NAMT
95-013

# A Relation Between Shannon-Weaver Entropy and "Theoretical Dimension" for Classes of Smooth Functions

Aurelija Trgo
Carnegie Mellon University

Mladen V. Wickerhauser
Washington University

Research Report No. 95-NA-013

July 1995

Sponsors

U.S. Army Research Office
Research Triangle Park
NC 27709

National Science Foundation
1800 G Street, N.W.
Washington, DC 20550

# A Relation Between Shannon–Weaver Entropy and "Theoretical Dimension" for Classes of Smooth Functions*

AURELIJA TRGO†
Center for Nonlinear Analysis and Department of Mathematics
Carnegie Mellon University, Pittsburgh, PA

MLADEN V. WICKERHAUSER
Department of Mathematics
Washington University in St. Louis, MO

May 1995

**Abstract**

Suppose that an infinite sequence is produced by independent trials of a random variable with a fixed distribution. The Shannon–Weaver entropy of the sequence determines the minimum bit rate needed to transmit the values of the sequence. We show that if the source distribution is highly concentrated, as is commonly observed in practice, then its entropy is equal to the logarithm of the theoretical dimension of the sequence. We conclude that the best-basis algorithm, which minimizes this theoretical dimension over a library of transformations, both chooses the transformation that yields best compression and also gives an estimate of the compression rate.

## 1 Model

We need to define some basic objects. First, suppose that $\rho = \rho(t)$ is a *probability density function*, *i.e.*, a real valued, nonnegative, integrable function defined on $[0,1]$ which satisfies $\int_0^1 \rho(t)\,dt = 1$. For each measurable subset $E \subset [0,1]$ we define the associated *probability measure* by

$$P\{E\} \stackrel{\text{def}}{=} \int_E \rho(t)\,dt. \tag{1}$$

*Research supported by NSF, AFOSR, and the Southwestern Bell Telephone Company
†Work partially supported by the Army Research Office through the Center for Nonlinear Analysis

For technical reasons, we will assume that the density function $\rho$ is continuous and strictly positive on $(0,1)$.

Secondly, fix $1 \ll N < \infty$ and define *(uniform) quantization to $N$ values* by the formula

$$Q_N(x) \stackrel{\text{def}}{=} \lfloor Nx \rfloor / N. \tag{2}$$

If $x \in [0,1)$ then $Q_N(x) \in \{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}\}$.

Now suppose that $\{x_m : m = 1,2,\ldots\}$ is a sequence of independent trials of a random variable $x \in [0,1)$ whose density function is $\rho$. If the sequence is replaced by a quantized version of itself, namely $\{Q_N(x_m) : m = 1,2,\ldots\}$, then the root-mean-square error or *distortion* per sequence element will have the following expected value:

$$D_N \stackrel{\text{def}}{=} \left(E\{|x_m - Q_N(x_m)|^2\}\right)^{1/2} = \left(\int_0^1 |t - Q_N(t)|^2 \rho(t)\, dt\right)^{1/2}. \tag{3}$$

Since the terms in the sequence are independent and identically distributed random variables, the distortion is independent of $m$. The sequence of quantized values thus produced will have the following probability distribution function:

$$p_n \stackrel{\text{def}}{=} P\left\{Q_N(x_m) = \frac{n-1}{N}\right\} = \int_{\frac{n-1}{N}}^{\frac{n}{N}} \rho(t)\, dt; \qquad n = 1,2,\ldots,N. \tag{4}$$

Again, each $p_n$ is independent of $m$. Shannon's theorem [2] states that the expected number of bits per element required to encode this quantized sequence cannot be less than the entropy of the distribution, defined below:

$$H_N \stackrel{\text{def}}{=} -\sum_{n=1}^{N} p_n \log p_n \tag{5}$$

As before, $H_N$ is independent of $m$.

We obtain a rate-distortion curve for the sequence by plotting $10 \log D_N$ against $H_N$. We use $10 \log D_N$ so that the distortion units are decibels relative to a unit signal amplitude. The number of quantization intervals $N$ parameterizes the curve. It remains to estimate $H_N$ and $D_N$ from $\rho$.

Since we are assuming that $\rho$ is continuous, we may use the mean value theorem to estimate $p_n = \frac{1}{N}\rho(\xi_n)$, where $\xi_n \in (\frac{n-1}{N}, \frac{n}{N})$. Therefore,

$$H_N = -\sum_{n=1}^{N} \frac{1}{N}\rho(\xi_n) \log\left[\frac{1}{N}\rho(\xi_n)\right] = \log N - \sum_{n=1}^{N} \frac{1}{N}\rho(\xi_n) \log \rho(\xi_n) \tag{6}$$

The second term is evidently a Riemann sum approximating $-\int_0^1 \rho(t)\log\rho(t)\,dt$, which we may call the *source entropy* $\mathcal{H}(\rho)$. The $\log N$ term is present because at super fine quantizations the less significant digits contain most of the information even though they have almost no connection with $\rho$.

Likewise, we can estimate

$$D_N^2 = \sum_{n=1}^{N} \int_{\frac{n-1}{N}}^{\frac{n}{N}} \left| t - \frac{n-1}{N} \right|^2 \rho(t)\,dt \le \frac{1}{N^2} \sum_{n=1}^{N} p_n = \frac{1}{N^2}. \tag{7}$$

Hence $10\log D_N \le -10\log N$. Unfortunately, no lower bound exists for $D_N$, since even a continuous density $\rho$ can be arbitratrily concentrated at the values $0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}$. However, if $\rho$ is continuous then we can compute the asymptotic behavior of $D_N$ as $N \to \infty$:

$$\lim_{n\to\infty} N^2 D_N^2 = N^3 \int_0^{1/N} t^2\,dt = \frac{1}{3} \quad \Rightarrow \quad D_n \sim \frac{1}{\sqrt{3}\,N}. \tag{8}$$

Combining Equations 6 and 8 shows that

$$10\log D_n \sim -10\log N - 5\log 3 \sim -10H_N + 10\mathcal{H}(\rho) - 5\log 3, \qquad \text{as } N \to \infty. \tag{9}$$

Thus the rate-distortion curve is asymptotic to a line of slope $-10$ with an intercept at $10\mathcal{H}(\rho) - 5\log 3$. Shifting the curve to the left improves the rate-distortion relationship in the sense that the same transmission quality is obtained at a lower bit rate. Such a shift is accomplished by reducing $\mathcal{H}(\rho)$, or equivalently by transforming the sequence $\{x_m\}$ so that it appears to come from a lower-entropy source.

# 2 Relations

Fix $1 \ll M < \infty$ and suppose that $\{x_1, \ldots, x_M\}$ is a sequence of $M$ Bernoulli trials of the random variable with density $\rho$ defined in Equation 1 above. Let $\{x_1^*, \ldots, x_M^*\}$ be the decreasing rearrangement of the sequence $\{x_m\}$. That is,

$$x_0^* \stackrel{\text{def}}{=} 1 \ge x_1^* \ge x_2^* \ge \cdots x_M^* \ge 0 \stackrel{\text{def}}{=} x_{M+1}^*.$$

This decreasing rearrangement is uniquely defined, and it determines a decreasing step function $x^* = x^*(t)$ on the interval $[0,1]$ as follows:

$$x^*(t) = x_m^*, \quad \text{if } \tfrac{m-1}{M} < t \le \tfrac{m}{M}; \qquad x^*(0) = 1. \tag{10}$$

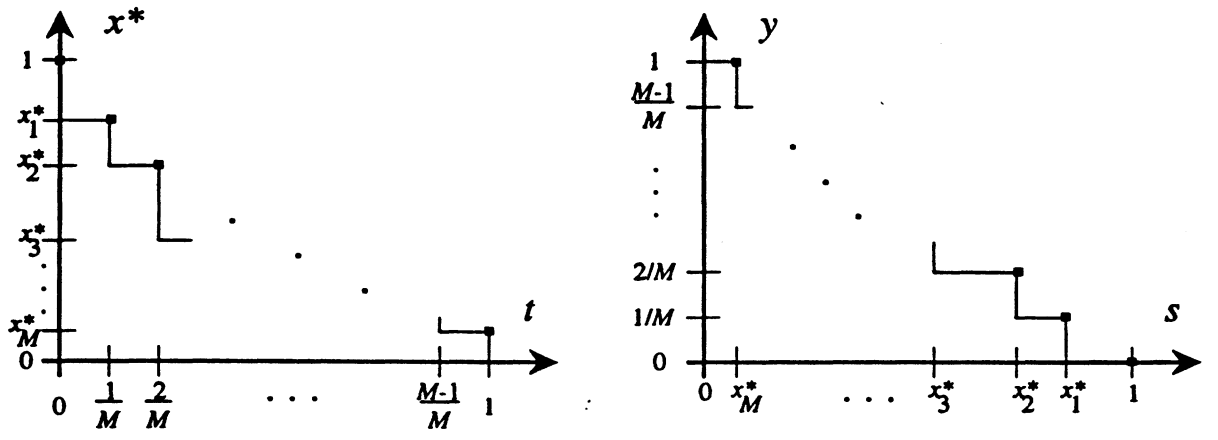


Figure 1: Two step functions determined by $\{x_1, x_2, \ldots x_M\}$.

The same sequence determines another step function as follows:

$$y(s) = \frac{m}{M}, \quad \text{if } x^*_{m+1} < s \leq x^*_m; \qquad y(1) = 0. \tag{11}$$

Examples of these two step functions are plotted in Figure 1.

The step functions $y$ and $x^*$ are approximate inverses, in the sense that

$$y\left(x^*(t)\right) = Q_M(t); \qquad x^*\left(y(s)\right) = \max\{x^*_m : x^*_m \leq s\} \stackrel{\text{def}}{=} Q_x(s). \tag{12}$$

Thus $y$ inverts $x^*$ up to the precision of the $M$-bin uniform quantization, while $x^*$ inverts $y$ up to the precision of the generally nonuniform quantization defined by the monotonic sequence $\{x^*_m\}$.

Now $y(s)$ is just $1/M$ times the number of values of $m \in \{1, 2, \ldots, M\}$ for which $x_m \geq s$, so we can compute its expectation in terms of the density $\rho$:

$$Ey(s) = \sum_{k=1}^{M} \frac{k}{M} P\{k \text{ values of } \{x_m\} \text{ lie in } [s,1]\} = \sum_{k=1}^{M} \frac{k}{M} \binom{M}{k} (1-q_s)^{M-k} q_s^k = q_s, \tag{13}$$

where $q_s = \int_s^1 \rho(t)\, dt$, and we have used the identity $\frac{k}{M}\binom{M}{k} = \binom{M-1}{k-1}$ to collapse the sum. Note that the expectation is independent of $M$.

The assumption that $\rho$ is continuous and positive implies that $Ey(s) = \int_s^1 \rho(t)\, dt$ is decreasing and continuously differentiable. Thus

$$\frac{d}{ds} Ey(s) = -\rho(s) < 0, \tag{14}$$

and $Ey$ has a continuously differentiable inverse function which we may call $z = z(t)$:

$$z(Ey(s)) = s; \qquad Ey(z(t)) = t; \qquad \frac{d}{ds}Ey(s) = \frac{1}{z'(Ey(s))}. \tag{15}$$

Combining Equations 14 and 15 allows us to compute the source entropy in terms of $z'$:

$$\mathcal{H}(\rho) = -\int_0^1 \rho(s)\log\rho(s)\,ds = \int_0^1 \left[\frac{1}{z'(Ey(s))}\right]\log\left[\frac{-1}{z'(Ey(s))}\right]ds = \int_0^1 \log[-z'(t)]\,dt. \tag{16}$$

In the last step, we substituted $s \leftarrow z(t)$ and then simplified.

It remains to relate $z$ with $x^*$. The idea is that $y$ is the "inverse" of $x^*$, while $z$ is the inverse of $Ey$. We claim that $z \approx x^*$ and

$$\int_0^1 \log[-z'(t)]\,dt \approx \sum_{m=1}^{M} \log[-\Delta x_m^*], \tag{17}$$

where $\Delta x_m^* \stackrel{\text{def}}{=} x_m^* - x_{m-1}^*$ for $m = 1, 2, \ldots, M$ is the difference between successive values in the decreasing rearrangement.

Finally, suppose that the values in the sequence $\{x_1, \ldots, x_m\}$ are concentrated near 0 in such a way that the decreasing rearrangement decreases exponentially or by some power law. Namely, suppose we choose constants $0 < A \le B$ and $0 < a \le b$ such that for all $m = 1, 2, \ldots, M$, we have

$$A\,(x_m^*)^a \le -\Delta x_m^* \le B\,(x_m^*)^b. \tag{18}$$

Then we can estimate

$$M\log A + a\sum_{m=1}^{M}\log[x_m^*] \le \mathcal{H}(\rho) \le M\log B + b\sum_{m=1}^{M}\log[x_m^*]. \tag{19}$$

But since the two sums are independent of the order of summation, we can dispense with the decreasing rearrangement and write the estimate as follows:

$$M\log A + a\sum_{m=1}^{M}\log[x_m] \le \mathcal{H}(\rho) \le M\log B + b\sum_{m=1}^{M}\log[x_m]. \tag{20}$$

# 3 Theoretical Dimension

Although $I(x) \stackrel{\text{def}}{=} \sum_{m=1}^{M}\log x_m$ is not an additive information cost function in the sense of [1], it can be replaced by any of the expressions below:

- $\sum_{m=1}^{M} \log(1 + x_m/\epsilon)$: Regard $\epsilon > 0$ as a roundoff error.
- $\left(\sum_{m=1}^{M} |x_m|^\epsilon\right)^{1/\epsilon}$: With $0 < \epsilon \ll 1$ this approximates the $L^0$ or counting norm, which in turn is an approximation for $I(x)$.
- $-\sum_{m=1}^{M} |x_m|^2 \log |x_m|^2$: This is the *entropy* functional discussed in [1]. It is the linear approximation to the $L^0$ norm of a sequence $\{x_m\}$ with unit $L^2$ norm, using the derivative of $L^p$ norm with respect to $p$ to obtain the differential.

The last of these is monotonic with the *theoretical dimension* $d(x)$, which is defined in Reference [1] as follows:

$$d(x) \stackrel{\text{def}}{=} \exp\left\{-\sum_{m=1}^{M} \frac{|x_m|^2}{\|x\|^2} \log \frac{|x_m|^2}{\|x\|^2}\right\}. \tag{21}$$

The idea is that $I(x)$ is minimized, whenever any one of these expressions is minimized. Now suppose that we have a particular sequence and a library of transforms containing some in which the transformed sequence has the "rapid decrease" property of Equation 18. Then choosing that transform which minimizes any one of these information cost functions produces a coefficient sequence which appears to come from the lowest-entropy source. In particular, if $\{x_m\}$ are samples of a smooth oscillatory function, and $\mathcal{B} \subset \mathbf{O}(M)$ is a family of smooth orthogonal wavelet packet transformations of $\mathbf{R}^M$, and $Bx$ denotes the coefficient sequence produced by applying $B \in \mathcal{B}$ to $\{x_m\}$, then $I(Bx)$ and $d(Bx)$ will have the same minimum $B^* \in \mathcal{B}$, and if each $Bx$ is regarded as Bernoulli trials from an unknown source density, $B^*x$ will look like it comes from the lowest-entropy source.

# 4 Example

We consider a simple family of source densities which produce sequences with the "rapid decrease" porperty.

Suppose that the source distribution is $\rho(t) = (\alpha + 1)t^\alpha$, where $-1 < \alpha < 0$ to insure that $\rho$ is integrable and concentrated near $t = 0$. The coefficient is chosen to insure that $\int \rho = 1$. Then $Ey(s) = \int_s^1 \rho(t)\,dt = 1 - s^{\alpha+1}$, so

$$z(t) = (1-t)^{1/(\alpha+1)} \quad \Rightarrow \quad z'(t) = \frac{-1}{\alpha+1}(1-t)^{-\alpha/(\alpha+1)} = \frac{-1}{\alpha+1} z(t)^{-\alpha}. \tag{22}$$

The relation between $z'$ and $z$ implies that

$$\mathcal{H}(\rho) = -\alpha \int_0^1 \log[z(t)]\ dt - \log(\alpha + 1). \tag{23}$$

# 5 Acknowledgment

The second author (MVW) wishes to thank David Donoho for pointing out the relationship between $L^0$ norm and the entropy information cost.

# References

[1] Ronald R. Coifman and Mladen Victor Wickerhauser. Entropy based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 32:712–718, March 1992.

[2] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication.* The University of Illinois Press, Urbana, 1964.