Universal Randomized Guessing Subjected to Distortion

Asaf Cohen[*] and Neri Merhav[†]

December 27, 2021

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 $\boldsymbol{x}$ of length n. Bob wishes to guess $\boldsymbol{x}$, yet he is satisfied with finding any sequence $\hat{\boldsymbol{x}}$ which is within a given distortion D from $\boldsymbol{x}$. 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 ρ–th moment) has several applications in information security, source and channel coding. Guessing subject to a distortion constraint is especially useful 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 $\boldsymbol{x}$, 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.

1. Introduction

Consider the problem of guessing a realization of a random n–vector $\boldsymbol{X}$ over an alphabet $\mathcal{X}$ subject to a distortion measure d. Specifically, assume Alice has such a realization $\boldsymbol{x}$. Bob wishes to

[*] A. Cohen is with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer Sheva 8410501, Israel (email: coasaf@bgu.ac.il).

[†] N. Merhav is with Andrew and Erna Viterbi Faculty of Electrical and Computer Engineering, Technion-Israel Institute of Technology (IIT), Haifa 3200003, Israel (e-mail: merhav@ee.technion.ac.il).

guess the value of $\boldsymbol{x}$, yet is content with discovering any n–vector $\hat{\boldsymbol{x}}$ over $\hat{\mathcal{X}}$ which is within a distortion D away from $\boldsymbol{x}$. He thus submits to Alice a sequence of *queries* of the form "Is $\hat{\boldsymbol{x}}_i$ close to $\boldsymbol{x}$ within distortion D?", $i = 1, 2, \ldots$, until receiving a positive answer. Let $G(\boldsymbol{x})$ denote the *guesswork*, namely, the smallest i for which $\hat{\boldsymbol{x}}_i$ is close enough to $\boldsymbol{x}$, in the sense of achieving an average distortion smaller than or equal to D under the measure $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$. Clearly, Bob is interested in devising a strategy which minimizes $G(\boldsymbol{x})$ in some sense. If $\boldsymbol{x}$ is a realization of a random vector, Bob may wish to minimize the expected value of $G(\boldsymbol{X})$ or some other moment of it. However, in many cases, Bob will have, for example, only limited knowledge (if any) on the distribution of $\boldsymbol{X}$ or the exact distortion required for success. Moreover, $\boldsymbol{x}$ 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 $\boldsymbol{x}$ 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, 2, 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, 6, 7, 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 2.

Recent trends in information security raised further attention to guesswork problems, adopting it as a proxy to password strength [13, 14, 15, 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, 19, 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 bare some level of distortion between the trained data and the tested data* (unlike 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–complex[1] 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 where 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 [28, 21]. 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.*

1.1. 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

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

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 source[2] 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 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 2, we thoroughly discuss related work. Section 3 includes the required notation and the problem statement. Section 4 includes the main results. Section 5 includes three important lemmas, which stand at the basis of the proofs. Indeed, the achievability result for memoryless sources easily follows from the first lemma, and matches a known lower bound. The achievability result for noisy guessing also follows directly, this time from the second lemma in Section 5. Its converse, however, is new and appears separately in Section 6. For sources with memory, both the achievability and the converse are more involved, hence their complete proofs appear in Section 7. Finally, the appendix includes additional proofs of some technical lemmas and claims for the first three setups, as well as the proofs for the achievability and converse parts of the fourth setup, in which Bob tries to guess an individual sequence.

2. Related Work

Guesswork was information–theoretically formalized by Massey [35]. Later, Arikan [2] computed the exponential rate of guesswork for memoryless sources. Specifically, it is was proved to be the Rènyi 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,

[2] A 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 4.3.

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 ρ. 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.

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[3]. 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.

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 1 also stress out the fact that the sequences to be guessed usually have an unknown distribution, and might contain highly dependent measurements

[3] From a practical viewpoint, this process applies in cases where a pre-compiled list of usernames and passwords is used - “hard-coding" the guessing strategy [41, 42]. This should be compared to distributed brute–force attacks [38, 43], and the asynchronous, randomized strategies we suggest herein.

[22, 24, 26].

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].

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 which can be applied to guesswork were given in [54] by Courtade 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, 56, 50, 43, 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.

3. Notation and Problem Statement

3.1. 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 $\boldsymbol{X} = (X_1, \dots, X_n)$ may take a specific vector value $\boldsymbol{x} = (x_1, \dots, x_n)$ in $\mathcal{X}^n$, the n–th order Cartesian power of $\mathcal{X}$, 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 $\boldsymbol{E}\{\cdot\}$, and $\mathcal{I}\{\cdot\}$ will denote the indicator function. The entropy of a distribution Q on $\mathcal{X}$ 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 $R(D, Q)$.

$\exp_2\{v\}$ denotes 2^v and log denotes $\log_2$. For two positive sequences a_n and b_n, $a_n \doteq b_n$ will stand for equality on the exponential scale, that is, $\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$. Similarly, $a_n \overset{\cdot}{\leq} b_n$ 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, $\boldsymbol{x} \in \mathcal{X}^n$, namely, $a_n = a_n(\boldsymbol{x})$ and $b_n = b_n(\boldsymbol{x})$, the notation $a_n(\boldsymbol{x}) \doteq b_n(\boldsymbol{x})$ means that the asymptotic convergence is uniform, namely,

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

The empirical distribution of a sequence $\boldsymbol{x} \in \mathcal{X}^n$, which will be denoted by $\hat{P}_{\boldsymbol{x}}$, is the vector of relative frequencies $\hat{P}_{\boldsymbol{x}}(x)$ of each symbol $x \in \mathcal{X}$ in $\boldsymbol{x}$. 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 $\hat{P}_{\boldsymbol{x}}$, which is the empirical entropy of $\boldsymbol{x}$, will be denoted by $\hat{H}_{\boldsymbol{x}}(X)$. Analogously, to stress out a distribution, e.g., Q, under which an information measure is calculated, we use subscripts as well, e.g., $H_Q(X)$. 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 $Q_{YY'} = Q_Y Q_{Y'|Y}$, which will be abbreviated by Q. Moreover, when the Y−marginal of such a distribution Q will be the empirical distribution of a specific $\boldsymbol{y}$, we will use the notation Q_Y and not $\hat{P}_{\boldsymbol{y}}$, if the specific $\boldsymbol{y}$ is clear from the context.

The *type class* of $\boldsymbol{x} \in \mathcal{X}^n$, i.e., the set of all vectors in $\mathcal{X}^n$ with empirical distribution $\hat{P}_{\boldsymbol{x}}$, will be denoted by $\mathcal{T}(\boldsymbol{x})$. Similarly, $\mathcal{T}(Q)$ will denote the type class of a specific (empirical) distribution Q. E.g., $\mathcal{T}(Q_{X|Y}|\boldsymbol{y})$ will denote the set vectors $\boldsymbol{x} \in \mathcal{X}^n$ whose relative frequencies equal $Q_{X|Y}$ for the specific $\boldsymbol{y}$.

3.2. Guessing Subject to a Distortion Measure

Alice selects a random n–vector $\boldsymbol{X}$, drawn from a finite alphabet source P. Bob, which is unaware of the realization of $\boldsymbol{X}$, submits a sequence of guesses in the form of yes/no queries: "Is $\boldsymbol{X}$ close within distortion D to $\hat{\boldsymbol{x}}_1$?", "Is $\boldsymbol{X}$ close within distortion D to $\hat{\boldsymbol{x}}_2$?", and so on, where $\hat{\boldsymbol{x}}_i \in \hat{\mathcal{X}}^n$ for each i, until receiving a positive answer[4]. Specifically, Alice gives a positive answer iff Bob's guess is *within an average distortion* D of Alice's realization, that is

$$\frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{x}_i) \le D, \tag{2}$$

where $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ is a single–letter distortion measure. With a slight abuse of notation, we may write $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$ for the n–letter distortion, that is, $\sum_{i=1}^n d(x_i, \hat{x}_i)$. Given $\boldsymbol{x} \in \mathcal{X}^n$, let $\mathcal{S}(\boldsymbol{x}) = \{\hat{\boldsymbol{x}} : \ d(\boldsymbol{x}, \hat{\boldsymbol{x}}) \le nD\}$. Clearly, for a given realization $\boldsymbol{x}$, Bob's goal is to guess a sequence $\hat{\boldsymbol{x}} \in \mathcal{S}(\boldsymbol{x})$.

When considering the deterministic setting, a *guessing list*, $\mathcal{G}_n$, is an ordered list of all members of $\hat{\mathcal{X}}^n$, that is, $\mathcal{G} = \{\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \ldots, \hat{\boldsymbol{x}}_{|\hat{\mathcal{X}}|^n}\}$, and it is associated with a *guessing function*, $G(\boldsymbol{x})$, which is the function that maps $\mathcal{X}^n$ onto $\{1, 2, \ldots, |\hat{\mathcal{X}}|^n\}$ by assigning to each $\boldsymbol{x} \in \mathcal{X}^n$ the smallest integer i for which $\hat{\boldsymbol{x}}_i$ satisfies the distortion constraint (2). Namely, $G(\boldsymbol{x})$ is the number of guesses required until success, using $\mathcal{G}_n$, when $\boldsymbol{X} = \boldsymbol{x}$. In this setting, the goal is to devise a guessing list $\mathcal{G}_n$ that minimizes a certain moment of $G(\boldsymbol{X})$, namely, $\boldsymbol{E}\{G^\rho(\boldsymbol{X})\}$, 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 $\tilde{P}(\hat{\boldsymbol{x}})$. This setting consumes less memory (compared to deterministic guessing which stores the list $\mathcal{G}_n$) and needs no synchronization. Note that, in general, $\tilde{P}(\hat{\boldsymbol{x}})$ may depend on i,

[4] Similar to the common definition in rate–distortion theory, $\hat{\mathcal{X}}$, the reconstruction alphabet, may be different from the source alphabet $\mathcal{X}$. Clearly, the scenario in which $\hat{\mathcal{X}} = \mathcal{X}$ is a special case.

where $i-1$ is the number of guesses submitted thus far. However, such a dependence *will require* memory and synchronization between multiple guessers. More importantly, we will later see that a distribution which is independent of *i still achieves the converse bounds for all setups considered* in this paper, hence there is no significant gain in considering a sequence of guessing distributions, indexed by the number of guesses thus far.

Clearly, in this setting, the goal is to devise a *guessing distribution $\tilde{P}$* that minimizes a certain moment of the guesswork. However, we will also be interested in devising a distribution which is independent of the source Alice has, the moment of the guesswork to be minimized, the distortion measure $d(\cdot,\cdot)$ and the distortion level D.

For a given $\boldsymbol{x}$, the expected number of guesses until a sequence with a small enough distortion is guessed is independent of the underlying distribution used to draw $\boldsymbol{x}$, namely, $\boldsymbol{E}\{G^\rho|\boldsymbol{x}\}$ depends solely on the distribution used to guess, $\tilde{P}(\hat{\boldsymbol{x}})$, and the set $\mathcal{S}(\boldsymbol{x})$. In fact, for $\boldsymbol{E}\{G^\rho|\boldsymbol{x}\}$ we have the following result, which directly utilizes [17, Lemma 1].

Lemma 1.

$$\boldsymbol{E}\{G^\rho|\boldsymbol{x}\} = \sum_{k=1}^{\infty} k^\rho \tilde{P}[\mathcal{S}(\boldsymbol{x})](1-\tilde{P}[\mathcal{S}(\boldsymbol{x})])^{k-1} \doteq \frac{1}{(\tilde{P}[\mathcal{S}(\boldsymbol{x})])^\rho}. \tag{3}$$

The expectation in Lemma 1 is with respect to the guessing distribution, yet for a given source sequence $\boldsymbol{x}$. When $\boldsymbol{x}$ is random, the source distribution has to be taken into account. We can now clearly define the main goal of this paper: devise universal guessing distributions, which achieve the smallest possible *guesswork exponent.* That is, we will be interested in the limit $\lim_{n\to\infty}\frac{1}{n}\log\boldsymbol{E}\{G^\rho\}$, where the expectation in $\boldsymbol{E}\{G^\rho\}$ is over both the source distribution and any randomization in the guessing strategy. Specifically, we will devise universal distributions, compute their resulting exponent $\boldsymbol{E}\{G^\rho\}$, and prove matching converse results, showing that no better exponent can be achieved.

3.3. Universal Guessing Distributions

We will utilize two important universal distributions and later show that these distributions, used for the respective source models, will suffice to achieve the universality goals stated above. A key step in the process will be to bound *the probability of the set $\mathcal{S}(\boldsymbol{x})$ under these distributions.* This is done in Section 5.

Specifically, for memoryless sources, we consider a sequence of randomized guesses, all drawn independently from the universal distribution,

$$\tilde{P}(\hat{\boldsymbol{x}}) = \frac{\exp_2\{-n\hat{H}_{\hat{\boldsymbol{x}}}(\hat{X})\}}{\sum_{\hat{\boldsymbol{z}}\in\hat{\mathcal{X}}^n}\exp_2\{-n\hat{H}_{\hat{\boldsymbol{z}}}(\hat{Z})\}}, \tag{4}$$

for all $\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}^n$. Note that this distribution satisfies $\tilde{P}(\hat{\boldsymbol{x}})\doteq 2^{-n\hat{H}_{\hat{\boldsymbol{x}}}(\hat{X})}$, 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

$$\hat{\boldsymbol{x}} = 01101101110010111 \tag{5}$$

is parsed as

$$0, 1, 10, 11, 01, 110, 010, 111. \tag{6}$$

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

$$\tilde{P}(\hat{\boldsymbol{x}}) = \frac{2^{-LZ(\hat{\boldsymbol{x}})}}{\sum_{\hat{\boldsymbol{z}} \in \hat{\mathcal{X}}^n} 2^{-LZ(\hat{\boldsymbol{z}})}}. \tag{7}$$

Note that since LZ is a uniquely decodable code, we have $\sum_{\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}^n} 2^{-LZ(\hat{\boldsymbol{x}})} \leq 1$ and hence $\tilde{P}(\hat{\boldsymbol{x}}) \geq 2^{-LZ(\hat{\boldsymbol{x}})}$.

4. 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.

4.1. Memoryless Sources

Theorem 1. *For any memoryless source P over $\mathcal{X}$, if sequences are selected independently at random according to the distribution $\tilde{P}(\hat{\boldsymbol{x}}) \doteq 2^{-n\hat{H}_{\hat{\boldsymbol{x}}}(\hat{X})}$, then, the ρ-th guesswork moment, subject to a distortion constraint D, satisfies*

$$\boldsymbol{E}\{G^\rho\} \doteq \exp_2\left\{n \max_Q [\rho R(D,Q) - D(Q\|P)]\right\}. \tag{8}$$

The proof of Theorem 1 follows from Lemma 3 and appears right after it, in Section 5.

Note that Theorem 1 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 $\mathcal{X}$, the best achievable ρ-th guesswork moment, subject to a distortion constraint D, is achievable by randomized guessing and we have*

$$\lim_{n\to\infty} \frac{1}{n} \log \boldsymbol{E}\{G^\rho\} = \max_Q [\rho R(D,Q) - D(Q\|P)]. \tag{9}$$

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.

4.2. Noisy Guessing for Memoryless Sources

We consider the following scenario, studied first in [51] for $D = 0$. Alice draws a random n–vector, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, from a discrete memoryless source (DMS), P, of a finite alphabet, $\mathcal{Y}$. Bob, who is unaware of the realization of $\boldsymbol{Y}$, sequentially submits to Alice a (possibly, infinite) sequence of guesses, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$, where each $\boldsymbol{x}_i$ is a vector of length n, whose components take on values in a finite alphabet, $\mathcal{X}$. Before arriving to Alice, each guess, $\boldsymbol{x}_i$, undergoes a discrete memoryless channel (DMC), defined by a matrix of single–letter input–output transition probabilities, $W = \{W(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$. Let $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ be the corresponding noisy versions of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$, 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, $d(\boldsymbol{Y}, \boldsymbol{Y}_i) \le nD$. Clearly, the number of guesses, G, until the first successful guess, is a random variable that depends on the source vector $\boldsymbol{Y}$ and the guesses, $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$.. It is given by

$$G = G(\boldsymbol{Y}, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots) = \sum_{k=1}^{\infty} k \cdot \mathcal{I}\{d(\boldsymbol{Y}, \boldsymbol{Y}_k) \le nD\} \cdot \prod_{i=1}^{k-1} [1 - \mathcal{I}\{d(\boldsymbol{Y}, \boldsymbol{Y}_i) \le nD\}]. \tag{10}$$

For a given list of guesses, $\mathcal{G}_n = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\}$, $\boldsymbol{x}_i \in \mathcal{X}^n$, $i = 1, 2, \ldots$, the ρ–th moment of G is given by

$$\boldsymbol{E}_{\mathcal{G}_n}\{G^\rho\} = \sum_{\boldsymbol{y} \in \mathcal{Y}^n} P(\boldsymbol{y}) \cdot \sum_{k=1}^{\infty} k^\rho \cdot W(\mathcal{S}(\boldsymbol{y})|\boldsymbol{x}_k) \cdot \prod_{i=1}^{k-1} [1 - W(\mathcal{S}(\boldsymbol{y})|\boldsymbol{x}_i)], \tag{11}$$

where

$$W(\boldsymbol{y}|\boldsymbol{x}) = \prod_{t=1}^{n} W(y_t|x_t) \tag{12}$$

and $\mathcal{S}(\boldsymbol{y}) = \{\boldsymbol{y}' :\ d(\boldsymbol{y}, \boldsymbol{y}') \le nD\}$. Randomized guessing lists (where the deterministic guesses, $\{\boldsymbol{x}_i\}$, are replaced by random ones, $\{\boldsymbol{X}_i\}$) are allowed as well. In this case, eq. (11) would include also an expectation w.r.t. the randomness of the guesses.

Let P, W, and $\rho \ge 0$ be given. For two given distributions, Q_X and Q_Y, defined on $\mathcal{X}$ and $\mathcal{Y}$, respectively, define

$$\Gamma(Q_X, Q_Y) = \inf\{D(\tilde{Q}_{Y|X} \| W | Q_X) :\ (Q_X \odot \tilde{Q}_{Y|X})_Y = Q_Y\}, \tag{13}$$

where

$$D(Q_{Y|X} \| P_{Y|X} | Q_X) = \sum_{x \in \mathcal{X}} Q_X(x) \sum_{y \in \mathcal{Y}} Q_{Y|X}(y|x) \log \frac{Q_{Y|X}(y|x)}{P_{Y|X}(y|x)}$$

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., $\sum_{x \in \mathcal{X}} Q_X(x) \tilde{Q}_{Y|X}(y|x) = Q_Y(y)$ for all $y \in \mathcal{Y}$. Next, define

$$\Gamma(Q_Y) = \inf_{Q_X} \Gamma(Q_X, Q_Y) = \inf_{Q_{X|Y}} D(Q_{Y|X} \| W | Q_X). \tag{14}$$

Finally, we define

$$E_W(\rho) = \sup_{Q_Y} \left\{ \rho \hat{R}_W(D, Q_Y) - D(Q_Y \| P) \right\}, \tag{15}$$

where

$$\hat{R}_W(D, Q_Y) \stackrel{\triangle}{=} \inf_{\{Q_{Y'|Y}:\ \boldsymbol{E}_Q d(Y,Y') \leq D\}} [I_Q(Y;Y') + \Gamma(Q_{Y'})]. \tag{16}$$

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 $\mathcal{Y}$, and Bob's guesses, before arriving to Alice, undergo a discrete memoryless channel (DMC) $W = \{W(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$. Then, the optimal guesswork ρ-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 ρ, and satisfies*

$$\lim_{n\to\infty} \frac{1}{n} \log \boldsymbol{E}\{G^\rho\} = E_W(\rho). \tag{17}$$

The proof of the direct part of Theorem 2 follows from Lemma 5 and appears right after it, in Section 5. The proof of the converse part is given in Section 6.

Similar to (8), for the clean memoryless case, the exponent in (15) 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, Γ, 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 (15) has two alternative forms, which facilitate numerical evaluation and give an operational meaning.

Lemma 2. *The function $E_W(\rho)$ in (15) has the following two alternative forms,*

$$\begin{aligned} E_W(\rho) &= \sup_{s\geq 0} \inf_V \left\{ \log\left(\sum_y \frac{P(y)}{\left[\sum_x V(x) U_s(x,y)\right]^\rho} \right) - \rho s D \right\} & (18) \\ &= \sup_{s\geq 0} \inf_{M\in\mathcal{CH}(W)} \left\{ \log\left(\sum_y \frac{P(y)}{\left[\sum_{y'} M(y') e^{s[D-d(y,y')]}\right]^\rho} \right) \right\}, & (19) \end{aligned}$$

where the infimum on V is over all probability distributions on $\mathcal{X}$, $U_s(x,y) \stackrel{\triangle}{=} \sum_{y'} W(y'|x) e^{-sd(y,y')}$, $M(y') \stackrel{\triangle}{=} \sum_x V(x) W(y'|x)$ and $\mathcal{CH}(W)$ is the convex hull of $\{W(\cdot|x),\ x \in \mathcal{X}\}$.

Lemma 2 is proved in Appendix A.1. Note that the form (18) can be easily evaluated for small alphabets and simple channels (e.g., binary alphabets, Hamming distortion and the binary symmetric channel). The form (19) has an operational meaning. To see this, first recall the result in [38, eq. (19)], stating that $\boldsymbol{E}\{V_\rho(X)\} = \sum_{x\in\mathcal{X}} \frac{P(x)}{\hat{P}(x)^\rho}$, where $V_\rho(X) = \binom{G(X)+\rho-1}{\rho}$ and $\binom{n}{k}$ is the generalized binomial coefficient ([38, eq. (11)-(12)]). As $\boldsymbol{E}\{V_\rho(X)\}$ approximates the ρ–th moment of the guesswork up to a constant factor, $\sum_{x\in\mathcal{X}} \frac{P(x)}{\hat{P}(x)^\rho}$ also has the interpretation of a guesswork's ρ–th moment, when the source distribution is $P(x)$ and the guessing is randomized according to $\hat{P}(x)$. A similar interpretation was given in [51, Section V.2]. In the problem at hand, $P(y)$, at the numerator of (19), is the source distribution. To see that the bracketed expression at the denominator (without the power of ρ) can be viewed as the *success probability of a single guess under some guessing strategy*, note that $M(y')$ is the channel output distribution when its input is i.i.d. according to V. Thus the bracketed expression at the denominator can be thought of as the Chernoff bound for the probability of a single successful guess within distortion D. To conclude this discussion, note that the minimizing V has the operative meaning of the optimal (non–universal) i.i.d. random guessing distribution.

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 $Q_{Y'}$ such that $\Gamma(Q_{Y'}) = 0$ for any Q_Y (or, at least, for the dominant one), then the guessing exponent is the same as in the clean–channel case.*

4.3. Sources with Memory

Our model for sources with memory will be that of Ψ-mixing sources. Consider a stationary source $\{X_i\}_{i=-\infty}^{i=\infty}$ over a probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{F}_j^l$ denote the σ-field of events generated by the random variables $\{X_m\}_{j\leq m\leq l}$. A *measure of dependence*, $\Psi(k)$, for the source $\{X\}$ is defined by

$$\Psi(k) = \sup_{A\in\mathcal{F}^0_{-\infty}, B\in\mathcal{F}^\infty_k, P(A)>0, P(B)>0} \left| \frac{P(A\cap B)}{P(A)P(B)} - 1 \right|. \tag{20}$$

A stationary source is said to be Ψ*-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 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 ρ as

$$E_P(D, \rho) = \liminf_{n\to\infty} \sup_{Q^n} \frac{1}{n} \left[\rho R_n(D, Q^n) - D\left(Q^n||P^n\right)\right]. \tag{21}$$

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, $\liminf_{n\to\infty} \sup_{Q^n} \frac{1}{n} R_n(D, Q^n)$ has the same property.*

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

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

$$\limsup_{n\to\infty} \frac{1}{n} \log \boldsymbol{E}\{G^\rho\} = E_P(D, \rho). \tag{22}$$

The randomized guessing strategy is based on the incremental parsing procedure of LZ78. It was described in Section 3.3. For the proof of Theorem 3, Section 5.3 includes a key result, focusing on the probability this distribution assigns to sequences satisfying the distortion constraint. Then, using this result, Section 7.1 completes the details of the direct. Section 7.2 includes the matching converse.

4.4. 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, 64, 65, 66, 67, 68].

A finite–state guessing machine is defined by the tuple $(\mathcal{U}, \hat{\mathcal{X}}, \mathcal{Z}, \Delta, f, g)$, where $\mathcal{U}$ is its input alphabet, assumed binary, $\hat{\mathcal{X}}$ is the output alphabet, $\mathcal{Z}$ is a finite set of states, with $|\mathcal{Z}| = s$, $\Delta : \mathcal{Z} \to \{0, 1, 2, \ldots\}$ defines the number of input bits used at each state, and f, g are the output and next–state functions, that is, $f : \mathcal{Z} \times \mathcal{U}^* \to \hat{\mathcal{X}}$ and $g : \mathcal{Z} \times \mathcal{U}^* \to \mathcal{Z}$. When a binary sequence $u_1, u_2, \ldots,$ $u_i \in \mathcal{U}$, is fed to the guessing machine, the output is a sequence $\hat{x}_1, \hat{x}_2, \ldots,$ $\hat{x}_i \in \hat{\mathcal{X}}$, according to the following recursive equations:

$$\begin{aligned} t_i &= t_{i-1} + \Delta(z_i), \qquad t_0 = 0, & (23)\\ v_i &= (u_{t_{i-1}+1}, u_{t_{i-1}+2}, \ldots, u_{t_i}), & (24)\\ \hat{x}_i &= f(z_i, v_i), & (25)\\ z_{i+1} &= g(z_i, v_i). & (26) \end{aligned}$$

The guessing game is thus as follows: Alice has an individual sequence $\boldsymbol{x}$ 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 $\mathcal{Z}$ and producing output symbols in $\hat{\mathcal{X}}$. When the output is of length n Bob submits it to Alice and asks "is the sequence $\hat{\boldsymbol{x}}$ within distortion nD of $\boldsymbol{x}$?" If the answer is affirmative, the game ends. Otherwise, the machine is restarted to its initial state (say, some $z_1 \in \mathcal{Z}$) and the game repeats.

For a finite–state guessing machine F, denote by $G_F(\boldsymbol{x})$ the (random, due to the random input to the machine) number of guesses until $\boldsymbol{x}$ 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 ρ–th moment of $G_F(\boldsymbol{x})$, 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 $\hat{\boldsymbol{x}}$. The compression ration of E is defined by $\rho_E(\hat{\boldsymbol{x}}) = \frac{L_E(b^n)}{n}$, and the finite–state compresibility of $\hat{\boldsymbol{x}}$ is defined by

$$\rho_K(\hat{\boldsymbol{x}}) = \min_{E \in \mathcal{E}(K)} \rho_E(\hat{\boldsymbol{x}}), \tag{27}$$

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

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

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

$$\boldsymbol{E}\{G_F^{\rho}|\boldsymbol{x}\} \stackrel{\cdot}{\geq} \left[\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\leq nD\}} 2^{-LZ(\hat{\boldsymbol{x}})}\right]^{-\rho}. \tag{28}$$

(ii) Alternatively, we have

$$\boldsymbol{E}\{G_F^{\rho}|\boldsymbol{x}\} \stackrel{\cdot}{\geq} \left[\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\leq nD\}} \exp_2\left\{-n\left(\rho_{K(l)}(\hat{\boldsymbol{x}}) - \frac{\log\left(2s^3e\right)}{l}\right)\right\}\right]^{-\rho}, \tag{29}$$

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* (28) *is achieved using the guessing distribution* (7)*. (ii) Assume l divides n and consider the distribution*

$$\tilde{P}(\hat{\boldsymbol{x}}) = \prod_{i=0}^{n/l-1}\left[\frac{2^{-LZ(\hat{x}_{il+1}^{il+l})}}{\sum_{\hat{x}^l\in\hat{\mathcal{X}}^l} 2^{-LZ(\hat{x}^l)}}\right], \tag{30}$$

which is implementable with a finite–state machine with at most $l\alpha^l$ states for any n. Then

$$\boldsymbol{E}\{G_{LZ}^{\rho}|\boldsymbol{x}\} \stackrel{\cdot}{\leq} \left[\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\leq nD\}} \exp_2\left\{-n\Big[\rho_K(\hat{\boldsymbol{x}}) + \frac{\log(4K^2)}{(1-\epsilon(l))\log l} + \frac{K^2\log(4K^2)}{l} + \epsilon(l)\Big]\right\}\right]^{-\rho}, \tag{31}$$

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

The complete proof of Theorem 4 is given in Appendix A.7.

At this point, a few remarks are in order. First, note that the universal guessing distribution in (7) 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. Moreover, while this distribution 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[5], as well as a block–wise version (30) which still asymptotically achieves the lower bound.

5. Three Lemmas on the Probability of $\mathcal{S}(x)$ Under Universal Guessing Distributions

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 $\tilde{P}(\hat{\boldsymbol{x}})$, evaluating

[5] This can be done without directly computing the complex sum in the denominator, but, rather, feeding an LZ decoder with uniform i.i.d. bits [17, Section V.D]. Thus, these algorithms implement (7) directly and efficiently, without the need for the block-wise implementation in (30).

the probability of hitting the distortion-achieving set, $\tilde{P}[\mathcal{S}(\boldsymbol{x})]$, and, finally, taking into account the actual source distribution P, computing the guesswork moment, $\boldsymbol{E}\{G^\rho\} = \boldsymbol{E}_P \boldsymbol{E}\{G^\rho|\boldsymbol{x}\}$.

Evaluating $\tilde{P}[\mathcal{S}(\boldsymbol{x})]$ is a key technical challenge. In this section, we include three lemmas which indeed evaluate (or at least lower bound) $\tilde{P}[\mathcal{S}(\boldsymbol{x})]$ for the three major scenarios in this paper: memoryless sources (Lemma 3), noisy guessing (Lemma 4) and sources with memory (Lemma 5). In fact, in the first and third 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.

5.1. Memoryless Sources

For memoryless sources, we have the following.

Lemma 3. *Assume sequences of length n are chosen independently at random according to the distribution $\tilde{P}(\hat{\boldsymbol{x}}) \doteq 2^{-n\hat{H}_{\hat{\boldsymbol{x}}}(\hat{X})}$ in (4). Then, for any $\boldsymbol{x} \in \mathcal{X}^n$*

$$\tilde{P}[\mathcal{S}(\boldsymbol{x})] \doteq 2^{-nR(D,\hat{P}_{\boldsymbol{x}})}, \tag{32}$$

where $R(D, \hat{P}_{\boldsymbol{x}})$ is the rate distortion function of $\hat{P}_{\boldsymbol{x}}$.

Proof.

$$\begin{aligned}
\tilde{P}[\mathcal{S}(\boldsymbol{x})] &= \sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}}) \le nD\}} \tilde{P}(\hat{\boldsymbol{x}}) && (33)\\
&\doteq \sum_{\{\mathcal{T}(Q_{\hat{X}|X}|\boldsymbol{x}):\ \boldsymbol{E}_Q d(X,\hat{X}) \le D\}} |\mathcal{T}(Q_{\hat{X}|X}|\boldsymbol{x})| \cdot 2^{-n\hat{H}_{\hat{\boldsymbol{x}}}(\hat{X})} && (34)\\
&\doteq \max_{\{Q_{\hat{X}|X}:\ \boldsymbol{E}_Q d(X,\hat{X}) \le D\}} 2^{nH_Q(\hat{X}|X)} \cdot 2^{-nH_Q(\hat{X})} && (35)\\
&= \max_{\{Q_{\hat{X}|X}:\ \boldsymbol{E}_Q d(X,\hat{X}) \le D\}} 2^{-nI_Q(X;\hat{X})} && (36)\\
&= \exp_2\left\{-n \min_{\{Q_{\hat{X}|X}:\ \boldsymbol{E}_Q d(X,\hat{X}) \le D\}} I_Q(X;\hat{X})\right\} && (37)\\
&= 2^{-nR(D,\hat{P}_{\boldsymbol{x}})}, && (38)
\end{aligned}$$

where the last equality is since the minimum is taken over all conditional types $Q_{\hat{X}|X}$, conditioned on the specific $\boldsymbol{x}$, satisfying the distortion constraint. □

Lemma 3 is the key step in proving the direct result for memoryless sources. In fact, using Lemma 1 of Section 3 together with Lemma 3, the direct result in Theorem 1 easily follows:

Proof of Theorem 1.

$$\boldsymbol{E}\{G^\rho\} = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \cdot \boldsymbol{E}\{G^\rho|\boldsymbol{x}\} \tag{39}$$

$$\doteq \sum_{\boldsymbol{x}} 2^{-n\left(\hat{H}_{\boldsymbol{x}}(X)+D(\hat{P}_{\boldsymbol{x}}\|P)\right)} \cdot 2^{n\rho R(D,\hat{P}_{\boldsymbol{x}})} \tag{40}$$

$$= \sum_{Q\in\mathcal{P}_n} |\mathcal{T}_Q| 2^{-n\left(H_Q(X)+D(Q\|P)\right)} \cdot 2^{n\rho R(D,Q)} \tag{41}$$

$$\stackrel{\cdot}{\leq} \sum_{Q\in\mathcal{P}_n} 2^{nH_Q(X)} \cdot 2^{-n\left(H_Q(X)+D(Q\|P)\right)} \cdot 2^{n\rho R(D,Q)} \tag{42}$$

$$\doteq \exp_2\left\{n \max_Q [\rho R(D,Q) - D(Q\|P)]\right\}. \tag{43}$$

□

5.2. Noisy Guessing

Analogously to Lemma 3, we now wish to evaluate the probability of $\mathcal{S}(\boldsymbol{y})$ 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 $\mathcal{S}(\boldsymbol{y})$ under the distribution induced by the guessing distribution (4).

Lemma 4. *Assume that sequences of length n are chosen independently at random according to the distribution $\tilde{P}(\boldsymbol{x}) \doteq 2^{-n\hat{H}_{\boldsymbol{x}}(X)}$ in* (4)*, then sent through the memoryless channel $W(\boldsymbol{y}|\boldsymbol{x})$. Denote by Q the distribution induced on the $\boldsymbol{y}$–vectors. Then,*

$$Q[\mathcal{S}(\boldsymbol{y})] \doteq \exp_2\{-n\hat{R}_W(D, \hat{P}_{\boldsymbol{y}})\}, \tag{44}$$

where $\hat{R}_W(D, \hat{P}_{\boldsymbol{y}})$ is given in (16)*.*

Proof. Guessing using the distribution in (4) induces the following distribution on the $\boldsymbol{y}$–vectors:

$$Q(\boldsymbol{y}) \doteq \sum_{\boldsymbol{x}\in\mathcal{X}^n} \exp_2\{-n\hat{H}_{\boldsymbol{x}}(X)\} \cdot W(\boldsymbol{y}|\boldsymbol{x}) \tag{45}$$

$$\doteq \sum_{\mathcal{T}(Q_{X|Y}|\boldsymbol{y})} |\mathcal{T}(Q_{X|Y}|\boldsymbol{y})| \cdot \exp_2\{-n[H_Q(X) + H_Q(Y|X) + D(Q_{Y|X}\|W|Q_X)]\} \tag{46}$$

$$\doteq \max_{Q_{X|Y}} \exp_2\{n[H_Q(X|Y) - H_Q(X) - H_Q(Y|X) - D(Q_{Y|X}\|W|Q_X)]\} \tag{47}$$

$$= \exp_2\{-nH_Q(Y)\} \cdot \max_{Q_{X|Y}} \exp_2\{-nD(Q_{Y|X}\|W|Q_X)\} \tag{48}$$

$$= \exp_2\{-n[H_Q(Y) + \Gamma(Q_Y)]\}. \tag{49}$$

Note that Q above is the distribution whose Y–marginal is $\hat{P}_{\boldsymbol{y}}$, hence $H_Q(Y)$ is, in fact, $\hat{H}_{\boldsymbol{y}}(Y)$. Thus, given that $\boldsymbol{Y} = \boldsymbol{y}$, the probability of a single successful guess within distortion D is given by

$$\sum_{\{\boldsymbol{y}':\ d(\boldsymbol{y},\boldsymbol{y}')\leq nD\}} Q(\boldsymbol{y}') \tag{50}$$

$$\doteq \sum_{\{\boldsymbol{y}':\ d(\boldsymbol{y},\boldsymbol{y}')\leq nD\}} \exp_2\{-n[H_Q(Y') + \Gamma(Q_{Y'})]\} \tag{51}$$

$$\doteq \max_{\{Q_{Y'|Y}:\ \boldsymbol{E}_Q d(Y,Y')\leq D\}} |\mathcal{T}(Q_{Y'|Y}|\boldsymbol{Y} = \boldsymbol{y})| \cdot \exp_2\{-n[H_Q(Y') + \Gamma(Q_{Y'})]\} \tag{52}$$

$$\doteq \max_{\{Q_{Y'|Y}\colon \boldsymbol{E}_Q d(Y,Y')\le D\}} \exp_2\{nH_Q(Y'|Y)\}\cdot\exp_2\{-n[H_Q(Y')+\Gamma(Q_{Y'})]\} \tag{53}$$

$$\doteq \exp_2\left\{-n \min_{\{Q_{Y'|Y}\colon \boldsymbol{E}_Q d(Y,Y')\le D\}} [I_Q(Y;Y')+\Gamma(Q_{Y'})]\right\} \tag{54}$$

$$= \exp_2\{-n\hat{R}_W(D,Q_Y)\}. \tag{55}$$

Remembering that in the above Q_Y is empirical distribution of the specific $\boldsymbol{y}$, that is, $\hat{P}_{\boldsymbol{y}}$, completes the proof. □

The direct result for noisy guessing under distortion now easily follows.

Proof of Theorem 2 (Direct Part). Applying Lemma 1 on the result in Lemma 4, we have

$$\boldsymbol{E}\{G^\rho|\boldsymbol{Y}=\boldsymbol{y}\} \doteq \exp_2\{n\rho\hat{R}_W(D,\hat{P}_{\boldsymbol{y}})\}. \tag{56}$$

The proof now follows the same steps as in (39)–(43), resulting in

$$\boldsymbol{E}\{G^\rho\} = \sum_{\boldsymbol{y}} P(\boldsymbol{y})\cdot\boldsymbol{E}\{G^\rho|\boldsymbol{Y}=\boldsymbol{y}\} \tag{57}$$

$$\doteq \exp_2\{n\max_{Q_Y}[\rho\hat{R}_W(D,Q_Y)-D(Q_Y\|P)]\}, \tag{58}$$

which completes the proof of this achievability result. □

5.3. Sources with Memory

Before presenting the main lemma in this section, a few definitions are required.

With proper scaling and gaps, Ψ-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 $K \gg k$. We further assume that $K+k$ divides n and denote $\frac{n}{K+k}=m$. Hence,

$$x_1^n = x_1^K, x_{K+1}^{K+k}, \ldots, x_{n-K-k+1}^{n-k}, x_{n-k+1}^n \tag{59}$$

$$\triangleq \underline{x}_1, \underline{g}_1, \ldots, \underline{x}_m, \underline{g}_m, \quad \underline{x}_i \in \mathcal{X}^K, \underline{g}_i \in \mathcal{X}^k, 1\le i\le m, \tag{60}$$

where we refer to the K-tuples $\underline{x}_i$ as *blocks*, and to the k-tuples $\underline{g}_i$ 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_1^n as the vector of empirical frequencies for each possible block of size K in x_1^n, disregarding the content of the gaps. That is, denoting by $N_{\boldsymbol{x}}(a_1^K)$ the number of occurrences of $a_1^K \in \mathcal{X}^K$ in the blocks $\{\underline{x}_i\}_{i=1}^m$ of $\boldsymbol{x}$, we have

$$\hat{P}_{\boldsymbol{x}}^K \triangleq \left(\frac{N_{\boldsymbol{x}}(a_1^K)}{m}, a_1^K \in \mathcal{X}^K\right). \tag{61}$$

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, $T^K(P^K) = \left\{\boldsymbol{x} \in \mathcal{X}^n : \hat{P}^K_{\boldsymbol{x}} = P^K\right\}$. Finally, denoting by $\mathcal{T}^K$ the set of all possible K-types P^K, clearly,

$$\left|\mathcal{T}^K\right| \leq (m+1)^{|\mathcal{X}|^K}. \tag{62}$$

The following lemma generalizes Lemma 3 to guessing sequences with memory based on LZ parsing, via the distribution in (7).

Lemma 5. *Assume sequences of length n are chosen independently at random according to the distribution $\tilde{P}(\hat{\boldsymbol{x}}) \doteq 2^{-LZ(\hat{\boldsymbol{x}})}$ in (7). Then, for any $\delta, \epsilon > 0$ there exists a sequence $\epsilon_m \to 0$ as $m \to \infty$ such that*

$$\tilde{P}[\mathcal{S}(\boldsymbol{x})] \stackrel{\cdot}{\geq} 2^{-\nu_{K+k}(n)}(1-2\epsilon_m)\exp_2\left\{-m\left[R_K(D', \hat{P}^K_{\boldsymbol{x}}) + \epsilon\right]\right\}, \tag{63}$$

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

To prove Lemma 5 we first, in Proposition 1 below, consider a specific *block–memoryless distribution*, induced by the distribution which achieves the rate–distortion function $R_K(D, \hat{P}^K_{\boldsymbol{x}})$, 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, $\underline{\mathrm{S}}(\boldsymbol{x}) \subset \mathcal{S}(\boldsymbol{x})$, to be defined precisely later, to the rate-distortion function of a block memoryless source. Then, we upper bound this block–memoryless distribution using the code length of the LZ encoder. Finally, we connect the two results to prove Lemma 5.

Proposition 1. *Fix $\boldsymbol{x} \in \mathcal{X}^n$ with a K-type $\hat{P}^K_{\boldsymbol{x}}$. 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 $\hat{x} \in \hat{\mathcal{X}}^K$ and $x \in \mathcal{X}^k$ the conditional distribution achieving $R_K(D', \hat{P}^K_{\boldsymbol{x}})$. Finally, denote by Q^K the resulting distribution on the construction alphabet $\hat{\mathcal{X}}^K$, that is, $Q^K(\hat{x}) = \sum_x P^K(\hat{x}|x)\hat{P}^K_{\boldsymbol{x}}(x)$. For any $\epsilon > 0$, the following holds*

$$\sum_{\hat{\boldsymbol{s}}_1^m \in (\hat{\mathcal{X}}^K)^m:\ \sum_i d(\underline{x}_i, \hat{\boldsymbol{s}}_i) \leq mK(D-\delta_2)} \prod_{i=1}^m Q^K(\hat{\boldsymbol{s}}_i) \geq (1-2\epsilon_m)\exp_2\left\{-m\left[R_K(D', \hat{P}^K_{\boldsymbol{x}}) + \epsilon\right]\right\}. \tag{64}$$

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

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

$$\sum_{\hat{\boldsymbol{x}} \in \mathcal{S}(\boldsymbol{x})} Q^n(\hat{\boldsymbol{x}}) \geq (1-2\epsilon_n)\exp_2\left\{-n\left[R(D^-, \hat{P}_{\boldsymbol{x}}) + \epsilon\right]\right\}, \tag{65}$$

where $Q^n(\hat{\boldsymbol{x}}) = \prod_{i=1}^n Q(\hat{x})$ and $\epsilon_n \to 0$ as $n \to \infty$ for all $\boldsymbol{x} \in \mathcal{X}^n$. To prove (65), we follow [69, eq. (3.2.8)-(3.2.10)]. Define

$$\mathcal{T}(\boldsymbol{x}) = \left\{\hat{\boldsymbol{x}} : \log\frac{P^n(\hat{\boldsymbol{x}}|\boldsymbol{x})}{Q^n(\hat{\boldsymbol{x}})} \leq n\left[R(D^-, \hat{P}_{\boldsymbol{x}}) + \epsilon\right]\right\}, \tag{66}$$

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

$$\sum_{\hat{\boldsymbol{x}}\in\mathcal{S}(\boldsymbol{x})} Q^n(\hat{\boldsymbol{x}}) \geq \sum_{\hat{\boldsymbol{x}}\in\mathcal{S}(\boldsymbol{x})\cap\mathcal{T}(\boldsymbol{x})} Q^n(\hat{\boldsymbol{x}}) \tag{67}$$

$$\geq \exp_2\left\{-n\left[R(D^-, \hat{P}_{\boldsymbol{x}}) + \epsilon\right]\right\} \sum_{\hat{\boldsymbol{x}}\in\mathcal{S}(\boldsymbol{x})\cap\mathcal{T}(\boldsymbol{x})} P^n(\hat{\boldsymbol{x}}|\boldsymbol{x}) \tag{68}$$

$$\geq \exp_2\left\{-n\left[R(D^-, \hat{P}_{\boldsymbol{x}}) + \epsilon\right]\right\} \left[\sum_{\hat{\boldsymbol{x}}\in\mathcal{S}(\boldsymbol{x})} P^n(\hat{\boldsymbol{x}}|\boldsymbol{x}) - \sum_{\hat{\boldsymbol{x}}\in\mathcal{T}^c(\boldsymbol{x})} P^n(\hat{\boldsymbol{x}}|\boldsymbol{x})\right]. \tag{69}$$

To conclude the proof of (65), we have to show that there exists a sequence $\epsilon_n \to 0$ such that both $\sum_{\hat{\boldsymbol{x}}\in\mathcal{S}(\boldsymbol{x})} P^n(\hat{\boldsymbol{x}}|\boldsymbol{x}) \geq 1-\epsilon_n$ and $\sum_{\hat{\boldsymbol{x}}\in\mathcal{T}^c(\boldsymbol{x})} P^n(\hat{\boldsymbol{x}}|\boldsymbol{x}) \leq \epsilon_n$ uniformly in $\boldsymbol{x}$. To this end, we have the following two claims, whose proofs are given in Appendix A.

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

Claim 2. *There exists* ϵ''_n *such that for all* $\boldsymbol{x}$, $\sum_{\hat{\boldsymbol{x}}\in\mathcal{T}^c(\boldsymbol{x})} P^n(\hat{\boldsymbol{x}}|\boldsymbol{x}) \leq \epsilon''_n$ *with* $\epsilon''_n \to 0$ *as* $n \to \infty$.

Taking $\epsilon_n = \max(\epsilon'_n, \epsilon''_n)$ completes the proof of (65).

To complete the proof of Proposition 1, note that (65) 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 $\hat{\mathcal{X}}^K$. However, we take this a step further, and apply it to a sequence $\boldsymbol{x}$ of length n over $\mathcal{X}$, using a *memoryless distribution* over K-tuples and the K-type of $\boldsymbol{x}$. Namely, given any $\boldsymbol{x} \in \mathcal{X}^n$, apply (65) on the sequence $(\underline{x}_1, \ldots, \underline{x}_m)$ (that is, only the K-tuple blocks of $\boldsymbol{x}$, 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 $\hat{P}^K_{\boldsymbol{x}}$. Thus, setting a per-super-letter distortion level $K(D-\delta_2)$, the proposition follows. □

Using Proposition 1 above, we can now prove Lemma 5.

Proof. (Lemma 5) Let $M^{K+k}(\cdot)$ be any probability distribution on $\hat{\mathcal{X}}^{K+k}$ and

$$\prod_{i=1}^m M^{K+k}\left(\hat{x}_{(i-1)(K+k)+1}^{i(K+k)}\right) \tag{70}$$

be the product distribution on $\hat{\mathcal{X}}^n$. We first have Claim 3 below, which can be viewed as an extension of Ziv's inequality [34, Theorem 1], and is proved in Appendix A.

Claim 3. *For any sequence* $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}^n$ *and any distribution* $M^{K+k}(\cdot)$ *on* $\hat{\mathcal{X}}^{K+k}$*, the LZ code length (in bits),* $LZ(\hat{\boldsymbol{x}})$*, satisfies*

$$LZ(\hat{\boldsymbol{x}}) \leq -\log\left[\prod_{i=1}^m M^{K+k}\left(\hat{x}_{(i-1)(K+k)+1}^{i(K+k)}\right)\right] + \nu_{K+k}(n), \tag{71}$$

where $\nu_{K+k}(n)$ *depends on* $|\hat{\mathcal{X}}|$ *yet satisfies* $\lim_{n\to\infty} \frac{1}{n}\nu_{K+k}(n) = O\left(\frac{1}{K}\right)$ *for any fixed* K *and* $k < K$.

Similar to (59), consider the partition of a sequence $\hat{\boldsymbol{x}}$ of length n over $\hat{\mathcal{X}}$ into non-overlapping blocks of length K, followed by gaps of length k

$$\hat{\boldsymbol{x}} = \underline{\hat{x}}_1, \underline{\hat{g}}_1, \dots, \underline{\hat{x}}_m, \underline{\hat{g}}_m. \tag{72}$$

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

$$\underline{\mathcal{S}}(\boldsymbol{x}) = \left\{ \hat{\boldsymbol{x}} : \underline{\hat{g}}_i = (\hat{x}_0, \dots, \hat{x}_0) \text{ for all } i,\ \sum_i d(\underline{x}_i, \underline{\hat{x}}_i) \le mK(D - \delta_2) \right\}. \tag{73}$$

That is, $\underline{\mathrm{S}}(\boldsymbol{x})$ includes only sequences in $\mathcal{S}(\boldsymbol{x})$ 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

$$\begin{aligned}
\tilde{P}[\mathcal{S}(\boldsymbol{x})] &\ge \sum_{\hat{\boldsymbol{x}} \in \mathcal{S}(\boldsymbol{x})} 2^{-LZ(\hat{\boldsymbol{x}})} && (74)\\
&\ge 2^{-\nu_{K+k}(n)} \sum_{\hat{\boldsymbol{x}} \in \mathcal{S}(\boldsymbol{x})} \prod_{i=1}^{m} M^{K+k}\left(\hat{x}_{(i-1)(K+k)+1}^{i(K+k)}\right) && (75)\\
&\ge 2^{-\nu_{K+k}(n)} \sum_{\hat{\boldsymbol{x}} \in \mathcal{S}(\boldsymbol{x})} \prod_{i=1}^{m} Q^{K}\left(\hat{x}_{(i-1)(K+k)+1}^{i(K+k)-k}\right) \mathbb{1}_{\left\{\hat{x}_{i(K+k)-k+1}^{i(K+k)} = (\hat{x}_0, \dots \hat{x}_0)\right\}} && (76)\\
&= 2^{-\nu_{K+k}(n)} \sum_{\hat{\boldsymbol{x}} \in \underline{\mathcal{S}}(\boldsymbol{x})} \prod_{i=1}^{m} Q^{K}\left(\hat{x}_{(i-1)(K+k)+1}^{i(K+k)-k}\right) && (77)\\
&= 2^{-\nu_{K+k}(n)} \sum_{\hat{\boldsymbol{s}}_1^m \in (\hat{\mathcal{X}}^K)^m :\ \sum_i d(\underline{x}_i, \hat{\boldsymbol{s}}_i) \le mK(D-\delta_2)} \prod_{i=1}^{m} Q^{K}\left(\hat{\boldsymbol{s}}_i\right) && (78)\\
&\ge 2^{-\nu_{K+k}(n)} (1 - 2\epsilon_m) \exp_2\left\{ -m \left[R_K(D', \hat{P}^K_{\boldsymbol{x}}) + \epsilon \right] \right\}, && (79)
\end{aligned}$$

where (75) is by Claim 3 and the last inequality follows from Proposition 1. □

6. Converse for Noisy Guesses

The direct part of Theorem 2 was proved in Section 5.2 via Lemma 4. In this section, we prove the converse part of Theorem 2. We begin from a simple preparatory lemma.

Lemma 6. *Let $\mathcal{T}(Q_Y)$ be a given type class of $\boldsymbol{y}$–vectors, and let $\boldsymbol{x} \in \mathcal{T}(Q_X)$ be given. Then,*

$$\frac{1}{|\mathcal{T}(Q_Y)|} \sum_{\boldsymbol{y} \in \mathcal{T}(Q_Y)} W[\mathcal{S}(\boldsymbol{y}) | \boldsymbol{x}] \overset{\cdot}{\le} \exp_2\{-n\hat{R}_W(D, Q_Y)\} \qquad \forall\ \boldsymbol{x} \in \mathcal{X}^n. \tag{80}$$

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 $\boldsymbol{x}$ is drawn from the universal distribution used in the direct results, Lemma 6, which is later utilized in the *converse result*, does not assume any distribution on $\boldsymbol{x}$. Yet, it includes averaging over $\boldsymbol{y} \in \mathcal{T}(Q_Y)$.

Proof.

$$\frac{1}{|\mathcal{T}(Q_Y)|}\sum_{\boldsymbol{y}\in\mathcal{T}(Q_Y)} W[\mathcal{S}(\boldsymbol{y})|\boldsymbol{x}] \tag{81}$$

$$= \frac{1}{|\mathcal{T}(Q_Y)|}\sum_{\boldsymbol{y}\in\mathcal{T}(Q_Y)}\sum_{\{\boldsymbol{y}':\ d(\boldsymbol{y},\boldsymbol{y}')\le nD\}} W(\boldsymbol{y}'|\boldsymbol{x}) \tag{82}$$

$$= \frac{1}{|\mathcal{T}(Q_Y)|}\sum_{\{\mathcal{T}(Q_{YY'|X}|\boldsymbol{x}):\ \boldsymbol{E}_Q d(Y,Y')\le D\}} |\mathcal{T}(Q_{YY'|X}|\boldsymbol{x})| \tag{83}$$

$$\cdot \exp_2\{-n[H_Q(Y'|X)+D(Q_{Y'|X}\|W|Q_X)]\} \tag{84}$$

$$\doteq \exp_2\Bigg\{-n\min_{\{Q_{YY'|X}:\ \boldsymbol{E}_Q d(Y,Y')\le D\}}[H_Q(Y)-H_Q(Y,Y'|X) \tag{85}$$

$$+ H_Q(Y'|X)+D(Q_{Y'|X}\|W|Q_X)]\Bigg\} \tag{86}$$

$$= \exp_2\left\{-n\min_{\{Q_{YY'|X}:\ \boldsymbol{E}_Q d(Y,Y')\le D\}}[H_Q(Y)-H_Q(Y|X,Y')+D(Q_{Y'|X}\|W|Q_X)]\right\} \tag{87}$$

$$= \exp_2\left\{-n\min_{\{Q_{YY'|X}:\ \boldsymbol{E}_Q d(Y,Y')\le D\}}[I_Q(Y;X,Y')+D(Q_{Y'|X}\|W|Q_X)]\right\} \tag{88}$$

$$\le \exp_2\Bigg\{-n\min_{\{Q_{Y'|Y}:\ \boldsymbol{E}_Q d(Y,Y')\le D\}}[I_Q(Y;Y') \tag{89}$$

$$+ \min_{\{Q_{Y'|X}:(Q_X\odot Q_{Y'|X})_{Y'}=Q_{Y'}\}} D(Q_{Y'|X}\|W|Q_X)]\Bigg\} \tag{90}$$

$$= \exp_2\left\{-n\min_{\{Q_{Y'|Y}:\ \boldsymbol{E}_Q d(Y,Y')\le D\}}[I_Q(Y;Y')+\Gamma(Q_X,Q_{Y'})]\right\} \tag{91}$$

$$\le \exp_2\left\{-n\min_{\{Q_{Y'|Y}:\ \boldsymbol{E}_Q d(Y,Y')\le D\}}[I_Q(Y;Y')+\Gamma(Q_{Y'})]\right\} \tag{92}$$

$$= \exp_2\{-n\hat{R}_W(D,Q_Y)\}. \tag{93}$$

□

The proof of the converse is now as follows.

Proof of Theorem 2 (Converse Part). Denote by $k(Q_Y)$ a positive integer, whose value, possibly depending on the type of $\boldsymbol{y}$, Q_Y, will be defined later. We have the following chain of inequalities, whose explanations are given below.

$$\boldsymbol{E}\{G^\rho\} = \sum_{Q_Y} P[\mathcal{T}(Q_Y)]\boldsymbol{E}\{G^\rho|\boldsymbol{y}\in\mathcal{T}(Q_Y)\} \tag{94}$$

$$\begin{aligned}
&\doteq \sum_{Q_Y} 2^{-nD(Q_Y||P)}\boldsymbol{E}\{G^\rho|\boldsymbol{y}\in\mathcal{T}(Q_Y)\} && (95)\\
&\geq \sum_{Q_Y} 2^{-nD(Q_Y||P)}k(Q_Y)^\rho\cdot\Pr\{G>k(Q_Y)|\boldsymbol{y}\in\mathcal{T}(Q_Y)\} && (96)\\
&\geq \sum_{Q_Y} 2^{-nD(Q_Y||P)}k(Q_Y)^\rho \sum_{\boldsymbol{y}\in\mathcal{T}(Q_Y)}\frac{1}{|\mathcal{T}(Q_Y)|}\Pr\{G>k(Q_Y)|\boldsymbol{y}\} && (97)\\
&\geq \sum_{Q_Y} 2^{-nD(Q_Y||P)}k(Q_Y)^\rho \sum_{\boldsymbol{y}\in\mathcal{T}(Q_Y)}\frac{1}{|\mathcal{T}(Q_Y)|}\left(1-\sum_{i=1}^{k(Q_Y)}W(\mathcal{S}(\boldsymbol{y})|\boldsymbol{x}_i)\right) && (98)\\
&= \sum_{Q_Y} 2^{-nD(Q_Y||P)}k(Q_Y)^\rho\left[1-\sum_{i=1}^{k(Q_Y)}\sum_{\boldsymbol{y}\in\mathcal{T}(Q_Y)}\frac{1}{|\mathcal{T}(Q_Y)|}W(\mathcal{S}(\boldsymbol{y})|\boldsymbol{x}_i)\right] && (99)\\
&\stackrel{\cdot}{\geq} \sum_{Q_Y} 2^{-nD(Q_Y||P)}k(Q_Y)^\rho\left[1-\sum_{i=1}^{k(Q_Y)}2^{-n\hat{R}_W(D,Q_Y)}\right] && (100)\\
&= \sum_{Q_Y} 2^{-nD(Q_Y||P)}k(Q_Y)^\rho\left[1-k(Q_Y)2^{-n\hat{R}_W(D,Q_Y)}\right] && (101)\\
&\doteq \sum_{Q_Y} 2^{-nD(Q_Y||P)}2^{\rho n\left(\hat{R}_W(D,Q_Y)-\epsilon\right)}\left[1-2^{-n\epsilon}\right] && (102)\\
&\doteq \exp_2\left\{n\max_{Q_Y}\left[\rho\hat{R}_W(D,Q_Y)-D(Q_Y||P)-\rho\epsilon\right]\right\}. && (103)
\end{aligned}$$

In the above chain, (96) is since $\boldsymbol{E}G^\rho = \sum_{k=1}^\infty k^\rho\cdot\Pr\{G=k\}\geq k^\rho\cdot\Pr\{G>k\}$. (97) is since $\boldsymbol{Y}$ is drawn from a DMS, hence the conditioning $\boldsymbol{Y}\in\mathcal{T}(Q_Y)$ means that $\boldsymbol{Y}$ is uniformly distributed within $\mathcal{T}(Q_Y)$, which explains the uniform averaging across $\mathcal{T}(Q_Y)$. (98) is since $\Pr\{G>k|\boldsymbol{y}\} = 1-\Pr\{\cup_{i=1}^k \text{ guess i is successful}|\boldsymbol{y}\}\geq 1-\sum_{i=1}^k\Pr\{\text{guess i is successful}|\boldsymbol{y}\}$, and $\Pr\{\text{guess i is successful}|\boldsymbol{y}\} = W(\mathcal{S}(\boldsymbol{y})|\boldsymbol{x}_i)$. (100) is due to Lemma 6. (102) is by choosing $k(Q_Y) = \lceil 2^{n\left(\hat{R}(D,Q_Y)-\epsilon\right)}\rceil$ for some arbitrarily small ϵ. Since the above is true for any $\epsilon>0$, we have

$$\boldsymbol{E}\{G^\rho\}\stackrel{\cdot}{\geq}\exp_2\left\{n\max_{Q_Y}\left[\rho\hat{R}_W(D,Q_Y)-D(Q_Y||P)\right]\right\}, \tag{104}$$

which completes the proof of the converse. □

7. Direct and Converse for Sources with Memory

7.1. Direct

We can now assert the following theorem, which gives an upper bound on the guesswork exponent, using a K-tuples parsing strategy.

Theorem 5. *For any stationary Ψ-mixing source P over $\mathcal{X}$, if guesses are selected independently at random according to the distribution (7), then, for $K\gg k$, the ρ-th guesswork moment, subject to a distortion constraint D, satisfies*

$$\limsup_{n\to\infty}\frac{1}{n}\log\boldsymbol{E}\{G^\rho\}\stackrel{\cdot}{\leq}\liminf_{K\to\infty}\sup_{Q^K}\frac{1}{K}\left[\rho R_K(D,Q^K)-D\left(Q^K||P^K\right)\right]. \tag{105}$$

Proof. We first bound the expected guesswork moment in terms of the guessing distribution $\tilde{P}$ via Lemma 1, and further bound $\tilde{P}[\mathcal{S}(\boldsymbol{x})]$ via Lemma 5. Specifically,

$$\begin{aligned}
\boldsymbol{E}\{G^\rho\} &= \boldsymbol{E}\{\boldsymbol{E}\{G^\rho|\boldsymbol{X}\}\} && (106)\\
&\doteq \boldsymbol{E}\left\{\tilde{P}[\mathcal{S}(\boldsymbol{X})]^{-\rho}\right\} && (107)\\
&= \sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x})\tilde{P}[\mathcal{S}(\boldsymbol{x})]^{-\rho} && (108)\\
&\leq 2^{\rho\nu_{K+k}(n)} \sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x})\frac{1}{(1-2\epsilon_m)^\rho}\exp_2\left\{\rho m\left[R_K(D',\hat{P}^K_{\boldsymbol{x}})+\epsilon\right]\right\} && (109)\\
&= \frac{2^{\rho\nu_{K+k}(n)}}{(1-2\epsilon_m)^\rho}\sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x})\exp_2\left\{\rho m\left[R_K(D',\hat{P}^K_{\boldsymbol{x}})+\epsilon\right]\right\}, && (110)
\end{aligned}$$

where (107) is due to Lemma 1 and (109) is due to Lemma 5.

While it is tempting to replace the sum over $\boldsymbol{x}\in\mathcal{X}^n$ 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 $\boldsymbol{x}\in\mathcal{X}^n$, whose blocks are denoted by $\underline{x}_i, 1\leq i\leq m$ (see (59)), define the *gaps oblivious* set as

$$G^K_{\boldsymbol{x}} = \{\tilde{\boldsymbol{x}}\in\mathcal{X}^n : \underline{\tilde{x}}_i = \underline{x}_i, 1\leq i\leq m\}. \tag{111}$$

That is, $G^K_{\boldsymbol{x}}$ includes all sequences which differ from $\boldsymbol{x}$ only in the gaps between the blocks. $G^K_{\boldsymbol{x}}$ satisfies the following bounds, proved in Appendix A.5.

Lemma 7. *For any stationary Ψ-mixing source Q and any $\boldsymbol{x}\in\mathcal{X}^n$,*

$$\begin{aligned}
Q\left(G^K_{\boldsymbol{x}}\right) &\leq (1+\epsilon_k)^{m-1}\exp_2\left\{-m\left(D\left(\hat{P}^K_{\boldsymbol{x}}||Q^K\right)+\hat{H}^K_{\boldsymbol{x}}(X^K)\right)\right\} && (112)\\
Q\left(G^K_{\boldsymbol{x}}\right) &\geq (1-\epsilon_k)^{m-1}\exp_2\left\{-m\left(D\left(\hat{P}^K_{\boldsymbol{x}}||Q^K\right)+\hat{H}^K_{\boldsymbol{x}}(X^K)\right)\right\}, && (113)
\end{aligned}$$

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

Note that if the K-th order marginal of Q satisfies $Q^K=\hat{P}^K_{\boldsymbol{x}}$, the divergences in Lemma 7 are zero and we have,[6]

$$\begin{aligned}
\hat{P}^K_{\boldsymbol{x}}\left(G^K_{\boldsymbol{x}}\right) &\leq (1+\epsilon_k)^{m-1}2^{-m\hat{H}^K_{\boldsymbol{x}}(X^K)} && (114)\\
\hat{P}^K_{\boldsymbol{x}}\left(G^K_{\boldsymbol{x}}\right) &\geq (1-\epsilon_k)^{m-1}2^{-m\hat{H}^K_{\boldsymbol{x}}(X^K)}. && (115)
\end{aligned}$$

Denote by $\left|T^K(P^K)\right|_G$ the number of *distinct* sets $G^K_{\boldsymbol{x}}$ in $T^K(P^K)$. The following corollary, proved in Appendix A.6, bounds the number of such sets within a K-type class $T^K(P^K)$.

[6] $\hat{P}^K_{\boldsymbol{x}}\left(G^K_{\boldsymbol{x}}\right)$ 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.

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

We are now able to further bound $\boldsymbol{E}\{G^\rho\}$. Continuing from (110),

$$
\begin{aligned}
\boldsymbol{E}\{G^\rho\} &\overset{\cdot}{\leq} \frac{2^{\rho\nu_{K+k}(n)}}{(1-2\epsilon_m)^\rho} \sum_{T^K \in \mathcal{T}^K} \sum_{G^K_{\boldsymbol{x}} \in T^K} P(G^K_{\boldsymbol{x}}) 2^{\rho m\left[R_K(D',\hat{P}^K_{\boldsymbol{x}})+\epsilon\right]} && (116)\\
&\leq \frac{2^{\rho\nu_{K+k}(n)}(1+\epsilon_k)^{m-1}}{(1-2\epsilon_m)^\rho} \sum_{T^K \in \mathcal{T}^K} \sum_{G^K_{\boldsymbol{x}} \in T^K} 2^{m\left(-D\left(\hat{P}^K_{\boldsymbol{x}}||P^K\right)-\hat{H}^K_{\boldsymbol{x}}(X^K)\right)} 2^{\rho m\left[R_K(D',\hat{P}^K_{\boldsymbol{x}})+\epsilon\right]} && (117)\\
&\leq \frac{2^{\rho\nu_{K+k}(n)}(1+\epsilon_k)^{m-1}}{(1-\epsilon_k)^{m-1}(1-2\epsilon_m)^\rho} && (118)\\
&\qquad \sum_{T^K \in \mathcal{T}^K} 2^{m\hat{H}^K_{\boldsymbol{x}}(X^K)} 2^{m\left(-D\left(\hat{P}^K_{\boldsymbol{x}}||P^K\right)-\hat{H}^K_{\boldsymbol{x}}(X^K)\right)} 2^{\rho m\left[R_K(D',\hat{P}^K_{\boldsymbol{x}})+\epsilon\right]} && (119)\\
&= \frac{2^{\rho\nu_{K+k}(n)}(1+\epsilon_k)^{m-1}}{(1-\epsilon_k)^{m-1}(1-2\epsilon_m)^\rho} \sum_{T^K \in \mathcal{T}^K} 2^{-mD\left(\hat{P}^K_{\boldsymbol{x}}||P^K\right)} 2^{\rho m\left[R_K(D',\hat{P}^K_{\boldsymbol{x}})+\epsilon\right]} && (120)\\
&\leq \frac{2^{\rho\nu_{K+k}(n)}(1+\epsilon_k)^{m-1}}{(1-\epsilon_k)^{m-1}(1-2\epsilon_m)^\rho} \left|\mathcal{T}^K\right| \max_{\hat{P}^K_{\boldsymbol{x}}} 2^{-mD\left(\hat{P}^K_{\boldsymbol{x}}||P^K\right)} 2^{\rho m\left[R_K(D',\hat{P}^K_{\boldsymbol{x}})+\epsilon\right]} && (121)\\
&\leq \frac{2^{\rho\nu_{K+k}(n)}(1+\epsilon_k)^m\,(m+1)^{|\mathcal{X}|^K}}{(1-\epsilon_k)^m(1-2\epsilon_m)^\rho} 2^{\max_{\hat{P}^K_{\boldsymbol{x}}} m\left[\rho R_K(D',\hat{P}^K_{\boldsymbol{x}})+\rho\epsilon-D\left(\hat{P}^K_{\boldsymbol{x}}||P^K\right)\right]} && (122)\\
&= \exp_2\left\{\max_{\hat{P}^K_{\boldsymbol{x}}} n\left[\rho\frac{1}{K+k}R_K(D',\hat{P}^K_{\boldsymbol{x}}) - \frac{1}{K+k}D\left(\hat{P}^K_{\boldsymbol{x}}||P^K\right)\right.\right. && (123)\\
&\qquad \left.\left. +\,\rho\left(\epsilon - \frac{\log(1-2\epsilon_m)-\nu_{K+k}(n)}{n}\right) + \frac{\log\left(\frac{1+\epsilon_k}{1-\epsilon_k}\right)}{K+k} + \frac{|\mathcal{X}|^K\log(m+1)}{n}\right]\right\}. && (124)
\end{aligned}
$$

In the above chain, (117) is due to Lemma 7, while (119) is due to (114); (122) follows from (62) and since $\frac{1-\epsilon_k}{1+\epsilon_k} \leq 1$.

Now, fix $K \gg k$. Since $D' = D - \delta_2 - \frac{\delta}{K}$, with $\delta_2 = \frac{k(D_{max}-D)}{K}$, and since the above is true for any $\delta, \epsilon > 0$, taking the limit over n, we have

$$
\limsup_{n\to\infty} \frac{1}{n}\log \boldsymbol{E}\{G^\rho\} \overset{\cdot}{\leq} \max_{\hat{P}^K_{\boldsymbol{x}}} \frac{1}{K}\left[\rho R_K(D,\hat{P}^K_{\boldsymbol{x}}) - D\left(\hat{P}^K_{\boldsymbol{x}}||P^K\right) + O(1)\right], \qquad (125)
$$

where we have also used the continuity of $R_K(D,\hat{P}^K_{\boldsymbol{x}})$ in D. Finally, the r.h.s of (125) is at least as large if the maximization is done over all distributions on $\mathcal{X}^K$, rather than only (fractional) types. Thus, enlarging the optimization domain, then taking a limit over K, we have

$$
\limsup_{n\to\infty} \frac{1}{n}\log \boldsymbol{E}\{G^\rho\} \overset{\cdot}{\leq} \liminf_{K\to\infty} \sup_{Q^K} \frac{1}{K}\left[\rho R_K(D,Q^K) - D\left(Q^K||P^K\right)\right], \qquad (126)
$$

which completes the proof. □

7.2. 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*, $\mathcal{G}_n$, which is an ordered list of all members of $\hat{\mathcal{X}}^n$. That is, $\mathcal{G}_n = \{\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \ldots, \hat{\boldsymbol{x}}_{|\hat{\mathcal{X}}|^n}\}$. $\mathcal{G}_n$ is associated with a *guessing function*, $G(\boldsymbol{x})$, which is the function that maps $\mathcal{X}^n$ to the integers, by assigning each $\boldsymbol{x} \in \mathcal{X}^n$ the smallest integer k for which $d(\boldsymbol{x}, \hat{\boldsymbol{x}}_k) \leq nD$, $\hat{\boldsymbol{x}}_k \in \mathcal{G}_n$, namely, $\hat{\boldsymbol{x}}_k$ is the first element of $\mathcal{G}_n$ which is in $\mathcal{S}(\boldsymbol{x})$. Thus, $G(\boldsymbol{x})$ is the number of guesses required until successfully meeting the distortion criteria, using $\mathcal{G}_n$, when $\boldsymbol{X} = \boldsymbol{x}$. Note that any randomized guessing strategy can be implemented with such a list. Note also that for finite alphabet, and assuming that for all $x \in \mathcal{X}$ we have $d_{\min} = \min_{\hat{x} \in \hat{\mathcal{X}}} d(x, \hat{x}) = 0$, one can easily append any list $\mathcal{G}_n$ such that the range of $G(\boldsymbol{x})$ is finite. Thus, the following converse bounds from below the best achievable exponent of the ρ-th guessing moment.

Theorem 6. *For any source P with an K-letter marginal P^K the guesswork ρ-th moment exponent at distortion level D is lower bounded by*

$$\liminf_{K\to\infty} \inf_{\mathcal{G}_K} \frac{1}{K} \log \boldsymbol{E}_{P^K}\{G^\rho(\boldsymbol{X})\} \geq \liminf_{K\to\infty} \sup_{Q^K} \frac{1}{K}\left[\rho R_K(D, Q^K) - D(Q^K \| P^K)\right], \tag{127}$$

where P^K is any distribution on $\mathcal{X}^K$.

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

$$p_i = \frac{c(\delta)}{i^{1+\delta}} \tag{128}$$

where $\delta > 0$ and

$$c(\delta) = \left[\sum_{i=1}^{\infty} \frac{1}{i^{1+\delta}}\right]^{-1}. \tag{129}$$

Define $\mathcal{I}_i^K = \{\boldsymbol{x} : G(\boldsymbol{x}) = i\}$. Using the Shannon code for the distribution in (128), for any distribution Q^K on $\mathcal{X}^K$, we have

$$\begin{aligned} \boldsymbol{E}_{Q^K} l(\boldsymbol{x}) &= \sum_i Q^K(\mathcal{I}_i^K) \lceil -\log p_i \rceil & (130)\\ &\leq 1 - \sum_i Q^K(\mathcal{I}_i^K) \log p_i & (131)\\ &= 1 - \sum_i Q^K(\mathcal{I}_i^K) \log c(\delta) + (1+\delta) \sum_i Q^K(\mathcal{I}_i^K) \log i & (132)\\ &= 1 - \log c(\delta) + (1+\delta) \sum_{\boldsymbol{x}} Q^K(\boldsymbol{x}) \log G(\boldsymbol{x}). & (133) \end{aligned}$$

However, since $\boldsymbol{E}_{Q^K} l(\boldsymbol{x})$ 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,

$$\sum_{\boldsymbol{x}} Q^K(\boldsymbol{x}) \log G(\boldsymbol{x}) \geq \frac{\boldsymbol{E}_{Q^K} l(\boldsymbol{x}) - 1 + \log c(\delta)}{1+\delta} \tag{134}$$

$$\geq \frac{R_K(D,Q^K) - 1 + \log c(\delta)}{1+\delta}. \tag{135}$$

Hence, for any source P with a marginal distribution P^K and any Q^K, we have

$$\begin{aligned}
\boldsymbol{E}_{P^K}\{G^\rho(\boldsymbol{X})\} &= \sum_{\boldsymbol{x}\in\mathcal{X}^K} P^K(\boldsymbol{x})G^\rho(\boldsymbol{x}) && (136)\\
&= \sum_{\boldsymbol{x}\in\mathcal{X}^K} Q^K(\boldsymbol{x}) \exp_2\left\{-\log\frac{Q^K(\boldsymbol{x})}{P^K(\boldsymbol{x})G^\rho(\boldsymbol{x})}\right\} && (137)\\
&\geq \exp_2\left\{-\sum_{\boldsymbol{x}\in\mathcal{X}^K} Q^K(\boldsymbol{x})\log\frac{Q^K(\boldsymbol{x})}{P^K(\boldsymbol{x})G^\rho(\boldsymbol{x})}\right\} && (138)\\
&= \exp_2\left\{-D(Q^K\|P^K) + \rho\sum_{\boldsymbol{x}\in\mathcal{X}^K} Q^K(\boldsymbol{x})\log G(\boldsymbol{x})\right\} && (139)\\
&\geq \exp_2\left\{-D(Q^K\|P^K) + \rho\frac{R_K(D,Q^K) - 1 + \log c(\delta)}{1+\delta}\right\}, && (140)
\end{aligned}$$

where (138) follows from Jensen's inequality. Since the above holds for any Q^K, we have, for any $\delta > 0$,

$$\inf_{\mathcal{G}_K}\frac{1}{K}\log\boldsymbol{E}_{P^K}\{G^\rho(\boldsymbol{X})\} \geq \sup_{Q^K}\frac{1}{K}\left[\frac{\rho}{1+\delta}R_K(D,Q^K) - D(Q^K\|P^K) + O\left(\rho\frac{\log c(\delta)-1}{1+\delta}\right)\right]. \tag{141}$$

At this point, a remark on δ is in order. Note that δ 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 $\delta\to\infty$, $\log c(\delta)$ can be approximated by $-\log\delta$. Yet, due to the independence in K, we have

$$\liminf_{K\to\infty}\inf_{\mathcal{G}_K}\frac{1}{K}\log\boldsymbol{E}_{P^K}\{G^\rho(\boldsymbol{X})\} \geq \liminf_{K\to\infty}\sup_{Q^K}\frac{1}{K}\left[\frac{\rho}{1+\delta}R_K(D,Q^K) - D(Q^K\|P^K)\right], \tag{142}$$

for any $\delta > 0$, which completes the proof. □

A. Appendix - Additional Proofs and Technical Claims

A.1. Proof of Lemma 2

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

$$\begin{aligned}
\Gamma(Q_Y) &= \inf_{Q_{X|Y}} D(Q_{Y|X}\|W|Q_X) && (143)\\
&= \inf_{Q_{X|Y}}\sum_{x,y} Q_{XY}(x,y)\log\frac{Q_{Y|X}(y|x)}{W(y|x)} && (144)\\
&= \inf_{Q_{X|Y}}\sum_{x,y} Q_{XY}(x,y)\log\frac{Q_{X|Y}(x|y)Q_Y(y)}{Q_X(x)W(y|x)} && (145)
\end{aligned}$$

$$= -H_Q(Y) + \inf_{Q_{X|Y}} \sum_{x,y} Q_{XY}(x,y) \log \frac{Q_{X|Y}(x|y)}{Q_X(x)W(y|x)} \tag{146}$$

$$= -H_Q(Y) + \inf_V \inf_{Q_{X|Y}} \sum_y Q_Y(y) \sum_x Q_{X|Y}(x|y) \log \frac{Q_{X|Y}(x|y)}{V(x)W(y|x)} \tag{147}$$

$$= -H_Q(Y) + \inf_V \left\{ -\sum_y Q_Y(y) \log \left[\sum_x V(x)W(y|x) \right] \right\}. \tag{148}$$

In the above chain, (147) is true since

$$\sum_{x,y} Q_{XY}(x,y) \log \frac{Q_{X|Y}(x|y)}{V(x)W(y|x)} - \sum_{x,y} Q_{XY}(x,y) \log \frac{Q_{X|Y}(x|y)}{Q_X(x)W(y|x)} \tag{149}$$

$$= \sum_{x,y} Q_{XY}(x,y) \log \frac{Q_X(x)}{V(x)} \geq 0 \tag{150}$$

with equality when $V(x) = Q_X(x)$ for all x. (148) follows by noting that

$$\sum_x Q_{X|Y}(x|y) \log \frac{Q_{X|Y}(x|y)}{V(x)W(y|x)} = \sum_x Q_{X|Y}(x|y) \log \frac{[\sum_{\hat{x}} V(\hat{x})W(y|\hat{x})]Q_{X|Y}(x|y)}{[\sum_{\hat{x}} V(\hat{x})W(y|\hat{x})]V(x)W(y|x)} \tag{151}$$

$$= D\left(Q_{X|Y}||\tilde{V}_{X|Y}\right) - \log \left[\sum_{\hat{x}} V(\hat{x})W(y|\hat{x}) \right], \tag{152}$$

where $\tilde{V}_{X|Y}(x|y) = \frac{1}{\sum_{\hat{x}} V(\hat{x})W(y|\hat{x})} V(x)W(y|x)$ and the expression is clearly minimized by choosing $Q_{X|Y} = \tilde{V}_{X|Y}$. Now,

$$\hat{R}_W(D, Q_Y) \tag{153}$$

$$= \inf_{\{Q_{Y'|Y}:\ \boldsymbol{E}_Q d(Y,Y') \leq D\}} [H_Q(Y') - H_Q(Y'|Y) + \Gamma(Q_{Y'})] \tag{154}$$

$$= \inf_{\{Q_{Y'|Y}:\ \boldsymbol{E}_Q d(Y,Y') \leq D\}} \inf_V \left\{ -H_Q(Y'|Y) - \sum_{y'} Q_{Y'}(y') \left[\log \sum_x V(x)W(y'|x) \right] \right\} \tag{155}$$

$$= \inf_{Q_{Y'|Y}} \sup_{s \geq 0} \inf_V \left[\sum_y Q_Y(y) \sum_{y'} Q_{Y'|Y}(y'|y) \log \frac{Q_{Y'|Y}(y'|y)}{\sum_x V(x)W(y'|x)} + \right. \tag{156}$$

$$\left. s\Big(\sum_y Q_Y(y) \sum_{y'} Q_{Y'|Y}(y'|y) d(y,y') - D \Big) \right] \tag{157}$$

$$= \inf_{Q_{Y'|Y}} \inf_V \sup_{s \geq 0} \sum_y Q_Y(y) \sum_{y'} Q_{Y'|Y}(y'|y) \left[\log \frac{Q_{Y'|Y}(y'|y)}{\sum_x V(x)W(y'|x)} + sd(y,y') - sD \right] \tag{158}$$

$$= \inf_V \inf_{Q_{Y'|Y}} \sup_{s \geq 0} \sum_y Q_Y(y) \sum_{y'} Q_{Y'|Y}(y'|y) \left[\log \frac{Q_{Y'|Y}(y'|y)}{e^{-sd(y,y')} \sum_x V(x)W(y'|x)} - sD \right] \tag{159}$$

$$= \inf_V \sup_{s \geq 0} \inf_{Q_{Y'|Y}} \sum_y Q_Y(y) \sum_{y'} Q_{Y'|Y}(y'|y) \left[\log \frac{Q_{Y'|Y}(y'|y)}{e^{-sd(y,y')} \sum_x V(x)W(y'|x)} - sD \right] \tag{160}$$

$$= \inf_{V} \sup_{s\geq 0} \left\{ -\sum_{y} Q_Y(y) \log \left[\sum_{y'} e^{-sd(y,y')} \sum_{x} V(x) W(y'|x) \right] - sD \right\} \tag{161}$$

$$= \inf_{V} \sup_{s\geq 0} \left\{ -\sum_{y} Q_Y(y) \log \left[\sum_{x} V(x) U_s(x,y) \right] - sD \right\} \tag{162}$$

$$= \sup_{s\geq 0} \inf_{V} \left\{ -\sum_{y} Q_Y(y) \log \left[\sum_{x} V(x) U_s(x,y) \right] - sD \right\}, \tag{163}$$

where (156) is due to the Lagrange multipliers analogous form for the optimization problem. (160) is since the utility function is convex in $Q_{Y'|Y}$ and concave (in fact, linear) in s. (161) uses a similar technique to (151)-(152), namely, the expression

$$\sum_{y'} Q_{Y'|Y}(y'|y) \log \frac{Q_{Y'|Y}(y'|y)}{e^{-sd(y,y')} \sum_x V(x) W(y'|x)} \tag{164}$$

can be viewed as a divergence minus a normalization term. In (162) we have defined

$$U_s(x,y) \stackrel{\triangle}{=} \sum_{y'} W(y'|x) e^{-sd(y,y')} \tag{165}$$

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

$$\begin{aligned}
E(\rho) &= \sup_{Q_Y} [\rho \hat{R}(D, Q_Y) - D(Q_Y \| P)] && (166)\\
&= \sup_{Q_Y} \sup_{s\geq 0} \inf_{V} \left\{ \sum_{y} Q_Y(y) \log \frac{P(y)}{Q_Y(y) \left[\sum_x V(x) U_s(x,y)\right]^{\rho}} - \rho s D \right\} && (167)\\
&= \sup_{s\geq 0} \sup_{Q_Y} \inf_{V} \left\{ \sum_{y} Q_Y(y) \log \frac{P(y)}{Q_Y(y) \left[\sum_x V(x) U_s(x,y)\right]^{\rho}} - \rho s D \right\} && (168)\\
&= \sup_{s\geq 0} \inf_{V} \sup_{Q_Y} \left\{ \sum_{y} Q_Y(y) \log \frac{P(y)}{Q_Y(y) \left[\sum_x V(x) U_s(x,y)\right]^{\rho}} - \rho s D \right\} && (169)\\
&= \sup_{s\geq 0} \inf_{V} \left\{ \log \left(\sum_{y} \frac{P(y)}{\left[\sum_x V(x) U_s(x,y)\right]^{\rho}} \right) - \rho s D \right\} && (170)\\
&= \sup_{s\geq 0} \inf_{M \in \mathcal{CH}(W)} \left\{ \log \left(\sum_{y} \frac{P(y)}{\left[\sum_{y'} M(y') e^{-sd(y,y')}\right]^{\rho}} \right) - \rho s D \right\} && (171)\\
&= \sup_{s\geq 0} \inf_{M \in \mathcal{CH}(W)} \left\{ \log \left(\sum_{y} \frac{P(y)}{\left[\sum_{y'} M(y') e^{s[D-d(y,y')]}\right]^{\rho}} \right) \right\}. && (172)
\end{aligned}$$

In the above chain, (169) is since the utility is convex in $V(\cdot)$ and concave in $Q_Y(\cdot)$ and (170), again, uses a similar technique to (151)-(152). For (171), remember that $U_s(x,y) \stackrel{\triangle}{=} \sum_{y'} W(y'|x) e^{-sd(y,y')}$ 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. (171) thus replaces the minimization over V by minimization over M, which is limited to $\mathcal{CH}(W)$, that is, the convex hull of $\{W(\cdot|x),\ x \in \mathcal{X}\}$.

A.2. Proof of Claim 1

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

$$\sum_{\hat{\boldsymbol{x}}\in\mathcal{S}(\boldsymbol{x})} P^n(\hat{\boldsymbol{x}}|\boldsymbol{x}) = P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\left(d(\boldsymbol{x},\hat{\boldsymbol{x}}) \le nD\right) \tag{173}$$

$$= P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\left(d(\boldsymbol{x},\hat{\boldsymbol{x}}) - nD^- \le n\delta\right) \tag{174}$$

$$= P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\left(\sum_{i=1}^n d(x_i,\hat{x}_i) - nD^- \le n\delta\right) \tag{175}$$

$$= P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\left(\sum_{x\in\mathcal{X}}\sum_{i:x_i=x} d(x,\hat{x}_i) - nD^- \le n\delta\right) \tag{176}$$

$$= P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\left(\sum_{x\in\mathcal{X}}\frac{1}{n}\sum_{i:x_i=x} d(x,\hat{x}_i) - D^- \le \delta\right). \tag{177}$$

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

$$D^-(x) \triangleq \sum_{\hat{x}\in\hat{\mathcal{X}}} P(\hat{x}|x)d(x,\hat{x}), \tag{178}$$

we have $\sum_{x\in\mathcal{X}} \hat{P}_{\boldsymbol{x}}(x)D^-(x) = D^-$. Thus, denoting by $N_{\boldsymbol{x}}(x)$ the number of occurrences of x in $\boldsymbol{x}$, and analogously $N_{\boldsymbol{x}\hat{\boldsymbol{x}}}(x,\hat{x})$ the number of occurrences of the pair $(x,\hat{x})$, we have

$$\sum_{\hat{\boldsymbol{x}}\in\mathcal{S}(\boldsymbol{x})} P^n(\hat{\boldsymbol{x}}|\boldsymbol{x}) = P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\left(\sum_{x\in\mathcal{X}}\left[\frac{N_{\boldsymbol{x}}(x)}{n}\sum_{\hat{x}\in\hat{\mathcal{X}}}\frac{N_{\boldsymbol{x}\hat{\boldsymbol{x}}}(x,\hat{x})}{N_{\boldsymbol{x}}(x)}d(x,\hat{x}_i) - \hat{P}_{\boldsymbol{x}}(x)D^-(x)\right] \le \delta\right) \tag{179}$$

$$= P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\left(\sum_{x\in\mathcal{X}}\hat{P}_{\boldsymbol{x}}(x)\left[\sum_{\hat{x}\in\hat{\mathcal{X}}}\frac{N_{\boldsymbol{x}\hat{\boldsymbol{x}}}(x,\hat{x})}{N_{\boldsymbol{x}}(x)}d(x,\hat{x}_i) - D^-(x)\right] \le \delta\right) \tag{180}$$

$$\ge P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\left(\forall_x\left[\sum_{\hat{x}\in\hat{\mathcal{X}}}\frac{N_{\boldsymbol{x}\hat{\boldsymbol{x}}}(x,\hat{x})}{N_{\boldsymbol{x}}(x)}d(x,\hat{x}_i) - D^-(x)\right] \le \delta\right) \tag{181}$$

$$\ge P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\left(\forall_x\left|\frac{N_{\boldsymbol{x}\hat{\boldsymbol{x}}}(x,\hat{x})}{N_{\boldsymbol{x}}(x)} - P(\hat{x}|x)\right| \le \frac{\delta}{|\hat{\mathcal{X}}|D_{max}}\right) \tag{182}$$

$$\ge 1 - \epsilon'_n \tag{183}$$

for some ϵ'_n such that $\epsilon'_n \to 0$ as $n \to \infty$, where the last inequality is guaranteed by the law of large numbers, as each entry in $\hat{\boldsymbol{x}}$ was drawn independently according to $P(\hat{x}|x)$. Note that since the law of large numbers is applied on independent drawings from $P(\hat{x}|x)$, $\epsilon'_n \to 0$ uniformly in $\boldsymbol{x}$.

A.3. Proof of Claim 2

Following the same notation as in the proof of Claim 1, we have

$$\sum_{\hat{\boldsymbol{x}}\in\mathcal{T}^c(\boldsymbol{x})} P^n(\hat{\boldsymbol{x}}|\boldsymbol{x}) = P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\left(\log\frac{P^n(\hat{\boldsymbol{x}}|\boldsymbol{x})}{Q^n(\hat{\boldsymbol{x}})} > n\left[R(D^-,\hat{P}_{\boldsymbol{x}}) + \epsilon\right]\right) \tag{184}$$

$$
\begin{aligned}
&= P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{P(\hat{x}_i|x_i)}{Q(\hat{x}_i)} > R(D^-, \hat{P}_{\boldsymbol{x}}) + \epsilon\right) && (185)\\
&= P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{P(\hat{x}_i|x_i)}{Q(\hat{x}_i)} > \sum_{x\in\mathcal{X}}\hat{P}_{\boldsymbol{x}}(x)\sum_{\hat{x}\in\hat{\mathcal{X}}}P(\hat{x}|x)\log\frac{P(\hat{x}|x)}{Q(\hat{x})} + \epsilon\right) && (186)\\
&= P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\Bigg(\sum_{x\in\mathcal{X}}\frac{N_{\boldsymbol{x}}(x)}{n}\sum_{\hat{x}\in\hat{\mathcal{X}}}\frac{N_{\boldsymbol{x}\hat{\boldsymbol{x}}}(x,\hat{x})}{N_{\boldsymbol{x}}(x)}\log\frac{P(\hat{x}|x)}{Q(\hat{x})} && (187)\\
&\qquad\qquad > \sum_{x\in\mathcal{X}}\hat{P}_{\boldsymbol{x}}(x)\sum_{\hat{x}\in\hat{\mathcal{X}}}P(\hat{x}|x)\log\frac{P(\hat{x}|x)}{Q(\hat{x})} + \epsilon\Bigg) && (188)\\
&= P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\left(\sum_{x\in\mathcal{X}}\hat{P}_{\boldsymbol{x}}(x)\sum_{\hat{x}\in\hat{\mathcal{X}}}\left[\frac{N_{\boldsymbol{x}\hat{\boldsymbol{x}}}(x,\hat{x})}{N_{\boldsymbol{x}}(x)} - P(\hat{x}|x)\right]\log\frac{P(\hat{x}|x)}{Q(\hat{x})} > \epsilon\right) && (189)\\
&\le P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\left(\sum_{x\in\mathcal{X}}\hat{P}_{\boldsymbol{x}}(x)\sum_{\hat{x}\in\hat{\mathcal{X}}}\left[\frac{N_{\boldsymbol{x}\hat{\boldsymbol{x}}}(x,\hat{x})}{N_{\boldsymbol{x}}(x)} - P(\hat{x}|x)\right] > \frac{\epsilon}{\log\frac{1}{\min_{\hat{x}}Q(\hat{x})}}\right) && (190)\\
&\le P^n_{\hat{\boldsymbol{x}}|\boldsymbol{x}}\left(\max_{x,\hat{x}}\left[\frac{N_{\boldsymbol{x}\hat{\boldsymbol{x}}}(x,\hat{x})}{N_{\boldsymbol{x}}(x)} - P(\hat{x}|x)\right] > \frac{\epsilon}{|\hat{\mathcal{X}}|\log\frac{1}{\min_{\hat{x}}Q(\hat{x})}}\right) && (191)\\
&\le \epsilon''_n, && (192)
\end{aligned}
$$

where, again, the last inequality is due to the law of large numbers, $\frac{N_{\boldsymbol{x}\hat{\boldsymbol{x}}}(x,\hat{x})}{N_{\boldsymbol{x}}(x)} \to P(\hat{x}|x)$ 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 $\boldsymbol{x}$.

A.4. Proof of Claim 3

The proof is based on viewing any block-memoryless probability assignment for $\hat{\boldsymbol{x}}$, followed by a Shannon code, as a finite-state encoder for $\hat{\boldsymbol{x}}$. 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 $\hat{\mathcal{X}}^{K+k}$, and assume the sequence $\hat{\boldsymbol{x}}$ is divided into blocks of length $K+k$ and encoded using a Shannon code for $M^{K+k}(\cdot)$. The resulting compression ratio satisfies

$$
\begin{aligned}
\rho_{K+k} &= \frac{1}{n\log|\hat{\mathcal{X}}|}\sum_{i=1}^{m}\left\lceil -\log M^{K+k}\left(\hat{x}_{(i-1)(K+k)+1}^{i(K+k)}\right)\right\rceil && (193)\\
&\le \frac{1}{n\log|\hat{\mathcal{X}}|}\left[-\sum_{i=1}^{m}\log M^{K+k}\left(\hat{x}_{(i-1)(K+k)+1}^{i(K+k)}\right) + m\right] && (194)\\
&= \frac{1}{n\log|\hat{\mathcal{X}}|}\left[-\log\prod_{i=1}^{m}M^{K+k}\left(\hat{x}_{(i-1)(K+k)+1}^{i(K+k)}\right) + m\right]. && (195)
\end{aligned}
$$

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

$$\rho_{K+k} \geq \frac{c(\hat{\boldsymbol{x}})+s^2}{n\log|\hat{\mathcal{X}}|}\log\left(\frac{c(\hat{\boldsymbol{x}})+s^2}{4s^2}\right)+\frac{2s^2}{n\log|\hat{\mathcal{X}}|}. \tag{196}$$

Consequently,

$$\begin{aligned}
-\log\prod_{i=1}^{m} M^{K+k}\left(\hat{x}_{(i-1)(K+k)+1}^{i(K+k)}\right)+m &\geq (c(\hat{\boldsymbol{x}})+s^2)\log\left(\frac{c(\hat{\boldsymbol{x}})+s^2}{4s^2}\right)+2s^2 &(197)\\
&\geq (c(\hat{\boldsymbol{x}})+1+s^2-1)\log\left(\frac{2|\hat{\mathcal{X}}|}{2|\hat{\mathcal{X}}|}\frac{c(\hat{\boldsymbol{x}})+1}{4s^2}\right)+2s^2 &(198)\\
&= (c(\hat{\boldsymbol{x}})+1)\log\left(2|\hat{\mathcal{X}}|(c(\hat{\boldsymbol{x}})+1)\right) &(199)\\
&\quad+(s^2-1)\log\left(2|\hat{\mathcal{X}}|(c(\hat{\boldsymbol{x}})+1)\right) &(200)\\
&\quad+2s^2-(c(\hat{\boldsymbol{x}})+s^2)\log\left(8|\hat{\mathcal{X}}|s^2\right). &(201)
\end{aligned}$$

Finally, by [34, Theorem 2], $LZ(\hat{\boldsymbol{x}}) \leq (c(\hat{\boldsymbol{x}})+1)\log\left(2|\hat{\mathcal{X}}|(c(\hat{\boldsymbol{x}})+1)\right)$, thus

$$LZ(\hat{\boldsymbol{x}}) \leq -\log\prod_{i=1}^{m} M^{K+k}\left(\hat{x}_{(i-1)(K+k)+1}^{i(K+k)}\right)+m+(c(\hat{\boldsymbol{x}})+s^2)\log\left(8|\hat{\mathcal{X}}|s^2\right). \tag{202}$$

Setting $\nu_{K+k}(n) = m+(c(\hat{\boldsymbol{x}})+s^2)\log\left(8|\hat{\mathcal{X}}|s^2\right)$, $m = \frac{n}{K+k}$ and since $c(\hat{\boldsymbol{x}}) \leq \frac{n\log|\hat{\mathcal{X}}|}{(1-\epsilon_n)\log n}$ ([62, Theorem 2]), we obtain the required result.

A.5. Proof of Lemma 7

Since Q is Ψ-mixing, for any two block $\underline{x}_i, \underline{x}_{i+1}$ which are k symbols apart, we have

$$\left|\frac{Q(\underline{x}_i,\underline{x}_{i+1})}{Q(\underline{x}_i)Q(\underline{x}_{i+1})}-1\right| \leq \epsilon_k \tag{203}$$

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

$$\begin{aligned}
Q\left(G_{\boldsymbol{x}}^K\right) &= \sum_{\tilde{\boldsymbol{x}}\in\mathcal{X}^n:\underline{\tilde{x}}_i=\underline{x}_i,1\leq i\leq m} Q(\tilde{\boldsymbol{x}}) &(204)\\
&= Q(\underline{x}_1,\ldots,\underline{x}_m) &(205)\\
&\leq (1+\epsilon_k)^{m-1}\prod_{i=1}^{m} Q^K(\underline{x}_i) &(206)\\
&= (1+\epsilon_k)^{m-1}2^{-m\left(D\left(\hat{P}_{\boldsymbol{x}}^K||Q^K\right)+\hat{H}_{\boldsymbol{x}}^K(X^K)\right)}. &(207)
\end{aligned}$$

The opposite inequality follows in a similar manner.

A.6. Proof of Corollary 2

Let P be any stationary Ψ-mixing distribution with a K-th order marginal P^K. Then,

$$1 \geq P\left(T^K(P^K)\right) \tag{208}$$

$$= \sum_{G_{\boldsymbol{x}}^K \in T^K(P^K)} P^K\left(G_{\boldsymbol{x}}^K\right) \tag{209}$$

$$\geq \left|T^K(P^K)\right|_G (1-\epsilon_k)^{m-1} 2^{-m\hat{H}_{\boldsymbol{x}}^K(X^K)}. \tag{210}$$

A.7. Converse and Direct for Individual Sequences and Finite–State Guessing

A.7.1. 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 $\boldsymbol{x}$, we first generalize the success probability of the finite–state machine, $P(\boldsymbol{x})$ in the notation of [48] when one has to guess the exact sequence $\boldsymbol{x}$, 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 ρ–th moment satisfies the following lower bound

$$\boldsymbol{E}\{G_F^\rho|\boldsymbol{x}\} \geq \frac{2^{-\rho}}{e^2} \exp_2\left\{-\rho \log P(\boldsymbol{x})\right\}, \tag{211}$$

where $P(\boldsymbol{x})$ is the probability that the machine output is $\boldsymbol{x}$. Note that similar to Lemma 1, the bound above is based on bounding the ρ–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

$$\boldsymbol{E}\{G_F^\rho|\boldsymbol{x}\} \geq \frac{2^{-\rho}}{e^2} \exp_2\left\{-\rho \log P[\mathcal{S}(\boldsymbol{x})]\right\}, \tag{212}$$

where $\mathcal{S}(\boldsymbol{x})$, the success set, was defined below (2). We thus have

$$\boldsymbol{E}\{G_F^\rho|\boldsymbol{x}\} \geq \frac{2^{-\rho}}{e^2} \exp_2\left\{-\rho \log \sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}}) \leq nD\}} P(\hat{\boldsymbol{x}})\right\}. \tag{213}$$

$P(\hat{\boldsymbol{x}})$, the probability that the machine outputs $\hat{\boldsymbol{x}}$, can be viewed as a probability assignment on the sequences $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}^n$. As such, $-\log P(\hat{\boldsymbol{x}})$ cannot be too small compared to the code length of an LZ encoder. This is similar to the argument behind Claim 3 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

$$-\log P(\hat{\boldsymbol{x}}) \geq c(\hat{\boldsymbol{x}}) \log c(\hat{\boldsymbol{x}}) - \frac{n \log[4K^2(l)] \log|\hat{\mathcal{X}}|}{(1-\epsilon_n)\log n} - K^2(l)\log[4K^2(l)] - m\log(2s^3 e), \tag{214}$$

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 $-\log P(\hat{\boldsymbol{x}}) \stackrel{\cdot}{\geq} LZ(\hat{\boldsymbol{x}})$. Specifically, with $l = \log(\log n)$ and since $K(l)$ is bounded by $l\alpha^l$, we have

$$\frac{n \log[4K^2(l)] \log |\hat{\mathcal{X}}|}{(1-\epsilon_n)\log n} + K^2(l)\log[4K^2(l)] + m\log(2s^3e) = o(n) \tag{215}$$

and hence

$$\begin{aligned} \boldsymbol{E}\{G_F^\rho|\boldsymbol{x}\} &\stackrel{\cdot}{\geq} \exp_2\left\{-\rho \log \sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\leq nD\}} 2^{-LZ(\hat{\boldsymbol{x}})}\right\} &(216)\\ &= \left[\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\leq nD\}} 2^{-LZ(\hat{\boldsymbol{x}})}\right]^{-\rho}, &(217)\end{aligned}$$

which proves part (i) of the converse result.

For part (ii), instead of (214), we lower bound $-\log P(\hat{\boldsymbol{x}})$ using the finite–state compressability of $\hat{\boldsymbol{x}}$, that is, by [48, eq. (31)],

$$-\log P(\hat{\boldsymbol{x}}) \geq n\rho_{K(l)}(\hat{\boldsymbol{x}}) - m\log(2s^3e). \tag{218}$$

Continuing from (213), we have

$$\begin{aligned} \boldsymbol{E}\{G_F^\rho|\boldsymbol{x}\} &\geq \frac{2^{-\rho}}{e^2}\exp_2\left\{-\rho\log \sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\leq nD\}} \exp_2\left\{-n\rho_{K(l)}(\hat{\boldsymbol{x}}) + m\log\left(2s^3e\right)\right\}\right\} &(219)\\ &= \frac{2^{-\rho}}{e^2}\left[\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\leq nD\}} \exp_2\left\{-n\rho_{K(l)}(\hat{\boldsymbol{x}}) + m\log\left(2s^3e\right)\right\}\right]^{-\rho}, &(220)\end{aligned}$$

which is true for any integers m and l such that $ml = n$. This completes the proof of the converse.

A.7.2. Direct

For the direct, we first note that the universal distribution given in (7), that is $\tilde{P}(\hat{\boldsymbol{x}}) \doteq 2^{-LZ(\hat{\boldsymbol{x}})}$, results in

$$\tilde{P}[\mathcal{S}(\boldsymbol{x})] \doteq \sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\leq nD\}} 2^{-LZ(\hat{\boldsymbol{x}})}. \tag{221}$$

Thus, invoking Lemma 1 yields

$$\boldsymbol{E}\{G_{LZ}^\rho|\boldsymbol{x}\} \doteq \left[\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\leq nD\}} 2^{-LZ(\hat{\boldsymbol{x}})}\right]^{-\rho} \tag{222}$$

for the specific universal distribution in (7). Clearly, this meets the converse result. However, the distribution in (7) cannot be implemented with a finite–state machine whose number of states is independent of n. To asymptotically achieve (222) 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 (30). This distribution can be implemented with a finite–state machine with $s(l)$ states, and for such a finite–state guessing machine we have

$$\boldsymbol{E}\{G_{LZ}^{\rho}|\boldsymbol{x}\} \doteq \left[\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} \prod_{i=0}^{n/l-1}\left[\frac{2^{-LZ(\hat{x}_{il+1}^{il+l})}}{\sum_{\hat{x}^l\in\hat{\mathcal{X}}^l} 2^{-LZ(\hat{x}^l)}}\right]\right]^{-\rho} \tag{223}$$

$$\doteq \left[\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} \prod_{i=0}^{n/l-1} 2^{-LZ(\hat{x}_{il+1}^{il+l})}\right]^{-\rho} \tag{224}$$

$$= \left[\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} \exp_2\left\{-\sum_{i=0}^{n/l-1} LZ(\hat{x}_{il+1}^{il+l})\right\}\right]^{-\rho} \tag{225}$$

$$\le \left[\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} \exp_2\left\{-\sum_{i=0}^{n/l-1} c(\hat{x}_{il+1}^{il+l})\log c(\hat{x}_{il+1}^{il+l}) - n\epsilon(l)\right\}\right]^{-\rho} \tag{226}$$

$$\le \left[\sum_{\{\hat{\boldsymbol{x}}:\ d(\boldsymbol{x},\hat{\boldsymbol{x}})\le nD\}} \exp_2\left\{-n\left[\rho_K(\hat{\boldsymbol{x}}) + \frac{\log(4K^2)}{(1-\epsilon(l))\log l} + \frac{K^2\log(4K^2)}{l} + \epsilon(l)\right]\right\}\right]^{-\rho}, \tag{227}$$

where $\epsilon(l) \to 0$ as $l \to \infty$, (226) is by [48, eq. (33)] and (227) is by [48, eq. (38)].

References

[1] J. M. Wozencraft, "Sequential decoding for reliable communication," Ph.D. dissertation, Research Laboratory of Electronics, Massachusetts Institute of Technology, May 1957.

[2] E. Arikan, "An inequality on guessing and its application to sequential decoding," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 99–105, January 1996.

[3] C. E. Pfister and W. G. Sullivan, "Rényi entropy, guesswork moments, and large deviations," *IEEE Transactions on Information Theory*, vol. 50, no. 11, pp. 2794–2800, November 2004.

[4] E. Arikan and N. Merhav, "Joint source-channel coding and guessing with application to sequential decoding," *IEEE Transactions on Information Theory*, vol. 44, no. 5, pp. 1756–1769, 1998.

[5] W. Szpankowski and S. Verdú, "Minimum expected length of fixed-to-variable lossless compression without prefix constraints," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4017–4025, July 2011.

[6] I. Kontoyiannis and S. Verdú, "Optimal lossless data compression: Non-asymptotics and asymptotics," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 777–795, February 2014.

[7] T. A. Courtade and S. Verdú, "Cumulant generating function of codeword lengths in optimal lossless compression," in *2014 IEEE International Symposium on Information Theory*, Honolulu, HI, USA, June 2014, pp. 2494–2498.

[8] O. Kosut and L. Sankar, "Asymptotics and non-asymptotics for universal fixed-to-variable source coding," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3757–3772, June 2017.

[9] M. A. Kumar, A. Sunny, A. Thakre, A. Kumar, and G. D. Manohar, "Are guessing, source coding, and tasks partitioning birds of a feather?" *arXiv preprint arXiv:2012.13707*, 2020.

[10] C. Bunte and A. Lapidoth, "Encoding tasks and Rényi entropy," *IEEE Transactions on Information Theory*, vol. 60, no. 9, pp. 5065–5076, 2014.

[11] E. Arikan and N. Merhav, "Guessing subject to distortion," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1041–1056, May 1998.

[12] K. R. Duffy, J. Li, and M. Médard, "Capacity-achieving guessing random additive noise decoding," *IEEE Transactions on Information Theory*, vol. 65, no. 7, pp. 4023–4040, 2019.

[13] M. Bishop and D. V. Klein, "Improving system security via proactive password checking," *Computers & Security*, vol. 14, no. 3, pp. 233–249, 1995.

[14] M. D. Amico, P. Michiardi, and Y. Roudier, "Password strength: An empirical analysis," in *Proceedings of IEEE Infocom*, San Diego, CA, USA, March 2010, pp. 1–9.

[15] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, and J. Lopez, "Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms," in *2012 IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, May 2012, pp. 523–537.

[16] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman, "Of passwords and people: Measuring the effect of password-composition policies," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* New York, NY, USA: ACM, 2011, pp. 2595–2604.

[17] N. Merhav and A. Cohen, "Universal randomized guessing with application to asynchronous decentralized brute–force attacks," *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 114–129, 2020.

[18] D. Malone and K. Maher, "Investigating the distribution of password choices," in *Proceedings of the 21st international conference on World Wide Web.* Lyon, France: ACM, 2012, pp. 301–310.

[19] J. Yan, A. Blackwell, R. Anderson, and A. Grant, "Password memorability and security: empirical results," *IEEE Security Privacy*, vol. 2, no. 5, pp. 25–31, September 2004.

[20] D. Vishwakarma and C. E. V. Madhavan, "Efficient dictionary for salted password analysis," in *2014 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, Bangalore, India, January 2014, pp. 1–6.

[21] Z. Rui and Z. Yan, "A survey on biometric authentication: Toward secure and privacy-preserving identification," *IEEE Access*, vol. 7, pp. 5994–6009, 2019.

[22] A. Page, A. Kulkarni, and T. Mohsenin, "Utilizing deep neural nets for an embedded ecg-based biometric authentication system," in *2015 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Atlanta, GA, USA, 2015, pp. 1–4.

[23] W. contributors, "Qrs complex," online; accessed 29-September-2021. [Online]. Available: https://en.wikipedia.org/wiki/QRS_complex

[24] M. Hammad, Y. Liu, and K. Wang, "Multimodal biometric authentication systems using convolution neural network based on different level fusion of ecg and fingerprint," *IEEE Access*, vol. 7, pp. 26 527–26 542, 2019.

[25] S. Aziz, M. U. Khan, Z. A. Choudhry, A. Aymin, and A. Usman, "ECG-based biometric authentication using empirical mode decomposition and support vector machines," in *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. Vancouver, BC, Canada: IEEE, 2019, pp. 0906–0912.

[26] A. Boles and P. Rad, "Voice biometrics: Deep learning-based voiceprint authentication system," in *2017 12th System of Systems Engineering Conference (SoSE)*, Waikoloa, HI, USA, 2017, pp. 1–6.

[27] Z. Meng, M. U. B. Altaf, and B.-H. F. Juang, "Active voice authentication," *Digital Signal Processing*, vol. 101, p. 102672, 2020.

[28] A. Mahfouz, T. M. Mahmoud, and A. S. Eldin, "A survey on behavioral biometric authentication on smartphones," *Journal of information security and applications*, vol. 37, pp. 28–37, 2017.

[29] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," *arXiv preprint arXiv:1412.5068*, 2014.

[30] N. Carlini, G. Katz, C. Barrett, and D. L. Dill, "Provably minimally-distorted adversarial examples," *arXiv preprint arXiv:1709.10207*, 2017.

[31] J. Bi and T. Zhang, "Support vector classification with input data uncertainty," *Advances in neural information processing systems*, vol. 17, no. 1, pp. 161–168, 2005.

[32] Y.-L. Tsai, C.-Y. Hsu, C.-M. Yu, and P.-Y. Chen, "Non-singular adversarial robustness of neural networks," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, pp. 3840–3844.

[33] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *arXiv preprint arXiv:1704.01155*, 2017.

[34] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Transactions on Information Theory*, vol. 24, no. 5, pp. 530–536, September 1978.

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

[36] S. Kuzuoka, "Asynchronous guessing subject to distortion," *arXiv preprint arXiv:2012.04901*, 2020.

[37] ——, "Asynchronous guessing subject to distortion," in *Proceedings 2021 IEEE International Symposium on Information Theory*, Melbourne, Australia, July 2021, pp. 2008–2012.

[38] S. Salamatian, W. Huleihel, A. Beirami, A. Cohen, and M. Médard, "Why botnets work: Distributed brute-force attacks need no synchronization," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 9, pp. 2288–2299, 2019.

[39] M. J. Weinberger, J. Ziv, and A. Lempel, "On the optimal asymptotic performance of universal ordering and of discrimination of individual sequences," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 380–385, March 1992.

[40] A. Beirami, R. Calderbank, M. Christiansen, K. Duffy, A. Makhdoumi, and M. Médard, "A geometric perspective on guesswork," in *2005 53rd Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, September 2015, pp. 941–948.

[41] J. Owens and J. Matthews, "A study of passwords and methods used in brute-force SSH attacks," in *USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, San Francisco, CA, USA, 2008.

[42] E. Tirado, B. Turpin, C. Beltz, P. Roshon, R. Judge, and K. Gagneja, "A new distributed brute-force password cracking technique," in *International Conference on Future Network Systems and Security*, Paris, France, 2018, pp. 117–127.

[43] S. Salamatian, W. Huleihel, A. Beirami, A. Cohen, and M. Médard, "Centralized vs decentralized targeted brute-force attacks: Guessing with side-information," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3749–3759, 2020.

[44] R. Sundaresan, "Guessing under source uncertainty," *IEEE Transactions on Information Theory*, vol. 53, no. 1, pp. 269–287, January 2007.

[45] D. Malone and W. G. Sullivan, "Guesswork and entropy," *IEEE Transactions on Information Theory*, vol. 50, no. 3, pp. 525–526, March 2004.

[46] M. K. Hanawal and R. Sundaresan, "Guessing revisited: A large deviations approach," *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 70–78, January 2011.

[47] M. M. Christiansen and K. R. Duffy, "Guesswork, large deviations, and shannon entropy," *IEEE Transactions Information Theory*, vol. 59, no. 2, pp. 796–802, February 2013.

[48] N. Merhav, "Guessing individual sequences: generating randomized guesses using finite–state machines," *IEEE Transactions on Information Theory*, vol. 66, no. 5, pp. 2912–2920, 2020.

[49] J. Bonneau, "The science of guessing: Analyzing an anonymized corpus of 70 million passwords," in *2012 IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, May 2012, pp. 538–552.

[50] M. M. Christiansen, K. R. Duffy, F. du Pin Calmon, and M. Médard, “Guessing a password over a wireless channel (on the effect of noise non-uniformity),” in *2013 Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, November 2013, pp. 51–55.

[51] N. Merhav, “Noisy guesses,” *IEEE Transactions on Information Theory*, vol. 66, no. 8, pp. 4796–4803, 2020.

[52] M. M. Christiansen, K. R. Duffy, F. du Pin Calmon, and M. Médard, “Multi-user guesswork and brute force security,” *IEEE Transactions on Information Theory*, vol. 61, no. 12, pp. 6876–6886, December 2015.

[53] A. Beirami, R. Calderbank, K. Duffy, and M. Médard, “Quantifying computational security subject to source constraints, guesswork and inscrutability,” in *2015 IEEE International Symposium on Information Theory (ISIT)*, Hong Kong, China, June 2015, pp. 2757–2761.

[54] T. A. Courtade and S. Verdú, “Variable-length lossy compression and channel coding: Non-asymptotic converses via cumulant generating functions,” in *2014 IEEE International Symposium on Information Theory*, Honolulu, HI, USA, June 2014, pp. 2499–2503.

[55] I. Sason, “Tight bounds on the Rényi entropy via majorization with applications to guessing and compression,” *Entropy*, vol. 20, no. 12, p. 896, 2018.

[56] R. Sundaresan, “Guessing under source uncertainty with side information,” in *2006 IEEE International Symposium on Information Theory*, Seattle, WA, USA, July 2006, pp. 2438–2440.

[57] I. Sason and S. Verdú, “Improved bounds on lossless source coding and guessing moments via Rényi measures,” *IEEE Transactions on Information Theory*, vol. 64, no. 6, pp. 4323–4346, June 2018.

[58] Y. Yona and S. Diggavi, “The effect of bias on the guesswork of hash functions,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, Aachen, Germany, June 2017, pp. 2248–2252.

[59] N. Ardimanov, O. Shayevitz, and I. Tamo, “Minimum guesswork with an unreliable oracle,” *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7528–7538, 2020.

[60] N. Weinberger and O. Shayevitz, “Guessing with a bit of help,” *Entropy*, vol. 22, no. 1, p. 39, 2020.

[61] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: John Wiley & Sons, 2006.

[62] A. Lempel and J. Ziv, “On the complexity of finite sequences,” *IEEE Transactions on information theory*, vol. 22, no. 1, pp. 75–81, 1976.

[63] M. Feder, “Gambling using a finite state machine,” *IEEE Transactions on Information Theory*, vol. 37, no. 5, pp. 1459–1465, 1991.

[64] M. Feder, N. Merhav, and M. Gutman, “Universal prediction of individual sequences,” *IEEE Transactions on Information Theory*, vol. 38, no. 4, pp. 1258–1270, July 1992.

[65] T. Weissman, N. Merhav, and A. Somekh-Baruch, “Twofold universal prediction schemes for achieving the finite-state predictability of a noisy individual binary sequence,” *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1849–1866, 2001.

[66] D. Modha and D. de Farias, “Finite-state rate-distortion for individual sequences,” in *Proceedings of the IEEE International Symposium on Information Theory*, Chicago, IL, USA, 2004, pp. 562–.

[67] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. Weinberger, “Universal discrete denoising: known channel,” *IEEE Transactions on Information Theory*, vol. 51, no. 1, pp. 5–28, 2005.

[68] N. Merhav, “Perfectly secure encryption of individual sequences,” *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1302–1310, 2013.

[69] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, ser. Prentice-Hall electrical engineering series. Prentice-Hall, 1971.