On the Minimum Entropy of a Mixture of Unimodal and Symmetric Distributions

T.-L. Chen
Institute of Statistical Science
Academia Sinica
Taipei, Taiwan
S. Geman
Division of Applied Mathematics
Brown University
Providence, RI

***Abstract*—Progressive encoding of a signal generally involves an estimation step, designed to reduce the entropy of the residual of an observation over the entropy of the observation itself. Oftentimes the conditional distributions of an observation, given already-encoded observations, are well fit within a class of symmetric and unimodal distributions (e.g. the two-sided geometric distributions in images of natural scenes, or symmetric Paretian distributions in models of financial data). It is common practice to choose an estimator that centers, or aligns, the modes of the conditional distributions, since it is common sense that this will minimize the entropy, and hence the coding cost of the residuals. But with the exception of a special case, there has been no rigorous proof. Here we prove that the entropy of an arbitrary mixture of symmetric and unimodal distributions is minimized by aligning the modes. The result generalizes to unimodal and rotation-invariant distributions in R^n. We illustrate the result through some experiments with natural images.**

***Index Terms*—Entropy coding, LOCO, lossless image compression, mixture distributions, predictive coding, symmetric distributions, unimodal distributions.**

I. Introduction

It is generally assumed that the entropy of a mixture of symmetric unimodal densities is minimized by aligning the modes of the component densities. This comes up in various compression applications in which a multi-variate signal is progressively encoded, one variable at a time, conditioned upon the "past" or "context," as represented by the already encoded variables ("predictive encoding"). Oftentimes these conditional distributions are unimodal and symmetric, as is the case for images encoded in raster order (cf. [1], [2], [3], [4]), or in models of price fluctuations of securities wherein the conditional returns are well approximated as unimodal and symmetric (cf. [5], [6], [7]). Consider two random variables: $X \in R$ which is to be encoded, and $Y \in R^m$, a function of the past, meaning the already-encoded variables. Assume that X has a conditional density $p(x|y)$, given any $Y = y$, which is symmetric and unimodal. Given a good predictor of X based on Y, call it $g(Y)$, an expedient and much-used approach to coding $X = x$, given $Y = y$, is to code $x - g(y)$ under the distribution of $X - g(Y)$. The alternative, which is optimal, is to code x under the conditional probability given y, but this can be impractical as it involves a knowledge of the conditional distribution for every value of y. (Of course the range of Y can be partitioned into a relatively small number of more-or-less homogeneous categories – often referred to as "contexts," for better results, but this only moves the prediction problem discussed here to the equivalent problem for each category. See, for example, our experiments in §III, or the popular and highly efficient lossless compression standard known as LOCO – [8], [9].)

Which predictor yields the minimum average coding cost? Since $X - g(Y)$ has density $\int_y p(x - g(y)|y)dF(y)$, the problem of choosing an optimal g is the problem of shifting the components of a mixture of symmetric unimodal densities so as to minimize entropy. It is easy to believe that $g(y) = \text{median}\, p(\cdot|y)$ is the best choice (cf. [10], [11], and [12]), but we were unable to come up with an easy proof. Janžura et al. [13] have a nice proof for the case of *finite* mixtures (i.e. Y takes one of a finite number of possible values), but even that is not elementary. In any case, we provide here a proof that imposes no further restriction on $p(x|y)$ and no restriction at all on Y. Possibly, the method of proof, which employs a "function rearrangement" (cf. [14]) to reduce the problem to one of comparing entropies of *monotone decreasing densities on R^+*, even when x itself is multivariate, may be of some independent interest for other, related, entropy type problems, such as the analysis of the Minimum Entropy Error Principle for estimation ([13], [15], [16], [17]).

Concerning the case when x is multivariate ($x \in R^n$), it is tempting to conjecture that the same result holds: $g(y) = \text{median}\, p(\cdot|y)$ minimizes the entropy of $X - g(Y)$, provided that, for every y, $p(x|y)$ is even with respect to its mode (i.e. $p(\alpha + x|y) = p(\alpha - x|y)$, where $\alpha = \text{median}\, p(\cdot|y) \in R^n$). But this is wrong, as demonstrated by Otáhal [16], who constructs a mixture of three indicator functions of rectangles (so $n = 2$), rotated with respect to each other, which achieves its minimum entropy when the medians are not aligned. Otáhal proves that finite mixtures of unimodal *isotropic* (rotation invariant) densities, on the other hand, do achieve minimum entropy when the medians coincide. As we shall see, the result also

Supported in part by the National Science Foundation under grants ITR-0427223 and IIS-0423031, and the Office of Naval Research under grant N000140610749.

holds for arbitrary mixtures.

Section II contains the statement of our result and its proof. Section III illustrates the result with some empirical experiments on real images.

II. Theorem: Align the Modes

It is hard to avoid the possibility of infinite (differential) entropies if we want to treat arbitrary arrangements of the modes, through arbitrary $g(y)$. But the theorem can still be stated in full generality if we agree on the following extension of H:

Definition 1: Given a density function $f(x)$, $x \in R^n$, we say that the entropy, $H(f)$ "exists in the extended sense" if either

i. $\int_{R^n} 1_{f\leq 1} f \log(1/f) < +\infty$ and $\int_{R^n} 1_{f>1} f\log(1/f) = -\infty$, in which case we write $H(f) = -\infty$; or

ii. $\int_{R^n} 1_{f\leq 1} f \log(1/f) = +\infty$ and $\int_{R^n} 1_{f>1} f\log(1/f) > -\infty$, in which case we write $H(f) = +\infty$; or

iii. $\int_{R^n} 1_{f\leq 1} f \log(1/f) < +\infty$ and $\int_{R^n} 1_{f>1} f\log(1/f) > -\infty$, in which case $H(f) \in (-\infty, \infty)$.

Definition 2: Given random variables $X \in R^n$, $Y \in R^m$ and given, for every value $Y = y$, a conditional density $p(x|y)$, we say that $p(x|y)$ is CSUM (conditionally symmetric and unimodal) if $p(x|y)$ is symmetric (rotation invariant, when $n > 1$) and unimodal in x for every y.

Theorem 1: Assume that $p(x|y)$ is CSUM, and let $\mu(y) = \text{median}\, p(\cdot|y)$. If $H(X - \mu(Y))$ exists (in the extended sense) then $H(X - \mu(Y)) \leq H(X - g(Y))$ for all $g : R^m \to R$ for which $H(X - g(Y))$ also exists (in the extended sense).

Remark. Whether or not $H(X-\mu(Y))$ or $H(X-g(Y))$ exists, both $X-\mu(Y)$ and $X-g(Y)$ have absolutely continuous distributions:

$$p^\mu(x) = \int_{R^m} p(x + \mu(y)|y)dF(y)$$

is the density of $X - \mu(Y)$, and

$$p^g(x) = \int_{R^m} p(x + g(y)|y)dF(y)$$

is the density of $X - g(Y)$.

In an effort to make the proof more transparent, we provide here, first, a kind of road map of the development:

Outline of Proof. The task is to show $H(p^g) \geq H(p^\mu)$.

i. Start with a special case: $p(x|y)$ is continuous in x for every y, and uniformly bounded in x and y.

ii. Replace p^g and p^μ by uni-variate and non-increasing functions m^g and m^μ on $[0,\infty)$ that behave just like p^g and p^μ when it comes to integration:

$$\int_{R^n} G(p^g(x))dx = \int_0^\infty G(m^g(\xi))d\xi$$

and

$$\int_{R^n} G(p^\mu(x))dx = \int_0^\infty G(m^\mu(\xi))d\xi$$

$\forall G : R \to R$, including $G(z) = z\log z$ (and hence $H(p^g) = H(m^g)$ and $H(p^\mu) = H(m^\mu)$). This can always be done through a transformation of the occupation measures ("distributions") of p^g and p^μ, respectively.

iii. Show that by virtue of the alignment of modes,

$$\int_0^{x_o} m^\mu(x)dx \geq \int_0^{x_o} m^g(x)dx \quad \forall x_o \tag{1}$$

iv. Show that (1) implies $H(m^\mu) \leq H(m^g)$ and hence $H(p^\mu) \leq H(p^g)$.

v. Extend to arbitrary (but CSUM) $p(x|y)$ by approximating $p(\cdot|y)$, for each y, with a suitable sequence of continuous and bounded CSUM functions.

Proof of Theorem. The proof is given for the uni-variate ($n = 1$) case, but is essentially identical in the multi-variate ($n > 1$) case. In the few spots where the generalization requires explanation, remarks are made accordingly.

It will be convenient to further extend the definition of entropy to non-negative L^1 functions:

$$H(f) = \int_R f(x)\log\frac{1}{f(x)}dx \in [-\infty, \infty]$$

whether or not $\int f(x) = 1$, provided that either $\int 1_{f\leq 1} f\log(1/f) < +\infty$ or $\int 1_{f>1} f\log(1/f) > -\infty$. In the discrete case, for $f_k > 0$, $k = 1, 2, \ldots$, $\sum_k f_k < \infty$,

$$H(f) = \sum_{k=1}^{\infty} f_k \log\frac{1}{f_k} \in (-\infty, \infty]$$

Mostly, we will work with the negative of H, which we denote by $\tilde{H}$.

Without loss of generality, we can assume that $p(x|y)$ has median at $x = 0$, $\forall y$, since otherwise we could replace $p(x|y)$ by $p(x - \mu(y)|y)$, where $\mu(y) = \text{median}\, p(\cdot|y)$, and work instead with conditional densities centered at $x = 0$.

Now fix $g : R \to R$. Since $X - g(Y)$ has density

$$p^g = \int_{R^m} p(x - g(y)|y)dF(y)$$

our task is to show that $\tilde{H}(p^g) \leq \tilde{H}(p^o)$, where $p^o = \int_{R^m} p(x|y)dF(y)$, provided that $\tilde{H}(p^g)$ exists in the extended sense. Most of the work is in handling the following special case, which we state as a proposition:

Proposition 1: Assume that $f(x|y)$ is:

i. non-negative, continuous, and integrable in x for each $y \in R^m$;

ii. symmetric (rotation invariant for $n > 1$) around $x = 0$ and unimodal for each $y \in R^m$;

iii. uniformly bounded in (x, y).

Then for any $g : R^m \to R$

$$\tilde{H}(f^g) \leq \tilde{H}(f^o) \in [-\infty, \infty)$$

where

$$f^g = \int_{R^m} f(x - g(y)|y)dF(y) \quad \& \quad f^o = \int_{R^m} f(x|y)dF(y)$$

and $F(y)$ is a probability distribution function on R^m.[1]

Remark. Observe that $\int f^g dx = \int f^o dx$ (just change the order of integration), and $\int f^o dx \leq \sup_{(x,y)} f(x|y) < \infty$, but possibly $\int f^o dx \neq 1$.

Proof of Proposition. The main idea of the proof is to "rearrange" f^g and f^o, using their respective occupation measures, into non-negative non-increasing functions on $[0, \infty)$, which are easier to work with and whose entropies are easier to compare.

Lemma 1: Let $h : R \to [0, \infty)$ be bounded, continuous, and integrable, and define $O_h(z)$ (the "occupation measure") by

$$O_h(z) = \lambda\{x : h(x) \geq z\},$$

finite or infinite, for all $z \geq 0$ (where λ is Lebesgue measure).

(a) Define $m^h(x) = \sup\{z : O_h(z) \geq x\}$, $x \in (0, \infty)$, and $m^h(0) = \sup_x h(x)$. Then $m^h(x)$ is continuous and non-increasing on $[0, \infty)$ and $m^h(x) \to 0$ as $x \to \infty$.

(b) For any function $G : [0, \infty) \to R$ with $\int_R |G(h(x))|dx < \infty$,

$$\int_R G(h(x))dx = \int_0^\infty G(m^h(x))dx$$

(c) For any $x_o \in [0, \infty)$

$$\int_0^{x_o} m^h(x)dx = \sup_{A:\lambda(A)=x_o} \int_A h(x)dx$$

The transformation $h \to m^h$ is what Hardy, Littlewood, and Pólya [14] called the "rearrangement" of h. In that we assume more about h, an demand more of m^h, we have included a complete proof of Lemma 1 in Appendix A.

Our task is to prove $\tilde{H}(f^g) \leq \tilde{H}(f^o)$. In light of Lemma 1, we can compare, instead, two monotonic functions. Both $f^g(x)$ and $f^o(x)$ are bounded (by the uniform bound on $f(x|y)$), continuous, and integrable, and hence Lemma 1 can be applied to each function. For notational convenience, use m^g to represent m^{f^g} and m^o to represent m^{f^o}. Since $f^g(x)$ and $f^o(x)$ are bounded

$$-\infty \leq \tilde{H}(f^g), \tilde{H}(f^o) < \infty,$$

and result (b) of Lemma 1 then easily extends to $G(z) = z \log z$ (separate G on $z > 1$ from G on $z \leq 1$), to get

$$\tilde{H}(m^g) = \tilde{H}(f^g) \in [-\infty, \infty) \,\&\, \tilde{H}(m^o) = \tilde{H}(f^o) \in [-\infty, \infty)$$

and the task is now to show $\tilde{H}(m^g) \leq \tilde{H}(m^o)$.

This is most easily accomplished via a discretization of m^g and m^o, justified by the following lemma:

Lemma 2: Let $f : [0, \infty) \to [0, \infty)$ be continuous and non-increasing, with $\int_0^\infty f(x)dx < \infty$. For every $N, k \in \{1, 2, \ldots\}$ define

$$f_k^N = N \int_{\frac{k-1}{N}}^{\frac{k}{N}} f(x)dx$$

Then

$$\frac{1}{N} \tilde{H}(\{f_k^N\}_{k=1}^\infty) \overset{N \to \infty}{\longrightarrow} \tilde{H}(f) \in [-\infty, \infty)$$

The result is standard fare, at least when $f(x) \log f(x)$ is Riemann integrable (cf. [18]). We want to accommodate $\tilde{H}(f) = -\infty$ as well; a proof is included in Appendix A.

In order to show $\tilde{H}(m^g) \leq \tilde{H}(m^o)$ (and hence $\tilde{H}(f^g) \leq \tilde{H}(f^o)$), we use Lemma 2 to generate a discrete approximation of m^g and m^o: let

$$a_k^N = N \int_{\frac{k-1}{N}}^{\frac{k}{N}} m^g(x)dx \quad and \quad b_k^N = N \int_{\frac{k-1}{N}}^{\frac{k}{N}} m^o(x)dx \quad (2)$$

for all $k, N \in \{1, 2, \ldots\}$. Obviously, a_k^N and b_k^N are non-increasing in k. By an application of Lemma 2, $\tilde{H}(m^g) \leq \tilde{H}(m^o)$ can be established by proving

$$-\infty \leq \tilde{H}(\{a_k^N\}_{k=1}^\infty) \leq \tilde{H}(\{b_k^N\}_{k=1}^\infty) \quad (3)$$

for all N. Equation (3) is based on a final lemma, which contains the main idea of the proof of the Theorem, and justifies the use of function rearrangements (Lemma 1):

Lemma 3: Define $\{a_k^N\}_{k=1}^\infty$ and $\{b_k^N\}_{k=1}^\infty$ as in equation (2). Then

(a)

$$\sum_{k=1}^\infty a_k^N = \sum_{k=1}^\infty b_k^N < \infty$$

(b)

$$\sum_{k=1}^m a_k^N \leq \sum_{k=1}^m b_k^N \quad \forall m = 1, 2, \ldots$$

Proof of Lemma 3.
(1) Immediate from the definition and the fact that $\int m^g = \int f^g = \int f^o = \int m^o$.

(2) Follows immediately if we can show

$$\int_0^{x_o} m^g(x)dx \leq \int_0^{x_o} m^o(x)dx$$

for every $x_o \in [0, \infty)$.

Start with Lemma 1(c):

$$\int_0^{x_o} m^g(x)dx = \sup_{A:\lambda(A)=x_o} \int_A f^g(x)dx$$

and

$$\int_0^{x_o} m^o(x)dx = \sup_{A:\lambda(A)=x_o} \int_A f^o(x)dx$$

[1] Later, functions $f(x|y)$ with these properties will be used to approximate, from below, the more general functions $p(x|y)$ of the Theorem. Hence we do *not* assume that $\int f(x|y)dx = 1$ for every, or even almost every, y.

Then observe that

$$\begin{aligned}
&\sup_{A:\lambda(A)=x_o}\int_A f^g(x)dx\\
&= \sup_{A:\lambda(A)=x_o}\int_A\int_{R^m} f(x-g(y)|y)dF(y)dx\\
&= \sup_{A:\lambda(A)=x_o}\int_{R^m}\int_A f(x-g(y)|y)dxdF(y)\\
&\leq \int_{R^m}\int_{-x_o/2}^{x_o/2} f(x|y)dxdF(y)\\
&= \int_{-x_o/2}^{x_o/2}\int_{R^m} f(x|y)dF(y)dx\\
&= \sup_{A:\lambda(A)=x_o}\int_A\int_{R^m} f(x|y)dF(y)dx\\
&= \sup_{A:\lambda(A)=x_o}\int_A f^o(x)dx
\end{aligned}$$

Q.E.D. (Lemma 3)

Remark. In the multi-variate case ($x \in R^n$), $\int_{-x_o/2}^{x_o/2}$ is replaced by an integral over the n-dimensional sphere, centered at the origin.

The remaining task in the proof of Proposition 1 is to apply Lemma 3 to get verification of equation 3. What follows greatly improves on our original argument, which was long-winded and pedestrian. We are indebted to one of the anonymous referees for pointing us in a much more efficient direction:

Fix N and define probabilities on $\{1, 2, \ldots\}$ by

$$\tilde{a}_k = \frac{a_k^N}{\sum_{l=1}^{\infty} a_l^N}, \quad \tilde{b}_k = \frac{b_k^N}{\sum_{l=1}^{\infty} b_l^N}$$

Since $\sum_{l=1}^{\infty} a_l^N = \sum_{l=1}^{\infty} b_l^N$, (3) is equivalent to

$$-\infty \leq \tilde{H}(\{\tilde{a}_k\}_{k=1}^{\infty}) \leq \tilde{H}(\{\tilde{b}_k\}_{k=1}^{\infty})$$

For any non-decreasing sequence $g_k \in [0, \infty]$ (with the convention $0 \cdot \infty = 0$):

$$\begin{aligned}
\infty \quad &\geq \quad \sum_{k=1}^{\infty} \tilde{a}_k g_k\\
&= \quad \sum_{k=1}^{\infty} \tilde{a}_k \int_0^{g_k} dx\\
&= \quad \sum_{k=1}^{\infty} \int_0^{\infty} \tilde{a}_k 1_{x\leq g_k} dx\\
&= \quad \int_0^{\infty} \sum_{k:g_k\geq x} \tilde{a}_k dx\\
\underset{\text{(lemma 3,b)}}{\geq} \quad & \quad \int_0^{\infty} \sum_{k:g_k\geq x} \tilde{b}_k dx\\
= \cdots = \quad & \quad \sum_{k=1}^{\infty} \tilde{b}_k g_k\\
&\geq \quad 0
\end{aligned}$$

In particular, take $g_k = -\log \tilde{a}_k$. Then $\tilde{H}(\{\tilde{a}_k\}_{k=1}^{\infty}) = -\sum_{k=1}^{\infty} \tilde{a}_k g_k$ and if $\sum_{k=1}^{\infty} \tilde{a}_k g_k = \infty$ then we are done. If not:

$$\begin{aligned}
-\infty \quad &< \quad \tilde{H}(\{\tilde{a}_k\}_{k=1}^{\infty})\\
&= \quad -\sum_{k=1}^{\infty} \tilde{a}_k g_k\\
&\leq \quad -\sum_{k=1}^{\infty} \tilde{b}_k g_k\\
&= \quad \sum_{k=1}^{\infty} \tilde{b}_k \log \tilde{a}_k\\
&\leq \quad \sum_{k=1}^{\infty} \tilde{b}_k \log \tilde{b}_k\\
&= \quad \tilde{H}(\{\tilde{b}_k\}_{k=1}^{\infty})
\end{aligned}$$

(The last inequality is essentially Jensen's, i.e. the relative-entropy inequality, extended because $\sum_{k=1}^{n} \tilde{b}_k \log \frac{\tilde{a}_k}{\tilde{b}_k}$ might not converge: $\log x \leq x - 1 \Rightarrow \limsup \sum_{k=1}^{n} \tilde{b}_k \log \frac{\tilde{a}_k}{\tilde{b}_k} \leq 0$, and since $\sum_{k=1}^{n} \tilde{b}_k \log \tilde{a}_k > -\infty$, $\sum_{k=1}^{n} \tilde{b}_k \log \tilde{b}_k$, which is decreasing in n, has a limit and the limit exceeds $\sum_{k=1}^{\infty} \tilde{b}_k \log \tilde{a}_k$.)

Q.E.D. (Proposition 1)

What remains to be done is to remove the conditions of continuity and uniform boundedness imposed in Proposition 1. One way to accomplish this is to approximate $p(x|y)$ by a sequence of functions $\{f_n(x|y)\}$ that satisfy the conditions of Proposition 1, and then to make sure that the entropy is continuous in the approximation. Following this plan, for each $n = 1, 2, \ldots$ and each $y \in R^m$, define

$$f_n(x|y) = n \int_x^{x+\frac{1}{n}} \min(n, p(z|y))dz \quad \forall x \in [0, \infty)$$

and $f_n(x|y) = f_n(-x|y)$ for $x \in (-\infty, 0)$. (If x is multi-variate, use the same construction along any line emanating from the origin, and then use rotation invariance to complete the definition.) Here is a summary of the properties of $\{f_n(x|y)\}$; all are easily verified.

- **i.** $f_n(x|y)$ is symmetric around $x = 0$ and unimodal, for every y and every n;
- **ii.** $f_n(x|y)$ is non-negative, continuous, and integrable in x for every y and every n;
- **iii.** $|f_n(x|y)| \leq n$ for every (x, y) and every n;
- **iv.** $f_n(x|y) \leq p(x|y)$ and $f_n(x|y) \leq f_{n+1}(x|y)$ for every (x, y) and every n;
- **v.** For every y, $f_n(x|y) \to p(x|y)$ $a.s.$ dx, as $n \to \infty$ (in fact, at every point of continuity of $p(\cdot|y)$), and hence $f_n(x|y) \to p(x|y)$ $a.s$ $dx \times dF(y)$

Now consider

$$f_n^o(x) \doteq \int_{R^m} f_n(x|y)dF(y)$$

and

$$f_n^g(x) \doteq \int_{R^m} f_n(x - g(y)|y)dF(y)$$

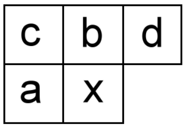

Fig. 1. The median edge detector, m.e.d., predicts intensity x from the intensities a, b, and c. The residual can then be directly encoded, or conditionally encoded based on the "context." The JPEG lossless coding standard LOCO uses a, b, c, and d to define the context.

According to Proposition 1, $\tilde{H}(f_n^g) \leq \tilde{H}(f_n^o)$ for every n. The technicalities involved in concluding that therefore $\tilde{H}(p^g) \leq \tilde{H}(p^o)$ are in Appendix B, which then completes the proof of the Theorem.

III. Experiments with Images

Predictive image-encoding schemes typically visit pixels in raster-scan order. Ignoring picture boundaries, a much-used proxy for the "past" of a given pixel "x" is the triple consisting of pixels to the immediate left of x, immediately above x, and, diagonally, immediately above and to the left of x: a, b, and c, respectively, in Figure 1.

The idea is that the intensities at a, b, and c (themselves denoted, ambiguously, a, b, and c for convenience) can be used to make a good first guess at the intensity at x (itself denoted by x). If $g(a, b, c)$ is a good predictor of x, then it might be expected that coding $X - g(A, B, C)$ is more efficient than coding just X (where we have used upper-case letters to distinguish random variables from observations). Formally

$$\begin{aligned} H(X|\text{raster-order past}) &\leq H(X|A, B, C) \\ &= H(X - g(A, B, C)|A, B, C) \\ &\leq H(X - g(A, B, C)) \end{aligned}$$

Since a Shannon code of $X|$raster-order past is optimal, but impractical, the idea is to choose g to minimize $H(X - g(A, B, C))$.

The connection to the result derived in §II is through the common observation that conditional distributions on intensities, such as those of X given A, B, and C, are typically symmetric and unimodal in real images. Since the distribution of $X - g(A, B, C)$ is a mixture of the conditional distributions of $X - g(a, b, c)$, given $A = a$, $B = b$, and $C = c$, mixed with respect to the joint distribution of A, B, and C, we are faced with exactly the problem addressed in §II, with $Y = (A, B, C) \in R^3$. As an illustration of the theorem we will examine, empirically, the entropy of $X - g(A, B, C)$ for different predictors g on a set of six images borrowed from the image library made public as part of the JPEG standard [19]. We wish to emphasize that our experiments are not meant as a step towards an improved image-compression scheme, as we are well aware of the many practical issues involving complexity of encoding and decoding, proper handling of quantization errors, run coding, and so-on. Instead, we hope to illustrate some connections between natural scene statistics and the use of median-type estimators in progressive image encoding.

The lossless compression algorithm LOCO ([8], [9]), as well as an earlier scheme by Martucci [20], use what Weinberg et al. [8] call the *median edge detector* (m.e.d.): $g_{\text{m.e.d.}}(a, b, c) = \text{median}(a, b, a + b - c)$, which can also be written as

$$g_{\text{m.e.d.}}(a, b, c) = \begin{cases} \min(a, b) & \text{if } c \geq \max(a, b) \\ \max(a, b) & \text{if } c \leq \min(a, b) \\ a + b - c & \text{otherwise} \end{cases}$$

It would be of interest, for the purpose of testing the assumptions of symmetry and unimodality, as well as to illustrate the result of §II, to compare $H(X - g_{\text{m.e.d.}}(A, B, C))$ to $H(X - g_{\text{opt}}(A, B, C))$, where $g_{\text{opt}}(a, b, c)$ is the actual median of the conditional distribution of X given $A = a$, $B = b$, and $C = c$.

Evidently, the proposed experiment requires a knowledge of $P(X = x|A = a, B = b, C = c)$ for every value of x, a, b, and c, which is hard to come by, even for a modest eight-bit pixel depth. For our experiments, we made an additional assumption about the nature of image statistics (supported to a degree by the resulting demonstration that $H(X - g_{\text{opt}}(A, B, C)) < H(X - g_{\text{m.e.d.}}(A, B, C))$ – see below): we will call $P(X|Y)$, $X \in R$, $Y \in R^m$, *shift invariant* if $P(X = x|Y = y) = P(X = x + s|Y = y + s)$ for every scalar s, where $y + s$ represents the addition of s to each component of y. To the extent that $P(X|A, B, C)$ is shift invariant, it can be estimated efficiently from the empirical tri-variate distribution of $(X - A, B - A, C - A)$ as follows:

$$\begin{aligned} &P(X = x|A = a, B = b, C = c) \\ &= P(X = x + s|A = a + s, B = b + s, C = c + s) \quad \forall\ s \end{aligned}$$

and hence, also, for the right-hand side mixed under any distribution on s. Using, in particular, $P(A - a = s|B - A = b - a, C - A = c - a)$:

$$\begin{aligned} &P(X = x|A = a, B = b, C = c) \\ &= \sum_s \{P(X = x + s|A = a + s, B = b + s, C = c + s) \\ &\times P(A - a = s|B - A = b - a, C - A = c - a)\} \\ &= \sum_s \left\{ \frac{P(X = x + s, A = a + s, B = b + s, C = c + s)}{P(A = a + s, B = b + s, C = c + s)} \right. \\ &\left. \times \frac{P(A - a = s, B - A = b - a, C - A = c - a)}{P(B - A = b - a, C - A = c - a)} \right\} \end{aligned}$$

Since

$$\sum_s P(X = x + s, A = a + s, B = b + s, C = c + s)$$

$$= \sum_s P(A - a = s, X - A = x - a, B - A = b - a, C - A = c - a)$$

$$= P(X - A = x - a, B - A = b - a, C - A = c - a)$$

and

$$P(A = a + s, B = b + s, C = c + s)$$

$$= P(A - a = s, B - A = b - a, C - A = c - a)$$

Fig. 2. Six ISO/JPEG test images – see [19]

we have, finally,

$$\begin{aligned} &P(X = x|A = a, B = b, C = c) \qquad (4)\\ &= \frac{P(X - A = x - a, B - A = b - a, C - A = c - a)}{P(B - A = b - a, C - A = c - a)} \end{aligned}$$

For each of the six images in Figure 2 we used equation 4 to compute the empirical median[2] of

$$P(X|A = a, B = b, C = c)$$

("$g_{\text{opt}}(a, b, c)$"), and then computed $H(X-g_{\text{m.e.d.}}(A, B, C))$ and $H(X-g_{\text{opt}}(A, B, C))$ using the empirical distributions on $X-g_{\text{m.e.d.}}(A, B, C)$ and $X-g_{\text{opt}}(A, B, C)$, respectively. In each case, $H(X-g_{\text{opt}}(A, B, C)) < H(X-g_{\text{m.e.d.}}(A, B, C))$ (see Figure 3), as might be expected from the development in §II together with the observation that, typically, conditional distributions of images are nearly symmetric, unimodal, and shift invariant.

Image	m.e.d.	mod	opt
Gold	4.72	4.68	4.60
Hotel	4.74	4.68	4.48
Water	3.69	3.55	3.46
Woman	4.89	4.85	4.76
Cmpnd1	1.91	1.88	1.68
Tools	5.63	5.62	5.41

Fig. 3. Entropies of empirical residual distributions, for each of the six images in Figure 2, and for each of three predictors (m.e.d.: median edge detector; mod: modified median edge detector; opt: empirical conditional median).

The middle column in the table of Figure 3 is based on a third predictor, g_{mod}, which we devised by a simple modification of the median edge detector. We observed that $g_{\text{m.e.d.}}$ is an excellent approximation of the conditional median (g_{opt}) when $|a-b| > 15$, but less accurate at smaller values of $|a-b|$. An ad hoc correction was made for $|a-b| \leq 15$, defining what we will call the modified predictor, g_{mod}: $g_{\text{mod}} = g_{\text{m.e.d.}}$ whenever $|a - b| > 15$, and

$$\begin{aligned} &g_{\text{mod}}(a, b, c) =\\ &= \begin{cases} [\frac{a+b+\min(a,b)}{3}] & \text{if } c \geq \max(a, b)\\ [\frac{a+b+\max(a,b)}{3}] & \text{if } c \leq \min(a, b)\\ [0.6 * a + 0.6 * b - 0.2 * c] & \text{otherwise,} \end{cases} \end{aligned}$$

whenever $|a-b| \leq 15$, where $[t]$ denotes the integer nearest to t. In all six test images, $H(X-g_{\text{mod}}(A, B, C))$ fell between $H(X - g_{\text{m.e.d.}}(A, B, C))$ and $H(X - g_{\text{opt}}(A, B, C))$.

Putting aside practical considerations, it is better to code $X - g(A, B, C)$ under a conditional distribution than to code it directly – conditional entropies never exceed unconditional entropies. It is obviously impossible to condition on the entire "past" (already-encoded pixels), but not impossible to divide the past into categories, or "contexts," within which (i.e. conditioned on which) $X-g(A, B, C)$ may have significantly lower entropy. Complexity grows with the number of categories, so when it comes to a practical implementation, there is a tradeoff. The highly efficient lossless compression scheme LOCO defines 365 contextual categories, based upon the four contextual pixels labeled a, b, c, and d in Figure 1; see [8] for details.

We repeated our experiment using the same three predictors ($g_{\text{m.e.d.}}$, g_{mod}, and g_{opt}), but comparing, instead, the context-conditioned entropies, under the particular contextual categories defined in LOCO. With $Cxt(a, b, c, d) \in \{1, 2, \ldots, 365\}$ representing the LOCO categories:

$$\begin{aligned} &H(X|\text{raster-order past})\\ &\leq H(X|A, B, C, D)\\ &= H(X - g(A, B, C)|A, B, C, D)\\ &\leq H(X - g(A, B, C)|Cxt(A, B, C, D)) \end{aligned}$$

To the extent that g_{opt} is still a good estimate of the median, now conditioned on $Cxt(a, b, c, d)$, the theorem of §II would still apply and we would expect $H(X - g_{\text{opt}}(A, B, C)|Cxt(A, B, C, D))$ to improve on the corresponding conditional residual entropies under the estimators $g_{\text{m.e.d.}}$ and g_{mod}. Of course all of the entropies are improved by context, but at the same time the orderings are preserved, in every case, as can be seen by comparing the tables in Figures 3 and 4.

Image	m.e.d.	mod	opt
Gold	4.46	4.42	4.36
Hotel	4.40	4.34	4.19
Water	3.59	3.49	3.43
Woman	4.45	4.41	4.36
Cmpnd1	1.13	1.13	1.03
Tools	5.25	5.23	5.07

Fig. 4. Context-conditioned entropies of empirical residual distributions, using the 365 context categories defined in LOCO [8]. Notation as in Figure 3.

[2] When the median fell between two integers, we chose the more populated of the two values.

APPENDIX A: PROOF OF LEMMAS 1 AND 2

Proof of Lemma 1. First, we note that $O_h(z)$ is left continuous: Let $B_z = \{x : h(x) \geq z\}$, and take $0 < z_1 \leq z_2 \leq \ldots$, $\tilde{z}$, such that $z_n \to \tilde{z}$. Then $B_{z_n} \supseteq B_{z_{n+1}}$ $\forall n$ and

$$B_{\tilde{z}} = \bigcap_{n=1}^{\infty} B_{z_n}$$

Also, $z_1 > 0 \Rightarrow \lambda\{x : h(x) \geq z_1\} < \infty$ (since h is integrable), and hence, by continuity of measures,

$$O_h(z_n) = \lambda(B_{z_n}) \to \lambda(B_{\tilde{z}}) = O_h(\tilde{z})$$

so that $O_h(z)$ is left continuous.

Furthermore, we claim that $O_h(z)$ is strictly decreasing on $[0, Z_M]$, where $Z_M = \sup_x h(x)$. Choose $z_1, z_2 \in [0, Z_M]$ with $z_1 < z_2$. Then

$$\begin{aligned} [z_1, \infty) &= [z_1, z_2) \cup [z_2, \infty) \\ \Rightarrow h^{-1}[z_1, \infty) &= h^{-1}[z_1, z_2) \cup h^{-1}[z_2, \infty) \\ \Rightarrow O_h(z_1) &= \lambda(h^{-1}[z_1, z_2)) + O_h(z_2) \end{aligned}$$

and it is enough to show $\lambda(h^{-1}[z_1, z_2)) > 0$. Since h is continuous $h^{-1}(z_1, z_2)$ is open, and therefore $\lambda(h^{-1}[z_1, z_2))$ $\geq \lambda(h^{-1}(z_1, z_2)) > 0$.

(a) m^h is clearly non-increasing. For any $x \in [0, \infty)$, let $A_x = \{z : O_h(z) \geq x\}$. Since $O_h(0) = +\infty$, $0 \in A_x$ $\forall x \in [0, \infty)$, so A_x is never empty. Fix $\tilde{x} \in [0, \infty)$ and let $\tilde{z} = m^h(\tilde{x})$. Claim $O_h(\tilde{z}) \geq \tilde{x}$ (i.e. $\tilde{z} \in A_{\tilde{x}}$). Take $z_n \in A_{\tilde{x}}$, $n = 1, 2, \ldots \ni$ $z_n \uparrow \tilde{z}$. Then $O_h(z_n) \downarrow O_h(\tilde{z})$, since O_h is left continuous. Hence $O_h(\tilde{z}) \geq \tilde{x}$.

Fix $x_o \in [0, \infty)$ and choose $\{x_n^+\}$ non-decreasing, $x_n^+ \to x_o$ and $\{x_n^-\}$ non-increasing, $x_n^- \to x_o$. Let $z_n^+ = m^h(x_n^+)$, $z_n^- = m^h(x_n^-)$. Then $\{z_n^+\}$ is non-increasing; let $z_o^+ = \lim z_n^+$. And $\{z_n^-\}$ is non-decreasing and furthermore bounded by Z_M; let $z_o^- = \lim z_n^-$. For continuity it is enough to show that both $z_o^+ = m^h(x_o)$ and $z_o^- = m^h(x_o)$.

$\underline{z_o^+ = m^h(x_o)}$ $z_n^+ \in A_{x_n^+}$ and $z_n^+ \geq z_o^+ \Rightarrow z_o^+ \in A_{x_n^+}$ $\forall n$. Thus $O_h(z_o^+) \geq x_n^+$ $\forall n \Rightarrow O_h(z_o^+) \geq x_o$, since $x_n^+ \to x_o$. Hence $z_o^+ \in A_{x_o}$. What's more, if $z > z_o^+$ then $z > z_n^+$ $\forall n$ large $\Rightarrow z \notin A_{x_n^+}$ $\forall n$ large $\Rightarrow z \notin A_{x_o}$ (since $x_o \geq x_n^+ \Rightarrow$ $A_{x_o} \subseteq A_{x_n^+}$). Hence $z_o^+ = m^h(x_o)$. Similarly, if $x_n^+ \uparrow \infty$, then $z_o^+ \in A_{x_n^+}$ $\forall n$ and hence $O_h(z_o^+) \geq x_n^+$ $\forall n$, so $O_h(z_o^+) =$ $+\infty$. But h integrable $\Rightarrow O_h(z) < \infty$ $\forall z > 0$. Thus $z_o^+ = 0$, and $\lim_{x \to \infty} m^h(x) = 0$.

$\underline{z_o^- = m^h(x_o)}$ Since (i) z_n^- is non-decreasing, (ii) $O_h(z_n^-) \geq$ x_n^- $\forall n$, and (iii) O left continuous:

$$O_h(z_o) = \lim_{n \to \infty} O(z_n^-) \geq \lim_{n \to \infty} x_n^- = x_o$$

Hence $z_o^- \in A_{x_o}$.

Let $z \in A_{x_o}$ and suppose $z > z_o^-$. Choose $\tilde{z} \in (z_o^-, z)$. Then O_h strictly decreasing $\Rightarrow O_h(\tilde{z}) > x_o \Rightarrow O_h(\tilde{z}) \geq x_n^-$ $\forall n$ large $\Rightarrow \tilde{z} \leq z_n^-$ $\forall n$ large. But $z_n^- \leq z_o^- \Rightarrow \tilde{z} \leq z_o^-$, which contradicts $\tilde{z} \in (z_o^-, z)$. Hence $\forall z \in A_{x_o}$, $z \leq z_o^-$, i.e. $z_o^- = m^h(x_o)$.

(b) Start with $G(z) = 1_{z \geq a}$. Then

$$\int_R G(h(x))dx = \lambda\{x : h(x) \geq a\} = O_h(a)$$

and

$$\int_0^\infty G(m^h(x))dx = \lambda\{x : m^h(x) \geq a\}$$

Hence it would be enough to show

$$m^h(x) \geq a \Leftrightarrow x \leq O_h(a) \tag{A-1}$$

Choose x_o such that $m^h(x_o) \geq a$. Recall (from the proof of part (a)) that if $z_o = \sup\{z : O_h(z) \geq x\}$ then $O_h(z_o) \geq x$. So $O_h(m^h(x_o)) \geq x_o$. Since O_h is non-increasing: $x_o \leq$ $O_h(m^h(x_o)) \leq O_h(a)$. Now choose $x_o \leq O_h(a)$. Then $a \in$ $\{z : O_h(z) \geq x_o\} \Rightarrow m^h(x_o) = \sup\{z : O_h(z) \geq x_o\} \geq a$. This proves (A-1).

Now extend by the usual arguments: first to $G(z) = 1_{z \in [a,b)}$ (by noting that $1_{z \in [a,b)} = 1_{z \geq a)} - 1_{z \geq b}$); then to G of the form

$$G(z) = \sum_{i=1}^{n} 1_{z \in [a_i, b_i)}$$

then to $G(z) = 1_{z \in B}$, B Borel (using Monotone Class Theorem); and finally to arbitrary G (using monotone approximation by simple functions).

(c) Let $\tilde{x}_o = \inf\{x : m^h(x) = m^h(x_o)\}$. Observe, first, that

$$\begin{aligned} \lambda\{x : h(x) = m^h(x_o)\} &= \int_{-\infty}^{\infty} 1_{h(x) = m^h(x_o)} dx \\ &= \int_0^\infty 1_{m^h(x) = m^h(x_o)} dx \\ &= \lambda\{x : m^h(x) = m^h(x_o)\} \\ &\geq x_o - \tilde{x}_o \end{aligned}$$

(by application of (b) with $G(z) = 1_{z = m^h(x_o)}$ and by virtue of the monotonicity of m^h). Similarly,

$$\begin{aligned} &\int_0^{x_o} m^h(x)dx \\ &\quad = \int_0^{\tilde{x}_o} m^h(x)dx + \int_{\tilde{x}_o}^{x_o} m^h(x)dx \\ &\quad = \int_0^\infty m^h(x) 1_{m^h(x) > m^h(x_o)} dx + (x_o - \tilde{x}_o) m^h(x_o) \\ &\quad = \int_{-\infty}^{\infty} h(x) 1_{h(x) > m^h(x_o)} dx + (x_o - \tilde{x}_o) m^h(x_o) \end{aligned} \tag{A-2}$$

(again by application of (b), this time with $G(z) = z 1_{z > m^h(x_o)}$, and again using the monotonicity of m^h).

Let $B^+ = \{x : h(x) > m^h(x_o)\}$ and choose $B^o \subseteq \{x : h(x) = m^h(x_o)\}$ such that $\lambda(B^o) = x_o - \tilde{x}_o$. Then

$$\begin{aligned} \lambda(B^+) &= \int_{-\infty}^{\infty} 1_{h(x) > m^h(x_o)} dx \\ &= \int_0^\infty 1_{m^h(x) > m^h(x_o)} dx \\ &= \tilde{x}_o \end{aligned}$$

by the monotonicity and continuity of m^h, and therefore (see equation (A-2))

$$\int_0^{x_o} m^h(x)dx = \int_{B^+ \cup B_o} h(x)dx$$

with $\lambda(B^+ \cup B_o) = \tilde{x}_o + x_o - \tilde{x}_o = x_o$.

Now fix $A \subseteq R$ with $\lambda(A) = x_o$, and write $A = A^+ \cup A^o \cup A^-$ where

$$\begin{array}{rcl} A^+ & = & \{x \in A : h(x) > m^h(x_o)\} \\ A^o & = & \{x \in A : h(x) = m^h(x_o)\} \\ A^- & = & \{x \in A : h(x) < m^h(x_o)\} \end{array}$$

Finally, observe that

$$\int_A h(x)dx = \int_{A^+} h(x)dx + \int_{A^o} h(x)dx + \int_{A^-} h(x)dx$$

can be increased by transferring mass from A^o and A^- to $B^+ \backslash A^+$ until all of B^+ is accounted for, and then moving the remaining mass in A^- (if any) to $\{x : h(x) = m^h(x_o)\}$. Hence

$$\int_A h(x)dx \leq \int_{B^+ \cup B^o} h(x)dx = \int_0^{x_o} m^h(x)dx$$

Q.E.D. (Lemma 1)

Proof of Lemma 2. Let $x_o = \inf\{x : f(x) \leq 1/e\}$, and let $k_o = k_o(N) = \lfloor x_o N \rfloor$ (greatest integer less than or equal to $x_o N$). Then $x_o \in [\frac{k_o}{N}, \frac{k_o+1}{N})$ and

$$\begin{array}{rcll} f(x) & > & \dfrac{1}{e} & \forall\, x \in [0, \dfrac{k_o}{N}) \\ f(x) & \leq & \dfrac{1}{e} & \forall\, x \in [\dfrac{k_o+1}{N}, \infty) \end{array}$$

(although, possibly, $k_o = 0$).

Evidently, $\sup_{N,k} f_k^N \leq f(0) < \infty$, and

$$\begin{array}{rcll} f_k^N & > & \dfrac{1}{e} & \forall\, k \leq k_o \\ f_k^N & \leq & \dfrac{1}{e} & \forall\, k \geq k_o + 2 \end{array}$$

Hence

$$\sum_{k=1}^{k_o} \frac{1}{N} f_k^N \log f_k^N \overset{N\to\infty}{\longrightarrow} \int_0^{x_o} f(x) \log f(x) dx < \infty$$

(Riemann approximation) since

$$f(\frac{k-1}{N}) \geq f_k^N \geq f(\frac{k}{N})$$

and $|z \log z|$ is bounded by $\max(1/e, f(0) \log f(0))$ on $z > 1/e$.

Obviously, both $\frac{1}{N} f_{k_o+1}^N \log f_{k_o+1}^N$ and $\frac{1}{N} f_{k_o+2}^N \log f_{k_o+2}^N$ are negligible as $N \to \infty$.

Now since $z \log z$ is negative and decreasing on $[0, 1/e]$

$$\begin{aligned} &\int_{\frac{k-2}{N}}^{\frac{k-1}{N}} f(x) \log f(x) dx \\ &\quad \leq \frac{1}{N} f_k^N \log f_k^N \\ &\quad \leq \int_{\frac{k}{N}}^{\frac{k+1}{N}} f(x) \log f(x) dx \\ &\quad \leq 0 \end{aligned}$$

for all $k \geq k_o + 3$. Hence

$$\begin{aligned} &\int_{\frac{k_o+1}{N}}^{\infty} f(x) \log f(x) dx \\ &\quad \leq \sum_{k=k_o+3}^{\infty} \frac{1}{N} f_k^N \log f_k^N \\ &\quad \leq \int_{\frac{k_o+3}{N}}^{\infty} f(x) \log f(x) dx \end{aligned}$$

all of which is less than or equal to zero. Taking the $N \to \infty$ limit

$$\sum_{k=k_o+3}^{\infty} \frac{1}{N} f_k^N \log f_k^N \to \int_{x_o}^{\infty} f(x) \log f(x) dx \in [-\infty, 0]$$

and, putting together the pieces,

$$-\infty \leq \frac{1}{N} \sum_{k=1}^{\infty} f_k^N \log f_k^N \to \tilde{H}(f) \in [-\infty, \infty)$$

Q.E.D. (Lemma 2)

APPENDIX B: Approximation of $\tilde{H}(p^o)$ by $\tilde{H}(f_n^o)$

It remains to show that $\tilde{H}(f_n^g) \leq \tilde{H}(f_n^o)$ for every n implies $\tilde{H}(p^g) \leq \tilde{H}(p^o)$.

We will show that $\tilde{H}(f_n^o) \to \tilde{H}(p^o)$; the argument for $\tilde{H}(f_n^g) \to \tilde{H}(p^g)$ is identical. Obviously $f_n^o(x)$ is non-decreasing in n for every x and, furthermore, bounded by $p^o(x)$. Hence $f_n^o(x)$ has a limit (possibly $+\infty$ at $x = 0$) for every x. We claim that $f_n^o \uparrow p^o$ $a.s.\ dx$. By dominated convergence

$$\begin{aligned} &\int_R |p^o(x) - f_n^o(x)| dx \\ &\quad = \int_R | \int_{R^m} (p(x|y) - f_n(x|y)) dF(y)| dx \\ &\quad \leq \int_R \int_{R^m} |(p(x|y) - f_n(x|y))| dF(y) dx \\ &\quad \to 0 \end{aligned}$$

so $f_n^o(x) \to p^o(x)$ in L^1, and hence, in light of the fact that $f_n^o(x)$ has an almost sure limit, $f_n^o \uparrow p^o$ $a.s.\ dx$.

As for the limit of $\tilde{H}(f_n^o)$, there are three cases to consider: $\tilde{H}(p^o) = -\infty$, $\tilde{H}(p^o) \in (-\infty, \infty)$, and $\tilde{H}(p^o) = +\infty$.

Suppose, first, that $\tilde{H}(p^o) \in (-\infty, \infty)$. Then

$$\begin{aligned}
|f_n^o(x)\log f_n^o(x)| &= |1_{\{f_n^o(x)\le\frac{1}{e}\}}1_{\{p^o(x)\le\frac{1}{e}\}}f_n^o(x)\log f_n^o(x) \\
&+ 1_{\{f_n^o(x)\le\frac{1}{e}\}}1_{\{p^o(x)>\frac{1}{e}\}}f_n^o(x)\log f_n^o(x) \\
&+ 1_{\{f_n^o(x)\in(\frac{1}{e},1]\}}f_n^o(x)\log f_n^o(x) \\
&+ 1_{\{f_n^o(x)>1\}}f_n^o(x)\log f_n^o(x)| \\
&\le 1_{\{p^o(x)\le\frac{1}{e}\}}|p^o(x)\log p^o(x)| \\
&+ 1_{\{p^o(x)>\frac{1}{e}\}}(1/e) \\
&+ p^o(x) \\
&+ 1_{\{p^o(x)>1\}}p^o(x)\log p^o(x) \\
&\in L^1
\end{aligned}$$

Hence, by dominated convergence,

$$\begin{aligned}
\lim_{n\to\infty}\tilde{H}(f_n^o) &= \lim_{n\to\infty}\int f_n^o(x)\log f_n^o(x)dx \\
&= \int \lim_{n\to\infty} f_n^o(x)\log f_n^o(x)dx \\
&= \int p^o(x)\log p^o(x)dx \\
&= \tilde{H}(p^o)
\end{aligned}$$

If on the other hand $\tilde{H}(p^o) = -\infty$, then for every $\epsilon > 0$, $1_{\{p^o(x)>\epsilon\}}p^o(x)\log p^o(x) \in L^1$, and

$$\lim_{\epsilon\downarrow 0}\int 1_{\{p^o(x)>\epsilon\}}p^o(x)\log p^o(x)dx = -\infty \qquad \text{(A-1)}$$

Fix $\epsilon < \frac{1}{e}$. Since $1_{\{f_n^o(x)>\epsilon\}} \to 1_{\{p^o(x)>\epsilon\}}$, and since the previous bound on the integrand, restricted to $\{f_n^o(x) > \epsilon\}$ (and hence also to $\{p^o(x) > \epsilon\}$) is again in L^1,

$$\begin{aligned}
\limsup_{n\to\infty}\tilde{H}(f_n^o) &= \limsup_{n\to\infty}\int f_n^o(x)\log f_n^o(x)dx \\
&\le \limsup_{n\to\infty}\int 1_{\{f_n^o(x)>\epsilon\}}f_n^o(x)\log f_n^o(x)dx \\
&= \int 1_{\{p^o(x)>\epsilon\}}p^o(x)\log p^o(x)dx
\end{aligned}$$

Hence, by virtue of (A-1), $\lim_{n\to\infty}\tilde{H}(f_n^o) = -\infty = \tilde{H}(p^o)$.

If, finally, $\tilde{H}(p^o) = +\infty$, then

$$\begin{aligned}
&\int f_n^o(x)\log f_n^o(x)dx \\
&\quad = \int 1_{\{f_n^o(x)\le 1\}}f_n^o(x)\log f_n^o(x)dx \\
&\quad + \int 1_{\{f_n^o(x)>1\}}f_n^o(x)\log f_n^o(x)dx
\end{aligned}$$

The first term is dominated, exactly as in the case $\tilde{H}(p^o) \in (-\infty, \infty)$, and therefore

$$\begin{aligned}
&\lim_{n\to\infty}\int 1_{\{f_n^o(x)\le 1\}}f_n^o(x)\log f_n^o(x)dx \\
&\quad = \int 1_{\{p^o(x)\le 1\}}p^o(x)\log p^o(x)dx > -\infty
\end{aligned}$$

What's more, $0 \le 1_{\{f_n^o>1\}}f_n^o\log f_n^o \uparrow 1_{\{p^o>1\}}p^o\log p^o$. And therefore, by monotone convergence,

$$\begin{aligned}
&\int 1_{\{f_n^o(x)>1\}}f_n^o(x)\log f_n^o(x)dx \\
&\quad \to \int 1_{\{p^o(x)>1\}}p^o(x)\log p^o(x)dx = +\infty
\end{aligned}$$

and hence, again, $\tilde{H}(f_n^o) \to \tilde{H}(p^o)$.

By the same arguments, $\tilde{H}(f_n^g) \to \tilde{H}(p^g)$, and we conclude that

$$\tilde{H}(p^g) = \lim_{n\to\infty}\tilde{H}(f_n^g) \le \lim_{n\to\infty}\tilde{H}(f_n^o) = \tilde{H}(p^o) \in [-\infty, \infty]$$

Acknowledgment

We would like to thank Wei Zhang for her careful reading and helpful comments on an earlier draft, and the anonymous referees for uniformly good suggestions about organization, clarification, and simplification.

References

[1] A. Netravali and J. O. Limb. Picture coding: A review. *Proc. IEEE*, Vol. 68, pp. 366-406, 1980.

[2] L.A. Lehmann and A. Macovski. Data compression of X-ray images by adaptive DPCM coding. *SPIE Digital Radiology Conference*, Stanford University, September 1981, pp. 314-317.

[3] P. Howard and J.S. Vitter. Error modeling for hierarchical lossless image compression. *IEEE Data Compression Conference*, Snowbird UT, pp. 269-278, 1992.

[4] G. Langdon and M. Manohar. Centering of context-dependent components of prediction error distributions of images. *SPIE Applications of Digital Image Processing XVI*, Vol. 2028, pp. 26-32, 1993.

[5] R.F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, Vol. 50, pp. 987-1007, 1982.

[6] T.P. Bollerslev. A conditionally heteroskedastic time series model for security prices and rates of return data. *Review of Economics and Statistics*, Vol. 69, pp. 542-547, 1987.

[7] Balaban, Ercan, Ouenniche, Jamal, Politou, and Danae. A note on return distribution of UK stock indices. *Applied Economics Letters*, Vol. 12, pp.573-576, 2005.

[8] M. Weinberger, G. Seroussi, and G. Sapiro. LOCO-I: A low complexity, context-based, lossless image compression algorithm. Proc. IEEE Data Compression Conference, Snowbird, Utah, March-April 1996.

[9] M. Weinberger, G. Seroussi, and G. Sapiro. The LOCO-I lossless image compression algorithm: principles and standardization into JPEG-LS. Hewlett-Packard Laboratories Technical Report No. HPL-98-193R1, November 1998, revised October 1999. *IEEE Trans. Image Processing*, Vol. 9, issue 8, August 2000, pp. 1309-1324.

[10] N. Farvardin and J.W. Modestino. Adaptive buffer-instrumented entropy-coded quantizer performance for memoryless sources. *IEEE Trans. Inf. Theory*, Vol. IT-32, pp. 9-22, 1986.

[11] X. Wu. Lossless compression of continuous-tone images via context selection, quantization, and modeling. *IEEE Trans. Image Processing*, Vol. IP-6, pp. 656-664, May 1997.

[12] K. Popat. Lossy compression of grayscale document images by adaptive-offset quantization. *Proceedings of IS&T/SPIE Electronic Imaging 2001: Document Recognition and Retrieval VIII*, January, 2001.

[13] M. Janžura, T. Koski, and A. Otáhal. Minimum entropy of error principle in estimation. *Information Sciences*, Vol. 79, pp. 123-144, 1994.

[14] G.H. Hardy, J.E. Littlewood, and G. Pólya. Inequalities. Cambridge University Press, 1934.

[15] H.L. Weidemann and E.B. Stear. Entropy analysis of parameter estimation. *Information and Control.*, vol. 14, pp. 493-506, 1969.

[16] A. Otáhal. Minimum entropy of error estimate for multi-dimensional parameter and finite-state-space observations. *Kybernetika*, Vol. 31, pp. 331-335, 1995.

[17] M. Janžura, T. Koski, and A. Otáhal. Minimum entropy of error estimation for discrete random variables. *IEEE Trans. Inf. Theory*, Vol. IT-42, pp. 1193-1201, 1996.

[18] T.M. Cover and J.A. Thomas. Elements of Information Theory. Second Edition, John Wiley & Sons, Inc., 2006.

[19] *Digital Compression and Coding of Continuous Tone Still Images – Requirements and Guidelines*, Sept. 1993. ISO/IEC 10918-1, ITU Recommend. T.81.

[20] S.A. Martucci. Reversible compression of HDTV images using median adaptive prediction and arithmetic coding. In *Proc. IEEE Int. Symp. Circuits Syst.*, pp. 1310-1313, 1990.

BIOGRAPHIES

Ting-Li Chen received his Bachelor of Science and Master of Science degrees in Mathematics from the National Taiwan University, Taiwan, in 1994 and 1996, and a Ph.D. degree in Applied Mathematics from Brown University in 2005. Ting-Li Chen joined the Division of Statistical Science at Academia Sinica, Taiwan, as an Assistant Research Fellow in 2006. His research interests are in computational probability and statistics, image processing and analysis, and clustering and classification.

Stuart Geman received a Bachelor of Science degree in Physics from the University of Michigan in 1971, a Masters degree in Physiology from Dartmouth Medical School in 1973, and a Ph.D. degree in Mathematics from MIT in 1977. Stuart Geman joined the faculty at Brown University in 1977, where he is currently the James Manning Professor in the Division of Applied Mathematics. His research interests are in probability and statistics, stochastic processes, and computer and biological vision.