Universal Randomized Guessing Subject to Distortion

Asaf Cohen\textsuperscript{a}, Senior Member, IEEE, and Neri Merhav\textsuperscript{b}, Fellow, IEEE

Abstract—In this paper, we consider the problem of guessing a sequence subject to a distortion constraint. Specifically, we assume the following game between Alice and Bob: Alice has a sequence $x$ of length $n$. Bob wishes to guess $x$, yet he is satisfied with finding any sequence $\hat{x}$ which is within a given distortion $D$ from $x$. Thus, he successively submits queries to Alice, until receiving an affirmative answer, stating that his guess was within the required distortion. Finding guessing strategies which minimize the number of guesses (the guesswork), and analyzing its properties (e.g., its $\rho$-th moment) has several applications in information security, source and channel coding. Guessing subject to a distortion constraint is especially useful in applications in information security, source and channel coding.

We suggest a randomized guessing strategy which is asymptotically optimal when considering contemporary biometrically-secured systems, where the “password” which protects the data is not a single, fixed vector but rather a ball of feature vectors centered at some $x$, and any feature vector within the ball results in acceptance. We formally define the guessing problem under distortion in four different setups: memoryless sources, guessing through a noisy channel, sources with memory and individual sequences. We suggest a randomized guessing strategy which is asymptotically optimal for all setups and is five-fold universal, as it is independent of the source statistics, the channel, the moment to be optimized, the distortion measure and the distortion level.

Index Terms—Guesswork, universal guessing, asynchronous guessing, universal distribution, rate–distortion theory.

I. INTRODUCTION

CONSIDER the problem of guessing a realization of a random $n$–vector $X$ over an alphabet $\mathcal{X}$ subject to a distortion measure $d$. Specifically, assume Alice has such a realization $x$. Bob wishes to guess the value of $x$, yet he is content with discovering any $n$–vector $\hat{x}$ over $\mathcal{X}$ which is within a distortion $D$ away from $x$. He thus submits to Alice a sequence of queries of the form “Is $\hat{x} _i$ close to $x$ within distortion $D(\cdot)$, $i = 1, 2, \ldots$, until receiving a positive answer. Let $G(x)$ denote the guesswork, namely, the smallest $i$ for which $\hat{x} _i$ is close enough to $x$, in the sense of achieving an average distortion smaller than or equal to $D$ under the measure $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. Clearly, Bob is interested in devising a strategy which minimizes $G(x)$ in some sense. If $x$ is a realization of a random vector, Bob may wish to minimize the expected value of $G(X)$ or some other moment of it. However, in many cases, Bob will have, for example, only limited knowledge (if any) on the distribution of $X$ or the exact distortion required for success.

Moreover, $x$ might be a single, fixed individual sequence, without any underlying probability distribution. Bob might also have limited resources, preventing him from remembering which queries were already submitted, or, alternatively, it might be the case where multiple “Bobs” try to guess $x$ simultaneously, without any coordination. To make the matter even worse, Bob’s queries might also reach Alice indirectly, e.g., through a noisy channel. Thus, a guessing strategy which is universal and efficient in the senses mentioned above is highly desirable. In other words, we seek a strategy which is independent of the source distribution, the memory structure, the moment to be optimized, the distortion measure and the distortion value, the channel between Bob and Alice, and, finally, a strategy which requires no memory of previously submitted queries and can be applied by multiple guessers simultaneously, yet, despite all the above, asymptotically achieves the optimal guessing performance.

Guesswork problems have numerous applications in information theory and several other fields. Coding applications were first given in [1]–[3] by Wozencraft, Arikan and Pfister and Sullivan, respectively. In [4], Arikan and Merhav used a guessing decoder for joint source–channel coding. Fixed-to-variable source coding without a prefix constraint, or one-shot coding, were considered by several groups in [5]–[8]. In fact, recently Kumar et al. [9] nicely tied several source coding, guessing and task partitioning problems [10] to the same minimization problem, resulting in simple proofs for several known results in the area. Guessing can also be seen as a method to quantify the complexity of a rate–distortion encoder, that is, the number of guesses until a vector within distortion $D$ of the source vector is found corresponds to the number of candidate codebook vectors the encoder examines until it finds an adequate codeword [11]. Another important application for guessing is directly guessing the noise over a channel for capacity-achieving decoding. Duffy et al. [12] suggested that the receiver can simply order noise sequences from most likely to least likely, then subtract these noise sequences from the received signal. The first noise sequence subtracted which results in a codeword corresponds to...
ML decoding. We review additional information-theoretic results in Section II.

Recent trends in information security raised further attention to guesswork problems, adopting it as a proxy to password strength [13]–[16]. A detailed discussion is given in [17, Section I], supporting the applicability of guesswork analysis, and its insights, to practical problems. It is important to note, though, that when focusing on such applications, one cannot limit the analysis to a simple, i.i.d. case with a known distribution, as sequences used in information security applications tend to have memory [18]–[20] and their exact distribution is rarely known [18]. More importantly, due to the many difficulties in managing passwords, contemporary authentication systems use different kinds of biometric measures [21]. In such systems, a feature vector is extracted from the biometric data (e.g., fingerprint, ECG, iris photo or even a sequence of keystroke timings) and fed to a previously–trained unit (e.g., a neural network) which either accepts or rejects the data. In what follows, we argue that not only such a feature vector might have a complex and unknown distribution, any system which either accepts or rejects it must have some level of distortion between the trained data and the tested data (unless ordinary passwords), as such feature vectors are extracted from real–world, human metrics and behaviours.

In [22], Page et al. suggested ECG-based biometric authentication. The raw QRS–complex1 of the signal was classified using a neural network, with only minor filtering as pre–processing. Raw ECG signals were used in [24] as 2D images. Again a (convolution) neural network was used for classification. ECG signals were used for authentication in [25] as well. This time, several features, such as variance, skewness and bandwidth occupied were extracted, and a Support Vector Machine (SVM) was used for classification. SVM was also used for voice biometrics in [26]. Additional techniques for voice authentication are discussed in [27], including both traditional MAP-based techniques as well as modern Deep Neural Networks. Detailed surveys can also be found in [21], [28]. The underlying assumption in all is that slight variations in the input data to the authentication system should not result in a different decision. Again, this should be contrasted with passwords, which require a perfect match for acceptance.

In fact, the classification literature is focused on constructing robust neural networks and SVMs, such that slightly distorted input would not distort the classification. Indeed, slightly distorted inputs at selected places which cause the classifier to output different classes are known in the literature as adversarial examples, and quite a few techniques were suggested to combat them (and output the same class as the non–distorted input), including adding noise [29], [30]. Bi and Zhang gave motivating examples in [31], referring to classifying sentences from speech recognition, or classifying images after an image processing phase. In both, it is clear that the initial recognition or processing might introduce errors to the input, and multiple instances of the input will result in multiple, slightly different, feature vectors, yet all should be mapped to the same class (or have the same accept or reject result). Hence, the classifier should be robust to such differences (distortions, in the context of this work), in the sense of producing the same class as the non-distorted input. Moreover, the noise model considered in [31] is i.i.d. with zero mean, which fits the distortion model we consider in this paper. A deterministic formulation is given in [32] by Tsai et al., where one wishes the output vector of the neural network for a perturbed input to be close to the original output, if the perturbed input vector is in a ball around the original one. In [33], Xu et al. suggested feature squeezing of the input data before learning or classification. For images, the technique reduces to bit depth reduction, which essentially means, again, that a “ball” of different, yet close, input vectors (images) is treated as one for both learning and classification, hence, in the context of guessing, any guess within this ball should be successful.

A. Main Contributions

In this paper, we suggest randomized, five-fold universal guessing strategies for guessing subject to a distortion measure. Specifically, we start with memoryless sources, and give a randomized, universal guessing distribution which achieves the optimal guesswork exponent for any memoryless source distribution, any moment (of the guesswork) and any distortion measure and level. As guessing sums up to drawing sequences independently from a given universal distribution, this method is inherently distributed, allowing multiple and asynchronous guessers to guess independently, yet asymptotically achieve the minimal number of queries. We continue to the problem of guessing through a noisy channel. In this case, again, the guesser (Bob) wishes to toss a sequence which is within a given distortion from the true word, yet now guesses pass through a noisy (memoryless) channel before arriving at Alice. We show that now the universal guessing distribution is not only universal in the distribution of the source, the moment and the distortion, but is also universal in the channel transition probability matrix. We also intuitively motivate the resulting guesswork exponent, and show how it implicitly includes the optimal, non-universal guessing distribution in an alternative expression.

We then revisit the problem for sources with memory. We give a guessing distribution, based on the Lempel–Ziv 78 (LZ78, [34]) parsing rule, which is shown to be universal for any Ψ-mixing source2 and any moment or distortion level. For the direct part, we first formally define a block–memoryless approximation for such sources. We then derive the exponent achieved by the universal guessing distribution, and conclude with a matching converse.

Last but not least, we turn to the fourth setup, in which the sequence to be guessed is an individual sequence, without any underlying probability distribution. We show that the LZ-based

1These are the typical downward–upward–downward deflections usually seen on an ECG signal [23].

2A wide family of sources with memory, which includes Markov models, Unifilar sources, etc. Roughly speaking, these are sources which might have long-term dependencies, yet the dependency between any two blocks decays with the distance between the blocks. A formal definition is given in Section IV-C.
guessing distribution discussed above is asymptotically optimal, in the sense of achieving the best possible guesswork exponent compared to any finite-state guessing machine.

Since the distribution suggested for memoryless sources can be viewed as a special case of the distribution for sources with memory, the results of the above four setups reveal that a single guessing distribution, is, in fact, asymptotically optimal for all setups, and hence universal in a multitude of dimensions: the source distribution and memory, the distortion measure and level and the memoryless channel the guesses might pass through. Moreover, this distribution has a simple and efficient implementation, which, in short, requires only applying an LZ decoder on a sequence of i.i.d. uniform bits. Finally, we also note that the guesswork exponents under this distribution meet very general converse bounds, which do not assume randomized guessing and, in fact, allow for a wide variety of guessing strategies.

The rest of this paper is organized as follows. In Section II, we thoroughly discuss related work. Section III includes the required notation and the problem statement. Section IV includes the main results, while Section V includes their proofs, except for the proofs of some technical lemmas and claims, which are deferred to the Appendix. Section VI concludes the paper.

II. RELATED WORK

Guesswork was information-theoretically formalized by Massey [35]. Later, Arikan [2] computed the exponential rate of guesswork for memoryless sources. Specifically, it was proved to be the Rényi entropy of order $\frac{1}{2}$. Arikan and Merhav [11] were then the first to study guesswork under a distortion constraint, and devised a strategy which is universal and asymptotically optimal for memoryless sources, both in the source distribution and the moment of the guesswork. Recently, randomized guessing subject to distortion was studied by Kuzuoka in [36], [37]. It was shown that for memoryless sources, the optimal guesswork moment can be achieved by randomized guessing. Yet, the distribution used was the tilted distribution similar to [38], thus depending on both the source distribution and the moment $\rho$. Hence, for memoryless sources, the strategy in [11] is independent of the distribution and the moment, yet depends on the distortion measure and value, while the tilted distribution in [36], [37] is independent of the distortion measure and value, yet depends on the source distribution and moment.

A. Ordering and Universality

Indeed, since the strategy suggested in [11] is to sort the sequences to be guessed in an increasing order of the empirical rate-distortion function, the method is independent of the true source distribution and the moment order. This universal strategy extends the strategy for guessing without distortion, which sorts the sequences in an increasing order of their empirical entropy. Such orderings were, in fact, found useful in other applications, e.g., Weinberger et al. [39] ordered strings by the size of their type-class before assigning them codewords in a fixed-to-variable source coding scheme, Kosut and Sankar used it in [8] and Beirami et al. [40] showed that the dominating type in guesswork (the position of a given string in the list) is the largest among all types whose elements are more likely than the given string. Note that in all these cases, sequences are sorted, then guessed one after the other, using, e.g., a list of sequences. This process is inherently centralized. However, all methods hint towards the structure of universal guessing distributions, which assign probabilities to sequences in an analogous manner. Such distributions played a key role in [17] and will play a role here as well. To conclude the discussion on universality, Sundaresan [44] also considered guessing under source uncertainty. The redundancy as a function of the radius of the family of possible distributions was defined and quantified. As expected ([11]), for discrete memoryless sources this redundancy tends to zero as the length of the sequence grows without bound.

B. Sources With Memory

Guesswork for Markov processes was first studied by Malone and Sullivan [45], and later extended by Pfister and Sullivan in [3]. Hanawal and Sundaresan proposed a large deviations approach [46], generalizing results from [2] and [45]. In [47], again via large deviations, Christiansen et al. proposed an approximation to the distribution of the guesswork. Recently, Merhav and Cohen [17] considered universal randomized guessing for sources with memory, and showed that an LZ78 [34] parsing procedure can be utilized to construct such a universal, asymptotically optimal distribution. Motivated by these results, Merhav [48] showed that, in fact, an LZ78 decoder fed by purely random bits is asymptotically optimal, not only for sources with memory, but also for individual sequences, compared to any other finite-state machine. The above works, however, considered only the lossless case, i.e., with no distortion allowed.

The need to account for sources with memory in guessing problems was also verified experimentally. Dell’Amico et al. [14] evaluated the probability of guessing passwords using dictionary-based, grammar-free and Markov chain strategies, using existing data sets of passwords. Guessing strategies which account for memory performed better. Moreover, the need to tune memory parameters was clear, thus highlighted the urgency for universal strategies. Bonneau [49] also acknowledged the difficulty of coping with passwords from unknown distributions. Needless to say, biometric authentication methods mentioned in Section I also stress out the fact that the sequences to be guessed usually have an unknown distribution, and might contain highly dependent measurements [22], [24], [26].

C. Noisy Channels

Guessing over a noisy channel was studied by Christiansen et al. in [50]. Therein, the sequence to be
guessed was assumed i.i.d., but was passed through a channel before reaching the guesser. Thus, if the channel introduces erasures, the goal is to fill the missing gaps. In a sense, this model was extended by Salamatian et al. in [43], where multiple guessers tried to fill the missing gaps, or correct errors introduced by the channel, using either a centralized or a decentralized approach. On the other hand, submitting guesses through a noisy channel was considered by Merhav in [51].

D. A Multitude of Models

The current literature on guesswork is growing, with numerous new results on various aspects of the problem. Christiansen et al. [52] considered a multi-user case, where an adversary has to guess \( U \) out of \( V \) strings. Beirami et al. [53] further defined the inscrutability of a source as the number of guesses for the above problem, and the inscrutability rate as its exponential growth rate. The same paper also showed that ordering strings by their type-size in ascending order is a universal guessing strategy. Yet, again, both [52] and [53] considered a single attacker, with the ability to create a list of strings and guess one after the other. Non-asymptotic results considered a single attacker, with the ability to create a list of strings and guess one after the other. Non-asymptotic results which can be applied to guesswork were given in [54] by Courtrand and Verdú. In [55], Sason derived bounds on the Rényi entropy of a function of a random variable, and applied them to derive non-asymptotic bounds on the difference between the exponent in guessing the original random variable, and that of guessing the (possibly non one-to-one) function. Several works considered guessing with side information [2], [43], [50], [56], [57]. Yona and Diggavi [58] considered the problem of guessing a word which has a similar hash function as the source word - a highly practical scenario as passwords are rarely stored as is, and only a hash is used. Ardimanov et al. [59] considered the problem of guessing with the use of an oracle, which the guesser can use before the guessing game begins. On the same line of work, Weinberger and Shayevitz [60] considered the problem where the guesser can receive information from a helper which saw the true word through a memoryless channel.

III. NOTATION AND PROBLEM STATEMENT

A. Basic Notation

Throughout, random variables will be denoted by capital letters, their realizations will be denoted by the corresponding lower case letters, and their alphabets will be denoted by calligraphic letters. Random vectors and their realizations will be denoted in the bold. For example, the random vector \( X = (X_1, \ldots, X_n) \) may take a specific vector value \( x = (x_1, \ldots, x_n) \) in \( X^n \), the \( n \)-th order Cartesian power of \( X \), which is the alphabet of each component of this vector. Sources will be denoted by the letters \( P \) or \( Q \). The expectation operator will be denoted by \( E \{ \cdot \} \), and \( \mathbb{I} \{ \cdot \} \) will denote the indicator function. The entropy of a distribution \( Q \) on \( X \) will be denoted by \( H_Q(X) \) where \( X \) designates a random variable drawn by \( Q \). \( D(Q||P) \) will denote the relative entropy between \( Q \) and \( P \). The rate-distortion function of a memoryless source \( Q \) at distortion level \( D \) will be denoted by \( R(D,Q) \).

\[ \exp_2 \{ \cdot \} \text{ denotes } 2^n \text{ and } \log \text{ denotes } \log_2. \] For two positive sequences \( \{ a_n \} \) and \( \{ b_n \} \), \( a_n = b_n \) will stand for equality on the exponential scale, that is, \( \lim_{n \to \infty} \frac{\log a_n}{b_n} = 0 \). Similarly, \( a_n \leq b_n \) means that \( \limsup_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} \leq 0 \), and so on. When both sequences depend on a vector \( x \in \mathcal{X}^n \), namely, \( a_n = a_n(x) \) and \( b_n = b_n(x) \), the notation \( a_n(x) = b_n(x) \) means that the asymptotic convergence is uniform, namely,

\[ \lim_{n \to \infty} \max_{x \in \mathcal{X}^n} \left| \frac{1}{n} \log \frac{a_n(x)}{b_n(x)} \right| = 0. \]

The empirical distribution of a sequence \( x \in \mathcal{X}^n \), which will be denoted by \( \hat{P}_x \), is the vector of relative frequencies \( \hat{P}_x(x) \) of each symbol \( x \in \mathcal{X} \) in \( x \). Information measures associated with empirical distributions will be denoted with ‘hats’ and subscripted by the sequences from which they are induced. For example, the entropy associated with \( \hat{P}_x \), which is the empirical entropy of \( x \), will be denoted by \( \hat{H}_x(X) \). Analogously, to stress out a distribution, e.g., \( Q \), under which an information measure is calculated, we use subscripts as well, e.g., \( \hat{H}_Q(X) \). With a slight abuse of notation, when clear from the context, we use similar notation for multi-variate measures, e.g., \( I_Q(Y;Y') \) denotes the mutual information under the joint distribution \( Q_{Y,Y'} = Q_Y Q_{Y'|Y} \), which will be abbreviated by \( Q \). Moreover, when the \( Y \)-marginal of such a distribution \( Q \) will be the empirical distribution of a specific \( y \), we will use the notation \( Q_y \) and not \( \hat{P}_y \), if the specific \( y \) is clear from the context.

The type class of \( x \in \mathcal{X}^n \), i.e., the set of all vectors in \( \mathcal{X}^n \) with empirical distribution \( \hat{P}_x \), will be denoted by \( \mathbf{T}(x) \). Similarly, \( \mathbf{T}(Q) \) will denote the type class of a specific (empirical) distribution \( Q \). E.g., \( \mathbf{T}(Q_{X|Y}|y) \) will denote the set vectors \( x \in \mathcal{X}^n \) whose relative frequencies equal \( Q_{X|Y} \) for the specific \( y \). As common in the literature, \( \mathcal{P}_n \) denotes the set of all types with denominator \( n \).

B. Guessing Subject to a Distortion Measure

Alice selects a random \( n \)-vector \( \vec{x} \), drawn from a finite alphabet source \( P \). Bob, which is unaware of the realization of \( X \), submits a sequence of guesses in the form of yes/no queries: “Is \( X \) close within distortion \( D \) to \( \vec{x}_1 \)?”, “Is \( X \) close within distortion \( D \) to \( \vec{x}_2 \)?”, and so on, where \( \vec{x}_i \in \mathcal{X}^n \) for each \( i \), until receiving a positive answer. Specifically, Alice gives a positive answer if Bob’s guess is within an average distortion \( D \) of Alice’s realization, that is

\[ \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i) \leq D, \]

where \( d: \mathcal{X} \times \hat{X} \to \mathbb{R}^+ \) is a single-letter distortion measure. We denote \( D_{\max} = \max_{x,\hat{x}} d(x, \hat{x}) \) and assume \( D_{\max} < \infty \). With a slight abuse of notation, we may write \( d(x, \hat{x}) \) for the \( n \)-letter distortion, that is, \( \sum_{i=1}^{n} d(x_i, \hat{x}_i) \). Given \( x \in \mathcal{X}^n \), let \( \mathcal{S}(x) = \{ \hat{x} : d(x, \hat{x}) \leq nD \} \). Clearly, for a given realization \( x \), Bob’s goal is to guess a sequence \( \hat{x} \in \mathcal{S}(x) \).

\(^1\)Similar to the common definition in rate-distortion theory, \( \hat{X} \), the reconstruction alphabet, may be different from the source alphabet \( X \). Clearly, the scenario in which \( \hat{X} = X \) is a special case.
When considering the deterministic setting, a guessing list, \( G_n \), is an ordered list of all members of \( X^n \), that is, \( G = \{ x_1, x_2, \ldots, x_{|X|^n} \} \), and it is associated with a guessing function, \( G(x) \), which is the function that maps \( X^n \) onto \( \{ 1, 2, \ldots, |X|^n \} \) by assigning to each \( x \in X^n \) the smallest integer \( i \) for which \( x_i \) satisfies the distortion constraint (1). Namely, \( G(x) \) is the number of guesses required until success, using \( G_n \), when \( X = x \). In this setting, the goal is to devise a guessing list \( G_n \) that minimizes a certain moment of \( G(X) \), namely, \( E(G^\rho(X)) \), where \( \rho > 0 \) is a given positive real.

In the randomized setting, on which we focus in this work, the guesser simply sequentially submits a sequence of random guesses, each one drawn independently according to a certain probability distribution \( \tilde{P}(x) \). This setting consumes less memory (compared to deterministic guessing which is independent of the source Alice has, the moment of the source models, will suffice to achieve the universality goals stated above. Specifically, for memoryless sources, we consider a sequence of randomized guesses, all drawn independently from the universal distribution,

\[
\tilde{P}(x) = \frac{\exp \{-nH(x)\}}{\sum_{z \in X^n} \exp \{-nH(z)\}},
\]

for all \( x \in X^n \). Note that this distribution satisfies \( \tilde{P}(x) \to 2^{-nH(x)} \), and was used in [17] for \( D = 0 \).

For sources with memory, we suggest a guessing distribution based on the incremental parsing procedure of LZ78 similar to [17]. The incremental parsing procedure of [34] (see also [61, Subsection 13.4.2]) is a sequential procedure for parsing a sequence, such that each new parsed phrase is the shortest string that has not been obtained before as a phrase. For example, the binary string

\[
\hat{x} = 01101101110010111
\]

is parsed as

\[
0, 1, 10, 11, 01, 110, 010, 111.
\]

For a string \( \hat{x} \), let \( c(\hat{x}) \) denote the number of such distinct phrases. In the example above, \( c(\hat{x}) = 8 \). In general, for a binary sequence of length \( n \), \( c(\hat{x}) \leq n/(1 - \epsilon_n) \log n \), where \( \epsilon_n \to 0 \) as \( n \to \infty \) ([62, Theorem 2]). Roughly speaking, when coding \( \hat{x} \) losslessly, one may simply describe, for each new phrase, the location of its previously appearing prefix, and the new symbol. This results in a code length of about \( c(\hat{x}) \log c(\hat{x}) \) bits. Specifically, let \( LZ(\hat{x}) \) be the \( LZ \) code length (in bits) of the sequence \( \hat{x} = \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n \). We define a probability distribution \( P \) over \( X^n \) as follows.

\[
P(\hat{x}) = \frac{2^{-LZ(\hat{x})}}{\sum_{\hat{z} \in X^n} 2^{-LZ(\hat{z})}}.
\]

Note that since \( LZ \) is a uniquely decodable code, we have \( \sum_{\hat{z} \in X^n} 2^{-LZ(\hat{z})} \leq 1 \) and hence \( P(\hat{x}) \geq 2^{-LZ(\hat{x})} \).
While the distribution in (3) is conceptually more complex than that in (2), it is, in fact, easily approximated without directly computing the exponentially-long sum in the denominator, but, rather, by feeding an LZ decoder with uniform i.i.d. bits. See [17, Section V.D] for algorithms to implement (3) directly and efficiently. Moreover, it is also superior to (2) in the sense of always assigning (asymptotically) a higher probability. Specifically, by Ziv’s inequality ([34, Theorem 1]; see [61, eq. (13.125)] for the version below), for any \( \nu \) to maximizes \( \prod_{i=1}^{n} M_0(\hat{x}_i) \), which is the empirical distribution of \( \hat{x} \). We thus have

\[
\frac{c(\hat{x}) \log c(\hat{x})}{n} \leq -\frac{1}{n} \log \prod_{i=1}^{n} \hat{P}_\nu(\hat{x}_i) + \epsilon_k(n),
\]

where \( \epsilon_k(n) \to 0 \) as \( n \to \infty \). This is, a fortiori, true for any memoryless distribution \( M_k \), and thus also for the one which maximizes \( \prod_{i=1}^{n} M_0(\hat{x}_i) \), which is the empirical distribution of \( \hat{x} \). We thus have

\[
\frac{c(\hat{x}) \log c(\hat{x})}{n} \leq -\frac{1}{n} \log \prod_{i=1}^{n} \hat{P}_\nu(\hat{x}_i) + \epsilon_1(n).
\]

As a result

\[
\exp_2 \{-LZ(\hat{x})\} \geq \exp_2 \left\{ -n \left( \frac{1}{n} \log \prod_{i=1}^{n} \hat{P}_\nu(\hat{x}_i) + \epsilon_1(n) \right) \right\} = \exp_2 \{-nH_\nu(\hat{x})\}. \tag{4}
\]

To conclude the discussion, we also derive an extension of Ziv’s inequality, tailored for block–memoryless distributions which will be used in the sequel. It is proved in Appendix A.

Claim 1: Assume a fixed integer \( K, k < K \), and that \( K+k \) divides \( n \). Denote \( m = \frac{n}{K+k} \). For any sequence \( \hat{x} \in \hat{X}^n \) and any distribution \( M^{K+k}(\cdot) \) on \( \hat{X}^{K+k} \), the LZ code length (in bits),\( LZ(\hat{x}) \), satisfies

\[
LZ(\hat{x}) \leq -\log \left[ \prod_{i=1}^{m} M^{K+k}(\hat{x}^{i(K+k)} \hat{x}^{i+1}(K+k+1)) \right] + \nu_{K+k}(n),
\]

where \( \nu_{K+k}(n) \) depends on \( |\hat{X}| \) yet satisfies \( \lim_{n \to \infty} \frac{1}{n} \nu_{K+k}(n) = O \left( \frac{1}{n} \right) \).

IV. MAIN RESULTS

In this section, we summarize the main results in this paper. We begin with memoryless sources. Then, we provide the results for guessing through a noisy channel, followed by the results for sources with memory. Finally, we conclude with results for guessing an individual sequence.

A. Memoryless Sources

Theorem 1: For any memoryless source \( P \) over \( X \), if sequences are selected independently at random according to the distribution \( \hat{P}(\hat{x}) = 2^{-nH_\nu(\hat{x})} \), then, the \( \rho \)-th guesswork moment, subject to a distortion constraint \( D \), satisfies

\[
E \{ G^\rho \} \geq \exp_2 \left\{ n \max_Q [\rho R(D,Q) - D(Q||P)] \right\}. \tag{5}
\]

The proof of Theorem 1 appears in Section V.A. Note that the theorem meets the exponential order of the lower bound in [11]. Hence, guessing using the distribution in Theorem 1 is asymptotically optimal, and we have the following.

Corollary 1: For any memoryless source \( P \) over \( X \), the best achievable \( \rho \)-th guesswork moment, subject to a distortion constraint \( D \), is achievable by randomized guessing and we have

\[
\lim_{n \to \infty} \frac{1}{n} \log E \{ G^\rho \} = \max_Q [\rho R(D,Q) - D(Q||P)].
\]

Clearly, the exponent in Corollary 1 is similar to that in [11], yet it can be achieved in a randomized manner, with no memory (to hold a list) or synchronization.

B. Noisy Guessing for Memoryless Sources

We consider the following scenario, studied first in [51] for \( D = 0 \). Alice draws a random \( n \)-vector, \( Y = (Y_1, \ldots, Y_n) \), from a discrete memoryless source (DMS), \( P \), of a finite alphabet, \( Y \). Bob, who is unaware of the realization of \( Y \), sequentially submits to Alice a (possibly, infinite) sequence of guesses, \( x_1, x_2, \ldots \), where each \( x_i \) is a vector of length \( n \), whose components take on values in a finite alphabet, \( X \). Before arriving to Alice, each guess, \( x_i \), undergoes a discrete memoryless channel (DMC), defined by a matrix of single–letter input–output transition probabilities, \( W = \{ W(y|x), x \in X, y \in Y \} \). Let \( Y_1, Y_2, \ldots \) be the corresponding noisy versions of \( x_1, x_2, \ldots \), after being corrupted by the DMC, \( W \). Alice sequentially examines the noisy guesses and she returns to Bob an affirmative feedback upon the first match within distortion \( D \), that is, \( d(Y, Y_i) \leq nD \). Clearly, the number of guesses, \( G \), until the first successful guess, is a random variable that depends on the source vector \( Y \) and the guesses, \( Y_1, Y_2, \ldots \). It is given by

\[
G = G(Y, Y_1, Y_2, \ldots) = \infty \sum_{k=1}^{\infty} k \cdot I\{ d(Y, Y_k) \leq nD \} \cdot \prod_{i=1}^{k-1} \left[ 1 - I\{ d(Y, Y_i) \leq nD \} \right].
\]

For a given list of guesses, \( \mathcal{G}_n = \{ x_1, x_2, \ldots \}, x_i \in X^n, i = 1, 2, \ldots \), the \( \rho \)-th moment of \( G \) is given by

\[
E_{\mathcal{G}_n} \{ G^\rho \} = \sum_{y \in Y^n} P(y) \cdot \sum_{k=1}^{\infty} k^\rho \cdot W(S(y)|x_k) \cdot \prod_{i=1}^{k-1} [1 - W(S(y)|x_i)], \tag{6}
\]

where

\[
W(y|x) = \prod_{l=1}^{n} W(y_l|x_l)
\]

and \( S(y) = \{ y' : d(y, y') \leq nD \} \). Randomized guessing lists (where the deterministic guesses, \{ \( x_i \) \}, are replaced by random ones, \{ \( X_i \) \}) are allowed as well. In this case, eq. (6) would include also an expectation w.r.t. the randomness of the guesses.
Let \( P, W, \) and \( \rho \) be given. For two given distributions, \( Q_X \) and \( Q_Y, \) defined on \( X \) and \( Y, \) respectively, define
\[
\Gamma(Q_X, Q_Y) = \inf\{ D(\tilde{Q}_Y|X||W|Q_X) : (Q_X \odot \tilde{Q}_Y|X)_Y = Q_Y \},
\]
where
\[
D(\tilde{Q}_Y|X||W|Q_X) = \sum_{x \in X} Q_X(x) \sum_{y \in Y} \tilde{Q}_Y(x|y|x) \log \frac{\tilde{Q}_Y(x|y|x)}{P_Y(x|y|x)}
\]
and the notation \((Q_X \odot \tilde{Q}_Y|X)_Y = Q_Y\) means that the \( Y\)-marginal induced by the given \( Q_X \) and by \( \tilde{Q}_Y|X \) is constrained to be the given \( Q_Y, \) i.e., \( \sum_{x \in X} Q_X(x) \tilde{Q}_Y(x|y|x) = Q_Y(y) \) for all \( y \in Y. \) Next, define
\[
\Gamma(Q_Y) = \inf_{Q_X} \Gamma(Q_X, Q_Y) = \inf_{Q_X} D(\tilde{Q}_Y|X||W|Q_X).
\]
Finally, we define
\[
E_W(\rho) = \sup_{Q_Y} \left\{ \rho \tilde{R}_W(D, Q_Y) - D(\tilde{Q}_Y|X||P) \right\},
\]
where
\[
\tilde{R}_W(D, Q_Y) \triangleq \inf_{\{Q_{Y'}|Y'\} \subseteq I_Q(Y;Y') + \Gamma(Q_{Y'})} \sup_{\gamma \in \mathcal{M}(\gamma)} \left[ \sum_{X \in X} Q_X(x) \tilde{Q}_Y(x|y|x) \right].
\]
We argue that \( E_W(\rho) \) is the best achievable guessing exponent.

**Theorem 2:** Assume Alice draws a sequence of length \( n \) from a memoryless source \( P \) over \( X, \) and Bob’s guesses, before arriving to Alice, undergo a discrete memoryless channel (DMC) \( W = \{W(y|x), x \in X, y \in Y\}. \) Then, the optimal guessing \( \rho \)-th moment exponent at distortion level \( D \) is achievable using a randomized guessing strategy, which is universal in \( P, W, d(\cdot, \cdot), D \) and \( \rho, \) and satisfies
\[
\lim_{n \to \infty} \frac{1}{n} \log E[G^\rho] = E_W(\rho).
\]
The proof of Theorem 2 is given in Section V-B.

Similar to (5), for the clean memoryless case, the exponent in (7) has a clear structure of a rate–distortion function, minus the divergence between the distribution for which the function is computed, and the original distribution. The rate–distortion function, however, is not simply the minimal value of the mutual information subject to the distortion constraint, but, rather, includes a correction term, \( \Gamma, \) to reflect the penalty the guesser suffers due to the noisy channel. In fact, this is the same correction term as in [51], when no distortion is allowed.

The exponent in (7) has two alternative forms, which facilitate numerical evaluation and give an operational meaning.

**Lemma 2:** The function \( E_W(\rho) \) in (7) has the following two alternative forms.
\[
E_W(\rho) = \sup_{s \geq 0} \inf_V \left\{ \log \left( \sum_y P(y) \frac{V(x)}{\sum_x V(x)U(s(x,y))} \right) - \rho sD \right\},
\]
where the infimum on \( V \) is over all probability distributions
\[
\sum_x V(x)W(y|x) \delta_{sd(y,y')} = M(y') \Gamma(D,P,Q_Y).\]

A stationary source is said to be \( \Psi \)-mixing if \( \Psi(k) \to 0 \) as \( k \to \infty. \)

Before presenting the main result for mixing sources, a few additional definitions are in order. For any positive integer \( n, \) denote by \( R_n(D, P^n) \) the rate–distortion function of a block–memoryless source of block size \( n \) and a distribution \( P^n. \)

**Remark 1 (A Sufficient Condition for No–Noise–Penalty):** It is interesting to identify the scenario where there is no loss due to the channel \( W. \) If the test channel, \( Q_{Y'|Y}, \) that achieves the ordinary rate–distortion function, \( R(D, Q_Y), \) happens to induce a marginal distribution \( Q_{Y'|Y} \) such that \( \Gamma(Q_{Y'|Y}) = 0 \) for any \( Q_Y \) (or, at least, for the dominant one), then the guessing exponent is the same as in the clean–channel case.

**C. Sources With Memory**

Our model for sources with memory will be that of \( \Psi \)-mixing sources. Consider a stationary source \( \{X_m\}_{m=1}^{\infty} \) over a probability space \((\Omega, \mathcal{F}, P). \) Let \( \mathcal{F}_j \) denote the \( \sigma \)-field of events generated by the random variables \( \{X_m\}_{1 \leq m \leq j}. \) A measure of dependence, \( \Psi(k), \) for the source \( \{X\} \) is defined by
\[
\Psi(k) = \sup_{A \in \mathcal{F}_0, B \in \mathcal{F}_k, P(A) > 0, P(B) > 0} \left| \frac{P(A \cap B)}{P(A)P(B)} - 1 \right|.
\]
over $\mathcal{X}^n$, at a per–letter distortion level $D$ (that is, $nD$ per-block). Note that we do not normalize the rate by $n$, hence $R_n(D, P^n)$ represents the number of bits required per block of size $n$. Now, for any stationary source $P$ with a marginal distribution $P^n$ define the guessing exponent at distortion level $D$ and moment $\rho$ as

$$E_P(D, \rho) = \liminf_{n \to \infty} \frac{1}{n} \log \sup_{Q^n} \left[ \rho R_n(D, Q^n) - D \left(Q^n || P^n\right) \right].$$

Remark 2: For each $n$, $\sup_{Q^n} \frac{1}{n} R_n(D, Q^n)$ is monotonic in $D$ and hence continuous everywhere with the possible exception of countably many points. Thus, $\lim_{n \to \infty} \sup_{Q^n} \frac{1}{n} R_n(D, Q^n)$ has the same property.

The main result for sources with memory is the following theorem.

Theorem 3: For any stationary $\Psi$–mixing source $P$, the optimal guesswork $\rho$–th moment exponent at distortion level $D$ is achievable using a randomized guessing strategy, which is universal in $P, d(\cdot, \cdot), D$ and $\rho$, and satisfies

$$\limsup_{n \to \infty} \frac{1}{n} \log E\{G^\rho\} = E_P(D, \rho).$$

The randomized guessing strategy is based on the incremental parsing procedure of LZ78. It was described in Section III–C. The complete proof of Theorem 3 is given in Section V–C.

D. Guessing an Individual Sequence

In [48], the guessing game was extended to guessing an individual sequence, that is, a single, deterministic sequence without any underlying probability model. In such a setup, it is unrealistic to measure performance against any guessing strategy, and one tries to compete against any finite–state guessing machine. This is, in fact, the common practice in many other setups involving individual sequences, including compression, gambling, prediction, denoising and encryption [34, 63]–[68].

A finite–state guessing machine is defined by the tuple $(U, \mathcal{X}, Z, \Delta, f, g)$, where $U$ is its input alphabet, assumed binary, $\mathcal{X}$ is the output alphabet, $Z$ is a finite set of states, with $|Z| = s$, $\Delta : Z \to \{0, 1, 2, \ldots \}$ defines the number of input bits used at each state, and $f, g$ are the output and next–state functions, that is, $f : Z \times U^* \to \mathcal{X}$ and $g : Z \times U^* \to Z$.

When a binary sequence $u_1, u_2, \ldots, u_l \in U$, is fed to the guessing machine, the output is a sequence $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_l \in \mathcal{X}$, according to the following recursive equations:

$$t_i = t_{i-1} + \Delta(z_i), \quad t_0 = 0,$$

$$v_i = (u_{t_{i-1}+1}, u_{t_{i-1}+2}, \ldots, u_t),$$

$$\hat{x}_i = f(z_i, u_i),$$

$$z_{i+1} = g(z_i, u_i).$$

The guessing game is thus as follows: Alice has an individual sequence $x$ of length $n$. Bob is equipped with a finite–state guessing machine and unlimited number of uniform i.i.d. Bits. The machine consumes bits, while passing through states in $Z$ and producing output symbols in $\mathcal{X}$. When the output is of length $n$ Bob submits it to Alice and asks “is the sequence $\hat{x}$ within distortion $nD$ of $x$?” If the answer is affirmative, the game ends. Otherwise, the machine is restarted to its initial state (say, some $z_1 \in Z$) and the game repeats.

For a finite–state guessing machine $F$, denote by $G_F(x)$ the (random, due to the random input to the machine) number of guesses until $x$ is guessed with distortion at most $nD$. Our main result in this context is the following theorem, which, on the one hand, lower bounds the $\rho$–th moment of $G_F(x)$, yet, on the other, states that this lower bound is asymptotically achievable with a guessing distribution which can be implemented with a finite–state machine whose number of states is independent of the length $n$.

To present the result, we first remind the reader of the notion of finite–state compressibility [34]. We briefly follow the notation in [48, Section III.A]. Let $L_E(b^n)$ be the total length of the binary strings $b_1, \ldots, b_n$ produced by an information lossless, finite–state encoder $E$ with at most $K$ states as it processes the string $x$. The compression ratio of $E$ is defined by $\rho_E(\hat{x}) = \frac{L_E(x)}{n}$, and the finite–state compressibility of $\hat{x}$ is defined by

$$\rho_K(\hat{x}) = \min_{E \in E(K)} \rho_E(\hat{x}),$$

where $E(K)$ is the set of all information lossless finite–state encoders with at most $K$ states.

Theorem 4: Denote by $E\{G_F^\rho|x\}$ the $\rho$–th moment of the guesswork for a finite–state machine $F$ with at most $s$ states, given the individual sequence $x$. Then:

Converse part. (i) For any fixed $s$ independent of $n$ we have

$$E\{G_F^\rho|x\} \geq \left[ \sum_{\{d(\hat{x}, x) \leq nD\}} 2^{-LZ(\hat{x})} \right]^{-\rho}. \quad (11)$$

(ii) Alternatively, we have

$$E\{G_F^\rho|x\} \geq \left[ \sum_{\{d(\hat{x}, x) \leq nD\}} \exp_2 \left\{ -n \left( \rho_K(\hat{x}) - \frac{\log(2s^3e)}{l} \right) \right\} \right]^{-\rho}$$

for any $l$ which divides $n$ and $K(l)$ denoting the number of states of a machine implementing a Shannon code on blocks of size $l$.

Direct part. (i) The lower bound in (11) is achieved using the guessing distribution (3). (ii) Assume $l$ divides $n$ and consider the distribution

$$\hat{P}(\hat{x}) = \prod_{i=0}^{n/2l-1} \frac{2^{-LZ(\hat{x}_{2i+1})}}{\sum_{\hat{x} \in \mathcal{X}^l} 2^{-LZ(\hat{x})}}, \quad (12)$$

which is implementable with a finite–state machine with at most $l|\mathcal{X}|^l$ states for any $n$. Then

$$E\{G_{F,LZ}^\rho|x\}$$

$$\leq \left[ \sum_{\{d(\hat{x}, x) \leq nD\}} \exp_2 \left\{ -n \left( \rho_K(\hat{x}) + \frac{\log(4K^2)}{(1 - \epsilon(l)) \log l} + \frac{K^2 \log(4K^2)}{l} + \epsilon(l) \right) \right\} \right]^{-\rho},$$

where $\epsilon(l) \to 0$ as $l \to \infty$. 


The complete proof of Theorem 4 is given in Section V-D.

At this point, a few remarks are in order. First, note that the universal guessing distribution in (3) is thus not only five-fold universal, it is also asymptotically optimal for four different setups: guessing memoryless sources, guessing memoryless sources through a noisy channel, guessing sources with memory and guessing an individual sequence. To see this, note that the memoryless sources in the setup of Theorem 1 are, in fact, Ψ-mixing, hence Theorem 3 holds for this case directly. However, as mentioned in Section III-C, the distribution in (3) is more capable in a very wide sense, and thus Theorem 2 also holds if one guesses according to (3) rather than (2). The details are given in Remark 3 in the sequel, when the details of the direct part of Theorem 2 are clear. Finally, while the universal distribution in (3) is not implementable with a finite-state machine as \( n \) grows without bound, it has both efficient algorithms which allow easily sampling from this distribution ([17, Section V.D]), as well as a block-wise version (12) which still asymptotically achieves the lower bound.

V. PROOFS

When proving direct (achievability) results, the focus is on designing good guessing distributions, and then analysing the expected guesswork under the relevant source distribution. Thus, roughly speaking, a direct result includes three steps: introducing a guessing distribution \( \hat{P}(x) \), evaluating the probability of hitting the distortion-achieving set, \( \hat{P}[S(x)] \), and, finally, taking into account the actual source distribution \( P \), computing the guesswork moment, \( E\{G^p\} = E_{\hat{P}}[E\{G^p|X\}] \).

Evaluating \( \hat{P}[S(x)] \) for the three probabilistic scenarios is a key technical challenge. Indeed, the proofs of the direct results for memoryless sources, noisy guessing and sources with memory are based on three lemmas (Lemma 3, Lemma 4 and Lemma 6, respectively), which evaluate (or at least lower bound) \( \hat{P}[S(x)] \) for the three scenarios. In fact, in the first and second scenarios, these lemmas will almost directly result in the achievability proofs. For sources with memory, the analysis of the achievability is more involved and will require additional steps. Interestingly, the three lemmas also have a similar structure, relating the exponential rate of the hitting probability to some rate–distortion–like function, depending on the scenario considered.

In the individual sequences scenario, however, we do not give a rate–distortion interpretation to \( \hat{P}[S(x)] \), but, rather, focus on a converse result, lower bounding the guesswork moment for \( x \) using \( LZ(\hat{x}) \), which then easily matches the direct result.

A. Proof of Theorem 1

As mentioned, we first evaluate \( \hat{P}[S(x)] \) for the distribution \( \hat{P} \) given in (2).

Lemma 3: Assume sequences of length \( n \) are chosen independently at random according to the distribution \( \hat{P}(x) \equiv 2^{-n\bar{H}(x)} \) in (2). Then, for any \( x \in \mathcal{X}^n \)
\[
\hat{P}[S(x)] \leq 2^{-nR(D, \hat{P}_x)},
\]
where \( R(D, \hat{P}_x) \) is the rate–distortion function of \( \hat{P}_x \).

Proof:
\[
\hat{P}[S(x)] = \sum_{\{\hat{x}: d(\hat{x}, x) \leq nD\}} \hat{P}(\hat{x}) = \sum_{\{T(Q_{\hat{x}|X}|x)\leq n\bar{H}(\hat{x})\}} |T(Q_{\hat{x}|X}|x)| \cdot 2^{-n\bar{H}(\hat{x})}(
\hat{x})
\]

\[
\leq \max_{\{Q_{\hat{x}|X}: E_Q(d(X,X)\leq D)\}} 2^nH_Q(X|X), 2^{-nH_Q(\hat{x})}
\]

\[
= \max_{\{Q_{\hat{x}|X}: E_Q(d(X,X)\leq D)\}} 2^{-nI_Q(X;\hat{x})}
\]

\[
= \exp_2 \left\{-n \min_{\{Q_{\hat{x}|X}: E_Q(d(X,X)\leq D)\}} I_Q(X;\hat{x}) \right\}
\]

\[
= 2^{-nR(D, \hat{P}_x)},
\]
where the last equality is since the minimum is taken over all conditional types \( Q_{\hat{x}|X} \), conditioned on the specific \( x \), satisfying the distortion constraint.

We can now complete the proof of Theorem 1. We have,
\[
E\{G^p\} = \sum_x P(x) \cdot E\{G^p|x\}
\]

\[
= \sum_x 2^{-n(\bar{H}(x) + D(\hat{P}_x||P))} \cdot 2^{nR(D, \hat{P}_x)}
\]

\[
= \sum_{Q \in \mathcal{P}_n} |T_Q| \cdot \sum_{Q \in \mathcal{P}_n} 2^nH_Q(X), 2^{-nH_Q(\hat{x})} \cdot 2^{nR(D, Q)}
\]

\[
= \exp_2 \left\{n \max_Q [\rho R(D, Q) - D(Q||P)] \right\},
\]
where (14) is due to Lemma 1 and Lemma 3.

B. Proof of Theorem 2

1) Direct: Analogously to Lemma 3, we now wish to evaluate the probability of \( S(y) \) under the guessing distribution we use. However, since guessing is done through a noisy channel, each guess induces a distribution on the reconstruction words. A fortiori, when using a guessing distribution. The following lemma evaluates the probability of \( S(y) \) under the distribution induced by the guessing distribution (2).

Lemma 4: Assume that sequences of length \( n \) are chosen independently at random according to the distribution \( \hat{P}(x) \equiv 2^{-n\bar{H}(x)} \) in (2), then sent through the memoryless channel \( W(y|x) \). Denote by \( I^n \) the \( n \)-letter distribution induced on the \( y \)-vectors. Then,
\[
I^n[S(y)] = \exp_2\{-n\hat{R}_W(D, \hat{P}_y)\},
\]
where \( \hat{R}_W(D, \hat{P}_y) \) is given in (8).

Proof: Guessing using the distribution in (2) induces the following distribution on the \( y \)-vectors:
\[
I^n(y)
\]

\[
\sum_{x \in \mathcal{X}^n} \exp_2\{-n\bar{H}(x)\} \cdot W(y|x)
\]

(17)
\[ \sum_{T(Q_X|Y)} |T(Q_X|Y)| \cdot \exp_{2}\{-n[H_Q(X) + H_Q(Y|X) + D(Q_Y|X) W(Q_X)]\} \]
\[ \leq \max_{Q_{Y|X}} \exp_{2}\{n [H_Q(Y|X) - H_Q(X) - H_Q(Y|X) - D(Q_Y|X) W(Q_X)]\} \]
\[ = \exp_{2}\{-nH_Q(Y)\} \cdot \max_{Q_{Y|X}} \exp_{2}\{-nD(Q_Y|X) W(Q_X)\} \]
\[ = \exp_{2}\{-n(H_Q(Y) + (Q(Y))}\}. \]

Note that \( Q \) above is the distribution whose \( Y \)-marginal is \( \hat{P}_y \), hence \( H_Q(Y) \) is, in fact, \( \hat{H}_Q(Y) \). Thus, given that \( Y = y \), the probability of a single successful guess within distortion \( D \) is given by
\[ \sum_{\{y': d(y,y') \leq nD\}} \pi^n(y') \]
\[ = \sum_{\{y': d(y,y') \leq nD\}} \exp_{2}\{-n[H_Q(Y') + (Q(Y'))]\} \]
\[ = \max_{Q_{Y|Y}} \exp_{2}\{n H_Q(Y')\} \cdot \exp_{2}\{-nH_Q(Y') + (Q(Y'))\} \]
\[ = \exp_{2}\{-n \\min_{Q_{Y|Y}}\{I_Q(Y;Y') + (Q(Y'))\}\} \]
\[ = \exp_{2}\{-n \hat{R}_W(D, Q_Y)\}. \]

Remembering that in the above \( Q_Y \) is empirical distribution of the specific \( y \), that is, \( \hat{P}_Y \), completes the proof. \( \square \)

The direct result for noisy guessing under distortion now easily follows. This time, applying Lemma 1 on the result in Lemma 4, we have
\[ E\{G^n|Y = y\} = \exp_{2}\{n \hat{R}_W(D, \hat{P}_Y)\}. \]

Pursuing the same steps as in (13)–(15), we have
\[ E\{G^n\} = \sum_y P(y) \cdot E\{G^n|Y = y\} \]
\[ = \exp_{2}\{n \max_{Q_Y} \hat{R}_W(D, Q_Y) - D(Q_Y|P)\}, \]
which completes the proof of the direct.

Remark 3: Note that guessing according to the universal distribution (3), rather than (2), will yield an equivalent direct result. To see this, note that the guessing distribution is used only in (17). By (4), changing this distribution to (3) will give
\[ I^n(y) \geq \sum_{x \in \mathbb{X}^n} \exp_{2}\{-nH(x)\} \cdot W(y|x) \]

The rest of the direct will continue verbatim, resulting in an upper bound which meets the converse.

2) Converse: We begin from a simple preparatory lemma.

Lemma 5: Let \( T(Q_Y) \) be a given type class of \( y \)-vectors, and let \( x \in T(Q_Y) \) be given. Then,
\[ \frac{1}{|T(Q_Y)|} \sum_{y \in T(Q_Y)} W[S(y)|x] \leq \exp_{2}\{-n \hat{R}_W(D, Q_Y)\}. \]

Note that the lemma above is, in a sense, a matching upper bound to the result in Lemma 4. However, while Lemma 4 assumes that the vector \( x \) is drawn from the universal distribution used in the direct results, Lemma 5, which is later utilized in the converse result, does not assume any distribution on \( x \). Yet, it includes averaging over \( y \in T(Q_Y) \).

Proof:
\[ \frac{1}{|T(Q_Y)|} \sum_{y \in T(Q_Y)} W[S(y)|x] \]
\[ = \frac{1}{|T(Q_Y)|} \sum_{y \in T(Q_Y)} \sum_{d(y,y') \leq nD} W(y'|x) \]
\[ = \frac{1}{|T(Q_Y)|} \sum_{T(Q_{Y|Y'})} \sum_{d(y,y') \leq nD} |T(Q_{Y|Y'}|x)| \]
\[ = \exp_{2}\left\{-n \\min_{Q_{Y|Y'}}\{I_Q(Y;Y') + (Q(Y'))\}\right\} \]
\[ - H_Q(Y, Y'|X) + H_Q(Y'|X) + D(Y|X) W(Q_X) \]
\[ = \exp_{2}\left\{-n \\min_{Q_{Y|Y'}}\{I_Q(Y;Y')\right\} \]
\[ + D(Q_Y|X) W(Q_X) \}
\[ \leq \exp_{2}\left\{-n \\min_{Q_{Y|Y'}}\{I_Q(Y;Y')\}ight\} \]
\[ + \max_{Q_{Y|Y}:Q_{Y|Y'}=Q_V} D(Q_Y|X) W(Q_X) \}
\[ = \exp_{2}\left\{-n \\min_{Q_{Y|Y'}}\{I_Q(Y;Y')\}ight\} \]
\[ + \Gamma(Q_X, Q_Y) \}
\[ \leq \exp_{2}\left\{-n \\min_{Q_{Y|Y'}}\{I_Q(Y;Y') + (Q(Y'))\}\right\} \]
\[ = \exp_{2}\{-n \hat{R}_W(D, Q_Y)\}. \]

\( \square \)
of \( y, Q_Y \), will be defined later. We have the following chain of inequalities, whose explanations are given below.

\[
E\{G^\rho\} = \sum_{Q_Y} P[\mathcal{T}(Q_Y)] E\{G^\rho|y \in T(Q_Y)\}
\]

\[
= \sum_{Q_Y} 2^{-nD(Q_Y)\rho} E\{G^\rho|y \in T(Q_Y)\}
\]

\[
\geq \sum_{Q_Y} 2^{-nD(Q_Y)\rho} k(Q_Y)^\rho \cdot \Pr\{G > k(Q_Y)|y \in T(Q_Y)\}
\]

\[
= \sum_{Q_Y} 2^{-nD(Q_Y)\rho} k(Q_Y)^\rho \sum_{y \in T(Q_Y)} \frac{1}{|T(Q_Y)|} \Pr\{G > k(Q_Y)|y\}
\]

\[
\geq \sum_{Q_Y} 2^{-nD(Q_Y)\rho} k(Q_Y)^\rho \left[1 - \sum_{i=1}^{k(Q_Y)} \frac{1}{|T(Q_Y)|} W(S(y)|x_i)\right]
\]

\[
\geq \sum_{Q_Y} 2^{-nD(Q_Y)\rho} k(Q_Y)^\rho \left[1 - \sum_{i=1}^{k(Q_Y)} 2^{-n\tilde{R}_W(D,Q_Y)}\right]
\]

\[
= \sum_{Q_Y} 2^{-nD(Q_Y)\rho} k(Q_Y)^\rho \left[1 - k(Q_Y)\right] 2^{-n\tilde{R}_W(D,Q_Y)}
\]

\[
= \sum_{Q_Y} 2^{-nD(Q_Y)\rho} k(Q_Y)^\rho \left[1 - k(Q_Y)\right] 2^{-n\tilde{R}_W(D,Q_Y)}
\]

\[
= \exp_2 \left\{ n \max_{Q_Y} \left[\rho \tilde{R}_W(D,Q_Y) - D(Q_Y\|P) - \rho \epsilon\right]\right\}
\]

In the above chain, (18) is since \( EGP^\rho = \sum_{k=1}^{\infty} k^\rho \cdot \Pr\{G = k\} \geq k^\rho \cdot \Pr\{G > k\} \), (19) is since \( Y \) is drawn from a DMS, hence the conditioning \( Y \in T(Q_Y) \) means that \( Y \) is uniformly distributed within \( T(Q_Y) \), which explains the uniform averaging across \( T(Q_Y) \). (20) is since \( \Pr\{G > k|y\} = 1 - \Pr\left\{ \bigcup_{i=1}^{k(Q_Y)} \text{guess i is successful}\right\} \geq 1 - \sum_{i=1}^{k(Q_Y)} \Pr\{\text{guess i is successful}|y\} \), and \( \Pr\{\text{guess i is successful}|y\} = W(S(y)|x_i) \). (21) is due to Lemma 5. (22) is by choosing \( k(Q_Y) = \left[2^n(\tilde{R}_W(D,Q_Y) - \epsilon)\right] \) for some arbitrarily small \( \epsilon \). Since the above is true for any \( \epsilon > 0 \), we have

\[
E\{G^\rho\} \geq \exp_2 \left\{ n \max_{Q_Y} \left[\rho \tilde{R}_W(D,Q_Y) - D(Q_Y\|P)\right]\right\}
\]

which completes the proof of the converse.

### C. Proof of Theorem 3

In the converse, we derive a lower bound which holds for any guessing strategy, fixed or random. Whether Bob uses a fixed list prepared in advance, or a certain distribution to draw sequences from, eventually the result of this process is submitting queries one after the other. If Bob uses a distribution, then any time the game starts his queries might form a different list, and each such list might have a different probability of being used, yet eventually, for each game and any distribution Bob uses, nature draws a list for him to use.

Hence, a lower bound which holds for any list, holds for both deterministic as well as randomized guessing strategies.

In the direct, on the other hand, we show there exists a strategy which asymptotically achieves the converse. We do this by suggesting a randomized strategy, due to its many advantages.

1) **Converse:** For a converse theorem, we do not assume that the guesser is restricted to any randomized guessing strategy. In fact, a guesser is allowed to use a guessing list, \( G_n \), which is an ordered list of all members of \( \mathcal{X}^n \). That is, \( G_n = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{G_n}\} \). In the list, each element of \( G_n \) is associated with a guessing function, \( G(x) \), which is the function that maps \( \mathcal{X}^n \) to the integers, by assigning each \( x \in \mathcal{X}^n \) the smallest integer \( k \) for which \( d(x, \hat{x}_k) \leq nD, \hat{x}_k \in G_n \), namely, \( \hat{x}_k \) is the first element of \( G_n \) which is in \( S(x) \). Thus, \( G(x) \) is the number of guesses required until successfully meeting the distortion criteria, using \( G_n \), when \( X = x \). Note that if Bob does use a randomized strategy, any realization of this (random) process will eventually result in a list of queries which Bob will submit. While each realization (list) may be different, since the lower bound we derive in herein applies for any list, it applies for any realization, hence definitely applies for the average over all such realizations, regardless of the distribution from which they were drawn. Note also that for finite alphabet, and assuming that for all \( x \in \mathcal{X} \) we have \( d_{\min} = \min_{\hat{x} \in \mathcal{X}} d(x, \hat{x}) = 0 \), one can easily append any list \( G_n \) such that the range of \( G(x) \) is finite.

Thus, for a converse result, we wish to show that for any source \( P \) with a \( K \)-letter marginal \( P^K \), the guesswork \( \rho \)-th moment exponent at distortion level \( D \) is lower bounded by

\[
\liminf_{K \to \infty} \inf_{Q^K} \frac{1}{K} \log E_{P^K}\{G^\rho(X)\} \geq \liminf_{K \to \infty} \sup_{Q^K} \frac{1}{K} \left[\rho R_K(D,Q^K) - D(Q^K\|P^K)\right],
\]

where \( P^K \) is any distribution on \( X^K \). Note that the expectation in the l.h.s. of (23) is only with respect to the source distribution, \( P^K \), as the bound holds for any deterministic list.

To this end, following [11], the key tool in the converse is to view \([ -\log G(x) ]\) as the length of a rate–distortion code for \( x \in \mathcal{X}^K \). Specifically, assume a guesser indeed has a list \( G \) such that for any \( x \in \mathcal{X}^K \), there is at least one item in the list, say \( \hat{x}_k \) such that \( d(x, \hat{x}_k) \leq K \). Then, this list can be viewed as a rate–distortion codebook. To encode a word \( x \), the encoder will simply use the smallest integer \( k \) for which \( d(x, \hat{x}_k) \leq K \), with an appropriate code for the integers. A decoder will simply output \( \hat{x}_k \). Consider the following distribution on the
positive integers:

\[ p_i = \frac{c(\delta)}{i^{1+\delta}} \]  

(24)

where \( \delta > 0 \) and

\[ c(\delta) = \left[ \sum_{i=1}^{\infty} \frac{1}{i^{1+\delta}} \right]^{-1}. \]

Define \( I^K_i = \{ x : G(x) = i \} \). Using the Shannon code for the distribution in (24), for any distribution \( Q^K \) on \( \mathcal{X}^K \), we have

\[
E_{Q^K} l(X) = \sum_i Q^K(i) \log p_i 
\leq 1 - \sum_i Q^K(i) \log c(\delta) + (1+\delta) \sum_i Q^K(i) \log i 
= 1 - \log c(\delta) + (1+\delta) \sum_x Q^K(x) \log G(x).
\]

Hence, provided that \( E_{Q^K} l(X) \) is the expected length of a rate–distortion code for blocks of length \( K \) of the source \( Q^K \) at a (per–letter) distortion level \( D \), we have,

\[
\sum_x Q^K(x) \log G(x) \geq \frac{E_{Q^K} l(X) - 1 + \log c(\delta)}{1 + \delta}
\]

(25)

where (25) follows from Jensen’s inequality. Since the above holds for any \( Q^K \), we have, for any \( \delta > 0 \),

\[
\inf_{Q^K} \frac{1}{K} \log E_{Q^K} \{ G^P(X) \}
\geq \sup_{Q^K} \frac{1}{K} \left[ \frac{\rho}{1 + \delta} R_K(D, Q^K) - D(Q^K\| P^K) + O \left( \frac{\log c(\delta) - 1}{1 + \delta} \right) \right].
\]

At this point, a remark on \( \delta \) is in order. Note that \( \delta \) is a parameter of a code for the integers, regardless of \( K \) and the guessing strategy used. \( c(\delta) \) is a normalizing constant, with \( c(\delta) \to \infty \) as \( \delta \to 0 \). In fact, \( c(\delta) = \zeta(1 + \delta) \), where \( \zeta(\cdot) \) is the Riemann Zeta function, satisfying \( \lim_{\delta \to 0} \delta \zeta(1 + \delta) = 1 \). Hence at the limit of \( \delta \to 0 \), \( \log c(\delta) \) can be approximated by \( -\log \delta \). Yet, due to the independence in \( K \), we have

\[
\lim_{K \to \infty} \inf_{Q^K} \frac{1}{K} \log E_{Q^K} \{ G^P(X) \}
\geq \lim_{K \to \infty} \sup_{Q^K} \frac{1}{K} \left[ \frac{\rho}{1 + \delta} R_K(D, Q^K) - D(Q^K\| P^K) \right],
\]

for any \( \delta > 0 \), which completes the proof of (23).

2) Direct: A key ingredient in the direct proof is that with proper scaling and gaps, \( \Psi \)-mixing sources can be approximated by block-memoryless sources, for which one can use the method of types over the super–alphabet of the blocks. Note, however, that directly applying the method of types to consecutive blocks of source symbols will not result in a good approximation, as the blocks are not independent, hence gaps between the blocks should be introduced. To make this precise, assume a source \( n \)-tuple is divided to consecutive, non-overlapping tuples of length \( K + k \). We assume both \( K \) and \( k \) are fixed, yet \( K \gg k \). Remember that \( K + k \) divides \( n \) and \( n/K \). Hence,

\[
x_1^n = x_1^K, x_{K+1}^{K+k}, \ldots, x_{n-k+1}^{n-K}, x_{n-k+1}^{n-k+1} \equiv \tilde{x}_1, \tilde{g}_1, \ldots, \tilde{g}_m, \tilde{g}_m, \tilde{x} \in \mathcal{X}^K, \tilde{g}_i \in \mathcal{X}^k, 1 \leq i \leq m,
\]

(26)

where we refer to the \( K \)-tuples \( \tilde{x} \) as blocks, and to the \( k \)-tuples \( \tilde{g}_i \) as gaps. Analogously to the type of a memoryless sequence, herein, we define the (generalized, due to the gaps) \( K \)-type of a sequence \( x_i^t \) as the set of empirical frequencies for each possible block of size \( K \) in \( x_i^t \), disregarding the content of the gaps. That is, denoting by \( N_{\mathcal{X}}(a_1^K) \) the number of occurrences of \( a_1^K \in \mathcal{X}^K \) in the blocks \( \{ \tilde{x}_i \}_i \subseteq \mathcal{X}^K \), we have

\[
\hat{P}_{\mathcal{X}}^{K} \triangleq \left( \frac{N_{\mathcal{X}}(a_1^K)}{m}, a_1^K \in \mathcal{X}^K \right).
\]

The \( K \)-type class of \( P^K \), denoted by \( T^K(P^K) \), is the set of all sequences with the same \( K \)-type \( P^K \), that is, \( T^K(P^K) = \{ x \in \mathcal{X}^n : \hat{P}_{\mathcal{X}}^{K} = P^K \} \). Finally, denoting by \( T^K \) the set of all possible \( K \)-types \( P^K \), clearly,

\[
| T^K | \leq (m + 1)^{| \mathcal{X} |^K}.
\]

(27)

We are now ready to prove the direct. We wish to show that for any stationary \( \Psi \)-mixing source \( P \) over \( \mathcal{X} \), if guesses are selected independently at random according to the distribution (3), then, for \( K \gg k \), the \( \rho \)-th guesswork moment, subject to a distortion constraint \( D \), satisfies

\[
\lim_{n \to \infty} \frac{1}{n} \log E \{ G^P \}
\leq \lim_{K \to \infty} \sup_{Q^K} \frac{1}{K} \left[ \rho R_K(D, Q^K) - D(Q^K\| P^K) \right].
\]

(28)

Note that the l.h.s. of (28), unlike that in (23), includes an expectation with respect to both the source distribution and the randomization in the guessing strategy. This will also be clear from (29) below, which uses a smoothing theorem to compute this expectation.
To prove (28), we first describe the expected guesswork moment in terms of the guessing distribution $\hat{P}$ via Lemma 1. Specifically,

\[
E\{G^\rho\} = E\{E\{G^\rho|X\}\} = E\{\hat{P}[S(X)]^{-\rho}\} = \sum_{x \in \mathcal{X}^n} P(x)\hat{P}[S(x)]^{-\rho}.
\]

We now wish to lower bound $\hat{P}[S(x)]$, generalizing Lemma 3 to guessing using the distribution in (3). This is summarized by the following lemma.

**Lemma 6:** Assume sequences of length $n$ are chosen independently at random according to the distribution $\hat{P}(\hat{x}) = 2^{-LZ(\hat{x})}$ in (3). Then, for any $\delta, \epsilon > 0$ there exists a sequence $\epsilon_m \to 0$ as $m \to \infty$ such that

\[
\hat{P}[S(x)] \geq 2^{-\nu_{K+k}(n)}(1 - 2\epsilon_m)\exp \left\{-m\left[R_K(D', \hat{P}_{x}) + \epsilon\right]\right\},
\]

where $D' = D - \delta_2 - \frac{\delta}{K}$, $\delta_2 = \frac{k(D_{max} - D)}{K}$, $\hat{P}_{x}$ is the $K$-type of $x$, and for any fixed $K$ and $k < K$, $\lim_{n \to \infty} \frac{1}{m}\nu_{K+k}(n) = O\left(\frac{1}{K}\right)$.

**Proof:** To prove the lemma we first, in (31) below, consider a specific block–memoryless distribution, induced by the distribution which achieves the rate–distortion function $R_K(D, \hat{P}_{x})$, and connect the probability of hitting the distortion–achieving blocks under this distribution to the rate–distortion function. In a sense, this relates the probability of a slightly smaller set, $\mathcal{S}(x) \subset S(x)$, to be defined precisely later, to the rate–distortion function of a block memoryless source.

Fix $x \in \mathcal{X}^n$ with a $K$-type $\hat{P}_{x}$. For any $\delta > 0$, set $D' = D - \delta_2 - \frac{\delta}{K}$, with $\delta_2 = \frac{k(D_{max} - D)}{K}$ and denote by $P^K(\hat{x}|x)$, with $\hat{x} \in \mathcal{X}^K$ and $x \in \mathcal{X}^n$ the conditional distribution achieving $R_K(D', \hat{P}_{x})$. Finally, denote by $Q^K$ the resulting distribution on the alphabet $\mathcal{X}^K$, that is, $Q^K(\hat{x}) = \sum_x P^K(\hat{x}|x)\hat{P}_{x}|x)$. For any $\epsilon > 0$, the following holds

\[
\sum_{\hat{s}_m^i \in \mathcal{X}^K: \sum d(\hat{s}_m^i, \hat{s}_i) \leq mK(D - \delta_2)} Q^K(\hat{s}_i) \geq (1 - 2\epsilon_m)\exp \left\{-m\left[R_K(D', \hat{P}_{x}) + \epsilon\right]\right\}.
\]

(31)

The proof of (31) is technical, hence it is deferred to Appendix C.

Let $M^{K+k}(\cdot)$ be any probability distribution on $\mathcal{X}^{K+k}$ and

\[
\prod_{i=1}^{m} M^{K+k}(\hat{x}^{(K+k)}(\hat{x}^{(i-1)}(K+k)+1))
\]

be the product distribution on $\mathcal{X}^n$. Similar to (26), consider the partition of a sequence $\hat{x}$ of length $n$ over $\hat{X}$ into non-overlapping blocks of length $K$, followed by gaps of length $k$

\[
\hat{x} = \hat{x}_1, \hat{g}_1, \ldots, \hat{x}_m, \hat{g}_m.
\]

Next, define $\mathcal{S}(x) \subset S(x)$ as follows

\[
\mathcal{S}(x) = \{\hat{x} : \hat{g}_i = (\hat{x}_0, \ldots, \hat{x}_0)\text{ for all } i, \sum_i d(\hat{x}_i, \hat{x}_i) \leq mK(D - \delta_2)\}.
\]

That is, $\mathcal{S}(x)$ includes only sequences in $S(x)$ for which the gaps include only $\hat{x}_0$ (for some fixed symbol $\hat{x}_0$), yet the distortion constraint is satisfied. Consequently, we have

\[
\hat{P}[S(x)] \geq \sum_{\hat{x} \in \mathcal{S}(x)} 2^{-LZ(\hat{x})}
\]

\[
\geq 2^{-\nu_{K+k}(n)} \sum_{\hat{x} \in \mathcal{S}(x)} \prod_{i=1}^{m} M^{K+k}(\hat{x}^{(i-1)}(K+k)+1)
\]

\[
= 2^{-\nu_{K+k}(n)} \sum_{\hat{x} \in \mathcal{S}(x)} \prod_{i=1}^{m} Q^K(\hat{x}^{(i-1)}(K+k)+1)
\]

\[
= 2^{-\nu_{K+k}(n)} \sum_{\hat{x} \in \mathcal{S}(x)} \prod_{i=1}^{m} Q^K(\hat{x}^{(i-1)}(K+k)+1)
\]

\[
= 2^{-\nu_{K+k}(n)} \sum_{\hat{x} \in \mathcal{S}(x)} \prod_{i=1}^{m} Q^K(\hat{x}^{(i-1)}(K+k)+1)
\]

\[
\geq 2^{-\nu_{K+k}(n)}(1 - 2\epsilon_m)\exp \left\{-m\left[R_K(D', \hat{P}_{x}) + \epsilon\right]\right\}.
\]

In the above chain, (33) is by Claim 1, (34) is since (33) is valid for any probability distribution $M^{K+k}(\cdot)$ on $\hat{X}^{K+k}$, hence we can choose it to be $Q^K(\hat{x}^{K}) \mathbb{I}[\hat{x}^{K} = (\hat{x}_0, \ldots, \hat{x}_0)]$, where $Q^K(\cdot)$ is the specific output distribution induced by the rate–distortion achieving distribution, as given in (31). The last inequality thus follows from (31).

Using Lemma 6, we can now continue from (30) to complete the proof of the direct.

\[
E\{G^\rho\} \geq \sum_{\hat{x} \in \mathcal{X}^n} P(\hat{x})\hat{P}[S(\hat{x})]^{-\rho}
\]

\[
\geq 2^{-\nu_{K+k}(n)} \sum_{\hat{x} \in \mathcal{S}(x)} \prod_{i=1}^{m} Q^K(\hat{x}^{(i-1)}(K+k)+1)
\]

\[
= 2^{-\nu_{K+k}(n)} \sum_{\hat{x} \in \mathcal{S}(x)} \prod_{i=1}^{m} Q^K(\hat{x}^{(i-1)}(K+k)+1)
\]

\[
= 2^{-\nu_{K+k}(n)} \sum_{\hat{x} \in \mathcal{S}(x)} \prod_{i=1}^{m} Q^K(\hat{x}^{(i-1)}(K+k)+1)
\]

\[
= 2^{-\nu_{K+k}(n)} \sum_{\hat{x} \in \mathcal{S}(x)} \prod_{i=1}^{m} Q^K(\hat{x}^{(i-1)}(K+k)+1)
\]

While it is tempting to replace the sum over $x \in \mathcal{X}^n$ with the usual sum over types, unlike the memoryless case, sequences in $T^K(P^K)$ are not equiprobable. Fortunately, dividing $T^K(P^K)$ into subsets of sequences, which differ only in the content of the gaps between the blocks, and with the right choice of $k$, results in asymptotically equiprobable subsets, whose probability can be described similarly to that of memoryless sequences. Specifically, for any $x \in \mathcal{X}^n$, whose blocks are denoted by $x_i, 1 \leq i \leq m$ (see (26)), define the gaps oblivious set as

\[
G^K = \{x \in \mathcal{X}^n : x_i = x_i, 1 \leq i \leq m\}.
\]
That is, $G^K_{x}$ includes all sequences which differ from $x$ only in the gaps between the blocks. $G^K_{x}$ satisfies the following bounds, proved in Appendix D.

**Lemma 7:** For any stationary $\Psi$-mixing source $Q$ and any $x \in X^n$, 

$$(1 - \epsilon_k)^{m-1} \exp \left\{ -m \left( D \left( \hat{P}^K_{x} \| Q^K \right) + \hat{H}^K_{x}(X^K) \right) \right\} \leq Q \left( G^K_{x} \right) \leq (1 + \epsilon_k)^{m-1} \exp \left\{ -m \left( D \left( \hat{P}^K_{x} \| Q^K \right) + \hat{H}^K_{x}(X^K) \right) \right\},$$

where $Q^K(X)$ is the $K$-th order marginal distribution of $Q$, $\hat{H}^K_{x}(X^K)$ is the empirical entropy associated with the empirical distribution $\hat{P}^K_{x}$ and $\epsilon_k \to 0$ as $k \to \infty$.

Note that if the $K$-th order marginal of $Q$ satisfies $Q^K = \hat{P}^K_{x}$, the divergences in Lemma 7 are zero and we have: \footnote{$\hat{P}^K_{x}$ ($G^K_{x}$) is the probability of a set of sequences, rather than the probability of a single sequence. Sequences in the subset are not necessarily equiprobable, yet the probability of the whole set is equal to that of other sets of sequences, as long as the sets are defined for the same $K$-type $P^K$.}

\begin{align}
\hat{P}^K_{x} \left( G^K_{x} \right) & \leq (1 + \epsilon_k)^{m-1} - 2m \hat{H}^K_{x}(X^K) \quad (38) \\
\hat{P}^K_{x} \left( G^K_{x} \right) & \geq (1 - \epsilon_k)^{m-1} - 2m \hat{H}^K_{x}(X^K). \quad (39)
\end{align}

Denote by $|T^K(P^K)|_G$ the number of distinct sets $G^K_{x}$ in $T^K(P^K)$. The following corollary, proved in Appendix E, bounds the number of such sets with a $K$-type class $T^K(P^K)$.

**Corollary 2:** $|T^K(P^K)|_G \leq (1 - \epsilon_k)^{-m+1} \exp \left( -m \hat{H}^K_{x}(X^K) \right)$.

We are now able to further bound $E(G^0)$. Continuing from (36),

$$E(G^0) \leq \frac{2^{\rho m K + \epsilon(n)}}{(1 - \epsilon_k)^{m-1} - 2m \hat{H}^K_{x}(X^K)} \sum_{T^K \in T^K} \sum_{G^K_{x} \in T^K} P(G^K_{x}) 2^{m \left[ R_K(D', \hat{P}^K_{x}) + \epsilon \right]}$$

\begin{align}
& \leq \frac{2^{\rho m K + \epsilon(n)} (1 + \epsilon_k)^{m-1}}{(1 - \epsilon_k)^{m-1} - 2m \hat{H}^K_{x}(X^K)} \sum_{T^K \in T^K} \sum_{G^K_{x} \in T^K} 2^{m \left( -D(D', \hat{P}^K_{x}) - \hat{H}^K_{x}(X^K) \right)} 2^{m \left[ R_K(D', \hat{P}^K_{x}) + \epsilon \right]} \quad (40)
\end{align}

$$\leq \frac{2^{\rho m K + \epsilon(n)} (1 + \epsilon_k)^{m-1}}{(1 - \epsilon_k)^{m-1} - 2m \hat{H}^K_{x}(X^K)} \sum_{T^K \in T^K} \sum_{G^K_{x} \in T^K} 2^{-m D(D', \hat{P}^K_{x})} 2^{m \left[ R_K(D', \hat{P}^K_{x}) + \epsilon \right]} \quad (41)
$$

Finally, the $\epsilon$-type

$$\leq \frac{2^{\rho m K + \epsilon(n)} (1 + \epsilon_k)^{m-1}}{(1 - \epsilon_k)^{m-1} - 2m \hat{H}^K_{x}(X^K)} \left[ T^K \right] \max_{\hat{P}^K_{x}} 2^{-m D(D', \hat{P}^K_{x})} 2^{m \left[ R_K(D', \hat{P}^K_{x}) + \epsilon \right]} \quad (42)
$$

In the above chain, (40) is due to Lemma 7, while (41) is due to Corollary 2; (42) follows from (27) and since $\frac{1}{1 + \epsilon_k} \leq 1$.

Now, fix $K \gg k$. Since $D' = D - \delta - \frac{\epsilon}{K}$, with $\delta = \frac{k(D_{\max} - D)}{K}$, and since the above is true for any $\delta, \epsilon > 0$, taking the limit over $n$, we have

$$\lim sup_{n \to \infty} \frac{1}{n} \log E(G^0) \leq \max_{K} \frac{1}{K} \left[ \rho R_K(D, \hat{P}^K_{x}) - D \left( \hat{P}^K_{x} \| P^K \right) + O(1) \right], \quad (43)
$$

which where we have also used the continuity of $R_K(D, \hat{P}^K_{x})$ in $D$. Finally, the r.h.s of (43) is at least as large if the maximization is done over all distributions on $X^K$, rather than only (fractional) types. Thus, enlarging the optimization domain, then taking a limit over $K$, we have

$$\lim sup_{n \to \infty} \frac{1}{n} \log E(G^0) \leq \lim inf_{K \to \infty} \frac{1}{K} \left[ \rho R_K(D, Q^K) - D \left( Q^K \| P^K \right) \right], \quad (44)
$$

which completes the proof of (28).

**D. Proof of Theorem 4**

1) Converse: The converse result is similar to the converse in [48], with the exception of a different notion of a success probability. Specifically, for any sequence $x$, we first generalize the success probability of the finite-state machine, $P(x)$ in the notation of [48] when one has to guess the exact sequence $x$, to a success probability in terms of guessing under a distortion constraint. That is, by [48, eq. (21)], for any finite-state guessing machine $F$, which is fed by i.i.d. random bits, the $\rho$-th moment satisfies the following lower bound

$$E(G^0_{x} | x) \geq \frac{2^{-\rho}}{c^2} \exp \left\{ -\rho \log P(x) \right\},$$

where $P(x)$ is the probability that the machine output is $x$. Note that similar to Lemma 1, the bound above is based on bounding the $\rho$-th moment of a Geometric random variable, i.e., repeated independent trials, where success is declared when the machine outputs the correct sequence. Thus, analogously, when guessing subject to a distortion measure $d$ and value $D$, each trail is independent as well (as the machine is
restarted after each guess), yet success is declared when the machine outputs a sequence within the designated distortion. Hence, we have

\[
E\{G^\rho_P|\hat{x}\} \geq \frac{2^{-\rho}}{e^2} \exp_2 \left\{ -\rho \log P[S(\hat{x})]\right\},
\]

where \(S(\hat{x})\), the success set, was defined below (1). We thus have

\[
E\{G^\rho_P|\hat{x}\} \geq \frac{2^{-\rho}}{e^2} \exp_2 \left\{ -\rho \log \sum_{\{\hat{x}:d(\hat{x},\bar{x}) \leq nD\}} P(\hat{x})\right\}.
\]

(44)

\(P(\hat{x})\), the probability that the machine outputs \(\hat{x}\), can be viewed as a probability assignment on the sequences \(\hat{x} \in \hat{X}^n\). As such, \(-\log P(\hat{x})\) cannot be too small compared to the code length of an LZ encoder. This is similar to the argument behind Claim 1 used on a block–memoryless distribution. In the context of the current discussion, however, it is easier to harness [48, equations (31) and (32)], resulting in

\[
-\log P(\hat{x}) \geq c(\hat{x}) \log c(\hat{x}) - \frac{n \log [4K^2(l)] \log |\hat{X}|}{(1 - \epsilon_n) \log n} - K^2(l) \log [4K^2(l)] - m \log (2s^3 e),
\]

(45)

where \(s\) is the number of states, \(m\) and \(l\) are integers such that \(ml = n\) and \(K(l)\) is the number of states of a machine implementing a Shannon code on blocks of size \(l\). Thus, choosing \(l\) to grow to infinity slow enough, we have

\[
-\log P(\hat{x}) \geq LZ(\hat{x}).
\]

Specifically, with \(l = \log(\log n)\) and since \(K(l)\) is bounded by \(l|\hat{X}|\), we have

\[
\frac{n \log [4K^2(l)] \log |\hat{X}|}{(1 - \epsilon_n) \log n} + K^2(l) \log [4K^2(l)] + m \log (2s^3 e) = o(n)
\]

and hence

\[
E\{G^\rho_P|\hat{x}\} \geq \exp_2 \left\{ -\rho \log \sum_{\{\hat{x}:d(\hat{x},\bar{x}) \leq nD\}} 2^{-LZ(\hat{x})}\right\} = \left[ \sum_{\{\hat{x}:d(\hat{x},\bar{x}) \leq nD\}} 2^{-LZ(\hat{x})}\right]^{-\rho},
\]

which proves part (i) of the converse result.

For part (ii), instead of (45), we lower bound \(-\log P(\hat{x})\) using the finite–state compressibility of \(\hat{x}\), that is, by [48, eq. (31)],

\[
-\log P(\hat{x}) \geq n\rho_{K(l)}(\hat{x}) - m \log (2s^3 e).
\]

Continuing from (44), we have

\[
E\{G^\rho_P|\hat{x}\} \geq \frac{2^{-\rho}}{e^2} \exp_2 \left\{ -\rho \log \sum_{\{\hat{x}:d(\hat{x},\bar{x}) \leq nD\}} 2^{-LZ(\hat{x})}\right\}.
\]

(46)

for the specific universal distribution in (3). Clearly, this meets the converse result. However, the distribution in (3) cannot be implemented with a finite–state machine whose number of states is independent of \(n\). To asymptotically achieve (46) with such a finite–state machine, we invoke the solution in [48] for the same problem (as the same distribution was sought). That is, we consider the distribution in (12). This distribution can be implemented with a finite–state machine with \(s(l)\) states, and for such a finite–state guessing machine we have

\[
E\{G^\rho_{LZ}|\hat{x}\}
\]

\[
= \left[ \sum_{\{\hat{x}:d(\hat{x},\bar{x}) \leq nD\}} 2^{-LZ(\hat{x})}\right]^{-\rho}.
\]

(47)

where \(\epsilon(l) \rightarrow 0\) as \(l \rightarrow \infty\), (47) is by [48, eq. (33)] and (48) is by [48, eq. (38)].
VI. CONCLUSION

We considered the problem of guessing subject to a distortion constraint. Specifically, assuming one player, Alice, has a sequence \( x \), unknown to the other player, Bob, Bob’s goal is to guess a sequence \( \hat{x} \), which is within a given distortion \( D \) from \( x \). We analyzed the problem above under four different setups. The first assuming \( x \) is drawn from a memoryless source. In the second we further assumed that Bob’s guesses are not sent directly to Alice, but, rather, are passed through a discrete memoryless channel. In the third, we assumed \( x \) is drawn from a much more general family of sources, that is, \( \Psi \)-mixing sources, which includes different families of sources with memory. Last but not least, we also considered the case where Alice’s sequence is an individual sequence, without any probability distribution.

For all four setups we suggested randomized guessing strategies, analyzed their performance and proved their universality in the source statistics, the channel, the guesswork moment to be optimized, the distortion measure and the distortion value. Moreover, we showed that, in fact, a single guessing strategy is asymptotically optimal for all setups. The strategy is based on feeding an LZ decoder with random bits, and hence can be very efficiently implemented, with asymptotically optimal performance for all setups and parameters considered.

APPENDIX A

ADDITIONAL PROOFS AND TECHNICAL CLAIMS

A. Proof of Claim 1

The proof is based on viewing any block-memoryless probability assignment for \( \hat{x} \), followed by a Shannon code, as a finite-state encoder for \( \hat{x} \). As such, its code length cannot be significantly shorter than the length of the code assigned by the LZ encoder [34, Theorem 1]. It is similar to the line of argument used in [48, eq. (32)].

Let \( M^{K+k}(\cdot) \) be any distribution on \( \hat{X}^{K+k} \), and assume the sequence \( \hat{x} \) is divided into blocks of length \( K + k \) and encoded using a Shannon code for \( M^{K+k}(\cdot) \). The resulting compression ratio satisfies

\[
\rho_{K+k} = \frac{1}{n \log |X|} \sum_{i=1}^{m} \left[ -\log M^{K+k}(\hat{x}^{i(K+k)}_{(i-1)(K+k)+1}) \right] \\
\leq \frac{1}{n \log |X|} \sum_{i=1}^{m} \log M^{K+k}(\hat{x}^{i(K+k)}_{(i-1)(K+k)+1}) + m \\
= \frac{1}{n \log |X|} \left[ \log \prod_{i=1}^{m} M^{K+k}(\hat{x}^{i(K+k)}_{(i-1)(K+k)+1}) + m \right].
\]

The above encoder can clearly be implemented using a finite-state machine of \( s(K+k) \) states, where \( s(\cdot) \) is independent of \( n \). Hence, by [34, Theorem 1], dropping the dependence on \( K + k \) for ease of notation, we have

\[
\rho_{K+k} \geq \frac{c(\hat{x}) + s^2}{n \log |X|} \log \left( \frac{c(\hat{x}) + s^2}{4s^2} \right) + \frac{2s^2}{n \log |X|}. \]

Consequently,

\[
-\log \prod_{i=1}^{m} M^{K+k}(\hat{x}^{i(K+k)}_{(i-1)(K+k)+1}) + m \\
\geq (c(\hat{x}) + s^2) \log \left( \frac{c(\hat{x}) + s^2}{4s^2} \right) + 2s^2 \\
\geq (c(\hat{x}) + 1 + s^2 - 1) \log \left( \frac{2|\hat{X}|}{2|x|} + 1 \right) + 2s^2 \\
= (c(\hat{x}) + 1) \log \left( \frac{2|\hat{X}|(c(\hat{x}) + 1)}{4s^2} \right) + \log \left( \frac{2|\hat{X}|(c(\hat{x}) + 1)}{4s^2} + 2s^2 - (c(\hat{x}) + s^2) \log \left( \frac{8|\hat{X}|s^2}{|x|} \right) \right).
\]

Finally, by [34, Theorem 2], \( LZ(\hat{x}) \leq (c(\hat{x}) + 1) \log \left( \frac{2|\hat{X}|(c(\hat{x}) + 1)}{4s^2} \right) \), thus

\[
LZ(\hat{x}) \leq -\log \prod_{i=1}^{m} M^{K+k}(\hat{x}^{i(K+k)}_{(i-1)(K+k)+1}) + m + (c(\hat{x}) + s^2) \log \left( \frac{8|\hat{X}|s^2}{|x|} \right),
\]

Setting \( \nu_{K+k}(n) = m + (c(\hat{x}) + s^2) \log \left( \frac{8|\hat{X}|s^2}{|x|} \right) \), we obtain the required result.

B. Proof of Lemma 2

First, we derive an alternative expression to \( \Gamma(Q_Y) \). We have

\[
\Gamma(Q_Y) = \inf_{Q_{X|Y}} D(Q_{Y|X} \| W(Q_X)) \\
= \inf_{Q_{X|Y}} \sum_{x,y} Q_{XY}(x,y) \log \frac{Q_{Y|X}(y|x)}{W(y|x)} \\
= \inf_{Q_{X|Y}} \sum_{x,y} Q_{XY}(x,y) \log \frac{Q_{Y|X}(y|x)Q_Y(y)}{Q_X(x)W(y|x)} \\
= -H_Q(Y) + \inf_{Q_{X|Y}} \sum_{x,y} Q_{XY}(x,y) \log \frac{Q_{Y|X}(x|y)Q_Y(y)}{Q_X(x)W(y|x)} \\
= -H_Q(Y) + \inf_{Q_{X|Y}} \sum_{y} Q_Y(y) \sum_{x} Q_{X|Y}(x|y) \log \frac{Q_{Y|X}(x|y)}{V(x)W(y|x)} \quad (49) \\
= -H_Q(Y) + \inf_{Q_{X|Y}} \left\{ -\sum_{y} Q_Y(y) \log \left( \sum_{x} V(x)W(y|x) \right) \right\} \quad (50)
\]

In the above chain, (49) is true since

\[
\sum_{x,y} Q_{XY}(x,y) \log \frac{Q_{Y|X}(x|y)}{V(x)W(y|x)}
\]
\[- \sum_{x,y} Q_{XY}(x, y) \log \frac{Q_{XY}(x|y)}{Q_X(x)W(y|x)} = \sum_{x,y} Q_{XY}(x, y) \log \frac{Q_X(x)}{V(x)} \geq 0 \]

with equality when \( V(x) = Q_X(x) \) for all \( x \). (50) follows by noting that

\[
\sum_x Q_{XY}(x|y) \log \frac{Q_{XY}(x|y)}{V(x)W(y|x)} = \sum_x Q_{XY}(x|y) \log \frac{\sum_{\hat{x}} V(\hat{x})W(y(\hat{x})]}{\sum_{\hat{x}} V(\hat{x})W(y|\hat{x})} V(x)W(y|x)
\]

\[
= D \left( Q_{XY} \parallel \tilde{V}_{XY} \right) - \log \left[ \sum_{\hat{x}} V(\hat{x})W(y|\hat{x}) \right], \tag{52}
\]

where \( \tilde{V}_{XY}(x|y) = \frac{1}{\sum_x V(x|y(x))} V(x)W(y|x) \) and the expression is clearly minimized by choosing \( Q_{XY} = \tilde{V}_{XY} \).

Now, \( \hat{R}_W(D, Q_Y) \)

\[
\hat{R}_W(D, Q_Y) = \inf_{Q_{XY} \leq D} \left[ H_Q(Y') - H_Q(Y'|Y) + \Gamma(Q_{Y'}) \right]
\]

\[
= \inf_{Q_{XY} \leq D} \inf_V \left\{ - H_Q(Y'|Y) - \sum_{y'} Q_{Y'} \left[ \log \sum_x V(x)W(y'|x) \right] \right\}
\]

\[
= \inf_{Q_{XY} \leq D} \sup_{s \geq 0} \inf_V \left[ \sum_y Q_Y(y) \sum_{y'} Q_{Y'|Y}(y'|y) \log \frac{Q_{Y'|Y}(y'|y)}{\sum_x V(x)W(y'|x)} \right. \\
\left. + s \left( \sum_y Q_Y(y) \sum_{y'} Q_{Y'|Y}(y'|y) d(y, y') - D \right) \right]\tag{53}
\]

\[
= \inf_{Q_{XY} \leq D} \inf_V \sup_{s \geq 0} \left[ \sum_y Q_Y(y) \sum_{y'} Q_{Y'|Y}(y'|y) \right. \\
\left. \log \frac{Q_{Y'|Y}(y'|y)}{\sum_x V(x)W(y'|x)} + sd(y, y') - sD \right] \tag{54}
\]

\[
= \inf_{Q_{XY} \leq D} \inf_V \sup_{s \geq 0} \left[ \sum_y Q_Y(y) \sum_{y'} Q_{Y'|Y}(y'|y) \right. \\
\left. \log \frac{e^{-sd(y, y')}}{\sum_x V(x)W(y'|x)} - sD \right] \tag{55}
\]

\[
= \inf_{s \geq 0} \left\{ - \sum_y Q_Y(y) \right. \left. \log \left[ \sum_{y'} e^{-sd(y, y')} \sum_x V(x)W(y'|x) \right] - sD \right\} \tag{56}
\]

\[
= \inf_{s \geq 0} \left\{ - \sum_y Q_Y(y) \frac{\sum_x V(x)W(y|x)}{V(x)} - sD \right\}
\]

\[
= \inf_{s \geq 0} \left\{ - \sum_y Q_Y(y) \frac{\sum_x V(x)U_s(x, y)}{V(x)} - sD \right\}
\]

\[
= \inf_{s \geq 0} \left\{ - \sum_y Q_Y(y) \frac{\sum_x V(x)U_s(x, y)}{V(x)} - sD \right\}, \tag{57}
\]

where (53) is due to the Lagrange multipliers analogous form for the optimization problem. (54) is due to the Minimax theorem, as the function within the squared brackets is linear (hence concave) in \( s \) and convex in \( V \). (55) is since the utility function is convex in \( Q_{Y'|Y} \) and concave (in fact, linear) in \( s \). (56) uses a similar technique to (51)-(52), namely, the expression

\[
\sum_y Q_{Y'|Y}(y'|y) \log \frac{Q_{Y'|Y}(y'|y)}{e^{-sd(y, y')} \sum_x V(x)W(y'|x)}
\]

can be viewed as a divergence minus a normalization term.

In (57) we have defined

\[
U_s(x, y) \triangleq \sum_{y'} W(y'|x)e^{-sd(y, y')}
\]

and the last equality follows from the convexity in \( V(x) \). Finally,

\[
E_W(\rho)
= \sup_{Q_X} [\hat{R}_W(D, Q_Y) - D(Q_Y \parallel P)]
\]

\[
= \sup_{Q_X} \inf_{s \geq 0} \left\{ \sum_y Q_Y(y) \frac{P(y)}{Q_Y(y)} \frac{\sum_x V(x)U_s(x, y)}{V(x)} - sD \right\}
\]

\[
= \sup_{s \geq 0} \inf_Q \left\{ \sum_y Q_Y(y) \frac{P(y)}{Q_Y(y)} \frac{\sum_x V(x)U_s(x, y)}{V(x)} - sD \right\}
\]

\[
= \sup_{s \geq 0} \inf_Q \left\{ \sum_y Q_Y(y) \frac{P(y)}{Q_Y(y)} \frac{\sum_x V(x)U_s(x, y)}{V(x)} - sD \right\}
\]

\[
= \sup_{s \geq 0} \inf_{Q \in CH(W)} \left\{ \log \left( \sum_y P(y) \frac{\sum_x V(x)U_s(x, y)}{\sum_x V(x)U_s(x, y)} \right) - sD \right\}
\]

\[
= \sup_{s \geq 0} \inf_{Q \in CH(W)} \left\{ \log \left( \sum_y P(y) \frac{\sum_x M(y)e^{-sd(y, y')}}{\sum_x M(y)e^{-sd(y, y')}} \right) - sD \right\}
\]

\[
= \sup_{s \geq 0} \inf_{Q \in CH(W)} \left\{ \log \left( \sum_y P(y) \frac{M(y')e^{-sd(y, y')}}{\sum_x M(y')e^{-sd(y, y')}} \right) - sD \right\}.
\]
In the above chain, (58) is since the utility is convex in $V(\cdot)$ and concave in $Q_Y(\cdot)$ and (59), again, uses a similar technique to (51)-(52). For (60), remember that $U_s(x, y) = \sum_{y} W(y|x)e^{-sd(y, y')}$ and define $M(y') = \sum_{x} V(x)W(y'|x)$. In fact, $M(y')$ is the channel output distribution when its input is i.i.d. according to $V$. (60) thus replaces the minimization over $V$ by minimization over $M$, which is limited to $C\mathcal{H}(W)$, that is, the convex hull of $\{W(\cdot|x), x \in X\}$.

C. Proof of equation (31)

We first state a result for memoryless sources, then extend it to a block–memoryless scenario, minding the gaps between the blocks.

Fix $x \in X^m$ with a type $\hat{P}_x$. For any $\delta > 0$, set $D^- = D - \delta$ and denote by $P(\hat{x}|x)$ the conditional distribution achieving $R(D^-, \hat{P}_x)$. Finally, denote by $Q$ the result distribution on the construction alphabet, that is, $Q(\hat{x}) = \sum_{x} P(\hat{x}|x)\hat{P}_x(x)$. For any $\epsilon > 0$, the following holds

$$\sum_{\hat{x} \in S(\hat{x})} Q^n(\hat{x}) \geq (1 - 2\epsilon_n) \exp_2 \left\{ -n \left[ R(D^-, \hat{P}_x) + \epsilon \right] \right\}, \tag{61}$$

where $Q^n(\hat{x}) = \prod_{i=1}^n Q(\hat{x}_i)$ and $\epsilon_n \to 0$ as $n \to \infty$ for all $x \in X^n$. To prove (61), we follow [69, eq. (3.2.8)-(3.2.10)]. Define

$$T(x) = \left\{ \hat{x} : \log \frac{P^n(\hat{x}|x)}{Q^n(\hat{x})} \leq n \left[ R(D^-, \hat{P}_x) + \epsilon \right] \right\},$$

where $P^n(\hat{x}|x) = \prod_{i=1}^n P(\hat{x}_i|x_i)$, $P(\hat{x}_i|x_i)$ and $Q(\hat{x})$ being the optimal test channel and reconstruction distribution for $R(D^-, \hat{P}_x)$, respectively. We have

$$\sum_{\hat{x} \in S(\hat{x})} Q^n(\hat{x}) \geq \sum_{\hat{x} \in S(\hat{x}) \cap T(x)} Q^n(\hat{x}) \geq \exp_2 \left\{ -n \left[ R(D^-, \hat{P}_x) + \epsilon \right] \right\} \sum_{\hat{x} \in S(\hat{x}) \cap T(x)} P^n(\hat{x}|x) \geq \exp_2 \left\{ -n \left[ R(D^-, \hat{P}_x) + \epsilon \right] \right\} \left[ \sum_{\hat{x} \in S(\hat{x})} P^n(\hat{x}|x) - \sum_{\hat{x} \in T^c(x)} P^n(\hat{x}|x) \right].$$

To conclude the proof of (61), we have to show that there exists a sequence $\epsilon_n \to 0$ such that both $\sum_{\hat{x} \in S(\hat{x})} P^n(\hat{x}|x) \geq 1 - \epsilon_n$ and $\sum_{\hat{x} \in T^c(x)} P^n(\hat{x}|x) \leq \epsilon_n$ uniformly in $x$. To this end, we have the following two claims, whose proofs are given in Appendix C1 and Appendix C2 below.

Claim 2: There exists $\epsilon'_n$ such that for all $x$, $\sum_{\hat{x} \in S(\hat{x})} P^n(\hat{x}|x) \geq 1 - \epsilon'_n$ with $\epsilon'_n \to 0$ as $n \to \infty$.

Claim 3: There exists $\epsilon''_n$ such that for all $x$, $\sum_{\hat{x} \in T^c(x)} P^n(\hat{x}|x) \leq \epsilon''_n$ with $\epsilon''_n \to 0$ as $n \to \infty$.

Taking $\epsilon_n = \max(\epsilon'_n, \epsilon''_n)$ completes the proof of (61).

To complete the proof of (31), note that (61) can be easily extended to block–memoryless sources, simply by applying it to sequences of length $m$ over $\mathcal{X}^K$, with the proper definition on a type over $\mathcal{X}^K$ and output distribution $Q^K$ on $\mathcal{X}^K$. However, we take this a step further, and apply it to a sequence $x$ of length $n$ over $\mathcal{X}$, using a memoryless distribution over $K$-tuples and the $K$-type of $x$. Namely, given any $x \in X^n$, apply (61) on the sequence $(\hat{x}_1, \ldots, \hat{x}_n)$ (that is, only the $K$-tuple blocks of $x$, without the $k$-tuple gaps). This sequence, which includes $m$ blocks of length $K$ over $\mathcal{X}$, when viewed as an $m$-tuple over $\mathcal{X}^K$ has a type $\hat{P}_x^K$. Thus, setting a per-letter distortion level $K(D - \delta_2)$, (31) follows.

1) Proof of Claim 2: To ease notation, we write $P^n(\hat{x}|x) = \prod_{i=1}^n P(\hat{x}_i|x_i)$ as $P^n(\hat{x}|x)$ in what follows. We have

$$\sum_{\hat{x} \in S(\hat{x})} P^n(\hat{x}|x) = P^n(\hat{x}|x) \left( d(\hat{x}, \hat{x}) \leq nD \right)$$

$$= \sum_{\hat{x} \in S(\hat{x})} P^n(\hat{x}|x) \left( d(\hat{x}, \hat{x}) - nD \leq n\delta \right)$$

$$= \sum_{\hat{x} \in S(\hat{x})} P^n(\hat{x}|x) \left( \sum_{i=1}^n d(\hat{x}_i, \hat{x}_i) - nD \leq n\delta \right)$$

$$= \sum_{\hat{x} \in S(\hat{x})} P^n(\hat{x}|x) \left( \sum_{i=1}^n \frac{1}{n} \sum d(\hat{x}_i, \hat{x}_i) - D \leq \delta \right).$$

Since $P(\hat{x}|x)$ is the conditional distribution achieving $D^-$, denoting $D^-(x) = \sum_{\hat{x} \in \mathcal{X}} P(\hat{x}|x)d(\hat{x}, \hat{x})$,

$$\sum_{\hat{x} \in S(\hat{x})} P^n(\hat{x}|x) \left( \sum_{i=1}^n \frac{1}{n} \sum d(\hat{x}_i, \hat{x}_i) - D \leq \delta \right).$$

Since $P^n(\hat{x}|x)$ is the conditional distribution achieving $D^-$, denoting

$$D^-(x) = \sum_{\hat{x} \in \mathcal{X}} P(\hat{x}|x)d(\hat{x}, \hat{x}),$$

we have $\sum_{x \in \mathcal{X}} \hat{P}(x) D^-(x) = D^-$.

Thus, denoting by $N_x(\hat{x})$ the number of occurrences of $x$ in $\mathcal{X}$, and analogously $N_{x\hat{x}}(\hat{x})$ the number of occurrences of the pair $(x, \hat{x})$, we have

$$\sum_{\hat{x} \in \mathcal{X}} P^n(\hat{x}|x) \left( \sum_{i=1}^n \frac{1}{n} \sum d(\hat{x}_i, \hat{x}_i) - D^-(x) \leq \delta \right)$$

$$= \sum_{\hat{x} \in \mathcal{X}} \hat{P}(x) \left( \sum_{i=1}^n \frac{1}{n} \sum d(\hat{x}_i, \hat{x}_i) - \hat{P}(x) D^-(x) \leq \delta \right)$$

$$\geq \sum_{\hat{x} \in \mathcal{X}} \hat{P}(x) \left( \sum_{i=1}^n \frac{1}{n} \sum d(\hat{x}_i, \hat{x}_i) - \hat{P}(x) D^-(x) \leq \delta \right)$$

$$\geq \sum_{\hat{x} \in \mathcal{X}} \hat{P}(x) \left( \sum_{i=1}^n \frac{1}{n} \sum d(\hat{x}_i, \hat{x}_i) - \hat{P}(x) D^-(x) \leq \delta \right)$$

\text{(63)}
where, again, the last inequality is due to the law of large numbers, \( \frac{N_{\hat{x}}(x, 2)}{N_{\hat{x}}(x)} \rightarrow P(\hat{x}|x) \) for all \( x, \hat{x} \), since the alphabets \( \mathcal{X}, \hat{\mathcal{X}} \) are finite and, without loss of generality, since \( \min_{\hat{x}} Q(\hat{x}) > 0 \). Again, the convergence is uniform in \( x \).

### D. Proof of Lemma 7

Since \( Q \) is \( \Psi \)-mixing, for any two block \( \mathbf{z}_1, \mathbf{z}_1+1 \) which are \( k \) symbols apart, we have

\[
\frac{Q(\mathbf{z}_1, \mathbf{z}_1+1)}{Q(\mathbf{z}_1)} - 1 \leq \epsilon_k
\]

for some \( \epsilon_k \to 0 \) as \( k \to \infty \). Hence,

\[
Q(G^K_x) = \sum_{x \in \mathcal{X}^n : \mathbf{z}_i = x, 1 \leq i \leq m} Q(\hat{x})
\]

\[
\leq (1 + \epsilon_k)^{m-1} \prod_{i=1}^{m} Q^K(\mathbf{z}_i)
\]

\[
= (1 + \epsilon_k)^{m-1} 2^{-m(D(P^K_x||Q^K)+\delta^K_x(X^K))}.
\]

The opposite inequality follows in a similar manner.

### E. Proof of Corollary 2

Let \( P \) be any stationary \( \Psi \)-mixing distribution with a \( K \)-th order marginal \( P^K \). Then,

\[
1 \geq P(T^K(P^K)) = \sum_{G^K_x \in \mathcal{T}^K(P^K)} P^K(G^K_x)
\]

\[
\geq |T^K(P^K)| G(1 - \epsilon_k)^{m-1} 2^{-m \bar{H}^K_x(X^K)}.
\]

**ACKNOWLEDGMENT**

The authors would like to thank the Associate Editor and the anonymous referees for their valuable and constructive comments. Their suggestions improved the article significantly.

### REFERENCES

[1] J. M. Wozencraft, “Sequential decoding for reliable communication,” Ph.D. dissertation, Res. Lab. Electron., Massachusetts Inst. Technol., Cambridge, MA, USA, May 1957.

[2] E. Arikan, “An inequality on guessing and its application to sequential decoding,” *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 99–105, Jan. 1996.

[3] C. E. Pfister and W. G. Sullivan, “Rényi entropy, guesswork moments, and large deviations,” *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2794–2800, Nov. 2004.

[4] E. Arikan and N. Merhav, “Joint source-channel coding and guessing with application to sequential decoding,” *IEEE Trans. Inf. Theory*, vol. 44, no. 5, pp. 1756–1769, Sep. 1998.

[5] W. Szpankowski and S. Verdú, “Minimum expected length of fixed-to-variable lossless compression without prefix constraints,” *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4017–4025, Jul. 2011.

[6] I. Kontoyiannis and S. Verdú, “Optimal lossless data compression: Non-asymptotics and asymptotics,” *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 777–795, Feb. 2014.

[7] T. Courtade and S. Verdú, “Cumulant generating function of codeword lengths in optimal lossless compression,” in *Proc. IEEE Int. Symp. Inf. Theory*, Honolulu, HI, USA, Jul. 2014, pp. 2494–2498.
Neri Merhav (Fellow, IEEE) was born in Haifa, Israel, in March 1957. He received the B.Sc., M.Sc., and D.Sc. degrees in electrical engineering from the Technion—Israel Institute of Technology in 1982, 1985, and 1988, respectively. From 1988 to 1990, he was with AT&T Bell Laboratories, Murray Hill, NJ, USA. Since 1990, he has been with the Electrical Engineering Department, Technion, where he is currently the Irving Shepard Professor. During 1994 to 2000, he was also serving as a Consultant for the Hewlett-Packard Laboratories—Israel (HPL-I). His research interests include information theory, statistical communications, and statistical signal processing. He is especially interested in the areas of lossless/lossy source coding and prediction/filtering, relationships between information theory and statistics, detection, estimation, as well as in the area of Shannon Theory, including topics in joint source-channel coding, source/channel simulation, and coding with side information with applications to information hiding and watermarking systems. Another recent research interest concerns the relationships between information theory and statistical physics. He is currently on the Editorial Board of *Foundations and Trends in Communications and Information Theory*. He was a co-recipient of the 1993 Paper Award of the IEEE Information Theory Society. He also received the 1994 American Technion Society Award for Academic Excelllence and the 2002 Technion Henry Taub Prize for Excellence in Research. He also served as the Co-Chairperson for the Program Committee of the 2001 IEEE International Symposium on Information Theory. From 1996 to 1999, he served as an Associate Editor for the Special Issue on Source Coding of the IEEE TRANSACTIONS ON INFORMATION THEORY and has been an Associate Editor for the Special Issue on Shannon Theory of the IEEE TRANSACTIONS ON INFORMATION THEORY since 2017.

Asaf Cohen (Senior Member, IEEE) received the B.Sc. (Hons.), M.Sc. (Hons.), and Ph.D. degrees from the Department of Electrical Engineering, Technion—Israel Institute of Technology, in 2001, 2003, and 2007, respectively. From 1998 to 2000, he was with the IBM Research Laboratory, Haifa, where he was working on distributed computing. Between 2007 and 2009, he was a Postdoctoral Scholar with the California Institute of Technology and a Visiting Scientist with the Massachusetts Institute of Technology between 2015 and 2016. He is currently an Associate Professor with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Israel. His research interests are information theory, learning, and coding. In particular, he is interested in networks information theory, networks coding and coding in general, networks security and anomaly detection, statistical signal processing with applications to detection and estimation, and sequential decision-making. He received several honors and awards, including the Viterbi Postdoctoral Scholarship, the Dr. Philip Marlin Prize for Computer Engineering in 2000, the Student Paper Award from IEEE Israel in 2006, and the Ben-Gurion University Excellence in Teaching Award in 2014. He served as a Technical Program Committee Member for ISIT, ITW, and VTC for several years; and is currently an Associate Editor for *Information Theory and Network Coding, Physical Layer Security*, and the Special Issue on Source/Channel Coding and Cross-Layer Design of the IEEE TRANSACTIONS ON COMMUNICATIONS.