A Probabilistic Upper Bound on Differential Entropy

Erik Learned-Miller
Department of Computer Science
University of Massachusetts, Amherst
140 Governors Drive
Amherst, MA 01003
Email: elm@cs.umass.edu

Joseph DeStefano,
Department of Math and Computer Science
College of the Holy Cross
Worcester, MA 01610
Email: jdestefa@holycross.edu

***Abstract*— A novel probabilistic upper bound on the entropy of an unknown one-dimensional distribution, given the support of the distribution and a sample from that distribution, is presented. No knowledge beyond the support of the unknown distribution is required. Previous distribution-free bounds on the cumulative distribution function of a random variable given a sample of that variable are used to construct the bound. A simple, fast, and intuitive algorithm for computing the entropy bound from a sample is provided.**

I. Introduction

The differential entropy of a distribution [9] is a quantity employed ubiquitously in communications, statistical learning, physics, and many other fields. Let X be a one-dimensional random variable with absolutely continuous distribution $F(x)$ and density $f(x)$. The differential entropy of X is defined to be

$$h(X) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx. \tag{1}$$

Since the entropy of X depends only upon its density (if it exists), we also write $h(f) \equiv h(X)$.

It is well known [3] that the entropy of a distribution with support $[x_L, x_R]$ is at most $\log(x_R - x_L)$, which is the entropy of the distribution that is uniform over the support. Given a sample of size n from an unknown distribution with this support, we cannot rule out with certainty the possibility that this sample was drawn from the uniform distribution over this interval. Thus, we cannot hope to improve a deterministic upper bound on the entropy over such an interval when nothing more is known about the distribution.

However, given a sample from an unknown distribution, we can say that it is *unlikely* to have been drawn from a distribution with entropy greater than some value. In this paper, we formalize this notion and give a specific, probabilistic upper bound for the entropy of an unknown distribution using both the support of the distribution and a sample of this distribution. To our knowledge, this is the first non-trivial upper bound on differential entropy which incorporates information from a sample and can be applied to any one-dimensional probability distribution with a density.

In this work, we restrict our analysis of distributions F whose entropy we wish to bound to those with densities and finite support. For some distributions without densities, such as discrete distributions and mixtures of absolutely continuous distributions with finite support and discrete distributions, the definition of differential entropy can be extended to be $-\infty$. Since our algorithm, presented below, returns an extended real number, it also returns a valid upper bound for these latter types of distributions. We do not address the class of singular distributions (e.g. Cantor distributions), whose entropies are undefined.

II. The Bound

Let $X_1, X_2, ..., X_n$ be i.i.d. real-valued random variables with distribution function $F(x) = \text{Prob}\{X_1 \leq x\}$ and density $f(x)$.[1] Denote the standard empirical distribution function by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}},$$

[1] We will assume for the remainder of the paper that $n \geq 3$, as this will simplify certain analyses.

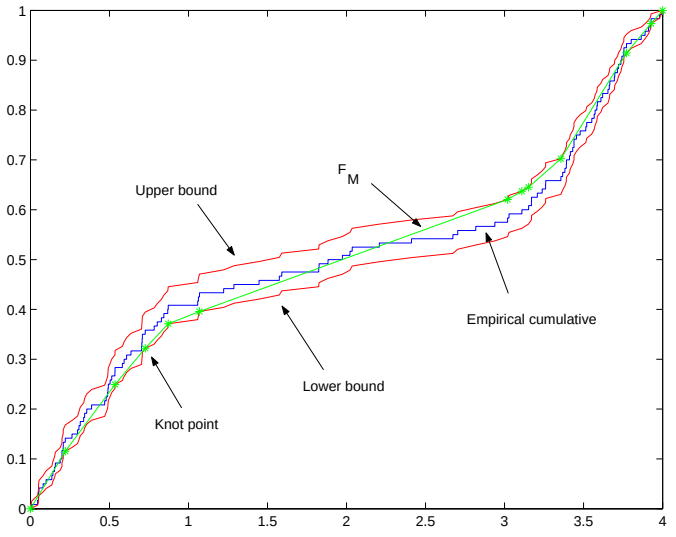

Fig. 1. This figure shows the maximum entropy cumulative distribution F_M which fits the constraints of the Massart inequality for the given empirical cumulative distribution and some confidence level α. Notice that the cumulative is piecewise linear, implying a piecewise constant density function. With probability at least α, the true cumulative distribution F has entropy less than or equal to this maximum entropy distribution.

where $I_{\{E\}}$ is the indicator function which takes a value of 1 when E is true, and 0 otherwise.

Consider a sample of size n and the order statistics[2] Z_1 through Z_n of that sample. We assume that the distribution has finite support and that we know this support. For ease of exposition, we label the left support Z_0 and the right support Z_{n+1} making the support values act like additional order statistics of the sample. But this is done merely for notational convenience and does not imply in any way that these are real samples of the random variable.

It will also be useful to refer to the (extended) entropy of a distribution directly as a function of its cdf. For distributions with densities, discrete distributions, and mixture distributions with cdf G, we defined the extended entropy, as a function of G, to be

$$h(G(x)) = \begin{cases} h(\frac{dG}{dx}) & \text{when } G \text{ has a density,} \\ -\infty & \text{otherwise.} \end{cases} \quad (2)$$

We start with a bound due to Dvoretzky, Kiefer, and Wolfowitz [5], and whose constant was determined by Massart [7], on the supremum of the distance between the empirical n-sample cumulative, $F_n(x)$, and the true distribution:

$$P(\sup_x |F(x) - F_n(x)| > \varepsilon) \leq 2e^{-2n\varepsilon^2} \equiv 1 - \alpha. \quad (3)$$

Thus, with probability *at least* α, the true cumulative does not differ from the empirical cumulative by more than ε. This is a distribution-free bound. That is, it holds for any one-dimensional probability distribution. For background on such bounds and their uses, see the text by Devroye et al. [4].

For a given empirical distribution F_n, consider the family $\mathcal{C}(F_n)$ of cumulative distribution functions $G(x)$ that satisfy the condition

$$\sup_x |G(x) - F_n(x)| \leq \varepsilon. \quad (4)$$

Let h^* be the supremum of the entropies of these functions, i.e.

$$h^* = \sup_{G \in \mathcal{C}(F_n)} h(G(x)).$$

We present an algorithm that constructs a distribution $F_M(x)$ whose entropy attains this supremum. Furthermore, we show that this distribution is unique. In other words, the algorithm constructs the unique distribution with maximum entropy that satisfies condition (4). This establishes a probabilistic upper bound on the entropy of the distribution from which the original sample was drawn.

Figure 1 illustrates some of the basic ideas of the paper. The central piecewise constant curve is a typical empirical cdf. The outer curves are confidence bounds for the true cumulative distribution based upon (3).[3] The piecewise linear curve F_M between the bounds shows the maximum entropy distribution within the bounds. Note that this maximum entropy distribution follows the same path that would be followed by a string which has been threaded through the "tube" provided by the upper and lower bounding curves and then pulled tight. For this reason, we call the algorithm which generates this curve the *string-tightening algorithm*.

III. Approach

We proceed as follows:

1) Consider again a sample of size n from an unknown one-dimensional distribution. We handle

[2]The order statistics $Z_1, Z_2, ..., Z_n$ of a sample $X_1, X_2, ..., X_n$ are simply the values in the sample arranged in non-decreasing order. Hence, Z_1 is the minimum sample value, Z_2 the next largest value, and so on.

[3]More precisely, the outer curves are slight modifications of the envelope suggested by (3) which restrict the curve to obey the bound *and* to be piecewise linear and continuous between sample points. We show in Lemma 3 that any entropy maximizing cdf must be piecewise linear and continuous.

separately the cases in which there are no duplicate sample values and in which there are duplicates. We address the latter case in Appendix B. When there are no duplicate sample values, we show that any entropy-maximizing distribution must be continuous.

2) For a given sample, at each sample point, we define sets of pairs of points called *pegs*, between which a continuous cdf must pass to obey the bound.
3) Now consider a distribution $G(x)$ with continuous cdf that is a candidate for the maximum entropy distribution. We refer to the values $G(Z_i)$ of the cdf evaluated at the samples Z_i as the *critical ordinates* of $G(x)$. Likewise, we refer to the ordered pairs $(Z_i, G(Z_i))$ as the *critical points* of $G(x)$. We show that among cdfs with a particular set of critical points, the piecewise linear cdf maximizes the entropy.
4) Thus, to find a globally entropy-maximizing cdf, it is enough to consider piecewise linear functions, since for any cdf that is not piecewise linear, there exists another admissible one that is piecewise linear and has larger differential entropy. The piecewise linear cdfs can be parameterized by their critical ordinates. Let the vector of critical ordinates be represented by $\Theta = [\theta_1 \theta_2 ... \theta_n]^T$. We show that the entropy of these piecewise linear cdfs is a strictly convex function of Θ and that finding the Θ that maximizes the entropy is a convex optimization problem over a closed convex set. Hence it has a unique maximum.
5) We then show that the entropy maximizing cdf, in addition to being piecewise linear, must bend upward only at upper pegs and downward only at lower pegs.
6) Since only a finite number of cdfs satisfy this condition, it suffices to consider each of these candidates for the maximum entropy distribution. We provide a simple graph algorithm that efficiently finds the maximum entropy distribution from this finite set, and show how to calculate the entropy of the resulting distribution.

Now we provide details.

A. Sample point bounds

For a desired confidence level α, we can compute a corresponding ε from (3) that meets that level:

$$\varepsilon = \sqrt{-\frac{\ln \frac{1-\alpha}{2}}{2n}}. \tag{5}$$

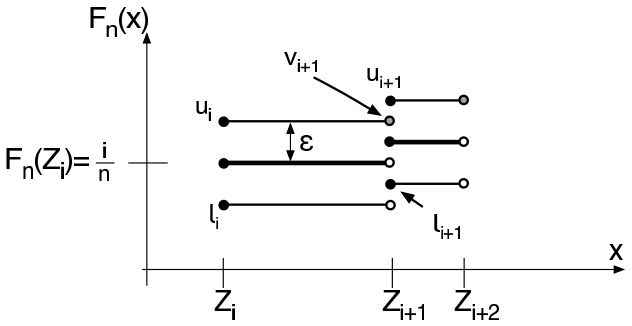

Fig. 2. This figure shows, in detail, a portion of an empirical distribution and the plus-or-minus ε bounds of (4). The empirical cumulative distribution is shown by the thicker horizontal lines. The left end of each segment of the distribution is shown as a closed circle, and the right end with an open circle, to denote the semi-open interval spanned by each segment. The upper and lower bounds are shown by the thinner horizontal lines. Note in particular the points $\mathbf{u}_i, \mathbf{l}_i, \mathbf{u}_{i+1}$, and $\mathbf{l}_{i+1}$ at the order statistic locations Z_i and Z_{i+1}. These points bracket the function (according to the bound) at the sample values. Furthermore, since any entropy maximizing distribution must be continuous (Lemma 2), at the sample value Z_{i+1} an entropy maximizing distribution must also have an ordinate value at or below the point $\mathbf{v}_{i+1}$. We refer to the lower bracketing points $\mathbf{l}_i$ and the tighter upper bracketing points $\mathbf{v}_i$ as *pegs*. These pegs are the only locations at which the maximum entropy cumulative distribution can potentially bend (Lemma 5).

We conclude that with probability α, the true distribution lies within this ε of the empirical distribution at all x.

The upper bound on the ordinate at sample point Z_i is given by

$$u_i = \min(\frac{i}{n} + \varepsilon, 1).$$

Similarly, the lower bound is given by

$$l_i = \max(\frac{i}{n} - \varepsilon, 0).$$

As shown in Figure 2, we also define the *points* or ordered pairs corresponding to these bounds as

$$\mathbf{u}_i = (Z_i, u_i)$$

and

$$\mathbf{l}_i = (Z_i, l_i),$$

for $1 \le i \le n$. Additionally, we define $u_0 = l_0 = 0$, $u_{n+1} = l_{n+1} = 1$, $\mathbf{u}_0 = \mathbf{l}_0 = (Z_0, 0)$ and $\mathbf{u}_{n+1} = \mathbf{l}_{n+1} = (Z_{n+1}, 1)$.

B. Existence of certain cdfs

Lemma 1: For a sample with no repeated values and a given confidence α, there exists at least one cdf which satisfies condition (4) and that is continuous and is linear between successive order statistics (and hence is piecewise linear).

Proof: Note that for a confidence level α, ε is always larger than the step size of the cdf, since

$$\begin{aligned}\varepsilon &\geq \sqrt{-\frac{\ln\frac{1-0}{2}}{2n}} \\ &= \sqrt{\frac{\ln 2}{2}}\cdot\frac{1}{\sqrt{n}} \\ &> \frac{1}{n}, \quad \forall n \geq 3.\end{aligned}$$

Given this relationship between ε and $\frac{1}{n}$, we note that the cdf which is linear between order statistics and which connects $\mathbf{l}_i$ to $\mathbf{l}_{i+1}$ satisfies condition (4) at every point. (See Figure 2.) ■

C. Continuity of cdfs

Lemma 2: For samples without repeated values, among cdfs that meet condition (4), a cdf with discontinuities cannot maximize entropy.

Proof: The differential entropy of any cdf with discontinuities is $-\infty$. Since by Lemma 1 there always exists a cdf without discontinuities that satisfies (4), there is always at least one cdf that will have entropy greater than any non-continuous cdf. ■

Hence, in searching for a cdf with maximimum entropy, it is enough to consider only continuous cdfs.

D. Pegs

The points $\mathbf{u}_i$ and $\mathbf{l}_i$ defined above represent the straightforward application of (3) at the sample points. Using Lemma 2 we can tighten these bounds for the entropy maximizing distribution, if it exists. In particular, referring again to Figure 2, at a sample point Z_{i+1}, while condition (4) allows the cdf to pass through a point $\mathbf{a} = (Z_{i+1}, b)$, where $u_i < b \leq u_{i+1}$, such a curve cannot be an entropy maximizing cdf since it will have a discontinuity at Z_{i+1}. This discontinuity would be unavoidable for a cdf containing $\mathbf{a}$, since to the left of the sample Z_{i+1} the curve is upper bounded by u_i. Thus, *at a sample point* Z_i, any continuous cdf is upper bounded by the value

$$v_i = \min(\frac{i}{n} + \varepsilon - \frac{1}{n}, 1),$$

for $1 \leq i \leq n$. As before we define $v_0 = 0$ and $v_{n+1} = 1$ for notational convenience. We also define the *points*

$$\mathbf{v}_i = (Z_i, v_i)$$

for $0 \leq i \leq n+1$. Figure 2 shows the location of $\mathbf{v}_{i+1}$.

Together, we refer to the points $\mathbf{v}_i$ and $\mathbf{l}_i$ at the sample points as *pegs*, and they will play a key role. As we have just shown, at each sample point, any entropy maximizing cdf must pass between each pair of pegs (or pass through one of the pegs). It follows immediately that unless each critical point $(Z_i, G(Z_i))$ of a cdf $G(x)$ falls between (or on) the pegs $\mathbf{v}_i$ and $\mathbf{l}_i$, such a cdf cannot be the maximum we are seeking. We refer to critical points which fall between (or on) pegs as *admissible critical points* and to their ordinates as *admissible critical ordinates*.

E. Piecewise linearity of cdfs

Lemma 3: Among the cdfs with a particular set of admissible critical ordinates, the one which is piecewise linear both satisfies condition (4) and maximizes the entropy.

Proof: Assuming all are admissible, let the critical ordinates of a cdf at Z_i be denoted θ_i. Then the set of critical ordinates for a cdf can be encoded as a parameter vector $\Theta = \{\theta_0, \theta_1, ..., \theta_{n+1}\}$, where θ_0 and θ_{n+1} are 0 and 1 by definition, but the other θ_i can be chosen to maximize entropy. Now consider the set of possible cdfs with a particular set of critical ordinates Θ and a corresponding set of critical points.

Because $\varepsilon > \frac{1}{n}$ and the v_i upper bound cdfs with admissible critical ordinates at the sample points, condition (4) admits any cdf that is linear between successive critical points (see Figure 2).

We next show for any set of critical ordinates Θ, the cdf with those critical ordinates that maximizes entropy is piecewise linear between the corresponding critical points.

Note that the entropy function for a cdf $G(x)$ is separable into integrals over the interval of interest $[Z_i, Z_{i+1}]$ and the remainder of the real line $\overline{[Z_i, Z_{i+1}]}$:

$$\begin{aligned}h(G) = &- \int_{Z_i}^{Z_{i+1}} g(x)\log g(x)\,dx \\ &- \int_{\overline{[Z_i,Z_{i+1}]}} g(x)\log g(x)\,dx.\end{aligned}$$

Because of this separability, conditioned on specific values for critical ordinates θ_i and θ_{i+1}, the cdf must maximize each of the terms above separately.

Focussing on the first term, letting $\mathcal{G}$ be the set of all continuous monotonic non-decreasing functions over $[Z_i, Z_{i+1}]$, and with $g(x) = \frac{dG(x)}{dx}$, we have

$$\max_{G\in\mathcal{G}}\left[-\int_{Z_i}^{Z_{i+1}} g(x)\log g(x)dx\right] \quad (6)$$

$$= \max_{G\in\mathcal{G}}\left[-\int_{Z_i}^{Z_{i+1}} g(x)\left[\log g(x)-\log(C)\right]dx\right] \quad (7)$$

$$= \max_{G\in\mathcal{G}}\left[-\int_{Z_i}^{Z_{i+1}} g(x)\log\frac{g(x)}{C}dx\right] \quad (8)$$

$$= \max_{G\in\mathcal{G}}\left[-\int_{Z_i}^{Z_{i+1}} \frac{g(x)}{C}\log\frac{g(x)}{C}dx\right]. \quad (9)$$

.

The last expression is just the entropy of the distribution $d(x)=\frac{g(x)}{C}$, which for the right choice of C is a properly normalized probability distribution over $[Z_i,Z_{i+1}]$. It is well-known [3] that $d(x)$ must be uniform (excluding a set of measure 0) to maximize entropy over a finite interval. This in turn, implies that $g(x)$ must be uniform between Z_i and Z_{i+1} to maximize (6). Hence, G must be linear between Z_i and Z_{i+1} to be an entropy maximizing distribution. ■

As we can now restrict our search for entropy maximizing distributions to those which are piecewise linear, it will be useful to define the following envelope curves. Let $F_{\mathbf{v}}$ and $F_{\mathbf{l}}$ be the piecewise linear cdfs connecting the points $\mathbf{v}_i$ and $\mathbf{l}_i$ respectively. Note that these curves (shown in Figure 1) represent a tighter envelope which must be obeyed by any entropy maximizing cdf than the envelope defined by condition (4).

F. Existence and uniqueness of solution

Given piecewise linearity, the set of remaining candidates for the cdf with maximum entropy is parameterized by Θ, the vector of critical ordinates. That is, to maximize entropy, we should evaluate

$$\arg\sup_{\Theta\in\Theta^*} h(F_\Theta(x)), \quad (10)$$

where F_Θ is a piecewise linear cdf depending only on Θ and Θ^* is the set of all possible Θ. That is, Θ^* is the subset of all ordered $n-$tuples in $[0,1]$ which satisfy the constraints on the critical ordinates. See Appendix A for a formal definition of Θ^*.

Lemma 4: The solution to the optimization problem (10) exists and is unique.

Proof: As shown in Appendix A, the set Θ^* of feasible Θ is a closed convex set, and the function $h(F_\Theta(x))$ is strictly concave in Θ. Therefore the optimization problem (10) has a unique maximum. See the text by Rockafellar for more on convex optimization [8]. ■

G. "String" bends only at pegs

Let F_M be the unique entropy maximizing distribution. As shown above, F_M should be piecewise linear, with any "bends", or changes in slope, occurring only at the critical points. Intuitively, using the string-tightening analogy, as the string is tightened, one might guess that that these slope changes can occur only at the pegs, which we prove here.

Lemma 5: An increase (decrease) in slope of F_M can occur only at the upper (lower) peg of a sample.

Proof: By contradiction. Define the points $\mathbf{a},\mathbf{b}$, and $\mathbf{c}$ to be (Z_{i-1},a), (Z_i,b), and (Z_{i+1},c) respectively, with $a\le b\le c$ and $1\le i\le n$. Suppose that there are two connected segments of F_M, $\overline{\mathbf{ab}}$ and $\overline{\mathbf{bc}}$. Now suppose that $b<v_i$ (it is below the upper peg) and the point $\mathbf{b}$ is below the line segment $\overline{\mathbf{ac}}$. That is, the slope of the cdf increases at $\mathbf{b}$.

Then there is an interval $[Z_i-\delta,Z_i+\delta]$, $\delta>0$, where the line segment $\overline{(Z_i-\delta,F_M(Z_i-\delta))(Z_i+\delta),F_M(Z_i+\delta))}$ lies entirely between $F_{\mathbf{l}}$ and $F_{\mathbf{v}}$, the lower and upper envelope curves defined above. The argument of Lemma 3 shows that this segment maximizes the entropy on $[Z_i-\delta,Z_i+\delta]$, and thus F_M, being maximal, cannot pass through the point $\mathbf{b}$, contradicting the assumption. A similar argument applies for a decrease in slope. ■

IV. The String-Tightening Algorithm

Thus F_M is completely described by the sequence of pegs that it touches, which we call the *knot points*. Since the entropy function is separable at knot points into independent sums, and there are a finite number of pegs that can act as knot points, the search for the set of knot points which give the maximum entropy cdf can be formulated as a shortest path problem.[4]

In particular, consider a directed graph whose vertices consist of all the upper pegs $\mathbf{v}_i$ and lower pegs $\mathbf{l}_i$, with $0\le i\le n+1$, and in which there is a directed edge from a peg $\mathbf{p}=(p_x,p_y)$ to another peg $\mathbf{q}=(q_x,q_y)$ if and only if

- $p_x<q_x$,
- $p_y\le q_y$, and
- the line segment $\overline{\mathbf{pq}}$ is between (or coincident with) the upper and lower bounds $F_{\mathbf{v}}$ and $F_{\mathbf{l}}$ for piecewise linear cdfs.

Notice that this graph cannot contain cycles, so it is a directed acyclic graph (DAG). Hence we can apply the single-source shortest path DAG algorithm [2] to find

[4]We thank one of the anonymous reviewers for suggesting this approach to the optimization problem.

the "least costly" path from the first peg $\mathbf{v}_0$ to the last peg $\mathbf{v}_{n+1}$.

The only remaining detail is to define a weight w for each edge of the graph equal to the negative of the portion of the entropy function that edge is responsible for:

$$w(\mathbf{p},\mathbf{q}) = (q_y - p_y)\log\frac{q_y - p_y}{q_x - p_x}.$$

With this set up, a (possibly non-unique) shortest path through the graph can be found in time $O(V+E)$, where V is the number of graph vertices (in this case, $2n+2$) and E is the number of edges. While it is possible that the path itself that maximizes entropy is not unique (this can occur if three or more pegs are colinear), we are guaranteed by Lemma 4 that there is a unique optimal cdf corresponding to all shortest paths.

The segments connecting the knots form F_M. Its entropy, which is our probability-α bound on the entropy of the true distribution F, is just the negative of the weight of the shortest path computed by the algorithm. Writing the K final knots as (a_i, b_i), the entropy can be written as

$$h(F_M) = -\sum_{i=1}^{K-1}(b_{i+1} - b_i)\log\frac{b_{i+1} - b_i}{a_{i+1} - a_i}. \qquad (11)$$

A. Examples

We shall refer to the process of defining the edges of the graph, running the shortest path algorithm, and computing the entropy of the resulting graph, together, as the string-tightening algorithm. Figure 3 compares the 95% confidence bound produced by the string-tightening algorithm for several distributions (shown in Figure 4) to the true entropy, obtained by numerical integration. The naïve bound $\log(Z_{n+1} - Z_0)$ is also plotted at the top of each graph. Note that the distributions have been truncated to a finite interval in each case.

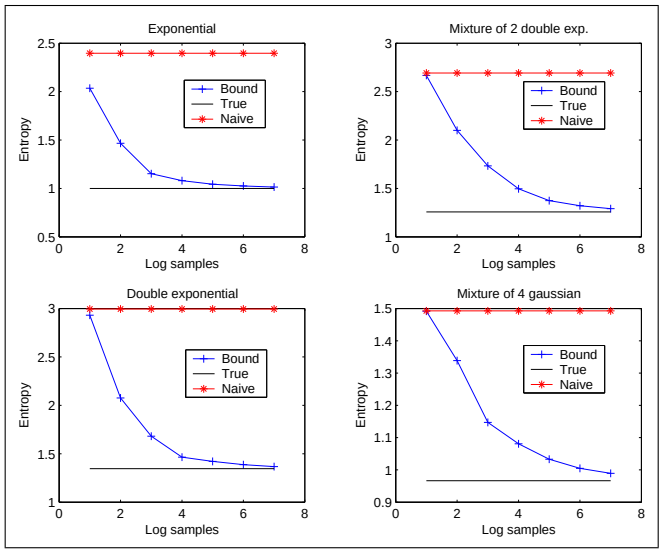

Fig. 3. The 95% confidence bound quickly becomes much tighter than the naïve bound.

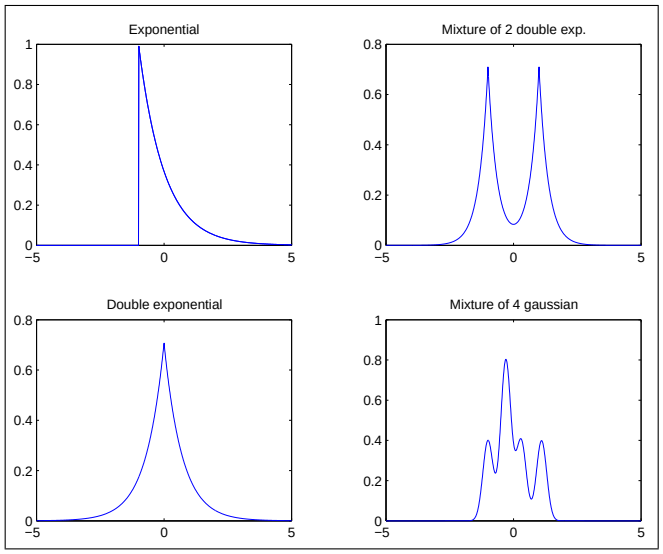

Fig. 4. The four distributions used in the comparisons in Figure 3.

V. Alternative Bounds

The bound on the distribution provided by condition (4) allows for the same degree of uncertainty at all points in the empirical cdf. Intuitively, it seems we should be able to bound the distribution more tightly near the ends of the support than in the middle. For empirical support of this intuition, we ran 10000 experiments with 100 random samples each from a known distribution, and recorded which of the order statistics were outside the bound (4) for $\alpha = 0.95$. The histogram of this data in Figure 5 clearly suggests that the bound provided by (4) is not as tight as it could be near the ends of the distribution. One would expect that a bound that was as tight as possible everywhere would miss equally often at all points.

As an alternative bound, we use the fact that for samples X_i from a continuous distribution $F(x)$, the values $F(X_i)$ are uniformly distributed on $[0,1]$ [6]. Therefore the random variable $F(Z_i)$ has the same distribution as the i-th order statistic of a uniform variate, i.e., it is beta distributed with parameters i and $n-i+1$ [1]. Its mean is $\frac{i}{n+1}$. In particular, this means that for each i, and

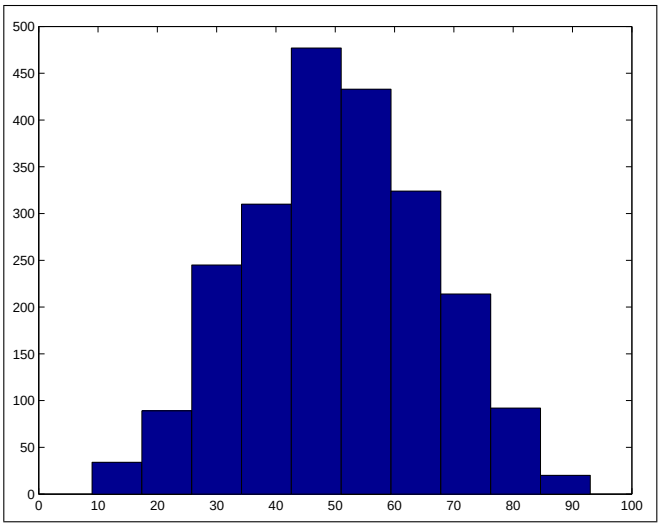

Fig. 5. The bound provided by (4) is too loose near the edges. The histogram shows that in simulations, the bound is violated more frequently near the center of the cdf. The horizontal axis gives the height (as a percentage) of the cumulative distribution where the bound violation occurred, and the vertical axis gives the number of violations.

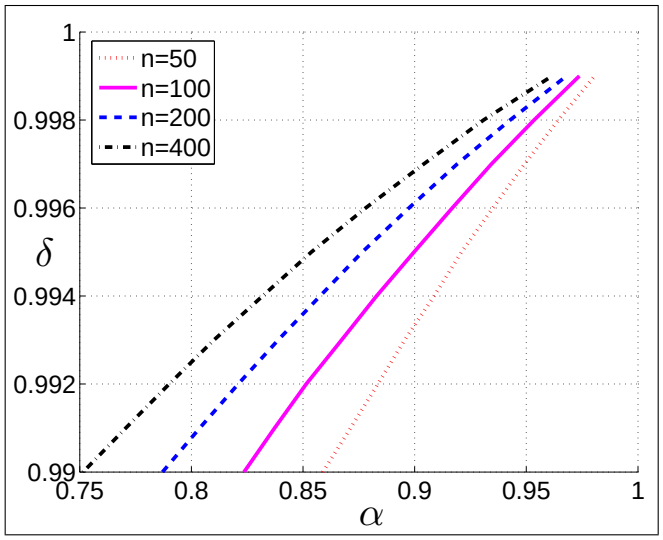

Fig. 6. Relationship between confidence bounds on the difference between a cdf and the individual order statistics of a sample (given by δ) and the confidence α that all order statistics will be "close" to the cdf. These curves were obtained by Monte Carlo simulation.

$\frac{1}{2} \leq \delta \leq 1$,

$$P(F(Z_i) \in (\beta^{-1}_{i,n-i+1}(\frac{1-\delta}{2}), \beta^{-1}_{i,n-i+1}(\frac{1+\delta}{2}))) = \delta, \quad (12)$$

where $\beta^{-1}_{i,n-i+1}$ is the inverse cdf of the beta distribution with parameters i and $n-i+1$. These bounds are tighter when i is near 0 or n, and looser when i is near $\frac{n}{2}$. By design, these bounds will be violated equally often at all sample locations.

To use this information in setting confidence bounds on a cdf, we need to calculate the probability *for all i* of $F(Z_i)$ being within the intervals defined by the inverse beta cdfs. For fixed n and δ, let $a_i = \beta^{-1}_{i,n-i+1}(\frac{1-\delta}{2})$ and $b_i = \beta^{-1}_{i,n-i+1}(\frac{1+\delta}{2})$. Then we define α_{order} such that

$$\alpha_{order} \equiv P(\forall i \, F(Z_i) \in (a_i, b_i)), \quad (13)$$

i.e., the probability that every point on the true cumulative falls within the bounds provided by a_i and b_i.

While it appears to be computationally intractable to calculate the value of δ which leads to a particular confidence α_{order} (or vice versa), we can estimate α_{order} as a function of n and δ by repeatedly drawing samples of size n from a known distribution (say, uniform) and examining how frequently the intervals (a_i, b_i) are violated *for at least one value of i*. The fraction of violations over repeated trials is an estimate of $(1-\alpha_{order})$.

Figure 6 shows the relationship between the value of δ and α_{order} for four values of n. Each value of α_{order} was estimated by drawing one million samples of size n from a uniform distribution and evaluating whether any order statistics in a given sample extended beyond the interval (a_i, b_i) for any point.

For example, to use this information for bounding entropy, we note from Figure 6 that for $n = 100$ and $\delta = 0.998$,

$$P(\forall i \, F(Z_i) \in (a_i, b_i)) = \alpha_{order} > 0.95. \quad (14)$$

Hence using the order statistic bounds as the pegs (i.e., taking $\mathbf{l}_i = a_i$ and $\mathbf{v}_i = b_i$), we obtain an envelope similar to that defined by (4) on the empirical distribution at the sample points, with greater than 95% confidence.[5] While these bounds on $F(Z_i)$ hold only at sample points, by noting that when $F(Z_i) \in (a_i, b_i)$, we have by the monotonicity of the cdf that for $w \in (Z_{i-1}, Z_i]$, $F(w) < b_i$. Also, when $F(Z_i) \in (a_i, b_i)$, for $w \in [Z_i, Z_{i+1})$, $F(w) > a_i$. This allows us to extend the bounding technique to elements of the domain other than just the sample points.

As an example, Figure 7 illustrates the bounds provided by the order statistics and our simulations, and compares them to the bounds provided by condition (4). For this example, $n = 100$ and $\alpha = \alpha_{order} = 0.975$. For clarity, we show the bounds relative to the true cumulative (in this case a uniform distribution) rather

[5]The conditions under which these new bounds provide a unique maximum entropy distribution are slightly different, but similar, to the arguments presented for the Massart bound, and are omitted for brevity.

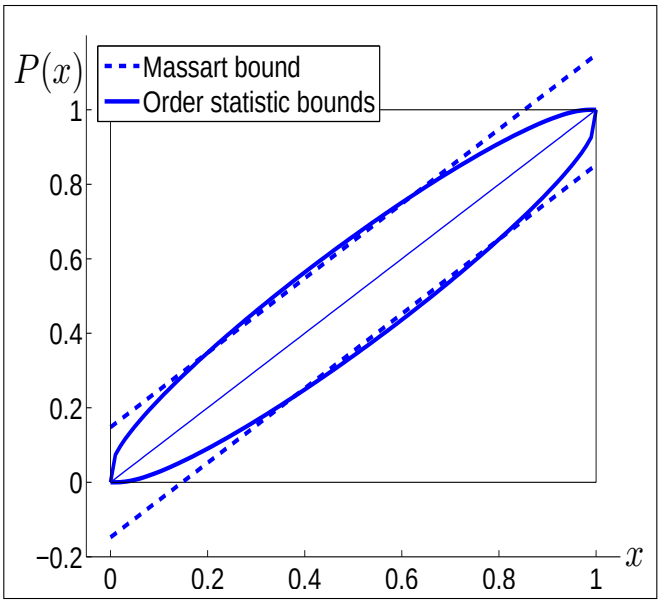

Fig. 7. Empirically generated bounds (solid lines) are tighter in some places and looser in others than those provided by Massart (dashed lines).

than an empirical cdf. Note that while the order statistic bounds are slightly looser at the fiftieth percentile, they are substantially tighter near the ends of the distribution. In fact, substantial portions of the Massart bound are uninformative since they extend beyond the $[0,1]$ interval. We leave an assessment of which set of bounds are more practically useful to future work.

VI. Conclusion

We have shown how distribution-free bounds on the cumulative distributions of unknown one-dimensional probability densities can be used to give sample-based probabilistic bounds on the entropies of distributions with known support. As an alternative to providing the support of the distribution, one can provide bounds on the mean log probability density of the tails of a distribution, and still provide similar bounds. We leave this topic to future work.

We have provided a simple algorithm to compute this bound exactly from samples taken from the unknown distribution. A by-product of the algorithm is an explicit representation of F_M, the distribution that achieves the computed bound. The simple form of F_M makes it convenient for use in resampling applications.

Appendix A. Proof of Lemma 4

Given continuity and piecewise linearity, the set of remaining candidates for the distribution with maximum entropy is parameterized by Θ, the vector of critical ordinates. Thus, to maximize entropy, we should evaluate

$$\arg\sup_{\Theta\in\Theta^*} h(F_\Theta(x)),$$

where Θ^* is the set of all possible Θ and $F_\Theta(x)$ is the piecewise linear cdf associated with the critical ordinates Θ. We wish to show that this optimization problem has a unique maximum.

Proof: The set Θ^* can be characterized by a set of linear inequalities:

$$\theta_i \leq v_i, \tag{15}$$
$$\theta_i \geq l_i, \tag{16}$$

for $0 \leq i \leq n+1$, and

$$\theta_i \leq \theta_{i+1} \tag{17}$$

for $0 \leq i \leq n$.

Inequalities (15) and (16) are the restrictions imposed by the pegs discussed above. The last set of inequalities (17) encodes the fact that the cdf must be non-decreasing. Together, these inequalities define Θ^* to be a closed convex set.

Next we wish to show that the function $h(F_\Theta(x))$ is a strictly concave function of Θ. With $F_\Theta(x)$ piecewise linear, $f_\Theta(x)$ is piecewise constant, and in particular

$$f_\Theta(x) = \frac{\theta_{i+1}-\theta_i}{Z_{i+1}-Z_i},$$

on the interval $[Z_i, Z_{i+1})$, for $0 \leq i \leq n$.

Hence, we can write

$$\begin{aligned} h(F_\Theta(x)) &= h(f_\Theta(x)) && (18)\\ &= -\sum_{i=0}^{n}\int_{Z_i}^{Z_{i+1}} f_\Theta(x)\log f_\Theta(x)\,dx && (19)\\ &= -\sum_{i=0}^{n}(\theta_{i+1}-\theta_i)\log\frac{\theta_{i+1}-\theta_i}{Z_{i+1}-Z_i} && (20)\\ &= -\sum_{i=0}^{n}(\theta_{i+1}-\theta_i)\log\frac{\theta_{i+1}-\theta_i}{d_i}, && (21)\end{aligned}$$

using $d_i = Z_{i+1}-Z_i$.

Since any given θ_i only affects two terms in this sum, we have that, for $1 \leq i \leq n$,

$$\begin{aligned}&\frac{\partial h}{\partial\theta_i}\\ =\;& \frac{\partial}{\partial\theta_i}-\left[(\theta_i-\theta_{i-1})\log\frac{\theta_i-\theta_{i-1}}{d_{i-1}}+(\theta_{i+1}-\theta_i)\log\frac{\theta_{i+1}-\theta_i}{d_i}\right]\\ =\;& -(d_{i-1}+\log\frac{\theta_i-\theta_{i-1}}{d_{i-1}}-d_i-\log\frac{\theta_{i+1}-\theta_i}{d_i}).\end{aligned}$$

The second partial derivatives give

$$\frac{\partial^2 h}{\partial \theta_i^2} = -\left(\frac{d_{i-1}}{\theta_i - \theta_{i-1}} + \frac{d_i}{\theta_{i+1} - \theta_i}\right)$$

for $1 \leq i \leq n$ and

$$\frac{\partial^2 h}{\partial \theta_i \, \partial \theta_{i-1}} = \frac{\partial^2 h}{\partial \theta_{i-1} \, \partial \theta_i} \quad (22)$$

$$= \frac{d_{i-1}}{\theta_i - \theta_{i-1}} \quad (23)$$

for $2 \leq i \leq n$. All of the other second derivatives are zero.

To show that $h(F_\Theta)$ is strictly concave, it suffices to show that the *nxn* matrix $\mathbf{A}$ of second partial derivatives (the Hessian) is such that the quadratic form

$$\mathbf{x}^T \mathbf{A} \mathbf{x} < 0$$

for all vectors $\mathbf{x}$, i.e., that A is negative definite.

Expanded in terms of components, the quadratic form can be written

$$\begin{aligned}
& \sum_{i=1}^{n} \frac{\partial^2 h}{\partial \theta_i^2} x_i^2 + \sum_{i=2}^{n} \frac{\partial^2 h}{\partial \theta_i \, \theta_{i-1}} x_i x_{i-1} \\
= & \sum_{i=1}^{n} -\left(\frac{d_{i-1}}{\theta_i - \theta_{i-1}} + \frac{d_i}{\theta_{i+1} - \theta_i}\right) x_i^2 + \sum_{i=2}^{n} \frac{d_{i-1}}{\theta_i - \theta_{i-1}} x_i x_{i-1} \\
= & -\frac{d_0}{\theta_1 - \theta_0} - \frac{d_n}{\theta_{n+1} - \theta_n} \\
& \qquad - \sum_{i=2}^{n} (x_{i-1}^2 - 2 x_{i-1} x_i + x_i^2) \frac{d_i}{\theta_{i+1} - \theta_i} \\
= & -\frac{d_0}{\theta_1 - \theta_0} - \frac{d_n}{\theta_{n+1} - \theta_n} - \sum_{i=2}^{n} (x_{i-1} - x_i)^2 \frac{d_i}{\theta_{i+1} - \theta_i} \\
< & \; 0.
\end{aligned}$$

The last inequality follows since the d_i's and the differences in successive θ_i's are uniformly positive. Hence, $\mathbf{A}$ is negative definite and $h(F_\Theta)$ is strictly concave.

Any strictly concave function defined over a closed convex set has a unique maximum [8]. ■

Appendix B: Samples with Duplicate Points

The lemmas above and the string-tightening algorithm are developed assuming that the sample of size n of the unknown distribution contains no duplicate values. Here we sketch the necessary changes to arguments to address the case in which there are one or more duplicate values in the sample.

Given that there are duplicated values in the sample, there are two cases to consider. In the first case, condition (4) is loose enough so that there still exist distributions which are are continuous and satisfy the condition. For example, if n is large enough and ε is small enough so that $\varepsilon > \frac{2}{n}$ then a single duplicated point will still admit continuous F_M, and essentially all of the arguments of the main line of reasoning still hold.

The other case is that the duplicated points force *all* cdfs which obey condition (4) to be discontinuous. In other words, if there are sufficiently many duplicated points in a sample, there will be no continuous cdf which satisfies (4). This occurs when $\varepsilon < \frac{k}{n}$, where k is the maximum number of replications of any single value in the sample. In this case, the Massart inequality ensures that with high probability, the true distribution has the entropy of a distribution with discontinuous cdf, i.e. an entropy of $-\infty$. Note that in this case, in general, there is no unique entropy maximizing distribution.

References

[1] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *A First Course in Order Statistics*. John Wiley & Sons, 1992.

[2] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. MIT Press, 1990.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[4] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer Verlag, 1996.

[5] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of a sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 27:642–669, 1956.

[6] E. B. Manoukian. *Modern Concepts and Theorems of Mathematical Statistics*. New York: Springer-Verlag, 1986.

[7] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

[8] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[9] C. E. Shannon. The mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423,623–656, Jul,Oct 1948.