Critical Behavior in Lossy Source Coding

Amir Dembo, Ioannis Kontoyiannis

Abstract— **The following critical phenomenon was recently discovered. When a memoryless source is compressed using a variable-length fixed-distortion code, the fastest convergence rate of the (pointwise) compression ratio to $R(D)$ is either $O(\sqrt{n})$ or $O(\log n)$. We show it is always $O(\sqrt{n})$, except for discrete, uniformly distributed sources.**

Keywords— **Redundancy, rate-distortion theory, lossy data compression**

I. Introduction

SUPPOSE that data is produced by a stationary memoryless source $\{X_n \,;\, n \geq 1\}$, so that the X_i are independent and identically distributed (IID) random variables with common distribution P. We will assume throughout that the X_i take values in the *source alphabet* A, where A is a subset of $\mathbb{R}$, and that the *reproduction alphabet* $\hat{A}$ is a finite subset of $\mathbb{R}$, say $\hat{A} = \{a_1, a_2, \ldots, a_k\}$.

The main objective of data compression is to find efficient approximate representations for data $x_1^n = (x_1, x_2, \ldots, x_n)$ generated from the source $X_1^n = (X_1, X_2, \ldots, X_n)$. Specifically, we wish to represent each source string x_1^n by a corresponding string $y_1^n = (y_1, y_2, \ldots, y_n)$ taking values in the reproduction alphabet $\hat{A}$, so that the distortion between each x_1^n and its representation lies within some fixed allowable range. For our purposes, distortion is measured by a family of single-letter distortion measures,

$$\rho_n(x_1^n, y_1^n) = \frac{1}{n}\sum_{i=1}^{n} \rho(x_i, y_i) \qquad x_1^n \in A^n,\ \ y_1^n \in \hat{A}^n, \quad (1)$$

where $\rho : A \times \hat{A} \to [0, \infty)$ is a fixed nonnegative function.

We consider *variable-length block codes operating at a fixed distortion level*, that is, codes C_n defined by triplets (B_n, ϕ_n, ψ_n) where:

- (*a*) B_n is a subset of $\hat{A}^n$ called the *codebook*;
- (*b*) $\phi_n : A^n \to B_n$ is a map called the *encoder*;
- (*c*) $\psi_n : B_n \to \{0,1\}^*$ is a prefix-free representation of the elements of B_n by finite-length binary strings.

For a fixed distortion level $D \geq 0$, the code $C_n = (B_n, \phi_n, \psi_n)$ is said to *operate at distortion level* D [8] if it encodes each source string with distortion D or less:

$$\rho_n(x_1^n, \phi_n(x_1^n)) \leq D \quad \text{for all} \ \ x_1^n \in A^n.$$

A. Dembo is with the Department of Mathematics and the Department of Statistics, Stanford University, Stanford, CA 94305. Email: amir@stat.stanford.edu. I. Kontoyiannis is with the Department of Statistics, Purdue University, 1399 Mathematical Sciences Building, W. Lafayette, IN 47907-1399. Email: yiannis@stat.purdue.edu.

A.D.'s research was supported in part by NSF grant #DMS-0072331. I.K.'s research was supported in part by NSF grant #0073378-CCR and by a grant from the Purdue Research Foundation.

Our main quantity of interest here is the description length of a block code C_n, expressed by its length function $\ell_n : A^n \to \mathbb{N}$:

$$\ell_n(x_1^n) = \ \text{length of} \ [\psi_n(\phi_n(x_1^n))].$$

Broadly speaking, the smaller the description length, the better the code.

Shannon's celebrated source coding theorem states that, for an arbitrary sequence of block codes $\{C_n = (B_n, \phi_n, \psi_n) \,;\, n \geq 1\}$ operating at distortion level D, the expected compression ratio $E[\ell_n(X_1^n)]/n$ is asymptotically bounded below by the rate-distortion function $R(D)$:

$$\liminf_{n\to\infty} \frac{E[\ell_n(X_1^n)]}{n} \geq R(D) \quad \text{bits per symbol.}$$

Moreover, Shannon showed that there exist codes achieving the above lower bound with equality; see Shannon's 1959 paper [11] or Berger's classic text [4]. A stronger version of Shannon's theorem was proved by Kieffer in 1991 [8], where it is shown that the rate-distortion function is a *pointwise* asymptotic lower bound for $\ell_n(X_1^n)$:

$$\liminf_{n\to\infty} \frac{\ell_n(X_1^n)}{n} \geq R(D) \quad \text{with prob. 1.} \qquad (2)$$

In [8] it is also demonstrated that the bound in (2) can be achieved with equality.

The following refinement to Kieffer's result was recently given in [10]:

(POINTWISE REDUNDANCY): *For any sequence of block codes $\{C_n\}$ with associated length functions $\{\ell_n\}$, operating at distortion level D,*

$$\ell_n(X_1^n) \quad \geq \quad nR(D) + \sum_{i=1}^{n} f(X_i) - 2\log n$$

eventually, with prob. 1, (3)

where $f : A \to \mathbb{R}$ is a bounded function depending on P and D but not *on the codes $\{C_n\}$, such that $E_P[f(X_1)] = 0$. Moreover, there exist codes $\{C_n, \ell_n\}$ that achieve*

$$\ell_n(X_1^n) \quad \leq \quad nR(D) + \sum_{i=1}^{n} f(X_i) + 5\log n$$

eventually, with prob. 1. (4)

[*cf.* Theorems 4 and 5 and eq. (18) in [10]; above and throughout the paper, 'log' denotes the logarithm taken to base 2 and '$\log_e$' denotes the natural logarithm.] The function f is defined precisely in Section III; here we just mention the following interpretation: If we write $\tilde{f}(x) =$

$f(x)+R(D)$, then $\tilde{f}$ can be expressed in a natural way in terms of familiar information theoretic quantities. In particular, $E(\tilde{f}(X_1))=R(D)$, its variance $\sigma^2=\mathrm{Var}(\tilde{f}(X_1))$ is the "minimal coding variance" of the source with distribution P [10], and in the case of lossless compression (as $D\downarrow 0$), $\tilde{f}(x)$ reduces to $-\log P(x)$.

The above result says that, for any source distribution P and any sequence of codes $\{C_n\}$ operating at distortion level D, the "pointwise redundancy" in the description lengths of the codes C_n, namely, the difference between $\ell_n(X_1^n)$ and the optimum $nR(D)$ bits, is essentially bounded below by the sum of the IID, bounded, zero-mean random variables $f(X_i)$. So there are two possibilities:

- Either the random variables $f(X_i)$ are non-constant, in which case the best achievable pointwise redundancy rate will be of order $O(\sqrt{n})$ (by the central limit theorem and the upper and lower bounds in (3) and (4));
- or the random variables $f(X_i)$ are equal to zero with probability one, in which case the best achievable pointwise redundancy is no more than $(5\log n)$ bits, eventually (by (4)).

To be more precise, in the first case when the random variables $f(X_i)$ are not constant, the central limit theorem implies that the term $\sum_{i=1}^n f(X_i)$ is of order $O(\sqrt{n})$ in probability, and therefore, by (3) and (4), the best achievable pointwise redundancy will also be of order $O(\sqrt{n})$ in probability. [In a similar fashion, the law of the iterated logarithm implies that the *pointwise* fluctuations of the best achievable pointwise redundancy will be of order $O(\sqrt{n\log_e\log_e n})$; see [10, Section I] for a more detailed discussion. Also the contrast between the pointwise and the expected redundancy rate is interpreted and commented on in [10, Remark 3, p.139].]

Our purpose in this paper is to characterize exactly when each one of the above two cases occurs, namely, when the minimal pointwise redundancy is $O(\sqrt{n})$ and when it is $O(\log n)$. In the next section we show that it is almost never the case that $f(X_1)=0$ with probability one, so the minimal pointwise redundancy is typically of order $\sqrt{n}$. In particular, in the common case when the X_i take values in a finite alphabet $A=\hat{A}$, then (under mild conditions) we show that $f(X_1)=0$ with probability one if and only if the X_i are uniformly distributed.

Before stating our main results (Theorems 1, 2 and 3 in the next section) in detail, we recall the following representative examples from [9] and [10].

Example 1 (Lossless Compression) For a source $\{X_n\}$ with distribution P on the finite alphabet A, a lossless code C_n is a prefix-free map $\psi_n : A^n\to\{0,1\}^*$. [Or, to be pedantic, in our setting a lossless code is a code operating at distortion level $D=0$ with respect to Hamming distortion.] In this case the function f has the simple form

$$f(x)=-\log P(x)-H(P) \tag{5}$$

where $H(P)=E_P[-\log P(X_1)]$ is the entropy of P, and the lower bound (3) is simply

$$\begin{aligned}\ell_n(X_1^n) &\geq nH(P)+\sum_{i=1}^n f(X_i)-2\log n\\ &= -\log P(X_1^n)-2\log n \qquad (6)\\ &\qquad\text{eventually, with prob. 1.}\end{aligned}$$

The lower bound (6) is a well-known information-theoretic fact called Barron's lemma (see [2][3] and the discussion in [10]). It says that the description lengths $\ell_n(X_1^n)$ of an arbitrary sequence of codes are (eventually with probability 1) bounded below by the idealized Shannon code lengths $-\log P(X_1^n)$, up to terms of order $\log n$. From (5) it is obvious that $f(X_1)=0$ with probability one if and only if P is the uniform distribution on A.

Example 2 (Binary Source, Hamming Distortion) This is the simplest non-trivial lossy example. Suppose $\{X_n\}$ is a binary source with Bernoulli(p) distribution for some $p\in(0,1/2]$. Let $A=\hat{A}=\{0,1\}$ and take ρ to be the Hamming distortion measure, $\rho(x,y)=0$ when $x=y$, and equal to 1 otherwise. For each fixed $D\in(0,p)$ it is shown in [10] that

$$f(x)=-\log\left(\frac{P(x)}{1-D}\right)-E_P\left[-\log\left(\frac{P(X_1)}{1-D}\right)\right],$$

from which it is again obvious that $f(X_1)=0$ with probability one if and only if $p=1/2$, i.e., if and only if P is the uniform distribution on $A=\{0,1\}$.

In a third example presented in [10] it is also found that $f(X_1)=0$ with probability one if and only if P is the uniform distribution, and the natural question is raised as to whether this pattern persists in general. In the next section we answer this question by showing (in Theorem 1 and Corollary 1) that for a source distribution P on a finite alphabet, $f(X_1)$ can be equal to zero with probability one for at most finitely many distortion levels D, *unless* P is the uniform distribution and ρ is a "permutation" distortion measure. In Theorems 2 and 3 and in Corollary 2 the continuous case is considered, and it is shown that when P is a continuous distribution it essentially never happens that $f(X_1)=0$ with probability one. Section III contains the proofs of Theorems 1, 2 and 3 and Corollaries 1 and 2.

II. Results

Suppose that the source alphabet A is an arbitrary (Borel) subset of $\mathbb{R}$, and let P be a (Borel) probability measure on $\mathbb{R}$, supported on A (the special cases when P is purely discrete or purely continuous are considered separately below). Let $\hat{A}=\{a_1,a_2,\ldots,a_k\}$ be the finite reproduction alphabet of size k. Given an arbitrary, bounded, nonnegative function $\rho : A\times\hat{A}\to[0,M]$ (for some finite constant M), define a sequence of single-letter distortion measures $\rho_n : A^n\times\hat{A}^n\to[0,M]$ as in (1). Throughout the paper, we make the usual assumption:

$$\sup_{x\in A}\min_{y\in\hat{A}}\rho(x,y)=0. \tag{7}$$

[See, e.g., [4, p.26] or [5, Ch.13, ex.4]; if (7) is not satisfied, for example when A is an interval of real numbers, $\hat{A}$ is a finite set, and $\rho(x,y) = (x-y)^2$, we may consider the distortion measure $\rho'(x,y) = \rho(x,y) - \min_{z\in\hat{A}} \rho(x,z)$ instead.] For $D \geq 0$, the *rate-distortion function* of a memoryless source with distribution P is

$$R(D) = \inf_{(X,Y)} I(X;Y) \tag{8}$$

where the infimum is over all jointly distributed random variables (X,Y) with values in $A\times\hat{A}$ such that X has distribution P and $E[\rho(X,Y)] \leq D$; $I(X;Y)$ denotes the mutual information (in bits) between X and Y (see [4] for more details). Under our assumptions, the rate-distortion function $R(D)$ is a convex, nonincreasing function of $D \geq 0$, and it is finite for all D.

For a fixed distribution P on A, let

$$D_{\max} = D_{\max}(P) = \min_{y\in\hat{A}} E_P[\rho(X,y)]$$

and recall that $R(D) = 0$ for $D \geq D_{\max}$ (see, e.g., Proposition 1 in Section III). In order to avoid the trivial case when $R(D)$ is identically zero we assume that $D_{\max} > 0$, and from now on we restrict our attention to the interesting range of distortion levels $D \in (0, D_{\max})$.

A. The Discrete Case: $A = \hat{A}$

We first consider the most common case where the source $\{X_n\}$ takes values in a finite alphabet $A = \hat{A} = \{a_1, a_2, \ldots, a_k\}$. Suppose that $\{X_n\}$ are IID with common distribution P on A, and assume, without loss of generality, that $P_i = P(a_i) > 0$ for all $i = 1, \ldots, k$. Given a distortion measure ρ, write ρ_{ij} for $\rho(a_i, a_j)$. We assume throughout this section that ρ is symmetric, i.e., that $\rho_{ij} = \rho_{ji}$ for all i, j, and also that $\rho_{ij} = 0$ if and only if $i = j$. We call ρ a *permutation distortion measure*, if all rows of the matrix $(\rho_{ij})_{i,j=1,\ldots,k}$ are permutations of one another (which, by symmetry, is equivalent to saying that all columns are permutations of one another).

Recall that the minimal pointwise redundancy is of order $O(\log n)$ if and only if $f(X_1) = 0$ with probability one; otherwise it is $O(\sqrt{n})$. Our first result says that the rate cannot be $O(\log n)$ for many distortion levels D, unless the distribution P is uniform in which case the rate is $O(\log n)$ for all distortion levels D.

Theorem 1:
(a) If P is the uniform distribution on A and ρ is a permutation distortion measure, then $f(X_1) = 0$ with probability one for all $D \in (0, D_{\max})$.
(b) If $f(X_1) = 0$ with probability one for a sequence of distortion values $D_n \in (0, D_{\max})$ such that $D_n \downarrow 0$, then P is the uniform distribution and ρ is a permutation distortion measure, and therefore $f(X_1) = 0$ with probability one for *all* $D \in (0, D_{\max})$.

As we mentioned above, the rate-distortion function $R(D)$ is convex for $D \in (0, D_{\max})$. If it is *strictly* convex (as it is usually the case – see the discussion in [4, Chapter 2]), then Theorem 1 can be strengthened to the following.

Corollary 1: Suppose $R(D)$ is strictly convex over the range $D \in (0, D_{\max})$. If $f(X_1) = 0$ with probability one for infinitely many $D \in (0, D_{\max})$ then P is the uniform distribution and ρ is a permutation distortion measure, and therefore $f(X_1) = 0$ with probability one for *all* $D \in (0, D_{\max})$.

Remark. In the examples presented in the previous section it turned out that either $f(X_1) = 0$ with probability one for *all* D, or it was never the case. But it may happen that $f(X_1) = 0$ with probability one only for a few isolated values of D, while P is *not* the uniform distribution. Such an example is given after Lemma 3 in Section III-B.

B. The Continuous Case: $A = \mathbb{R}$

Here we take $A = \mathbb{R}$ and we assume that the distribution P of the source has a positive density g (with respect to Lebesgue measure), or, more generally, that there exists a (nonempty) open interval $I \subset \mathbb{R}$ on which P has an absolutely continuous component with density g such that $g(x) > 0$ for $x \in I$. Since the reproduction alphabet $\hat{A} = \{a_1, a_2, \ldots, a_k\}$ is finite, given a distortion measure ρ we can write $r_j(x) = \rho(x, a_j)$ for all $1 \leq j \leq k$ and all $x \in A$. We assume that for all j the functions r_j are continuous on I. For convenience we also define, for $j = 0$, $r_j(x) \equiv 0$ on I.

Our next result gives a sufficient condition on the distortion measures r_j, under which the best redundancy rate in (2) can *never* be $O(\log n)$.

Theorem 2: If for every $\lambda < 0$ the functions

$$e^{\lambda r_j(\cdot)}, \quad j = 0, 1, \ldots, k$$

are linearly independent on I, then $f(X_1)$ cannot be equal to zero with probability one for any distortion level $D \in (0, D_{\max})$.

Next we provide a somewhat simpler set of conditions, under which we get a weaker conclusion. Theorem 3 says that the best redundancy rate in (2) cannot be $O(\log n)$ for many distortion levels D.

Theorem 3: Under either one of the following two conditions, $f(X_1)$ cannot be equal to zero with probability one for distortion levels $D > 0$ arbitrarily close to zero.
(a) There exist (distinct) points $\{x_0, x_1, \ldots, x_k\}$ in I such that, for all $0 \leq i \neq j \leq k$, with $j \neq 0$, we have $r_j(x_j) > r_j(x_i)$.
(b) There exist (distinct) points $\{x_0, x_1, \ldots, x_k\}$ in I such that, for every permutation π of the indices $\{0, 1, \ldots, k\}$ with π not equal to the identity, we have

$$\sum_{j=0}^{k} r_j(x_j) \neq \sum_{j=0}^{k} r_j(x_{\pi(j)}).$$

Although the conditions of Theorems 2 and 3 may seem unusual, they are natural and generally easy to verify. To illustrate this, we present below two simple examples.

Example 3 (Mean-Squared Error) Suppose P has a positive density on the interval $I=[-2,2]$, let $\hat{A}$ consist of the two reproduction points ± 1, and let ρ be the mean-squared error distortion measure. Recall that, to satisfy (7), $\rho(x,y)$ is actually defined by

$$\rho(x,y)=(x-y)^2-\min\{(x-1)^2,(x+1)^2\}.$$

The corresponding distortion functions $r_1(x)=\rho(x,-1)$ and $r_2(x)=\rho(x,+1)$ are shown in Figure 1. Here, condition (a) of Theorem 3 is easily seen to hold with $x_0=0$, $x_1=2$ and $x_2=-2$.

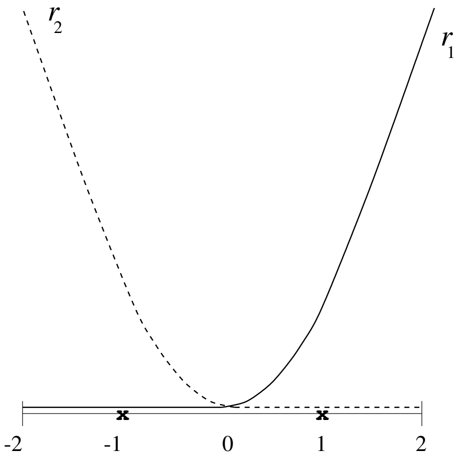

Fig. 1. Distortion measure in Example 3. Reproduction points are shown as $\mathbf{x}$'s.

Example 4 (L^1 Distance) Suppose P has a positive density on the interval $I=[0,6]$, let $\hat{A}=\{1,3,5\}$, and take ρ to be the normalized L^1 distance $|x-y|$ adjusted so that (7) is satisfied; the resulting functions $r_j(\cdot)$ are shown in Figure 2. Here it is easy to verify that the condition of Theorem 2 is satisfied, i.e., that the functions $\{e^{\lambda r_j(\cdot)}\ ;\ 0\le j\le 3\}$ are linearly independent on I. For this it suffices to observe that $e^{\lambda r_1}$ and $e^{\lambda r_3}$ are linearly independent on $[2,4]$ (essentially because the functions $e^{\lambda x}$ and $e^{-\lambda x}$ are linearly independent on $[0,2]$), and that $e^{\lambda r_2}$ is not constant outside $[2,4]$.

Like in the discrete case, under some additional assumptions on the rate-distortion function $R(D)$, it is possible to get a stronger version of Theorem 3:

Corollary 2. Suppose $R(D)$ is differentiable and strictly convex on $(0,D_{\max})$. Under either one of the assumptions (a) and (b) in Theorem 3, there can be at most finitely many $D\in(0,D_{\max})$ such that $f(X_1)=0$ with probability one.

Remark. Under somewhat more restrictive assumptions on the distortion measure ρ, it is possible to prove that, for any P with a continuous component as above, there can be at most $k(k+1)/2$ distortion levels D for which $f(X_1)=0$ with probability one. Since the proof of this slightly stronger result relies on an argument different from the ones used to prove Theorems 2 and 3, we omit it here.

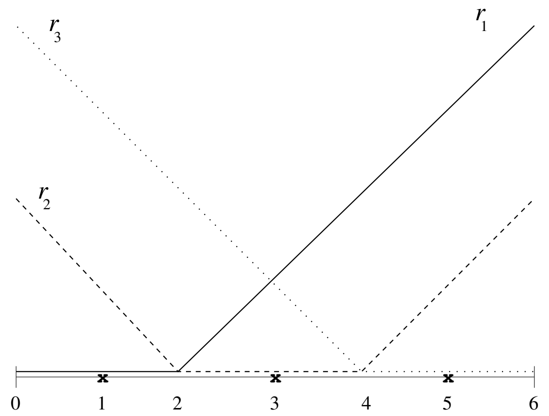

Fig. 2. Distortion measure in Example 4. Reproduction points are shown as $\mathbf{x}$'s.

III. Proofs

A. Preliminaries

Before giving the proofs of Theorems 1, 2 and 3, we recall some definitions and notation from [10] and give the precise form of the function f (see equation (12) below).

Let P be a source distribution on A, and let Q be an arbitrary probability mass function on $\hat{A}$. Write X for a random variable with distribution P on A, and Y for an independent random variable with distribution Q on $\hat{A}$. Let $S=\{a\in\hat{A}\ :\ Q(a)>0\}$ be the support of Q and define

$$\begin{aligned} D_{\min}^{P,Q} &= E_P\left[\min_{a\in S}\rho(X,a)\right] \\ D_{\max}^{P,Q} &= E_{P\times Q}\left[\rho(X,Y)\right]. \end{aligned}$$

For $\lambda\le 0$, let

$$\Lambda_{P,Q}(\lambda)\;=\;E_P\left[\log_e E_Q\left(e^{\lambda\rho(X,Y)}\right)\right],$$

and for $D\ge 0$ write $\Lambda^*_{P,Q}$ for the Fenchel-Legendre transform of $\Lambda_{P,Q}$,

$$\Lambda^*_{P,Q}(D)=\sup_{\lambda\le 0}\,[\lambda D-\Lambda_{P,Q}(\lambda)].$$

We also define

$$R(P,Q,D)\;=\;\inf_{(X,Z)}\,[I(X;Z)+H(Q_Z\|Q)]$$

where $H(R\|Q)=\sum_{a\in\hat{A}}R(a)\log[R(a)/Q(a)]$ denotes the relative entropy (in bits) between R and Q, Q_Z denotes the distribution of Z, and the infimum is over all jointly distributed random variables (X,Z) with values in $A\times\hat{A}$ such that X has distribution P and $E[\rho(X,Z)]\le D$. In view of (8), we clearly have

$$R(D)=\inf_{\text{all }Q}R(P,Q,D). \tag{9}$$

In Lemma 1 and Proposition 1 below we summarize some useful properties of $\Lambda_{P,Q}$, $\Lambda^*_{P,Q}$ and $R(P,Q,D)$ (see Lemma 1 and Propositions 1 and 2 in [10]).

Lemma 1:
(i) $\Lambda_{P,Q}$ is infinitely differentiable on $(-\infty,0)$, and $\Lambda''_{P,Q}(\lambda)\geq 0$ for all $\lambda\leq 0$.
(ii) If $D\in(D^{P,Q}_{\min},D^{P,Q}_{\max})$ then there exists a unique $\lambda<0$ such that $\Lambda'_{P,Q}(\lambda)=D$ and $\Lambda^*_{P,Q}(D)=\lambda D-\Lambda_{P,Q}(\lambda)$.

Proposition 1:
(i) For all $D\geq 0$,

$$R(P,Q,D)=\inf_W E_P\left[H(W(\cdot|X)\|Q(\cdot))\right],$$

where the infimum is over all probability measures W on $A\times\hat{A}$ such that the A-marginal of W equals P and $E_W[\rho(X,Y)]\leq D$.
(ii) For all $D\geq 0$, $R(P,Q,D)=(\log e)\Lambda^*_{P,Q}(D)$.
(iii) For $0<D<D_{\max}$ we have $0<R(D)<\infty$, whereas for $D\geq D_{\max}$, $R(D)=0$.
(iv) For every $D\in(0,D_{\max})$ there exists a $Q=Q^*$ on $\hat{A}$ achieving the infimum in (9), and $D\in(D^{P,Q^*}_{\min},D^{P,Q^*}_{\max})$.

For any distribution P on A and any distortion level $D\in(0,D_{\max}(P))$, by Proposition 1 we can pick a Q^* achieving the infimum in (9) so that $R(D)=R(P,Q^*,D)$ and also $D\in(D^{P,Q^*}_{\min},D^{P,Q^*}_{\max})$, so by Lemma 1 we can pick a $\lambda^*<0$ with

$$\begin{aligned}\lambda^*D-\Lambda_{P,Q^*}(\lambda^*) &= \Lambda^*_{P,Q^*}(D)\\ &= (\log_e 2)R(P,Q^*,D)\\ &= (\log_e 2)R(D).\end{aligned}\tag{10}$$

Note also that

$$\lambda^*\to-\infty \quad \text{as} \quad D\to 0 \tag{11}$$

(see the Appendix for a short proof). Finally we can define the function f, for $x\in A$,

$$f(x)\triangleq(\log e)\left[\lambda^*D-\log_e E_{Q^*}\left(e^{\lambda^*\rho(x,Y)}\right)\right]-R(D). \tag{12}$$

Since $E_P[f(X_1)]=0$, $f(X_1)=0$ with probability one if and only if

$$\sum_{j=1}^k Q^*(a_j)e^{\lambda^*\rho(x,a_j)}=\text{Constant}, \text{ for } P\text{-almost all } x. \tag{13}$$

Next we give an useful interpretation for the constant λ^* in the representation of $R(D)$ in (10): If $R(D)$ is differentiable at D, then λ^* is proportional to its slope at D; Lemma 2 is proved in the Appendix.

Lemma 2: For any $D\in(0,D_{\max})$:
(i) We have $(\log_e 2)R(D)=\sup_{\lambda\leq 0}[\lambda D-\Gamma(\lambda)]$, where $\Gamma(\lambda)=\sup_Q\Lambda_{P,Q}(\lambda)$.
(ii) Let λ^* be chosen as in (10). If $R(\cdot)$ is differentiable at D, then $\lambda^*=(\log_e 2)R'(D)$.

B. Proofs in the Discrete Case

For the proof of Theorem 1 we will need the following lemma. It easily follows from Theorem 3.7 in Chapter 2 of [6] (see the Appendix). Recall the notation $P_i=P(a_i)$ and $\rho_{ij}=\rho(a_i,a_j)$.

Lemma 3: A probability mass function Q^* on A achieves the infimum in (9) if and only if there exists a $\lambda^*<0$ such that the following all hold:
(a) $\Lambda'_{P,Q^*}(\lambda^*)=D$.
(b) If we define, for $a_i,a_j\in A$,

$$W(a_i,a_j)=P_iQ^*(a_j)\frac{e^{\lambda^*\rho_{ij}}}{\sum_{j'}Q^*(a_{j'})e^{\lambda^*\rho_{ij'}}}$$

then the second marginal of W is Q^*.
(c) If $Q^*(a_j)=0$ for some j, then

$$\sum_i P_i\frac{e^{\lambda^*\rho_{ij}}}{\sum_{j'}Q^*(a_{j'})e^{\lambda^*\rho_{ij'}}}\leq 1.$$

Example 5: Here we present a simple example illustrating the fact that it may happen that $f(X_1)=0$ for a few isolated values D even when P is not uniform. Take $A=\hat{A}=\{0,1,2\}$, let $\alpha=\log_e[3e/(4-e)]$, and consider the distortion measure

$$(\rho_{ij})=\begin{pmatrix}0 & 1 & \alpha\\ 1 & 0 & \alpha\\ \alpha & \alpha & 0\end{pmatrix}.$$

Then with $P=Q^*=(4/13,4/13,5/13)$ and $\lambda^*=-1$, it is straightforward to check that condition (b) of Lemma 3 holds (condition (c) is irrelevant here), and also (13) is satisfied. Therefore, at $D=\Lambda'_{P,Q^*}(\lambda^*)\approx 0.43$, we must have $f(X_1)=0$ with probability one. [Note, also, that the distortion measure used here is not a permutation distortion measure.]

Proof of Theorem 1, (a): Suppose ρ is a permutation distortion measure and P is the uniform distribution on A, $P_i=1/k$ for all $i=1,\ldots,k$. First we claim that for any $D\in(0,D_{\max})$ we can take Q^* to also be uniform. With $Q^*(a_j)=1/k$ for all j, it suffices to find $\lambda^*<0$ satisfying (a) and (b) of Lemma 3 (part (c) is irrelevant here). We have $D^{P,Q^*}_{\min}=0$ and

$$D^{P,Q^*}_{\max}=\sum_{i,j}\frac{1}{k}\frac{1}{k}\rho_{ij}=\frac{1}{k}\Sigma$$

where $\Sigma\triangleq\sum_i\rho_{ij}$, which is independent of j (since ρ is a permutation). Also by the permutation property, $D_{\max}=\min_j E_P[\rho(X,a_j)]=\min_j\sum_i(1/k)\rho_{ij}=(1/k)\Sigma$. Choose and fix a $D\in(0,D_{\max})$, and pick $\lambda^*<0$ as in (10) so that Lemma 3 (a) holds. With this λ^* and Q^* being uniform let W^* be as in Lemma 3 (b); then

$$\sum_i W^*(a_i,a_j)=\sum_i\frac{1}{k}\frac{1}{k}\frac{e^{\lambda^*\rho_{ij}}}{\sum_{j'}\frac{1}{k}e^{\lambda^*\rho_{ij'}}}=\frac{1}{k}\sum_i\frac{e^{\lambda^*\rho_{ij}}}{\sum_{j'}e^{\lambda^*\rho_{ij'}}}.$$

But the sum in the denominator above

$$\sum_{j'} e^{\lambda^* \rho_{ij'}} \quad \text{is independent of } i \tag{14}$$

because ρ is a permutation, so $\sum_i W^*(a_i, a_j) = 1/k = Q^*(a_j)$, and (b) is satisfied. This proves that we can take Q^* to be uniform. Now simply multiplying (14) by $1/k$ we obtain (13), and this implies that $f(x) = 0$ for all $x \in A$. Since $D \in (0, D_{\max})$ was arbitrary, we are done. □

Proof of Theorem 1, (b): Let D_n, $n \geq 1$, be a sequence of of distortion values in $(0, D_{\max})$ for which $f(X_1) = 0$ with probability one, and such that $D_n \downarrow 0$. For each D_n, we can pick Q_n and a $\lambda_n < 0$ as in (10) such that $R(D_n) = (\log e)\Lambda^*_{P,Q_n}(\lambda_n)$ and $D_n = \Lambda'_{P,Q_n}(\lambda_n)$. Let

$$\widetilde{D} = (\min_i P_i)(\min_{i \neq j} \rho_{ij}) > 0.$$

Then for all n large enough so that $D_n < \widetilde{D}$, we must have $Q_n(a_i) > 0$ for all i (otherwise it is trivial to check that $\Lambda'_{P,Q_n}(\lambda) \geq \widetilde{D}$ for any $\lambda < 0$, contradicting the choice of λ_n). From now on we restrict attention to these large enough n's. As discussed above, $f(X_1) = 0$ with probability one if and only if condition (13) holds, which, in this case, becomes

$$\sum_{j=1}^{k} Q_n(a_j) e^{\lambda_n \rho_{ij}} \quad \text{is independent of } i. \tag{15}$$

By Lemma 3 (b) we have that for all j

$$\sum_i P_i \frac{e^{\lambda_n \rho_{ij}}}{\sum_{j'} Q_n(a_{j'}) e^{\lambda^* \rho_{ij'}}} = 1,$$

but by (15) the denominator is independent of i so

$$\sum_i P_i e^{\lambda_n \rho_{ij}} = c_n, \quad \text{independent of } j. \tag{16}$$

By (11), $\lambda_n \to -\infty$ as $n \to \infty$, so letting $n \to \infty$ yields $P_j = \lim_n c_n$ for all j, so P is the uniform distribution (recall our assumption that $\rho_{ij} = 0$ if and only if $i = j$). Moreover, from (16) it follows that

$$\sum_i e^{\lambda_n \rho_{ij}} = k c_n, \quad \text{independent of } j. \tag{17}$$

To show that ρ is a permutation, fix two arbitrary indices $j \neq j'$ and reorder the vectors $(\rho_{1j}, \ldots, \rho_{kj})$ and $(\rho_{1j'}, \ldots, \rho_{kj'})$ so that their elements are nondecreasing. Let $(\sigma_1, \ldots, \sigma_k)$ and $(\sigma'_1, \ldots, \sigma'_k)$ be the corresponding ordered vectors. Then $\sigma_1 = \sigma'_1 = 0$ and (17) implies that

$$\sum_{i=2}^{k} e^{\lambda_n (\sigma_i - \sigma'_2)} = \sum_{i=2}^{k} e^{\lambda_n (\sigma'_i - \sigma'_2)}.$$

Next we show that if $\sigma_2 \neq \sigma'_2$, say $\sigma_2 > \sigma'_2$, we get a contradiction. Since $\sigma_i - \sigma'_2 > 0$ for all $i \geq 2$, the left-hand-side above tends to 0 as $n \to \infty$, but the right-hand-side is ≥ 1. Therefore $\sigma_2 = \sigma'_2$. Continuing inductively, $\sigma_i = \sigma'_i$ for all i, so $(\rho_{1j}, \ldots, \rho_{kj})$ and $(\rho_{1j'}, \ldots, \rho_{kj'})$ are permutations of one another. Since j and j' were arbitrary, this completes the proof. □

Proof of Corollary 1: As before, let D_n, $n \geq 1$, be a sequence of of distortion values in $(0, D_{\max})$ for which $f(X_1) = 0$ with probability one, and let Q_n and $\lambda_n < 0$ be chosen such that $R(D_n) = (\log e)\Lambda^*_{P,Q_n}(\lambda_n)$. Since $R(D)$ is differentiable on $(0, D_{\max})$ (see [4, Theorem 2.5.1]), from Lemma 2 we get that $\lambda_n = (\log_e 2) R'(D_n)$. Moreover, since we assume that $R(D)$ is strictly convex on $(0, D_{\max})$, the λ_n are all distinct.

If the sequence $\{\lambda_n\}$ is unbounded, i.e., it has a subsequence that tends to $-\infty$, then we can proceed exactly as in the proof of Theorem 1. So assume that the sequence $\{\lambda_n\}$ is bounded. Since for each n, $R(P, Q_n, D_n) = R(D_n) > 0$, there must be a subset S of $\{1, 2, \ldots, k\}$ of size N, say, with $N = |S| \geq 2$, such that infinitely many of the Q_n are supported on $\{a_j : j \in S\}$. Without loss of generality we can relabel the elements of A so that $S = \{1, 2, \ldots, N\}$. If $N = k$ then we can again repeat the argument in the proof of Theorem 1.

Assuming $N \leq k - 1$, we proceed to get a contradiction. Since $f(x) = 0$ with probability one, condition (13) implies that

$$\sum_{j=1}^{k} Q_n(a_j) e^{\lambda_n \rho_{ij}} = \sum_{j=1}^{N} Q_n(a_j) e^{\lambda_n \rho_{ij}} = c_n, \ \text{ for } \textit{all } i.$$

Defining $\rho_{i0} = 0$ for all i, and letting $T(\lambda)$ denote the $(N+1) \times (N+1)$ matrix with entries $\exp(\lambda \rho_{ij})$ for $i = 1, 2, \ldots, N+1$ and $j = 0, 1, \ldots, N$, the above conditions imply that

$$T(\lambda_n) \begin{pmatrix} -c_n \\ Q_n(a_1) \\ \vdots \\ \vdots \\ Q_n(a_N) \end{pmatrix} = \mathbf{0} \in \mathbb{R}^{N+1}.$$

Therefore $\det(T(\lambda_n)) = 0$ for all λ_n. The sequence $\{\lambda_n\}$ is bounded so it must have an accumulation point, and since $\det(T(\lambda))$ is an analytic function of λ it can only have isolated zeroes unless it is identically zero (see, e.g., the discussion in [1, Section 4.3.2]). So here we must have that $\det(T(\lambda)) \equiv 0$ for all $\lambda \leq 0$. But as $\lambda \to -\infty$, $T(\lambda)$ converges to the matrix

$$T_\infty = \begin{pmatrix} 1 & & & \\ 1 & & \mathbf{I}_N & \\ \vdots & & & \\ 1 & 0 & \cdots & 0 \end{pmatrix}$$

which has determinant equal to 1 or -1 ($\mathbf{I}_N$ denotes the $N \times N$ identity matrix), and this provides the desired contradiction. □

C. Proofs in the Continuous Case

Proof of Theorem 2: We argue by contradiction. Suppose $f(X_1) = 0$ with probability one for some $D \in (0, D_{\max})$. Choose a Q^* and a $\lambda^* < 0$ as in (10). Then (13) implies that

$$\sum_{j=1}^{k} Q^*(a_j) e^{\lambda^* r_j(x)} = \text{Constant}, \quad \text{for } P\text{-almost all } x,$$

but since P has an absolutely continuous component with positive density on I, and since the functions $r_j(\cdot)$ are assumed to be continuous, this holds for all $x \in I$, and therefore contradicts the linear independence assumption of Theorem 2. □

Proof of Theorem 3: First we observe that condition (a) immediately implies condition (b). Therefore it suffices to show that if condition (b) holds, $f(X_1)$ cannot be equal to zero with probability one for distortion levels $D > 0$ arbitrarily close to zero. We proceed as in the proof of Corollary 1. Assuming that there is a sequence D_n, $n \geq 1$, of distortion values in $(0, D_{\max})$ for which $f(X_1) = 0$ with probability one, and such that $D_n \downarrow 0$, we will derive a contradiction. Pick Q_n and $\lambda_n < 0$ such that $R(D_n) = (\log e)\Lambda^*_{P,Q_n}(\lambda_n)$. By (13),

$$\sum_{j=1}^{k} Q_n(a_j) e^{\lambda_n r_j(x)} = c_n, \quad \text{for } P\text{-almost all } x \in I. \tag{18}$$

Since P has an absolutely continuous component with positive density on I, and since the functions $r_j(\cdot)$ are assumed to be continuous, (18) holds for all $x \in I$. In particular, for the points $x_0, \ldots, x_k$ in condition (b), (18) becomes

$$\widetilde{T}(\lambda_n)\,(-c_n, Q_n(a_1), \ldots, Q_n(a_k))' = \mathbf{0} \in \mathbb{R}^{k+1},$$

where $\widetilde{T}(\lambda)$ is the $(k+1) \times (k+1)$ matrix with entries $\exp(\lambda r_j(x_i))$, $0 \leq i, j \leq k$, and v' denotes the transpose of a vector v. Therefore, since the entries of the vector $(Q_n(a_1), \ldots, Q_n(a_k))$ sum to 1, it follows that $\det(\widetilde{T}(\lambda_n)) = 0$ for all n, or, equivalently,

$$\begin{aligned} \det(\widetilde{T}(\lambda_n)) &= \sum_{\pi} (-1)^{\mathrm{sign}(\pi)} e^{\lambda_n \sum_{j=0}^{k} r_j(x_{\pi(j)})} \\ &= \sum_{\pi} (-1)^{\mathrm{sign}(\pi)} e^{\lambda_n s_\pi} = 0, \end{aligned} \tag{19}$$

where the sums are taken over all permutations π of the set $\{0, 1, \ldots, k\}$, and the constants s_π are given by $\sum_{j=0}^{k} r_j(x_{\pi(j)})$. Therefore, for any real number $s \geq 0$, we must have that

$$\sum_{\pi\,:\, s_\pi = s} (-1)^{\mathrm{sign}(\pi)} = 0. \tag{20}$$

To see this, let $\{s(1), s(2), \ldots\}$ be the (finite) increasing sequence of all possible values for the constants s_π. Then (19) implies that

$$\sum_{\pi\,:\, s_\pi = s(1)} (-1)^{\mathrm{sign}(\pi)} e^{\lambda_n s(1)} + \sum_{\pi\,:\, s_\pi > s(1)} (-1)^{\mathrm{sign}(\pi)} e^{\lambda_n s_\pi} = 0.$$

By (11), $\lambda_n \to -\infty$ as $n \to \infty$, so multiplying both sides by $e^{-\lambda_n s(1)}$ and letting $n \to \infty$ yields (20) with $s = s(1)$. Continuing this way with $s(2)$, then $s(3)$ and so on, proves (20) for all s.

But now notice that condition (b) implies that, if π^* denotes the identity permutation, then $s_\pi \neq s_{\pi^*}$ for all other permutations π. Therefore, taking $s = s_{\pi^*}$ in (20) we get the desired contradiction. □

Proof of Corollary 2: Let D_n, $n \geq 1$, be a sequence of distortion values in $(0, D_{\max})$ for which $f(X_1) = 0$ with probability one, and pick Q_n and $\lambda_n < 0$ as in the proof of Theorem 3. If the sequence $\{\lambda_n\}$ is unbounded, we can repeat the exact same proof as for Theorem 3. So assume that $\{\lambda_n\}$ is bounded. Since we also assume that $R(D)$ is differentiable and strictly convex, it follows from Lemma 2 that the $\lambda_n = (\log_e 2)R'(D_n)$ are all distinct. Proceeding as in the proof of Theorem 3, we get that $\det(\widetilde{T}(\lambda)) = 0$ for all $\lambda = \lambda_n$. The sequence $\{\lambda_n\}$ is bounded so it must have an accumulation point, and $\det(\widetilde{T}(\lambda))$ is an analytic function of λ. Therefore, arguing as in the proof of Corollary 1, $\det(T(\lambda)) \equiv 0$ for all $\lambda \leq 0$. So we can find a sequence $\lambda'_m \to -\infty$ for which $\det(\widetilde{T}(\lambda'_m)) = 0$. With λ'_m in place of λ_n, the argument proceeds exactly as in the proof of Theorem 3. □

Acknowledgment

We wish to thank the anonymous Referees for their comments, which helped improve the presentation of our paper.

Appendix

Proof of (11): Suppose (11) is false. Then it is possible to pick a constant $K < \infty$ and a sequence of $D_n \in (0, D_{\max})$ with corresponding $\lambda^*_n < 0$, such that $D_n \to 0$ as $n \to \infty$ but $\lambda^*_n \geq -K$ for all n. Let Q^*_n achieve (9) with $D = D_n$, so that

$$\Lambda'_{P,Q^*_n}(\lambda^*_n) = D_n. \tag{21}$$

For each n, recalling that $\rho(x, y) \leq M$ for all x, y,

$$\begin{aligned} \Lambda'_{P,Q_n}(\lambda^*_n) &= E_P\left[\frac{E_{Q_n}\left(\rho(X,Y)e^{\lambda^*_n \rho(X,Y)}\right)}{E_{Q_n}\left(e^{\lambda^*_n \rho(X,Y)}\right)}\right] \\ &\geq E_P\left[E_{Q_n}\left(\rho(X,Y)e^{\lambda^*_n \rho(X,Y)}\right)\right] \\ &\geq E_{Q_n}\left[E_P\left(\rho(X,Y)e^{-KM}\right)\right] \\ &\geq e^{-KM} D_{\max}, \end{aligned}$$

which is bounded away from zero. Since the $D_n \downarrow 0$, this contradicts (21). □

Proof of Lemma 2: Part (i) immediately follows from the minimax representation in [10, Lemma 2]. For (ii) note that, since $\Lambda_{P,Q}(\lambda)$ is continuous and convex in λ (Lemma 1), $\Gamma(\lambda)$ is lower semicontinuous and convex. Then by convex duality (see, e.g., Lemma 4.5.8 in [7]), it follows that $\Gamma(\lambda) = \sup_{x \geq 0}[\lambda x - (\log_e 2)R(x)]$. For $D \in (0, D_{\max})$

and λ^* as in (10), we have

$$\Gamma(\lambda^*) = \lambda^* D - (\log_e 2) R(D) = \sup_{x \geq 0} [\lambda^* x - (\log_e 2) R(x)].$$

But since $R(\cdot)$ is convex and (by assumption) differentiable at D, it must be that the derivative of $[\lambda^* x - (\log_e 2) R(x)]$ vanishes at $x = D$, i.e., $\lambda^* = (\log_e 2) R'(D)$. □

Proof of Lemma 3: First suppose that for some $\lambda^* < 0$, (a), (b) and (c) all hold. For $i = 1, \ldots, k$, let

$$B_i = \frac{P_i}{\sum_j Q^*(a_j) e^{\lambda^* \rho_{ij}}}.$$

Then (b) and (c) imply that equations (3.19) and (3.20) in [6, p. 145] are satisfied with $\delta = -\lambda^*$, so by [6, Theorem 3.7] equation (3.18) is satisfied by W^*. This, together with Lemma 3.1 in [6, Chapter 2] implies that

$$R(D) = H(W \| P \times W_Y^*)$$

where W_Y^* is the second marginal of W^*. But $W_Y^* = Q^*$, so $R(D) = E_P[H(W^*(\cdot|X) \| Q^*(\cdot))]$, and by the definition of W^* and Proposition 1, $E_P[H(W^*(\cdot|X) \| Q^*(\cdot))] = R(P, Q^*, D)$.

Conversely, suppose Q^* achieves the infimum in (9). Then by Lemma 1 there is a (unique) $\lambda^* < 0$ such that (a) holds, and letting W^* be defined as in (b) we also have

$$\begin{aligned} R(D) &\stackrel{(a)}{=} R(P, Q^*, D) \\ &\stackrel{(b)}{=} H(W^* \| P \times Q^*) \\ &\stackrel{(c)}{=} H(W^* \| P \times W_Y^*) + H(W_Y^* \| Q^*) \\ &\stackrel{(d)}{\geq} H(W^* \| P \times W_Y^*) \\ &\stackrel{(e)}{\geq} R(D) \end{aligned}$$

where (a) follows by assumption; (b) from (10), Proposition 1 and the definition of W^*; (c) by the chain rule for relative entropy (see [5, Theorem 2.5.3]); (d) is because relative entropy is nonnegative; and (e) follows from the definition of $R(D)$ in (8). Therefore $H(W_Y^* \| Q^*) = 0$, implying (b). Finally note that the above argument shows that W^* achieves $R(D)$. Then by Theorem 3.7 in [6, p. 145] W^* satisfies equation (3.18) of [6, p. 145] with $\delta = -\lambda^*$, and by the uniqueness of the constants B_i and equation (3.19) of [6, p. 145] we get (c). □

References

[1] L.V. Ahlfors, *Complex Analysis*, McGraw-Hill, New York, 1953.

[2] P.H. Algoet, *Log-Optimal Investment*, Ph.D. thesis, Dept. of Electrical Engineering, Stanford University, 1985.

[3] A.R. Barron, *Logically Smooth Density Estimation*, Ph.D. thesis, Dept. of Electrical Engineering, Stanford University, 1985.

[4] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall Inc., Englewood Cliffs, NJ, 1971.

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

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

[7] A. Dembo and O. Zeitouni, *Large Deviations Techniques And Applications. Second Edition*, Springer-Verlag, New York, 1998.

[8] J.C. Kieffer, "Sample converses in source coding theory," *IEEE Trans. Inform. Theory*, vol. 37, no. 2, pp. 263–268, 1991.

[9] I. Kontoyiannis, "Second-order noiseless source coding theorems," *IEEE Trans. Inform. Theory*, vol. 43, no. 4, pp. 1339–1341, July 1997.

[10] I. Kontoyiannis, "Pointwise redundancy in lossy data compression and universal lossy data compression," *IEEE Trans. Inform. Theory*, vol. 46, no. 1, pp. 136–152, January 2000.

[11] C.E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec.*, vol. part 4, pp. 142–163, 1959, Reprinted in D. Slepian (ed.), *Key Papers in the Development of Information Theory*, IEEE Press, 1974.

Ioannis Kontoyiannis was born in Athens, Greece, in 1972. He received the B.Sc. degree in mathematics in 1992 from Imperial College (University of London), and in 1993 he obtained a distinction in Part III of the Cambridge University Pure Mathematics Tripos. In 1997 he received the M.S. degree in statistics, and in 1998 the Ph.D. degree in in electrical engineering, both from Stanford University. Between June and December 1995 he worked at IBM Research, on a satellite image processing and compression project, funded by NASA and IBM. He has been with the Department of Statistics at Purdue University (and also, by courtesy, with the Department of Mathematics, and the School of Electrical and Computer Engineering) since 1998. During the 2000-01 academic year he is visiting the Applied Mathematics Division of Brown University. His research interests include data compression, applied probability, statistical genetics, nonparametric statistics, entropy theory of stationary processes and random fields, and ergodic theory.

Amir Dembo received the B.Sc. (Summa Cum Laude), and D.Sc. degrees in electrical engineering from the Technion-Israel Institute of Technology, Haifa in 1980, 1986 respectively. During 1980-1985 he was a Senior Research Engineer with the Israel Defense Forces. In 1986/1987 he visited AT&T Bell Laboratories and the Applied Mathematics Division of Brown University. Since 1988 he is at Stanford University, where he is presently an associate professor jointly in the Mathematics and Statistics departments. During 1994-1996 he was a professor of Electrical Engineering at the Technion, Israel. He authored or co-authored more than 70 technical publications, including the book Large Deviations Techniques and Applications (Second Edition, Springer-Verlag, 1998, with O. Zeitouni). From 1994-2000 he was an associate editor for the Annals of Probability, and is currently serving on the editorial boards of the Annals of Applied Probability and of the Electronic Journal of Probability. He has worked in a number of areas including information and communication theory, signal processing and estimation theory. His current research interests are in probability theory and its applications to statistical physics, to queueing and communication systems, and to biomolecular data analysis.