}^n)} \left\{ \mathbb{E}_{Q_n}\left[ -\left(\frac{\rho}{1+\rho}\right) \log P_n(X^n) \right] - D(Q_n \parallel P_n) \right\}$$

and then applying (9) with $-(\rho/(1+\rho)\log P_n(X^n))$ in place of $\rho L_n(X^n)$ to get the scaled Rényi entropy.

Normalise both (10) and (13) by $n$ and let $n \to \infty$ to deduce that (8) exists if and only if the limiting normalised Rényi entropy rate exists. This concludes the proof.

### B. Proof of Proposition 2

This is a straightforward application of Varadhan's theorem [15] on asymptotics of integrals. Recall that $\nu_n$ is the distribution of the information spectrum $n^{-1}\log P_n(X^n)$. Define $F(t) = \beta t$. Since the $(\nu_n : n \in \mathbb{N})$ sequence satisfies the LDP with rate function $I$, Varadhan's theorem (see Ellis [7, Th. II.7.1.b]) states that if

$$\lim_{M\to\infty} \limsup_{n\to\infty} \frac{1}{n} \log \int_{t \geq \frac{M}{\beta}} \exp\{n\beta t\}\, d\nu_n(t) = -\infty \quad (14)$$

then the limit

$$\lim_{n\to\infty} \frac{1}{n} \log \int_{\mathbb{R}} \exp\{n\beta t\}\, \nu_n(dt) = \sup_{t\in\mathbb{R}} \{\beta t - I(t)\} \quad (15)$$

holds. The integral on the left side in (15) can be simplified by defining the finite cardinality set

$$A_n = \{-n^{-1}\log P_n(x^n) : \forall x^n \in \mathbb{X}^n\} \subset \mathbb{R}$$

and by observing that

$$\begin{aligned}
&\int_{\mathbb{R}} \exp\{n\beta t\}\nu_n(dt) \\
&\quad = \sum_{t\in A_n} \exp\{n\beta t\} \sum_{x^n : P_n(x^n) = \exp\{-nt\}} P_n(x^n) \\
&\quad = \sum_{x^n} P_n(x^n)^{1-\beta} \\
&\quad = \sum_{x^n} P_n(x^n)^{\frac{1}{1+\rho}} = \exp\left\{\beta H_{1/(1+\rho)}(P_n)\right\}.
\end{aligned}$$

Take logarithms, normalise by $n$, take limits, and apply (15) to get the desired result. It therefore remains to prove (14).

The event $\{t \geq \frac{M}{\beta}\}$ occurs if and only if $\{P_n(x^n) \leq \exp\{\frac{-nM}{\beta}\}$. The integral in (14) can therefore be written as

$$\begin{aligned}
&\sum_{t\in A_n, t\geq \frac{M}{\beta}} \sum_{x^n : P_n(x^n) = \exp\{-nt\}} \exp\{n\beta t\} P_n(x^n) \\
&\quad = \sum_{x^n : P_n(x^n) \leq \exp\{\frac{-nM}{\beta}\}} P_n(x^n)^{\frac{1}{1+\rho}} \\
&\quad \leq |\mathbb{X}|^n \cdot \exp\left\{\frac{-nM}{\beta(1+\rho)}\right\}.
\end{aligned}$$

The sequence in $n$ on the left side of (14) is then

$$\log|\mathbb{X}| - \frac{M}{\beta(1+\rho)},$$

a constant sequence. Take the limit as $M \to \infty$ to verify (14). This concludes the proof.

## 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, and by the University Grants Commission under Grant Part (2B) UGC-CAS-(Ph.IV).

## References

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

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

[3] D. Mallone and W. G. Sullivan, "Guesswork and entropy," *IEEE Trans. Inform. Theory*, vol. 50, no. 4, pp. 525–526, Mar. 2004.

[4] E. Pfister and W. G. Sullivan, "Rényi entropy, guesswork moments, and large deviations," *IEEE Trans. Inform. Theory*, vol. 50, no. 11, pp. 2794–2800, Nov. 2004.

[5] 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.

[6] T. S. Han, *Information-Spectrum Methods in Information Theory*. New York: Springer-Verlog, 2003.

[7] R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics*. New York: Springer-Verlag, 1985.

[8] ——, "The theory of large deviations and applications to statistical mechanics," Oct. 2006, Lectures for the International Seminar on Extreme Events in Complex Dynamics, Dresden, Germany.

[9] S. Natarajan, "Large deviations, hypotheses testing, and source coding for finite Markov chains," *IEEE Trans. Inform. Theory*, vol. 31, no. 3, pp. 360–365, May 1985.

[10] E. Seneta, *Non-negative Matrices: An Introduction to Theory and Applications*. London: George Allen & Unwin Ltd., 1973.

[11] F. den Hollander, *Large Deviations*. Rhode Island: American Mathematical Society, 2003.

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

[13] P. Dupuis and R.S.Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*. New York: John Wiley & Sons, 1997.

[14] I. Joó and L. L. Stachó, "A note on Ky Fan's minimax theorem," *Acta Math. Acad. Sci. Hungar.*, vol. 39, pp. 401–407, 1982.

[15] S. R. S. Varadhan, "Asymptotic probabilities and differential equations," *Comm. Pure Appl. Math.*, vol. 19, pp. 261–286, 1966.