

# The Information Lost in Erasures

Sergio Verdú, *Fellow, IEEE*, and Tsachy Weissman, *Senior Member, IEEE*

**Abstract**—We consider sources and channels with memory observed through erasure channels. In particular, we examine the impact of sporadic erasures on the fundamental limits of lossless data compression, lossy data compression, channel coding, and denoising.

We define the *erasure entropy* of a collection of random variables as the sum of entropies of the individual variables conditioned on all the rest. The erasure entropy measures the information content carried by each symbol knowing its context. The erasure entropy rate is shown to be the minimal amount of bits per erasure required to recover the lost information in the limit of small erasure probability. When we allow recovery of the erased symbols within a prescribed degree of distortion, the fundamental tradeoff is described by the erasure rate–distortion function which we characterize. We show that in the regime of sporadic erasures, knowledge at the encoder of the erasure locations does not lower the rate required to achieve a given distortion. When no additional encoded information is available, the erased information is reconstructed solely on the basis of its context by a denoiser. Connections between erasure entropy and discrete denoising are developed. The decrease of the capacity of channels with memory due to sporadic memoryless erasures is also characterized in wide generality.

**Index Terms**—Channel coding, channels with memory, data compression, discrete denoising, entropy, erasure channels, Markov processes, rate–distortion theory, Shannon theory.

## I. INTRODUCTION

### A. Scope

THE memoryless erasure channel in which symbols are replaced by a special erasure symbol independently and with probability  $\epsilon$  is a very useful abstraction for various types of data loss or low reliability reception, and plays a fundamental role in channel coding theory and the information theory of noisy channels.

Questions of engineering interest that arise when information is erased can be divided into two major categories according to whether the recovery of the erased information is lossless or lossy.

Problems in lossless (or almost lossless in the usual sense) recovery include the following.

Manuscript received April 30, 2007; revised July 17, 2008; current version published October 22, 2008. This work was supported by the U.S. National Science Foundation (NSF) under Grants CCR-0312839, CCF-0728445 and an NSF CAREER award. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Seattle, WA, July 2006.

S. Verdú is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: verdu@princeton.edu).

T. Weissman is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Technion City, Haifa 32000, Israel, on leave from the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: tsachy@stanford.edu).

Communicated by Y. Steinberg, Associate Editor for Shannon Theory.

Color versions of Figures 1 and 3 in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TIT.2008.929968

1. Adding redundancy prior to transmission of nonredundant data for protection against erasures as well as other noisy channel uncertainty. A particular interesting class of problems is the analysis of the capacity of channels formed by concatenating a noisy channel and an erasure channel.
2. The source is connected directly to the erasure channel, and the goal is to convey efficiently the erased information to the receiver knowing the erasure locations, taking advantage of the redundancy of the source. The fundamental measure of compression efficiency is the entropy of the source conditioned on the erasure channel output.
3. Conveying the erased information to the receiver not knowing the erasure locations. This is a Slepian–Wolf data compression setup and therefore the compression rate does not suffer because of the ignorance of the erasure locations.
4. The problems in items 2 and 3 where the source undergoes a concatenation of noisy channel and erasure channel.

In lossy recovery, where a distortion measure gauges the quality of the reproduction of the erased information, problems of interest include the following.

5. The source is connected directly to the erasure channel, and the goal is to convey an approximation of the erased information to the receiver knowing the erasure locations. In this case, the fundamental tradeoff of compression efficiency is represented by a conditional rate–distortion function given the nonerased information.
6. The same as in item 5 but with a compressor which is ignorant of the locations of the erasures. Since those are known only at the decompressor, the fundamental tradeoff is given by a Wyner–Ziv rate–distortion function.
7. If the erasure channel output is available to the compressor but not the decompressor, and the nonerased source is available to neither, the fundamental tradeoff of rate versus reproduction fidelity is given by a rate–distortion function for lossy compression of noisy (in this case partially erased) data.
8. Denoising (or more concretely “derasing”) the output of the erasure channel. In this case, no coding (either adding or eliminating redundancy) is allowed, and thus the denoiser relies entirely on the redundancy of the source. The minimum achievable distortion at the output of the erasure channel is called the “derasurability.”

In each of the above cases we face problems of analysis of either channel capacity, lossless data compression rate, rate–distortion functions, or derasurability. Often, these problems turn out to elude explicit analysis for even simple models unless we turn to the asymptotic case of sporadic erasures. In this regime, the fundamental limits are given by new information measures we introduce in this paper: erasure entropy and erasure mutual information.

## B. Erasure Entropy

The entropy of a source  $X_1, \dots, X_n$  is equal to the sum of the conditional entropies of each symbol given all preceding (or all succeeding) symbols. As the number of symbols grows without bound, the minimum compression rate converges for almost all source realizations to the limiting per-symbol entropy (entropy rate) provided that the source is stationary and ergodic.

Conditioning on the past *or* the future leads to the same measure of information content. However, what if we condition on both the past *and* the future? We define the *erasure entropy* of a collection of random variables as the sum of the individual entropies conditioned on all the other variables. The erasure entropy rate of a stationary random process is equal to the entropy of any of its values conditioned on all past and future values, or equivalently, the decrease in entropy that ensues by eliminating one value from the ensemble.

Erasure entropy rate of a stationary source is strictly lower than the conventional entropy rate (unless the source is memoryless, in which case they are identical). Regarding images or other data indexed by multidimensional sets, the representation of entropy as a sum of conditional entropies given the past requires an artificial definition of a “past,” while erasure entropy does not suffer from that drawback.

What is erasure entropy good for? Here are some illustrative applications.

- If one of the symbols in a text is erased, erasure entropy quantifies the number of bits it takes to convey the erased symbol knowing the rest of the text.
- For the simulation of a Markov random field via Gibbs sampling, each pixel value is generated by its conditional distribution given the other pixels in its Markov neighborhood, using as values for the neighboring pixels those that were generated in the previous iteration. This simulation gives, in the limit of many iterations (and of large image), a sample from the desired Markov random field (e.g., [11]). The number of random bits per pixel per iteration required for this simulation is equal to the erasure entropy of the field.
- The information content in the samples of a continuous-time Markov process is given by the entropy of a Markov chain. If the process is then sampled at twice the rate, the additional information content of the new samples is given by the erasure entropy of a Markov chain.
- The counterpart of the Markov order estimation problem in Markov random fields has been studied in [16], which finds the erasure entropy useful in the proof of consistency of an estimator of the Markov neighborhood size.
- Landauer [28] put forward the principle that the entropy of a symbol physically stored in a computer memory and the increase in the thermodynamical entropy of the overall system when the symbol is erased are equal modulo a scale factor (Boltzmann’s constant). According to that principle, when the stored information source has memory, the increase in thermodynamical entropy incurred by an erasure should be proportional to the erasure entropy rather than to the conventional Shannon entropy.

In the regime of sporadic erasures, we show in this paper that the erasure channel emerges as a convenient paradigm to

obtain Shannon-theoretic operational characterizations of erasure entropy. These characterizations are related to the minimal amount of additional (encoded) information required to recover the erased information either losslessly, almost losslessly, or within a prescribed degree of distortion.

In a follow-up to this paper, [43] studies the problem of universal estimation of erasure entropy rate.

## C. Organization

Section II introduces erasure information measures: We begin in Section II-A by establishing some basic properties of erasure entropy and erasure entropy rate, as well as examples of explicit computation. In the case of a  $k$ th-order Markov process we show the relationship between conventional entropy rate and erasure entropy rate. In Sections II-B and -C we introduce the related notions of erasure divergence and erasure mutual information, and explore some of their properties.

The basic operational characterization of erasure entropy is obtained by considering a memoryless erasure channel where the destination wants to recover the erased symbols losslessly or almost losslessly. As shown in Section III, the amount of information required per erasure is lower-bounded by the erasure entropy, a bound that becomes tight for small erasure probability.

In Section IV, we examine the setup where the erasures are allowed to be reconstructed within a certain distortion, and in particular, we analyze the tradeoff of distortion versus the amount of information per erasure that an encoder needs to provide upon observation of both the clean source and the location of the erasures. For vanishing erasure probability the fundamental limit is shown in Section IV-A to be given by the erasure rate–distortion function defined as the minimal mutual information between a source symbol and its reconstruction conditioned on all other source symbols, where the minimization is over all conditional distributions of the reconstruction symbol given the source symbol and its contexts that satisfy the average distortion constraint. In Section IV-B, we show that in the regime of sporadic erasures, the rate required to achieve a given distortion does not increase if the encoder is unaware of the location of the erasures. This surprising result is obtained by showing that the memoryless Wyner–Ziv rate distortion function is the same as in the case when both compressor and decompressor have access to the erased version of the source. In Section IV-C, we explore the counterpart of the Shannon lower bound and conditions for its tightness. The erasure rate–distortion function of a general stationary binary source under Hamming loss is given an explicit characterization in Section IV-D, and shown to admit a water-flooding interpretation. Section IV-E examines the form of the erasure rate–distortion function for a couple of additional canonical processes and fields, including Gaussian sources with Euclidean distortion and the Ising model. Section IV-F develops an upper bound on the rate distortion function of the binary symmetric Markov source in terms of its erasure rate–distortion function. Section IV-G deals with the case where the compressor has no access to the erased symbols and the nonerased symbols must be reproduced almost losslessly.

In Section V, we study the decrease in channel capacity due to erasures of the channel outputs, and show that the erasure

mutual information emerges naturally in the answer to this question in the regime of sporadic erasures. In contrast to the cases of lossless and lossy compression, where the fundamental limits in the case of sporadic nonerasures are given by the conventional Shannon entropy and rate distortion functions, in the problem of channel coding, the unavailability of most outputs leads to a fundamental limit different from the conventional maximal mutual information rate.

Relationships between erasure information measures and discrete denoising [39] are revealed in Section VI. Tight bounds on the minimum per-symbol distortion achievable by a non-causal denoiser which only has access to the discrete memoryless channel output are given in Section VI-A in terms of the erasure entropy rate. Interestingly, for a causal denoiser the same bounds hold upon replacing erasure entropy rate by Shannon's entropy rate [12]. Section VI-B shows that erasure divergence plays a role in quantifying the loss due to mismatched denoising, analogous to that shown in [31] to be played by standard divergence in quantifying the loss due to mismatched filtering (causal denoising).

## II. ERASURE INFORMATION MEASURES

In this section, we define erasure entropy, erasure divergence, and erasure mutual information. These measures coincide with the conventional quantities when the collections contain only one element (i.e.,  $n = 1$  below).

### A. Erasure Entropy

*Definition 1:* The erasure entropy of a collection of discrete random variables  $\{X_1, \dots, X_n\}$  is

$$H^-(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{\setminus i}) \quad (1)$$

where

$$X_{\setminus i} = \{X_j, j = 1, \dots, n, j \neq i\}. \quad (2)$$

Using the chain rule of entropy we can express the erasure entropy in terms of unconditional entropies as

$$H^-(X_1, \dots, X_n) = nH(X_1, \dots, X_n) - \sum_{i=1}^n H(X_{\setminus i}). \quad (3)$$

In addition, analogously to the conventional entropy, we define the erasure entropy rate as the limiting normalized erasure entropy, i.e., the limit of the arithmetic mean of the conditional entropies of each symbol given all preceding and succeeding symbols.

*Definition 2:* The erasure entropy rate of a process  $\mathbf{X} = \{X_i\}_{-\infty}^{\infty}$  is

$$H^-(\mathbf{X}) = \limsup_{n \rightarrow \infty} \frac{1}{n} H^-(X_1, \dots, X_n). \quad (4)$$

*Theorem 1:* For any collection of discrete random variables  $\{X_1, \dots, X_n\}$

$$H^-(X_1, \dots, X_n) \leq H(X_1, \dots, X_n) \quad (5)$$

with equality if and only if  $\{X_1, \dots, X_n\}$  are independent.

*Proof:*

$$\begin{aligned} & H(X_1, \dots, X_n) - H^-(X_1, \dots, X_n) \\ &= \sum_{i=1}^n [H(X_i | X_1^{i-1}) - H(X_i | X_1^{i-1}, X_{i+1}^n)] \quad (6) \end{aligned}$$

$$= \sum_{i=1}^n I(X_i; X_{i+1}^n | X_1^{i-1}) \quad (7)$$

$$\geq 0. \quad (8)$$

The condition for equality in (5) is obtained by induction using equality in (8).  $\square$

According to (7), the difference between Shannon and erasure entropy is the information that the present provides about the future knowing the past, or equivalently that the present provides about the past knowing the future, since in the proof of Theorem 1 we can interchange the future  $X_{i+1}^n$  and the past  $X_1^{i-1}$ . Denoting the reversed process  $\overleftarrow{X}_k = X_{n-k+1}$ ,  $k = 1, \dots, n$

$$H(X_1, \dots, X_n) - H^-(X_1, \dots, X_n) = I(\overleftarrow{X}^n \rightarrow X^n) \quad (9)$$

where the ‘‘directed information’’ [29] is defined as

$$I(W^n \rightarrow Y^n) = \sum_{i=1}^n I(Y_i; W_1^i | Y_1^{i-1}). \quad (10)$$

Expressed in terms of unconditional conventional entropies, the difference in (9) is referred to as the *dual total correlation* in [25].

Note that unlike entropy, erasure entropy is not associative. For example,  $H^-(X_1, X_2, X_3) \leq H^-((X_1, X_2), X_3)$ , with strict inequality unless  $X_1 - X_3 - X_2$ . Also, unlike entropy, erasure entropy is not invariant to one-to-one transformations of the collection of random variables.

*Theorem 2:* For any stationary process

$$\frac{1}{n} H^-(X_1, \dots, X_n) \leq \frac{1}{n-1} H^-(X_1, \dots, X_{n-1}). \quad (11)$$

*Proof:* The proof is quite different from the analogous result for conventional entropy [19]. According to the definition of erasure entropy

$$\begin{aligned} & \frac{1}{n} H^-(X_1, \dots, X_n) - \frac{1}{n} H(X_n | X_1^{n-1}) \\ &= \sum_{j=1}^{n-1} \left\{ \left( \frac{j-1}{n-1} - \frac{j-1}{n} \right) + \left( \frac{j}{n} - \frac{j-1}{n-1} \right) \right\} \\ & \quad \times H(X_j | X_1^{j-1} X_{j+1}^n) \quad (12) \end{aligned}$$

$$\begin{aligned} &= \sum_{j=1}^{n-2} \left( \frac{j}{n-1} - \frac{j}{n} \right) H(X_j | X_0^{j-1} X_{j+1}^{n-1}) \\ & \quad + \sum_{j=1}^{n-1} \left( \frac{j}{n} - \frac{j-1}{n-1} \right) H(X_j | X_1^{j-1} X_{j+1}^n) \quad (13) \end{aligned}$$

$$\begin{aligned} &\leq \sum_{j=1}^{n-2} \left( \frac{j}{n-1} - \frac{j}{n} \right) H(X_j|X_1^{j-1}X_{j+1}^{n-1}) \\ &\quad + \sum_{j=1}^{n-1} \left( \frac{j}{n} - \frac{j-1}{n-1} \right) H(X_j|X_1^{j-1}X_{j+1}^{n-1}) \end{aligned} \quad (14)$$

$$\begin{aligned} &= \sum_{j=1}^{n-1} \left\{ \left( \frac{j}{n-1} - \frac{j}{n} \right) + \left( \frac{j}{n} - \frac{j-1}{n-1} \right) \right\} \\ &\quad \times H(X_j|X_1^{j-1}X_{j+1}^{n-1}) \\ &\quad - \left( \frac{n-1}{n-1} - \frac{n-1}{n} \right) H(X_{n-1}|X_1^{n-2}) \end{aligned} \quad (15)$$

$$\begin{aligned} &= \frac{1}{n-1} \sum_{j=1}^{n-1} H(X_j|X_1^{j-1}X_{j+1}^{n-1}) \\ &\quad - \frac{1}{n} H(X_{n-1}|X_1^{n-2}) \end{aligned} \quad (16)$$

$$= \frac{1}{n-1} H^-(X_1, \dots, X_{n-1}) - \frac{1}{n} H(X_n|X_1^{n-1}) \quad (17)$$

$$\leq \frac{1}{n-1} H^-(X_1, \dots, X_{n-1}) - \frac{1}{n} H(X_n|X_1^{n-1}) \quad (18)$$

where

- (12)  $\Leftarrow$  the term in braces is simply  $\frac{1}{n}$ ;
- (13)  $\Leftarrow$  stationarity;
- (14)  $\Leftarrow$  dropping the conditioning variables  $X_0$  and  $X^n$ ;
- (17)  $\Leftarrow$  stationarity and the definition of erasure entropy;
- (18)  $\Leftarrow$  adding the conditioning variable  $X_1$ .  $\square$

*Theorem 3:* For any stationary process

$$H^-(\mathbf{X}) = \lim_{n \rightarrow \infty} \frac{1}{n} H^-(X_1, \dots, X_n) \quad (19)$$

$$= \lim_{k \rightarrow \infty} H(X_0|X_{-k}^{-1}, X_1^k) \quad (20)$$

$$= H(X_0|X_{-\infty}^{-1}, X_1^\infty). \quad (21)$$

*Proof:* Definition 2 and Theorem 2 imply (19). To show (20), first note that

$$\begin{aligned} \frac{1}{n} H^-(X_1, \dots, X_n) &= \frac{1}{n} \sum_{j=1}^n H(X_j|X_1^{j-1}X_{j+1}^n) \\ &\geq \frac{1}{n} \sum_{j=1}^n H(X_j|X_{j-n}^{j-1}X_{j+1}^n) \quad (22) \\ &= H(X_0|X_{-n}^{-1}X_1^n) \quad (23) \end{aligned}$$

where (22) follows by introducing further conditioning random variables and (23) follows from stationarity. The reverse inequality only holds asymptotically: Choose a positive integer  $k$ ; then for  $n > k$

$$\begin{aligned} &\frac{1}{n} H^-(X_1, \dots, X_n) \\ &= \frac{1}{n} \left\{ \sum_{j=1}^k + \sum_{j=k+1}^{n-k} + \sum_{j=n-k+1}^n \right\} H(X_j|X_1^{j-1}X_{j+1}^n) \\ &\leq \frac{2k}{n} \log |\mathcal{A}| + \frac{1}{n} \sum_{j=k+1}^{n-k} H(X_j|X_{j-k}^{j-1}X_{j+1}^n) \quad (24) \end{aligned}$$

$$= \frac{2k}{n} \log |\mathcal{A}| + \frac{n-2k}{n} H(X_0|X_{-k}^{-1}X_1^k) \quad (25)$$

where  $|\mathcal{A}|$  stands for the cardinality of the alphabet, (24) results from upper-bounding conditional entropy by the logarithm of the cardinality of the alphabet and from removing some of the conditioning random variables; and (25) follows from stationarity. Taking the limit  $n \rightarrow \infty$ , we see that  $\leq$  holds in (20). Finally, (21) follows from the bounded convergence theorem and the version of the martingale convergence theorem in [8] that implies

$$P_{X_0|X_{-k}^{-1}, X_1^k} \xrightarrow{k \rightarrow \infty} P_{X_0|X_{-\infty}^{-1}, X_1^\infty} \text{ a.s.} \quad (26)$$

$\square$

Theorem 1 implies that a collection of random variables has zero erasure entropy if it has zero entropy. The converse is of course not true: if  $X_1 = X_2$  a.s. then  $H(X_1, X_2) = H(X_1)$  whereas  $H^-(X_1, X_2) = 0$ . Similarly,  $H^-(\mathbf{X}) = 0$  does not necessarily imply  $H(\mathbf{X}) = 0$  as the following stationary ergodic example reveals.

*Example 1:* Let  $W_i$  be independent and identically distributed (i.i.d.) with positive entropy. Let  $Y_{2i} = Y_{2i+1} = W_i$ . Construct now  $\{X_i\}$  by letting  $X_i = Y_{i+U}$  where

$$P[U = 0] = P[U = 1] = \frac{1}{2}$$

and  $U$  is independent of  $\{W_i\}$ . The source  $\{X_i\}$  is stationary and ergodic with  $H(\mathbf{X}) = \frac{1}{2}H(W)$ . On the other hand,  $H^-(\mathbf{X}) = 0$  because it is possible to decide the value of  $U$  with vanishing error probability by observing a sufficiently long sample path of  $\{X_i\}$  and, for each  $i$ ,  $X_i$  is a deterministic function of  $(X_{i-1}, X_{i+1}, U)$ .

Markov chains provide simple examples of computation of erasure entropy rates.

*Example 2:* Let  $\mathbf{X}$  be a first-order homogeneous binary Markov chain with  $P_{X_1|X_0}(0|1) = P_{X_1|X_0}(1|0) = p$ . Then

$$H(\mathbf{X}) = h(p) \triangleq p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}, \quad (27)$$

and noticing that  $H(X_2|X_0) = h(2p(1-p))$ ,

$$H^-(\mathbf{X}) = h^-(p) \triangleq 2h(p) - h(2p(1-p)). \quad (28)$$

The entropy and erasure entropy of the first-order homogeneous binary symmetric Markov chain are shown in Fig. 1. It is interesting to note that in Example 2

$$\lim_{p \rightarrow 0} \frac{H(\mathbf{X})}{H^-(\mathbf{X})} = \infty.$$

The result in Example 2 can be checked by particularizing the following formula.

*Theorem 4:* For a homogeneous  $k$ th-order Markov source

$$H(\mathbf{X}) = \frac{H^-(\mathbf{X}) + H(X_1, \dots, X_k|X_{-1}, \dots, X_{-k})}{k+1}. \quad (29)$$

*Proof:* Let  $n \geq k$ . The following holds for all arguments in the state space, which we drop for the sake of brevity:

$$P_{X_0|X_{-n}^{-1}, X_1^n} P_{X_1^n|X_{-n}^{-1}} = P_{X_0^n|X_{-n}^{-1}}. \quad (30)$$

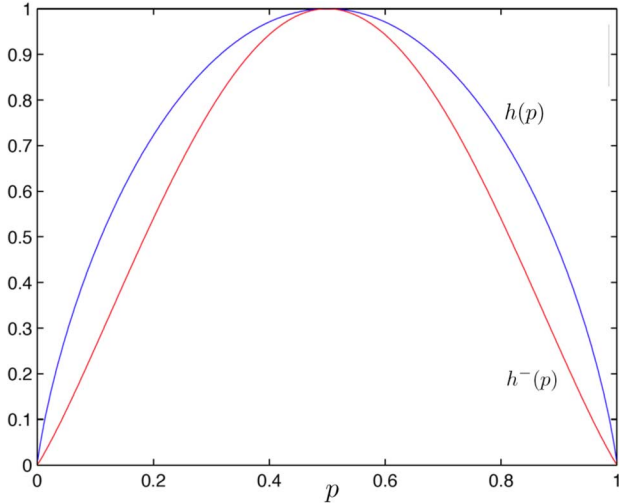


Fig. 1. Entropy rate and erasure entropy rate of a binary Markov chain with transition probability  $p$ .

Using the  $k$ th-order Markov property on both sides we can write

$$P_{X_0|X_{-n}^{-1}X_1^n}P_{X_1^k|X_{-k}^{-1}X_{k+1}^n}P_{X_{k+1}^n|X_1^k} = P_{X_{k+1}^n|X_1^k} \prod_{\ell=0}^k P_{X_\ell|X_{\ell-k}^{\ell-1}}. \quad (31)$$

Dropping the common term  $P_{X_{k+1}^n|X_1^k}$ , taking logarithms of both sides, averaging with respect to the joint distribution  $P_{X_{-n}^{-1}X_1^n}$ , and taking the limit as  $n \rightarrow \infty$ , we obtain the desired result (29) in view of Theorem 3. Note that the proof shows that for a homogeneous  $k$ th-order Markov source  $\mathbf{X}$ , if  $k < n$

$$H^-(\mathbf{X}) = H(X_0|X_{-n}^{-1}, X_1^n). \quad (32)$$

□

For a general stationary Markov random field, the erasure entropy is equal to the conditional entropy of the symbol given its neighborhood. This conditional entropy plays a role in [16].

In parallel with Definition 1, it will be useful to consider the conditional erasure entropy.

**Definition 3:** The conditional erasure entropy of a collection of discrete random variables  $\{X_1, \dots, X_n\}$  given a random object  $Z$  is

$$H^-(X_1, \dots, X_n|Z) = \sum_{i=1}^n H(X_i|X_{\setminus i}, Z). \quad (33)$$

In particular notice that

$$H^-(X_i|X_1^{i-1}) = H(X_i|X_1^{i-1}). \quad (34)$$

The following counterpart of the Shannon–Macmillan–Breiman theorem [9] is useful in the universal estimation of erasure entropy [43].

**Theorem 5:** For a stationary ergodic finite-alphabet process  $\mathbf{X}$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log 1/P_{X_i|X_{\setminus i}}(X_i|X_{\setminus i}) = H^-(\mathbf{X}) \text{ a.s.} \quad (35)$$

*Proof:* Define the functions  $f_i : \mathbb{R}^n \mapsto \mathbb{R}$

$$f_i(a, b_1^{i-1}, c_1^{n-i}) = \log 1/P_{X_0|X_{-i}^{-1}X_1^{n-i}}(a|b_1^{i-1}, c_1^{n-i})$$

and

$$f(a, b_1^\infty, c_1^\infty) = \log 1/P_{X_0|X_{-\infty}^{-1}X_1^\infty}(a|b_1^\infty, c_1^\infty).$$

By the ergodic theorem [35], it follows that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f(X_i, X_{-\infty}^{i-1}, X_{i+1}^\infty) = H^-(\mathbf{X}) \text{ a.s.} \quad (36)$$

Thus, by stationarity of  $\mathbf{X}$  we need to show that since  $\mathbf{X}$  is ergodic

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n |f_i(X_i, X_1^{i-1}, X_{i+1}^n) - f(X_i, X_{-\infty}^{i-1}, X_{i+1}^\infty)| = 0 \text{ a.s.} \quad (37)$$

The convergence in (37) can be shown using the martingale convergence theorem following the same steps as in the proof of the Shannon–Macmillan–Breiman theorem in [35, pp. 259–262]. □

## B. Erasure Divergence

For the joint distributions  $P_{X_1, \dots, X_n}$  and  $Q_{X_1, \dots, X_n}$ , divergence satisfies the chain rule

$$\begin{aligned} D(P_{X_1, \dots, X_n} \parallel Q_{X_1, \dots, X_n}) \\ = \sum_{i=1}^n D(P_{X_i|X^{i-1}} \parallel Q_{X_i|X^{i-1}} | P_{X^{i-1}}), \end{aligned} \quad (38)$$

an identity which is relevant to the following definition.

**Definition 4:** For the distributions  $P_{X_1, \dots, X_n}$  and  $Q_{X_1, \dots, X_n}$ , the erasure divergence  $D^-$  is defined in terms of the conditional divergence as

$$\begin{aligned} D^-(P_{X_1, \dots, X_n} \parallel Q_{X_1, \dots, X_n}) \\ = \sum_{i=1}^n D(P_{X_i|X_{\setminus i}} \parallel Q_{X_i|X_{\setminus i}} | P_{X_{\setminus i}}). \end{aligned} \quad (39)$$

Note that if  $Q_{X_1, \dots, X_n}$  is the distribution of i.i.d. equiprobable random variables on a finite set  $\mathcal{A}$ , then

$$D^-(P_{X_1, \dots, X_n} \parallel Q_{X_1, \dots, X_n}) = n \log |\mathcal{A}| - H^-(X_1, \dots, X_n). \quad (40)$$

In the spirit of (3), erasure divergence can be written in terms of the unconditional divergence by means of the following formula.

**Theorem 6:**

$$\begin{aligned} D^-(P_{X_1, \dots, X_n} \parallel Q_{X_1, \dots, X_n}) \\ = nD(P_{X_1, \dots, X_n} \parallel Q_{X_1, \dots, X_n}) - \sum_{i=1}^n D(P_{X_{\setminus i}} \parallel Q_{X_{\setminus i}}). \end{aligned} \quad (41)$$

*Proof:* By the telescoping property of divergence we can write for any  $i \in \{1, \dots, n\}$

$$\begin{aligned} D(P_{X_1, \dots, X_n} \parallel Q_{X_1, \dots, X_n}) \\ = D(P_{X_i|X_{\setminus i}} | Q_{X_i|X_{\setminus i}} | P_{X_{\setminus i}}) + D(P_{X_{\setminus i}} \parallel Q_{X_{\setminus i}}). \end{aligned} \quad (42)$$

Upon summing both sides with respect to  $i \in \{1, \dots, n\}$ , the result follows in view of (39).  $\square$

Erasure divergence may be strictly larger or strictly smaller than conventional divergence. For example, let  $n = 2$ ,  $P_{X_1} = Q_{X_1}$ ,  $P_{X_2} = Q_{X_2}$ ; then, using (41) we see that the erasure divergence is twice the conventional divergence. When  $X_1 = X_2$  under both  $P$  and  $Q$ , the erasure divergence is zero, while the conventional divergence is positive if  $P \neq Q$ .

The conditional erasure divergence is then defined in parallel with the unconditional case.

*Definition 5:* For the distributions  $P_Z$ ,  $P_{X_1, \dots, X_n|Z}$ , and  $Q_{X_1, \dots, X_n|Z}$ , the conditional erasure divergence is

$$D^-(P_{X_1, \dots, X_n|Z} \parallel Q_{X_1, \dots, X_n|Z} | P_Z) = \sum_{i=1}^n D(P_{X_i|X_{\setminus i}, Z} \parallel Q_{X_i|X_{\setminus i}, Z} | P_{X_{\setminus i}, Z}). \quad (43)$$

The special case of (43) where  $Q_{X_1, \dots, X_n|Z} = P_{X_1, \dots, X_n}$  is particularly important and merits Section II-C.

C. Erasure Mutual Information

*Definition 6:* The erasure mutual information between a random object  $X$  and  $(Y_1, \dots, Y_n)$  is

$$I^-(X; Y_1, \dots, Y_n) = D^-(P_{Y_1, \dots, Y_n|X} \parallel P_{Y_1, \dots, Y_n} | P_X) \quad (44)$$

$$= \sum_{i=1}^n I(X; Y_i | Y_{\setminus i}). \quad (45)$$

From the natural generalization of Theorem 6 to conditional erasure divergence, we get the following representation for erasure mutual information.

*Theorem 7:*

$$I^-(X; Y_1, \dots, Y_n) = nI(X; Y_1, \dots, Y_n) - \sum_{i=1}^n I(X; Y_i). \quad (46)$$

*Proof:* Sum for  $i \in \{1, \dots, n\}$

$$I(X; Y_1, \dots, Y_n) = I(X; Y_{\setminus i}) + I(X; Y_i | Y_{\setminus i}) \quad (47)$$

and use (45).  $\square$

Erasure mutual information may be smaller or larger than mutual information

$$I(X; Y_1, \dots, Y_n) - I^-(X; Y_1, \dots, Y_n) = \sum_{i=1}^n I(X; Y_i | Y_1^{i-1}) - I(X; Y_i | Y_{\setminus i}) \quad (48)$$

*Example 3:* If  $X = Y_1 = \dots = Y_n$ , then

$$I(X; Y_1, \dots, Y_n) = H(X) \quad (49)$$

$$I^-(X; Y_1, \dots, Y_n) = 0. \quad (50)$$

*Example 4:* If  $X = Y_1 \oplus Y_2 \oplus \dots \oplus Y_n$  where  $Y_1, \dots, Y_n$  are i.i.d. equiprobable on  $\{0, 1\}$ , then

$$I(X; Y_1, \dots, Y_n) = 1 \text{ bit} \quad (51)$$

$$I^-(X; Y_1, \dots, Y_n) = n \text{ bits}. \quad (52)$$

*Example 5:* For finite-alphabet random variables

$$I^-(Y_1, \dots, Y_n; Y_1, \dots, Y_n) = H^-(Y_1, \dots, Y_n). \quad (53)$$

*Example 6:* If  $Y_i = X + N_i$  where  $N_i, i = 1, \dots, n$  are independent Gaussian with unit variance, independent of  $X$  which is also Gaussian with variance  $\gamma$ , then

$$I(X; Y_1, \dots, Y_n) = \frac{1}{2} \log(1 + n\gamma) \quad (54)$$

$$I^-(X; Y_1, \dots, Y_n) = \frac{n}{2} \log\left(1 + \frac{\gamma}{1 + (n-1)\gamma}\right). \quad (55)$$

*Definition 7:* If the random processes  $(\mathbf{X}, \mathbf{Y})$  are jointly stationary, then the erasure mutual information rate is defined as

$$I^-(\mathbf{X}, \mathbf{Y}) = \lim_{n \rightarrow \infty} \frac{1}{n} I^-(X_1, \dots, X_n; Y_1, \dots, Y_n) = I(\mathbf{X}; Y_0 | Y_0). \quad (56)$$

*Example 7:* If  $Y_i = X_i + N_i$  where  $X_i$  and  $N_i$  are independent stationary Gaussian processes with power spectral densities  $S_X(f)$  and  $S_N(f)$ , respectively, then the erasure mutual information rate is

$$I^-(\mathbf{X}, \mathbf{Y}) = \frac{1}{2} \log\left(\frac{\int_{-1/2}^{1/2} S_N^{-1}(f) df}{\int_{-1/2}^{1/2} (S_X(f) + S_N(f))^{-1} df}\right). \quad (57)$$

If  $Y_1, \dots, Y_n$  are discrete random variables, it is easy to check that the erasure mutual information can be expressed as

$$I^-(X; Y_1, \dots, Y_n) = H^-(Y_1, \dots, Y_n) - H^-(Y_1, \dots, Y_n | X). \quad (58)$$

Note that  $I^-$  is not symmetric in its arguments. For example, if  $X_1 = X_2 = Y_1$  is binary equiprobable independent of  $Y_2$ , then  $I^-(X_1, X_2; Y_1, Y_2) = 1$  bit and  $I^-(Y_1, Y_2; X_1, X_2) = 0$ .

In parallel with the conventional mutual information, erasure mutual information satisfies the following.

*Theorem 8:* If  $(Y_1, \dots, Y_n)$  are independent, then

$$I^-(X_1, \dots, X_n; Y_1, \dots, Y_n) \geq \sum_{i=1}^n I^-(X_i; Y_i) \quad (59)$$

with equality if and only if, conditioned on  $X_i$ ,  $Y_i$  is independent of the remaining random variables, for all  $i = 1, \dots, n$ .

*Proof:* Using the chain rule of mutual information twice, we can write

$$\begin{aligned} & I(Y_i; X_i) + I(Y_i; X_{\setminus i}, Y_{\setminus i} | X_i) \\ &= I(Y_i; X_1^n, Y_{\setminus i}) \\ &= I(Y_i; Y_{\setminus i}) + I(Y_i; X_1^n | Y_{\setminus i}) \\ &= I(Y_i; X_1^n | Y_{\setminus i}) \end{aligned} \quad (60)$$

where (60) follows from the assumed independence of  $(Y_1, \dots, Y_n)$ . The difference between the left and right sides of (59) is equal to

$$\begin{aligned} I^-(X_1^n; Y_1^n) - \sum_{i=1}^n I^-(X_i; Y_i) \\ = \sum_{i=1}^n I(X_1^n; Y_i | Y_{\setminus i}) - I(X_i; Y_i) \\ = \sum_{i=1}^n I(Y_i; X_{\setminus i}, Y_{\setminus i} | X_i) \end{aligned} \quad (61)$$

where (61) follows from (60).  $\square$

*Theorem 9:* If conditioned on  $X_i$ ,  $Y_i$  is independent of the remaining random variables, for all  $i = 1, \dots, n$ , then

$$I^-(X_1, \dots, X_n; Y_1, \dots, Y_n) \leq \sum_{i=1}^n I^-(X_i; Y_i) \quad (62)$$

with equality if and only if  $(Y_1, \dots, Y_n)$  are independent.

*Proof:* Invoking the chain rule twice, we can write

$$\begin{aligned} I(Y_i; X_i | Y_{\setminus i}) + I(Y_i; Y_{\setminus i}) &= I(Y_i; X_i, Y_{\setminus i}) \\ &= I(Y_i; X_i) + I(Y_i; Y_{\setminus i} | X_i) \\ &= I(Y_i; X_i) \end{aligned} \quad (63)$$

where (63) follows from the assumption. The difference between the right and left sides of (62) is equal to

$$\begin{aligned} \sum_{i=1}^n I^-(X_i; Y_i) - I^-(X_1^n; Y_1^n) \\ = \sum_{i=1}^n I(X_i; Y_i) - I(X_1^n; Y_i | Y_{\setminus i}) \end{aligned} \quad (64)$$

$$\begin{aligned} = \sum_{i=1}^n I(X_i; Y_i) - I(Y_i; X_i | Y_{\setminus i}) \\ - I(Y_i; X_{\setminus i} | Y_{\setminus i}, X_i) \end{aligned} \quad (65)$$

$$= \sum_{i=1}^n I(X_i; Y_i) - I(Y_i; X_i | Y_{\setminus i}) \quad (66)$$

$$= \sum_{i=1}^n I(Y_i; Y_{\setminus i}) \quad (67)$$

where (66) follows from the assumption of the theorem and (67) follows from (63).  $\square$

Note that the condition in Theorem 9 and for the equality in Theorem 8 is satisfied when  $(Y_1, \dots, Y_n)$  are the output of a memoryless channel whose input is  $(X_1, \dots, X_n)$ .

In general, erasure mutual information does not satisfy a data processing property: let  $X$  be independent of  $Z_1, \dots, Z_n$  conditioned on  $Y_1, \dots, Y_n$ , then it does not follow that

$$I^-(X; Y_1, \dots, Y_n) \geq I^-(X; Z_1, \dots, Z_n). \quad (68)$$

For example, let  $X = Z_1 = Y_1 = \dots = Y_n$  and  $Z_2, \dots, Z_n$  independent of all other random variables. In this case,  $I^-(X; Y_1, \dots, Y_n) = 0$ , and  $I^-(X; Z_1, \dots, Z_n) = H(X)$ .

### III. LOSSLESS COMPRESSION

For jointly distributed processes

$$\mathbf{X} = (\dots, X_{-1}, X_0, X_1, \dots)$$

and

$$\mathbf{Z} = (\dots, Z_{-1}, Z_0, Z_1, \dots)$$

let  $H(\mathbf{X}|\mathbf{Z})$  denote the conditional entropy rate defined by

$$H(\mathbf{X}|\mathbf{Z}) = \limsup_{n \rightarrow \infty} \frac{1}{n} H(X^n | Z^n).$$

*Theorem 10:* Suppose that the source  $\mathbf{X}$  goes through a discrete memoryless erasure channel with erasure probability  $e$ , and denote the output process by  $\mathbf{Z}$ . If  $\mathbf{X}$  is stationary

$$H(\mathbf{X}|\mathbf{Z}) \geq eH^-(\mathbf{X}), \quad \text{for all } e \in [0, 1] \quad (69)$$

$$H(\mathbf{X}|\mathbf{Z}) = eH^-(\mathbf{X}) + o(e) \quad (70)$$

where  $o(x)/x \rightarrow 0$  as  $x \rightarrow 0$ .

The following lemma will be used to prove Theorem 10.

*Lemma 1:* Suppose that the source  $\mathbf{X}$  goes through a discrete memoryless erasure channel with erasure probability  $e > 0$ , and denote the output process by  $\mathbf{Z}$ . If  $\mathbf{X}$  is stationary then, for every  $k$

$$\begin{aligned} \frac{H(\mathbf{X}|\mathbf{Z})}{e} &= H(X_0 | X_{-\infty}^{-1}, X^k, Z_{k+1}^{\infty}) \\ &+ e \sum_{i=1}^k I(X_0; X_i | X_{-\infty}^{-1}, X^{i-1}, Z_{i+1}^{\infty}). \end{aligned} \quad (71)$$

*Proof:* Since

$$\begin{aligned} H(\mathbf{X}|\mathbf{Z}) &= H(X_0 | X_{-\infty}^{-1}, \mathbf{Z}) \\ &= H(X_0 | X_{-\infty}^{-1}, Z_0^{\infty}) \\ &= eH(X_0 | X_{-\infty}^{-1}, Z_1^{\infty}) \end{aligned} \quad (72)$$

we have

$$\frac{H(\mathbf{X}|\mathbf{Z})}{e} = H(X_0 | X_{-\infty}^{-1}, Z_1^{\infty}) \quad (73)$$

$$= eH(X_0 | X_{-\infty}^{-1}, Z_2^{\infty}) + (1-e)H(X_0 | X_{-\infty}^{-1}, X_1, Z_2^{\infty}) \quad (74)$$

$$= eI(X_0; X_1 | X_{-\infty}^{-1}, Z_2^{\infty}) + H(X_0 | X_{-\infty}^{-1}, X_1, Z_2^{\infty}) \quad (75)$$

proving (71) for  $k = 1$ . Proceeding by induction, assuming the validity of (71) for  $k$ , we have

$$\begin{aligned} \frac{H(\mathbf{X}|\mathbf{Z})}{e} &= H(X_0 | X_{-\infty}^{-1}, X^k, Z_{k+1}^{\infty}) \\ &+ e \sum_{i=1}^k I(X_0; X_i | X_{-\infty}^{-1}, X^{i-1}, Z_{i+1}^{\infty}) \\ &= eH(X_0 | X_{-\infty}^{-1}, X^k, Z_{k+2}^{\infty}) \\ &+ (1-e)H(X_0 | X_{-\infty}^{-1}, X^{k+1}, Z_{k+2}^{\infty}) \\ &+ e \sum_{i=1}^k I(X_0; X_i | X_{-\infty}^{-1}, X^{i-1}, Z_{i+1}^{\infty}) \end{aligned}$$

$$\begin{aligned}
 &= eI(X_0; X_{k+1}|X_{-\infty}^{-1}, X^k, Z_{k+2}^\infty) \\
 &\quad + H(X_0|X_{-\infty}^{-1}, X^{k+1}, Z_{k+2}^\infty) \\
 &\quad + e \sum_{i=1}^k I(X_0; X_i|X_{-\infty}^{-1}, X^{i-1}, Z_{i+1}^\infty) \\
 &= H(X_0|X_{-\infty}^{-1}, X^{k+1}, Z_{k+2}^\infty) \\
 &\quad + e \sum_{i=1}^{k+1} I(X_0; X_i|X_{-\infty}^{-1}, X^{i-1}, Z_{i+1}^\infty),
 \end{aligned}$$

establishing the validity of (71) for  $k + 1$ .  $\square$

*Proof:* (of Theorem 10): For a set of indices  $S$ , use  $X_S$  to denote  $\{X_i\}_{i \in S}$ . For an arbitrary realization  $Z^n = z^n$ , define  $S = S(z^n) = \{i_1, \dots, i_{|S|}\} \subset \{1, \dots, n\}$  as the collection of all the indices for which the output is erased, namely  $z_{i_j} = e$ . Then

$$H(X^n|Z^n = z^n) = H(X_S|X_{S^c}) \quad (76)$$

$$= \sum_{j=1}^{|S|} H(X_{i_j}|X_{i_1}, \dots, X_{i_{j-1}}, X_{S^c}) \quad (77)$$

$$\geq \sum_{j=1}^{|S|} H(X_{i_j}|X_{i_j}) \quad (78)$$

$$\geq |S|H^-(\mathbf{X}) \quad (79)$$

where the inequalities follow from the decrease of entropy upon further conditioning. Dividing by  $n$  and averaging with respect to  $z^n$ , we get (69) since by definition  $E[|S|] = ne$ . Note that (69) does not require that the channel be memoryless.

To prove (70), use Lemma 1 to obtain, for every  $k$

$$\frac{H(\mathbf{X}|\mathbf{Z})}{e} \leq ek \log |\mathcal{X}| + H(X_0|X_{-\infty}^{-1}, X^k). \quad (80)$$

Thus

$$\limsup_{e \rightarrow 0} \frac{H(\mathbf{X}|\mathbf{Z})}{e} \leq H(X_0|X_{-\infty}^{-1}, X^k), \quad (81)$$

completing the proof by the arbitrariness of  $k$ .  $\square$

The result in Theorem 10 holds even if the erasures have memory: The above proof is readily verified to imply (69) for any erasure channel whose expected fraction of erasures converges to  $e$ . In the Appendix, we show that to establish (70), it is enough to assume that the erasure process is stationary and that

$$\lim_{e \rightarrow 0} P[Z_1 \neq e, \dots, Z_k \neq e | Z_0 = e] = 1, \quad \text{for any } k > 0 \quad (82)$$

a condition which makes erasure bursts unlikely in the sporadic erasure regime.

If  $\mathbf{X}$  is ergodic, the Slepian–Wolf theorem [37], and its extension to stationary and ergodic sources [13], give an operational characterization for  $H(\mathbf{X}|\mathbf{Z})$ , namely, the information rate that an encoder needs to supply to the observer of  $\mathbf{Z}$  in order to recover the erased symbols almost losslessly even if the output of the channel  $\mathbf{Z}$  is not available to the encoder. Having  $\mathbf{Z}$  available to the encoder does not save any rate but enables strictly lossless recovery of the erasures. It also simplifies the achieving schemes, cf. [10].

In the regime of sporadic nonerasures, the fundamental limit is given by the entropy rate even if the erasures have memory:

*Theorem 11:* If the process  $\{\mathbf{X}\}$  is stationary, and  $\{\mathbf{Z}\}$  is the output of a stationary erasure channel with erasure rate  $e$  driven by  $\{\mathbf{X}\}$ , then

$$\lim_{e \rightarrow 1} H(\mathbf{X}|\mathbf{Z}) = H(\mathbf{X}). \quad (83)$$

*Proof:* Fix an arbitrary integer  $k$ , and define the random variables

$$V_\ell = \prod_{j=k\ell+1}^{k\ell+k} 1\{Z_j = e\}. \quad (84)$$

Note that since the channel is stationary,  $\{V_\ell\}$  are identically distributed with probabilities satisfying

$$ke \leq kP[V_\ell = 1] + (k-1)(1 - P[V_\ell = 1]) \quad (85)$$

which implies that regardless of the value of  $k$

$$\lim_{e \rightarrow 1} P[V_\ell = 1] = 1. \quad (86)$$

For convenience, restrict attention to  $n = mk$  for some integer  $m$ . We can lower-bound

$$H(X^n|Z^n) = H(X_1^k, X_{k+1}^{2k}, \dots, X_{n-k+1}^n|Z^n) \quad (87)$$

$$\geq H(\{X_{\ell k+1}^{\ell k+k} : \ell = 0, \dots, m-1, V_\ell = 1\}|Z^n) \quad (88)$$

$$\geq \sum_{\ell=0}^{m-1} P[V_\ell = 1] H(\{X_{\ell k+1}^{\ell k+k}|X_1^{\ell k}, X_{(\ell+1)k+1}^n\}) \quad (89)$$

$$\geq P[V_\ell = 1] \frac{n}{k} H(X_1^k|X_{-\infty}^0, X_{k+1}^\infty) \quad (90)$$

where (90) follows from the stationarity of  $\mathbf{X}$ . Dividing both sides by  $n$  and letting  $n \rightarrow \infty$ , we obtain

$$\lim_{e \rightarrow 1} H(\mathbf{X}|\mathbf{Z}) \geq \frac{1}{k} H(X_1^k|X_{-\infty}^0, X_{k+1}^\infty) \lim_{e \rightarrow 1} P[V_\ell = 1] \quad (91)$$

$$= \frac{1}{k} H(X_1^k|X_{-\infty}^0, X_{k+1}^\infty). \quad (92)$$

But since  $k$  is arbitrary, the left side of (91) is lower-bounded by the limit of the right side of (92), which is equal to the entropy rate.  $\square$

A general expression for  $H(\mathbf{X}|\mathbf{Z})$  as a function of  $e$  appears to be a challenging problem. For a Markov chain we have the following.

*Theorem 12:* For a general stationary first-order Markov chain, and  $0 \leq e < 1$

$$H(\mathbf{X}|\mathbf{Z}) = (1-e)S(e), \quad (93)$$

where

$$S(e) = \sum_{k=1}^{\infty} e^k H(X_1|X_0, X_{k+1}). \quad (94)$$

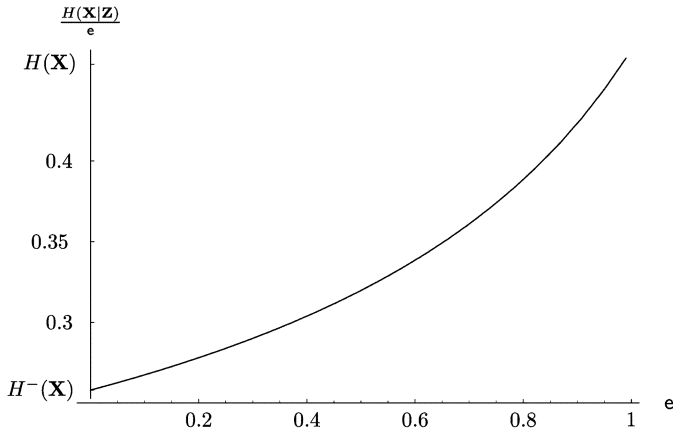


Fig. 2.  $\frac{H(\mathbf{X}|\mathbf{Z})}{e}$  as a function of erasure rate for the binary symmetric Markov chain with transition probability 0.1.  $H(\mathbf{X}) = h(0.1) = 0.469$  and  $H^-(\mathbf{X}) = h^-(0.1) = 0.258$ .

*Proof:* Using the Markov property, we can compute  $H(\mathbf{X}|\mathbf{Z})$  by considering each string of consecutive erasures separately. Since the mean length of a string of erasures is  $1/(1-e)$

$$H(\mathbf{X}|\mathbf{Z}) = (1-e)^2 \sum_{k=1}^{\infty} e^k \sum_{j=1}^k H(X_j|X_0^{j-1}, X_{k+1}) \quad (95)$$

$$= (1-e)^2 \sum_{t=1}^{\infty} e^t H(X_1|X_0, X_t) \sum_{k=t}^{\infty} e^{k-t} \quad (96)$$

$$= (1-e)S(e) \quad (97)$$

where (97) follows because the summation over  $k$  is equal to  $(1-e)^{-1}$ .  $\square$

Note that (97) extends the first-order expansion of  $H(\mathbf{X}|\mathbf{Z})$  around  $e=0$  (characterized for a general process in Theorem 10) to any number of terms when the process is first-order Markov. If  $\mathbf{X}$  is a first-order symmetric binary Markov chain with transition probability  $p$ ,  $S(e)$  is defined as in (98) at the bottom of the page. A plot of  $\frac{H(\mathbf{X}|\mathbf{Z})}{e}$  (number of required bits of description per erasure) as a function of  $e$  for the binary symmetric Markov source is given in Fig. 2. As is to be expected,  $\lim_{e \downarrow 0} \frac{H(\mathbf{X}|\mathbf{Z})}{e} = H^-(\mathbf{X})$ ,  $\lim_{e \uparrow 1} \frac{H(\mathbf{X}|\mathbf{Z})}{e} = H(\mathbf{X})$ , and  $\frac{H(\mathbf{X}|\mathbf{Z})}{e}$  is increasing with  $e$ .

Other operational characterizations of erasure entropy in the setting of lossless compression are possible as the following example illustrates.

*Example 8:* Let  $\{Y_t\}_{t \in [0, T]}$  be a random telegraph signal: a binary valued continuous-time Markov process with both transition rates equal to  $\lambda$ ; thus, the switching times are a Poisson point process of rate  $\lambda$ . Suppose  $n$  uniformly spaced samples of the signal  $\{Y_t\}_{t \in \{T/n, 2T/n, \dots, T\}}$  are to be losslessly stored. The

sampled process is a first-order symmetric binary Markov chain with transition probability

$$p = P[\text{Poisson}(\lambda T/n) \text{ is odd}] = \frac{1 - e^{-2\lambda T/n}}{2}. \quad (99)$$

Storage of  $\{Y_t\}_{t \in \{T/n, 2T/n, \dots, T\}}$  requires (for large  $n$ ) essentially  $h\left(\frac{1-e^{-2\lambda T/n}}{2}\right)$  bits/sample. Suppose now that we require a higher precision approximation of the random telegraph signal by sampling it at twice the rate to obtain  $\{Y_t\}_{t \in \{T/2n, T/n, 3T/(2n), \dots, T\}}$ . Given the knowledge of  $\{Y_t\}_{t \in \{T/n, 2T/n, \dots, T\}}$ , it is not necessary to double the storage requirements for the new  $n$  samples. It suffices to spend  $h\left(\frac{1-e^{-\lambda T/n}}{2}\right)$  bits/sample.

To conclude this section, we note that for other (nonerasure) channels, the behavior of the conditional input entropy given the output when the channel is almost noiseless can be quite different from (70). Consider, for example, a finite-alphabet discrete memoryless channel (DMC) whose transition probability matrix has the form  $\mathbf{I} - \delta \mathbf{M}$ , where  $\mathbf{M}$  is a square matrix whose rows sum to 0 and whose off-diagonal entries are nonpositive (so that  $\mathbf{I} - \delta \mathbf{M}$  is a *bona fide* channel matrix for  $\delta > 0$  sufficiently small). Then, denoting the simplex of distributions on a finite alphabet that assign to all letters probability at least  $\varepsilon$  by  $\mathcal{M}_\varepsilon$ , we have the following.

*Theorem 13:* Let  $\mathbf{X}$  be a stationary finite-alphabet process satisfying the positivity condition

$$P_{X_0|X_0} \in \mathcal{M}_\varepsilon \text{ a.s.} \quad (100)$$

for some  $\varepsilon > 0$ , and let  $\mathbf{Z}$  denote its noisy version when corrupted by the DMC with identical input/output alphabets whose channel matrix is  $\mathbf{I} - \delta \mathbf{M}$ . Then

$$\lim_{\delta \rightarrow 0} \frac{H(\mathbf{X}|\mathbf{Z})}{\delta \log \frac{1}{\delta}} = E[M(X_0, X_0)]. \quad (101)$$

The proof of Theorem 13 can be found in Appendix B. We note the following.

1. The condition (100) is rather benign, satisfied by any Markov process of any order with no restricted sequences, any dithered process that can be represented as some other process going through a DMC of positive but arbitrarily small transition probabilities, etc.
2. The limit in (101) depends on the process  $\mathbf{X}$  only through its first-order marginal.
3. The limit in (101) depends on the DMC only through the diagonal of the transition probability matrix.
4. For a binary source corrupted by a binary symmetric channel (BSC) with crossover probability  $\delta$ , (101) becomes

$$\lim_{\delta \rightarrow 0} \frac{H(\mathbf{X}|\mathbf{Z})}{\delta \log \frac{1}{\delta}} = 1. \quad (102)$$

$$S(e) = \sum_{k=1}^{\infty} e^k \left[ h\left(p \frac{1 + (1-2p)^k}{1 - (1-2p)^{k+1}}\right) \frac{1 - (1-2p)^{k+1}}{2} + h\left(p \frac{1 - (1-2p)^k}{1 + (1-2p)^{k+1}}\right) \frac{1 + (1-2p)^{k+1}}{2} \right]. \quad (98)$$

IV. LOSSY COMPRESSION

A problem that has received some interest in the literature is that of reconstruction of erased information in audio, image, and video applications. Universal algorithms have been developed based on the nonerased information and the redundancy of the source, e.g., [6]. This problem can be cast as a special case of the denoising setting, fully developed for memoryless channels in [39] (see also Section VI). Inevitably, in the absence of additional information it is only possible to restore a distorted version of the missing information. At the other extreme, the additional amount of information to recreate the erased information losslessly was explored in Section III. In this section, we explore the fundamental limits of the intermediate setting where both some additional information and some reconstruction distortion are allowed.

A. Erasure Rate–Distortion Function

As in Section III, suppose that the source  $\mathbf{X}$  goes through a discrete memoryless erasure channel with erasure probability  $e$ , and denote the output process by  $\mathbf{Z}$ . An encoder that knows the realization of  $\mathbf{X}$  and the location of the erasures wants to spend a rate  $R$  per expected erasure to obtain a distortion  $D$  under some distortion criterion.

More formally, a scheme for block length  $n$  and rate  $R$  consists of an encoder, which is a mapping  $T : \mathcal{X}^n \times \mathcal{Z}^n \rightarrow \{1, \dots, \lfloor 2^{neR} \rfloor\}$ , and a decoder, which is a sequence of mappings  $\{\hat{X}_i\}_{i=1}^n$ , where  $\hat{X}_i : \{1, \dots, \lfloor 2^{neR} \rfloor\} \times \mathcal{Z}^n \rightarrow \hat{\mathcal{X}}$ . The scheme operates as follows: the encoder maps the source and erasure sequences  $(X^n, Z^n)$  into an index  $T = T(X^n, Z^n)$ , and the decoder generates a reconstruction  $\hat{X}^n = (\hat{X}_1, \dots, \hat{X}_n)$ , where  $\hat{X}_i = \hat{X}_i(T, Z^n)$ . The performance of a scheme can be measured by its expected distortion per erased symbol according to a given distortion measure  $\rho : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathbb{R}$ .

*Definition 8:* A rate–distortion pair  $(R, D)$  is  $e$ -achievable if for every  $\varepsilon > 0$  and sufficiently large  $n$  there exists a scheme for block length  $n$  and rate  $R$  with

$$E \left[ \frac{1}{|\{1 \leq i \leq n : Z_i = e\}|} \sum_{1 \leq i \leq n : Z_i = e} \rho(X_i, \hat{X}_i) \right] \leq D + \varepsilon. \tag{103}$$

The rate–distortion function  $R_e(D)$  is the infimum of rates  $R$  such that  $(R, D)$  is  $e$ -achievable.  $R_e(D)$  is the minimum amount of information required per erasure to achieve expected distortion of  $D$  per erasure.

The setup in this subsection is one of lossy source coding with side information  $\mathbf{Z}$  available to both encoder and decoder. Hence

$$R_e(D) = \frac{1}{e} R_{X|Z}(D \cdot e) \tag{104}$$

where  $R_{X|Z}(\cdot)$  is the conditional rate–distortion function for encoding the source  $\mathbf{X}$  in the presence of side information  $\mathbf{Z}$ . The  $1/e$  and  $e$  factors in (104) are due to the fact that  $R_{X|Z}(D)$

corresponds to rate in bits per source symbol, rather than per erased symbol as in Definition 8.

*Definition 9:* For a stationary source  $\mathbf{X}$ , define the erasure rate–distortion function

$$R^-(D) = \min I(Y_0; X_0 | X_{\setminus 0}) \tag{105}$$

$$= H^-(\mathbf{X}) - \max H(X_0 | X_{\setminus 0}, Y_0) \tag{106}$$

where the optimization in (105)–(106) is over all  $P_{Y_0|X_{\setminus 0}^\infty}$  such that  $E[\rho(X_0, Y_0)] \leq D$ .

It follows immediately from the conventional rate–distortion theorem that for an i.i.d. source with distribution  $P_X$ ,  $R(P_X, D) = R_e(D) = R^-(D)$ , where

$$R(P_X, D) = \min_{\substack{P_{Y|X}: \\ E\rho(X,Y) \leq D}} I(X; Y). \tag{107}$$

For every positive integer  $k$ , define now

$$R_k(D) = \min I(Y_0; X_0 | X_{-k}^{-1}) \tag{108}$$

$$R_k^-(D) = \min I(Y_0; X_0 | X_{-k}^{-1}, X_1^k) \tag{109}$$

where the minima in (108) and (109) are, respectively, over all  $P_{Y_0|X_{-k}^0}$  and  $P_{Y_0|X_{-k}^k}$  such that  $E[\rho(X_0, Y_0)] \leq D$ . It is a well-known consequence of the martingale convergence theorem that the simplex-valued random vectors  $P_{X_0|X_{-k}^{-1}, X_1^k}$  converge in distribution to  $P_{X_0|X_{-\infty}^{-1}, X_1^\infty}$  (in fact, this follows from the almost sure convergence noted in (26)). Coupled with the uniform continuity of the mutual information  $I(X; Y)$  as a function of the distribution of  $X, Y$  (in the finite-alphabet setting), this convergence implies

$$\lim_{k \rightarrow \infty} R_k^-(D) = R^-(D). \tag{110}$$

We can now state the counterpart of Theorem 1 in the lossy case.

*Theorem 14:* Let  $R(D)$  and  $R^-(D)$  be the rate–distortion and erasure rate–distortion functions of a stationary source  $\mathbf{X}$ . Then

$$R(D) \geq R^-(D). \tag{111}$$

*Proof:* If  $S' = f(S)$  is some function of  $S$ , then the conditional rate–distortion function of  $X$  given  $S$  is upper-bounded by that of  $X$  given  $S'$  (since the latter corresponds to lossy coding based on less information at encoder and decoder). Since  $X_{-k}^{-1}$  is a function of  $(X_{-k}^{-1}, X_1^k)$ , which is a function of  $X_{\setminus 0}$ , we obtain

$$R_k(D) \geq R_k^-(D) \geq R^-(D). \tag{112}$$

To conclude the proof, it suffices to show then, by (112), that

$$R(D) \geq \lim_{k \rightarrow \infty} R_k(D) \tag{113}$$

where the limit exists since  $R_k(D)$  is monotone nonincreasing in  $k$ . Towards this end, fix  $k$  and an arbitrary sequence of rate–distortion codes of block lengths  $n > k$ , rates  $\leq R$ ,

and distortions  $\leq D$  for the source  $\mathbf{X}$ . Letting  $Y^n$  denote the reconstruction of the  $n$ -block code

$$nR \geq H(Y^n) \quad (114)$$

$$\geq I(X^n, Y^n) \quad (115)$$

$$= H(X^n) - H(X^n|Y^n) \quad (116)$$

$$= \sum_{i=1}^n H(X_i|X^{i-1}) - H(X_i|X^{i-1}, Y^n) \quad (117)$$

$$\geq \sum_{i=k+1}^n H(X_i|X^{i-1}) - H(X_i|X_{i-k}^{i-1}) + H(X_i|X_{i-k}^{i-1}) - H(X_i|X^{i-1}, Y^n) \quad (118)$$

$$= H(X^n) - H(X^k) - (n-k)H(X_0|X_{-k}^{-1}) + \sum_{i=k+1}^n H(X_i|X_{i-k}^{i-1}) - H(X_i|X^{i-1}, Y^n) \quad (119)$$

$$\geq H(X^n) - H(X^k) - (n-k)H(X_0|X_{-k}^{-1}) + \sum_{i=k+1}^n H(X_i|X_{i-k}^{i-1}) - H(X_i|X_{i-k}^{i-1}, Y_i) \quad (120)$$

$$= H(X^n) - H(X^k) - (n-k)H(X_0|X_{-k}^{-1}) + \sum_{i=k+1}^n I(X_i; Y_i|X_{i-k}^{i-1}) \quad (121)$$

$$\geq H(X^n) - H(X^k) - (n-k)H(X_0|X_{-k}^{-1}) + \sum_{i=k+1}^n R_k(E\rho(X_i, Y_i)) \quad (122)$$

$$\geq H(X^n) - H(X^k) - (n-k)H(X_0|X_{-k}^{-1}) + (n-k)R_k \left( \frac{1}{n-k} \sum_{i=k+1}^n E\rho(X_i, Y_i) \right) \quad (123)$$

where

- (119)  $\Leftarrow$  stationarity of  $\mathbf{X}$ ;
- (120)  $\Leftarrow$  data processing inequality;
- (122)  $\Leftarrow$  definition of  $R_k(D)$  (recall (108));
- (123)  $\Leftarrow$  convexity of  $R_k(D)$ .

Considering the limits of the normalized expressions on both sides of (123) implies that if  $(R, D)$  is an achievable rate–distortion pair for the source  $\mathbf{X}$  then

$$R \geq H(\mathbf{X}) - H(X_0|X_{-k}^{-1}) + R_k(D). \quad (124)$$

Hence

$$R(D) \geq H(\mathbf{X}) - H(X_0|X_{-k}^{-1}) + R_k(D) \quad (125)$$

which implies (113) when taking  $k \rightarrow \infty$ .  $\square$

The next result, which can be thought of as the analogue of Theorem 10 for the lossy case, shows that  $R^-(D)$  is the function

characterizing the best achievable rate–distortion tradeoff in the rare erasure regime.

*Theorem 15:* If  $\mathbf{X}$  is a stationary ergodic source then

$$\lim_{e \rightarrow 0} R_e(D) = R^-(D). \quad (126)$$

The proof of Theorem 15, which can be found in Appendix C, is somewhat analogous to that of Theorem 10, though the details are more cumbersome. The main idea is the following: According to (104), to prove Theorem 15 one must show that, for small  $e$ ,  $R_{X|Z}(De) \approx e \cdot R^-(D)$  or, equivalently

$$R_{X|Z}(D) \approx e \cdot R^-(D/e). \quad (127)$$

Rate–distortion theory for stationary ergodic sources implies (128) shown at the bottom of the page. As is the case with the conditional rate distortion function, the minimum in (128) can be performed greedily for every value of  $z^n$  separately. On those  $z^n$  that have about  $n \cdot e$  erasures, most of which are at a distance of at least  $k$  symbols from other erased symbols, it follows from the definitions of  $R^-$  and  $R_k^-$  that

$$eR^-(D/e) \lesssim \frac{1}{n} \min I(X^n; Y^n|Z^n = z^n) \lesssim eR_k^-(D/e). \quad (129)$$

For  $k$  fixed and large  $n$ , the remaining  $z^n$ 's have negligible probability, so we get overall

$$eR^-(D/e) \lesssim \frac{1}{n} \min I(X^n; Y^n|Z^n) \lesssim eR_k^-(D/e) \quad (130)$$

which leads to (127) in the limits of large  $n$ , then large  $k$ , and then small  $e$ .

In parallel with Theorem 11, we have the following result (whose proof follows an idea similar to that of Theorem 11).

*Theorem 16:* If the process  $\{\mathbf{X}\}$  is stationary, and  $\{\mathbf{Z}\}$  is the output of a stationary erasure channel with erasure rate  $e$  driven by  $\{\mathbf{X}\}$ , then

$$\lim_{e \rightarrow 1} R_e(D) = R(D). \quad (131)$$

### B. Erasures Unbeknownst to Encoder

The previous subsection was dedicated to the case where the erasures were not only seen by the decoder, but also known at the encoder. In this subsection, we examine the fundamental limits when the encoder does not know the location of the erasures (as would be the case in most audio, image, and video applications). Analogously to the way we defined  $R_e(D)$  in the previous subsection, define  $R_{WZ,e}(D)$  as the infimum of rates  $R$  such that  $(R, D)$  is achievable, where achievability is defined as in the previous subsection, the only difference being that here the encoder is a mapping of the form  $T = T(X^n)$  rather than  $T = T(X^n, Z^n)$ . The subscript in  $R_{WZ,e}(D)$  stands for ‘‘Wyner–Ziv,’’ which is appropriate since our setting is one of

$$R_{X|Z}(D) = \lim_{n \rightarrow \infty} \frac{1}{n} \min_{P_{Y^n|X^n, Z^n}: \sum_{i=1}^n E[1_{\{Z_i=e\}} \cdot \rho(X_i, Y_i)] \leq nD} I(X^n; Y^n|Z^n). \quad (128)$$

lossy compression with side information (the erased sequence) available at the decoder only, as considered in the seminal paper [40].

We assume throughout this subsection that  $\mathcal{X} = \hat{\mathcal{X}}$  and that the distortion measure satisfies, for all  $x, d(x, \hat{x}) \geq 0$  for all  $\hat{x}$  with equality for some  $\hat{x}$ . This assumption is by no means essential to the derivations that follow, but simplifies the statement of the results. The main result of this subsection is the following.

*Theorem 17:* If  $\mathbf{X}$  is a stationary ergodic source then

$$\lim_{\epsilon \rightarrow 0} R_{WZ, \epsilon}(D) = R^-(D). \quad (132)$$

Coupled with Theorem 15 of the previous subsection, Theorem 17 tells us that, as in the lossless setting of Section III, there is no penalty for the encoder’s ignorance of the erasure locations. Note that unlike in the lossless setting, where the absence of encoder side information has been known since the work by Slepian and Wolf [13], [37] to entail no penalty, in the lossy setting this absence has been known since the publication of [40] to, in general, entail a nonnegligible cost in the fundamental limit. It is thus surprising that in the case of sporadic erasures there is no such cost, regardless of the (stationary ergodic) source.

To get a feel for why Theorem 17 should hold, let us consider first the case of a memoryless source. For an arbitrary joint distribution  $P_{XZ}$ , Wyner and Ziv characterized in [40] the fundamental tradeoff between rate and distortion for compressing the source  $X$  based on side information  $Z$  at the decoder (more precisely, for compressing  $X^n$  with decoder side information  $Z^n$  where  $X_i, Z_i$  are i.i.d. drawings of the pair  $X, Z$ , in the limit of large  $n$ ). The Wyner–Ziv rate distortion function is given by [40]

$$\begin{aligned} R_{WZ}(D) &= \min_{Ed(X, \hat{X}(W, Z)) \leq D} I(X; W) - I(Z; W) \\ &= \min_{Ed(X, \hat{X}(W, Z)) \leq D} I(X; W|Z) \end{aligned} \quad (133)$$

where  $W - X - Z$ , and  $\hat{X}(W, Z)$  denotes the optimal estimate (in the sense of minimizing the expected distortion) of  $X$  based on  $W$  and  $Z$ . Up to now, there are only three cases where (133) is known explicitly: the doubly binary symmetric setting [40], the Gaussian setting [41], and the binary source with side information equal to the source observed in additive Gaussian noise [36]. In those cases, encoder side information is useful to lower the required rate for a given distortion. In Theorem 18, we proceed to identify a fourth case where (133) can be solved, and in which encoder side information is shown not to lower the required rate.

Letting  $R_{X|Z}(D)$  denote the conditional rate distortion function corresponding to the availability of the side information at the encoder as well, clearly

$$R_{X|Z}(D) \leq R_{WZ}(D) \quad (134)$$

where the inequality may, in general, be strict. As it turns out, however, when  $X$  and  $Z$  are the input and output of an erasure channel, there is no penalty in the fundamental rate–distortion

tradeoff for absence of the side information at the encoder. More specifically,<sup>1</sup> we have the following.

*Theorem 18:* Let  $Z$  be the output of a memoryless erasure channel with erasure probability  $\epsilon$  whose input is a memoryless source  $X$ . Then

$$R_{WZ}(D) = R_{X|Z}(D) = \epsilon R_X(D/\epsilon) \quad (135)$$

where  $R_X(D)$  is the rate–distortion function of the source  $X$ .

*Proof:* The second equality in (135) follows directly from noting that when the erasures are known to both encoder and decoder, the problem reduces to regular rate–distortion coding of the source symbols that were erased. It will thus suffice to show that  $R_{WZ}(D) = \epsilon R_X(D/\epsilon)$ :

$$R_{WZ}(D) = \min_{Ed(X, \hat{X}(W, Z)) \leq D} I(X; W|Z) \quad (136)$$

$$= \min_{\epsilon Ed(X, \hat{X}(W)) \leq D} \epsilon I(X; W) \quad (137)$$

$$= \min_{Ed(X, \hat{X}) \leq D/\epsilon} \epsilon I(X; \hat{X}) \quad (138)$$

$$= \epsilon R_X(D/\epsilon) \quad (139)$$

where (136) follows from (133) and (137) follows from the facts that  $I(X; W|Z) = \epsilon I(X; W)$  and

$$\begin{aligned} Ed(X, \hat{X}(W, Z)) &= \epsilon E[d(X, \hat{X}(W, Z))|Z = e] \\ &\quad + (1 - \epsilon) E[d(X, \hat{X}(W, Z))|Z = X] = \epsilon E[d(X, \hat{X}(W))] \end{aligned}$$

where the right-most equality is due to the optimality of  $\hat{X}$  (and our assumption on the distortion measure) which implies  $d(X, \hat{X}(W, X)) = 0$ .  $\square$

From Theorem 18 it is a small additional step to deduce that the Wyner–Ziv rate distortion function for an arbitrarily varying source, that has fraction  $p(s)$  of its components drawn (independently) from the random variable  $X_s$ , is  $\epsilon \sum_s p(s) R_{X_s}(D/\epsilon)$ . That is, for such a source too, there is no penalty for ignoring the location of the erasures at the encoder. Returning to the case of sources with memory, in the rare erasure regime, our problem is essentially one of rate distortion coding where the role of the state of the  $i$ th sequence component is played by the context  $X_{\setminus i}$ . Thus, as Theorem 17 asserts, in the regime of sporadic erasures, the compressor’s ignorance of their location does not hurt. For the formal proof of Theorem 17, which we now sketch, we utilize the informational characterization of the Wyner–Ziv rate distortion function for sources with memory.

*Sketch of Proof of Theorem 17:* Note that only the direct part

$$\limsup_{\epsilon \rightarrow 0} R_{WZ, \epsilon}(D) \leq R^-(D) \quad (140)$$

needs to be proven since the other direction is immediate from the fact that  $R_{WZ, \epsilon}(D) \geq R_\epsilon(D)$  and Theorem 15. Analogously as in (104)

$$R_{WZ, \epsilon}(D) = \frac{1}{\epsilon} R_{WZ, X|Z}(D \cdot \epsilon) \quad (141)$$

<sup>1</sup>A similar result was obtained independently and contemporaneously in [34].

where  $R_{WZ, X|Z}(\cdot)$  is the Wyner–Ziv rate–distortion function for encoding the source  $\mathbf{X}$  in the presence of side information  $\mathbf{Z}$  available only at the decoder.  $R_{WZ, X|Z}(D)$  can be expressed as (142), shown at the bottom of the page, which follows by a straightforward extension of the achievability argument in [40] to stationary ergodic sources (the converse is trivial for this “multiletter” characterization). For fixed  $k \ll 1/e \ll n$  construct  $P_{W^n|X^n}$  by letting the components of  $W_i$  be conditionally independent given  $X^n$ , with  $P_{W_i|X^n} = P_{W_i|X_{i-k}^{i+k}} = P_{Y_0|X_{-k}^k}$ , where the rightmost conditional probability is the one that achieves the minimum in the definition of  $R_k^-(D)$ . Under this  $P_{W^n|X^n}$ , on those  $z^n$  that have about  $n \cdot e$  erasures, most of which are at a distance of at least  $k$  symbols from other erased symbols, we have

$$\frac{1}{n} I(X^n; W^n | Z^n = z^n) \lesssim e R_k^-(D) \quad (143)$$

where  $\lesssim$  is an approximate inequality that becomes increasingly precise as  $e \rightarrow 0$ . Since the remaining  $z^n$ 's have negligible probability, we get overall

$$\frac{1}{n} I(X^n; W^n | Z^n) \lesssim e R_k^-(D). \quad (144)$$

As for the distortion, the “symbol-by-symbol” reconstruction mapping

$$\hat{X}_i(W_i, Z_i) = \begin{cases} Z_i, & \text{if } Z_i \neq e \\ W_i, & \text{if } Z_i = e \end{cases}$$

would give

$$E\rho(X^n, \hat{X}^n(W^n, Z^n)) \approx eD, \quad (145)$$

implying, when combined with (144), that  $R_{WZ, X|Z}(eD) \lesssim e R_k^-(D)$  or equivalently

$$\frac{1}{e} R_{WZ, X|Z}(eD) \lesssim R_k^-(D). \quad (146)$$

The combination of (141) and (146) implies

$$R_{WZ, e}(D) \lesssim R_k^-(D) \quad (147)$$

i.e.,

$$\limsup_{e \rightarrow 0} R_{WZ, e}(D) \leq R_k^-(D) \quad (148)$$

implying (140) by the arbitrariness of  $k$ .  $\square$

The formal proof of Theorem 17 makes the foregoing arguments precise and proceeds in a path parallel to that of the proof of Theorem 15 in Appendix C, where the limit in (128) is replaced any the limit (142).

To conclude this subsection we note that, similarly to Theorem 16 (and in fact as a corollary of it), if the process  $\{\mathbf{X}\}$

is stationary, and the erasure channel is stationary with erasure rate  $e$ , then

$$\lim_{e \rightarrow 1} R_{WZ, e}(D) = R(D). \quad (149)$$

### C. Shannon Lower Bound for $R^-(D)$

For simplicity, assume here that  $\mathcal{X}$  is either finite or  $\mathcal{X} = \mathbb{R}$ . Also assume  $\hat{\mathcal{X}} = \mathcal{X}$  and that  $d$  is a difference distortion measure (i.e.,  $d(x, \hat{x}) = d(x - \hat{x})$ ), where, when  $\mathcal{X}$  is finite, addition and subtraction of elements is modulo the size of the alphabet (for some assumed ordering of the elements of  $\mathcal{X}$ ). This is the setting in which the Shannon lower bound (SLB) applies. The SLB (e.g., [4]) states that for any stationary and ergodic process  $\mathbf{X}$

$$R(D) \geq H(\mathbf{X}) - \phi(D) \quad (150)$$

where  $\phi(D)$  is the maximum-entropy function defined by

$$\phi(D) = \max_{N: E[d(N)] \leq D} H(N), \quad (151)$$

the maximization being over random variables  $N$  taking values in  $\mathcal{X}$  (and  $H(N)$  stands for differential entropy if  $\mathcal{X}$  is not countable). Note the concavity of  $\phi$ , which is a consequence of the concavity of entropy. Equality in (150) holds if and only if  $\mathbf{X}$  has the decomposition

$$X_i = Y_i + N_i \quad (152)$$

where  $\mathbf{N}$  is an i.i.d. process with components achieving the maximum in (151), independent of the process  $\mathbf{Y}$ .

We now proceed to develop a parallel bound for  $R^-(D)$ . To this end, let  $N_D$  denote the achiever of the maximum in (151) and define the set of all distributions that can be attained at the output of an additive-noise channel with noise  $N_D$  for some input distribution

$$\mathcal{S}(D) = \{P \in \mathcal{M}(\mathcal{X}) : \exists P_Y \in \mathcal{M}(\mathcal{X}) \text{ s.t. } P = P_Y * P_{N_D}\}, \quad (153)$$

where  $*$  denotes (discrete—when  $\mathcal{X}$  is discrete) convolution. The SLB for  $R^-(D)$  is given by the following.

*Theorem 19:* For any stationary source  $\mathbf{X}$

$$R^-(D) \geq H^-(\mathbf{X}) - \phi(D) \quad (154)$$

with equality if and only if

$$P_{X_0|X_{\setminus 0}} \in \mathcal{S}(D) \text{ a.s.} \quad (155)$$

Note that whenever the source  $\mathbf{X}$  has the decomposition in (152), it certainly satisfies (155), since any distribution satisfying (152) satisfies also

$$P_{X_0|X_{\setminus 0}} = P_{Y_0|X_{\setminus 0}} * P_{N_d} \text{ a.s.} \quad (156)$$

$$R_{WZ, X|Z}(D) = \lim_{n \rightarrow \infty} \frac{1}{n} \min_{P_{W^n|X^n}: E\rho(X^n, \hat{X}^n(W^n, Z^n)) \leq D} I(X^n; W^n | Z^n) \quad (142)$$

Thus, Theorem 19 implies the following observation.

*Observation 1:* When the SLB for  $R(D)$  holds with equality, the SLB for  $R^-(D)$  holds with equality as well.

The converse of Observation 1 does not hold. As examples below show, the SLB for  $R^-(D)$  may hold with equality even when the SLB for  $R(D)$  does not. Furthermore, there are many cases (cf. [21], [22]) where the SLB for  $R(D)$  is known to hold with equality for a distortion region of the form  $D \leq D^*$ . On the other hand, in such cases the threshold value for tightness of the SLB for  $R^-(D)$  (which by Observation 1 is larger than  $D^*$ ) is explicitly characterizable, as in examples to follow.

*Proof of Theorem 19:* Under any  $P_{Y_0|X}$  such that  $E d(X_0 - Y_0) \leq D$  we have

$$I(X_0; Y_0|X_{\setminus 0}) = H^-(\mathbf{X}) - H(X_0|X_{\setminus 0}, Y_0) \tag{157}$$

$$= H^-(\mathbf{X}) - H(X_0 - Y_0|X_{\setminus 0}, Y_0) \tag{158}$$

$$\geq H^-(\mathbf{X}) - H(X_0 - Y_0|X_{\setminus 0}) \tag{159}$$

$$= H^-(\mathbf{X}) - \int H(X_0 - Y_0|x_{\setminus 0})dP(x_{\setminus 0}) \tag{160}$$

$$\geq H^-(\mathbf{X}) - \int \phi(E[d(X_0 - Y_0)|x_{\setminus 0}])dP(x_{\setminus 0}) \tag{161}$$

$$\geq H^-(\mathbf{X}) - \phi(E[d(X_0 - Y_0)]) \tag{162}$$

$$\geq H^-(\mathbf{X}) - \phi(D) \tag{163}$$

where

- (159)  $\Leftarrow$  conditioning reduces entropy;
- (161)  $\Leftarrow$  definition of  $\phi$ ;
- (162)  $\Leftarrow$  concavity;
- (163)  $\Leftarrow$  monotonicity.

This proves (154). To prove the condition for equality assume without loss of generality that  $D$  is a point of increase of  $\phi$ , i.e., that  $\phi(D') < \phi(D)$  for  $D' < D$ .<sup>2</sup> Inequalities (161)–(163) are then seen to hold with equality if and only if

$$P_{X_0 - Y_0|x_{\setminus 0}} = P_{ND}, \quad \text{for } P_{X_{\setminus 0}} - \text{almost every } x_{\setminus 0}. \tag{164}$$

Inequality (159) holds with equality if and only if the Markov relationship  $(X_0 - Y_0) - X_{\setminus 0} - Y_0$  holds or, equivalently, if

$$X_0 - Y_0 \text{ and } Y_0 \text{ are independent given } X_{\setminus 0}. \tag{165}$$

By definition of  $\mathcal{S}(D)$ , the existence of a conditional distribution  $P_{Y_0|X}$  under which both (164) and (165) hold is equivalent to the requirement that the conditional distribution  $P_{X_0|x_{\setminus 0}}$  belong to  $\mathcal{S}(D)$  for  $P_{X_{\setminus 0}}$ -almost every  $x_{\setminus 0}$ .  $\square$

### D. $R^-(D)$ for Binary Sources and Hamming Distortion

Consider  $R^-(D)$  for a binary source, under Hamming loss. For  $p \in [0, 1]$ , let  $R_b(p, D)$  denote the rate distortion function of the Bernoulli( $p$ ) source

$$R_b(p, D) = \max\{h(p) - h(D), 0\}. \tag{166}$$

The following theorem presents  $R^-(D)$  explicitly in parametric form.

<sup>2</sup>To see that this entails no loss of generality note that  $D$  does not satisfy the increase requirement occurs only in the finite alphabet setting, only for  $D > d_{\max} = \max_x d(x)$ , for which the assertion of the theorem holds trivially.

*Theorem 20:* Let  $\mathbf{X}$  be a binary stationary source and define the  $[0, 1/2]$ -valued random variable

$$U = \min\{P(X_0 = 1|X_{\setminus 0}), P(X_0 = 0|X_{\setminus 0})\}. \tag{167}$$

The curve  $R^-$  for the source  $\mathbf{X}$  is given in parametric form by

$$D = D(\Delta) = E[\min\{U, \Delta\}] \tag{168}$$

and

$$R = R(\Delta) = E[R_b(U, \Delta)] = E[\max\{h(U) - h(\Delta), 0\}] \tag{169}$$

where the parameter  $\Delta$  varies in  $[0, 1/2]$ .

Note that the representation given in the theorem is amenable to a water-flooding interpretation as follows: Writing explicitly

$$U(x_{\setminus 0}) = \min\{P(X_0 = 1|x_{\setminus 0}), P(X_0 = 0|x_{\setminus 0})\} \tag{170}$$

the parametric representation in the theorem can equivalently be given as

$$R^-(D) = \int [h(U(x_{\setminus 0})) - h(D_{x_{\setminus 0}})]dP_{X_{\setminus 0}}(x_{\setminus 0}) \tag{171}$$

where  $D_{x_{\setminus 0}}$ , the ‘‘distortion spectrum,’’ is given by

$$D_{x_{\setminus 0}} = \begin{cases} \Delta, & \text{if } U(x_{\setminus 0}) \geq \Delta \\ U(x_{\setminus 0}), & \text{otherwise} \end{cases} \tag{172}$$

and where we choose the ‘‘water level’’  $\Delta$  so that the total distortion is  $D$ :

$$D = \int D_{x_{\setminus 0}}dP_{X_{\setminus 0}}(x_{\setminus 0}). \tag{173}$$

Note that here  $D$ , the ‘‘volume’’ of the water at level  $\Delta$ , is obtained as a weighted sum, according to  $P_{X_{\setminus 0}}$ . This water flooding interpretation is illustrated in Fig. 3. The ‘‘spectrum’’ of the process is the collection  $\{U(x_{\setminus 0})\}_{x_{\setminus 0}}$ , where the higher the spectral value  $U(x_{\setminus 0})$  the higher the source entropy at the context  $x_{\setminus 0}$  (and the more rate will be required to cover it to within a given distortion). Regarding the water flooding interpretation, we make the following remarks.

1.  $D(\Delta) = E[U]$  and  $R(\Delta) = 0$  for  $\Delta \geq \text{esssup } U$ .<sup>3</sup> Thus,  $\Delta$  need only be varied in  $[0, \text{esssup } U]$  to obtain the whole  $R^-$  curve.
2. Theorem 19, applied to the binary case, implies that the SLB for  $R^-(D)$  is tight if and only if  $D \in [0, \text{essinf } U]$ . This is consistent with Theorem 20 as, for  $0 \leq \Delta \leq \text{essinf } U$ ,  $D(\Delta) = \Delta$  and  $R(\Delta) = H^-(\mathbf{X}) - h(\Delta)$ .
3. When  $\mathbf{X}$  is a  $k$ th-order Markov source,  $U$ , as defined in (167), is discrete, assuming at most  $2^{2k}$  different values. The characterization in Theorem 20 gives  $R^-(D)$  explicitly for any such source (cf. Example 9 below). This is in contrast to the case for  $R(D)$ , which is not explicitly known even for the binary symmetric first-order Markov process [22].
4. When  $\mathbf{X}$  is not Markov, e.g., a hidden Markov process,  $U$  may have no point masses and, in fact, have a singular distribution. In such cases, the distribution of  $U$  can be arbitrarily precisely approximated by expressing it as a solution to an integral equation, of the type obtained by Black-

<sup>3</sup>The fact that  $R^-(D) = 0$  for  $D \geq E[U]$  is to be expected since an error rate equal to  $E[U]$  can be achieved by reconstructing the erased information solely on the basis of its context, with no additional encoded information.

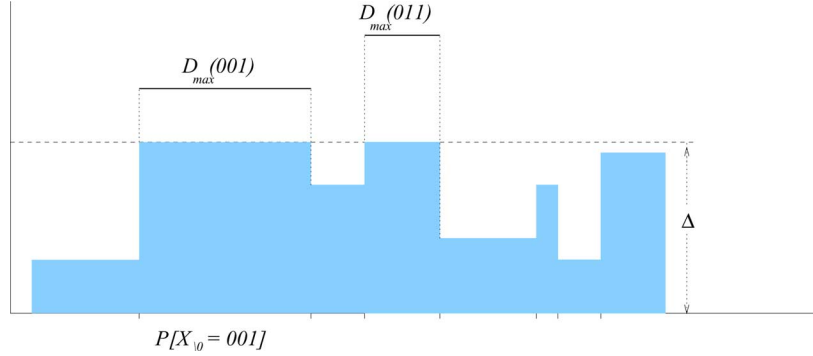


Fig. 3. The water flooding solution for  $R^-(D)$  of a binary source. The  $x$ -axis presents the different effective values of  $x_{\setminus 0}$  (this example may correspond to a symmetric second-order Markov sources whose 16 contexts are reduced to 8).

well in [7]. This then leads to arbitrarily precise approximations for  $R^-(D)$  via the characterization in Theorem 20.

*Proof of Theorem 20:* Note first that for any  $\Delta \in [0, 1/2]$ , both  $\min\{u, \Delta\}$  and  $R_b(u, \Delta)$  are bounded and continuous functions of  $0 \leq u \leq 1/2$ . It follows that, for an arbitrarily distributed  $U \in [0, 1/2]$ , there exists a sequence of discrete, finite-alphabet, random variables  $U_n \in [0, 1/2]$ , such that both  $E[\min\{U_n, \Delta\}] \rightarrow E[\min\{U, \Delta\}]$  and  $E[R_b(U_n, \Delta)] \rightarrow E[R_b(U, \Delta)]$ . It will thus suffice to prove the assertion assuming  $U$  is discrete, with a finite support. Assume then that  $U$  is distributed as

$$U = u_i \text{ w.p. } p_i, \quad 1 \leq i \leq n \quad (174)$$

where  $0 \leq u_i \leq 1/2$ , and  $p_i > 0$  with  $\sum_{i=1}^n p_i = 1$ . It follows from the first part of Lemma 3, similarly as in Observation 2 (Appendix), that

$$R^-(D) = \min \sum_{i=1}^n R_b(u_i, D_i) p_i \quad (175)$$

where the minimum is over all  $\{D_i\}_{i=1}^n$  satisfying

$$\sum_{i=1}^n D_i p_i = D. \quad (176)$$

Since  $R_b(u_i, D_i) = \max\{h(p) - h(D_i), 0\}$ , we obtain, equivalently,

$$R^-(D) = \min_{\{D_i\}_{i=1}^n: \sum_{i=1}^n D_i p_i = D, D_i \leq u_i} \sum_{i=1}^n [h(u_i) - h(D_i)] p_i. \quad (177)$$

The Lagrange multiplier functional is

$$J(D_1, \dots, D_n) = \sum_{i=1}^n [h(u_i) - h(D_i)] p_i + \lambda \left( \sum_{i=1}^n D_i p_i \right) \quad (178)$$

from which the Kuhn–Tucker conditions are readily checked to be satisfied by

$$D_i = \min\{u_i, \Delta\}, \quad (179)$$

where  $\Delta$  is chosen so that (176) holds. In other words, the relation between  $D$  and  $\Delta$  is given by

$$D = \sum_{i=1}^n D_i p_i = \sum_{i=1}^n p_i \min\{u_i, \Delta\} = E[\min\{U, \Delta\}]. \quad (180)$$

On the other hand, under  $\{D_i\}_{i=1}^n$  of (179)

$$\begin{aligned} \sum_{i=1}^n R_b(u_i, D_i) p_i &= \sum_{1 \leq i \leq n: u_i \leq \Delta} R_b(u_i, D_i) p_i \\ &\quad + \sum_{1 \leq i \leq n: u_i > \Delta} R_b(u_i, D_i) p_i \\ &= \sum_{1 \leq i \leq n: u_i \leq \Delta} R_b(u_i, u_i) p_i \\ &\quad + \sum_{1 \leq i \leq n: u_i > \Delta} R_b(u_i, \Delta) p_i \\ &= \sum_{1 \leq i \leq n: u_i \leq \Delta} R_b(u_i, \Delta) p_i \\ &\quad + \sum_{1 \leq i \leq n: u_i > \Delta} R_b(u_i, \Delta) p_i \quad (181) \\ &= \sum_{i=1}^n R_b(u_i, \Delta) p_i \\ &= E[R_b(U, \Delta)] \quad (182) \end{aligned}$$

where (181) follows since  $R_b(u_i, u_i) = R_b(u_i, \Delta)$  for  $u_i \leq \Delta$ .  $\square$

*Example 9:* Consider the binary symmetric Markov source, as in Example 2, with transition probability  $p \in [0, 1]$ . Let  $p_{\min} = \min\{p, 1 - p\}$ . In this case,  $U$  in (167) is distributed as

$$U = \begin{cases} \frac{p_{\min}^2}{p^2 + (1-p)^2}, & \text{w.p. } p^2 + (1-p)^2 \\ 1/2, & \text{w.p. } 2p(1-p). \end{cases} \quad (183)$$

Consequently,  $D(\Delta)$  in (168) is given by

$$D(\Delta) = \begin{cases} \Delta, & \text{for } 0 \leq \Delta < \frac{p_{\min}^2}{p^2 + (1-p)^2} \\ p_{\min}^2 + \Delta 2p(1-p), & \text{for } \frac{p_{\min}^2}{p^2 + (1-p)^2} \leq \Delta < 1/2 \\ p_{\min}, & \text{for } \Delta = 1/2 \end{cases} \quad (184)$$

while  $R(\Delta)$  in (169) is given by

$$R(\Delta) = \begin{cases} h^-(p_{\min}) - h(\Delta), & \text{for } 0 \leq \Delta < \frac{p_{\min}^2}{p^2 + (1-p)^2} \\ 2p(1-p)[1 - h(\Delta)], & \text{for } \frac{p_{\min}^2}{p^2 + (1-p)^2} \leq \Delta < 1/2 \\ 0, & \text{for } \Delta = 1/2. \end{cases} \quad (185)$$

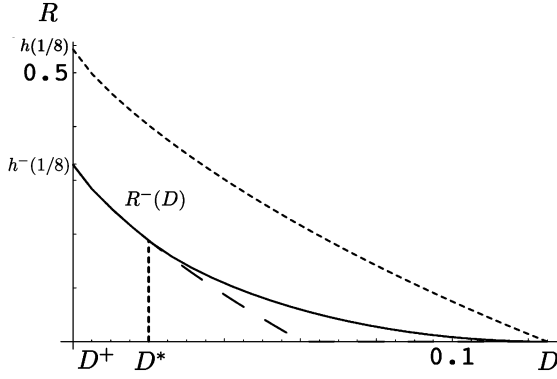


Fig. 4. Binary symmetric Markov source with  $p_{\min} = 1/8$ . Solid curve is  $R^-(D)$ ; Dashed curve is the SLB for  $R^-(D)$ . Dotted curve is the rate-distortion function of the Bernoulli(1/8) source, which is the SLB of the binary symmetric Markov source with  $p_{\min} = 1/8$ .

Solving for  $R$  as a function of  $D$  gives (186) at the bottom of the page. A plot of  $R^-(D)$  for the case  $p_{\min} = 1/8$  is given in Fig. 4. The SLB for  $R^-(D)$  is tight up to  $D^* = \frac{p_{\min}^2}{p^2 + (1-p)^2} = 0.02$ , whereas the SLB for  $R(D)$  is tight only up to

$$D^+ = \frac{1}{2} \left( 1 - \sqrt{1 - ((1/8)/(7/8))^2} \right) \approx 0.0051$$

(cf. [21]). In this example, erasure entropy and entropy are  $h^-(1/8) \approx 0.329$  and  $h(1/8) \approx 0.544$ , respectively. Note that, in agreement with Observation 1,  $D^+ < D^*$ .  $R(D)$  for this source is not explicitly known.

E.  $R^-(D)$  for Other Sources

1)  $R^-(D)$  for Gaussian Sources: Let  $\mathbf{X}$  be a stationary Gaussian process with a bounded and strictly positive spectral density  $S_{\mathbf{X}}(e^{jw})$ . Then

$$R^-(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma_x^2}{D}, & \text{for } 0 \leq D \leq \sigma_x^2 \\ 0, & D > \sigma_x^2 \end{cases} \quad (187)$$

where

$$\sigma_x^2 = \left[ \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{dw}{S_{\mathbf{X}}(e^{jw})} \right]^{-1}. \quad (188)$$

To see why this follows from Theorem 19 note that in this case  $X_0$ , conditioned on  $X_{\setminus 0}$ , is, with probability one, Gaussian with variance  $\sigma_x^2$ . So, in particular,  $P_{X_0|X_{\setminus 0}} \in \mathcal{S}(D)$  a.s. for every  $D \leq \sigma_x^2$ . Thus,  $R^-(D)$  satisfies the SLB with equality in the whole range of positive rates.

In comparison, the rate-distortion function in this case is well known to be given by water-pouring (cf, e.g., [5]) and satisfies

$$R(D) \begin{cases} = \frac{1}{2} \log \frac{\sigma_x^2}{D}, & \text{for } 0 \leq D \leq D^* \\ > \frac{1}{2} \log \frac{\sigma_x^2}{D}, & D > D^* \end{cases} \quad (189)$$

where

$$\sigma_x^2 = \exp \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln S_{\mathbf{X}}(e^{jw}) dw \right\} \quad (190)$$

and  $D^*$ , the point until which the SLB is tight, is given by  $D^* = \min_w S_{\mathbf{X}}(e^{jw})$ . Indeed, in compliance with Observation 1

$$\sigma_x^2 = \left[ \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{dw}{S_{\mathbf{X}}(e^{jw})} \right]^{-1} \geq \min_w S_{\mathbf{X}}(e^{jw}) = D^*. \quad (191)$$

Further, equality in (191) holds if and only if  $S_{\mathbf{X}}(e^{jw})$  is constant with frequency, implying that, for Gaussian sources, the curves  $R(D)$  and  $R^-(D)$  are identical if and only if  $\mathbf{X}$  is memoryless.

2)  $R^-(D)$  for Random Fields: The foregoing framework and results for  $R^-(D)$  carry over to the case where  $\mathbf{X}$  is a stationary (shift invariant) random field (with index set  $\mathbb{Z}^d$ ). Markov random fields (MRFs) can be characterized by the set of conditional distributions  $\{P_{X_i|x(N(i))}\}_{x(N(i))}$ , where  $N(i)$  is the neighborhood of  $i$ , i.e., the smallest subset of  $\mathbb{Z}^d$ , that does not contain  $i$ , for which the Markov relation  $X_i - X(N(i)) - X(\mathbb{Z}^d \setminus (N(i) \cup \{i\}))$  holds.

Example 10: Let  $\mathbf{X}$  be a shift-invariant binary MRF specified by  $\{P_{X_0|x(N(0))}\}_{x(N(0))}$ . Then Theorem 19 implies that  $R^-(D)$ , under Hamming loss, is given by  $H^-(\mathbf{X}) - h(D)$  for all

$$0 \leq D \leq \text{essinf} \min\{P_{X_0=1|X(N(0))}, P_{X_0=0|X(N(0))}\}$$

and is strictly larger than  $H^-(\mathbf{X}) - h(D)$  for larger  $D$ .

A key point to note is that, when the MRF satisfies the benign (and easily verifiable) positivity condition  $P(x(N(0))) > 0$  for all  $x(N(0))$

$$\text{essinf} \min\{P_{X_0=1|X(N(0))}, P_{X_0=0|X(N(0))}\} = \inf_{x(N(0))} \min\{P_{X_0=1|x(N(0))}, P_{X_0=0|x(N(0))}\},$$

where the right-hand side depends only on  $\{P_{X_0|x(N(0))}\}_{x(N(0))}$ , and not on the probabilities  $P(x(N(0)))$ .<sup>4</sup> Thus, the threshold for the tightness of the SLB for  $R^-(D)$  is explicitly obtained for any MRF. This is in contrast to  $R(D)$ , whose threshold for the tightness of the SLB is not known even for the simplest binary MRFs, cf. [24], [42]. Furthermore, the explicit form of  $R^-(D)$  at distortions larger than the SLB threshold (and thus at all distortions) can, in principle, be obtained via the prescription in Theorem 20. This prescription requires the distribution of  $U = \min\{P_{X_0=1|X(N(0))}, P_{X_0=0|X(N(0))}\}$  which is un-

<sup>4</sup>Typically, a stationary MRF is given in terms of its specification  $\{P_{X_0|x(N(0))}\}_{x(N(0))}$ , but the probabilities  $P(x(N(0)))$  are hard to compute and, in fact, known only for very few MRFs.

$$R^-(D) = \begin{cases} h^-(p_{\min}) - h(D), & \text{for } 0 \leq D \leq \frac{p_{\min}^2}{p^2 + (1-p)^2} \\ 2p(1-p) \left[ 1 - h\left(\frac{D - p_{\min}^2}{2p(1-p)}\right) \right], & \frac{p_{\min}^2}{p^2 + (1-p)^2} < D \leq p_{\min} \\ 0 & \text{otherwise.} \end{cases} \quad (186)$$

fortunately known for very few MRFs, but can be readily approximated to yield approximations of  $R^-(D)$ . We next illustrate this for the Ising model with no external field.

#### F. $R^-(D)$ for the Ising Model With No External Field

Consider the Ising model on  $\mathbb{Z}^2$ , with no external field [20], [23]. The energy function is of the form

$$E = -\beta \sum_{\langle i,j \rangle} x_i x_j \quad (192)$$

where  $x_i \in \{-1, 1\}$  and the summation is over nearest neighbor pairs. We wish to obtain  $R^-(D)$  (under Hamming loss) for this field. Symmetry implies that, for this field, the random variable  $U = \min\{P_{X_0=1|X(N(0))}, P_{X_0=0|X(N(0))}\}$  assumes one of three possible values, according to whether all the sites in the (four nearest neighbor) neighborhood  $X(N(0))$  are the same, one differs from the remaining three, or two have one value and two the other. Specifically

$$U = \begin{cases} \frac{e^{-4|\beta|}}{e^{-4|\beta|} + e^{4|\beta|}}, & \text{w.p. } p_1(\beta) \\ \frac{e^{-2|\beta|}}{e^{-2|\beta|} + e^{2|\beta|}}, & \text{w.p. } p_2(\beta) \\ \frac{1}{2}, & \text{w.p. } 1 - p_1(\beta) - p_2(\beta) \end{cases} \quad (193)$$

where  $p_1(\beta)$  is the probability that all the sites of  $X(N(0))$  share the same value and  $p_2(\beta)$  is the probability that one site in  $X(N(0))$  differs from the remaining three. It is now a direct application of the waterpouring characterization in Theorem 20 to deduce  $R_{\beta}^-(D)$  shown in (194) at the bottom of the page, where we let

$$\alpha_1(\beta) = \frac{e^{-4|\beta|}}{e^{-4|\beta|} + e^{4|\beta|}} \quad \text{and} \quad \alpha_2(\beta) = \frac{e^{-2|\beta|}}{e^{-2|\beta|} + e^{2|\beta|}}$$

and  $H^-(\beta)$  denotes the erasure entropy of this field, given explicitly by

$$H^-(\beta) = p_1(\beta)h(\alpha_1(\beta)) + p_2(\beta)h(\alpha_2(\beta)) + 1 - p_1(\beta) - p_2(\beta).$$

Note that this provides  $R_{\beta}^-(D)$  in closed form, up to  $p_1(\beta)$  and  $p_2(\beta)$  which are unknown explicitly but can be numerically approximated for any value of  $\beta$  to arbitrary precision using, e.g., Markov chain Monte Carlo methods [3].

#### G. Upper Bound on $R(D)$ Via $R^-(D)$ for Markov Source

Consider the following lossy compression scheme for a first-order Markov source: lossless compression of every other source symbol, and then rate-distortion coding of the remaining

half of the source symbols. Assuming both the lossless and lossy parts of this scheme are done optimally, achieving overall distortion  $D$  with this scheme requires a rate

$$R_{\text{ub}}(D) = \frac{1}{2} [R^-(2D) + H(X_2|X_0)] \quad \text{for } 0 \leq D \\ \leq \frac{1}{2} E \left[ \min_{\hat{x}} E[\rho(X_0, \hat{x})|X_{-1}, X_1] \right] \quad (195)$$

where the subscript in  $R_{\text{ub}}(D)$  signifies that this is an upper bound on the rate-distortion function of the source. Note that the distortion level

$$D = \frac{1}{2} E \left[ \min_{\hat{x}} E[\rho(X_0, \hat{x})|X_{-1}, X_1] \right]$$

corresponds to zero rate coding of the second subsequence. Thus, the suggested scheme is not relevant for distortion values exceeding that level. Of course, higher distortion working points can be achieved by time-sharing with other schemes, such as the trivial zero-rate scheme. Note also that  $R^-(0) = H(X_0|X_{-1}, X_1)$  and hence

$$R_{\text{ub}}(0) = \frac{1}{2} [H(X_0|X_{-1}, X_1) + H(X_2|X_0)] \\ = \frac{1}{2} [H(X_0|X_{-1}, X_1) + H(X_1|X_{-1})] \\ = \frac{1}{2} H(X_0, X_1|X_{-1}) \\ = \frac{1}{2} [H(X_0|X_{-1}) + H(X_1|X_0, X_{-1})] \\ = \frac{1}{2} [H(X_0|X_{-1}) + H(X_1|X_0)] \\ = H(X_1|X_0) \\ = H(\mathbf{X}). \quad (196)$$

Evidently, the suggested scheme is optimal in the low-distortion limit.

For a concrete example, consider the binary symmetric Markov source, under Hamming loss, for which (195), which we denote below by  $R_{\text{ub}}(D)$ , assumes the form

$$R_{\text{ub}}(D) = \frac{1}{2} [R^-(2D) + h(2p(1-p))] \quad (197)$$

$$= \frac{1}{2} [h^-(p_{\min}) - h(2D) + h(2p(1-p))] \quad (198)$$

$$= h(p) - \frac{1}{2} h(2D) \quad (199)$$

$$= h(p) - D \log 1/D \\ + o(D \log 1/D) \quad \text{as } D \rightarrow 0 \quad (200)$$

$$R_{\beta}^-(D) = \begin{cases} H^-(\beta) - h(D), & 0 \leq D \leq \alpha_1(\beta) \\ p_2(\beta) \left[ h(\alpha_2(\beta)) - h\left(\frac{D - p_1(\beta)\alpha_1(\beta)}{1 - p_1(\beta)}\right) \right] \\ + (1 - p_1(\beta) - p_2(\beta)) \left[ 1 - h\left(\frac{D - p_1(\beta)\alpha_1(\beta)}{1 - p_1(\beta)}\right) \right], & \alpha_1(\beta) < D \leq \alpha_2(\beta)(1 - p_1(\beta)) + p_1(\beta)\alpha_1(\beta) \\ (1 - p_1(\beta) - p_2(\beta)) \left[ 1 - h\left(\frac{D - p_1(\beta)\alpha_1(\beta) - p_2(\beta)\alpha_2(\beta)}{1 - p_1(\beta) - p_2(\beta)}\right) \right], & \alpha_2(\beta)(1 - p_1(\beta)) + p_1(\beta)\alpha_1(\beta) < D \\ 0, & \leq \alpha_1(\beta)p_1(\beta) + \alpha_2(\beta)p_2(\beta) + \frac{1 - p_1(\beta) - p_2(\beta)}{2} \\ & D > \alpha_1(\beta)p_1(\beta) + \alpha_2(\beta)p_2(\beta) + \frac{1 - p_1(\beta) - p_2(\beta)}{2} \end{cases} \quad (194)$$

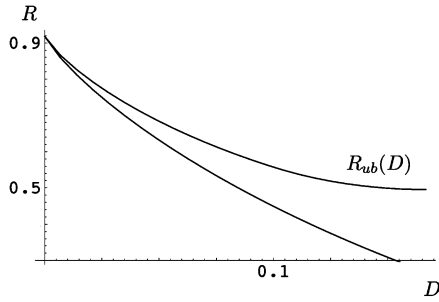


Fig. 5. Binary symmetric Markov chain with transition probability 1/3. The upper bound on the rate–distortion function curve is  $R_{ub}(D)$  from (195), and the lower curve is the SLB.

where  $p$  is the source transition probability and  $R^-(D)$  is given in (186). On the other hand, the SLB gives  $R(D) \geq h(p) - h(D)$ , implying

$$R(D) = h(p) - D \log 1/D + o(D \log 1/D), \text{ as } D \rightarrow 0. \tag{201}$$

Thus,  $R_{ub}(D)$  is optimal in the low-distortion regime not only in the sense, implied by (196), that  $\lim_{D \rightarrow 0} R_{ub}(D) = R(0) = H(\mathbf{X})$ , but in the stronger sense of attaining the second-order term in the expansion of  $R(D)$  around 0. Fig. 5 shows (197) and the SLB.

H. Almost Lossless Compression of Nonerased Symbols

To conclude this section, we briefly consider the setting where a source  $\mathbf{X}$  is connected to a memoryless erasure channel with erasure rate  $\epsilon$ . The compressor only has access to the output of the erasure channel  $\mathbf{Z}$ , and the decompressor must output a reproduction  $\hat{\mathbf{X}}$  so that any symbol  $X_i$  that has not been erased must be reproduced almost losslessly, i.e.,

$$\lim_{n \rightarrow \infty} P \left( \bigcup_{i=1}^n \{Z_i = X_i \neq \hat{X}_i\} \right) = 0. \tag{202}$$

For symbols that have been erased, we naturally must allow lossy reproduction since the compressor does not have access to  $\mathbf{X}$ . The problem is to find the optimum rate–distortion tradeoff when (202) is satisfied and the distortion is gauged as in (103) by

$$E \left[ \frac{1}{|\{1 \leq i \leq n : Z_i = e\}|} \sum_{1 \leq i \leq n : Z_i = e} \rho(X_i, \hat{X}_i) \right].$$

The simplest case is a binary nonredundant source with Hamming distortion. This setup is identical to a conventional memoryless rate–distortion problem where the source has alphabet  $\{0, e, 1\}$  and the reproduction alphabet is  $\{0, 1\}$  with distortion  $\rho(0,0) = \rho(1,1) = 0$ ;  $\rho(e,1) = \rho(e,0) = 1/2$ ,  $\rho(1,0) = \rho(0,1) = 1$ . No distortion smaller than 1/2 per erasure (or  $\epsilon/2$  per reproduced symbol) is achievable since even if the decompressor were to have access to  $\mathbf{Z}$ , it does not have any information about those symbols of  $\mathbf{X}$  that have been erased. Moreover,

because of the requirement (202), distortion higher than  $\epsilon/2$  per reproduced symbol is not allowed. Reproducing the nonerased symbols almost losslessly requires rate  $1 - \epsilon$  bits per reproduced symbol. In fact, solving for the rate–distortion function in this problem, we see that the optimum achievable point is  $(R, D) = (1 - \epsilon, 1/2)$  (where distortion is gauged per erased symbol). Thus, there is no penalty for the location of the erasures being unknown to the decompressor.

What about sources with memory? For low  $\epsilon$ , we argue that the optimum rate distortion point for a stationary ergodic source is

$$(R, D) = (H(\mathbf{X}), D_{max}), \tag{203}$$

where  $D_{max}$  is the lowest value of  $D$  for which  $R^-(D) = 0$ , namely

$$D_{max} = E \left[ \min_{\hat{x}} E [\rho(X_0, \hat{x}) | X_{\setminus 0}] \right]. \tag{204}$$

To see that the pair in (203) is achievable note that separate description of the nonerased source and the erasure pattern by the compressor requires rate no larger than  $H(\mathbf{X}) + h(\epsilon)$ . The decompressor, now knowing the erased sequence, employs a decoder from the setting of Section IV-A corresponding to zero rate (as we assume no additional description rate from the encoder). The achieved distortion is then  $D_{max}^{(\epsilon)}$ , where  $D_{max}^{(\epsilon)}$  is the lowest value of  $D$  for which  $R_e(D) = 0$ . Thus, for any value of  $0 \leq \epsilon \leq 1$ , the pair

$$(H(\mathbf{X}) + h(\epsilon), D_{max}^{(\epsilon)}) \tag{205}$$

is achievable. Noting that, as  $\epsilon \rightarrow 0$ ,  $h(\epsilon) \rightarrow 0$  and, by Theorem 15,  $D_{max}^{(\epsilon)} \rightarrow D_{max}$ , shows that the rate distortion pair in (203) is achievable in the small  $\epsilon$  regime.

For a lower bound, consider a genie-aided scheme where the decompressor knows the erasure locations. The compressor needs to convey the nonerased symbols (but not the location of the erased ones), which requires rate  $R$  lower-bounded by  $H(\mathbf{X}) - \epsilon \log |\mathcal{X}|$ , since the savings relative to the case where the erased symbols would also need to be described cannot exceed  $\epsilon \log |\mathcal{X}|$ . The expected distortion in estimating each erased symbol is lower-bounded by  $D_{max}$  (the distortion achieved by a genie that estimates every erased symbol on the basis of all other source symbols rather than only on the nonerased ones). Thus, one can do no better than (203) in the low erasure regime.

V. CHANNEL CAPACITY

In this section, we study the decrease in channel capacity due to erasures of the channel outputs. The capacity of the concatenation of a noisy channel with an erasure channel with erasure rate  $\epsilon$  is denoted by  $C(\epsilon)$  (Fig. 6). Thus,  $C(0) = C$ , the capacity of the noisy channel. Throughout, the erasures are independent of the noisy channel. For simplicity, we restrict our attention to finite-alphabet channels without cost constraints.

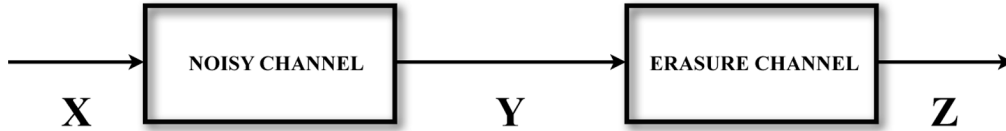


Fig. 6. Noisy channel observed through an erasure channel.

We first deal with the case where the noisy channel is memoryless.<sup>5</sup>

*Theorem 21:* Suppose that the erasures are independent of the inputs and outputs of the noisy channel and that the proportion of erasures converges in probability to  $e$  (the erasure channel may have memory), and let the noisy channel be memoryless with capacity  $C$ . Then

$$C(e) = C - eC. \quad (206)$$

This result remains true if the encoder has noncausal (and hence causal) knowledge of the location of the erasures.

*Proof:* To show the converse under noncausal knowledge of the location of the erasures at the encoder, we note that it is immaterial what the encoder chooses to send at the erased symbols. At the decoder, the erasures are discarded and the non-erased outputs are fed to the noisy channel decoder. Thus, the problem is equivalent to one of coding for the noisy channel except that now the effective coding length (i.e., the number of nonerased symbols) is a random fraction of the number of transmitted symbols. Still, by the definition of  $e$ , the rate of the resulting scheme converges to  $R(1 - e)$ , where  $R$  is the rate of the noisy channel code. The result follows from the conventional converse of the memoryless channel. To show the achievability of (206), in the absence of any knowledge at the encoder of the location of the erasures, it is enough to follow the conventional random coding reasoning with codebooks whose rate is  $(1 - e)C - \epsilon$  and a decoder that simply discards the erasures. Averaging with respect to the codebook selection, the noisy channel randomness and the erasures, the block error probability vanishes just as it does in the case where the erasures are memoryless. Thus, there must exist a good codebook (independent of the erasure realization), similarly as is the case in the absence of the noisy channel.  $\square$

Theorem 21 gives an example of a channel which may have memory, but for which noiseless feedback does not increase capacity.<sup>6</sup> The memoryless Gaussian channel followed by a memoryless erasure channel has been shown to have the capacity of the Gaussian channel times  $(1 - e)$  [27]. It is tempting to generalize Theorem 21 to the case where both the noisy channel and the erasure channel have memory. However, in that situation,  $(1 - e)C$  need not be either an upper or a lower bound.

*Example 11:* Consider a binary channel where  $Y_i = Y_{i-1} = X_i$  when  $i$  is even, and suppose that every other output

<sup>5</sup>The special case of this result where the erasure process is stationary and weakly mixing is given in [18].

<sup>6</sup>See [1] and [2] for other examples of channels with memory for which feedback does not increase capacity.

symbol is erased. Then, regardless of the probability that the erased symbols are even or odd,  $C(0.5) = C = \frac{1}{2}$  bit.

*Example 12:* Suppose  $N_i$  is a sequence of fair coin flips independent of the input, and consider the binary channel where  $Y_{2i} = X_{2i} \oplus N_i$  and  $Y_{2i+1} = N_i$ . The capacity of this channel is  $C = \frac{1}{2}$  bit, while with the same erasure channel as in Example 11, we get  $C(0.5) = 0$ .

In the remainder of this section, we restrict our attention to the memoryless erasure channel. Before giving the capacity of the cascade of noisy channel and erasure channel we give the following auxiliary result.

*Theorem 22:* Let  $(\mathbf{X}, \mathbf{Y})$  be jointly stationary finite-alphabet processes and let  $\mathbf{Z}$  be the output of a memoryless erasure channel with erasure rate  $e$  driven by  $\mathbf{Y}$ . Then

$$I(\mathbf{X}; \mathbf{Y}|\mathbf{Z}) = eI^-(\mathbf{X}; \mathbf{Y}) + o(e). \quad (207)$$

*Proof:* Because of stationarity and the expression for erasure mutual information rate (56), we need to show

$$I(\mathbf{X}; Y_0|Y_{-\infty}^{-1}, \mathbf{Z}) = eI(\mathbf{X}; Y_0|Y_{\setminus 0}) + o(e). \quad (208)$$

Now, fix an arbitrary integer  $k$ . Since the process of erasures is i.i.d. independent of  $\mathbf{Y}$  we can write for any erasure rate

$$\begin{aligned} & \frac{1}{e}I(\mathbf{X}; Y_0|Y_{-\infty}^{-1}, \mathbf{Z}) \\ &= I(\mathbf{X}; Y_0|Y_{-\infty}^{-1}, Z_0 = e, Z_1^\infty) \end{aligned} \quad (209)$$

$$= I(\mathbf{X}; Y_0|Y_{-\infty}^{-1}, Z_1^\infty) \quad (210)$$

$$\begin{aligned} & \geq I(\mathbf{X}; Y_0|Y_{-\infty}^{-1}, Y_1^k, Z_1 \neq e, \dots, Z_k \neq e, Z_{k+1}^\infty) \\ & \quad \times P[Z_1 \neq e, \dots, Z_k \neq e] \end{aligned} \quad (211)$$

$$= (1 - e)^k I(\mathbf{X}; Y_0|Y_{-\infty}^{-1}, Y_1^k, Z_{k+1}^\infty) \quad (212)$$

$$= I(\mathbf{X}; Y_0|Y_{-\infty}^{-1}, Y_1^k, Z_{k+1}^\infty) + o(1) \quad (213)$$

for any integer  $k$ . So we must have that

$$\lim_{e \rightarrow 0} \frac{1}{e}I(\mathbf{X}; Y_0|Y_{-\infty}^{-1}, \mathbf{Z}) \geq \lim_{k \rightarrow \infty} I(\mathbf{X}; Y_0|Y_{-\infty}^{-1}, Y_1^k, Z_{k+1}^\infty) \quad (214)$$

$$= I(\mathbf{X}; Y_0|Y_{\setminus 0}). \quad (215)$$

On the other hand, again fixing an arbitrary integer  $k$

$$I(\mathbf{X}; Y_0|Y_{-\infty}^{-1}, Z_1^\infty) \quad (216)$$

$$= H(Y_0|Y_{-\infty}^{-1}, Z_1^\infty) - H(Y_0|\mathbf{X}, Y_{\setminus 0}, Z_1^\infty) \quad (217)$$

$$\leq H(Y_0|Y_{-\infty}^{-1}, Z_1^\infty) - H(Y_0|\mathbf{X}, Y_{\setminus 0}) \quad (218)$$

$$\leq H(Y_0|Y_{-\infty}^{-1}, Z_1^k) - H(Y_0|\mathbf{X}, Y_{\setminus 0}) \quad (219)$$

$$\leq H(Y_0|Y_{-\infty}^{-1}, Y_1^k)(1 - e)^k - H(Y_0|\mathbf{X}, Y_{\setminus 0}) \quad (220)$$

$$+ (1 - (1 - e)^k)H(\mathbf{Y}) \quad (221)$$

Thus

$$\begin{aligned} & \lim_{e \rightarrow 0} \frac{1}{e}I(\mathbf{X}; Y_0|Y_{-\infty}^{-1}, \mathbf{Z}) \\ & \leq H(Y_0|Y_{-\infty}^{-1}, Y_1^k) - H(Y_0|\mathbf{X}, Y_{-\infty}^{-1}, Y_1^\infty). \end{aligned} \quad (222)$$

So we must have

$$\begin{aligned} & \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} I(\mathbf{X}; Y_0 | Y_{-\infty}^{-1}, \mathbf{Z}) \\ & \leq \lim_{k \rightarrow \infty} H(Y_0 | Y_{-\infty}^{-1}, Y_1^k) \\ & - H(Y_0 | \mathbf{X}, Y_{-\infty}^{-1}, Y_1^\infty) \end{aligned} \quad (223)$$

$$= I(\mathbf{X}; Y_0 | Y_0). \quad (224)$$

□

*Theorem 23:* Assume that the erasure channel is memoryless, and that the noisy channel has finite input/output alphabets, is stationary, information stable, with capacity given by

$$C = \lim_{n \rightarrow \infty} \frac{1}{n} \max_{X^n} I(X^n; Y^n) \quad (225)$$

which is achieved by a unique stationary process  $\hat{\mathbf{X}} = \hat{X}_{-\infty}^\infty$

$$C = I(\hat{\mathbf{X}}; \hat{\mathbf{Y}}) \quad (226)$$

$$= I(\hat{\mathbf{X}}; \hat{Y}_0 | \hat{Y}_{-\infty}^{-1}). \quad (227)$$

Then

$$C(\epsilon) = C - \epsilon I^-(\hat{\mathbf{X}}; \hat{\mathbf{Y}}) + o(\epsilon). \quad (228)$$

*Proof:* The general framework in [38] implies that the capacity of the concatenation of the information stable noisy channel and the DMC is given by

$$C(\epsilon) = \lim_{n \rightarrow \infty} \frac{1}{n} \max_{X^n} I(X^n; Z^n) \quad (229)$$

$$= \lim_{n \rightarrow \infty} \frac{1}{n} I(X^{(\epsilon), n}; Z^{(\epsilon), n}) \quad (230)$$

$$= I(\mathbf{X}^{(\epsilon)}; Z_0^{(\epsilon)} | Z_{-\infty}^{(\epsilon), -1}) \quad (231)$$

where  $\mathbf{X}^{(\epsilon)}$  is a stationary process. Note that by definition  $\mathbf{X}^{(0)} = \hat{\mathbf{X}}$ ,  $\mathbf{Y}^{(0)} = \hat{\mathbf{Y}}$ ; furthermore, the response of the erasure channel to  $\hat{\mathbf{Y}}$  is denoted by  $\hat{\mathbf{Z}} = \mathbf{Z}^{(0)}$ . Since the channel input  $X^n$  and the erasure channel output  $Z^n$  are conditionally independent given the noisy channel output  $Y^n$

$$I(X^n; Z^n) = I(X^n; Y^n) - I(X^n; Y^n | Z^n). \quad (232)$$

Considering the normalized limits of the two sides of (232) implies that, when  $\mathbf{X}$  is stationary

$$I(\mathbf{X}; Z_0 | Z_{-\infty}^{-1}) = I(\mathbf{X}; Y_0 | Y_{-\infty}^{-1}) - I(\mathbf{X}; Y_0 | Y_{-\infty}^{-1}, \mathbf{Z}). \quad (233)$$

We can lower-bound the capacity of the cascade of the noisy channel and the erasure channel by

$$C(\epsilon) = I(\mathbf{X}^{(\epsilon)}; Z_0^{(\epsilon)} | Z_{-\infty}^{(\epsilon), -1}) \quad (234)$$

$$\geq I(\hat{\mathbf{X}}; \hat{Z}_0 | \hat{Z}_{-\infty}^{-1}) \quad (235)$$

$$= I(\hat{\mathbf{X}}; \hat{Y}_0 | \hat{Y}_{-\infty}^{-1}) - I(\hat{\mathbf{X}}; \hat{Y}_0 | \hat{Y}_{-\infty}^{-1}, \hat{\mathbf{Z}}) \quad (236)$$

$$= C - I(\hat{\mathbf{X}}; \hat{Y}_0 | \hat{Y}_{-\infty}^{-1}, \hat{\mathbf{Z}}), \quad (237)$$

$$= C - \epsilon I^-(\hat{\mathbf{X}}; \hat{\mathbf{Y}}) + o(\epsilon) \quad (238)$$

where

- (234) = (231);
- (235)  $\Leftarrow$   $\hat{\mathbf{X}}$  is stationary;

- (236)  $\Leftarrow$  (233);
- (237)  $\Leftarrow$  (231);
- (238)  $\Leftarrow$  (207)

On the other hand, we can similarly upper-bound

$$C(\epsilon) = I(\mathbf{X}^{(\epsilon)}; Z_0^{(\epsilon)} | Z_{-\infty}^{(\epsilon), -1}) \quad (239)$$

$$= I(\mathbf{X}^{(\epsilon)}; Y_0^{(\epsilon)} | Y_{-\infty}^{(\epsilon), -1}) - I(\mathbf{X}^{(\epsilon)}; Y_0^{(\epsilon)} | Y_{-\infty}^{(\epsilon), -1}, \mathbf{Z}^{(\epsilon)}) \quad (240)$$

$$\leq \max_{\mathbf{X}} I(\mathbf{X}; Y_0 | Y_{-\infty}^{-1}) - I(\mathbf{X}^{(\epsilon)}; Y_0^{(\epsilon)} | Y_{-\infty}^{(\epsilon), -1}, \mathbf{Z}^{(\epsilon)}) \quad (241)$$

$$= C - I(\mathbf{X}^{(\epsilon)}; Y_0^{(\epsilon)} | Y_{-\infty}^{(\epsilon), -1}, \mathbf{Z}^{(\epsilon)}). \quad (242)$$

$$\leq C - \epsilon I(\mathbf{X}^{(\epsilon)}; Y_0^{(\epsilon)} | Y_0^{(\epsilon)}) + o(\epsilon) \quad (243)$$

$$= C - \epsilon I(\hat{\mathbf{X}}; \hat{Y}_0 | \hat{Y}_0) + o(\epsilon) \quad (244)$$

$$= C - \epsilon I^-(\hat{\mathbf{X}}; \hat{\mathbf{Y}}) + o(\epsilon) \quad (245)$$

where

- (240)  $\Leftarrow$  (233);
- (242)  $\Leftarrow$  the scope of channels considered in the theorem;
- (243)  $\Leftarrow$  the chain (209)–(213) holds even if the input process  $\mathbf{X}$  is allowed to depend on  $\epsilon$ ;
- (244)  $\Leftarrow$  the uniqueness of  $\hat{\mathbf{X}}$  as the capacity achieving input for the case  $\epsilon = 0$ . □

In the regime of sporadic nonerasures, the capacity vanishes linearly with the nonerasure rate:

*Theorem 24:* Assume that the erasure channel is memoryless, and that the noisy channel is information stable with capacity given by (225). Then

$$\lim_{\epsilon \rightarrow 1} \frac{C(\epsilon)}{1 - \epsilon} = \lim_{n \rightarrow \infty} \frac{1}{n} \max_{X^n} \sum_{k=1}^n I(X^n; Y_k). \quad (246)$$

*Proof:* Let  $V_k = 1\{Z_k = e\}$ . Since  $V_k$  are i.i.d., independent of the channel input and  $P[V_k = 1] = \epsilon$ , we can write for any  $X^n$

$$I(X^n; Z^n) = I(X^n; Z^n | V^n) \quad (247)$$

$$= \epsilon^{n-1} (1 - \epsilon) \sum_{k=1}^n I(X^n; Y_k) + \sum_{\ell=2}^n \epsilon^{n-\ell} (1 - \epsilon)^\ell$$

$$\times \sum_{S \subset \{1, \dots, n\}, |S|=\ell} I(X^n; Y_S). \quad (248)$$

The second term in the right side of (248) can be upper-bounded by

$$\begin{aligned} & \sum_{\ell=2}^n \epsilon^{n-\ell} (1 - \epsilon)^\ell \sum_{S \subset \{1, \dots, n\}, |S|=\ell} I(X^n; Y_S) \\ & \leq \log |\mathcal{B}| \sum_{\ell=2}^n \binom{n}{\ell} \epsilon^{n-\ell} (1 - \epsilon)^\ell \ell \end{aligned} \quad (249)$$

$$= [(1 - \epsilon)n - \epsilon^{n-1}(1 - \epsilon)n] \log |\mathcal{B}| \quad (250)$$

$$= (1 - \epsilon^{n-1})(1 - \epsilon)n \log |\mathcal{B}| \quad (251)$$

where  $\mathcal{B}$  is the output alphabet. Denote the right side of (246) by  $C^*$ . Fix  $\epsilon > 0$  and let  $n_0$  be such that for all  $n > n_0$

$$C^* - \epsilon \leq \frac{1}{n} \max_{X^n} \sum_{k=1}^n I(X^n; Y_k) \leq C^* + \epsilon. \quad (252)$$

Putting together (229), (248), (251), and choosing  $n = n_0 + 1$

$$(C^* - \epsilon)e^{n_0} \leq \frac{C(e)}{1 - e} \quad (253)$$

$$\leq (C^* + \epsilon)e^{n_0} + (1 - e^{n_0}) \log |\mathcal{B}|. \quad (254)$$

In view of the arbitrariness of  $\epsilon$ , (246) follows by taking the limit as  $e \uparrow 1$  of (253) and (254).  $\square$

An application of Theorems 23 and 24 is the discrete symmetric channel with memory:

*Theorem 25:* Consider a discrete channel whose input/output alphabet is a finite field  $\mathcal{A}$  endowed with addition  $\oplus$

$$Y_i = X_i \oplus N_i \quad (255)$$

where the stationary ergodic error process  $\mathbf{N}$  has entropy rate  $H(\mathbf{N})$  and erasure entropy rate  $H^-(\mathbf{N})$ . Then

$$C = \log |\mathcal{A}| - H(\mathbf{N}) \quad (256)$$

$$C(e) = (1 - e)C - e(H(\mathbf{N}) - H^-(\mathbf{N})) + o(e) \quad (257)$$

$$C(e) = (1 - e)(\log |\mathcal{A}| - H(N_1)) + o(1 - e). \quad (258)$$

*Proof:* The capacity without erasures (256) is well known (e.g., [38]). In the absence of erasures, for every  $n$ , independent equiprobable inputs  $\hat{X}^n$  uniquely maximize the mutual information, yielding capacity (256). For those inputs, applying (58) we obtain

$$I^-(X^n; Y^n) = n - H^-(N_1, \dots, N_n) \quad (259)$$

and (258) follows from (225) and the definition of erasure entropy rate. To show (258), we apply Theorem 24 and use

$$\max_{X^n} I(X^n; Y_k) = \log |\mathcal{A}| - H(N_k) \quad (260)$$

and the stationarity of the error process.  $\square$

Note that Theorem 25 provides a counterexample to the statement that if the erasures are memoryless then

$$(1 - e)C(0) \leq C(e).$$

## VI. DENOISING

The setting of Section IV dealt with the scenario where the source to be encoded is corrupted by a memoryless erasure channel, and the decoder's task is to recover the erased symbols to within the lowest possible distortion on the basis of its observations, and the additional description of the source from the encoder. In this section, we look at the denoising problem, where there is no additional description of the source by an encoder, but the corruption may be by a channel other than the erasure channel. We develop relationships between the information measures introduced in Section II, in particular

erasure entropy and divergence, and the fundamental limits of denoising.

Discrete denoising deals with the minimization of the distortion achieved by an algorithm that observes the output of the channel but, in contrast to the settings in Sections III and IV, has no other information on the input realization.

Concretely, an  $n$ -block denoiser  $\hat{X}^n$  is a sequence of mappings  $\hat{X}^n = \{\hat{X}_t(\cdot)\}_{t=1}^n$ , where  $\hat{X}_t(\cdot)$  takes the noisy  $n$ -tuple  $Y^n$  into its reconstruction  $\hat{X}_t(Y^n)$ . We denote the per-symbol cumulative loss of the denoiser  $\hat{X}^n$  when observing the noisy sequence  $y^n$  while the underlying clean one is  $x^n$  by

$$L_{\hat{X}^n}(x^n, y^n) = \frac{1}{n} \sum_{t=1}^n \Lambda(x_t, \hat{X}_t(y^n)). \quad (261)$$

Letting  $\mathcal{D}_n$  denote the set of all  $n$ -block denoisers, we define the *denoisability* of the process pair  $\mathbf{X}, \mathbf{Y}$  by

$$\mathbb{D}(\mathbf{X}, \mathbf{Y}) = \limsup_{n \rightarrow \infty} \min_{\hat{X}^n \in \mathcal{D}_n} E[L_{\hat{X}^n}(X^n, Y^n)]. \quad (262)$$

A filter is a causal denoiser, i.e., one whose  $t$ th reconstruction  $\hat{X}_t$  depends on  $Y^n$  only through  $Y^t$ . Letting  $\mathcal{F}_n$  denote the subset of  $\mathcal{D}_n$  consisting of all the filters, we define the *filterability* of the process pair  $\mathbf{X}, \mathbf{Y}$  by

$$\mathbb{F}(\mathbf{X}, \mathbf{Y}) = \limsup_{n \rightarrow \infty} \min_{\hat{X}^n \in \mathcal{F}_n} E[L_{\hat{X}^n}(X^n, Y^n)]. \quad (263)$$

The limit suprema both in (262) and in (263) are, by a standard subadditivity argument, in fact limits when  $\mathbf{X}, \mathbf{Y}$  are jointly stationary. We also use the notation  $\mathbb{D}(\mathbf{X}, \mathbf{\Pi})$  and  $\mathbb{F}(\mathbf{X}, \mathbf{\Pi})$  to denote  $\mathbb{D}(\mathbf{X}, \mathbf{Y})$  and  $\mathbb{F}(\mathbf{X}, \mathbf{Y})$  when  $\mathbf{Y}$  is the output of the DMC with transition matrix  $\mathbf{\Pi}$  whose input is  $\mathbf{X}$ .

### A. Relationship Between $\mathbb{D}(\mathbf{X}, \mathbf{Y})$ and $H^-(\mathbf{Y})$

Intuitively, one can expect that the higher the entropy rate of a noise-corrupted process the more difficult it is to estimate its components. This intuition was made precise for some specific filtering problems in [12] which, among other things, established a relationship between the filterability of the process pair  $\mathbf{X}, \mathbf{Y}$ , when  $\mathbf{Y}$  is the BSC-corrupted version of  $\mathbf{X}$ , and the entropy rate of the noisy process  $\mathbf{Y}$ . In this subsection, we show that an analogous relationship holds for the denoisability of the pair  $\mathbf{X}, \mathbf{Y}$ , when the entropy rate is replaced by the erasure entropy rate. We do this in the generality of a stationary noise-free process  $\mathbf{X}$  corrupted by a DMC with full row-rank  $\mathbf{\Pi}$ , which is the setting we assume throughout the remainder of this section.

A distribution  $P$  on a noisy (channel output)  $n$ -tuple  $Y^n$  will be said to be "*bona fide*" if there exists a distribution on  $X^n$  that gives rise to  $P$  when corrupted by the channel, i.e., for all  $y^n$

$$P(y^n) = \sum_{x^n} P_{X^n}(x^n) \prod_{i=1}^n \mathbf{\Pi}(x_i, y_i). \quad (264)$$

Invertibility of the channel implies uniqueness of the  $P_{X^n}$  satisfying (264) for a given *bona fide*  $P$ . Furthermore, this invertibility guarantees the existence of an inverse (linear) transformation to the one in (264) that maps a *bona fide*  $P$  into the

channel input distribution  $P_{X^n}$  (cf. [17], [45] for its explicit form). Thus, to determine whether a given  $P$  is *bona fide* we can simply apply it to the inverse transformation and verify that all the components of the resulting probability vector are non-negative (the components will sum to 1 even if  $P$  is not *bona fide*).

Consider first the problem of estimating a single random variable  $X$  on the basis of its DMC( $\mathbf{\Pi}$ )-corrupted observation  $Y$ . Note that, due to the invertibility of the channel, the distribution of  $Y$ ,  $P_Y$ , uniquely determines the channel input distribution  $P_X$  and, hence, the joint distribution of  $X$ ,  $Y$ . Thus, the minimum attainable expected loss for such an estimation problem, under the loss function  $\Lambda$ , can be expressed as a function of the channel output distribution, which we denote by  $f_{\mathbf{\Pi}}(P_Y)$ . Specifically,  $f_{\mathbf{\Pi}}(P_Y)$  is expressed as

$$f_{\mathbf{\Pi}}(P_Y) = \min_{\hat{X}(\cdot)} E[\Lambda(X, \hat{X}(Y))] \quad (265)$$

$$= \sum_y P_Y(y) \min_{\hat{x}} E[\Lambda(X, \hat{x})|Y = y] \quad (266)$$

$$= \sum_y P_Y(y) \min_{\hat{x}} \sum_x P_{X|Y}(x|y) \Lambda(x, \hat{x}) \quad (267)$$

$$= \sum_y \min_{\hat{x}} \sum_x P_X(x) \mathbf{\Pi}(x, y) \Lambda(x, \hat{x}) \quad (268)$$

$$= \sum_y \min_{\hat{x}} \sum_x [(\mathbf{\Pi}\mathbf{\Pi}^T)^{-1} \mathbf{\Pi} P_Y](x) \mathbf{\Pi}(x, y) \Lambda(x, \hat{x}) \quad (269)$$

where the expectation on the right-hand side of (265) is under the (unique) distribution of  $X$ ,  $Y$  consistent with  $P_Y$ ,  $[(\mathbf{\Pi}\mathbf{\Pi}^T)^{-1} \mathbf{\Pi} P_Y](x)$  in (269) stands for the  $x$ th component of the (column) vector  $(\mathbf{\Pi}\mathbf{\Pi}^T)^{-1} \mathbf{\Pi} P_Y$ , and equality (269) follows from the relationship  $P_X = (\mathbf{\Pi}\mathbf{\Pi}^T)^{-1} \mathbf{\Pi} P_Y$  (cf. [39, Sec. 4]). The expression in (269) is the explicit form of  $f_{\mathbf{\Pi}}(P_Y)$ . Define now

$$\varepsilon_{\mathbf{\Pi}} = \min_{a,b} \max_{P_Y \in \mathcal{C}(\mathbf{\Pi})} |f_{\mathbf{\Pi}}(P_Y) - [aH(P_Y) + b]| \quad (270)$$

where  $\mathcal{C}(\mathbf{\Pi}) \subseteq \mathcal{M}(\mathcal{Y})$  denotes the set of all *bona fide* channel output distributions (of a single-channel output symbol)

$$\mathcal{C}(\mathbf{\Pi}) = \left\{ \mathbf{\Pi}^T P_X : P_X \in \mathcal{M}(\mathcal{X}) \right\}. \quad (271)$$

$\varepsilon_{\mathbf{\Pi}}$  quantifies the extent to which the (single-letter) channel denoisability can be approximated by an affine function of the channel output entropy.

*Example 13:* With slight abuse of notation let  $f_{\delta}$  and  $\varepsilon_{\delta}$  stand for  $f_{\mathbf{\Pi}}$  and  $\varepsilon_{\mathbf{\Pi}}$ , when  $\mathbf{\Pi}$  is the BSC of crossover probability  $\delta < 1/2$ . Specializing (269) and (270) to this case gives

$$f_{\delta}(P_Y) = \min \left\{ \frac{\min\{P_Y(1), P_Y(0)\} - \delta}{1 - 2\delta}, \delta \right\} \quad (272)$$

and

$$\varepsilon_{\delta} = \min_{a,b} \max_{\delta \leq \alpha \leq 1/2} |f_{\delta}(\alpha) - [ah(\alpha) + b]| \quad (273)$$

where  $h$  denotes the binary entropy function. It is easy to bound  $\varepsilon_{\delta}$  for specific values of  $\delta$ . For example,  $\varepsilon_{0.25} < 0.03$ .

For any stationary process  $\mathbf{X}$ , it is shown in [12] that

$$|\mathbb{F}(\mathbf{X}, \delta) - [a_{\delta}^* H(\mathbf{Y}) + b_{\delta}^*]| \leq \varepsilon_{\delta}.$$

This fact is used in [12] to bound the sensitivity of the filtering performance to the order in which a multidimensional data array is scanned into a one-dimensional signal. As the following theorem shows, a similar bound holds for the denoising problem, upon replacing entropy rate with erasure entropy rate.

*Theorem 26:* Let  $a_{\mathbf{\Pi}}^*$ ,  $b_{\mathbf{\Pi}}^*$  be achievers of the minimum in (270). For any stationary process  $\mathbf{X}$

$$|\mathbb{D}(\mathbf{X}, \mathbf{\Pi}) - [a_{\mathbf{\Pi}}^* H^-(\mathbf{Y}) + b_{\mathbf{\Pi}}^*]| \leq \varepsilon_{\mathbf{\Pi}}.$$

*Proof:* Consider

$$\min_{\hat{X}_t(\cdot), 1 \leq t \leq n} E \frac{1}{n} \sum_{t=1}^n \Lambda(X_t, \hat{X}_t(Y^n)) \quad (274)$$

$$= \frac{1}{n} \sum_{t=1}^n \min_{\hat{X}_t(\cdot)} E \Lambda(X_t, \hat{X}_t(Y^n)) \quad (275)$$

$$= \frac{1}{n} \sum_{t=1}^n \min_{\hat{X}_t(\cdot)} E \left[ E \left( \Lambda(X_t, \hat{X}_t(Y^n)) | Y_{\setminus t} \right) \right] \quad (276)$$

$$= \frac{1}{n} \sum_{t=1}^n E \left[ \min_{\hat{X}_t(\cdot)} E \left( \Lambda(X_t, \hat{X}_t(Y_t)) | Y_{\setminus t} \right) \right] \quad (277)$$

$$= \frac{1}{n} \sum_{t=1}^n E \left[ f_{\mathbf{\Pi}}(P_{Y_t | Y_{\setminus t}}) \right] \quad (278)$$

$$\leq \frac{1}{n} \sum_{t=1}^n E \left[ a_{\mathbf{\Pi}}^* H(P_{Y_t | Y_{\setminus t}}) + b_{\mathbf{\Pi}}^* + \varepsilon_{\mathbf{\Pi}} \right] \quad (279)$$

$$= \frac{1}{n} \sum_{t=1}^n a_{\mathbf{\Pi}}^* H(Y_t | Y_{\setminus t}) + b_{\mathbf{\Pi}}^* + \varepsilon_{\mathbf{\Pi}} \quad (280)$$

$$= a_{\mathbf{\Pi}}^* \frac{1}{n} H^-(Y^n) + b_{\mathbf{\Pi}}^* + \varepsilon_{\mathbf{\Pi}} \quad (281)$$

where

- (278)  $\Leftarrow$  definition of  $f_{\mathbf{\Pi}}$ ;
- (279)  $\Leftarrow$  definitions of  $a_{\mathbf{\Pi}}^*$ ,  $b_{\mathbf{\Pi}}^*$ , and  $\varepsilon_{\mathbf{\Pi}}$ ;
- (280)  $\Leftarrow$   $E[H(P_{Y_t | Y_{\setminus t}})] = H(Y_t | Y_{\setminus t})$ .

Taking the limits of both sides of (281) gives

$$\mathbb{D}(\mathbf{X}, \mathbf{\Pi}) \leq a_{\mathbf{\Pi}}^* H^-(\mathbf{Y}) + b_{\mathbf{\Pi}}^* + \varepsilon_{\mathbf{\Pi}}. \quad (282)$$

A similar argument, where the inequality in (279) would be reversed upon replacement of  $\varepsilon_{\mathbf{\Pi}}$  by  $-\varepsilon_{\mathbf{\Pi}}$  would lead to

$$\mathbb{D}(\mathbf{X}, \delta) \geq a_{\mathbf{\Pi}}^* H^-(\mathbf{Y}) + b_{\mathbf{\Pi}}^* - \varepsilon_{\mathbf{\Pi}} \quad (283)$$

which completes the proof when combined with (282).  $\square$

Thus, the entropy and erasure entropy determine the filterability and denoisability, respectively, to within  $\varepsilon_{\mathbf{\Pi}}$ . In particular, two noisy processes with the same erasure entropy rate can differ in their denoisability by no more than  $2\varepsilon_{\mathbf{\Pi}}$ .

### B. Mismatched Denoising: the Role of Erasure Divergence

Let  $P$  be a *bona fide* distribution of a noisy channel  $n$ -tuple  $Y^n$ . Throughout the remainder of this section expectations are assuming that the channel output sequence distribution is  $P$ . Thus, we write  $E[L_{\hat{X}^n}(X^n, Y^n)]$  to denote the expected loss of the filter/denoiser  $\hat{X}^n$  when the noisy sequence  $Y^n$  is distributed according to  $P$  (which uniquely determines the joint

distribution of  $X^n, Y^n$ ). Let further  $L_P^f(x^n, y^n)$  denote<sup>7</sup> the normalized loss incurred with input  $x^n$  and output  $y^n$  attained by a filter that is optimal for the noisy source  $P$  in the sense of achieving the minimum  $\min_{\hat{X}^n \in \mathcal{F}_n} E[L_{\hat{X}^n}(X^n, Y^n)]$ , i.e.,

$$E[L_P^f(X^n, Y^n)] = \min_{\hat{X}^n \in \mathcal{F}_n} E_P L_{\hat{X}^n}(X^n, Y^n). \quad (284)$$

It was shown in [31] that, for *bona fide*  $P, Q$

$$E[L_Q^f(X^n, Y^n) - L_P^f(X^n, Y^n)] \leq \sqrt{2} \Lambda_{max} K_{\Pi} \sqrt{\frac{1}{n} D(P||Q)} \quad (285)$$

where  $K_{\Pi}$  is the squared Frobenius norm of  $\Pi^{-1}$ . The implication of inequality (285) is that if one assumes the noisy source to be  $Q$ , and operates optimally under this assumption, then one's performance will be close to optimum, provided  $Q$  is close to the true noisy source distribution in the sense of normalized divergence. This result is the filtering analogue of the result on mismatched prediction in [30]. The bound in (285) motivates the following approach for the construction of a universal filter: find a probability assignment for the noisy source  $Q$ , which is universal in the sense that  $\frac{1}{n} D(P||Q) \rightarrow 0$  for the sources in the uncertainty set. Inequality (285) guarantees that the filter induced by the "source"  $Q$  (i.e., which is optimal for that source) is a universal filter. For example, the filter in [33] can be thought of as the filter induced by the Lempel–Ziv (LZ) probability assignment which, in turn, is induced by the LZ data compression scheme [46] known to be universal with respect to stationary sources or, equivalently, to satisfy  $\frac{1}{n} D(P||Q) \rightarrow 0$  for all stationary  $P$ .

We shall now see that erasure divergence plays a key role in bounding the loss due to denoising a source using a denoiser which was tailored for a different source. This role is analogous to the role played by regular divergence in the filtering problem (cf. (285)). Analogously to  $L_P^f(x^n, y^n)$ , let  $L_P^d(x^n, y^n)$  denote<sup>8</sup> the normalized loss of a denoiser which is optimal for the noisy source  $P$  in the sense of achieving the minimum  $\min_{\hat{X}^n \in \mathcal{D}_n} EL_{\hat{X}^n}(X^n, Y^n)$ , i.e.,

$$E[L_P^d(X^n, Y^n)] = \min_{\hat{X}^n \in \mathcal{D}_n} EL_{\hat{X}^n}(X^n, Y^n). \quad (286)$$

*Theorem 27:* For any pair  $P, Q$  of bona fide distributions on  $Y^n$

$$E[L_Q^d(X^n, Y^n) - L_P^d(X^n, Y^n)] \leq \sqrt{2} \Lambda_{max} K_{\Pi} \sqrt{\frac{1}{n} D^-(P||Q)}. \quad (287)$$

Like (285), the implication of Theorem 27, whose proof is deferred to the Appendix, is that if one finds a probability assignment for the noisy source  $Q$ , which is universal in the sense that  $\frac{1}{n} D^-(P||Q) \rightarrow 0$  for the sources in the uncertainty set, then the denoiser induced by  $Q$  will be a universal denoiser. Unlike the case with filtering, however, even when a universal  $Q$  is found, obtaining the induced denoiser can be a computational challenge. Whereas  $Q$  is likely to be specified in terms of

the conditional distributions  $Q_{X_t|X^{t-1}}$ , which lend themselves to a simple derivation of the induced filter, the induced denoiser requires the conditional distributions  $Q_{X_t|X^t}$ . The problem of computing  $Q_{X_t|X^t}$  and approximations thereof, as induced by sequential probability assignments, was considered in [32], [44].

## APPENDIX

### A. Proof of (70) Assuming the Condition in (82)

In this proof, for a set of indices  $S$ , we use  $X(S)$  to denote  $\{X_i\}_{i \in S}$ . For  $A_n \subseteq \{1, \dots, n\}$ , denote for brevity  $A_n^c = \{1, \dots, n\} \setminus A_n$  and  $A_n[k] = \{i \in A_n : \min\{|i-j| : j \neq i, j \in A_n\} > k\}$  and note that for any positive integer  $k$  small enough that  $2k+1 < n$

$$H(X(A_n)|X(A_n^c)) = \sum_{i \in A_n} H(X_i|X(A_n^c), \{X_j : j \in A_n, j < i\}) \quad (A1)$$

$$\leq \sum_{i \in A_n[k]} H(X_i|X_{i-k}^{i-1}, X_{i+1}^{i+k}) + \sum_{i \in A_n \setminus A_n[k]} H(X_i) \quad (A2)$$

$$\leq |A_n| H(X_0|X_{-k}^{-1}, X_1^k) + |A_n \setminus A_n[k]| H(X_0). \quad (A3)$$

For  $\varepsilon > 0$  and positive integer  $k$  let  $\mathcal{A}_n^{\varepsilon, k, e}$  denote the collection of subsets  $A_n \subseteq \{1, \dots, n\}$  that satisfy the following two properties:

$$1. \quad e - \varepsilon \leq \frac{|A_n|}{n} \leq e + \varepsilon; \quad (A4)$$

$$2. \quad |A_n \setminus A_n[k]| \leq |A_n|(\alpha_e(k) + \varepsilon) \quad (A5)$$

where  $\alpha_e(k) = e \sum_{\ell=0}^k (1-e)^\ell$ .

Note that, by (82)

$$P(\{1 \leq i \leq n : Z_i = e\} \in \mathcal{A}_n^{\varepsilon, k, e}) \rightarrow 1 \text{ as } n \rightarrow \infty \quad (A6)$$

and (A4) and (A5) when combined with the upper bound (A3), imply that

$$\begin{aligned} \frac{1}{n} H(X(A_n)|X(A_n^c)) &\leq (e + \varepsilon) \cdot H(X_0|X_{-k}^{-1}, X_1^k) + (e + \varepsilon) \\ &\quad \times (\alpha_e(k) + \varepsilon) \cdot H(X_0) \quad \forall A_n \in \mathcal{A}_n^{\varepsilon, k, e}. \end{aligned} \quad (A7)$$

Recalling that  $S(Z^n) = \{1 \leq i \leq n : Z_i = e\}$ , we can write  $\frac{1}{n} H(X^n|Z^n)$  as shown in (A8)–(A10) at the top of the following page, where the inequality follows from (A7). Taking the limit as  $n \rightarrow \infty$  of both sides of (A10) gives, when combined with (A6)

$$H(\mathbf{X}|\mathbf{Z}) \leq (e + \varepsilon) \cdot H(X_0|X_{-k}^{-1}, X_1^k) + (e + \varepsilon)(\alpha_e(k) + \varepsilon) \cdot H(X_0)$$

implying, by the arbitrariness of  $\varepsilon > 0$ , that

$$H(\mathbf{X}|\mathbf{Z}) \leq e \cdot H(X_0|X_{-k}^{-1}, X_1^k) + e \alpha_e(k) \cdot H(X_0). \quad (A11)$$

Noting that  $\lim_{e \rightarrow 0} \alpha_e(k) = 0$ , we obtain

$$\limsup_{e \rightarrow 0} \frac{H(\mathbf{X}|\mathbf{Z})}{e} \leq H(X_0|X_{-k}^{-1}, X_1^k) \quad (A12)$$

<sup>7</sup>The superscript  $f$  indicates that this is a filtering loss.

<sup>8</sup>The superscript  $d$  indicates that this is a denoising loss.

$$\frac{1}{n}H(X^n|Z^n) = \frac{1}{n} \sum_{A_n \subseteq \{1, \dots, n\}} H(X(A_n)|X(A_n^c))P[S(Z^n) = A_n] \tag{A8}$$

$$= \frac{1}{n} \sum_{A_n \in \mathcal{A}_n^{\varepsilon, k, e}} H(X(A_n)|X(A_n^c))P[S(Z^n) = A_n] + \frac{1}{n} \sum_{A_n \notin \mathcal{A}_n^{\varepsilon, k, e}} H(X(A_n)|X(A_n^c))P[S(Z^n) = A_n] \tag{A9}$$

$$\leq [(e + \varepsilon) \cdot H(X_0|X_{-k}^{-1}, X_1^k) + (e + \varepsilon)(\alpha_e(k) + \varepsilon) \cdot H(X_0)] P[S(Z^n) \in \mathcal{A}_n^{\varepsilon, k, e}] + H(X_0)P[S(Z^n) \notin \mathcal{A}_n^{\varepsilon, k, e}] \tag{A10}$$

implying, by the arbitrariness of  $k$ ,

$$\limsup_{\varepsilon \rightarrow 0} \frac{H(\mathbf{X}|\mathbf{Z})}{e} \leq H^-(\mathbf{X}). \tag{A13}$$

**B. Proof of Theorem 13**

The proof is based on the following lemma.

*Lemma 2:* Consider the single-letter problem where  $Z$  is the noisy version of  $X$  corrupted by the DMC with channel matrix  $\mathbf{I} - \delta\mathbf{M}$ . Then

$$\lim_{\delta \rightarrow 0} \frac{H(X|Z)}{\delta \log \frac{1}{\delta}} = E[M(X, X)] \tag{A14}$$

where, for any  $\varepsilon > 0$ , the convergence in (A14) is uniform over  $P_X \in \mathcal{M}_\varepsilon$ .

*Proof:* For  $\mathbf{v} \in [0, \infty)^{|\mathcal{X}|-1}$  let  $P_{\mathbf{v}, \delta}$  be a parametrized family of distributions on  $\{1, \dots, |\mathcal{X}|\}$  that satisfies

$$P_{\mathbf{v}, \delta}(i) = \begin{cases} 1 - \delta\|\mathbf{v}\|_1 + o(\delta), & \text{if } i = i_0, \\ \mathbf{v}(i)\delta + o(\delta), & \text{if } i \neq i_0 \end{cases} \tag{A15}$$

for any  $i_0 \in \mathcal{X}$ . It is easy to check that

$$\lim_{\delta \rightarrow 0} \frac{H(P_{\mathbf{v}, \delta})}{\delta \log \frac{1}{\delta}} = \|\mathbf{v}\|_1 \tag{A16}$$

where the convergence in (A16) is uniform over  $\|\mathbf{v}\|_1 \leq B$ , for any  $B > 0$ .

Returning to the problem at hand note that

$$P_Z(b) = P_X(b) - \delta \sum_{a \in \mathcal{X}} P_X(a)M(a, b). \tag{A17}$$

Since  $P_X \in \mathcal{M}_\varepsilon$ , we can write

$$P_{X|Z}(a|b) = \begin{cases} 1 - \delta M(b, b) + o(\delta), & a = b \\ -\delta M(a, b) \frac{P_X(a)}{P_X(b)} + o(\delta), & a \neq b. \end{cases} \tag{A18}$$

Applying (A16) to (A18) we obtain

$$\lim_{\delta \rightarrow 0} \frac{H(X|Z = b)}{\delta \log \frac{1}{\delta}} = M(b, b). \tag{A19}$$

Averaging the numerator in the left side of (A19) with respect to  $P_Z$  and using (A17) we obtain (A14) uniformly in  $\mathcal{M}_\varepsilon$ .

*Proof of Theorem 13:* To obtain the corresponding behavior of  $H(\mathbf{X}|\mathbf{Z})$ , for a stationary input that satisfies (100), we can simply notice that

$$H(\mathbf{X}|\mathbf{Z}) = H(X_0|X_{-\infty}^{-1}, Z_0^\infty) \tag{A20}$$

and proceed as we did in the proof of Lemma 2 with  $Z_0$  taking the role of  $Z$  and substituting  $P_X$  by  $P_{X_0|X_{-\infty}^{-1} = a_{-\infty}^{-1}, Z_1^\infty = b_1^\infty}$  which also belongs to  $\mathcal{M}_\varepsilon$ . Thus, we obtain

$$\lim_{\delta \rightarrow 0} \frac{H(X_0|X_{-\infty}^{-1} = a_{-\infty}^{-1}, Z_0^\infty = b_0^\infty)}{\delta \log \frac{1}{\delta}} = M(b_0, b_0). \tag{A21}$$

The average with respect to the distribution in (A17) where  $P_X$  is substituted by  $P_{X_0|X_{-\infty}^{-1} = a_{-\infty}^{-1}, Z_1^\infty = b_1^\infty}$ , is

$$\begin{aligned} \lim_{\delta \rightarrow 0} \frac{H(X_0|Z_0, X_{-\infty}^{-1} = a_{-\infty}^{-1}, Z_1^\infty = b_1^\infty)}{\delta \log \frac{1}{\delta}} \\ = E[M(X_0, X_0)|X_{\setminus 0} = (a_{-\infty}^{-1}, b_1^\infty)]. \end{aligned} \tag{A22}$$

Finally, averaging with respect to the infinite past/future of the input the desired result follows.

**C. Proof of Theorem 15**

For  $P_{X^n} \in \mathcal{M}(\mathcal{X}^n)$ , let  $R(P_{X^n}, D)$  denote the rate–distortion function in (107) under the association  $P_X \leftrightarrow P_{X^n}$  and  $\rho(X, Y) \leftrightarrow \rho(X^n, Y^n)$ , i.e.,

$$\begin{aligned} R(P_{X^n}, nD) = \min\{I(X^n; Y^n) \\ : E\rho(X^n, Y^n) \leq nD, X^n \sim P_{X^n}\} \end{aligned} \tag{A23}$$

where  $\rho(X^n, Y^n) = \sum_{i=1}^n \rho(X_i, Y_i)$ . For  $P_{X, S} \in \mathcal{M}(\mathcal{X} \times \mathcal{S})$ , let  $R_{SI}(P_{X, S}, D)$  denote the conditional rate distortion function when the source and side information are i.i.d.  $\sim P_{X, S}$

$$\begin{aligned} R_{SI}(P_{X, S}, D) = \min\{I(X; Y|S) \\ : E\rho(X, Y, S) \leq D, (X, S) \sim P_{X, S}\}. \end{aligned} \tag{A24}$$

The following lemma collects some observations that will be used in the proof of Theorem 15.

*Lemma 3:*  $R_{SI}(P_{X, S}, D)$  has the following properties.

1.

$$\begin{aligned} R_{SI}(P_{X, S}, D) \\ = \min_{\{D_s\}_s \in \mathcal{S}: \sum_s P_S(s)D_s = D} \sum_s P_S(s)R_s(P_{X|S}, D_s) \end{aligned}$$

where  $R_s$  is the rate–distortion function defined in (107), under distortion function  $\rho(\cdot, \cdot, s)$ .

2. For fixed  $P_{X|S}$ ,  $R_{SI}(P_S \times P_{X|S}, D)$  is uniformly continuous as a function of  $P_S$  in the following sense: There exists  $\delta(\varepsilon)$  (dependent on  $|\mathcal{X}|, |\mathcal{S}|, \rho$ ) such that  $\lim_{\varepsilon \rightarrow 0} \delta(\varepsilon) = 0$  and

$$\begin{aligned} |R_{SI}(P_S \times P_{X|S}, D) - R_{SI}(P'_S \times P_{X|S}, D)| \leq \delta(\varepsilon), \\ \text{for all } D \text{ and all } P_S, P'_S \text{ with } \|P_S - P'_S\| \leq \varepsilon. \end{aligned}$$

3. Let  $X_1, \dots, X_n$  be independent,  $X_i \sim P_{X|S=s_i}$ , where  $s_i \in \mathcal{S}$ ,  $s^n = (s_1, \dots, s_n)$  being a deterministic sequence. Letting  $p_{s^n}$  denote the empirical distribution of  $s^n$

$$\frac{1}{n} R(P_{X^n}, nD) = R_{SI}(p_{s^n} \times P_{X|S}, D).$$

The first and third properties in the lemma are direct consequences of the definitions of the functions  $R_{SI}$ ,  $R_s$ , and  $R$ . The second property follows from the uniform continuity of the mutual information  $I(X; S)$  in the distribution  $P_{X,S}$  when  $X$  and  $S$  take values in finite alphabets (cf., e.g., [15]).

Before we proceed to the proof of Theorem 15, we make the following observation.

*Observation 2:* For every  $k$ ,  $R_k^-(D)$ , as defined in (108), satisfies

$$R^-(D) = \min \sum_{x_{-k}^{-1}, x_1^k} P(x_{-k}^{-1}, x_1^k) R(P_{X_0|x_{-k}^{-1}, x_1^k}, D_{x_{-k}^{-1}, x_1^k}) \quad (\text{A25})$$

where the minimum is over all distortion values indexed by contexts  $\{D_{x_{-k}^{-1}, x_1^k}\}$  that yield an overall distortion  $D$  when averaged over the contexts, i.e.,

$$\sum_{x_{-k}^{-1}, x_1^k} P(x_{-k}^{-1}, x_1^k) D_{x_{-k}^{-1}, x_1^k} = D.$$

To verify the validity of the observation note that  $R_k^-(D)$  is nothing but  $R_{SI}(P_{X,S}, D)$  in (A24) with the association  $X \leftrightarrow X_0$  and  $S \leftrightarrow (X_{-k}^{-1}, X_1^k)$ . Equality (A25) is then a consequence of the first part of Lemma 3.

We are now in a position to prove Theorem 15.

*Proof of Theorem 15:* Throughout the proof let  $\mathbf{X}$  be a stationary and ergodic source. Noting that, by the weak law of large numbers,  $\frac{|\{1 \leq i \leq n: Z_i \neq e\}|}{n} \rightarrow e$  in probability, it follows that replacing the condition in (103) with

$$\frac{1}{ne} \sum_{i=1}^n E \left[ 1_{\{Z_i=e\}} \cdot \rho(X_i, \hat{X}_i) \right] \leq D + \varepsilon \quad (\text{A26})$$

will result in the same rate distortion function  $R_e(D)$ . Evidently

$$R_e(D) = \frac{1}{e} R_{X|Z}(De) \quad (\text{A27})$$

where  $R_{X|Z}(\cdot)$  is the conditional rate-distortion function for encoding the source  $\mathbf{X}$  in the presence of side information  $\mathbf{Z}$ , under the distortion function  $d(x, \hat{x}, z) = 1_{\{z=e\}} \cdot \rho(x, \hat{x})$  (the  $1/e$  factor on the right-hand side is due to the fact that  $R_{X|Z}(D)$  corresponds to rate in bits per source symbol, rather than per erased symbol as in the definition of  $R_e(D)$ ). Rate-distortion theory for stationary ergodic sources [4], [19] implies (A28) at the bottom of the page. Now, for any fixed  $k$ , see (A29)–(A41) on the top of the following page, where

- (A30)  $\Leftarrow$  definition of  $R_{SI}$  (in (A24)) with the association  $X \leftrightarrow X^n$ ,  $Y \leftrightarrow Y^n$ ,  $S \rightarrow Z^n$  and  $\rho(X, Y, S) \leftrightarrow \sum_{i=1}^n 1_{\{Z_i=e\}} \cdot \rho(X_i, Y_i)$ .

- (A31)  $\Leftarrow$  second item in Lemma 3;
- (A32)  $\Leftarrow$  taking  $D_{z^n} = D$  for all  $z^n$  in lieu of the  $\{D_{z^n}\}$  that achieve the minimum.
- (A33)  $\Leftarrow$  letting  $P_{X(\{1 \leq i \leq n: z_i=e\})|X_i=z_i, 1 \leq i \leq n: z_i \neq e}$  denote the conditional distribution of  $X(\{1 \leq i \leq n: z_i=e\})$  given the event  $\{X_i = z_i, 1 \leq i \leq n: z_i \neq e\}$ , and noting that  $P_{X^n|z^n} = P_{X^n|X_i=z_i, 1 \leq i \leq n: z_i \neq e}$  (thus, in particular, conditioned on  $\{X_i = z_i, 1 \leq i \leq n: z_i \neq e\}$ ,  $X(\{1 \leq i \leq n: z_i \neq e\})$  is deterministic).
- (A34) holds for any set  $B_n^{\varepsilon, k, e} \subseteq \mathcal{Z}^n$ .
- (A35)  $\Leftarrow$  defining  $B_n^{\varepsilon, k, e}$  as the set of  $z^n$ 's for which:
  - 1.

$$\{1 \leq i \leq n: z_i = e\} \in \mathcal{A}_n^{\varepsilon, k, e} \quad (\text{A42})$$

where  $\mathcal{A}_n^{\varepsilon, k, e}$  was defined in the proof of Theorem 10 as the collection of subsets of  $\{1, \dots, n\}$  satisfying (A4) and (A5),

2. for all  $(x_{-k}^{-1}, x_1^k) \in \mathcal{X}^{2k}$ , see (A43) at the bottom of the following page

Note, in particular, that the two properties defining  $\mathcal{A}_n^{\varepsilon, k, e}$  (A.4), (A.5) imply that if  $z^n \in B_n^{\varepsilon, k, e}$  then see (A44) also the bottom of the following page implying (A45) at the bottom of the following page, since omitting no more than  $(e + \varepsilon)n(\alpha_e(k) + \varepsilon)$  terms from the mutual information defining the rate-distortion function in (A45) can decrease it by no more than  $(e + \varepsilon)n(\alpha_e(k) + \varepsilon) \log |\mathcal{X}|$ . Inequality (A45) accounts for (A35). We also note, for future reference, that the law of large numbers (which implies (A6)), combined with the ergodicity of  $\mathbf{X}$  implies

$$\lim_{n \rightarrow \infty} P(Z^n \in B_n^{\varepsilon, k, e}) = 1. \quad (\text{A46})$$

- (A37)  $\Leftarrow$  letting  $Q_{z^n}$  denote the empirical distribution shown in (A47) at the bottom of the following page, and noting that for all  $z^n$  we get (A48) at the top of the subsequent page, since the right-hand side corresponds to a suboptimal solution from the feasible set associated with the optimization problem defining  $R(\cdot)$  on the left-hand side, whereby the reconstruction symbols are independent with a conditional distribution, given the source, that depends on a window of radius  $k$  source symbols.
- (A38)  $\Leftarrow$  by definition of  $B_n^{\varepsilon, k, e}$ ,  $z^n \in B_n^{\varepsilon, k, e}$  implies (A42) which, in turn, implies

$$\begin{aligned} & |\{1 \leq i \leq n: z_i = e, \min\{|i-j|: j \neq i, z_j = e\} > k\}| \\ & \leq |\{1 \leq i \leq n: z_i = e\}| \leq n(e + \varepsilon). \end{aligned}$$

- (A39)  $\Leftarrow$  by definition of  $B_n^{\varepsilon, k, e}$ ,  $z^n \in B_n^{\varepsilon, k, e}$  implies (A43) or, equivalently,  $\|Q_{z^n} - P_{X_{-k}^{-1}, X_1^k}\| \leq \varepsilon$ . Thus (A38) follows from the second item in Lemma 3 with the association  $P_S \leftrightarrow P_{X_{-k}^{-1}, X_1^k}$  and  $P'_S \leftrightarrow Q_{z^n}$ ;
- (A41)  $\Leftarrow$   $R_k^-(D)$ : as defined in (108), is nothing but  $R_{SI}(P_{X,S}, D)$  with the association  $X \leftrightarrow X_0$  and  $S \leftrightarrow (X_{-k}^{-1}, X_1^k)$ .

$$R_{X|Z}(D) = \lim_{n \rightarrow \infty} \frac{1}{n} \min_{P_{Y^n|X^n, Z^n}: \sum_{i=1}^n E[1_{\{Z_i=e\}} \cdot \rho(X_i, Y_i)] \leq nD} I(X^n; Y^n|Z^n). \quad (\text{A28})$$

$$\frac{1}{n} P(y^n | x^n, z^n) : \sum_{i=1}^n \min_{E[1_{\{Z_i=e\}} \cdot \rho(X_i, Y_i)] \leq nD} I(X^n; Y^n | Z^n) \tag{A29}$$

$$= \frac{1}{n} R_{SI}(P_{X^n, Z^n}, nD) \tag{A30}$$

$$= \frac{1}{n} \min_{\{D_{z^n}\} : \sum_{z^n} P_{Z^n}(z^n) D_{z^n} = D} \sum_{z^n} P_{Z^n}(z^n) R(P_{X^n | z^n}, nD_{z^n}) \tag{A31}$$

$$\leq \frac{1}{n} \sum_{z^n} P_{Z^n}(z^n) R(P_{X^n | z^n}, nD) \tag{A32}$$

$$= \frac{1}{n} \sum_{z^n} P_{Z^n}(z^n) R(P_{X(\{1 \leq i \leq n: z_i=e\}) | \{X_i=z_i, 1 \leq i \leq n: z_i \neq e\}}, nD) \tag{A33}$$

$$\leq \frac{1}{n} \sum_{z^n \in B_n^{\varepsilon, k, e}} P_{Z^n}(z^n) R(P_{X(\{1 \leq i \leq n: z_i=e\}) | \{X_i=z_i, 1 \leq i \leq n: z_i \neq e\}}, nD) + P(Z^n \notin B_n^{\varepsilon, k, e}) \log |\mathcal{X}| \tag{A34}$$

$$\leq \frac{1}{n} \sum_{z^n \in B_n^{\varepsilon, k, e}} P_{Z^n}(z^n) [R(P_{X(\{1 \leq i \leq n: z_i=e, \min\{|i-j|: j \neq i, z_j=e\}} > k) | \{X_i=z_i, 1 \leq i \leq n: z_i \neq e\}}, nD) \tag{A35}$$

$$+ (e + \varepsilon)n(\alpha_e(k) + \varepsilon) \log |\mathcal{X}|] + P(Z^n \notin B_n^{\varepsilon, k, e}) \log |\mathcal{X}|$$

$$\leq \frac{1}{n} \sum_{z^n \in B_n^{\varepsilon, k, e}} P_{Z^n}(z^n) R(P_{X(\{1 \leq i \leq n: z_i=e, \min\{|i-j|: j \neq i, z_j=e\}} > k) | \{X_i=z_i, 1 \leq i \leq n: z_i \neq e\}}, nD) \tag{A36}$$

$$+ [(e + \varepsilon)(\alpha_e(k) + \varepsilon) + P(Z^n \notin B_n^{\varepsilon, k, e})] \log |\mathcal{X}|$$

$$\leq \frac{1}{n} \sum_{z^n \in B_n^{\varepsilon, k, e}} P_{Z^n}(z^n) |\{1 \leq i \leq n : z_i = e, \min\{|i - j| : j \neq i, z_j = e\} > k\}| \tag{A37}$$

$$\times R_{SI} \left( Q_{z^n} \times P_{X_0 | X_{-k}^{-1}, X_1^k}, \frac{nD}{|\{1 \leq i \leq n : z_i = e, \min\{|i - j| : j \neq i, z_j = e\} > k\}|} \right) \tag{A38}$$

$$+ [(e + \varepsilon)(\alpha_e(k) + \varepsilon) + P(Z^n \notin B_n^{\varepsilon, k, e})] \log |\mathcal{X}|$$

$$\leq (e + \varepsilon) \left[ \sum_{z^n \in B_n^{\varepsilon, k, e}} P_{Z^n}(z^n) R_{SI} \left( Q_{z^n} \times P_{X_0 | X_{-k}^{-1}, X_1^k}, \frac{D}{e + \varepsilon} \right) \right] \tag{A39}$$

$$+ [(e + \varepsilon)(\alpha_e(k) + \varepsilon) + P(Z^n \notin B_n^{\varepsilon, k, e})] \log |\mathcal{X}|$$

$$\leq (e + \varepsilon) \sum_{z^n \in B_n^{\varepsilon, k, e}} P_{Z^n}(z^n) \left[ R_{SI} \left( P_{X_{-k}^{-1}, X_1^k} \times P_{X_0 | X_{-k}^{-1}, X_1^k}, \frac{D}{e + \varepsilon} \right) + \delta(\varepsilon) \right] \tag{A40}$$

$$+ [(e + \varepsilon)(\alpha_e(k) + \varepsilon) + P(Z^n \notin B_n^{\varepsilon, k, e})] \log |\mathcal{X}|$$

$$= (e + \varepsilon) \left[ R_k^- \left( \frac{D}{e + \varepsilon} \right) + \delta(\varepsilon) \right] + [(e + \varepsilon)(\alpha_e(k) + \varepsilon) + P(Z^n \notin B_n^{\varepsilon, k, e})] \log |\mathcal{X}| \tag{A41}$$

$$\left| \frac{|\{1 \leq i \leq n : z_i = e, \min\{|i - j| : j \neq i, z_j = e\} > k, (z_{i-k}^{i-1}, z_{i+1}^{i+k}) = (x_{-k}^{-1}, x_1^k)\}|}{|\{1 \leq i \leq n : z_i = e, \min\{|i - j| : j \neq i, z_j = e\} > k\}|} - P_{X_{-k}^{-1}, X_1^k}(x_{-k}^{-1}, x_1^k) \right| \leq \varepsilon. \tag{A43}$$

$$|\{1 \leq i \leq n : z_i = e\}| \leq |\{1 \leq i \leq n : z_i = e, \min\{|i - j| : j \neq i, z_j = e\} > k\}| + (e + \varepsilon)n(\alpha_e(k) + \varepsilon) \tag{A44}$$

$$\begin{aligned} & R(P_{X(\{1 \leq i \leq n: z_i=e\}) | \{X_i=z_i, 1 \leq i \leq n: z_i \neq e\}}, nD) \\ & \leq R(P_{X(\{1 \leq i \leq n: z_i=e, \min\{|i-j|: j \neq i, z_j=e\}} > k) | \{X_i=z_i, 1 \leq i \leq n: z_i \neq e\}}, nD) \\ & \quad + (e + \varepsilon)n(\alpha_e(k) + \varepsilon) \log |\mathcal{X}| \end{aligned} \tag{A45}$$

$$Q_{z^n}(x_{-k}^{-1}, x_1^k) = \frac{|\{1 \leq i \leq n : z_i = e, \min\{|i - j| : j \neq i, z_j = e\} > k, (z_{i-k}^{i-1}, z_{i+1}^{i+k}) = (x_{-k}^{-1}, x_1^k)\}|}{|\{1 \leq i \leq n : z_i = e, \min\{|i - j| : j \neq i, z_j = e\} > k\}|} \tag{A47}$$

$$\begin{aligned}
& R \left( P_X(\{1 \leq i \leq n: z_i = e, \min\{|i-j|: j \neq i, z_j = e\} > k\}) | \{X_i = z_i, 1 \leq i \leq n: z_i \neq e\}, nD) \right. \\
& \leq |\{1 \leq i \leq n: z_i = e, \min\{|i-j|: j \neq i, z_j = e\} > k\}| \\
& \quad \times R_{SI} \left( Q_{z^n} \times P_{X_0|X_{-k}^{-1}, X_1^k}, \frac{nD}{|\{1 \leq i \leq n: z_i = e, \min\{|i-j|: j \neq i, z_j = e\} > k\}|} \right) \tag{A48}
\end{aligned}$$

$$\begin{aligned}
R_e(D) & \leq \lim_{n \rightarrow \infty} \left\{ \frac{e + \varepsilon}{e} \left[ R_k^- \left( \frac{De}{e + \varepsilon} \right) + \delta(\varepsilon) \right] + \left[ \frac{e + \varepsilon}{e} (\alpha_e(k) + \varepsilon) + \frac{1}{e} P(Z^n \notin B_n^{\varepsilon, k, e}) \right] \log |\mathcal{X}| \right\} \\
& = \frac{e + \varepsilon}{e} \left[ R_k^- \left( \frac{De}{e + \varepsilon} \right) + \delta(\varepsilon) \right] + \frac{e + \varepsilon}{e} (\alpha_e(k) + \varepsilon) \log |\mathcal{X}|, \tag{A49}
\end{aligned}$$

$$\frac{1}{n} I(X^n; Y^n | Z^n) = \frac{1}{n} [H(X^n | Z^n) - H(X^n | Y^n, Z^n)] \tag{A54}$$

$$= \frac{1}{n} \sum_{i=1}^n H(X_i | X^{i-1}, Z^n) - H(X_i | X^{i-1}, Y^n, Z^n) \tag{A55}$$

$$\geq \frac{1}{n} \sum_{i=1}^n H(X_i | X^{i-1}, X_{i+1}^n, Z^n) - H(X_i | X^{i-1}, Y_i, Z_i^n) \tag{A56}$$

$$\geq \frac{1}{n} \sum_{i=1}^n H(X_i | X_{-\infty}^{i-1}, X_{i+1}^\infty, Z_i) - H(X_i | X^{i-1}, Y_i, Z_i^{i+k}) \tag{A57}$$

$$= H(X_0 | X_{-\infty}^{-1}, X_1^\infty, Z_0) - \frac{1}{n} \sum_{i=1}^n H(X_i | X^{i-1}, Y_i, Z_i^{i+k}) \tag{A58}$$

$$= H^-(\mathbf{X}) \cdot e - \frac{1}{n} \sum_{i=1}^n H(X_i | X^{i-1}, Y_i, Z_{i+1}^{i+k}, Z_i = e) \cdot e \tag{A59}$$

$$\geq e \left[ H^-(\mathbf{X}) - \frac{1}{n} \sum_{i=1}^n \psi \left( \frac{1}{e} \frac{E[1_{\{Z_i=e\}} \cdot \rho(X_i, Y_i)]}{1 - ek} \right) - ek \log |\mathcal{X}| \right] \tag{A60}$$

$$\geq e \left[ H^-(\mathbf{X}) - \psi \left( \frac{1}{e} \frac{\frac{1}{n} \sum_{i=1}^n E[1_{\{Z_i=e\}} \cdot \rho(X_i, Y_i)]}{1 - ek} \right) - ek \log |\mathcal{X}| \right] \tag{A61}$$

$$\geq e \left[ H^-(\mathbf{X}) - \psi \left( \frac{1}{e} \frac{De}{1 - ek} \right) - ek \log |\mathcal{X}| \right] \tag{A62}$$

The combination of (A27), (A28), and (A41) gives (A49) at the top of the page, where the equality is due to (A46). The arbitrariness of  $\varepsilon > 0$ , continuity of  $R_k^-(\cdot)$ , and fact that  $\lim_{\varepsilon \rightarrow 0} \delta(\varepsilon)$  imply

$$R_e(D) \leq R_k^-(D) + \alpha_e(k) \log |\mathcal{X}| \tag{A50}$$

which, since  $\lim_{e \rightarrow 0} \alpha_e(k) = 0$ , implies

$$\limsup_{e \rightarrow 0} R_e(D) \leq R_k^-(D) \tag{A51}$$

and, therefore, by the arbitrariness of  $k$  and (110)

$$\limsup_{e \rightarrow 0} R_e(D) \leq R^-(D). \tag{A52}$$

For the lower bound fix any joint distribution of  $(X^n, Y^n, Z^n)$  (i.e., a conditional  $P_{Y^n | X^n, Z^n}$ ) under which

$$\sum_{i=1}^n E[1_{\{Z_i=e\}} \cdot \rho(X_i, Y_i)] \leq nDe. \tag{A53}$$

Then, for fixed  $n$  and  $k < n$ , we get (A54)–(A62) at the top of the page, where

- (A57)  $\Leftrightarrow H(X_i | X^{i-1}, X_{i+1}^n, Z^n) \geq H(X_i | X^{i-1}, X_{i+1}^\infty, Z^n) = H(X_i | X^{i-1}, X_{i+1}^\infty, Z_i)$ ;
- (A58)  $\Leftarrow$  stationarity;
- (A59)  $\Leftarrow H(X_0 | X_{-\infty}^{-1}, X_1^\infty, Z_0) = H(X_0 | X_{-\infty}^{-1}, X_1^\infty) \cdot e = H^-(\mathbf{X}) \cdot e$  and

$$\begin{aligned}
& H(X_i | X^{i-1}, Y_i, Z_i^{i+k}) \\
& = H(X_i | X^{i-1}, Y_i, Z_{i+1}^{i+k}, Z_i = e) \cdot e; \tag{A63}
\end{aligned}$$

- (A60)  $\Leftarrow$  defining

$$\psi(D) = \max_{P_{Y_0 | X_{-\infty}^{-1}}} \max_{E\rho(X_0, Y_0) \leq D} H(X_0 | X_{-\infty}^{-1}, X_1^\infty, Y_0) \tag{A64}$$

$$H(X_i|X^{i-1}, Y_i, Z_{i+1}^{i+k}, Z_i = e) \tag{A65}$$

$$\leq H(X_i|X^{i-1}, Y_i, X_{i+1}^{i+k})P(Z_{i+1}^{i+k} = X_{i+1}^{i+k}) + P(Z_{i+1}^{i+k} \neq X_{i+1}^{i+k}) \log |\mathcal{X}| \tag{A66}$$

$$\leq \psi \left( E [\rho(X_i, Y_i)|Z_i = e, Z_{i+1}^{i+k} = X_{i+1}^{i+k}] \right) P(Z_{i+1}^{i+k} = X_{i+1}^{i+k}) + P(Z_{i+1}^{i+k} \neq X_{i+1}^{i+k}) \log |\mathcal{X}| \tag{A67}$$

$$= \psi \left( \frac{1}{e} E [1_{\{Z_i=e\}} \cdot \rho(X_i, Y_i)|Z_{i+1}^{i+k} = X_{i+1}^{i+k}] \right) P(Z_{i+1}^{i+k} = X_{i+1}^{i+k}) + P(Z_{i+1}^{i+k} \neq X_{i+1}^{i+k}) \log |\mathcal{X}| \tag{A68}$$

$$\leq \psi \left( \frac{1}{e} E [1_{\{Z_i=e\}} \cdot \rho(X_i, Y_i)|Z_{i+1}^{i+k} = X_{i+1}^{i+k}] \right) + ek \log |\mathcal{X}| \tag{A69}$$

$$\leq \psi \left( \frac{1}{e} \frac{E [1_{\{Z_i=e\}} \cdot \rho(X_i, Y_i)]}{P(Z_{i+1}^{i+k} = X_{i+1}^{i+k})} \right) + ek \log |\mathcal{X}| \tag{A70}$$

$$\leq \psi \left( \frac{1}{e} \frac{E [1_{\{Z_i=e\}} \cdot \rho(X_i, Y_i)]}{1 - ek} \right) + ek \log |\mathcal{X}| \tag{A71}$$

and noting that, for  $i > k$ , we have (A65)–(A71) at the top of the page, where (A67) follows by the definition of  $\psi$ , (A70) follows since

$$E [1_{\{Z_i=e\}} \cdot \rho(X_i, Y_i)|Z_{i+1}^{i+k} = X_{i+1}^{i+k}] \leq \frac{E [1_{\{Z_i=e\}} \cdot \rho(X_i, Y_i)]}{P(Z_{i+1}^{i+k} = X_{i+1}^{i+k})} \tag{A72}$$

and in the other steps we assume  $ek < 1$  and use the crude inequality  $P(Z_{i+1}^{i+k} \neq X_{i+1}^{i+k}) \leq ek$  (which follows from the union bound);

- (A61)  $\Leftarrow$  concavity of  $\psi$ ;
- (A62)  $\Leftarrow$  monotonicity of  $\psi$  and (A.53);

Thus, (A27), (A28), and (A62) imply

$$R_e(D) = \frac{1}{e} R_{X|Z}(De) \geq H^-(\mathbf{X}) - \psi \left( \frac{D}{1 - ek} \right) - ek \log |\mathcal{X}|. \tag{A73}$$

The continuity of  $\psi$  now further implies

$$\liminf_{\varepsilon \rightarrow 0} R_e(D) \geq H^-(\mathbf{X}) - \psi(D) = R^-(D) \tag{A74}$$

where the equality follows as in (106). Combining (A74) with (A52) completes the proof.  $\square$

*D. Proof of Theorem 27*

We use the following lemma, on the penalty of mismatched estimation, which follows by specializing [31, Lemma 4], to the case  $\varepsilon = 0$ .

*Lemma 4:* Consider the problem of estimating the random variable  $X$  on the basis of its DMC ( $\mathbf{\Pi}$ )-corrupted observation  $Y$ . Let  $\hat{X}^{P_Y}(\cdot)$  denote the optimal estimator in the sense of minimizing the expected loss  $E_{P_Y} \Lambda(X, \hat{X}(Y))$ , where  $E_{P_Y}$  denotes expectation when the noisy observation  $Y$  is distributed

according to  $P_Y$  (which uniquely determines the joint distribution of  $X, Y$  due to the invertibility of  $\mathbf{\Pi}$ ). Then, for any  $Q_Y$

$$E_{P_Y} \Lambda(X, \hat{X}^{Q_Y}(Y)) - E_{P_Y} \Lambda(X, \hat{X}^{P_Y}(Y)) \leq \Lambda_{max} K_{\mathbf{\Pi}} \|P_Y - Q_Y\|_1.$$

*Proof of Theorem 27:* Let  $\{\hat{X}_t^P(\cdot)\}_{t=1}^n$  denote the denoiser achieving the minimum in (286), i.e., which is optimal for the noisy source  $P$ . Then we get (A75)–(A82) at the top of the following page where

- (A78)  $\Leftarrow$  Lemma 4;
- (A79)  $\Leftarrow$  Pinsker’s inequality;
- (A80)  $\Leftarrow$  Jensen’s inequality.  $\square$

REFERENCES

- [1] F. Alajaji, “Feedback does not increase the capacity of discrete channels with additive noise,” *IEEE Trans. Inf. Theory*, vol. 41, no. 2, pp. 546–549, Mar. 1995.
- [2] V. Anantharam and S. Verdú, “Bits through queues,” *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 4–18, Jan. 1996.
- [3] B. A. Berg, *Markov Chain Monte Carlo Simulations and Their Statistical Analysis*. Singapore: World Scientific, 2004.
- [4] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
- [5] T. Berger and J. Gibson, “Lossy source coding,” *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2693–2723, Oct. 1998.
- [6] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, “Image inpainting,” in *Proc. SIGGRAPH 2000*, New Orleans, LA, Jul. 2000, pp. 417–424.
- [7] D. Blackwell, “The entropy of functions of finite-state Markov chains,” in *Trans. 1st Prague Conf. Information Theory: Statistical Decision Functions, Random Processes*, Prague, Czechoslovakia, 1957, pp. 13–20.
- [8] L. Breiman, *Probability*, *SIAM: Society for Industrial and Applied Mathematics*, Reprint ed. Philadelphia, PA: SIAM, 1992.
- [9] L. Breiman, “The individual ergodic theorem of information theory,” *Ann. Math. Stat.*, vol. 28, pp. 809–811, 1957.
- [10] H. Cai, S. Kulkarni, and S. Verdú, “An algorithm for universal lossless compression with side information,” *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4008–4016, Sep. 2006.
- [11] G. Casella and E. I. George, “Explaining the Gibbs sampler,” *Amer. Statistician*, vol. 46, pp. 167–174, 1992.

$$EL_Q^d(X^n, Y^n) - EL_P^d(X^n, Y^n) \quad (A75)$$

$$= \frac{1}{n} \sum_{t=1}^n E \left[ \Lambda(X_t, \hat{X}_t^Q(Y^n)) - \Lambda(X_t, \hat{X}_t^P(Y^n)) \right] \quad (A76)$$

$$= \frac{1}{n} \sum_{t=1}^n E \left[ E \left[ \Lambda(X_t, \hat{X}_t^Q(Y^n)) - \Lambda(X_t, \hat{X}_t^P(Y^n)) | Y_{\setminus t} \right] \right] \quad (A77)$$

$$\leq \frac{\Lambda_{max} K_{\Pi}}{n} \sum_{t=1}^n E \| P_{Y_t | Y_{\setminus t}} - Q_{Y_t | Y_{\setminus t}} \|_1 \quad (A78)$$

$$\leq \frac{\sqrt{2} \Lambda_{max} K_{\Pi}}{n} \sum_{t=1}^n E \sqrt{D(P_{Y_t | Y_{\setminus t}} \| Q_{Y_t | Y_{\setminus t}})} \quad (A79)$$

$$\leq \sqrt{2} \Lambda_{max} K_{\Pi} \sqrt{\frac{1}{n} \sum_{t=1}^n ED(P_{Y_t | Y_{\setminus t}} \| Q_{Y_t | Y_{\setminus t}})} \quad (A80)$$

$$= \sqrt{2} \Lambda_{max} K_{\Pi} \sqrt{\frac{1}{n} \sum_{t=1}^n D(P_{Y_t | Y_{\setminus t}} \| Q_{Y_t | Y_{\setminus t}} | P_{Y_{\setminus t}})} \quad (A81)$$

$$= \sqrt{2} \Lambda_{max} K_{\Pi} \sqrt{\frac{1}{n} D^-(P \| Q)}, \quad (A82)$$

- [12] A. Cohen, N. Merhav, and T. Weissman, "Scanning and sequential decision making for multi-dimensional data: Part II—The noisy case," *IEEE Trans. Inf. Theory*, to be published.
- [13] T. M. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 2, pp. 226–228, Mar. 1975.
- [14] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, pp. 146–158, Feb. 1975.
- [15] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
- [16] I. Csiszár and Z. Talata, "Consistent estimation of the basic neighborhood of Markov random fields," *Ann. Statist.*, vol. 34, no. 1, pp. 123–145, Feb. 2006.
- [17] A. Dembo and T. Weissman, "Universal denoising for the finite-input general-output channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1507–1517, Apr. 2005.
- [18] S. N. Diggavi and M. Grossglauser, "Information transmission over a finite buffer channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1226–1237, Mar. 2006.
- [19] R. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
- [20] H. O. Georgii, *Gibbs Measures and Phase Transitions*. Berlin–New York: Walter de Gruyter, 1988.
- [21] R. M. Gray, "Information rates of autoregressive processes," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 4, pp. 412–421, Jul. 1970.
- [22] R. M. Gray, "Rate distortion functions for finite-state finite-alphabet Markov sources," *IEEE Trans. Inf. Theory*, vol. IT-17, no. 2, pp. 127–134, Mar. 1971.
- [23] X. Guyon, *Random Fields on a Network*. New York: Springer-Verlag, 1995.
- [24] B. Hajek and T. Berger, "A decomposition theorem for binary Markov random fields," *Ann. Probab.*, vol. 15, pp. 1112–1125, 1987.
- [25] T. S. Han, "Linear dependence structure of the entropy space," *Info. Contr.*, vol. 29, pp. 337–368.
- [26] S. Jalali and T. Weissman, "New bounds on the rate-distortion function of a binary Markov source," in *proc. IEEE Int. Symp. Information Theory*, Nice, France, Jun. 2007, pp. 571–575.
- [27] D. Julian, "Erasure networks," in *Proc. IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, Jun. 2002, p. 138.
- [28] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM J. Res. Devel.*, vol. 5, pp. 183–191, 1961.
- [29] J. L. Massey, "Causality, feedback, and directed information," in *Proc. Int. Symp. Information Theory and its Applications*, Honolulu, HI, Nov. 1990, pp. 303–305.
- [30] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, Oct. 1998.
- [31] T. Moon and T. Weissman, "Discrete universal filtering via hidden Markov modeling," *IEEE Trans. Inf. Theory*, vol. 54, no. 2, pp. 692–708, Feb. 2008.
- [32] E. Ordentlich, M. Weinberger, and T. Weissman, "Efficient pruning of multi-directional context trees with applications to universal denoising and compression," in *Proc. 2004 IEEE Information Theory Workshop*, San Antonio, TX, Oct. 2004.
- [33] E. Ordentlich, T. Weissman, M. Weinberger, A. Somekh-Baruch, and N. Merhav, "Discrete universal filtering through incremental parsing," in *Proc. Data Compression Conf. (DCC 2004)*, Snowbird, UT, Mar. 2004.
- [34] E. Perron, S. Diggavi, and E. Telatar, "The Kaspi Rate-Distortion Problem with Encoder Side-Information: Binary Erasure Case," EPFL, 2007, LICOS-Rep. 2006-004.
- [35] K. Petersen, *Ergodic Theory*. Cambridge, U.K.: Cambridge Univ. Press, 1983, vol. 2, Cambridge Studies in Advanced Mathematics.
- [36] S. Shamai (Shitz), S. Verdú, and R. Zamir, "Systematic lossy source-channel coding," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 564–579, Mar. 1998.
- [37] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 4, pp. 471–480, Jul. 1973.
- [38] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.
- [39] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal discrete denoising: Known channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 5–28, Jan. 2005.
- [40] A. D. Wyner and J. Ziv, "The rate distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.
- [41] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder II: General sources," *Inf. Contr.*, vol. 38, pp. 60–80, 1978.
- [42] Z. Ye and T. Berger, *Information Measures for Discrete Random Fields*. Beijing and New York: Science, 1998.
- [43] J. Yu and S. Verdú, "Universal estimation of erasure entropy," in *Proc. 2006 IEEE Int. Symp. Information Theory*, Seattle, WA, Jul. 2006, pp. 2358–2362.
- [44] J. Yu and S. Verdú, "Schemes for bidirectional modeling of discrete stationary sources," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4789–4807, Nov. 2006.
- [45] R. Zhang and T. Weissman, "Discrete denoising for channels with memory," *Commun. in Inf. and Syst.*, vol. 5, no. 2, pp. 257–288, 2005.
- [46] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 5, pp. 530–536, Sep. 1978.