Robust Entropy Estimation via Greedy Approximation to Minimal k-Point Random Graphs

Alfred O. Hero and Olivier Michel

Abstract— **Let $\mathcal{X}_n = \{x_1, \ldots, x_n\}$, be an i.i.d. sample having multivariate distribution P which is the mixture: $(1-\epsilon)P_1 + \epsilon P_o$, $0 \leq \epsilon \ll 1$. We derive robust estimates of the Renyi entropy of any order $\nu \in (0,1)$ of the distribution P_1. The estimator is constructed by growing a minimum weight graph over a subset k of the points $\mathcal{X}_n$, called a minimal k-point graph, and using the log of the minimum weight function as an entropy estimator. We prove several strong convergence results for a class of polynomial time greedy approximations to the minimal k-point graph constructed on a uniform resolution-$1/m$ partition of the sample space. Specifically, for $x_i \in \mathbb{R}^d$ and power weighted edge lengths with power exponent $0 < \gamma < d$, we show that the log of the total edge weight is a strongly consistent estimator of the conditional Renyi entropy of order $\nu = (d-\gamma)/d$ given $x \in A$ where A is an entropy maximizing subset of coverage probability $\alpha = k/n$. We also obtain tight bounds on convergence rate as a function of resolution $1/m$. An expression for the influence curve (IC) of the greedy entropy estimator is obtained along with an asymptotic approximation to the estimator variance. The form of the IC indicates that the estimator has robustness properties in $\mathbb{R}^d$ which are analogous to those of rank order statistics in one dimension. Several examples are given to quantitatively illustrate the robustness of the technique.**

I. Introduction

In this paper we treat the problem of estimation of the entropy of a multidimensional density function based on n i.i.d. samples. In particular, we consider estimation of the p-th order Renyi entropy of a distribution function P based on a random sample from the mixture distribution $(1-\epsilon)P_1 + \epsilon P_o$ where ϵ is an unknown mixture parameter between zero and one and P_o is a high entropy contaminating distribution. Entropy estimation has been of interest for pattern analysis, process complexity assessment, model identification, fractal models, and other applications where invariance to scale, translation and other invertible transformations is desired in the discriminant. Another application is in vector quantization where Renyi entropy is related to asymptotic quantizer distortion via Bennett's integral [1], [2]. Possible applications of our results include: estimation of Lyapounov exponents in non-linear models [3], [4], multimodality image registration using mutual information matching criteria [5], stopping criteria for regression and classification trees and adaptive tree structured vector quantization (TSVQ) [6].

The most direct method for entropy estimation is to use density estimation techniques, e.g. kernel estimation or smoothed histograms, and subsequently extract the entropy via functional transformation. Direct methods suffer from poor conditioning due to the complications arising from estimating very high dimensional distribution functions with finite data. Here we propose a different method based on computing the weight of minimal graphs incuding: spanning trees, traveling salesman tours, and steiner trees spanning through portions of the data points. The principal results presented here are as follows:

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 [7] 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 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. The minimal k-point graph entropy estimator with power weighted edge criterion with exponent γ is shown to converge a.s. with rate $n^{-\nu}$ to the maximum Renyi entropy of order $\nu = (d-\gamma)/d \in (0,1)$, where d is the sample space dimension. For $\nu > 1/2$ this rate is better than the standard $1/\sqrt{n}$ rate of kernel based histogram estimators.
5. Our asymptotic results hold for a very general class of graphs constructed by minimizing an edge weight function which is a quasiadditive functional. This class includes the optimal Euclidean traveling salesman tour, the minimal spanning tree, the Steiner tree, and the two population minimal matching graph.
6. When the density follows a mixture model the minimal k-point graph algorithm effectively rejects samples from the contaminating distribution. Influence curves studies are presented which quantitatively establish that the greedy minimal k-point graph construction generates a robust estimator of distribution entropy.
7. The influence curve study provides a threshold rule on the graph's edge weight function which we apply to select an optimal proportional number of points $(n-k)/n = 1-\alpha$. This threshold specifies the number of points which should be rejected from the sample to achieve maximum outlier rejection rate while preserving inliers.
8. 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 $k \to \infty$. Under this condition, the total weight of the exact minimal k-point graph and the greedy approximation converge to the same asymptotic limit, i.e. the greedy approximation is asymptotically optimal. However, we can only conjecture that this condition is satisfied for the k-point minimal spanning tree.

The outline of the paper is as follows. In Section II-A we review Euclidean minimal spanning graphs, including those arising from the solution of the traveling salesman problem, the minimal spanning tree, and the Steiner tree. In Section II-B we review the theory of quasiadditive functionals which were used by Redmond and Yukich to prove a general asymptotic theo-

[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 and NIH Grants R01-CA54362 and R01-CA59412.

rem 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 giving a partition approximation for such graphs which satisfy a tightly coverable property. 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 VI by a study of quantitative robustness of the greedy approximation via the influence function. We conclude in Section VII with a set of examples.

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 [8].

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 ψ be an edge weight function which satisfies $\psi(|e|) \geq 0$. The total weight $L(\mathcal{X})$ of 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 [9] 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"

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

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 [10]. Let $\mathcal{T}(\mathcal{X}_n)$ denote the possible sets of edges in the class graphs of degree 2 which span $\mathcal{X}_n$. The weight of the minimal TSP tour is specified by

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

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 [11], [12], two sample matching [13],pattern recognition [14], clustering [15], nonparametric regression [16] and testing for randomness [17]. 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 [18]. Steiner tree problems are NP-hard but have been of interest for minimum area routing problems, e.g. in VLSI layout [19]. 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_{ST} can be studied using the more general concept of quasiadditive euclidean functionals introduced by Redmond and Yukich [20].

B. Quasiadditive Euclidean Functionals

Let F be a finite subset of $\mathbb{R}^d$, i.e. a set of points. Define the following conditions on a real valued set function L [20]:

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

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

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

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

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

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

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

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 conditions, superadditivity and continuity.

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

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

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. Redmond and Yukich [20, Thm. 1.3] show that there exist dual functionals and that L is quasiadditive for a large number of minimal graph problems including the minimal TSP tour, the MST, the Steiner tree, and the two population minimal matching problem. They also prove the following strong convergence theorem[2].

Theorem 1: Let L be a continuous quasiadditive Euclidean functional with power-exponent γ, 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

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

In Theorem **??** $\beta_{L,\gamma}$ is a constant which only depends on the functional L and the graph optimality criterion, e.g. 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 the well known Beardwood, Halton and Hammersley Theorem [21].

III. Minimal k-point Euclidean Graphs

Assume that we are given a set $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ of n points in $\mathbf{R}^d$. 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 traveling salesman problem for visiting at least k out of n cities, and other combinatorial optimization problems.

[2] Redmond and Yukich actually prove an even stronger convergence (complete convergence) of the functional $L_n = L(\mathcal{X}_n)$ for the case $\gamma = 1$. Steele [9] proves the a.s. limit for $\gamma \neq 1$ in the context of the MST.

For example, the Euclidean *minimal k-point spanning tree* (k-MST) is the minimum weight tree spanning any k of the points. Denote the k vertices of the k-MST by $\mathcal{X}^*_{n,k}$. The weight of the k-MST is

$$\begin{aligned} L_{\mathrm{kMST}}(\mathcal{X}^*_{n,k}) &= \min_{\mathcal{X}_{n,k}, \mathcal{T}(\mathcal{X}_{n,k})} \sum_{e \in \mathcal{T}(\mathcal{X}_{n,k})} |e|^\gamma \\ &= \min_{\mathcal{X}_{n,k}} L_{\mathrm{MST}}(\mathcal{X}_{n,k}). \end{aligned}$$

As pointed out in Eppstein [22] the planar k-MST problem was shown to be NP-complete by Zelikovsky and Lozevanu [23] and Ravi, Sundaram, Marathe, Rosenkrantz and Ravi [7]. Ravi etal proposed a polynomial time approximation algorithm for the planar k-MST with approximation ratio $O(k^{1/4})$ which has been successively improved to $O(\log(k))$ by Garg and Hochbaum [24], $O(\log(k)/\log\log(n)$ by Eppstein [22], $O(1)$ by Blum, Chalasani and Vempala [25], $2\sqrt{2}$ by Mitchell [26], 3 by Garg [27], and $1+\epsilon$ by Arora [18]. Arora has recently obtained a linear time $1+\epsilon$ algorithm for multi-dimensional k-point TSP, MST and Steiner problems [28].

A. A Tight Cover Property

Let $\mathcal{Q}^m$ be a uniform partition of $[0,1]^d$ into m^d cubes 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 algorithms for which the vertices of the minimal graph can be covered by a small number of partition cells Q_i. Define $\mathrm{card}(D) = \sum_{i=1}^n I_D x_i$ the number of points of $\mathcal{X}_n$ falling in a set $D \in [0,1]^d$.

Definition 2 (**Tightly Coverable Graphs**) Let $\mathcal{Q}^m$ be a sequence of uniform partitions of $[0,1]^d$ of resolution $1/m$, $m = 1, 2, \ldots$, and let $\sigma(\mathcal{Q}^m)$ denote the sigma-algebra generated by the partition cells in $\mathcal{Q}^m$. Let G_n be 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 $D^m_k = \cap_{\{C \in \sigma(\mathcal{Q}^m): \mathcal{X}_{n,k} \in C\}} D$ 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 $m > M$

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

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

It is not difficult to see that the weight functional $L(\mathcal{X}^*_{n,k})$ for the k-MST graph neither satisfies the subadditivity property nor the continuity property. Hence, the elegant methods of Steele [9], Redmond and Yukich [20], and Rhee [29] cannot be directly applied.

Consider the class (sigma algebra) $\sigma(\mathcal{Q}^m)$ of sets C of resolution $1/m$ for which the probability $P(C)$ is greater than α. Out of this class we define C^m_α to be a set for which $L(C \cap \mathcal{X}_n)$

is minimized

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

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

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

With these definitions we have:

Lemma 1: Let L be a continuous subadditive functional with power exponent γ as in Theorem 1. If $\text{card}(C_\alpha^m)\geq k$

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

Proof

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

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

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

$$\begin{aligned} L(\mathcal{X}_n\cap C_\alpha^m) &\geq L(\mathcal{X}_{n,k}^{C_\alpha^m}) - C_3\left(\text{card}(C_\alpha^m)-k\right)^{(d-\gamma)/d} \\ &\geq L(\mathcal{X}_{n,k}^*) - C_3\left(\text{card}(C_\alpha^m)-k\right)^{(d-\gamma)/d}. \end{aligned} \tag{6}$$

Furthermore, again by continuity,

$$\left|L(\mathcal{X}_n\cap D_k^m) - L(\mathcal{X}_{n,k}^*)\right| \leq C_3\left(\text{card}(D_k^m)-k\right)^{(d-\gamma)/d}, \tag{7}$$

so that

$$L(\mathcal{X}_n\cap D_k^m) \leq L(\mathcal{X}_{n,k}^*) + C_3\left(\text{card}(D_k^m)-k\right)^{(d-\gamma)/d}. \tag{8}$$

Hence, combining (6) and (8)

$$\begin{aligned} &L(\mathcal{X}_n\cap D_k^m) - L(\mathcal{X}_n\cap C_\alpha^m) \\ &\leq C_3\left[\left(\text{card}(D_k^m)-k\right)^{(d-\gamma)/d} + \left(\text{card}(C_\alpha^m)-k\right)^{(d-\gamma)/d}\right] \end{aligned} \tag{9}$$

On the other hand, by definition of C_α^m the left hand side of inequality (9) is greater than zero so that

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

which when combined with (7) gives

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

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

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. 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 thus 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 is still an open problem.

IV. Limit Theorem for a k-MST 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 [7] for points on the plane. Their algorithm (generalized to unit hypercubes of d dimension, $d>1$, and arbitrary subadditive edge weight function L) is performed in three stages: 1) specify a uniform partition $\mathcal{Q}^m$ of $[0,1]^d$ having m^d cells Q_i of resolution $1/m$; 2) find the smallest subset $B_k^m=\cup_i Q_i$ of partition elements containing at least k points; 3) on this smallest subset select 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, which 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. Similarly to [30], [7] 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

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

At termination of the algorithm $j=\tilde{q}\leq m^d$ and we have a minimal subset $B_{\lfloor\alpha n\rfloor}^m \stackrel{\text{def}}{=} B=\cup_{i=1}^{\tilde{q}} Q_{(i)}$ containing at least k points. Let $\mathcal{X}_{n,k}^{G_m}$ be 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 α.

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 *mininal resolution*-$1/m$ *set of probability at least* α. The class of all such sets is denoted $\mathcal{A}_\alpha^m$ and, as shown in

the following construction, all sets in $\mathcal{A}_\alpha^m$ have identical coverage probabilities $p_{\mathcal{A}_\alpha^m} \geq \alpha$.

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 $I - K$ identical values of φ_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 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 $\mathcal{A}_\alpha^m$

$$\overline{\mathcal{A}_\alpha^m} = \cap_{A_i \in \mathcal{A}_\alpha^m} A_i \tag{10}$$

and the associated residual set

$$\begin{aligned} \partial\mathcal{A}_\alpha^m &= \cup_{A_i \in \mathcal{A}_\alpha^m} A_i - \cap_{A_i \in \mathcal{A}_\alpha^m} A_i \\ &= \cup_{A_i, A_j \in \mathcal{A}_\alpha^m} A_i A_j^c. \end{aligned} \tag{11}$$

The total variation $v(Q)$ over a rectangle $Q \subset [0,1]^d$ of a function g on $\mathbb{R}^d$ is defined as [31]

$$v(Q) = \limsup_{\{z_i\} \in Q} \sum_i |g(z_i) - g(z_{i-1})| = v(Q), \tag{12}$$

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 ϕ 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 [9] and [20] 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 η on $f(x)$ (see remark below). The main result of this section is the following asymptotic theorem.

Theorem 2: 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 $A \subset [0,1]^d$. 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

$$\limsup_{n\to\infty} \left| L(\mathcal{X}_{n,\lfloor \alpha n \rfloor}^{G_m})/n^{(d-\gamma)/d} - \beta_{L,\gamma} \int_{A_\alpha^m} f^{(d-\gamma)/d}(x)dx \right| < \delta, \quad (a.s.), \tag{13}$$

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

$$\begin{aligned} \delta &= 2m^{-d} \sum_{i=1}^{m^d} v(Q_i \cap \partial\mathcal{A}_\alpha^m) + C_3(p_{\mathcal{A}_\alpha^m} - \alpha)^{(d-\gamma)/d} \\ &= O(m^{\gamma-d}), \end{aligned} \tag{14}$$

and $p_{\mathcal{A}_\alpha^m}$ is the coverage probability of sets in $\mathcal{A}_\alpha^m$. Furthermore, the bound (13) holds pointwise when $L(\mathcal{X}_{n,\lfloor \alpha n \rfloor}^{G_m})$ is replaced by $E[L(\mathcal{X}_{n,\lfloor \alpha n \rfloor}^{G_m})]$.

The following result relates the integral over a set A in $\mathcal{A}_\alpha^m$ in (13) to a constrained minimum over $A \in \sigma(\mathcal{Q}^m)$.

Lemma 2: Under the assumptions of Theorem 2

$$\int_{A_\alpha^m} f^{(d-\gamma)/d}(x)dx = \min_{A \in \sigma(\mathcal{Q}^m): P(A) \geq \alpha} \int_A f^{(d-\gamma)/d}(x)dx + O(m^{-d}). \tag{15}$$

Proof

First recall that by construction of $\mathcal{A}_\alpha^m$ the coverage probability of any set in $\mathcal{A}_\alpha^m$ satisfies for some q: $\sum_{i=1}^{q-1} \varphi_{(i)} < \alpha$ and $\sum_{i=1}^{q} \varphi_{(i)} = p_{\mathcal{A}_\alpha^m} \geq \alpha$, and $0 \leq p_{\mathcal{A}_\alpha^m} - \alpha \leq \varphi_{(q)}$, 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 (15) 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 $A_\alpha^m \in \mathcal{A}_\alpha^m$ and for any η satisfying $m^d\varphi_{(q-1)} < \eta \leq m^d\varphi_{(q)}$: if $x \in A_\alpha^m$ then $\tilde{f}(x) \geq \eta$, or, equivalently, $\tilde{f}^{(d-\gamma)/d}(x) - \lambda\tilde{f}(x) \leq 0$ where $\lambda = m^{-\gamma}/\eta^{\gamma/d}$. With $I_{A_\alpha^m}(x)$ the indicator function of A_α^m this implies that for any $A \in \sigma(\mathcal{Q}^m)$

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

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

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

or

$$\begin{aligned} \int_{A_\alpha^m} \tilde{f}^{(d-\gamma)/d}(x)dx - \int_A \tilde{f}^{(d-\gamma)/d}(x)dx \\ \leq \lambda \left(\int_{A_\alpha^m} \tilde{f}(x)dx - \int_A \tilde{f}(x)dx \right). \end{aligned} \tag{16}$$

Now, as $P(A_\alpha^m) = p_{\mathcal{A}_\alpha^m}$, if $P(A) \geq \alpha$ then the right hand side of this inequality is upper bounded by $\lambda(p_{\mathcal{A}_\alpha^m} - \alpha) \leq \varphi_{(q)} = m^{-d}\theta_{(q)}$. Hence, minimizing both sides of the inequality (16) over A we obtain

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

Since, obviously,

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

the lemma is established. □

We next show how Theorem 2 can be related to estimation of Renyi 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

$$\begin{aligned}\lim_{m\to\infty}\int_{A_\alpha^m} f^{(d-\gamma)/d} &= \inf_{A\in\mathcal{B}:P(A)\geq\alpha}\int_A f^{(d-\gamma)/d} \\ &= \inf_{A\in\mathcal{B}:P(A)=\alpha}\int_A f^{(d-\gamma)/d}. \quad (17)\end{aligned}$$

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. The Renyi entropy of $f(x|A)$ of order $\nu\in(0,1)$ is defined as

$$R_\nu(f|A)=\frac{1}{\nu-1}\log\int f^\nu(x|A)dx. \quad (18)$$

This is also called the conditional Renyi entropy given A. As $\nu-1<0$ maximization of $R_\nu(f|A)$ over A is equivalent to minimization of the integral in (18). Let A_o be the probability-at-least-α Borel subset of $[0,1]^d$ which maximizes $R_\nu(f|A)$

$$R_\nu(f|A_o)=\sup_{\{A\in\mathcal{B}:P(A)\geq\alpha\}} R_\nu(f|A). \quad (19)$$

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

$$\hat{R}_\nu \stackrel{\text{def}}{=} \frac{1}{\nu-1}\log L(\mathcal{X}_{n,\lfloor\alpha n\rfloor}^{G_m})/(\lfloor\alpha n\rfloor)^\nu+\log\beta_{L,\gamma} \quad (20)$$

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 maximum conditional Renyi entropy $R_\nu(f|A_o)$ of order $\nu\in(0,1)$ as $m,n\to\infty$.

Before developing the proof of Theorem 2 we make the following remarks.

1. The bound δ 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-α sets contains only one set $A_1=[0,1]^d$. Therefore, $p_{\mathcal{A}_\alpha^m}-\alpha=0$, $\partial\mathcal{A}_\alpha^m=\phi$, the empty set, 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 μ_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 negligable weight. Thus, by the strong law of large numbers, $\alpha_s n$ of the n points can be included in the graph at negligable 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 $\mathcal{A}_{\alpha-\alpha_s}^m$ of resolution $1/m$ subsets with coverage probability at least $\alpha-\alpha_s$. If $\alpha_s\geq\alpha$ then $\mathcal{A}_\alpha^m$ is replaced by the empty set and $L(\mathcal{X}_{\lfloor\alpha n\rfloor}^{G_m})$ 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 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_k^m 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_k^m to form $\mathcal{X}_{n,k}^{G_m}$. This implies that for large n all methods of elimination are equivalent, including simply randomly rejecting points from B_k^m until exactly k points remain.
4. In smooth estimation problems the factor required to ensure convergence of a parameter estimator is typically $1/\sqrt{n}$. In contrast Theorem 2 says that the stabilization factor for $L(\mathcal{X}_{n,\lfloor\alpha n\rfloor}^{G_m})$ is $1/n^{(d-\gamma)/d}$, which is larger than the standard $1/\sqrt{n}$ factor for $\gamma>d/2$ and smaller for $\gamma<d/2$. Thus, for large m the estimate (20) of the Renyi entropy of order ν for $\nu<\frac{1}{2}$ converges more slowly than for the case $\nu>\frac{1}{2}$.
5. The bound δ of Theorem 2 decays to zero as a function of resolution $1/m$ at overall rate $O(m^{\gamma-d})$. The first term $2C_3m^{-d}\sum_{i=1}^{m^d}v(Q_i\cap\partial A_\alpha^m)\leq 2C_3v([0,1])m^{-d}$ decays as m^{-d} and and is due to non-uniqueness of the resolution-$1/m$ subsets $A_\alpha^m\in\mathcal{A}_\alpha^m$ all of which have identical coverage probability but over which f may have different variation. When f is the blocked distribution [9] $f(x)=m^{-d}\cdot\sum_{i=1}^{m^d}\varphi_i I_{Q_i}(x)$ over the resolution-$1/m$ cells this term is equal to zero. The second additive term $C_3(p_{\mathcal{A}_\alpha^m}-\alpha)^{(d-\gamma)/d}$ is due to the overshoot of coverage probability by the subsets $A_\alpha^m\in\mathcal{A}_\alpha^m$. This term is zero when it so happens that the α 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 of the bound δ in Theorem 2 is fastest for small γ. Since small γ also gives a smaller stabilization factor (see previous remark) this suggests that estimates of Renyi entropy of order $\nu>\frac{1}{2}$ using edge weight power exponents $\gamma<d/2$ can be expected to have better asymptotic performance than $\nu<\frac{1}{2}$ and $\gamma>d/2$.
6. Since $\sup_{x\in Q_{(q)}}f^{(d-\gamma)/d}(x)\leq v([0,1]^d)$ and $\sum_{i=1}^{m^d}v(Q_i\cap\partial\mathcal{A}_\alpha^m)\leq v([0,1]^d)$ we can weaken the a.s. bound in Theorem 2 by using the result (37), which was shown in the proof of Theorem 2. This results in the following
$$\delta\leq\left[2m^{-d}+C_3m^{\gamma-d}\right]v([0,1]^d). \quad (21)$$
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 Renyi entropy $R_\nu=R_\nu(f|A_o)$ is desired such that
$$|L(\mathcal{X}_{n,\lfloor\alpha n\rfloor}^{G_m})/(\lfloor\alpha n\rfloor)^\nu-\beta_{L,\gamma}\exp\{-(1-\nu)R_\nu)|<\epsilon$$
the weakened bound (21) 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-α defined in (19) which maximizes Renyi entropy of order ν is independent of ν and can be constructed by a simple water pouring procedure. To see this define the Lagrangian

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

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 satisify $\rho(A_o + \Delta A, \lambda) - \rho(A_o, \lambda) \leq 0$ for A_o to minimize $\rho(A,\lambda)$ and hence maximize 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

$$A_o = \{x : f(x) \geq \eta\}. \tag{22}$$

where, if possible, $\eta = \lambda^{-d/q} \geq 0$ is selected to satisfy $P(A_o) = \alpha$. Hence, in this case, the conditional density $f(x|A_o)$ in (20) is obtained by truncating $f(x)$ wherever it falls below η and renormalizing to obtain a valid probability density integrating to 1 over $[0,1]^d$. See Fig. **??** 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 η in (22) to satisfy $P(A_o) = \alpha$. Otherwise, we need to slightly modify the definition (22) of A_o. Let η be such that the set $\{x : f(x) < \eta\}$ has probability α_-, the set $\{x : f(x) = \eta\}$ has probability $\alpha_+ - \alpha_- > 0$ and assume that $\alpha \in (\alpha_-, \alpha_+)$. Then defining

$$A_o = \{x : f(x) \geq \eta\} \cup C, \tag{23}$$

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

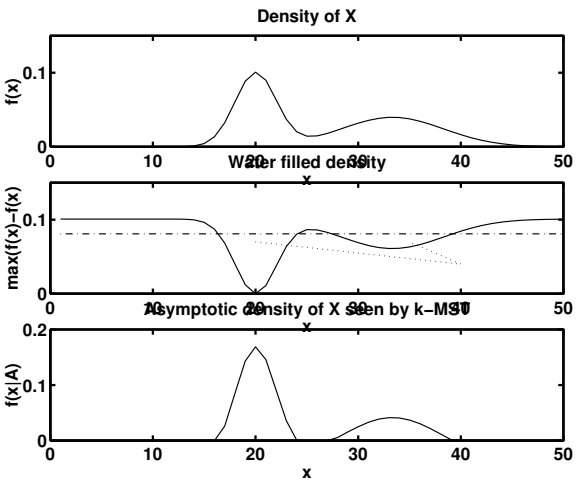

Fig. 1. *Water pouring construction of* $f(x|A_o)$

8. The maximum entropy set A_o in (19) 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 α maximizes entropy. In this case the asssertion 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 $area(A) = \alpha$. 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 Renyi entropy estimator (20) 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.[32] 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 contaminating density f_0 and ϵ is small. Then, since f_1 increasingly dominates f_o as ϵ decreases, for small ϵ a suitable threshold η (α) exists for which: $f(x|A_o) \approx f_1(x)$. Thus (20) can be viewed as a robust estimator of the Renyi entropy of the nominal density f_1. This is the main motivation of this work and will be explored further in the next section.

To prove Theorem 2 we will need several lemmas.

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

Define the m^d independent random variables $N_i = \text{card}(Q_i)$ 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 αn points which satisfies the two conditions:

$$n^{-1} \sum_{i=1}^{\tilde{q}-1} N_{(i)} \quad < \quad \alpha \tag{24}$$

$$n^{-1} \sum_{i=1}^{\tilde{q}} N_{(i)} \quad \geq \quad \alpha. \tag{25}$$

Define $(i)_\varphi$ the index function which establishes a correspondance between rank ordered probabilities $\varphi_{(1)} \geq, \ldots, \geq \varphi_{(m^d)}$ and the cells Q_i which support each of these 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

$$E_n(\epsilon) \quad = \quad \left\{\mathcal{X}_n : n^{-1} \sum_{i=1}^{q-1} N_{(i)_\varphi} \leq \alpha - \epsilon\right\} \tag{26}$$

$$F_n \quad = \quad \left\{\mathcal{X}_n : n^{-1} \sum_{i=1}^{q} N_{(i)_\varphi} \geq \alpha\right\}. \tag{27}$$

Comparing these equations to (24) and (25) 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 $p_y \stackrel{\text{def}}{=} P(y_j = 1) = P(x_i \in A_\alpha^m) - P(x_i \in Q_{(q)_\varphi}) = \sum_{i=1}^{q-1} \varphi_{(i)}$ and $p_z \stackrel{\text{def}}{=} P(z_j = 1) = P(x_i \in A_\alpha^m) = \sum_{i=1}^{q} \varphi_{(i)}$. Then we have the equivalent form for (26) and (27)

$$E_n(\epsilon) = \left\{ Y_n : n^{-1} \sum_{j=1}^{n} y_j \leq \alpha - \epsilon \right\} \tag{28}$$

$$F_n = \left\{ Z_n : n^{-1} \sum_{j=1}^{n} z_j \geq \alpha \right\}. \tag{29}$$

Let δ be defined as the smallest non-zero value of φ_i, $i = 1, \ldots, m^d$. Then by definition of A_α^m we have

$$p_y < \alpha \text{ and } p_z \geq \alpha + \delta. \tag{30}$$

From Sanov's theorem [33], [34]

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

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 $\{p_1, 1-p_2\}$ and $\{p_2, 1-p_2\}$, $p_1, p_2 \in [0,1]$. Furthermore, from (30) 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 $B_{\lfloor \alpha n \rfloor}^m$ 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 $\mathcal{A}_\alpha^m$.

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^ν be of bounded variation over $[0,1]$ 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

First note that as t^ν is a concave 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

$$\begin{aligned} &\int_A [\tilde{f}^\nu(x) - f^\nu(x)]dx \qquad (31) \\ &= m^{-d} \sum_{i:Q_i \cap A \neq \phi}^{m^d} \left[\left(m^d \int_{Q_i} f(x)dx \right)^\nu - m^d \int_{Q_i} f^\nu(x)dx \right]. \end{aligned}$$

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 [31], 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 (31) and the definition (12) of v

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

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 with $L_n(B_{\lfloor \alpha n \rfloor}^m) \stackrel{\text{def}}{=} L(\mathcal{X}_n \cap B_{\lfloor \alpha n \rfloor}^m)$

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

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

The following is a straightforward extension of [20, Eq. (1.2)] proven therein for the case $A = [0,1]^d$ and $\gamma = 1$.

Lemma 6: Assume that f is bounded 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 (10), (11) of $\mathcal{A}_\alpha^m$, $\partial\mathcal{A}_\alpha^m$, and the triangle inequality

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

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

$$\begin{aligned}\left|\int_A f^\nu - \int_{A'} f^\nu\right| &\le 2\max_{A\in\mathcal{A}_\alpha^m} m^{-d}\sum_{i=1}^{m^d} v(Q_i\cap A\cap\partial\mathcal{A}_\alpha^m) \\ &\le 2m^{-d}\sum_{i=1}^{m^d} v(Q_i\cap\partial\mathcal{A}_\alpha^m). \qquad (32)\end{aligned}$$

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 $n > n_0$

$$\left|L_n(A)/n^{(d-\gamma)/d} - \beta_{L,\gamma}\int_A f^{(d-\gamma)/d}(x)dx\right| \le \epsilon, \quad (a.s.) \qquad (33)$$

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 $n > n_2$

$$\min_{A\in\mathcal{A}_\alpha^m} L_n(A) \le L_n(B_{\lfloor\alpha n\rfloor}^m) \le \max_{A\in\mathcal{A}_\alpha^m} L_n(A), \quad (a.s). \qquad (34)$$

Now choosing $n_3 = \max(n_2, n_1)$ it follows from (33) and (34) that for all $n > n_3$

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

Applying the bound (32) we have for arbitrary $A \in \mathcal{A}_\alpha^m$ and all $n > n_3$

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

Since ϵ 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 $n > n_o$, $P(B_{\lfloor\alpha n\rfloor}^m \in \mathcal{A}_\alpha^m) \ge \epsilon$ and

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

for any $A \in \mathcal{A}_\alpha^m$. Furthermore, by continuity of quasiadditive functionals, L_n is bounded:

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

and therefore for $n > n_o$

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

Combining the above and applying (32) yields for $n > n_o$

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

Since ϵ 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. $\mathcal{X}_{n,k}^{G_m}$, drawn from $B_{\lfloor\alpha n\rfloor}^m$.

Lemma 7: Let $\mathcal{X}_{n,\lfloor\alpha n\rfloor}$ be any $\lfloor\alpha n\rfloor$ points selected from $B_{\lfloor\alpha n\rfloor}^m$. Then, with $L_n(B_{\lfloor\alpha n\rfloor}^m) \stackrel{\mathrm{def}}{=} L_n(\mathcal{X}_n\cap B_{\lfloor\alpha n\rfloor}^m)$

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

and

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

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

Proof

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

$$\begin{aligned}&\left|L_n(B_{\lfloor\alpha n\rfloor}^m) - L(\mathcal{X}_{n,\lfloor\alpha n\rfloor})\right|/n^{(d-\gamma)/d} \qquad (35) \\ &\le C_3\left[\frac{\mathrm{card}\left(B_{\lfloor\alpha n\rfloor}^m\right) - \lfloor\alpha n\rfloor}{n}\right]^{(d-\gamma)/d} \\ &\le C_3\left[n^{-1}\mathrm{card}\left(B_{\lfloor\alpha n\rfloor}^m\right) - \alpha\right]^{(d-\gamma)/d} + C_3 n^{-(d-\gamma)/d}.\end{aligned}$$

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

$$\begin{aligned}&P\left(n^{-1}\mathrm{card}(B_{\lfloor\alpha n\rfloor}^m) - \alpha \ge \epsilon\right) \\ &= P\left(n^{-1}\sum_{i=1}^n z_i \ge \epsilon + \alpha\right) \\ &\le (n+1)^2\exp\left\{-nK(\epsilon+\alpha, p_{\mathcal{A}_\alpha^m})\right\}\end{aligned}$$

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

$$P\left(\limsup\left\{\mathcal{X}_n : n^{-1}\text{card}(B^m_{\lfloor \alpha n \rfloor}) - \alpha \geq p_{\mathcal{A}_\alpha^m} - \alpha\right\}\right) = 0. \tag{36}$$

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 (36) and (35).

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 $n > n_o$

$$\left| L(\mathcal{X}_{n,\lfloor \alpha n \rfloor})/n^{(d-\gamma)/d} - \beta_{L,\gamma} \int_{A_\alpha^m} \tilde{f}^{(d-\gamma)/d}(x)dx \right|$$

$$< 2m^{-d} \sum_{i=1}^{m^d} v(Q_i \cap \partial\mathcal{A}_\alpha^m) + C_3(p_{\mathcal{A}_\alpha^m} - \alpha)^{(d-\gamma)/d}, \quad (a.s.)$$

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

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 $A \in \mathcal{A}_\alpha^m$ 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

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

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

$$0 \leq C_3(p_{\mathcal{A}_\alpha^m} - \alpha)^{(d-\gamma)/d} \leq C_3 m^{\gamma-d} \sup_{x \in Q_i} f^{(d-\gamma)/d}(x). \tag{37}$$

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

□

V. 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 maximizes Renyi entropy is

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

and the associated maximum entropy conditional density is

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

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 α

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

and the Renyi estimator converges a.s. to the log linear function $\hat{R}_\nu = \log \alpha$.

Next assume that f is a multivariate Gaussian density with mean μ 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 maximum 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

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

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

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

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 the graph weight curves as a function of α.

VI. Influence Functions

Influence functions have been used to study quantitative robustness of estimators to outliers and other contaminating densities for over thirty years [35]. These functions provide a quantitative measure of outlier sensitivity of an estimator. An unbounded influence curve 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 curves (see Figure 2).

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

Here we give the influence function for the normalized greedy minimal k-point graph weight $L(\mathcal{X}^{G_m}_{n,k})/(\lfloor \alpha n \rfloor)^{(d-\gamma)/d}$ described in Section IV. The influence function motivates the use of the weight function 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.

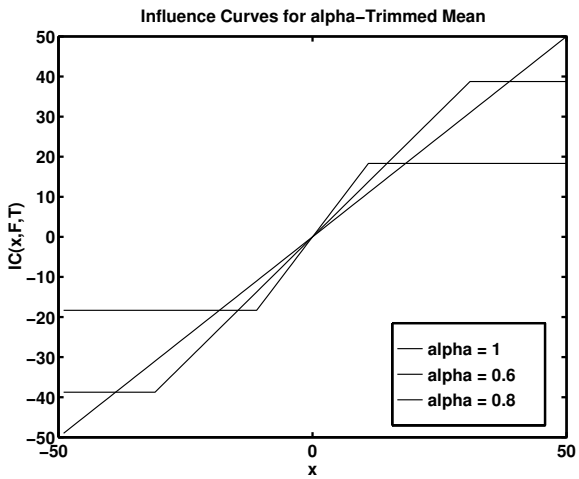

Fig. 2. *2in*

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

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

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 [36]

$$\text{IC}(x_o) = \lim_{s \to 0} \frac{T(\ (1-s)P + s\delta_{x_o}) - T(P)}{s}. \tag{38}$$

where δ_{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 δ_{x_o}. Thus $\text{IC}(x_o)$ can be used to probe the asymptotic sensitivity of the estimator T_n to localized perturbations of P.

The influence function can also be used to approximate asymptotic estimator variance [36] via the formula

$$n\text{var}(T_n) \ \rightarrow \ \int \text{IC}^2(x) f(x) dx$$

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 (17) which we thus identify as $T(P)$.

In the Appendix (Lemma 8) we show that $\text{IC}(x_o)$ has the following form for densities f having no flat spots (cf Remark 2 of Section IV)

$$\text{IC} = \begin{cases} \frac{\alpha-1}{\alpha^\nu}\zeta(\alpha) - \nu \int f^\nu(x|A_o)dx + \frac{\nu}{\alpha} f^{\nu-1}(x_o|A_o), & x_o \in A_o \\ \frac{\alpha}{\alpha^\nu}\zeta(\alpha) - \nu \int f^\nu(x|A_o)dx, & x_o \in A_o \end{cases} \tag{39}$$

where ζ is a non-negative function and $A_o = \{x : f(x) \geq \eta\}$ is the entropy maximizing set of probability α.

Note that the influence function (39) 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 $f^{1-\nu} = f^{-|\nu-1|}$ 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 IC 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 3 where IC 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 [13] in which they proposed the MST as a natural generalization of one dimensional rank order statistical tests of Smirnov and Wald Wolfowitz.

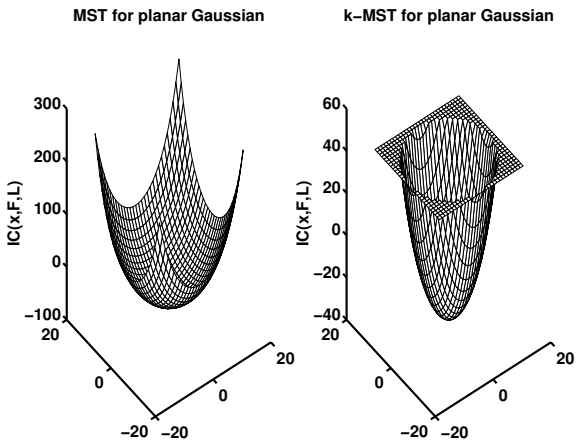

Fig. 3. *2in*

MST and k-MST influence curves for Gaussian density on the plane.

VII. Numerical Examples

In Fig. 4 is a set of 100 random points drawn from the uniform distribution $f_o(x)$ over a rectangle in the plane $[0,15] \times [0,15]$. Figure 5 shows a graph of the Euclidean lengths ($\alpha = \frac{p-1}{p}$) of the k-minimum MST for $k \in \{0, 1, \ldots, 75\}$ computed from these points. Note that this curve is well approximated by a straight line with constant slope (regression line not shown). Figure 6 shows a set of 100 random points on the plane drawn from the mixture distribution $(1-\epsilon)f_1(x) + \epsilon f_o(x)$ where, $\epsilon = 1/2$, f_o is uniform as above and $f_1(x)$ is a density whose constant contours are annuli centered at the point $(7.5, 7.5)$ with mode occurring at radius 4 and of standard width

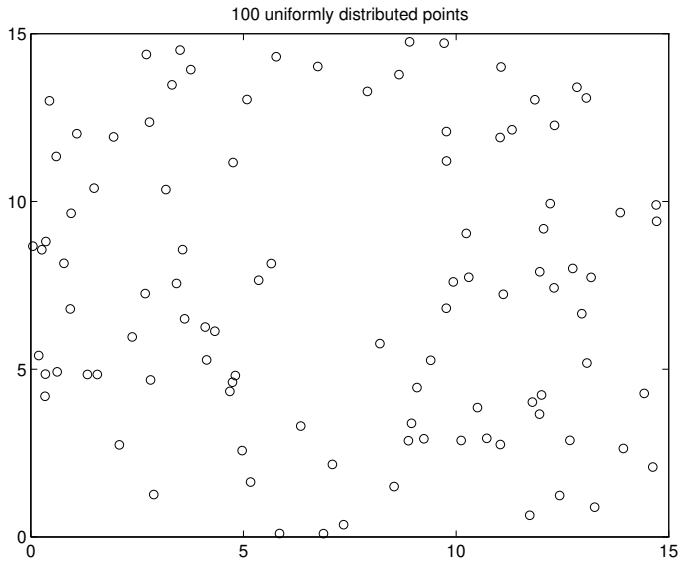

Fig. 4. 100 points drawn from the uniform distribution on the plane.

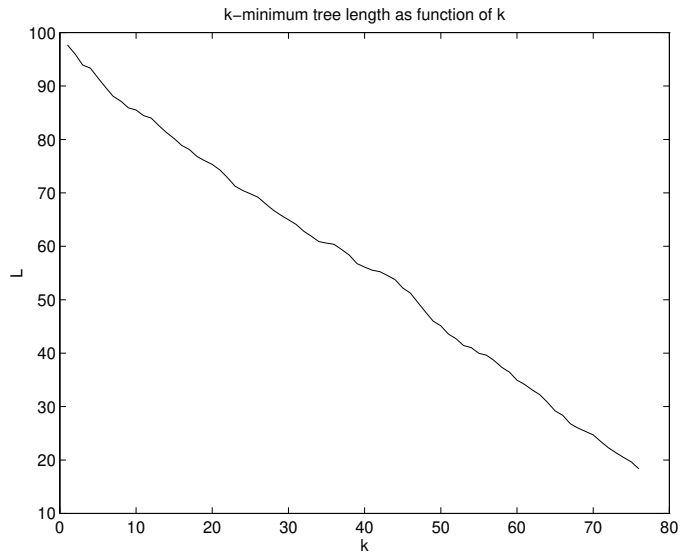

Fig. 5. Length of k-minimum MST plotted as a function of k for points shown in Fig. 4 drawn from uniform distribution.

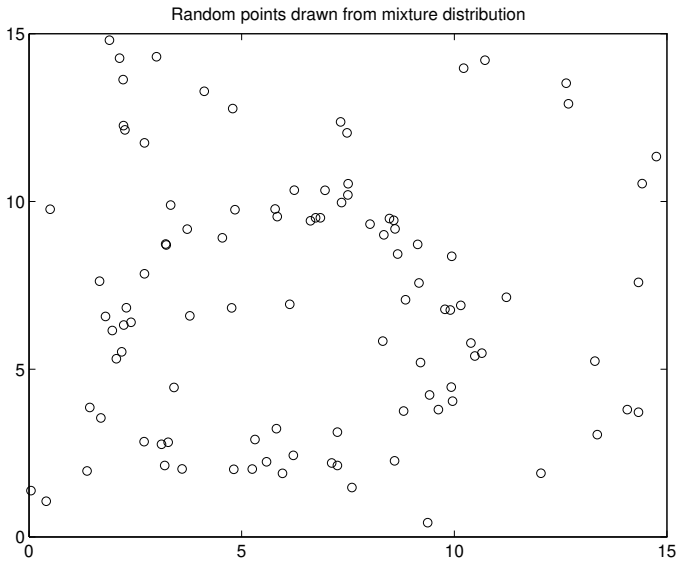

Fig. 6. 100 points drawn from mixture density $\frac{1}{2}f_1 + \frac{1}{2}f_o$ on the plane. Here f_o is uniform and f_1 is as shown in figure Fig. ??

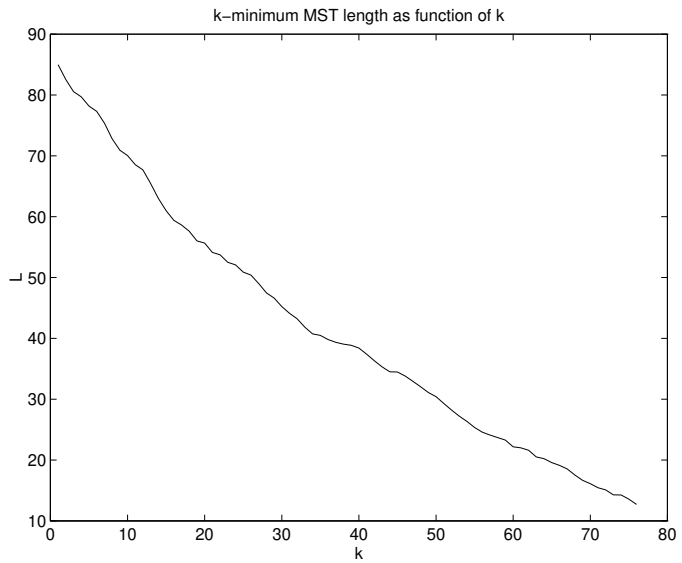

Fig. 7. Length of the k-minimum MST plotted as a function of k for mixture data.

unity about the mode (see FIg. ??. Roughly half of the points in Fig. 6 arise from f_1 and the other half from f_o. Figure 7 is a plot of the length of the k-MST as a function of k over the range $k \in \{0, 1, \ldots, 75\}$. Note the presence of a breakpoint or knee in this curve which occurs in the vicinity of $k = 34$. In Fig. 9 we show the corresponding 34 node spanning tree. Observe that when k is chosen as the location of the knee of the length curve the k-minimum spanning tree effectively extracts the points arising from the annular density f_1 and eliminates most of those arising from the uniform noise background f_o.

The Renyi entropy of any fractional order α of the uniform distribution is zero while that of the annular distribution is given in Table ?? for fractional orders $\alpha = 1/2, 1/3, 2/3, 1/4, 3/4$. By performing multiple (50) experiments we generated estimates of the Renyi entropy of these fractional orders from the power weighted length of the full MST. The estimator means and variances are shown in Figs ? for a range of mixture parameters $\epsilon \in [0, 1]$. Fig ? shows corresponding robust estimator curves obtained by using the k-minimum MST and a visual "breakpoint selection" rule explained above. Thus a more or less unbiased estimate results from appying the k-minimum MST Renyi entropy estimation method, however at the expense of an increase in variance due to reduction in sample statistics from elimination of from 25 to 35 percent of the nodes.

VIII. Conclusion

We have given strong asymptotic convergence results for k-MST and greedy approximations to the k-MST. These convergence results indicate that k-MST's provide natural extensions of one dimensional rank order statistics to multiple dimensions.

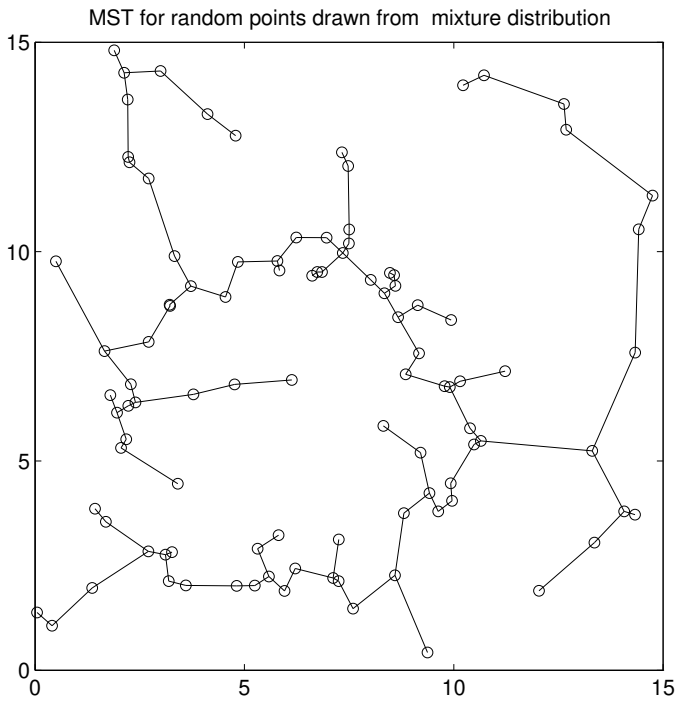

Fig. 8. The minimal spanning tree for points shown in Fig. 6 drawn from mixture distribution.

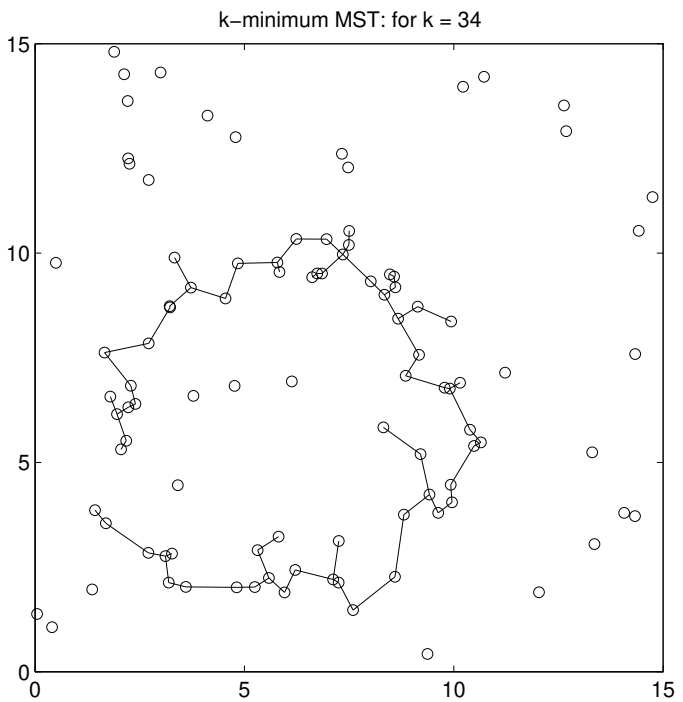

Fig. 9. The k-minimum MST for $k = 34$ for points shown in Fig. 6 drawn from mixture distribution.

Numerical examples were given to illustrate the theory. As a practical problem we will need to develop better methds for finding the breakpoint of the length curve in order to select the best pruning, i.e. the best minimal k-point MST, of the MST. A piecewise linear least squares fit of this curve does not give a very good estimate of this breakpoint. We expect a piecewise exponential fit to provide better results. In this regard it would be very useful to characterize the statistical distribution of the L_n^α curve for various values of α. This is a topic for future work.

IX. Appendix

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 maximum entropy subset of probability α defined in (22) then $g_\nu(\eta) = \int_{A_o} f^\nu(x) dx$ and $g_1(\eta) = P(A_o) = \alpha$.

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. Assume α, η have the relation $\alpha = g_1(\eta)$. Let δ_{x_o} be a uniform distribution over a hypercube with center x_o and volume Δ^d, where Δ is smaller than the modulus of continuity of f. Then the influence function (38) of $T_n = L(\mathcal{X}^{G_m}_{n,\lfloor \alpha n \rfloor})/(\lfloor \alpha n \rfloor)^{(d-\gamma)/d}$ has the form

$$\mathrm{IC}(x_o) = \begin{cases} \frac{\alpha-1}{\alpha^\nu}\zeta(\alpha) - \nu \int f^\nu(x|A_o) + \frac{\nu}{\alpha} f^{\nu-1}(x_o|A_o), & x_o \in A_o \\ \frac{\alpha}{\alpha^\nu}\zeta(\alpha) - \nu \int f^\nu(x|A_o) dx, & x_o \in A_o \end{cases} \tag{41}$$

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

$$\zeta(\alpha) = \frac{g_\nu^{'}\left(g_1^{-1}(\alpha)\right)}{g_1^{'}\left(g_1^{'}(\alpha)\right)}$$

is non-negative and $A_o = \{x : f(x) \geq \eta\}$ is the entropy maximizing set of probability α.

Proof

First we define an entropy maximizing set A_s of probability α analogously to definition (22) of A_o

$$A_s = \{x : f_s \geq \eta\}$$

where η satisfies

$$\int_{A_s} f_s = \alpha. \tag{42}$$

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

$$\begin{aligned} \alpha &= (1-s)\int_{A_s} f + s\int_{A_s} \\ &= \begin{cases} (1-s)\int_{A_s} f + s, & (1-s)f(x_o) + s\Delta^{-d} \geq \eta \\ (1-s)\int_{A_s} f, & o.w. \end{cases}. \end{aligned}$$

Therefore

$$\int_{A_s} f = \begin{cases} \frac{\alpha - s}{1-s}, & (1-s)f(x_o) + s\Delta^{-d} \geq \eta \\ \frac{\alpha}{1-s}, & o.w. \end{cases}. \tag{43}$$

On the other hand, using the definition $g_1(a) = \int_{\{x: f(x) \geq a\}} f$ and the assumption that Δ is less than the modulus of continuity of f, which implies that over the support of δ_{x_o} the density $f(x) \approx f(x_o)$, it is seen that

$$\int_{A_s} f = \begin{cases} g_1(\eta/(1-s)) + \Delta^d f(x_o), & 0 < \eta - (1-s)f(x_o) \leq s\Delta^{-d} \\ g_1(\eta/(1-s)), & o.w. \end{cases} . \quad (44)$$

Hence, combining (43) and (44) we have an equation for the threshold $\eta = \eta(s)$ which guarantees (42)

$$\eta = \begin{cases} (1-s)g_1^{-1}\left(\frac{\alpha-s}{1-s} - \Delta^d f(x_o)\right), & 0 < \eta - (1-s)f(x_o) \leq s\Delta^{-d} \\ (1-s)g_1^{-1}\left(\frac{\alpha-s}{1-s}\right), & (1-s)f(x_o) \geq \eta \\ (1-s)g_1^{-1}\left(\frac{\alpha}{1-s}\right), & o.w. \end{cases} \quad (45)$$

where we have used the obvious fact that, as $\{x : f(x) = a\}$, $a > 0$, has measure zero, $g_1(a)$ is strictly monotone decreasing and thus has an inverse $g_1^{-1}(b)$.

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

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

and note the following identities

$$\int_{A_s} f^\nu = \begin{cases} g_\nu\left(\eta/(1-s)\right) + \Delta^{-d\nu} f^\nu(x_o), & 0 < \eta - (1-s)f(x_o) \leq s\Delta^{-d} \\ g_\nu\left(\eta/(1-s)\right), & o.w. \end{cases}$$

$$\int_{A_s} \delta_{x_o} f^{\nu-1} = \begin{cases} f^{\nu-1}(x_o), & (1-s)f(x_o) + s\Delta^{-d} \geq \eta \\ 0, & o.w. \end{cases} . \quad (46)$$

To simplify the notation we define the partition E_0, E_1, E_2 of $\mathbb{R}^d$

$$\begin{aligned} E_0(s) &= \{x : 0 < \eta - (1-s)f(x) \leq s\Delta^{-d}\} \\ E_1(s) &= \{x : (1-s)f(x) \geq \eta\} \\ E_2(s) &= \{x : (1-s)f(x) + s\Delta^{-d} < \eta\}. \end{aligned}$$

Putting (46) and (46) together and using (45) we obtain

$$\begin{aligned} \int_{A_s} f_s^\nu = (1 - s\nu) g_\nu \Big(g_1^{-1} \Big(& \frac{\alpha - s}{1-s} - \Delta^{-d} f(x_o) I_{E_0}(x_o) \\ & + \frac{s}{1-s} I_{E_2}(x_0) \Big) \Big) s\nu f^{\nu-1}(x_o) I_{E_0 \cup E_1}(x_o) \\ & + + \delta^{-d\nu} f^\nu(x_o) I_{E_0}(x_o) + o(s) \end{aligned} \quad (47)$$

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

$$\begin{aligned} \mathrm{IC}(x_o) &= \lim_{s \downarrow 0} \frac{1}{s} \left[\int f_s^\nu(x|A_s) dx - \int f^\nu(x|A_o) dx \right] \\ &= \frac{1}{\alpha^\nu} \cdot \lim_{s \downarrow 0} \frac{1}{s} \left[\int_{A_s} f_s^\nu(x) dx - \int_{A_o} f^\nu(x) dx \right] \end{aligned} \quad (48)$$

First observe that $E_0 = E_0(s)$ converges to the empty set as $s \to 0$. Hence, in what follows we can safely neglect terms in (47) which are multiplied by I_{E_0}. Next observe that, as $\{x : f(x) = a\}$ has measure zero g_v, g_1 and g_1^{-1} are differentiable a.e. Furthermore, applying the chain rule to the identity $g_1(g_1^{-1}(q(s))) = q(s)$, for any differentiable function $q(s)$

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

Identifying $q(s) = (\alpha - s)/(1-s)$ over $E_1(s)$ and $q(s) = \alpha/(1-s)$ over $E_2(s)$, observing that $E_1(0) = A_o$ and $E_2(0) = A_o^c$, and $o(s)/s \to 0$, it can be verified after some algebra that the limit (48) takes the form

$$\begin{aligned} \mathrm{IC}(x_o) = & \frac{1}{\alpha^\nu} \frac{g_\nu'\left(g_1^{-1}(\alpha)\right)}{g_1'\left(g_1'(\alpha)\right)} \cdot \left[(\alpha - 1) I_{A_o}(x_o) + \alpha I_{A_o^c}(x_o)\right] \\ & - \frac{\nu}{\alpha^\nu} g_\nu\left(g^{-1}(\alpha)\right) + \frac{\nu}{\alpha^\nu} f^{\nu-1}(x_o) I_{A_o}(x_o) \end{aligned} .$$

Finally, using $g_\nu\left(g^{-1}(\alpha)\right) = \int_{A_o} f^\nu$ and the definition of the conditional density $f(x|A_o)$ we obtain the assertion of Lemma 8.

References

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

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

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

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

[5] 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.

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

[7] 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.

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

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

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

[11] 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.

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

[13] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests," *Annals of Statistics*, no. 4, pp. 697–717, 1979.

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

[15] C. Zahn, "Graph-theoretical methods for detecting and describing Gestalt clusters," *IEEE Trans. on Computers*, pp. 68–86, 1971.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

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

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

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

[24] 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.

[25] A. Blum, P. Chalasani, and S. Vempala, "A constant-factor approximation for the k-MST problem in the plane,", pp. 163–171, 1995.

[26] 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.

[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] 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.

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

[30] 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.

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

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

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

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

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

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