Information Cost Functions

Hrvoje Šikić[1]
Department of Mathematics, University of Zagreb, Croatia
E-mail: hsikic@math.hr

and

Mladen Victor Wickerhauser[2]
Department of Mathematics, Washington University in St. Louis, Missouri
E-mail: victor@math.wustl.edu

A best orthogonal basis for a vector is selected from a library to minimize a cost function of the expansion coefficients. How it depends on the cost function, and under what conditions it provides the fastest nonlinear approximation, are still open questions which we partially answer in this paper.

Squared expansion coefficients may be considered a discrete probability density function, or pdf. We apply some inequalities for pdfs to obtain three positive results and two counterexamples. We use the notion of subexponentiality, derived from the classical proof of an entropy inequality, to derive a number of curious inequalities relating different information costs of a single pdf. We then generalize slightly the classical result that one pdf majorizes another if it is cheaper with respect to a large-enough set of information cost functions. Finally, we present inequalities that bracket any information cost for a pdf between two functions of norms of the pdf, plus a counterexample showing that our result has a certain optimality. Another counterexample shows that, unfortunately, the set of norm-type pdfs is not large enough to imply majorization.

We conclude that all information cost functions are weakly comparable to norms, but this is not quite enough to guarantee in general that the cheapest-norm pdf majorizes.

Key Words: best basis, concave, entropy, majorization, nonlinear approximation, pdf, rearrangement inequality, Schur functional, subexponential, wavelet packet library

[1] Supported in part by Southwestern Bell Telephone Company
[2] Supported in part by NSF and AFOSR

1. INTRODUCTION

Our ultimate goal is to obtain some estimates for the rate of approximation by partial sums of orthogonal functions. These yield existence and uniqueness results for fast divide-and-conquer algorithms that choose a best orthogonal basis.

Suppose we have a finite-energy signal to approximate. Given a collection of orthonormal bases, it is desirable to choose one that concentrates the signal's energy, namely, that has two properties:

- only a relatively tiny number of expansion coefficients are non-negligible;
- the individually negligible coefficients add up to a negligible sum.

The *fastest approximation basis* of the collection is the one for which the squared expansion coefficients, when rearranged into decreasing order, decrease most rapidly. Comparison of rates of decrease may be done with the classical notion of *majorization*, and we will say that the fastest approximation basis majorizes all others in the collection. However, to find that basis with an efficient divide-and-conquer strategy, it is necessary to avoid rearrangement.

There are classical inequalities that estimate rates of decrease without rearrangement. They use various *entropies*, or *information cost functions*. Some of these are described in seminal work by Hardy, Littlewood and Pólya [4, 5], and more recent contributions by Rényi [8], Aczél and Daróczy [1], and Marshall and Olkin [7]. The basis which minimizes a particular entropy or information cost function is called the *best basis* for that function.

This paper presents three results. In Section 2, we introduce the notion of subexponentiality, derived from the classical proof of an entropy inequality, and obtain four useful lemmas and a number of curious inequalities relating different information costs of a single pdf. In Section 3, we give a proof that one pdf majorizes another if and only if it is cheaper with respect to any information cost functions. Our proof reformulates a classical result of Hardy, Littlewood and Pólya, and is valid for infinitely-supported pdfs. Finally, in Section 4 we present two inequalities that bracket any information cost for a pdf between two simpler functions of the pdf, plus a counterexample showing that our result has a certain optimality.

These methods apply to the still open question of the existence of a fastest approximation basis within a library. The best basis for a single information cost function yields the sole candidate, and we hope to use the inequalities bracketing information cost functions to decide whether that candidate indeed majorizes all others.

2. INEQUALITIES RELATING COST FUNCTIONS

Let $p = \{p_n\}$ be a (discrete) probability density function, or *pdf*: $0 \leq p_n \leq 1$ and $\sum_n p_n = 1$. Let $\mathbf{M} = \{n : p_n > 0\}$ and write $M = \#\mathbf{M}$, its cardinality, if $\mathbf{M}$ is finite. Denote the positive reals $\{x > 0\}$ by $\mathbf{R}^+$.

An *additive information cost function* H is a real-valued functional defined on pdfs by $H(p) = \sum_n f(p_n)$, where $f : [0, 1] \to \mathbf{R}$ is nonnegative, concave and satisfies $f(0) = 0$. This is a special case of a *Schur concave functional* [1, 5, 7].

We have entropy in mind as a starting point: $f(t) = t \log(1/t)$. To generalize it, let f be given and define $\ell(t) \overset{\text{def}}{=} f(t)/t$ for $t > 0$. We may assume that f is right continuous at 0:

$$\lim_{t\to 0+} f(t) = 0. \tag{1}$$

This does not imply that $\ell(t)$ is right continuous at 0, and some of the most interesting examples have $\ell(t) \to \infty$ as $t \to 0$. The following notation will be used to indicate which ℓ defines the additive information cost function:

$$H = H(\ell, p) \overset{\text{def}}{=} \sum_n p_n \ell(p_n). \tag{2}$$

This article investigates some of the properties of such H.

If $\mathbf{M}$ is a singleton, then $H = \ell(1)$ is trivial to compute and does not involve any properties of ℓ on $(0,1)$ or $(1,\infty)$. Therefore, it will always be assumed that $\mathbf{M}$ contains at least two elements, and thus that

$$0 < p_n < 1, \qquad \text{for all } n \in \mathbf{M}. \tag{3}$$

Definition 2 and assumption 3 suggest that only the restriction of ℓ to $(0,1)$ matters. However, in many of the inequalities that follow, expressions like $\ell(1/p_n)$ will occur, necessitating an extension of ℓ to $(1,\infty)$. There is no *a priori* relation between the extension and the restriction to $(0,1)$.

Considering only $n \in \mathbf{M}$ avoids the need to define $\ell(0)$. However, comparing pdfs with ℓ satisfying $\ell(0+) = \infty$ will sometimes force evaluating $\ell(0) \overset{\text{def}}{=} \infty$. Because of Eq. 1, arithmetic in such cases will obey the conventions $0 \cdot \ell(0) = 0$; $x \cdot (\pm\infty) = \pm\infty$ for $x > 0$; $x + (\pm\infty) = \pm\infty$ for any real number x.

2.1. Basic machinery

We will say that ℓ is nonnegative, decreasing, or convex if $\ell|_{(0,1)}$ has these properties. We will call ℓ *concavable* if $t \mapsto t\ell(t)$ is a concave function on $(0,1)$, and *d-subexponential* if $\ell(x^d) \le d\,\ell(x)$ for a given $d \in \mathbf{R}$ and all $x \in (0,1)$.

A pdf p is called *$(1+d)$-summable* if $\sum_{n\in\mathbf{M}} p_n^{1+d} < \infty$ for a given $d \in \mathbf{R}$, and *$p\,\ell(p^d)$-summable* if $\sum_{n\in\mathbf{M}} p_n\, \ell(p_n^d)$ is finite for a given $d \in \mathbf{R}$. Every pdf is $(1+d)$-summable for all $d \ge 0$. Finite support implies $(1+d)$-summability and $p\ell(p^d)$-summability for any d. Conversely, if p is $(1+d)$-summable for some $d \le -1$, then $\mathbf{M}$ must be finite. Hence, for infinite $\mathbf{M}$, $(1+d)$-summability is unknown only for $d \in (-1,0)$.

A d-subexponential ℓ for a fixed d is needed to prove some of the inequalities below, but it is not a very strong condition. For example, ℓ is 0-subexponential if and only if $\ell(1) \le 0$, while $\ell(1) = 0$ in many interesting examples. Likewise, ℓ is 1-subexponential if and only if $\ell(t) \le \ell(t)$, which is no restriction at all. It is a restriction to have d-subexponential ℓ for $d \notin \{0,1\}$, or for more than one d. Also, there is a significant difference between $d > 0$ and $d < 0$: in the first case, only ℓ's behavior on $(0,1)$ matters, while the second influences the behavior of ℓ on all of $\mathbf{R}^+$. If D is a subset of $\mathbf{R}$ and ℓ is d-subexponential for every $d \in D$, it will be said that ℓ is D-subexponential.

Examples:

- $\ell(t) = \frac{1}{2} - t$ is convex on $(0,1)$, concavable, and decreasing, though not nonnegative.

- $\ell(t) = 1 - t^2$ is nonnegative, decreasing, and concavable, though not convex on $(0, 1)$.
- $\ell(t) = 1/t^2$ is nonnegative, decreasing, and convex on $(0, 1)$, though not concavable.
- Let $a > 1$ be a fixed real number. Define $\ell(t) = -\log_a(t)$; then ℓ is nonnegative, decreasing, convex on $(0, 1)$, convex on $(1, \infty)$, concavable, and $\mathbf{R}$-subexponential.

The hypothesis of subexponentiality allows us to estimate cost functions in terms of simpler functionals, but it is a strong assumption that implies behavior comparable to the logarithm function, at least on part of the domain.

The following several results explain D-subexponentiality in more detail. We begin with a version of a classical uniqueness result:

Lemma 2.1. *If $\ell : (0, 1) \to [0, \infty)$ satisfies $\ell(t^d) = d\ell(t)$ for all $d > 0$ and all $0 < t < 1$, then either $\ell \equiv 0$ or $\ell = -\log_a$ for some $a > 1$.*

Proof. Obviously $\ell \equiv 0$ works, so suppose $\ell \not\equiv 0$. Let $t_0 \in (0, 1)$ be a point such that $\ell(t_0) > 0$. Then for domain $d \in \mathbf{R}^+$, the map $d \mapsto t_0^d$ has range $(0, 1)$, and $d \mapsto d\ell(t_0)$ has range $(0, \infty)$. In particular, there is some $d_0 \in \mathbf{R}^+$ such that $d_0\ell(t_0) = 1$. But then $t_0^{d_0} \in (0, 1)$, and $\ell(t_0^{d_0}) = 1$. Now put $a = 1/t_0^{d_0}$; then $a > 1$ and $\ell(1/a) = 1$. It remains to show that $\ell(t) = -\log_a(t)$ for all $t \in (0, 1)$. But for any $t \in (0, 1)$, there is a unique $d > 0$ such that $t = (1/a)^d = a^{-d}$, and thus $\log_a t = -d = -d \cdot 1 = -d\ell(1/a) = -\ell(t)$. ■

Proposition 2.1. *If ℓ is nonnegative and $\mathbf{R}^+$-subexponential, then either $\ell \equiv 0$ or $\ell = -\log_a$ on $(0, 1)$ for some $a > 1$.*

Proof. For $d > 0$, ℓ is both d-subexponential and $1/d$-subexponential, which implies that $\ell(t^d) = d\ell(t)$ for all $t \in (0, 1)$ and $d > 0$. The result follows from Lemma 2.1. ■

Proposition 2.2. *If ℓ is nonnegative, $(0, 1)$-subexponential, and not identically zero on $(0, 1)$, then there exists $T \in (0, 1)$ and $K > 0$ such that*

$$\ell(t) \leq -K \ln t, \quad \textit{for every } t \in (T, 1); \tag{4}$$

$$\ell(t) \geq -K \ln t, \quad \textit{for every } t \in (0, T). \tag{5}$$

In particular, $\lim_{t \to 0^+} \ell(t) = \infty$ *and* $\lim_{t \to 1^-} \ell(t) = 0$.

Proof. The limit evaluations follow from Eqs. 4 and 5. By the hypotheses, there is some $T \in (0, 1)$ with $\ell(T) > 0$. For every $t \in (0, T)$ there exists $d \in (0, 1)$ such that $t^d = T$, namely $d = \frac{\ln T}{\ln t}$. Since ℓ is $(0, 1)$-subexponential, $\ell(T) \leq d\ell(t)$, so that, for every $t \in (0, T)$,

$$\ell(t) \geq -\ln t \frac{\ell(T)}{-\ln T},$$

which establishes Eq. 5 with $K = \frac{\ell(T)}{-\ln T}$. Similarly, for $t \in (T,1)$, choose $d = \frac{\ln t}{\ln T} \in (0,1)$ to get $T^d = t$. Since ℓ is d-subexponential, this implies $\ell(t) \le d\ell(T)$ and thus,

$$\ell(t) \le -\ln t \frac{\ell(T)}{-\ln T} = -K \ln t,$$

which proves Eq. 4. ■

There can be $(0,1)$-subexponentiality without logarithms, however. Although Proposition 2.2 forces $(0,1)$-subexponential ℓ to be bounded by the logarithm, it is less restrictive than Proposition 2.1. Consider the class of functions ℓ described in the following lemma:

Lemma 2.2. *Fix $\alpha > 0$. Then the function $\ell(t) = t^{-\alpha} - 1$ is strictly positive on $(0,1)$, decreasing and convex on $\mathbf{R}^+$, and $(0,1)$-subexponential. Also, ℓ is concavable if $0 < \alpha < 1$.*

Proof. That ℓ is positive, decreasing, and convex is evident. Now define

$$g(t) \overset{\text{def}}{=} d\ell(t) - \ell(t^d) = d(t^{-\alpha} - 1) - (t^{-\alpha d} - 1).$$

This is a continuously differentiable function on $(0,1]$, satisfying $g(1) = 0$, and having a strictly negative derivative on $(0,1)$:

$$g'(t) = -\alpha d t^{-\alpha-1} + \alpha d t^{-\alpha d-1} = -\alpha d (t^{-\alpha d-1}) \left[t^{\alpha(d-1)} - 1\right] = -(+)[+] < 0.$$

Thus, $g(t) > 0$ for all $t \in (0,1)$, establishing that ℓ is $(0,1)$-subexponential.

Finally, if $\alpha < 1$, then $f(t) = t\ell(t) = t^{1-\alpha} - t$ is concave, since $f''(t) = -\alpha(1-\alpha)t^{-1-\alpha} < 0$. ■

Notice also that $(0,1)$-subexponentiality does not imply convexity on $(0,1)$. Take ℓ_0 to be one of the example $(0,1)$-subexponential functions, that is, logarithm or $t^{-\alpha} - 1$. Define, for fixed $A \in (0,1)$, a function ℓ by

$$\ell(t) = \begin{cases} \ell_0(t), & \text{if } t \in (0,A); \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to check that ℓ is $(0,1)$-subexponential, but is not convex on $(0,1)$.

Consider the case of negative d. If only one exponent d is involved, there are many examples: Any function ℓ defined on $(0,1)$ can be extended to $(1,\infty)$ so as to be d-subexponential. For $x \in (1,\infty)$, there is a unique $t \in (0,1)$ such that $t^d = x$, and to have a d-subexponential ℓ it suffices to define $\ell(x) = r$, where $r \in \mathbf{R}$ is any number satisfying $r \le d\ell(t)$.

However, there are many restrictions on D-subexponential functions if $D \subset (-\infty,0)$ is not a singleton:

Proposition 2.3. *Suppose that ℓ is nonnegative, not identically zero on $(0,1)$, and $(-1,0)$-subexponential. Then there exist $\epsilon > 0$ and $K > 0$ such that*

$$\ell(t) \le K(-\ln t), \qquad \textit{for every } t \in (0,\epsilon).$$

Proof. By the assumptions, there exists $x \in (0,1)$ such that $\ell(x) > 0$. Take any $d \in (-1,0)$; then $y = x^d$ is in $(1,\infty)$, $\ell(y) \in \mathbf{R}$, and $\ell(y) \leq d\ell(x)$. This implies that $-\infty < \ell(y) < 0$. Consider $T = 1/y \in (0,1)$. For any $t \in (0,T)$ there exists a unique $d \in (-1,0)$ such that $t^d = y$, namely $d = \ln y/\ln t$. It follows that $\ell(y) \leq d\ell(t)$. But since $d < 0$,

$$\ell(t) \leq \frac{1}{d}\ell(y) = \frac{-\ell(y)}{\ln y}(-\ln t),$$

which proves the result with $K = -\ell(y)/\ln y > 0$. ■

Proposition 2.3 implies that it is not possible to extend $\ell(t) = t^{-\alpha} - 1$ to $(1,\infty)$ so that ℓ becomes $(-1,0)$-subexponential. However, it can be done for a single exponent $d < 0$.

Propositions 2.2 and 2.3 together show that if ℓ is nonnegative, not identically zero, and $(-1,0) \cup (0,1)$-subexponential, then $\ell(t) \sim -\ln(t)$ as $t \to 0+$.

2.2. Applications

We can derive inequalities for additive information cost functions H using the properties mentioned above. For example, it is obvious that if ℓ is nonnegative, then $H(\ell,p) \geq 0$ for every pdf p. A probabilistic interpretation of H aids in the derivations. Fix a pdf p and consider a probability space $(\Omega, 2^{\Omega}, P)$, where $\Omega = \mathbf{M}$ and P is defined by $P(\{n\}) = p_n$ for every $n \in \mathbf{M}$. Define a random variable $X : \Omega \to \mathbf{R}$ by

$$X(n) = p_n. \tag{6}$$

Then the following results hold:

1. If ℓ is nonnegative, then the expectation $E[\ell(X)]$ lies in $[0,\infty]$, and
$$E[\ell(X)] = H(\ell,p)$$
2. If $d \geq 0$, then $E\left[X^d\right]$ lies in $[0,1]$, and
$$E\left[X^d\right] = \sum_{n\in\mathbf{M}} p_n^{1+d}. \tag{7}$$
3. If $d < 0$ and p is $(1+d)$-summable, then $E\left[X^d\right]$ is finite and Eq. 7 is valid.
4. If $d \in \mathbf{R}$ and p is $p\,\ell(p^d)$-summable, then $E\left[\ell(X^d)\right]$ is finite and
$$E\left[\ell(X^d)\right] = \sum_{n\in\mathbf{M}} p_n \ell\left(p_n^d\right).$$

Proposition 2.4. *If ℓ is nonnegative, -1-subexponential, and convex on $(1,\infty)$, and if p is $(1+d)$-summable for some $d \leq -1$, then p is finitely supported and $0 \leq H \leq -\ell(M)$.*

Proof. First note that $H \geq 0$ since ℓ is nonnegative. Also, $\mathbf{M}$ must be finite. Thus, $\sum_{n\in\mathbf{M}} p_n \ell(p_n^d)$ is finite, so both $E\left[\ell(X^{-1})\right]$ and $E\left[X^{-1}\right]$ are finite. The convexity of ℓ on $(1,\infty)$ permits application of Jensen's inequality, yielding

$$\ell(M) = \ell\left(\sum_{n\in\mathbf{M}} p_n^{1+(-1)}\right) = \ell\left(E[X^{-1}]\right) \leq E\left[\ell(X^{-1})\right] = \sum_{n\in\mathbf{M}} p_n \ell\left(p_n^{-1}\right).$$

But ℓ is -1-subexponential, so $\ell\left(p_n^{-1}\right) \leq -\ell(p_n)$. Thus, $\ell(M) \leq -\sum_n p_n \ell(p_n) = -H(\ell, p)$. ■

Note that the upper bound in Proposition 2.4 is not sharp, since there may be many -1-subexponential extensions of ℓ. Taking the minimal one, $\ell(t) \overset{\text{def}}{=} -\ell(1/t)$ for $t > 1$, gives a sharp upper bound attained by $p_n = 1/M$ for all $n \in \mathbf{M}$.

Proposition 2.5. *If ℓ is nonnegative and convex on $(0,1)$, then $0 \leq \ell\left(\sum_n p_n^2\right) \leq H(\ell,p)$. Equality holds on the right if $p_n = 1$ for a single index n.*

Proof. If $H = \infty$ the conclusion holds trivially. If H is finite, then $H = E\left[\ell(X)\right]$ is finite. By Eq. 7, $E[X] = \sum_n p_n^2 \in (0,1)$, so Jensen's inequality implies $0 \leq \ell\left(E[X]\right) \leq E\left[\ell(X)\right] = H$. ■

Proposition 2.6. *Fix $d > 0$. If ℓ is nonnegative, convex on $(0,1)$, and d-subexponential, then*

$$0 \leq \frac{1}{d}\ell\left(\sum_n p_n^{1+d}\right) \leq H(\ell,p).$$

Proof. As in the previous proof, notice that p is $(1+d)$-summable and $\sum_n p_n^{1+d} = E\left[X^d\right] \in (0,1)$. Nonnegative ℓ and $d > 0$ imply the left-hand inequality. It remains to prove the right-hand inequality in the case $H < \infty$. But then $E\left[\ell(X)\right] = H$ is finite. Likewise, $E\left[\ell(X^d)\right] \leq dE\left[\ell(X)\right] = dH$ is also finite, by d-subexponentiality. Jensen's inequality applies and yields $\ell\left(E[X^d]\right) \leq E\left[\ell(X^d)\right]$. Since $\ell\left(E[X^d]\right) = \ell\left(\sum_n p_n^{1+d}\right)$ and $d > 0$, the result follows. ■

Proposition 2.7. *Fix $d > 0$. If ℓ is convex on $(1,\infty)$, and $-d$-subexponential, and if both $\{p_n^{1-d}\}$ and $\{p_n \ell(p_n^{-d})\}$ are summable, then*

$$H(\ell,p) \leq -\frac{1}{d}\ell\left(\sum_{n\in\mathbf{M}} p_n^{1-d}\right).$$

Proof. Since ℓ is convex on $(1,\infty)$ and $1 < E\left[X^{-d}\right] < \infty$ by the summability assumption, Jensen's inequality applies. Together with $-d$-subexponentiality, it yields

$$-dH = -dE\left[\ell(X)\right] \geq E\left[\ell(X^{-d})\right] \geq \ell\left(E[X^{-d}]\right) = \ell\left(\sum_n p_n^{1-d}\right).$$

Division by $-d < 0$ gives the result. ∎

2.3. Examples

Applying the results of the previous section to three specific examples gives proofs of some curious inequalities, and enables comparison of different cost functions in common use.

2.3.1. $\ell(t) = -\log_a(t);\ a > 1.$

This ℓ is nonnegative, decreasing, convex on $(0,1)$ and $(1,\infty)$, concavable, and **R**-subexponential, so all the results apply.

If p is a finitely-supported pdf, then Proposition 2.4 yields the following classical result:

$$0 \le \sum_{n\in\mathbf{M}} (-p_n)\log_a(p_n) \le \log_a(M). \tag{8}$$

Both estimates are sharp. Equality holds on the left if and only if $p_i = 1$ for a single i with $p_n = 0$ for all $n \neq i$. Equality holds on the right if and only if $p_1 = p_2 = \cdots = p_N = \frac{1}{N}$.

Proposition 2.5 gives another lower bound for every pdf p:

$$-\log_a\left(\sum_{n\in\mathbf{M}} p_n^2\right) \le \sum_{n\in\mathbf{M}} (-p_n)\log_a(p_n). \tag{9}$$

This is an improvement on Eq. 8: whenever $\mathbf{M}$ contains at least two elements, $\sum_{n\in\mathbf{M}} p_n^2 < 1$ and so $-\log_a\left(\sum_{n\in\mathbf{M}} p_n^2\right) > 0$. It is also sharp, for equality holds in the extreme cases $p_n = 1$ for a single n and $p_n = 1/M$ for all $n \in \mathbf{M}$.

More generally, Proposition 2.6 implies, for any $d > 0$ and all pdfs p,

$$0 \le -\frac{1}{d}\log_a\left(\sum_{n\in\mathbf{M}} p_n^{1+d}\right) \le \sum_{n\in\mathbf{M}} (-p_n)\log_a(p_n). \tag{10}$$

This lower bound is, in a sense, the best possible. Namely, for finitely supported p the following limit exists:

$$\lim_{d\to 0+} -\frac{1}{d}\log_a\left(\sum_{n\in\mathbf{M}} p_n^{1+d}\right) = \sum_{n\in\mathbf{M}} (-p_n)\log_a(p_n). \tag{11}$$

In the case $a = 2$, the expression on the left is known as the *Rényi entropy* $I_\alpha(p)$, where $\alpha = d + 1$ (see [8], p.468).

Finally, Proposition 2.7 implies that for $d > 0$ and any $(1-d)$-summable pdf p for which $\{p_n \log_a(p_n)\}$ is summable, the following inequality holds:

$$\sum_{n\in\mathbf{M}} (-p_n)\log_a(p_n) \le \frac{1}{d}\log_a\left(\sum_{n\in\mathbf{M}} p_n^{1-d}\right). \tag{12}$$

For finitely supported p, the following limit holds as well:

$$\lim_{d\to 0+} \frac{1}{d}\log_a\left(\sum_{n\in\mathbf{M}} p_n^{1-d}\right) = \sum_{n\in\mathbf{M}} (-p_n)\log_a(p_n). \tag{13}$$

2.3.2. $\ell(t) = t^{-\alpha} - 1;\ t \in (0,1);\ 0 < \alpha < 1.$

This ℓ is nonnegative, decreasing, convex on $(0,1)$, $(0,1)$-subexponential, and concavable. The results in this subsection do not depend on concavability, but it is used in the next chapter.

Propositions 2.6 and 2.5 give the following lower bound for $H(\ell, p)$, for every pdf p and every $d \in (0,1)$:

$$0 \leq \frac{1}{d}\left[\left(\sum_{n\in\mathbf{M}} p_n^{1+d}\right)^{-\alpha} - 1\right] \leq \left(\sum_{n\in\mathbf{M}} p_n^{1-\alpha}\right) - 1. \tag{14}$$

If scaled by the constant factor $(2^\alpha - 1)^{-1}$, the right-hand side of Eq. 14 is equal to the entropy of degree $1-\alpha$ introduced in [1], pp. 184–185. In the same reference, there is a useful discussion of the behavior of this functional as $\alpha \to 0$.

Negative subexponentiality may be used to obtain upper bounds for $H(t^{-\alpha}-1, p)$. As already shown, this is not possible for a single function ℓ and all negative d, but it can be achieved for every particular d using a suitable extension of ℓ to $(1,\infty)$.

For the case $d = -1$, to get -1-subexponentiality ℓ may be extended as follows:

$$\ell(x) = 1 - x^\alpha, \qquad \text{if } x > 1. \tag{15}$$

Since $0 < \alpha < 1$, the extension is convex on $(1,\infty)$. Thus, Proposition 2.4 shows that for every finitely supported p,

$$0 \leq \sum_{n\in\mathbf{M}} p_n^{1-\alpha} \leq M^\alpha \tag{16}$$

The upper bound is sharp: equality holds on the right if $p_n = 1/M$ for all $n \in \mathbf{M}$. In this example, that is the only case that achieves equality, as shown by the following argument: $t^{1-\alpha}$ lies below its tangent line at $t = 1$, so $t^{1-\alpha} \leq (1-\alpha)t + \alpha$ with equality if and only if $t = 1$. Putting $t = p_n/q_n$, where q is any other pdf supported on $\mathbf{M}$, multiplying the inequality by q_n, and summing over $n \in M$, we see that

$$\sum_{n\in\mathbf{M}} q_n^\alpha p_n^{1-\alpha} \leq 1,$$

with equality if and only $p_n = q_n$ for all $n \in \mathbf{M}$. Choosing $q_n = 1/M$ for all $n \in \mathbf{M}$ shows that $\sum_{n\in\mathbf{M}} p_n^{1-\alpha} \leq M^\alpha$, with equality if and only if $p_n = q_n = 1/M$ for all $n \in \mathbf{M}$.

More generally, for each $-d \in (-1, 0)$, the following extension of $\ell(x) = x^{-\alpha} - 1$ will be $-d$-subexponential:

$$\ell(x) = -d\left[x^{\alpha/d} - 1\right], \qquad \text{if } x > 1. \tag{17}$$

To have convexity as well, it is necessary that

$$-d < -\alpha < 0. \tag{18}$$

Eqs. 17 and 18 imply that if a pdf is $(1-d)$-summable, it is also $(1-\alpha)$-summable and $p\ell(p)$-summable. Hence, for $0 < \alpha \leq d < 1$ and every pdf p for which $\{p_n^{1-d}\}$

is summable, the following inequality holds:

$$\left(\sum_{n \in \mathbf{M}} p_n^{1-\alpha} \right)^{d} \leq \left(\sum_{n \in \mathbf{M}} p_n^{1-d} \right)^{\alpha}. \tag{19}$$

2.3.3. $\ell(t) = -t^{\beta-1} \log_a t$; $t \in (0,1)$; $a > 1$; $0 < \beta < 1$.

This ℓ is nonnegative, decreasing, convex on $(0,1)$ and $(0,1)$-subexponential. It is also concavable for $\beta \geq \frac{1}{2}$. If $a = 2$ and p is finitely supported, then

$$H(\ell, p) = {}_{(0,\beta)}H(p) \cdot \left(\sum_{n \in \mathbf{M}} p_n^{\beta} \right)^{-1}, \tag{20}$$

where ${}_{(0,\beta)}H(p)$ is the entropy of order $(0,\beta)$ defined by Aczél and Daróczy (see [1], p. 192). Proposition 2.6 applies (and, as a special case, so does Proposition 2.5) to give, for every pdf p and for every $d > 0$,

$$0 \leq -\frac{1}{d} \left(\sum_{n \in \mathbf{M}} p_n^{1+d} \right)^{\beta-1} \log_a \left(\sum_{n \in \mathbf{M}} p_n^{1+d} \right) \leq H(\ell, p) = - \sum_{n \in \mathbf{M}} p_n^{\beta} \log_a p_n. \tag{21}$$

Next, consider $d = -1$. If ℓ is extended to $(1, \infty)$ by

$$\ell(t) = -t^{1-\beta} \log_a t, \qquad \text{if } t > 1, \tag{22}$$

then it is a -1-subexponential function. However, for ℓ to be convex on $(1, \infty)$, it is necessary that $\beta \geq \frac{1}{2}$. Then Proposition 2.4 applies to give the following inequality for finitely-supported p, $\beta \geq \frac{1}{2}$, and ℓ defined by Eq. 22:

$$0 \leq \sum_{n \in \mathbf{M}} p_n^{\beta} \log_a (1/p_n) \leq M^{1-\beta} \log_a M. \tag{23}$$

More generally, for any $0 < \beta < 1$ and $d \leq 2(\beta - 1) < 0$, the extension of ℓ defined by

$$\ell(t) = -t^{\frac{\beta-1}{d}} \log_a t, \qquad \text{if } t > 1, \tag{24}$$

is d-subexponential and convex on $(1, \infty)$. Therefore, Proposition 2.7 applies. Moreover, any pdf which is $(1+d)$-summable will be $p\ell(p^d)$-summable: since $p_n^d \in (1, \infty)$ for all $n \in \mathbf{M}$, we compute $p_n \ell(p_n^d) = -dp_n^{\beta} \log_a p_n$. But also, since $\beta \geq 1 + \frac{d}{2}$, we have

$$p_n^{\beta} \leq p_n^{1+\frac{d}{2}} = p_n^{1+d} p_n^{-\frac{d}{2}},$$

and since $p_n^{-\frac{d}{2}} |\log_a p_n|$ is bounded on $(0,1)$ for any $d < 0$, we see that $\sum_n p_n^{\beta} |\log_a p_n|$ is bounded whenever $\sum_n p_n^{1+d}$ is bounded. With a change of d's sign for clarity, we conclude that for any $0 < \beta < 1$ and any $d \geq 2(1 - \beta) > 0$, we have

$$\sum_{n \in \mathbf{M}} p_n^{\beta} \log_a (1/p_n) \leq \frac{1}{d} \left(\sum_{n \in \mathbf{M}} p_n^{1-d} \right)^{\frac{1-\beta}{d}} \log_a \left(\sum_{n \in \mathbf{M}} p_n^{1-d} \right), \tag{25}$$

for any $(1-d)$-summable pdf p.

3. INEQUALITIES FOR COMPARING PDFS

Estimates of additive information cost functions use some elementary properties of nonnegative concave functions. Recall that $f = f(t)$ is concave if and only if its domain is a convex set and if, for all $0 \leq \theta \leq 1$ and all x, y in the domain of f,

$$f\left(\theta x + (1-\theta)y\right) \geq \theta f(x) + (1-\theta)f(y). \tag{26}$$

For twice differentiable f, this condition is equivalent to $f''(x) \leq 0$ for all x in the domain of f. The properties we need, which hold even in the non-differentiable case, are contained in the following two classical results:

Lemma 3.1. *The function $f = f(t)$ is concave if and only if for any numbers a, b, c, d in its domain satisfying $a < b$, $c < d$, $a \leq c$, and $b \leq d$, the following inequality holds:*

$$\frac{f(b)-f(a)}{b-a} \geq \frac{f(d)-f(c)}{d-c}.$$

Corollary 3.1. *If $f = f(t)$ is concave and nonnegative on $(0,1)$, then $f(t)/t$ is a non-increasing nonnegative function on $(0,1)$.* ■

No conclusion may be drawn about the concavity or convexity of $f(t)/t$, though: for the concave nonnegative "hat function,"

$$f(t) = \begin{cases} t, & \text{if } 0 \leq t < \frac{1}{2}; \\ 1-t, & \text{if } \frac{1}{2} \leq t \leq 1, \end{cases} \tag{27}$$

the function $f(t)/t$ is neither concave nor convex.

3.1. Information cost algebra

Suppose that H is an additive information cost function. Then we have several purely algebraic results applicable to the comparison of pdfs by H:

1. Pdfs form a *convex set*: If p and q are pdfs and $0 \leq \theta \leq 1$, then $\theta p + (1-\theta)q$ is a pdf, and $H(\theta p + (1-\theta)q) \geq \theta H(p) + (1-\theta)H(q)$. More generally, if q^n is a pdf for all n, and $\sum_n \alpha_n = 1$ with $\alpha_n \in [0,1]$ for each n, then $p = \sum_n \alpha_n q^n$ is a pdf, and $H(p) \geq \sum_n \alpha_n H(q^n)$.

2. A *stochastic* linear operator $Ax_i \stackrel{\text{def}}{=} \sum_j a_{ij}x_j$, for which $\sum_j a_{ij} = 1$, maps pdfs to pdfs. A *doubly stochastic* linear operator A, for which $\sum_j a_{ij} = 1 = \sum_i a_{ij}$, increases information cost: $H(Ap) \geq H(p)$ for any p, H.

3. The *tensor product* $p \otimes q \stackrel{\text{def}}{=} \{p_i q_j\}$ is a pdf, and $H(p \otimes q) \geq \max\left(H(p), H(q)\right)$. This result is sharp: equality holds for $f(t) = t$ and any p, or for any f if $p_i = q_j = 1$ for just a single index pair i, j.

(i) The entropy information cost function, defined by $f(t) = t\log(1/t)$, satisfies the stronger condition $H(p \otimes q) = H(p) + H(q)$. See [9], p. 277.

(ii) If H is defined by $f(t) = t^\alpha - t$ with fixed $0 < \alpha < 1$, then $H(p \otimes q) \geq H(p)H(q)$, with equality if and only if $p_j = q_i = 1$ for a single pair i, j.

(iii) The function $f(t) = t^r$, with any r, satisfies $f(q_i p_j) = f(q_i)f(p_j)$ and gives the stronger result $H(p \otimes q) = H(p)H(q)$. Of course, only $0 < r \leq 1$ gives an information cost functional.

Item 2. can be recovered from a theorem by Markus [6]. Earlier, Hardy, Littlewood and Pólya [4] used the result that if p is a *finite* discrete pdf, and A is a doubly stochastic matrix, then $H(Ap) \geq H(p)$ for every additive information cost function H.

3.2. Rearrangement inequalities

Define the *nonincreasing rearrangement* of a pdf p to be another pdf p^* with the following two properties:

1. For all $t > 0$, $\#\{n : p_n > t\} = \#\{n : p_n^* > t\}$;
2. If $i \geq j$, then $p_i^* \leq p_j^*$.

The map $p \mapsto p^*$ is equivalent to an index permutation, which is unfortunately not uniquely defined in general. Note that nonincreasing rearrangements preserve additive information cost functions, as do all other index permutations:

$$H(p^*) = \sum_{i=1}^{\infty} f(p_i^*) = \sum_{i=1}^{\infty} f(p_i) = H(p). \tag{28}$$

Define the *partial sums* of a sequence p to be

$$Sp_n = \sum_{k=1}^{n} p_k; \qquad Sp_0 \stackrel{\text{def}}{=} 0. \tag{29}$$

If p is a pdf, then Sp takes values in the interval $[0,1]$, with $\lim_{n\to\infty} Sp_n = 1$. Of greatest interest is the sequence of partial sums of the nonincreasing rearrangement of a pdf, namely Sp^*. This is easily shown to be "concave" in the sense that $2Sp_j^* \geq Sp_{j-1}^* + Sp_{j+1}^*$, whenever all the indices are valid.

Suppose that p and q are two pdfs. It will be said that $Sp^* \geq Sq^*$ if $Sp_n^* \geq Sq_n^*$ for all n. In the case of finitely supported p and q, this corresponds to the notion of *majorization* defined in [7]. We will use the same notation for all pdfs, including those with infinite support. That is, we will say that p *majorizes* q, and write $p \succ q$ or equivalently $q \prec p$, if we mean that $Sp^* \geq Sq^*$. Majorization implies the following inequalities, through a standard summation-by-parts:

Lemma 3.2. *Suppose that ℓ is nonincreasing on $(0,1)$. If p and q are pdfs and $p \succ q$, then*

1.$\sum_i p_i \ell(p_i) \leq \sum_i q_i^* \ell(p_i^*)$, *and*
2.$\sum_i q_i \ell(q_i) \geq \sum_i p_i^* \ell(q_i^*)$.

■

It is well known (see [9], p. 278, for the standard proof) that if p and q are pdfs, then $\sum_k p_k \log(1/p_k) \leq \sum_k p_k \log(1/q_k)$, with equality if and only if $p_k = q_k$ for all k. Applying this with Lemma 3.2 to the nonincreasing rearrangements p^* and q^* yields $\sum_k p_k^* \log(1/p_k^*) \leq \sum_k p_k^* \log(1/q_k^*) \leq \sum_k q_k \log(1/q_k)$, which is a classical result:

Corollary 3.2. *If $f(t) = t \log(1/t)$, and p, q are two pdfs with $p \succ q$, then $H(p) \leq H(q)$.* ■

This argument depends on special properties of log, but the result generalizes to all additive information cost functions H, as we shall now show. We need some technical lemmas, whose straightforward proofs we omit:

Lemma 3.3. *Suppose that $f = f(t)$ is concave and nonnegative on $(0,1)$. Then the following are true:*

1.*The extension of f to $[0,1]$ defined by $f(0) = f(1) \stackrel{\text{def}}{=} 0$ is also concave and nonnegative.*

2.*If there exists a discrete pdf p such that $p_k > 0$ for all k and $\sum_k f(p_k) < \infty$, then $\lim_{x \to 0} f(x) = 0$.*

3.*If $\lim_{x \to 0} f(x) = 0$, then there exists some $\delta > 0$ such that f is nondecreasing on $[0, \delta]$.* ■

Lemma 3.4. *Suppose that $\{a_k\}$ and $\{h_k\}$ are real sequences satisfying the following conditions:*

1.*$\sum_{k=0}^{n} h_k \geq 0$ for all $n \geq 0$;*

2.*$\{a_k\}$ is nondecreasing, namely, $a_k \leq a_{k+1}$ for all $k \geq 0$;*

3.*$\lim_{n \to \infty} a_n (h_0 + h_1 + \cdots + h_n) = 0$.*

Then $\sum_k a_k h_k \leq 0$. ■

Corollary 3.3. *Suppose that $\{a_k\}$ and $\{h_k\}$ are real sequences satisfying the following conditions:*

1.*$\sum_{k=0}^{n} h_k \geq 0$ for all $n \geq 0$;*

2.*$\{a_k\}$ is nondecreasing, namely, $a_k \leq a_{k+1}$ for all $k \geq 0$;*

3.*$\lim_{n \to \infty} \sum_{k=0}^{n} h_k = 0$;*

4.*$\sum_{k=0}^{\infty} a_k h_k$ converges absolutely.*

Then $\sum_k a_k h_k \leq 0$. ■

We can now prove the main theorem of this section: the majorizing pdf always costs less.

Theorem 3.1. *If $p \succ q$, then $H(p) \leq H(q)$ for every additive information cost function H.*

Proof. It may be assumed without loss of generality that $p^* \neq q^*$ and $H(q) = H(q^*)$ is finite.

Put $m_k = \min\{p_k^*, q_k^*\}$ and $M_k = \max\{p_k^*, q_k^*\}$; then $0 \leq m_k \leq M_k \leq 1$. Since the sequences $\{m_k\}$ and $\{M_k\}$ are nonincreasing, $M_{k+1} \leq M_k$ and $m_{k+1} \leq m_k$. Since $p^* \neq q^*$, there must be a least integer j such that $p_j^* \neq q_j^*$, and thus $m_j \neq M_j$. Now let f be the concave function defining H by $H(p) = \sum_n f(p_n)$, and put

$$a_k = \begin{cases} \dfrac{f(M_j) - f(m_j)}{M_j - m_j}, & \text{if } k < j; \\ \dfrac{f(M_k) - f(m_k)}{M_k - m_k}, & \text{if } m_k < M_k; \\ a_{k'}, & \text{if } m_k = M_k \text{ and } k > j, \end{cases}$$

where k' is the greatest index less than k for which $m_{k'} < M_{k'}$. The sequence $\{a_k\}$ is thus well-defined, and by Lemma 3.1 is nondecreasing.

Put $h_k = p_k^* - q_k^*$; then $\sum_{k=0}^n h_k \geq 0$ for all $n \geq 0$, and $\lim_{n\to\infty} \sum_{k=0}^n h_n = 0$.

Now $f(p_k^*) = f(q_k^*) + a_k h_k$, since either $M_k = p_k^*$ and $m_k = q_k^*$ or else $M_k = q_k^*$ and $m_k = p_k^*$. Thus, $|a_k h_k| \leq f(p_k^*) + f(q_k^*)$, and $\sum_k a_k h_k$ will converge absolutely if both $H(p) < \infty$ and $H(q) < \infty$. The latter is true by assumption, and the former is a consequence of the following lemma:

Lemma 3.5. *Suppose H is an additive information cost function. If $p \succ q$ and $H(q) < \infty$, then $H(p) < \infty$.*

Proof. Without loss of generality, we may assume that $q_k > 0$ for all k. Using Lemma 3.3, we can rewrite $H(p)$ as follows:

$$\sum_k f(p_k) = \sum_{k \leq m} f(p_k^*) + \sum_{\substack{k > m \\ p_k^* \leq q_k^*}} f(p_k^*) + \sum_{\substack{k > m \\ p_k^* > q_k^*}} f(p_k^*).$$

The first two sums are finite. To show that the third sum is finite, we apply the estimate $f(p_n^*) < p_n^* \dfrac{f(q_n^*)}{q_n^*}$, Lemma 3.1, and summation by parts. ■

To complete the proof of the theorem, consider the finite sums

$$\sum_{k=0}^n f(p_k^*) = \sum_{k=0}^n f(q_k^*) + \sum_{k=0}^n a_k h_k.$$

All sums converge absolutely, so $\sum_{k=0}^\infty a_k h_k \leq 0$ by Corollary 3.3. We conclude that $H(p) = \sum_{k=0}^\infty f(p_k^*) \leq \sum_{k=0}^\infty f(q_k^*) = H(q)$. ■

Remark. Theorem 3.1 was proved for *finite* sequences and arbitrary concave f by Hardy, Littlewood and Pólya [4]. Given $p \succ q$, they used Muirhead's algorithm to find a doubly stochastic matrix A such that $q = Ap$, then applied Proposition 2.. The proof is constructive and builds A in a number of steps not greater than the lengths of p and q, but we know of no proof that Muirhead's algorithm works in the case of infinitely supported p and q. Instead, our proof avoids using doubly stochastic operators to characterize majorization. In a subsequent search of the literature, we found a paper by Fuchs [3], in which the idea we use was applied to the simpler case of finite sequences, where no question of convergence arises.

Hardy, Littlewood and Pólya [4] also proved a partial converse to Theorem 3.1. A pdf which always measures least in cost majorizes all others in the collection:

THEOREM 3.2. *If $\{p_1, \dots, p_n\}$ and $\{q_1, \dots, q_n\}$ are finite pdfs, and $\sum_{k=1}^n f(p_k) \leq \sum_{k=1}^n f(q_k)$ for all concave functions f, then $p \succ q$.* ■

Actually, their converse only requires that $\sum_{k=1}^n f(p_k) \leq \sum_{k=1}^n f(q_k)$ for a sufficiently large subclass of concave functions f. In particular, if the inequality holds for $f_T(t) = t - [t-T]_+$ for all $0 \leq T \leq 1$, then $p \succ q$. The same subclass works if the sequences are infinite. We may even use somewhat weaker hypotheses to prove our converse to Theorem 3.1. For $T > 0$ and $a > b \geq 0$, define a function $h_{T,a,b} : [0,\infty) \to [0,\infty)$ by

$$h_{T,a,b}(t) = at + (b-a)[t-T]_+. \tag{30}$$

This function is continuous, concave, nondecreasing and piecewise linear, with slope a from $h_{T,a,b}(0) = 0$ to $h_{T,a,b}(T) = aT$, and smaller slope b thereafter.

LEMMA 3.6. *If, for every $T \in (0,1)$, there exist $a = a(T)$ and $b = b(T)$, with $a > b > 0$, such that*

$$\sum_n h_{T,a,b}(p_n) \leq \sum_n h_{T,a,b}(q_n),$$

then $p \succ q$. ■

The proof is a slight modification of the one in [4], so we omit it. An immediate consequence is our converse:

THEOREM 3.3. *If $H(p) \leq H(q)$ for every additive information cost function H, then $p \succ q$.* ■

It is natural to examine other subsets of the concave nonnegative functions, to see if they can replace the "threshold" functions $h_{T,a,b}$ as additive information cost functions that imply majorization. One such class is the functions $f(t) = t^\alpha$, where $\alpha \in (0,1)$, which were studied in the first section. Unfortunately, not all classes of costs suffice, as there is the following negative result:

LEMMA 3.7. *There exist two pdfs p and q such that $\sum_n p_n^\alpha \leq \sum_n q_n^\alpha$ for all $0 < \alpha < 1$, yet $p \not\succ q$.*

Proof. We first establish that, for sufficiently small $1 > \epsilon > 0$, the following inequality holds for all $0 \leq \alpha \leq 1$:

$$h(\alpha) \overset{\text{def}}{=} (1+\epsilon)^\alpha + 2(\frac{1}{2})^\alpha(1-\epsilon)^\alpha - 2 \geq 0. \tag{31}$$

Since $h(0) = 1$, $h(1) = 0$, and h is differentiable on $(0,1)$, it suffices to show that $h'(\alpha) \leq 0$ on $(0,1)$. But

$$\begin{aligned} h'(\alpha) &= (1+\epsilon)^\alpha \log(1+\epsilon) + 2(\frac{1}{2})^\alpha (1-\epsilon)^\alpha \log\left(\frac{1-\epsilon}{2}\right) \\ &\leq (1+\epsilon)\log(1+\epsilon) + (1-\epsilon)\log\left(\frac{1-\epsilon}{2}\right) \quad \overset{\mathrm{def}}{=} d(\epsilon), \end{aligned} \tag{32}$$

since $\log(1+\epsilon) > 0$ and $(1+\epsilon) > (1+\epsilon)^\alpha$, while $\log(\frac{1-\epsilon}{2}) < 0$ and $(1-\epsilon) < 2(\frac{1}{2})^\alpha(1-\epsilon)^\alpha$. But d is continuous and $d(0) = \log\frac{1}{2} < 0$, so $d(\epsilon) < 0$ for all sufficiently small $\epsilon > 0$.

With Inequality 31 established, we construct the counterexample pdfs. Let $\epsilon > 0$ satisfy 31, and define $a = \frac{1}{2}(1+\epsilon)$ and $b = \frac{1}{2}(1-a) = \frac{1}{2}(\frac{1-\epsilon}{2})$. Then $a+b+b = 1$, so $p = (a, b, b, 0, 0, \ldots)$ is a (finitely-supported) pdf. Furthermore, $a > b$, so $p = p^*$. Let $q = (\frac{1}{2}, \frac{1}{2}, 0, 0, \ldots)$; this is another nonincreasing, finitely-supported pdf. Since $a > \frac{1}{2}$ but $a + b < \frac{1}{2} + \frac{1}{2}$, neither $p \succ q$ nor $q \succ p$. However, for any $\alpha \in (0,1)$,

$$\begin{aligned} \sum_n p_n^\alpha - \sum_n q_n^\alpha &= a^\alpha + 2b^\alpha - 2(\frac{1}{2})^\alpha \\ &= \left(\frac{1}{2}\right)^\alpha (1+\epsilon)^\alpha + 2\left(\frac{1}{2}\right)^\alpha \left(\frac{1-\epsilon}{2}\right)^\alpha - 2\left(\frac{1}{2}\right)^\alpha \\ &= \left(\frac{1}{2}\right)^\alpha h(\alpha) \quad \geq 0. \end{aligned}$$

■

4. CONSEQUENCES FOR THE BEST-BASIS ALGORITHM

Let $\mathcal{B}$ be a *library*, or collection of orthonormal bases for a separable Hilbert space X with norm $\|\cdot\|$. A vector $x \in X$, expanded in a basis $B = \{b_n \in X : n = 0, 1, 2, \ldots\}$ of $\mathcal{B}$, is represented by a sequence $\{c_n : n = 0, 1, 2, \ldots\}$ of expansion coefficients:

$$x = \sum_{n=0}^{\infty} c_n b_n \quad \text{in the sense that} \quad \lim_{N\to\infty} \|x - \sum_{n=0}^{N} c_n b_n\| = 0. \tag{33}$$

Orthonormality implies that $\|x\|^2 = \sum_n |c_n|^2$, so the sequence $p = p(x, B)$ defined by $p_n = |c_n|^2/\|x\|^2$, $n = 0, 1, 2, \ldots$, is a discrete probability density function, or pdf, associated to x and the basis B.

Let p^* be the nonincreasing rearrangement of p as defined in Section 3.2 above, and write $\{c_n^*\}$ and $\{b_n^*\}$ for the corresponding rearrangements of the expansion coefficients and basis vectors, respectively. Then majorization can be used to compare rates of approximation in X by truncated expansions: if $p(x, B) \succ p(x, B')$, then for every $N = 0, 1, 2, \ldots$, $\sum_{n=0}^{N} c_n^* b_n^*$ is a better approximation to x than $\sum_{n=0}^{N} c_n'^* b_n'^*$.

A *best basis* $B \in \mathcal{B}$ for a fixed x is one satisfying $p(x, B) \succ p(x, B')$ for any $B' \in \mathcal{B}$. It evidently gives fastest approximation in norm by partial sums $\sum_{n=0}^{N} c_n^* b_n^*$. One way to achieve data compression is to describe $x \in X$ using just the largest expansion coefficients $\{c_n^* : n = 0, \ldots, N\}$ in its best basis, plus a code defining the basis.

Wavelet packet bases constructed from a finite-depth multi-resolution analysis of X form a discrete, in fact finite, library $\mathcal{B}$ whose members are the many combinations of relatively few pieces. With decomposability comes a low-complexity divide-and-conquer algorithm for finding the minimizing basis for a fixed information cost function H, and also for coding it [2]. Reference [9], pages 310ff, describes the wavelet packet algorithm in detail.

By Theorem 3.1, minimizing any single H locates the sole candidate for best basis. Since $H(p) = H(p^*)$, this candidate can be identified without rearrangement. By Theorem 3.3, that candidate is in fact a best basis if it minimizes sufficiently many information cost functions.

In this section, we prove that all information costs for a pdf p are bracketed between two values that depend only on $\sum_k p_k^2$ and $\sum p_k^{\frac{2}{3}}$. The latter is an information cost function, so it is minimal at the candidate majorizer p. We then apply the result to wavelet packet libraries, to obtain an algorithm to decide when a candidate basis is a best basis. Whether it exists depends on $\mathcal{B}$ and x, but the exact conditions are yet to be found. We obtain a partial answer, concluding that the cheapest pdf of a discrete set, determined by a single information cost function, must be cheapest for all the $h_{T,a,b}$ cost functions of Eq. 30 with T in an open interval. In special cases, this allows us to deduce from a single cost evaluation that the minimal pdf majorizes.

4.1. Legendre transforms

To apply the results of Section 2, we require convexity. Given a real-valued function $\ell = \ell(t)$ on an open interval I, define the *Legendre transform* of ℓ as follows:

$$\tilde{\ell}(t) = \sup\{at + b : \forall s \in I, as + b \leq \ell(s)\}. \tag{34}$$

We put $\tilde{\ell}(t) = -\infty$ if the set is empty. This has several well-known basic properties:

LEMMA 4.1. *Either $\tilde{\ell}(t) = -\infty$ for all $t \in I$, or else $\tilde{\ell}$ is finite and convex on I, and satisfies $-\infty \leq \tilde{\ell}(t) \leq \ell(t)$ for all $t \in I$. Furthermore, $\tilde{\ell}$ is the greatest convex function below ℓ, in the sense that if $c = c(t)$ is convex and $c(t) \leq \ell(t)$ for all $t \in I$, then $c(t) \leq \tilde{\ell}(t)$ for all $t \in I$.* ■

Note that if ℓ is convex, then $\tilde{\ell} = \ell$.

LEMMA 4.2. *If $\ell = \ell(t)$ is nonnegative and nonincreasing on an interval I, then $\tilde{\ell}$ is also nonnegative and nonincreasing on I. Furthermore, if the left endpoint of I is finite and $\ell(t)$ is positive at some $t \in I$, then $\tilde{\ell}$ is not identically zero.* ■

4.2. Comparability of information costs

Using the Legendre transform $\tilde{\ell}$, we can generalize the results of Section 2 to get a pair of inequalities bracketing any information cost function $H(\ell, p)$, regardless of the convexity or subexponentiality of ℓ:

Theorem 4.1. *Let H be any additive information cost function determined by concave nonnegative $f = f(t)$, and put $\ell(t) = f(t)/t$ for $0 < t < 1$. For any $d \in (0,1)$, and any pdf p which is $1-d$-summable, we have the inequalities*

$$\tilde{\ell}\left(\sum_n p_n^2\right) \leq H(\ell, p) \leq -\frac{1}{d}\tilde{\ell}\left(\sum_n p_n^{1-d}\right),$$

where $\tilde{\ell}$ is the Legendre transform of ℓ on $(0,1)$ and the Legendre transform of the $-d$-subexponential extension of ℓ on $(1,\infty)$.

Proof. Zeros in the sequence $\{p_n\}$ can be ignored in all sums, and if $p_n = 1$ for a unique n, then the sums each reduce to a single term, and the inequalities follow from the definitions of ℓ and $\tilde{\ell}$:

$$\tilde{\ell}(1-) \leq \ell(1-) = -\frac{1}{d}\ell((1-)^{-d}) = -\frac{1}{d}\ell(1+) \leq -\frac{1}{d}\tilde{\ell}(1+).$$

This ordering also holds for the continuous extensions of $\ell(t)$ and $\tilde{\ell}(t)$ to $t = 1$, and also to the extensions $\ell(1) = \tilde{\ell}(1) \overset{\text{def}}{=} 0$. We may thus assume without loss of generality that $0 < p_n < 1$ for all n.

To get the lower bound, we use Lemmas 4.1 and 4.2 to conclude that $\tilde{\ell}$ is finite, nonnegative, and convex on $(0,1)$. Then, by Proposition 2.5,

$$0 \leq \tilde{\ell}\left(\sum_n p_n^2\right) \leq H(\ell, p).$$

For the upper bound, first note that $1 < \sum_n p_n^{1-d} < \infty$. The $-d$-subexponential extension of ℓ to $1 < s < \infty$ is defined by $\ell(s) = -d\ell(s^{-1/d})$, so as to satisfy $\ell(t^{-d}) = -d\ell(t)$ for $t \in (0,1)$. Let $\tilde{\ell}$ be the Legendre transform of this ℓ on $(1,\infty)$ and put

$$\ell_1(t) \overset{\text{def}}{=} \begin{cases} \ell(t), & \text{if } t \in (0,1); \\ \tilde{\ell}(t), & \text{if } t \in (1,\infty). \end{cases}$$

By Lemma 4.1, there are two possibilities. Either $\tilde{\ell}(t) = -\infty$ for $t > 1$, and the upper bound holds trivially. Otherwise, $\ell_1(t)$ is finite and convex for all $t > 1$, and so there must exist a, b such that $\ell_1(s) \geq as + b$ for all $s > 1$. The assumption that $\sum_n p^{1-d}$ is finite gives a finite lower bound $\sum_n p_n \ell_1(p_n^{-d}) \geq \sum_n p_n[ap_n^{-d} + b] = b + a\sum_n p^{1-d}$, so the series $\sum_n p_n \ell_1(p_n^{-d})$, in which all the terms are negative, must converge monotonically to a finite sum. Finally, since ℓ_1 satisfies $\ell_1(t^{-d}) = \tilde{\ell}(t^{-d}) \leq \ell(t^{-d}) = -d\ell(t) = -d\ell_1(t)$, it is $-d$-subexponential on $(1,\infty)$, and we get the upper bound,

$$H(\ell, p) \leq -\frac{1}{d}\ell_1\left(\sum_n p_n^{1-d}\right) = -\frac{1}{d}\tilde{\ell}\left(\sum_n p_n^{1-d}\right),$$

by applying Proposition 2.7. ■

An infinite upper bound is possible, as shown by Eq. 17: for $\alpha \in (0,1)$ and $\ell(t) = t^{-\alpha} - 1$ on $0 < t < 1$, any $-d$-subexponential extension must satisfy $\ell(x) \leq$

$-d\left[x^{\alpha/d}-1\right]$ on $1<x<\infty$. It is easy to check that if $d<\alpha$, then $\tilde{\ell}(t)=-\infty$ for all $t\in(1,\infty)$.

4.3. Example application: wavelet packet best bases

The main fact that we need here is that wavelet packet bases form a discrete, indeed finite, subset of the orthonormal bases of X.

We specialize Theorem 4.1 with the example class of information cost functions with $f(t)=h_{T,a,b}(t)$ as in Eq. 30. Then $\ell(t)=f(t)/t$ is nonnegative and nonincreasing:

$$\ell(t)=\begin{cases} a, & \text{if } 0\le t<T,\\ b+(a-b)T/t, & \text{if } T\le t\le 1.\end{cases} \tag{35}$$

The lower bound function comes from the Legendre transform on $(0,1)$, which by Lemma 4.2 is nonincreasing. We compute it explicitly:

$$\tilde{\ell}(t)=\begin{cases} a-(1-T)(a-b)t, & \text{if } 0\le t\le 1\le 2T,\\ a-\frac{1}{4T}(a-b)t, & \text{if } 0\le t\le 2T<1,\\ \ell(t), & \text{if } 2T<t\le 1.\end{cases} \tag{36}$$

The $-d$-subexponential extension of ℓ to $(1,\infty)$ will be

$$\ell(s)=\begin{cases} -d[b+(a-b)Ts^{1/d}], & \text{if } 1\le s\le T^{-d},\\ -da, & \text{if } T^{-d}<s<\infty,\end{cases} \tag{37}$$

and the upper bound based on its Legendre transform will therefore be piecewise linear and nondecreasing:

$$-\frac{1}{d}\tilde{\ell}(s)=\begin{cases} b+(a-b)T+\frac{(1-T)(a-b)}{T^{-d}-1}(s-1), & \text{if } 1\le s\le T^{-d},\\ a, & \text{if } T^{-d}<s<\infty.\end{cases} \tag{38}$$

Now fix $d=\frac{1}{3}$, so that Theorem 4.1 becomes

$$\tilde{\ell}\left(\sum_n p_n^2\right)\le H(\ell,p)\le -3\tilde{\ell}\left(\sum_n p_n^{2/3}\right). \tag{39}$$

Suppose that p and q are $1-d=2/3$-summable pdfs, with $\sum_k p_k^{\frac{2}{3}}\overset{\text{def}}{=}x\ge 1$ and $\sum_k q_k^{\frac{2}{3}}=x+\delta$ for some $\delta>0$. To show that $H(\ell,p)\le H(\ell,q)$ for all ℓ in some class, it suffices to show that

$$-3\tilde{\ell}\left(\sum_k p_k^{\frac{2}{3}}\right)\le\tilde{\ell}\left(\sum_k q_k^2\right). \tag{40}$$

But Hölder's inequality implies that $1=\left(\sum q_k^2\right)\left(\sum q_k^{2/3}\right)^3$, so $\sum q_k^2=(x+\delta)^{-3}$. It therefore suffices to show that

$$-3\tilde{\ell}(x)\le\tilde{\ell}\left((x+\delta)^{-3}\right), \tag{41}$$

for all ℓ in the class. But for fixed x and δ, this holds for all information cost functions $t\ell(t) = h_{T,a,b}(t)$ with $(x+\delta)^{-3} \leq T \leq x^{-3}$, since then

$$\frac{T}{(x+\delta)^{-3}} \geq 1 = T + (1-T) \geq T + \frac{1-T}{T^{-1/3}-1}(x-1).$$

Multiplying both sides by $a-b>0$ and adding b gives the result.

In a discrete library, there will always be some separation $\delta > 0$ between the minimum value $H(p)$ and the next lowest value $H(q)$, so there will always be a nonempty open interval $\mathcal{T}$ of values T such that all information cost functions $h_{T,a,b}$ with $T \in \mathcal{T}$ are cheapest at p.

4.4. Sharpness of the result

We might ask whether it is possible to find a better upper bound than the one in Theorem 4.1, one that avoids subexponential extensions and thus has a simpler dependence on d, and is always finite if $H(\ell, p)$ is finite. For example, is it possible to have an estimate of the form

$$H(\ell,p) \leq C\ell \circ f\left(\sum_n p_n^\beta\right), \tag{42}$$

where $f : \mathbf{R} \to (0,1)$ satisfies $0 < r \leq f(t) < 1$ for all $t \in [\frac{1}{2}, 2]$, and C and β are some fixed positive numbers. The idea is to map $\sum_n p_n^\beta$ back into the domain $(0,1)$ of ℓ, while making sure the upper bound avoids the potential infinity at $\ell(0)$ at least for $\beta \sim 1$. But no such estimate can hold for all concavable nonnegative nonincreasing ℓ, as the following shows:

Lemma 4.3. *Inequality 42 must fail for $\ell(t) = t^{-\alpha} - 1$ with some $\alpha \in (0,1)$ and some pdf p.*

Proof. Let $\{\epsilon_j : j = 1, 2, \ldots\} \subset (0,1)$ be a decreasing sequence satisfying

$$(1-\epsilon_j)^\beta \geq \frac{1}{2}; \qquad j^{1-\beta}\epsilon_j^\beta \leq 1, \quad \text{for all } j = 1, 2, \ldots \tag{43}$$

Define p^j by

$$p_1^j = 1-\epsilon_j; \qquad p_2^j = p_3^j = \cdots = p_{j+1}^j = \frac{\epsilon_j}{j}; \qquad p_n^j = 0, \text{ for } n > j+1.$$

Then p^j is a pdf, and for it and the given ℓ,

$$H(\ell, p^j) = (1-\epsilon_j)^{1-\alpha} + j^\alpha \epsilon_j^{1-\alpha} - 1.$$

On the other hand,

$$X \overset{\text{def}}{=} \sum_n \left[p_n^j\right]^\beta = (1-\epsilon_j)^\beta + j^{1-\beta}\epsilon_j^\beta.$$

Eq. 43 implies that $\frac{1}{2} \leq X \leq 2$, so $0 < r \leq f(X) < 1$. Since ℓ is nondecreasing, the right-hand side of Inequality 42 is smaller than $C\ell(r)$, which is uniformly bounded above by $\frac{C}{r}$ for all $\alpha \in (0,1)$. Thus, taking $\alpha_j \geq \frac{1}{2}$ sufficiently close to 1 so that $\epsilon_j^{1-\alpha_j} \geq \frac{1}{2}$, and using this α_j to define ℓ^j, we have

$$H(\ell^j, p^j) \geq (-1) + \frac{1}{2}\sqrt{j},$$

which increases without bound as $j \to \infty$. ■

Since $0 \leq \tilde{\ell} \leq \ell$, Inequality 42 cannot be made to hold by using Legendre transforms, either.

REFERENCES

1. J. Aczél and Z. Daróczy. "On Measures of Information and Their Characterizations," Academic Press, New York, 1975.
2. R. R. Coifman and M. V. Wickerhauser. Entropy based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 32:712–718, March 1992.
3. L. Fuchs. A new proof of an inequality of Hardy–Littlewood–Pólya. *Mat. Tidsskr.*, B:53–54, 1947.
4. G. H. Hardy, J. E. Littlewood, and G. Pólya. Some simple inequalities satisfied by convex functions. *Messenger of Mathematics*, 58:145–152, 1929.
5. G. H. Hardy, J. E. Littlewood, and G. Pólya. "Inequalities," Cambridge University Press, Cambridge, England, 1959.
6. A. S. Markus. Eigenvalues and singular values of the sum and product of linear operators. *Uspehi Mat. Nauk*, 19(4):93–123, 1964. Russian Math. Surveys 19:91–120, 1964.
7. A. W. Marshall and I. Olkin. "Inequalities: Theory of Majorization and Its Applications," Number 143 in Mathematics in Science and Engineering. Academic Press, New York, 1979.
8. A. Rényi. "Wahrscheinlichkeitsrechnung," Veb Deutscher Verlag der Wissenschaften, Berlin, 1962.
9. M. V. Wickerhauser. "Adapted Wavelet Analysis from Theory to Software," AK Peters, Ltd., Wellesley, Massachusetts, 1994.