Erasure Entropy

Sergio Verdú
Department of Electrical Engineering
Princeton University
Princeton, NJ 08544
verdu@princeton.edu

Tsachy Weissman
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
tsachy@stanford.edu

***Abstract*— 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 rate of a source is defined as the limit of the normalized erasure entropy. The erasure entropy measures the information content carried by each symbol knowing its context. In the setup of a source observed through an erasure channel, we offer an operational characterization of erasure entropy rate as the minimal amount of bits per erasure required to recover the erased 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. 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 also explored.**

Keywords: Shannon Theory, Entropy, Data Compression, Rate-Distortion Theory, Discrete Denoising, Markov processes, Erasure Channels.

I. Introduction

The entropy of a source $\{X_1, \ldots X_n\}$ is equal to the sum of the conditional entropies of each symbol given all preceding (or all succeeding) symbols. The minimum expected number of bits of a compressed version of $\{X_1, \ldots X_n\}$ is equal to the entropy plus at most one bit. Moreover, as the number of symbols grows without bound, the minimum compression length 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?

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

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

where

$$X_{\setminus i} = \{X_j, j = 1, \ldots n, j \neq i\}. \tag{2}$$

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

$$H^-(\mathbf{X}) = \limsup_{n \to \infty} \frac{1}{n} H^-(X_1, \ldots X_n). \tag{3}$$

Erasure entropy is strictly lower than the conventional entropy (unless the source is memoryless, in which case they are identical). As we will see, there are processes with zero erasure entropy and nonzero entropy. 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? What properties does the entropy of a symbol conditioned on both the past and the future have?

For example, 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.

A cornerstone in the theory of reversible computing was put forward by Landauer [1] establishing a proportionality (Boltzmann's constant) between the entropy of a symbol (physically stored in a computer memory for example) and the increase in the thermodynamical entropy of the overall system when the symbol is erased. When the stored information source has memory, the increase in thermodynamical entropy incurred by an erasure is proportional to the erasure entropy rather than to the conventional Shannon entropy.

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. With probability one, this simulation gives, in the limit of many iterations (and of large image) a sample from the desired Markov random field (e.g. [2]). The number of random bits per pixel per iteration required for this simulation is equal to the erasure entropy of the field.

In the regime of low erasure rate, the erasure channel emerges as a convenient paradigm to obtain nontrivial 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 the ab-

1-4244-0504-1/06/$20.00 ©2006 IEEE

sence of any additional information it is impossible to perfectly recover the erased symbols. However, the recent body of work on *discrete denoising*, starting with [3], has shown efficient universal algorithms that exploit the information in the context of the erased (or, in general, contaminated) symbols in order to achieve the minimal distortion that would be feasible by an algorithm with perfect knowledge of the source statistics. We show in this paper that, at least for binary symmetric channels and bit-error rate distortion, the best achievable noncausal denoising performance and the erasure entropy are tightly coupled. In a follow-up to this paper, [4] studies the problem of universal estimation of erasure entropy rate.

The rest of the paper is organized as follows. Section II shows some basic properties of erasure entropy, as well as examples of explicit computation. In the case of a kth order Markov process we show the relationship between conventional entropy rate and erasure entropy rate. 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 that observes both the clean source and the location of the erasures needs to provide. For vanishing erasure probability the fundamental limit is 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 marginal conditional distributions of the output given all inputs that satisfy the average distortion constraint. In Section IV we also explore the counterpart of the Shannon lower bound and conditions for its tightness. The erasure rate-distortion function is determined for Gaussian sources with Euclidean distortion and for binary sources. Some relationships between erasure entropy and discrete denoising [3] are revealed in Section V. For the binary symmetric channel, very tight bounds on the minimum error probability achievable by a noncausal denoiser which only has access to the binary symmetric channel output are given 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 [5].

Additional related results, as well as proofs of all results stated below, are given in [6].

II. Basic Properties

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

$$H^-(X_1, \ldots X_n) \le H(X_1, \ldots X_n) \tag{4}$$

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

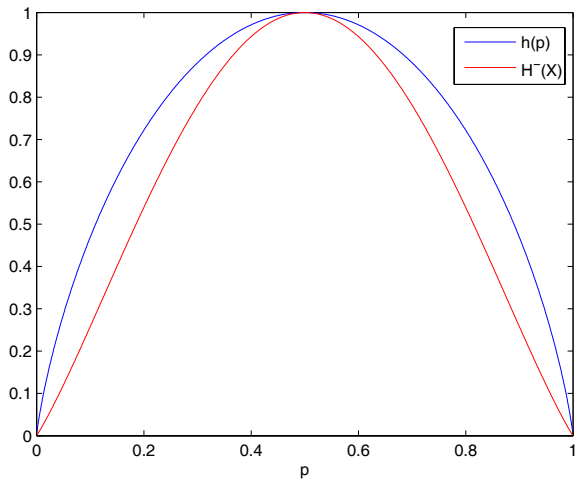

Fig. 1. Entropy rate and erasure entropy rate of a binary Markov chain with transition probability p.

Note that unlike entropy, erasure entropy is not associative. For example, $H^-(X_1, X_2, X_3) \le H^-((X_1, X_2), X_3)$, 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

$$\begin{aligned} H^-(\mathbf{X}) &= \lim_{n\to\infty} \frac{1}{n} H^-(X_1, \ldots X_n) && (5)\\ &= \lim_{n\to\infty} H(X_0 | X_{-n}^{-1}, X_1^n) && (6)\\ &= H(X_0 | X_{-\infty}^{-1}, X_1^{\infty}). && (7) \end{aligned}$$

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 example reveals.

Example 1: Let Z_i be iid with $Z_i \stackrel{d}{=} Z$. Let $Y_{2i} = Y_{2i+1} = Z_i$. 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 $\{Z_i\}$. The source $\{X_i\}$ is stationary and ergodic. Its entropy rate, $H(\mathbf{X})$, is readily seen to be given by $\frac{1}{2}H(Z)$. On the other hand, the erasure entropy of $\{X_i\}$ is 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).

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

$$H(\mathbf{X}) = h(p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p} \tag{8}$$

whereas,

$$H^-(\mathbf{X}) = h^-(p) \stackrel{\text{def}}{=} 2h(p) - h(2p(1-p)) \tag{9}$$

The entropy and erasure entropy of the first-order homogeneous binary symmetric Markov chain are shown in Figure 1.

It is interesting to note that in Example 2

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

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

Theorem 3: For a homogeneous kth-order Markov source

$$H(\mathbf{X})=\frac{H^-(\mathbf{X})+H(X_1,\dots X_k|X_{-1},\dots,X_{-k})}{k+1}. \quad (10)$$

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

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

Theorem 4: 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,

$$H(\mathbf{X}|\mathbf{Z}) \geq \mathsf{e}H^-(\mathbf{X}) \text{ for all } \mathsf{e}\in[0,1] \quad (11)$$

$$H(\mathbf{X}|\mathbf{Z}) = \mathsf{e}H^-(\mathbf{X})+o(\mathsf{e}), \quad (12)$$

where $o(\mathsf{e})/\mathsf{e}\to 0$ as $\mathsf{e}\to 0$.

If $\mathbf{X}$ is ergodic, the Slepian-Wolf theorem [7], and its extension to stationary and ergodic sources [8], 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. [9].

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

Example 3: 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 λ; thus, the switching times are a Poisson point process of rate λ. 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

$$p=P[\text{Poisson}(\lambda T/n) \text{ is odd}]=\frac{1-e^{-2\lambda T/n}}{2}. \quad (13)$$

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.

IV. Lossy Compression

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 erasure to obtain a distortion D under some distortion criterion.

More formally, a scheme for blocklength n and rate R consists of an encoder, which is a mapping $T:\mathcal{X}^n\times\mathcal{Z}^n\to\{1,\dots,\lfloor 2^{n\mathsf{e}R}\rfloor\}$, and a decoder, which is a sequence of mappings $\{\hat{X}_i\}_{i=1}^n$, where $\hat{X}_i:\{1,\dots,\lfloor 2^{n\mathsf{e}R}\rfloor\}\times\mathcal{Z}^n\to\hat{\mathcal{X}}$. 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 is measured by its expected distortion per erased symbol according to a given distortion measure $\rho:\mathcal{X}\times\hat{\mathcal{X}}\to\mathbb{R}$.

A rate distortion pair (R,D) is achievable if for every $\varepsilon>0$ and sufficiently large n there exists a scheme for blocklength n and rate R with

$$E\left[\frac{1}{|\{1\le i\le n:Z_i=e\}|}\sum_{1\le i\le n:Z_i=e}\rho(X_i,\hat{X}_i)\right]\le D+\varepsilon. \quad (14)$$

The rate distortion function $R_{\mathsf{e}}(D)$ is the infimum of rates R such that (R,D) is achievable. $R_{\mathsf{e}}(D)$ is the minimum amount of information required per erasure to achieve expected distortion of D per erasure.

Definition 3: For a stationary source $\mathbf{X}$, define the *erasure rate-distortion function*

$$R^-(D) = \min I(X_0;Y_0|X_{\setminus 0}) \quad (15)$$

$$= H^-(\mathbf{X})-\max H(X_0|X_{\setminus 0},Y_0), \quad (16)$$

where the minimum in (15) is over all $P_{Y_0|X_{-\infty}^{\infty}}$ such that $E[\rho(X_0,Y_0)]\le D$.

Theorem 5: If $\mathbf{X}$ is a stationary ergodic Markov source, of arbitrary order, then

$$\lim_{\mathsf{e}\to 0}R_{\mathsf{e}}(D)=R^-(D). \quad (17)$$

B. 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. [10]) states that for any stationary and ergodic process $\mathbf{X}$

$$R(D)\ge H(\mathbf{X})-\phi(D), \quad (18)$$

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

$$\phi(D) = \max_{N:E[d(N)]\leq D} H(N), \tag{19}$$

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 ϕ, which is a consequence of the concavity of entropy. Equality in (18) holds if and only if $\mathbf{X}$ has the decomposition

$$X_i = Y_i + N_i, \tag{20}$$

where $\mathbf{N}$ is an iid process with components achieving the maximum in (19), 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 (19) and define

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

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

Theorem 6: For any stationary source $\mathbf{X}$

$$R^-(D) \geq H^-(\mathbf{X}) - \phi(D) \tag{21}$$

with equality if and only if

$$P_{X_0|X_{\setminus 0}} \in \mathcal{S}(D) \quad a.s. \tag{22}$$

Note that whenever the source $\mathbf{X}$ has the decomposition in (20), it certainly satisfies (22), since (20) in particular exhibits a joint distribution under which

$$P_{X_0|X_{\setminus 0}} = P_{Y_0|X_{\setminus 0}} * P_{N_d} \quad a.s. \tag{23}$$

Thus Theorem 6 implies that when the distortion level D is such that the SLB for $R(D)$ holds with equality, the SLB for $R^-(D)$ also holds with equality. The opposite implication does not hold as the examples below show. Furthermore, there are cases (cf. [11]) where the SLB for $R(D)$ is known to hold with equality for a distortion region of the form $D \leq D^*$, where D^* is not explicitly known, while the larger threshold value for tightness of the SLB for $R^-(D)$ may be explicitly characterizable, as in examples to follow.

C. $R^-(D)$ for Binary Sources with Hamming Distortion

Consider now $R^-(D)$ for a binary source, under Hamming loss. As we now show, $R^-(D)$ can be given rather explicitly in parametric form for a general process. For $p \in [0,1]$, let $R_b(p,D)$ denote the rate distortion function of the Bernoulli(p) source:

$$R_b(p,D) = \begin{cases} h(p) - h(D) & \text{for } 0 \leq D \leq \min\{p, 1-p\} \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

Theorem 7: Let $\mathbf{X}$ be a binary stationary source and define the $[0,1/2]$-valued random variable

$$U = \min\{P(X_0 = 1|X_{\setminus 0}), P(X_0 = 0|X_{\setminus 0})\}. \tag{25}$$

The erasure rate-distortion function $R^-(D)$ for the source $\mathbf{X}$ is given in parametric form by

$$D(\Delta) = E[\min\{U, \Delta\}] = \int_{[0,\Delta]} u dF_U(u) + \Delta[1 - F_U(\Delta)] \tag{26}$$

and

$$R(\Delta) = E[R_b(U, \Delta)] = \int_{(\Delta,1/2]} h(u)\, dF_U(u) - h(\Delta)[1 - F_U(\Delta)], \tag{27}$$

where F_U is the CDF of U and $\Delta \in [0, \operatorname{esssup} U]$.

Remarks:

- Theorem 6, applied to the binary case, implies that the SLB for $R^-(D)$ is tight for $0 \leq D \leq \operatorname{essinf} U$. This is consistent with Theorem 7 as, for $0 \leq \Delta \leq \operatorname{essinf} U$, $D(\Delta) = \Delta$ and $R(\Delta) = H^-(\mathbf{X}) - h(\Delta)$.
- If $\mathbf{X}$ is a kth-order Markov source, U, as defined in (25), is discrete, assuming at most 2^{2k} different values. The characterization in Theorem 7 gives $R^-(D)$ *explicitly* for any such source. This is in contrast to the case for $R(D)$, which is not explicitly known even for the binary symmetric first-order Markov process [11].
- If $\mathbf{X}$ 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 approximated arbitrarily precisely by expressing it as a solution to an integral equation, of the type obtained by Blackwell in [12]. This then leads to arbitrarily precise approximations for $R^-(D)$ via the characterization in Theorem 7.

Example 4: 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 (25) is distributed as

$$U = \begin{cases} \frac{p_{min}^2}{p^2+(1-p)^2} & \text{w.p. } p^2 + (1-p)^2 \\ 1/2 & \text{w.p. } 2p(1-p). \end{cases} \tag{28}$$

With the distribution in (28), it is possible to solve the parametric equation in Theorem 7 to yield

$$R^-(D) = \begin{cases} h^-(p_{min}) - h(D) & \text{for } 0 \leq D \leq \frac{p_{min}^2}{p^2+(1-p)^2} \\ 2p(1-p)\left[1 - h\left(\frac{D-p_{min}^2}{2p(1-p)}\right)\right] & \frac{p_{min}^2}{p^2+(1-p)^2} < D \leq p_{min} \\ 0 & \text{otherwise.} \end{cases} \tag{29}$$

A plot of the erasure rate-distortion function $R^-(D)$ for the binary Markov chain with $p_{min} = 1/8$ is given in Figure 2. The SLB for $R^-(D)$ is tight up to $D^* = 0.02$, whereas the SLB for $R(D)$ is tight in a smaller unknown region. In this example, erasure entropy and entropy are $h^-(1/8) \approx 0.329$, and $h(1/8) \approx 0.544$, respectively.

$R^-(D)$ for Gaussian Sources

Example 5: Let $\mathbf{X}$ be a stationary Gaussian process with a bounded and strictly positive spectral density $S_{\mathbf{X}}(e^{jw})$. Then,

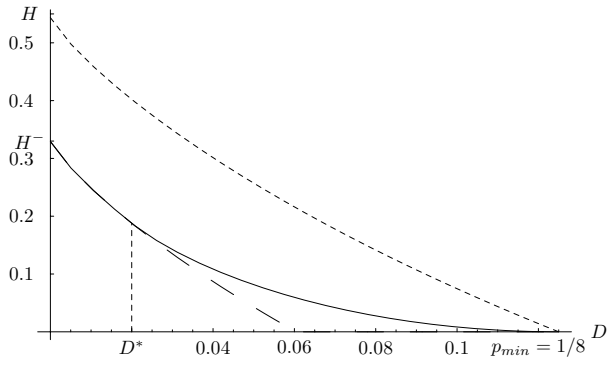

Fig. 2. Solid curve is $R^-(D)$ for the binary symmetric Markov source with $p_{min} = 1/8$. The dashed curve is the SLB for $R^-(D)$. The upper dotted curve is $R_b\,(1/8, D)$, the rate distortion function of the Bernoulli(1/8) source, which is also the SLB for the rate distortion function of the binary symmetric Markov source with $p_{min} = 1/8$.

under Euclidean distortion,

$$R^-(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2_{x^-}}{D} & \text{for } 0 \le D \le \sigma^2_{x^-} \\ 0 & D > \sigma^2_{x^-}, \end{cases} \tag{30}$$

where

$$\sigma^2_{x^-} = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{dw}{S_{\mathbf{X}}(e^{jw})}\right]^{-1}. \tag{31}$$

To see why this follows from Theorem 6 note that in this case X_0, conditioned on $X_{\setminus 0}$, is, with probability one, Gaussian of variance $\sigma^2_{x^-}$. So, in particular, $P_{X_0|X_{\setminus 0}} \in \mathcal{S}(D)$ a.s. for every $D \le \sigma^2_{x^-}$.

V. 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. DUDE [3] is a noncausal discrete universal denoiser which, upon knowledge of the channel transition matrix, exploits the information in the context of the noisy symbols to achieve the same performance as if it knew the statistics of the input. Empirically, it was observed in [3] that the compressibility of a denoised source is a good indication of the quality of reconstruction.

Denoising for binary symmetric channels under bit-error rate distortion turns out to be tightly coupled to entropy (in the causal setting) and to erasure entropy (in the noncausal setting). Defining

$$f_\delta(\alpha) = \min\left\{\frac{\alpha-\delta}{1-2\delta}, \delta\right\}, \tag{32}$$

let

$$\varepsilon_\delta = \min_{a,b}\max_{\delta\le\alpha\le 1/2}\left|f_\delta(\alpha) - [ah(\alpha)+b]\right|, \tag{33}$$

where h denotes the binary entropy function (in bits). Let a^*_δ, b^*_δ be the achievers of the minimum in (33).

For a stationary binary process $\mathbf{X} = \{X_t\}$ corrupted by a BSC(δ) let $\mathbb{F}(\mathbf{X}, \delta)$ and $\mathbb{D}(\mathbf{X}, \delta)$ denote, respectively, the filterability and denoisability (i.e., the minimum bit error rate in reproducing $\mathbf{X}$ from its noise-corrupted observation in the causal and non-causal case). For any stationary binary process $\mathbf{X} = \{X_t\}$ and $\delta \in [0, 1/2)$, it can be shown that

$$\left|\mathbb{F}(\mathbf{X}, \delta) - [a^*_\delta H(\mathbf{Z}) + b^*_\delta]\right| \le \varepsilon_\delta,$$

where $\mathbf{Z}$ denotes the response of the BSC(δ) to $\mathbf{X}$. This fact is used in [5] to bound the sensitivity of the filtering performance to the order in which a multi-dimensional 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 8: For any stationary binary process $\mathbf{X} = \{X_t\}$ and $\delta \in [0, 1/2)$,

$$\left|\mathbb{D}(\mathbf{X}, \delta) - \left[a^*_\delta H^-(\mathbf{Z}) + b^*_\delta\right]\right| \le \varepsilon_\delta$$

Thus, the entropy and erasure entropy determine the filterability and denoisability respectively to within ε_δ. This also implies that two noisy processes with the same erasure entropy rate can differ in their denoisability by no more than $2\varepsilon_\delta$. For example, $\varepsilon_{0.25} < 0.03$.

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

$$D^-(P_{X_1,\ldots X_n} \,||\, Q_{X_1,\ldots X_n}) = \sum_{i=1}^{n} D\left(P_{X_i|X_{\setminus i}} \parallel Q_{X_i|X_{\setminus i}} | P_{X_{\setminus i}}\right),$$

The erasure divergence is shown in [6], among other things, to play a key role in bounding the loss due to denoising a source using a denoiser which was tailored for a different source.

References

[1] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM J. Res. Dev.* vol. 5, pp. 183-191, 1961

[2] G. Casella and E. I. George, "Explaining the Gibbs sampler," *The American Statistician*, 46:167-174, 1992.

[3] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú and M. Weinberger, "Universal Discrete Denoising: Known Channel," *IEEE Trans. Information Theory*, vol. 51, no. 1, pp. 5-28, Jan. 2005.

[4] J. Yu and S. Verdú, "Universal Estimation of Erasure Entropy," *Proc. 2006 IEEE Int Symp. on Information Theory,* Seattle, WA, July 9-14, 2006

[5] A. Cohen, N. Merhav and T. Weissman, "Universal Scanning and Sequential Decision Making for Multi-Dimensional Data," *2006 IEEE Int Symp. on Information Theory,* Seattle, WA, July 9-14, 2006

[6] S. Verdú and T. Weissman, "Erasure Entropy," in preparation.

[7] D. Slepian and J. K. Wolf, "Noiseless Coding of correlated information sources," *IEEE Trans. Inform. Theory,*, vol. IT-19, pp. 471-480, 1973.

[8] T. M. Cover, A proof of the data compression theorem of Slepian and Wolf for ergodic sources, vol. IT-22, pp. 226228, Mar. 1975.

[9] H. Cai, S. Kulkarni and S. Verdú, "A Universal Lossless Compressor with Side Information based on Context Tree Weighting," *2005 IEEE Int. Symp. Information Theory*, Adelaide, Australia, Sep. 2005.

[10] T. Berger, "Rate Distortion Theory," Prentice-Hall, Englewood Cliffs, NJ, 1971

[11] R. M. Gray. "Rate distortion functions for finite-state finite-alphabet Markov sources," *IEEE Trans. Inform. Theory*, IT-17(2):127–134, March 1971.

[12] D. Blackwell, "The entropy of functions of finite-state Markov chains," "Trans. First Prague Conf. Information Theory: Statistical Decision Functions, Random Processes," pp. 1320, 1957.