Joint Source-Channel Coding and Guessing with Application to Sequential Decoding

Erdal Arikan, *Senior Member, IEEE,* and Neri Merhav, *Senior Member, IEEE*

***Abstract*— We extend our earlier work on guessing subject to distortion to the joint source-channel coding context. We consider a system in which there is a source connected to a destination via a channel and the goal is to reconstruct the source output at the destination within a prescribed distortion level with respect to (w.r.t.) some distortion measure. The decoder is a *guessing* decoder in the sense that it is allowed to generate successive estimates of the source output until the distortion criterion is met. The problem is to design the encoder and the decoder so as to minimize the average number of estimates until successful reconstruction. We derive estimates on nonnegative moments of the number of guesses, which are asymptotically tight as the length of the source block goes to infinity. Using the close relationship between guessing and sequential decoding, we give a tight lower bound to the complexity of sequential decoding in joint source-channel coding systems, complementing earlier works by Koshelev and Hellman. Another topic explored here is the probability of error for list decoders with exponential list sizes for joint source-channel coding systems, for which we obtain tight bounds as well. It is noteworthy that optimal performance w.r.t. the performance measures considered here can be achieved in a manner that separates source coding and channel coding.**

***Index Terms*— Guessing, joint source-channel coding, list decoding, rate distortion, sequential decoding.**

I. Introduction

CONSIDER the joint source-channel coding system in Fig. 1 where a source is connected to a destination via a channel and the goal is to reconstruct the source output at the destination within a prescribed per-letter distortion D with respect to (w.r.t.) some distortion measure d. The source generates a random vector $\boldsymbol{U} = (U_1, \cdots, U_N)$ which is encoded into a channel input vector $\boldsymbol{X} = (X_1, \cdots, X_K)$ and sent over the channel. The decoder observes the channel output $\boldsymbol{Y} = (Y_1, \cdots, Y_K)$ and generates successive "guesses" (reconstruction vectors), $\hat{\boldsymbol{U}}_1$, $\hat{\boldsymbol{U}}_2$, and so on, until a guess $\hat{\boldsymbol{U}}_i$ is produced such that $d(\boldsymbol{U}, \hat{\boldsymbol{U}}_i) \leq ND$. At each step, the decoder is informed by a genie whether the present guess $\hat{\boldsymbol{U}}_j$ satisfies $d(\boldsymbol{U}, \hat{\boldsymbol{U}}_j) \leq ND$, but receives no other information about the value of $d(\boldsymbol{U}, \hat{\boldsymbol{U}}_j)$. We shall refer to this type of decoder as a *guessing decoder* and denote the number of guesses until successful reconstruction (which is a random variable) by $G_N(\boldsymbol{U}|\boldsymbol{Y})$ in the sequel.

The main aim of this paper is to determine the best attainable performance of the above system under the performance goal of minimizing the average decoding complexity, as measured by the moments $\boldsymbol{E}[G_N(\boldsymbol{U}|\boldsymbol{Y})^\rho]$, $\rho > 0$. We also study the closely related problem of finding tight bounds on the probability $\Pr[G_N(\boldsymbol{U}|\boldsymbol{Y}) > e^{NL}]$ that an exponentially large number of guesses will be required until successful reconstruction. We have two motivations for studying these problems. First, the present model extends the basic search model treated in [2], where the problem was to guess the output of a source in the absence of any coded information supplied via a channel. Second, and on the more applied side, the guessing decoder model is suitable for studying the computational complexity of *sequential decoding*, which is a decoding algorithm of practical interest. Indeed, through this method, we are able to solve a previously open problem relating to the *cutoff rate* of sequential decoding in joint source-channel coding systems.

In the remainder of this introduction, we shall outline the results of this paper more precisely. We begin by pointing out the relationship of the present joint source-channel guessing framework to earlier work on guessing. In [2], we considered a guessing problem which is equivalent to the rather special case of the joint source-channel guessing problem where there is no channel (i.e., the decoder receives no coded information about $\boldsymbol{U}$ before guessing begins). There, the number of guesses was denoted by $G_N(\boldsymbol{U})$ and an asymptotic quantity called the *guessing exponent* was defined as

$$E(D, \rho) = \lim_{N\to\infty} \frac{1}{N} \min_{G_N} \ln \boldsymbol{E}[G_N(\boldsymbol{U})^\rho] \tag{1}$$

for $\rho \geq 0$, provided that the limit exists. It was shown that, for any discrete memoryless source (DMS) P and additive (single-letter) distortion measure d

$$E(D, \rho) = \max_{Q} [\rho R(D, Q) - D(Q||P)] \tag{2}$$

where Q ranges over all probability mass functions (PMF's) on the source alphabet, $R(D, Q)$ is the rate-distortion function of a source with PMF Q, and $D(Q||P)$ is the relative entropy function.

The asymptotic quantity of interest in this paper is the *joint source-channel guessing exponent* defined, whenever the limit

Manuscript received March 11, 1997; revised February 28, 1998. The work of N. Merhav was supported in part by the Israel Science Foundation administered by the Israel Academy of Sciences and Humanities. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Ulm, Germany, June–July 1997.

E. Arikan is with the Electrical–Electronics Engineering Department, Bilkent University, 06533 Ankara, Turkey (e-mail: arikan@ee.bilkent.edu.tr).

N. Merhav was with Hewlett-Packard Laboratories, Palo Alto, CA, USA. He is now with the Department of Electrical Engineering and HP-ISC, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: merhav@ee.technion.ac.il).

Publisher Item Identifier S 0018-9448(98)04933-5.

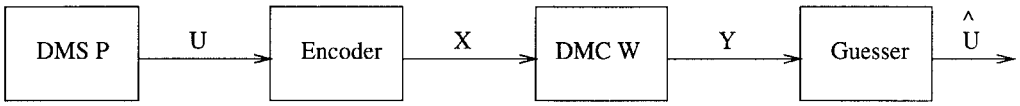

Fig. 1. Joint source-channel coding and guessing system.

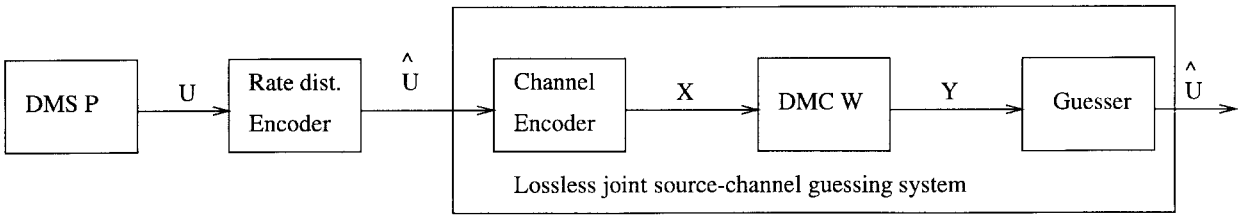

Fig. 2. Separation of source coding and channel coding.

exists, as

$$E_{\rm sc}(D,\rho)=\lim_{N\to\infty}\frac{1}{N}\min_{e_N,G_N}\ln \boldsymbol{E}[G_N(\boldsymbol{U}|\boldsymbol{Y})^\rho] \quad (3)$$

where e_N denotes an encoding function that maps source sequences of length N into channel sequences of length K. In letting $N\to\infty$, we set $K=\lceil \lambda N\rceil$ for some constant λ that represents the ratio of the channel signaling rate to source symbol rate. The main result of this paper is that for any DMS P, discrete memoryless channel (DMC) W, and single-letter distortion measure d, the joint source-channel guessing exponent has a single-letter form given by

$$E_{\rm sc}(D,\rho)=[E(D,\rho)-\lambda E_0(\rho)]^+ \quad (4)$$

where $E_0(\rho)$ is the Gallager function for W [9] and

$$[x]^+\stackrel{\Delta}{=}\max\{0,x\}.$$

Thus the exponent $E_{\rm sc}(D,\rho)$ is determined by the difference of a source-related term, $E(D,\rho)$, and a channel-related term, $\lambda E_0(\rho)$; the channel term $\lambda E_0(\rho)$ represents the potential benefit of having a channel. This result indicates that the ρth moment of $G_N(\boldsymbol{U}|\boldsymbol{Y})$ for any such system must grow exponentially in the source blocklength N if $E(D,\rho)>\lambda E_0(\rho)$. Conversely, for $E(D,\rho)<\lambda E_0(\rho)$, the ρth moment can be kept from growing exponentially in N by suitable design of the encoder and the decoder.

We prove (4) in Sections III and IV. The proof exhibits a separation principle for such systems in the sense that an optimal encoder can be built as a two-stage device: the first stage maps the source output vector to a rate-distortion codeword, independently of the channel characteristics; while the second stage encodes the rate-distortion codeword into a channel codeword, independently of the source statistics. The guesser then essentially aims to recover the rate-distortion codeword in a lossless manner (Fig. 2).

The joint source-channel guessing problem that we consider here is also closely related to another guessing problem considered in [2], namely, guessing with *uncoded* side-information (as opposed to *coded* side-information of the present context). In the case of uncoded side-information, the pair $(\boldsymbol{U},\boldsymbol{Y})$ has a given joint PMF that is not subject to design in any manner. In the case of coded side-information, however, the joint PMF of $(\boldsymbol{U},\boldsymbol{Y})$ is influenced by the choice of the encoder e_N; thus it is subject to design, at least partially. In [2], a single-letter expression was given for the exponent $E_{U|Y}(D,\rho)$ that applies to guessing with (uncoded) side-information. It is immediate from the definitions that

$$E_{\rm sc}(D,\rho)=\lim_{N\to\infty}\{N^{-1}\min_{e_N}E_{\boldsymbol{U}|\boldsymbol{Y}}(D,\rho)\} \quad (5)$$

where the exponent on the right-hand side applies to an ensemble $(\boldsymbol{U},\boldsymbol{Y})$ such that $\boldsymbol{U}$ is a source block of length N, $\boldsymbol{Y}$ is a channel output block of length $K=\lceil\lambda N\rceil$, and the joint PMF $P(\boldsymbol{u},\boldsymbol{y})$ is given by the product of the source probability $P(\boldsymbol{u})$ and the channel transition probability $W(\boldsymbol{y}|e_N(\boldsymbol{u}))$. Unfortunately, the term on the right-hand side of (5) is not in a single-letter form. The main accomplishment in this paper is to give a single-letter form for $E_{\rm sc}(D,\rho)$.

Next, we explain the relationship of guessing to coding, specifically to list decoding and sequential decoding, and outline our results in this regard. Recall that a list decoder generates a fixed number, $\ell\geq 1$, of guesses (estimates) and a decoding failure is said to occur if none of the guesses approximates the source output within the desired distortion level. On the other hand, a guessing decoder is fully determined by the sequence of guesses $\mathcal{G}_N(\boldsymbol{Y})=\{\hat{\boldsymbol{U}}_1,\hat{\boldsymbol{U}}_2,\cdots\}$ that it would generate if at each stage of guessing the desired distortion criterion remained unmet. So, a guessing decoder may be viewed conceptually as a list decoder, whose output is the possibly infinite list $\mathcal{G}_N(\boldsymbol{Y})$. A list-$\ell$ decoder can be obtained from a guessing decoder by truncating the list $\mathcal{G}_N(\boldsymbol{Y})$ to its first ℓ elements. For $\ell=1$, we have ordinary decoding and the usual performance criterion is to have the average distortion satisfy $\boldsymbol{E}[d(\boldsymbol{U},\hat{\boldsymbol{U}}_1)]\leq ND$. This is the original setting for the joint source-channel coding problem and Shannon's joint source-channel coding theorem (see, e.g., [9, Theorem 9.2.2, p. 449]) addresses the conditions under which this requirement can be met. For $\ell\geq 1$, a common performance criterion is the probability $\Pr[G_N(\boldsymbol{U}|\boldsymbol{Y})>\ell]$ that none of the first ℓ guesses meet the desired distortion threshold. The best attainable performance under this criterion has been studied by Csiszár [4] for $\ell=1$ as $N\to\infty$; however, the exact asymptotic performance remains unknown.

In this paper, we are interested in the performance of list decoders with exponential list sizes, $\ell=e^{NL}$, $L>0$, for

which we obtain an exact asymptotic result. Specifically, we define the *source-channel list-error exponent* as

$$F_{\rm sc}(L, D) = \lim_{N\to\infty} \sup_{e_N, G_N} -\frac{1}{N} \ln \Pr[G_N(\boldsymbol{U}|\boldsymbol{Y}) > e^{NL}] \tag{6}$$

whenever the limit exists. (In taking the limit, we set $K = \lceil \lambda N \rceil$.) In Section V, we prove that for any DMS P, DMC W, additive distortion measure d, and $L > 0$

$$F_{\rm sc}(L, D) = \min_{R\geq L} [F(R, D) + \lambda E_{\rm sp}[(R-L)/\lambda]] \tag{7}$$

where $F(R, D)$ is Marton's source-coding exponent [15], and $E_{\rm sp}(\cdot)$ is the sphere-packing exponent [9, p. 157] for W.

List decoders with exponential list sizes are not practical; however, bounds on the probability of error for such decoders may have applications to the analysis of concatenated and hierarchical coding systems. In fact, an immediate application of these results is given in Section VI, where we obtain a lower bound to the distribution of computation in sequential decoding.

As stated before, one of our main motivations for studying joint source-channel guessing systems is for its suitability as a model for sequential decoding. We now summarize our results in this regard. Sequential decoding is a decoding algorithm for tree codes invented by Wozencraft [18]. The use of sequential decoding in joint source-channel coding systems was proposed by Koshelev [14] and Hellman [12]. The attractive feature of sequential decoding, in this context, is the possibility of generating a D-admissible reconstruction sequence, with an *average* computational complexity that grows only linearly with N, the length of the source sequence. To be more precise, let C_N denote the amount of computation by the sequential decoder to reconstruct the first N source symbols within distortion level ND. Then, C_N is a random variable, which depends on the level of channel noise, as well as the specific tree code that is used and also the source and channel parameters. For practical applications, it is desirable to have $\boldsymbol{E}[C_N]/N$, the average complexity per reconstructed source digit, bounded independently of N. Koshelev [14] studied this problem for the lossless case $(D = 0)$ and gave a sufficient condition; in our notation, he showed that if $E(0, 1) < \lambda E_0(1)$ then it is possible to have $\boldsymbol{E}[C_N]/N$ bounded (independently of N). Our interest in this paper is in converse results, i.e., necessary conditions for the possibility of having a bounded $\boldsymbol{E}[C_N]/N$.

In Section VI, we point out a close connection between guessing and sequential decoding, and prove, as a simple corollary to (4), that for any DMS P, DMC W, and additive distortion measure d, $\boldsymbol{E}[C_N^\rho]$ must grow exponentially with N (thus $\boldsymbol{E}[C_N^\rho]/N$ cannot be bounded) if

$$E(D, \rho) > \lambda E_0(\rho). \tag{8}$$

For the special case $D = 0$ and $\rho = 1$, this result complements Koshelev's result, showing that his sufficient condition is also necessary. This result also generalizes the converse result in [1], where lossless guessing $(D = 0)$ was considered for an equiprobable message ensemble. These issues are discussed further in Section VI.

The remainder of this paper is organized as follows. In Section II, we define the notation and give a more formal definition of the guessing problem. The single-letter form (4) is proved in Section III for the lossless case $D = 0$, and in Section IV for the lossy case $D > 0$. In Section V, we prove the single-letter form (7) for the source-channel list-error exponent. In Section VI, we apply the results about guessing to sequential decoding. Section VII concludes the paper by summarizing the results and stating some open problems. We also discuss in Section VII the possibility of using a stochastic encoder in place of e_N and show that there is no advantage to be gained.

II. Problem Statement: Notation, and Definitions

We assume, unless otherwise specified, that the system in Fig. 1 has the following properties. The source is a DMS with a PMF P over a finite alphabet $\mathcal{U}$. The channel is a DMC with finite input alphabet $\mathcal{X}$, finite output alphabet $\mathcal{Y}$, and transition probability matrix W. The reconstruction alphabet $\hat{\mathcal{U}}$ is finite as well. The distortion measure d is a single-letter measure, i.e., it is a function $d\colon \mathcal{U}\times\hat{\mathcal{U}} \to [0, \infty)$, which is extended to $\mathcal{U}^N \times \hat{\mathcal{U}}^N$ by setting $d(\boldsymbol{u}, \hat{\boldsymbol{u}}) = \sum_{n=1}^N d(u_n, \hat{u}_n)$, $\boldsymbol{u} = (u_1, \cdots, u_N)$, $\hat{\boldsymbol{u}} = (\hat{u}_1, \cdots, \hat{u}_N)$. Also, for each $u \in \mathcal{U}$, there exists some $\hat{u} \in \hat{\mathcal{U}}$ such that $d(u, \hat{u}) = 0$.

Throughout, scalar random variables will be denoted by capital letters and their realizations by the respective lower case letters. Random vectors will be denoted by boldface capital letters and their realizations by lower case boldface letters. Thus e.g., $\boldsymbol{U} = (U_1, \cdots, U_N)$ will denote a random vector, while $\boldsymbol{u} = (u_1, \cdots, u_N)$ a realization of $\boldsymbol{U}$. PMF's of scalar random variables will be denoted by upper case letters, e.g., P, P', Q, S. For random vectors, we will denote the PMF's by upper case letters indexed by the length of the vector, e.g., P_N, P'_N, etc. We will omit the index N for product-form PMF's; e.g., we write $P(\boldsymbol{u})$ instead of $P_N(\boldsymbol{u})$ when P_N is a product-form PMF. The probability of an event A w.r.t. a probability measure P' will be denoted by $P'(A)$. When the underlying probability measure is specified unambiguously, we also use a notation such as $\Pr(\boldsymbol{u}, \boldsymbol{y})$ to denote the joint PMF of $\boldsymbol{U}$ and $\boldsymbol{Y}$, or $\Pr(\hat{\boldsymbol{u}}, A)$ to denote the probability of joint occurrence of $\hat{\boldsymbol{U}} = \hat{\boldsymbol{u}}$ and an event A. The expectation operation is denoted by $\boldsymbol{E}[\cdot]$.

For a given vector $\boldsymbol{x} \in \mathcal{A}^N$, the empirical PMF is defined as $Q_{\boldsymbol{x}} = \{Q_{\boldsymbol{x}}(x);\, x \in \mathcal{A}\}$, where $Q_{\boldsymbol{x}}(x) = N_{\boldsymbol{x}}(x)/N$, $N_{\boldsymbol{x}}(x)$ being the number of occurrences of the letter x in the vector $\boldsymbol{x}$. The type class $T_{\boldsymbol{x}}$ of $\boldsymbol{x}$ is the set of all vectors $\boldsymbol{x'} \in \mathcal{A}^N$ such that $Q_{\boldsymbol{x'}} = Q_{\boldsymbol{x}}$. When we need to attribute a type class to a certain PMF Q rather than to a vector, we shall use the notation T_Q.

In the same manner, for sequence pairs $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{A}^N \times \mathcal{B}^N$, the joint empirical PMF is the matrix $Q_{\boldsymbol{xy}} = \{Q_{\boldsymbol{xy}}(x, y);\, x \in \mathcal{A}, y \in \mathcal{B}\}$, where $Q_{\boldsymbol{xy}}(x, y) = N_{\boldsymbol{xy}}(x, y)/N$, $N_{\boldsymbol{xy}}(x, y)$ being the number of joint occurrences of $x_i = x$ and $y_i = y$. For a stochastic matrix $\{V(y|x)\colon x \in \mathcal{A}, y \in \mathcal{B}\}$, the V-shell $T_V(\boldsymbol{x})$ of a sequence $\boldsymbol{x} \in \mathcal{A}^N$ is the set of sequences $\boldsymbol{y} \in \mathcal{B}^N$ such that $Q_{\boldsymbol{xy}}(x, y) = Q_{\boldsymbol{x}}(x)V(y|x)$ for all x and y.

Next, we recall the definitions of some information-theoretic functions that appear in the paper. For a PMF Q over an alphabet $\mathcal{A}$, the entropy of Q is defined as

$$H(Q) = -\sum_{x\in\mathcal{A}} Q(x)\ln Q(x) \tag{9}$$

and its Rényi entropy of order $\alpha > 0$, $\alpha \neq 1$, as [16]

$$H_\alpha(Q) = \frac{1}{1-\alpha}\ln\sum_{x\in\mathcal{A}} Q(x)^\alpha. \tag{10}$$

Sometimes we write $H(X)$ and $H_\alpha(X)$ to denote the entropy functions for a random variable X. For two PMF's Q and Q' on a common alphabet $\mathcal{A}$, the relative entropy function is

$$D(Q\|Q') = \sum_{x\in\mathcal{A}} Q(x)\ln[Q(x)/Q'(x)]. \tag{11}$$

For a stochastic matrix $\{V(y|x);\ x\in\mathcal{A}, y\in\mathcal{B}\}$, and a PMF Q on $\mathcal{A}$, the mutual information function is defined as

$$I(Q, V) = \sum_{y\in\mathcal{B}}\sum_{x\in\mathcal{A}} Q(x)V(y|x)\ln[V(y|x)/V'(y)] \tag{12}$$

where

$$V'(y) = \sum_{x\in\mathcal{A}} Q(x)V(y|x).$$

The rate-distortion function $R(D, Q)$ for a DMS Q on $\mathcal{U}$, w.r.t. a single-letter distortion measure d on $\mathcal{U}\times\hat{\mathcal{U}}$, is defined as

$$R(D, Q) = \min_V I(Q, V) \tag{13}$$

where the minimum is taken over all stochastic matrices V such that

$$\sum_{u\in\mathcal{U}}\sum_{\hat{u}\in\hat{\mathcal{U}}} Q(u)V(\hat{u}|u)\,d(u, \hat{u}) \le D. \tag{14}$$

Marton's source-coding exponent $F(R, D)$ for a DMS P is given by

$$F(R, D) = \min_{Q:\, R(D,Q)\ge R} D(Q\|P). \tag{15}$$

For a DMC W, we recall the following definitions. The channel capacity is defined as $C = \max_S I(S, W)$, where the maximum is over all PMF's on the channel input alphabet. Gallager's auxiliary functions are defined as

$$E_0(\rho, S) = -\ln\sum_y\left[\sum_x S(x)W(y|x)^{1/(1+\rho)}\right]^{1+\rho} \tag{16}$$

for any PMF S on the channel input alphabet and any $\rho \ge 0$; and

$$E_0(\rho) = \max_S E_0(\rho, S). \tag{17}$$

The sphere-packing exponent function is defined as

$$E_{\rm sp}(R) = \sup_{\rho\ge 0}[E_0(\rho) - \rho R]. \tag{18}$$

Next, we define the guessing problem more precisely. For $\hat{\boldsymbol{u}} \in \hat{\mathcal{U}}^N$, let

$$B(\hat{\boldsymbol{u}}, D) \stackrel{\Delta}{=} \{\boldsymbol{u}\in\mathcal{U}^N : d(\boldsymbol{u}, \hat{\boldsymbol{u}}) \le ND\}.$$

Definition 1: A D-admissible guessing strategy for the set of sequences $\mathcal{U}^N$ is an ordered list $\mathcal{G}_N = \{\hat{\boldsymbol{u}}_1, \hat{\boldsymbol{u}}_2, \cdots\}$ of vectors in $\hat{\mathcal{U}}^N$ such that

$$\bigcup_i B(\hat{\boldsymbol{u}}_i, D) = \mathcal{U}^N. \tag{19}$$

In other words, $\mathcal{G}_N$ is an ordered covering of the set $\mathcal{U}^N$ by the "D-spheres" $B(\hat{\boldsymbol{u}}_i, D)$.

Definition 2: The guessing function $G_N(\cdot)$ induced by a D-admissible guessing strategy $\mathcal{G}_N$, is the function that maps each $\boldsymbol{u}\in\mathcal{U}^N$ into a positive integer, which is the index j of the first guessing word $\hat{\boldsymbol{u}}_j \in \mathcal{G}_N$ such that $d(\boldsymbol{u}, \hat{\boldsymbol{u}}_j) \le ND$.

We now extend these definitions to the case where some *side-information* vector $\boldsymbol{y}\in\mathcal{Y}^K$ is provided.

Definition 3: A D-admissible guessing strategy for $\mathcal{U}^N$ with side-information space $\mathcal{Y}^K$ is a collection $\{\mathcal{G}_N(\boldsymbol{y})\}$ such that for each $\boldsymbol{y}\in\mathcal{Y}^K$, $\mathcal{G}_N(\boldsymbol{y})$ is a guessing strategy for $\mathcal{U}^N$ in the sense of Definition 1.

Definition 4: The guessing function $G_N(\cdot|\cdot)$ induced by a D-admissible guessing strategy with side-information, $\{\mathcal{G}_N(\boldsymbol{y})\}$, is the function that maps each $\boldsymbol{u}\in\mathcal{U}^N$ and $\boldsymbol{y}\in\mathcal{Y}^K$ into a positive integer, $G_N(\boldsymbol{u}|\boldsymbol{y})$, which is the index j of the first guessing word $\hat{\boldsymbol{u}}_j\in\mathcal{G}_N(\boldsymbol{y})$ such that $d(\boldsymbol{u}, \hat{\boldsymbol{u}}_j) \le ND$.

We shall omit the subscript N from the guessing functions and simply write $G(\cdot|\cdot)$, etc., when there is no room for ambiguity.

Notice that the above definitions make no reference to a probability measure. In the context of joint source-channel guessing, we regard $\mathcal{U}^N$ as the sample space for the source vector $\boldsymbol{U}$, and $\mathcal{Y}^K$ as that for the channel output vector $\boldsymbol{Y}$. The joint PMF for $\boldsymbol{U}, \boldsymbol{Y}$ is given by $\Pr(\boldsymbol{u}, \boldsymbol{y}) = P(\boldsymbol{u})W(\boldsymbol{y}|e_N(\boldsymbol{u}))$ where $e_N: \mathcal{U}^N \to \mathcal{X}^K$ is the encoding function. The decoder observes the channel output realization $\boldsymbol{y}$ and employs a guessing strategy $\mathcal{G}_N(\boldsymbol{y})$ to find a D-admissible reconstruction of the source realization $\boldsymbol{u}$. Under such a strategy $G_N(\boldsymbol{U}|\boldsymbol{Y})$ equals the random number of guesses until a D-admissible reconstruction $\hat{\boldsymbol{U}}$ of $\boldsymbol{U}$ is found.

Throughout, $o(N)$ will denote a positive quantity that goes to zero as N goes to infinity.

III. The Lossless Source-Channel Guessing Exponent

In this section, we consider the source-channel guessing problem for the lossless case $D = 0$, i.e., the case where the reconstruction alphabet is the same as the source alphabet and we desire exact reconstruction of the source output. This case is of interest in its own right. Also, the general lossy guessing problem $(D > 0)$ is reduced to the lossless one by an argument given in the next section.

For lossless guessing, a guessing strategy $\{\mathcal{G}(\boldsymbol{y})\}$ that generates its guesses in decreasing order of *a posteriori* probabilities $\Pr[\boldsymbol{U} = \boldsymbol{u}|\boldsymbol{Y} = \boldsymbol{y}]$ achieves the minimum possible value for the moments $\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho]$ of the associated guessing function. This is easily seen by simply writing

$$\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho] = \sum_{\boldsymbol{y}}\Pr[\boldsymbol{Y}=\boldsymbol{y}]\sum_{\boldsymbol{u}}\Pr[\boldsymbol{U}=\boldsymbol{u}|\boldsymbol{Y}=\boldsymbol{u}]G(\boldsymbol{u}|\boldsymbol{y})^\rho. \tag{20}$$

Note that such an optimal ordering of guesses depends on the encoder e_N since the joint PMF is given by

$$\Pr[\boldsymbol{U}=\boldsymbol{u}, \boldsymbol{Y}=\boldsymbol{y}] = P(\boldsymbol{u})W(\boldsymbol{y}|e_N(\boldsymbol{u})).$$

The fact that optimal guessing strategy is known for the lossless case facilitates the characterization of the associated guessing exponent, denoted by $E_{\rm sc}(\rho)$. Our main result in this section is the following singe-letter expression for this exponent.

Theorem 1: For any DMS P and DMC W, the lossless joint source-channel guessing exponent is given by

$$E_{\rm sc}(\rho) = [\rho H_{1/(1+\rho)}(P) - \lambda E_0(\rho)]^+. \tag{21}$$

Since the proof of (21) is rather lengthy, it is deferred to the Appendix. In fact, in the Appendix we prove a stronger form of Theorem 1, which applies to sources with memory as well. Since the proofs for lossy guessing require the treatment of sources with memory (as the coded channel input may not be memoryless), we state this stronger result for future reference as the following proposition.

Proposition 1: For any discrete source with a possibly nonmemoryless PMF P_N for the first N source letters, and any fixed $\rho \geq 0$, there exists a lossless guessing function $G(\boldsymbol{U}|\boldsymbol{Y})$ such that

$$\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho] \leq c(\rho)\exp\{N[\rho H_{1/(1+\rho)}(P_N)/N - \lambda E_0(\rho)]^+\} \tag{22}$$

where $c(\rho)$ is a constant, independent of the source and channel, and of the length N. Conversely, for any guessing function $G(\boldsymbol{U}|\boldsymbol{Y})$ and $\rho \geq 0$

$$\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho] \geq \exp\{N[\rho H_{1/(1+\rho)}(P_N)/N - \lambda E_0(\rho) - o(N)]^+\}. \tag{23}$$

Proposition 1 implies, in particular, that for a memoryless source P, the ρth moment of $G(\boldsymbol{U}|\boldsymbol{Y})$ can be kept below the constant $c(\rho)$ for all $N \geq 1$ if

$$H_{1/(1+\rho)}(P) < \lambda E_0(\rho)/\rho. \tag{24}$$

(This cannot be deduced from (21) since it leaves open the possibility of subexponential growth of the moment.) Conversely, it follows directly from (21) that if

$$H_{1/(1+\rho)}(P) > \lambda E_0(\rho)/\rho \tag{25}$$

then $E_{\rm sc}(\rho) > 0$ and the ρth moment of $G(\boldsymbol{U}|\boldsymbol{Y})$ must go to infinity exponentially in N.

Since $H_{1/(1+\rho)}(P)$ is increasing and $E_0(\rho)/\rho$ is decreasing as functions of $\rho > 0$, the term $H_{1/(1+\rho)}(P) - \lambda E_0(\rho)/\rho$ is minimized in the limit as $\rho \to 0$ (this is proved formally below), with the limiting value $H(P) - \lambda C$, where C is the capacity of W. Thus we conclude that if $H(P) > \lambda C$, then $\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho]$ must go to infinity exponentially in N for all $\rho > 0$. Conversely, if $H(P) < \lambda C$, then there exists a $\rho > 0$ such that, for any given N, it is possible to have $\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho] \leq c(\rho)$ by a suitable choice of the encoder and the guessing strategy.

It is interesting that the conditions $H(P) < \lambda C$ and $H(P) > \lambda C$ are also the conditions for the validity of the direct and converse parts, respectively, of Shannon's joint source-channel coding theorem for the lossless case [3, p. 216]. This suggests an underlying strong relationship between the problems of i) being able to keep $\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho]$ from growing exponentially in N, for some $\rho > 0$, and ii) being able to make the probability of error $\Pr[G(\boldsymbol{U}|\boldsymbol{Y}) > 1]$ arbitrarily small as $N \to \infty$. However, we have found no simple argument that would explain why the conditions for the two problems are identical. We propose this as a topic for further consideration.

We end this section by discussing monotonicity and convexity properties for the function $E_{\rm sc}(\rho)$. It is clear from the definition that $E_{\rm sc}(\rho)$ must be a nondecreasing function of $\rho \geq 0$. This property, and further properties of $E_{\rm sc}(\rho)$, can be obtained analytically by considering the form (21). For this, we refer to Lemma 1 (see the Appendix), which states that, for any fixed PMF S

$$f(\rho, S) \triangleq \rho H_{1/(1+\rho)}(P) - \lambda E_0(\rho, S)$$

is a convex function, which is strictly increasing in the range of $\rho > 0$ where $f(\rho, S) > 0$. We have

$$E_{\rm sc}(\rho) = [\min_S f(\rho, S)]^+ = \min_S [f(\rho, S)]^+.$$

Since the minimum of a family of increasing functions is increasing, it follows that $E_{\rm sc}(\rho)$ is increasing in the range where it is positive.

As for convexity, $E_{\rm sc}(\rho)$ is convex whenever $E_0(\rho) = \min_S E_0(\rho, S)$ is concave; this is true in particular for those channels where the minimum is achieved by the same S for all $\rho \geq 0$, such as the binary-symmetric channel. There are channels, however, for which $E_0(\rho)$ is not concave [9], and hence it is possible to construct examples for which $E_{\rm sc}(\rho)$ is not convex. (For example, take P as the uniform distribution on a binary alphabet so that $\rho H_{1/(1+\rho)}(P) = \rho \ln(2)$. Let $E_0(\rho)$ be nonconcave. Then, for λ large enough, $E_{\rm sc}(\rho)$ will be nonconvex.)

IV. The Lossy Source-Channel Guessing Exponent

We are now in a position to prove the main result of this paper.

Theorem 2: For any DMS P, DMC W, and single-letter distortion measure d, the joint source-channel guessing exponent $E_{\rm sc}(D, \rho)$ has a single-letter form given by

$$E_{\rm sc}(D, \rho) = [E(D, \rho) - \lambda E_0(\rho)]^+. \tag{26}$$

Proof:

Direct Part: We need to show

$$E_{\rm sc}(D, \rho) \leq [E(D, \rho) - \lambda E_0(\rho)]^+.$$

To obtain an upper bound on the minimum attainable $\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho]$, we consider a two-stage source-channel coding scheme (Fig. 2). In the first stage, the source output $\boldsymbol{U}$ is encoded into a rate-distortion codeword $\hat{\boldsymbol{U}}$ such that

$d(\boldsymbol{U}, \hat{\boldsymbol{U}}) \leq ND$. In the second stage, a joint source-channel guessing scheme is employed, aiming at lossless recovery of $\hat{\boldsymbol{U}}$. The details are as follows.

The encoding of $\boldsymbol{U}$ into a channel input block $\boldsymbol{X}$ is dependent on the type of $\boldsymbol{U}$. Let $f_Q : T_Q \rightarrow \mathcal{C}_Q$ be a rate-distortion encoder for the type class $T_Q \subset \mathcal{U}^N$ such that $d(\boldsymbol{u}, f_Q(\boldsymbol{u})) \leq ND$ for each $\boldsymbol{u} \in T_Q$ and the codebook $\mathcal{C}_Q$ has size $e^{N(R(D,Q)+o(N))}$. Such an encoder exists by the type-covering lemma [5, p. 150]. Let $g_Q : \mathcal{C}_Q \rightarrow \mathcal{X}^K$ denote a channel encoder that maps the codebook $\mathcal{C}_Q$ into channel codewords. The two-stage encoder first checks the type of $\boldsymbol{U}$, and if $\boldsymbol{U} \in T_Q$, then the encoding functions f_Q and g_Q are applied to generate the channel input block $\boldsymbol{X} = g_Q(f_Q(\boldsymbol{U}))$.

The guesser in the system does not know in advance the type of $\boldsymbol{U}$. To overcome this difficulty, we employ a D-admissible guessing strategy $\{\mathcal{G}(\boldsymbol{y});\ \boldsymbol{y} \in \mathcal{Y}^K\}$ for $\mathcal{U}^N$ which interlaces the guesses by a family of D-admissible guessing strategies $\{\mathcal{G}_Q(\boldsymbol{y}); \boldsymbol{y} \in \mathcal{Y}^K\}$ for $\mathcal{U}^N$, indexed by types Q over $\mathcal{U}^N$. To be precise, let $Q_1, \cdots, Q_v$ be an enumeration of the types. For any fixed $\boldsymbol{y} \in \mathcal{Y}^K$, the interlaced guessing strategy $\{\mathcal{G}(\boldsymbol{y})\}$ generates its guesses in rounds. In the first round, the first guesses by $\mathcal{G}_{Q_i}(\boldsymbol{y})$, $i = 1, \cdots, v$, are generated, respectively; in the second round, the second guesses are generated, and so on. (If at some round, there are no more guesses by some $\mathcal{G}_Q(\boldsymbol{y})$, dummy guesses are inserted.) Let $G(\boldsymbol{u}|\boldsymbol{y})$, $G_Q(\boldsymbol{u}|\boldsymbol{y})$ be the guessing functions for $\mathcal{G}(\boldsymbol{y})$, $\mathcal{G}_Q(\boldsymbol{y})$, respectively. Due to interlacing, we have $G(\boldsymbol{u}|\boldsymbol{y}) \leq vG_Q(\boldsymbol{u}|\boldsymbol{y})$ for all $\boldsymbol{u}$, $\boldsymbol{y}$, and Q, hence

$$\begin{aligned}
&\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho] \\
&\quad = \sum_Q \Pr[\boldsymbol{U} \in T_Q]\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho|\boldsymbol{U} \in T_Q] && (27)\\
&\quad \leq \sum_Q \Pr[\boldsymbol{U} \in T_Q]v^\rho \boldsymbol{E}[G_Q(\boldsymbol{U}|\boldsymbol{Y})^\rho|\boldsymbol{U} \in T_Q] && (28)\\
&\quad \leq v^{\rho+1} \max_Q\{P(T_Q)\boldsymbol{E}[G_Q(\boldsymbol{U}|\boldsymbol{Y})^\rho|\boldsymbol{U} \in T_Q]\}. && (29)
\end{aligned}$$

Next, we specify $\mathcal{G}_Q(\boldsymbol{y})$ so that it is an "efficient" guesser when $\boldsymbol{U} \in T_Q$. For this, we suppose that the first $|\mathcal{C}_Q|$ guesses by $\mathcal{G}_Q(\boldsymbol{y})$ consist of an enumeration of the elements of $\mathcal{C}_Q$ in descending order of the conditional probabilities $\Pr[\hat{\boldsymbol{U}} = \hat{\boldsymbol{u}}|\boldsymbol{U} \in T_Q, \boldsymbol{Y} = \boldsymbol{y}]$; the remaining guesses are immaterial so long as they are chosen to ensure the validity of the hypothesis that $\mathcal{G}_Q(\boldsymbol{y})$ is D-admissible for $\mathcal{U}^N$. Observe that $\mathcal{G}_Q(\boldsymbol{y})$ is also a lossless guessing strategy for $\mathcal{C}_Q$; furthermore, due to the way it has been specified, it is optimal as a lossless guessing strategy for $\mathcal{C}_Q$, in the sense of minimizing the conditional moments $\boldsymbol{E}[G_Q(\hat{\boldsymbol{U}}|\boldsymbol{Y})^\rho|\boldsymbol{U} \in T_Q]$, for all $\rho \geq 0$. (It is important to note that $G_Q(\hat{\boldsymbol{U}}|\boldsymbol{Y})$ denotes the guessing function associated with $\mathcal{G}_Q(\boldsymbol{y})$, when the latter is regarded as a lossless guessing strategy for $\mathcal{C}_Q$. Whereas, $G_Q(\boldsymbol{U}|\boldsymbol{Y})$ denotes the guessing function when $\mathcal{G}_Q(\boldsymbol{y})$ is regarded as a D-admissible guessing strategy for $\mathcal{U}^N$.)

Now, we observe that

$$G_Q(\boldsymbol{u}|\boldsymbol{y}) \leq G_Q(f_Q(\boldsymbol{u})|\boldsymbol{y}), \qquad \text{for all } \boldsymbol{u} \in T_Q \tag{30}$$

where we may have strict inequality if $d(\boldsymbol{u}, \hat{\boldsymbol{u}}') \leq ND$ for some $\hat{\boldsymbol{u}}' \in \mathcal{C}_Q$ such that

$$G_Q(\hat{\boldsymbol{u}}'|\boldsymbol{y}) < G_Q(f_Q(\boldsymbol{u})|\boldsymbol{y})$$

(i.e., when $\boldsymbol{u}$ falls in the D-sphere of a codeword $\hat{\boldsymbol{u}}'$ that precedes $f_Q(\boldsymbol{u})$ in the order they are generated by $\mathcal{G}_Q(\boldsymbol{y})$). Taking expectations of both sides of (30) w.r.t. the conditional probability measure $\Pr[\boldsymbol{U} = \boldsymbol{u}, \hat{\boldsymbol{U}} = \hat{\boldsymbol{u}}, \boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{U} \in T_Q]$ (note that this conditional PMF equals zero unless $\hat{\boldsymbol{u}} = f_Q(\boldsymbol{u})$), we obtain

$$\boldsymbol{E}[G_Q(\boldsymbol{U}|\boldsymbol{Y})^\rho|\boldsymbol{U} \in T_Q] \leq \boldsymbol{E}[G_Q(\hat{\boldsymbol{U}}|\boldsymbol{Y})^\rho|\boldsymbol{U} \in T_Q]. \tag{31}$$

By Proposition 1, we know that the channel encoder g_Q can be chosen so that

$$\begin{aligned}
&\boldsymbol{E}[G_Q(\hat{\boldsymbol{U}}|\boldsymbol{Y})^\rho|\boldsymbol{U} \in T_Q] \\
&\qquad \leq c(\rho)\ \exp\{N[\rho H_{1/(1+\rho)}(P_N)/N - \lambda E_0(\rho)]^+\} && (32)
\end{aligned}$$

where P_N is the conditional PMF of $\hat{\boldsymbol{U}}$ given $\boldsymbol{U} \in T_Q$, i.e., P_N is a PMF on $\mathcal{C}_Q$ with

$$P_N(\hat{\boldsymbol{u}}) = \Pr[\hat{\boldsymbol{U}} = \hat{\boldsymbol{u}}|\boldsymbol{U} \in T_Q] \sum_{\boldsymbol{u} \in T_Q:\, f_Q(\boldsymbol{u}) = \hat{\boldsymbol{u}}} P(u)/P(T_Q). \tag{33}$$

The Rényi entropy $H_{1/(1+\rho)}(P_N)$ is upper-bounded by

$$\ln |C_Q| = N[R(D, Q) + o(N)]$$

so

$$\begin{aligned}
&\boldsymbol{E}[G_Q(\hat{\boldsymbol{U}}|\boldsymbol{Y})^\rho|\boldsymbol{U} \in T_Q] \\
&\qquad \leq c(\rho)\ \exp\{N[\rho R(D, Q) - \lambda E_0(\rho) + o(N)]^+\}. && (34)
\end{aligned}$$

Now recalling that $P(T_Q) \leq \exp[-ND(Q||P)]$ [5, p. 32], we have

$$\begin{aligned}
&\max_Q\{P(T_Q)\boldsymbol{E}[G_Q(\hat{\boldsymbol{U}}|\boldsymbol{Y})^\rho|\boldsymbol{U} \in T_Q]\} && (35)\\
&\qquad \leq c(\rho)\ \exp\{N \max_Q[-D(Q||P) + \rho R(D, Q) \\
&\qquad\qquad - \lambda E_0(\rho) + o(N)]^+\} && (36)\\
&\qquad = c(\rho)\ \exp\{N[E(D, \rho) - \lambda E_0(\rho) + o(N)]^+\} && (37)
\end{aligned}$$

where the last line follows by (2), proved in [2]. Substituting this into (29) and noting that $v \leq (1 + N)^{|\mathcal{U}|}$, we have the proof that $E_{\text{sc}}(D, \rho) \leq [E(D, \rho) - \lambda E_0(\rho)]^+$.

Converse Part: We need to show

$$E_{\text{sc}}(D, \rho) \geq [E(D, \rho) - \lambda E_0(\rho)]^+.$$

Consider an arbitrary D-admissible joint source-channel guessing strategy $\{\mathcal{G}(\boldsymbol{y}): \boldsymbol{y} \in \mathcal{Y}^K\}$ for $\mathcal{U}^N$, with associated guessing function $G(\boldsymbol{u}|\boldsymbol{y})$. Let $\boldsymbol{U}_Q$, $\boldsymbol{Y}_Q$ denote the random variables whose joint PMF equals the conditional PMF of $\boldsymbol{U}$, $\boldsymbol{Y}$ given $\boldsymbol{U} \in T_Q$; i.e., $\boldsymbol{U}_Q$ has a PMF which is uniform on T_Q, and Y_Q is the channel output random variable when the channel codeword for $\boldsymbol{U}_Q$ is transmitted. Then

$$\begin{aligned}
\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho] &= \sum_Q P(T_Q)\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho|\boldsymbol{U} \in T_Q] && (38)\\
&= \sum_Q P(T_Q)\boldsymbol{E}[G(\boldsymbol{U}_Q|\boldsymbol{Y}_Q)^\rho]. && (39)
\end{aligned}$$

Next we lower-bound the moments of $G(\boldsymbol{U}_Q|\boldsymbol{Y}_Q)$. For any fixed $\boldsymbol{y}$, let $\mathcal{G}_{0,Q}(\boldsymbol{y})$ be a guessing strategy for T_Q which is obtained from $\mathcal{G}(\boldsymbol{y})$ as follows. For each guess $\hat{\boldsymbol{u}}$ produced by $\mathcal{G}(\boldsymbol{y})$, $\mathcal{G}_{0,Q}(\boldsymbol{y})$ produces, successively, the elements of the set $B(\hat{\boldsymbol{u}}, D) = \{\boldsymbol{u} \in T_Q : d(\boldsymbol{u}, \hat{\boldsymbol{u}}) \le ND\}$. Clearly, $\mathcal{G}_{0,Q}(\boldsymbol{y})$ is lossless for T_Q, and has an associated guessing function that satisfies the bound

$$G_{0,Q}(\boldsymbol{u}|\boldsymbol{y}) \le B_{\max} G(\boldsymbol{u}|\boldsymbol{y}), \qquad \text{for each } \boldsymbol{u} \in T_Q \tag{40}$$

where

$$B_{\max} \triangleq \max_{\hat{\boldsymbol{u}} \in \hat{\mathcal{U}}^N} |B(\hat{\boldsymbol{u}}, D)|.$$

It is known [4] and also shown in the Appendix that

$$B_{\max} \le \exp\{N[H(Q) - R(D, Q) + o(N)]\}. \tag{41}$$

Now, by Proposition 1, and since

$$H_{1/(1+\rho)}(\boldsymbol{U}_Q) = (1/N) \ln |T_Q| \ge H(Q) - o(N)$$

(for the inequality, see, e.g., [5, p. 30]), we have

$$E[G_{0,Q}(\boldsymbol{U}_Q|\boldsymbol{Y}_Q)^\rho] \ge \exp\{N[\rho H(Q) - \lambda E_0(\rho) - o(N)]^+\}. \tag{42}$$

Combining (39)–(42), and using the bound

$$P(T_Q) \ge \exp\{-N[D(Q||P) + o(N)]\}$$

[5, p. 32], we obtain

$$\begin{aligned} \boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho] &\ge \sum_Q \exp\{N[-D(Q||P) + \rho R(D, Q) \\ &\qquad - \lambda E_0(\rho) - o(N)]^+\} \qquad (43) \\ &\ge \max_Q \exp\{N[-D(Q||P) + \rho R(D, Q) \\ &\qquad - \lambda E_0(\rho) - o(N)]^+\} \qquad (44) \\ &= \exp\{N[E(D, \rho) - \lambda E_0(\rho) - o(N)]^+\}. \qquad (45) \end{aligned}$$

This completes the proof of the converse part. □

As mentioned in the Introduction, the special case of Theorem 2 for $\lambda = 0$, which corresponds to having no channel, was proved in [2].

Further insight into Theorem 2 can be gained by studying the properties of the function $E_{\rm sc}(D, \rho)$.

Proposition 2: The joint source-channel guessing exponent function $E_{\rm sc}(D, \rho)$ has the following properties.

a) For fixed $\rho > 0$, $E_{\rm sc}(D, \rho)$ is a convex function of $D \ge 0$, which is strictly decreasing in the range where it is positive. There is a finite D_0, given by the solution of $E(D_0, \rho) = \lambda E_0(\rho)$, such that $E_{\rm sc}(D, \rho) = 0$ for $D \ge D_0$. For any distortion measure such that there exists no reconstruction symbol which is at distance zero to more than one source symbol, we have

$$E_{\rm sc}(0, \rho) = [\rho H_{1/(1+\rho)}(P) - \lambda E_0(\rho)]^+.$$

b) For fixed $D \ge 0$, $E_{\rm sc}(D, \rho)$ is a continuous function of $\rho \ge 0$, which is strictly increasing in the range where it is positive. We have $E_{\rm sc}(D, \rho) > 0$ for all $\rho > 0$ if and only if $R(D, P) > \lambda C$, where C is the channel capacity. The function $E_{\rm sc}(D, \rho)$ is convex in ρ whenever $E_0(\rho)$ is concave.

Proof: For the most part, this proposition is straightforward and we omit the full proof. We only mention that in part a), the convexity and monotone decreasing property of $E_{\rm sc}(D, \rho)$ as a function of D follow from the fact, proved in [2], that for fixed $\rho > 0$, $E(D, \rho)$ is a strictly decreasing, convex function of D in the range where it is positive.

For part b), we recall the fact, shown in [2], that $E(D, \rho)$ is a convex function of $\rho \ge 0$ (for fixed D). Since $E_0(\rho, S)$ is a concave function of $\rho \ge 0$ for any fixed PMF S [9, p. 142], $e(\rho, S) \triangleq E(D, \rho) - \lambda E_0(\rho, S)$ is a convex function of $\rho \ge 0$. By convexity and the fact that $e(0, S) = 0$, the function $e(\rho, S)$ is strictly increasing in the range of ρ where $e(\rho, S) > 0$. (This last statement is proved in the same manner as in the proof of Lemma 1.) Since $E_{\rm sc}(D, \rho)$ is given by $\min_S [e(\rho, S)]^+$, it is also strictly increasing where it is positive (the minimum of a family of increasing functions is increasing).

We have $E_{\rm sc}(D, \rho) > 0$ for all $\rho > 0$ if $e(\rho, S) > 0$ for all $\rho > 0$ and all S. Since $e(\rho, S)$ is a convex function of ρ with $e(0, S) = 0$, $e(\rho, S) > 0$ for all $\rho > 0$ if and only if $e'(0, S) > 0$. But

$$\begin{aligned} e'(0, S) &= \lim_{\rho \to 0} [E(D, \rho) - \lambda E_0(\rho)]/\rho \\ &= R(D, P) - \lambda I(S, W). \end{aligned}$$

It follows that $E(D, \rho) > 0$ for all $\rho > 0$ if and only if $R(D, P) > \lambda C = \lambda \max_S I(S, W)$. Since $E(D, \rho)$ is convex, it is clear that $E_{\rm sc}(D, \rho)$ is convex whenever $E_0(\rho)$ is concave. (However, $E_{\rm sc}(D, \rho)$ is in general nonconvex, as shown for the lossless case $D = 0$ in the previous section.) This completes the proof.

It is interesting that, as we have just proved, if $R(D, P) > \lambda C$, then $E_{\rm sc}(D, \rho) > 0$ for all $\rho > 0$, and hence, $\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho]$ must go to infinity as N goes to infinity for all $\rho > 0$. Conversely, if $R(D, P) < \lambda C$, then there exists a $\rho > 0$ such that it is possible to keep $\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho]$ from growing exponentially in N. The conditions $R(D, P) < \lambda C$ and $R(D, P) > \lambda C$ are also the conditions for the validity of the direct and converse parts, respectively, of Shannon's joint source-channel coding theorem [9, p. 449] for the lossy case. This is analogous to the problem already mentioned in the lossless case, and the same type of remarks apply.

V. Source-Channel List Decoding Exponent

The aim of this section is to prove the following result.

Theorem 3: For any DMS P, DMC W, $L > 0$, and $D \ge 0$, the source-channel list-error exponent is given by

$$F_{\rm sc}(L, D) = \min_{R \ge L} [F(R, D) + \lambda E_{\rm sp}[(R - L)/\lambda]]. \tag{46}$$

Before we give the proof, we wish to comment on some aspects of this theorem. We remark that for the special case $\lambda = 0$, determining $F_{\rm sc}(L, D)$ is equivalent to determining the "error exponent in source coding with a fidelity criterion," a problem solved by Marton [15]. In this problem, one is interested in the probability that a rate-distortion codebook $\mathcal{C} \subset \hat{\mathcal{U}}^N$ of size e^{NL} contains no codeword which is within distance ND of the random vector $\boldsymbol{U} \in \mathcal{U}^N$ produced by a DMS P. Marton's exponent $F(L, D)$ is the best attainable exponential rate of decay of this probability as $N \to \infty$. Indeed, for $\lambda = 0$, we have $F_{\rm sc}(L, D) = F(L, D)$, in agreement with Marton's result.

It will be noted that the case $L = 0$ is excluded from the theorem. For $L = 0$, we have a list of size 1, independent of N. As mentioned in the Introduction, list-of-1 decoding in joint source-channel coding systems was considered by Csiszár; and the error exponent remains only partially known. We also note that if $L = 0$ is interpreted as the list size going to infinity at a subexponential rate, then the theorem holds also for $L = 0$. We do not prove this statement, since subexponential list sizes are not of interest in the present work.

Finally, we wish to re-iterate that though list-decoders with exponential list sizes are not viable in applications, the above theorem serves as a tool to find bounds on the distribution of computation in sequential decoding, as shown in the next section.

Proof of Theorem 3:

Direct Part: We need to show

$$F_{\rm sc}(L, D) \geq \min_{R \geq L} \{F(R, D) + \lambda E_{\rm sp}[(R - L)/\lambda]\}.$$

To obtain an upper bound on the minimum attainable probability of list decoding error, we consider a two-stage encoding scheme and an interlaced guessing strategy, just as in the proof of Theorem 2. Then, for any fixed $\boldsymbol{y}$, among the first e^{NL} guesses by $\mathcal{G}(\boldsymbol{y})$, there are at least $\lfloor v^{-1}e^{NL} \rfloor$ guesses by each $\mathcal{G}_Q(\boldsymbol{y})$. So, we have (writing $v^{-1}e^{NL}$ in place of $\lfloor v^{-1}e^{NL} \rfloor$ for notational convenience)

$$\begin{aligned}
&\Pr[G(\boldsymbol{U}|\boldsymbol{Y}) > e^{NL}] \\
&\quad \leq \sum_Q P(T_Q) \Pr[G_Q(\boldsymbol{U}|\boldsymbol{Y}) > v^{-1}e^{NL} | \boldsymbol{U} \in T_Q] && (47) \\
&\quad = \sum_Q P(T_Q) \Pr[G_Q(\boldsymbol{U}_Q|\boldsymbol{Y}_Q) > v^{-1}e^{NL}] && (48) \\
&\quad \leq \sum_Q P(T_Q) \Pr[G_Q(\hat{\boldsymbol{U}}_Q|\boldsymbol{Y}_Q) > v^{-1}e^{NL}] && (49)
\end{aligned}$$

where $\boldsymbol{U}_Q$, $\hat{\boldsymbol{U}}_Q$, $\boldsymbol{Y}_Q$ are random variables whose joint PMF equals the conditional joint PMF of $\boldsymbol{U}$, $\hat{\boldsymbol{U}}$, $\boldsymbol{Y}$, given $\boldsymbol{U} \in T_Q$. To be precise

$$\begin{aligned}
&\Pr[\boldsymbol{U}_Q = \boldsymbol{u}, \hat{\boldsymbol{U}}_Q = \hat{\boldsymbol{u}}, \boldsymbol{Y}_Q = \boldsymbol{y}] \\
&\quad = \Pr[\boldsymbol{U} = \boldsymbol{u}, \hat{\boldsymbol{U}} = \hat{\boldsymbol{u}}, \boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{U} \in T_Q] && (50) \\
&\quad = \begin{cases} P(\boldsymbol{u})W(\boldsymbol{y}|g_Q(\hat{\boldsymbol{u}}))/P(T_Q), & \boldsymbol{u} \in T_Q \\ & \hat{\boldsymbol{u}} = f_Q(\boldsymbol{u}) \\ 0, & \text{otherwise.} \end{cases} && (51)
\end{aligned}$$

Note, in particular, that $\hat{\boldsymbol{U}}_Q$ is a random variable over the rate-distortion code $\mathcal{C}_Q$. The event $G(\hat{\boldsymbol{U}}_Q|\boldsymbol{Y}_Q) > v^{-1}e^{NL}$ may be interpreted as an error event in a communication system with a message ensemble $\hat{\boldsymbol{U}}_Q$ of rate

$$H(\hat{\boldsymbol{U}}_Q)/K \leq (1/K) \ln|\mathcal{C}_Q| \leq [R(D, Q) + o(N)]/\lambda$$

and with a list-decoder of list-rate

$$(1/K) \ln[v^{-1}e^{NL}] = [L - o(N)]/\lambda.$$

By a well-known random-coding bound on the best attainable probability of error for list-decoders [7], [17], the channel encoder g_Q can be chosen so that

$$\begin{aligned}
&\Pr[G(\hat{\boldsymbol{U}}_Q|\boldsymbol{Y}_Q) > v^{-1}e^{NL}] \\
&\qquad \leq \exp\{-N[\lambda E_{\rm sp}[(R(D, Q) - L)/\lambda] - o(N)]\}. && (52)
\end{aligned}$$

By (49) and (52), and using the fact that

$$P(T_Q) \leq \exp[-N(D(Q||P)]$$

we now have

$$\begin{aligned}
&\Pr[G(\boldsymbol{U}|\boldsymbol{Y}) > e^{NL}] \\
&\quad \leq \sum_Q \exp\{-N[D(Q||P) + \lambda E_{\rm sp}[(R(D, Q) - L)/\lambda] \\
&\qquad - o(N)]\} && (53) \\
&\quad \leq v \exp\{-N \min_Q [D(Q||P) + \lambda E_{\rm sp}[(R(D, Q) - L)/\lambda] \\
&\qquad - o(N)]\} && (54) \\
&\quad = \exp\{-N \min_R [F(R, D) + \lambda E_{\rm sp}[(R - L)/\lambda] \\
&\qquad - o(N)]\} && (55)
\end{aligned}$$

where in the last line, the term v was absorbed by $o(N)$, and we used the following equality:

$$\begin{aligned}
&\min_R \{F(R, D) + \lambda E_{\rm sp}[(R - L)/\lambda]\} \\
&\quad = \min_R \min_{Q:\, R(D, Q) \geq R} \{D(Q||P) + \lambda E_{\rm sp}[(R - L)/\lambda]\} && (56) \\
&\quad = \min_Q \min_{R \leq R(D, Q)} \{D(Q||P) + \lambda E_{\rm sp}[(R - L)/\lambda]\} && (57) \\
&\quad = \min_Q \{D(Q||P) + \lambda E_{\rm sp}[(R(D, Q) - L)/\lambda]\}. && (58)
\end{aligned}$$

In (58), we made use of the monotone decreasing property of $E_{\rm sp}$. Note that since $E_{\rm sp}(\cdot)$ is infinite for negative arguments and $F(R, D)$ is infinite for

$$R > R_{\max}(D) \stackrel{\Delta}{=} \max_Q R(D, Q)$$

the minimum over R in (55) can be restricted to the range $[L, R_{\max}(D)]$, provided, of course, that $L < R_{\max}(D)$. This justifies the use of min rather than inf in the minimization over R. (For $L \geq R_{\max}(D)$, the probability of failure can be trivially made zero). This completes the proof of the direct part.

Converse Part: We need to show

$$F_{\rm sc}(L, D) \le \min_{R \ge L} \{F(R, D) + \lambda E_{\rm sp}[(R-L)/\lambda]\}.$$

We follow the method Csiszár [4] used in lower-bounding $\Pr[G(\boldsymbol{U}|\boldsymbol{Y}) > 1]$. Let $\{\mathcal{G}(\boldsymbol{y})\}$ be an arbitrary D-admissible guessing strategy for $\mathcal{U}^N$, and $G(\boldsymbol{u}|\boldsymbol{y})$ the associated guessing function. As proved in Appendix C, each guess $\hat{\boldsymbol{u}} \in \hat{\mathcal{U}}^N$ by $\mathcal{G}(\boldsymbol{y})$ covers, within distortion level ND, at most

$$\exp\{N[H(Q) - R(D, Q) + o(N)]\}$$

elements of T_Q. Thus e^{NL} guesses cover at most

$$\exp\{N[L + H(Q) - R(D, Q) + o(N)]\}$$

elements of T_Q. Thus conditional on $\boldsymbol{U} \in T_Q$, $G(\boldsymbol{U}|\boldsymbol{Y}) > e^{NL}$ corresponds to making an error with a list size of at most

$$\exp\{N[L + H(Q) - R(D, Q) + o(N)]\}.$$

So, by the sphere-packing lower bound for list decoding [17], we have

$$\Pr[G(\boldsymbol{U}|\boldsymbol{Y}) > e^{NL} | \boldsymbol{U} \in T_Q] \ge \exp[-N[\lambda E_{\rm sp}[(R(D, Q) - L)/\lambda] + o(N)]]. \quad (59)$$

(Note that the argument of $E_{\rm sp}$ is obtained as the difference of the source rate $H(Q)/\lambda$ and the list rate $[L + H(Q) - R(D, Q)]/\lambda$.) Since $P(T_Q) \ge \exp\{-N[D(Q||P) + o(N)]\}$, we obtain

$$\Pr[G(\boldsymbol{U}|\boldsymbol{Y}) > e^{NL}] \ge \exp\{-N \min_Q [D(Q||P) + \lambda E_{\rm sp}[(R(D, Q) - L)/\lambda] + o(N)]\} \quad (60)$$

which completes the proof in view of (56)–(58).

VI. Application to Sequential Decoding

Sequential decoding is a search algorithm introduced by Wozencraft [18] for finding the transmitted path through a tree code. Well-known versions of sequential decoding are due to Fano [6], Zigangirov [19], and Jelinek [10]. The computational effort in sequential decoding is a random variable, depending on the transmitted sequence, the received sequence, and the exact search algorithm. Our aim in this section is to exploit the relationship between guessing and sequential decoding to obtain converse (unachievability) results on the performance of sequential decoders.

Koshelev [14] and Hellman [12] considered using a convolutional encoder for joint source-channel encoding and a sequential decoder at the receiver for lossless recovery ($D = 0$) of the source output sequence. For the class of Markov sources, Koshelev showed that the expected computation per correctly decoded digit in such a system can be kept bounded if the Rényi entropy of order $1/2$ for the source, $\lim_{N\to\infty} H_{1/2}(P_N)/N$, is smaller than $\lambda E_0(1)$. Here, P_N denotes the joint probability distribution for the first N source letters. In this section, we first prove a converse result which complements Koshelev's achievability result. Subsequently, we prove a converse for the lossy case.

Consider an arbitrary discrete source (not necessarily Markovian) with distribution P_N for the first N source letters. Consider an arbitrary tree code that maps source sequences into channel input sequences so that at each step the encoder receives n source symbols and emits $k = \lambda n$ channel input symbols. Thus each node of the tree has $|\mathcal{U}|^n$ branches emanating from it, and each branch is labeled with k channel symbols. Consider the set of nodes at a fixed level, N source symbols (or $K = \lambda N$ channel symbols) into the tree code. Each node at this level is associated in a one-to-one manner with a sequence $\boldsymbol{u}$ of length N in the source ensemble. Only one of these nodes lies on the channel sequence that actually gets transmitted in response to the source output realization; we call this node the *correct node*. The correct node at level N is a random variable, which we identify and denote by $\boldsymbol{U}$, the first N symbols of the source. We let $\boldsymbol{X}$ denote the channel input sequence of length K corresponding to the correct node $\boldsymbol{U}$, and $\boldsymbol{Y}$ the channel output sequence of length K that is received when $\boldsymbol{X}$ is transmitted.

Now we use an idea due to Jacobs and Berlekamp [13] to relate guessing to sequential decoding. Any sequential decoder, applied to the above tree code, begins its search at the origin and extends it branch by branch eventually to examine a node $\boldsymbol{u}'$ at level N, possibly going on to explore nodes beyond $\boldsymbol{u}'$. We assume that if $\boldsymbol{U} \ne \boldsymbol{u}'$, i.e., if $\boldsymbol{u}'$ is not the correct node at level N, then the decoder eventually retraces its steps back to below level N and proceeds to examine a second node $\boldsymbol{u}''$ at level N. If $\boldsymbol{U} \ne \boldsymbol{u}''$, then eventually a third node at level N is examined, and so on. Thus for any given realization $\boldsymbol{y}$ of $\boldsymbol{Y}$, we have an ordering of the nodes at level N, in which a node $\boldsymbol{u}$ is preceded by those nodes that the sequential decoder examines before $\boldsymbol{u}$, when $\boldsymbol{u}$ is the correct node. We let $G(\boldsymbol{u}|\boldsymbol{y})$ denote the position of $\boldsymbol{u}$ in this ordering when $\boldsymbol{Y} = \boldsymbol{y}$. (By definition of sequential decoding, the value $G(\boldsymbol{u}|\boldsymbol{y})$ is well-defined in the sense that, for any fixed sequential decoder and fixed tree code, the order in which nodes at level N are examined does not depend on the portion of the channel output sequence beyond level K; it depends only on $\boldsymbol{y}$.)

Clearly, $G(\boldsymbol{u}|\boldsymbol{y})$ is a lower bound to the number of computational steps performed by the sequential decoder in decoding the first N symbols of the transmitted sequence, when $\boldsymbol{U} = \boldsymbol{u}$ and $\boldsymbol{Y} = \boldsymbol{y}$. Let C_N denote the (random) number of steps by the sequential decoder to correctly decode the first N source symbols. Then, lower bounds to the moments $\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho]$ constitute lower bounds to $\boldsymbol{E}[C_N^\rho]$. By Proposition 1

$$\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho] \ge \exp\{N[\rho H_{1/(1+\rho)}(P_N)/N - \lambda E_0(\rho) - o(N)]^+\}. \quad (61)$$

So, if

$$\limsup_{N\to\infty} \rho H_{1/(1+\rho)}(P_N)/N > \lambda E_0(\rho)$$

then $\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho]$ grows exponentially with N (for some subsequence), and so does $\boldsymbol{E}[C_N^\rho]$. In particular, if

$$\limsup_{N\to\infty} H_{1/2}(P_N)/N > \lambda E_0(1)$$

then the average computation per correctly decoded digit is unbounded and sequential decoding cannot be used in practice.

We summarize this converse result as follows.

Proposition 3: Suppose a discrete source, with distribution P_N for the first N source letters, is encoded, using a tree code, into the input of a DMC W at a rate of λ channel symbols per source symbol, and a sequential decoder is used at the receiver. Let C_N be the amount of computation by the sequential decoder to correctly decode the first N source symbols. Then, the ρth moment of C_N grows exponentially with N if the "source rate"

$$\limsup_{N\to\infty} H_{1/(1+\rho)}(P_N)/N$$

exceeds λ times the channel "cutoff rate" $E_0(\rho)/\rho$.

This result complements Koshelev's result [14], mentioned above. Note that it applies for any $\rho \geq 0$, while Koshelev was concerned only with $\rho = 1$. We also note that this result generalizes the converse in [1], where the source was restricted to be a DMS with equiprobable letters.

Next we consider the lossy case. First, we need to make precise what successful guessing means in this case, since we are dealing here with piecemeal generation of a reconstruction sequence of indefinite length. We shall insist that for any realization $u_1, u_2, \cdots$ of the source sequence, the system eventually produces a reconstruction sequence $\hat{u}_1, \hat{u}_2, \cdots$ such that $d(u_1, \cdots, u_N;\ \hat{u}_1, \cdots, \hat{u}_N) \leq ND$ for all $N \geq N_0$, where N_0 is a constant independent of the source and reconstruction sequences. This means that we desire to have a reconstruction sequence that stays close to the source sequence, with the possible exception of a finite initial segment.

As in the lossless case, the tree encoder receives successive blocks of n symbols from the source and for each such block emits $k = \lambda n$ channel input symbols. The sequential decoder works in the usual manner, generating a guess at each node it visits. The guess associated with a node at level N is a reconstruction block $\hat{\boldsymbol{u}} = (\hat{u}_1, \cdots, \hat{u}_N)$ of length N, which stays fixed throughout. We assume a prefix property for the guesses in the sense that the guess at a node is the prefix of the guesses at its descendants.

Fix $N \geq N_0$. For any source block $\boldsymbol{u} = (u_1, \cdots, u_N)$ and channel output block $\boldsymbol{y} = (y_1, \cdots, y_N)$, let $G(\boldsymbol{u}|\boldsymbol{y})$ denote the number of nodes at level N visited by the sequential decoder before it first generates a guess $\hat{\boldsymbol{u}} = (\hat{u}_1, \cdots, \hat{u}_N)$ satisfying $d(\boldsymbol{u}, \hat{\boldsymbol{u}}) \leq ND$. It is possible that the sequential decoder subsequently revises its first D-admissible guess $\hat{\boldsymbol{u}}$ at level N, but eventually it must settle for some D-admissible guess if it ever produces a D-admissible reconstruction of the entire source sequence. In any case, $G(\boldsymbol{u}|\boldsymbol{y})$ is a lower bound to the number of computational steps by the sequential decoder until it settles for its final D-admissible guess about the source block $\boldsymbol{u}$, when $\boldsymbol{y}$ is the channel output block. Now assuming that the source in the system is a DMS, we have by Theorem 2

$$\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho] \geq \exp\{N[E(D, \rho) - \lambda E_0(\rho) - o(N)]\}. \quad (62)$$

We thus obtain the following converse result on the computational complexity of sequential decoding.

Proposition 4: Suppose a DMS P is encoded, using a tree code, into the input of a DMC W at a rate of λ channel symbols per source symbol, and a sequential decoder is used at the receiver. Let C_N be the amount of computation by the sequential decoder to generate a D-admissible reconstruction of the first N source letters. Then, for any $\rho > 0$, the moment $\boldsymbol{E}[C_N^\rho]$ must grow exponentially with N if $E(D, \rho) > \lambda E_0(\rho)$.

This result exhibits the operational significance of the functions $E(D, \rho)/\rho$ and $E_0(\rho)/\rho$. Note that as $\rho \to 0$, $E(D, \rho)/\rho \to R(D, P)$ and $E_0(\rho)/\rho \to C$, leading to the expected conclusion that if $R(D, P) > \lambda C$, then $\boldsymbol{E}[C_N^\rho]/N$ must go to infinity as N increases, for all $\rho > 0$.

We conjecture that a direct result complementing Proposition 4 can be proved. In other words, we conjecture that there exists a system, employing tree coding and sequential decoding, for which $\boldsymbol{E}[C_N^\rho]/N$ is bounded independently of N, for any given $\rho > 0$ satisfying $E(D, \rho) < \lambda E_0(\rho)$. The proof of such a direct result would be lengthy and will not be pursued here.

As a final remark, we note that the lower bound in Section V on the probability of list decoding error directly yields the following lower bound on the distribution of computation in sequential decoding:

$$\Pr[C_N \geq e^{NL}] \geq \exp\{-N[F_{\rm sc}(L, D) + o(N)]\}. \quad (63)$$

This is a generalization of the result in [13] about the Paretian behavior of the distribution of computation in sequential decoding.

VII. Conclusions

We considered the joint source-channel coding and guessing problem, and gave single-letter characterizations for the guessing exponent $E_{\rm sc}(D, \rho)$ and the list-error exponent $F_{\rm sc}(L, D)$ for the case where the source and channel are finite and memoryless. We applied the results to sequential decoding and gave a tight lower bound to moments of computation, which, in the lossless case, established the tightness of Koshelev's achievability result.

The results suggest that, as far as the ρth moment of the guessing effort is concerned, the quantity $E(D, \rho)/\rho$ can be interpreted as the effective rate of a DMS, and $E_0(\rho)/\rho$ as the effective capacity (cutoff rate) of a DMC. The operational significance of these information measures has emerged in connection with sequential decoding.

One may consider extending the joint source-channel guessing framework that we studied here by allowing *stochastic encoders* with the goal of improving the guessing performance. By a stochastic encoder we mean an encoder that maps any specific source output block $\boldsymbol{u}$ to a channel input block $\boldsymbol{x}$ with a certain probability $V(\boldsymbol{x}|\boldsymbol{u})$, where V is a transition probability matrix that characterizes the stochastic encoder. A deterministic encoder is a special case of a stochastic encoder for which $V(\boldsymbol{x}|\boldsymbol{u})$ takes the values 0 or 1 only. Now we show by a straightforward argument that stochastic encoders offer no advantage over deterministic ones. By a well-known fact, any stochastic encoder V can be written as a convex combination

of a number of deterministic encoders $\{V_i\}$

$$V(\boldsymbol{x}|\boldsymbol{u}) = \sum_i p_i V_i(\boldsymbol{x}|\boldsymbol{u}) \tag{64}$$

where $p_i > 0$ and $\sum_i p_i = 1$. In light of this, encoding by V may be seen as a two-stage process. First, one draws a sample from a random variable Z that takes the value i with probability p_i. The sample value of Z indicates which of the deterministic encoders V_i is to be used in the second stage. Now, consider two guessers for a system employing such a stochastic encoder. The first guesser observes only the channel output $\boldsymbol{Y}$ and tries to recover the source block $\boldsymbol{U}$ as best it can. The second guesser observes the random variable Z in addition to $\boldsymbol{Y}$. Suppose both guessers employ optimal strategies for their respective situations so as to minimize the ρth moment of the number of guesses. It is clear that any guessing strategy available to the first guesser is also available to the second. So, the second guesser can do no worse than the first, and we have

$$E[G^*(\boldsymbol{U}|\boldsymbol{Y})^\rho] \geq E[G^*(\boldsymbol{U}|\boldsymbol{Y}, Z)^\rho] \tag{65}$$

$$= \sum_i p_i E[G^*(\boldsymbol{U}|\boldsymbol{Y}, Z = i)^\rho] \tag{66}$$

$$\geq \min_i \{EG^*(\boldsymbol{U}|\boldsymbol{Y}, Z = i)^\rho] \tag{67}$$

where all guessing functions are optimal ones, i.e., they achieve the minimum possible value for the ρth moment (in particular, $G^*(\boldsymbol{U}|\boldsymbol{Y}, Z = i)$ is an optimal guessing function for the encoder V_i). This shows that the performance achieved by using a stochastic encoder cannot be better than that achievable by deterministic encoders.

A topic left unexplored in this paper is whether there exist universal guessing schemes, for which the encoder and the guessing strategy are designed without knowledge of the source and channel statistics and yet achieve the best possible performance. Other topics that may be studied further are the problems mentioned at the end of Sections III and IV, and the conjecture stated at the end of Section VI.

Appendix A
Proof of Proposition 1

We carry out the proof for an arbitrary finite-alphabet source with distribution P_N for the first N source letters. Note that this proof also covers Theorem 1 by taking P_N as a product-form distribution.

Direct Part: Fix an arbitrary encoder e_N. Let $\Pr(\boldsymbol{u}, \boldsymbol{y})$ denote the joint probability assignment

$$\Pr(\boldsymbol{u}, \boldsymbol{y}) = P_N(\boldsymbol{u}) W(\boldsymbol{y}|e_N(\boldsymbol{u})). \tag{A.1}$$

We use a guessing strategy $\{\mathcal{G}_N(\boldsymbol{y})\}$ such that $\mathcal{G}_N(\boldsymbol{y})$ generates its guesses in descending order of the probabilities $\Pr(\boldsymbol{u}, \boldsymbol{y})$. We let $G(\boldsymbol{U}|\boldsymbol{Y})$ denote the associated guessing function. By Gallager's method [8], we have for any $\rho \geq 0$

$$G(\boldsymbol{u}|\boldsymbol{y}) \leq \sum_{\boldsymbol{u}'} \left[\frac{\Pr(\boldsymbol{u}', \boldsymbol{y})}{\Pr(\boldsymbol{u}, \boldsymbol{y})}\right]^{1/(1+\rho)}. \tag{A.2}$$

Thus

$$\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho] \leq \sum_{\boldsymbol{u}, \boldsymbol{y}} \Pr(\boldsymbol{u}, \boldsymbol{y}) \left\{\sum_{\boldsymbol{u}'} \left[\frac{\Pr(\boldsymbol{u}', \boldsymbol{y})}{\Pr(\boldsymbol{u}, \boldsymbol{y})}\right]^{1/(1+\rho)}\right\}^\rho$$

$$= \sum_{\boldsymbol{y}} \left[\sum_{\boldsymbol{u}} \Pr(\boldsymbol{u}, \boldsymbol{y})^{1/(1+\rho)}\right]^{1+\rho}. \tag{A.3}$$

Now, we employ a technique used in the sequential decoding literature to upper-bound the moments of computation [11]. Fix $\rho > 0$ and let n be the integer satisfying $n - 1 < \rho \leq n$. Then

$$\left[\sum_{\boldsymbol{u}} \Pr(\boldsymbol{u}, \boldsymbol{y})^{1/(1+\rho)}\right]^{1+\rho}$$

$$= \left[\sum_{\boldsymbol{u}_1} \cdots \sum_{\boldsymbol{u}_{n+1}} \Pr(\boldsymbol{u}_1, \boldsymbol{y})^{1/(1+\rho)} \cdots \Pr(\boldsymbol{u}_{n+1}, \boldsymbol{y})^{1/(1+\rho)}\right]^{(1+\rho)/(1+n)} \tag{A.4}$$

$$= \left[\sum_{\mathcal{S}} \sum_{\boldsymbol{u}_1} \sum_{\boldsymbol{u}_2 \neq \boldsymbol{u}_1} \cdots \sum_{\boldsymbol{u}_{|\mathcal{S}|} \neq \boldsymbol{u}_1, \cdots, \boldsymbol{u}_{|\mathcal{S}|-1}} \prod_{i=1}^{|\mathcal{S}|} \cdot \Pr(\boldsymbol{u}_i, \boldsymbol{y})^{m_i/(1+\rho)}\right]^{(1+\rho)/(1+n)} \tag{A.5}$$

$$= \left[\sum_{\mathcal{S}} \alpha_{\mathcal{S}}(\boldsymbol{y})\right]^{(1+\rho)/(1+n)} \tag{A.6}$$

$$\leq \sum_{\mathcal{S}} \alpha_{\mathcal{S}}(\boldsymbol{y})^{(1+\rho)/(1+n)}. \tag{A.7}$$

In (A.5), we rewrote the summation in terms of partitions $\mathcal{S} = \{S_1, \cdots, S_{|\mathcal{S}|}\}$ of the set $\{1, \cdots, n+1\}$. Each element S_i of a partition denotes the group of sums on the right-hand side of (A.4) whose indexes $\boldsymbol{u}_j$, $j \in S_i$, are restricted to remain identical (as they range through the set of all possible source blocks). In (A.5), m_i denotes the cardinality of S_i. Note that since sums belonging to different S_i's must assume distinct $\boldsymbol{u}_i$ values, we have the restriction $\boldsymbol{u}_i \neq \boldsymbol{u}_1, \cdots, \boldsymbol{u}_{i-1}$ in (A.5). Equation (A.6) defines the notation $\alpha_{\mathcal{S}}(\boldsymbol{y})$, and (A.7) follows by a variant of Jensen's inequality [9, ineq. (f), p. 523].

Before we proceed, we illustrate the above partitioning by an example. Suppose $n = 2$. Then, there are five partitions: $\mathcal{S}_0 = \{\{1, 2, 3\}\}$, $\mathcal{S}_1 = \{\{1, 2\}, \{3\}\}$, $\mathcal{S}_2 = \{\{1, 3\}, \{2\}\}$, $\mathcal{S}_3 = \{\{2, 3\}, \{1\}\}$, $\mathcal{S}_4 = \{\{1\}, \{2\}, \{3\}\}$; and, any sum of the form

$$\sum_i \sum_j \sum_k a_i a_j a_k$$

with indexes running through a common set, can be written as the sum of the sums $\sum_i \sum_{j \neq i} \sum_{k \neq i, j} a_i a_j a_k$, $\sum_i \sum_{j \neq i} a_i^2 a_j$ (repeated three times), and $\sum_i a_i^3$.

To continue with the proof, let $\mathcal{S}_0$ denote the trivial partition which has only one element, i.e., $|\mathcal{S}_0| = 1$ and $m_1 = n + 1$.

We shall treat this partition separately. By the same variant of Jensen's inequality mentioned above, we have

$$\sum_{\boldsymbol{y}} \alpha_{\mathcal{S}_0}(\boldsymbol{y})^{(1+\rho)/(1+n)}$$

$$= \sum_{\boldsymbol{y}} \left[\sum_{\boldsymbol{u}} \Pr(\boldsymbol{u}, \boldsymbol{y})^{(1+n)/(1+\rho)} \right]^{(1+\rho)/(1+n)} \tag{A.8}$$

$$\leq \sum_{\boldsymbol{y}} \sum_{\boldsymbol{u}} \Pr(\boldsymbol{u}, \boldsymbol{y}) \tag{A.9}$$

$$= 1. \tag{A.10}$$

Combining (A.3), (A.7), and (A.10), we obtain

$$\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^{\rho}] \leq 1 + \sum_{\mathcal{S} \neq \mathcal{S}_0} \sum_{\boldsymbol{y}} \alpha_{\mathcal{S}}(\boldsymbol{y})^{(1+\rho)/(1+n)}. \tag{A.11}$$

We shall now consider choosing the encoder e_N at random. Specifically, we suppose that each source block $\boldsymbol{u}$ is assigned the codeword $\boldsymbol{x}$ with probability $S^*(\boldsymbol{x})$, independently of all other codeword assignments. The PMF S^* is of product form with single-letter distribution S^* chosen so as to achieve the maximum in (17). Denoting expectation w.r.t. the random code ensemble by an overline, we have

$$\overline{\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^{\rho}]} \leq 1 + \sum_{\mathcal{S} \neq \mathcal{S}_0} \sum_{\boldsymbol{y}} \overline{\alpha_{\mathcal{S}}(\boldsymbol{y})^{(1+\rho)/(1+n)}} \tag{A.12}$$

$$\leq 1 + \sum_{\mathcal{S} \neq \mathcal{S}_0} \sum_{\boldsymbol{y}} \overline{\alpha_{\mathcal{S}}(\boldsymbol{y})}^{(1+\rho)/(1+n)} \tag{A.13}$$

where (A.13) is by Jensen's inequality. Now we can write (A.14)–(A.17) shown at the bottom of this page, where (A.15) is by the independence of codeword assignments to distinct messages, and (A.16) is simply by removing the restriction $\boldsymbol{u}_i \neq \boldsymbol{u}_1, \cdots, \boldsymbol{u}_{i-1}$.

Now define

$$\beta_i(\boldsymbol{y}) = \left[\sum_{\boldsymbol{u}} \overline{\Pr(\boldsymbol{u}, \boldsymbol{y})^{i/(1+\rho)}} \right]^{(1+\rho)/(1+n)} \tag{A.18}$$

and use (A.13) and (A.17) to write

$$\overline{\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^{\rho}]} \leq 1 + \sum_{\mathcal{S} \neq \mathcal{S}_0} \sum_{\boldsymbol{y}} \prod_{i=1}^{|\mathcal{S}|} \beta_{m_i}(\boldsymbol{y}) \tag{A.19}$$

$$\leq 1 + \sum_{\mathcal{S} \neq \mathcal{S}_0} \prod_{i=1}^{|\mathcal{S}|} \left[\sum_{\boldsymbol{y}} \beta_{m_i}(\boldsymbol{y})^{(1+n)/m_i} \right]^{m_i/(1+n)} \tag{A.20}$$

where (A.20) is by Hölder's inequality (note that $\sum_i m_i = n+1$). Now,

$$\sum_{\boldsymbol{y}} \beta_i(\boldsymbol{y})^{(1+n)/i}$$

$$= \sum_{\boldsymbol{y}} \left[\sum_{\boldsymbol{u}} \overline{\Pr(\boldsymbol{u}, \boldsymbol{y})^{i/(1+\rho)}} \right]^{(1+\rho)/i} \tag{A.21}$$

$$= \sum_{\boldsymbol{y}} \left[\sum_{\boldsymbol{u}} \sum_{\boldsymbol{x}} S^*(\boldsymbol{x}) P_N(\boldsymbol{u})^{i/(1+\rho)} W(\boldsymbol{y}|\boldsymbol{x})^{i/(1+\rho)} \right]^{(1+\rho)/i} \tag{A.22}$$

$$= \left[\sum_{\boldsymbol{u}} P_N(\boldsymbol{u})^{i/(1+\rho)} \right]^{(1+\rho)/i} \cdot \sum_{\boldsymbol{y}} \left[\sum_{\boldsymbol{x}} S^*(\boldsymbol{x}) W(\boldsymbol{y}|\boldsymbol{x})^{i/(1+\rho)} \right]^{(1+\rho)/i} \tag{A.23}$$

$$= \exp\{N[\rho_i H_{1/(1+\rho_i)}(P_N)/N - \lambda E_0(\rho_i, S^*)]\} \tag{A.24}$$

where we have defined $\rho_i = (1 + \rho - i)/i$. Note that for $1 \leq i \leq n$, we have $0 < \rho_i \leq \rho$.

For shorthand, let us write

$$f(r) = r H_{1/(1+r)}(P_N)/N - \lambda E_0(r, S^*). \tag{A.25}$$

To continue we need the following fact which is proved in Appendix B.

Lemma 1: $f(r)$ is a convex function of $r \geq 0$; $f(0) = 0$; and $f(r)$ is increasing in the range where it is positive.

Now we consider two cases. Case $f(\rho) \leq 0$: Then, for all $i = 1, \cdots, n$, we have $f(\rho_i) \leq 0$, and by (A.24) $\sum_{\boldsymbol{y}} \beta_i(\boldsymbol{y})^{(1+n)/i} \leq 1$.Using this in (A.20) (note that $1 \leq m_i \leq n$ for $\mathcal{S} \neq \mathcal{S}_0$), we obtain

$$\overline{\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^{\rho}]} \leq 1 + \sum_{\mathcal{S} \neq \mathcal{S}_0} \prod_{i=1}^{|\mathcal{S}|} \left[\sum_{\boldsymbol{y}} \beta_{m_i}(\boldsymbol{y})^{(1+n)/m_i} \right]^{m_i/(1+n)} \tag{A.26}$$

$$\overline{\alpha_{\mathcal{S}}(\boldsymbol{y})} = \overline{\sum_{\boldsymbol{u}_1} \sum_{\boldsymbol{u}_2 \neq \boldsymbol{u}_1} \cdots \sum_{\boldsymbol{u}_{|\mathcal{S}|} \neq \boldsymbol{u}_1, \cdots, \boldsymbol{u}_{|\mathcal{S}|-1}} \prod_{i=1}^{|\mathcal{S}|} \Pr(\boldsymbol{u}_i, \boldsymbol{y})^{m_i/(1+\rho)}} \tag{A.14}$$

$$= \sum_{\boldsymbol{u}_1} \sum_{\boldsymbol{u}_2 \neq \boldsymbol{u}_1} \cdots \sum_{\boldsymbol{u}_{|\mathcal{S}|} \neq \boldsymbol{u}_1, \cdots, \boldsymbol{u}_{|\mathcal{S}|-1}} \prod_{i=1}^{|\mathcal{S}|} \overline{\Pr(\boldsymbol{u}_i, \boldsymbol{y})^{m_i/(1+\rho)}} \tag{A.15}$$

$$\leq \sum_{\boldsymbol{u}_1} \sum_{\boldsymbol{u}_2} \cdots \cdots \sum_{\boldsymbol{u}_{|\mathcal{S}|}} \prod_{i=1}^{|\mathcal{S}|} \overline{\Pr(\boldsymbol{u}_i, \boldsymbol{y})^{m_i/(1+\rho)}} \tag{A.16}$$

$$= \prod_{i=1}^{|\mathcal{S}|} \sum_{\boldsymbol{u}} \overline{\Pr(\boldsymbol{u}, \boldsymbol{y})^{m_i/(1+\rho)}} \tag{A.17}$$

$$\leq 1+\sum_{\mathcal{S}\neq\mathcal{S}_0}\prod_{i=1}^{|\mathcal{S}|}1 \tag{A.27}$$

$$= c(\rho) \tag{A.28}$$

where $c(\rho)$ has been defined as the number of partitions $\mathcal{S}$.

Case $f(\rho)>0$: Now, for all $i=1,\cdots,n$, $f(\rho)\geq f(\rho_i)$, and by (A.24)

$$\sum_{\boldsymbol{y}}\beta_i(\boldsymbol{y})^{(1+n)/i}\leq\exp[Nf(\rho)].$$

Using this in (A.20), and recalling that

$$\sum_i m_i = n+1$$

we obtain

$$\overline{\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho]}$$

$$\leq 1+\sum_{\mathcal{S}\neq\mathcal{S}_0}\prod_{i=1}^{|\mathcal{S}|}\left[\sum_{\boldsymbol{y}}\beta_{m_i}(\boldsymbol{y})^{(1+n)/m_i}\right]^{m_i/(1+n)} \tag{A.29}$$

$$\leq 1+\sum_{\mathcal{S}\neq\mathcal{S}_0}\prod_{i=1}^{|\mathcal{S}|}\exp[Nf(\rho)m_i/(1+n)] \tag{A.30}$$

$$\leq c(\rho)\exp[Nf(\rho)]. \tag{A.31}$$

Combining (A.28) and (A.31), we conclude that

$$\overline{\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho]}\leq c(\rho)\exp\{N[f(\rho)]^+\}.$$

Thus there must be an encoder such that the resulting joint source-channel guessing scheme satisfies

$$\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho]\leq c(\rho)\exp\{N[f(\rho)]^+\}.$$

This completes the proof of the direct part.

Converse: Fix an arbitrary encoder e_N and an arbitrary guessing scheme $G(\boldsymbol{U}|\boldsymbol{Y})$. Let

$$\Pr(\boldsymbol{u},\boldsymbol{y}) = P_N(\boldsymbol{u})W(\boldsymbol{y}|e_N(\boldsymbol{u})).$$

By [1, Theorem 1]

$$\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho]\geq(1+N\ln|\mathcal{U}|)^{-\rho}\sum_{\boldsymbol{y}}\cdot\left[\sum_{\boldsymbol{u}}\Pr(\boldsymbol{u},\boldsymbol{y})^{1/(1+\rho)}\right]^{1+\rho}. \tag{A.32}$$

Now

$$\sum_{\boldsymbol{y}}\left[\sum_{\boldsymbol{u}}\Pr(\boldsymbol{u},\boldsymbol{y})^{1/(1+\rho)}\right]^{1+\rho}$$

$$=\left[\sum_{\boldsymbol{u}}P_N(\boldsymbol{u})^{1/(1+\rho)}\right]^{1+\rho}\cdot\sum_{\boldsymbol{y}}\left[\sum_{\boldsymbol{u}}P'_N(\boldsymbol{u})W(\boldsymbol{y}|e_N(\boldsymbol{u}))^{1/(1+\rho)}\right]^{1+\rho} \tag{A.33}$$

$$=\exp[\rho H_{1/(1+\rho)}(P_N)]\cdot\sum_{\boldsymbol{y}}\left[\sum_{\boldsymbol{x}}P''_N(\boldsymbol{x})W(\boldsymbol{y}|\boldsymbol{x})^{1/(1+\rho)}\right]^{1+\rho} \tag{A.34}$$

$$\geq\exp\{N[\rho H_{1/(1+\rho)}(P_N)/N-\lambda E_0(\rho)]\} \tag{A.35}$$

where

$$P'_N(\boldsymbol{u})=\frac{P_N(\boldsymbol{u})^{1/(1+\rho)}}{\sum_{\boldsymbol{u}'}P_N(\boldsymbol{u}')^{1/(1+\rho)}} \tag{A.36}$$

and

$$P''_N(\boldsymbol{x})=\sum_{\boldsymbol{u}\in\mathcal{U}^N:\,e_N(\boldsymbol{u})=\boldsymbol{x}}P'_N(\boldsymbol{u}). \tag{A.37}$$

Equation (A.35) follows by the parallel channels theorem [8, Theorem 5]. Thus

$$\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})]^\rho]\geq\exp\{N[\rho H_{1/(1+\rho)}(P_N)/N-\lambda E_0(\rho)-o(N)]\}. \tag{A.38}$$

This, together with the obvious fact that $\boldsymbol{E}[G(\boldsymbol{U}|\boldsymbol{Y})^\rho]\geq 1$, completes the proof.

Appendix B
Proof of Lemma 1

First, $rH_{1/(1+r)}(P')$ is convex in $r>0$ for any distribution P' since

$$g(r)\triangleq\left[\sum_u P(u)^{1/(1+r)}\right]^{1+r}$$

satisfies, by Hölder's inequality [9, ineq. (b), p. 522],

$$g(r_1)^\alpha g(r_2)^{1-\alpha}\geq g(\alpha r_1+(1-\alpha)r_2)$$

for any $r_1>0$, $r_2>0$, and $0<\alpha<1$. Since it is also known that $E_0(\rho,S)$ is a concave function of $\rho\geq 0$ [9, p. 142], the convexity of $f(r)$ follows.

That $f(0)=0$ is due to $E_0(0,S)=0$ [9, p. 142]. Thus the function $f(r)$ starts at 0 and may dip to negative values initially; then, it will become positive (excluding trivial cases) for r large enough. To see that $f(r)$ is increasing in the range where it is positive, consider any $0<r_1<r_2$ such that $f(r_1)>0$, $f(r_2)>0$. Let $\alpha=r_1/r_2$. Then, by convexity, $(1-\alpha)f(0)+\alpha f(r_2)\geq f(r_1)$. But $f(0)=0$, so we have $f(r_2)\geq(r_2/r_1)f(r_1)>f(r_1)$.

Appendix C
Upper Bound on $B_{\max}$

We wish to upper-bound the size of

$$B(\hat{\boldsymbol{u}},D)=\{\boldsymbol{u}\in T_Q: d(\boldsymbol{u},\hat{\boldsymbol{u}})\leq ND\}$$

for arbitrary $\hat{\boldsymbol{u}}\in\hat{\mathcal{U}}^N$. Let $\hat{Q}$ denote the type of $\hat{\boldsymbol{u}}$, i.e, suppose $\hat{\boldsymbol{u}}\in T_{\hat{Q}}\subset\hat{\mathcal{U}}^N$. Consider the sets

$$S_V(\hat{\boldsymbol{u}},D)\triangleq B(\hat{\boldsymbol{u}},D)\cap T_V(\hat{\boldsymbol{u}}).$$

$T_V(\hat{\boldsymbol{u}})$ is empty unless the shell V is consistent with the marginal compositions, i.e.,

$$Q(x)=\sum_{\hat{u}}\hat{Q}(\hat{u})V(x|\hat{u}).$$

Assume henceforth that V is consistent in this sense. We have [5, p. 31]

$$|T_V(\hat{\boldsymbol{u}})|\leq\exp\{N[H(Q)-I(\hat{Q},V)]\}. \tag{A.39}$$

Now, note that $S_V(\hat{\boldsymbol{u}}, D)$ is empty unless

$$d(\hat{Q}, V) \stackrel{\Delta}{=} \sum_{\hat{u}, x} \hat{Q}(\hat{u})V(x|\hat{u})d(x, \hat{u}) \le D.$$

However, if $d(\hat{Q}, V) \le D$, then we have by definition, $R(D, Q) \le I(\hat{Q}, V)$, and hence by (A.39)

$$|T_V(\hat{\boldsymbol{u}})| \le \exp\{N[H(Q) - R(D, Q)]\}. \tag{A.40}$$

The proof is now completed as follows.

$$\begin{aligned}
&|B(\hat{\boldsymbol{u}}, D)| \\
&\quad = \sum_V |S_V(\hat{\boldsymbol{u}}, D)| && \text{(A.41)} \\
&\quad \le \sum_{V:\, d(\hat{Q}, V) \le D} |T_V(\hat{\boldsymbol{u}})| && \text{(A.42)} \\
&\quad \le \sum_{V:\, d(\hat{Q}, V) \le D} \exp\{N[H(Q) - R(D, Q)]\} && \text{(A.43)} \\
&\quad = \exp\{N[H(Q) - R(D, Q) + o(N)]\} && \text{(A.44)}
\end{aligned}$$

where in the last line we made use of the fact that the number of shells V grows polynomially in N.

ACKNOWLEDGMENT

The authors are grateful to I. Csiszár and V. Balakirsky for enlightening discussions.

REFERENCES

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

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

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

[4] I. Csiszár, "On the error exponent of source channel transmission with a distortion threshold," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 823–828, Nov. 1982.

[5] I. Csiszár and J. Körner, *Information Theory: Coding Systems for Discrete Memoryless Systems.* New York: Academic, 1981.

[6] R. M. Fano, "A heuristic discussion of sequential decoding," *IEEE Trans. Inform. Theory*, vol. IT-9, pp. 66–74, Jan. 1963.

[7] G. D. Forney, Jr., "Exponential error bounds for erasure, list and decision feedback schemes," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 206–220, Mar. 1968.

[8] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 3–18, Jan. 1965.

[9] ———, *Information Theory and Reliable Transmission.* New York: Wiley, 1968.

[10] F. Jelinek, "A fast sequential decoding algorithm using a stack," *IBM J. Res. Develop.*, vol. 13, pp. 675–685, 1969.

[11] T. Hashimoto and S. Arimoto, "Computational moments for sequential decoding of convolutional codes," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 584–591, Sept. 1979.

[12] M. Hellman, "Convolutional source encoding," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 651–656, Nov. 1975.

[13] I. M. Jacobs and E. R. Berlekamp, "A lowerbound to the distribution of computation for sequential decoding," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 167–174, Apr. 1967.

[14] V. N. Koshelev, "Direct sequential encoding and decoding for discrete sources," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 340–343, May 1973.

[15] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 197–199, 1974.

[16] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. Probability* (Berkeley, CA, 1961), vol. 1, pp. 547–561.

[17] C. E. Shannon, R. G. Gallager, and E. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels," *Inform. Contr.*, vol. 10, pp. 65–103, Jan. 1967.

[18] J. M. Wozencraft and B. Reiffen, *Sequential Decoding.* Cambridge, MA: MIT Press, 1961.

[19] K. Zigangirov, "Some sequential decoding procedures," *Probl. Pered. Inform.*, vol. 2, pp. 13–25, 1966.