The Empirical Distribution of Rate-Constrained Source Codes

Tsachy Weissman, Erik Ordentlich
HP Laboratories Palo Alto
HPL-2003-253
December 8th , 2003*

E-mail: tsachy@stanford.edu, eord@hpl.hp.com

rate-distortion theory, sample converse, denoising, empirical distribution

Let $\mathbf{X} = (X_{1, \ldots})$ be a stationary ergodic finite-alphabet source, X^n denote its first n symbols, and Y^n be the codeword assigned to X^n by a lossy source code. The empirical kth-order joint distribution $\hat{Q}^k[X^n, Y^n](x^k, y^k)$ is defined as the frequency of appearances of pairs of k-strings (x^k, y^k) along the pair (X^n, Y^n). Our main interest is in the sample behavior of this (random) distribution. Letting $I(Q^k)$ denote the mutual information $I(X^k; Y^k)$ when $(X^k, Y^k) \sim Q^k$ we show that *for any* (sequence of) lossy source code(s) of rate $\leq R$

$$\limsup_{n\to\infty} \frac{1}{k} I(\hat{Q}^k[X^n, Y^n]) \leq R + \frac{1}{k} H(X_1^k) - \overline{H}(\mathbf{X}) \text{ a.s.}$$

where $\overline{H}(\mathbf{X})$ denotes the entropy rate of $\mathbf{X}$. This is shown to imply, for a large class of sources including all i.i.d. sources and all sources satisfying the Shannon lower bound with equality, that for any sequence of codes which is *good* in the sense of asymptotically attaining a point on the rate distortion curve

$$\hat{Q}^k[X^n, Y^n] \stackrel{d}{\Rightarrow} P_{X^k, \hat{Y}^k} \text{ a.s.},$$

whenever $P_{X^k, \hat{Y}^k}$ is the unique distribution attaining the minimum in the definition of the kth-order rate distortion function. Further consequences of these results are explored. These include a simple proof of Kieffer's sample converse to lossy source coding, as well as pointwise performance bounds for compression-based denoisers.

* Internal Accession Date Only
Approved for External Publication

The Empirical Distribution of Rate-Constrained Source Codes

Tsachy Weissman* Erik Ordentlich†

November 16, 2003

Abstract

Let $\mathbf{X} = (X_1, \ldots)$ be a stationary ergodic finite-alphabet source, X^n denote its first n symbols, and Y^n be the codeword assigned to X^n by a lossy source code. The empirical kth-order joint distribution $\hat{Q}^k[X^n, Y^n](x^k, y^k)$ is defined as the frequency of appearances of pairs of k-strings (x^k, y^k) along the pair (X^n, Y^n). Our main interest is in the sample behavior of this (random) distribution. Letting $I(Q^k)$ denote the mutual information $I(X^k; Y^k)$ when $(X^k, Y^k) \sim Q^k$ we show that *for any* (sequence of) lossy source code(s) of rate $\leq R$

$$\limsup_{n\to\infty} \frac{1}{k} I\left(\hat{Q}^k[X^n, Y^n]\right) \leq R + \frac{1}{k} H(X_1^k) - \overline{H}(\mathbf{X}) \quad a.s.,$$

where $\overline{H}(\mathbf{X})$ denotes the entropy rate of $\mathbf{X}$. This is shown to imply, for a large class of sources including all i.i.d. sources and all sources satisfying the Shannon lower bound with equality, that for any sequence of codes which is *good* in the sense of asymptotically attaining a point on the rate distortion curve

$$\hat{Q}^k[X^n, Y^n] \stackrel{d}{\Rightarrow} P_{X^k, \tilde{Y}^k} \quad a.s.,$$

whenever $P_{X^k, \tilde{Y}^k}$ is the unique distribution attaining the minimum in the definition of the kth-order rate distortion function. Further consequences of these results are explored. These include a simple proof of Kieffer's sample converse to lossy source coding, as well as pointwise performance bounds for compression-based denoisers.

1 Introduction

Loosely speaking, a rate distortion code sequence is considered good for a given source if it attains a point on its rate distortion curve. The existence of good code sequences with empirical distributions close to those achieving the minimum mutual information in the definition of the rate distortion function is a consequence of the random coding argument at the heart of the achievability part of rate distortion theory. It turns out, however, in ways that we quantify in this work, that *any* good code sequence must have this property.

This behavior of the empirical distribution of good rate distortion codes is somewhat analogous to that of good channel codes which was characterized by Shamai and Verdú in [SV97]. Defining the kth-order empirical distribution of a channel code as the proportion of k-strings anywhere in the codebook equal to every given k-string, this empirical distribution was shown to converge to the capacity-achieving channel input distribution. The analogy is more than merely qualitative. For example, the growth rate of k with n where this convergence was shown in [SV97] to break down will be seen to have an analogue in our setting. A slight difference between the problems is that in the setting of [SV97] the *whole* codebook of a good code was shown to be well-behaved in the sense that the empirical distribution

*T. Weissman is with the Electrical Engineering Department, Stanford University, CA 94305 USA (e-mail: tsachy@stanford.edu). Part of this work was done while T. Weissman was visiting Hewlett-Packard Laboratories, Palo Alto, CA, USA. T. Weissman was partially supported by NSF grants DMS-0072331 and CCR-0312839.

†E. Ordentlich is with Hewlett-Packard Laboratories, Palo Alto, CA 94304 USA (e-mail: eord@hpl.hp.com).

is defined as an average over *all* codewords. In the source coding analogue, it is clear that no meaningful statement can be made on the empirical distribution obtained when averaging and giving equal weights to *all* codewords in the codebook. The reason is that any approximately good codebook will remain approximately good when appending to it a polynomial number of additional codebooks of the same size, so essentially any empirical distribution induced by the new codebook can be created while maintaining the "goodness" of the codebook. We therefore adopt in this work what seems to be a more meaningful analogue to the empirical distribution notion of [SV97], namely the empirical distribution of the codeword corresponding to the particular source realization, or the joint empirical distribution of source realization and its associated codeword. We shall show several strong probabilistic senses in which the latter entities converge to the distributions attaining the minimum in the associated rate distortion function.

The basic phenomenon we quantify in this work is that a rate constraint on a sequence of codes imposes a severe limitation on the mutual information associated with its empirical distribution. This is independent of the notion of distortion, or goodness of a code. The described behavior of the empirical distribution of good codes is a consequence of it.

Our work consists of two main parts. The first part, Section 3, considers the distribution of the pair of k-tuples obtained by looking at the source-reconstruction pair (X^n, Y^n) through a k-window at a uniformly sampled location. More specifically, we look at the distribution of $(X_I^{I+k-1}, Y_I^{I+k-1})$, where I is chosen uniformly from $\{1, \ldots, n-k+1\}$ (independently of everything else). We refer to this distribution as the "kth-order marginalization of the distribution of (X^n, Y^n)". In subsection A we show that the normalized mutual information between X_I^{I+k-1} and Y_I^{I+k-1} is esentially[1] upper bounded by $R + \frac{1}{k}H(X^k) - \overline{H}(\mathbf{X})$, R denoting the rate of the code and $\overline{H}(\mathbf{X})$ the source entropy rate. This is seen in subsection B to imply the convergence in distribution of $(X_I^{I+k-1}, Y_I^{I+k-1})$ to the joint distribution attaining $R(X^k, D)$ (the kth-order rate distortion function) whenever it is unique and satisfies the relation $R(X^k, D) = R(\mathbf{X}, D) + \frac{1}{k}H(X^k) - \overline{H}(\mathbf{X})$ (with $R(\mathbf{X}, D)$ denoting the rate distortion function of the source). This includes all i.i.d. sources, as well as the large family of sources for which the Shannon lower bound holds with equality (cf. [Gra70, Gra71, BG98] and references therein). In subsection C we show that convergence in any reasonable sense cannot hold for *any* code sequence when k increases with n such that $k/n > R(\mathbf{X}, D)/\overline{H}(\tilde{\mathbf{Y}})$, where $\overline{H}(\tilde{\mathbf{Y}})$ is the entropy rate of the reconstruction process attaining the minimum mutual information defining the rate distortion function. In subsection D we apply the results to obtain performance bounds for compression-based denoising. We extend results from [Don02] by showing that any additive noise distribution induces a distortion measure such that if optimal lossy compression of the noisy signal is performed under it, at a distortion level matched to the level of the noise, then the marginalized joint distribution of the noisy source and the reconstruction converges to that of the noisy source and the underlying clean source. This is shown to lead to bounds on the performance of such compression-based denoising schemes.

Section 4 presents the primary part of our work, which consists of pointwise analogues to the results of Section 3. More specifically, we look at properties of the (random) empirical kth-order joint distribution $\hat{Q}^k[X^n, Y^n]$ induced by the source-reconstruction pair of n-tuples (X^n, Y^n). The main result of subsection A (Theorem 7) asserts that

[1] Up to an $o(1)$ term, independent on the particular code.

$R+\frac{1}{k}H(X^k)-\overline{H}(\mathbf{X})$ is not only an upper bound on the normalized mutual information between X_I^{I+k-1} and Y_I^{I+k-1} (as established in Section 3), but is essentially[2], with probability one, an upper bound on the normalized mutual information under $\hat{Q}^k[X^n, Y^n]$. This is seen in subsection B to imply the sample converse to lossy source coding of [Kie91], avoiding the use of the ergodic theorem in [Kie89]. In subsection C this is used to establish the almost sure convergence of $\hat{Q}^k$ to the joint distribution attaining the kth-order rate distortion function of the source, under the conditions stipulated for the convergence in the setting of Section 3. In subsection D we apply these almost sure convergence results to derive a pointwise analogue of the performance bound for compression-based denoising derived in subsection 3.D. In subsection E we show that a simple post-processing "derandomization" scheme performed on the output of the previously analyzed compression-based denoisers results in essentially optimum denoising performance.

The empirical distribution of good lossy source codes was first considered by Kanlis, Khudanpur and Narayan in [KSP96]. For memoryless sources they showed that any good sequence of codebooks must have an exponentially non-negligible fraction of codewords which are typical with respect to the $R(D)$-achieving output distribution. It was later further shown in [Kan97] that this subset of typical codewords carries most of the probabilistic mass in that the probability of a source word having a non-typical codeword is negligible. That work also sketched a proof of the fact that the source word and its reconstruction are, with high probability, jointly typical with respect to the joint distribution attaining $R(D)$. More recently, Donoho established distributional properties of good rate distortion codes for certain families of processes and applied them to performance analysis of compression based denoisers [Don02]. The main innovation in the parts of the present work related to good source codes is in considering the pointwise behavior of any kth-order joint empirical distribution for general stationary ergodic processes.

Other than sections 3 and 4 which were detailed above, we introduce notation and conventions in Section 2, and summarize the paper with a few of its open directions in Section 5.

2 Notation and Conventions

$\mathcal{X}$ and $\mathcal{Y}$ will denote, respectively, the source and reconstruction alphabets which we assume throughout to be finite. We shall also use the notation x^n for $(x_1, \ldots, x_n)$ and $x_i^j = (x_i, \ldots x_j)$.

We denote the set of probability measures on a finite set $\mathcal{A}$ by $\mathcal{M}(\mathcal{A})$. For $P, P' \in \mathcal{M}(\mathcal{A})$ we let $\|P - P'\| = \max_{a\in\mathcal{A}} |P(a) - P'(a)|$. $B(P, \delta)$ will denote the l_∞ ball around P, i.e.,

$$B(P, \delta) = \{P' : \|P' - P\| \le \delta\}. \tag{1}$$

For $\{P_n\}$, $P_n, P \in \mathcal{M}(\mathcal{A})$, $P_n \stackrel{d}{\Rightarrow} P$, $P_n \Rightarrow P$, and $P_n \to P$ will all stand for

$$\lim_{n\to\infty} \|P_n - P\| = 0. \tag{2}$$

For $Q \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$, $I(Q)$ will denote the mutual information $I(X;Y)$ when $(X, Y) \sim Q$. Similarly, for $Q^k \in \mathcal{M}(\mathcal{X}^k \times \mathcal{Y}^k)$, $I(Q^k)$ will stand for $I(X^k;Y^k)$ when $(X^k, Y^k) \sim Q^k$. For $P \in \mathcal{M}(\mathcal{A})$ and a stochastic matrix W from $\mathcal{A}$ into $\mathcal{B}$, $P \otimes W \in \mathcal{M}(\mathcal{A} \times \mathcal{B})$ is the distribution on (A, B) under which $\Pr(A = a, B = b) = P(a)W(b|a)$.

[2] In the limit of large n.

$H_{P\otimes W}(X|Y)$ and $H(W|P)$ will both denote $H(X|Y)$ when $(X,Y)\sim P\otimes W$. $I(P;W)$ will stand for $I(X;Y)$ when $(X,Y)\sim P\otimes W$. When dealing with expectations of functions or with functionals of random variables, we shall sometimes subscript the distributions of the associated random variables. Thus, for example, for any $f:\mathcal{A}\to\mathbb{R}$ and $P\in\mathcal{M}(\mathcal{A})$ we shall write $E_P f(A)$ for the expectation of $f(A)$ when A is distributed according to P.

$\mathcal{M}_n(\mathcal{A})$ will denote the subset of $\mathcal{M}(\mathcal{A})$ consisting of n-th-order empirical types, i.e., $P\in\mathcal{M}_n(\mathcal{A})$ if and only if $P(a)$ is an integer multiple of $1/n$ for all $a\in\mathcal{A}$. For any $P\in\mathcal{M}(\mathcal{X})$, $\mathcal{C}_n(P)$ will denote the set of stochastic matrices (conditional distributions) from $\mathcal{X}$ into $\mathcal{Y}$ for which $P\otimes W\in\mathcal{M}_n(\mathcal{X}\times\mathcal{Y})$.

For $P\in\mathcal{M}(\mathcal{A})$ we let T_P^n, or simply T_P, denote the associated type class, i.e., the set of sequences $u^n\in\mathcal{A}^n$ with

$$\frac{1}{n}|\{1\le i\le n: u_i=a\}|=P(a)\quad\forall a\in\mathcal{A}. \tag{3}$$

$\mathcal{M}_n(\mathcal{A})$ will denote the set of $P\in\mathcal{M}(\mathcal{A})$ for which $T_P^n\neq\emptyset$. For $\delta\ge 0$, $T_{[P]_\delta}^n$, or simply $T_{[P]_\delta}$, will denote the set of sequences $u^n\in\mathcal{A}^n$ with

$$\left|\frac{1}{n}|\{1\le i\le n: u_i=a\}|-P(a)\right|\le\delta\quad\forall a\in\mathcal{A}. \tag{4}$$

For a random element, P subscripted by the element will denote its distribution. Thus, for example, for the process $\mathbf{X}=(X_1,X_2,\ldots)$, P_{X_1}, P_{X^n} and $P_{\mathbf{X}}$ will denote, respectively, the distribution of its first component, the distribution of its first n components, and the distribution of the process itself. If, say, (X_0,Z_i^j) are jointly distributed then $P_{X_0|z_i^j}\in\mathcal{M}(\mathcal{X})$ will denote the conditional distribution of X_0 given $Z_i^j=z_i^j$. This will also hold for the cases with $i=-\infty$ and\or $j=\infty$ by assuming $P_{X_0|z_i^j}$ to be a regular version of the conditional distribution of X_0 given Z_i^j, evaluated at z_i^j. For a sequence of random variables $\{X_n\}$, $X_n\stackrel{d}{\Rightarrow}X$ will stand for $P_{X_n}\stackrel{d}{\Rightarrow}P_X$.

For $x^n\in\mathcal{X}^n$ and $y^n\in\mathcal{Y}^n$ we let $\hat{Q}^k[x^n]\in\mathcal{M}(\mathcal{X}^k)$, $\hat{Q}^k[y^n]\in\mathcal{M}(\mathcal{Y}^k)$, and $\hat{Q}^k[x^n,y^n]\in\mathcal{M}(\mathcal{X}^k\times\mathcal{Y}^k)$ denote the k-th order empirical distributions defined by

$$\hat{Q}^k[x^n](u^k)=\frac{1}{n-k+1}|\{0\le i\le n-k: x_{i+1}^{i+k}=u^k\}|, \tag{5}$$

$$\hat{Q}^k[y^n](v^k)=\frac{1}{n-k+1}|\{0\le i\le n-k: y_{i+1}^{i+k}=v^k\}|, \tag{6}$$

and

$$\hat{Q}^k[x^n,y^n](u^k,v^k)=\frac{1}{n-k+1}|\{0\le i\le n-k: x_{i+1}^{i+k}=u^k, y_{i+1}^{i+k}=v^k\}|. \tag{7}$$

$\hat{Q}^k[y^n|x^n]$ will denote the conditional distribution $P_{Y^k|X^k}$ when $(X^k,Y^k)\sim\hat{Q}^k[x^n,y^n]$. In accordance with notation defined above, for example, $E_{\hat{Q}^k[x^n,y^n]}f(X^k,Y^k)$ will denote expectation of $f(X^k,Y^k)$ when $(X^k,Y^k)\sim\hat{Q}^k[x^n,y^n]$.

Definition 1 *A fixed-rate n-block code is a pair (C_n,ϕ_n) where $C_n\subseteq\mathcal{Y}^n$ is the code-book and $\phi_n:\mathcal{X}^n\to C_n$. The rate of the block-code is given by $\frac{1}{n}\log|C_n|$. The rate of a sequence of block codes is defined by*

$$\limsup_{n\to\infty}\frac{1}{n}\log|C_n|.$$

$Y^n=Y^n(X^n)=\phi_n(X^n)$ will denote the reconstruction sequence when the n-block code is applied to the source sequence X^n.

Definition 2 *A variable-rate code for n-blocks is a triple* (C_n, ϕ_n, l_n) *with* C_n *and* ϕ_n *as in Definition 1. The code operates by mapping a source n-tuple* X^n *into* C_n *via* ϕ_n*, and then encoding the corresponding member of the code-book (denoted* Y^n *in Definition 1) using a uniquely decodable binary code. Letting* $l_n(X^n)$ *denote the associated length function, the rate of the code is defined by* $E\frac{1}{n}l_n(X^n)$ *and the rate of a sequence of codes for the source* $\mathbf{X} = (X_1, X_2, \ldots)$ *is defined by* $\limsup_{n\to\infty} E\frac{1}{n}l_n(X^n)$.

Note that the notion of a code in the above definitions does not involve a distortion measure according to which goodness of reconstruction is judged.

We assume throughout a given single-letter loss function $\rho : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ satisfying

$$\forall x \in \mathcal{X} \quad \exists y \in \mathcal{Y} \quad \text{s.t.} \quad \rho(x, y) = 0. \tag{8}$$

For $x^n \in \mathcal{X}^n$ and $y^n \in \mathcal{Y}^n$ we let $\rho_n(x^n, y^n) = \frac{1}{n}\rho(x_i, y_i)$. We shall also let $\delta_{Y|X}$ denote a conditional distribution of Y given X with the property that for all $x \in \mathcal{X}$

$$\sum_{y\in\mathcal{Y}} \delta_{Y|X}(y|x)\rho(x, y) = 0 \tag{9}$$

(note that (8) implies existence of at least one such conditional distribution). The rate distortion function associated with the random variable $X \in \mathcal{X}$ is defined by

$$R(X, D) = \min_{E\rho(X,Y)\le D} I(X; Y), \tag{10}$$

where the minimum is over all joint distributions of the pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ consistent with the given distribution of X. Letting ρ_n denote the (normalized) distortion measure between n-tuples induced by ρ,

$$\rho_n(x^n, y^n) = \frac{1}{n}\sum_{i=1}^{n} \rho(x_i, y_i), \tag{11}$$

the rate distortion function of $X^n \in \mathcal{X}^n$ is defined by

$$R(X^n, D) = \min_{E\rho_n(X^n,Y^n)\le D} \frac{1}{n} I(X^n; Y^n). \tag{12}$$

The rate distortion function of the stationary ergodic process $\mathbf{X} = (X_1, X_2, \ldots)$ is defined by

$$R(\mathbf{X}, D) = \lim_{n\to\infty} R(X^n, D). \tag{13}$$

Note that our assumptions on the distortion measure, together with the assumption of finite source alphabets, combined with its well-known convexity, imply that $R(\mathbf{X}, D)$ (as a function of $D \in [0, \infty)$) is

1. Non-negative valued and bounded.
2. Continuous.
3. Strictly decreasing in the range $0 \le D \le D^{\mathbf{X}}_{max}$, where $D^{\mathbf{X}}_{max} \stackrel{\triangle}{=} \min\{D : R(\mathbf{X}, D) = 0\}$ (and identically 0 for $D \ge D^{\mathbf{X}}_{max}$),

properties that will be tacitly relied on in proofs of some of the results. The set $\left\{(R(\mathbf{X},D),D) : 0 \le D \le D^{\mathbf{X}}_{max}\right\}$ will be referred to as the "rate distortion curve". The distortion rate function is the inverse of $R(\mathbf{X},\cdot)$, which we formally define (to account for the trivial regions) as

$$D(\mathbf{X},R) = \begin{cases} \text{unique value of } D \text{ such that } R(\mathbf{X},D) = R & \text{for } R \in (0, R(\mathbf{X},0)] \\ D^{\mathbf{X}}_{max} & \text{for } R = 0 \\ 0 & \text{for } R > R(\mathbf{X},0). \end{cases} \tag{14}$$

3 Distributional Properties of Rate-Constrained Codes

Throughout this section "codes" can be taken to mean either fixed- or variable-rate codes.

A Bound on the Mutual Information Induced by a Rate-Constrained Code

The following result and its proof are implicit in the proof of the converse of [DW03, Theorem 1].

Theorem 1 *Let $\mathbf{X} = (X_1, X_2 \ldots)$ be stationary and $Y^n = Y^n(X^n)$ be the reconstruction of an n-block code of rate $\le R$. Let J be uniformly distributed on $\{1,\ldots,n\}$, and independent of $\mathbf{X}$. Then*

$$I(X_J;Y_J) \le R + H(X_1) - \frac{1}{n}H(X^n). \tag{15}$$

In particular $I(X_J;Y_J) \le R$ when $\mathbf{X}$ is memoryless.

Proof:

$$\begin{aligned} H(Y^n) &\ge I(X^n;Y^n) \\ &= H(X^n) - H(X^n|Y^n) \\ &= H(X^n) - nH(X_1) + \sum_{i=1}^{n}[H(X_i) - H(X_i|X^{i-1},Y^n)] \\ &\ge H(X^n) - nH(X_1) + \sum_{i=1}^{n}[H(X_i) - H(X_i|Y_i)] \\ &= H(X^n) - nH(X_1) + \sum_{i=1}^{n} I(X_i;Y_i) \\ &\ge H(X^n) - nH(X_1) + nI(X_J;Y_J), \end{aligned} \tag{16}$$

where the last inequality follows from the facts that (by stationarity) for all i $X_J \stackrel{d}{=} X_i$, that $P_{Y_J|X_J} = \frac{1}{n}\sum_{i=1}^{n} P_{Y_i|X_i}$, and the convexity of the mutual information in the conditional distribution (cf. [CT91, Theorem 2.7.4]). The fact that Y^n is the reconstruction of an n-block code of rate $\le R$ implies $H(Y^n) \le nR$, completing the proof when combined with (16). □

Remark: It is readily verified that (15) remains true even when the requirement of a stationary $\mathbf{X}$ is relaxed to the requirement of equal first order marginals. To see that this cannot be relaxed much further note that when the source is an individual sequence the right side of (15) equals R. On the other hand, if this individual sequence is non-constant one can take a code of rate 0 consisting of one codeword whose joint empirical distribution with the source sequence has positive mutual information, violating the bound.

Theorem 2 *Let $\mathbf{X} = (X_1, X_2 \ldots)$ be stationary and $Y^n = Y^n(X^n)$ be the reconstruction of an n-block code of rate $\le R$. Let $1 \le k \le n$ and J be uniformly distributed on $\{1, \ldots, n-k+1\}$, and independent of $\mathbf{X}$. Then*

$$\frac{1}{k} I\left(X_J^{J+k-1}; Y_J^{J+k-1}\right) \le \frac{n}{n-k} R + \frac{1}{k} H(X^k) - \frac{1}{n} H(X^n) + \frac{2k}{n} \log |\mathcal{X}|. \tag{17}$$

Proof: Let, for $1 \le j \le k$, $S_j = \{1 \le i \le n : i = j \bmod k\}$. Note that $\frac{n}{k} - 1 \le |S_j| \le \frac{n}{k}$. Let $P^{(j)}_{X^k, Y^k}$ be the distribution defined by

$$P^{(j)}_{X^k, Y^k}(x^k, y^k) = \frac{1}{|S_j|} \sum_{i \in S_j} \Pr\{X_i^{i+k-1} = x^k, Y_i^{i+k-1} = y^k\}. \tag{18}$$

Note, in particular, that

$$P_{X_J^{J+k-1}, Y_J^{J+k-1}} = \sum_{j=1}^{k} \frac{|S_j|}{n} P^{(j)}_{X^k, Y^k}. \tag{19}$$

Now

$$\begin{aligned} nR &\ge H(Y^n) \\ &\ge H\left(Y_j^{j+k|S_j|-1}\right) \\ &\ge H\left(X_j^{j+k|S_j|-1}\right) - |S_j| H(X^k) + |S_j| I\left(P^{(j)}_{X^k, Y^k}\right), \end{aligned} \tag{20}$$

where the first inequality is due to the rate constraint of the code and the last one follows from the bound established in (16) with the assignment $n \to |S_j|$, $(X^n, Y^n) \to \left(X_j^{j+k|S_j|-1}, Y_j^{j+k|S_j|-1}\right)$, $(X_i, Y_i) \to \left(X_{j+(i-1)k}^{j+ik-1}, X_{j+(i-1)k}^{j+ik-1}\right)$. Rearranging terms

$$\begin{aligned} I\left(P^{(j)}_{X^k, Y^k}\right) &\le \frac{n}{|S_j|} R + H(X^k) - \frac{1}{|S_j|} H\left(X_j^{j+k|S_j|-1}\right) \\ &\le k\left[\frac{n}{n-k} R + \frac{1}{k} H(X^k) - \frac{1}{n} H(X^n) + \frac{2k}{n} \log |\mathcal{X}|\right], \end{aligned} \tag{21}$$

where in the second inequality we have used the fact that

$$H(X^n) \le H\left(X_j^{j+k|S_j|-1}\right) + H(X^{j-1}, X_{j+k|S_j|}^n) \le H\left(X_j^{j+k|S_j|-1}\right) + 2k \log |\mathcal{X}|.$$

Inequality (17) now follows from (21), (19), the fact that $P_{X_J^{J+k-1}} = P^{(j)}_{X^k}$ (both equaling the distribution of a source k-tuple) for all $1 \le j \le k$, and the convexity of the mutual information in the conditional distribution. □

An immediate consequence of Theorem 2 is

Corollary 1 *Let $\mathbf{X} = (X_1, X_2 \ldots)$ be stationary and $\{Y^n(\cdot)\}$ be a sequence of block codes of rate $\le R$. For $1 \le k \le n$ let $J^{(n,k)}$ be uniformly distributed on $\{1, \ldots, n-k+1\}$, and independent of $\mathbf{X}$. Consider the pair $(X^n, Y^n) = (X^n, Y^n(X^n))$ and denote*

$$\left(\tilde{X}^{(n,k)}, \tilde{Y}^{(n,k)}\right) = \left(X_{J^{(n,k)}}^{J^{(n,k)}+k-1}, Y_{J^{(n,k)}}^{J^{(n,k)}+k-1}\right). \tag{22}$$

Then

$$\limsup_{k \to \infty} \limsup_{n \to \infty} \frac{1}{k} I\left(\tilde{X}^{(n,k)}; \tilde{Y}^{(n,k)}\right) \le R. \tag{23}$$

Proof: Theorem 2 implies

$$\limsup_{n \to \infty} \frac{1}{k} I\left(\tilde{X}^{(n,k)}; \tilde{Y}^{(n,k)}\right) \le R + \frac{1}{k} H(X^k) - \overline{H}(\mathbf{X}), \tag{24}$$

completing the proof by letting $k \to \infty$. □

B The Empirical Distribution of Good Codes

The converse in rate distortion theory (cf. [Ber71], [Gal68]) asserts that if $\mathbf{X}$ is stationary ergodic and (R, D) is a point on its rate distortion curve then any code sequence of rate $\leq R$ must satisfy

$$\liminf_{n\to\infty} E\rho_n(X^n, Y^n(X^n)) \geq D. \tag{25}$$

The direct part, on the other hand, guarantees the existence of codes that are good in the following sense.

Definition 3 *Let $\mathbf{X}$ be stationary ergodic, $0 \leq D \leq D_{max}^{\mathbf{X}}$, and (R, D) be a point on its rate distortion curve. The sequence of codes (with associated mappings $\{Y^n(\cdot)\}$) will be said to be good for the source $\mathbf{X}$ at rate R (or at distortion level D) if it has rate $\leq R$ and*

$$\limsup_{n\to\infty} E\rho_n(X^n, Y^n(X^n)) \leq D. \tag{26}$$

Note that the converse to rate distortion coding implies that the limit supremum in the above definition is actually a limit and the rate of the corresponding good code sequence is $= R$.

Theorem 3 *Let $\{Y^n(\cdot)\}$ correspond to a good code sequence for the stationary ergodic source $\mathbf{X}$ and $\left(\tilde{X}^{(n,k)}, \tilde{Y}^{(n,k)}\right)$ be defined as in Corollary 1. Then:*

1. *For every k*

$$\lim_{n\to\infty} E\rho_k\left(\tilde{X}^{(n,k)}, \tilde{Y}^{(n,k)}\right) = D, \tag{27}$$

$$R(X^k, D) \leq \liminf_{n\to\infty} \frac{1}{k} I\left(\tilde{X}^{(n,k)}; \tilde{Y}^{(n,k)}\right) \leq \limsup_{n\to\infty} \frac{1}{k} I\left(\tilde{X}^{(n,k)}; \tilde{Y}^{(n,k)}\right) \leq R(\mathbf{X}, D) + \frac{1}{k} H(X^k) - \overline{H}(\mathbf{X}) \tag{28}$$

   so, in particular,

$$\lim_{k\to\infty} \liminf_{n\to\infty} \frac{1}{k} I\left(\tilde{X}^{(n,k)}; \tilde{Y}^{(n,k)}\right) = \lim_{k\to\infty} \limsup_{n\to\infty} \frac{1}{k} I\left(\tilde{X}^{(n,k)}; \tilde{Y}^{(n,k)}\right) = R(\mathbf{X}, D). \tag{29}$$

2. *If it also holds that $R(X^k, D) = R(\mathbf{X}, D) + \frac{1}{k} H(X_1^k) - \overline{H}(\mathbf{X})$ then*

$$\lim_{n\to\infty} \frac{1}{k} I\left(\tilde{X}^{(n,k)}; \tilde{Y}^{(n,k)}\right) = R(X^k, D). \tag{30}$$

3. *If, additionally, $R(X^k, D)$ is uniquely achieved (in distribution) by the pair $(\tilde{X}^k, \tilde{Y}^k)$ then*

$$\left(\tilde{X}^{(n,k)}, \tilde{Y}^{(n,k)}\right) \stackrel{d}{\Rightarrow} (\tilde{X}^k, \tilde{Y}^k) \quad as \quad n \to \infty. \tag{31}$$

The condition in the second item clearly holds for any memoryless source. It holds, however, much beyond memoryless sources. Any source for which the Shannon lower bound holds with equality is readily seen to satisfy $R(X^k, D) = R(\mathbf{X}, D) + \frac{1}{k} H(X_1^k) - \overline{H}(\mathbf{X})$ for all k. This is a rich family of sources that includes Markov processes, hidden Markov processes, auto-regressive sources, and Gaussian processes (in the continuous alphabet setting), cf. [Gra70, Gra71, BG98, EM02, WM03].

The first part of Theorem 3 will be seen to follow from Theorem 2. Its last part will be a consequence of the following lemma, whose proof is given in Appendix A.

Lemma 1 *Let $R(D)$ be uniquely achieved by the joint distribution $P_{X,Y}$. Let further $\{P_{X,Y}^{(n)}\}$ be a sequence of distributions on (X,Y) satisfying*

$$P_X^{(n)} \stackrel{d}{\Rightarrow} P_X, \tag{32}$$

$$\lim_{n\to\infty} I\left(P_{X,Y}^{(n)}\right) = R(D), \tag{33}$$

and

$$\lim_{n\to\infty} E_{P_{X,Y}^{(n)}} \rho(X,Y) = D. \tag{34}$$

Then

$$P_{X,Y}^{(n)} \stackrel{d}{\Rightarrow} P_{X,Y}. \tag{35}$$

Proof of Theorem 3: For any code sequence $\{Y^n(\cdot)\}$, an immediate consequence of the definition of $\left(\tilde{X}^{(n,k)}, \tilde{Y}^{(n,k)}\right)$ is that

$$\lim_{n\to\infty} \left[E\rho_k\left(\tilde{X}^{(n,k)}, \tilde{Y}^{(n,k)}\right) - E\rho_n(X^n, Y^n)\right] = 0. \tag{36}$$

Combined with the fact that $\{Y^n(\cdot)\}$ is a *good* code this implies

$$\lim_{n\to\infty} E\rho_k\left(\tilde{X}^{(n,k)}, \tilde{Y}^{(n,k)}\right) = \lim_{n\to\infty} E\rho_n(X^n, Y^n) = D, \tag{37}$$

proving (27). To prove (28) note that since $\{Y^n(\cdot)\}$ is a sequence of codes for $\mathbf{X}$ at rate $R(\mathbf{X}, D)$, Theorem 2 implies

$$\limsup_{n\to\infty} \frac{1}{k} I\left(\tilde{X}^{(n,k)}; \tilde{Y}^{(n,k)}\right) \le R(\mathbf{X}, D) + \frac{1}{k} H(X^k) - \overline{H}(\mathbf{X}). \tag{38}$$

On the other hand, since $\tilde{X}^{(n,k)} \stackrel{d}{=} X^k$, it follows from the definition of $R(X^k, \cdot)$ that

$$\frac{1}{k} I\left(\tilde{X}^{(n,k)}; \tilde{Y}^{(n,k)}\right) \ge R\left(X^k, E\rho_k\left(\tilde{X}^{(n,k)}, \tilde{Y}^{(n,k)}\right)\right), \tag{39}$$

implying

$$\liminf_{n\to\infty} \frac{1}{k} I\left(\tilde{X}^{(n,k)}, \tilde{Y}^{(n,k)}\right) \ge R(X^k, D) \tag{40}$$

by (37) and the continuity of $R(X^k, \cdot)$. This complete the proof of (28). Displays (29) and (30) are immediate consequences. The convergence in (31) is a direct consequence of (30) and Lemma 1 (applied with the assignment $X \to X^k$, $Y \to Y^k$ and $\rho \to \rho_k$). □

Remark: Note that to conclude the convergence in (31) in the above proof a weaker version of Lemma 1 would have sufficed, that assumes $P_X^{(n)} = P_X$ for all n instead of $P_X^{(n)} \stackrel{d}{\Rightarrow} P_X$ in (32). This is because clearly, by construction of $\tilde{X}^{(n,k)}$, $\tilde{X}^{(n,k)} \stackrel{d}{=} X^k$ for all n and $k \le n$. The stronger form of Lemma 1 given will be instrumental in the derivation of a pointwise analogue to (31) in Section 4 (third item of Theorem 9).

C A Lower Bound on the Convergence Rate

Assume a stationary ergodic source $\mathbf{X}$ that satisfies, for each k, $R(X^k, D) = R(\mathbf{X}, D) + \frac{1}{k} H(X_1^k) - \overline{H}(\mathbf{X})$ and for which $R(X^k, D)$ is uniquely achieved by the pair $(\tilde{X}^k, \tilde{Y}^k)$. Theorem 3 implies that any good code sequence for $\mathbf{X}$ at rate $R(\mathbf{X}, D)$ must satisfy

$$(\tilde{X}^{(n,k)}, \tilde{Y}^{(n,k)}) \stackrel{d}{\Rightarrow} (\tilde{X}^k, \tilde{Y}^k) \quad \text{as} \quad n \to \infty \tag{41}$$

(recall Corollary 1 for the definition of the pair $(\tilde{X}^{(n,k)}, \tilde{Y}^{(n,k)})$). This implies, in particular,

$$\frac{1}{k}H(\tilde{Y}^{(n,k)}) \longrightarrow \frac{1}{k}H(\tilde{Y}^{k}) \quad \text{as } n \to \infty. \tag{42}$$

It follows from (42) that for k_n increasing slowly enough with n

$$\frac{1}{k_n}[H(\tilde{Y}^{(n,k_n)}) - H(\tilde{Y}^{k_n})] \longrightarrow 0 \quad \text{as } n \to \infty. \tag{43}$$

Display (43) is equivalent to

$$\frac{1}{k_n}H(\tilde{Y}^{(n,k_n)}) \longrightarrow \overline{H}(\tilde{\mathbf{Y}}) \quad \text{as } n \to \infty, \tag{44}$$

where we denote $\lim_{k\to\infty} \frac{1}{k}H(\tilde{Y}^{k})$ by $\overline{H}(\tilde{\mathbf{Y}})$ (a limit which exists[3] under our hypothesis that $R(X^k, D)$ is uniquely achieved by $(\tilde{X}^k, \tilde{Y}^k)$ for every k). The rate at which k_n may be allowed to increase while maintaining the convergence in (44) may depend on the particular code sequence (through which $(\tilde{X}^{(n,k)}, \tilde{Y}^{(n,k)})$ are defined). We now show, however, that if k_n increases more rapidly than a certain fraction of n, then the convergence in (44) fails for *any* code sequence. To see this note that for any k

$$\begin{aligned}
\frac{1}{k}H(\tilde{Y}^{(n,k)}) &= \frac{1}{k}H(Y_{J^{(n,k)}}^{J^{(n,k)}+k-1}) && (45)\\
&\le \frac{1}{k}H(Y_{J^{(n,k)}}^{J^{(n,k)}+k-1}, J^{(n,k)}) && (46)\\
&\le \frac{1}{k}H(Y^n, J^{(n,k)}) && (47)\\
&\le \frac{1}{k}H(Y^n) + \frac{1}{k}H(J^{(n,k)}) && (48)\\
&\le \frac{1}{k}H(Y^n) + \frac{\log n}{k}. && (49)
\end{aligned}$$

Now, since Y^n is associated with a good code,

$$\frac{1}{n}H(Y^n) \to R(\mathbf{X}, D). \tag{50}$$

Letting $k_n = \alpha n$, (49) and (50) imply

$$\limsup_{n\to\infty} \frac{1}{k_n}H(\tilde{Y}^{(n,k_n)}) \le \frac{1}{\alpha}R(\mathbf{X}, D), \tag{51}$$

excluding the possibility that (44) holds whenever

$$\alpha > \frac{R(\mathbf{X}, D)}{\overline{H}(\tilde{\mathbf{Y}})}. \tag{52}$$

This is analogous to the situation in the empirical distribution of good channel codes where it was shown in [SV97, Section III] for the memoryless case that convergence is excluded whenever $k_n = \alpha n$ for $\alpha > \frac{C}{H}$, where H stands for the entropy of the capacity-achieving channel input distribution.

To sum up, we see that convergence takes place whenever $k = O(1)$ (Theorem 3), and does not take place whenever $k = \alpha n$ for α satisfying (52). For k growing with n at the intermediate rates convergence depends on the particular code sequence, and examples can be constructed both for convergence and for lack thereof (analogous to [SV97, Examples 4,5]).

[3] In particular, for a memoryless source $\overline{H}(\tilde{\mathbf{Y}}) = H(\tilde{Y}_1)$. For sources for which the Shannon lower bound is tight it is the entropy rate of the source which results in $\mathbf{X}$ when corrupted by the max-entropy white noise tailored to the corresponding distortion measure and level.

D Applications to Compression-Based Denoising

The point of view that compression may facilitate denoising was put forth by Natarajan in [Nat95]. It is based on the intuition that the noise constitutes that part of a noisy signal which is least compressible. Thus, lossy compression of the noisy signal, under the right distortion measure and at the right distortion level, should lead to effective denoising. Compression-based denoising schemes have since been suggested and studied under various settings and assumptions (cf. [Nat93, Nat95, NKH98, CYV97, JY01, Don02, TRA02, Ris00, TPB99] and references therein). In this subsection we consider compression-based denoising when the clean source is corrupted by additive white noise. We will give a new performance bound for denoisers that optimally lossily compress the noisy signal (under distortion measure and level induced by the noise).

For simplicity, we restrict attention throughout this section to the case where the alphabets of the clean-, noisy-, and reconstructed-source are all equal to the M-ary alphabet $\mathcal{A} = \{0, 1, \ldots, M-1\}$. Addition and subtraction between elements of this alphabet should be understood modulo-M throughout. The results we develop below have analogues for cases where the alphabet is the real line.

We consider the case of a stationary ergodic source $\mathbf{X}$ corrupted by additive "white" noise $\mathbf{N}$. That is, we assume N_i are i.i.d.$\sim N$, independent of $\mathbf{X}$, and that the noisy observation sequence $\mathbf{Z}$ is given by

$$Z_i = X_i + N_i. \tag{53}$$

By "channel matrix" we refer to the Toeplitz matrix whose, say, first row is $(\Pr(N=0), \ldots, \Pr(N=M-1))$. We assume that $\Pr(N=a) > 0$ for all $a \in \mathcal{A}$ and associate with it a difference distortion measure $\rho^{(N)}$ defined by

$$\rho^{(N)}(a) = \log \frac{1}{\Pr(N=a)}. \tag{54}$$

We shall omit the superscript and let ρ stand for $\rho^{(N)}$ throughout this section.

Fact 1

$$\max\{H(X) : X \text{ is } \mathcal{A}\text{-valued and } E\rho(X) \le H(N)\}$$

is uniquely achieved (in distribution) by N.

Proof: N is clearly in the feasible set by the definition of ρ. Now, for any $\mathcal{A}$-valued X in the feasible set with $X \stackrel{d}{\neq} N$

$$\begin{aligned} H(N) &\ge E\rho(X) \\ &= \sum_a P_X(a) \log \frac{1}{\Pr(N=a)} \\ &> H(X), \end{aligned}$$

where the strict inequality follows from $D(P_X \| P_N) > 0$. □

Theorem 4 *Let $R(Z^k, \cdot)$ denote the k-th order rate distortion function of $\mathbf{Z}$ under the distortion measure in (54). Then*

$$R(Z^k, H(N)) = \frac{1}{k} H(Z^k) - H(N). \tag{55}$$

Furthermore, it is uniquely achieved (in distribution) by the pair (Z^k, X^k) whenever the channel matrix is invertible.

Proof: Let (Z^k, Y^k) achieve $R(Z^k, H(N))$. Then

$$\begin{aligned} kR(Z^k, H(N)) &= I(Z^k; Y^k) \\ &= H(Z^k) - H(Z^k|Y^k) \\ &\geq H(Z^k) - H(Z^k - Y^k|Y^k) \\ &\geq H(Z^k) - H(Z^k - Y^k) \\ &\geq H(Z^k) - kH(N), \end{aligned} \tag{56}$$

where the last inequality follows by the fact that $E\rho_n(Z^k, Y^k) \leq H(N)$ (otherwise (Z^k, Y^k) would not be an achiever of $R(Z^k, H(N))$) and Fact 1. The first part of the theorem follows since the pair (Z^k, X^k) satisfies $E\rho_n(Z^k, X^k) = H(N)$ and is readily seen to satisfy all the inequalities in (56) with equality. For the uniqueness part note that it follows from the chain of inequalities in (56) that if (Z^k, Y^k) achieve $R(Z^k, H(N))$ then Z^k is the output of the memoryless additive channel with noise components $\sim N$ whose input is Y^k. The invertibility of the channel matrix guarantees uniqueness of the distribution of the Y^k satisfying this for any given output distribution (cf., e.g., [WOS+03]). □

Let now $\Lambda : \mathcal{A} \times \mathcal{A} \to [0, \infty)$ be the loss function according to which denoising performance is measured. For $u^n, v^n \in \mathcal{A}^n$ let $\Lambda_n(u^n, v^n) = \frac{1}{n}\sum_{i=1}^n \Lambda(u_i, v_i)$. Define $\Psi : \mathcal{M}(\mathcal{A}) \times \mathcal{M}(\mathcal{A}) \to [0, \infty)$ by

$$\Gamma(P, Q) \stackrel{\triangle}{=} \max_{(U,V):U\sim P, V\sim Q} E\Lambda(U, V). \tag{57}$$

and $\Psi : \mathcal{M}(\mathcal{A}) \to [0, \infty)$ by

$$\Psi(P) \stackrel{\triangle}{=} \Gamma(P, P) = \max_{(U,V):U\sim P, V\sim P} E\Lambda(U, V). \tag{58}$$

It will be convenient to state the following (implied by an elementary continuity argument) for future reference

Fact 2 *Let $\{R_n\}$, $\{Q_n\}$ be sequences in $\mathcal{M}(\mathcal{A})$ with $R_n \to P$ and $Q_n \to P$. Then*

$$\lim_{n\to\infty} \Gamma(R_n, Q_n) = \Psi(P).$$

Example 1 *In the binary case, $M = 2$, when Λ denotes Hamming distortion, it is readily checked that, for $0 \leq p \leq 1$*

$$\Psi(\text{Bernoulli}(p)) = 2\min\{p, 1-p\} = 2\phi(p), \tag{59}$$

where

$$\phi(p) = \min\{p, 1-p\} \tag{60}$$

is the well-known Bayesian envelope for this setting [Han57, Sam63, MF98]. The achieving distribution in this case (assuming, say, $p \leq 1/2$) is

$$\Pr(U = 0, V = 0) = 1 - 2p, \quad \Pr(U = 0, V = 1) = p, \quad \Pr(U = 1, V = 0) = p, \quad \Pr(U = 0, V = 0) = 0. \tag{61}$$

Theorem 5 *Let $\{Y^n(\cdot)\}$ correspond to a good code sequence for the noisy source $\mathbf{Z}$ at distortion level $H(N)$ (under the difference distortion measure in (54)) and assume that the channel matrix is invertible. Then*

$$\limsup_{n\to\infty} E\Lambda_n(X^n, Y^n(Z^n)) \leq E\Psi\left(P_{X_0|Z_{-\infty}^{\infty}}\right). \tag{62}$$

Note, in particular, that for the case of a binary source corrupted by a BSC, the optimum distribution dependent denoising performance is $E\phi\left(P_{X_0|Z_{-\infty}^{\infty}}\right)$ (cf. [WOS+03, Section 5]), so the right side of (62) is, by (59), precisely twice the expected fraction of errors made by the optimal denoiser (in the limit of many observations). Note that this performance can be attained universally (i.e., with no a priori knowledge of the distribution of the noise-free source $\mathbf{X}$) by employing a universal rate distortion code on the noisy source, e.g., the fixed distortion version of the Yang-Kieffer codes[4] [YK96, Section III].

Proof of Theorem 5: Similarly as in Corollary 1, for $1 \le k \le \frac{n-1}{2}$ let $J^{(n,k)}$ be uniformly distributed on $\{k+1, \ldots, n-k+1\}$, and independent of $(\mathbf{X}, \mathbf{Z})$. Consider the triple $(X^n, Z^n, Y^n) = (X^n, Z^n, Y^n(Z^n))$ and denote

$$\left(X^{[n,k]}{}_{-k}^{k}, Z^{[n,k]}{}_{-k}^{k}, Y^{[n,k]}{}_{-k}^{k}\right) = \left(X_{J^{(n,k)}-k}^{J^{(n,k)}+k}, Z_{J^{(n,k)}-k}^{J^{(n,k)}+k}, Y_{J^{(n,k)}-k}^{J^{(n,k)}+k}\right). \tag{63}$$

Note first that

$$E\Lambda\left(X^{[n,k]}{}_0, Y^{[n,k]}{}_0\right) = \frac{1}{n-2k}\sum_{i=k+1}^{n-k+1} E\Lambda(X_i, Y_i) \tag{64}$$

so

$$E\Lambda_n(X^n, Y^n(Z^n)) - \frac{2k\Lambda_{max}}{n} \le E\Lambda\left(X^{[n,k]}{}_0, Y^{[n,k]}{}_0\right) \le \frac{n}{n-2k} E\Lambda_n(X^n, Y^n(Z^n)) \tag{65}$$

and, in particular,

$$\lim_{n\to\infty}\left[E\Lambda_n(X^n, Y^n(Z^n)) - E\Lambda\left(X^{[n,k]}{}_0, Y^{[n,k]}{}_0\right)\right] = 0. \tag{66}$$

Also, by the definition of Γ,

$$\begin{aligned} E\Lambda\left(X^{[n,k]}{}_0, Y^{[n,k]}{}_0\right) &= \sum_{z_{-k}^k \in \mathcal{A}^k} E\left[\Lambda\left(X^{[n,k]}{}_0, Y^{[n,k]}{}_0\right) \middle| Z^{[n,k]}{}_{-k}^{k} = z_{-k}^k\right] \Pr\left(Z^{[n,k]}{}_{-k}^{k} = z_{-k}^k\right) \\ &\le \sum_{z_{-k}^k \in \mathcal{A}^k} \Gamma\left(P_{X^{[n,k]}{}_0|Z^{[n,k]}{}_{-k}^{k}=z_{-k}^k}, P_{Y^{[n,k]}{}_0|Z^{[n,k]}{}_{-k}^{k}=z_{-k}^k}\right) \Pr\left(Z^{[n,k]}{}_{-k}^{k} = z_{-k}^k\right). \end{aligned} \tag{67}$$

Now, clearly $\left(X^{[n,k]}{}_{-k}^{k}, Z^{[n,k]}{}_{-k}^{k}\right) \stackrel{d}{=} (X_{-k}^k, Z_{-k}^k)$ so, in particular, $\forall z_{-k}^k \in \mathcal{A}^k$

$$P_{X^{[n,k]}{}_0|Z^{[n,k]}{}_{-k}^{k}=z_{-k}^k} = P_{X_0|Z_{-k}^k=z_{-k}^k}. \tag{68}$$

On the other hand, from the combination of Theorem 3 and Theorem 4 it follows, since $\{Y^n(\cdot)\}$ corresponds to a good code sequence, that

$$\left(Y^{[n,k]}{}_{-k}^{k}, Z^{[n,k]}{}_{-k}^{k}\right) \stackrel{d}{\Rightarrow} (X_{-k}^k, Z_{-k}^k) \quad \text{as } n \to \infty. \tag{69}$$

and therefore, in particular, $\forall z_{-k}^k \in \mathcal{A}^k$

$$P_{Y^{[n,k]}{}_0|Z^{[n,k]}{}_{-k}^{k}=z_{-k}^k} \longrightarrow P_{X_0|Z_{-k}^k=z_{-k}^k} \quad \text{as } n \to \infty. \tag{70}$$

The combination of (68), (70), and Fact 2 implies that $\forall z_{-k}^k \in \mathcal{A}^k$

$$\Gamma\left(P_{X^{[n,k]}{}_0|Z^{[n,k]}{}_{-k}^{k}=z_{-k}^k}, P_{Y^{[n,k]}{}_0|Z^{[n,k]}{}_{-k}^{k}=z_{-k}^k}\right) \longrightarrow \Psi\left(P_{X_0|Z_{-k}^k=z_{-k}^k}\right) \quad \text{as } n \to \infty. \tag{71}$$

[4]This may conceptually motivate employing a universal lossy source code for denoising. Implementation of such a code, however, is too complex to be of practical value. It seems even less motivated in light of the universally optimal and practical scheme in [WOS+03].

Thus we obtain

$$\begin{aligned}\limsup_{n\to\infty} E\Lambda_n(X^n, Y^n(Z^n)) &\overset{(a)}{=} \limsup_{n\to\infty} E\Lambda\left(X^{[n,k]}{}_0, Y^{[n,k]}{}_0\right)\\ &\overset{(b)}{\leq} \sum_{z^k_{-k}\in\mathcal{A}^k} \Psi\left(P_{X_0|Z^k_{-k}=z^k_{-k}}\right)\Pr\left(Z^k_{-k}=z^k_{-k}\right)\\ &= E\Psi\left(P_{X_0|Z^k_{-k}}\right), \end{aligned} \tag{72}$$

where (a) follows from (66) and (b) from combining (67) with (71) (and the fact that ${Z^{[n,k]}}^k_{-k} \overset{d}{=} Z^k_{-k}$). The fact that $\lim_{k\to\infty} P_{X_0|Z^k_{-k}} = P_{X_0|Z^\infty_{-\infty}}$ a.s. (martingale convergence), the continuity of Ψ, and the bounded convergence theorem imply

$$\lim_{k\to\infty} E\Psi\left(P_{X_0|Z^k_{-k}}\right) = E\Psi\left(P_{X_0|Z^\infty_{-\infty}}\right), \tag{73}$$

which completes the proof when combined with (72). □

4 Sample Properties of the Empirical Distribution

Throughout this section "codes" should be understood in the fixed-rate sense of Definition 1.

A Pointwise Bounds on the Mutual Information Induced by a Rate-Constrained Code

The first result of this section is the following.

Theorem 6 *Let $\{Y^n(\cdot)\}$ be a sequence of block codes with rate $\leq R$. For any stationary ergodic* $\mathbf{X}$

$$\limsup_{n\to\infty} I\left(\hat{Q}^1[X^n, Y^n(X^n)]\right) \leq R + H(X_1) - \overline{H}(\mathbf{X}) \quad a.s. \tag{74}$$

In particular, when $\mathbf{X}$ *is memoryless,*

$$\limsup_{n\to\infty} I\left(\hat{Q}^1[X^n, Y^n(X^n)]\right) \leq R \quad a.s. \tag{75}$$

The following lemma is at the heart of the proof of Theorem 6.

Lemma 2 *Let $Y^n(\cdot)$ be an n-block code of rate $\leq R$. Then, for every $P \in \mathcal{M}_n(\mathcal{X})$ and $\eta > 0$,*

$$\left|\left\{x^n \in T_P : I\left(\hat{Q}^1[x^n, Y^n(x^n)]\right) > R + \eta\right\}\right| \leq e^{n[H(P)+\varepsilon_n-\eta]}, \tag{76}$$

where $\varepsilon_n = \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)$.

Lemma 2 quantifies a remarkable limitation that the rate of the code induces: for any $\eta > 0$, only an exponentially negligible fraction of the source sequences in any type can have an empirical joint distribution with their reconstruction with mutual information exceeding the rate by more than η.

Proof of Lemma 2: For every $W \in \mathcal{C}_n(P)$ let

$$S(P, W) = \{x^n : (x^n, Y^n(x^n)) \in T_{P\otimes W}\} \tag{77}$$

and note that

$$|S(P,W)| \le e^{nR} e^{n[H_{P\otimes W}(X|Y)]}, \tag{78}$$

since there are no more than e^{nR} different values of $Y^n(x^n)$ and, by [CK81, Lemma 2.13], no more than $e^{n[H_{P\otimes W}(X|Y)]}$ different sequences in $S(P,W)$ that can be mapped to the same value of Y^n. It follows that

$$I(P;W) \le R+\eta \ \ \forall W \in \mathcal{C}_n(P) : |S(P,W)| \ge e^{n[H(P)-\eta]}. \tag{79}$$

Now, by definition,

$$\left\{x^n \in T_P : I\left(\hat{Q}^1[x^n, Y^n(x^n)]\right) > R+\eta\right\} = \bigcup_{W\in\mathcal{C}_n(P):I(P;W)>R+\eta} S(P,W) \tag{80}$$

so

$$\begin{aligned}
\left|\left\{x^n \in T_P : I\left(\hat{Q}^1[x^n, Y^n(x^n)]\right) > R+\eta\right\}\right| &\le \sum_{W\in\mathcal{C}_n(P):I(P;W)>R+\eta} |S(P,W)| \\
&\le \sum_{W\in\mathcal{C}_n(P):I(P;W)>R+\eta} e^{n[H(P)-\eta]} \qquad (81)\\
&\le |\mathcal{C}_n(P)| e^{n[H(P)-\eta]} \\
&\le (n+1)^{|\mathcal{X}||\mathcal{Y}|} e^{n[H(P)-\eta]} \\
&\le e^{n[H(P)+\varepsilon_n-\eta]}, \qquad (82)
\end{aligned}$$

where (81) follows from (79) and (82) from the definition of ε_n. □

We shall also make use of the following "converse to the AEP" (proof of which is deferred to Appendix B).

Lemma 3 *Let* $\mathbf{X}$ *be stationary ergodic and* $A_n \subseteq \mathcal{X}^n$ *an arbitrary sequence satisfying*

$$\limsup_{n\to\infty} \frac{1}{n} \log |A_n| < \overline{H}(\mathbf{X}). \tag{83}$$

Then, with probability one, $X^n \in A_n$ *for only finitely many* n.

Proof of Theorem 6: Fix an arbitrary $\eta > 0$ and $\xi > 0$ small enough so that

$$\max_{P\in B(P_{X_1},\xi)} H(P) \le H(X_1) + \eta/2 \tag{84}$$

(the continuity of the entropy functional guarantees the existence of such a $\xi > 0$). Lemma 2 (with the assignment $\eta \to H(X_1) - \overline{H}(\mathbf{X}) + \eta$) guarantees that for every n and $P \in \mathcal{M}_n(\mathcal{X})$,

$$\left|\left\{x^n \in T_P : I\left(\hat{Q}^1[x^n, Y^n(x^n)]\right) > R_n + H(X_1) - \overline{H}(\mathbf{X}) + \eta\right\}\right| \le e^{n[H(P)+\varepsilon_n - H(X_1)+\overline{H}(\mathbf{X})-\eta]}, \tag{85}$$

where R_n denotes the rate of $Y^n(\cdot)$. Thus

$$\begin{aligned}
&\left|\left\{x^n \in T_{[P_{X_1}]_\xi} : I\left(\hat{Q}^1[x^n, Y^n(x^n)]\right) > R_n + H(X_1) - \overline{H}(\mathbf{X}) + \eta\right\}\right| \qquad (86)\\
\le\ & \exp\left\{n\left[\max_{P\in B(P_{X_1},\xi)} H(P) + 2\varepsilon_n - H(X_1) + \overline{H}(\mathbf{X}) - \eta\right]\right\} \qquad (87)\\
\le\ & \exp\{n[2\varepsilon_n + \overline{H}(\mathbf{X}) - \eta/2]\}, \qquad (88)
\end{aligned}$$

where (87) follows from (85) and from the (obvious) fact that $T_{[P_{X_1}]_\xi}$ is a union of less than $e^{n\varepsilon_n}$ strict types, and (88) follows from (84). Denoting now the set in (86) by A_n, i.e.,

$$A_n = \left\{x^n \in T_{[P_{X_1}]_\xi} : I\left(\hat{Q}^1[x^n, Y^n(x^n)]\right) > R_n + H(X_1) - \overline{H}(\mathbf{X}) + \eta\right\}, \tag{89}$$

we have

$$\left\{x^n : I\left(\hat{Q}^1[x^n, Y^n(x^n)]\right) > R_n + H(X_1) - \overline{H}(\mathbf{X}) + \eta\right\} \subseteq T^c_{[P_{X_1}]_\xi} \cup A_n. \tag{90}$$

But, by ergodicity, $X^n \in T^c_{[P_{X_1}]_\xi}$ for only finitely many n with probability one. On the other hand, the fact that $|A_n| \le \exp\{n[2\varepsilon_n + \overline{H}(\mathbf{X}) - \eta/2]\}$ (recall (88)) implies, by Lemma 3, that $X^n \in A_n$ for only finitely many n with probability one. Thus we get

$$I\left(\hat{Q}^1[X^n, Y^n(X^n)]\right) \le R_n + H(X_1) - \overline{H}(\mathbf{X}) + \eta \ \text{ eventually } \ a.s. \tag{91}$$

But, by hypothesis, $\{Y^n(\cdot)\}$ has rate $\le R$ (namely $\limsup_{n\to\infty} R_n \le R$) so (91) implies

$$\limsup_{n\to\infty} I\left(\hat{Q}^1[X^n, Y^n(X^n)]\right) \le R + H(X_1) - \overline{H}(\mathbf{X}) + \eta \ \ a.s. \tag{92}$$

which completes the proof by the arbitrariness of $\eta > 0$. □

Theorem 6 extends to higher-order empirical distributions as follows:

Theorem 7 *Let $\{Y^n(\cdot)\}$ be a sequence of block codes with rate $\le R$. For any stationary ergodic $\mathbf{X}$ and $k \ge 1$*

$$\limsup_{n\to\infty} \frac{1}{k} I\left(\hat{Q}^k[X^n, Y^n(X^n)]\right) \le R + \frac{1}{k} H(X_1^k) - \overline{H}(\mathbf{X}) \ \ a.s. \tag{93}$$

In particular, when $\mathbf{X}$ is memoryless,

$$\limsup_{n\to\infty} \frac{1}{k} I\left(\hat{Q}^k[X^n, Y^n(X^n)]\right) \le R \ \ a.s. \tag{94}$$

Note that defining, for each $0 \le j \le k-1$, $\hat{Q}^{k,j}$ by[5]

$$\hat{Q}^{k,j}[x^n, y^n](u^k, v^k) = \frac{k}{n}\left|\left\{0 \le i \le \frac{n}{k} - 1 : x_{ik+j+1}^{(i+1)k+j} = u^k, y_{ik+j+1}^{(i+1)k+j} = v^k\right\}\right|, \tag{95}$$

it follows from Theorem 6 (applied to a kth-order super symbol) that when the source is ergodic in k-blocks[6]

$$\limsup_{n\to\infty} \frac{1}{k} I\left(\hat{Q}^{k,j}[X^n, Y^n(X^n)]\right) \le R + \frac{1}{k} H(X_1^k) - \overline{H}(\mathbf{X}) \ \ a.s. \tag{96}$$

Then, since for fixed k and large n $\hat{Q}^k \approx \frac{1}{k}\sum_{j=0}^{k-1} \hat{Q}^{k,j}$ and, by the k-block ergodictiy, $\hat{Q}^{k,j}[X^n] \approx \hat{Q}^{k,j}[X^n] \approx P_{X^k}$, it would follow from the continuity of $I(\cdot)$ and its convexity in the conditional distribution that, for large n, $I(\hat{Q}^k) \lessapprox \frac{1}{k}\sum_{j=0}^{k-1} I(\hat{Q}^{k,j})$, implying (93) when combined with (96). The proof that follows makes these continuity arguments precise and accommodates the possibility that the source is not ergodic in k-blocks.

Proof of Theorem 7: For each $0 \le j \le k-1$ denote the kth-order super source by $\mathbf{X}^{(k,j)} = \{X_{ik+j+1}^{(i+1)k+j}\}_i$. Lemma 9.8.2 of [Gal68] implies the existence of k events $A_0, \ldots, A_{k-1}$ with the following properties:

[5]Note that $\hat{Q}^{k,j}[x^n, y^n]$ may not be a bona fide distribution but is very close to one for fixed k and large n.

[6]When $\hat{Q}^{k,j}[x^n, y^n]$ is not a bona fide distribution $I\left(\hat{Q}^{k,j}[x^n, y^n]\right)$ should be understood as the mutual information under the normalized version of $\hat{Q}^{k,j}[x^n, y^n]$.

1. $P(A_i \cap A_j) = \begin{cases} 1/k & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$

2. For each $i, j \in \{0, \ldots, k-1\}$, $\mathbf{X}^{(k,j)}$ conditioned on A_i is stationary and ergodic.

3. $\mathbf{X}^{(k,j)}$ conditioned on A_i is equal in distribution to $\mathbf{X}^{(k,j+1)}$ conditioned on A_{i+1} (where the indices i, j are to be taken modulo k).

Note that

$$\left\| \hat{Q}^k[x^n, y^n] - \frac{1}{k} \sum_{j=0}^{k-1} \hat{Q}^{k,j}[x^n, y^n] \right\| \leq \delta_n^{(k)}, \tag{97}$$

where $\{\delta_n^{(k)}\}_n$ is a deterministic sequence with $\delta_n^{(k)} \to 0$ as $n \to \infty$. Theorem 6 applied to $\mathbf{X}^{(k,j)}$ conditioned on A_i gives

$$\limsup_{n\to\infty} I\left(\hat{Q}^{k,j}[X^n, Y^n(X^n)]\right) \leq Rk + H(X_{j+1}^{k+j}|A_i) - \overline{H}(\mathbf{X}^{(k,j)}|A_i) \ \ a.s. \ \text{ on } A_i, \tag{98}$$

where $H(X_{j+1}^{k+j}|A_i)$ and $\overline{H}(\mathbf{X}^{(k,j)}|A_i)$ denote, respectively, the entropy of the distribution of X_{j+1}^{k+j} when conditioned on A_i and the entropy rate of $\mathbf{X}^{(k,j)}$ when condition on A_i. Now, due to the fact that $\mathbf{X}^{(k,j)}$ conditioned on A_i is stationary ergodic we have, for each j,

$$\lim_{n\to\infty} \hat{Q}^{k,j}[X^n] = P_{X^k|A_i} \ \ a.s. \text{ on } A_i, \tag{99}$$

with $P_{X^k|A_i}$ denoting the distribution of X^k conditioned on A_i. Equation (99) implies, in turn, by a standard continuity argument (letting $\hat{Q}^{k,j}[y^n|x^n]$ be defined analogously as $\hat{Q}^k[y^n|x^n]$),

$$\lim_{n\to\infty} \left\| \hat{Q}^{k,j}[X^n, Y^n(X^n)] - P_{X^k|A_i} \otimes \hat{Q}^{k,j}[Y^n(X^n)|X^n] \right\| = 0 \ \ a.s. \text{ on } A_i. \tag{100}$$

Combined with (97), (100) implies also

$$\lim_{n\to\infty} \left\| \hat{Q}^k[X^n, Y^n(X^n)] - P_{X^k|A_i} \otimes \left[\frac{1}{k} \sum_{j=0}^{k-1} \hat{Q}^{k,j}[Y^n(X^n)|X^n] \right] \right\| = 0 \ \ a.s. \text{ on } A_i \tag{101}$$

implying, in turn, by the continuity of the mutual information as a function of the joint distribution

$$I(\hat{Q}^k[X^n, Y^n(X^n)]) - I\left(P_{X^k|A_i} \otimes \left[\frac{1}{k} \sum_{j=0}^{k-1} \hat{Q}^{k,j}[Y^n(X^n)|X^n] \right] \right) \overset{n\to\infty}{\longrightarrow} 0 \ \ a.s. \text{ on } A_i. \tag{102}$$

On the other hand, by the convexity of the mutual information in the conditional distribution [CT91, Theorem 2.7.4], we have for all k, n (and all sample paths)

$$I\left(P_{X^k|A_i} \otimes \left[\frac{1}{k} \sum_{j=0}^{k-1} \hat{Q}^{k,j}[Y^n(X^n)|X^n] \right] \right) \leq \frac{1}{k} \sum_{j=0}^{k-1} I\left(P_{X^k|A_i} \otimes \hat{Q}^{k,j}[Y^n(X^n)|X^n] \right). \tag{103}$$

Using (99) and the continuity of the mutual information yet again gives for all j

$$I\left(P_{X^k|A_i} \otimes \hat{Q}^{k,j}[Y^n(X^n)|X^n] \right) - I\left(\hat{Q}^{k,j}[X^n, Y^n(X^n)] \right) \overset{n\to\infty}{\longrightarrow} 0 \ \ a.s. \text{ on } A_i. \tag{104}$$

Consequently, a.s. on A_i,

$$\begin{aligned}
\limsup_{n\to\infty} I(\hat{Q}^k[X^n, Y^n(X^n)]) &\leq \limsup_{n\to\infty} \frac{1}{k}\sum_{j=0}^{k-1} I\left(\hat{Q}^{k,j}[X^n, Y^n(X^n)]\right) && (105)\\
&\leq Rk + \frac{1}{k}\sum_{j=0}^{k-1}[H(X_{j+1}^{k+j}|A_i) - \overline{H}(\mathbf{X}^{(k,j)}|A_i)] && (106)\\
&= Rk + \frac{1}{k}\sum_{j=0}^{k-1}[H(X_1^k|A_{i-j \bmod k}) - \overline{H}(\mathbf{X}^{(k,1)}|A_{i-j \bmod k})] && (107)\\
&= Rk + \frac{1}{k}\sum_{j=0}^{k-1}[H(X_1^k|A_j) - \overline{H}(\mathbf{X}^{(k,1)}|A_j)], && (108)
\end{aligned}$$

where (105) follows by combining (102) - (104), (106) follows from (98), and (107) follows from the third property of the events A_i recalled above. Since (108) does not depend on i we obtain

$$\limsup_{n\to\infty} I(\hat{Q}^k[X^n, Y^n(X^n)]) \leq Rk + \frac{1}{k}\sum_{j=0}^{k-1}[H(X_1^k|A_j) - \overline{H}(\mathbf{X}^{(k,1)}|A_j)] \quad a.s. \tag{109}$$

Defining the $\{0, \ldots, k-1\}$-valued random variable J by

$$J = i \text{ on } A_i, \tag{110}$$

it follows from the first property of the sets A_i recalled above that

$$\frac{1}{k}\sum_{j=0}^{k-1} H(X_1^k|A_j) = H(X_1^k|J) \leq H(X_1^k) \tag{111}$$

and that

$$\begin{aligned}
\frac{1}{k}\sum_{j=0}^{k-1}\overline{H}(\mathbf{X}^{(k,1)}|A_j) &= \overline{H}(\mathbf{X}^{(k,1)}|J)\\
&= k\lim_{n\to\infty}\frac{1}{n}H(X^n|J)\\
&= k\lim_{n\to\infty}\frac{1}{n}[H(X^n, J) - H(J)]\\
&= k\lim_{n\to\infty}\frac{1}{n}H(X^n, J)\\
&\geq k\lim_{n\to\infty}\frac{1}{n}H(X^n) = k\overline{H}(\mathbf{X}). && (112)
\end{aligned}$$

Combining (109) with (111) and (112) gives (93). □

A direct consequence of (93) is the following:

Corollary 2 *Let $\{Y^n(\cdot)\}$ be a sequence of block codes with rate $\leq R$. For any stationary ergodic $\mathbf{X}$*

$$\limsup_{k\to\infty}\limsup_{n\to\infty}\frac{1}{k} I\left(\hat{Q}^k[X^n, Y^n(X^n)]\right) \leq R \quad a.s. \tag{113}$$

B Sample Converse in Lossy Source Coding

One of the main results of [Kie91] is the following.

Theorem 8 (Theorem 1 in [Kie91]) *Let $\{Y^n(\cdot)\}$ be a sequence of block codes with rate $\leq R$. For any stationary ergodic* $\mathbf{X}$

$$\liminf_{n\to\infty} \rho_n(X^n, Y^n(X^n)) \geq D(\mathbf{X}, R) \quad a.s. \tag{114}$$

The proof in [Kie91] was valid for general source and reconstruction alphabets, and was based on the powerful ergodic theorem in [Kie89]. We now show how Corollary 2 can be used to give a simple proof, valid in our finite-alphabet setting.

Proof of Theorem 8: Note first that by a standard continuity argument and ergodicity, for each k,

$$\lim_{n\to\infty} \left\| \hat{Q}^k[X^n, Y^n(X^n)] - P_{X^k} \otimes \hat{Q}^k[Y^n(X^n)|X^n] \right\| = 0 \quad a.s. \tag{115}$$

implying, in turn, by the continuity of the mutual information as a function of the joint distribution

$$I(\hat{Q}^k[X^n, Y^n(X^n)]) - I\left(P_{X^k} \otimes \hat{Q}^k[Y^n(X^n)|X^n]\right) \stackrel{n\to\infty}{\longrightarrow} 0 \quad a.s. \tag{116}$$

Fix now $\varepsilon > 0$. By Corollary 2, with probability one there exists a k large enough that

$$\frac{1}{k} I(\hat{Q}^k[X^n, Y^n(X^n)]) \leq R + \varepsilon \text{ for all sufficiently large } n, \tag{117}$$

implying, by (116), with probability one existence of k large enough that

$$\frac{1}{k} I\left(P_{X^k} \otimes \hat{Q}^k[Y^n(X^n)|X^n]\right) \leq R + 2\varepsilon \text{ for all sufficiently large } n. \tag{118}$$

By the definition of $D_k(\mathbf{X}, \cdot)$ it follows that for all such k and n

$$E_{P_{X^k} \otimes \hat{Q}^k[Y^n(X^n)|X^n]} \rho_k(X^k, Y^k) \geq D_k(\mathbf{X}, R + 2\varepsilon) \geq D(\mathbf{X}, R + 2\varepsilon), \tag{119}$$

implying, by the continuity property (115), with probability one the existence of k large enough that

$$E_{\hat{Q}^k[X^n, Y^n(X^n)]} \rho_k(X^k, Y^k) \geq D(\mathbf{X}, R + 2\varepsilon) - \varepsilon \quad \text{for all sufficiently large } n. \tag{120}$$

By the definition of $\hat{Q}^k$ it is readily verified that for any k (and all sample paths)[7]

$$\left| E_{\hat{Q}^k[X^n, Y^n(X^n)]} \rho_k(X^k, Y^k) - \rho_n(X^n, Y^n) \right| \leq \varepsilon_n^{(k)}, \tag{121}$$

where $\{\varepsilon_n^{(k)}\}_n$ is a deterministic sequence with $\varepsilon_n^{(k)} \to 0$ as $n \to \infty$. Combined with (120) this implies

$$\liminf_{n\to\infty} \rho_n(X^n, Y^n) \geq D(\mathbf{X}, R + 2\varepsilon) - \varepsilon \quad \text{a.s.}, \tag{122}$$

completing the proof by the arbitrariness of ε and the continuity of $D(\mathbf{X}, \cdot)$. □

[7] In fact, we would have pointwise equality $E_{\hat{Q}^k[x^n, y^n]} \rho_k(X^k, Y^k) = \rho_n(x^n, y^n)$ had $\hat{Q}^k$ been slightly modified from its definition in (2) to $\hat{Q}^k[x^n, y^n](u^k, v^k) = \frac{1}{n} |\{1 \leq i \leq n : x_{i+1}^{i+k} = u^k, y_{i+1}^{i+k} = v^k\}|$ with the indices understood modulo n.

C Sample Behavior of the Empirical Distribution of Good Codes

In the context of Theorem 8 we make the following sample analogue of Definition 3.

Definition 4 *Let $\{Y^n(\cdot)\}$ be a sequence of block codes with rate $\leq R$ and let $\mathbf{X}$ be a stationary ergodic source. $\{Y^n(\cdot)\}$ will be said to be a pointwise good sequence of codes for the stationary ergodic source $\mathbf{X}$ in the strong sense at rate R if*

$$\limsup_{n\to\infty} \rho_n(X^n, Y^n(X^n)) \leq D(\mathbf{X}, R) \quad a.s. \tag{123}$$

Note that Theorem 8 implies that the limit supremum in the above definition is actually a limit, the inequality in (123) actually holds with equality, and the rate of the corresponding good code sequence is $= R$. The bounded convergence theorem implies that a pointwise good sequence of codes is also good in the sense of Definition 3. The converse, however, is not true [Wei]. The existence of pointwise good code sequences is a known consequence of the existence of good code sequences [Kie91, Kie78]. In fact, there exist pointwise good code sequences that are universally good for all stationary and ergodic sources [Ziv72, Ziv80, YK96]. Henceforth the phrase "good codes" should be understood in the sense of Definition 4, even when omitting "pointwise".

The following is the pointwise version of Theorem 3.

Theorem 9 *Let $\{Y^n(\cdot)\}$ correspond to a pointwise good code sequence for the stationary ergodic source $\mathbf{X}$ at rate $R(\mathbf{X}, D)$. Then:*

1. *For every k*

$$\lim_{n\to\infty} E_{\hat{Q}^k[X^n, Y^n(X^n)]} \rho_k(X^k, Y^k) = D \quad a.s., \tag{124}$$

$$R(X^k, D) \leq \liminf_{n\to\infty} \frac{1}{k} I\left(\hat{Q}^k[X^n, Y^n(X^n)]\right) \leq \limsup_{n\to\infty} \frac{1}{k} I\left(\hat{Q}^k[X^n, Y^n(X^n)]\right) \leq R(\mathbf{X}, D) + \frac{1}{k} H(X^k) - \overline{H}(\mathbf{X}) \quad a.s., \tag{125}$$

   so in particular

$$\lim_{k\to\infty} \liminf_{n\to\infty} \frac{1}{k} I\left(\hat{Q}^k[X^n, Y^n(X^n)]\right) = \lim_{k\to\infty} \limsup_{n\to\infty} \frac{1}{k} I\left(\hat{Q}^k[X^n, Y^n(X^n)]\right) = R(\mathbf{X}, D) \quad a.s. \tag{126}$$

2. *If it also holds that $R(X^k, D) = R(\mathbf{X}, D) + \frac{1}{k} H(X_1^k) - \overline{H}(\mathbf{X})$ then*

$$\lim_{n\to\infty} \frac{1}{k} I\left(\hat{Q}^k[X^n, Y^n(X^n)]\right) = R(X^k, D) \quad a.s. \tag{127}$$

3. *If, additionally, $R_k(\mathbf{X}, D)$ is uniquely achieved by the pair $(\tilde{X}^k, \tilde{Y}^k)$ then*

$$\hat{Q}^k[X^n, Y^n(X^n)] \Rightarrow P_{\tilde{X}^k, \tilde{Y}^k} \quad a.s. \tag{128}$$

Proof of Theorem 9: $\{Y^n(\cdot)\}$ being a good code sequence implies

$$\lim_{n\to\infty} \rho_n(X^n, Y^n(X^n)) = D \quad a.s., \tag{129}$$

implying (124) by (121). Reasoning similarly as in the proof of Theorem 8, it follows from (115) that

$$\lim_{n\to\infty} E_{P_{X^k}\otimes\hat{Q}^k[Y^n(X^n)|X^n]}\rho_k(X^k,Y^k) = \lim_{n\to\infty} E_{\hat{Q}^k[X^n,Y^n(X^n)]}\rho_k(X^k,Y^k) = D \quad a.s., \tag{130}$$

implying

$$\liminf_{n\to\infty}\frac{1}{k}I(\hat{Q}^k[X^n,Y^n(X^n)]) \geq \liminf_{n\to\infty}\frac{1}{k}I\left(P_{X^k}\otimes\hat{Q}^k[Y^n(X^n)|X^n]\right) \geq R_k(\mathbf{X},D) \quad a.s., \tag{131}$$

where the left inequality follows from (116) and the second one by (130) and the continuity of $R_k(\mathbf{X},\cdot)$. The upper bound in (125) follows from Theorem 7. Displays (126) and (127) are immediate consequences. The convergence in (128) follows directly from combining (127), (124), the fact that, by ergodicity,

$$\hat{Q}^k[X^n] \Rightarrow P_{\tilde{X}^k} \quad a.s., \tag{132}$$

and Lemma 1 (applied to the k-th order super-symbol). □

D Application to Compression-Based Denoising

Consider again the setting of subsection 3.D where the stationary ergodic source $\mathbf{X}$ is corrupted by additive noise $\mathbf{N}$ of i.i.d. components. The following is the almost sure version of Theorem 5.

Theorem 10 *Let $\{Y^n(\cdot)\}$ correspond to a good code sequence for the noisy source $\mathbf{Z}$ at distortion level $H(N)$ (under the difference distortion measure in (54)) and assume that the channel matrix is invertible. Then*

$$\limsup_{n\to\infty}\Lambda_n(X^n,Y^n(Z^n)) \leq E\Psi\left(P_{X_0|Z_{-\infty}^{\infty}}\right) \quad a.s. \tag{133}$$

Proof: The combination of Theorem 4 with the third item of Theorem 9 gives

$$\hat{Q}^k[Z^n,Y^n(Z^n)] \Rightarrow P_{Z^k,X^k} \quad a.s. \tag{134}$$

On the other hand, joint stationarity and ergodicity of the pair $(\mathbf{X},\mathbf{Z})$ implies

$$\hat{Q}^k[X^n,Z^n] \Rightarrow P_{X^k,Z^k} \quad a.s. \tag{135}$$

Arguing analogously as in the proof of Theorem 5 (this time for the sample empirical distribution[8] $\hat{Q}^k[X^n,Z^n,Y^n]$ instead of the distribution of the triple in (63)) leads, taking say $k=2m+1$, to

$$\limsup_{n\to\infty}\Lambda_n(X^n,Y^n(Z^n)) \leq E\Psi\left(P_{X_0|Z_{-m}^{m}}\right) \quad \text{a.s.}, \tag{136}$$

implying (62) upon letting $m\to\infty$. □

[8] We extend the notation by letting $\hat{Q}^k[x^n,z^n,y^n]$ stand for the k-th-order empirical distribution of the triple (X^k,Z^k,Y^k) induced by (x^n,z^n,y^n).

E Derandomizing for Optimum Denoising

The reconstruction associated with a good code sequence has what seems to be a desirable property in the denoising setting of the previous subsection, namely, that the marginalized distribution, for any finite k, of the noisy source and reconstruction is essentially distributed like the noisy source with the underlying clean signal, as n becomes large. Indeed, this property was enough to derive an upper bound on the denoising loss (theorems 5 and 10) which, for example, for the binary case was seen to be within a factor of 2 from the optimum.

We now point out, using the said property, that essentially optimum denoising can be attained by a simple "post-processing" procedure. The procedure is to fix an m and evaluate, for each $z_{-m}^m \in \mathcal{A}^{2m+1}$ and $y \in \mathcal{A}$

$$\hat{Q}^{2m+1}[Z^n, Y^n(Z^n)](z_{-m}^m, y) \stackrel{\triangle}{=} \sum_{y_{-m}^{-1}, y_1^m} \hat{Q}^{2m+1}[Z^n, Y^n(Z^n)](z_{-m}^m, y_{-m}^{-1} y y_1^m).$$

In practice, the computation of $\hat{Q}^{2m+1}[Z^n, Y^n(Z^n)](z_{-m}^m, y)$ can be done quite efficiently and sequentially by updating counts for the various $z_{-m}^m \in \mathcal{A}^{2m+1}$ as they appear along the noisy sequence (cf. [WOS+03, Section 3]). Define now the n-block denoiser $\hat{X}^{[m],n}(Z^n)$ by letting the reconstruction symbol at location $m+1 \le i \le n-m$ be given by

$$\hat{X}_i = \arg\min_{\hat{x}} \sum_y \hat{Q}^{2m+1}[Z^n, Y^n(Z^n)](z_{i-m}^{i+m}, y)\Lambda(y, \hat{x}) \tag{137}$$

(and can be arbitrarily defined for i-s outside that range). Note that $\hat{X}_i$ is nothing but the Bayes response to the conditional distribution of Y_{m+1} given $Z^{2m+1} = z_{i-m}^{i+m}$ induced by the joint distribution $\hat{Q}^{2m+1}[Z^n, Y^n(Z^n)]$ of (Z^{2m+1}, Y^{2m+1}). However, from the conclusion in (134) we know that this converges almost surely to the conditional distribution of X_{m+1} given $Z^{2m+1} = z_{i-m}^{i+m}$. Thus, $\hat{X}_i$ in (137) is a Bayes response to a conditional distribution that converges almost surely (as $n \to \infty$) to the true conditional distribution of X_i conditioned on Z_{i-m}^{i+m}. It thus follows from continuity of the performance of Bayes responses (cf., e.g., [Han57, Equation (14)], [WOS+03, Lemma1]) and (135) that

$$\lim_{n\to\infty} \Lambda_n(X^n, \hat{X}^{[m],n}(Z^n)) = E\phi\left(P_{X_0|Z_{-m}^m}\right) \quad \text{a.s.}, \tag{138}$$

with ϕ denoting the Bayes envelope associated with the loss function Λ defined, for $P \in \mathcal{M}(\mathcal{A})$, by

$$\phi(P) = \min_{\hat{x}\in\mathcal{A}} \sum_{x\in\mathcal{A}} \Lambda(x, \hat{x})P(x). \tag{139}$$

Letting $m \to \infty$ in (138) gives

$$\lim_{m\to\infty} \lim_{n\to\infty} \Lambda_n(X^n, \hat{X}^{[m],n}(Z^n)) = E\phi\left(P_{X_0|Z_{-\infty}^\infty}\right) \quad \text{a.s.}, \tag{140}$$

where the right side is the asymptotic optimum distribution-dependent denoising performance (cf. [WOS+03, Section 5]). Note that $\hat{X}^{[m],n}$ can be chosen independently of a source, e.g., taking the Yang-Kieffer codes of [YK96] (followed by the postprocessing detailed in (137)). Thus, (140) implies that the post-processing step leads to essentially asymptotically optimum denoising performance. Bounded convergence implies also

$$\lim_{m\to\infty} \lim_{n\to\infty} E\Lambda_n(X^n, \hat{X}^{[m],n}(Z^n)) = E\phi\left(P_{X_0|Z_{-\infty}^\infty}\right) \tag{141}$$

implying, in turn, the existence of a deterministic sequence $\{m_n\}$ such that for $\hat{X}^n = \hat{X}^{[m_n],n}$

$$\lim_{n\to\infty} E\Lambda_n(X^n, \hat{X}^{[m_n],n}(Z^n)) = E\phi\left(P_{X_0|Z_{-\infty}^{\infty}}\right). \tag{142}$$

The sequence $\{m_n\}$ for which (142) holds, however, may depend on the particular code sequence $\{Y^n(\cdot)\}$, as well as on the distribution of the active source $\mathbf{X}$.

5 Conclusion and Open Questions

In this work we have seen that a rate constraint on a source code places a similar limitation on the mutual information associated with its k-th order marginalization, for any finite k, as n becomes large. This was shown to be the case both in the distributional sense of Section 3 and the pointwise sense of Section 4. This was also shown to imply, in various quantitative senses, that the empirical distribution of codes that are good (in the sense of asymptotically attaining a point on the rate distortion curve) becomes close to that attaining the minimum mutual information problem associated with the rate distortion function of the source. For a source corrupted by additive "white" noise this was shown to imply that, under the right distortion measure, and at the right distortion level, the joint empirical distribution of the noisy source and reconstruction associated with a good code tends to "imitate" the joint distribution of the noisy and noise-free source. This property led to performance bounds (both in expectation and pointwise) on the denoising performance of such codes. It was also seen to imply the existence of a simple postprocessing procedure that, when applied to these compression-based denoisers, results in essentially optimum denoising performance.

One of the salient questions left open in the context of the third item in Theorem 3 (respectively, Theorem 9) is whether the convergence in distribution (in some appropriately defined sense) continues to hold in cases beyond those required in the second item. Another interesting question, in the context of the postprocessing scheme of subsection 4.E, is whether there exists a universally good code sequence (for the noisy source sequence under the distortion measure and level associated with the given noise distribution) and a corresponding growth rate for m under which (142) holds simultaneously for all stationary ergodic underlying noise-free sources.

Appendix

A Proof of Lemma 1

We shall use the following elementary fact from analysis.

Fact 3 *Let $f : \mathbb{D} \to \mathbb{R}$ be continuous and $\mathbb{D}$ be compact. Assume further that $\min_{z\in\mathbb{D}} f(z)$ is uniquely attained by some $z_* \in \mathbb{D}$. Let $z_n \in \mathbb{D}$ satisfy*

$$\lim_{n\to\infty} f(z_n) = f(z_*)\left(= \min_{z\in\mathbb{D}} f(z)\right).$$

Then

$$\lim_{n\to\infty} z_n = z_*.$$

Equipped with this, we prove Lemma 1 as follows. Let $P_{Y|X}^{(n)}$ denote the conditional distribution of Y given X induced by $P_{X,Y}^{(n)}$ and define

$$\hat{P}_{X,Y}^{(n)} = P_X \otimes P_{Y|X}^{(n)}. \tag{A.1}$$

It follows from (32) that

$$\|\hat{P}_{X,Y}^{(n)} - P_{X,Y}^{(n)}\| \to 0 \tag{A.2}$$

and therefore also, by (34),

$$\lim_{n\to\infty} E_{\hat{P}_{X,Y}^{(n)}} \rho(X,Y) = D. \tag{A.3}$$

Define further $\tilde{P}_{X,Y}^{(n)}$ by

$$\tilde{P}_{X,Y}^{(n)} = \alpha_n \hat{P}_{X,Y}^{(n)} + (1-\alpha_n) P_X \otimes \delta_{Y|X}, \tag{A.4}$$

where

$$\alpha_n = \min\left\{ \frac{D}{E_{\hat{P}_{X,Y}^{(n)}} \rho(X,Y)}, 1 \right\}. \tag{A.5}$$

By construction,

$$\tilde{P}_X^{(n)} = P_X \quad \text{and} \quad E_{\tilde{P}_{X,Y}^{(n)}} \rho(X,Y) \le D \quad \forall n. \tag{A.6}$$

Also, (A.3) implies that $\alpha_n \to 1$ and hence $\|\hat{P}_{X,Y}^{(n)} - \tilde{P}_{X,Y}^{(n)}\| \to 0$ implying, when combined with (A.2), that

$$\|\tilde{P}_{X,Y}^{(n)} - P_{X,Y}^{(n)}\| \to 0 \tag{A.7}$$

implying, in turn, by (33) and uniform continuity of $I(\cdot)$,

$$\lim_{n\to\infty} I\left(\tilde{P}_{X,Y}^{(n)}\right) = R(D). \tag{A.8}$$

Now, the set $\left\{P'_{X,Y} : P'_X = P_X, E_{P'_{X,Y}} \rho(X,Y) \le D\right\}$ is compact, $I(\cdot)$ is continuous, and $R(D)$ is uniquely achieved by $P_{X,Y}$. So (A.6), (A.8), and Fact 3 imply

$$\tilde{P}_{X,Y}^{(n)} \stackrel{d}{\Rightarrow} P_{X,Y}, \tag{A.9}$$

implying (35) by (A.7). □

B Proof of Lemma 3

Inequality (83) implies the existence of $\eta > 0$ and n_0 such that

$$|A_n| \le e^{n[\overline{H}(\mathbf{X})-2\eta]} \quad \forall n \ge n_0. \tag{A.10}$$

Defining

$$G_{n,\eta} = \left\{ x^n : \left| -\frac{1}{n} \log P_{X^n}(x^n) - \overline{H}(\mathbf{X}) \right| \le \eta \right\}, \tag{A.11}$$

we have for all $n \geq n_0$

$$\begin{aligned}
\Pr(X^n \in G_{n,\eta} \cap A_n) &= \sum_{x^n \in G_{n,\eta} \cap A_n} P_{X^n}(x^n) && \text{(A.12)} \\
&\leq \sum_{x^n \in G_{n,\eta} \cap A_n} e^{n[-\overline{H}(\mathbf{X})+\eta]} && \text{(A.13)} \\
&\leq |A_n| e^{n[-\overline{H}(\mathbf{X})+\eta]} && \text{(A.14)} \\
&\leq e^{-n\eta}. && \text{(A.15)}
\end{aligned}$$

Thus, by the Borel-Cantelli lemma, with probability one, $X^n \in G_{n,\eta} \cap A_n$ for only finitely many n. On the other hand, by the Shannon-McMillan-Breiman Theorem (cf., e.g., [CT91, Theorem 15.7.1]), with probability one, $X^n \in G_{n,\eta}^c$ for only finitely many n. The result now follows since $A_n \subseteq G_{n,\eta}^c \cup (G_{n,\eta} \cap A_n)$. □

References

[Ber71] T. Berger. *Rate-Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, Englewood Cliffs, N.J., 1971.

[BG98] T. Berger and J. D. Gibson. Lossy source coding. *IEEE Trans. Inform. Theory*, 44(6):2693–2723, October 1998.

[CK81] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, New York, 1981.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[CYV97] G. Chang, B. Yu, and M. Vetterli. Bridging compression to wavelet thresholding as a denoising method. *Proc. Conf. Inf. Sciences and Systems*, 1997.

[Don02] D. Donoho. The Kolmogorov sampler. January 2002. (available at: `http://www-stat.stanford.edu/donoho/`).

[DW03] A. Dembo and T. Weissman. The minimax distortion redundancy in noisy source coding. *IEEE Trans. Inform. Theory*, 49(11), November 2003.

[EM02] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Trans. Inform. Theory*, June 2002.

[Gal68] R. G. Gallager. *Information Theory and Reliable Communication*. Wiley, New York, 1968.

[Gra70] R. M. Gray. Information rates of autoregressive processes. *IEEE Trans. Info. Theory*, IT-16:412–421, July 1970.

[Gra71] R. M. Gray. Rate distortion functions for finite-state finite-alphabet markov sources. *IEEE Trans. Inform. Theory*, IT-17(2):127–134, March 1971.

[Han57] J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games, Ann. Math. Study*, III(39):97–139, 1957. Princeton University Press.

[JY01] R. Jörnstein and B. Yu. Adaptive quantization. *Technical Report, Department of statistics*, 2001. UC Berkeley.

[Kan97] A. Kanlis. *Compression and Transmission of Information at Multiple Resolutions*. PhD thesis, Electrical Engineering Department, University of Maryland at College Park, August 1997.

[Kie78] J. C. Kieffer. A unified approach to weak universal source coding. *IEEE Transactions on Information Theory*, IT–24(6):674–682, November 1978.

[Kie89] J. C. Kieffer. An almost sure convergence theorem for sequences of random variables selected from log-convex sets. In *Proc. Conf. Almost Everywhere Convergence in Probab. Ergodic Theory II*, Evanston, IL, 1989. Northwestern Univ.

[Kie91] J. C. Kieffer. Sample converses in source coding theory. *IEEE Trans. Info. Theory*, IT-37:263–268, March 1991.

[KSP96] A. Kanlis, S.Khudanpur, and P.Narayan. Typicality of a good rate-distortion code. *Problems of Information Transmission (Problemy Peredachi Informatsii)*, January 1996. Special issue in honor of M. S. Pinsker.

[MF98] N. Merhav and M. Feder. Universal prediction. *IEEE Trans. Inform. Theory*, IT-44(6):2124–2147, October 1998.

[Nat93] B. Natarajan. Filtering random noise via data compression. *Data Compression Conference, DCC '93*, pages 60–69, 1993.

[Nat95] B. Natarajan. Filtering random noise from deterministic signals via data compression. *IEEE Trans. Signal Proc.*, 43(11):2595–2605, November 1995.

[NKH98] B. Natarajan, K. Konstantinides, and C. Herley. Occam filters for stochastic sources with application to digital images. *IEEE Trans. Signal Proc.*, 46:1434–1438, November 1998.

[Ris00] J. Rissanen. MDL denoising. *IEEE Trans. Inform. Theory*, IT-46:2537–2543, November 2000.

[Sam63] E. Samuel. An empirical Bayes approach to the testing of certain parametric hypotheses. *Ann. Math. Statist.*, 34(4):1370–1385, 1963.

[SV97] S. Shamai and S. Verdu. The empirical distribution of good codes. *IEEE Trans. Inform. Theory*, 43(3):836–846, May 1997.

[TPB99] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proc. 37th Allerton Conference on Communication and Computation*, 1999.

[TRA02] I. Tabus, J. Rissanen, and J. Astola. Normalized maximum likelihood models for boolean regression with application to prediction and classification in genomics. *Computational and Statistical Approaches to Genomics*, March 2002.

[Wei] T. Weissman. Not all universal source codes are pointwise universal. In preparation.

[WM03] T. Weissman and N. Merhav. On competitive prediction and its relationship to rate-distortion theory. 49(12), December 2003. IEEE Trans. Inform. Theory.

[WOS+03] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger. Universal discrete denoising: Known channel. 2003. Submitted to IEEE Trans. Inform. Theory (available at: `http://www.hpl.hp.com/techreports/2003/HPL-2003-29.pdf`).

[YK96] E. H. Yang and J. Kieffer. Simple universal lossy data compression schemes derived from the lempel-ziv algorithm. *IEEE Trans. Inform. Theory*, 42:239–245, 1996.

[Ziv72] J. Ziv. Coding of sources with unknown statistics - part ii: Distortion relative to a fidelity criterion. *IEEE Trans. Inform. Theory*, IT-18:389–394, May 1972.

[Ziv80] J. Ziv. Distortion-rate theory for individual sequences. *IEEE Trans. Inform. Theory*, IT-26(2):137–143, March 1980.