Entropy **2010**, *12*, 338-374; doi:10.3390/e12030338

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Estimation of an Entropy-based Functional

Brian N. Maurizi

7576 Dale Ave., Saint Louis, Missouri 63117, USA; E-Mail: brian.maurizi@gmail.com;
Tel.: (314)-308-5393

Received: 30 December 2009; in revised form: 8 February 2010 / Accepted: 24 February 2010 / Published: 3 March 2010

Abstract: Given a function f from $[0,1]$ to the real line, we consider the (nonlinear) functional $\mathtt{h}$ obtained by evaluating the continuous entropy of the "density function" of f. Motivated by an application in signal processing, we wish to estimate $\mathtt{h}(f)$. Our main tool is a decomposition of $\mathtt{h}$ into two terms, which each have favorable scaling properties. We show that, if functions f and g satisfy a regularity condition, then the smallness of $\|f-g\|_\infty$ and $\|f'-g'\|_\infty$, along with some basic control on derivatives of f and g, is sufficient to imply that $\mathtt{h}(f)$ and $\mathtt{h}(g)$ are close.

Keywords: entropy; differential entropy; Shannon entropy; entropy estimation; nonlinear functional; signal processing

---

1. Introduction

We define the continuous entropy of a probability density ρ to be

$$\mathtt{h}(\rho) = -\int_{u\in\mathbb{R}} \rho(u)\log[\rho(u)]du$$

(where the base of the logarithm is $e = 2.71828...$). Stripped of its interpretation, continuous entropy is just a particular way of measuring how "spread out" a probability density is. In this paper, we are motivated by a novel application of this measurement in signal processing.

An ultrasound probe generates a short acoustic pulse which travels through the medium of interest (in medical applications, this medium is the tissue of the patient). As the pulse travels through the medium, features within the medium cause some of the pulse to be reflected back towards the probe,

and the strength of the reflection contains information about the feature at that location. The signal of interest is the intensity of this reflected pulse that arrives back at the probe over time. This signal can be divided into many short windows of time; after re-scaling the time axis, one of these time windows can be represented by the interval $[0, 1]$, and the signal over that time window can be represented by a real-valued function f on the interval $[0, 1]$. So, $f(t)$ represents the intensity of the reflected pulse that arrives back at the probe at the time $t \in [0, 1]$.

The ultrasound probe can only measure $f(t)$ at a discrete set of values of t, and this measurement can be corrupted in various ways. Therefore, one step in the processing of an ultrasound signal is to reconstruct f from the measurements of the ultrasound probe. Then, a functional is applied to f to obtain a single number, the idea being that this number contains all the relevant information about the reflected signal over that particular window in time. A standard functional in the industry is the "energy" in the signal, $\int_0^1 |f|^2$, or (more often) its logarithm. However, there is a series of papers by Hughes and others ([1–7]) demonstrating the utility of using the continuous entropy of the "density function" of f as the functional, instead of the log-energy.

Suppose that f is a real-valued function on $[0, 1]$. By "density function" of f, we mean the following: Define the measure μ_f on $\mathbb{R}$ by $\mu_f(E) = |f^{-1}(E)|$, where $|A|$ denotes the Lebesgue measure of the set $A \subset [0, 1]$. Suppose that μ_f is given by the density ρ_f, *i.e.*,

$$\mu_f(E) = \int_E \rho_f$$

This density ρ_f is the "density function" of f. We can think of $[0, 1]$ as a probability space and f as a function on that probability space (*i.e.*, a random variable) and then ρ_f is literally the density of the "random variable" f.

Another way to think of ρ_f is as follows: Suppose that we consider the uniform distribution $U(0, 1)$. Suppose we choose a point t at random from $[0, 1]$ according to $U(0, 1)$, and we calculate $f(t)$. The density ρ_f gives the relative frequency of different outcomes of this experiment. Also, intuitively we can say that

$$\rho_f(u) = \sum_{\{t: f(t)=u\}} \frac{1}{|f'(t)|}$$

(although we will not attempt to formalize this definition for general f). Let us abuse notation and write

$$\mathtt{h}(f) = \mathtt{h}(\rho_f)$$

The functional of interest is $f \to \mathtt{h}(f)$.

The effectiveness of using the $\mathtt{h}$ functional has been shown in several different settings in ultrasound signal processing, ranging from defect detection in plexiglass [1] to detection of contrast agents in the tissue of live animals ([5,7]). In some situations, the $\mathtt{h}$ functional produced images of objects which, because the material was engineered by the experimenter, were known to be present but were not detected by the log-energy functional.

The interpretation of what is being measured by $\mathtt{h}(f)$ in this context is not yet clear. However, given the utility of the technique, we believe the area merits investigation. We would like to answer the question

"How well can we estimate $\mathtt{h}(f)$?"

In a real-world setting we will perhaps receive some samples of f corrupted by noise, and we would like to estimate $\mathtt{h}(f)$ (as we described, the processing of an ultrasound signal is an example of this). Therefore, we seek a solution to the following problem, where the "regularity conditions" and the form of the "partial information" is not specified:

Estimation Problem 1. *Let f be a function on $[0,1]$ that satisfies some regularity conditions, and suppose we are given some "partial information" on f, such as samples corrupted by noise. We wish to estimate $\mathtt{h}(f)$ with an estimator $\hat{\mathtt{h}}$, and give a quantitative bound on the error $|\mathtt{h}(f) - \hat{\mathtt{h}}|$, such that the bound is invariant under the scaling $f \to \lambda f$ for $\lambda \neq 0$.*

We seek this scale invariance because, if $\hat{\mathtt{h}}$ is any reasonable estimator of $\mathtt{h}(f)$, the true difference $\mathtt{h}(f) - \hat{\mathtt{h}}$ does not change under the scaling $f \to \lambda f$ for $\lambda \neq 0$, so we do not want our bound on that difference to change under this scaling either. (This scaling issue is discussed more in section 5.).

Our main result will not give a specific prescription for solving this problem. Instead, we reduce it to a much more tractable problem, namely the problem of taking noisy samples from f and producing an approximating function g which is "close" to f in some quantifiable way.

However, our concern at the moment is to relate problem 1 to the existing literature, and we notice that problem 1 is similar to the standard entropy estimation problem (for an overview of this problem, see Beirlant *et al.* [8]):

Estimation Problem 2 (Entropy Estimation)**.** *Let X be a random variable with unknown density function ρ. We make N independent observations of X. From these, we wish to estimate the continuous entropy of ρ.*

There is a large literature concerning the entropy estimation problem, so one might hope to solve problem 1 by applying some of the standard methods in the field to the density ρ_f. However, this is not what we will do. Since we are departing from traditional methods, we would like to provide some indication that new methods may in fact be necessary; before working to develop new methods to solve a problem, one would like to know whether the problem is obviously solved with readily available standard methods. In section 2., we review the current literature with regard to how it applies to densities such as ρ_f. This review is not comprehensive; the purpose is only to show that the density ρ_f presents some difficulty for current methods of estimation of $\mathtt{h}(\rho)$.

Some of the difficulties that we find, stated in general terms, are the following. First, the methods we have seen often involve intermediate parameters that truncate the tails of integrals, truncate the unbounded pieces of ρ, or perform other types of operations that enable a splitting into "good" and "bad" regions. In view of the scale-invariance that we seek, these intermediate parameters will all become involved and their rates of growth will need to be related in a way that produces the desired result and remains scale-invariant. This task, for a density such as ρ_f, seems to be non-trivial. Second, many of the results which might apply to ρ_f use methods which are non-quantitative: theorems from measure theory such as Lebesgue Differentiation, the Lebesgue Dominated Convergence Theorem, Egoroff's Theorem, and the Borel-Cantelli Lemma. Replacing these theorems with quantitative estimates will introduce more difficulties. Finally, and perhaps most importantly from an intuitive point of view, there seems to be a general progression from "easy" to "hard" as the assumptions on ρ become less stringent, moving from

differentiable to continuous to bounded to L^2, and so on. The density ρ_f seems to fall on the "hard" end of this continuum. This does not correspond to the progression of difficulty that a practitioner of signal processing would expect to see when they attempt to solve problem 1. Even an extremely well behaved function (such as $f(t) = \sin(2\pi t)$) will produce a density ρ_f that will be ranked as "badly behaved." A signal processor would expect that the problem would be "easy" for well behaved functions f, and get harder as the function f becomes more badly behaved.

In short, the readily available methods that we are aware of do not seem appropriate for the problem we are presented with. Fortunately, we believe that estimating $\mathtt{h}(f)$ is easier than solving the entropy estimation problem for ρ_f, for the simple reason that we are able to take advantage of the "good behavior" of f in the time domain, instead of only having access to the "bad behavior" of the density ρ_f. For example, suppose that in a real-world setting we receive sampled values of f at specific values of t, say $\{t_j\}$ (not to be confused with the values of t where f actually has critical points, which later will be referred to as $\{t_i\}$). Neglecting noise (purely for the simplicity of this explanation), we receive the data $\{(t_j, f(t_j))\}$. We intend to use *all* the information available to us, *i.e.* the time series $\{(t_j, f(t_j))\}$, not just the (unordered) values $\{f(t_j)\}$. For instance, we may want to estimate f' or f'', which will certainly make use of the indicated time series. We will *not* attempt to apply a standard entropy estimation technique directly to the values $\{f(t_j)\}$.

One strategy would be to first construct an approximating function g (using the observations $\{(t_j, f(t_j))\}$), and then calculate $\mathtt{h}(g)$ and use this as our estimate of $\mathtt{h}(f)$. In Hughes *et al.* [7], this method is applied to real-time imaging calculations in a laboratory setting; instead of using just the values $\{f(t_j)\}$ directly, they first construct an approximating function g and then calculate the desired entropy value using g. In [7], they are estimating a variant of the Rényi entropy of ρ_f, not $\mathtt{h}(f)$ as we have defined it, but the same authors have demonstrated the utility of $\mathtt{h}(f)$. Incidentally, the quantity that is eventually computed in [7] is

$$\sum_i \frac{1}{|g''(s_i)|}$$

where the $\{s_i\}$ are the critical points of g, and the authors comment [7] that "while this involves use of the second derivative of $f(t)$ at its critical points, which can be expected to increase noise in the processing chain output, surprisingly the resulting signal processing scheme does not sacrifice sensitivity." This gives some assurance that our result, which requires control on multiple derivatives of f, is not without practical application.

Our goal in this paper is to estimate the difference $\mathtt{h}(f) - \mathtt{h}(g)$ in terms of quantities involving f and g (not ρ_f or ρ_g). As we mentioned, this reduces the problem at hand, problem 1, to a much more tractable question of function approximation given noisy data. We note, for completeness, that a specific prescription for solving problem 1, using methods similar to our methods here, was given in Maurizi [9], but we believe that the approach we present now will be clearer than the approach given in [9].

As mentioned, in Section 2., we review the current literature with regard to how it applies to densities such as ρ_f. After some definitions and notation in Section 3., we will highlight a useful identity in section 4. which may be of independent interest, and we discuss some issues regarding the scaling $f \to \lambda f$ for $\lambda \neq 0$ in Section 5.. In Section 6. we present a proposition showing why the regularity assumption (Definition 6) we introduce is necessary at least in some form, if not the specific form given here. Our

regularity assumption on f and g, defined precisely in Section 7., essentially prevents f and g from becoming "too flat" by preventing their first and second derivatives from simultaneously vanishing, this is quantified with a parameter δ.

Section 7. contains our main result, Theorem 1, which states that if f and g satisfy the regularity assumption with parameter δ, then the difference $\mathtt{h}(f) - \mathtt{h}(g)$ can be bounded by the metrics $\|f - g\|_\infty, \|f' - g'\|_\infty$, and the value δ, along with factors that control the overall sizes of f and g and their derivatives.

The proof of the main result is outlined in Section 8. and carried out in the subsequent sections.

2. Background

We now look to the current literature, seeking a method to estimate $\mathtt{h}(f)$ with quantitative bounds on the error, in a manner that is invariant under the scaling $f \to \lambda f$ for $\lambda \neq 0$.

The first sign that new methods may be necessary is that, from the viewpoint of entropy estimation, the densities ρ_f are very badly behaved. A non-degenerate critical point in f will induce an asymptote of the form $1/\sqrt{x}$ in ρ_f, so we are in general dealing with density functions that are unbounded, not L^2, and have discontinuities in the interior of their support. These properties, and other features of the density ρ_f, are discussed in [6,7].

Many of the standard results in entropy estimation do not apply to such badly behaved densities. Several theorems in entropy estimation (such as Goria *et al.* [10], van Es [11], Joe [12], Levit [13], and more recently Leonenko *et al.* [14]) use the assumption that ρ is bounded. In Tsybakov and van der Meulen [15] and Eggermont and LaRiccia [16], the assumption that ρ is twice differentiable was needed. In Dmitriev and Tarasenko [17] and Ahmad and Lin [18], ρ was assumed to have a bounded derivative. In Hall and Morton [19] ρ was assumed to have a continuous derivative. In Mokkadem [20], a "distributional" equivalent of $\rho'' \in L^1$ was required.

There are methods of solving problem 2 that could shed light on densities such as ρ_f. In Kozachenko and Leonenko [21], Vasicek [22], and Györfi and van der Meulen [23], methods are developed which solve the entropy estimation problem 2 defined above and impose only mild restrictions on ρ.

In Kozachenko and Leonenko [21], they consider densities on $\mathbb{R}^m$, for simplicity we will consider the case $m = 1$. Note that they use different notation; we will continue to use ρ to stand for the density function, whereas they use f. A nearest-neighbor estimator is used, and only fairly mild conditions are imposed on the density ρ: For some $\epsilon > 0$, the following two equations must hold:

$$\int |\log \rho(x)|^{1+\epsilon} \rho(x) dx < \infty \tag{1}$$

$$\int |\log |x - y|\,|^{1+\epsilon} \rho(x)\rho(y) dx dy < \infty \tag{2}$$

(these are equations numbered (3) and (4) in [21]). Densities such as our ρ_f will "typically" satisfy these constraints (for instance, the function $f(t) = t^2$ on $[0, 1]$ has density $\rho_f(u) = (1/2)u^{-1/2}, u \in (0, 1]$ and this density satisfies these constraints). The authors then prove that the estimator $\mathtt{h}_N$ computed from N

independent samples of ρ satisfies $\mathbb{E}(\mathtt{h}_N) \to \mathtt{h}(\rho)$ as $N \to \infty$. We would like to discuss some features of their proof, so for convenience of our reader we attempt to sketch it here: They first show that

$$\mathtt{h}_N = (1/N)\sum_{i=1}^{N} \zeta_i$$

Due to their dependence on N, the ζ_i could also be written $\zeta_{i,N}$. The ζ_i are identically distributed random variables, so $\mathbb{E}(\mathtt{h}_N) = \mathbb{E}(\zeta_i)$ and we focus just on ζ_i. They consider the cumulative distribution function $F_{N,x}(u)$ of e^{ζ_i} conditioned on the case when the i-th sample equals x, *i.e.*, $X_i = x$. Let $\nu(y,r) = \{x \in \mathbb{R} : |x-y| < r\}$; one computes that

$$F_{N,x}(u) = 1 - \left(1 - \int_{\nu(x,\frac{u}{2\gamma(N-1)})} \rho(y)dy\right)^{N-1}$$

where $\log\gamma = .5772\ldots$ is the Euler constant (this is their equation (8)). We have

$$\left|\nu\left(x, \frac{u}{2\gamma(N-1)}\right)\right| = \frac{u}{\gamma(N-1)}$$

so we can write

$$\int_{\nu(x,\frac{u}{2\gamma(N-1)})} \rho(y)dy = \left(\frac{u}{\gamma(N-1)}\right) \frac{1}{|\nu(x,\frac{u}{2\gamma(N-1)})|}\int_{\nu(x,\frac{u}{2\gamma(N-1)})} \rho(y)dy$$

This means that, by the Lebesgue Differentiation Theorem, since $\nu(x,\frac{u}{2\gamma(N-1)})$ shrinks to x as $N \to \infty$, we have

$$\frac{1}{|\nu(x,\frac{u}{2\gamma(N-1)})|}\int_{\nu(x,\frac{u}{2\gamma(N-1)})} \rho(y)dy \longrightarrow \rho(x)$$

and so $F_{N,x}(u) \to 1 - e^{-\rho(x)u/\gamma}$ (this is their equation (8)). They define

$$F_x(u) = 1 - e^{-\rho(x)u/\gamma}$$

Let the random variable $\xi_{N,x}$ have the cumulative distribution function $F_{N,x}$ and the random variable ξ_x have the cumulative distribution function F_x. We can compute

$$\mathbb{E}(\log\xi_x) = -\log\rho(x)$$

and so one might hope that, since $F_{N,x} \to F_x$, we have E($\log\xi_{N,x}) \to \mathbb{E}(\log\xi_x) = -\log\rho(x)$ \$ because that would mean that $\mathbb{E}(\zeta_i^{(N)}|X_i = x) \to -\log\rho(x)$ (this is their equation (9)). The bulk of their proof is to show that this is in fact true. The proof is then completed by taking the pointwise result

$$\mathbb{E}(\zeta_i^{(N)}|X_i = x) \to -\log\rho(x)$$

and extending it to a convergence of integrals,

$$\mathbb{E}\mathtt{h}_N = \mathbb{E}(\zeta_i^{(N)}) = \int \mathbb{E}(\zeta_i^{(N)}|X_i = x)\rho(x)dx \to \int(-\log\rho(x)\ )\rho(x)dx = \mathtt{h}(\rho)\ \ \$$$

This is done in their equation (21).

There are several rates of convergence that would have to be quantified: First, a quantitative bound on

$$\left| \frac{1}{|\nu(x, \frac{u}{2\gamma(N-1)})|} \int_{\nu(x, \frac{u}{2\gamma(N-1)})} \rho(y)dy \quad - \quad \rho(x) \right|$$

would be needed. Next, we would need to quantify how the rate of convergence of $F_{N,x} \to F_x$ translates into the convergence $\mathbb{E}(\log \xi_{N,x}) \to \mathbb{E}(\log \xi_x)$. This rate would need to be somehow uniform in x; their constants C_1 (from their equation (20)) and therefore C_2 (at the bottom of [21, p. 99]) depend on x in an unspecified way. Some sort of uniformity is needed because the convergence $\mathbb{E}(\log \xi_{N,x}) \to \mathbb{E}(\log \xi_x)$ must be translated into the convergence

$$\int \mathbb{E}(\log \xi_{N,x})\rho(x)dx \to \int \mathbb{E}(\log \xi_x)\rho(x)dx$$

Quantifying these various rates of convergence appears to be nontrivial. We also note that subsequent research which has worked with estimators analogous to the one used in [21], such as [15], [10], and [14], has required more assumptions on the density in question.

In Györfi and van der Meulen [23], histogram-based density estimators ρ_N (computed from the sampled values $\{X_i\}$) are used. The stated results make no assumptions on ρ other than the finiteness of $\mathtt{h}(\rho)$. They do not give a quantitative rate of convergence, and non-quantitative methods in the proof (such as the Borel-Cantelli Lemma) suggest that a quantitative result would not necessarily follow trivially from their methods. Also, choosing the proper bin size and grid placement for a histogram is a continual problem. For unbounded densities such as ρ_f, any bin size which can capture the very tight grouping of sample points near an asymptote of ρ_f will then be too narrow and cause a "choppy" estimate in areas where ρ_f is not particularly large. One way of addressing this general drawback of histograms has been to adjust the "coarseness" of the estimate depending on how tightly the sample points are grouped, such as nearest-neighbor or sample-spacing-type estimators. We could also accept inefficiency near the critical points, but since we seek quantitative bounds this is not desirable. In practice, we believe that estimating the entropy of a density such as ρ_f with a histogram estimator will mean very large inefficiencies near the critical points.

In Vasicek [22], no assumptions are made on ρ other than $\int u^2\rho(u) < \infty$, which is satisfied by ρ_f. Let $\{X_{(i)}\}$ be the order statistics of the sample $\{X_i\}$, and let F be the cumulative distribution function of ρ. They use a sample-spacing estimator with a parameter m specifying the number of "neighbors" that will be considered when estimating the value of the density ρ near $X_{(i)}$. Their conclusion is that their estimator $\mathtt{h}'_{m,N}$ converges in probability to $\mathtt{h}(\rho)$ as $N \to \infty$, as long as $m \to \infty, m/N \to 0$. One key question, of course, is how m is chosen. They state [22] that "an optimal choice of m for a given $[N]$, however, depends on the (unknown) $[\rho]$. In general, the smoother the density $[\rho]$, the larger is the optimal value of m." Therefore, one would expect that we would need to choose relatively small values of m since ρ_f is very far from smooth, while still having $m \to \infty$. A key step in the proof is that

$$\frac{F(X_{(i+m)}) - F(X_{(i-m)})}{X_{(i+m)} - X_{(i-m)}} \qquad \left(\text{which is equal to } \frac{1}{X_{(i+m)} - X_{(i-m)}} \int_{X_{(i-m)}}^{X_{(i+m)}} \rho \right)$$

will equal $\rho(x')$ for some $x' \in (X_{(i-m)}, X_{(i+m)})$, whenever ρ is positive and continuous over the interval $(X_{(i-m)}, X_{(i+m)})$. This step is essentially using the Lebesgue Differentiation Theorem, so the accuracy of

$$\frac{1}{X_{(i+m)} - X_{(i-m)}} \int_{X_{(i-m)}}^{X_{(i+m)}} \approx \rho(x')$$

simultaneously over the entire domain of ρ, will need to be quantified. The tradeoff between needing m to be small in order to capture the "fine-scale" behavior of ρ_f near an asymptote, while needing m to be large in order to carry through the rest of the proof, appears non-trivial. A discussion of results which use a fixed m, versus $m \to \infty$ as in [22], is in Tsybakov and van der Meulen [15]. We note also that results using similar techniques, which have achieved conclusions stronger than the "convergence in probability" shown here (such as van Es [11]) have required more restrictions on ρ.

There are other results, such as those in Godavarti and Hero [24], Csiszár [25], and Rényi [26], which do not specifically solve problem 2, but which give insight into how one might produce a value that is "close" to $\mathtt{h}(\rho)$ given only some information about ρ. These results are not "stochastic"; the approximations for $\mathtt{h}(\rho)$ that will be considered are not constructed by sampling from ρ.

In Csiszár [25], the only assumption is that the entropy integral exists, there are no other assumptions on the density ρ. Their results are explained in terms of a general measurable space X, and in their presentation the density ρ is not the central object, it simply arises as the Radon-Nikodym derivative of a probability measure on X (in their notation this probability measure is μ), with respect to a σ-finite measure on X (in their notation this σ-finite measure is λ). We will assume that the measurable space is $\mathbb{R}$, the σ-finite measure is Lebesgue measure on $\mathbb{R}$, and the probability measure is given by the density ρ. In their Theorem 1, they show that $\mathtt{h}(\rho)$ equals the infimum value of a set of "approximate entropies," where the density ρ is replaced with its average value on each set in a partition $\{A_i\}$ of $\mathbb{R}$, and then the entropy of the resulting "approximate density" is calculated. If we define the characteristic function of a set S, χ_S, by

$$\chi_S(u) = \begin{cases} 1 & \text{if } u \in S \\ 0 & \text{if } u \notin S \end{cases}$$

then the "approximate density" is:

$$\rho_{\{A_i\}}(u) = \sum_i \frac{1}{|A_i|} \left(\int_{A_i} \rho \right) \chi_{A_i}(u) \qquad (\text{this is} \approx \rho)$$

and the "approximate entropy" is:

$$\begin{aligned} \mathtt{h}(\rho_{\{A_i\}}) &= -\int \rho_{\{A_i\}} \log \rho_{\{A_i\}} \\ &= -\sum_i \left(\int_{A_i} \rho \right) \log \left(\frac{1}{|A_i|} \int_{A_i} \rho \right) \qquad (\text{this is} \approx \mathtt{h}(\rho)) \end{aligned}$$

They prove this result by choosing a specific set of approximate entropy values $\mathtt{h}_\epsilon$, where in fact $\mathtt{h}_\epsilon$ is easily seen to be within ϵ of the true value $\mathtt{h}(\rho)$. This is certainly quantitative in the sense that the approximating value, $\mathtt{h}_\epsilon$, is known to be within a certain explicit amount of the true value. The question then becomes whether we can calculate the value $\mathtt{h}_\epsilon$. The approximating values $\mathtt{h}_\epsilon$ are obtained,

essentially, by the definition of the Lebesgue integral of $\log \rho$ (with respect to the measure given by ρ) as a limit of integrals of simple functions (for this definition, see for example [27]. For a probability measure μ on $\mathbb{R}$ and a function $g \in L^1(\mu)$, if we choose ϵ and we let

$$A_i = \{x \in \mathbb{R} : g(x) \in [\epsilon i, \epsilon(i+1)\ )\ \}\ , \quad i \in \mathbb{Z}$$

then it is obvious that $\sum_i (\epsilon i)\mu(A_i)$ is within ϵ of $\int g d\mu$. In [25], $\log \rho$ plays the role of g, the measure μ is given by ρ, and we have

$$\mathtt{h}_\epsilon = \sum_i (\epsilon i)\mu(A_i) \ \text{ is within } \epsilon \text{ of } \int \log \rho d\mu = \int_{\mathbb{R}} \rho \log \rho$$

In order to calculate the value of $\sum_i (\epsilon i)\mu(A_i)$, we need to calculate $\mu(A_i)$, which is

$$\mathbb{P}[\log \rho \in [\epsilon i, \epsilon(i+1)\ )\ ] = \int_{\{\log \rho \in [\epsilon i, \epsilon(i+1)\ )\}} \rho$$

However, we do not know ρ, so we do not know the exact locations on $\mathbb{R}$ when $\log \rho$ is going to fall in a certain range. Of course, as we mentioned, the purpose of the $\mathtt{h}_\epsilon$ construction is an existence proof, not necessarily as a calculational or estimation tool.

We turn now to Rényi [26]. Rényi considers density functions ρ, the only requirement on ρ being that the discrete entropy of the "integer compression" of ρ is finite, by which we mean (suppose X is a random variable having density ρ):

$$-\sum_{j \in \mathbb{Z}} \mathbb{P}[X \in [j, j+1]\ ] \log\ \mathbb{P}[X \in [j, j+1] < \infty \tag{3}$$

This is not equivalent to the existence of the entropy integral, as he shows in [26]. This admits a wide class of densities, including ρ_f for the functions f we will consider. The key tool will be to consider successively finer step-function approximations of ρ, the partition set being the grid with "step size" $1/N$. Let us define $\rho_{[1/N]}$ by

$$\rho_{[1/N]} = \sum_{j \in \mathbb{Z}} \chi_{[j/N,(j+1)/N)} \left(\frac{1}{1/N} \int_{j/N}^{(j+1)/N} \rho \right)$$

Note that $\rho_{[1/N]}$ is essentially the "true" histogram with the grid $\{j/N\}_{j=-\infty}^{\infty}$ (an example of the approximations $\rho_{\{A_i\}}$ considered in Csiszár [25]). Rényi proves that, if the "integer compression" of ρ has finite discrete entropy, then

$$\lim_{N \to \infty} \mathtt{h}(\rho_{[1/N]}) = -\int \rho \log \rho$$

One sign that this result will not be optimal for a density such as ρ_f is that Rényi first proves this convergence for bounded ρ. He introduces an intermediate parameter L (which specifies the point at which the tails of the sums and integrals will be cut off), proves the pointwise convergence result

$$\rho_{[1/N]}(x) \log \rho_{[1/N]}(x) \to \rho(x) \log \rho(x) \qquad \text{a.e. } x$$

(see his equation (35)) and then applies the "Theorem of Lebesgue" [26] (we believe he is referring to the Dominated Convergence Theorem) over the interval $[-L, L]$ to obtain

$$\int_{-L}^{L} \rho_{[1/N]} \log \rho_{[1/N]} \to \int_{-L}^{L} \rho \log \rho \qquad \text{as } N \to \infty$$

The tails of the integral are bounded using our numbered equation (3), which completes the proof in the case of bounded ρ. For unbounded ρ he introduces another intermediate parameter A and considers the truncated function ρ_A (which is bounded):

$$\rho_A(u) = \begin{cases} \rho(u) & \text{if } \rho(u) \leq A \\ 0 & \text{if } \rho(u) > A \end{cases}$$

He applies the result for bounded ρ to ρ_A, and then he bounds the difference

$$\int \rho_{[1/N]} \log \rho_{[1/N]} - \int (\rho_A)_{[1/N]} \log (\rho_A)_{[1/N]}$$

using expressions involving A and L (but, importantly, *not* involving N, see his equations (48), (50) and (53)). Therefore, his equations (56), (57), (58) and (59) show how the parameters L, A and N interact to produce the result. To make this result quantitative would entail balancing these various rates of convergence, which seems to be non-trivial. Finally, being essentially a histogram-type method (the bin sizes depend only on N), it would face the difficulties that come along with histograms that we covered in our discussion of Györfi and van der Meulen [23].

We turn now to Godavarti and Hero [24]. In Theorem 4 of [24], the convergence $\mathtt{h}(\rho_N) \to \mathtt{h}(\rho)$ is proved under fairly general conditions. The only assumptions are that $\rho_N \to \rho$ pointwise and that there exist a constant L and some $\kappa > 1$ such that the following statements are true:

$$\int \rho |\log \rho|^{\kappa} \leq L \tag{4}$$

and for all N,

$$\int \rho_N |\log \rho_N|^{\kappa} \leq L \tag{5}$$

Densities such as ρ_f typically satisfy equation (4). Furthermore, if we require f to satisfy the regularity assumption we introduce later (Definition 6), and if an approximation method is used to produce g_N given N samples from f, it is reasonable to expect that ρ_f and $\{\rho_{g_N}\}$ would satisfy equations (4) and (5) for some L. The necessary pointwise convergence is assured by any reasonable method of producing g_N, and then Theorem 4 of [24] in fact proves the result $\mathtt{h}(g_N) \to \mathtt{h}(f)$ (although we will not attempt to formalize the above argument). This is essentially what we prove in our Theorem 1, with the key difference that our result is quantitative. This certainly makes one wonder if the result in Theorem 4 of [24] could in fact be easily adapted to obtain a quantitative proof of our Theorem 1, thus not needing our methods. Currently, we believe the answer is no. Or, looked at another way, we believe that a quantitative version of their result for the problem we are examining would end up confronting the same issues we faced and might end up resembling our methods.

Since [24] does not present a quantitative result, we look to their proof to see if such a result can be found in their methods. Note that in [24], "N" has a meaning different from how we have been

using it, and "i" is the variable which corresponds to our "N." We will use the notation from [24] for the moment, for the convenience of the reader who might want to refer back to [24] during our short discussion below; we hope confusion can be avoided. The main ingredients of the proof in [24] are several variables (N, K, i) which must become "sufficiently large" and the application of theorems such as Egoroff's Theorem and Lebesgue Dominated Convergence. One of the chief advantages of these powerful theorems from measure theory is that they can bootstrap pointwise convergence into stronger conclusions and bypass potentially forbidding quantitative relationships. Also, the "essence of the proof" (as they state in [24]) is to find a set A_ϵ such that everything happening on A_ϵ^c (the "bad" set) is negligible, and the desired convergence does in fact take place on A_ϵ. The set A_ϵ, by construction, is a set on which ρ and all but finitely many of the ρ_i are bounded above by N and bounded away from zero by $1/N$. There is a tradeoff between the size of N (larger N means a smaller "bad" set, but less control on the "good" set) and the other parameters K and i. Therefore, to obtain a quantitative result, the rates of growth of these parameters must be balanced, and the use of the measure-theoretic theorems mentioned above must be replaced with concrete estimates; this would appear to be non-trivial.

There is another issue of concern in [24], a scaling issue. As mentioned in problem 1, the scaling $f \to \lambda f$ for $\lambda \neq 0$ should essentially have no effect on the problem. Note that this scaling induces the standard L^1 scaling

$$\rho(u) \to |\lambda|^{-1}\rho(u/\lambda)$$

so a method which proves a bound analogous to ours should respect this scaling. The assumption by Godavarti and Hero (in equation (4) of the present article) involves a constant, L, and the specific value of L is used in the proof (for example, we need N, K large enough that (essentially) $L/|\log K|^{\kappa-1}$ and $L/|\log N|^{\kappa-1}$ are sufficiently small). The scaling $\rho(u) \to |\lambda|^{-1}\rho(u/\lambda)$ transforms the integral in (4) into

$$\int \rho \left| \log|\lambda|^{-1} + \log\rho \right|^{\kappa}$$

which means in turn that a different value of L (and therefore different values of K and N) will be needed. Therefore, in order to prove a result similar to our Theorem 1, this scaling issue will need to be resolved as well.

3. Definitions and Notation

We will use the following notation:

Definition 1. *For real numbers a, b, we define*

$$\langle a, b\rangle = [\min\{a,b\}, \max\{a,b\}] = \begin{cases} [a,b] & \text{if } a \le b \\ [b,a] & \text{if } a > b \end{cases}$$

For subsets $S, T \subset \mathbb{R}$, we define

$$S + T = \{s + t : s \in S, t \in T\}$$

So, for example,

$$\langle a, b\rangle + [-2, 3] = [\min\{a,b\} - 2, \max\{a,b\} + 3]$$

As we defined in the Introduction,

Definition 2. *For a measurable function $h : [0,1] \to \mathbb{R}$, we define the measure μ_h on $\mathbb{R}$ by*

$$\mu_h(E) = |h^{-1}(E)|$$

We will have occasion to refer to the "monotone pieces" of a function, so we make the following definition. We do not want to consider functions which have critical points at either 0 or 1, for technical reasons and to avoid more cumbersome notation, so we define:

Definition 3. *For a differentiable function h with a finite set of critical points $t_1 < \cdots < t_k$, and with $h'(0), h'(1) \neq 0$, we define $t_0 = 0, t_{k+1} = 1$ and for $j = 0, \ldots, k$ we define $h_j = h|_{[t_j, t_{j+1}]}$.*

For any function h, we will abbreviate the domain and range of h:

Definition 4. *We denote the domain of h by Dom(h), and the range of h by Range(h)*

Recall the definition of the (discrete) Shannon entropy, H, of a finite probability mass distribution $\vec{p} = \{p_i\}_{i=1}^n$:

$$H(\vec{p}) = -\sum_i p_i \log p_i$$

Let h be a differentiable function with finitely many critical points and $h'(0), h'(1) \neq 0$. Suppose that μ_h is absolutely continuous. Then we know that the measure given by $\mu_{h_i}(E) = |h_i^{-1}(E)|$ is absolutely continuous; let its density be ρ_{h_i}. For a fixed u in the range of h, note that u is in the range of at least one, possibly several, of the h_i. For those i such that u is in the range of h_i we see that ρ_{h_i} is non-zero, and we know that

$$\sum_i \rho_{h_i}(u) = \rho_h(u)$$

So, we define the probability mass distribution

$$\vec{p}_h(u) = \left\{ \frac{\rho_{h_i}(u)}{\rho_h(u)} \right\}_{i=0}^k \tag{6}$$

Note that the number of non-zero entries of $\vec{p}_h(u)$ equals the number of i such that u is in the range of h_i.

4. A Useful Identity

Assume that f is a differentiable function with a finite set of critical points $t_1 < \cdots < t_k$ and $f'(0), f'(1) \neq 0$. The following identity is the centerpiece of our method for proving our main result (a helpful comment [28] led us to this identity):

$$\mathrm{h}(f) = \int_{t=0}^1 \log |f'(t)| dt - \int_{t=0}^1 H\Big(\vec{p}_f(f(t))\Big) dt \tag{7}$$

In the subsequent sections of this article, to simplify a discussion we will sometimes refer to the terms on the right hand side by (L) and (E), corresponding to the "logarithm" term and the "entropy" term. To prove (7), note the following:

$$\begin{aligned}\sum[\rho_{f_i}\log\rho_{f_i}]-\rho_f\log\rho_f &= \sum[\rho_{f_i}\log\rho_{f_i}-\rho_{f_i}\log\rho_f]\\ &= \sum\rho_{f_i}\log(\rho_{f_i}/\rho_f)\\ &= \rho_f\sum(\rho_{f_i}/\rho_f)\log(\rho_{f_i}/\rho_f)\\ &= -\rho_f H(\vec{p}_f(u))\end{aligned}$$

Integrating both sides, we have

$$\int\sum[\rho_{f_i}\log\rho_{f_i}]-\int\rho_f\log\rho_f=-\int\rho_f H(\vec{p}_f(u))$$

or

$$\sum_i\int_{u\in Range(f_i)}\rho_{f_i}(u)\log\rho_{f_i}(u)\,du+\mathtt{h}(f)=-\int\rho_f H(\vec{p}_f(u))$$

The identity $\int_u\rho_{f_i}(u)g(u)du=\int_{t\in Dom(f_i)}g(\ f_i(t)\ )dt$, and the formula

$$\rho_{f_i}(u)=\frac{1}{\left|f_i'\left(f_i^{-1}(u)\ \right)\right|}$$

are used on the first term to obtain

$$\sum_i\int_{u\in Range(f_i)}\rho_{f_i}(u)\log\rho_{f_i}(u)\,du=\sum_i\int_{t\in Dom(f_i)}\log\rho_{f_i}(\ f_i(t)\ )dt=\int_{t=0}^1\log\left(\frac{1}{|f_i'(t)|}\right)dt$$

The identity $\int_u\rho_f(u)g(u)du=\int_{t=0}^1 g(\ f(t)\ )dt$ is used on the last term to obtain

$$-\int_u\rho_f H(\vec{p}_f(u))du=-\int_{t=0}^1 H\Big(\vec{p}_f(f(t))\Big)dt$$

and so (7) is proved.

This identity is equivalent to a standard identity for Shannon entropy. Suppose we have a probability space Ω, a finite set S, and a random variable $X:\Omega\to S$. Furthermore, suppose we have an "indicator" function

$$\Delta:\Omega\longrightarrow\mathbb{N}$$

which takes a finite number of values on Ω. We think of Δ as specifying a partition of Ω, *i.e.*,

$$\Omega=\Delta^{-1}(1)\ \cup\ \Delta^{-1}(2)\cup\ \ldots\ =\Omega_1\ \cup\ \Omega_2\ \cup\ldots$$

If X maps the Ω_i to disjoints parts of S, *i.e.*,

$$X(\Omega_i)\cap X(\Omega_j)=\emptyset\ \text{ for } i\neq j$$

then we would have

$$\begin{aligned}
H(X) &= -\sum_{s\in S} \mathbb{P}[X=s]\log \mathbb{P}[X=s] \\
&= -\sum_i \sum_{s\in X(\Omega_i)} \mathbb{P}[X=s]\log \mathbb{P}[X=s] \\
&= -\sum_i \sum_{s\in X(\Omega_i)} (\mathbb{P}(\Omega_i)\mathbb{P}[X=s|\Omega_i])\log(\mathbb{P}(\Omega_i)\mathbb{P}[X=s|\Omega_i]) \\
&= -\sum_i \mathbb{P}(\Omega_i) \sum_{s\in X(\Omega_i)} \mathbb{P}[X=s|\Omega_i]\Big(\log(\mathbb{P}(\Omega_i)+\log \mathbb{P}[X=s|\Omega_i])\Big) \\
&= -\sum_i \mathbb{P}(\Omega_i)\log(\mathbb{P}(\Omega_i) - \sum_i \mathbb{P}(\Omega_i) \sum_{s\in X(\Omega_i)} \mathbb{P}[X=s|\Omega_i]\log \mathbb{P}[X=s|\Omega_i] \\
&= H(\Delta) + H(X|\Delta)
\end{aligned}$$

where $H(X|\Delta)$ is the conditional entropy of X given Δ (see [29]). In other words, $H(\Delta)+H(X|\Delta)$ would be the Shannon entropy of X if the images of the Ω_i under X did not overlap. If they did overlap, this would be an overestimation of the entropy of X, and in general the true value is given by subtracting a term, $H(\Delta|X)$, which represents the amount of "overlap":

$$H(X) = H(\Delta) + H(X|\Delta) - H(\Delta|X) \tag{8}$$

This identity follows from the "chain rule" for entropy [29].

In our case, the probability space is $[0,1]$, the set is $\mathbb{R}$, the random variable is $f:[0,1]\to\mathbb{R}$, and the indicator function on the probability space is the function

$$\Delta(t) = i \quad \text{if } t\in[t_i,t_{i+1}]$$

which (essentially) partitions $[0,1]:[0,1]=[0,t_1]\cup[t_1,t_2]\cup\ldots\cup[t_k,1]$. As above,

$$H(\Delta)+\mathtt{h}(f|\Delta) = \int_{t=0}^{1} \log|f'(t)|dt$$

measures the entropy of f as if the different "monotone laps" of f did not overlap. If they do overlap then this is an overestimate, and the true value is given by subtracting the term representing the amount of "overlap":

$$\mathtt{h}(\Delta|f) = \int_{t=0}^{1} H\Big(\vec{p}_f(f(t))\Big)dt$$

In this way, (7) is analogous to (8).

5. Scaling

For simplicity, we consider $\lambda>0$ (the case $\lambda<0$ is no different). For $\lambda>0$, we have $\rho_{\lambda f}(u) = \lambda^{-1}\rho_f(u/\lambda)$, and so $\mathtt{h}(\lambda f)=\mathtt{h}(f)+\log\lambda$. This means

$$\mathtt{h}(\lambda f)-\mathtt{h}(\lambda g) = \mathtt{h}(f)-\mathtt{h}(g)$$

So, the difference $\mathtt{h}(f) - \mathtt{h}(g)$ is invariant under the scaling $f \to \lambda f$. However, if we take $[a, b] \subset \mathbb{R}$, and we consider

$$\int_a^b \rho_f \log \rho_f - \rho_g \log \rho_g$$

then, scaling by the factor λ, we have

$$\int_{\lambda a}^{\lambda b} \rho_{\lambda f} \log \rho_{\lambda f} - \rho_{\lambda g} \log \rho_{\lambda g} = \left[\int_a^b \rho_f \log \rho_f - \rho_g \log \rho_g\right] + (\log \lambda) \left[\int_a^b \rho_f - \rho_g\right]$$

Since the second term on the right hand side of this equation is not necessarily zero, we see that an integral over just a part of the domain may not be scale-invariant.

Turning our attention to the identity (7), we see that for $\lambda > 0$ we have:

$$\int_{t=0}^1 \log |(\lambda f)'(t)| dt = \int_{t=0}^1 \log |f'(t)| dt + \log \lambda$$
$$\int_{t=0}^1 H\Big(\vec{p}_{\lambda f}\big((\lambda f)(t)\big)\Big) dt = \int_{t=0}^1 H(\vec{p}_f(f(t))) dt$$

Perhaps more importantly, each term can be divided along its respective axis, and each individual piece scales similarly:

$$\int_{t_a}^{t_b} \log |(\lambda f)'(t)| dt = \int_{t_a}^{t_b} \log |f'(t)| dt + (t_b - t_a) \log \lambda$$
$$\int_{t_a}^{t_b} H\Big(\vec{p}_{\lambda f}\big((\lambda f)(t)\big)\Big) dt = \int_{t_a}^{t_b} H(\vec{p}_f(f(t))) dt$$

Our main strategy takes advantage of precisely this. Recall the definitions for (L) and (E) that followed (7); both (L) and (E) will be subdivided along the t axis.

6. A Counter-Example

We will denote the ith derivative of f (or g) by $f^{(i)}$ (or $g^{(i)}$). One might hope that a theorem along the following lines could be proved, C being an absolute constant:

$$|\mathtt{h}(f) - \mathtt{h}(g)| \leq C \sum_{i=0}^{K} \|f^{(i)} - g^{(i)}\|_\infty$$

This would show that the $\mathtt{h}$-functional is a continuous map in the norm topology of some Banach space of differentiable functions. One can see immediately that this result is not possible, just due to scaling: If we let $f_n = (1/n)f, g_n = (1/n)g$ then $|\mathtt{h}(f_n) - \mathtt{h}(g_n)| = |\mathtt{h}(f) - \mathtt{h}(g)|$ while the right hand side converges to zero as $n \to \infty$. However, we might hope that we could prove a result such as

$$|\mathtt{h}(f) - \mathtt{h}(g)| \leq \left[C \sum_{i=0}^{K} \|f^{(i)} - g^{(i)}\|_\infty\right] / \min\Big\{ \{\|f^{(i)}\|_\infty, \|g^{(i)}\|_\infty\}_{i=0}^K \Big\} \tag{9}$$

In other words, perhaps this scaling issue is the only problem and the $\mathtt{h}$-functional is continuous when restricted to points f and g which are bounded away from zero in this Banach space. This is not the case. In fact, suppose that we ask even less of the result we seek. Suppose that we seek some $r > 0$, $c > 0$, and

some function Γ of $2K+2$ variables which is bounded on compact subsets of the open positive orthant $(\mathbb{R}_+)^{2K+2}$, and we just want the following bound to hold: Whenever

$$\Big[\sum_{i=0}^{K}\|f^{(i)}-g^{(i)}\|_\infty^r\Big]\ \Gamma(\|f^{(0)}\|_\infty,\ldots,\|f^{(K)}\|_\infty,\|g^{(0)}\|_\infty,\ldots,\|g^{(K)}\|_\infty)\le c$$

is true, we have

$$|\mathtt{h}(f)-\mathtt{h}(g)|\le\Big[\sum_{i=0}^{K}\|f^{(i)}-g^{(i)}\|_\infty^r\Big]\ \Gamma(\|f^{(0)}\|_\infty,\ldots,\|f^{(K)}\|_\infty,\|g^{(0)}\|_\infty,\ldots,\|g^{(K)}\|_\infty) \tag{10}$$

In other words, whenever the differences $\{\|f^{(i)}-g^{(i)}\|_\infty^r\}_{i=0}^K$, after being multiplied by the (large) value Γ, are sufficiently small, then we have a quantitative estimate on $\mathtt{h}(f)-\mathtt{h}(g)$.

This is still not possible. Note that (9) is a special case of (10). Note also that our main result, Theorem 1, is almost a special case of (10) with $K=3$, the only additional element in Theorem 1 being the regularity condition (Definition 6) and the involvement of the parameter from that condition, δ.

The impossibility of a result such as (10) can be seen just looking at monotone functions, in which case the $\mathtt{h}$ functional is identical to the functional $f\to\int\log|f'|$. The reason is quite straightforward: The functional $f\to\int\log|f'|$ will become large if a function is extremely flat even for a short distance, whereas any $\|\cdot\|_\infty$-type norm will not necessarily measure a large difference between a function that is extremely flat and a function that is merely somewhat flat. We formalize this in the following proposition:

Proposition 5. *Fix K and ϵ. There exists a constant C_K depending only on K, and monotone functions f,g, such that*

$$\begin{aligned}
&\forall i\le K,\quad f^{(i)},g^{(i)}\ \textit{are continuous}\\
&\forall i\le K,\|f^{(i)}-g^{(i)}\|_\infty\le\epsilon\\
&\forall i\le K,\|f^{(i)}\|_\infty,\|g^{(i)}\|_\infty\ge 1\\
&\forall i\le K,\|f^{(i)}\|_\infty,\|g^{(i)}\|_\infty\le C_K
\end{aligned}$$

but

$$\left|\int\log|f'|-\log|g'|\right|\ge 1$$

Proof. Let $f_0(t)=t^N$ and $g_0(t)=t^M$; we are free to choose M and N independently, and the idea is to let $M>>N>>K$. Define

$$f(t)=f_0(t)\ \text{ on }[0,1/2]\,,\qquad f(t)=2^{K+1}(t-1/2)^{K+1}+\sum_{i=0}^{K}\frac{f^{(i)}(1/2)}{i!}(t-1/2)^i\ \text{ on }[1/2,1]$$

$$g(t)=g_0(t)\ \text{ on }[0,1/2]\,,\qquad g(t)=2^{K+1}(t-1/2)^{K+1}+\sum_{i=0}^{K}\frac{g^{(i)}(1/2)}{i!}(t-1/2)^i\ \text{ on }[1/2,1]$$

We see that f and g were chosen to have $f^{(i)},g^{(i)}$ continuous for $i\le K$. If we choose N and M so that $K<N/2, N<M/2$, we see that $\forall i\le K$, $\alpha\le 1/2$,

$$g_0^{(i)}(\alpha)\le M^K\alpha^{M/2}\ \text{ and }\ f_0^{(i)}(\alpha)\ge\alpha^N$$

so once we choose sufficiently large M so that $M^K\alpha^{M/2} \leq \alpha^N$ for all $\alpha \leq 1/2$, we have $g_0^{(i)}(\alpha) \leq f_0^{(i)}(\alpha)\ \ \forall i \leq K, \alpha \leq 1/2$ and therefore $g^{(i)}(t) \leq f^{(i)}(t)\ \ \forall i \leq K, t \in [0,1]$.

Note that, on $[0,1/2]$ we have $f^{(i)}(t) \leq f^{(i)}(1/2) \leq N^K(1/2)^{N/2}$. Choose sufficiently large N so that

$$N^K(1/2)^{N/2} \leq \epsilon/\Big[(K+1)K^K\Big]$$

This certainly means $|f^{(i)} - g^{(i)}| \leq \epsilon$ on $[0,1/2]$. On $[1/2,1]$ we just need $\sum_{i=0}^K \frac{f^{(i)}(1/2)}{i!}(t-1/2)^i$ and all of its derivatives to be smaller than ϵ and then we will have $\|f^{(i)} - g^{(i)}\|_\infty \leq \epsilon\ \ \forall i \leq K$. We see that

$$\Big[\sum_{i=0}^K \frac{f^{(i)}(1/2)}{i!}(t-1/2)^i\Big]^{(j)}(t) \leq (K+1)\big[\max_{i\leq K} f^{(i)}(1/2)\big]K^K \leq \epsilon$$

and so we have $\|f^{(i)} - g^{(i)}\|_\infty \leq \epsilon\ \ \forall i \leq K$.

By the above calculation, we have the immediate bound

$$f^{(i)}(t) \leq f^{(i)}(1) \leq 2^{K+1}(K+1)! + (K+1)K^K\big[\max_{i\leq K} f^{(i)}(1/2)\big]$$

and we see that the following is true, with a constant C'_K depending only on K:

$$\max_{i\leq K} f^{(i)}(1/2) \leq N^K 2^{-(N-K)} \leq C'_K$$

which means $\|f^{(i)}\|_\infty \leq C_K$ (and the same is true for $g^{(i)}$ since $g^{(i)} \leq f^{(i)}$).

Finally, observing that $f^{(i)}(1) \geq 2^{K+1}(1/2)^{K+1} + \big(\sum \cdots\big) \geq 1$, and noting that the same is true for g, we have $\forall i \leq K, \|f^{(i)}\|_\infty, \|g^{(i)}\|_\infty \geq 1$.

So, f and g meet the criteria of the proposition. At this point, let us "fix" the value of N that we have arrived at. Since $f', g' > 0$ and $g' \leq f'$, we have

$$\begin{aligned}\left|\int \log|f'| - \log|g'|\right| &= \int_0^1 \log f' - \log g' \\ &\geq \int_0^{1/2} \log f' - \log g' \\ &= (1/2)[\log N - \log M] + [(N-1)-(M-1)]\left[\int_0^{1/2} \log t\right]\end{aligned}$$

and as $M \to \infty$ this expression grows without bound, proving the proposition. □

7. Main Result

Suppose that we have functions $f, g : [0,1] \to \mathbb{R}$. We will assume that each function has three continuous derivatives; we do not attempt to determine the precise smoothness required. The quantities $\|f^{(3)}\|_\infty, \|g^{(3)}\|_\infty$ will be used frequently, and the reader might be worried, glancing at Theorem 1, that the estimate becomes "worse" as $\|f^{(3)}\|_\infty, \|g^{(3)}\|_\infty$ become small, certainly a puzzling property. In fact we only need upper bounds on $|f'''|, |g'''|$, not *least* upper bounds, so the reader may substitute any upper bounds on $|f'''|, |g'''|$ in place of $\|f^{(3)}\|_\infty, \|g^{(3)}\|_\infty$ as long as it is done consistently. Notice that this is still scale invariant (if A is an upper bound on $|f'''|$ then λA is an upper bound on $|(\lambda f)'''|$). We will use $\|f^{(3)}\|_\infty, \|g^{(3)}\|_\infty$ for simplicity.

The quantities which actually bring $\mathtt{h}(f)$ and $\mathtt{h}(g)$ close together are $\|f-g\|_\infty$ and $\|f'-g'\|_\infty$; the "smallness" of these quantities, relative to the size of the third derivatives of f and g, will be quantified by a parameter ϵ.

However, Proposition 5 means that the "smallness" of $\|f-g\|_\infty$ and $\|f'-g'\|_\infty$ alone cannot imply the result we are seeking. We will impose an additional regularity assumption.

Definition 6 (The Set $A(\delta)$)**.** *Suppose that $h:[0,1]\longrightarrow\mathbb{R}$. For $0<\delta\leq 1$, we say $h\in A(\delta)$ if h, h', h'' and $h^{(3)}$ are continuous, and the following two conditions hold:*

$$|h'(t)|+|h''(t)|\geq\delta\|h^{(3)}\|_\infty \qquad \forall t\in[0,1] \tag{11}$$

$$|h'(t)|\geq\delta\|h^{(3)}\|_\infty \qquad \forall t\in[0,\delta]\cup[1-\delta,1] \tag{12}$$

We will require that f and g are in the set $A(\delta)$. Intuitively, the requirement for a function to be in $A(\delta)$ is a quantitative statement that its first and second derivatives do not simultaneously vanish, *i.e.*, it does not have any "flat spots". The second half of the condition, equation (12), is not as natural; its purpose is simply to eliminate some uninteresting technical difficulties at the endpoints of $[0,1]$.

Lastly, we will need bounds on the ratios

$$\frac{\max\|f'\|_\infty,\|g'\|_\infty}{\min\left\{\|f^{(3)}\|_\infty,\|g^{(3)}\|_\infty\right\}} \quad\text{and}\quad \frac{\max\|f''\|_\infty,\|g''\|_\infty}{\min\left\{\|f^{(3)}\|_\infty,\|g^{(3)}\|_\infty\right\}}$$

With these assumptions, we can bound the difference between $\mathtt{h}(f)$ and $\mathtt{h}(g)$:

Theorem 1 (Main Result)**.** *Suppose that $f,g:[0,1]\to\mathbb{R}$ are each members of $A(\delta)$, and for some $\epsilon\leq 2^{-20}\delta^8$ we have*

$$\|f-g\|_\infty\leq\epsilon\min\left\{\|f^{(3)}\|_\infty,\|g^{(3)}\|_\infty\right\} \tag{13}$$

$$\|f'-g'\|_\infty\leq\epsilon^{3/4}\min\left\{\|f^{(3)}\|_\infty,\|g^{(3)}\|_\infty\right\} \tag{14}$$

Furthermore, let C_1,C_2 be constants such that $C_1,C_2\geq 3$ and

$$C_1^{-1}\leq\frac{\max\|f'\|_\infty,\|g'\|_\infty}{\min\left\{\|f^{(3)}\|_\infty,\|g^{(3)}\|_\infty\right\}}\leq C_1\ ,\qquad \frac{\max\|f''\|_\infty,\|g''\|_\infty}{\min\left\{\|f^{(3)}\|_\infty,\|g^{(3)}\|_\infty\right\}}\leq C_2$$

Then

$$|\mathtt{h}(f)-\mathtt{h}(g)|\leq CC_1C_2\log\left(C_1\delta^{-1}\epsilon^{-1/4}\right)\epsilon^{1/4}/\delta^4$$

where C is an absolute constant.

Our main result can be thought of as a converse to Proposition 5: If the functions f and g do not have any extreme "flat spots" then (essentially) the smallness of $\|f-g\|_\infty$ and $\|f'-g'\|_\infty$ are sufficient to imply that $\mathtt{h}(f)$ and $\mathtt{h}(g)$ are close.

Notice that the result is scale invariant: the assumptions and conclusion are not altered in the least if we multiply both f and g by the same non-zero constant.

8. Overview of Proof

The proof begins by using the "useful identity" from Section 4. to split the difference $\mathtt{h}(f) - \mathtt{h}(g)$ into two terms:

$$\mathtt{h}(f) - \mathtt{h}(g) = \int_{t=0}^{1} \log|f'(t)| - \log|g'(t)| \; - \; \int_{t=0}^{1} H\Big(\vec{p}_f\big(f(t)\big)\Big) - H\Big(\vec{p}_g\big(g(t)\big)\Big)$$

As in Section 4., we abbreviate the first term by (L) and the second term by (E). Next, we wish to split the domain of integration into a "good" set and a "bad" set. However, this is done differently for (L) and (E).

The regions of $[0,1]$ that are causing the problems are the places where the derivatives of f and g are very small (or zero). Let us denote this region by B^*. When we are dealing with just $\log|f'|, \log|g'|$, it is just the set B^* that will constitute the "bad" set. Therefore, we write

$$\int \log|f'(t)| - \log|g'(t)| = \int_{B^*} \log|f'(t)| - \log|g'(t)| + \int_{(B^*)^c} \log|f'(t)| - \log|g'(t)|$$

or $(L) = (LB) + (LG)$.

When dealing with the expression $H\Big(\vec{p}_f\big(f(t)\big)\Big)$, it is of course the same set B^* that causes the trouble, but the affected areas spread beyond just those values of t which are in B^*. If we have a value of t such that the function value $f(t)$ is in the image (by f) of some *other* point t' with f having a small derivative at t', then the point t will be a "problem" for $H\Big(\vec{p}_f\big(f(t)\big)\Big)$, since the finite probability mass distribution $\vec{p}_f\big(f(t)\big)$ involves all points t' such that $f(t') = f(t)$. In other words, the behavior of $H\Big(\vec{p}_f\big(f(t)\big)\Big)$ is not determined by just a neighborhood of t. This makes sense if we recall that this term is trying to capture the "overlap" in function values that might happen between widely separated regions of the domain of f.

We can think of this in the following way: the "bad" set B^* is going to "infect" a region of the u-axis (the set $\mathcal{I}$ from Definition 16) and then any points in $[0,1]$ which map to that "infected" area will also be "infected". The set G^* (defined precisely in Definition 16) is the portion of $[0,1]$ that is not "infected" at all, *i.e.*,. it is "good". If $t \in G^*$, this means that there is no point $t' \in B^*$ for which $f(t') = f(t)$ or $g(t') = g(t)$ (the statement is slightly stronger than this, but that is the idea).

To summarize: We divide $[0,1]$ into a "bad" region B^*, a "good" region G^*, and a "neutral" region $(B^*)^c \cap (G^*)^c$. For (L), it is B^* that is "bad" and everything else is "good," while for (E) it is only G^* that is "good" and everything else is "bad".

Therefore, for purposes of $H\Big(\vec{p}_f\big(f(t)\big)\Big)$ the bad set will be $(G^*)^c$ and so we write

$$\int H\Big(\vec{p}_f\big(f(t)\big)\Big) - H\Big(\vec{p}_g\big(g(t)\big)\Big) = \int_{(G^*)^c} H\Big(\vec{p}_f\big(f(t)\big)\Big) - H\Big(\vec{p}_g\big(g(t)\big)\Big) + \int_{G^*} H\Big(\vec{p}_f\big(f(t)\big)\Big) - H\Big(\vec{p}_g\big(g(t)\big)\Big)$$

or $(E) = (EB) + (EG)$.

Note that each piece $(LG), (LB), (EG), (EB)$ by itself is scale invariant. As discussed in section 5., some care must be taken to split $\mathtt{h}(f) - \mathtt{h}(g)$ into scale invariant pieces.

In preparation for estimating the size of each of the four pieces, in Sections 9. and 10. we develop some results pursuing the implications of the regularity assumption. These implications are both quantitative and qualitative. The quantitative implications are pursued in Section 9.. In Section 10., we define the space $E(\tau)$ (see Definition 11), which is meant to abstract just some "qualitative" or "incidence" properties which follow from the quantitative results, and we prove some results about $E(\tau)$.

In section 11., we apply the results from Sections 9. and 10. to f and g. In brief, the bound on each piece is as follows.

Looking first at (L), the bound for (LG) is immediate, so we concentrate on (LB). The regularity assumption (Definition 6) will imply that critical points of f and g roughly coincide; we consider the portion of B^* near just one critical point, call this B^*_j. As $\epsilon \to 0$, certainly $|B^*_j| \to 0$, but B^*_j is where $|f'|$ or $|g'|$ are small (or zero), so $\log|f'|, \log|g'|$ can be unbounded and we cannot say that (LB) is negligible just because $|B^*_j|$ is small. However, we observe that f can be approximated by a parabola p_f near a critical point, and the same is true for g. We have

$$\int_{B^*_j} \log|f'| = \int_{B^*_j} \log|p'_f| + \int_{B^*_j} \log|f'/p'_f|$$

and it is only $\log|p'_f|$ which is unbounded; the estimate of (LB) is completed by proving a bound on

$$\int_{B^*_j} \log|p'_f| - \int_{B^*_j} \log|p'_g|$$

We turn next to (EB). This piece is bounded by "crude" size estimates, the main point being that the set $(G^*)^c$ is small. The integrand,

$$H\Big(\vec{p}_f\big(f(t)\big)\Big) - H\Big(\vec{p}_g\big(g(t)\big)\Big)$$

is bounded in size by the logarithm of one plus the number of critical points of f (the regularity assumption will imply that f and g have the same number of critical points). The size of the set $(G^*)^c$ is bounded just by the number of components of B^*, the size of B^*, and some basic data about the functions f and g.

Finally, we have (EG). We in fact only need some "incidence" data about the functions f and g to prove the result here (this data is summarized in Definition 11). The bound on this piece proceeds essentially as follows: Once we consider only $t \in G^*$, we can say

$$\vec{p}_f\big(f(t)\big) = \{p_i\}_{i=0}^k\ , \quad \vec{p}_g\big(g(t)\big) = \{q_i\}_{i=0}^k$$

and we need to bound $|\sum_{i=0}^k p_i \log p_i - q_i \log q_i|$. On G^*, we will see that p_i, q_i are bounded below, so we use the inequality

$$|p_i \log p_i - q_i \log q_i| \le [\ 1 + \max|\log p_i|, |\log q_i|\ ]\ |p_i - q_i|$$

and the proof is completed by estimating the difference

$$|p_i - q_i| = \left|\frac{\rho_{f_i}}{\rho_f} - \frac{\rho_{g_i}}{\rho_g}\right|$$

which ultimately reduces to the fact that $f'\Big(f_j^{-1}(\ f(t)\ )\Big)$ and $g'\Big(g_j^{-1}(\ g(t)\ )\Big)$ are close together because $|f(t)-g(t)| \leq \|f-g\|_\infty$, and f_j^{-1}, g_j^{-1} are going to be close together when evaluated on points that are close, and finally f', g' differ by at most $\|f'-g'\|_\infty$ when evaluated on the same point, so they are going to be close together when evaluated on points that are close.

In the following sections, the functions "f" and "g" will always refer to the two specific functions assumed by Theorem 1. We will use h and l to refer to generic functions.

9. The Space $A(\delta)$

In the following, we assume $h, l \in A(\delta)$.

Lemma 7. *If* $|h''(t_0)| \geq A > 0$ *and* $|t-t_0| \leq A/(2\|h^{(3)}\|_\infty)$*, then* $|h''(t)| \geq A/2$.

$|h''(t)-h''(t_0)| \leq \|h^{(3)}\|_\infty |t-t_0|$.

Lemma 8. *Distinct zeros of* h' *are separated by a distance of at least* $\delta/2$*. This also means* h' *has at most* $2/\delta$ *zeros on*$[0,1]$.

$h'(t_0)=0$ implies that $|h''(t_0)| \geq \delta\|h^{(3)}\|_\infty$, and so

$$|t-t_0| \leq \delta/2 \implies |t-t_0| \leq (\delta\|h^{(3)}\|_\infty)/(2\|h^{(3)}\|_\infty)$$

and then Lemma 7 implies $|h''(t)| > 0$. Since h'' is continuous, h'' does not change sign between t and t_0, so a zero of h' must be further than $\delta/2$ away from t_0.

Lemma 9. *If* $\gamma \leq \delta^2/16$ *and* $|h'(t_0)| \leq \gamma\|h^{(3)}\|_\infty$ *then* $\exists t, |t-t_0| \leq 4\gamma/\delta$ *with* $h'(t)=0$.

By definition of $A(\delta)$, we know that $t_0 \in (\delta, 1-\delta)$ and $|h''(t_0)| \geq (\delta/2)\|h^{(3)}\|_\infty$. By Lemma 7 we know that h'' has the same sign on

$$[t_0-\delta/4, t_0+\delta/4]$$

and $|h''|$ is greater than $(\delta/4)\|h^{(3)}\|_\infty$ on that interval. Without loss of generality, suppose that

$$0 \leq h'(t_0) \leq \gamma\|h^{(3)}\|_\infty \ \ , \ \ h''(t_0) < 0$$

Let $t = t_0 + 4\gamma/\delta$. We see that $4\gamma/\delta \leq \delta/4$ which means $t \in [0,1]$ and $t \in [t_0-\delta/4, t_0+\delta/4]$. So, on $[t_0, t]$, we have $h'' \leq -(\delta/4)\|h^{(3)}\|_\infty$. Therefore,

$$\begin{aligned} h'(t) &= h'(t_0) + \int_{t_0}^{t} h'' \\ &\leq \gamma\|h^{(3)}\|_\infty + \int_{t_0}^{t} -(\delta/4)\|h^{(3)}\|_\infty \\ &= \gamma\|h^{(3)}\|_\infty - (4\gamma/\delta)\ (\delta/4)\ \|h^{(3)}\|_\infty \\ &= 0 \end{aligned}$$

So $h'(t_0) \geq 0$ and $h'(t) \leq 0$ and it is proved.

Lemma 10. *If*

$$\|h' - l'\|_\infty < \left(\delta^2/16\right) \min\left\{\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty\right\}$$

then

$$h'(t_0) = 0 \quad \Longrightarrow \quad \exists t' \ \textit{with} \ l'(t') = 0, \ \ |t' - t_0| \leq 4 \frac{\|h' - l'\|_\infty}{\delta \min\left\{\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty\right\}} \tag{15}$$

$$nonumber \tag{16}$$

$$l'(s_0) = 0 \quad \Longrightarrow \quad \exists s' \ \textit{with} \ h'(s') = 0, \ \ |s' - s_0| \leq 4 \frac{\|h' - l'\|_\infty}{\delta \min\left\{\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty\right\}}$$

This is a corollary of Lemma 9; if $h'(t_0) = 0$ then

$$|l'(t_0)| < \left(\delta^2/16\right) \min\left\{\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty\right\}$$

and we apply Lemma 9 to l. We proceed analogously if we assume $l'(t_0) = 0$.

10. The Space $E(\tau)$

Definition 11. *We say that $h : [0,1] \longrightarrow \mathbb{R}$ is a member of $E(\tau)$ if $h, h', h'', h^{(3)}$ are continuous and*

$$\left[t_1 \neq t_2 \textit{ and } h'(t_1) = h'(t_2) = 0\right] \implies |t_1 - t_2| > 2\tau \tag{17}$$

$$t \in [0, \tau] \cup [1 - \tau, 1] \implies h'(t) \neq 0 \tag{18}$$

Definition 12. *For $h, l \in E(\tau)$, we say $h \sim l$ if the following two statements are true:*

$$\left[h'(t) = 0 \implies \exists t', |t - t'| < \tau, l'(t') = 0\right]$$
$$\left[l'(s) = 0 \implies \exists s', |s - s'| < \tau, h'(s') = 0\right]$$

The set $E(\tau)$ just isolates some of the "incidence" properties of the critical points of h and l; we will make use of it because of the following two lemmas which connect it to $A(\delta)$:

Lemma 13. *If $h \in A(\delta)$ then there exists τ such that $h \in E(\tau)$.*

This follows directly from the definition of $A(\delta)$ and Lemma 8.

Lemma 14. *If $h, l \in A(\delta)$ and*

$$\|h' - l'\|_\infty < \left(\delta^2/16\right) \min\left\{\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty\right\}$$

then there exists τ such that $h, l \in E(\tau)$ and $h \sim l$ in $E(\tau)$.

This follows directly from Lemma 10 and Lemma 14.

Lemma 15. *If $h, l \in E(\tau)$ and $h \sim l$, then h and l have the same number of critical points, and if $h' = 0$ at $t_1 < \cdots < t_k$ and $l' = 0$ at $s_1 < \cdots < s_k$ then*

$$\forall i = 1, \ldots, k-1, \ \max\{t_i, s_i\} < \min\{t_{i+1}, s_{i+1}\}$$

Define the function Φ,

$$\Phi : \{ \text{ critical points of } h \ \} \longrightarrow \{ \text{ critical points of } l \ \}$$

as follows: For a critical point t of h, let

$$\Phi(t) = t' \qquad \text{where } |t' - t| < \tau \ , \ l'(t') = 0$$

It is easy to check that Φ is well defined and is, in fact, a bijection. It follows that $s_i = \Phi(t_i)$, and so the second conclusion follows from the fact that $|t_{i+1} - t_i| > 2\tau$, $|t_i - s_i| < \tau$, $|t_{i+1} - s_{i+1}| < \tau$.

Definition 16. *For functions h and l, and $B \subset [0,1]$, let*

$$\mathcal{I}_{h,l}(B) = \bigcup_{t \in B} \langle h(t), l(t) \rangle + [-\|h-l\|_\infty, \|h-l\|_\infty]$$
$$G_{h,l}(B) = [0,1] \setminus \left(h^{-1}(\mathcal{I}) \cup l^{-1}(\mathcal{I}) \right)$$

(recall the notation from Definition 1).

Lemma 17. *Suppose $h, l \in E(\tau)$, and $h \sim l$. By Lemma 15 suppose they have critical points $t_1 < \cdots < t_k$ and $s_1 < \cdots < s_k$ respectively, and define $h_i, l_i, t_0, s_0, t_{k+1}, s_{k+1}$ as in Definition 3. Suppose we have a set B such that*

$$B \supset \{0\} \cup \{1\} \cup \bigcup_i \langle t_i, s_i \rangle$$

Let $\mathcal{I} = \mathcal{I}_{h,l}(B)$ and $G = G_{h,l}(B)$ be defined as in Definition 16. Fix $i \in \{0, \ldots, k\}$ and $t \in G$. Then the following four statements are equivalent:

$$\begin{array}{ll} 1) & h(t) \in \textit{Range}(h_i) \\ 2) & h(t) \in \textit{Range}(l_i) \\ 3) & l(t) \in \textit{Range}(h_i) \\ 4) & l(t) \in \textit{Range}(l_i) \end{array}$$

Suppose that 1) holds and $h(t) \in \text{Range}(h_i)$. Without loss of generality, suppose that $h_i' > 0$. We know

$$h(t) \notin \left[\min_{\langle t_i, s_i \rangle} \{h, l\} - \|h-l\|_\infty, \max_{\langle t_i, s_i \rangle} \{h, l\} + \|h-l\|_\infty \right]$$

since that set is in $\mathcal{I}$ and $t \in G$. Since $h(t) \geq h(t_i)$, we know $h(t) \geq \max_{\langle t_i, s_i \rangle} \{h, l\} + \|h-l\|_\infty$. Also,

$$l(t) \notin \left[\min_{\langle t_i, s_i \rangle} \{h, l\} - \|h-l\|_\infty, \max_{\langle t_i, s_i \rangle} \{h, l\} + \|h-l\|_\infty \right]$$

for the same reason, and we cannot have $l(t) < \min_{\langle t_i, s_i \rangle} \{h, l\} - \|h-l\|_\infty$ since that would imply

$$|h(t) - l(t)| > \|h-l\|_\infty$$

So, $l(t) \geq \max_{\langle t_i, s_i \rangle} \{h, l\} + \|h-l\|_\infty$. By similar logic,

$$h(t), l(t) \leq \min_{\langle t_{i+1}, s_{i+1} \rangle} \{h, l\} - \|h-l\|_\infty$$

So, we see that

$$\max_{\langle t_i, s_i \rangle} \{h, l\} + \|h - l\|_\infty \leq h(t), l(t) \leq \min_{\langle t_{i+1}, s_{i+1} \rangle} \{h, l\} - \|h - l\|_\infty$$

So, the left hand side is less than or equal to the right hand side and we have $h(t), l(t) \in \text{Range}(h_i), \text{Range}(l_i)$. This proves

$$1) \implies 2), 3), 4)$$

The other implications are seen in the same manner.

Lemma 18. *Under the same assumptions as Lemma 17, for $t \in G$ we have*

$$\langle h(t), l(t) \rangle \subset \mathcal{I}^c$$

and so

$$h^{-1}\Big(\langle h(t), l(t) \rangle\Big), l^{-1}\Big(\langle h(t), l(t) \rangle\Big) \subset B^c$$

This follows from the observation that $\mathcal{I}$ contains no intervals of length less than $\|h - l\|_\infty$, and $h(t), l(t) \notin \mathcal{I}$.

Lemma 19. *Under the same assumptions as Lemma 17, if we have $t \in G$, $u \in \langle h(t), l(t) \rangle$ and $h(t) \in$ Range(h_j) then*

$$\langle h_j^{-1}(u), l_j^{-1}(u) \rangle \subset \textit{Dom}(h_j) \cap \textit{Dom}(l_j) \cap B^c$$

and therefore neither h' nor l' changes sign on $\langle h_j^{-1}(u), l_j^{-1}(u) \rangle$.

We know by Lemma 17 that $u \in \text{Range}(h_j) \cap \text{Range}(l_j)$. Suppose $h_j(t_1) = u$. Then $t_1 \notin B$ by Lemma 18. From Lemma 15 we see that $\text{Dom}(h_j) \cap B^c = \text{Dom}(l_j) \cap B^c$. So,

$$\begin{aligned} & t_1 \in \text{Dom}(h_j) \cap B^c \\ \implies & t_1 \in \text{Dom}(l_j) \\ \implies & h_j^{-1}(u) \in \text{Dom}(h_j) \cap \text{Dom}(l_j) \cap B^c \end{aligned}$$

The same is true for $l_j^{-1}(u)$, and so $\langle h_j^{-1}(u), l_j^{-1}(u) \rangle \subset \text{Dom}(h_j) \cap \text{Dom}(l_j) \cap B^c$.

11. Applying the Results to f and g

We now return to the two specific functions f and g, which have the properties stated in the assumptions of Theorem 1. We introduce a variable, μ, which will allow us to divide $[0, 1]$ into areas where either f' or g' are "small" according to μ, specifically

$$|f'(t)| \leq \mu \|f^{(3)}\|_\infty \text{ or } |g'(t)| \leq \mu \|g^{(3)}\|_\infty$$

and areas where both f' and g' are "large" according to μ, specifically

$$|f'(t)| \geq \mu \|f^{(3)}\|_\infty \text{ and } |g'(t)| \geq \mu \|g^{(3)}\|_\infty$$

At the end of the proof, we will optimize μ; it will turn out that $\mu = \epsilon^{1/4}$ is optimal.

We state here some restrictions on the values of μ that we will consider:

$$\epsilon^{3/4} < \mu \leq (1/32)\delta^2 \tag{19}$$

Notice that, since we will set $\mu = \epsilon^{1/4}$, we will require $\epsilon \leq 2^{-20}\delta^8$.

We know that $f, g \in A(\delta)$, and we also know that $\|f' - g'\|_\infty < \delta^2/16$ by (14) and (19). This means that we can apply the results of section 9. to f and g, and with Lemmas 13 and 14 we see that we can also apply the results of section 10..

With this in mind, by Lemma 15 let us suppose that $f' = 0$ at $t_1^* < \cdots < t_k^*$ and $g' = 0$ at $s_1^* < \cdots < s_k^*$. Next, we want to construct a region covering the critical points of f and g, and extending beyond the critical points far enough to cover the entire area where f' or g' is "small" (according to μ).

Definition 20. *For $i = 1, \ldots, k$, let $J_i^* = [\min\{t_i^*, s_i^*\} - 4\mu/\delta, \max\{t_i^*, s_i^*\} + 4\mu/\delta]$.*

We see that the critical points of f and g are contained in the J_i^*. Furthermore, suppose that $|f'(t_0)| \leq \mu\|f^{(3)}\|_\infty$. By Lemma 9, we see that the distance from t_0 to a critical point of f is no more than $4\mu/\delta$, which means that t_0 is contained in some J_i^*. The same is true for g, and therefore we have

$$\{t : |f'(t)| \leq \mu\|f^{(3)}\|_\infty \text{ or } |g'(t)| \leq \mu\|g^{(3)}\|_\infty\} \subset \bigcup_{i=1}^{k} J_i^* \tag{20}$$

For purposes of the expression $H\Big(\vec{p}_f\big(f(t)\big)\Big)$, the "bad" set will also include the endpoints of the interval, so we define the sets B^* and G^* as follows.

Definition 21.

$$B^* = \{0\} \cup \{1\} \cup \bigcup_{i=1}^{k} J_i^*$$
$$G^* = G_{f,g}(B^*)$$

Notice that $G^* \cap B^* = \emptyset$, but we do not necessarily have $G^* \cup B^* = [0, 1]$.

We are now ready to state how we will decompose the expression $\mathtt{h}(f) - \mathtt{h}(g)$ in order to estimate it. First, recall by equation (7) that

$$\mathtt{h}(f) - \mathtt{h}(g) = \int_t \log|f'(t)| - \log|g'(t)| \; - \left(\int_t H\Big(\vec{p}_f\big(f(t)\big)\Big) - H\Big(\vec{p}_g\big(g(t)\big)\Big) \right)$$

We will divide $[0, 1]$ in different ways to estimate the different integrals:

$$\text{For } \int_t \log|f'(t)| - \log|g'(t)| \text{ we write } [0,1] = \bigcup J_i^* \cup \Big(\bigcup J_i^*\Big)^c$$
$$\text{For } \int_t H\Big(\vec{p}_f\big(f(t)\big)\Big) - H\Big(\vec{p}_g\big(g(t)\big)\Big) \text{ we write } [0,1] = G^* \cup (G^*)^c$$

This allows us to decompose our main expression, $\mathtt{h}(f) - \mathtt{h}(g)$, into the following four pieces:

Definition 22.

$$(LB) = \int_{\bigcup J_i^*} \log|f'(t)| - \log|g'(t)|$$
$$(LG) = \int_{\left(\bigcup J_i^*\right)^c} \log|f'(t)| - \log|g'(t)|$$
$$(EB) = \int_{(G^*)^c} H\Big(\vec{p}_f\big(f(t)\big)\Big) - H\Big(\vec{p}_g\big(g(t)\big)\Big)$$
$$(EG) = \int_{G^*} H\Big(\vec{p}_f\big(f(t)\big)\Big) - H\Big(\vec{p}_g\big(g(t)\big)\Big)$$

Next, we have four propositions, each estimating one of the above expressions in a certain setting. Each proposition will be applicable to the functions f and g, and the sets J_i^*, B^* and G^*, but it will be helpful to state and prove the propositions using only the relevant information in each case.

In order to not interrupt our exposition, we have placed the proofs of these propositions in the appendix.

Proposition 23 (Used to bound (LB))**.** *Suppose that $h, l \in A(\delta)$ and*

$$\|h' - l'\|_\infty < \left(\delta^2/16\right) \min\left\{\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty\right\}$$

By Lemma 15, let their critical points be $t_1 < \cdots < t_k$ and $s_1 < \cdots < s_k$ respectively. Let

$$J_i = [\min\{t_i, s_i\} - a, \max\{t_i, s_i\} + a]$$

and suppose that we know $\forall i, |t_i - s_i| + a \leq \delta/4$. Then

$$\left|\int_{\bigcup J_i} \log|h'(t)| - \log|l'(t)|\right| \leq \sum_i \left[(|t_i - s_i| + 2a)\frac{\|h'' - l''\|_\infty}{\delta \min\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty} + (8/\delta)(|t_i - s_i| + 2a)^2 \right]$$

Proposition 24 (Used to bound (LG))**.** *Suppose that h and l have three continuous derivatives on $[0, 1]$. Let*

$$X = \left\{t : |h'(t)| \geq \mu\|h^{(3)}\|_\infty \text{ and } |l'(t)| \geq \mu\|l^{(3)}\|_\infty\right\}$$

Then

$$\left|\int_X \log|h'(t)| - \log|l'(t)|\right| \leq \frac{\|h' - l'\|_\infty}{\mu \min\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty}$$

This follows immediately from Lemma 27.

Proposition 25 (Used to bound (EB))**.** *Suppose that h, l have three continuous derivatives and have finite numbers, k_h and k_l, of critical points respectively, and $h'(0), l'(0), h'(1), l'(1) \neq 0$. Suppose that we have closed intervals $\{L_\alpha\}_{\alpha\in\mathcal{A}}$ and we define $B = \bigcup_{\alpha\in\mathcal{A}} L_\alpha$ and $G = G_{h,l}(B)$ as in Definition 16. Then*

$$|G^c| \leq (k_h + k_l + 2)\left(4\,|\mathcal{A}|\,\|h - l\|_\infty + \sum_\alpha \left(\max_{L_\alpha}|h'|, |l'|\right)|L_\alpha|\right)\left(\min_{B^c}|h'|, |l'|\right)^{-1} + |B|$$

Proposition 26 (Used to bound (EG))**.** *Suppose that $h, l \in E(\tau)$ and $h \sim l$ in $E(\tau)$. By Lemma 15, let their critical points be $t_1 < \cdots < t_k$ and $s_1 < \cdots < s_k$ respectively. Suppose we have a set B such that*

$$B \supset \{0\} \cup \{1\} \cup \bigcup_i \langle t_i, s_i \rangle$$

and define $G = G_{h,l}(B)$ as in Definition 16. Then

$$\begin{aligned}
&\left| \int_G H\Big(\vec{p}_h\big(h(t)\big)\Big) - H\Big(\vec{p}_l\big(l(t)\big)\Big) \right| \\
&\leq (k+1)(k+2)\left[1 + \left|\log \frac{\min_{B^c} |h'|, |l'|}{(k+1)\max \|h'\|_\infty, \|l'\|_\infty}\right|\right] (\max \|h'\|_\infty, \|l'\|_\infty) \left(\min_{B^c} |h'|, |l'|\right)^{-2} \\
&\times \left\{ \|h' - l'\|_\infty + 2\|h - l\|_\infty \left(\min_{B^c} |h'|, |l'|\right)^{-1} \frac{\|h''\|_\infty + \|l''\|_\infty}{2} \right\}
\end{aligned}$$

Now, we apply these propositions to f and g. We will abbreviate

$$\Gamma_1 = \frac{\max \|f'\|_\infty, \|g'\|_\infty}{\min \{\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty\}}, \Gamma_2 = \frac{\max \|f''\|_\infty, \|g''\|_\infty}{\min \{\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty\}}$$

The assumptions of Theorem 1 mean that

$$C_1^{-1} \leq \Gamma_1 \leq C_1, \Gamma_2 \leq C_2$$

and without loss of generality we may assume $C_1, C_2 \geq 3$.

We know that $\|f' - g'\|_\infty < \delta^2/16$ by (14) and (19), and therefore by Lemma 10 and equation (14) we have

$$\forall i, |t_i^* - s_i^*| \leq 4\epsilon^{3/4}/\delta \tag{21}$$

To apply Proposition 23 (with $a = 4\mu/\delta$), we must have

$$4\epsilon^{3/4}/\delta + 4\mu/\delta \leq \delta/4$$

and this is implied by (19). So, with some simplifying to arrive at the second equation, we have

$$\begin{aligned}
\big|(LB)\big| &\leq \sum_i \left[(|t_i - s_i| + 2 \cdot 4\mu/\delta) \frac{\|f'' - g''\|_\infty}{\delta \min \|f^{(3)}\|_\infty, \|g^{(3)}\|_\infty} + (8/\delta)(|t_i - s_i| + 2 \cdot 4\mu/\delta)^2 \right] \\
&\leq k(4\epsilon^{3/4} + 8\mu)\delta^{-2} \left[\frac{\|f'' - g''\|_\infty}{\min \|f^{(3)}\|_\infty, \|g^{(3)}\|_\infty} + (8/\delta)(4\epsilon^{3/4} + 8\mu) \right] \\
&\leq 12k\mu/\delta^2 \left[\Gamma_2 + 96\mu/\delta\right] \\
&\leq 12kC_2 \cdot 97\mu/\delta^2
\end{aligned} \tag{22}$$

Next, looking to Proposition 24, recalling equation (20) we have

$$\left(\bigcup_{i=1}^k J_i^* \right)^c \subset \left\{ t : |f'(t)| \geq \mu \|f^{(3)}\|_\infty \text{ and } |g'(t)| \geq \mu \|g^{(3)}\|_\infty \right\}$$

and so we can apply Proposition 24 to f and g, which gives

$$\begin{aligned}\big|(LG)\big| &\leq \frac{\|f'-g'\|_\infty}{\mu \min \|f^{(3)}\|_\infty, \|g^{(3)}\|_\infty} \\ &\leq \epsilon^{3/4}/\mu \end{aligned} \tag{23}$$

Looking to Proposition 25, we have $B^* = \{0\} \cup \{1\} \cup \bigcup_{i=1}^k J_i^* = \bigcup L_\alpha$. Equation (20) means that

$$B^* \supset \left\{t : |h'(t)| \leq \mu \|h^{(3)}\|_\infty \text{ or } |l'(t)| \leq \mu \|l^{(3)}\|_\infty\right\}$$

which means

$$\left(\min_{(B^*)^c} |h'|, |l'|\right)^{-1} \leq \left[\mu \min\left\{\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty\right\}\right]^{-1}$$

In the notation of Proposition 25, $k_f, k_g = k$ and $|\mathcal{A}| = k+2$, and aside from $\{0\}$ and $\{1\}$ which have measure zero, $|J_i| \leq (4\epsilon^{3/4}/\delta) + 4\mu/\delta \leq 8\mu/\delta$. Also, note that

$$(\max_{J_i} |h'|, |l'|) \leq (\max \|f''\|_\infty, \|g''\|_\infty) \cdot |J_i| \leq 8\mu/\delta (\max \|f''\|_\infty, \|g''\|_\infty)$$

which means

$$\begin{aligned}|G^c| &\leq (k_h + k_l + 2)\left(4|\mathcal{A}|\|h-l\|_\infty + \sum_\alpha \left(\max_{L_\alpha} |h'|, |l'|\right)|L_\alpha|\right)\left(\min_{B^c} |h'|, |l'|\right)^{-1} + |B| \\ &\leq 2(k+1)\frac{4(k+2)\epsilon \min\left\{\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty\right\} + k(8\mu/\delta)^2 \max \|f''\|_\infty, \|g''\|_\infty}{\mu \min\left\{\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty\right\}} + k(8\mu/\delta) \\ &\leq 8(k+2)^2\epsilon/\mu + 2^7 k(k+1)\mu/\delta^2 (\Gamma_2) + 8k\mu/\delta \\ &\leq 2^7(k+2)^2 C_2(\epsilon/\mu + \mu/\delta^2 + \mu/\delta) \\ &\leq 2^8(k+2)^2 C_2(\epsilon/\mu + \mu/\delta)\end{aligned}$$

This means

$$(EB) \leq 2^8(k+2)^2 C_2(\epsilon/\mu + \mu/\delta)\log(k+1) \tag{24}$$

Looking to Proposition 26, by Lemma 14 and Definition 21 we can apply Proposition 26 to f and g, so we have

$$\begin{aligned}&\left|\int_{G^*} H\Big(\vec{p}_f(f(t))\Big) - H\Big(\vec{p}_g(g(t))\Big)\right| \\ &\leq (k+1)(k+2)\left[1 + \left|\log \frac{\min_{B^c} |f'|, |g'|}{(k+1)\max \|f'\|_\infty, \|g'\|_\infty}\right|\right](\max \|f'\|_\infty, \|g'\|_\infty)\left(\min_{B^c} |f'|, |g'|\right)^{-2} \\ &\quad \times \left\{\|f'-g'\|_\infty + 2\|f-g\|_\infty \left(\min_{B^c} |f'|, |g'|\right)^{-1} \frac{\|f''\|_\infty + \|g''\|_\infty}{2}\right\} \\ &\leq (k+2)^2\left[1 + \left|\log \frac{\mu \min\left\{\|f^{(3)}\|_\infty, \|g^{(3)}\|_\infty\right\}}{(k+1)\max \|f'\|_\infty, \|g'\|_\infty}\right|\right](\max \|f'\|_\infty, \|g'\|_\infty)\left(\mu \min\left\{\|f^{(3)}\|_\infty, \|g^{(3)}\|_\infty\right\}\right)^{-2} \\ &\quad \times \left\{\epsilon^{3/4} \min\left\{\|f^{(3)}\|_\infty, \|g^{(3)}\|_\infty\right\} + 2\epsilon \max \|f''\|_\infty, \|g''\|_\infty\right\} \\ &\leq (k+2)^2 \Gamma_1\Big(\epsilon^{3/4}/\mu^2 + 2\epsilon\Gamma_2\Big)\left[1 + \left|\log \frac{\mu}{(k+1)\Gamma_1}\right|\right] \\ &\leq 6C_1C_2(k+2)^2 \log\left(C_1(k+1)\mu^{-1}\right)\epsilon^{3/4}/\mu^2\end{aligned} \tag{25}$$

Simplifying the logarithm term in the last line was the reason for assuming $C_1 \geq 3$.

Combining equations (22), (23), (24), (25), and letting C denote an absolute constant, we have

$$|\mathtt{h}(f) - \mathtt{h}(g)| \leq CC_1C_2(k+2)^2 \log\left(C_1(k+1)\mu^{-1}\right) \left[\mu/\delta^2 + \epsilon^{3/4}/\mu + \epsilon/\mu + \mu/\delta + \epsilon^{3/4}/\mu^2\right]$$

Examining just the last expression in the product, again letting C be an absolute constant, we have

$$\mu/\delta^2 + \epsilon^{3/4}/\mu + \epsilon/\mu + \mu/\delta + \epsilon^{3/4}/\mu^2 \leq C(\mu/\delta^2 + \epsilon^{3/4}/\mu^2)$$

Treating δ as a fixed parameter, this expression is optimized if $\mu = \epsilon^{1/4}$, and using Lemma 8 we see that $k \leq 2/\delta$. Therefore, with C being another absolute constant, we have

$$|\mathtt{h}(f) - \mathtt{h}(g)| \leq CC_1C_2 \log\left(C_1\delta^{-1}\epsilon^{-1/4}\right) \epsilon^{1/4}/\delta^4$$

Acknowledgements

We would like to thank the referees, and Michael S. Hughes, for many helpful comments.

References

1. Hughes, M. A comparison of shannon entropy versus signal energy for acoustic detection of artificially induced defects in plexiglass. *J. Acoust. Soc. Am.* **1992**, *91*, 2272–2275.
2. Hughes, M. Analysis of digitized waveforms using shannon entropy. *J. Acoust. Soc. Am.* **1993**, *93*, 892–906.
3. Hughes, M. Analysis of digitized waveforms using shannon entropy II. High-speed algorithms based on Green's functions. *J. Acoust. Soc. Am.* **1994**, *95*, 2582–2588.
4. Hughes, M.; Marsh, J.; Hall, C.; Savy, D.; Scott, M.; Allen, J.; Lacy, E.; Carradine, C.; Lanza, G.; Wickline, S. Characterization of digital waveforms using thermodynamic analogs: Applications to detection of material defects. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **2005**, *52*, 555–1564.
5. Hughes, M.; Marsh, J.; Zhang, H.; Woodson, A.; Allen, J.; Lacy, E.; Carradine, C.; Lanza, G.; Wickline, S. Characterization of digital waveforms using thermodynamic analogs: Detection of contrast-targeted tissue *in vivo*. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **2006**, *53*, 1609–1616.
6. Hughes, M.; Mc Carthy, J.; Marsh, J.; Arbeit, J.; Neumann, R.; Fuhrhop, R.; Wallace, K.; Znidersic, D.; Maurizi, B.; Baldwin, S.; Lanza, G.; Wickline, S. Properties of an entropy-based signal receiver with an application to ultrasonic molecular imaging. *J. Acoust. Soc. Am.* **2007**, *121*, 3542–3557.
7. Hughes, M.; Mc Carthy, J.; Wickerhauser, M.V.; Marsh, J.; Arbeit,J.; Fuhrhop, R.; Wallace, K.; Thomas, T.; Smith, J.; Agyem, K.; Lanza, G.; Wickline, S. Real-time calculation of the limiting form of the Renyi entropy applied to detection of subtle changes in scattering architecture. *J. Acoust. Soc. Am.* **2009**, *126*, 2350–2358.
8. Beirlant, J.; Dudewicz, E.; Gyorfi, L.; van der Meulen, E. Nonparametric entropy estimation: An overview. *Int. J. Math. Stat. Sci.* **1997**, *6*, 17–39.

9. Maurizi, B. Noise Sensitivity of An Entropy-Based Signal Receiver. Ph.D. thesis, Washington University in Saint Louis, May 2008.
10. Goria, M.; Leonenko, N.; Mergel, V.; Inverardi, P. A new class of random vector entropy estimators and its applications in Testing Statistical Hypotheses. *J. Nonparametr. Statist.* **2005**, *17*, 277–297.
11. van Es, B. Estimating functionals related to a density by a class of statistics based on spacings. *Scand. J. Stat.* **1992**, *19*, 61–72.
12. Joe, H. Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Stat. Math.* **1989**, *41*, 683–697.
13. Levit, B. Asymptotically efficient estimation of nonlinear functionals. *Probl. Inform. Transm.* **1978**, *14*, 204–209.
14. Leonenko, N.; Pronzato, L.; Savani, V. A class of renyi information estimators for multidimensional densities. *Ann. Statist* **2008**, *36*, 2153–2182.
15. Tsybakov, A.; van der Meulen, E. Root-n consistent estimators of entropy for densities with unbounded support. *Scand. J. Stat.* **1996**, *23*, 75–83.
16. Eggermont, P.; LaRiccia, V. Best asymptotic normality of the kernel density entropy estimator for smooth densities. *IEEE Trans. Inf. Theory* **1999**, *45*, 1321–1326.
17. Dmitriev, Y.; Tarasenko, F. On the estimation of functionals of the probability density and its derivatives. *Theory Probab. Appl.* **1973**, *18*, 628–633.
18. Ahmad, I.; Lin, P. A nonparametric estimation of the entropy for absolutely continuous distributions. *IEEE Trans. Inf. Theory* **1976**, *22*, 372–375.
19. Hall, P.; Morton, S. On the estimation of entropy. *Ann. Inst. Stat. Math.* **1993**, *45*, 69–88.
20. Mokkadem, A. Estimation of the entropy and information for absolutely continuous random variables. *IEEE Trans. Inf. Theory* **1989**, *35*, 193–196.
21. Joe, H. Sample estimate of the entropy of a random vector. *Ann. Inst. Stat. Math.* **1989**, *41*, 83–697.
22. Vasicek, O. A test for normality based on sample entropy. *J. Roy. Statist. Soc. Ser. B.* **1976**, *38*, 54–59.
23. Gyorfi, L.; van der Meulen, E. Density-Free convergence properties of various estimators of entropy. *Comput. Stat. Data Anal.* **1987**, *5*, 425–436.
24. Godavarti, M.; Hero, A. Convergence of differential entropies. *IEEE Trans. Inf. Theory* **2004**, *50*, 171–176.
25. Csiszár, I. On generalized entropy. *Studia Sci. Math. Hungar.* **1969**, *4*, 401–419.
26. Rényi, A. On the dimension and entropy of probability distributions. *Acta Math. Acad. Sci. Hungar.* **1959**, *10*, 193–215.
27. Rudin, W. *Real and Complex Analysis*, 3rd Edition; McGraw-Hill Book Company: New York, NY, USA, 1987.
28. O'Sullivan, J.A. Washington University in Saint Louis School of Engineering and Applied Science, Saint Louis, MS, USA. Personal Communication, 2007.
29. Cover, T.; Thomas, J. *Elements of Information Theory*, 2nd Edition; Wiley-Interscience [John Wiley and Sons]: Hoboken, NJ, USA, 2006.

12. Appendix

12.1. Miscellaneous Technical Lemmas

Lemma 27. *If $a \geq \bar{a} > 0$ and $b \geq \bar{b} > 0$, then*

$$\frac{1}{1+|a-b|/\bar{a}} \leq a/b \leq 1+|a-b|/\bar{b}$$

and therefore

$$\begin{aligned} |\log(a/b)| &\leq \log\left(1+\frac{|a-b|}{\min \bar{a},\bar{b}}\right) \\ &\leq \frac{|a-b|}{\min \bar{a},\bar{b}} \end{aligned}$$

Lemma 28. *If $a_i, b_i \geq 0$ and $\sum_{i=1}^n a_i = a > 0, \sum_{i=1}^n b_i = b > 0$ then*

$$|a_i/a - b_i/b| \leq \min(a,b)^{-1}(n+1)\max_i |a_i - b_i|$$

$$\begin{aligned} |a_i/a - b_i/b| &= (1/a)|a_i - b_i(a/b)| \\ &\leq \min(a,b)^{-1}\Big[|a_i - b_i| + |b_i - b_i(a/b)|\Big] \\ &= \min(a,b)^{-1}\Big[|a_i - b_i| + (b_i/b)|b-a|\Big] \\ &\leq \min(a,b)^{-1}\Big[|a_i - b_i| + \sum_j |a_j - b_j|\Big] \\ &\leq \min(a,b)^{-1}(n+1)\max_i |a_i - b_i| \end{aligned}$$

12.2. Proofs of Main Propositions

Proof of Proposition 23

Proposition 23 states: Suppose that $h, l \in A(\delta)$ and

$$\|h' - l'\|_\infty < \left(\delta^2/16\right)\min\left\{\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty\right\}$$

By Lemma 15, let their critical points be $t_1 < \cdots < t_k$ and $s_1 < \cdots < s_k$ respectively. Let

$$J_i = [\min\{t_i, s_i\} - a, \max\{t_i, s_i\} + a]$$

and suppose that we know $\forall i, |t_i - s_i| + a \leq \delta/4$. Then

$$\left|\int_{\bigcup J_i} \log|h'(t)| - \log|l'(t)|\right| \leq \sum_i \left[(|t_i - s_i| + 2a)\frac{\|h'' - l''\|_\infty}{\delta \min \|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty} + (8/\delta)(|t_i - s_i| + 2a)^2\right]$$

Let us fix i and define the Taylor polynomials

$$\begin{aligned} p_h(t) &= h(t_i) + (h''(t_i)/2)(t - t_i)^2 \\ p_l(s) &= l(s_i) + (l''(s_i)/2)(s - s_i)^2 \end{aligned}$$

We see that

$$\left|\frac{h'(t)}{p_h'(t)} - 1\right| = \left|\frac{h'(t) - p_h'(t)}{p_h'(t)}\right| = \left|\frac{h'''(c)(t-t_i)^2}{2}\frac{1}{p_h'(t)}\right|$$

and we know $|p_h'(t)| = |h''(t_i)||t - t_i| \geq \delta\|h^{(3)}\|_\infty|t - t_i|$, so the above expression is

$$\begin{aligned}
&\leq \left|\frac{\|h^{(3)}\|_\infty(t-t_i)^2}{2}\frac{1}{\delta\|h^{(3)}\|_\infty|t-t_i|}\right| \\
&\leq |t - t_i|/(2\delta)
\end{aligned}$$

Note that, since $|t - t_i| \leq |t_i - s_i| + a$, by assumption the above expression is $\leq 1/2$. Therefore,

$$|\log\left|\frac{h'(t)}{p_h'(t)}\right|| \leq |t - t_i|/\delta$$

and so

$$\left|\int_{J_i} \log|h'(t)| - \log|p_h'(t)|\right| \leq (|t_i - s_i| + 2a)^2/\delta$$

This is also true for $\int_{J_i} \log|l'(t)| - \log|p_l'(t)|$, so it remains to bound $\int_{J_i} \log|p_h'(t)| - \log|p_l'(t)|$.

We will make some abbreviations to aid our exposition in this section: Let

$$\alpha = h''(t_i)\ , \qquad \beta = l''(s_i)$$

so we have $p_h'(t) = \alpha(t - t_i)$ and $p_l'(s) = \beta(s - s_i)$. Without loss of generality, suppose $t_i < s_i$ and abbreviate $\gamma = (s_i - t_i)/2$. Define $t_c = (t_i + s_i)/2$ so that we have

$$J_i = [t_c - \eta, t_c + \eta]$$

where $\eta = |t_i - s_i|/2 + a$. Now, we have

$$\begin{aligned}
&\int_{J_i} \log|p_h'(t)| - \log|p_l'(t)| \\
&\quad= \int_{t=t_c-\eta}^{t_c+\eta} \log|\alpha(t-t_i)| - \int_{t=t_c-\eta}^{t_c+\eta} \log|\beta(t-s_i)| \\
&\quad= \int_{t=t_c-t_i-\eta}^{t_c-t_i+\eta} \log|\alpha t| - \int_{t=t_c-s_i-\eta}^{t_c-s_i+\eta} \log|\beta t| \\
&\quad= \int_{\gamma-\eta}^{\gamma+\eta} \log|\alpha t| - \int_{-\gamma-\eta}^{-\gamma+\eta} \log|\beta t| \\
&\quad= \int_{\gamma-\eta}^{-\gamma+\eta} \log|\alpha t| - \log|\beta t| + \int_{-\gamma+\eta}^{\gamma+\eta} \log|\alpha t| - \int_{-\gamma-\eta}^{\gamma-\eta} \log|\beta t| \\
&\quad= \log(|\alpha|/|\beta|)(2\eta - 2\gamma) + 2\gamma\log|\alpha| - 2\gamma\log|\beta| + \int_{\eta-\gamma}^{\eta+\gamma} \log|t| - \int_{-\eta-\gamma}^{-\eta+\gamma} \log|t| \\
&\quad= 2\eta\log(|h''(t_i)|/|l''(s_i)|)
\end{aligned}$$

Now, $|h''(t_i)| \geq \delta\|h^{(3)}\|_\infty$ and $|l''(s_i)| \geq \delta\|l^{(3)}\|_\infty$, so by Lemma 27 we have

$$2\eta\,|\log(|h''(t_i)|/|l''(s_i)|)| \leq 2\eta\frac{\|h'' - l''\|_\infty}{\delta\min\{\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty\}}$$

¿From this, we conclude

$$\left|\int_{J_i} \log|p_h'(t)| - \log|p_l'(t)|\right| \leq (|t_i - s_i| + 2a)\,\frac{\|h'' - l''\|_\infty}{\delta \min\{\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty\}}$$

and therefore

$$\begin{aligned}
&\left|\int_{\bigcup J_i} \log|h'(t)| - \log|l'(t)|\right| \\
&\leq \sum_i \left|\int_{J_i} \log|h'(t)| - \log|p_h'(t)|\right| + \left|\int_{J_i} \log|p_h'(t)| - \log|p_l'(t)|\right| + \left|\int_{J_i} \log|p_l'(t)| - \log|l'(t)|\right| \\
&\leq \sum_i \left[\ (|t_i - s_i| + 2a)\,\frac{\|h'' - l''\|_\infty}{\delta \min\{\|h^{(3)}\|_\infty, \|l^{(3)}\|_\infty\}} + 2(|t_i - s_i| + 2a)^2/\delta\right]
\end{aligned}$$

Proof of Proposition 25

Proposition 25 states: Suppose that h, l have three continuous derivatives and have finite numbers, k_h and k_l, of critical points respectively, and $h'(0), l'(0), h'(1), l'(1) \neq 0$. Suppose that we have closed intervals $\{L_\alpha\}_{\alpha\in\mathcal{A}}$ and we define $B = \bigcup_{\alpha\in\mathcal{A}} L_\alpha$ and $G = G_{h,l}(B)$ as in Definition 16. Then

$$|G^c| \leq (k_h + k_l + 2)\left(4|\mathcal{A}|\|h - l\|_\infty + \sum_\alpha \left(\max_{L_\alpha} |h'|, |l'|\right)|L_\alpha|\right)\left(\min_{B^c} |h'|, |l'|\right)^{-1} + |B|$$

We begin by noting that

$$\begin{aligned}
G^c &= h^{-1}(\mathcal{I}) \cup l^{-1}(\mathcal{I}) \\
&\subset \left(h^{-1}(\mathcal{I}) \cap B^c\right) \cup \left(l^{-1}(\mathcal{I}) \cap B^c\right) \cup B \\
&= \left(\bigcup_j \left[h_j^{-1}(\mathcal{I}) \cap B^c\right]\right) \cup \left(\bigcup_j \left[l_j^{-1}(\mathcal{I}) \cap B^c\right]\right) \cup B
\end{aligned} \tag{26}$$

(recall Definition 3). For a monotone function λ, we can measure $|\lambda^{-1}(A)|$ by looking at $\int_A |(\lambda^{-1})'|$. In this case, we see that

$$h_j^{-1}(\mathcal{I}) \cap B^c = h_j^{-1}\big(\mathcal{I} \cap h_j(B^c)\big)$$

and therefore

$$\begin{aligned}
\left|h_j^{-1}(\mathcal{I}) \cap B^c\right| &= \int_{\mathcal{I}\cap h_j(B^c)} \left|(h_j^{-1})'(u)\right| du \\
&= \int_{\mathcal{I}\cap h_j(B^c)} \left|h'\left(h_j^{-1}(u)\right)\right|^{-1} du \\
&\leq \int_{\mathcal{I}\cap h_j(B^c)} \left(\min_{B^c} |h'|, |l'|\right)^{-1} du \\
&\leq |\mathcal{I}| \left(\min_{B^c} |h'|, |l'|\right)^{-1}
\end{aligned}$$

The same is true for l_j. By (26), this means

$$|G^c| \leq (k_h + k_l + 2)|\mathcal{I}| \left(\min_{B^c} |h'|, |l'|\right)^{-1} + |B|$$

So, it remains to estimate $|\mathcal{I}|$. We have

$$\begin{aligned}\mathcal{I} &= \bigcup_{t\in B}\langle h(t), l(t)\rangle + [-\|h-l\|_\infty, \|h-l\|_\infty] \\ &= \bigcup_{\alpha}\bigcup_{t\in L_\alpha}\langle h(t), l(t)\rangle + [-\|h-l\|_\infty, \|h-l\|_\infty]\end{aligned}$$

We now make the following observation:

$$\bigcup_{t\in L_\alpha}\langle h(t), l(t)\rangle + [-\|h-l\|_\infty, \|h-l\|_\infty] = \left[\min_{L_\alpha}\{h,l\} - \|h-l\|_\infty, \max_{L_\alpha}\{h,l\} + \|h-l\|_\infty\right]$$

This is a straightforward verification, the only non-trivial observation being that

$$\bigcup_{t\in L_\alpha}\langle h(t), l(t)\rangle + [-\|h-l\|_\infty, \|h-l\|_\infty]$$

is connected. We see that the size of this interval is bounded by

$$\begin{aligned}&\max_{L_\alpha}\{h,l\} - \min_{L_\alpha}\{h,l\} + 2\|h-l\|_\infty \\ &\qquad \le \left(\min\left\{\max_{L_\alpha}\{h\} - \min_{L_\alpha}\{h\}\ ,\ \max_{L_\alpha}\{l\} - \min_{L_\alpha}\{l\}\right\} + 2\|h-l\|_\infty\right) + 2\|h-l\|_\infty \\ &\qquad \le |L_\alpha|\max_{L_\alpha}\left\{|h'|, |l'|\right\} + 4\|h-l\|_\infty\end{aligned}$$

and so

$$|\mathcal{I}| \le \sum_\alpha\left(|L_\alpha|\max_{L_\alpha}\left\{|h'|, |l'|\right\} + 4\|h-l\|_\infty\right)$$

This means

$$\begin{aligned}|G^c| &\le (k_h + k_l + 2)|\mathcal{I}|\left(\min_{B^c}|h'|, |l'|\right)^{-1} + |B| \\ &\le (k_h + k_l + 2)\left(4|\mathcal{A}|\|h-l\|_\infty + \sum_\alpha\left(\max_{L_\alpha}|h'|, |l'|\right)|L_\alpha|\right)\left(\min_{B^c}|h'|, |l'|\right)^{-1} + |B|\end{aligned}$$

Proof of Proposition 26

Proposition 26 states: Suppose that $h, l \in E(\tau)$ and $h \sim l$ in $E(\tau)$. By Lemma 15, let their critical points be $t_1 < \cdots < t_k$ and $s_1 < \cdots < s_k$ respectively. Suppose we have a set B such that

$$B \supset \{0\} \cup \{1\} \cup \bigcup_i \langle t_i, s_i\rangle$$

and define $G = G_{h,l}(B)$ as in Definition 16. Then

$$\begin{aligned}&\left|\int_G H\Big(\vec{p}_h\big(h(t)\big)\Big) - H\Big(\vec{p}_l\big(l(t)\big)\Big)\right| \\ &\le (k+1)(k+2)\left[1 + \left|\log\frac{\min_{B^c}|h'|, |l'|}{(k+1)\max\|h'\|_\infty, \|l'\|_\infty}\right|\right](\max\|h'\|_\infty, \|l'\|_\infty)\left(\min_{B^c}|h'|, |l'|\right)^{-2} \\ &\times\left\{\|h'-l'\|_\infty + 2\|h-l\|_\infty\left(\min_{B^c}|h'|, |l'|\right)^{-1}\frac{\|h''\|_\infty + \|l''\|_\infty}{2}\right\}\end{aligned}$$

Let us abbreviate

$$A_1 = \int_G H\Big(\vec{p}_h\big(h(t)\big)\Big) - H\Big(\vec{p}_l\big(l(t)\big)\Big) \tag{27}$$

We have

$$A_1 = \int_G \sum_{i=0}^{k} \left[\frac{\rho_{h_i}(h(t))}{\rho_h(h(t))} \log\left\{ \frac{\rho_{h_i}(h(t))}{\rho_h(h(t))} \right\} - \frac{\rho_{l_i}(l(t))}{\rho_l(l(t))} \log\left\{ \frac{\rho_{l_i}(l(t))}{\rho_l(l(t))} \right\} \right]$$

Lemma 17 shows that, for $t \in G$, we have $[h(t) \in \text{Range}(h_i)] \Leftrightarrow [l(t) \in \text{Range}(l_i)]$, and therefore

$$A_1 = \sum_{i=0}^{k} \int_{\substack{t\in G:\\ h(t)\in\text{Range}(h_i)}} \left[\frac{\rho_{h_i}(h(t))}{\rho_h(h(t))} \log\left\{ \frac{\rho_{h_i}(h(t))}{\rho_h(h(t))} \right\} - \frac{\rho_{l_i}(l(t))}{\rho_l(l(t))} \log\left\{ \frac{\rho_{l_i}(l(t))}{\rho_l(l(t))} \right\} \right]$$

Then, estimating $|x_2 \log x_2 - x_1 \log x_1| \leq \Big[1 + \max\{|\log x_2|, |\log x_1|\}\Big]\ |x_2 - x_1|$, we see that

$$|A_1| \leq \sum_{i=0}^{k} \int_{\substack{t\in G:\\ h(t)\in\text{Range}(h_i)}} \left[1 + \max\left\{ \left|\log \frac{\rho_{h_i}(h(t))}{\rho_h(h(t))}\right|, \left|\log \frac{\rho_{l_i}(l(t))}{\rho_l(l(t))}\right| \right\}\right] \left| \frac{\rho_{h_i}(h(t))}{\rho_h(h(t))} - \frac{\rho_{l_i}(l(t))}{\rho_l(l(t))} \right|$$

We observe that

$$\min_{t\in[0,1]} \{\rho_{h_i}(h(t)), \rho_{l_i}(l(t))\} \geq \frac{1}{\max|h'|, |l'|}$$

Also note that, for $t \in G$, Lemma 18 means that $\forall i, h_i^{-1}(h(t))\ ,\ l_i^{-1}(l(t)) \notin B$, and therefore we have

$$\max_{t\in G} \{\rho_h(h(t)), \rho_l(l(t))\} \leq (k+1) \left[\min_{B^c} |h'|, |l'|\right]^{-1}$$

This means

$$\max_{i,t\in G} \max\left\{ \left|\log \frac{\rho_{h_i}(h(t))}{\rho_h(h(t))}\right|, \left|\log \frac{\rho_{l_i}(l(t))}{\rho_l(l(t))}\right| \right\} \leq \left|\log \frac{\min_{B^c}|h'|, |l'|}{(k+1)\max|h'|, |l'|}\right|$$

Abbreviating

$$A_2 = \frac{\rho_{h_i}(h(t))}{\rho_h(h(t))} - \frac{\rho_{l_i}(l(t))}{\rho_l(l(t))} \qquad t \in G,\ \ h(t) \in \text{Range}(h_i) \tag{28}$$

we have

$$|A_1| \leq \sum_{i=0}^{k} \int_{\substack{t\in G:\\ h(t)\in\text{Range}(h_i)}} \left[1 + \left|\log \frac{\min_{B^c}|h'|, |l'|}{(k+1)\max|h'|, |l'|}\right|\right] |A_2| \tag{29}$$

By Lemma 28,

$$|A_2| \leq (k+2)[\max|h'|, |l'|] \max_{\substack{t\in G\\ j:h(t)\in\text{Range}(h_j)}} |\rho_{h_j}(h(t)) - \rho_{l_j}(l(t))|$$

Then, using the observations

$$x, y > 0 \implies |1/x - 1/y| \leq [\min x, y]^{-2}|x - y|\ \ , \qquad \big||a| - |b|\big| \leq |a - b|$$

and noting that $\Big[t \in G, h(t) \in \text{Range}(h_j)\Big] \implies h_j^{-1}(h(t)), l_j^{-1}(l(t)) \notin B$, we have

$$\begin{aligned}
|A_2| &\leq (k+2)[\max|h'|,|l'|] \max_{\substack{t\in G\\ j:h(t)\in\text{Range}(h_j)}} \left| \frac{1}{\left|h'\Big(h_j^{-1}(h(t))\Big)\right|} - \frac{1}{\left|l'\Big(l_j^{-1}(l(t))\Big)\right|} \right| \\
&\leq (k+2)[\max|h'|,|l'|] \left[\min_{t\notin B} |h'(t)|, |l'(t)|\right]^{-2} \max_{\substack{t\in G\\ j:h(t)\in\text{Range}(h_j)}} \left| h'\Big(h_j^{-1}(h(t))\Big) - l'\Big(l_j^{-1}(l(t))\Big) \right| \\
&\leq (k+2) \frac{\max|h'|,|l'|}{[\min_{B^c}|h'|,|l'|]^2} \max_{\substack{t\in G\\ j:h(t)\in\text{Range}(h_j)}} \left| h'\Big(h_j^{-1}(h(t))\Big) - l'\Big(l_j^{-1}(l(t))\Big) \right|
\end{aligned}$$

Using the triangle inequality, we have

$$|h'\Big(h_j^{-1}(h(t))\Big) - l'\Big(l_j^{-1}(l(t))\Big)| \leq \|h'-l'\|_\infty + \frac{\|h''\|_\infty + \|l''\|_\infty}{2} |h_j^{-1}(h(t)) - l_j^{-1}(l(t))|$$

Abbreviating

$$A_3 = h_j^{-1}(h(t)) - l_j^{-1}(l(t)) \qquad t \in G, \;\; h(t) \in \text{Range}(h_j) \;\; , \tag{30}$$

we have

$$|A_2| \leq (k+2) \frac{\max|h'|,|l'|}{[\min_{B^c}|h'|,|l'|]^2} \max_{\substack{t\in G\\ j:h(t)\in\text{Range}(h_j)}} \left(\|h'-l'\|_\infty + \frac{\|h''\|_\infty + \|l''\|_\infty}{2} |A_3| \right) \tag{31}$$

Next, recalling Lemma 17 (which verifies that the expressions in the first following inequality are well-defined) we have:

$$\begin{aligned}
|A_3| &\leq |h_j^{-1}(h(t)) - h_j^{-1}(l(t))| + |h_j^{-1}(l(t)) - l_j^{-1}(l(t))| \\
&\leq \Big[\max_{\substack{t\in G\\ j:h(t)\in\text{Range}(h_j)\\ u\in\langle h(t),l(t)\rangle}} |(h^{-1})'(u)| \Big] \|h-l\|_\infty + \max_{\substack{t\in G\\ u\in\langle h(t),l(t)\rangle}} |h_j^{-1}(u) - l_j^{-1}(u)|
\end{aligned}$$

We see that

$$\begin{aligned}
\max_{\substack{t\in G\\ j:h(t)\in\text{Range}(h_j)\\ u\in\langle h(t),l(t)\rangle}} |(h^{-1})'(u)| &= \max_{\substack{t\in G\\ j:h(t)\in\text{Range}(h_j)\\ u\in\langle h(t),l(t)\rangle}} \left|h'(h_j^{-1}(u))\right|^{-1} \\
&= \Big[\min_{\substack{t\in G\\ j:h(t)\in\text{Range}(h_j)\\ u\in\langle h(t),l(t)\rangle}} \left|h'(h_j^{-1}(u))\right| \Big]^{-1}
\end{aligned}$$

Now, we can apply Lemma 18 to conclude that this is

$$\leq \left[\min_{B^c} |h'|, |l'|\right]^{-1}$$

Abbreviating

$$A_4 = h_j^{-1}(u) - l_j^{-1}(u) \qquad t \in G, u \in \langle h(t), l(t)\rangle \tag{32}$$

we have

$$|A_3| \leq \left[\min_{B^c} |h'|, |l'|\right]^{-1} \|h-l\|_\infty + \max_{\substack{t\in G\\ u\in\langle h(t),l(t)\rangle}} |A_4| \tag{33}$$

Turning to A_4, we observe that

$$\int_{\langle h_j^{-1}(u), l_j^{-1}(u)\rangle} |h'| \geq \left[\min_{\langle h_j^{-1}(u), l_j^{-1}(u)\rangle} |h'|\right] |h_j^{-1}(u) - l_j^{-1}(u)|$$

By Lemma 19, since $t \in G, u \in \langle h(t), l(t)\rangle$ we have

$$\langle h_j^{-1}(u), l_j^{-1}(u)\rangle \subset B^c$$

and therefore

$$\begin{aligned}\int_{\langle h_j^{-1}(u), l_j^{-1}(u)\rangle} |h'| &\geq \left[\min_{t\notin B} |h'(t)|, |l'(t)|\right] |h_j^{-1}(u) - l_j^{-1}(u)| \\ &\geq \left[\min_{B^c} |h'|, |l'|\right] |h_j^{-1}(u) - l_j^{-1}(u)|\end{aligned}$$

Lemma 19 also tells us that h' does not change sign on $\langle h_j^{-1}(u), l_j^{-1}(u)\rangle$, which means

$$\begin{aligned}\int_{\langle h_j^{-1}(u), l_j^{-1}(u)\rangle} |h'| &= \left|\int_{\langle h_j^{-1}(u), l_j^{-1}(u)\rangle} h'\right| \\ &= \left|h(h_j^{-1}(u)) - h(l_j^{-1}(u))\right| \\ &= \left|u - h(l_j^{-1}(u))\right| \\ &= \left|l(l_j^{-1}(u)) - h(l_j^{-1}(u))\right| \\ &\leq \|h - l\|_\infty\end{aligned}$$

Therefore,

$$\begin{aligned}|A_4| &= |h_j^{-1}(u) - l_j^{-1}(u)| \\ &\leq \left[\min_{B^c} |h'|, |l'|\right]^{-1} \int_{\langle h_j^{-1}(u), l_j^{-1}(u)\rangle} |h'| \\ &\leq \left[\min_{B^c} |h'|, |l'|\right]^{-1} \|h - l\|_\infty \qquad (34)\end{aligned}$$

Combining equations (27), (28), (29), (30), (31), (32), (33), (34), the proposition is proved.