Universal Detection of Messages via Finite–State Channels *

Neri Merhav†

April 16, 1999

Abstract

We propose a universal, asymptotically optimum decision rule for deciding whether or not an observed sequence generated at the output of an unknown finite–state channel corresponds to a given channel–input message. The hypothesized message is tested against a certain alternative message or several alternative messages under the Neyman–Pearson criterion.

Index Terms: universal decoding, finite–state channels, Lempel–Ziv algorithm, Neyman–Pearson, watermarking.

*This research was supported by the Fund for the Promotion of Research at the Technion.

†N. Merhav is with the Electrical Engineering Department, Technion – I.I.T. Haifa 32000, Israel. E-mail: merhav@ee.technion.ac.il

I. Introduction

Consider the following detection problem: We are given two possible channel input sequences $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ of length n, an unknown channel from a class $\mathcal{W}_M$ of finite-state channels with a known number of states M, and a channel output sequence $\boldsymbol{y}$ of length n as well. We wish to decide whether the transmitted sequence was $\boldsymbol{x}$ or $\tilde{\boldsymbol{x}}$ under the Neyman–Pearson criterion of uniformly minimizing the probability of error given that $\boldsymbol{x}$ was transmitted s.t. the constraint that the probability of error given $\tilde{\boldsymbol{x}}$ is guaranteed to be less than $2^{-\lambda n}$, where λ is a given positive constant.

There are a few examples of situations where such a problem may arise. One example is watermark detection, or even more concretely, resolving rightful ownerships of watermarked data (see, e.g., [2] and references therein). The owner of a certain data set $\boldsymbol{x}_0$ (image, video clip, music, text, etc.) wishes to make it publicly available in the Internet, while maintaining good protection against illegal duplication and possible competing copyright ownership claims that could be made by forgers. To this end, the owner embeds into the data set an imperceivable watermark $\boldsymbol{w}$, which conveys a certain message for identification and authentication, and thereby generates the watermarked version $\boldsymbol{x}$ of the data, which should be very close (in some reasonable sense) to the original data set $\boldsymbol{x}_0$ in order to maintain quality. An active attacker processes the watermarked data set $\boldsymbol{x}$ and generates a third version $\boldsymbol{y}$ of the data set within limited additional distortion, in attempt to invalidate the watermark without perceivably further degrading quality. The attack action, which may be deterministic or randomized, is modeled by a finite–state channel, where the finite number of states designates a certain limitation on the memory and/or computational resources that the attacker may have at his diposal. Furthermore, the attacker (or a third party) may make an attempt to claim that the forged data set $\boldsymbol{y}$ actually results from processing (or attacking) yet another watermarked version $\tilde{\boldsymbol{x}}$ of the data set $\boldsymbol{x}_0$, in which a different watermark $\tilde{\boldsymbol{w}}$ is embedded. The message $\tilde{\boldsymbol{w}}$ is planted to seemingly authenticate the illegal party as the owner of this data set. The problem at hand, as described in the first paragraph above, is therefore, to make a decision regarding the real owner (or to prove ownership "beyond any reasonable doubt") without full knowledge of the attack strategy, while keeping good balance between the probabilities of errors of the two kinds.

The above discussion, as well as the results presented in the sequel, extend to the case of multiple alternatives (rather than only one alternative $\tilde{\boldsymbol{x}}$), designating more than one competing ownership claim. Our asymptotic results, however, will hold under the assumption that the number of alternatives is held fixed while $n \to \infty$.

In the case of multiple alternatives, an additional application example naturally arises, and that is: communication across the identification channel (see, e.g., [1], [4], [9] and references therein). The identification communicatiom problem differs from the classical transmission problem in that the receiver, rather than carrying out full decoding, is interested only in a yes/no answer to the question: "Was it message number i that was transmitted?" where the number i is given to the decoder but not pre-specified for the encoder design. For example, in a multiuser network that consists of a broadcast channel, there might be several users that receive the same transmission, which is intended only to one of them. If the communication protocol requires a header that specifies a pointer (or address) of the intended destination, then every receiver will first have to read this header and decide whether or not the message is for him/her. Again, this should be done with good tradeoff between the false-alarm and mis-detection rates, thus giving rise to the Neyman–Pearson criterion in the problem we consider.

In this paper, we propose a universal, asymptotically optimum decision rule for the above–defined detection problem, where the probabilistic characterization of the channel is apriori unknown. The proposed decision rule, as well as our analysis techniques, combine ideas and results from both universal hypothesis testing [3], [7], [13], [14] and universal decoding for finite–state channels [5],[10].

1 Statement of the Problem and Main Result

1.1 Statement of the Problem

Let $\boldsymbol{x} = (x_1, ..., x_n)$ designate a channel input sequence whose elements $\{x_i\}$ take on values in a finite alphabet $\mathcal{X}$. The space of channel input sequences of length n will be denoted by $\mathcal{X}^n$. Similarly, let $\boldsymbol{y} = (y_1, ..., y_n)$ denote a channel output sequence whose elements $\{y_i\}$ take on values in a finite alphabet $\mathcal{Y}$, and let $\mathcal{Y}^n$ denote the set of all possible channel output sequences.

A finite–state channel W with M states (or, an M-state channel) is a probabilistic device that, when fed with an input sequence $\boldsymbol{x}$, generates an output sequence $\boldsymbol{y}$ while going through a sequence of states $\boldsymbol{s} = (s_1, ..., s_n, s_{n+1})$, where each s_i, $1 \leq i \leq n+1$, takes on values in a finite set $\mathcal{S}$ of size M, according to the following probability law

$$\Pr\{\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{X}=\boldsymbol{x}\} \stackrel{\triangle}{=} W^n(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\boldsymbol{s}\in\mathcal{S}^n} W^n(\boldsymbol{y},\boldsymbol{s}|\boldsymbol{x}) \tag{1}$$

where

$$W^n(\boldsymbol{y},\boldsymbol{s}|\boldsymbol{x}) = \prod_{i=1}^{n} W(y_i, s_{i+1}|x_i, s_i), \tag{2}$$

with $W(y,s'|x,s)$ being a conditional probability distribution on $\mathcal{Y}\times\mathcal{S}$ for every $(x,s)\in\mathcal{X}\times\mathcal{S}$, and s_1 being an arbitrary initial state. The conditional probability of an event $A\in\mathcal{Y}^n$ in the channel output space, given the input $\boldsymbol{x}$, will be denoted by $W^n(A|\boldsymbol{x})$.

Given two possible input sequences $\boldsymbol{x}=(x_1,...,x_n)$, $\tilde{\boldsymbol{x}}=(\tilde{x}_1,...,\tilde{x}_n)$, and an output sequence $\boldsymbol{y}=(y_1,...,y_n)$ generated by W with either $\boldsymbol{x}$ or $\tilde{\boldsymbol{x}}$ as input, our task is to decide which one of these input sequences has actually generated the given $\boldsymbol{y}$. The channel W is known to belong to some subclass $\mathcal{W}$ of the class of M-state channels, but is otherwise unknown (and therefore, the likelihood ratio test cannot be implemented).

Following earlier works on universal hypothesis testing (see, e.g., [3],[7],[11] [13]), our concrete goal is the following: We wish to partition $\mathcal{Y}^n$ into two complementary subsets Λ and Λ^c (depending on $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$), where a decision is made in favor of $\boldsymbol{x}$ if and only if $\boldsymbol{y}\in\Lambda$. Under the constraint that the probability of error of the first kind $W^n(\Lambda^c|\boldsymbol{x})$ is less than $2^{-\lambda n}$ for all $W\in\mathcal{W}$, and a given exponential rate $\lambda>0$, we would like to minimize the probability of error of the second kind $W^n(\Lambda|\tilde{\boldsymbol{x}})$, uniformly for every $W\in\mathcal{W}$.
Two comments are in order:

1. The set $\mathcal{W}$ reflects our prior knowledge on the channel. For example, if we have in mind the watermarking application, as discussed in the Introduction, then $\mathcal{W}$ can be naturally defined as a subclass of M-state channels for which the expected distortion between the input and output sequences is less than some prescribed level D, for some distortion measure (see, also [8],[6]). As explained earlier, this reflects the fact that the attacker would like to destory the watermark without causing noticeable degradation in quality.

2. We have forumlated the problem as a binary hypothesis testing problem only for the sake of simplicity. As mentioned in the Introduction and as we shall see later, the extension to the case of many alternative hypotheses is conceptually straightforward.

1.2 Main Result

We now describe the proposed decision rule and state a theorem on its asymptotic optimality in the above defined sense.

Since our decision rule employs Ziv's universal decoding metric for finite–state channels [5],[10], we next present the definition of this decoding metric for the sake of completeness. Given two finite–alphabet sequences, $\boldsymbol{y}$ and $\boldsymbol{z}$ of length n, let us apply the incremental parsing procedure of the Lempel–Ziv algorithm [12] to the sequence of pairs $((y_1, z_1), (y_2, z_2), ..., (y_n, z_n))$. According to this procedure, $(\boldsymbol{y}, \boldsymbol{z})$ is parsed sequentially into phrases, where each new phrase is the shortest string that has not appeared earlier as a parsed phrase. Thus all phrases are distinct with a possible exception of the last phrase, which might be incomplete. Let $c(\boldsymbol{y}, \boldsymbol{z})$ denote the number of distinct phrases. For example,[1] if

$$\begin{aligned} \boldsymbol{y} &= 0 \mid 1 \mid 0\,0 \mid 0\,1 \mid \\ \boldsymbol{z} &= 0 \mid 1 \mid 0\,1 \mid 0\,1 \mid \end{aligned}$$

then $c(\boldsymbol{y}, \boldsymbol{z}) = 4$. Let $c(\boldsymbol{z})$ denote the resulting number of distinct phrases of $\boldsymbol{z}$, and let $\boldsymbol{z}(l)$ denote the lth distinct $\boldsymbol{z}$–phrase, $l = 1, 2, ..., c(\boldsymbol{z})$. In the above example, $c(\boldsymbol{z}) = 3$. Denote by $c_l(\boldsymbol{y}|\boldsymbol{z})$ the number of occurrences of $\boldsymbol{z}(l)$ in the parsing of $\boldsymbol{z}$, or equivalently, the number of distinct $\boldsymbol{y}$-phrases that jointly appear with $\boldsymbol{z}(l)$. Clearly, $\sum_{l=1}^{c(\boldsymbol{z})} c_l(\boldsymbol{y}|\boldsymbol{z}) = c(\boldsymbol{y}, \boldsymbol{z})$. In the above example, $\boldsymbol{z}(1) = 0$, $\boldsymbol{z}(2) = 1$, $\boldsymbol{z}(3) = 01$, $c_1(\boldsymbol{y}|\boldsymbol{z}) = c_2(\boldsymbol{y}|\boldsymbol{z}) = 1$, and $c_3(\boldsymbol{y}|\boldsymbol{z}) = 2$. Now, Ziv's universal decoding metric is defined as

$$U(\boldsymbol{y}|\boldsymbol{z}) = \sum_{l=1}^{c(\boldsymbol{z})} c_l(\boldsymbol{y}|\boldsymbol{z}) \log c_l(\boldsymbol{y}|\boldsymbol{z}). \tag{3}$$

Returning to our problem, given $\boldsymbol{x}$, $\tilde{\boldsymbol{x}}$ and $\boldsymbol{y}$ of length n, let $U(\boldsymbol{y}|\boldsymbol{x}, \tilde{\boldsymbol{x}})$ be defined as $U(\boldsymbol{y}|\boldsymbol{z})$, where $\boldsymbol{z} = (z_1, ..., z_n)$, $z_i = (x_i, \tilde{x}_i) \in \mathcal{X}^2$, $i = 1, ..., n$. (This means that $U(\boldsymbol{y}|\boldsymbol{x}, \tilde{\boldsymbol{x}})$ is calculated from the joint incremental parsing of all three sequences $\boldsymbol{x}$, $\tilde{\boldsymbol{x}}$, and $\boldsymbol{y}$.) Finally, for given $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$, define

$$\Lambda \stackrel{\triangle}{=} \left\{ \boldsymbol{y} : \max_{W \in \mathcal{W}} \log W^n(\boldsymbol{y}|\boldsymbol{x}) + U(\boldsymbol{y}|\boldsymbol{x}, \tilde{\boldsymbol{x}}) + \lambda n \geq 0 \right\}. \tag{4}$$

Our main result is the asymptotic optimality of a decision rule that decides in favor of $\boldsymbol{x}$ if and only if $\boldsymbol{y} \in \Lambda$. This is summarized in the following theorem.

[1] The same example appears in [10].

Theorem 1 *(a) For every $W \in \mathcal{W}$,*

$$W^n(\Lambda^c|\boldsymbol{x}) \leq 2^{-\lambda n + o(n)}$$

.

(b) If Ω is a subset of $\mathcal{Y}^n$ (which may depend on $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$) that satisfies

$$W^n(\Omega^c|\boldsymbol{x}) \leq 2^{-\lambda' n} \quad \forall W \in \mathcal{W},$$

for some $\lambda' > \lambda$, then $\Omega \supseteq \Lambda$ for all sufficiently large n.

The proof appears in Section 3.

Discussion Part (a) of the theorem tells us that the decision rule defined by Λ satisfies the constraint regarding the error probability of the first kind for every possible channel in the class $\mathcal{W}$. Part (b) tells that for every competing decision rule defined by Ω (i.e., a decision rule that classifies $\boldsymbol{y}$ as emerging from $\boldsymbol{x}$ iff $\boldsymbol{y} \in \Omega$) that also satisfies the constraint of part (a), with the slight modification that the weak inequality is replaced by a strong one, gives a higher probability of error uniformly for every W and $\tilde{\boldsymbol{x}}$, because $\Omega \supseteq \Lambda$ implies $W^n(\Omega|\tilde{\boldsymbol{x}}) \geq W^n(\Lambda|\tilde{\boldsymbol{x}})$. Thus, the proposed desicion rule Λ asymptotically minimizes the probability of error of the second kind subject to the aforementioned constraint on the probability of error of the first kind.

Note that the decision rule Λ has a form similar to those proposed in earilier works [3], [7], [13] in the sense that the sum of the maximum log-likelihood and the Lempel-Ziv complexity is compared to a threshold. Here there is the additional ingredient that the Lempel-Ziv complexity term is replaced by the universal decoding metric that incorporates the "side information" $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$. Observe also that in our definition of $U(\boldsymbol{y}|\boldsymbol{x}, \tilde{\boldsymbol{x}})$, the roles of the channel input and channel output sequences have been interchanged compared to the definition of the universal decoding metric in [10], where a random coding regime was considered. Another interesting difference between the two metrics is that here the joint parsing includes *both* the input $\boldsymbol{x}$ corresponding to the null hypothesis and the alternative $\tilde{\boldsymbol{x}}$, and not a comparison between two quantities, one of which depending on $\boldsymbol{x}$, and the other on $\tilde{\boldsymbol{x}}$.

To gain some intuition regarding the rationale behind this decision rule, it is perhaps illuminating to confine attention, for a moment, to the class $\mathcal{W}$ of memoryless channels, i.e., $M = 1$. In this case, $\log \max_W W(\boldsymbol{y}|\boldsymbol{x}) = -n\hat{H}(Y|X)$, where $\hat{H}(Y|X)$ is the empirical conditional entropy of a random variable Y given another random variable X, where X and Y are jointly distributed according to the empirical joint probability distribution induced by $(\boldsymbol{x}, \boldsymbol{y})$. On the other hand, $U(\boldsymbol{y}|\boldsymbol{x}, \tilde{\boldsymbol{x}})$ is an estimate of (and in fact, for $M = 1$, can be replaced by) $\hat{H}(Y|X, \tilde{X})$, the conditional empirical entropy of Y given X and $\tilde{X}$ corresponding to the empirical joint probability distribution of $(\boldsymbol{x}, \tilde{\boldsymbol{x}}, \boldsymbol{y})$. It follows then, that our decision rule, in this case, essentially compares the empirical conditional mutual information $\hat{I}(\tilde{X}; Y|X)$ to the threshold λ. Now, the intuition is quite clear: If $\boldsymbol{x}$ is indeed the true channel input, then typically, $\tilde{\boldsymbol{x}}$ and $\boldsymbol{y}$ will look like independent processes given $\boldsymbol{x}$, and so, $\hat{I}(\tilde{X}; Y|X)$ is likely to be small. If, on the other hand, $\tilde{\boldsymbol{x}}$ was the input that generated $\boldsymbol{y}$, then $\tilde{\boldsymbol{x}}$ and $\boldsymbol{y}$ will typically seem to be dependent, whereas $\boldsymbol{x}$ will be empirically independent of them. The assymetric roles played by $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ are a result of the assymetry of the Neyman–Pearson criterion w.r.t. the two kinds of errors.

Finally, when multiple alternative channel input sequences are given $\tilde{\boldsymbol{x}}^{(1)}, \tilde{\boldsymbol{x}}^{(2)}, ..., \tilde{\boldsymbol{x}}^{(m)}$, the corresponding generalization of our decision rule is as follows: Accept the hypothesis that $\boldsymbol{x}$ has been transmitted if and only if

$$\log \max_{W \in \mathcal{W}} W^n(\boldsymbol{y}|\boldsymbol{x}) + U(\boldsymbol{y}|\boldsymbol{x}, \tilde{\boldsymbol{x}}^{(1)}, ..., \tilde{\boldsymbol{x}}^{(m)}) + \lambda n > 0. \tag{5}$$

The proof of Theorem 1 extends straightforwardly to show that, provided that m is a positive integer that does not grow with n, this decision rule continues to satisfy the constraint of part (a) of Theorem 1 and at the same time it asymptotically minimizes all the probabilitiies of error of the second kind, i.e., $W(\text{error}|\tilde{\boldsymbol{x}}^{(i)})$, $i = 1, ..., m$. The assumption that m is fixed independently of n (which is also essential for practical implementability of this decision rule), might be acceptable in watermarking applications where relatively short messages are embedded in very large images, thus requiring very low coding rates.

Proof of Theorem 1

We begin with part (a) of the theorem, showing that $W^n(\Lambda^c|\boldsymbol{x}) \leq 2^{-\lambda n + o(n)}$ for all $W \in \mathcal{W}_M$.

$$\begin{aligned} W^n(\Lambda^c|\boldsymbol{x}) &= \sum_{\boldsymbol{y} \in \Lambda^c} W^n(\boldsymbol{y}|\boldsymbol{x}) \\ &\leq \sum_{\boldsymbol{y} \in \Lambda^c} 2^{\max_{W \in \mathcal{W}_M} \log W^n(\boldsymbol{y}|\boldsymbol{x})} \end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{\boldsymbol{y}\in\Lambda^c} 2^{-\lambda n - U(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}})} \\
&\leq 2^{-\lambda n} \sum_{\boldsymbol{y}\in\mathcal{Y}^n} 2^{-U(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}})} \\
&= 2^{-\lambda n + o(n)},
\end{aligned}
\tag{6}
$$

where the last step follows from Kraft's inequality and the fact that $U(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}}) + o(n)$ is a length function of a legitimate lossless code for $\boldsymbol{y}$ when $(\boldsymbol{x},\tilde{\boldsymbol{x}})$ serves as side information available at both encoder and decoder [10, Lemma 2].

Turning to part (b), let k divide n, and let us parse $\boldsymbol{x}$, $\tilde{\boldsymbol{x}}$, and $\boldsymbol{y}$ into n/k blocks of length k, where the ith block starts at position $(i-1)k+1$ and ends at ik, $i = 1, 2, ..., n/k$. Let $\boldsymbol{x}_i$, $\tilde{\boldsymbol{x}}_i$ and $\boldsymbol{y}_i$ denote the ith blocks of $\boldsymbol{x}$, $\tilde{\boldsymbol{x}}$, and $\boldsymbol{y}$, respectively. Let $\boldsymbol{s}^k$ denote the sequence of states at the block boundaries, i.e., $(s_0, s_k, s_{2k}, ..., s_n)$. Suppose next that a genie provides us with $\boldsymbol{s}^k$ in addition to the channel output $\boldsymbol{y}$. Obviously, the original class of decision rules, that are formed by partitioning $\mathcal{Y}^n$ into two complementary subsets, is a subset of the class of decision rules that partitions the product space $\mathcal{Y}^n \times \mathcal{S}^{n/k+1}$ of pairs $(\boldsymbol{y}, \boldsymbol{s}^k)$: While the latter can be considered as a partition of $\mathcal{Y}^n$ that depends on $\boldsymbol{s}^k$, in the former, this partition is the same for all $\boldsymbol{s}^k$. Stated differently, decision rules that partition $\mathcal{Y}^n$ into (Ω, Ω^c) and are based on $\boldsymbol{y}$ only, can be formally viewed as partitions of $\mathcal{Y}^n \times \mathcal{S}^{n/k+1}$ with the form $(\Omega \times \mathcal{S}^{n/k+1}, \Omega^c \times \mathcal{S}^{n/k+1})$. Thus, if we can prove our claim w.r.t. the wider class of decision rules, it will certainly hold for the original class as well.

Let the conditional type class $T_k(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}},\boldsymbol{s}^k)$ denote the set of sequences $\boldsymbol{y}'$ generated by permuting blocks $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ for which $s_{(i-1)k} = s_{(j-1)k}$, $s_{ik} = s_{jk}$, $\boldsymbol{x}_i = \boldsymbol{x}_j$, and $\tilde{\boldsymbol{x}}_i' = \tilde{\boldsymbol{x}}_j'$. Note that $W^n(\boldsymbol{y}, \boldsymbol{s}^k|\boldsymbol{x}) = \prod_{i=1}^{n/k} W^n(\boldsymbol{y}_i, s_{ik}|\boldsymbol{x}_i, s_{(i-1)k})$, and a similar equality holds for $W^n(\boldsymbol{y}, \boldsymbol{s}^k|\tilde{\boldsymbol{x}})$. Thus, for any $W \in \mathcal{W}_M$, $W^n(\boldsymbol{y}', \boldsymbol{s}^k|\boldsymbol{x}) = W^n(\boldsymbol{y}, \boldsymbol{s}^k|\boldsymbol{x})$ and $W^n(\boldsymbol{y}', \boldsymbol{s}^k|\tilde{\boldsymbol{x}}) = W^n(\boldsymbol{y}, \boldsymbol{s}^k|\tilde{\boldsymbol{x}})$ whenever $\boldsymbol{y}' \in T_k(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}},\boldsymbol{s}^k)$. We now observe that without loss of asymptotic optimality (in the error exponent sense), $\{T_k(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}},\boldsymbol{s}^k)\}$ can be considered as "atoms" of decision rules over $\mathcal{Y}^n \times \mathcal{S}^{n/k+1}$. In other words, we may assume that each conditional type class $T_k(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}},\boldsymbol{s}^k)$ is either entirely in Ω or entirely in Ω^c. To see why this is true, consider an arbitrary decision rule, and modify it so as to satisfy this property by transferring $\boldsymbol{y}$–vectors to the other decision region whenever more than half of the vectors of $T_k(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}},\boldsymbol{s}^k)$ are already in the other region. This modification at most doubles (and hence does not affect exponents of) either kind of error probability $W^n(\Omega^c|\boldsymbol{x})$ or $W^n(\Omega|\tilde{\boldsymbol{x}})$ because typical sequences are equiprobable under both measures as explained earlier.

Consider then an arbitrary decision rule over $\mathcal{Y}^n \times \mathcal{S}^{n/k+1}$ for which each decision region is a union of type classes $\{T_k(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}},\boldsymbol{s}^k)\}$. The constraint on the probability of error of the first kind implies that there exists some $\epsilon > 0$ such that

$$\begin{aligned} 2^{-(\lambda+\epsilon)n} &\geq \max_{W\in\mathcal{W}_M} \sum_{(\boldsymbol{y},\boldsymbol{s}^k)\in\Omega^c} W^n(\boldsymbol{y},\boldsymbol{s}^k|\boldsymbol{x}) \\ &\geq |T_k(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}},\boldsymbol{s}^k)| \max_{W\in\mathcal{W}_M} W^n(\boldsymbol{y},\boldsymbol{s}^k|\boldsymbol{x}) \end{aligned} \tag{7}$$

for any $(\boldsymbol{y},\boldsymbol{s}^k) \in \Omega^c$. We next further lower bound the right–most side of the last expression by an expression that is independent of $\boldsymbol{s}^k$. This can be done by using the following lemmas.

Lemma 1 *For every $\boldsymbol{x}$, $\tilde{\boldsymbol{x}}$, $\boldsymbol{y}$, and $\boldsymbol{s}^k$,*

$$|T_k(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}},\boldsymbol{s}^k)| \geq 2^{U(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}})-n\epsilon_1(n,k)}, \tag{8}$$

where $\epsilon_1(n,k)$ is independent of $\boldsymbol{x}$, $\tilde{\boldsymbol{x}}$, $\boldsymbol{y}$, and $\boldsymbol{s}^k$, and $\lim_{k\to\infty}\lim_{n\to\infty}\epsilon_1(n,k) = 0$.

Lemma 2 *For every $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{s}^k$,*

$$\max_{W\in\mathcal{W}_M} W^n(\boldsymbol{y},\boldsymbol{s}^k|\boldsymbol{x}) \geq 2^{-n\epsilon_2(k)} \max_{W\in\mathcal{W}_M} W^n(\boldsymbol{y}|\boldsymbol{x}), \tag{9}$$

where $\epsilon_1(k)$ is independent of $\boldsymbol{x}$, $\tilde{\boldsymbol{x}}$, and $\boldsymbol{s}^k$, and $\lim_{k\to\infty}\epsilon_2(k) = 0$.

Once these two lemmas are proven, then by combining them with eq. (7), we could conclude that any $(\boldsymbol{y},\boldsymbol{s}^k)$ in Ω^c is also in $\Lambda^c \times \mathcal{S}^{n/k}$ for sufficiently large n and k such that $\epsilon_1(n,k)+\epsilon_2(k) \leq \epsilon$. Thus, $\Lambda \times \mathcal{S}^{n/k}$ is a subset of Ω. Now, the proof of Lemma 2 is exactly the same as that of [13, Lemma 2]. The proof of Lemma 1 is similar to that of Lemma 1 of [13], but since there are a few technical differences, we provide it here.

Proof of Lemma 1. Given $\boldsymbol{x}$, $\tilde{\boldsymbol{x}}$, $\boldsymbol{y}$, and $\boldsymbol{s}^k$, let us define the empirical joint probability mass function

$$\hat{P}(x^k,\tilde{x}^k,y^k,s,s') = \frac{k}{n}\sum_{i=1}^{n/k}\delta(\boldsymbol{x}_i = x^k, \tilde{\boldsymbol{x}}_i = \tilde{x}^k, \boldsymbol{y}_i = y^k, s_{(i-1)k} = s, s_{ik} = s'), \tag{10}$$

where $\delta(\cdot)$ is the indicator function of an event, and $x^k \in \mathcal{X}^k$, $\tilde{x}^k \in \mathcal{X}^k$, $y^k \in \mathcal{Y}^k$, $s,s' \in \mathcal{S}$. We shall also denote the unnormalized sum on the right–hand side of the last equation by $n(x^k,\tilde{x}^k,y^k,s,s')$, symbolizing the joint count of the vector $(x^k,\tilde{x}^k,y^k,s,s')$. Joint counts of subsets of coordinates of this vector (marginal counts) will be denoted similarly by $n(\cdot)$ with an argument denoting the event being counted, e.g., $n(x^k,y^k)$ is the joint count of x^k and

y^k, and so on. Information measures like entropy and mutual information, that are induced by the joint PMF $\hat{P}$, will henceforth be denoted by $\hat{H}(\cdot)$ and $\hat{I}(\cdot)$, respectively, where the 'arguments' are denoted by capital letters that symbolize the corresponding random variables governed by $\hat{P}$. By using the inequalities $m(\log m - \log e) \le m! \le m \log m$ (m – positive integer), we get

$$
\begin{aligned}
\log |T_k(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}},\boldsymbol{s}^k)| &= \log \prod_{x^k,\tilde{x}^k,s,s'} \frac{n(x^k,\tilde{x}^k,s,s')!}{\prod_{y^k} n(x^k,\tilde{x}^k,y^k,s,s')!} \\
&= \sum_{x^k,\tilde{x}^k,s,s'} \left[\log n(x^k,\tilde{x}^k,s,s')! - \sum_{y^k} \log n(x^k,\tilde{x}^k,y^k,s,s')!\right] \\
&\ge \sum_{x^k,\tilde{x}^k,s,s'} \Big[n(x^k,\tilde{x}^k,s,s')(\log n(x^k,\tilde{x}^k,s,s') - \log e) - \\
&\qquad \sum_{y^k} n(x^k,\tilde{x}^k,y^k,s,s') \log n(x^k,\tilde{x}^k,y^k,s,s')\Big] \\
&= \frac{n}{k}[\hat{H}(Y^k|X^k,\tilde{X}^k,S,S') - \log e] \\
&\ge \frac{n}{k}[\hat{H}(Y^k|X^k,\tilde{X}^k) - \hat{I}(Y^k;S,S'|X^k,\tilde{X}^k) - \log e] \\
&\ge \frac{n}{k}[\hat{H}(Y^k|X^k,\tilde{X}^k) - \hat{H}(S,S'|X^k,\tilde{X}^k) - \log e] \\
&\ge \frac{n}{k}[\hat{H}(Y^k|X^k,\tilde{X}^k) - \log(M^2 e)]. \qquad (11)
\end{aligned}
$$

The final step would be to show that $\frac{n}{k}\hat{H}(Y^k|X^k,\tilde{X}^k)$ is essentially lower bounded by $U(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}})$. This will be done by showing that analogously to the Lempel–Ziv complexity [12], $U(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}})$ is roughly a lower bound to the compression ratio attainable by a finite–state machine when compressing $\boldsymbol{y}$ in the presence of $(\boldsymbol{x},\tilde{\boldsymbol{x}})$ as side information at both encoder and decoder.

To this end, let us first generalize the concept of finite–state compressibility [12] so as to incorporate side-information. For the sake of convenience, let us again use the notation $z_i = (x_i, \tilde{x}_i)$, and $\boldsymbol{z} = (\boldsymbol{x},\tilde{\boldsymbol{x}}) = (z_1, z_2, ..., z_n)$. A K-state encoder E with side information is defined by a set of six objects $(\Sigma, \mathcal{B}, \mathcal{Y}, \mathcal{Z}, f, g)$, where Σ is a finite set of K states, $\mathcal{B}$ is a finite set of binary words (possibly of different lengths, including the null word for idling), $\mathcal{Y}$ is the finite alphabet of the source to be compressed, $\mathcal{Z}$ is a finite alphabet of side information (in our case, $\mathcal{Z} = \mathcal{X}^2$), $f : \Sigma \times \mathcal{Y} \times \mathcal{Z} \to \mathcal{B}$ is the encoder output function, and $g : \Sigma \times \mathcal{Y} \times \mathcal{Z} \to \Sigma$ is the next–state function. When an input sequence $(y_1, y_2, ...)$ and a side information sequence $(z_1, z_2, ...)$ are fed together, sequentially into $E = (\Sigma, \mathcal{Y}, \mathcal{Z}, \mathcal{B}, f, g)$, the encoder ouputs a sequence of binary words $(b_1, b_2, ...)$, $b_i \in \mathcal{B}$, while going through a sequence of states $(\sigma_1, \sigma_2, ...)$, $\sigma_i \in \Sigma$, according to

$$b_i = f(\sigma_i, y_i, z_i), \quad \sigma_{i+1} = g(\sigma_i, y_i, z_i), \quad i = 1, 2, ... \tag{12}$$

where σ_i is the state of E at time instant i. The decoder, on the other hand, receives the pair sequence $((b_1, z_1), (b_2, z_2), ...)$ and reconstructs the source sequence $(y_1, y_2, ...)$.

A finite–state encoder E with side information is said to be *information lossless* (IL) if for all $\sigma_1 \in \Sigma$ and all $(\boldsymbol{y}, \boldsymbol{z}) \in \mathcal{Y}^n \times \mathcal{Z}^n$, $n \geq 1$, the quadruple $(\sigma_1, \sigma_{n+1}, \boldsymbol{b}, \boldsymbol{z})$ uniquely determines $\boldsymbol{y}$, where σ_{n+1} and $\boldsymbol{b} = (b_1, ..., b_n)$ are obtained by iterating eq. (12) with σ_1, $\boldsymbol{y}$, and $\boldsymbol{z}$ as inputs. The length function associated with E is defined as $L_E(\boldsymbol{y}|\boldsymbol{z}) = \sum_{i=1}^n L(b_i)$, where $L(b_i)$ is the length of the binary string $b_i \in \mathcal{B}$.

Consider the incremental parsing of the pair sequence $(\boldsymbol{y}, \boldsymbol{z}) = ((y_1, z_1), ...(y_n, z_n))$ and let $c(\boldsymbol{z})$ and $c_l(\boldsymbol{y}|\boldsymbol{z})$, $l = 1, ..., c(\boldsymbol{z})$, be defined as in Section 2. We are now ready to state a lower bound to $L_E(\boldsymbol{y}|\boldsymbol{z})$ that is analogous to [12, Theorem 1]: For any K-state IL encoder with side information and for every $(\boldsymbol{y}, \boldsymbol{z}) \in \mathcal{Y}^n \times \mathcal{Z}^n$, $n \geq 1$,

$$L_E(\boldsymbol{y}|\boldsymbol{z}) \geq \sum_{l=1}^{c(\boldsymbol{z})} [c_l(\boldsymbol{y}|\boldsymbol{z}) + K^2] \log \frac{c_l(\boldsymbol{y}|\boldsymbol{z})}{4K^2}. \tag{13}$$

The proof of eq. (13) is a straightforward extension of the proof of [12, Theorem 1]: Consider the $c_l(\boldsymbol{y}|\boldsymbol{z})$ distinct y-phrases associated with the lth distinct z-phrase, $1 \leq l \leq c(\boldsymbol{z})$. Let $L_E^l(\boldsymbol{y}|\boldsymbol{z})$ denote the total codeword length contributed by encoding these $c_l(\boldsymbol{y}|\boldsymbol{z})$ phrases, that is, $L_E^l(\boldsymbol{y}|\boldsymbol{z}) = \sum_{i \in \mathcal{I}_l} L(b_i)$, where $\mathcal{I}_l$ denotes the set of time instants at which some occurrence of the lth distinct z-phrase is visited. Since the corresponding z-phrases in these segments are all identical by construction, the proof of [12, Theorem 1] can be directely applied to the phrases of each $\mathcal{I}_l$, resulting in the inequality [12]

$$\begin{aligned} L_E^l(\boldsymbol{y}|\boldsymbol{z}) &\geq [c_l(\boldsymbol{y}|\boldsymbol{z}) + K^2] \log \frac{c_l(\boldsymbol{y}|\boldsymbol{z}) + K^2}{4K^2} + 2K^2 \\ &> [c_l(\boldsymbol{y}|\boldsymbol{z}) + K^2] \log \frac{c_l(\boldsymbol{y}|\boldsymbol{z})}{4K^2}, \end{aligned} \tag{14}$$

and so, eq. (13) is readily obtained by summation of both sides over $1 \leq l \leq c(\boldsymbol{z})$.

Consider next the Shannon code E_S of $\boldsymbol{y}$ given $\boldsymbol{z}$ based on k-blocks using an arbitrary conditional PMF $Q(y^k|z^k)$. According to this code, y^k is encoded, in the presence of z^k, using $\lceil -\log Q(y^k|z^k) \rceil$ bits, and so,

$$L_{E_S}(\boldsymbol{y}|\boldsymbol{z}) = \sum_{i=1}^{n/k} \lceil -\log Q(y^k|z^k) \rceil$$

$$
\begin{aligned}
&= \frac{n}{k} \sum_{y^k, z^k} \hat{P}(y^k, z^k) \lceil -\log Q(y^k|z^k) \rceil \\
&\leq \frac{n}{k} \sum_{y^k, z^k} \hat{P}(y^k, z^k) [-\log Q(y^k|z^k) + 1] \\
&\leq -\frac{n}{k} \sum_{y^k, z^k} \hat{P}(y^k, z^k) \log Q(y^k|z^k) + \frac{n}{k}. \qquad (15)
\end{aligned}
$$

It is easy to see that this code can be implemented by a finite-state encoder in the following manner: At the beginning of each block, the encoder is always at some fixed "initial state" symbolized by σ_0. At time instant t, the state σ_t is defined as

$$
((y_t, z_t), (y_{t-1}, z_{t-1}), ..., (y_{t-t_k+1}, z_{t-t_k+1})),
$$

where $t_k = t \bmod k$. The encoder outputs the null string whenever $t_k \neq 0$; when $t_k = 0$, it emits the Shannon codeword of the block just terminated. The total number of states is therefore $K = \sum_{j=0}^{k} q^j = (q^{k+1} - 1)/(q-1) \stackrel{\triangle}{=} S(q,k)$, where $q = |\mathcal{Y} \times \mathcal{Z}|$.

It is also easy to see that E_S is IL. For given positive integers i and j ($i < j$), suppose we are given σ_i, σ_{j+1}, $(z_i, z_{i+1}, ..., z_j)$ and $(b_i, b_{i+1}, ..., b_j)$. Then $(y_i, y_{i+1}, ..., y_j)$ can be reconstructed as follows. If time instants i and j fall in the same k-block then σ_{j+1} conveys full information on $(y_i, y_{i+1}, ..., y_j)$. Otherwise, we use the following procedure: The segment from time i until the end of the current block is reconstructed by decoding the codeword emitted at the end of this block. Similarly, if there are any additional blocks that are fully contained in the segment from i to j, they can also be reconstructed by decoding. Finally, the portion of last block until position j can be recovered again from the final state.

It now follows that $L_{E_S}(\boldsymbol{y}|\boldsymbol{z})$ must satisfy the lower bound (13) with $K = S(q,k)$, and so,

$$
-\frac{n}{k} \sum_{y^k, z^k} \hat{P}(y^k, z^k) \log Q(y^k|z^k) + \frac{n}{k} \geq \sum_{l=1}^{c(\boldsymbol{z})} [c_l(\boldsymbol{y}|\boldsymbol{z}) + S^2(q,k)] \log \frac{c_l(\boldsymbol{y}|\boldsymbol{z})}{4S^2(q,k)}. \qquad (16)
$$

Since this holds for every Q while the right-hand side is independent of Q, we may minimize the left-hand side w.r.t. Q and obtain

$$
\begin{aligned}
\frac{n}{k} \hat{H}(Y^k|Z^k) + \frac{n}{k} &\equiv \frac{n}{k} \hat{H}(Y^k|X^k, \tilde{X}^k) + \frac{n}{k} \\
&\geq \sum_{l=1}^{c(\boldsymbol{z})} [c_l(\boldsymbol{y}|\boldsymbol{z}) + S^2(q,k)] \log \frac{c_l(\boldsymbol{y}|\boldsymbol{z})}{4S^2(q,k)}. \qquad (17)
\end{aligned}
$$

Combining this with eq. (11), we get

$$
\log |T_k(\boldsymbol{y}|\boldsymbol{x}, \tilde{\boldsymbol{x}}, \boldsymbol{s}^k)| \geq \sum_{l=1}^{c(\boldsymbol{z})} [c_l(\boldsymbol{y}|\boldsymbol{z}) + S^2(q,k)] \log \frac{c_l(\boldsymbol{y}|\boldsymbol{z})}{4S^2(q,k)} - \frac{n}{k} \log(2M^2 e)
$$

$$\geq \quad U(\boldsymbol{y}|\boldsymbol{x},\tilde{\boldsymbol{x}}) - n\epsilon_1(n,k) \tag{18}$$

with

$$\epsilon_1(n,k) \stackrel{\triangle}{=} \max_{\boldsymbol{y},\boldsymbol{z}} \left\{ \frac{1}{n}[c(\boldsymbol{y},\boldsymbol{z}) + S^2(q,k)c(\boldsymbol{z})]\log[4S^2(q,k)] + \frac{\log(2M^2e)}{k} \right\} \tag{19}$$

However, it is well-known [12] that $\max_{\boldsymbol{y},\boldsymbol{z}} c(\boldsymbol{y},\boldsymbol{z})$ (and hence also $\max_{\boldsymbol{z}} c(\boldsymbol{z})$ which is smaller) cannot exceed $O(n/\log n)$, and so, the above expression of $\epsilon_1(n,k)$ satisfies the properties stated in the assertion of Lemma 1. This completes the proof of Lemma 1.

References

[1] R. Ahlswede and G. Dueck, "Identification via channels," *IEEE Trans. on Inform. Theory*, vol. 35, no. 1, pp. 15–29, January 1989.

[2] S. Craver, N. Memon, B.-L. Yeo, and M. M. Yeung, "Resolving rightful ownerships with invisible watermarking techniques: limitations, attacks, and implications," *IEEE J. Select. Areas in Commun.,* vol. 16, no. 4, pp. 573–586, May 1998.

[3] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically–observed statistics," *IEEE Trans. on Inform. Theory*, vol. 35, no. 2, pp. 401–408, March 1989.

[4] T. S. Han and S. Verdú, "New results in the theory of identification via channels," *IEEE Trans. on Inform. Theory*, vol. 38, no. 1, pp. 14–25, January 1992.

[5] A. Lapidoth and J. Ziv, "On the universality of the LZ-based decoding algorithm," *IEEE Trans. on Inform. Theory*, vol. 44, no. 5, pp. 1746—1755, September 1998.

[6] N. Merhav, "On random coding error exponents of watermarking codes," submitted for publication to *IEEE Trans. on Inform. Theory.*

[7] N. Merhav, M. Gutman, and J. Ziv, "On the estimation of the order of a Markov chain and universal data compression," *IEEE Trans. on Inform. Theory*, vol. 35, no. 5, pp. 1014–1019, September 1989.

[8] J. A. O'Sullivan, P. Moulin, and M. Ettinger, "Information thoeretic analysis of steganography," *Proc. 1998 International Symposium on Information Theory*, p. 298, MIT, Cambridge MA, August 1998.

[9] Y. Steinberg, "New converses in the theory of identification via channels," *IEEE Trans. on Inform. Theory*, vol. 44, no. 3, pp. 984–998, May 1998.

[10] J. Ziv, "Universal decoding for finite–state channels," *IEEE Trans. on Inform. Theory*, vol. 33, no. 4, pp. 453–460, July 1985.

[11] J. Ziv, "On classification with empirically–observed statisitics and universal data compression," *IEEE Trans. on Inform. Theory*, vol. 34, no. 2, pp. 278–286, March 1988.

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

[13] J. Ziv and N. Merhav, "Estimating the number of states of a finite–state source," *IEEE Trans. on Inform. Theory*, vol. 38, no. 1, pp. 61–65, January 1992.

[14] J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification," *IEEE Trans. on Inform. Theory*, vol. 39, no. 4, pp. 1270–1279, July 1993.