Asymptotic Relations Between Minimal Graphs and α-entropy

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

February, 2003

Abstract

This report 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 $f \in \Sigma_d(\beta, L)$ (the space of Holder continuous functions), $0 < \beta \leq 1$, the $\mathcal{L}_\kappa$-norm convergence rate is bounded above by $O\left(n^{-\alpha\beta/(\alpha\beta+1)\ 1/d)}\right)$. We obtain similar rate bounds for minimal graph approximations implemented by a progressive divide-and-conquer partitioning heuristic. We also obtain asymptotic lower bounds for the respective rates of convergence, using minimax techniques from nonparametric function estimation. In addition to Euclidean optimization problems, these results have application to non-parametric entropy and information divergence estimation; adaptive vector quantization; and pattern recognition.

Keywords: continuous quasi-additive functionals, combinatorial optimization, graph theory, progressive-resolution approximations, data partitioning heuristic, minimax rates, nonparametric 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. Recall that the Rényi entropy or α-entropy is defined as

$$H_\alpha(f) = \frac{1}{1-\alpha} \log \int f^\alpha(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \ .$$

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 [1] and Yukich [2] 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 [3, 4] when the vertices are uniformly distributed over $[0,1]^d$.

In the present report 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 over $[0,1]^d$, for which $f \in \Sigma_d(\beta, L)$, the space of Holder continuous functions, $0 < \beta \leq 1$, 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 $O\left(n^{-\alpha\beta/(\alpha\beta+1)\ 1/d)}\right)$ bound on the r.m.s. convergence. 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 [3, 4]. Furthermore, the rate constants derived here suggest that slower convergence occurs when either the (Rényi) entropy of the underlying density f or the constant L is large. We also derive lower bounds to the respective convergence rates by recasting the problem as that of estimating the Rényi entropy, or equivalently $\int f^\alpha(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, over the non-parametric class of densities $f \in \Sigma_d(\beta, L)$. For this, we use standard minimax techniques from non-parametric function estimation.

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 [5],

Ravi *et al.* [6], and Hero and Michel [7], 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. [8, 9, 7, 10]. 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 [11, 12]. Furthermore, as empirical versions of vector quantization can be cast as geometric location problems [13], the asymptotics of adaptive VQ may be studied within the present framework of minimal Euclidean graphs.

The outline of this report 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 upper bounds for such functionals with general Holder continuous density f. In Section 4 we extend these results to partitioned approximations. In Section 5 we derive lower bounds to the convergence rates. In Section 6 we make a brief comment about nonparametric estimation of the Rényi entropy. Finally, in section 7 we digress about the characterization of a density from its α-entropy, when the later is regarded as a function of α. We also give an extension of the convergence rate upper bounds to densities in a Sobolev class in Appendix B.

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 [1] and Yukich [2] 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 [3]. 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

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

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 $L_\gamma^{\mathrm{TSP}}(\mathcal{X}_n) = \min_{C \in \mathcal{C}} \sum_{e \in C} |e|^\gamma$. Construction of the TSP graph is NP-hard and arises in many different areas of operations research [14]. 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 $L_\gamma^{\mathrm{MST}}(\mathcal{X}_n) = \min_{T \in \mathcal{T}} \sum_{e \in T} |e|^\gamma$. MST's arise in areas including: pattern recognition [15]; clustering [16]; nonparametric regression [17] and testing for randomness [18]. 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 $L_\gamma^{k-\mathrm{NNG}}(\mathcal{X}_n) = \sum_{i=1}^n \sum_{e \in \mathcal{N}_{k,i}} |e|^\gamma$. The k-NNG arises in computational geometry [19], clustering and pattern recognition [20], spatial statistics [21], and adaptive vector quantization [22].

The following technical conditions on a Euclidean functional L_γ were defined in [3, 2].

- *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 of $[0,1]^d$; 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 [2].

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

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

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 [4, 3] 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 [2, 3] 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*:

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

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 [2, Ch. 5]. The $n^{-1/d}$ convergence rate bound is exact for $d = 2$.

3 Convergence Rate Upper Bounds for General Density

In this section we obtain convergence rate bounds for a general non-uniform Lebesgue density $f \in \Sigma_d(\beta, L)$. 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 [23, 1, 2]: we first establish 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 concave 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 concave 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 concave, 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 $1 \leq \gamma < d$ [2, 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 Holder Continuous Density Functions

Before extending Proposition 1 to general densities we will need to establish an approximation lemma for Holder continuous functions.

Recall that the Holder class $\Sigma_d(\beta,L)$ is defined by

$$\Sigma_d(\beta,L)=\left\{g:|g(\boldsymbol{z})-p_{\boldsymbol{x}}^{\lfloor\beta\rfloor}(\boldsymbol{z})|\leq L\,|\boldsymbol{x}-\boldsymbol{z}|^\beta,\boldsymbol{x},\boldsymbol{z}\in\mathbb{R}^d\right\}$$

where $p_{\boldsymbol{x}}^k(\boldsymbol{z})$ is the Taylor polynomial (multinomial) of g of order k expanded about the point $\boldsymbol{x}$, $|.|$ denotes a norm in $\mathbb{R}^d$ and $\lfloor\beta\rfloor$ is defined as the greatest integer strictly less than β. $\Sigma_d(1,L)$ is the set of Lipschitz functions with Lipschitz constant L and $\Sigma_d(\beta,L)$ contains increasingly smooth functions as β increases.

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 $L_1([0,1]^d)$ sense) these resolution-m block densities approximate functions in $W^{1,p}(\mathbb{R}^d)$.

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

$$\int_{\mathcal{S}}|\phi(\boldsymbol{x})-f(\boldsymbol{x})|\mathrm{d}\boldsymbol{x}\leq C_6\,L\,m^{-\beta}. \tag{20}$$

A proof of this lemma is given in Appendix A.

Remark. Lemma 2 shows how close, in $\mathcal{L}_1(\mathbb{R}^d)$ sense, a function $f\in\Sigma_d(\beta,L)$ can be approximated by its resolution-m block density. To extend the results in this and the following sections to other classes of functions, all that it is needed is an upper bound to the $\mathcal{L}_1$ approximation error similar to the one in equation 20. In Appendix B, we show how to do this

for densities in the Sobolev space $W^{1,p}(\mathbb{R}^d)$, $1 \leq p < \infty$. The importance of Sobolev spaces derives from the fact that it includes functions that are not differentiable in the usual (strong) sense.

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,

$$II \leq \beta_{L_\gamma,d} \int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})|^{\frac{d-\gamma}{d}} \mathrm{d}\boldsymbol{x} \leq \beta_{L_\gamma,d}\, C_6'\, L^{(d-\gamma)/d}\, m^{-\beta(d-\gamma)/d}, \tag{22}$$

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 $f \in \Sigma_d(\beta, L)$, $0 < \beta \leq 1$, 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*

$$\left| E[L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)] - E[L_\gamma(\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_n)] \right| / n^{\frac{d-\gamma}{d}} \leq C_3'\, C_6'\, L^{(d-\gamma)/d}\, m^{-\beta(d-\gamma)/d}, \tag{23}$$

where $C_3' = 2^{(2d-\gamma)/d} C_3$.

Proof:

As in equation (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 ([23], 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 $C_6\, L\, m^{-\beta}$ 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

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

This bound is finite under the assumptions that $f \in \Sigma_d(\beta, L)$ 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 $f \in \Sigma_d(\beta, L)$, $0 < \beta \leq 1$, 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

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

where $\alpha = \frac{d-\gamma}{d}$.

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 $m^{-\beta(d-\gamma)/d}$. 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

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

The solution $m = m(n)$ occurs when $(nm^{-d})^{-1/d} = m^{-\beta(d-\gamma)/d}$, or $m = n^{1/[d(\alpha\beta+1)]}$ (integer part) and, correspondingly, $m^{-\beta(d-\gamma)/d} = n^{-\frac{\alpha\beta}{\alpha\beta+1}\frac{1}{d}}$. 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 [2, Thm. 6.3] established by Rhee [24]:

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

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 $f \in \Sigma_d(\beta, L)$, $0 < \beta \leq 1$, 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)*

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

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 $f \in \Sigma_d(\beta, L)$, $0 < \beta \leq 1$, 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)*

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

Proof:

For any non-random constant μ, using Minkowski inequality, $[E|W+\mu|^\kappa]^{1/\kappa} \leq [E|W|^\kappa]^{1/\kappa} + |\mu|$. 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 $f \in \Sigma_d(\beta, L)$, $0 < \beta \leq 1$, 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)*

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

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 $f \in \Sigma_d(\beta, L)$ 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 convergence rate bound of $O\left(n^{-1/d}\right)$ used in (10), established by Redmond and Yukich [3], to the case $0 < \gamma < 1$.

2. It can be shown in analogous manner to the proof of the umbrella theorems of [2, 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 $[0,1]^d$ [2, Thm. 5.2]. Thus, the bound predicts slower worst case convergence rates for non-uniform densities.

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 [2, 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 [2, 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

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

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 [5], Ravi *etal* [25], Mitchell [26], and Arora [27], as ways to reduce computational complexity. The fixed partition approximation is a simple example whose convergence rate has been studied by Karp [5, 28], Karp and Steele [29] and Yukich [2] 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 [2, 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 [2, 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 $f \in \Sigma_d(\beta, L)$, $0 < \beta \leq 1$, 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 $b(m) = O(m^{d-\gamma})$*

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

Proof:

Start with

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

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

Analogously to the proof of [2, 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 $0 \leq u \leq 1$

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

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 $\left(|b(m) - C_1 m^{d-\gamma}| + C_2 m^{d-\gamma} + b(m)\right) / n^{(d-\gamma)/d} \leq B(nm^{-d})^{-(d-\gamma)/d}$ for some constant B

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

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

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

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

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 $f \in \Sigma_d(\beta, L)$, $0 < \beta \leq 1$, 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

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

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

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 $(d-\gamma-1)/(d-1)$. 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* [6] whose a.s. convergence was established by Hero and Michel [7]

(Note that the overly strong BV condition assumed in [7] 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 Convergence Rate Lower Bounds

In this section we derive lower bounds for the convergence rates of minimal graphs. Define

$$I_\alpha(f) = \int f^\alpha(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \; . \tag{38}$$

From sections 2 and 3, $L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)/n^{(d-\gamma)/d}$ is a (strongly) consistent estimator of $I_\alpha(f)$ for $\alpha = \frac{d-\gamma}{d}$. Thus, it is natural to recast our problem as that of estimating $I_\alpha(f)$ over the nonparametric class of densities $f \in \Sigma_d(\beta, L)$.

Let $\hat{I}_\alpha$ be an estimator of $I_\alpha(f)$ ($0 < \alpha < 1$) based on a sample of n i.i.d. observations from a density f. To access the "quality" of $\hat{I}_\alpha$ we adopt the usual (nonparametric) minimax risk criterion, i.e., we look at $\sup_{f \in \mathcal{F}} E|\hat{I}_\alpha - I_\alpha(f)|^p$, the worst case performance of $\hat{I}_\alpha$ over a known class of densities $\mathcal{F}$, for a choice of $p \geq 1$. Under this criterion it is natural to ask what is the minimum achievable risk for any estimator, i.e., what is

$$\inf_{\hat{I}_\alpha} \sup_{f \in \mathcal{F}} E|\hat{I}_\alpha - I_\alpha(f)|^p \; ,$$

where the infimum is taken over all estimators of $I_\alpha(f)$, as this quantifies the best performance possible for any estimator. Of course, as $L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)/n^\alpha$ is valid estimator of $I_\alpha(f)$, this will also yield a lower bound to the convergence rates of interest. The rest of this section is devoted to deriving these (asymptotic) bounds using standard minimax techniques.

5.1 Notation

In the following, we will take the class $\mathcal{F}$ as the set of multivariate Lebesgue densities defined on the unit cube $[0,1]^d$ ($d \geq 1$), belonging to the functional Holder class $\Sigma_d(\beta, L)$.

We will also use the affinity $\|P \wedge Q\|$ between measures P and Q defined by:

$$\|P \wedge Q\| = 1 - \frac{1}{2}\|P - Q\|_1 \tag{39}$$

where $\|P\|_1$ is the total variation norm of P defined as

$$\|P\|_1 = \sup_{|f|\leq 1} \left| \int f \, \mathrm{d}P \right|$$

and the supremum is taken over all measurable functions f bounded by 1. If P and Q are absolutely continuous w.r.t. a measure μ, with densities p and q, respectively, then $\|P - Q\|_1 = \int |p - q| \, \mathrm{d}\mu$. In this case, we will write $\|p - q\|_1$ for $\|P - Q\|_1$ and $\|p \wedge q\|$ for $\|P \wedge Q\|$. Also, write p^n as shorthand notation for $\prod_{i=1}^n p(\boldsymbol{x}_i)$, the density of the product measure $P^{\otimes n}$.

Finally, write $co(\mathcal{F})$ to denote the convex hull of $\mathcal{F}$.

5.2 Lower Bounds

In order to get lower bounds for the minimax risk, the usual technique is to build, for every n, a subset $\mathcal{F}_{0,n} \subset \mathcal{F}$ of finite cardinality, such that the problem of estimating $I_\alpha(f)$ over $\mathcal{F}_{0,n}$ is essentially as difficult as the full problem. Assouad's lemma or Fano's lemma are the commonly used tools to address such constructions ([30]). However, in the case of entropy estimation (as well as many other functional estimation problems, [31], [32]), these methods only give the trivial lower bound zero. We will thus rely on a result by Le Cam (see for example [31]) that relates the minimax risk to a testing problem between two sets of hypothesis, whose convex hulls are "well" separated in a total variation distance sense. Bellow is a simplified version of this result, suited for our needs (for a simple proof see [31]):

Lemma 3 *Let $\hat{I}$ be an estimator of $I(f)$[1] based on n i.i.d. observations from a density $f \in \mathcal{F}$. Suppose that there are subsets $\mathcal{G}_1$ and $\mathcal{G}_2$ of $\mathcal{G} = \{f^n : f \in \mathcal{F}\}$ that are 2δ-separated, in the sense that, $|I(f_1) - I(f_2)| \geq 2\delta$ for all $f_1^n \in \mathcal{G}_1$ and $f_2^n \in \mathcal{G}_2$. Then*

$$\sup_{f \in \mathcal{F}} E|\hat{I} - I(f)| \geq \delta \cdot \sup_{p_i \, \in \, co(\mathcal{G}_i)} \|p_1 \wedge p_2\| \, .$$

We will apply lemma 3 to the usual small perturbations of the uniform density, u, on $[0,1]^d$. Towards this goal, fix $g \in \Sigma_d(\beta, 1)$ with support in $[0,1]^d$ such that $\int g(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = 0$, $\int g^2(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \kappa_2 > 0$ and $|g(\boldsymbol{x})| \leq M$. Let $\{Q_j\}_{j=1}^{m^d}$ be the uniform resolution-m partition and $\{\boldsymbol{x}_j\}_{j=1}^{m^d}$ be the set of points in $[0,1]^d$ that translate each Q_j back to the origin, as

[1] From now on, we will omit the subscript α from $\hat{I}_\alpha$ and $I_\alpha(f)$, unless necessary.

defined in Sub-section 2.1. Let $g_j(\boldsymbol{x}) = g\left(m(\boldsymbol{x} - \boldsymbol{x}_j)\right)$. For $\lambda \in \Lambda = \{-1, 1\}^{m^d}$, define the perturbation of u as

$$f_\lambda(\boldsymbol{x}) = 1 + \sum_{j=1}^{m^d} \frac{L}{2}\, m^{-\beta}\, \lambda_j\, g_j(\boldsymbol{x}) \tag{40}$$

It is easy to see that $\int f_\lambda(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 1$, $f_\lambda \in \Sigma_d(\beta, L)$ and, for m large enough, $f \geq 0$. So (for m large enough) $f \in \mathcal{F}$.

We can now apply lemma 3 to the sets $\mathcal{G}_1 = \{u^n\}$ and $\mathcal{G}_2 = \{f_\lambda^n : \lambda \in \Lambda\}$. We will start by determining the 2δ-separation between $\mathcal{G}_1$ and $\mathcal{G}_2$. Consider the second order Taylor expansion

$$(1+y)^\alpha = 1 + \alpha y + \frac{1}{2}\,\alpha(\alpha-1)\,\xi^{\alpha-2} y^2$$

where ξ lies between 1 and $1+y$. This implies that

$$\begin{aligned}
\int f_\lambda^\alpha(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} - 1 &= \sum_{j=1}^{m^d} \int_{Q_j} \left(1 + \frac{L}{2}\, m^{-\beta}\, \lambda_j\, g_j(\boldsymbol{x})\kappa\right)^\alpha \mathrm{d}\boldsymbol{x} - 1 \\
&= \frac{1}{2}\left(\frac{L}{2}\right)^2 \alpha(\alpha-1)\, m^{-2\beta} \sum_{j=1}^{m^d} \int_{Q_j} \xi^{\alpha-2}(\boldsymbol{x}) g_j^2(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}\,,
\end{aligned} \tag{41}$$

where $1 - M\frac{L}{2}\, m^{-\beta} \leq \xi(\boldsymbol{x}) \leq 1 + M\frac{L}{2}\, m^{-\beta}$. Inserting these bounds in equation (41), we have

$$\begin{aligned}
\frac{1}{2}\left(\frac{L}{2}\right)^2 \alpha(\alpha-1)\,\kappa_2\,(1 - M\frac{L}{2}\, m^{-\beta})^{\alpha-2}\, m^{-2\beta} &\leq \int f_\lambda^\alpha(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} - 1 \\
&\leq \frac{1}{2}\left(\frac{L}{2}\right)^2 \alpha(\alpha-1)\,\kappa_2\,(1 + M\frac{L}{2}\, m^{-\beta})^{\alpha-2}\, m^{-2\beta}\,,
\end{aligned} \tag{42}$$

which essentially means that $\int f_\lambda^\alpha(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} - 1 \doteq m^{-2\beta}$. We can now use this result to conclude, for any $\lambda \in \Lambda$,

$$|I(f_\lambda^n) - I(u^n)| = \left| \int f_\lambda^\alpha(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} - 1 \right| \geq 2\, C\, m^{-2\beta}\,, \tag{43}$$

for some constant $C > 0$ and m large enough.

We now need to derive a lower bound for $\sup_{p_i \,\in\, co(\mathcal{G}_i)} \|p_1 \wedge p_2\|$. To this end, let $h_n = 2^{-m^d} \sum_{\lambda\in\Lambda} f_\lambda^n \ \in co(\mathcal{G}_2)$. The following lemma provides such a bound ([33]):

Lemma 4

$$\|u^n - h_n\|_1^2 \leq \exp\left\{\frac{1}{2}\, m^d \left[n \int \left(\frac{L}{2}\, m^{-\beta} g_1(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x}\right]^2\right\} - 1$$

A proof of this lemma is given in appendix A.

For our choice of g_1, lemma 4 simplifies to:

$$\|u^n - h_n\|_1^2 \leq \exp\left\{\frac{1}{2}\left(\frac{L}{2}\right)^4 \kappa_2^2\, n^2\, m^{-(4\beta+d)}\right\} - 1$$

Now, choosing $m = O\left(n^{-2/(4\beta+d)}\right)$, the optimum value that balances the rates in lemma 3, and g such that κ_2 is small enough, then there exists an $\epsilon > 0$ such that

$$\|u^n - h_n\|_1^2 \leq (2(1-\epsilon))^2 \,.$$

Hence, by equation (39),

$$\|u^n \wedge h_n\| \geq 1 - 2(1-\epsilon)/2 = \epsilon > 0 \,. \tag{44}$$

Finally, plugging equation (43) and (44), with the choice of $m = O\left(n^{-2/(4\beta+d)}\right)$, into lemma 3 and using Jensen's inequality, gives us the desired lower bound:

Proposition 7 *For $\mathcal{F} = \{f : f$ is a Lebesgue density on $[0,1]^d$ and $f \in \Sigma_d(\beta, L)\}$, $p \geq 1$ and n large enough, there exists a constant $c = c(\beta, L, d, \alpha) > 0$ such that*

$$\inf_{\hat{I}_\alpha} \sup_{f \in \mathcal{F}} \left[E|\hat{I}_\alpha - I_\alpha(f)|^p\right]^{1/p} \geq c\, n^{-\frac{4\beta}{4\beta+d}} \,, \tag{45}$$

where the supremum is taken over all estimators $\hat{I}_\alpha$ of $I_\alpha(f)$ based on n i.i.d. observations from density f.

We make the following comments about this proposition.

1. For sufficiently smooth densities, i.e., for $\beta \geq d/4$, $4\beta/(4\beta+d) \geq 1/2$, which is the usual rate of convergence for parametric problems. This suggests, using the extension of the efficiency concept to the nonparametric setting (Cramer-Rao type inequalities, ... to be verified), that the lower bound in Proposition 7 can be replaced by

$$\inf_{\hat{I}_\alpha} \sup_{f \in \mathcal{F}} \left[E|\hat{I}_\alpha - I_\alpha(f)|^p\right]^{1/p} \geq c\, n^{-\left(\frac{4\beta}{4\beta+d} \wedge \frac{1}{2}\right)}$$

2. The results of Proposition 7 agree with those obtained by Birgé and Massart in [34]. In there, they derive lower bounds on the minimax risk for the general problem of nonparametric estimation of a functional $T(f) = \int \varphi(f(x), f'(x), \ldots, f^{(k)}(x), x)\, \mathrm{d}x$ satisfying some smoothness conditions. They also show, that for $\beta \geq 2k + d/4$, the $\sqrt{n}$-rate is achievable. Kerkyacharian and Picard closed the problem in [35] by showing that the corresponding rates for $\beta < 2k + d/4$ are also achievable.

3. Are the rates in Proposition 7 achievable? (I think they are...)

Remark. If, instead of the Rényi entropy, we were interested in the Shannon entropy $H_1(f) = -\int f(\boldsymbol{x}) \log f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$, the same rates would be obtained. This can be seen by considering the second order Taylor expansion,

$$(1+y)\log(1+y) = y + \frac{1}{2}\,\xi^{-1}\,y^2$$

and following the same steps as for $I_\alpha(f)$. In [36], Laurent exhibits an efficient estimator of this entropy, for densities defined on a compact set of the real line with smoothness parameter $\beta \geq 1/4$, that achieves the $\sqrt{n}$-rate on densities bounded away from zero on their domain.

6 Performance of Minimal Graph and Plug-in Entropy Estimators

In this section we derive upper bounds for the maximum risk of plug-in estimators and minimal-graph based estimators of entropy.

We consider entropy estimates of the form $\hat{H}_\alpha = (1-\alpha)^{-1} \log \hat{I}_\alpha$, where $\hat{I}_\alpha$ is a consistent estimator of $I_\alpha(f)$. By a standard perturbation analysis of $\ln x$,

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

Thus, as $I_\alpha(f)$ is bounded away from zero uniformly over the class $\mathcal{F}$ (i.e., $\inf_{f\in\mathcal{F}} I_\alpha(f) > 0$) , the asymptotic rate of convergence of $\hat{H}_\alpha - H_\alpha(f)$, as a function of n, will be identical to that of $\hat{I}_\alpha - I_\alpha(f)$.

Let $\hat{f}$ be a density estimate of f based on n i.i.d. observations (from density f). We have the following upper bound for plug-in estimators $I_\alpha(\hat{f})$:

Proposition 8 *For $\mathcal{F}$ as defined in Proposition 7,*

$$\sup_{f\in\mathcal{F}} E\left|I_\alpha(\hat{f}) - I_\alpha(f)\right| \leq C_1 \, n^{-\frac{\alpha\beta}{2\beta+d}} \tag{46}$$

for $C_1 = C_1(\beta, L, d) > 0$.

Proof: The proof relies on the well known minimax rates for density estimation available in the literature (see, for example, [37]). Specifically, these rates are of order $O\left(n^{-\beta/(2\beta+d)}\right)$, i.e.,

$$\sup_{f\in\mathcal{F}} E\int |\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x} \leq C_1\, n^{-\frac{\beta}{2\beta+d}}$$

for the best estimators $\hat{f}$ (for example, wavelet thresholding based estimators).

Using the above result, the inequality $|a^\alpha - b^\alpha| \leq |a-b|^\alpha$ $(a, b \geq 0)$ and successive applications of Jensen's inequality yield the desired result,

$$\begin{aligned} E\left|I_\alpha(\hat{f}) - I_\alpha(f)\right| &\leq E\int \left|\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right|^\alpha \mathrm{d}\boldsymbol{x} \\ &\leq E\left[\int \left|\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right| \mathrm{d}\boldsymbol{x}\right]^\alpha \leq \left[E\int \left|\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right| \mathrm{d}\boldsymbol{x}\right]^\alpha \leq C_1\, n^{-\frac{\alpha\beta}{2\beta+d}} \end{aligned}$$

□

For $\hat{I}_\alpha$ denoting the minimal graph estimator of $I_\alpha(f)$, we have from Proposition 5 the following result:

Proposition 9 *For $\mathcal{F}$ as defined in Proposition 7, with $0 < \beta \leq 1$, $1/2 \leq \alpha \leq (d-1)/d$,*

$$\sup_{f\in\mathcal{F}} E\left|\hat{I}_\alpha - I_\alpha(f)\right| \leq C_2\, n^{-\frac{\alpha\beta}{\alpha\beta+1}\frac{1}{d}} \tag{47}$$

for $C_2 = C_2(\beta, L, d, \alpha) > 0$.

7 Notes on the Invertibility of the α-entropy

This section is somehow different in character from the previous sections. In here, we briefly digress about how the knowledge of $H_\alpha(f)$, for $\alpha \in G$, whith $G \in [0,1]$ being any open interval, can provide information about the density f.

We consider first the 1-dimensional case, i.e., f is a univariate Lebesgue density. Let $S_1, S_2, \ldots$ be the support regions of a monotonic decomposition of f such that the change of variable $y = \ln f(x)$ is (locally) invertible over each set S_i. Define also f_i^{-1} as the local inverse of $f(x)$ over $x \in S_i$. We thus have

$$\begin{aligned} I_\alpha(f) &= \int f^\alpha(x)\,\mathrm{d}x = \int \mathrm{e}^{\alpha \ln f(x)}\,\mathrm{d}x = \int \mathrm{e}^{\alpha y} \sum_i \left(\left|\frac{\mathrm{d}}{\mathrm{d}x}\ln f(x)\right|_{x=f_i^{-1}(\mathrm{e}^y)}\right)^{-1} 1_{S_i}(f_i^{-1}(\mathrm{e}^y))\,\mathrm{d}y \\ &= \int \mathrm{e}^{\alpha y} \sum_i \left|\frac{f(f_i^{-1}(\mathrm{e}^y))}{f'(f_i^{-1}(\mathrm{e}^y))}\right| 1_{\ln f(S_i)}(y)\,\mathrm{d}y\,. \end{aligned} \tag{48}$$

Equation (48) shows that $I_\alpha(f)$, as a function of α, is the Laplace transform of the function $g(y) = \sum_i \left| \frac{f(f_i^{-1}(\mathrm{e}^y))}{f'(f_i^{-1}(\mathrm{e}^y))} \right| 1_{\ln f(S_i)}(y)$. Of course many different densities f will result in the same g; just consider any location change of probability mass in f and the resulting entropy will remain the same.

Consider now the multivariate case. Without loss of generality, we only need to study the 2-dimensional case, as the general situation follows by induction. Write f as $f(x_1, x_2) = f(x_1|x_2)f(x_2)$, where $f(x_1|x_2)$ is the conditional density of X_1 given X_2 and $f(x_2)$ is the marginal density of X_2. Let $g(y_1|x_2)$ be the function g defined above with $f(x)$ replaced by $f(x_1|x_2)$. Proceeding in the same fashion as above, we have the following equalities:

$$\begin{aligned} I_\alpha(f) &= \int\int f^\alpha(x_1, x_2)\,\mathrm{d}x_1\mathrm{d}x_2 = \int\int f^\alpha(x_1|x_2) f^\alpha(x_2)\,\mathrm{d}x_1\mathrm{d}x_2 \\ &= \int\int \mathrm{e}^{\alpha y_1 + \alpha y_2} \sum_i g(y_1|f_i^{-1}(\mathrm{e}^{y_2}))\, g(y_2)\,\mathrm{d}y_1\mathrm{d}y_2 = \int\int \mathrm{e}^{\alpha y_1 + \alpha y_2}\, G(y_1, y_2)\,\mathrm{d}y_1\mathrm{d}y_2\ , \end{aligned} \tag{49}$$

where $G(y_1, y_2) = \sum_i g(y_1|f_i^{-1}(\mathrm{e}^{y_2}))\, g(y_2)$. Equation (49) shows that $I_\alpha(f)$ is the 2-D Laplace transform of the function $G(y_1, y_2)$, evaluated at the point (α, α). So, the knowledge of the multivariate α-entropy of a density, as a function of α, characterizes only the Laplace transform of the function G over the line $\alpha_1 = \alpha_2$ on the Laplace frequency plane.

8 Conclusion

In this report we have given rate of convergence bounds for length functionals of minimal-graphs satisfying continuous quasi-additivity, and briefly discussed their performance for entropy estimation. These results suggest that further exploration of minimal graphs for estimation of Rényi divergence, Rényi mutual information, and Rényi Jensen difference is justifed.

There are still many problems that remain to be studied. One such problem is the achievability of the minimax rates derived in section 5, in particular, the existence of practical estimators that achieve these rates. We believe this is a challenging problem as the techniques commonly used to address this problem yield only estimators of theoretical interest. One other problem is the derivation of convergence rate bounds for the k-MST, as this graph provides a robust entropy estimator. Also, to complete the results given in this report, it would be interesting to extend the rate bounds to smoother Holder continuous densities (i.e., $\beta > 1$). With regards to future applications, we feel that these methods can be applied in problems such as independent component analysis (ICA) or clustering techniques. Finally, establishing general weak

convergence results for these types of minimal graphs could have a significant impact in applications such as hypothesis testing and goodness of fit tests.

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

Proof of Lemma 2: By the mean value theorem, there exist points $\boldsymbol{\xi}_i \in Q_i$ such that

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

Note that, in what follows, $|.|$ means both the absolute value in $\mathbb{R}$ and any norm in $\mathbb{R}^d$. Using now the fact that $f \in \Sigma_d(\beta, L)$,

$$\int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} = \sum_{i=1}^{m^d} \int_{Q_i} |f(\boldsymbol{\xi}_i) - f(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \leq \sum_{i=1}^{m^d} \int_{Q_i} L \, |\boldsymbol{x} - \boldsymbol{\xi}_i|^\beta \mathrm{d}\boldsymbol{x} \, .$$

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

$$\int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \leq C \, L \, m^{-\beta} \, .$$

□

Proof of Lemma 4: This proof follows from [33]. Define

$$G_i(\boldsymbol{\lambda}) = G(\boldsymbol{X}_i, \boldsymbol{\lambda}) = \sum_{j=1}^{m^d} \frac{L}{2} m^{-\beta} \lambda_j g_j(\boldsymbol{X}_i) = \frac{L}{2} m^{-\beta} \boldsymbol{\lambda}^t \boldsymbol{g}(\boldsymbol{X}_i)$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{m^d})^t \in \Lambda$ and $\boldsymbol{g} = (g_1, \ldots, g_{m^d})^t$. Define also

$$\tau_i(\boldsymbol{\lambda}, \boldsymbol{\mu}) = E_{u^n} G_i(\boldsymbol{\lambda}) G_i(\boldsymbol{\mu})$$

for $\boldsymbol{\lambda}, \boldsymbol{\mu} \in \Lambda$. Note that, due to construction of g,

$$E_{u^n} G_i(\boldsymbol{\lambda}) = 0 \, , \tag{50}$$

and due to identically distributed samples assumption, $\tau_i(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \tau_1(\boldsymbol{\lambda}, \boldsymbol{\mu})$.

Now, rewrite h_n as:

$$\begin{aligned} h_n &= \sum_{\boldsymbol{\lambda} \in \Lambda} w_{\boldsymbol{\lambda}} \prod_{i=1}^{n} (1 + G_i(\boldsymbol{\lambda})) \\ &= \sum_{\boldsymbol{\lambda} \in \Lambda} w_{\boldsymbol{\lambda}} \left(1 + \sum_i G_i(\boldsymbol{\lambda}) + \sum_{i<j} G_i(\boldsymbol{\lambda}) G_j(\boldsymbol{\lambda}) + \sum_{i<j<k} G_i(\boldsymbol{\lambda}) G_j(\boldsymbol{\lambda}) G_k(\boldsymbol{\lambda}) + \ldots \right) \end{aligned}$$

where $w_{\boldsymbol{\lambda}} = 2^{-m^d}$. From a Bayesian perspective, the weights $w_{\boldsymbol{\lambda}}$ define a uniform prior probability on Λ.

Using Jensen's inequality,

$$\begin{aligned}\|h_n - u^n\|_1^2 &= \left(E_{u^n}|h_n - 1|\right)^2 \le E_{u^n}|h_n - 1|^2 \\ &= E_{u^n}\left\{\sum_{\boldsymbol{\lambda},\boldsymbol{\mu}\in\Lambda} w_{\boldsymbol{\lambda}} w_{\boldsymbol{\mu}} \left(\sum_i G_i(\boldsymbol{\lambda}) + \sum_{i<j} G_i(\boldsymbol{\lambda}) G_j(\boldsymbol{\lambda}) + \ldots\right)\left(\sum_i G_i(\boldsymbol{\mu}) + \right.\right. \\ &\qquad \left.\left. + \sum_{i<j} G_i(\boldsymbol{\mu}) G_j(\boldsymbol{\mu}) + \ldots\right)\right\} \end{aligned} \tag{51}$$

Expanding out the product in (51), due to independence and (50), only the terms where each factor $G_i(\boldsymbol{\lambda})$ is paired with a corresponding $G_i(\boldsymbol{\mu})$ will survive. All other terms with an isolated factor will be zero. The result is

$$\begin{aligned} E_{u^n}|h_n - 1|^2 &= \sum_{\boldsymbol{\lambda},\boldsymbol{\mu}\in\Lambda} w_{\boldsymbol{\lambda}} w_{\boldsymbol{\mu}} \left(\sum_i \tau_i(\boldsymbol{\lambda},\boldsymbol{\mu}) + \sum_{i<j} \tau_i(\boldsymbol{\lambda},\boldsymbol{\mu})\tau_j(\boldsymbol{\lambda},\boldsymbol{\mu}) + \ldots\right) \\ &= \sum_{\boldsymbol{\lambda},\boldsymbol{\mu}\in\Lambda} w_{\boldsymbol{\lambda}} w_{\boldsymbol{\mu}} \left(1 + \tau_1(\boldsymbol{\lambda},\boldsymbol{\mu})\right)^n - 1 \end{aligned} \tag{52}$$

Regarding the double sum in (52) as an expectation of a pair of independent random variables $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, each distributed according to a uniform prior in Λ, we get the following bound for the total variation norm:

$$\|h_n - u^n\|_1^2 \le E\left(1 + \tau_1(\boldsymbol{\lambda},\boldsymbol{\mu})\right)^n - 1 \le E\exp\{n\,\tau_1(\boldsymbol{\lambda},\boldsymbol{\mu})\} - 1\,, \tag{53}$$

where the last inequality comes from $\mathrm{e}^x \ge 1 + x$.

Now, note that the functions g_i have disjoint supports and, so, are orthogonal in the sense that $E_u g_i(\boldsymbol{X}_1) g_j(\boldsymbol{X}_1) = 0$, for $i \ne j$. Thus, we have

$$\tau_1(\boldsymbol{\lambda},\boldsymbol{\mu}) = \left(\frac{L}{2}\, m^{-\beta}\right)^2 \boldsymbol{\lambda}^t\, E_{u^n}\left\{\boldsymbol{g}(\boldsymbol{X}_1)\boldsymbol{g}^t(\boldsymbol{X}_1)\right\}\boldsymbol{\mu} = \sigma^2\,\boldsymbol{\lambda}^t\boldsymbol{\mu}\,,$$

with $\sigma^2 = \int \left(\frac{L}{2}\, m^{-\beta} g_1(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x}$. Equation (53) simplifies to

$$\|h_n - u^n\|_1^2 \le E\exp\{n\,\sigma^2\,\boldsymbol{\lambda}^t\boldsymbol{\mu}\} - 1\,.$$

The above expectation is easy to compute because the choice of a uniform prior on Λ makes the coordinates λ_i independent, taking values $+1$ and -1 with probability $1/2$:

$$E\exp\{n\,\sigma^2\,\boldsymbol{\lambda}^t\boldsymbol{\mu}\} = \left(\frac{1}{2}\,\mathrm{e}^{n\,\sigma^2} + \frac{1}{2}\,\mathrm{e}^{-n\,\sigma^2}\right)^{m^d} \le \exp\left\{\frac{1}{2}\, m^d\,(n\,\sigma^2)^2\right\}\,.$$

Lemma 4 now follows. □

B Appendix

In this Appendix we will introduce some concepts from the theory of Sobolev spaces and then show how to extend the previous results on convergence rate bounds to densities in the Sobolev class.

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 [38], 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 [38]. 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 ([38, Thm. 2.3.2]).

Let ϕ be the resolution-m block density approximation of f, as defined in section 3.2. 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 5 *For $1 \leq p < \infty$, let $f \in W^{1,p}(\mathbb{R}^d)$ have support $\mathcal{S} \subset [0,1]^d$. Then there exists a constant $C > 0$, independent of m, such that*

$$\int_{\mathcal{S}} |\phi(\boldsymbol{x}) - f(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \leq C m^{-1} (\|\mathrm{D}f\|_p + o(1)) . \tag{54}$$

Proof: First assume 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

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

Using the above results, Jensen inequality and 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 $W^{1,p}(\mathbb{R}^d)$ ([38, Thm. 2.3.2]), using the standard limiting argument the above inequality holds for $f \in W^{1,p}(\mathbb{R}^d)$. This establishes the desired result. □

Lemma 5 now provides the necessary result to extend the convergence rate bounds derived previously to the Sobolev case. As it can be seen from section 3.2, the $\mathcal{L}_1$ approximation error will influence the final rate upper bound only through the exponent β in equation (20). As the Sobolev approximation error (54) is similar to the Holder class case for $\beta = 1$, we immediately have the following proposition:

Proposition 10 *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)*

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

References

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

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

[3] 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.

[4] 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.

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

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

[7] A.O. 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.

[8] A. O. Hero, B. Ma, O. Michel, and J. D. Gorman, "Alpha-divergence for classification, indexing and retrieval," Tech. Rep. 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`.

[9] 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, Oct. 2000.

[10] A.O. 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.

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

[12] 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.

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

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

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

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

[17] David Banks, Michael Lavine, and H. Joseph 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. Joseph Newton, Ed., pp. 370–374. 1992.

[18] 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.

[19] 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.

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

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

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

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

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

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

[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*, 1996, pp. 402–408.

[27] 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.

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

[29] 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. Rinnooy Kan, and D. B. Shmoys, Eds., pp. 181–206. Wiley, New York, 1985.

[30] C. Huber, "Lower bounds for function estimation," in *Festschrift for Lucien Le Cam*, D. Pollard, E. Torgersem, and G. Yang, Eds., pp. 245–258. Springer-Verlag, New York, 1997.

[31] Bin Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*, D. Pollard, E. Torgersem, and G. Yang, Eds., pp. 423–435. Springer-Verlag, New York, 1997.

[32] O. Lepski, A. Nemirovski, and V. Spokoiny, "On estimation of the L_r norm of a regression function," *Probab. Theory Relat. Fields*, vol. 113, pp. 221–253, 1999.

[33] D. Pollard, "Asymptopia," `http://www.stat.yale.edu/~pollard/Asymptopia/`.

[34] L. Birgé and P. Massart, "Estimation of integral functionals of a density," *The Annals of Statistics*, vol. 23, no. 1, pp. 11–29, 1995.

[35] G. Kerkyacharian and D. Picard, "Estimating nonquadratic functionals of a density using Haar wavelets," *The Annals of Statistics*, vol. 24, no. 2, pp. 485–507, 1996.

[36] B. Laurent, "Efficient estimation of integral functionals of a density," *The Annals of Statistics*, vol. 24, no. 2, pp. 659–681, 1996.

[37] M. Neumann, "Multivariate wavelet thresholding: a remedy against the curse of dimensionality," Preprint no. 229, Weierstrass Institute, Berlin, 1996.

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