# Asymptotic theory of greedy approximations to minimal $K$-point random graphs

Alfred O. Hero and Olivier J.J. Michel

### Abstract

Let $\mathcal{X}_n = \{x_1, \ldots, x_n\}$, be an i.i.d. sample having multivariate distribution $P$. We derive a.s. limits for the power weighted edge weight function of greedy approximations to a class of minimal graphs spanning $k$ of the $n$ samples. The class includes minimal $k$-point graphs constructed by the partitioning method of Ravi, Sundaram, Marathe, Rosenkrantz and Ravi [43] where the edge weight function satisfies the quasi-additive property of Redmond and Yukich [45]. In particular this includes greedy approximations to the $k$-point minimal spanning tree ($k$-MST), Steiner tree ($k$-ST), and the traveling salesman problem ($k$-TSP). An expression for the influence function of the minimal weight function is given which characterizes the asymptotic sensitivity of the graph weight to perturbations in the underlying distribution. The influence function takes a form which indicates that the $k$-point minimal graph in $d > 1$ dimensions has robustness properties in $\mathbb{R}^d$ which are analogous to those of rank order statistics in one dimension. A direct result of our theory is that the log-weight of the $k$-point minimal graph is a consistent non-parametric estimate of the Rényi entropy of the distribution $P$. Possible applications of this work include: analysis of random communication network topologies, estimation of the mixing coefficient in $\epsilon$-contaminated mixture models, outlier discrimination and rejection, clustering and pattern recognition, robust non-parametric regression, two sample matching and image registration.

[1] Alfred Hero is with the University of Michigan at Ann Arbor, MI 48109-2122 (hero@eecs.umich.edu), and Olivier Michel is with the Ecole Normale Superieure, Lyon cedex 07, France (omichel@ens-lyon.fr). This research was supported in part by AFOSR grant F49620-97-0028.

## Figure List

# I. Introduction

Assume that one is given a set $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ of $n$ points in $\mathbb{R}^d$. Fix $k$ and denote by $\mathcal{X}_{n,k}$ a $k$-point subset of $\mathcal{X}_n$, $0 < k \leq n$. The elements of the subset $\mathcal{X}_{n,k}$ are distinct and there are $\binom{n}{k}$ possible $k$-point subsets of $\mathcal{X}_n$. The minimal k-point Euclidean graph problem is to find the subset of points $\mathcal{X}_{n,k}$ and the set of edges connecting these points such that the resultant graph has minimum total weight $L(\mathcal{X}_{n,k})$. This problem arises in competitive bidding for network routing contracts when some nodes may be left out of the connected network, the prize-collecting traveling salesman and Steiner tree problems for visiting at least $k$ out of $n$ cities, the minimum latency problem, and other combinatorial optimization problems [62], [12], [6], [44], [28].

For example, the Euclidean *minimal k-point spanning tree* (k-MST) is the minimum weight tree spanning any $k$ of the $n$ points. The planar $k$-MST problem was shown to be NP-complete by Zelikovsky and Lozevanu [62] and Ravi, Sundaram, Marathe, Rosenkrantz and Ravi [43]. Ravi *etal* proposed a polynomial time approximation algorithm for the planar $k$-MST ($d = 2$) with approximation ratio $O(k^{1/4})$ which has been successively improved to $O(\log(k))$ by Garg and Hochbaum [28], $O(\log(k)/\log\log(n))$ by Eppstein [23], $O(1)$ by Blum, Chalasani and Vempala [13], $2\sqrt{2}$ by Mitchell [40], 3 by Garg [27], and $1+\epsilon$ by Arora [3], [4].

While it is difficult to establish general and useful properties of optimal graphs for a fixed set of points $\mathcal{X}_n$, interesting properties of many classes of optimal graphs can be obtained by assuming that the $n$ points are random samples from a distribution $P$. In particular, the asymptotic behavior of the TSP, MST and Steiner trees ($k = n$) as $n$ tends to $\infty$ is now well understood in this stochastic setting. The recent books by Steele [54] and Yukich [60] provide excellent introductions to this subject. The main result of this paper is the derivation of a.s. limits for a class of greedy approximations to Euclidean minimal $k$-point graphs over a random set of $n$ points. This class of greedy approximations is the set of minimal $k$-point graphs constructed by the generalized method of Ravi *etal* [43] and where the edge weight function satisfies the quasi-additive property of Redmond and Yukich [45].

It directly follows from the asymptotic analysis that the log of the minimum graph weight is a strongly consistent and robust estimator of the order-$\nu$ Rényi entropy [47] of the multivariate distribution $P$, where $\nu \in (0,1)$ is a function of the sample space dimension and the edge weight power exponent. Alternatively, although we do not develop this extension here due to space limitations, by performing an appropriate measure transformation on the data space one obtains a robust non-parametric estimate of the Rényi divergence (a.k.a. relative entropy or Chernoff distance) between $P$ and a prespecified reference measure $P_o$. It is remarkable that the weight function of the minimal $k$-point graph provides a direct estimate of entropy which completely bypasses the difficult intermediate step of multivariate density estimation required by previous estimators.

The problem of entropy estimation has long been of interest to the engineering, physics, and statistics communities, e.g. see the recent paper by Beirlant *etal* [11] for a thorough overview of the topic of Shannon entropy estimation. The general entropy estimation problem is relevant to pattern analysis, process complexity assessment, model identification, tests of distributions, and other applications where invariance to scale, translation and other invertible transformations is desired in the discriminant [1], [36], [30]. The Rényi entropy estimation problem arises in adaptive vector quantizer design, where the entropy is more commonly called the the Panter-Dite factor and is related to the asymptotically optimal quantization cell density [29], [41]. Estimates of Rényi entropy have also been proposed for characterizing complexity of time-frequency distributions [58], [25], [39], [51], [9]. Other relevant entropy estimation applications are: estimation of Lyapounov exponents in non-linear dynamical models [22], [24], multi-modality image registration using mutual information matching criteria [57], stopping criteria for regression and classification trees [14], and testing for normality of a random data sample [56].

In addition to the aforementioned entropy estimation applications the greedy algorithm, and associated theory, presented here can be applied to robustification of existing minimal graph approaches to pattern recognition [61], [55]; clustering [32], [19]; non-parametric regression [7], [8]; testing for randomness of a data sample [33]; and testing particle distributions in electron photomicrographs [21]. Finally, in addition to the combinatorial optimization problems mentioned in the first paragraph, our results may be useful for asymptotic analysis of $k$-point minimal graph techniques proposed for optimizing communications subnetwork topologies and network provisioning [15], [52], [42], [35]; minimum area routing in VLSI circuits [16]; minimum cost pipeline interconnections for subnetworks of oil wells [44]; and minimum cost interconnections for cable TV subnetworks [20].

The principal theoretical results presented here are:

1. A polynomial time greedy algorithm for constructing an approximation to the minimal $k$-point graph and its edge weight function is presented which is a direct generalization of the algorithm of Ravi, Sundaram, Marathe, Rosenkrantz and Ravi [43] developed therein for minimal $k$-point minimum spanning tree approximation on the plane.
2. A tight a.s. asymptotic bound on the entropy estimation error is given which can be used to determine the required partition resolution to obtain a prescribed estimator error when a bound on the total variation (roughness) of the density function is known.
3. Zero asymptotic error is achieved when the density function is piecewise constant over the resolution $1/m$ partition cells of the greedy algorithm.
4. We give a condition, called a tightly coverable graph property, which holds when $k + o(k)$ of the vertices of the $k$-point graph can be covered by a resolution $1/m$ partition set (a.s.) as $k \to \infty$. This condition is satisfied for the greedy approximation by construction. If the exact minimal $k$-point graph satisfies this condition then the weight of the minimal $k$-point graph converges to the same asymptotic limit as the greedy approximation, i.e. the greedy approximation is asymptotically optimal.
5. A robust Rényi entropy estimator is proposed based on the log of the weight of the minimal $k$-point graph with edge weight exponent $\gamma$. This estimator is shown to converge a.s. to a conditional Rényi entropy of order $\nu = (d-\gamma)/d \in (0,1)$, where $d > \gamma$ is the sample space dimension. Inspired by the convergence rates established in [46], for $1/d \leq \nu < 1$ we predict that the rate of convergence of the non-parametric Rényi entropy estimator is $O(n^{-1/d})$.
6. Influence function studies are presented which quantitatively establish that the greedy minimal $k$-point graph construction generates a robust estimator of distribution entropy.
7. The asymptotic results presented hold for a very general class of graphs constructed by minimizing a total edge weight function which is a quasi-additive functional. This class includes the optimal Euclidean traveling salesman tour, the minimal spanning tree, the Steiner tree, and the two population minimal matching graph.

The outline of the paper is as follows. In Section II-A we review Euclidean minimal spanning graphs. In Section II-B we review the theory of quasi-additive functionals which were used by Redmond and Yukich to prove a general asymptotic theorem on the edge weight function of minimal spanning graphs. A minimal $k$-point graph is defined in Section III and in Section III-A we give a lemma which specifies a partition approximation under when the graph satisfies a tightly coverable assumption. Then in Section IV we treat the asymptotic theory of greedy approximations with a series of lemmas and convergence results. This is followed in Section VII by a study of quantitative robustness of the greedy approximation via the influence function.

## II. Background

Assume that $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ is a realization of $n$ i.i.d. random vectors where each $x_k$ takes values in $\mathbb{R}^d$ and has distribution $P$ with Lebesgue density $f$. Additional smoothness assumptions on $f$ will be required and will be given in the sequel. To simplify certain mathematical technicalities we will assume that the distribution is supported on the unit cube $[0,1]^d$. Any finitely supported distribution can be mapped to this domain by invertible linear transformation. Although we do not prove it, the restriction to finite support can undoubtedly be relaxed for densities satisfying the tail decay bounds of [49].

### A. *Minimal Euclidean Graphs*

An $n$-point (Euclidean) undirected graph $G$ is defined by a set of vertices $\mathcal{X} = \{x_1, \ldots, x_n\}$ and a set of edges $\mathcal{E} = \{e_{ij}\}$, where each edge $e_{ij} = (x_i, x_j)$ connects a pair of vertices $x_i, x_j$. If for two vertices $x$ and $y$ a graph $G$ has a sequence of edges $(x, x_{j_1}), (x_{j_1}, x_{j_2}), \ldots, (x_{j_p}, y)$ then $G$ is said to contain a path from $x$ to $y$. A graph is said to be connected if there exists a path between any pair of its vertices. If there exists a sequence of distinct edges which provide a path from any vertex back to itself the graph is said to contain a cycle. A graph which contains no cycles is an acyclic graph called a tree. By the span of a graph we mean the set of vertices which are connected by edges. The degree of a graph is the maximum number of edges which can be incident on any single vertex. The complete graph over $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ is a graph for which each pair of vertices are connected by an edge; such a graph has $\binom{n}{2}$ edges, is connected, has cycles, and is of degree $n$.

Different graphs can be compared based on their weights which are defined as follows. If $x = [[x]_1, \ldots, [x]_d]^T$ and $y = [[y]_1, \ldots, [y]_d]^T$ are two vertices connected by an edge $e$ we denote by $|e|$ the Euclidean length $\|x - y\| = \sqrt{\sum_{k=1}^d ([x]_k - [y]_k)^2}$ of the edge. Let $\psi$ be an edge weight function which satisfies $\psi(|e|) \geq 0$. The total weight $L_G(\mathcal{X})$ of a graph $G$ with edges $\{e\}$ and vertices $\mathcal{X}$ is defined as the sum of the edge weights

$$L_G(\mathcal{X}) = \sum_{e \in G} \psi(|e|). \tag{1}$$

While as in [53] the results of this paper might be extended to general weight functions which satisfy $\psi(|e|) \sim O(|e|^\gamma)$ as $|e|$ approaches 0, we will restrict our attention to the case of "power weighted edges" of exponential order $\gamma$

$$\psi(|e|) = |e|^\gamma, \qquad 0 < \gamma < d. \tag{2}$$

#### A.1 Euclidean Traveling Salesman Problem

In the Euclidean traveling salesman (TSP) problem the objective is to find a graph of minimum weight among those that visit each point in $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ exactly once. The resultant graph is called the *minimal TSP tour.* This problem is NP-hard and arises in many different areas of operations research [38]. Let $\mathcal{T}(\mathcal{X}_n)$ denote the possible sets of edges in the class of graphs of degree 2 which span $\mathcal{X}_n$. The weight of the minimal TSP tour is specified by

$$L_{\text{TSP}}(\mathcal{X}_n) = \min_{\mathcal{T}(\mathcal{X}_n)} \sum_{e \in \mathcal{T}(\mathcal{X}_n)} |e|^\gamma.$$

### A.2 Euclidean Minimal Spanning Tree Problem

In the Euclidean minimal spanning tree (MST) problem the objective is to find a graph of minimum weight which spans nodes $\mathcal{X}_n = \{x_1, \ldots, x_n\}$. This problem admits exact solutions which run in polynomial time and arises for $d = 2$ in VLSI circuit layout and network provisioning [35], [42], two sample matching [26],pattern recognition [55], clustering [61], nonparametric regression [8] and testing for randomness [33]. Let $\mathcal{M}(\mathcal{X}_n)$ denote the possible sets of edges in the class of acyclic graphs which span $\mathcal{X}_n$. The weight of the MST is specified by

$$L_{\mathrm{MST}}(\mathcal{X}_n) = \min_{\mathcal{M}(\mathcal{X}_n)} \sum_{e \in \mathcal{M}(\mathcal{X}_n)} |e|^\gamma.$$

### A.3 Euclidean Steiner Tree Problem

In the Euclidean Steiner tree (ST) problem a set of additional nodes $\mathcal{Y}$, called Steiner nodes, can be inserted into the MST problem to reduce the weight required to span the nodes $\mathcal{X}_n = \{x_1, \ldots, x_n\}$. The first formulation of this problem seems to be attributed to Gauss in the context of connecting 3 towns with a network of roads of minimum overall length [3]. Steiner tree problems are NP-hard but have been of interest for minimum area routing problems, e.g. in VLSI layout [16]. Let $\mathcal{S}(\mathcal{X}_n \cup \mathcal{Y})$ denote the possible sets of edges in the class of acyclic graphs which span $\mathcal{X}_n \cup \mathcal{Y}$, where the finite set $\mathcal{Y}$ is free. The weight of the minimal ST is specified by

$$L_{\mathrm{ST}}(\mathcal{X}_n) = \min_{\mathcal{Y}, \mathcal{S}(\mathcal{X}_n \cup \mathcal{Y})} \sum_{e \in \mathcal{S}(\mathcal{X}_n \cup \mathcal{Y})} |e|^\gamma.$$

The asymptotic behavior of each of the weight functions $L_{\mathrm{TSP}}, L_{\mathrm{MST}}$ and $L_{\mathrm{ST}}$ can be studied using the more general concept of quasi-additive Euclidean functionals introduced by Redmond and Yukich [45] and extended to the case of power weighted edges in [46] and [59].

## *B. Quasiadditive Euclidean Functionals*

Let $F$ be a finite subset of $[0,1]^d$, i.e. a set of points. Define the following conditions on a real valued set function $L$ [45], [46]:

**Null Condition**: $L(\phi) = 0$, where $\phi$ is the null set.

**Subadditivity**: There exists a constant $C_1$ with the following property: If $\mathcal{Q}^m = \{Q_i\}_{i=1}^{m^d}$ is a uniform partition of $[0,1]^d$ into $m^d$ cubes $Q_i$ each of edge length $m^{-1}$ and volume $m^{-d}$ and if $\{q_i\}_{i \le m^d}$ is the set of points in $[0,1]^d$ which translate each $Q_i$ back to the origin such that $Q_i - q_i = [0, m^{-1}]^d$, then

$$L(F) \le m^{-1} \sum_{i=1}^{m^d} L(m[(F \cap Q_i) - q_i]) + C_1 m^{d-\gamma}$$

**Superadditivity**: There exists a constant $C_2$ with the following property:

$$L(F) \ge m^{-1} \sum_{i=1}^{m^d} L(m[(F \cap Q_i) - q_i]) - C_2 m^{d-\gamma}$$

**Continuity**: There exists a constant $C_3$ such that for all finite subsets $F$ and $G$ of $[0,1]^d$

$$|L(F \cup G) - L(F)| \le C_3 \left(\mathrm{card}(G)\right)^{(d-\gamma)/d}$$

$L$ is said to be a *continuous subadditive functional* if it satisfies the null condition, subadditivity and continuity. $L$ is said to be a *continuous superadditive functional* if it satisfies the null condition, superadditivity and continuity.

*Definition 1:* A continuous subadditive functional $L$ is said to be a *quasi-additive functional* when there exists a continuous superadditive functional $L^*$ which satisfies $L(F)+1 \geq L^*(F)$ and the approximation property

$$|E[L(U_1,\ldots,U_n)] - E[L^*(U_1,\ldots,U_n)]| \leq C_4 n^{(d-\gamma-1)/d} \tag{3}$$

where $U_1,\ldots,U_n$ are i.i.d. uniform random vectors in $[0,1]^d$.

When such a functional $L^*$ exists it is called the dual of $L$. As shown in Redmond and Yukich [45, Thm. 1.3] and [46, Thm 2.3] duals can frequently be constructed by identifying a related boundary rooted graph over $\mathcal{X}_n$. In [46, Thm 2.3] it is shown that $L$ is quasi-additive for the following minimal graph problems: the minimal TSP tour, the MST, and the two population minimal matching problem. The following theorem is proven[2] in [46].

*Theorem 1:* Let $L$ be a quasi-additive Euclidean functional with power-exponent $\gamma$, and let $\mathcal{X}_n = \{x_1,\ldots,x_n\}$ be an i.i.d. sample drawn from a distribution on $[0,1]^d$ with an absolutely continuous component having (Lebesgue) density $f(x)$. Then

$$\lim_{n\to\infty} L(\mathcal{X}_n)/n^{(d-\gamma)/d} = \beta_{L,\gamma} \int f(x)^{(d-\gamma)/d} dx, \quad (a.s.) \tag{4}$$

In Theorem 1 $\beta_{L,\gamma}$ is a constant which only depends on $\gamma$ and the definition of the functional $L$, i.e. the graph optimality criterion (TSP, MST, or Steiner tree). In particular, $\beta_{L,\gamma}$ is independent of the distribution of the $x_i$'s. Theorem 1 is a generalization of Steele's work [53] which itself is a generalization of the well known Beardwood, Halton and Hammersley Theorem [10].

## III. Minimal $k$-point Euclidean Graphs

We denote by $\mathcal{X}_{n,k}$ a $k$-point subset of $\mathcal{X}_n$, $0 < k \leq n$. The minimal k-point Euclidean graph problem is to find the subset of points $\mathcal{X}_{n,k} = \mathcal{X}^*_{n,k}$ and the set of edges connecting these points such that the resultant graph has minimum total weight $L(\mathcal{X}^*_{n,k})$

$$L(\mathcal{X}^*_{n,k}) = \min_{\mathcal{X}_{n,k}} L(\mathcal{X}_{n,k}).$$

Define $k = \lfloor \alpha n \rfloor$ as the integer part of $\alpha n$. For the purposes of asymptotic analysis we will fix $\alpha \in (0,1)$ and study the behavior of $L(\mathcal{X}_{n,\lfloor \alpha n \rfloor})$ as $n \to \infty$. More generally define the weight functional $L_\alpha(F)$ as the $k$-minimal graph which spans $k = \lfloor \alpha \mathrm{card}(F) \rfloor$ of the points in the finite set $F$. Then $L(\mathcal{X}^*_{n,k}) = L_\alpha(\mathcal{X}_n)$. It is not difficult to see that $L_\alpha(F)$ neither satisfies the subadditivity property nor the continuity property. Hence, the elegant methods of Steele [53], Redmond and Yukich [45], and Rhee [48] cannot be directly applied.

### A. A Tight Cover Property

Let $\mathcal{Q}^m$ be a uniform partition of $[0,1]^d$ into $m^d$ cubes $Q_i$ of edge length $1/m$. The quantity $1/m$ is called the resolution of the partition. The following definition of a tightly coverable graph specifies a class of $k$-point graph

[2] In [46] Redmond and Yukich actually prove even stronger convergence (complete convergence) of the functional $L_n = L(\mathcal{X}_n)$ and give an asymptotic convergence rate.

algorithms for which the vertices of the minimal graph can be covered by a small number of partition cells $Q_i$.

*Definition 2* (**Tightly Coverable Graphs**) Let $\mathcal{Q}^m$, $m = 1, 2, \ldots$, be a sequence of uniform partitions of $[0,1]^d$ of resolution $1/m$ and let $\sigma(\mathcal{Q}^m)$ denote the sigma-algebra generated by the partition cells in $\mathcal{Q}^m$. Let $\mathcal{X}_n$ be $n$ i.i.d. uniform samples from $[0,1]^d$ and define $G_n$ as the complete graph spanning $\mathcal{X}_n$. For $\alpha \in [0,1]$ let $\mathcal{G}$ be an algorithm which constructs a subgraph of $G_n$ with $k = \lfloor \alpha n \rfloor$ vertices $\mathcal{X}_{n,k} \subset \mathcal{X}_n$. Define $D_k^m = \cap_{\{C \in \sigma(\mathcal{Q}^m) : \mathcal{X}_{n,k} \in C\}} C$ the minimum volume partition set which covers $\mathcal{X}_{n,k}$. The algorithm $\mathcal{G}$ is said to generate tightly coverable subgraphs if for any $\epsilon > 0$ there exists an $M$ such that for all $m > M$

$$\limsup_{n \to \infty} \left| \frac{\text{card}(\mathcal{X}_n \cap D^m_{\lfloor \alpha n \rfloor}) - \lfloor \alpha n \rfloor}{n} \right| \leq \epsilon, \qquad (a.s.)$$

Tightly coverable subgraphs have the property that the vertices $\mathcal{X}_{n,k}$ can be dissected from the rest of the points in $\mathcal{X}_n$ using a scalpel of resolution $1/m$. For an arbitrary distribution of vertices $\mathcal{X}_n$, this property would allow us to index $\mathcal{X}_{n,k}$ over the $\binom{n}{k}$ possible combinations of $k$-point sets of $\mathcal{X}_n$ by indexing the sets of vertices $D \cap \mathcal{X}_n$ over the partition-generated subsets $D \in \sigma(\mathcal{Q}^m)$ of $[0,1]^d$.

### *B. A Covering Lemma*

Consider the class (sigma algebra) $\sigma(\mathcal{Q}^m)$ of sets of resolution $1/m$. Out of this class we define $C_\alpha^m$ to be a set of probability at least $\alpha$ for which $L(C \cap \mathcal{X}_n)$ is minimized

$$C_\alpha^m = \text{argmin}_{C \in \sigma(\mathcal{Q}^m) : P(C) \geq \alpha} L(C \cap \mathcal{X}_n).$$

where $P(C) = P(x_i \in C)$. As in the tight cover definition, define $D_k^m$ as the minimum volume set in $\sigma(\mathcal{Q}^m)$ containing $\mathcal{X}^*_{n,k}$

$$D_k^m = \cap_{C \in \sigma(\mathcal{Q}^m) : C \supset \mathcal{X}^*_{n,k}} C.$$

Note that $\mathcal{X}^*_{n,k}$ is contained in set $D_k^m$ but may not be contained in $C_\alpha^m$.

With these definitions we have:

*Lemma 1:* Let $L$ be a quasi-additive functional with power exponent $\gamma$ as in Theorem 1. If $\text{card}(\mathcal{X}_n \cap C_\alpha^m) \geq k$

$$\begin{aligned} &\left| L(\mathcal{X}^*_{n,k}) - L(\mathcal{X}_n \cap C_\alpha^m) \right| / n^{(d-\gamma)/d} \leq \\ &C_3 \left[ \left( \frac{\text{card}(\mathcal{X}_n \cap C_\alpha^m) - k}{n} \right)^{(d-\gamma)/d} + 2 \left( \frac{\text{card}(\mathcal{X}_n \cap D_k^m) - k}{n} \right)^{(d-\gamma)/d} \right]. \end{aligned} \tag{5}$$

A proof of the lemma is given in Appendix A. When $k = \lfloor \alpha n \rfloor$ the first additive term on the right side of the inequality of Lemma 1 can be shown to converge a.s. to a term of order $O(m^{\gamma - d})$ using Lemma 6 and arguments similar to those used to prove Lemma 7 in the sequel. Thus if the minimal $k$-point graph can be shown to be tightly coverable, for any $\epsilon > 0$ there exists an $M$ such that: $\limsup_{n \to \infty} |L(\mathcal{X}^*_{n,k}) - L(X_n \cap C_\alpha^m)|/n^{(d-\gamma)/d} \leq \epsilon$ (a.s.) for all $m > M$. It would then be possible to show that $L(\mathcal{X}^*_{n,k})/n^{(d-\gamma)/d}$ would converge to the same a.s. limit as that of the greedy minimal $k$-point approximation given in the next section. Conversely, if the greedy approximation given below is not asymptotically equivalent to the exact minimal $k$-point graph then the latter graph does not satisfy the tightly coverable condition. The question whether the tightly coverable condition holds or not for the minimal $k$-point graph is an open problem.

## IV. Limit Theorem for $k$-point Greedy Approximation

Since the computation of the exact minimal $k$-point graph $\mathcal{X}^*_{n,k}$ has complexity which is exponential in the number of points $n$, the asymptotics of polynomial-time approximations are also of interest. Here we obtain asymptotic results for a greedy algorithm originally introduced by Ravi, Sundaram, Marathe, Rosenkrantz and Ravi [43] for constructing approximations to the $k$-MST on the plane. Their algorithm produces graphs which by construction satisfy the tightly coverable property introduced in the last section. Here we define a generalized version of their algorithm which constructs graphs in $d$ dimensions, $d > 1$, using arbitrary quasi-additive edge weight functions.

The algorithm is implemented in three steps: 1) the user specifies a uniform partition $\mathcal{Q}^m$ of $[0,1]^d$ having $m^d$ cells $Q_i$ of resolution $1/m$; 2) the algorithm finds the smallest subset $B^m_k = \cup_i Q_i$ of partition elements containing at least $k$ points; 3) out of this smallest subset the algorithm selects the $k$ points $\mathcal{X}_{n,k}$ which minimize $L(\mathcal{X}_{n,k})$. Stage 3 requires finding a $k$-point minimal graph on a much reduced set of points, that is typically only slightly larger than $k$ if $m$ is suitably chosen, which can be performed in polynomial time.

The smallest subset mentioned in Stage 2 of the algorithm is not unique. Figures 1 and 2 show an example with $m = 5$, $k = 17$ for which there are two possible smallest subsets, in this case both contain 18 points.

Similarly to [44], [43] we specify a small subset by the following greedy algorithm: i) find a reindexing $\{Q_{(i)}\}_{i=1}^{m^d}$ of the cells in $[0,1]^d$ ranked in decreasing order of the number of contained points, $\text{card}(\mathcal{X}_n \cap Q_{(1)}) \geq \ldots \geq \text{card}(\mathcal{X}_n \cap Q_{(m^d)})$ (if there are equalities arrange these in lexicographical order); ii) select the subset specified in Stage 2 by the recursion:

**Greedy Subset Selection Algorithm**

```
Initialize: $B = \phi$, $j = 1$
 Do until card$\{\mathcal{X}_n \cap B\} \geq k$
 $B = B \cup Q_{(j)}$
 End $j = j + 1$
```

At termination of the algorithm $j = \tilde{q} \leq m^d$ and we have a minimal subset $B^m_{\lfloor \alpha n \rfloor} \overset{\text{def}}{=} B = \cup_{i=1}^{\tilde{q}} Q_{(i)}$ containing at least $k$ points. Below we will use the notation $\mathcal{X}^{G_m}_{n,k}$ to denote the $k$ vertices of the graph found by the greedy algorithm.

It should not be surprising that as $n \to \infty$ the greedy subset selection algorithm should produce the smallest resolution-$1/m$ set $A$ of probability close to $\alpha = k/n$. Indeed, this is the basis for the asymptotic theorems stated below. Therefore we next specify a class of minimal subsets in the sigma-algebra $\sigma(\mathcal{Q}^m)$ which have coverage probability of at least $\alpha$.

Define the cell probabilities $\varphi_i = \int_{Q_i} f(x)dx$, $i = 1, \ldots, m^d$. If for any $C \in \sigma(\mathcal{Q}^m)$ satisfying $P(C) \geq \alpha$ the set $A \in \sigma(\mathcal{Q}^m)$ satisfies

$$P(C) \geq P(A) \geq \alpha,$$

then $A$ is called a *minimal resolution-*$1/m$ *set of probability at least* $\alpha$. The class of all such sets is denoted $\mathcal{A}^m_\alpha$ and, as shown in the following construction, all sets in $\mathcal{A}^m_\alpha$ have identical coverage probabilities $p_{\mathcal{A}^m_\alpha} \geq \alpha$.

The class $\mathcal{A}_\alpha^m$ can be generated by the following greedy algorithm: i) find a reindexing $\{Q_{(i)}\}_{i=1}^{m^d}$ of the cells in $\mathcal{Q}^m$ ranked in decreasing order of cell probabilities, $\varphi_{(1)} \geq \ldots \geq \varphi_{(m^d)}$; ii) select a subset $A$ using the Greedy Subset Selection Algorithm applied to the modified cell ordering prescribed in i). Assume the greedy algorithm terminates at iteration $j = q$. If $\varphi_{(q)} > \varphi_{(q+1)}$ then $A$ is the only set in the class $\mathcal{A}_\alpha^m$. Otherwise, let there be $K - I$ identical values of $\varphi_i$ satisfying $\varphi_{(I-1)} > \varphi_{(I)} = \ldots = \varphi_{(q)} = \ldots = \varphi_{(K)} > \varphi_{(K+1)}$ where $I \leq q$ and $K > q$. Then $\mathcal{A}_\alpha^m$ contains $S = \binom{K-I}{q-I+1}$ sets $\{A_i\}$ constructed by taking the last $q - I + 1$ cells that the greedy algorithm added to $A$ and exchanging them with any of the $q - I + 1$ possible combinations of cells in the set $\{Q_{(I)}, \ldots, Q_{(K)}\}$. Each of these sets $A_i$ is composed of an identical number $q$ of dissecting cells $\{Q_j^{A_i}\}_{j=1}^q$ in $\mathcal{Q}^m$ having identical sets of coverage probabilities $\{P(Q_j^{A_i})\}_{j=1}^q = \{\varphi_{(j)}\}_{j=1}^q$, and satisfying $P(A_i) = p_{\mathcal{A}_\alpha^m} \geq \alpha$, $i = 1, \ldots, S$.

Before developing the main result of this section we define some additional notation.

We will be interested in two special subsets generated by $\mathcal{A}_\alpha^m$. The interior $\overline{\mathcal{A}_\alpha^m}$ defined as the intersection of all $S$ sets in $\mathcal{A}_\alpha^m$

$$\overline{\mathcal{A}_\alpha^m} = \cap_{A_i \in \mathcal{A}_\alpha^m} A_i \tag{6}$$

and the associated residual set

$$\begin{aligned} \partial \mathcal{A}_\alpha^m &= \cup_{A_i \in \mathcal{A}_\alpha^m} A_i - \cap_{A_i \in \mathcal{A}_\alpha^m} A_i \\ &= \cup_{A_i, A_j \in \mathcal{A}_\alpha^m} A_i A_j^c. \end{aligned} \tag{7}$$

$\overline{\mathcal{A}_\alpha^m}$ is the "core" of the set $\mathcal{A}_\alpha^m$ and $\partial \mathcal{A}_\alpha^m$ is the "crust" of the set.

The total variation $v(Q)$ over a rectangle $Q \subset [0,1]^d$ of a function $g$ on $\mathbb{R}^d$ is defined as [50]

$$v(Q) = \limsup_{\{z_i\} \in Q} \sum_i |g(z_i) - g(z_{i-1})|, \tag{8}$$

where the limsup is taken over all countable subsets $\{z_1, z_2, \ldots,\}$ of points in $Q$. The function $g$ is said to have bounded variation over $Q$ if $v(Q) < \infty$. By convention, $v(\phi) = 0$ for $\phi$ the empty set.

To simplify the presentation we assume throughout that the distribution $P$ of each of the i.i.d. points in $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ is absolutely continuous with respect to Lebesgue measure and has a density $f(x)$. The results of [53] and [45] assert that the addition of singular components, e.g. delta functions, to the density does not change the asymptotics of $L(\mathcal{X}_n)$. Here it does change the asymptotics since the points of support of singular components entail zero edge weights and are therefore more attractive to include in the minimum $k$-point graph. However, the effect of singular components will only be to change the value of a threshold $\eta$ on $f(x)$ (see remark below). The main result of this section is the following asymptotic theorem.

*Theorem 2:* Let $\mathcal{X}_n$ be an i.i.d. sample from a distribution having Lebesgue density $f(x)$. Fix $\alpha \in [0,1]$, $\gamma \in (0,d)$. Let $f^{(d-\gamma)/d}$ be of bounded variation over $[0,1]^d$ and denote by $v(A)$ its total variation over a subset $A \subset [0,1]^d$. Let $L$ be a quasi-additive functional with power exponent $\gamma$ as in Theorem 1. Then, the total edge weight $L(\mathcal{X}_{n,k}^{G_m})$ of a $k$-point graph constructed by the resolution-$1/m$ greedy algorithm satisfies

$$\begin{aligned} \limsup_{n \to \infty} &\left| L(\mathcal{X}_{n,\lfloor \alpha n \rfloor}^{G_m})/n^{(d-\gamma)/d} - \beta_{L,\gamma} \int_{A_\alpha^m} f^{(d-\gamma)/d}(x) dx \right| \\ &< \delta, \quad (a.s.), \end{aligned} \tag{9}$$

where $A^m_\alpha \in \mathcal{A}^m_\alpha$ is any minimal resolution-$1/m$ set of probability at least $\alpha$,

$$\begin{aligned} \delta &= 2m^{-d}\beta_{L,\gamma}\sum_{i=1}^{m^d} v(Q_i \cap \partial\mathcal{A}^m_\alpha) + C_3(p_{\mathcal{A}^m_\alpha} - \alpha)^{(d-\gamma)/d} \\ &= O(m^{\gamma-d}), \end{aligned} \tag{10}$$

and $p_{\mathcal{A}^m_\alpha}$ is the coverage probability of sets in $\mathcal{A}^m_\alpha$. Furthermore, the bound (9) holds pointwise when $L(\mathcal{X}^{G_m}_{n,\lfloor\alpha n\rfloor})$ is replaced by $E[L(\mathcal{X}^{G_m}_{n,\lfloor\alpha n\rfloor})]$.

We prove Theorem 2 in Section V. It is of interest here to relate the integral in (9) to the Rényi entropy of the density $f(x)$. The key to this relation is the following lemma which relates the integral over a set $A$ in $\mathcal{A}^m_\alpha$ in (9) to a constrained minimum over $A \in \sigma(\mathcal{Q}^m)$.

*Lemma 2:* Under the assumptions of Theorem 2

$$\int_{A^m_\alpha} f^{(d-\gamma)/d}(x)dx = \min_{A\in\sigma(\mathcal{Q}^m):P(A)\geq\alpha} \int_A f^{(d-\gamma)/d}(x)dx \\ +O(m^{-d}). \tag{11}$$

*Proof of Lemma 2*

First recall that by construction of $\mathcal{A}^m_\alpha$ the coverage probability of any set in $\mathcal{A}^m_\alpha$ satisfies for some $q$: $\sum_{i=1}^{q-1}\varphi_{(i)} < \alpha$, $\sum_{i=1}^{q}\varphi_{(i)} = p_{\mathcal{A}^m_\alpha} \geq \alpha$, and $0 \leq p_{\mathcal{A}^m_\alpha} - \alpha < \varphi_{(q)}$, where $\varphi_{(1)} \geq \ldots \geq \varphi_{(m^d)}$, are the rank ordered cell probabilities.

In view of Lemma 4 it is sufficient to show that (11) holds for blocked densities of the form $f(x) = \tilde{f}(x) = \sum_{i=1}^{m^d}\theta_i I_{Q_i}(x)$. Observe that for any $A^m_\alpha \in \mathcal{A}^m_\alpha$ and for any $\eta$ satisfying $m^d\varphi_{(q-1)} < \eta \leq m^d\varphi_{(q)}$: if $x \in A^m_\alpha$ then $\tilde{f}(x) \geq \eta$. Equivalently, $\tilde{f}^{(d-\gamma)/d}(x) - \lambda\tilde{f}(x) \leq 0$ where $\lambda = \eta^{-\gamma/d}$. With $I_{A^m_\alpha}(x)$ the indicator function of $A^m_\alpha$ this implies that for any $A \in \sigma(\mathcal{Q}^m)$

$$I_{A^m_\alpha}(x)(\tilde{f}^{(d-\gamma)/d}(x) - \lambda\tilde{f}(x)) \leq I_A(x)(\tilde{f}^{(d-\gamma)/d}(x) - \lambda\tilde{f}(x)),$$

for all $x$. Therefore, integrating this inequality over $x \in [0,1]^d$

$$\int_{A^m_\alpha}\left(\tilde{f}^{(d-\gamma)/d}(x) - \lambda\tilde{f}(x)\right)dx \leq \int_A\left(\tilde{f}^{(d-\gamma)/d}(x) - \lambda\tilde{f}(x)\right)dx,$$

or

$$\int_{A^m_\alpha}\tilde{f}^{(d-\gamma)/d}(x)dx - \int_A \tilde{f}^{(d-\gamma)/d}(x)dx \\ \leq \lambda\left(\int_{A^m_\alpha}\tilde{f}(x)dx - \int_A \tilde{f}(x)dx\right). \tag{12}$$

Now, as $P(A^m_\alpha) = p_{\mathcal{A}^m_\alpha}$, if $P(A) \geq \alpha$ then the right side of this inequality is upper bounded by $\lambda(p_{\mathcal{A}^m_\alpha} - \alpha) \leq \varphi_{(q)} = m^{-d}\theta_{(q)}$. Hence, minimizing both sides of the inequality (12) over $A$ we obtain

$$\int_{A^m_\alpha}\tilde{f}^{(d-\gamma)/d}(x)dx \leq \min_{\{A\in\sigma(\mathcal{Q}^m):P(A)\geq\alpha\}}\int_A \tilde{f}^{(d-\gamma)/d}(x)dx \\ +O(m^{-d}).$$

Since, obviously,

$$\int_{A^m_\alpha}\tilde{f}^{(d-\gamma)/d}(x)dx \geq \min_{\{A\in\sigma(\mathcal{Q}^m):P(A)\geq\alpha\}}\int_A (\tilde{f}^{(d-\gamma)/d}(x)dx$$

the lemma is established. □

We can now show how Theorem 2 can be related to estimation of Rényi entropy. Using Lemma 2 and the fact that $\sigma(\mathcal{A}^m)$ converges to the class of Borel subsets $\mathcal{B}$ of $[0,1]^d$, it is an easy exercise to show that

$$\begin{aligned}\lim_{m\to\infty}\int_{A_\alpha^m} f^{(d-\gamma)/d}(x)dx &= \inf_{A\in\mathcal{B}:P(A)\geq\alpha}\int_A f^{(d-\gamma)/d}(x)dx \\ &= \inf_{A\in\mathcal{B}:P(A)=\alpha}\int_A f^{(d-\gamma)/d}(x)dx. \qquad (13)\end{aligned}$$

Now for any Borel set $A$ in $[0,1]^d$ having $P(A)>0$ define the conditional density $f(x|A)=f(x)/P(A)I_A(x)$ where $I_A(x)$ is the indicator function of $A$. For a continous density $f(x|A)$ the (differential) Rényi entropy of order $\nu\in(0,1)$ is defined as

$$R_\nu(f|A)=\frac{1}{1-\nu}\log\int f^\nu(x|A)dx. \qquad (14)$$

This is also called the conditional Rényi entropy given $A$. The Renyi entropy shares a number of properties with the Shannon entropy such as: it is concave as a function of the density (for $0<\nu<1$), it is maximized (over all densities with bounded support) for a uniform density.

As $1-\nu>0$ minimization of $R_\nu(f|A)$ over $A$ is equivalent to minimization of the integral in (14). Let $A_o$ be the probability-at-least-$\alpha$ Borel subset of $[0,1]^d$ which minimizes $R_\nu(f|A)$

$$R_\nu(f|A_o)=\inf_{\{A\in\mathcal{B}:P(A)\geq\alpha\}}R_\nu(f|A). \qquad (15)$$

For $\nu=(d-\gamma)/d$ define the following function of $L(\mathcal{X}_{n,\lfloor\alpha n\rfloor}^{G_m})$

$$\hat{R}_\nu \stackrel{\rm def}{=} \frac{1}{1-\nu}\left(\log L(\mathcal{X}_{n,\lfloor\alpha n\rfloor}^{G_m})/(\lfloor\alpha n\rfloor)^\nu-\log\beta_{L,\gamma}\right) \qquad (16)$$

An immediate consequence of Theorem 2 is the following.

*Theorem 3:* Under the assumptions of Theorem 2 $\hat{R}_\nu$ is a strongly consistent estimator of the minimum conditional Rényi entropy $R_\nu(f|A_o)$ of order $\nu\in(0,1)$ as $m,n\to\infty$.

Before developing the proof of Theorem 2 in the next section, we make the following remarks.

1. The bound $\delta$ in Theorem 2 is tight since it reduces to zero for the case $\alpha=1$, yielding the classical a.s. BHH limit theorem, Theorem 1, for minimal graphs spanning all $n$ points $\mathcal{X}_n$. Indeed, in this case, for arbitrary $m>0$ the class $\mathcal{A}_\alpha^m$ of resolution-$1/m$ probability-at-least-$\alpha$ sets contains the support set of $f$ and therefore satisfies $p_{\mathcal{A}_\alpha^m}-\alpha=0$, $v(\partial\mathcal{A}_\alpha^m)=0$, and $v(Q_i\cap\partial\mathcal{A}_\alpha^m)=0$ as required.
2. Theorems 2 and 3 are easily extended to the case where the density of $P$ contains singular components, e.g. delta functions. Specifically, let $P$ have the mixed density $f(x)dx+\mu_s$ where $dx$ is Lebesgue measure, $f(x)$ is the absolutely continuous component and $\mu_s$ is the singular component of $P$ relative to Lebesgue measure. Let the support of the singular measure be $A_s$ and let $\alpha_s=\mu_s([0,1]^d)=P(A_s)<\alpha$. Then, we know from Lemma 6 that $\lim_{n\to}L(\mathcal{X}_n\cap A_s)=0$ and hence points falling in the singular part $A_s$ of $[0,1]^d$ contribute negligible weight. Thus, by the strong law of large numbers, $\alpha_s n$ of the $n$ points can be included in the graph at negligible cost leaving only $(\alpha-\alpha_s)n$ points in $[0,1]^d-A_s$ whose edge weights are asymptotically significant. Therefore, in the case of a singular component with $\alpha_s<\alpha$, Theorem 2 holds with the class $\mathcal{A}_\alpha^m$

replaced by the class $\mathcal{A}^m_{\alpha-\alpha_s}$ of resolution $1/m$ subsets with coverage probability at least $\alpha - \alpha_s$. If $\alpha_s \geq \alpha$ then $\mathcal{A}^m_\alpha$ is replaced by the empty set and $L(\mathcal{X}^{G_m}_{\lfloor \alpha n \rfloor})$ converges to zero a.s. as $m, n \to \infty$. Likewise, it can be shown that when there is a singular component Lemma 2 holds with the minimization over the class $\{A \in \sigma(\mathcal{Q}^m) : P(A) \geq \alpha\}$ replaced by a minimization over the smaller class

$$\left\{ A \in \sigma(\mathcal{Q}^m) : \int_A f(x)dx \geq \max\{\alpha - \alpha_s, 0\} \right\}.$$

3. As can be seen from Lemma 7, Theorem 2 holds for any method of selection of $k = \lfloor \alpha n \rfloor$ points from the minimal subset $B^m_k$ of $\sigma(\mathcal{Q}^m)$ covering at least $k$ points. It does not depend on the precise way that a few points are eliminated from the set $B^m_k$ to form $\mathcal{X}^{G_m}_{n,k}$. This implies that for large $n$ all methods of elimination are equivalent, including simply randomly rejecting points from $B^m_k$ until exactly $k$ points remain.

4. In smooth estimation problems the normalization factor required to ensure convergence of a parameter estimator to a finite non-zero constant is typically $1/n$. In contrast Theorem 2 says that the stabilization factor for $L(\mathcal{X}^{G_m}_{n,\lfloor \alpha n \rfloor})$ is the larger quantity $1/n^{(d-\gamma)/d}$, i.e., $L(\mathcal{X}^{G_m}_{n,\lfloor \alpha n \rfloor})$ is less explosive as a function of $n$. On the other hand, inspired by [46, Thm. 2.5] which gives the tight convergence rate $|E[L_{MST}(\mathcal{X}_n)] - \beta_{L_{MST},\gamma} n^{(d-\gamma)/d}| = O\left(\max(1, n^{(d-\gamma-1)/d})\right)$ for the MST under the uniform distribution, we conjecture that the rate of convergence in the limsup of Theorem 2 is at best $O(1/n^{1/d})$ and this rate can be attained only when $\gamma \leq d-1$. This leads us to believe that the entropy estimator (16) will have fastest convergence when the Rényi order parameter $\nu$ is in the range $1/d \leq \nu < 1$.

5. The bound $\delta$ of Theorem 2 decays to zero as a function of resolution $1/m$ at overall rate $O(m^{\gamma-d})$. The first term $2m^{-d}\beta_{L,\gamma}\sum_{i=1}^{m^d} v(Q_i \cap \partial A^m_\alpha) \leq 2\beta_{L,\gamma} m^{-d} v([0,1])$ decays as $m^{-d}$ and is due to non-uniqueness of the resolution-$1/m$ subsets $A^m_\alpha \in \mathcal{A}^m_\alpha$ all of which have identical coverage probability but over which $f$ may have different amounts of variation. When $f$ is, in the terminology of [53], a "blocked distribution," $f(x) = m^{-d} \cdot \sum_{i=1}^{m^d} \varphi_i I_{Q_i}(x)$, over the resolution-$1/m$ cells this term is equal to zero. The second additive term $C_3(p_{\mathcal{A}^m_\alpha} - \alpha)^{(d-\gamma)/d}$ is due to the overshoot of coverage probability by the subsets $A^m_\alpha \in \mathcal{A}^m_\alpha$. This term is zero when it so happens that the $\alpha$ chosen in the greedy algorithm is exactly attainable by a $1/m$ resolution subset. However, this term decays to zero only as $m^{\gamma-d}$ and dominates the resolution convergence rate. Note that this implies that the rate of convergence in $m$ of the bound $\delta$ in Theorem 2 is fastest for small $\gamma$.

6. Since $\sup_{x \in Q_{(q)}} f^{(d-\gamma)/d}(x) \leq v([0,1]^d)$ and $\sum_{i=1}^{m^d} v(Q_i \cap \partial \mathcal{A}^m_\alpha) \leq v([0,1]^d)$ we can weaken the a.s. bound in Theorem 2 by using the result (33), which was shown in the proof of Theorem 2. This results in the following

$$\delta \leq \left[2\beta_{L,\gamma} m^{-d} + C_3 m^{\gamma-d}\right] v([0,1]^d). \tag{17}$$

This a.s. bound holds uniformly over the class of all density functions such that $f^{(d-\gamma)/d}$ has total variation less than or equal to $v([0,1]^d)$. Thus if an upper bound $\overline{v}$ on the total variation of an unknown density is available and a consistent estimate of conditional Rényi entropy $R_\nu = R_\nu(f|A_o)$ is desired such that

$$|L(\mathcal{X}^{G_m}_{n,\lfloor \alpha n \rfloor})/(\lfloor \alpha n \rfloor)^\nu - \beta_{L,\gamma}\exp\{-(1-\nu)R_\nu)| < \epsilon$$

the weakened bound (17) can be used to give a selection rule for the required partition resolution $1/m$

$$1/m \leq \frac{\epsilon}{(2 + C_3)\overline{v}}.$$

7. The Borel set $A_o$ of probability-at-least-$\alpha$ defined in (15) which minimizes Rényi entropy of order $\nu$ is independent of $\nu$ and can be constructed by a simple water filling procedure. To see this define the Lagrangian

$$\rho(A, \lambda) \quad \overset{\text{def}}{=} \quad \int_A f^{(d-\gamma)/d}(x)dx - \lambda\left(\int_A f(x)dx - \alpha\right)$$

for $\lambda > 0$. Consider an arbitrary Borel subset $\Delta A \subset [0,1]^d$ outside of $A_o$, $A_o \cap \Delta A = \phi$. By Kuhn-Tucker, $A_o$ must satisfy $\rho(A_o + \Delta A, \lambda) - \rho(A_o, \lambda) \leq 0$ for $A_o$ to minimize $\rho(A, \lambda)$ and hence minimize entropy. Here

$$\rho(A_o + \Delta A, \lambda) - \rho(A_o, \lambda) = \int_{\Delta A} \left(1 - \lambda f^{\gamma/d}\right) f^{(d-\gamma)/d},$$

which is negative when $A_o$ is defined by

$$A_o = \{x : f(x) \geq \eta\}. \tag{18}$$

where, if possible, $\eta = \lambda^{-d/\gamma} \geq 0$ is selected to satisfy $P(A_o) = \alpha$. Hence, in this case, the conditional density $f(x|A_o)$ in (16) is obtained by truncating $f(x)$ wherever it falls below $\eta$ and renormalizing to obtain a valid probability density integrating to 1 over $[0,1]^d$. See Fig. 3 for illustration.

When for any $a > 0$ the set $\{x : f(x) = a\}$ has (Lebesgue) measure zero $f(x)$ has no flat spots and it is always possible to select $\eta$ in (18) to satisfy $P(A_o) = \alpha$. Otherwise, we need to slightly modify the definition (18) of $A_o$. Let $\eta$ be such that the set $\{x : f(x) < \eta\}$ has probability $\alpha_-$, the set $\{x : f(x) = \eta\}$ has probability $\alpha_+ - \alpha_- > 0$ and assume that $\alpha \in (\alpha_-, \alpha_+)$. Then defining

$$A_o = \{x : f(x) \geq \eta\} \cup C, \tag{19}$$

where $C$ is an arbitrary Borel subset of $\{x : f(x) = \eta\}$ having $P(C) = \alpha - \alpha_-$, is an entropy minimizing subset of probability $\alpha$.

8. The minimum entropy set $A_o$ in (15) is not unique. For example any arbitrary probability zero set can be added to $A_o$ without affecting the entropy. A more interesting example occurs when $f$ is a uniform density for which case any set $A$ of area $\alpha$ minimizes entropy. In this case the assertion of Theorem 2 may come as a surprise since the largest distance between points in $A$ should be smallest for connected sets of small diameter, e.g., a sphere. However, let $A = \cup_{i=1}^{\infty} A_i$ be a countable union of disjoint sets $A_i$ and having vol$(A) = \alpha$. Note that only a single edge is needed between $A_i$ and $A_j$ to form a connected graph over any two sets. Thus in the limit of large $n$ the total edge weight of the graph is dominated by connections between points within each $A_i$ and not connections between different $A_i$. This is because the total edge weight depends more on the average edge weight than the maximum edge weight.

9. In the Rényi entropy estimator (16) the constant $\beta_{L,\gamma}$ is a bias offset which can in principle be computed offline as it does not depend on $f$. However, while upper and lower bounds are available, see e.g.[5] for MST bounds, analytic expressions for $\beta_{L,\gamma}$ are not available. Alternatively, for some estimation or classification problems only relative entropy may be needed, e.g. testing for different entropy rates between two populations via the ratio of $k$-point graph weight functionals, for which the bias offset need not be known.

10. Consider the case that $f = (1-\epsilon)f_1 + \epsilon f_o$ is a mixture of a nominal density $f_1$ of interest and a uniform contaminating density $f_0$. In order that $\epsilon$ be identifiable we assume that $\min_{x\in[0,1]^d} f_1(x) = 0$; this simply ensures that $f_1(x)$ have no uniform component. Then, since $f_1$ increasingly dominates $f_o$ as $\epsilon$ decreases, for small $\epsilon$ a suitable threshold $\eta$ $(\alpha)$ exists for which: $f(x|A_o) \approx f_1(x)$. Thus (16) can be viewed as a robust estimator of the Rényi entropy of the nominal density $f_1$.

## V. Proof of Theorem 2

Here we present a set of lemmas that are needed to prove Theorem 2. First we establish by Lemma 3 that any set $B_n^m$ obtained by the greedy algorithm belongs to class $\mathcal{A}_\alpha^m$ with probability close to one. Then it is shown in Lemma 4 that replacing $f(x)$ by its piecewise constant approximation leads to an approximation error to the integral in (9) that goes like $O(m^{-d})$. This allows us to establish in Lemma 5 that the length of a MST spanning all points in $B_n^m$ provides an estimate of $\int f^\nu(x)dx$. This result is then refined in Lemma 7 where it is shown that asymptotically the length of this MST increases at the same rate as the length of the $k$-MST spanning only $k = \lfloor \alpha n \rfloor$ of these points. It is then a simple matter to put Lemmas 5 and 7 together to prove Theorem 2.

*Lemma 3:* For given $\alpha \in [0,1]$ and a set of $n$ i.i.d. points $\mathcal{X}_n = [x_1, \ldots, x_n]^T$ let $B_n^m$ be the minimal cover of $\lfloor \alpha n \rfloor$ points with resolution-$1/m$ produced by the greedy subset selection algorithm. Then

$$P\left(\liminf_{n\to\infty} \{\mathcal{X}_n : B_n^m \in \mathcal{A}_\alpha^m\}\right) = 1.$$

*Proof of Lemma 3*

Define the $m^d$ independent random variables $N_i = \text{card}(\mathcal{X}_n \cap Q_i)$ of points in cell $Q_i$, $i = 1, \ldots, m^d$. By definition, the greedy algorithm gives a minimal cover $B_{\lfloor \alpha n \rfloor}^m$ containing at least $\alpha n$ points which satisfies the two conditions:

$$n^{-1}\sum_{i=1}^{\tilde{q}-1} N_{(i)} \quad < \quad \alpha \tag{20}$$

$$n^{-1}\sum_{i=1}^{\tilde{q}} N_{(i)} \quad \geq \quad \alpha. \tag{21}$$

Define $(i)_\varphi$ the index function which establishes a correspondence between rank ordered probabilities $\varphi_{(1)} \geq , \ldots, \geq \varphi_{(m^d)}$ and the cells $Q_i$ which support each of these probabilities: i.e. with this notation $P(x_i \in Q_{(i)_\varphi}) = \varphi_{(i)}$. For arbitrary $\epsilon > 0$ define the events $E_n(\epsilon)$ and $F_n$

$$E_n(\epsilon) \quad = \quad \left\{\mathcal{X}_n : n^{-1}\sum_{i=1}^{q-1} N_{(i)_\varphi} \leq \alpha - \epsilon\right\} \tag{22}$$

$$F_n \quad = \quad \left\{\mathcal{X}_n : n^{-1}\sum_{i=1}^{q} N_{(i)_\varphi} \geq \alpha\right\}. \tag{23}$$

Comparing these equations to (20) and (21) it will suffice to show $P(\liminf E_n \cap F_n) = 1$. Equivalently, since $P(\limsup E_n^c \cup F_n^c) \leq P(\limsup E_n^c) + P(\limsup F_n^c)$ we show that the latter two quantities are zero.

Define i.i.d. Bernouilli sequences $Y_n = \{y_1, \ldots, y_n\}$ and $Z_n = \{z_1, \ldots, z_n\}$ as

$$\begin{aligned} y_j &= I_{A_\alpha^m}(x_j) - I_{Q_{(q)_\varphi}}(x_j), \quad j = 1, \ldots, n \\ z_j &= I_{A_\alpha^m}(x_j), \quad j = 1, \ldots, n. \end{aligned}$$

and $p_y \overset{\text{def}}{=} P(y_j = 1) = P(x_i \in A_\alpha^m) - P(x_i \in Q_{(q)_\varphi}) = \sum_{i=1}^{q-1} \varphi_{(i)}$ and $p_z \overset{\text{def}}{=} P(z_j = 1) = P(x_i \in A_\alpha^m) = \sum_{i=1}^{q} \varphi_{(i)}$. Then we have the equivalent form for (22) and (23)

$$E_n(\epsilon) \quad = \quad \left\{Y_n : n^{-1}\sum_{j=1}^{n} y_j \leq \alpha - \epsilon\right\} \tag{24}$$

$$F_n \quad = \quad \left\{Z_n : n^{-1}\sum_{j=1}^{n} z_j \geq \alpha\right\}. \tag{25}$$

Let $\delta$ be defined as the smallest non-zero value of $\varphi_i$, $i = 1, \ldots, m^d$. Then by definition of $A_\alpha^m$ we have

$$p_y < \alpha \;\text{ and }\; p_z \geq \alpha + \delta. \tag{26}$$

From Sanov's theorem [17], [18]

$$\begin{aligned} P(E_n^c(\epsilon)) &\leq (n+1)^2 \ \exp\{-nK(\alpha-\epsilon, p_y)\} \\ P(F_n^c) &\leq (n+1)^2 \ \exp\{-nK(\alpha, p_z)\} \end{aligned}$$

where $K(p_1,p_2) = p_2 \ln p_1/p_2 + (1-p_2)\ln(1-p_2)/(1-p_1) \geq 0$ is the Kullback-Liebler distance between two Bernoulli probability distributions $\{p_1, 1-p_1\}$ and $\{p_2, 1-p_2\}$, $p_1, p_2 \in [0,1]$. Furthermore, from (26) and the fact that $K(p_1,p_2)$ is increasing in $|p_1 - p_2|$, we have for $\epsilon < (\alpha - p_y)/2$

$$\begin{aligned} P(E_n^c(\epsilon)) &\leq (n+1)^2 \exp\{-nK(\alpha-\epsilon, \alpha)\} \\ P(F_n^c) &\leq (n+1)^2 \exp\{-nK(\alpha, \alpha+\delta)\}. \end{aligned}$$

It is easily verified that for any $\rho > 0$

$$\sum_{n=1}^{\infty} (n+1)^2 e^{-n\rho} \leq \int_0^{\infty} (u+2)^2 e^{-u\rho} du = C_\rho,$$

where $C_\rho = 2(2\rho^2 + \rho + 1)/\rho^3$. Hence,

$$\begin{aligned} \sum_{n=1}^{\infty} P(E_n^c(\epsilon)) &\leq C_{K(\alpha-\epsilon,\alpha)} < \infty \\ \sum_{n=1}^{\infty} P(F_n^c) &\leq C_{K(\alpha,\alpha+\delta)} < \infty, \end{aligned}$$

and by Borel-Cantelli we have $P(\limsup E_n^c(\epsilon)) = 0$, $P(\limsup F_n^c) = 0$ and the lemma follows. □

While $B^m_{\lfloor \alpha n \rfloor}$ does not necessarily converge to any fixed set as $n \to \infty$, the preceeding lemma establishes that it converges to the equivalence class of sets defined by $\mathcal{A}^m_\alpha$.

The next result relates the error of a blocked distribution approximation of $\int f^{(d-\gamma)/d}(x)dx$ to the total variation of $f$.

*Lemma 4:* For $\nu \in [0,1]$ let $f^\nu$ be of bounded variation over $[0,1]^d$ and denote by $v(A)$ its total variation over any subset $A \in [0,1]^d$. Define the resolution $1/m$ block density approximation $\tilde{f}(x) = \sum_{i=1}^{m^d} \theta_i I_{Q_i}(x)$ where $\theta_i = m^d \int_{Q_i} f(x)dx$. Then for any $A \in \sigma(\mathcal{Q}^m)$

$$0 \leq \int_A [\tilde{f}^\nu(x) - f^\nu(x)]dx \leq m^{-d} \sum_{i=1}^{m^d} v(Q_i \cap A).$$

*Proof of Lemma 4*

First note that as $t^\nu$ is a convex cap function, by Jensen's inequality $|C|^{-1} \int_C f^\nu(x)dx \leq \left(|C|^{-1} \int_C f(x)dx\right)^\nu$ for any Borel set $C$ of positive volume $|C|$. The left side inequality of Lemma 4 now follows from the relation

$$\begin{aligned} &\int_A [\tilde{f}^\nu(x) - f^\nu(x)]dx \qquad (27) \\ &= m^{-d} \sum_{i: Q_i \cap A \neq \phi}^{m^d} \left[ \left( m^d \int_{Q_i} f(x)dx \right)^\nu - m^d \int_{Q_i} f^\nu(x)dx \right]. \end{aligned}$$

We next deal with the right side of the inequality in Lemma 4. As functions of bounded variation are continuous except at possibly a countable number of points [50], by the mean value theorem for each $Q_i$ there exist points $\xi_i \in Q_i$ and $\psi_i \in Q_i$ such that $\int_{Q_i} f(x)dx = f(\xi_i)m^{-d}$ and $\int_{Q_i} f^\nu(x)dx = f^\nu(\psi_i)m^{-d}$. Hence, using (27) and the definition (8) of $v$

$$\begin{aligned}
&\left|\int_A [\tilde{f}^\nu(x) - f^\nu(x)]dx\right| \\
&\quad = m^{-d}\left|\sum_{i:Q_i\cap A\neq\phi} (f^\nu(\xi_i) - f^\nu(\psi_i))\right| \\
&\quad \leq m^{-d}\sum_{i:Q_i\cap A\neq\phi} |f^\nu(\xi_i) - f^\nu(\psi_i)| \\
&\quad \leq m^{-d}\sum_{i:Q_i\cap A\neq\phi} v(Q_i) = m^{-d}\sum_{i=1}^{m^d} v(Q_i\cap A).
\end{aligned}$$

This establishes the Lemma. □

*Lemma 5:* Assume $f$ is of bounded total variation $v(Q_i)$ in each partition cell $Q_i \in \mathcal{Q}^m$. Let $A$ be any set in the class $\mathcal{A}_\alpha^m$. Then for any quasi-additive functional $L_n(B^m_{\lfloor\alpha n\rfloor}) \overset{\text{def}}{=} L(\mathcal{X}_n \cap B^m_{\lfloor\alpha n\rfloor})$

$$\begin{aligned}
\limsup_{n\to\infty} &\left|L_n(B^m_{\lfloor\alpha n\rfloor})/n^{(d-\gamma)/d} - \beta_{L,\gamma}\int_A f^{(d-\gamma)/d}(x)dx\right| \\
&< 2m^{-d}\beta_{L,\gamma}\sum_{i=1}^{m^d} v(Q_i\cap\partial\mathcal{A}_\alpha^m), \quad (a.s).
\end{aligned}$$

Furthermore, this same bound holds when $L_n(B^m_{\lfloor\alpha n\rfloor})$ is replaced by $E[L_n(B^m_{\lfloor\alpha n\rfloor})]$.

The following follows directly from Theorem 1

*Lemma 6:* Assume the conditions of Theorem 1 and let $A$ be an arbitrary Borel subset of $[0,1]^d$. Then for any quasi-additive functional $L_n(A) = L(\mathcal{X}_n \cap A)$

$$\lim_{n\to\infty} L_n(A)/n^{(d-\gamma)/d} = \beta_{L,\gamma}\int_A f^{(d-\gamma)/d}(x)dx, \quad (a.s.)$$

Furthermore, the above (a.s.) limit is the pointwise limit of $E[L_n(A)]/n^{(d-\gamma)/d}$ as $n \to \infty$.

*Proof of Lemma 5*

Let $\tilde{f}(x) = \sum_{i=1}^{m^d}\theta_i I_{Q_i}(x)$ be the blocked distribution approximation to $f(x)$ of Lemma 4. Now for any sets $A, A' \in \mathcal{A}_\alpha^m$ and $\nu = (d-\gamma)/d$, we have by the definitions (6), (7) of $\mathcal{A}_\alpha^m$, $\partial\mathcal{A}_\alpha^m$, and the triangle inequality

$$\begin{aligned}
\left|\int_A f^\nu - \int_{A'} f^\nu\right| &= \left|\int_{A\cap\partial\mathcal{A}_\alpha^m} f^\nu - \int_{A'\cap\partial\mathcal{A}_\alpha^m} f^\nu\right| \\
&\leq \left|\int_{A\cap\partial\mathcal{A}_\alpha^m} f^\nu - \int_{A\cap\partial\mathcal{A}_\alpha^m} \tilde{f}^\nu\right| \\
&\quad + \left|\int_{A'\cap\partial\mathcal{A}_\alpha^m} f^\nu - \int_{A'\cap\partial\mathcal{A}_\alpha^m} \tilde{f}^\nu\right| \\
&\quad + \left|\int_{A\cap\partial\mathcal{A}_\alpha^m} \tilde{f}^\nu - \int_{A'\cap\partial\mathcal{A}_\alpha^m} \tilde{f}^\nu\right|
\end{aligned}$$

By construction of $\mathcal{A}_\alpha^m$, the cell probabilities $\{\varphi_i\}_{i:Q_i\in A}$ and $\{\varphi_i\}_{i:Q_i\in A'}$ are identical so that the last term on the right side of the inequality is equal to zero. We therefore obtain by application of Lemma 4

$$\begin{aligned}\left|\int_A f^\nu - \int_{A'} f^\nu\right| &\leq 2\max_{A\in\mathcal{A}_\alpha^m} m^{-d}\sum_{i=1}^{m^d} v(Q_i\cap A\cap\partial\mathcal{A}_\alpha^m)\\ &\leq 2m^{-d}\sum_{i=1}^{m^d} v(Q_i\cap\partial\mathcal{A}_\alpha^m). \qquad (28)\end{aligned}$$

Next by Lemma 6 for any of the finite number of sets $A\in\mathcal{A}_\alpha^m$ and any $\epsilon>0$ there exists an integer $n_0=n_0(A)$ such that for all $n>n_0$

$$\left|L_n(A)/n^{(d-\gamma)/d} - \beta_{L,\gamma}\int_A f^{(d-\gamma)/d}(x)dx\right| \leq \epsilon, \quad (a.s.) \qquad (29)$$

Let $n_1$ be defined as the largest of the $\{n_0(A)\}_{A\in\mathcal{A}_\alpha^m}$. By Lemma 3 there exists an integer $n_2$ such that for all $n>n_2$

$$\min_{A\in\mathcal{A}_\alpha^m} L_n(A) \leq L_n(B_{\lfloor\alpha n\rfloor}^m) \leq \max_{A\in\mathcal{A}_\alpha^m} L_n(A), \quad (a.s). \qquad (30)$$

Now choosing $n_3=\max(n_2,n_1)$ it follows from (29) and (30) that for all $n>n_3$

$$\begin{aligned}&\beta_{L,\gamma}\min_{A\in\mathcal{A}_\alpha^m}\int_A f^{(d-\gamma)/d}(x)dx-\epsilon\\ &\leq L_n(B_{\lfloor\alpha n\rfloor}^m)/n^{(d-\gamma)/d}\\ &\leq \beta_{L,\gamma}\max_{A\in\mathcal{A}_\alpha^m}\int_A f^{(d-\gamma)/d}(x)dx+\epsilon, \quad (a.s.).\end{aligned}$$

Applying the bound (28) we have for arbitrary $A\in\mathcal{A}_\alpha^m$ and all $n>n_3$

$$\begin{aligned}&\left|L_n(B_{\lfloor\alpha n\rfloor}^m)/n^{(d-\gamma)}d - \beta_{L,\gamma}\int_A f^{(d-\gamma)/d}(x)dx\right|\\ &\leq 2m^{-d}\beta_{L,\gamma}\sum_{i=1}^{m^d} v(Q_i\cap\partial\mathcal{A}_\alpha^m)+\epsilon, \quad (a.s.).\end{aligned}$$

Since $\epsilon$ is arbitrary the a.s. limit of Lemma 5 follows.

It remains to show that the same bound also holds for the limit $E[L_n(B_{\lfloor\alpha n\rfloor}^m)]/n^{(d-\gamma)/d}$. It follows from Lemmas 3 and 6 that for any $\epsilon>0$ there exists $n_o$ such that for $n>n_o$, $P(B_{\lfloor\alpha n\rfloor}^m\in\mathcal{A}_\alpha^m)\geq\epsilon$ and

$$\left|E[L_n(A)]/n^{(d-\gamma)/d} - \beta_{L,\gamma}\int_A f^{(d-\gamma)/d}(x)dx\right| \leq \epsilon$$

for any $A\in\mathcal{A}_\alpha^m$. Furthermore, as $L_n$ is continuous it is bounded:

$$L_n(A)\leq C_3\left(\mathrm{card}(\mathcal{X}_n\cap A)\right)^{(d-\gamma)/d}\leq C_3 n^{(d-\gamma)/d}$$

and therefore for $n>n_o$

$$\begin{aligned}&\min_{A\in\mathcal{A}_\alpha^m} E[L_n(A)]-C_3 n^{(d-\gamma)/d}\epsilon\leq E[L_n(B_{\lfloor\alpha n\rfloor}^m)]\\ &\quad\leq\max_{A\in\mathcal{A}_\alpha^m} E[L_n(A)]+C_3 n^{(d-\gamma)/d}\epsilon, \quad (a.s).\end{aligned}$$

Combining the above and again applying (28) yields for $n > n_o$

$$\begin{aligned}&\left|E[L_n(B^m_{\lfloor \alpha n \rfloor})]/n^{(d-\gamma)/d} - \beta_{L,\gamma}\int_A f^{(d-\gamma)/d}(x)dx\right| \\ &\qquad \le (1+C_3)\epsilon + 2m^{-d}\beta_{L,\gamma}\sum_{i=1}^{m^d} v(Q_i \cap \partial \mathcal{A}^m_\alpha).\end{aligned}$$

Since $\epsilon$ is arbitrary we obtain the desired bound. □

We next extend Lemma 5 to a minimal graph constructed over any $k$ points $\mathcal{X}_{n,k}$, e.g. $\mathcal{X}^{G_m}_{n,k}$, drawn from $B^m_{\lfloor \alpha n \rfloor}$.

*Lemma 7:* Let $\mathcal{X}_{n,\lfloor \alpha n \rfloor}$ be any $\lfloor \alpha n \rfloor$ points selected from $B^m_{\lfloor \alpha n \rfloor}$. Then, for any quasi-additive functional $L_n(B^m_{\lfloor \alpha n \rfloor}) \overset{\text{def}}{=} L_n$ $B^m_{\lfloor \alpha n \rfloor})$

$$\begin{aligned}&\limsup_{n\to\infty}\left|L_n(B^m_{\lfloor \alpha n \rfloor}) - L(\mathcal{X}_{n,\lfloor \alpha n \rfloor})\right|/n^{(d-\gamma)/d} \\ &\qquad < C_3(p_{\mathcal{A}^m_\alpha} - \alpha)^{(d-\gamma)/d}, \quad (a.s.)\end{aligned}$$

and

$$\begin{aligned}&\limsup_{n\to\infty}\left|E[L_n(B^m_{\lfloor \alpha n \rfloor}) - E[L(\mathcal{X}_{n,\lfloor \alpha n \rfloor})]\right|/n^{(d-\gamma)/d} \\ &\qquad < C_3(p_{\mathcal{A}^m_\alpha} - \alpha)^{(d-\gamma)/d},\end{aligned}$$

where $p_{\mathcal{A}^m_\alpha} = P(A^m_\alpha)$ is the coverage probability of sets $A^m_\alpha$ in $\mathcal{A}^m_\alpha$.

*Proof of Lemma 7*

Firstly, note that from continuity of $L$ and the fact that $\alpha n - 1 \le \lfloor \alpha n \rfloor \le \alpha n$

$$\begin{aligned}&\left|L_n(B^m_{\lfloor \alpha n \rfloor}) - L(\mathcal{X}_{n,\lfloor \alpha n \rfloor})\right|/n^{(d-\gamma)/d} \\ &\le C_3\left[\frac{\text{card}\left(B^m_{\lfloor \alpha n \rfloor}\right) - \lfloor \alpha n \rfloor}{n}\right]^{(d-\gamma)/d} \\ &\le C_3\left[n^{-1}\text{card}\left(B^m_{\lfloor \alpha n \rfloor}\right) - \alpha\right]^{(d-\gamma)/d} + C_3 n^{-(d-\gamma)/d}.\end{aligned} \tag{31}$$

Next we establish an a.s. limit for the first additive term on the right side. Lemma 3 guarantees that there exists an $n_0$ such that $B^m_{\lfloor \alpha n \rfloor} \in \mathcal{A}^m_\alpha$ with probability arbitrarily close to one. Therefore, for $n > n_o$ and for any $\epsilon > 0$, by Sanov's theorem we have

$$\begin{aligned}&P\left(n^{-1}\text{card}(\mathcal{X}_n \cap B^m_{\lfloor \alpha n \rfloor}) - \alpha \ge \epsilon\right) \\ &\quad = P\left(n^{-1}\sum_{i=1}^n z_i \ge \epsilon + \alpha\right) \\ &\quad \le (n+1)^2 \exp\left\{-nK(\epsilon+\alpha, p_{\mathcal{A}^m_\alpha})\right\}\end{aligned}$$

where $z_j$'s are i.i.d. Bernoulli random variables with $P(z_j = 1) = p_{\mathcal{A}^m_\alpha} = \sum_{i=1}^q \varphi_{(i)}$, as defined in the proof of Lemma 3. Now $\rho \overset{\text{def}}{=} K(\epsilon + \alpha, p_{\mathcal{A}^m_\alpha}) > 0$ for any $\epsilon > p_{\mathcal{A}^m_\alpha} - \alpha$ and therefore, since $\sum_{n>n_o}(n+1)^2 \exp^{-n\rho} \le 2(2\rho^2 + \rho + 1)/\rho^3 < \infty$, by Borel-Cantelli we have

$$P\left(\limsup\left\{\mathcal{X}_n : n^{-1}\text{card}(\mathcal{X}_n \cap B^m_{\lfloor \alpha n \rfloor}) - \alpha \ge p_{\mathcal{A}^m_\alpha} - \alpha\right\}\right) = 0. \tag{32}$$

Since $C_3 n^{-(d-\gamma)/d}$ converges pointwise to zero as $n \to \infty$, the a.s. limit in the Lemma follows directly from (32) and (31).

Finally, as in the proof of 5, it can be shown that since $L(\mathcal{X}_n \cap B^m_{\lfloor \alpha n \rfloor}) - L(\mathcal{X}_{n,\lfloor \alpha n \rfloor})$ is bounded, $EL(\mathcal{X}_n \cap B^m_{\lfloor \alpha n \rfloor}) - EL(\mathcal{X}_{n,\lfloor \alpha n \rfloor}]$ satisfies the same asymptotic properties. □

We now have all the ingredients for the proof of Theorem 2.

*Proof of Theorem 2*

Combining Lemmas 5 and 7 and applying the triangle inequality we see that there exists an integer $n_o$ such that for all $n > n_o$

$$\begin{aligned}&\left| L(\mathcal{X}_{n,\lfloor \alpha n \rfloor})/n^{(d-\gamma)/d} - \beta_{L,\gamma} \int_{A^m_\alpha} \tilde{f}^{(d-\gamma)/d}(x)dx \right| \\ &\quad < 2m^{-d}\beta_{L,\gamma} \sum_{i=1}^{m^d} v(Q_i \cap \partial \mathcal{A}^m_\alpha) + C_3 (p_{\mathcal{A}^m_\alpha} - \alpha)^{(d-\gamma)/d}, \quad (a.s.)\end{aligned}$$

where $A^m_\alpha$ is any set in the class $\mathcal{A}^m_\alpha$.

It remains to show that $\delta = O(m^{\gamma-d})$. For this we recall as in the proof of Lemma 3 that the coverage probability of any set $A \in \mathcal{A}^m_\alpha$ is a sum of the rank ordered cell probabilities $P(Q_{(i)}) = \varphi_{(i)}$, $i = 1, \ldots, q$, where $\sum_{i=1}^{q-1} \varphi_{(i)} < \alpha$ and $\sum_{i=1}^{q} \varphi_{(i)} \geq \alpha$, and therefore

$$0 \leq P(A^m_\alpha) - \alpha \; = \; \sum_{i=1}^{q} \varphi_{(i)} - \alpha \leq \varphi_{(q)} = \int_{Q_{(q)}} f(x)dx = m^{-d} f(\xi)$$

where, by the mean value theorem, $\xi$ is a point in $Q_{(q)}$. This gives the order $m^{\gamma-d}$ bound on the second additive term in $\delta$ of Theorem 2

$$0 \leq C_3 (p_{\mathcal{A}^m_\alpha} - \alpha)^{(d-\gamma)/d} \leq C_3 m^{\gamma-d} \sup_{x \in Q_i} f^{(d-\gamma)/d}(x). \tag{33}$$

Since $\sup_{x\in Q_i} f^{(d-\gamma)/d}(x) \leq v([0,1]^d)$ and $\sum_{i=1}^{m^d} v(Q_i \cap \partial \mathcal{A}^m_\alpha) \leq v([0,1]^d)$ and $v([0,1]^d) < \infty$, Theorem 2 is established. □

□

## VI. Examples

We first assume that $f$ is a uniform density over the $d$-dimensional unit sphere $S(0,1)^d$. It is obvious that for $\alpha \in [0,1]$ a Borel subset $A_o$ which minimizes Rényi entropy is

$$A_o = \left\{ x : \|x\| \leq \left( \frac{\alpha}{|S(0,1)^d|} \right)^{\frac{1}{d}} \right\}$$

and the associated minimum entropy conditional density is

$$f(x|A_o) = \left\{ \begin{array}{ll} \frac{1}{\alpha}, & x \in A_o \\ 0, & o.w. \end{array} \right.$$

By Theorem 2 $L(\mathcal{X}^{G_m}_{n.\lfloor \alpha n \rfloor})/n^{(d-\gamma)/d}$ converges a.s. to a linear function of $\alpha$

$$\beta_{L,\gamma} \int f^{(d-\gamma)/d}(x|A_o)dx = \alpha \cdot \beta_{L,\gamma}.$$

Next assume that $f$ is a multivariate Gaussian density with mean $\mu$ and covariance $\sigma^2 I$ on $\mathbb{R}^d$. Note that, unlike the previous example, the support of $f$ is $\mathbb{R}^d$ which is not compact and cannot be mapped into $[0,1]^d$. However, in practice the range is finite and we can approximate by a truncated Gaussian density with compact support. The minimum entropy set for this case is

$$A_o = \left\{x : \|x\| \leq \sigma \sqrt{Q^{-1}_{\chi^2}(\alpha; d)}\right\}$$

where $Q^{-1}_{\chi^2}(\cdot; d)$ is the quantile function of a Chi-squared density with $d$ degrees of freedom. The associated conditional density is

$$f(x|A_o) = \left\{ \begin{array}{cc} \frac{1}{\alpha(2\pi\sigma)^{d/2}} e^{-\frac{\|x\|^2}{2\sigma^2}}, & x \in A_o \\ 0, & o.w. \end{array} \right. .$$

and $L(\mathcal{X}^{G_m}_n)/n^{(d-\gamma)/d}$ converges a.s. to the non-linear function of $\alpha$

$$\beta_{L,\gamma} \int f^{(d-\gamma)/d}(x|A_0)dx = Q_{\chi^2}(\nu Q^{-1}_{\chi^2}(\alpha; d)) \cdot (2\pi\sigma)^{\frac{\gamma}{2d}} \beta_{L,\gamma}$$

where $\nu = (d-\gamma)/d$.

These two examples suggest that the greedy $k$-point graph can be effectively used to discriminate between uniform and non-uniform densities based on plots of $L(\mathcal{X}^{G_m}_n)/n^{(d-\gamma)/d}$ as a function of $\alpha$.

## VII. Influence Functions

Influence functions have been used to study quantitative robustness of estimators to outliers and other contaminating densities for over thirty years [31]. These functions provide a quantitative measure of outlier sensitivity of an estimator. An unbounded influence functions implies that the effect of an outlier on the estimator can be very severe. Robust estimators, such as the trimmed mean estimator which rejects observations which exceed a given sample quantile, have bounded influence functions (see Figure 4).

Here we give the influence function for the normalized greedy minimal $k$-point graph weight $L(\mathcal{X}^{G_m}_{n,k})/(\lfloor \alpha n \rfloor)^{\frac{d-\gamma}{d}}$ described in Section IV. The form of this influence function motivates the use of the Renyi estimator (16) as a robust estimator of the entropy of a nominal density $f_1$ in the mixture model $f = (1-\epsilon) + f_1\epsilon$. It also establishes a kind of outlier robustness which is similar to that of rank order statistics for one dimensional observations. Finally, it gives an asymptotic approximation to the variance of $L(\mathcal{X}^{G_m}_{n,\lfloor \alpha n \rfloor})/(\lfloor \alpha n \rfloor)^{(d-\gamma)/d}$ which can be used to construct confidence intervals on finite sample accuracy.

Let $P_n$ be the empirical distribution function of the $n$ samples $\mathcal{X}_n = \{x_1, \ldots, x_n\}$

$$P_n(A) \overset{\text{def}}{=} \frac{1}{n} \int_A I_{x_i}(x)dx$$

for arbitrary Borel set $A$. For any statistic $T_n = T(P_n)$ converging a.s. to $T = T(P)$ the influence function (called an influence curve for one dimensional samples $x_i$) is defined as [34]

$$\text{IF}(x_o) = \lim_{s \to 0} \frac{T(\,(1-s)P + s\delta_{x_o}) - T(P)}{s}. \tag{34}$$

where $\delta_{x_o}$ is a concentrated distribution centered at $x_o \in \mathbb{R}^d$ and $s \in [0,1]$. For small $s$, $(1-s)P + s\delta_{x_o}$ is interpreted as a perturbed distribution resulting from exchanging $sn$ of the $n$ samples $x_i$ from distribution $P$ with $sn$ samples from the concentrated distribution $\delta_{x_o}$. Thus $\mathrm{IF}(x_o)$ can be used to probe the asymptotic sensitivity of the estimator $T_n$ to localized perturbations of $P$.

If the distribution of the estimator $T_n$ satisfies certain asymptotic conditions, not explored here, then the influence function can also be used to approximate asymptotic estimator variance [34] via the formula

$$n\mathrm{var}(T_n) \;\rightarrow\; \int \mathrm{IF}^2(x) f(x) dx$$

Now identifying $T_n = L^{G_m}_{n,\lfloor \alpha n \rfloor}/(\lfloor \alpha n \rfloor)^{(d-\gamma)/d}$ we have by Theorem 3 that $T_n$ converges a.s. to the integral (13) which we thus identify as $T(P)$.

In Appendix B (Lemma 8) we derive the influence function $\mathrm{IF}(x_o)$ for differentiable densities $f$ having no flat spots (cf Remark 2 of Section IV). The influence function specializes to the following form when one ignores behavior at the boundary of $A_o$

$$\mathrm{IF}(x_o)/\beta_{L,\gamma} = \begin{cases} -\frac{\nu}{\alpha^\nu} g_\nu\left(g^{-1}(\alpha)\right) + \alpha^{1-\nu}\, \frac{g_\nu^{'}\left(g_1^{-1}(\alpha)\right)}{g_1^{'}\left(g_1^{-1}(\alpha)\right)}, & x_o \notin A_o \\ -\frac{\nu}{\alpha^\nu} g_\nu\left(g^{-1}(\alpha)\right) + \frac{\alpha-1}{\alpha^\nu}\, \frac{g_\nu^{'}\left(g_1^{-1}(\alpha)\right)}{g_1^{'}\left(g_1^{-1}(\alpha)\right)} + f^{\nu-1}(x_o)\frac{\nu}{\alpha^\nu}, & x_o \in A_o \end{cases} \tag{35}$$

where $A_o = \{x : f(x) \geq \eta\}$ is the entropy minimizing set of probability $\alpha$ and $g_1$, $g_\eta$ are monotone functions defined in Appendix B.

The function $\mathrm{IF}(x_o)$ may take on positive or negative values for $x_o$ inside of $A_o$ while it takes on positive values outside of $A_o$ (observe that $f^{\nu-1} = f^{-|\nu-1|}$ increases without bound if the tails of $f$ decreases to zero). This can be explained as follows. By the theory developed in the Section IV we know that asymptotically the minimal $k$-point graph spans all points within $A_o$ and none of the points outside of $A_o$. Therefore exchanging a small number of vertices of the $k$-point graph within $A_o$ with a small number of points outside of $A_o$ necessarily increases the overall weight of the graph. On the other hand, the value of IF on the interior of $A_o$ corresponds to the effect on the weight due to perturbing the locations of a small number of vertices. Thus, depending on the direction of these perturbations the weight of the graph can either increase or decrease.

We illustrate these phenomena in Figure 5 where IF is plotted as a function of $x_o \in \mathbb{R}^2$ for the case of the bivariate Gaussian distribution considered in the previous subsection. Two cases are shown, the figure on the left is the influence function for $\alpha = 1$, i.e., for the minimal graph spanning all points (labeled MST), and the figure on the right is for $\alpha = 0.8$, i.e. for the minimal $k$-point graph (labeled $k$-MST) spanning only 80% of the $n$ points. Note that, as expected, the influence function is bounded for the $k$-point graph but unbounded for the graph spanning all $n$ points. This suggests that the greedy $k$-point minimal graph is a natural multi-dimensional extension of rank order statistical methods such as the trimmed mean. This complements the comments of Friedman and Rafsky [26] in which they proposed the MST as a natural generalization of the one-dimensional rank order Smirnov test statistic for testing equality of two multivariate distributions.

## VIII. Conclusion

We have given strong asymptotic convergence results for greedy approximations to minimal $k$-point graphs. These convergence results indicate that the weight function of minimal $k$-point graphs provide natural extensions of one dimensional rank order statistics to multiple dimensions. Our results also provide an interesting alternative to kernel or histogram methods of entropy estimation. An open problem is whether or not the exact $k$-point minimal graph satisfies the tight cover property stated in Definition 2 of Section III. Another open problem is the determination of the asymptotic distribution of the greedy $k$-point minimal graph weight and its rate of convergence. While percolation theory methods [2] and martingale convergence methods [37] can be used to establish CLT's for the MST their application to $k$-point minimal graphs, in particular the $k$-MST, appears more difficult. Resolution of this issue will be especially important for statistically significant utilisation of $k$-point minimal graphs for estimation, detection and discrimination applications.

**Acknowledgement**

The authors gratefully acknowledge the reviewers whose comments led to an improved presentation of our results.

## Appendix A

*Proof of Lemma 1*

Let $\mathcal{X}_{n,k}^{C_\alpha^m}$ be any $k$-points in $\mathcal{X}_n \cap C_\alpha^m$. Due to continuity of $L$ we have

$$\left|L(\mathcal{X}_n \cap C_\alpha^m) - L(\mathcal{X}_{n,k}^{C_\alpha^m})\right| \leq C_3 \left(\mathrm{card}(\mathcal{X}_n \cap C_\alpha^m) - k\right)^{(d-\gamma)/d}.$$

Hence, since $\mathcal{X}_{n,k}^*$ are the vertices of the minimal $k$-point graph

$$\begin{aligned} L(\mathcal{X}_n \cap C_\alpha^m) &\geq L(\mathcal{X}_{n,k}^{C_\alpha^m}) - C_3 \left(\mathrm{card}(\mathcal{X}_n \cap C_\alpha^m) - k\right)^{(d-\gamma)/d} \\ &\geq L(\mathcal{X}_{n,k}^*) - C_3 \left(\mathrm{card}(\mathcal{X}_n \cap C_\alpha^m) - k\right)^{(d-\gamma)/d}. \end{aligned} \tag{36}$$

Furthermore, again by continuity,

$$\left|L(\mathcal{X}_n \cap D_k^m) - L(\mathcal{X}_{n,k}^*)\right| \leq C_3 \left(\mathrm{card}(\mathcal{X}_n \cap D_k^m) - k\right)^{(d-\gamma)/d}, \tag{37}$$

so that

$$L(\mathcal{X}_n \cap D_k^m) \leq L(\mathcal{X}_{n,k}^*) + C_3 \left(\mathrm{card}(\mathcal{X}_n \cap D_k^m) - k\right)^{(d-\gamma)/d}. \tag{38}$$

Hence, combining (36) and (38)

$$\begin{aligned} & L(\mathcal{X}_n \cap D_k^m) - L(\mathcal{X}_n \cap C_\alpha^m) \\ \leq\ & C_3 \left[\left(\mathrm{card}(\mathcal{X}_n \cap D_k^m) - k\right)^{(d-\gamma)/d} + \left(\mathrm{card}(\mathcal{X}_n \cap C_\alpha^m) - k\right)^{(d-\gamma)/d}\right] \end{aligned} \tag{39}$$

On the other hand, by definition of $C_\alpha^m$ the left side of inequality (39) is greater than zero so that

$$\begin{aligned} & \left|L(\mathcal{X}_n \cap D_k^m) - L(\mathcal{X}_n \cap C_\alpha^m)\right| \\ \leq\ & C_3 \left[\left(\mathrm{card}(\mathcal{X}_n \cap D_k^m) - k\right)^{(d-\gamma)/d} + \left(\mathrm{card}(\mathcal{X}_n \cap C_\alpha^m) - k\right)^{(d-\gamma)/d}\right] \end{aligned}$$

which when combined with (37) gives

$$\begin{aligned}&\left|L(\mathcal{X}^*_{n,k}) - L(\mathcal{X}_n \cap C^m_\alpha)\right| \\ &\leq |L(\mathcal{X}_n \cap D^m_k) - L(\mathcal{X}_n \cap C^m_\alpha)| + \left|L(\mathcal{X}^*_{n,k}) - L(\mathcal{X}_n \cap D^m_k)\right| \\ &\leq C_3 \left[(\mathrm{card}(\mathcal{X}_n \cap C^m_\alpha) - k)^{(d-\gamma)/d} + 2\,(\mathrm{card}(\mathcal{X}_n \cap D^m_k) - k)^{(d-\gamma)/d}\right].\end{aligned}$$

Dividing both sides by $n^{(d-\gamma)/d}$ establishes the lemma. □

## Appendix B

For $\nu \in (0,1]$ define the function

$$g_\nu(\eta) = \int_{\{x: f(x) \geq \eta\}} f^\nu(x)dx. \tag{40}$$

Note that when $A_o$ is the minimum entropy subset of probability $\alpha$ defined in (18) then $g_\nu(\eta) = \int_{A_o} f^\nu(x)dx$ and $g_1(\eta) = P(A_o) = \alpha$. Under the assumption that the set $\{x : f(x) = c\}$ has measure zero, $g_\nu(\eta)$ is differentiable (a.e.) and monotone decreasing in $\eta \geq 0$. It therefore has an inverse function $g_\nu^{-1}(a)$ which is differentiable (a.e.) and monotone decreasing in $a \geq 0$.

Let $B_d(x_o, r)$ denote the open $d$-dimensional ball with center $x_o$ radius $r$. For given $\epsilon > 0$ the smoothness of a function $f$ can be quantified through its $\epsilon$-coefficient of variation defined as $\overline{r}_r(\epsilon) = \sup_r\{r : \sup_{x_o} v(B(x_o, r)) < \epsilon\}$ where $v(Q)$ is the total variation of $f$ over the set $Q$. A less stringent measure of smoothness is the $\epsilon$-modulus of continuity $r_f(\epsilon)$ which will be sufficient for the lemma below. $r_f(\epsilon)$ is defined as the maximum value (supremum) of $r$ such that $\sup_{x,x' \in B(x_o,r)} |f(x) - f(x')| < \epsilon$ uniformly for all $x_o$ in the support set of $f$. Note that $r_f(\epsilon) \geq \overline{r}_f(\epsilon)$.

*Lemma 8:* Let $f$ be a Lebesgue density over $\mathbb{R}^d$ and assume that for any $a > 0$ the set $\{x : f(x) = a\}$ has measure zero. As in Theorem let $\alpha, \eta$ satisfy the relation $\alpha = g_1(\eta)$ and let $\nu = (d - \gamma)/d \in (0, 1)$. For fixed $\epsilon > 0$ let $\delta_{x_o}$ denote the uniform distribution over the spheroid $S = B_d(x_o, \Delta)$ having center $x_o$ and radius $\Delta$, where $\Delta$ is smaller than the $\epsilon$-modulus of continuity of $f$ and $f^{\nu-1}$. The influence function (34) of $T_n = L(\mathcal{X}^{G_m}_{n,\lfloor\alpha n\rfloor})/(\lfloor \alpha n \rfloor)^{(d-\gamma)/d}$ has the form

$$\begin{aligned}\mathrm{IC}(x_o)/\beta_{L,\gamma} = &-\frac{\nu}{\alpha^\nu} g_\nu\left(g^{-1}(\alpha)\right) + \frac{(\alpha - \zeta_0 + f(x_o)\phi_0)}{\alpha^\nu}\, \frac{g_\nu'\left(g_1^{-1}(\alpha)\right)}{g_1'\left(g_1^{-1}(\alpha)\right)} \\ &+ f^{\nu-1}(x_o)\frac{\zeta_0 \nu}{\alpha^\nu} + f^\nu(x_o)\frac{\phi_0}{\alpha^\nu} + \phi_0 O(\epsilon)\end{aligned} \tag{41}$$

where, as defined in (46), $\zeta_0 = \zeta_s|_{s=0}$ is the proportion of the $d$-dimensional spheroid $B_d(x_o, \Delta)$ lying inside the entropy minimizing set $A_o$, and $\phi_0 = \xi_0' \mathrm{vol}\{S\}$ where $\xi_s$ is as defined in (48). When $f$ is continuously differentiable and $\nabla f(x_o) \neq 0$, $\phi_0$ has the explict form

$$\phi_0 = \begin{cases} \frac{V_{d-1}}{V_d}\left(1 - \left(\frac{\rho(x_o, \partial A_o)}{\Delta}\right)^2\right)^{d/2} \frac{1}{\|\nabla f(x_o)\|}, & \rho(x_o, \partial A_o) \leq \Delta \\ 0, & o.w. \end{cases}, \tag{42}$$

where $V_d$ is the volume of the $d$-dimensional unit spheroid, $\rho(x_o, \partial A_o)$ is the Euclidean distance between $x_o$ and the boundary $\partial A_o$ of $A_o$, and $\|\nabla f(x)\|$ is the norm of the gradient of $f$.

*Proof of Lemma 8*

First we define an entropy minimizing set $A_s$ of probability $\alpha$ analogously to the definition (18) of $A_o$

$$A_s = \{x : f_s \geq \eta_s\} \tag{43}$$

where $\eta_s$ satisfies

$$\int_{A_s} f_s dx = \alpha \tag{44}$$

Using $f_s = (1-s)f + s\delta_{x_o}$ relation (44) implies

$$\alpha \;=\; (1-s)\int_{A_s} f(x)dx + s\int_{A_s} \delta_{x_o}(x)dx \tag{45}$$

Let $S = B_d(x_o, \Delta)$ denote the support set of $\delta_{x_0}$ and let $\zeta_s, \xi_s, \tau_s$ be the proportional volumes in $S$ defined by:

$$\begin{aligned}
\zeta_s &= \frac{\mathrm{vol}\{x \in S : \eta_s \leq (1-s)f(x)\}}{\mathrm{vol}\{S\}} && (46)\\
\xi_s &= \frac{\mathrm{vol}\{x \in S : 0 < \eta_s - (1-s)f(x) \leq s\delta_{x_o}(x)\}}{\mathrm{vol}\{S\}} && (47)\\
\tau_s &= \frac{\mathrm{vol}\{(1-s)f(x) + s\delta_{x_o}(x) < \eta_s\}}{\mathrm{vol}\{S\}} = 1 - \zeta_s - \xi_s.
\end{aligned}$$

Observe that $\zeta_s$, $\xi_s$ and $\tau_s$ are functions of $x_o$, $\lim_{s\to 0}\zeta_s = \mathrm{vol}\{A_o \cap S\}/\mathrm{vol}\{S\}$, $\lim_{s\to 0}\xi_s = 0$, and $\zeta_s + \xi_s + \tau_s = 1$ (see figure VIII, for a graphical representation in the case $d = 1$). Due to the assumption that for any $c > 0$ the set $\{x : f(x) = c\}$ has measure zero it can be shown that (for $\mathrm{vol}\{S\} > 0$) the derivatives $\zeta_s^{'} = d\zeta/ds$ and $\xi_s^{'} = d\xi_s/ds$ are finite for $s \in [0, \delta)$, $\delta > 0$. Furthermore the two integrals in (44) have the representations

$$\int_{A_s} \delta_{x_o}(x)dx = (\zeta_s + \xi_s) \tag{48}$$

and

$$\int_{A_s} f(x)dx = \int_{(1-s)f \geq \eta_s} f(x)dx + \xi_s \mathrm{vol}\{S\}(f(x_o) + O(\epsilon)) \tag{49}$$

Equation (49) uses the fact that the radius of the support $S$ of $\delta_{x_o}$ is less than the $\epsilon$-modulus of continuity of $f$ which implies that over $x \in S$: $f(x) = f(x_o) + O(\epsilon)$. Therefore, using (48) and (49) in the relation (44), along with the definition (40) of $g_\nu$

$$\alpha = (1-s)g_1\left(\frac{\eta_s}{1-s}\right) + (1-s)\xi_s \mathrm{vol}\{S\}f(x_o) + (\zeta_s + \xi_s)s + \xi_s \mathrm{vol}\{S\}O(\epsilon) \tag{50}$$

Eq. (50) specifies, to order $O(\epsilon)$, the threshold $\eta_s$ which guarantees (44)

$$\eta_s = (1-s)g_1^{-1}\left(\frac{\alpha - (\xi_s + \zeta_s)s}{1-s} - \xi_s \mathrm{vol}\{S\}f(x_o)\right) + \xi_s \mathrm{vol}\{S\}O(\epsilon) \tag{51}$$

where to obtain (51) we have used continuity of the inverse function $g_1^{-1}$.

Next we express $\int_{A_s} f_s^\nu(x)dx$ to order $o(s)$

$$\int_{A_s} f_s^\nu(x)dx = (1-s\nu)\int_{A_s} f^\nu(x)dx + s\nu\int_{A_s} f^{\nu-1}(x)\delta_{x_o}(x)dx + o(s), \tag{52}$$

using the assumption on the $\epsilon$-modulus of continuity of $f^{\nu-1}$, and combining equations (51) and (52), we obtain

$$\begin{aligned}\int_{A_s} f_s^{\nu}(x)dx &= (1-s\nu)\int_{(1-s)f\geq\eta_s} f^{\nu}dx + (1-s\nu)\xi_s\text{vol}\{S\}f^{\nu}(x_o) + s\nu(\xi_s+\zeta_s)f^{\nu-1}(x_o) + o(s) && (53)\\ &= (1-s\nu)g_{\nu}\left(g_1^{-1}(\frac{\alpha-(\xi_s+\zeta_s)s}{1-s} - \xi_s\text{vol}\{S\}f(x_o))\right) + (1-s\nu)\xi_s\text{vol}\{S\}f^{\nu}(x_o) + && (54)\\ &\qquad s\nu f^{\nu-1}(x_o)(\xi_s+\zeta_s) + o(s) + \xi_s\text{vol}\{S\}O(\epsilon)\end{aligned}$$

To compute $\text{IC}(x_o)$ it remains to evaluate the limit

$$\begin{aligned}\text{IC}(x_o)/\beta_{L,\gamma} &= \lim_{s\downarrow 0}\frac{1}{s}\left[\int f_s^{\nu}(x|A_s)dx - \int f^{\nu}(x|A_o)dx\right] \\ &= \frac{1}{\alpha^{\nu}}\cdot\lim_{s\downarrow 0}\frac{1}{s}\left[\int_{A_s} f_s^{\nu}(x)dx - \int_{A_o} f^{\nu}(x)dx\right]\end{aligned} \tag{55}$$

Applying the chain rule to the identity $g_1(g_1^{-1}(q(s))) = q(s)$, for any differentiable function $q(s)$ we have the relation

$$\frac{d}{ds}g_1^{-1}(q(s)) = q^{'}(s)/g_1^{'}(g_1^{-1}(q(s))). \tag{56}$$

Identifying $q(s) = (\alpha-(\xi_s+\zeta_s)s)/(1-s) + \text{vol}\{S\}\xi_s f(x_o)$ and observing that $\xi_s = s\xi_0^{'} + o(s)$, $s(\xi_s^{'}+\zeta_s^{'}) = s(\xi_0^{'}+\zeta_0^{'}) + o(s)$, it is seen that $q^{'}(0) = \alpha - \zeta_0 + \xi_0^{'}\text{vol}\{S\}$. Therefore, after some algebra it can be verified that the limit (55) takes the form

$$\begin{aligned}\text{IC}(x_o)/\beta_{L,\gamma} = -\frac{\nu}{\alpha^{\nu}}g_{\nu}\left(g^{-1}(\alpha)\right) + \frac{(\alpha-\zeta_0+f(x_o)\xi_0^{'}\text{vol}\{S\})}{\alpha^{\nu}}\frac{g_{\nu}^{'}\left(g_1^{-1}(\alpha)\right)}{g_1^{'}\left(g_1^{-1}(\alpha)\right)} && (57)\\ +f^{\nu-1}(x_o)\frac{\zeta_0\nu}{\alpha^{\nu}} + f^{\nu}(x_o)\frac{\xi_0^{'}\text{vol}\{S\}}{\alpha^{\nu}} + \xi_0^{'}\text{vol}\{S\}O(\epsilon)\end{aligned}$$

It remains to establish that $\xi_0^{'}\text{vol}\{S\}$ is given by $\phi_0$ specified by expression (42). First, recall that $\xi_s$ is the relative volume of the region $\{x \in S : 0 < \eta_s - (1-s)f(x) \leq s\delta_{x_o}(x)\}$ (recall Figure VIII). This volume is zero when $S \cap A_s$ is empty, i.e. $\delta_x$ is entirely outside or entirely inside the region $A_s$. Therefore, in what follows we assume that $S\cap A_s$ is non-empty. Second, as $\Delta$ is less than the $\epsilon$-modulus of continuity and $f$ is differentiable, in $S$ the functions $(1-s)f(x)$ and $f_s(x) = (1-s)f(x) + s\delta_{x_o}(x)$ can be approximated to order $\epsilon$ by two parallel tangent hyperplanes since for any point $x_o^{'} \in S$: $\sup_{x\in S}|f(x) - f(x_o^{'}) - \nabla f(x_o^{'})(x - x_o^{'})| < 2\epsilon$. Let these hyperplanes be specified by the normal vector $\nabla f(x_o^{'})$ for a point $x_o^{'} \in S$ for which $\nabla f(x_o^{'}) \neq 0$. Existence of such a point $x_o^{'}$ is guaranteed by the hypothesis that for any constant $c > 0$ the set $\{x : f(x) = c\}$ has measure zero. The region $\{x \in S : 0 < \eta_s - (1-s)f(x) \leq s\delta_{x_o}(x)\}$ is therefore the intersection of a hyperslab of width $\Delta_o$ and the $d$-dimensional sphere $S = B_p(x_o, \Delta)$. $\Delta_o$ is specified by the intersection of the region sandwiched between the two tangent hyperplanes and the horizontal plane at level $\eta_s$ (see Figures VIII.a and VIII.b). By similarity of the two right triangles, having common edge along the ray $(\nabla f(x_o^{'}), 1)$, shown in Figure VIII.b, it is evident that the inner products $< (0, s/\text{vol}\{S\}), (\nabla f(x_o^{'}), 1) >$ and $< (\Delta_o\nabla f(x_o^{'})/\|\nabla f(x_o^{'})\|, 0), (\nabla f(x_o^{'}), 1) >$ are equal. Hence we have the relation:

$$\Delta_o = \frac{s}{\text{vol}\{S\}\|\nabla f(x_o^{'})\|}$$

As we will be taking the limit as $s \to 0$, we may assume that $s$ is sufficiently small to ensure that $\Delta_o \ll \Delta$. In this case the volume of the intersection of the hyperslab and $S = B_d(x_o, \Delta)$ is to order $o(s)$ equal to $\text{vol}\{B_{d-1}(x_1, \sqrt{\Delta^2 - \rho_s^2})\} \cdot \Delta_o$ where $\rho_s = \rho_s(x_o, \nabla A_s) = \rho(x_o, x_1)$ is the perpendicular distance between $x_o$ and the nearest face of the hyperslab (the point $x_1$ on the hyperslab is immaterial to the volume calculation), see Figure VIII. Therefore, defining $V_d$ as the volume of the $d$-dimensional unit sphere and noting that $\text{vol}\{S\} = V_d \Delta^d$:

$$\begin{aligned} \xi_s \text{vol}\{S\} &= V_{d-1} \left(\Delta^2 - \rho_s^2\right)^{d/2} \Delta_o & (58) \\ &= s \frac{V_{d-1}}{V_d} \left(1 - \left(\frac{\rho_s}{\Delta}\right)^2\right)^{d/2} \frac{1}{\|\nabla f(x_o')\|}. & (59) \end{aligned}$$

It can be shown that the derivative $d\rho_s/ds$ is uniformly bounded in the neighborhood of $s = 0$. Therefore, we readily obtain that the limiting value of the derivative $\lim_{s \to 0} d\xi_s/ds \text{vol}\{S\} = \xi_0' \text{vol}\{S\}$ is equal to the expression (42). □

## References

[1] I. Ahmad and P.-E. Lin, "A nonparametric estimation of the entropy for absolutely continuous distributions," *IEEE Trans. on Inform. Theory*, vol. IT-22, pp. 664–668, 1976.

[2] K. S. Alexander, "The RSW theorem for continuum percolation and the CLT for Euclidean minimal spanning trees," *Ann. Applied Probab.*, vol. 6, pp. 466–494, 1996.

[3] S. Arora, "Polynomial time approximation schemes for Euclidean TSP and other geometric problems," in *Proceedings of IEEE Symposium on Foundations of Computer Science*, pp. 2–11, Burlington, VT, 1996.

[4] S. Arora, "Nearly linear time approximation schemes for Euclidean TSP and other geometric problems," in *Proceedings of IEEE Symposium on Foundations of Computer Science*, 1997.

[5] F. Avram and D. Bertsimas, "The minimum spanning tree constant in geometrical probability and under the independent model: a unified approach," *Ann. Applied Probab.*, vol. 2, pp. 113–130, 1992.

[6] B. Awerbuch, Y. Azar, A. Blum, and S. Vempala, "Improved approximation guarantee for minimum weight k-trees and prize collecting traveling salesmen," in *Proc. 27th Annual ACM Symposium on Theory of Computing*, pp. 277–283, Las Vegas, NV, 1995.

[7] D. Banks, "The minimal spanning tree for nonparametric regression and structure discovery," in *Book of Abstracts of the 1996 Meeting of the Classification Society of North America*, p. 54, 1996.

[8] D. Banks, M. Lavine, and H. J. Newton, "The minimal spanning tree for nonparametric regression and structure discovery," in *Computing Science and Statistics. Proceedings of the 24th Symposium on the Interface*, H. J. Newton, editor, pp. 370–374, 1992.

[9] R. Baraniuk, P. Flandrin, A. J. E. M. Jensen, O. Michel : "Measuring Time Frequency information Content Using the Renyi Entropies.", submitted to *IEEE Trans. on Information Theory*, 1998.

[10] J. Beardwood, J. H. Halton, and J. M. Hammersley, "The shortest path through many points," *Proc. Cambridge Philosophical Society*, vol. 55, pp. 299–327, 1959.

[11] J. Beirlant, E. J. Dudewica, L. Györfi, and E. van der Meulen, "Nonparametric entropy estimation: an overview," *Intern. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, 1997.

[12] A. Blum, P. Chalasani, D. Coppersmith, B. Pulleybank, P. Raghavan, and M. Sudan, "The minumum latency problem," in *Proc. 26th Annual ACM Symposium on Theory of Computing*, pp. 163–172, Montreal, QUE, 1994.

[13] A. Blum, P. Chalasani, and S. Vempala, "A constant-factor approximation for the $k$-MST problem in the plane," in *Proc. of ACM-SIAM Symposium on Discrete Algorithms*, pp. 163–171, 1995.

[14] J. C. Chambers and T. J. Hastie, *Statistical models in S*, Wadsworth, Pacific Grove, CA, 1992.

[15] S. Y. Cheung and A. Kumar, "Efficient quorumcast routing algorithms," in *Proc. IEEE INFOCOM Conf. on Computer Communication*, volume 2, pp. 840–847, 1994.

[16] C. Chiang, M. Sarrafzadeh, and C. K. Wong, "Powerful global router: Based on Steiner min-max trees," in *IEEE International Conference on Computer-Aided Design*, pp. 2–5, Santa Clara, CA, 1989.

[17] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, New York, 1987.

[18] I. Csiszár and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, Orlando FL, 1981.

[19] A. Datta, H.-P. Lenhof, C. Schwarz, and M. Smid, "Static and dynamic algorithms for k-point clustering problems," in *Proc. 3rd Workshop on Algorthms and Data Structures*, volume 709 of *Lecture Notes in Computer Science*, pp. 265–276, Springer-Verlag, Berlin, 1994.

[20] C. W. Duln and A. Volgenant, "Some generalizations of the Steiner problem in graphs," *Networks*, vol. 17, pp. 353–364, 1987.

[21] C. Dussert, G. Rasigni, J. Palmari, and A. Llebaria, "Minimal spanning tree: a new approach for studying order and disorder," *Phys. Rev. B*, vol. 34, no. 5, pp. 3528–3531, 1986.

[22] J. Eckman and D. Ruelle, "Ergodic theory of chaos and strange attractors," *Rev. Mod. Phys*, vol. 57, pp. 617–656, 1985.

[23] D. Eppstein, "Faster geometric $k$-point MST approximation," Technical Report 95-13, Dept. of Information and Computer Science, University of California, Irvine, March 1995.

[24] J. Farmer, "Dimension, fractal measures and chaotic dynamics," in *Evolution of order and chaos*, pp. 228–246, 1982.

[25] P.Flandrin, R.G.Baraniuk, O. Michel : "Time-frequency complexity and informa tion." in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Proc essing - ICASSP'94*, Adelaide, Australia, Vol3, pp.329-332, 1994.

[26] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," *Annals of Statistics*, vol. 7, no. 4, pp. 697–717, 1979.

[27] N. Garg, "A 3-approximation for the minimum tree spanning $k$ vertices," in *Proc. of 37th Annual Symposium on Foundations of Computer Science*, pp. 302–309, Burlington, VT, 1996.

[28] N. Garg and D. S. Hochbaum, "An o(log$k$) approximation for the $k$ minimum spanning tree problem in the plane," in *Proc. of ACM-SIAM Symposium on Discrete Algorithms*, pp. 432–438, 1994.

[29] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. on Inform. Theory*, vol. IT-28, pp. 373–380, 1979.

[30] P. Hall and S. C. Morton, "On the estimation of entropy," *Ann. Inst. Statist. Math.*, vol. 45, pp. 69–88, 1993.

[31] F. R. Hampel, *Contributions to the theory of robust estimation*, PhD thesis, Univ. of California - Berkeley, 1968.

[32] J. Hartigan, *Clustering algorithms*, John Wiley and Sons, 1975.

[33] R. Hoffman and A. K. Jain, "A test of randomness based on the minimal spanning tree," *Pattern Recognition Letters*, vol. 1, pp. 175–180, 1983.

[34] P. J. Huber, *Robust Statistics*, Wiley, New York, 1981.

[35] A. Jain and J. Mamer, "Approximations for the random minimal spanning tree with applications to network provisioning," *Oper. Res.*, vol. 36, no. 4, pp. 575–584, 1988.

[36] H. Joe, "On the estimation of entropy and other functionals of a multivariate density," *Ann. Inst. Statist. Math.*, vol. 41, pp. 683–697, 1989.

[37] H. Kesten and S. Lee, "The central limit theorem for weighted minimal spanning trees on random points," *Ann. Applied Probab.*, vol. 6, pp. 495–527, 1996.

[38] E. L. Lawler, J. K. Lenstra, A. H. G. R. Kan, and D. B. Shmoys, *The traveling salesman problem*, Wiley, New York, 1985.

[39] O. Michel, R. Baraniuk, P. Flandrin : "Time-Frequency based distance and divergence measures", *Proc. IEEE International Time-Frequency and Time-Scale Analysis Symposiom*, Philadelphia, PA, pp.64-67,1994.

[40] J. Mitchell, "Guillotine subdivisions approximate polygonal subdivisions: a simple new method for the geometric $k$-MST problem," in *Proc. of ACM-SIAM Symposium on Discrete Algorithms*, pp. 402–408, 1996.

[41] D. N. Neuhoff, "On the asymptotic distribution of the errors in vector quantization," *IEEE Trans. on Inform. Theory*, vol. IT-42, pp. 461–468, March 1996.

[42] R. C. Prim, "Shortest connection networks and some generalizations," *Bell Syst. Tech. Journ.*, vol. 36, pp. 1389–1401, 1957.

[43] R. Ravi, M. Marathe, D. Rosenkrantz, and S. Ravi, "Spanning trees short or small," in *Proc. 5th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 546–555, Arlington, VA, 1994.

[44] R. Ravi, M. Marathe, D. Rosenkrantz, and S. Ravi, "Spanning trees – short or small," *SIAM Journal on Discrete Math*, vol. 9, pp. 178–200, 1996.

[45] C. Redmond and J. E. Yukich, "Limit theorems and rates of convergence for Euclidean functionals," *Ann. Applied Probab.*, vol. 4, no. 4, pp. 1057–1073, 1994.

[46] C. Redmond and J. E. Yukich, "Asymptotics for Euclidean functionals with power weighted edges," *Stochastic Processes and their Applications*, vol. 6, pp. 289–304, 1996.

[47] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, volume 1, pp. 547–561, 1961.

[48] W. T. Rhee, "A matching problem and subadditive Euclidean functionals," *Ann. Applied Probab.*, vol. 3, pp. 794–801, 1993.

[49] W. T. Rhee, "On the stochastic Euclidean traveling salesman problem for distributions with unbounded support," *Math. Oper. Res.*, vol. 18, pp. 292–299, 1993.

[50] F. Riesz and B. Sz.-Nagy, *Functional analysis*, Ungar, New York, 1955.

[51] T.-H. Sang and W. J. Williams, "Rényi information and signal-dependent optimal kernel design," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, volume 2, pp. 997–1000, 1995.

[52] A. Segev, "The node weighted Steiner problem," *Networks*, vol. 17, pp. 1–17, 1987.

[53] J. M. Steele, "Growth rates of euclidean minimal spanning trees with power weighted edges," *Ann. Probab.*, vol. 16, pp. 1767–1787, 1988.

[54] J. M. Steele, *Probability theory and combinatorial optimization*, volume 69 of *CBMF-NSF regional conferences in applied mathematics*, Society for Industrial and Applied Mathematics (SIAM), 1997.

[55] G. Toussaint, "The relative neighborhood graph of a finite planar set," *Pattern Recognition*, vol. 12, pp. 261–268, 1980.

[56] O. Vasicek, "A test for normality based on sample entropy," *J. Royal Statistical Society, Ser. B*, vol. 38, pp. 54–59, 1976.

[57] P. Viola and W. Wells, "Alignment by maximization of mutual information," in *Proc. of 5th Int. Conf. on Computer Vision, MIT*, volume 1, pp. 16–23, 1995.

[58] W. J. Williams, M. L. Brown, and A. O. Hero, "Uncertainty, information, and time-frequency distributions," in *Meeting of Intl. Soc. for Optical Engin. (SPIE)*, volume 1566, pp. 144–156, 1991.

[59] J. E. Yukich, "Worst case asymptotics for some classical optimization problems," *Combinatorica*, vol. 16, no. 4, pp. 575–586, 1996.

[60] J. E. Yukich, *Probability theory of classical Euclidean optimization*, volume 1675 of *Lecture Notes in Mathematics*, Springer-Verlag, Berlin, 1998.

[61] C. Zahn, “Graph-theoretical methods for detecting and describing Gestalt clusters,” *IEEE Trans. on Computers*, vol. C-20, pp. 68–86, 1971.

[62] A. A. Zelikovsky and D. D. Lozevanu, “Minimal and bounded trees,” in *Proc. of Tezele Congres XVIII Acad. Romano-Americaine*, pp. 25–26, Kishinev, 1993.

## Biographies

ALFRED O. HERO, III, (S '79, M '84, SM '96, F '97) was born in Boston, MA, USA in 1955. He received the B.S. (summa cum laude) from Boston University (1980) and the Ph.D. from Princeton University (1984), both in electrical engineering. He held the G.V.N. Lothrop Fellowship in Engineering at Princeton University. Since 1984 Alfred Hero has been with the Dept. of Electrical Engineering and Computer Science, University of Michigan - Ann Arbor, where he is currently Professor and Director of the Communications and Signal Processing Laboratory. He has held positions of Visiting Scientist at MIT Lincoln Laboratory, Lexington, MA (1987-89); Visiting Professor at l'Ecole Nationale de Techniques Avanceés (ENSTA), Paris, France (1991); William Clay Ford Fellow at the Ford Motor Company, Dearborn, MI (1993), Visiting Researcher at Ecole Normale Supérieure - Lyon (1999), and Ecole Nationale Supérieure de Télćommunications - Paris (1999). He has served as consultant for US government agencies and private industry. His present research interests are in the areas of detection and estimation theory, statistical signal and image processing, statistical pattern recognition, signal processing for communications, spatio-temporal processing, and biomedical signal and image analysis.

Alfred Hero is a member of Tau Beta Pi, the American Statistical Association, the New York Academy of Science, and Commission C of the International Union of Radio Science (URSI). In 1995 he received a Research Excellence Award from the College of Engineering at the University of Michigan. In 1999 he received a Best Paper Award from the IEEE Signal Processing Society. He was associate editor for the IEEE Transactions on Information Theory (1994-97); Chair of the IEEE SPS Statistical Signal and Array Processing Technical Committee (1996-98); and Treasurer of the IEEE SPS Conference Board (1997-2000). He was co-chair for the 1999 IEEE Information Theory Workshop and the 1999 IEEE Workshop on Higher Order Statistics. He served as publicity chair for the 1986 IEEE International Symposium on Information Theory and was general chair of the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing. He received the 1999 Meritorious Service Award from the IEEE Signal Processing Society.

OLIVIER J.J.MICHEL (S'84, M'85) was born in Mont Saint Martin, France, in 1963. He completed his studies at Ecole Normale Supérieure de Cachan, in the department of Applied Physics, where he received the "Agrégation de Physique" in 86. He received a Ph.D degree from University Paris-XI Orsay in 91, in signal processing. In 91, he joined the physics department at Ecole Normale Supérieure de Lyon, France, as an assistant professor. His research interest include non stationnary spectral analysis, array processing, non linear time series problems, information theory and dynamical systems studies, in close relationship with physical experiments in the field of chaos and hydrodynamical turbulence.

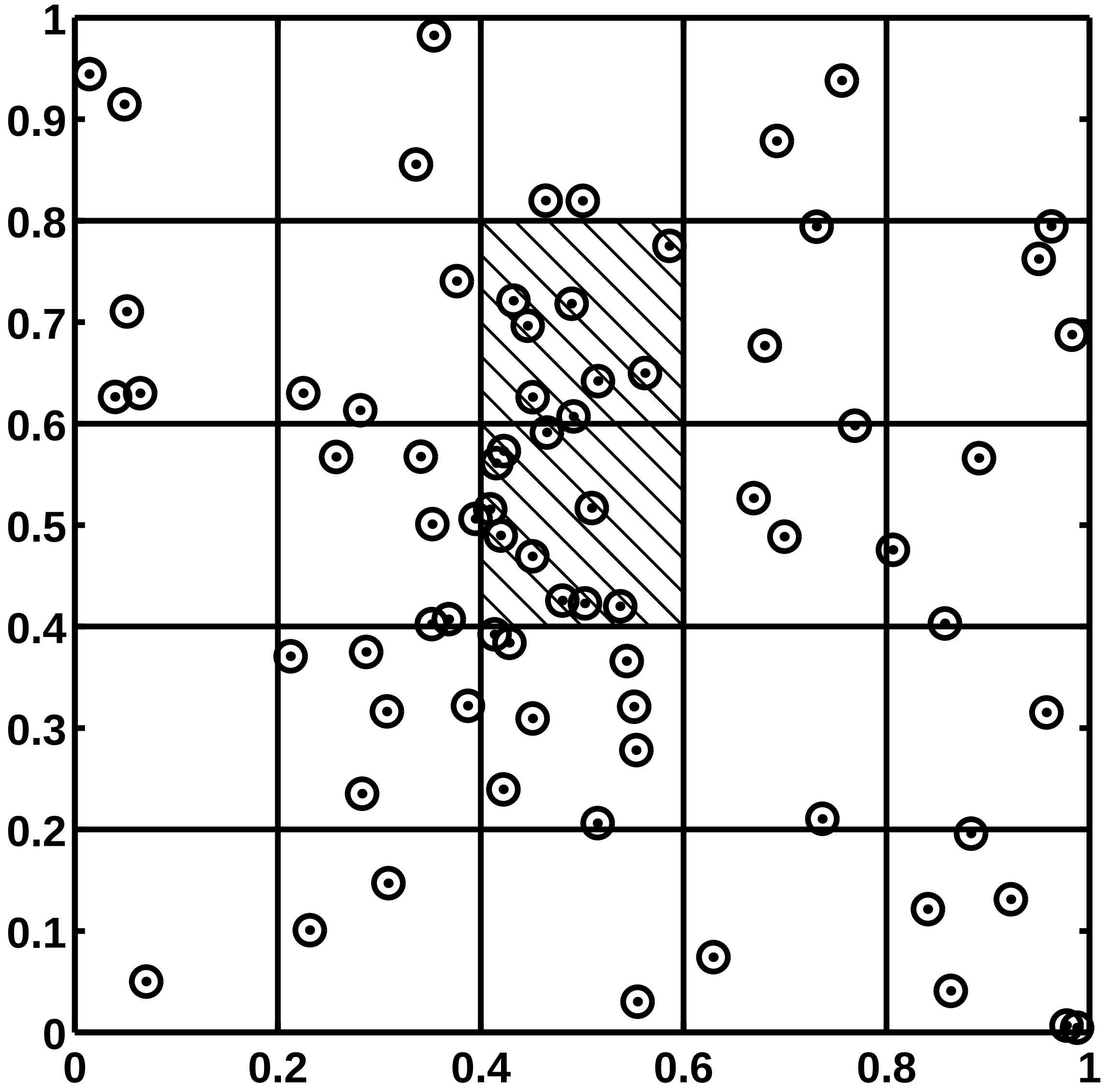

Fig. 1. *A sample of* 75 *points from the mixture density* $f(x) = 0.25 f_1(x) + 0.75 f_o(x)$ *where* $f_o$ *is a uniform density over* $[0, 1]^2$ *and* $f_1$ *is a bivariate Gaussian density with mean* $(1/2, 1/2)$ *and diagonal covariance* $\mathrm{diag}(0.01)$. *A smallest subset* $B_k^m$ *is the union of the two cross hatched cells shown for the case of* $m = 5$ *and* $k = 17$.

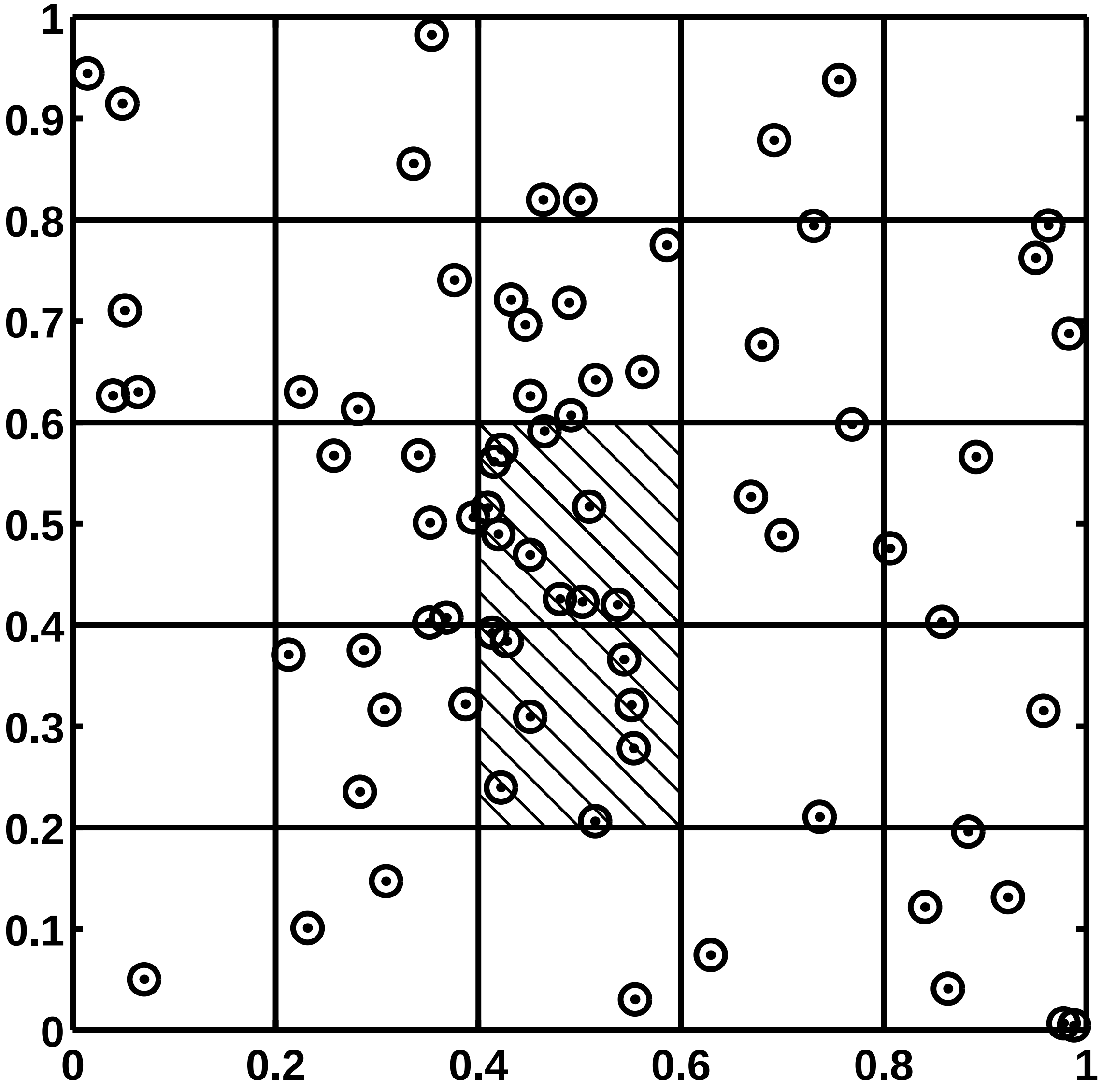

Fig. 2. *Another smallest subset* $B_k^m$ *containing at least* $k = 17$ *points for the mixture sample shown in Fig 1.*

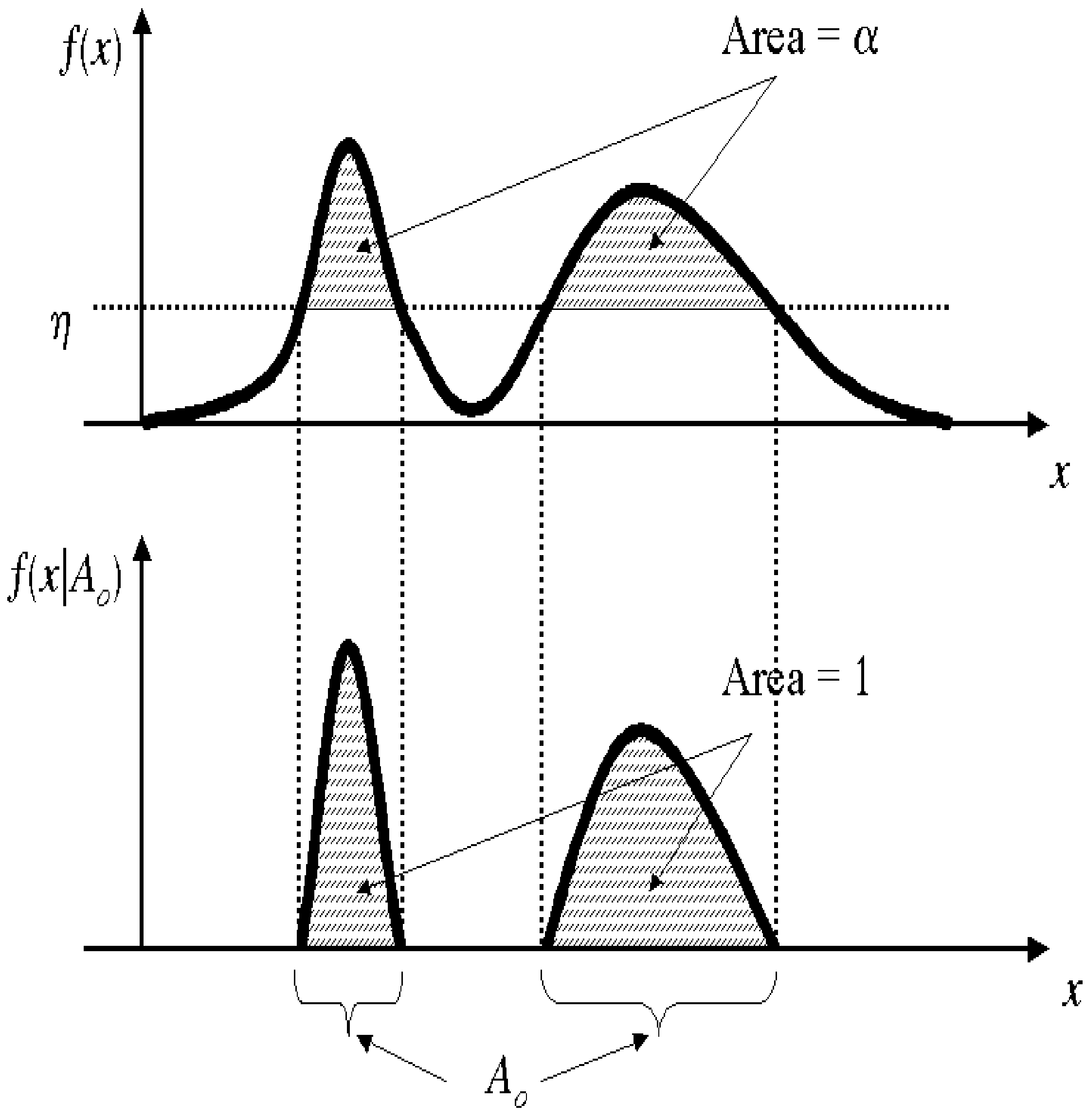

Fig. 3. *Water filling construction of $f(x|A_o)$. Region of support of $f(x|A_o)$ is $A_o = \{x : f(x) \geq \eta\}$ where $A_o, \eta$ are selected such that $\int_{A_o} f(x)dx = \alpha$.*

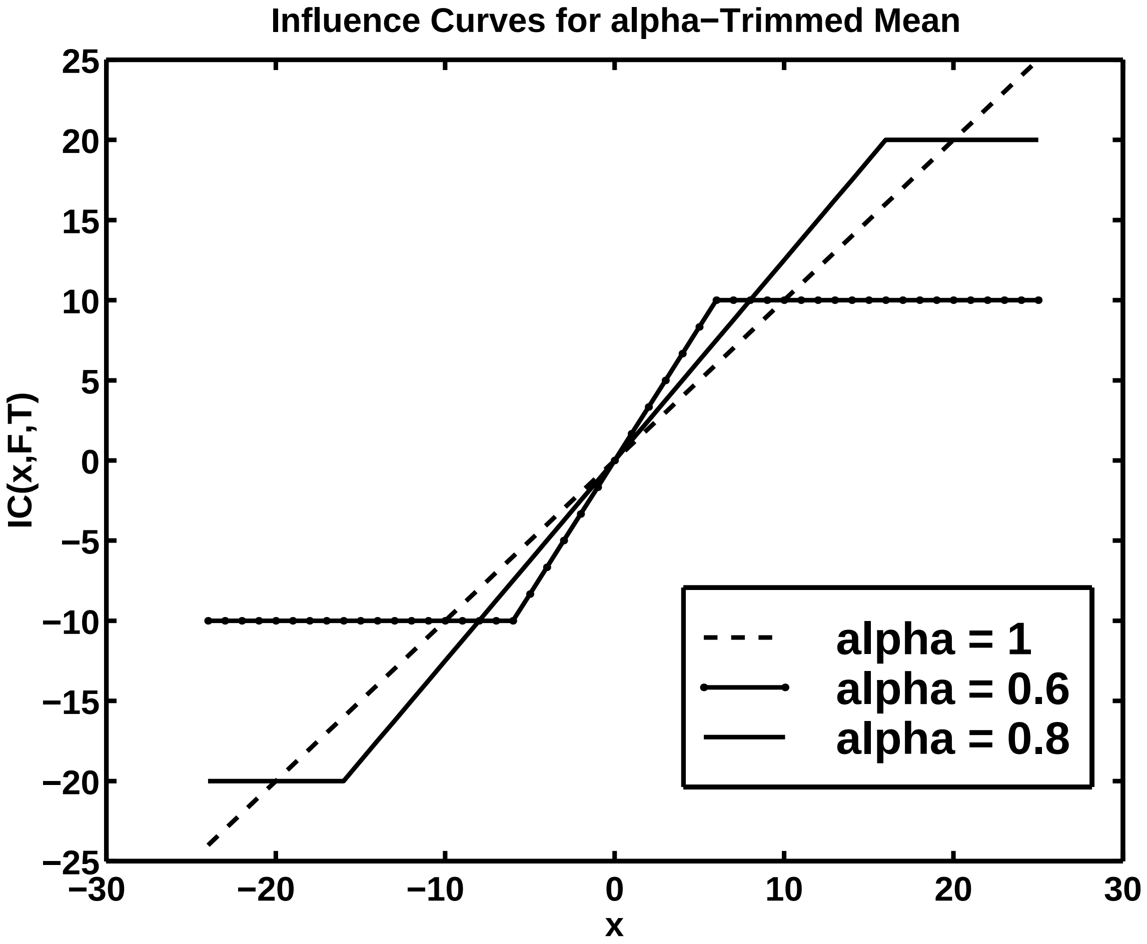

Fig. 4. *Trimmed mean influence curves for one dimensional observations and various trimming proportions $1-\alpha$. The trimmed mean estimator is a rank order statistic which robustifies the sample mean estimate by rejecting all samples whose values exceed either of the sample quantiles $1-\alpha/2$ and $\alpha/2$.*

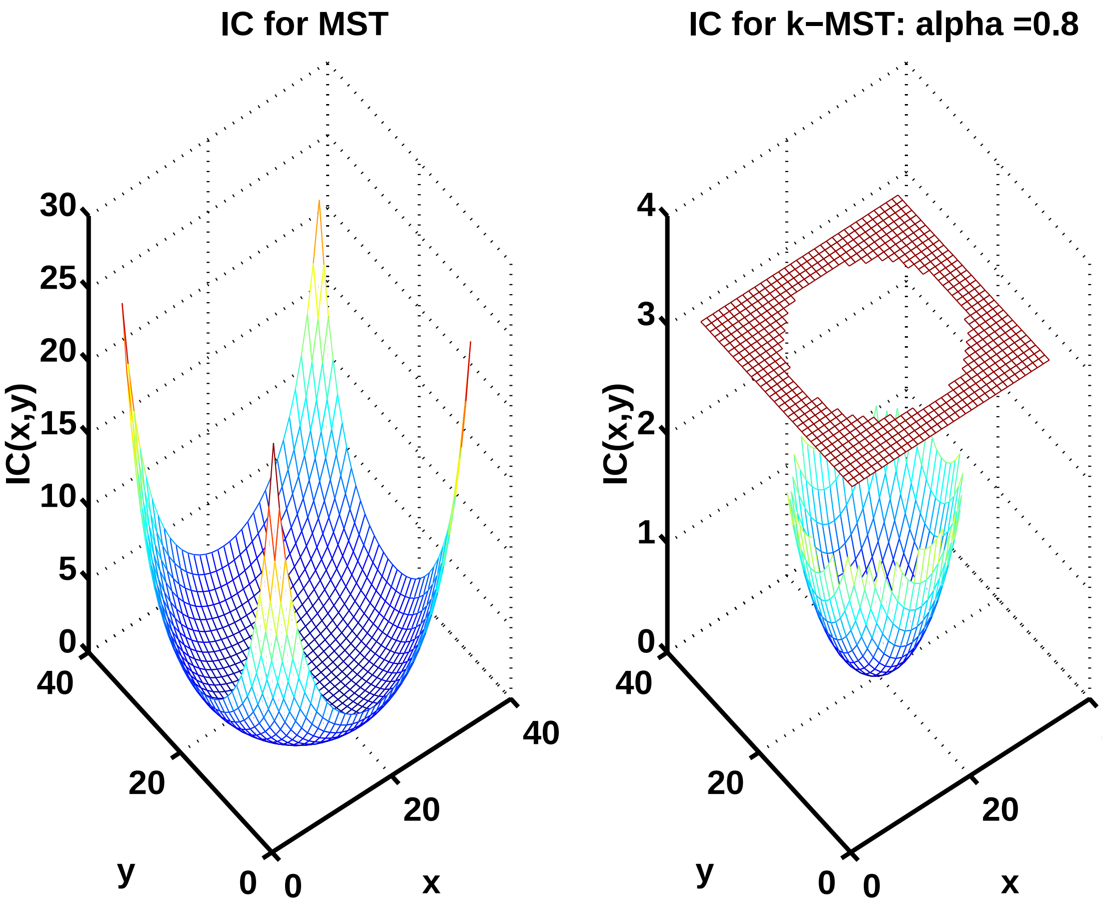

Fig. 5. *MST and k-MST influence functions for bivariate Gaussian density on the plane. MST influence function is unbounded.*

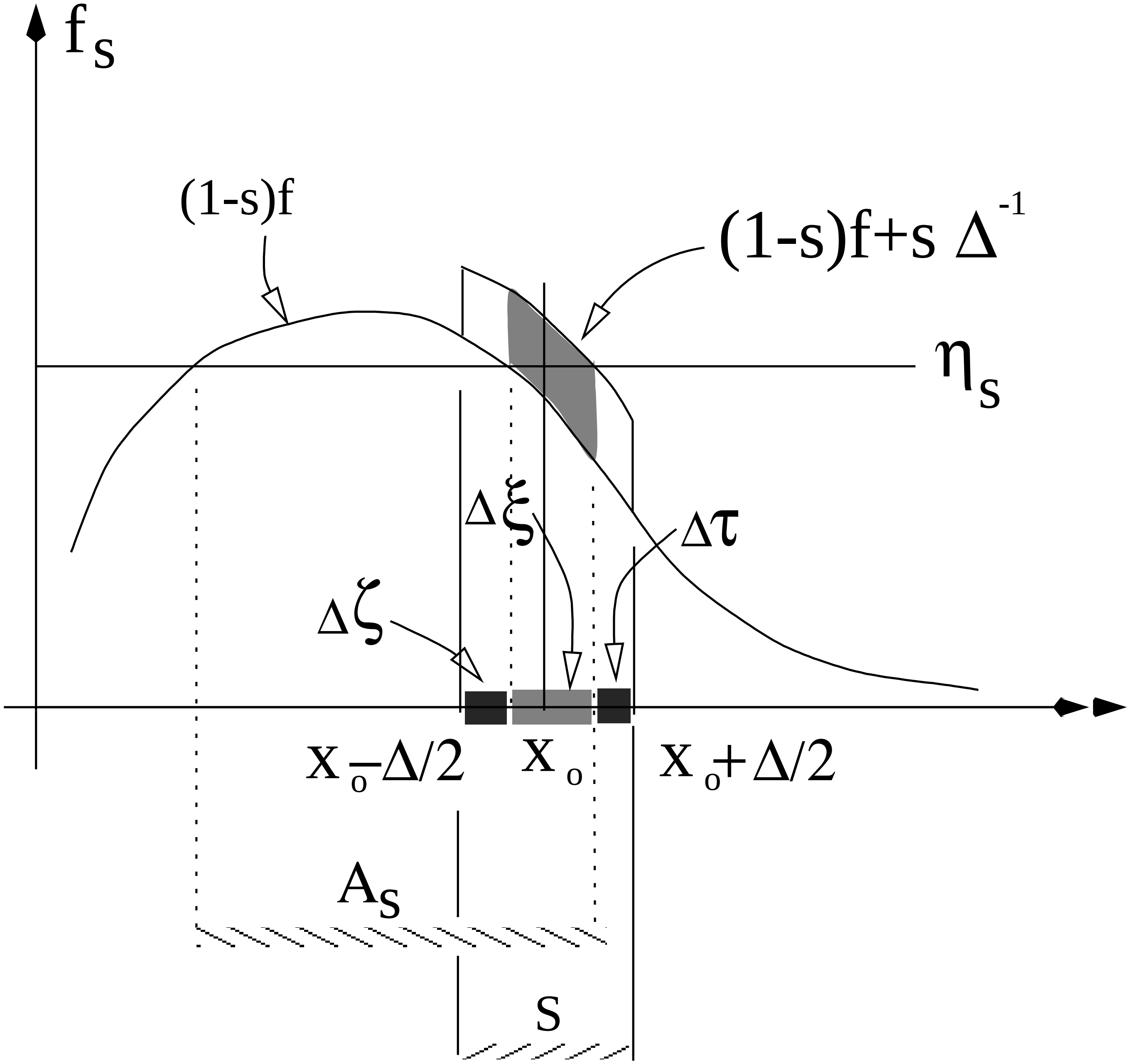

Fig. 6. Graphical illustration of the three constants $\zeta = \zeta_s$, $\xi = \xi_s$, and $\tau = \tau_s$ for the case $d = 1$. $\zeta_s + \xi_s + \tau_s = 1$ and $\zeta_s$ is proportional to the area of the region $S \cap A_s = \{x \in S : (1-s)f(x) \geq \eta\}$ and $\xi_s$ is proportional to the area of the region $\{x \in S : (1-s)f(x) < \eta \leq (1-s)f(x) + s\delta_{x_o}(x)\}$.

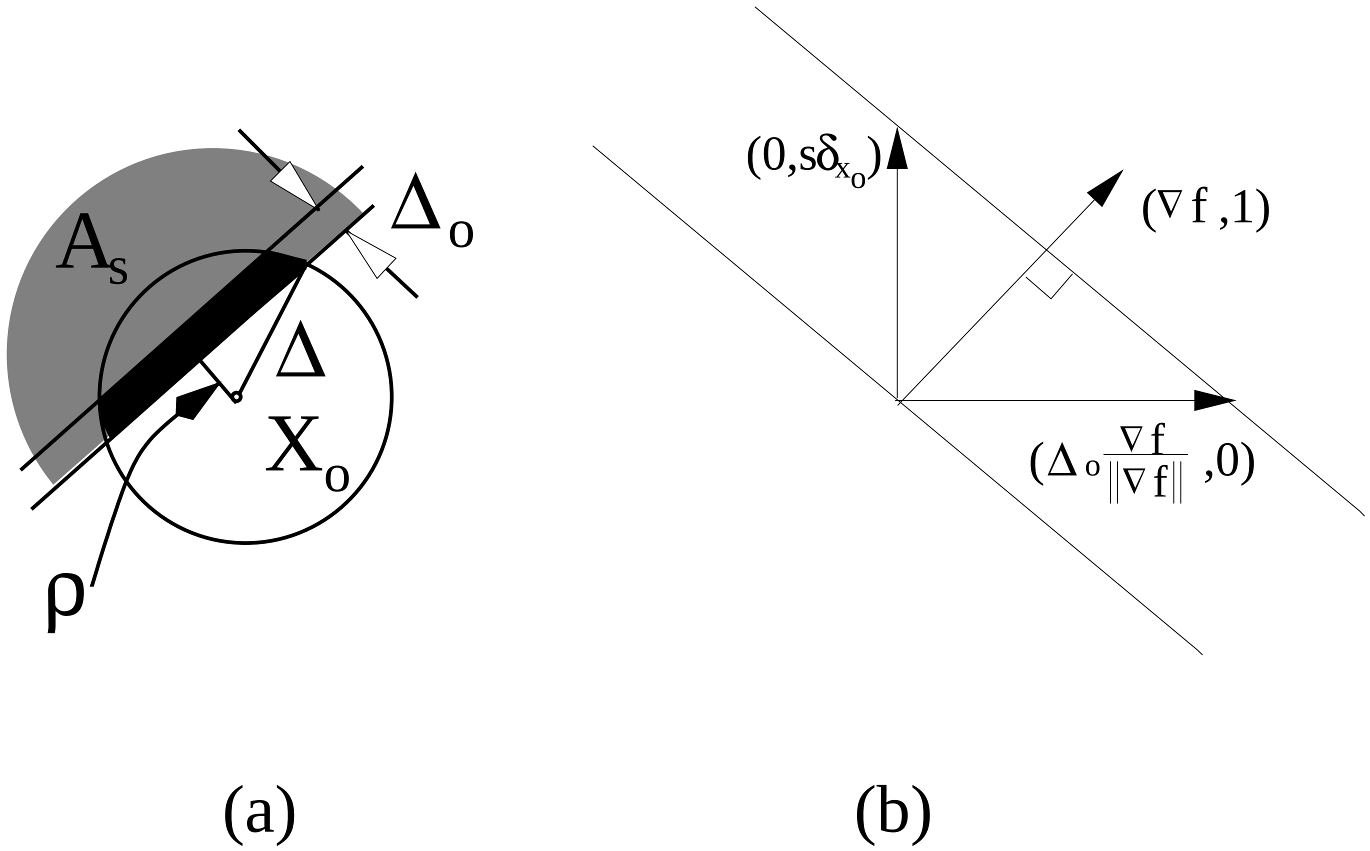

Fig. 7. Graphical illustration of the region $\{x \in S : 0 < \eta_s - (1-s)f(x) \leq s\delta_{x_o}(x)\}$ which is the intersection of the slab of width $\Delta_o$ and the spheroidal support of the uniform density $\delta_{x_o}$ shown in (a) for the case $d = 2$. Slab is at a distance $\rho$ from the center $x_o$ of the spheroid. The width $\Delta_o$ of the slab is determined by the intersection of the horizontal plane at level $\eta_s$ and the two parallel tangent hyperplanes to the surfaces $(1-s)f$ and $(1-s)f + s\delta_{x_o}$. In Figure (b) these are shown along with the normal vector $(\nabla f, 1)$ to these hyperplanes (shown as two parallel lines in (b)).