Convergence rates of minimal graphs with random vertices

Alfred O. Hero, Bing Ma

Submitted to IEEE Trans. on Information Theory - Aug. 2001

Aug 2001

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 $\gamma \in (0,d)$ 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 p-th mean (L_p) convergence rates of this functional. In particular we show that over the space of multivariate densities such that $f^{(d-\gamma)/d}$ is of bounded variation, the L_p convergence rate is bounded above by $n^{-1/(d+1)}$ when $d/2 \leq \gamma \leq d-1$. 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, over the bounded variation class considered, the maximum L_p error of a minimal-graph estimator of Rényi entropy converges faster than that of any plug-in estimator.

Keywords: continuous quasi-additive functionals, combinatorial optimization, graph theory, minimax convergence rates, 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. Bing Ma bing@intervideo.com was with the Dept. of EECS at UM and is now with Intervideo, Inc., Fremont, CA. This research was supported in part by AFOSR grant F49620-97-0028.

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), nearest neighbor graph (NNG), minimal matching graph (MMG), traveling salesman problem (TSP), and their power-weighted variants. See the recent books by Steele [35] and Yukich [38] 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 [28, 29] when the vertices are uniformly distributed over $[0,1]^d$.

In the present paper we obtain bounds on a.s. and p-th mean (L_p) convergence rates of power-weighted Euclidean weight functionals of order γ for general Lebesgue densities f for which $f^{(d-\gamma)/d}$ is of bounded variation. Here the dimension d is greater than one and $\gamma \in (0,d)$ 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 4, we obtain a $O(n^{-1/(d+1)})$ bound on the L_p convergence rate when $1 \leq \gamma \leq d/2$. As contrasted with $O(n^{-1/d})$ rate bound for uniform f, shown by Redmond and Yukich, this slower L_p rate of convergence has a rate constant which depends on the underlying density, indicating that fastest convergence occurs when f has low Rényi entropy of order $(d-\gamma-1)/d$ and the total variation $f^{(d-\gamma)/d}$ is small.

We also obtain r.m.s. 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 [17], Ravi *etal* [26], and Hero and Michel [14], 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 5 we obtain bounds on 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. [12, 22, 14, 13], 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 convergence rates than plug-in estimators, such as those discussed by Bierlant *etal* [4], based on function estimation. Specifically: over the space of densities f such that f^α is of bounded variation the worst case L_p convergence rate of the minimal graph estimator of Rényi entropy of order α is upper bounded by $O(n^{-1/(d+1)})$ while any plug-in estimator has minimax rate lower bounded by $O(n^{-1/(d+2)})$. Beyond the signal processing applications mentioned above, which are treated in [12], these results may have important practical implications in other areas including: 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 [10, 25], and entropy characterization of time-frequency signal representations [37, 3].

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 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 $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ as $n \to \infty$ has been of great interest. The monographs by Steele [35] and Yukich [38] 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 [28]. 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 weights functional 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 $L_\gamma^{TSP}(\mathcal{X}_n) = \min_{e \in \mathcal{C}} \sum_e |e|^\gamma$. Construction of the TSP graph is NP-hard and arises in many different areas of operations research [21]. 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 [36]; clustering [39]; nonparametric regression [2] and testing for randomness [15]. 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 $L_\gamma^{k-NNG}(\mathcal{X}_n) = \sum_{i=1}^n \min_{e \in \mathcal{N}_{k,i}} \sum_e |e|^\gamma$. The k-NNG arises in computational geometry [8], clustering and pattern recognition [33], spatial statistics [7], and adaptive vector quantization [11], among other areas.

The following technical conditions on a Euclidean functional L_γ were defined in [28, 38].

- *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 $[0,1]^d$

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

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

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

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

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

  where card(G) is the cardinality of the subset 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) $L_\gamma(F) + 1 \ge L^*_\gamma(F)$ for every finite subset F; and, 2) for i.i.d. uniform random vectors $U_1, \ldots, U_n$ over $[0,1]^d$,

$$\left|E[L_\gamma(U_1, \ldots, U_n)] - E[L^*_\gamma(U_1, \ldots, U_n)]\right| \le C_4 n^{(d-\gamma-1)/d} \tag{5}$$

with C_4 a finite constant. The condition (5) is called the "close-in-mean approximation" in [38].

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

$$\left|L_\gamma(F) - L^*_\gamma(F)\right| \le o\left([\text{card}(F)]^{(d-\gamma)/d}\right), \tag{6}$$

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 [29, 28] 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 (5) of the dual. These authors have displayed duals and shown continuous quasi-additivity and related properties for weight functionals of power weighted MST, Steiner tree, TSP, k-NNG and others.

In [38, 28] almost sure limits with a convergence rate upper bound of $n^{1/d}$ were obtained for continuous quasi-additive Euclidean functionals $L_\gamma(X_1, \ldots, X_n)$ under the assumption of uniformly distributed points $X_1, \ldots, X_n$ and an additional assumption that L_γ satisfies the "add-one bound"

- *Add-one bound*:

$$| \; E[L_\gamma(U_1, \ldots, U_{n+1})] - E[L_\gamma(U_1, \ldots, U_n)] \; | \leq C_5 n^{-\gamma/d}. \tag{7}$$

The MST length functional of order γ satisfies the add-one bound. A slightly weaker bound on convergence rate also holds when L_γ is merely continuous quasi-additive [38, Ch. 5]. The $n^{-1/d}$ 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. Our method of extension follows standard practice [34, 35, 38]: we first establish pointwise convergence rates of the mean $E[L_\gamma(X_1, \ldots, X_n)]/n^{(d-\gamma)/d}$ for piecewise constant densities and then extend to arbitrary densities. Then we use Rhee's concentration inequality to obtain a.s. and L_p convergence rates of $L_\gamma(X_1, \ldots, X_n)/n^{(d-\gamma)/d}$.

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 $z(u) \stackrel{\text{def}}{=} g(u_o) + \frac{g(u_o)}{u_o}(u - u_o)I(u \leq u_o)$ is a lower bound on g, where $I(u \leq u_o)$ is the indicator function of $u \leq u_o$. Hence,

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

$\square$

A density $f(x)$ 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 $\sum_{i=1}^{m^d} \phi_i m^{-d} = 1$,

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

where $1_Q(x)$ 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 $X_1, \ldots, X_n$ 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*

$$\left| E[L_\gamma(X_1, \ldots, X_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d} \int_{\mathcal{S}} f^{(d-\gamma)/d}(x)\, dx \right| \leq O\left((nm^{-d})^{-1/d} \right).$$

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

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

Proof

Let n_i denote the number of samples $\{X_1, \ldots, X_n\}$ falling into the partition cell Q_i and let $\{U_i\}_i$ denote an i.i.d. sequence of uniform points on $[0,1]^d$. By subadditivity, we have

$$\begin{aligned} L_\gamma(X_1, \ldots, X_n) \qquad &\leq m^{-\gamma} \sum_{i=1}^{m^d} L_\gamma\left(m[\{X_1, \ldots, X_n\} \cap Q_i - q_i]\right) + C_1 m^{d-\gamma} \\ &= m^{-\gamma} \sum_{i=1}^{m^d} L_\gamma(U_1, \ldots, U_{n_i}) + C_1 m^{d-\gamma} \end{aligned}$$

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,

$$E[L_\gamma(X_1, \ldots, X_n)] \qquad \leq m^{-\gamma} \sum_{i=1}^{m^d} E\left[E\left[L_\gamma(U_1, \ldots, U_{n_i})|\, n_i\right]\right] + C_1 m^{d-\gamma}. \tag{8}$$

For uniform samples $U_1, \ldots, U_n$ in $[0,1]^d$, $n > 0$, the following rate of convergence for quasi-additive edge functionals L_γ satisfying the add-one bound (7) has been established for $1 \leq \gamma < d$ [38, The. 5.2],

$$|E[L_\gamma(U_1, \ldots, U_n)] - \beta_{L_\gamma, d} n^{\frac{d-\gamma}{d}}| \leq K_1 n^{\frac{d-1-\gamma}{d}}, \tag{9}$$

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

Using the result (9) and subadditivity (8) on L_γ, for $1 \leq \gamma < d$ we have

$$\begin{aligned} E[L_\gamma(X_1, \ldots, X_n)] \quad &\leq \quad m^{-\gamma} \sum_{i=1}^{m^d} E\left[\beta_{L_\gamma, d} n_i^{\frac{d-\gamma}{d}} + K_1 n_i^{\frac{d-\gamma-1}{d}}\right] + C_1 m^{d-\gamma} \\ &= m^{-\gamma} \beta_{L_\gamma, d} n^{\frac{d-\gamma}{d}} \sum_{i=1}^{m^d} E\left[\left(\frac{n_i}{n}\right)^{\frac{d-\gamma}{d}}\right] + \; m^{-\gamma} K_1 n^{\frac{d-\gamma-1}{d}} \sum_{i=1}^{m^d} E\left[\left(\frac{n_i}{n}\right)^{\frac{d-\gamma-1}{d}}\right] + C_1 m^{d-\gamma}. \end{aligned} \tag{10}$$

Similarly for the dual L_γ^* it follows by superadditivity

$$\begin{aligned} &E[L_\gamma^*(X_1, \ldots, X_n)] \\ &\quad \geq m^{-\gamma} \beta_{L_\gamma, d} n^{\frac{d-\gamma}{d}} \sum_{i=1}^{m^d} E\left[\left(\frac{n_i}{n}\right)^{\frac{d-\gamma}{d}}\right] - \; m^{-\gamma} K_1 n^{\frac{d-\gamma-1}{d}} \sum_{i=1}^{m^d} E\left[\left(\frac{n_i}{n}\right)^{\frac{d-\gamma-1}{d}}\right] - C_2 m^{d-\gamma} \end{aligned} \tag{11}$$

for $1 \leq \gamma < d$.

We next develop lower and upper bounds on the expected values in (10) and (11). As the function $g(u) = u^\nu$ is monotone and concave over the range $u \geq 0$ for $0 < \nu < 1$, from Lemma 1

$$\left(\frac{n_i}{n}\right)^\nu \quad \geq \quad p_i^\nu - p_i^{\nu-1}\left|\frac{n_i}{n} - p_i\right|, \tag{12}$$

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 (12),

$$E\left[\left(\frac{n_i}{n}\right)^\nu\right] \quad \geq \quad p_i^\nu - p_i^{\nu-\frac{1}{2}}/\sqrt{n}. \tag{13}$$

By concavity, Jensen's inequality yields the upper bound

$$E\left[\left(\frac{n_i}{n}\right)^\nu\right] \leq E\left[\left(\frac{n_i}{n}\right)\right]^\nu = p_i^\nu \tag{14}$$

Under the hypothesis $1 \leq \gamma \leq d-1$ this upper bound can be substituted into expression (10) to obtain

$$\begin{aligned} & E[L_\gamma(X_1,\ldots,X_n)/n^{(d-\gamma)/d}] \\ & \leq \beta_{L_\gamma,d}\sum_{i=1}^{m^d}\phi_i^{\frac{d-\gamma}{d}}m^{-d} + \frac{K_1}{(nm^{-d})^{1/d}}\sum_{i=1}^{m^d}\phi_i^{\frac{d-\gamma-1}{d}}m^{-d} + \frac{C_1}{(nm^{-d})^{(d-\gamma)/d}} \\ & = \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(x)dx + \frac{K_1}{(nm^{-d})^{1/d}}\int_{\mathcal{S}} f^{(d-\gamma-1)/d}(x)dx + \frac{C_1}{(nm^{-d})^{(d-\gamma)/d}}. \end{aligned} \tag{15}$$

Applying the bounds (14) and (13) to (11) we obtain an analogous lower bound for the mean of the dual functional L_γ^*

$$\begin{aligned} & E[L_\gamma^*(X_1,\ldots,X_n)]/n^{(d-\gamma)/d} \\ & \geq \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(x)dx - \frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}}\int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(x)dx \\ & \quad -\frac{K_1}{(nm^{-d})^{1/d}}\int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(x)dx - \frac{C_2}{(nm^{-d})^{(d-\gamma)/d}} \end{aligned} \tag{16}$$

By definition of the dual,

$$E[L_\gamma(X_1,\ldots,X_n)]/n^{\frac{d-\gamma}{d}} \geq E[L_\gamma^*(X_1,\ldots,X_n)]/n^{\frac{d-\gamma}{d}} - n^{-\frac{d-\gamma}{d}} \tag{17}$$

which when combined with (16) and (15) yields the result

$$\left|\frac{E[L_\gamma(X_1,\ldots,X_n)]}{n^{\frac{d-\gamma}{d}}} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}dx\right| \leq \frac{K_1}{(nm^{-d})^{1/d}}\int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(x)dx + \frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}}\int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(x)dx + \frac{K_2}{(nm^{-d})^{(d-\gamma)/d}} + n^{-\frac{d-\gamma}{d}}, \tag{18}$$

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

3.2 Mean Convergence Rate for Density Functions of BV

The total variation $V(Q)$ of a function g on $\mathbf{R}^d$ over a set $Q \subset \mathbf{R}^d$ is defined as [32]

$$V(Q) = \sup_{\{z_i\}\in Q} \sum_i |g(z_i) - g(z_{i-1})|, \tag{19}$$

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

Denote the total variation of f^ν over a subset A of $[0,1]^d$ as $V_\nu(A)$. For $\{Q_i\}_{i=1}^{m^d}$ a uniform resolution-m partition of $[0,1]^d$ into cubes Q_i of volume m^{-d} define the resolution-m block density approximation $\phi(x) = \sum_{i=1}^{m^d} \phi_i\, 1_{Q_i}(x)$ of f, where $\phi_i = m^d \int_{Q_i} f(x)dx$. The following elementary result was established in [14].

Lemma 2 *For $\nu \in [0,1]$ let f^ν be of bounded variation over $[0,1]^d$. Then*

$$\int |\phi^\nu(x) - f^\nu(x)|dx \leq m^{-d} V_\nu^m \tag{20}$$

where $V_\nu^m \stackrel{\text{def}}{=} \sum_{i=1}^{m^d} V_\nu(Q_i)$ is the total variation of f over the resolution-m partition.

Applying this Lemma, the triangle inequality and (18)

$$\begin{aligned}
&\left|E[L_\gamma(X_1,\ldots,X_n)]/n^{\frac{d-\gamma}{d}} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(x)dx\right| \qquad (21)\\
&\leq \left|E[L_\gamma(X_1,\ldots,X_n)]/n^{\frac{d-\gamma}{d}} - \beta_{L_\gamma,d}\int_{\mathcal{S}} \phi^{\frac{d-\gamma}{d}}(x)dx\right| + \beta_{L_\gamma,d}\left|\int_{\mathcal{S}} \phi^{\frac{d-\gamma}{d}}(x)dx - \int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(x)dx\right|\\
&\leq \frac{K_1}{(nm^{-d})^{1/d}}\int_{\mathcal{S}} \phi^{\frac{d-\gamma-1}{d}}(x)dx + \frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}}\int_{\mathcal{S}} \phi^{\frac{1}{2}-\frac{\gamma}{d}}(x)dx
\end{aligned}$$

$$+\frac{K_2}{(nm^{-d})^{(d-\gamma)/d}} + n^{-\frac{d-\gamma}{d}} + \beta_{L_\gamma,d} m^{-d} V^m_{(d-\gamma)/d}$$
$$\leq \frac{K_1}{(nm^{-d})^{1/d}} \int_{\mathcal{S}} \phi^{\frac{d-\gamma-1}{d}}(x)dx + \frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}} \int_{\mathcal{S}} \phi^{\frac{1}{2}-\frac{\gamma}{d}}(x)dx$$
$$+\frac{K_2}{(nm^{-d})^{(d-\gamma)/d}} + n^{-\frac{d-\gamma}{d}} + \beta_{L_\gamma,d} m^{-d} V^m_{(d-\gamma)/d}. \tag{22}$$

This bound is finite under the assumptions that $\gamma \leq d-1$, $f^{(d-\gamma)/d}$ is of bounded variation over $[0,1]^d$ and $f^{1/2-\gamma/d}$ is integrable over $\mathcal{S}$.

The bound (22) on the mean deviation (21) is actually a family of bounds for different values of $m = 1, 2, \ldots$. By selecting m as the function of n which minimizes this bound we obtain the tightest bound among them.

Proposition 2 *Let $d \geq 2$, $1 \leq \gamma \leq d-1$. Assume that f is a Lebesgue density and that $f^{(d-\gamma)/d}$ is of bounded variation over $[0,1]^d$. Assume also that $f^{1/2-\gamma/d}$ is integrable. Then for any continuous quasi-additive functional L_γ of order γ satisfying the add-one bound*

$$\left| E[L_\gamma(X_1, \ldots, X_n)]/n^{\frac{d-\gamma}{d}} - \beta_{L_\gamma,d} \int_{\mathcal{S}} f^{(d-\gamma)/d}(x)dx \right| \leq C_{d,\gamma}\, n^{-1/(d+1)} \left(1 + o(1)\right)$$

where

$$C_{d,\gamma} = \begin{cases} \beta_{L_\gamma,d} V^m_{(d-\gamma)/d} + K_1 \int_{\mathcal{S}} f^{(d-\gamma-1)/d}(x)dx + K_2 I_{d-1}(\gamma), & d > 2 \\ (K_1 + \beta_{L_\gamma,d})\ \int_{\mathcal{S}} f^{(d-\gamma-1)/d}(x)dx + \beta_{L_\gamma,d} V^m_{(d-\gamma)/d} + K_2 I_{d-1}(\gamma), & d = 2 \end{cases}$$

where $I_{d-1}(\gamma) = 1$ if $\gamma = d-1$ and $I_{d-1}(\gamma) = 0$ otherwise.

Proof

The rates depending on m in (22) are $(nm^{-d})^{-1/d}$, $(nm^{-d})^{-1/2}$, $(nm^{-d})^{-(d-\gamma)/d}$ and m^{-d}. Without any loss in generality we can assume that $nm^{-d} > 1$. Thus, in the range $d \geq 1/2$ and $1 \leq \gamma \leq d-1$, the slowest of these rates are $(nm^{-d})^{-1/d}$ and m^{-d}. We obtain an m-independent bound on the mean deviation (21) by selecting $m = m(n)$ to be the sequence increasing in n which minimizes the maximum of these rates

$$m(n) = \text{argmin}_m \max\left\{m^{-d},\ (nm^{-d})^{-1/d}\right\}. \tag{23}$$

The solution $m = m(n)$ occurs when $(nm^{-d})^{-1/d} = m^{-d}$, or $m = n^{1/[d(d+1)]}$ (integer part) and, correspondingly, $nm^{-d} = n^{d/(d+1)}$ and $m^{-d} = (nm^{-d})^{-1/d} = n^{-1/(d+1)}$. Therefore,

$$\left| E[L_\gamma(X_1, \ldots, X_n)]/n^{\frac{d-\gamma}{d}} - \beta_{L_\gamma,d} \int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(x)dx \right|$$

$$\leq \frac{1}{n^{1/(d+1)}}\left[K_1\int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(x)dx + \beta_{L_\gamma,d}V^m_{(d-\gamma)/d}\right]$$
$$+\frac{\beta_{L_\gamma,d}}{n^{d/[2(d+1)]}}\int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(x)dx + \frac{K_2}{n^{(d-\gamma)/(d+1)}} + n^{-\frac{d-\gamma}{d}}.$$

This establishes Proposition 2. □

3.3 Concentration Bounds

Any Euclidean functional L_γ of order γ satisfying the continuity property (4) also satisfies the concentration inequality [38, Thm. 6.3] established by Rhee [31]:

$$P\left(|L_\gamma(X_1,\ldots,X_n) - E[L_\gamma(X_1,\ldots,X_n)]| > t\right) \leq C\exp\left(\frac{-(t/C_3)^{2d/(d-\gamma)}}{Cn}\right), \tag{24}$$

where C is a constant depending only on L_γ and d. It is readily verified that if $K > C_3C^{(d-\gamma)/(2d)}$ the right hand side of (24) 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

$$|L_\gamma(X_1,\ldots,X_n) - E[L_\gamma(X_1,\ldots,X_n)]| \leq O\left((n\ln n)^{(d-\gamma)/(2d)}\right) \quad (a.s.).$$

Therefore, combining this with Proposition 2 we obtain the a.s. bound

Proposition 3 *Let $d \geq 2$ and $1 \leq \gamma \leq d-1$. Assume that the Lebesgue density f supported on $\mathcal{S} \subset [0,1]^d$ is such that $f^{(d-\gamma)/d}$ is of bounded variation over $[0,1]^d$ and $f^{1/2-\gamma/d}$ is integrable over $\mathcal{S}$. Then for L_γ a continuous quasi-additive functional of order γ which satisfies the add-one bound*

$$\left|L_\gamma(X_1,\ldots,X_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(x)dx\right| \leq O\left(\max\left\{\left(\frac{\ln n}{n}\right)^{(d-\gamma)/(2d)},\ n^{-1/(d+1)}\right\}\right) \quad (a.s.).$$

The concentration inequality can also be used to bound L_p moments $E[|L_\gamma(X_1,\ldots,X_n) - E[L_\gamma(X_1,\ldots,X_n)]|^\kappa]^{1/\kappa}$, $\kappa = 1, 2, \ldots$. In particular, as for any r.v. Z: $E[|Z|] = \int_0^\infty P(|Z| > t)dt$, we have by (24)

$$\begin{aligned}
E\left[|L_\gamma(X_1,\ldots,X_n) - E[L_\gamma(X_1,\ldots,X_n)]|^\kappa\right] &= \int_0^\infty P\left(|L_\gamma(X_1,\ldots,X_n) - E[L_\gamma(X_1,\ldots,X_n)]| > t^{1/\kappa}\right)dt \\
&\leq C_3C\int_0^\infty \exp\left(\frac{-t^{2d/[\kappa(d-\gamma)]}}{Cn}\right)dt \\
&= K_\kappa n^{\kappa(d-\gamma)/(2d)},
\end{aligned} \tag{25}$$

where $K_\kappa = C_3C^{\kappa(d-\gamma)/(2d)+1}\int_0^\infty e^{-u^{2d/[\kappa(d-\gamma)]}}du$.

Combining the above with (22), we obtain

Proposition 4 *Let $d \geq 2$, $1 \leq \gamma \leq d-1$, and $\kappa > 0$. Assume that the Lebesgue density f supported on $\mathcal{S}$ is such that $f^{(d-\gamma)/d}$ is of bounded variation over $[0,1]^d$ and $f^{1/2-\gamma/d}$ is integrable over $\mathcal{S}$. Then for L_γ a continuous quasi-additive functional of order γ which satisfies the add-one bound*

$$E\left[\left|L_\gamma(X_1,\ldots,X_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(x)dx\right|^\kappa\right]^{1/\kappa} \leq \tag{26}$$

$$\frac{K_1}{(nm^{-d})^{1/d}}\int_{\mathcal{S}}\phi^{\frac{d-\gamma-1}{d}}(x)dx + \frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}}\int_{\mathcal{S}}\phi^{\frac{1}{2}-\frac{\gamma}{d}}(x)dx$$
$$+\frac{K_2}{(nm^{-d})^{(d-\gamma)/d}} + n^{-\frac{d-\gamma}{d}} + \beta_{L_\gamma,d}m^{-d}V^m_{(d-\gamma)/d} + K_\kappa^{1/\kappa}n^{-(d-\gamma)/(2d)}$$

Proof:

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

$$\begin{aligned} \mu &= E[L_\gamma(X_1,\ldots,X_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(x)dx \\ W &= (L_\gamma(X_1,\ldots,X_n) - E[L_\gamma(X_1,\ldots,X_n)])/n^{(d-\gamma)/d} \end{aligned}$$

and use (25) and (22) to establish Proposition 4. □

As the m-dependence of the bound of Proposition 4 is identical to that of the bias bound (22), minimization of the bound over $m = m(n)$ proceeds analogously to the proof of Proposition 2 and we obtain the following.

Corollary 1 *Let $d \geq 2$, $1 \leq \gamma \leq d-1$, and $\kappa > 0$. Assume that the Lebesgue density f supported on $\mathcal{S}$ is such that $f^{(d-\gamma)/d}$ is of bounded variation over $[0,1]^d$ and $f^{1/2-\gamma)/d}$ is integrable over $\mathcal{S}$. Then for L_γ a continuous quasi-additive functional of order γ which satisfies the add-one bound*

$$E\left[\left|L_\gamma(X_1,\ldots,X_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(x)dx\right|^\kappa\right]^{1/\kappa} \leq O\left(\max\left\{n^{-1/(d+1)}, n^{-(d-\gamma)/(2d)}\right\}\right). \tag{27}$$

3.4 Discussion

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

1. The bounds of Proposition 3 and Corollary 1 hold uniformly over the class of Lebesgue densities having $f^{(d-\gamma)/d}$ of bounded variation and integrable $f^{(d-\gamma)/d-1/2}$ over $[0,1]^d$. If $(d-\gamma)/d \in [1/2,(d-1)/d]$ then, as $\mathcal{S} \in [0,1]^d$ is bounded, this integrability condition is automatically satisfied.

2. A property of bounded variation classes of functions over $\mathbf{R}^d$, denoted BV, is that $f^{\alpha_0} \in \mathrm{BV}$ implies $f^{\alpha_1} \in \mathrm{BV}$ when $\alpha_0 \leq \alpha_1$. Thus the integrability assumption in Propositions 3 and 4 can be eliminated by replacing the BV and integrability hypotheses on f by the stronger assumption that $f^{1/2-\gamma/d}$ is of bounded variation over $[0,1]^d$.

3. It can be shown in analogous manner to the proof of the umbrella theorems of [38, Ch. 7] that if f is not a Lebesgue density then the convergence rates in Propositions 3 and 4 hold when the region of integration $\mathcal{S}$ is replaced by the support of the Lebesgue continuous component of f.

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 the case of the L_p bound of Proposition 4 the cell variation $V^{m_o}_{(d-\gamma)/d}$ is zero for $m \geq m_o$ and therefore the bound (26) is monotone increasing in m over this range. Therefore the sequence $m(n) = m_o$ minimizes the bound as $n \to \infty$ and the best rate bound is only of order $\max\left\{n^{-(d-\gamma)/(2d)},\ n^{-1/d}\right\}$. As for uniform density f the $O(n^{-1/d})$ bound on mean rate of convergence is tight [38, Sec. 5.3] for $d = 2$, we conclude that for $(d-\gamma)/d \geq 2/d$ the asymptotic rate of convergence of the left hand side of (27) is exactly $O(n^{-1/d})$ for piecewise constant f and $d = 2$.

5. The moment bound (27) is of order $n^{-1/(d+1)}$ for $d \geq (\gamma + 1 + \sqrt{(\gamma+1)^2 + 4\gamma})/2$ and of order $n^{-(d-\gamma)/(2d)}$ otherwise. As

$$\gamma + 1 \leq (\gamma + 1 + \sqrt{(\gamma+1)^2 + 4\gamma})/2 \leq \gamma + 2,$$

for $d \gg 2$ we conclude that the convergence rate is of order $n^{-1/(d+1)}$ except for a narrow range of γ contained inside the interval $(d-2, d)$.

6. By using a weaker rate of mean convergence bound [38, Thm. 5.1], which applies to all continuous quasi-additive functionals and uniform f, in place of (9) in the proof of Proposition 2, the mean deviation of any continuous quasi-additive functional L_γ from its a.s. limit can easily be shown to obey the bound

$$\left| E[L_\gamma(X_1, \ldots, X_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d} \int_{\mathcal{S}} f^{(d-\gamma)/d}(x)dx \right| \ \leq\ O\left(\max\left\{n^{-1/(d+1)},\ n^{-(d-\gamma)/(2d)}\right\}\right). \quad (28)$$

Since the term $n^{-(d-\gamma)/(2d)}$ already appears in the L_p bound of Proposition 4, Corollary 1 extends to any continuous quasi-additive functional L_γ including the MST, TSP, the minimal matching graph and the k-nearest neighbor graph functionals.

7. A tighter upper bound than Corollary 4 on the L_p convergence rate may be derived if a tighter m dependent analog to the concentration inequality (24) can be found.

4 Convergence Rates for Fixed Partition Approximations

Partitioning approximations to minimal graphs have been proposed by many authors, including Karp [17], Ravi *etal* [27], Mitchell [23], 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 [17, 18], Karp and Steele [19] and Yukich [38] 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 $b(m)$

$$L_\gamma^m(\mathcal{X}_n) = \sum_{i=1}^{m^d} L_\gamma(\mathcal{X}_n \cap Q_i) + b(m). \tag{29}$$

In this section we specify a bound on the m.s. convergence rate of the progressive-resolution approximation (29) and specify the optimal resolution sequence $\{m(n)\}_{n>0}$ which minimizes this bound. Our derivations are based on the approach of Yukich [38, Sec. 5.4] and rely on the concrete version of the pointwise closeness bound (6)

$$\left|L_\gamma(F) - L_\gamma^*(F)\right| \leq \begin{cases} C[\mathrm{card}(F)]^{(d-\gamma-1)/(d-1)}, & 1 \leq \gamma < d-1 \\ C \log \mathrm{card}(F), & \gamma = d-1 \neq 1 \\ C, & d-1 < \gamma < d \end{cases}, \tag{30}$$

for any finite $F \subset [0,1]^d$. This condition is satisfied by the MST, TSP and minimal matching function [38, Lemma 3.7].

We first obtain a fixed m bound on L_1 deviation of the $L_\gamma^m(\mathcal{X}_n)/n^{(d-\gamma)/d}$ from its limit.

Proposition 5 *Let $d \geq 2$ and $1 \leq \gamma < d-1$. Assume that the Lebesgue density f supported on $\mathcal{S} \subset [0,1]^d$ satisfies the properties that $f^{(d-\gamma)/d}$ is of bounded variation over $[0,1]^d$ and $f^{1/2-\gamma/d}$ are integrable over $\mathcal{S}$. Let $L_\gamma^m(\mathcal{X}_n)$ be defined as in (29) where L_γ is a continuous quasi-additive functional of order γ which satisfies the pointwise closeness bound*

(30). Then if $|b(m) - C_1 m^{d-\gamma}| \leq O(m^{d-\gamma})$

$$E\left[\left|L_\gamma^m(\mathcal{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(x)dx\right|\right] \\ \leq O\left(\max\left\{(nm^{-d})^{-\gamma/[d(d-1)]},\ m^{-d},\ n^{-(d-\gamma)/(2d)}\right\}\right). \tag{31}$$

Proof:

Start with

$$E\left[\left|L_\gamma^m(\mathcal{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(x)dx\right|\right] \quad \leq \tag{32}$$

$$E\left[\left|L_\gamma(\mathcal{X}_n)/n^{\frac{d-\gamma}{d}} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(x)dx\right|\right] + E\left[\left|L_\gamma^m(\mathcal{X}_n) - L_\gamma(\mathcal{X}_n)\right|\right]/n^{\frac{d-\gamma}{d}}. \tag{33}$$

Analogously to the proof of [38, Thm. 5.7], using the pointwise closeness bound (30) 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 $F \subset [0,1]^d$

$$b(m) - C_1 m^{d-\gamma} \leq L_\gamma^m(F) - L_\gamma(F) \leq m^{-\gamma} C \sum_{i=1}^{m^d} (\mathrm{card}(F \cap Q_i))^{(d-\gamma-1)/(d-1)}. \tag{34}$$

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

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

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 $|b(m) - C_1 m^{d-\gamma}| n^{-(d-\gamma)/d} \leq B(nm^{-d})^{-(d-\gamma)/d}$ for some constant B

$$E\left[\left|\frac{L_\gamma^m(\mathcal{X}_n) - L_\gamma(\mathcal{X}_n)}{n^{(d-\gamma)/d}}\right|\right] \quad \leq \quad (nm^{-d})^{-\gamma/[d(d-1)]} C \int_{\mathcal{S}} \phi^{(d-\gamma-1)/(d-1)}(x)dx + (nm^{-d})^{-(d-\gamma)/d} B. \tag{35}$$

Combining this with Proposition 4, with relaxed add-one bound condition (see comment 6 in Section 3.4), we can bound the right hand side of (33) to obtain

$$\begin{aligned} E\left[\left|L_\gamma^m(\mathcal{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(x)dx\right|\right] \quad \le& \\ & \frac{K_1}{(nm^{-d})^{1/d}}\int_{\mathcal{S}} \phi^{\frac{d-\gamma-1}{d}}(x)dx + \frac{\beta_{L_\gamma,d}}{(nm^{-d})^{1/2}}\int_{\mathcal{S}} \phi^{\frac{1}{2}-\frac{\gamma}{d}}(x)dx \\ & + \frac{K_2}{(nm^{-d})^{(d-\gamma)/d}} + n^{-\frac{d-\gamma}{d}} + \beta_{L_\gamma,d} m^{-d} V^m_{(d-\gamma)/d} + K_\kappa^{1/\kappa} n^{-(d-\gamma)/(2d)} \\ & + \frac{C}{(nm^{-d})^{\gamma/[d(d-1)]}}\int_{\mathcal{S}} \phi^{(d-\gamma-1)/(d-1)}(x)dx + (nm^{-d})^{-(d-\gamma)/d} B. \end{aligned} \tag{36}$$

Over the range $1 \le \gamma < d-1$ the dominant terms are as given in the Proposition. This establishes Proposition 5. $\square$.

Finally, by choosing $m = m(n)$ to minimize the maximum on the right hand side of the bound of Proposition 5 we have an analog to Corollary 1:

Corollary 2 *Let $d \ge 2$ and $1 \le \gamma < d-1$. Assume that the Lebesgue density f supported on $\mathcal{S}$ is such that $f^{(d-\gamma)/d}$ is of bounded variation over $[0,1]^d$ and $f^{1/2-\gamma/d}$ is integrable over $\mathcal{S}$. Let $L_\gamma^m(\mathcal{X}_n)$ be defined as in (29) with L_γ a continuous quasi-additive functional of order γ which satisfies the pointwise close bound (30). Then if $|b(m) - C_1 m^{d-\gamma}| \le O(m^{d-\gamma})$*

$$\begin{aligned} E\left[\left|L_\gamma^{m(n)}(X_1,\ldots,X_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(x)dx\right|\right] \quad \le& \\ & O\left(\max\left\{n^{-\beta/(1+\beta)},\ n^{-(d-\gamma)/(2d)}\right\}\right), \end{aligned} \tag{37}$$

where $\beta = \gamma/[d(d-1)]$. This rate is attained by choosing the progressive-resolution sequence $m = m(n) = n^{\frac{1}{d}\beta/(1+\beta)}$.

4.1 Discussion

We make the following remarks.

1. The function $\beta/(1+\beta)$ in Corollary 2 is increasing in $\beta = \gamma/[d(d-1)]$ over $1 \le \gamma < d-1$ and takes supremum of $1/(d+1)$ at $\gamma = d-1$. The rate $n^{-(d-\gamma)/(2d)}$ dominates in (37) when $\gamma > \gamma_o = d(d-1)(-d/(d-1) + \sqrt{(d/(d-1))^2 + 4/(d-1)})/2$. Thus, as might be expected, the partitioned approximation has a convergence rate (37) that is always worse than the rate bound (27) but improves as γ increases to γ_o.

2. In view of (36) the rate constant multiplying the asymptotic rate $n^{-\gamma/[d(d-1)]/(\gamma/[d(d-1)]+1)}$ is an increasing functions $\int_{\mathcal{S}} f^{(d-\gamma-1)/(d-1)}(x)dx$. Thus fastest asymptotic convergence can be expected for densities with small Rényi entropy of order $(d-\gamma-1)/(d-1)$.

3. It is more tedious but straightforward to show that the L_2 deviation $E\left[\left|L_\gamma^m(\mathcal{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(x)dx\right|^2\right]^{1/2}$ obeys the identical asymptotic rate bounds as in Proposition 5 and Corollary 2 with identical bound minimizing progressive-resolution sequence $m = m(n)$.

4. As pointed out in the proof of Proposition 5 the bound minimizing choice of the bias correction $b(m)$ of the progressive-resolution approximation (29) is $b(m) = C_1 m^{d-\gamma}$, where C_1 is the constant in the subaddivity condition (2). However, Proposition 5 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 (29) is a special case $k = n$ of the greedy approximation to the k-point minimal graph approximation introduced by Ravi *etal* [26] whose a.s. convergence was established by Hero and Michel [14]. Extension of Proposition 5 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 the convergence rate bounds derived above, Proposition 6 establishes asymptotic performance advantages of the minimal graph method of Rényi entropy estimation as contrasted to non-parametric density plug-in methods of entropy estimation. For concrete applications of Proposition 6 see Hero *etal* [12].

For a Lebesgue continuous multivariate density f the Rényi entropy of order α is defined as [30]:

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

Here, as above, 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 $H_1(f) = -\int f(x)\ln f(x)dx$ 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 consider 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

$$I_\alpha(f) = \int f^\alpha(z)dz.$$

Given non-parametric function estimates $\widehat{f^\alpha}_n$ of f^α based on the n i.i.d. observations $X_1,\ldots,X_n$ from f, define the function plug-in estimator $I_\alpha(\widehat{f^\alpha})$. Define the minimal-graph estimator $\hat{I}_\alpha = L_\gamma(X_1,\ldots,X_n)/(\beta_{L_\gamma,d}n^\alpha)$, where $\gamma \in (0,d)$ is selected such that $\alpha = (d-\gamma)/d$ and L_γ is continuous quasi-additive. A standard purturbation analysis of $\ln(z)$ establishes that for either of these estimators $\hat{I}_\alpha$

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

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

Define $\mathrm{BV}(C,d)$ as the class of functions on $[0,1]^d$ of bounded variation having total variation C.

Proposition 6 *Assume that the Lebesgue density f on $[0,1]^d$ is such that $f^\alpha \in \mathrm{BV}(C,d)$ where $\alpha \in [1/2,(d-1)/d]$, $d \geq 2$. Then, for $p = 1,2,\ldots$, and any plug-in estimator $I_\alpha(\widehat{f^\alpha})$*

$$\sup_{f^\alpha \in \mathrm{BV}(C,d)} E^{1/p}\left[\left|I_\alpha(\widehat{f^\alpha}) - I_\alpha(f)\right|^p\right] \;\geq\; O\left(n^{-1/(d+2)}\right), \tag{39}$$

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

$$\sup_{f^\alpha \in \mathrm{BV}(C,d)} E^{1/p}\left[\left|\hat{I}_\alpha - I_\alpha(f)\right|^p\right] \;\leq\; O\left(n^{-1/(d+1)}\right). \tag{40}$$

Proof:

The result (40) follows directly from Proposition 4, modified according to remarks 2 and 6 in Section 3.4. As for (39) the proof follows from well known results in non-parametric function estimation which we only sketch here. The reader is referred to Ibragimov and Has'minskii [16] or Korostolev and Tsybakov [20] for more details. Define the Hölder class $\Sigma_d(\kappa, C)$ of functions g on $[0,1]^d$

$$\Sigma_d(\kappa,C) = \left\{g(x) : |g(x) - p_x^{\lfloor\kappa\rfloor}(z)| \leq C\,\|x-z\|^\kappa\right\}$$

where $p_x^k(z)$ is the Taylor polynomial (multinomial) of g of order k expanded about the point x, $||x||$ denotes the L_2 norm and $\lfloor\kappa\rfloor$ is defined as the greatest integer strictly less than κ. $\Sigma_d(1, C)$ is the set of Lipschitz functions with Lipschitz constant C and $\Sigma_d(\kappa, C)$ contains increasingly smooth functions as κ increases.

For any estimator $\hat{g}_n$ of g based on i.i.d. samples $X_1, \ldots, X_n$ the minimax L_p integrated error over the Hölder class $\Sigma_d(\kappa, C)$ satisfies

$$\sup_{g\in\Sigma_d(\kappa,C)} E^{1/p}\left[\int(\hat{g}_n(x)-g(x))^p dx\right] = O\left(n^{-\kappa/(2\kappa+d)}\right). \tag{41}$$

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

$$\sup_{g\in\Sigma_d(\kappa,C)} E^{1/p}\left[\left|\int(\hat{g^\alpha}(x)-g(x)dx\right|^p\right] = O\left(n^{-\kappa/(2\kappa+d)}\right). \tag{42}$$

The inequality (39) follows from the fact that $\Sigma_d(1, C) \subset \mathrm{BV}(C, d)$.

Relation (41) implies that there exist positive constants C_1, C_2 such that for all $g \in \Sigma_d(\kappa, C)$

$$\limsup_{n\to\infty}\left|[\hat{g}(x)-g(x)]n^{\kappa/(2\kappa+d)}\right| < \infty, \quad (w.p.1), \tag{43}$$

except possibly on a subset of $[0, 1]^d$ of measure zero, and for some $g \in \Sigma_d(\kappa, C)$

$$\liminf_{n\to\infty}\left|[\hat{g}(x)-g(x)]n^{\kappa/(2\kappa+d)}\right| > 0,, \quad (w.p.1) \tag{44}$$

over some subset of $[0, 1]^d$ of positive measure. Therefore, letting $g = f^\alpha$, using relations (43) and (44), there exist finite constants C_1 and C_2 such that

$$E^{1/p}\left[\left|\int(\hat{g}(x)-g(x))dx\right|^p\right] \le C_1 n^{-\kappa/(2\kappa+d)}(1+o(1)),$$

for all $g \in \Sigma_d(\kappa, C)$, and there exists a function $g \in \Sigma_d(\kappa, C)$ such that

$$E^{1/p}\left[\left|\int g^{\alpha-1}(x)(\hat{g}(x)-g(x))dx\right|^p\right] \ge C_2 n^{-\kappa/(2\kappa+d)}(1+o(1))$$

Therefore,

$$C_2 n^{-\kappa/(2\kappa+d)}(1+o(1)) \le \sup_{g\in\Sigma_d(\kappa,C)} E^{1/p}\left[\left|\int(\hat{g}(x)-g(x))dx\right|^p\right] \le C_1 n^{-\kappa/(2\kappa+d)}(1+o(1))$$

which establishes (42) and the proof of Proposition 6 is completed. □

We make several comments in connection with Proposition 6.

1. By Proposition 4 assumption $\alpha \in [1/2, (d-1)/d]$ can be relaxed to $\alpha \in (0, (d-1)/d]$ with corresponding weakening of the rate bound (40) of the graph-based entropy estimator to $O\left(\max\{n^{-1/(d+1)}, n^{-\alpha/2}\}\right)$. Specializing to rms error ($p = 2$), the slower $O\left(n^{-\alpha/2}\right)$ rate occurs when the variance of the minimal graph estimator exceeds the bias squared.

2. The assumption $\alpha \le (d-1)/d$ prevents the application of the convergence rate bound (27) in Proposition 6 to minimal graph estimates of the Shannon entropy $(\alpha \to 1)$. In particular, we cannot use it to bound a minimal-graph analog to the plug-in estimation method proposed by Mokkadem [24] in which Shannon entropy is estimated by a sequence $\hat{I}_{\alpha_n}(\hat{f}_n)$ of α-entropy estimators where $\alpha_n < 1$ and $\lim_{n\to\infty} \alpha_n = 1$. As mentioned in Remark 4 of Section 3.4, relaxation of this assumption would require extension to $\gamma < 1$ of the fundamental convergence rate $O(n^{-1/d})$ in (9) established by Redmond and Yukich [28].

3. The partitioned minimal graph approximation (29) can be adapted to entropy estimation in an obvious way and an analog to Proposition 6 will hold with (27) replaced by the slower $O(n^{-\beta/(1+\beta)})$ rate bound.

4. If it is known *a priori* that the class of functions f is significantly smoother than the BV class assumed in Proposition 6 then density estimation methods can have much faster convergence. As an extreme example, if f is a piecewise constant block density over an *a priori* known partition, a histogram plug-in estimator will have faster L_p convergence rate $O(1/\sqrt{n})$ while the minimal graph estimator will only have $O(n^{-1/d})$ convergence rate. 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 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.

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] R. Baraniuk, P. Flandrin, A. J. E. M. Jensen, and O. Michel, “Measuring time frequency information content using the rényi entropies,” *IEEE Trans. on Inform. Theory*, vol. IT-47, no. 4, , April 2001.

[4] 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.

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

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

[12] A. O. Hero, B. Ma, and O. Michel, “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`.

[13] 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.

[14] 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.

[15] 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.

[16] I. A. Ibragimov and R. Z. Has’minskii, *Statistical estimation: Asymptotic theory*, Springer-Verlag, New York, 1981.

[17] 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.

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

[19] 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.

[20] A. P. Korostelev and A. B. Tsybakov, *Minimax theory of image reconstruction*, Springer-Verlag, New York, 1993.

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

[22] 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, CA, October 2000.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

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

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

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

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

[35] 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.

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

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

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

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