Convergence rates of minimal graphs with random vertices

Alfred O. Hero, Jose A. Costa, Bing Ma

Submitted to IEEE Trans. on Information Theory - June 2002.

June 2002

Corresponding author: Alfred O Hero III,
Dept of EECS,
Univ. of Michigan,
Ann Arbor, MI 48109-2122
734-763-0564 (Tel), 734-763-8041 (Fax)
http://www.eecs.umich.edu/~hero, hero@eecs.umich.edu

Abstract

This paper is concerned with power-weighted weight functionals associated with a minimal graph spanning a random sample of n points from a general multivariate Lebesgue density f over $[0,1]^d$. It is known that under broad conditions, when the functional applies power exponent $\gamma \in (1,d)$ to the graph edge lengths, the log of the functional normalized by $n^{(d-\gamma)/d}$ is a strongly consistent estimator of the Rényi entropy of order $\alpha = (d-\gamma)/d$. In this paper we investigate almost sure (a.s.) and $\mathcal{L}_\kappa$-norm (r.m.s. for $\kappa = 2$) convergence rates of this functional. In particular, when $1 \leq \gamma \leq d-1$, we show that over the space of compacted supported multivariate densities f such that $f \in W^{1,p}(\mathbb{R}^d)$ (the space of Sobolev functions) the $\mathcal{L}_\kappa$-norm convergence rate is bounded above by $O\left(n^{-\alpha\lambda(p)/(\alpha\lambda(p)+1)\ 1/d)}\right)$, where $\lambda(p) = 1$, if $1 \leq p \leq d$ and $\lambda(p) = d+1-d/p$, if $d < p < \infty$. We obtain similar rate bounds for minimal graph approximations implemented by a progressive divide-and-conquer partitioning heuristic. In addition to Euclidean optimization problems, these results have application to non-parametric entropy and information divergence estimation; adaptive vector quantization; and pattern recognition. As a concrete illustration, the bounds derived in this paper imply that the maximum r.m.s. error of a minimal-graph estimator of Rényi entropy converges faster than that of any plug-in estimator when the compactly supported multivariate density f lies in the Besov space $B^1_{p,1}(\mathbb{R}^d)$, with $p > d$.

Keywords: continuous quasi-additive functionals, combinatorial optimization, graph theory, progressive-resolution approximations, data partitioning heuristic, non-parametric entropy estimation.

Alfred Hero hero@eecs.umich.edu is with the Departments of Electrical Engineering and Computer Science (EECS), Biomedical Engineering, and Statistics at the University of Michigan, Ann Arbor, MI 48109-2122. Jose Costa jcosta@umich.edu is with the Dept. of EECS at the University of Michigan, Ann Arbor, MI 48109-2122. Bing Ma bingma2001@hotmail.com was with the Dept. of EECS at UM and is now with M-Vision Inc., Belleville MI, USA.
This research was supported in part by AFOSR grant F49620-97-0028. J. Costa was partially supported by Fundação para a Ciência e Tecnologia under the project SFRH/BD/2778/2000 and by a EECS Departmental Fellowship at UM.

1 Introduction

It has long been known that, under the assumption of n independent identically distributed (i.i.d.) vertices in $[0,1]^d$, the suitably normalized weight function of certain minimal graphs over d-dimensional Euclidean space converges almost surely (a.s.) to a limit which is a monotone function of the Rényi entropy of the multivariate density f of the random vertices. Graph constructions that satisfy this convergence property include: the minimal spanning tree (MST), k-nearest neighbors graph (k-NNG), minimal matching graph (MMG), traveling salesman problem (TSP), and their power-weighted variants. See the recent books by Steele [39] and Yukich [41] for introduction to this subject. An $O(n^{-1/d})$ bound on the almost sure (a.s.) convergence rate of the normalized weight functional of these and other minimal graphs was obtained by Redmond and Yukich [33, 34] when the vertices are uniformly distributed over $[0,1]^d$.

In the present paper we obtain bounds on a.s. and $\mathcal{L}_\kappa$-norm (r.m.s. for $\kappa = 2$) convergence rates of power-weighted Euclidean weight functionals of order γ for general Lebesgue densities f for which $f \in W^{1,p}([0,1]^d)$, the Sobolev space of smooth functions over $[0,1]^d$, and $f^{\frac{1}{2}-\frac{\gamma}{d}}$ is integrable. Here the dimension d is greater than one and $\gamma \in (1,d)$ is an edge exponent which is incorporated in the weight functional to taper the Euclidean distance between vertices of the graph (see next section for definitions). As a special case of Proposition 5, we obtain a $O\left(n^{-\alpha\lambda(p)/(\alpha\lambda(p)+1)\ 1/d)}\right)$ bound on the r.m.s. convergence rate when $1 \leq \gamma \leq d-1$, where $\lambda(p) = 1$, if $1 \leq p \leq d$ and $\lambda(p) = d+1-d/p$, if $d < p < \infty$. This bound implies a slower rate of convergence than the analogous $O(n^{-1/d})$ rate bound proven for uniform f by Redmond and Yukich [33, 34], although for large d the two rates coincide on the smoothest Sobolov class $W^{1,\infty}(\mathbb{R}^d)$ of densities. Furthermore, the rate constants derived here suggest that slower convergence occurs when either the (Rényi) entropy of the underlying density f or the ($\mathcal{L}_p$) norm of its (weak) derivative $\mathrm{D}f$ is large.

We also obtain $\mathcal{L}_\kappa$-norm convergence rate bounds for partitioned approximations to minimal graphs implemented by the following fixed partitioning heuristic: 1) dissect $[0,1]^d$ into a set of m^d cells of equal volumes $1/m^d$; 2) compute minimal graphs spanning the points in each non-empty cell; 3) stitch together these small graphs to form an approximation to the minimal graph spanning all of the points in $[0,1]^d$. Such heuristics have been widely adopted, e.g. see Karp [20], Ravi *etal* [31], and Hero and Michel [18], for examples. The computational advantage of this partitioned heuristic comes from its divide-and-conquer progressive-resolution strategy to an optimization whose complexity is non-linear in n: the partitioned algorithm only requires constructing minimal graphs on small cells each of which typically contains far fewer

than n points. In Proposition 6 we obtain bounds on $\mathcal{L}_\kappa$-norm convergence rate and specify an optimal "progressive-resolution sequence" $m = m(n)$, $n = 1, 2, \ldots$, for achieving these bounds.

A principal focus of our research on minimal graphs has been on the use of Euclidean functionals for signal processing applications such as image registration, pattern matching and non-parametric entropy estimation, see e.g. [15, 25, 18, 17], and the entropy estimation application considered in this paper reflects this focus. In particular we show that a Rényi entropy estimator constructed from a continuous quasi-additive minimal-graph, such as the MST or k-NNG, can have faster $\mathcal{L}_2$-norm convergence rates than plug-in estimators, such as those discussed by Bierlant *etal* [3] based on density function estimation. Such graph-based estimators were called entropic graph estimators in [16] and we show that their worst case converge rates are better than those of any plug-in estimator when $f \in B^1_{p,1}([0,1]^d)$, the Besov space of smooth functions over $[0,1]^d$, with $p > d$.

Beyond the signal processing applications mentioned above these results may have important practical implications in adaptive vector quantizer design, where the Rényi entropy is more commonly called the Panter-Dite factor and is related to the asymptotically optimal quantization cell density [12, 28]. Furthermore, as empirical versions of vector quantization can be cast as geometric location problems [14], the asymptotics of adaptive VQ may be studied within the present framework of minimal Euclidean graphs.

The outline of this paper is as follows. In Section 2 we briefly review Redmond and Yukich's unifying framework of continuous quasi-additive power-weighted edge functionals. In Section 3 we give convergence rate bounds for such functionals with general Lebesgue density f. In Section 4 we extend these results to partitioned approximations and in Section 5 we apply the results of Sections 3 and 4 to non-parametric entropy estimation.

2 Minimal Euclidean Graphs

Since the seminal work of Beardwood, Halton and Hammersley in 1959, the asymptotic behavior of the weight function of a minimal graph such as the MST and the TSP over i.i.d. random points $\mathcal{X}_n = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ as $n \to \infty$ has been of great interest. The monographs by Steele [39] and Yukich [41] provide two engaging presentations of ensuing research in this area. Many of the convergence results have been encapsulated in the general framework of continuous and quasi-additive Euclidean functionals recently introduced by Redmond and Yukich [33]. This framework allows one to

relatively simply obtain asymptotic convergence rates once a graph weight function has been shown to satisfy the required continuity and subadditivity properties. We follow this framework in this paper.

Let F be a finite subset of points in $[0,1]^d, d \geq 2$. A real-valued function L_γ defined on F is called a *Euclidean functional of order* γ if it is of the form

$$L_\gamma(F) = \min_{e \in \mathcal{E}} \sum_e |e(F)|^\gamma \tag{1}$$

where $\mathcal{E}$ is a set of graphs, e.g. spanning trees, over the points in F, e is an edge in the graph, $|e|$ is the Euclidean length of e, and γ is called the *edge exponent* or *power-weighting constant*. We assume throughout this paper that $0 < \gamma < d$.

2.1 Continuous Quasi-additive Euclidean Functionals

A weight functional $L_\gamma(\mathcal{X}_n)$ of a minimal graph on $[0,1]^d$ is a continuous quasi-additive functional if it can be closely approximated by the the sum of the weight functionals of minimal graphs constructed on a dense partition of $[0,1]^d$. Examples of quasi-additive graphs are the Euclidean traveling salesman (TSP) problem, the minimal spanning tree (MST), and the k-nearest neighbor graph (k-NNG). In the TSP the objective is to find a graph of minimum weight among the set $\mathcal{C}$ of graphs that visit each point in $\mathcal{X}_n$ exactly once. The resultant graph is called the *minimal TSP tour* and its weight is $L_\gamma^{\mathrm{TSP}}(\mathcal{X}_n) = \min_{e \in \mathcal{C}} \sum_e |e|^\gamma$. Construction of the TSP graph is NP-hard and arises in many different areas of operations research [24]. In the MST problem the objective is to find a graph of minimum weight among the graphs $\mathcal{T}$ which span the sample $\mathcal{X}_n$. This problem admits exact solutions which run in polynomial time and the weight of the MST is $L_\gamma^{\mathrm{MST}}(\mathcal{X}_n) = \min_{e \in \mathcal{T}} \sum_e |e|^\gamma$. MST's arise in areas including: pattern recognition [40]; clustering [42]; nonparametric regression [2] and testing for randomness [19]. The k-NNG problem consists of finding the set $\mathcal{N}_{k,i}$ of k-nearest neighbors of each point X_i in the set $\mathcal{X}_n - \{X_i\}$. This problem has exact solutions which run in linear-log-linear time and the weight is $L_\gamma^{k-\mathrm{NNG}}(\mathcal{X}_n) = \sum_{i=1}^n \min_{e \in \mathcal{N}_{k,i}} \sum_e |e|^\gamma$. The k-NNG arises in computational geometry [8], clustering and pattern recognition [37], spatial statistics [7], and adaptive vector quantization [13].

The following technical conditions on a Euclidean functional L_γ were defined in [33, 41].

- *Null condition*: $L_\gamma(\phi) = 0$, where ϕ is the null set.
- *Subadditivity*: Let $\mathcal{Q}^m = \{Q_i\}_{i=1}^{m^d}$ be a uniform partition of $[0,1]^d$ into m^d subcubes Q_i with edges parallel to

the coordinate axes having edge lengths m^{-1} and volumes m^{-d} and let $\{q_i\}_{i=1}^{m^d}$ be the set of points in $[0,1]^d$ that translate each Q_i back to the origin such that $Q_i - q_i$ has the form $m^{-1}[0,1]^d$. Then there exists a constant C_1 with the following property: for every finite subset F of $[0,1]^d$

$$L_\gamma(F) \leq m^{-\gamma} \sum_{i=1}^{m^d} L_\gamma\left(m[F \cap Q_i - q_i]\right) + C_1 m^{d-\gamma} \tag{2}$$

- *Superadditivity*: For the same conditions as above on Q_i, m, and q_i, there exists a constant C_2 with the following property:

$$L_\gamma(F) \geq m^{-\gamma} \sum_{i=1}^{m^d} L_\gamma\left(m[F \cap Q_i - q_i]\right) - C_2 m^{d-\gamma} \tag{3}$$

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

$$|L_\gamma(F \cup G) - L_\gamma(F)| \leq C_3(\text{card}(G))^{(d-\gamma)/d}, \tag{4}$$

where $\text{card}(G)$ is the cardinality of the subset G. Note that continuity implies

$$|L_\gamma(F) - L_\gamma(G)| \leq 2C_3(\text{card}(F \triangle G))^{(d-\gamma)/d}, \tag{5}$$

where $F \triangle G = (F \cup G) - (F \cap G)$ denotes the symmetric difference of sets F and G.

The functional L_γ is said to be a *continuous subadditive functional* of order γ if it satisfies the null condition, sudadditivity and continuity. L_γ is said to be a *continuous superadditive functional* of order γ if it satisfies the null condition, superadditivity and continuity.

For many continuous subadditive functionals L_γ on $[0,1]^d$ there exists a *dual* superadditive functional L_γ^*. The dual functional satisfies two properties: 1) $L_\gamma(F) + 1 \geq L_\gamma^*(F)$ for every finite subset F; and, 2) for i.i.d. uniform random vectors $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$ over $[0,1]^d$,

$$\left|E[L_\gamma(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)] - E[L_\gamma^*(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)]\right| \leq C_4 n^{(d-\gamma-1)/d} \tag{6}$$

with C_4 a finite constant. The condition (6) is called the *close-in-mean approximation* in [41].

A stronger condition which is useful for showing convergence of partitioned approximations is the *pointwise closeness* condition

$$\left|L_\gamma(F) - L_\gamma^*(F)\right| \leq o\left(\left[\text{card}(F)\right]^{(d-\gamma)/d}\right), \tag{7}$$

for any finite subset F of $[0,1]^d$.

A continuous subadditive functional L_γ is said to be a *continuous quasi-additive functional* if L_γ is continuous subadditive and there exists a continuous superadditive dual functional L_γ^*. We point out that the dual L_γ^* is not uniquely defined. It has been shown by Redmond and Yukich [34, 33] that the boundary-rooted version of L_γ, namely, one where edges may be connected to the boundary of the unit cube over which they accrue zero weight, usually has the requisite property (6) of the dual. These authors have displayed duals and shown continuous quasi-additivity and related properties for weight functionals of the power weighted MST, Steiner tree, TSP, k-NNG and others.

In [41, 33] almost sure limits with a convergence rate upper bound of $O\left(n^{-1/d}\right)$ were obtained for continuous quasi-additive Euclidean functionals $L_\gamma(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)$ under the assumption of uniformly distributed points $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$ and an additional assumption that L_γ satisfies the "add-one bound"

- *Add-one bound*:

$$| E[L_\gamma(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_{n+1})] - E[L_\gamma(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)] | \leq C_5 n^{-\gamma/d}. \tag{8}$$

The MST length functional of order γ satisfies the add-one bound. A slightly weaker bound on a.s. convergence rate also holds when L_γ is merely continuous quasi-additive [41, Ch. 5]. The $n^{-1/d}$ a.s. convergence rate bound is exact for $d = 2$.

3 Convergence Rate Bounds for General Density

In this section we obtain convergence rate bounds for a general non-uniform Lebesgue density f. For convenience we will focus on the case that L_γ is continuous quasi-additive and satisfies the add-one bound, although some of the following results can be established under weaker assumptions. Our method of extension follows common practice [38, 39, 41]: we first establish pointwise convergence rates of the mean $E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)]/n^{(d-\gamma)/d}$ for piecewise constant densities and then extend to arbitrary densities. Then we use a concentration inequality to obtain a.s. and $\mathcal{L}_\kappa$-norm convergence rates of $L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)/n^{(d-\gamma)/d}$.

3.1 Mean Convergence Rate for Block Densities

We will need the following elementary result for the sequel.

Lemma 1 *Let $g(u)$ be a continuously differentiable function of $u \in \mathbf{R}$ which is convex cap and monotone increasing over $u \geq 0$. Then for any $u_o > 0$*

$$g(u_o) - \frac{g(u_o)}{u_o}|\Delta| \leq g(u) \leq g(u_o) + g^{'}(u_o)|\Delta|$$

where $\Delta = u - u_o$ and $g^{'}(u) = dg(u)/du$.

Proof

Since $g(u)$ is convex cap the tangent line $y(u) \stackrel{\text{def}}{=} g(u_o) + g^{'}(u_o)(u - u_o)$ upper bounds g. Hence

$$g(u) \quad \leq \quad g(u_o) + g^{'}(u_o)|u - u_o|.$$

On the other hand, as g is monotone and convex cap, the function $z(u) \stackrel{\text{def}}{=} g(u_o) + \frac{g(u_o)}{u_o}(u - u_o)1_{\{u \leq u_o\}}$ is a lower bound on g, where $1_{\{u \leq u_o\}}$ is the indicator function of the set $\{u \leq u_o\}$. Hence,

$$g(u) \quad \geq \quad g(u_o) - \frac{g(u_o)}{u_o}|u - u_o|.$$

□

A density $f(\boldsymbol{x})$ over $[0,1]^d$ is said to be a block density with m^d levels if for some set of non-negative constants $\{\phi_i\}_{i=1}^{m^d}$ satisfying $\sum_{i=1}^{m^d} \phi_i m^{-d} = 1$,

$$f(\boldsymbol{x}) = \sum_{i=1}^{m^d} \phi_i 1_{Q_i}(\boldsymbol{x})$$

where $1_Q(\boldsymbol{x})$ is the set indicator function of $Q \subset [0,1]^d$ and $\{Q_i\}_{i=1}^{m^d}$ is the uniform partition of the unit cube $[0,1]^d$ defined above.

Proposition 1 *Let $d \geq 2$ and $1 \leq \gamma \leq d-1$. Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. sample points over $[0,1]^d$ whose marginal is a block density f with m^d levels and support $\mathcal{S} \subset [0,1]^d$. Then for any continuous quasi-additive Euclidean functional L_γ of order γ which satisfies the add-one bound (8)*

$$\left| E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d} \int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \right| \leq O\left((nm^{-d})^{-1/d}\right).$$

where $\beta_{L_\gamma,d}$ is a constant independent of f. A more explicit form for the bound on the right hand side is

$$O\left((nm^{-d})^{-1/d}\right) = \begin{cases} \frac{K_1 + C_4}{(nm^{-d})^{1/d}} \int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \ (1 + o(1)), & d > 2 \\ \frac{K_1 + C_4 + \beta_{L_\gamma,d}}{(nm^{-d})^{1/d}} \int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \ (1 + o(1)), & d = 2 \end{cases}.$$

Proof

Let n_i denote the number of samples $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ falling into the partition cell Q_i and let $\{\boldsymbol{U}_i\}_i$ denote an i.i.d. sequence of uniform points on $[0,1]^d$. By subadditivity, we have

$$\begin{aligned} L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) &\leq m^{-\gamma} \sum_{i=1}^{m^d} L_\gamma\left(m[\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} \cap Q_i - q_i]\right) + C_1 m^{d-\gamma} \\ &= m^{-\gamma} \sum_{i=1}^{m^d} L_\gamma(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_{n_i}) + C_1 m^{d-\gamma} \end{aligned}$$

since the samples in each partition cell Q_i are drawn independently from a conditionally uniform distribution given n_i. Note that n_i has a Binomial $B(n, \phi_i m^{-d})$ distribution.

Taking expectations on both sides of the above inequality,

$$E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)] \leq m^{-\gamma} \sum_{i=1}^{m^d} E\left[E\left[L_\gamma(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_{n_i}) | \, n_i\right]\right] + C_1 m^{d-\gamma}. \tag{9}$$

The following rate of convergence for quasi-additive edge functionals L_γ satisfying the add-one bound (8) has been established for $1 \leq \gamma < d$ [41, Thm. 5.2],

$$|E[L_\gamma(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)] - \beta_{L_\gamma, d} n^{\frac{d-\gamma}{d}}| \leq K_1 n^{\frac{d-1-\gamma}{d}}, \tag{10}$$

where K_1 is a function of C_1, C_3 and C_5.

Using the result (10) and subadditivity (9) on L_γ, for $1 \leq \gamma < d$ we have

$$\begin{aligned} E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)] &\leq m^{-\gamma} \sum_{i=1}^{m^d} E\left[\beta_{L_\gamma, d} n_i^{\frac{d-\gamma}{d}} + K_1 n_i^{\frac{d-\gamma-1}{d}}\right] + C_1 m^{d-\gamma} \\ &= m^{-\gamma} \beta_{L_\gamma, d} n^{\frac{d-\gamma}{d}} \sum_{i=1}^{m^d} E\left[\left(\frac{n_i}{n}\right)^{\frac{d-\gamma}{d}}\right] + m^{-\gamma} K_1 n^{\frac{d-\gamma-1}{d}} \sum_{i=1}^{m^d} E\left[\left(\frac{n_i}{n}\right)^{\frac{d-\gamma-1}{d}}\right] + C_1 m^{d-\gamma}. \end{aligned} \tag{11}$$

Similarly for the dual L_γ^* it follows by superadditivity (3) and the close-in-mean condition (6)

$$\begin{aligned} &E[L_\gamma^*(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)] \\ &\geq m^{-\gamma} \beta_{L_\gamma, d} n^{\frac{d-\gamma}{d}} \sum_{i=1}^{m^d} E\left[\left(\frac{n_i}{n}\right)^{\frac{d-\gamma}{d}}\right] - m^{-\gamma}(K_1 + C_4) n^{\frac{d-\gamma-1}{d}} \sum_{i=1}^{m^d} E\left[\left(\frac{n_i}{n}\right)^{\frac{d-\gamma-1}{d}}\right] - C_2 m^{d-\gamma} \end{aligned} \tag{12}$$

for $1 \leq \gamma < d$.

We next develop lower and upper bounds on the expected values in (11) and (12). As the function $g(u) = u^\nu$ is monotone and concave over the range $u \geq 0$ for $0 < \nu < 1$, from Lemma 1

$$\left(\frac{n_i}{n}\right)^\nu \quad \geq \quad p_i^\nu - p_i^{\nu-1}\left|\frac{n_i}{n} - p_i\right|, \tag{13}$$

where $p_i = \phi_i m^{-d}$. In order to bound the expectation of the above inequality we use the following bound

$$E\left[\left|\frac{n_i}{n} - p_i\right|\right] \leq \sqrt{E\left[\left|\frac{n_i}{n} - p_i\right|^2\right]} = \frac{1}{\sqrt{n}}\sqrt{p_i(1-p_i)} \leq \frac{\sqrt{p_i}}{\sqrt{n}}.$$

Therefore, from (13),

$$E\left[\left(\frac{n_i}{n}\right)^\nu\right] \quad \geq \quad p_i^\nu - p_i^{\nu-\frac{1}{2}}/\sqrt{n}. \tag{14}$$

By concavity, Jensen's inequality yields the upper bound

$$E\left[\left(\frac{n_i}{n}\right)^\nu\right] \leq \left[E\left(\frac{n_i}{n}\right)\right]^\nu = p_i^\nu \tag{15}$$

Under the hypothesis $1 \leq \gamma \leq d-1$ this upper bound can be substituted into expression (11) to obtain

$$\begin{aligned}
&E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)/n^{(d-\gamma)/d}] \\
&\quad \leq \beta_{L_\gamma,d} \sum_{i=1}^{m^d} \phi_i^{\frac{d-\gamma}{d}} m^{-d} + \frac{K_1}{(nm^{-d})^{1/d}} \sum_{i=1}^{m^d} \phi_i^{\frac{d-\gamma-1}{d}} m^{-d} + \frac{C_1}{(nm^{-d})^{(d-\gamma)/d}} \\
&\quad = \beta_{L_\gamma,d} \int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \frac{K_1}{(nm^{-d})^{1/d}} \int_{\mathcal{S}} f^{(d-\gamma-1)/d}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \frac{C_1}{(nm^{-d})^{(d-\gamma)/d}}.
\end{aligned} \tag{16}$$

Applying the bounds (15) and (14) to (12) we obtain an analogous lower bound for the mean of the dual functional L_γ^*

$$\begin{aligned}
&E[L_\gamma^*(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)]/n^{(d-\gamma)/d} \\
&\quad \geq \beta_{L_\gamma,d} \int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}} \int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\qquad - \frac{K_1 + C_4}{(nm^{-d})^{1/d}} \int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \frac{C_2}{(nm^{-d})^{(d-\gamma)/d}}
\end{aligned} \tag{17}$$

By definition of the dual,

$$E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)]/n^{\frac{d-\gamma}{d}} \geq E[L_\gamma^*(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)]/n^{\frac{d-\gamma}{d}} - n^{-\frac{d-\gamma}{d}} \tag{18}$$

which when combined with (17) and (16) yields the result

$$\left|\frac{E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]}{n^{\frac{d-\gamma}{d}}} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| \leq \frac{K_1+C_4}{(nm^{-d})^{1/d}}\int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}}\int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \frac{K_2}{(nm^{-d})^{(d-\gamma)/d}} + n^{-\frac{d-\gamma}{d}}, \tag{19}$$

where $K_2 = \max\{C_1, C_2\}$. This establishes Proposition 1. □

3.2 Mean Convergence Rate for Density Functions in Sobolev Spaces

Before extending Proposition 1 to general densities we will need to introduce some concepts from the theory of Sobolev spaces.

Let $\mathcal{L}_p(\mathbb{R}^d)$ be the space of measurable functions over $\mathbb{R}^d$ such that $\|f\|_p = (\int |f(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x})^{1/p} < \infty$. For f a real valued differentiable function over $\mathbb{R}^d$, let $\mathrm{D}_{x_j} f = \partial f/\partial x_j$ be the x_j-th partial derivative of f, and $\mathrm{D}f = [\partial f/\partial x_1, \ldots, \partial f/\partial x_d]$ be the gradient of f. The concept of derivative can be extended to non-differentiable functions. For $f \in \mathcal{L}_1(\mathbb{R}^d)$, g is called the x_j-th *weak derivative* of f [43], written as $g \stackrel{\text{def}}{=} \mathrm{D}_{x_j} f$ if

$$\int_{\mathbb{R}^d} f(\boldsymbol{x})\mathrm{D}_{x_j}\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = -\int_{\mathbb{R}^d} g(\boldsymbol{x})\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

for all functions φ infinitely differentiable with compact support. The weak derivative g is sometimes called the *generalized derivative* of f or *distributional derivative* of f. If f is differentiable, then its weak derivative coincides with the (usual) derivative.

We now define a function space whose members have weak derivatives lying in the $\mathcal{L}_p(\mathbb{R}^d)$ spaces [43]. For $p \geq 1$, define the *Sobolev space*

$$W^{1,p}(\mathbb{R}^d) = \mathcal{L}_p(\mathbb{R}^d) \cap \{f : \mathrm{D}_{x_j} f \in \mathcal{L}_p(\mathbb{R}^d),\ 1 \leq j \leq d\}\,.$$

The space $W^{1,p}$ is equipped with a norm

$$\|f\|_{1,p} = \|f\|_p + \|\mathrm{D}f\|_p\,.$$

The Sobolev space $W^{1,p}(\mathbb{R}^d)$ is a generalization of the space of continuously differentiable functions, in the sense that $W^{1,p}(\mathbb{R}^d)$ contains functions that do not have to be differentiable (in the usual sense), but can be approximated arbitrarily close in the $\|.\|_{1,p}$ norm by infinitely differentiable functions with compact support ([43, Thm. 2.3.2]).

For $\mathcal{Q}^m = \{Q_i\}_{i=1}^{m^d}$ a uniform resolution-m partition as defined in Sub-section 2.1, define the resolution-m block density approximation $\phi(\boldsymbol{x}) = \sum_{i=1}^{m^d} \phi_i 1_{Q_i}(\boldsymbol{x})$ of f, where $\phi_i = m^d \int_{Q_i} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$. The following lemma establishes how close (in $\mathcal{L}_1(\mathbb{R}^d)$ sense) these resolution-m block densities approximate functions in $W^{1,p}(\mathbb{R}^d)$.

Lemma 2 *For $1 \leq p < \infty$, let $f \in W^{1,p}(\mathbb{R}^d)$ have support $\mathcal{S} \in [0,1]^d$. Then there exists a constant $C_6 > 0$, independent of m, such that*

$$\int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})|\mathrm{d}\boldsymbol{x} \leq C_6 m^{-\lambda(p)}(\|\mathrm{D}f\|_p + o(1)), \tag{20}$$

where $\lambda(p) = 1$, if $1 \leq p \leq d$ and $\lambda(p) = d + 1 - d/p$, if $d < p < \infty$.

A proof of this lemma is given in Appendix A.

We can now return to the problem of finding convergence rate bounds on quasi-additive Euclidean functionals for non-uniform density f.Let $\{\tilde{\boldsymbol{X}}_i\}_{i=1}^n$ be i.i.d. random vectors having marginal Lebesgue density equal to the block density approximation ϕ. By the triangle inequality,

$$\begin{aligned}
&\left| E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)]/n^{\frac{d-\gamma}{d}} - \beta_{L_\gamma,d} \int_S f^{\frac{d-\gamma}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right| \\
&\quad \leq \left| E[L_\gamma(\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_n)]/n^{\frac{d-\gamma}{d}} - \beta_{L_\gamma,d} \int_S \phi^{\frac{d-\gamma}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right| + \beta_{L_\gamma,d} \left| \int_S \phi^{\frac{d-\gamma}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_S f^{\frac{d-\gamma}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right| \\
&\quad + \left| E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)] - E[L_\gamma(\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_n)] \right| / n^{\frac{d-\gamma}{d}} = I + II + III
\end{aligned} \tag{21}$$

Term I can be bounded by Proposition 1. To bound II, consider the following elementary inequality, which holds for $a, b \geq 0, 0 \leq \gamma \leq d$,

$$\left| a^{(d-\gamma)/d} - b^{(d-\gamma)/d} \right| \leq |a - b|^{(d-\gamma)/d},$$

and therefore, by Lemma 2 and Jensen's inequality,

$$II \leq \beta_{L_\gamma,d} \int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})|^{\frac{d-\gamma}{d}}\mathrm{d}\boldsymbol{x} \leq \beta_{L_\gamma,d}\, C_6' m^{-\lambda(p)(d-\gamma)/d} \left(\|\mathrm{D}f\|_p^{(d-\gamma)/d} + o(1) \right), \tag{22}$$

where $C_6' = C_6^{(d-\gamma)/d}$.

The following Proposition establishes an upper bound on term III in (21):

Proposition 2 *Let $d \geq 2$ and $1 \leq \gamma \leq d$. Assume $\{\boldsymbol{X}_i\}_{i=1}^n$ are i.i.d. random vectors over $[0,1]^d$ with density $f \in W^{1,p}(\mathbb{R}^d)$, $1 \leq p < \infty$, having support $\mathcal{S} \subset [0,1]^d$. Let $\{\tilde{\boldsymbol{X}}_i\}_{i=1}^n$ be i.i.d. random vectors with marginal Lebesgue*

density ϕ, the resolution-m block density approximation of f. Then, for any continuous quasi-additive Euclidean functional L_γ of order γ

$$\left|E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)] - E[L_\gamma(\tilde{\boldsymbol{X}}_1,\ldots,\tilde{\boldsymbol{X}}_n)]\right| / n^{\frac{d-\gamma}{d}} \leq C_3'\, C_6' m^{-\lambda(p)(d-\gamma)/d} \left(\|\mathrm{D}f\|_p^{(d-\gamma)/d} + o(1)\right), \tag{23}$$

where $\lambda(p)$ is defined in Lemma 2 and $C_3' = 2^{(2d-\gamma)/d} C_3$.

Proof:

As (21) we denote the left hand side of (23) by III. First invoke continuity (5) of L_γ

$$n^{(d-\gamma)/d} III \leq 2C_3 E\left[\text{card}\left(\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n\} \triangle \{\tilde{\boldsymbol{X}}_1,\ldots,\tilde{\boldsymbol{X}}_n\}\right)^{(d-\gamma)/d}\right].$$

To bound the right hand side of the above inequality we use an argument which is discussed and proved in ([38], Theorem 3). There it is shown that if ϕ approximates f in the $\mathcal{L}_1(\mathbb{R}^d)$ sense:

$$\int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \leq \varepsilon,$$

then, by standard coupling arguments, there exists a joint distribution P for the pair of random vectors $(\boldsymbol{X}, \tilde{\boldsymbol{X}})$ such that $P\{\boldsymbol{X} \neq \tilde{\boldsymbol{X}}\} \leq \varepsilon$. It then follows by Lemma 2 and the set inequality $\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n\} \triangle \{\tilde{\boldsymbol{X}}_1,\ldots,\tilde{\boldsymbol{X}}_n\} \subseteq \cup_{i=1}^n \{\boldsymbol{X}_i\} \triangle \{\tilde{\boldsymbol{X}}_i\}$ that

$$\begin{aligned} III &\leq 2C_3 E\left[\text{card}\left(\cup_{i=1}^n \{\boldsymbol{X}_i\} \triangle \{\tilde{\boldsymbol{X}}_i\}\right)^{(d-\gamma)/d}\right] / n^{(d-\gamma)/d} \\ &\leq 2C_3 E\left[\left(2\sum_{i=1}^n 1_{\{\boldsymbol{X}_i \neq \tilde{\boldsymbol{X}}_i\}}\right)^{(d-\gamma)/d}\right] / n^{(d-\gamma)/d} \\ &\leq 2C_3 (2nP\{\boldsymbol{X}_1 \neq \tilde{\boldsymbol{X}}_1\})^{(d-\gamma)/d} / n^{(d-\gamma)/d} \leq 2^{(2d-\gamma)/d} C_3 \varepsilon^{(d-\gamma)/d}, \end{aligned}$$

where the second inequality follows from the fact $\text{card}\left(\{\boldsymbol{X}_i\} \triangle \{\tilde{\boldsymbol{X}}_i\}\right) \in \{0, 2\}$. Finally, by Lemma 2 we can make ε as small as $C_6 m^{-\lambda(p)} \left(\|\mathrm{D}f\|_p + o(1)\right)$ and still ensure that ϕ be a block density approximation to f of resolution m. □

We can now substitute bounds (19), (22) and (23) in inequality (21) to obtain

$$\begin{aligned} &\left|E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d} \int_{\mathcal{S}} f(\boldsymbol{x})^{(d-\gamma)/d} \mathrm{d}\boldsymbol{x}\right| \\ &\quad \leq \frac{K_1 + C_4}{(nm^{-d})^{1/d}} \left(\int_{\mathcal{S}} f^{\frac{d-1-\gamma}{d}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + o(1)\right) + \frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}} \left(\int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + o(1)\right) \\ &\quad + \frac{K_2}{(nm^{-d})^{(d-\gamma)/d}} + \frac{1}{n^{(d-\gamma)/d}} + (\beta_{L_\gamma,d} + C_3')\, C_6' m^{-\lambda(p)(d-\gamma)/d} \left(\|\mathrm{D}f\|_p^{(d-\gamma)/d} + o(1)\right) \end{aligned} \tag{24}$$

This bound is finite under the assumptions that $f \in W^{1,p}(\mathbb{R}^d)$ with support in $\mathcal{S} \subset [0,1]^d$ and that $f^{\frac{1}{2}-\frac{\gamma}{d}}$ is integrable over $\mathcal{S}$.

The bound (24) is actually a family of bounds for different values of $m = 1, 2, \ldots$. By selecting m as the function of n that minimizes this bound, we obtain the tightest bound among them:

Proposition 3 *Let $d \geq 2$ and $1 \leq \gamma \leq d-1$. Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. random vectors over $[0,1]^d$ with density $f \in W^{1,p}(\mathbb{R}^d)$, $1 \leq p < \infty$, having support $\mathcal{S} \subset [0,1]^d$. Assume also that $f^{\frac{1}{2}-\frac{\gamma}{d}}$ is integrable over $\mathcal{S}$. Then, for any continuous quasi-additive Euclidean functional L_γ of order γ that satisfies the add-one bound (8)*

$$\left| E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d} \int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \leq O\left(n^{-r_1(d,\gamma,p)}\right),$$

where

$$r_1(d,\gamma,p) = \frac{\alpha\lambda(p)}{\alpha\lambda(p)+1} \, \frac{1}{d}$$

where $\alpha = \frac{d-\gamma}{d}$ and $\lambda(p)$ is defined in Lemma 2.

Proof: Without loss of generality assume that $nm^{-d} > 1$. In the range $d \geq 2$ and $1 \leq \gamma \leq d-1$, the slowest of the rates in (24) are $(nm^{-d})^{-1/d}$ and $m^{-\lambda(p)(d-\gamma)/d}$. We obtain an m-independent bound by selecting $m = m(n)$ to be the sequence increasing in n which minimizes the maximum of these rates

$$m(n) = \arg\min_m \max\left\{ (nm^{-d})^{-1/d}, m^{-\lambda(p)(d-\gamma)/d} \right\}.$$

The solution $m = m(n)$ occurs when $(nm^{-d})^{-1/d} = m^{-\lambda(p)(d-\gamma)/d}$, or $m = n^{1/[d(\alpha\lambda(p)+1)]}$ (integer part) and, correspondingly, $m^{-\lambda(p)(d-\gamma)/d} = n^{-\frac{\alpha\lambda(p)}{\alpha\lambda(p)+1}\frac{1}{d}}$. This establishes Proposition 3. □

3.3 Concentration Bounds

Any Euclidean functional L_γ of order γ satisfying the continuity property (4) also satisfies the concentration inequality [41, Thm. 6.3] established by Rhee [36]:

$$P\left(|L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) - E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)]| > t\right) \leq C \exp\left(\frac{-(t/C_3)^{2d/(d-\gamma)}}{Cn}\right), \tag{25}$$

where C is a constant depending only on the functional L_γ and d. It is readily verified that if $K > C_3 C^{(d-\gamma)/(2d)}$ the right hand side of (25) is summable over $n = 1, 2, \ldots$ when t is replaced by $K(n \ln n)^{(d-\gamma)/(2d)}$. Thus we have by

Borel-Cantelli

$$|L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n) - E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]| \leq O\left((n\ln n)^{(d-\gamma)/(2d)}\right) \quad (a.s.).$$

Therefore, combining this with Proposition 3 we obtain the a.s. bound

Proposition 4 *Let $d \geq 2$ and $1 \leq \gamma \leq d-1$. Assume $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ are i.i.d. random vectors over $[0,1]^d$ with density $f \in W^{1,p}(\mathbb{R}^d)$, $1 \leq p < \infty$, having support $\mathcal{S} \subset [0,1]^d$. Assume also that $f^{\frac{1}{2}-\frac{\gamma}{d}}$ is integrable over $\mathcal{S}$. Then, for any continuous quasi-additive Euclidean functional L_γ of order γ that satisfies the add-one bound (8)*

$$\left|L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| \leq O\left(\max\left\{\left(\frac{\ln n}{n}\right)^{(d-\gamma)/(2d)},\ n^{-r_1(d,\gamma,p)}\right\}\right) \quad (a.s.),$$

where $r_1(d,\gamma,p)$ is defined in Proposition 3.

The concentration inequality can also be used to bound the $\mathcal{L}_\kappa$ moments $E[|L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n) - E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]|^\kappa]^{1/\kappa}$, $\kappa = 1,2,\ldots$. In particular, as for any r.v. Z: $E[|Z|] = \int_0^\infty P(|Z| > t)dt$, we have by (25)

$$\begin{aligned} E\left[|L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n) - E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]|^\kappa\right] &= \int_0^\infty P\left(|L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n) - E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]| > t^{1/\kappa}\right)\mathrm{d}t \\ &\leq C_3 C\int_0^\infty \exp\left(\frac{-t^{2d/[\kappa(d-\gamma)]}}{Cn}\right)\mathrm{d}t \\ &= A_\kappa n^{\kappa(d-\gamma)/(2d)}, \end{aligned} \tag{26}$$

where $A_\kappa = C_3 C^{\kappa(d-\gamma)/(2d)+1}\int_0^\infty e^{-u^{2d/[\kappa(d-\gamma)]}}du$.

Combining the above with (24), we obtain

Proposition 5 *Let $d \geq 2$ and $1 \leq \gamma \leq d-1$. Assume $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ are i.i.d. random vectors over $[0,1]^d$ with density $f \in W^{1,p}(\mathbb{R}^d)$, $1 \leq p < \infty$, having support $\mathcal{S} \subset [0,1]^d$. Assume also that $f^{\frac{1}{2}-\frac{\gamma}{d}}$ is integrable over $\mathcal{S}$. Then, for any continuous quasi-additive Euclidean functional L_γ of order γ that satisfies the add-one bound (8)*

$$\begin{aligned} &E\left[\left|L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|^\kappa\right]^{1/\kappa} \\ &\quad\leq \frac{K_1+C_4}{(nm^{-d})^{1/d}}\left(\int_{\mathcal{S}} f^{\frac{d-1-\gamma}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + o(1)\right) + \frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}}\left(\int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + o(1)\right) \\ &\quad+ \frac{K_2}{(nm^{-d})^{(d-\gamma)/d}} + \frac{1}{n^{(d-\gamma)/d}} + (\beta_{L_\gamma,d}+C_3')C_6'\, m^{-\lambda(p)(d-\gamma)/d}\left(\|\mathrm{D}f\|_p^{(d-\gamma)/d} + o(1)\right) \\ &\quad+ A_\kappa^{1/\kappa} n^{-(d-\gamma)/(2d)} \end{aligned} \tag{27}$$

Proof:

For any non-random constant μ: $E[|W+\mu|^\kappa]^{1/\kappa} \leq E[|W|^\kappa]^{1/\kappa} + |\mu|$. Identify

$$\begin{aligned} \mu &= E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ W &= (L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n) - E[L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)])/n^{(d-\gamma)/d} \end{aligned}$$

and use (26) and (24) to establish Proposition 5. □

As the m-dependence of the bound of Proposition 5 is identical to that of the bias bound (24), minimization of the bound over $m = m(n)$ proceeds analogously to the proof of Proposition 3 and we obtain the following.

Corollary 1 *Let $d \geq 2$ and $1 \leq \gamma \leq d-1$. Assume $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ are i.i.d. random vectors over $[0,1]^d$ with density $f \in W^{1,p}(\mathbb{R}^d)$, $1 \leq p < \infty$, having support $\mathcal{S} \subset [0,1]^d$. Assume also that $f^{\frac{1}{2}-\frac{\gamma}{d}}$ is integrable over $\mathcal{S}$. Then, for any continuous quasi-additive Euclidean functional L_γ of order γ that satisfies the add-one bound (8)*

$$E\left[\left|L_\gamma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|^\kappa\right]^{1/\kappa} \leq O\left(n^{-r_1(d,\gamma,p)}\right), \tag{28}$$

where $r_1(d,\gamma,p)$ is defined in Proposition 3.

3.4 Discussion

It will be convenient to separate the discussion into the following points.

1. The bounds of Proposition 4 and Corollary 1 hold uniformly over the class of Lebesgue densities $f \in W^{1,p}(\mathbb{R}^d)$ with $\|\mathrm{D}f\|_p \leq C$ and integrable $f^{(d-\gamma)/d-1/2}$. If $\alpha = (d-\gamma)/d \in [1/2,(d-1)/d]$ then, as the support $\mathcal{S} \subset [0,1]^d$ is bounded, this integrability condition is automatically satisfied. To extend Proposition 4 and Corollary 1 to the range $\alpha \in ((d-1)/d,1)$ would require extension of the fundamental a.s. convergence rate bound of $O\left(n^{-1/d}\right)$ used in (10), established by Redmond and Yukich [33], to the case $0 < \gamma < 1$.

2. It can be shown in analogous manner to the proof of the umbrella theorems of [41, Ch. 7] that if f is not a Lebesgue density then the convergence rates in Propositions 4 and 5 hold when the region of integration $\mathcal{S}$ is replaced by the support of the Lebesgue continuous component of f.

3. The convergence rate bound satisfies $r_1(d,\gamma,p) < 1/d$, which corresponds to Redmond and Yukich's rate bound for the uniform density over $[0,1]^d$ [41, Thm. 5.2]. Thus, the bound predicts slower worst case convergence rates for non-uniform densities. However, as $p \to \infty$, the class of $f \in W^{1,p}(\mathbb{R}^d)$ becomes increasingly smooth and $r_1(d,\gamma,p) \to \frac{\alpha(d+1)}{\alpha(d+1)+1}\,\frac{1}{d}$, which for large d is very close to the $1/d$ rate bound.

4. When f is piecewise constant over a known partition of resolution $m = m_o$ faster rate of convergence bounds are available. For example, in Proposition 1 the bound in (19) is monotone increasing in m. Therefore the sequence $m(n) = m_o$ minimizes the bound as $n \to \infty$ and, proceeding in the same way as in the proof of Proposition 5, the best rate bound is of order $\max\left\{n^{-(d-\gamma)/(2d)},\ n^{-1/d}\right\}$. As the $O(n^{-1/d})$ bound on mean rate of convergence is tight [41, Sec. 5.3] for $d = 2$ and uniform density f, it is concluded that for $\alpha = (d-\gamma)/d \geq 2/d$ the asymptotic rate of convergence of the left hand side of (28) is exactly $O(n^{-1/d})$ for piecewise constant f and $d = 2$.

5. For $\alpha = (d-\gamma)) \geq 2/d$, it can be shown that the rate bound of Proposition 1 remains valid even if L_γ does not satisfy the "add-one bound." Thus, with $\alpha \geq 2/d$, Corollary 1 extends to any continuous quasi-additive functional L_γ including, in addition to the MST, the TSP, the minimal matching graph and the k-nearest neighbor graph functionals. As for the case $\alpha < 2/d$, we can use a weaker rate of mean convergence bound [41, Thm. 5.1], which applies to all continuous quasi-additive functionals and uniform f, in place of (10) in the proof of Proposition 1 to obtain

$$\left| E[L_\gamma(\boldsymbol{X_1},\ldots,\boldsymbol{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right| \ \leq\ O\left(n^{-\frac{\alpha}{d/\lambda(p)+2}}\right). \tag{29}$$

6. A tighter upper bound than Corollary 5 on the $\mathcal{L}_\kappa$-norm convergence rate may be derived if a better m-dependent analog to the concentration inequality (25) can be found.

4 Convergence Rates for Fixed Partition Approximations

Partitioning approximations to minimal graphs have been proposed by many authors, including Karp [20], Ravi *etal* [32], Mitchell [26], and Arora [1], as ways to reduce computational complexity. The fixed partition approximation is a simple example whose convergence rate has been studied by Karp [20, 21], Karp and Steele [22] and Yukich [41] in the context of a uniform density f.

Fixed partition approximations to a minimal graph weight function require specification of an integer resolution parameter m controlling the number of cells in the uniform partition $\mathcal{Q}^m = \{Q_i\}_{i=1}^m$ of $[0,1]^d$ discussed in Section 2. When m is defined as an increasing function of n we obtain a progressive-resolution approximation to $L_\gamma(\mathcal{X}_n)$. This approximation involves constructing minimal graphs of order γ on each of the cells Q_i, $i = 1, \ldots, m^d$, and the approximation $L_\gamma^m(\mathcal{X}_n)$ is defined as the sum of their weights plus a constant bias correction $b(m)$

$$L_\gamma^m(\mathcal{X}_n) = \sum_{i=1}^{m^d} L_\gamma(\mathcal{X}_n \cap Q_i) + b(m), \tag{30}$$

where $b(m)$ is $O\left(m^{d-\gamma}\right)$. In this section we specify a bound on the $\mathcal{L}_\kappa$-norm convergence rate of the progressive-resolution approximation (30) and specify the optimal resolution sequence $\{m(n)\}_{n>0}$ which minimizes this bound. Our derivations are based on the approach of Yukich [41, Sec. 5.4] and rely on the concrete version of the pointwise closeness bound (7)

$$\left|L_\gamma(F) - L_\gamma^*(F)\right| \leq \begin{cases} C[\mathrm{card}(F)]^{(d-\gamma-1)/(d-1)}, & 1 \leq \gamma < d-1 \\ C \log \mathrm{card}(F), & \gamma = d-1 \neq 1 \\ C, & d-1 < \gamma < d \end{cases}, \tag{31}$$

for any finite $F \subset [0,1]^d$. This condition is satisfied by the MST, TSP and minimal matching function [41, Lemma 3.7].

We first obtain a fixed-m bound on $\mathcal{L}_1$ deviation of $L_\gamma^m(\mathcal{X}_n)/n^{(d-\gamma)/d}$ from its a.s. limit.

Proposition 6 *Let $d \geq 2$ and $1 \leq \gamma < d-1$. Assume that the Lebesgue density $f \in W^{1,p}(\mathbb{R}^d)$, $1 \leq p < \infty$ has support $\mathcal{S} \subset [0,1]^d$. Assume also that $f^{1/2-\gamma/d}$ are integrable over $\mathcal{S}$. Let $L_\gamma^m(\mathcal{X}_n)$ be defined as in (30) where L_γ is a continuous quasi-additive functional of order γ which satisfies the pointwise closeness bound (31) and the add-one bound (8). Then if $b(m) = O(m^{d-\gamma})$*

$$\begin{aligned} E\left[\left|L_\gamma^m(\mathcal{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|\right] \\ \leq O\left(\max\left\{(nm^{-d})^{-\gamma/[d(d-1)]},\ m^{-\lambda(p)(d-\gamma)/d},\ n^{-(d-\gamma)/(2d)}\right\}\right), \end{aligned} \tag{32}$$

where $\lambda(p) = 1$ is defined in Lemma 2.

Proof:

Start with

$$E\left[\left|L_\gamma^m(\mathcal{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|\right] \quad \leq \tag{33}$$

$$E\left[\left|L_\gamma(\mathcal{X}_n)/n^{\frac{d-\gamma}{d}} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|\right] + E\left[\left|L_\gamma^m(\mathcal{X}_n) - L_\gamma(\mathcal{X}_n)\right|\right]/n^{\frac{d-\gamma}{d}}. \tag{34}$$

Analogously to the proof of [41, Thm. 5.7], using the pointwise closeness bound (31) one obtains a bound on the difference between the partitioned weight function $L_\gamma^m(F)$ and the minimal weight function $L_\gamma(F)$ for any finite $F \subset [0,1]^d$

$$b(m) - C_1 m^{d-\gamma} \leq L_\gamma^m(F) - L_\gamma(F) \leq m^{-\gamma} C \sum_{i=1}^{m^d} \left(\mathrm{card}(F \cap Q_i)\right)^{(d-\gamma-1)/(d-1)} + 1 + C_2 m^{d-\gamma} + b(m). \tag{35}$$

As usual let $\phi(\boldsymbol{x}) = \sum_{i=1}^{m^d} \phi_i m^{-d}$ be a block density approximation to $f(\boldsymbol{x})$. As $\{\mathcal{X}_n \cap Q_i\}_{i=1}^{m^d}$ are independent and $E[|Z|^u] \leq (E[|Z|])^u$ for $0 \leq u \leq 1$

$$\begin{aligned}
&E[\left|L_\gamma^m(\mathcal{X}_n) - L_\gamma(\mathcal{X}_n)\right|] \\
&\quad \leq m^{-\gamma} C \sum_{i=1}^{m^d} E\left[\left(\mathrm{card}(\mathcal{X}_n \cap Q_i)\right)^{(d-\gamma-1)/(d-1)}\right] + |b(m) - C_1 m^{d-\gamma}| + 1 + C_2 m^{d-\gamma} + b(m) \\
&\quad \leq m^{-\gamma} n^{(d-\gamma-1)/(d-1)} C \sum_{i=1}^{m^d} (\phi_i m^{-d})^{(d-\gamma-1)/(d-1)} + |b(m) - C_1 m^{d-\gamma}| + 1 + C_2 m^{d-\gamma} + b(m) \\
&\quad = m^{\gamma/(d-1)} n^{(d-\gamma-1)/(d-1)} C \sum_{i=1}^{m^d} \phi_i^{(d-\gamma-1)/(d-1)} m^{-d} + |b(m) - C_1 m^{d-\gamma}| + 1 + C_2 m^{d-\gamma} + b(m) \\
&\quad = m^{\gamma/(d-1)} n^{(d-\gamma-1)/(d-1)} C \int_{\mathcal{S}} \phi^{(d-\gamma-1)/(d-1)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + |b(m) - C_1 m^{d-\gamma}| + 1 + C_2 m^{d-\gamma} + b(m)
\end{aligned}$$

Note that the bias term $|b(m) - C_1 m^{d-\gamma}|$ can be eliminated by selecting $b(m) = C_1 m^{d-\gamma}$. Dividing through by $n^{(d-\gamma)/d}$, noting that $\left(|b(m) - C_1 m^{d-\gamma}| + C_2 m^{d-\gamma} + b(m)\right)/n^{(d-\gamma)/d} \leq B\left(nm^{-d}\right)^{-(d-\gamma)/d}$ for some constant B

$$E\left[\left|\frac{L_\gamma^m(\mathcal{X}_n) - L_\gamma(\mathcal{X}_n)}{n^{(d-\gamma)/d}}\right|\right] \quad \leq \quad (nm^{-d})^{-\gamma/[d(d-1)]} C \int_{\mathcal{S}} \phi^{(d-\gamma-1)/(d-1)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + (nm^{-d})^{-(d-\gamma)/d} B + n^{-(d-\gamma)/d}.$$

Combining this with Proposition 5 we can bound the right hand side of (34) to obtain

$$\begin{aligned}
&E\left[\left|L_\gamma^m(\mathcal{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|\right] \\
&\quad \leq \frac{K_1 + C_4}{(nm^{-d})^{1/d}}\left(\int_{\mathcal{S}} f^{\frac{d-1-\gamma}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + o(1)\right) + \frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}}\left(\int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + o(1)\right) \\
&\quad + \frac{K_2}{(nm^{-d})^{(d-\gamma)/d}} + \frac{2}{n^{(d-\gamma)/d}} + \left(\beta_{L_\gamma,d} + C_3'\right) C_6' m^{-\lambda(p)(d-\gamma)/d}\left(\|\mathrm{D}f\|_p^{(d-\gamma)/d} + o(1)\right) + A_1 n^{-(d-\gamma)/(2d)} \\
&\quad + \frac{C}{(nm^{-d})^{\gamma/[d(d-1)]}}\left(\int_{\mathcal{S}} f^{(d-\gamma-1)/(d-1)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + o(1)\right) + (nm^{-d})^{-(d-\gamma)/d} B.
\end{aligned} \tag{36}$$

Over the range $1 \leq \gamma < d-1$ the dominant terms are as given in the statement of Proposition 6. $\square$

Finally, by choosing $m = m(n)$ to minimize the maximum on the right hand side of the bound of Proposition 6 we have an analog to Corollary 1 for fixed partition approximations:

Corollary 2 *Let $d \geq 2$ and $1 \leq \gamma < d-1$. Assume that the Lebesgue density $f \in W^{1,p}(\mathbb{R}^d)$, $1 \leq p < \infty$ has support $\mathcal{S} \subset [0,1]^d$. Assume also that $f^{1/2-\gamma/d}$ is integrable over $\mathcal{S}$. Let $L_\gamma^m(\mathcal{X}_n)$ be defined as in (30) where L_γ is a continuous quasi-additive functional of order γ which satisfies the pointwise closeness bound (31) and the add-one bound (8). Then if $b(m) = O(m^{d-\gamma})$*

$$E\left[\left|L_\gamma^{m(n)}(\boldsymbol{X_1},\ldots,\boldsymbol{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|\right] \leq O\left(n^{-r_2(d,\gamma,p)}\right), \tag{37}$$

where

$$r_2(d,\gamma,p) = \frac{\alpha\lambda(p)}{\frac{d-1}{\gamma}\,\alpha\lambda(p)+1}\,\frac{1}{d},$$

where $\alpha = \frac{d-\gamma}{d}$ and $\lambda(p)$ is defined in Lemma 2. This rate is attained by choosing the progressive-resolution sequence $m = m(n) = n^{1/[d(\frac{d-1}{\gamma}\,\alpha\lambda(p)+1)]}$.

4.1 Discussion

We make the following remarks.

1. Under the assumed condition $\gamma < d-1$ in Corollary 2, $r_2(d,\gamma,p) \leq r_1(d,\gamma,p)$, where $r_1(d,\gamma,p)$ is defined in Corollary 1. Thus, as might be expected, the partitioned approximation has a $\mathcal{L}_\kappa$-norm convergence rate (37) that is always slower than the rate bound (28), and the slowdown increases as $(d-1)/\gamma$ increases.

2. In view of (36), up to a monotonic transformation, the rate constant multiplying the asymptotic rate $n^{-r_2(d,\gamma,p)}$ is an increasing function of $\int_{\mathcal{S}} f^{(d-\gamma-1)/(d-1)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, which is the Rényi entropy of f of order $(d-\gamma-1)/(d-1)$ (see (38) in the next section). Thus fastest convergence can be expected for densities with small Rényi entropy.

3. It is more tedious but straightforward to show that the $\mathcal{L}_2$ deviation $E\left[\left|L_\gamma^m(\mathcal{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|^2\right]^{1/2}$ obeys the identical asymptotic rate bounds as in Proposition 6 and Corollary 2 with identical bound minimizing progressive-resolution sequence $m = m(n)$.

4. As pointed out in the proof of Proposition 6 the bound minimizing choice of the bias correction $b(m)$ of the progressive-resolution approximation (30) is $b(m) = C_1 m^{d-\gamma}$, where C_1 is the constant in the subaddivity condition (2). However, Proposition 6 asserts that, for example, using $b(m) = Cm^{d-\gamma}$ with arbitrary scale constant C, or even using $b(m) = 0$, are asymptotically equivalent to the bound minimizing $b(m)$. This is important since the constant C_1 is frequently difficult to determine and depends on the specific properties of the minimal graph, which are different for the TSP, MST, etc.

5. The partitioned approximation (30) is a special case $k = n$ of the greedy approximation to the k-point minimal graph approximation introduced by Ravi *etal* [31] whose a.s. convergence was established by Hero and Michel [18] (Note that the overly strong BV condition assumed in [18] can be considerably weakened by replacing BV space with Sobolev space and applying Lemma 2 of this paper). Extension of Proposition 6 to greedy approximations to k-point graphs is an open problem.

5 Application to Entropy Estimation

In this section we apply the previous convergence results to non-parametric entropy estimation. In particular, using the convergence rate bounds derived above, Corollary 3 below establishes asymptotic performance advantages of the minimal graph estimator methods as contrasted to non-parametric density plug-in methods of entropy estimation. For concrete applications of Corollary 3 see Hero *etal* [15].

For a Lebesgue continuous multivariate density f the Rényi entropy of order α is defined as [35]:

$$H_\alpha(f) = (1-\alpha)^{-1} \ln \int f^\alpha(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{38}$$

To be consistent with previous sections of this paper, we restrict the support of f to a subset of $[0,1]^d$ and we only consider the range $\alpha \in (0,1)$. The Rényi entropy converges to the Shannon entropy $H_1(f) = -\int f(\boldsymbol{x}) \ln f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ in the limit as $\alpha \to 1$. As α becomes smaller the Rényi entropy tends to equalize the influence of the small amplitude regions, e.g. tails, and the large amplitude regions of f.

We treat entropy estimates of the form $\hat{H}_\alpha = (1-\alpha)^{-1} \ln \hat{I}_\alpha$, where $\hat{I}_\alpha$ is a consistent estimator of the integral

$$I(f^\alpha) = \int f^\alpha(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Given non-parametric function estimates $\widehat{f^\alpha}$ of f^α based on n i.i.d. observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from f, define the function plug-in estimator $I(\widehat{f^\alpha})$. Define the minimal-graph estimator $\hat{I}_\alpha = L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)/(\beta_{L_\gamma,d} n^\alpha)$, where $\gamma \in (0, d)$ is selected such that $\alpha = (d-\gamma)/d$ and L_γ is continuous quasi-additive. Let $\hat{H}_\alpha$ and $\hat{I}_\alpha$ denote estimates computed by either the plug-in or the minimal-graph estimators. A standard purturbation analysis of $\ln(z)$ establishes

$$|\hat{H}_\alpha - H_\alpha(f)| \quad = \quad \frac{1}{1-\alpha} \frac{|\hat{I}_\alpha - I(f^\alpha)|}{I(f^\alpha)} + o(|\hat{I}_\alpha - I(f^\alpha)|).$$

Thus as a function of n the asymptotic $\mathcal{L}_\kappa$-norm rate of convergence of $\hat{H}_\alpha - H_\alpha(f)$ will be identical to that of $\hat{I}_\alpha - I(f^\alpha)$.

Minimax rates of convergence for non-parametric function estimation have been extensively studied in the harmonic analysis literature. In these studies the Besov class $B^\sigma_{p,q}(\mathbb{R}^d)$ of smooth functions is widely adopted. Among other applications, Besov spaces play a central role in wavelets and wavelet shrinkage estimation [10]. The reader is referred to [23] for a general motivation on the use of Besov spaces. Before we can introduce the Besov class we need to define the following quantities. Define the finite difference of function f in the direction of $\boldsymbol{\tau} \in \mathbb{R}^d$ as:

$$\Delta^1_{\boldsymbol{\tau}} f(\boldsymbol{x}) \stackrel{\text{def}}{=} f(\boldsymbol{x}+\boldsymbol{\tau}) - f(\boldsymbol{x}) \, .$$

By induction, define the k-th order finite difference in the direction of $\boldsymbol{\tau}$ as

$$\Delta^k_{\boldsymbol{\tau}} f(\boldsymbol{x}) \stackrel{\text{def}}{=} \Delta^1_{\boldsymbol{\tau}} \Delta^{k-1}_{\boldsymbol{\tau}} f(\boldsymbol{x}) = \sum_{l=0}^{k} (-1)^{l+k} \binom{k}{l} f(\boldsymbol{x} + l\boldsymbol{\tau}) \, .$$

Following [30], let $1 \le p, q < \infty$, $\sigma > 0$ and k, ϱ be nonnegative integers satisfying the inequalities $k > \sigma - \varrho > 0$. One possible definition of the Besov space is the following: the function f belongs to the class $B^\sigma_{p,q}(\mathbb{R}^d)$ if $f \in \mathcal{L}_p(\mathbb{R}^d)$ and there exists partial weak derivatives $\mathrm{D}^{(\boldsymbol{s})} f = \partial^\varrho f / \partial x_1^{s_1} \ldots \partial x_d^{s_d}$ of order $\boldsymbol{s} = (s_1, \ldots, s_d)$ ($|\boldsymbol{s}| = s_1 + \ldots + s_d = \varrho$) such that the following seminorm is finite:

$$\|f\|_{b^\sigma_{p,q}} = \sum_{|\boldsymbol{s}|=\varrho} \left\{ \int_{\mathbb{R}^d} \left(\frac{\|\Delta^k_{\boldsymbol{\tau}} \mathrm{D}^{(\boldsymbol{s})} f\|_p}{|\boldsymbol{\tau}|^{\sigma-\varrho}} \right)^q \frac{\mathrm{d}\boldsymbol{\tau}}{|\boldsymbol{\tau}|^d} \right\} \, .$$

Finally, with minor abuse of notation, for scalar $C > 0$ define the Besov ball

$$B^\sigma_{p,q}(C) = \left\{ f \in B^\sigma_{p,q}(\mathbb{R}^d) : \|f\|_{\sigma,p,q} \stackrel{\text{def}}{=} \|f\|_p + \|f\|_{b^\sigma_{p,q}} \le C \right\}$$

Proposition 7 *Assume that the Lebesgue density f with support $\mathcal{S} \in [0,1]^d$ is such that $f \in B^\sigma_{p,q}(C)$. Then, for any plug-in estimator $I(\widehat{f^\alpha})$*

$$\sup_{f \in B^\sigma_{p,q}(C)} \left[E \left| I(\widehat{f^\alpha}) - I(f^\alpha) \right|^2 \right]^{1/2} \ge O\left(n^{-\sigma/(2\sigma+d)} \right) . \tag{39}$$

Proof:

The proof relies on well known results from non-parametric function estimation which we only sketch here. The reader is referred to [29] for more details.

For any estimator $\hat{g}_n$ of g based on i.i.d. samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ the minimax $\mathcal{L}_2$ integrated error over the Besov ball $B^{\sigma}_{p,q}(C)$ satisfies

$$\sup_{g \in B^{\sigma}_{p,q}(C)} \left[E \int (\hat{g}_n(\boldsymbol{x}) - g(\boldsymbol{x}))^2 \mathrm{d}\boldsymbol{x} \right]^{1/2} \geq O\left(n^{-\sigma/(2\sigma+d)} \right). \tag{40}$$

We show below that, for $\hat{g} = \widehat{f^{\alpha}}_n$ and $g = f^{\alpha}$, this implies

$$\sup_{g \in B^{\sigma}_{p,q}(C)} \left[E \left| \int (\hat{g}(\boldsymbol{x}) - g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right|^2 \right]^{1/2} \geq O\left(n^{-\sigma/(2\sigma+d)} \right). \tag{41}$$

The inequality (39) follows immediately from this.

Relation (40) implies that for all $g \in B^{\sigma}_{p,q}(C)$

$$\limsup_{n \to \infty} \left| [\hat{g}(\boldsymbol{x}) - g(\boldsymbol{x})] n^{\sigma/(2\sigma+d)} \right| < \infty, \quad (w.p.1), \tag{42}$$

except possibly on a subset of $[0,1]^d$ of measure zero, and for some $g \in B^{\sigma}_{p,q}(C)$

$$\liminf_{n \to \infty} \left| [\hat{g}(\boldsymbol{x}) - g(\boldsymbol{x})] n^{\sigma/(2\sigma+d)} \right| > 0, , \quad (w.p.1) \tag{43}$$

over some subset of $[0,1]^d$ of positive measure. Therefore, letting $g = f^{\alpha}$, using relations (42) and (43), there exist finite constants C_1 and C_2 such that

$$\left[E \left| \int (\hat{g}(\boldsymbol{x}) - g(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} \right|^2 \right]^{1/2} \leq C_1 n^{-\sigma/(2\sigma+d)} (1 + o(1)),$$

for all $g \in B^{\sigma}_{p,q}(C)$, and there exists a function $g \in B^{\sigma}_{p,q}(C)$ such that

$$\left[E \left| \int (\hat{g}(\boldsymbol{x}) - g(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} \right|^2 \right]^{1/2} \geq C_2 n^{-\sigma/(2\sigma+d)} (1 + o(1))$$

Therefore,

$$C_2 n^{-\sigma/(2\sigma+d)} (1 + o(1)) \leq \sup_{g \in B^{\sigma}_{p,q}(C)} \left[E \left| \int (\hat{g}(\boldsymbol{x}) - g(\boldsymbol{x})) \mathrm{d}\boldsymbol{x} \right|^2 \right]^{1/2} \leq C_1 n^{-\sigma/(2\sigma+d)} (1 + o(1))$$

which establishes (41) and the proof of Proposition 7 is completed. □

To compare the rate performance of minimal graph estimators to plug-in estimators, we will need the following result that relates functions in a Sobolev space to functions in a Besov class.

Lemma 3 *Let $p > d$ and let σ be a positive integer. Then $B^{\sigma}_{p,1}(\mathbb{R}^d) \subset W^{\sigma,p}(\mathbb{R}^d)$.*

Proof: See [4], theorems 18.8 and 18.9. □

Corollary 1 and Proposition 7 together with Lemma 3 provide a quantitative comparison between the worst case rates of convergence of both types of non-parametric estimators:

Corollary 3 *Let $d \geq 2$ and $\alpha \in [1/d, (d-1)/d]$. Assume that the Lebesgue density $f \in B^1_{p,1}(C)$, $p > d$, has support $\mathcal{S} \subset [0,1]^d$ and that $f^{\frac{1}{2}-\frac{\gamma}{d}}$ is integrable over $\mathcal{S}$. Then, for any plug-in estimator $I(\widehat{f^{\alpha}})$*

$$\sup_{f \in B^1_{p,1}(C)} \left[E \left| I(\widehat{f^{\alpha}}) - I(f^{\alpha}) \right|^2 \right]^{1/2} \geq O\left(n^{-\frac{1}{2+d}} \right), \tag{44}$$

while for the minimal-graph estimator $\hat{I}_{\alpha}$

$$\sup_{f \in B^1_{p,1}(C)} \left[E \left| \hat{I}_{\alpha} - I(f^{\alpha}) \right|^2 \right]^{1/2} \leq O\left(n^{-\frac{\alpha(d+1-d/p)}{\alpha(d+1-d/p)+1}\frac{1}{d}} \right). \tag{45}$$

We make several comments in connection with Corollary 3.

1. By equating the exponents in the rates bounds of Corollary 3, we find that for

$$\alpha \geq \frac{1}{2}\frac{d}{d+1-d/p}$$

the minimal graph estimators exhibit faster $\mathcal{L}_2$-norm convergence rates. In particular, we obtain that for f satisfying the conditions of the Corollary, the minimal graph-estimators always have faster r.m.s. convergence rate than the plug-in estimators for $\alpha \in [1/2, (d-1)/d]$.

2. The assumption $\alpha \leq (d-1)/d$ prevents the application of the convergence rate bound (28) in Proposition 7 to minimal graph estimates of the Shannon entropy, which would require $\alpha \to 1$. In particular, we cannot use it to bound a minimal-graph analog to the plug-in estimation method proposed by Mokkadem [27] in which Shannon

entropy is estimated by a sequence $\hat{I}(\hat{f}_n^{\alpha_n})$ of plug-in estimators where $\alpha_n < 1$ and $\lim_{n\to\infty} \alpha_n = 1$. As mentioned in Remark 1 of Section 3.4, relaxation of this assumption would require extending Redmond and Yukich's $O(n^{-1/d})$ convergence rates [33].

3. The partitioned minimal graph approximation (30) can be adapted to entropy estimation in an obvious way and an analog to Corollary 3 will hold with the right hand side of (45) replaced by the slower $O(n^{-r_2(d,\gamma,p)})$ rate bound.

4. If it is known *a priori* that the class of functions f is significantly smoother than the Besov class assumed in Corollary 3 then plug-in methods can have much faster convergence. As a rather extreme example, if f is a piecewise constant block density over an *a priori* known finite partition, a histogram plug-in estimator will have the faster r.m.s. convergence rate of $O(1/\sqrt{n})$ while the minimal graph estimator will only have $O(n^{-1/d})$. This dichotomy in entropy estimator convergence rates for smooth versus non-smooth density classes is analogous to well known behavior of minimax rates for non-parametric and semi-parametric estimation of general functionals, see work by Bickel and Ritov [5], Donoho and Low [9] and Birgé and Massart [6].

6 Conclusion

In this paper we have given rate of convergence bounds for length functionals of minimal-graphs satisfying continuous quasi-additivity. An application to entropy estimation was treated which established performance advantages of minimal graph estimators of entropy as contrasted with plug-in estimators. These results suggest that further exploration of minimal graphs for estimation of Rényi divergence, Rényi mutual information, and Rényi Jensen difference is justifed.

Future research should also include the extension of the rate of convergence bounds to smoother Sobolev space, i.e., densities with higher-order weak derivatives. This requires the derivation of new inequalities of the type stated by Lemma 2, which are similar to Sobolev- and Poincare- type inequalities. Also of interest is the extension of this work to densities with unbounded support.

Acknowledgment

The authors are grateful to Joseph Yukich for his close reading of an earlier version of this paper, in which he discovered an error, and to Andrew Nobel for his helpful suggestions on this work.

A Appendix

In this appendix we prove the approximation Lemma 2 that shows how close, in $\mathcal{L}_1(\mathbb{R}^d)$ sense, a function $f \in W^{1,p}(\mathbb{R}^d)$ can be approximated by its resolution-m block density. We follow a standard approach: first prove the inequalities for continuously differentiable functions and then extend the results to $W^{1,p}(\mathbb{R}^d)$ by using the fact that smooth functions are dense in $W^{1,p}(\mathbb{R}^d)$.

Proof of Lemma 2: First assume that $1 \leq p \leq d$ and that f is a continuously differentiable function. By the mean value theorem, there exist points $\boldsymbol{\xi_i} \in Q_i$ such that

$$\phi_i = m^d \int_{Q_i} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = f(\boldsymbol{\xi_i}) \, .$$

Also by the mean value theorem there exist points $\boldsymbol{\psi_i} \in Q_i$ such that

$$|f(\boldsymbol{x}) - f(\boldsymbol{\xi}_i)| = |\mathrm{D}f(\boldsymbol{\psi_i}) \cdot (\boldsymbol{x} - \boldsymbol{\xi}_i)|, \quad \boldsymbol{x} \in Q_i \, .$$

Note that, in what follows, $|.|$ means both the absolute value in $\mathbb{R}$ and any norm in $\mathbb{R}^d$. Using the above results, the Jensen inequality and the Cauchy-Schwarz inequality

$$\begin{aligned} \left(\int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \right)^p &\leq \int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} = \sum_{i=1}^{m^d} \int_{Q_i} |f(\boldsymbol{\xi_i}) - f(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} \\ &= \sum_{i=1}^{m^d} \int_{Q_i} |\mathrm{D}f(\boldsymbol{\psi_i}) \cdot (\boldsymbol{x} - \boldsymbol{\xi}_i)|^p \mathrm{d}\boldsymbol{x} \leq \sum_{i=1}^{m^d} |\mathrm{D}f(\boldsymbol{\psi_i})|^p \int_{Q_i} |\boldsymbol{x} - \boldsymbol{\xi}_i|^p \mathrm{d}\boldsymbol{x} \, . \end{aligned}$$

As $\boldsymbol{x}, \boldsymbol{\psi}_i \in Q_i$, a sub-cube with edge length m^{-1}: $\int_{Q_i} |\boldsymbol{x} - \boldsymbol{\xi}_i|^p \mathrm{d}\boldsymbol{x} = O(m^{-p-d})$. Thus, we have

$$\left(\int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \right)^p \leq C m^{-p} \sum_{i=1}^{m^d} |\mathrm{D}f(\boldsymbol{\psi_i})|^p m^{-d} \leq C m^{-p} \left(\int_{\mathcal{S}} |\mathrm{D}f(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x} + o(1) \right) \, .$$

Since smooth functions are dense in $W^{1,p}(\mathbb{R}^d)$ ([43, Thm. 2.3.2]), using the standard limiting argument the above inequality holds for $f \in W^{1,p}(\mathbb{R}^d)$. This establishes Lemma 2 for $1 \leq p \leq d$.

Now, let $d < p < \infty$ and assume f is a continuously differentiable function. This part of the proof closely follows the derivation of Morrey's inequality for Sobolev spaces [11].

Start with

$$f(\boldsymbol{x}) - \phi_i = f(\boldsymbol{x}) - f(\boldsymbol{\xi}_i) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} f(t\boldsymbol{x} + (1-t)\boldsymbol{\xi}_i) \mathrm{d}t = \int_0^1 \mathrm{D}f(t\boldsymbol{x} + (1-t)\boldsymbol{\xi}_i) \mathrm{d}t \cdot (\boldsymbol{x} - \boldsymbol{\xi}_i) \, .$$

By the Cauchy-Schwarz inequality

$$|f(\boldsymbol{x}) - \phi_i| \leq \int_0^1 |\mathrm{D}f(t\boldsymbol{x} + (1-t)\boldsymbol{\xi}_i)| \mathrm{d}t \; |\boldsymbol{x} - \boldsymbol{\xi}_i|.$$

Let $B(\boldsymbol{x}, r) \stackrel{\text{def}}{=} \{\boldsymbol{y} : |\boldsymbol{y} - \boldsymbol{x}| \leq r\}$ be the closed ball centered on $\boldsymbol{x}$ with radius r, and let $\partial B(\boldsymbol{x}, r) \stackrel{\text{def}}{=} \{\boldsymbol{y} : |\boldsymbol{y} - \boldsymbol{x}| = r\}$ be the boundary of $B(\boldsymbol{x}, r)$. Fix $0 < r \leq cm^{-1}$, for some $c > 0$, and integrate the above inequality with respect to $\boldsymbol{x}$ over $\partial B(\boldsymbol{\xi}_i, r)$:

$$\int_{\partial B(\boldsymbol{\xi}_i, r)} |f(\boldsymbol{x}) - \phi_i| \mathrm{d}S(\boldsymbol{x}) \leq \int_{\partial B(\boldsymbol{\xi}_i, r)} r \int_0^1 |\mathrm{D}f(t\boldsymbol{x} + (1-t)\boldsymbol{\xi}_i)| \mathrm{d}t \, \mathrm{d}S(\boldsymbol{x}),$$

where $\mathrm{d}S(\boldsymbol{x})$ denotes a differentiable surface element on $\partial B(\boldsymbol{x}, r)$. Now, make the following change of variables in the integral on the right hand side:

$$\boldsymbol{w} \stackrel{\text{def}}{=} t\boldsymbol{x} + (1-t)\boldsymbol{\xi}_i, \quad \tau \stackrel{\text{def}}{=} rt \,,$$

such that

$$|\boldsymbol{w} - \boldsymbol{\xi}_i| = t|\boldsymbol{x} - \boldsymbol{\xi}_i| = tr = \tau \,.$$

Then, by Tonelli's theorem,

$$\begin{aligned} \int_{\partial B(\boldsymbol{\xi}_i, r)} |f(\boldsymbol{x}) - \phi_i| \mathrm{d}S(\boldsymbol{x}) &\leq \int_0^r \int_{\partial B(\boldsymbol{\xi}_i, \tau)} r^{d-1} \frac{|\mathrm{D}f(\boldsymbol{w})|}{\tau^{d-1}} \mathrm{d}S(\boldsymbol{x}) \, \mathrm{d}\tau \\ &= \int_{B(\boldsymbol{\xi}_i, r)} r^{d-1} \frac{|\mathrm{D}f(\boldsymbol{w})|}{|\boldsymbol{w} - \boldsymbol{\xi}_i|^{d-1}} \mathrm{d}\boldsymbol{w} \,. \end{aligned}$$

Integrating with respect to r

$$\begin{aligned} \int_{B(\boldsymbol{\xi}_i, cm^{-1})} |f(\boldsymbol{x}) - \phi_i| \mathrm{d}\boldsymbol{x} &\leq \int_{B(\boldsymbol{\xi}_i, cm^{-1})} \frac{|\mathrm{D}f(\boldsymbol{w})|}{|\boldsymbol{w} - \boldsymbol{\xi}_i|^{d-1}} \mathrm{d}\boldsymbol{w} \int_0^{cm^{-1}} r^{d-1} \mathrm{d}r \\ &= Cm^{-d} \int_{B(\boldsymbol{\xi}_i, cm^{-1})} \frac{|\mathrm{D}f(\boldsymbol{w})|}{|\boldsymbol{w} - \boldsymbol{\xi}_i|^{d-1}} \mathrm{d}\boldsymbol{w} \,. \end{aligned}$$

Applying Holder's inequality to the right hand side and using the fact that $d < p < \infty$

$$\begin{aligned} &\int_{B(\boldsymbol{\xi}_i, cm^{-1})} |f(\boldsymbol{x}) - \phi_i| \mathrm{d}\boldsymbol{x} \\ &\quad \leq Cm^{-d} \left(\int_{B(\boldsymbol{\xi}_i, cm^{-1})} |\mathrm{D}f(\boldsymbol{w})|^p \mathrm{d}\boldsymbol{w} \right)^{1/p} \left(\int_{B(\boldsymbol{\xi}_i, cm^{-1})} \frac{1}{|\boldsymbol{w} - \boldsymbol{\xi}_i|^{(d-1)\frac{p}{p-1}}} \mathrm{d}\boldsymbol{w} \right)^{\frac{p-1}{p}} \\ &\quad = Cm^{-d} \left(m^{-d+(d-1)\frac{p}{p-1}} \right)^{\frac{p-1}{p}} \left(\int_{B(\boldsymbol{\xi}_i, cm^{-1})} |\mathrm{D}f(\boldsymbol{w})|^p \mathrm{d}\boldsymbol{w} \right)^{1/p} \\ &\quad = Cm^{-d-1+d/p} \left(\int_{B(\boldsymbol{\xi}_i, cm^{-1})} |\mathrm{D}f(\boldsymbol{w})|^p \mathrm{d}\boldsymbol{w} \right)^{1/p} \end{aligned}$$

Finally, we bound the $\mathcal{L}_1(\mathbb{R}^d)$ approximation error by using the above inequality, the elementary inequality $a^{1/p}+b^{1/p} \leq 2(a+b)^{1/p},\ a,b \geq 0$, and by choosing $c=\sqrt{d}$:

$$\begin{aligned}\int_S |f(\boldsymbol{x})-\phi(\boldsymbol{x})|\mathrm{d}\boldsymbol{x} &= \sum_{i=1}^{m^d}\int_{Q_i}|f(\boldsymbol{x})-\phi_i|\mathrm{d}\boldsymbol{x} \leq \sum_{i=1}^{m^d}\int_{B(\boldsymbol{\xi}_i,cm^{-1})}|f(\boldsymbol{x})-\phi_i|\mathrm{d}\boldsymbol{x} \\ &\leq Cm^{-d-1+d/p}\sum_{i=1}^{m^d}\left(\int_{B(\boldsymbol{\xi}_i,cm^{-1})}|\mathrm{D}f(\boldsymbol{w})|^p\mathrm{d}\boldsymbol{w}\right)^{1/p} \leq Cm^{-d-1+d/p}||\mathrm{D}f||_p\,.\end{aligned}$$

Note that C refers to different constants in the previous expressions. As before, extension to $f \in W^{1,p}(\mathbb{R}^d)$ follows by denseness of the smooth functions in $W^{1,p}(\mathbb{R}^d)$. □

References

[1] 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.

[2] 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.

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

[4] O. Besov, V. Il'in, and S. Nikol'skii, *Integral Representation of Functions and Imbedding Theorems*, volume II, John Wiley & Sons, 1979.

[5] P. Bickel and Y. Ritov, "Achieving information bounds in non and semiparametric estimation of non-linear functionals," *Annals of Statistics*, vol. 18, pp. 925–938, 1990.

[6] L. Birgé and P. Massart, "Estimation of integral functions of a density," *Annals of Statistics*, vol. 23, pp. 11–29, 1995.

[7] N. A. Cressie, *Statistics for spatial data*, Wiley, NY, 1993.

[8] M. T. Dickerson and D. Eppstein, "Algorithms for proximity problems in higher dimensions," *Comput. Geom. Theory and Appl.*, vol. 5, no. 5, pp. 277–291, 1996.

[9] D. L. Donoho, "Renormalization exponents and optimal pointwise rates of convergence," *Annals of Statistics*, vol. 20, pp. 944–970, 1992.

[10] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage: asymptopia?," *J. Royal Statistical Society, Ser. B*, vol. 57, pp. 301–309, 1995.

[11] L. C. Evans, *Partial Differential Equations*, Berkeley Mathematics Lecture Notes, 1994.

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

[13] A. Gersho and R. M. Gray, *Vector quantization and signal compression*, Kluwer, Boston MA, 1992.

[14] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*, Lecture Notes in Mathematics, Springer-Verlag, Berlin Heidelberg, 2000.

[15] A. O. Hero, B. Ma, O. Michel, and J. D. Gorman, "Alpha-divergence for classification, indexing and retrieval," Technical Report 328, Comm. and Sig. Proc. Lab. (CSPL), Dept. EECS, University of Michigan, Ann Arbor, May, 2001. `http://www.eecs.umich.edu/~hero/det_est.html`.

[16] A. Hero, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Processing Magazine*, To appear, Oct. 2002. `http://www.eecs.umich.edu/~hero/imag_proc.html`.

[17] A. Hero and O. Michel, "Estimation of Rényi information divergence via pruned minimal spanning trees," in *IEEE Workshop on Higher Order Statistics*, Caesaria, Israel, June 1999.

[18] A. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. on Inform. Theory*, vol. IT-45, no. 6, pp. 1921–1939, Sept. 1999.

[19] 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.

[20] R. M. Karp, "The probabilistic analysis of some combinatorial search algorithms," in *Algorithms and complexity: New directions and recent results*, J. F. Traub, editor, pp. 1–19, Academic Press, New York, 1976.

[21] R. M. Karp, "Probabilistic analysis of partitioning algorithms for the traveling salesman problem," *Oper. Res.*, vol. 2, pp. 209–224, 1977.

[22] R. M. Karp and J. M. Steele, "Probabilistic analysis of heuristics," in *The Traveling Salesman Problem: A guided tour of combinatorial optimization*, E. L. Lawler, J. K. Lenstra, A. H. G. R. Kan, and D. B. Shmoys, editors, pp. 181–206, Wiley, New York, 1985.

[23] G. Kerkyacharian and D. Picard, "Estimation de densité par méthod de noyau et d'ondelettes: les liens entre la géométrie du noyau et les constraintes de régularité," *C. R. Acad. Sci. Paris Sér. I Math.*, vol. 315, pp. 79–84, 1992.

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

[25] B. Ma, A. O. Hero, J. Gorman, and O. Michel, "Image registration with minimal spanning tree algorithm," in *IEEE Int. Conf. on Image Processing*, Vancouver, BC, October 2000.

[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] A. Mokkadem, "Estimation of the entropy and information of absolutely continuous random variables," *IEEE Trans. on Inform. Theory*, vol. IT-35, no. 1, pp. 193–196, 1989.

[28] 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.

[29] M. Neumann, "Multivariate wavelet thresholding: a remedy against the curse of dimensionality?," Technical Report Preprint 239, Weierstrass Inst. of Applied Analysis and Stochastics, Berlin, 1996. `citeseer.nj.nec.com/neumann96multivariate.html`.

[30] S. Nikol'skii, *Approximation of Functions of Several Variables and Imbedding Theorems*, Springer-Verlag, Berlin, Heidelberg, 1975.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

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

[37] B. D. Ripley, *Pattern recognition and neural networks*, Cambridge U. Press, 1996.

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

[39] 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.

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

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

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

[43] W. P. Ziemer, *Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation*, Graduate Texts in Mathematics, Springer-Verlag, New York, 1989.