Available online at www.sciencedirect.com

ScienceDirect

Journal of Approximation Theory 256 (2020) 105425

Journal of Approximation Theory

www.elsevier.com/locate/jat

Full Length Article

Bracketing numbers of convex and m-monotone functions on polytopes

Charles R. Doss[1]

School of Statistics, University of Minnesota, 224 Church St SE #313, Minneapolis, MN 55455, United States of America

Received 6 November 2017; received in revised form 23 April 2019; accepted 11 April 2020
Available online 20 April 2020

Communicated by Josef Dick

Abstract

We study bracketing covering numbers for spaces of bounded convex functions in the L_p norms. Bracketing numbers are crucial quantities for understanding asymptotic behavior for many statistical nonparametric estimators. Bracketing number upper bounds in the supremum distance are known for bounded classes that also have a fixed Lipschitz constraint. However, in most settings of interest, the classes that arise do not include Lipschitz constraints, and so standard techniques based on known bracketing numbers cannot be used. In this paper, we find upper bounds for bracketing numbers of classes of convex functions without Lipschitz constraints on arbitrary polytopes. Our results are of particular interest in many multidimensional estimation problems based on convexity shape constraints.

Additionally, we show other applications of our proof methods; in particular we define a new class of multivariate functions, the so-called m-monotone functions. Such functions have been considered mathematically and statistically in the univariate case but never in the multivariate case. We show how our proof for convex bracketing upper bounds also applies to the m-monotone case.

MSC: primary 52A41; 41A46; secondary 52A27; 52B11; 52C17; 62G20

Keywords: Bracketing entropy; Kolmogorov metric entropy; Convex functions; Convex polytope; Covering numbers; Nonparametric estimation; Convergence rates

1. Introduction and motivation

To quantify the size of an infinite dimensional set, the pioneering work of [34] studied the so-called metric entropy of the set, which is the logarithm of the metric covering number of the set. In this paper, we are interested in a related quantity, the bracketing entropy for

E-mail address: cdoss@stat.umn.edu.

[1] Supported by National Science Foundation grant DMS-1712664.

https://doi.org/10.1016/j.jat.2020.105425

a class of functions, which serves a similar purpose as metric entropy. Metric or bracketing entropies quantify the amount of information it takes to approximate any element of a set with a given accuracy $\epsilon > 0$. This quantity is important in many areas of statistics and information theory; in particular, the asymptotic behavior of empirical processes and thus of many statistical estimators is fundamentally tied to the entropy of related classes of functions under consideration [19].

Let $\mathcal{F}$ be a set of functions on some space $\mathcal{X}$ and let ρ be a metric on $\mathcal{F}$. Given a pair of functions l, u on $\mathcal{X}$, a *bracket* $[l, u]$ is the set of all functions $f : \mathcal{X} \to \mathbb{R}$ with $l \le f \le u$ pointwise. For $\epsilon > 0$, we say $[l, u]$ is an ϵ-bracket (for ρ) if $\rho(l, u) \le \epsilon$. Then the ϵ-bracketing number of $\mathcal{F}$, denoted $N_{[\,]}(\epsilon, \mathcal{F}, \rho)$, is the smallest integer N such that there exist ϵ-brackets $[l_i, u_i]$, $i = 1, \ldots, N$, such that for all $f \in \mathcal{F}$, $f \in [l_i, u_i]$ for some i. (We do not actually force $l_i, u_i \in \mathcal{F}$.) The bracketing entropy is the logarithm of the bracketing number. Like metric entropies, bracketing entropies are fundamentally tied to rates of convergence of certain estimators (see e.g., [5,41,42]). In this paper, we study the bracketing entropy of classes of convex functions. Our interest is motivated by the study of nonparametric estimation of functions satisfying convexity restrictions, such as the least-squares estimator of a convex or concave regression function on $\mathbb{R}^d$ (e.g., [30,39]), possibly in the high dimensional setting [45], or estimators of a log-concave or s-concave density (e.g., [14–16,32,33,40], among others). Entropy bounds, of the metric or bracketing type, are directly relevant for studying asymptotic behavior of estimators in these contexts.

Fix the dimension $d \in \{2, 3, \ldots\}$. Let $D \subset \mathbb{R}^d$ be a convex set, let $v_1, \ldots, v_d \in \mathbb{R}^d$, be linearly independent vectors, let $B, \Gamma_1, \ldots, \Gamma_d$ be positive reals, and let $\boldsymbol{v} = (v_1, \ldots, v_d)$ and $\boldsymbol{\Gamma} = (\Gamma_1, \ldots, \Gamma_d)$. For $f : D \to \mathbb{R}$, let $L_{p,D}(f) \equiv L_p(f) = \left(\int_D f(x)^p \, dx \right)^{1/p}$ for $1 \le p < \infty$, and let $L_\infty(f) = \sup_{x \in D} |f(x)|$. We will let $\mathcal{C}$ with various arguments denote different classes of convex functions. We let $\mathcal{C} \equiv \mathcal{C}_d$ be the class of convex functions on $\mathbb{R}^d$, where we consider all convex functions f to be defined on all of $\mathbb{R}^d$ and to take the value ∞ off of its *effective domain* $\operatorname{dom}(f) := \left\{ x \in \mathbb{R}^d : f(x) < \infty \right\}$ [37]. (This approach does not affect bracketing numbers.) For a function f and a set $D \subset \mathbb{R}^d$, we will use the notation $f : D \to \mathbb{R}$ to mean that $\operatorname{dom}(f) = D$ and we let $\mathcal{C}_d(D) \equiv \mathcal{C}(D)$ be the class of convex functions on $\mathbb{R}^d$ with $\operatorname{dom}(f) = D$. Then we let

$$\begin{aligned} \mathcal{C}(D, B, \boldsymbol{\Gamma}, \boldsymbol{v}) := \{ f \in \mathcal{C}(D) : L_\infty(f) \le B, &|f(x + \lambda v_i) - f(x)| \\ &\le \Gamma_i |\lambda| \text{ if } x, x + \lambda v_i \in D \} \end{aligned} \tag{1}$$

be the class of convex functions on D satisfying uniform boundedness and Lipschitz constraints given by B and $\boldsymbol{\Gamma}$. When $\{v_1, \ldots, v_n\}$ is the standard basis of $\mathbb{R}^d$, we just write $\mathcal{C}(D, B, \boldsymbol{\Gamma})$ for this class. If D is the hyperrectangle $\prod_{i=1}^d [a_i, b_i]$ (with $a_i < b_i$), then [7] and [20] (chapter 8) show that if $0 < \epsilon \le \epsilon_0$ (for some $\epsilon_0 > 0$) then

$$\log N\left(\epsilon, \mathcal{C}\left(\prod_{i=1}^d [a_i, b_i], B, \boldsymbol{\Gamma} \right), L_\infty \right) \le C \epsilon^{-d/2} \tag{2}$$

for a constant $C \equiv C_{D,B,\boldsymbol{\Gamma}}$. Here, $N(\epsilon, \mathcal{F}, \rho)$ is the ϵ-covering number of $\mathcal{F}$ in the metric ρ, which is defined to be the smallest number of balls of ρ-radius ϵ that cover $\mathcal{F}$, and $\log N(\epsilon, \mathcal{F}, \rho)$ is the corresponding metric entropy of $\mathcal{F}$, discussed in the first paragraph of this paper.

One would like to use (2) in the study of asymptotic properties of the statistical estimators discussed above. Unfortunately, the function classes that arise in those problems generally do not include Lipschitz constraints, and so the class $\mathcal{C}(D, B, \boldsymbol{\Gamma})$ is not of immediate use. Furthermore, it turns out that without Lipschitz constraints, the L_∞ covering or bracketing numbers are not bounded. Thus, instead of using the L_∞ distance, we may consider using the L_p distances, $1 \leq p < \infty$. Let $\mathcal{C}(D, B)$ be the class of convex functions on D with uniform bound B (and no Lipschitz constraints). Then [18] and [29] found bounds when $d = 1$ and $d > 1$, respectively, for metric entropies of $\mathcal{C}(D, B)$: they showed that $\log N\left(\epsilon, \mathcal{C}(D, B), L_p\right) \lesssim \epsilon^{-d/2}$, again with D a hyperrectangle and $1 \leq p < \infty$. Here $\lesssim$ means $\leq$ up to a constant which does not depend on ϵ (but does depend on D, B, and p). The $d = 1$ case (from [18]) was the fundamental building block in computing global rates of convergence of the univariate log-concave and s-concave MLEs in [14]. In the corresponding statistical problems when $d > 1$, the domain of the functions under consideration is not restricted to be a hyperrectangle but rather may be an arbitrary convex set D. Thus the results of [29] are not immediately applicable, and there is need for results on more general convex domains D with a more complicated boundary and no Lipschitz constraints.

In this paper we are indeed able to generalize the results of [29] considerably by finding bracketing entropy upper bounds for all (convex) polytopes D, attaining the bound

$$\log N_{[\,]}\left(\epsilon, \mathcal{C}(D, B), L_p\right) \lesssim \epsilon^{-d/2} \tag{3}$$

with $1 \leq p < \infty$, D a polytope, and $0 < B < \infty$; this result is given in Theorem 3.5. Note that we work with bracketing entropy rather than metric entropy. Bracketing entropies are larger than metric entropies for the L_p metrics,

$$N(\epsilon, \mathcal{F}, L_p) \leq N_{[\,]}(2\epsilon, \mathcal{F}, L_p), \quad \text{for } 1 \leq p \leq \infty, \text{ and } N(\epsilon, \mathcal{F}, L_\infty) = N_{[\,]}(2\epsilon, \mathcal{F}, L_\infty), \tag{4}$$

[42, p. 84], so our bracketing entropy bounds imply metric entropy bounds of the same order. Along the way, we also generalize the results of [7] to bound the L_∞ bracketing numbers of $\mathcal{C}(D, B, \boldsymbol{\Gamma})$ when D is arbitrary. One of the benefits of our method is its constructive nature. We initially study only simple polytopes (defined in Section 3.2) and in that case we pay careful attention to how the constants depend on D.

In Section 5, we consider two further applications of our methods and ideas. In Section 5.1 we define a new class of functions, the so-called multivariate m-monotone functions. In the univariate setting m-monotone functions have been studied mathematically ([43,44], and references therein) and statistically [1,2,26], but to the best of our knowledge there has been no consideration or even definition of m-monotone functions in the multivariate case. We define a class and show that our proof for the bracketing upper bound for convex functions applies to the case of m-monotone functions. This is given in Theorem 5.16.

In Section 5.2 we consider level set estimation (where the λ-level set of a function f is $\{x : f(x) = \lambda\}$). Nonparametric level set estimation has gained increasing attention in recent years, since it can capture very complex dependencies in a distribution or dataset. In Bayesian analysis, the level set of the posterior distribution is commonly used to form a credible set, and this level set often has to be estimated based on samples generated from the Markov chain Monte Carlo method. There are a large number of other settings where level set estimation arises; see, for instance, the introduction of [17]. Here, we consider convex level set estimation.

For a recent review of convexity-based methods in set estimation, see [8]. In Section 5.2, we present upper bounds for the so-called local entropy of level sets of convex functions. These upper bounds are an important step in proving that fast rates of convergence may be achievable when one is estimating a polytopal level set of a convex function.

During the course of the development of this paper, we became aware of the related work [27], which was developed simultaneously and separately from our paper. In [27], the authors demonstrate in their Theorem 1.6 that if D is a sphere then (3) fails when $p(d-1) > d$. This shows that if D is not a polytope the situation may be more complicated than when D is a polytope. They also find upper bounds of order $\epsilon^{-d/2}$ when D is a polytope. Their methods are quite different than ours and in particular they do not explicitly construct their bracketing set but rather rely on an algebraic relation (see their function $g(\cdot,\cdot)$ in their Section 2.5); our method on the other hand is explicitly constructive. Our constants differ from those of [27]. Our constants depend on the volume (measured in the appropriate dimension) of the faces of the polytope D, which is perhaps an interesting phenomenon (and is (distantly) reminiscent of the Minkowski–Steiner formula [23]). Besides the fact that our constants differ from those of [27] and reflect the geometry of D, the constructive nature of our approach enables consideration of other problems, not considered by [27], which we do in Section 5 (as described above).

This paper is organized as follows. In Section 2 we prove bounds for bracketing entropy of classes of convex functions with Lipschitz bounds, using the L_∞ metric. We use these to prove our main result, Theorem 3.5, for the bracketing entropy of classes of convex functions without Lipschitz bounds in the L_p metrics, $1 \le p < \infty$, which we do in Section 3. We defer some of the details of the proofs to Section 4. In Section 5 we study two more problems. In Section 5.1 we consider bracketing numbers related to univariate and multivariate m-monotone function classes. In Section 5.2 we consider local entropies related to level set estimation.

2. Bracketing with lipschitz constraints

If we have sets $D_i \subset \mathbb{R}^d$, $i = 1, \ldots, M$, for $M \in \mathbb{N}$, and $D \subseteq \cup_{i=1}^M D_i$ then for $\epsilon_i > 0$, $0 < p < \infty$, and any class of functions $\mathcal{F}$,

$$N_{[\,]}\left(\left(\sum_{i=1}^M \epsilon_i^p\right)^{1/p}, \mathcal{F}, L_p\right) \le \prod_{i=1}^M N_{[\,]}\left(\epsilon_i, \mathcal{F}|_{D_i}, L_p\right), \tag{5}$$

where, for a set G, we let $\mathcal{F}|_{\mathcal{G}}$ denote the class $\{f|_G : f \in \mathcal{F}\}$ where $f|_G$ is the restriction of f to the set G. We will apply (5) to a cover of D by sets G with the property that $\mathcal{C}\,(D,1)\,|_G \subseteq \mathcal{C}\,(G, 1, \boldsymbol{\Gamma})$ for some bounded vector $\boldsymbol{\Gamma}$, so that we can apply bracketing results for classes of convex functions with Lipschitz bounds. Thus, in this section, we develop the needed bracketing results for such Lipschitz classes, for arbitrary (bounded) convex domains D. Recall the definition of $\mathcal{C}\,(D, B, \boldsymbol{\Gamma}, \boldsymbol{v})$ and $\mathcal{C}\,(D, B, \boldsymbol{\Gamma})$ from (1). When we have Lipschitz constraints on convex functions, we will see that the situation for forming brackets for $\mathcal{C}\,(D, 1, \boldsymbol{\Gamma})$ with $D \subseteq [0,1]^d$ is essentially the same as for forming brackets for $\mathcal{C}\left([0,1]^d, 1, \boldsymbol{\Gamma}\right)$.

Theorem 3.2 from [29] gives the result of Theorem 2.1, stated below, when $D = \prod_{i=1}^d [a_i, b_i]$; we now extend it in Theorem 2.1 to the case of a general D. When we consider convex functions without Lipschitz constraints, we will partition D into sets that are contained in parallelotopes and apply Theorem 2.1 to those sets.

Theorem 2.1. *Let $a_i < b_i$ and let $D \subset \prod_{i=1}^d [a_i, b_i]$ be a convex set. Let $\boldsymbol{\Gamma} = (\Gamma_1, \ldots, \Gamma_d)$ and $0 < B, \Gamma_1, \ldots, \Gamma_d < \infty$. Then there exists a positive constant $c \equiv c_d$ such that*

$$\log N_{[\,]}\left(\epsilon\, \mathrm{Vol}_d(D)^{1/p}, \mathcal{C}\,(D, B, \boldsymbol{\Gamma}), L_p\right) \le \log N_{[\,]}\left(\epsilon, \mathcal{C}\,(D, B, \boldsymbol{\Gamma}), L_\infty\right) \tag{6}$$

$$\leq c\epsilon^{-d/2}\left(B+\sum_{i=1}^{d}\Gamma_i(b_i-a_i)\right)^{d/2} \tag{7}$$

for $\epsilon > 0$ *and* $p \geq 1$.

Here, $\mathrm{Vol}_d(D)$ is d-dimensional volume (Lebesgue measure) of the set D. The proof is given in [13]; we leave it out here due to space limitations.

3. Bracketing without lipschitz constraints

In the previous section we bounded bracketing entropy for classes of functions with Lipschitz constraints. In this section we remove those Lipschitz constraints. With Lipschitz constraints we could consider arbitrary domains D, but without the Lipschitz constraints we need more restrictions: now we will take D to be a *simple polytope* (defined below). We now define notation and assumptions we will use for the remainder of the document.

3.1. Notation and terminology

For $y, z \in \mathbb{R}^d$ let $\langle y, z\rangle := \sum_{i=1}^d y_i z_i$, let $\|z\|^2 := \langle z, z\rangle$, and for two sets $C, D \subset \mathbb{R}^d$, define the Hausdorff distance between them by

$$l_H(C,D) := \max\left(\sup_{x\in D}\inf_{y\in C}\|x-y\|, \sup_{y\in C}\inf_{x\in D}\|x-y\|\right).$$

Let $B_d(z,R) \equiv B(z,R) := \left\{x \in \mathbb{R}^d : \|x-z\| \leq R\right\}$.

We will consider only the case $d \geq 2$ since the result when $d = 1$ is given in [18]. Recall that for a convex set G, a set $F \subset G$ is a *face* of G if F is either $\emptyset$ (the empty set), G, or if $F = G \cap H$ for some supporting hyperplane H [37] of G. A set $F \subset G$ is a *facet* of G if F is a $(d-1)$-dimensional face (see e.g., [28]). We will focus on *simple* polytopes first (see Assumption 1). A simple polytope is one in which all $(d-k)$-dimensional faces (abbreviated "$(d-k)$-faces") of D have exactly k incident facets for $k \in \{0, \ldots, d\}$. The simple polytopes are dense in the class of all polytopes in the Hausdorff distance (page 82 of [28]). Any convex polytope can be triangulated into $O(n^{\lceil d/2\rceil})$ simplices (which are simple polytopes) if the polytope has n vertices (see e.g. [11]), and so we can translate our theorem into a result for a general polytope D; see Corollary 3.7. For two sets A and B let $A + B := \{a + b : a \in A, b \in B\}$. For a vector $v \in \mathbb{R}^d$, we let $[0, v] := \{\lambda v : \lambda \in [0, 1]\}$. For a set G, let $d^+(x, G, e) := \inf\{K \geq 0 : (x + Ke) \cap G \neq \emptyset\}$ (which may in general be infinite). For a point x, a set H, and a unit vector v, let

$$d(x,H,v) := \inf\{|k| : x + kv \in H\} = \min\left(d^+(x,H,v), d^+(x,H,-v)\right)$$

be the distance from x to H along the vector v, and for a set E, let $d(E, H, v) := \inf_{x\in E} d(x, H, v)$. We let ∂G be the boundary of G in $\mathbb{R}^d$ and we let $\partial_r G$ be the *relative boundary* of G, the set difference between the closure of G and the relative interior of G (e.g., page 44 of [37]). Let $\mathrm{Vol}_{d-k}(G)$ be the $(d-k)$-dimensional volume of G (and, in particular, $\mathrm{Vol}_0(G)$ is the number of elements in G).[2] For $a, b \in \mathbb{R}$, we let $a \vee b$ be the

[2] In general, Vol_{d-k} can be defined rigorously using the so-called $(d-k)$-dimensional Hausdorff measure. We will only need the $(d-k)$-dimensional volume of polytopes contained in affine spaces, and in such cases the definition is straightforward (and only requires Lebesgue measure).

maximum of a and b, and $a \wedge b$ be the minimum of a and b. For two vectors $e, v \in \mathbb{R}^d$ and a linear subspace V of $\mathbb{R}^d$, we write $e \perp v$ if $\langle e, v \rangle = 0$, we write $e \perp V$ if $e \perp v$ for all $v \in V$, and we let $V^\perp$ be the orthogonal complement linear subspace of V in $\mathbb{R}^d$.

3.2. Definitions and assumptions

In what follows, we will assume that D is a polytope, meaning that for some $N \in \mathbb{N}$, $D = \cap_{j=1}^N E_j$ where $E_j := \left\{x \in \mathbb{R}^d : \langle v_j, x \rangle \ge p_j \right\}$ are halfspaces with inner normal unit vectors v_j such that $v_i \neq v_j$ if $i \neq j$, and where $p_j \in \mathbb{R}$, for $j = 1, \dots, N$. Let $H_j := \left\{x \in \mathbb{R}^d : \langle x, v_j \rangle = p_j \right\}$ be the corresponding hyperplanes and let $F_j := H_j \cap D$ be the corresponding facets of D. For $k \in \{0, \dots, d\}$, we will define J_k to index the $(d-k)$-faces of D. First let $\tilde{J}_k := \left\{(j_1, \dots, j_k) \in \{1, \dots, N\}^k : j_1 < \dots < j_k \right\}$, and for $\boldsymbol{j} \in \tilde{J}_k$, let

$$G_{\boldsymbol{j}} = \cap_{\alpha=1}^k H_{j_\alpha} \cap D \text{ if } k \neq 0, \qquad \text{and let} \qquad G_{\boldsymbol{j}} = D \text{ if } k = 0.$$

Now let $J_0 = \{1\}$, and for $k \in \{1, \dots, d\}$, let $J_k := \left\{ \boldsymbol{j} \in \tilde{J}_k : G_{\boldsymbol{j}} \neq \emptyset \right\}$. The face $G_{\boldsymbol{j}}$, $\boldsymbol{j} \in J_k$, is $(d-k)$-dimensional and $H_{j_1} \cap D, \dots, H_{j_k} \cap D$ are the only facets of D containing $G_{\boldsymbol{j}}$, by Theorem 12.14 of [6]. Thus, by John's theorem, Theorem 5.22 [31], see also [3] or [4], there exists $x_{\boldsymbol{j}} \in G_{\boldsymbol{j}}$ such that $G_{\boldsymbol{j}} - x_{\boldsymbol{j}}$ contains a $(d-k)$-dimensional ellipsoid $A_{\boldsymbol{j}} - x_{\boldsymbol{j}}$ of maximal $(d-k)$-dimensional volume and such that

$$A_{\boldsymbol{j}} - x_{\boldsymbol{j}} \subset G_{\boldsymbol{j}} - x_{\boldsymbol{j}} \subset d(A_{\boldsymbol{j}} - x_{\boldsymbol{j}}). \tag{8}$$

Let $\gamma_{\boldsymbol{j},\alpha}/2 := d^+(x_{\boldsymbol{j}}, \partial_r A_{\boldsymbol{j}}, e_\alpha)$ be the radius of $A_{\boldsymbol{j}}$ in the direction $e_{\boldsymbol{j},\alpha}$, where $e_{\boldsymbol{j},k+1}, \dots, e_{\boldsymbol{j},d}$ are the orthonormal unit vectors given by the axes of the ellipsoid $A_{\boldsymbol{j}} - x_{\boldsymbol{j}}$. Let $E_{\boldsymbol{j}} := \operatorname{span}\left\{e_{\boldsymbol{j},k+1}, \dots, e_{\boldsymbol{j},d}\right\}$ be the linear space containing $G_{\boldsymbol{j}} - x_{\boldsymbol{j}}$. Let A be an integer and u a positive real number, and let

$$0 = \delta_0 < \delta_1 < \dots < \delta_A < u < \delta_{A+1} < \delta_{A+2} = \infty \tag{9}$$

be a sequence. This sequence as well as A and u will be specified in greater detail later. For $k \in \{1, \dots, d\}$, let $I_k := \{0, \dots, A\}^k$, and let $I_0 := \{A\}$. For $k \in \{1, \dots, d\}$, $\boldsymbol{i} = (i_1, \dots, i_k) \in I_k$, and $\boldsymbol{j} = (j_1, \dots, j_k) \in J_k$ let

$$G_{\boldsymbol{i},\boldsymbol{j}} := \left\{x \in D : \delta_{i_\alpha} \le d(x, H_{j_\alpha}, v_{j_\alpha}) \le \delta_{i_\alpha+1} \text{ for } \alpha = 1, \dots, N \right\} \tag{10}$$

where in the previous display for $\alpha > k$ we let $i_\alpha = A + 1$ and j_α take on the values in $\{1, \dots, N\} \setminus \{j_1, \dots, j_k\}$ (in any order). For the $k = 0$ case, let $G_{A,1} := \{x \in D : d(x, \partial D) \ge u\}$. These sets are not parallelotopes, since for $\alpha > k$, $\delta_{i_\alpha+1} = \infty$. However, for any $x \in G_{\boldsymbol{j}}$, $(G_{\boldsymbol{i},\boldsymbol{j}} - x) \cap \operatorname{span}\left\{v_{j_1}, \dots, v_{j_\beta}\right\}$, for $\beta \le k$, is contained in a β-dimensional parallelotope by construction; this will be used to understand the volume of $G_{\boldsymbol{i},\boldsymbol{j}}$. We will eventually define u such that $D \subset \bigcup_{k=0}^d \bigcup_{\boldsymbol{j} \in J_k, \boldsymbol{i} \in I_k} G_{\boldsymbol{i},\boldsymbol{j}}$ (see Lemma 3.3).

The setup for our first main results is summarized in the following assumption.

Assumption 1. Let $d \ge 2$, let the definitions of the above Section 3.2 hold, and let $D \subset \mathbb{R}^d$ be a simple convex polytope.

Additionally, define the support function for a convex set D to be, for $x \in \mathbb{R}^d$ with $\|x\| = 1$, $h(D, x) := \max_{d \in D} \langle d, x \rangle$. Then the width function is, for $\|u\| = 1$, $w(D, u) := h(D, u) + h(D, -u)$, which gives the distance between supporting hyperplanes of D with inner normal vectors u and $-u$, respectively, and let $\operatorname{diam}(D) := \sup_{\|u\|=1} w(D, u)$ be the diameter of D.

3.3. Main results

We want to bound the slope of functions $f \in \mathcal{C}(D,1)|_{G_{i,j}}$, so that we can apply bracketing bounds on convex function classes with Lipschitz bounds. Note that each $G_{i,j}$ is distance δ_{i_α} in the direction of v_{j_α} from H_{j_α}, which means that if $f \in \mathcal{C}(D,1)|_{G_{i,j}}$ then f has Lipschitz constant bounded by $2/\delta_{i_\alpha}$ along the direction v_{j_α}. However, the vectors v_{j_α} are not orthonormal, so the distance from $G_{i,j}$ along v_{j_α} to a hyperplane other than H_{j_α} may be smaller than δ_{i_α}.

Note that if $P \subset R \subset \mathbb{R}^d$ where R is a hyperrectangle and P is a parallelotope defined by vectors $v_1, \ldots, v_d$, then if A is a linear map with $v_1, \ldots, v_d$ as its eigenvectors (thus rescaling P), then AR will not necessarily still be a hyperrectangle, i.e. its axes may no longer be orthogonal. Thus, we cannot argue by simple scaling arguments that bracketing numbers for P scale with the lengths along the vectors v_i.

For each $G_{i,j}$ we will find an orthonormal basis such that $G_{i,j}$ is contained in a rectangle R whose axes are given by the basis and whose lengths along those axes (i.e., widths) are bounded by a constant times the width of one of the normal vectors v_{j_α}. Furthermore, the distance from R along each basis vector to ∂D will be bounded by the distance from $G_{i,j}$ along v_{j_α} to H_{j_α}. This will give us control of both the Lipschitz parameters and the widths corresponding to the basis, and thus control of the bracketing number for classes of convex functions. We rely on the following basic lemma.

Lemma 3.1. *If $f \in \mathcal{C}(D,B)$, $B > 0$, and $x \in D$ is such that $d(x, \partial D, e_\alpha) \geq \delta > 0$ then*

$$\left| \frac{\partial}{\partial x_i} f(x) \right| \leq \frac{2B}{\delta} \tag{11}$$

where the derivative stands for both the right and left derivative of f.

Proof. Let $z_1 = x - \gamma_1 e_\alpha$ and $z_2 = x + \gamma_2 e_\alpha$, $\gamma_1, \gamma_2 > 0$, both be elements of ∂D, so that by convexity we have for any $h \in [-\gamma_1, \gamma_2]$,

$$\frac{-2B}{\delta} \leq \frac{f(z_1) - f(z_1 + \delta e_\alpha)}{\delta} \leq \frac{f(x + h e_\alpha) - f(x)}{h} \leq \frac{f(z_2) - f(z_2 - \delta e_\alpha)}{\delta} \leq \frac{2B}{\delta}.$$

Thus, f satisfies a Lipschitz constraint in the direction of e_α. □

The following proposition constructs a basis and gives control for the basis elements in span $\{G_j\}$. For the basis elements perpendicular to span $\{G_j\}$, control is given by Lemmas 4.3 and 4.4 in Section 4.

Proposition 3.2. *Let Assumption 1 hold for a convex polytope D. For each $k \in \{0, \ldots, d\}$, $\boldsymbol{i} \in I_k$, $\boldsymbol{j} \in J_k$, and each $G_{\boldsymbol{i},\boldsymbol{j}}$, there is an orthonormal basis $\boldsymbol{e}_{\boldsymbol{i},\boldsymbol{j}} \equiv \boldsymbol{e} := (e_1, \ldots, e_d)$ of $\mathbb{R}^d$ such that for any $f \in \mathcal{C}(D,B)|_{G_{i,j}}$, f has Lipschitz constant $2B/\delta_{i_\alpha}$ in the direction e_α, where $\delta_{i_\alpha} = \delta_{A+1}$ if $k+1 \leq \alpha \leq d$. Furthermore, there exists a permutation π of $(1, \ldots, k)$ such that for $\alpha = 1, \ldots, k$, $e_{\boldsymbol{i},\boldsymbol{j},\alpha} \equiv e_\alpha$ satisfies*

$$e_\alpha \in \operatorname{span}\left\{v_{j_{\pi(1)}}, \ldots, v_{j_{\pi(\alpha)}}\right\}, \; e_\alpha \perp \operatorname{span}\left\{v_{j_{\pi(1)}}, \ldots, v_{j_{\pi(\alpha-1)}}\right\}, \;\; \text{and} \; \left\langle e_\alpha, v_{j_{\pi(\alpha)}} \right\rangle > 0, \tag{12}$$

and for $\alpha \in \{k+1, \ldots, d\}$, $e_\alpha \perp \operatorname{span}\left\{v_{j_{\pi(1)}}, \ldots, v_{j_{\pi(k)}}\right\} =: V$. In particular, we may take $e_{k+1}, \ldots, e_d$ to be the orthonormal unit axis vectors of $A_{\boldsymbol{j}} - x_{\boldsymbol{j}}$ as defined on Section 3.2. Thus it is immediate that neither V nor $V^\perp$ depends on $\boldsymbol{i}$.

Proof. Without loss of generality, for ease of notation we assume in this proof that $j_\beta = \beta$ for $\beta = 1, \dots, k$, and then that

$$\delta_{i_1} \le \delta_{i_2} \le \cdots \le \delta_{i_k} \le \delta_{i_{k+1}} = \cdots = \delta_{i_N},$$

where we let $i_\alpha = A + 1$ for $k < \alpha \le N$. That is, we assume that $H_1, \dots, H_k$ are the nearest hyperplanes to $G_{\boldsymbol{i},\boldsymbol{j}}$, in order of increasing distance; we then take π to be the identity. To define the orthonormal basis vectors, we will use a Gram–Schmidt orthonormalization, proceeding according to increasing distances from $G_{\boldsymbol{i},\boldsymbol{j}}$ to the hyperplanes H_j. Define $e_1 := v_1$ and for $1 < j \le k$, define e_j inductively by

$$e_j \in \operatorname{span}\left\{v_1, \dots, v_j\right\}, \; e_j \perp \operatorname{span}\left\{v_1, \dots, v_{j-1}\right\}, \; \left\langle e_j, v_j \right\rangle > 0, \text{ and } \|e_j\| = 1.$$

Let $e_{k+1}, \dots, e_d$ be orthonormal unit vectors given by the axes of the ellipsoid $A_{\boldsymbol{j}} - x_{\boldsymbol{j}}$. Note that these vectors form an orthonormal basis of $\operatorname{span}\{v_1, \dots, v_k\}^\perp$ because $\operatorname{span}\{e_{k+1}, \dots, e_d\} = \operatorname{span}(G_{\boldsymbol{j}} - x_{\boldsymbol{j}})$ is perpendicular to $\operatorname{span}\{v_1, \dots, v_k\}$ by definition. For $\alpha \in \{1, \dots, k\}$, for any $x \in G_{\boldsymbol{i},\boldsymbol{j}}$, since $d(x, H_\alpha, v)$ achieves its minimum when v is v_α,

$$\begin{aligned} &d(x, H_\alpha, e_\alpha) \ge d(x, H_\alpha, v_\alpha) \ge \delta_{i_\alpha}, \\ &d(x, H_j, e_\alpha) \ge d(x, H_j, v_j) \ge \delta_{i_j} \ge \delta_{i_\alpha}, \text{ for all } N \ge j > \alpha, \text{ and} \\ &d(x, H_j, e_\alpha) = \infty > \delta_{i_\alpha} \text{ for } j < \alpha, \end{aligned}$$

since $e_\alpha \perp \operatorname{span}\{v_1, \dots, v_{\alpha-1}\}$. Similarly, for $\alpha \in \{k+1, \dots, d\}$,

$$\begin{aligned} &d(x, H_j, e_\alpha) \ge d(x, H_j, v_j) \ge \delta_{A+1}, \text{ for all } N \ge j \ge k+1, \text{ and} \\ &d(x, H_j, e_\alpha) = \infty > \delta_{A+1} \text{ for } j \le k, \end{aligned}$$

since $e_\alpha \perp \operatorname{span}\{v_1, \dots, v_k\}$. Thus, we have $d(G_{\boldsymbol{i},\boldsymbol{j}}, H_j, e_\alpha) \ge \delta_{i_\alpha}$ for $\alpha \in \{1, \dots, d\}$ and for $j \in \{1, \dots, N\}$. That is, we have shown

$$d(G_{\boldsymbol{i},\boldsymbol{j}}, \partial D, e_\alpha) \ge \delta_{i_\alpha} \text{ for all } \alpha \in \{1, \dots, d\}. \tag{13}$$

Thus by (11), f has Lipschitz bound $2B/\delta_{i_\alpha}$ in the direction e_α. □

The next lemma is necessary for us to be able to apply (5). To state it, we first define some constants. For $k \in \{1, \dots, d\}$, let $d_{i,j,k} := d(E_i, F_j)$ where E_i, $i = 1, \dots, N_k$, is a $(d-k)$-face and F_j, $j = 1, \dots, N$, is a facet. Then let

$$r_D := \min\left\{d_{i,j,k} : d_{i,j,k} \ne 0, k \in \{1, \dots, d\}\right\} > 0. \tag{14}$$

Let

$$u \equiv u_D := r_D/2 \wedge 2^{-2(p+1)^2(p+2)} \wedge \min_{k \in \{1,\dots,d-1\}} \min_{j \in J_k, e \in E_{\boldsymbol{j}}} \frac{d^+(x_{\boldsymbol{j}}, \partial_r G_{\boldsymbol{j}}, e)}{L_{k,2}} \tag{15}$$

where for $k \in \{1, \dots, d-1\}$,

$$L_{k,2} := 1 \vee \max_{\boldsymbol{j} \in J_k} \max_{i \in \{1,\ldots,N\} \setminus \boldsymbol{j}} \sum_{\gamma=1}^{k} \frac{\left\langle \tilde{f}_{\boldsymbol{j},\gamma}, v_i \right\rangle}{\left\langle \tilde{f}_{\boldsymbol{j},\gamma}, v_{j_\gamma} \right\rangle}, \tag{16}$$

where $\tilde{f}_{\boldsymbol{j},\gamma}$ are defined in Proposition 4.2, and $E_{\boldsymbol{j}}$ is defined on Section 3.2.

Lemma 3.3. *Under Assumption 1, with u given in (15), we have*

$$D \subset \bigcup_{k=0}^{d} \bigcup_{\boldsymbol{j} \in J_k, \boldsymbol{i} \in I_k} G_{\boldsymbol{i},\boldsymbol{j}}.$$

Proof. Fix $x \in D$. We need to show that there are no more than d facets F such that $d(x, F) < u$. If $d(x, \partial D) \geq u$ then $x \in G_{A,1}$ (corresponding to $k = 0$), so we assume $d(x, \partial D) < u$. Then let $k_x := \max\{k \in \{1, \ldots, d\} : d(x, G) < u, \text{ some } (d-k)\text{-face } G\}$ and let G_x be any $(d - k_x)$-face such that the minimum is attained. Now for any facet F, if $d(x, F) < u$ then we also have $d(G_x, F) < 2u \leq r_D$. But this contradicts the definition of r_D unless $d(G_x, F) = 0$. Because G_x is nonempty, $G_x = G_{\boldsymbol{j}}$ for some $\boldsymbol{j} \in J_{k_x}$ (rather than $\boldsymbol{j} \in \tilde{J}_{k_x} \setminus J_{k_x}$). The distance from x to the boundary of G_x is no smaller than u, because otherwise we would contradict the maximality defining k_x since the boundary is given by $(d - (k_x + 1))$-faces. Thus the distance from x to any facet intersecting but not containing G_x is no smaller than u. Furthermore because D is simple, there are exactly $k_x \leq d$ facets containing G_x; and we have shown that the distance to every facet excluding these k_x is no smaller than u. Thus, G_x is unique and x lies in $G_{\boldsymbol{i},\boldsymbol{j}}$ for some $\boldsymbol{i} \in I_{k_x}$. □

The next lemma combines Lemmas 4.3 and 4.4 with Theorem 2.1. The statement depends on the constants $L_{k,1}$, $k \in \{1, \ldots, d\}$, and $L_{\boldsymbol{j},4}$, $\boldsymbol{j} \in J_k$. These depend only on D and are defined in (24) and (45).

Lemma 3.4. *Let Assumption 1 hold. Fix $k \in \{1, \ldots, d\}$, $\boldsymbol{i} \in I_k$, $\boldsymbol{j} \in J_k$. Then for any $p \geq 1$ and for $\epsilon > 0$,*

$$\begin{aligned} &\log N_{[\,]}\left(\epsilon \operatorname{Vol}_d(G_{\boldsymbol{i},\boldsymbol{j}})^{1/p}, \mathcal{C}(D,1)|_{G_{\boldsymbol{i},\boldsymbol{j}}}, L_p\right) \\ &\quad \leq c_d \epsilon^{-d/2} \left(1 + \frac{2d^2}{L_{\boldsymbol{j},4}} \max_{\alpha=1,\ldots,k} \frac{\delta_{i_\alpha+1}}{\delta_{i_\alpha}} + \sum_{\alpha=k+1}^{d} \frac{8 L_{k,1} \rho_{\boldsymbol{j},\alpha}}{u}\right)^{d/2}. \end{aligned} \tag{17}$$

Proof. Let

$$\boldsymbol{\Gamma}_{\boldsymbol{i}} := \left(\frac{2}{d(G_{\boldsymbol{i},\boldsymbol{j}}, \partial D, e_1)}, \ldots, \frac{2}{d(G_{\boldsymbol{i},\boldsymbol{j}}, \partial D, e_k)}, \frac{2}{u}, \ldots, \frac{2}{u}\right) \tag{18}$$

where $e_{\boldsymbol{i},\boldsymbol{j},\alpha} \equiv e_\alpha$, $\alpha = 1, \ldots, d$, is given by Proposition 3.2. Then

$$\mathcal{C}(D,1)|_{G_{\boldsymbol{i},\boldsymbol{j}}} \subset \mathcal{C}\left(G_{\boldsymbol{i},\boldsymbol{j}}, 1, \boldsymbol{\Gamma}_{\boldsymbol{i}}, \boldsymbol{e}\right) \tag{19}$$

where $\boldsymbol{e} = (e_1, \ldots, e_d)$. Let $\tilde{f}_{j_\gamma}$ be given by Lemma 4.1 applied to the k linearly independent unit normal vectors $v_{j_1}, \ldots, v_{j_k}$, and (as in that lemma, with "d_β" given by $(\delta_{i_\gamma+1} - \delta_{i_\gamma})$), let

$$f_{\boldsymbol{i},\boldsymbol{j},j_\gamma} \equiv f_{j_\gamma} := (\delta_{i_\gamma+1} - \delta_{i_\gamma}) \tilde{f}_{j_\gamma} / \left\langle \tilde{f}_{j_\gamma}, v_{j_\gamma} \right\rangle. \tag{20}$$

Let $P_{i,j} := \sum_{\gamma=1}^{k}[0, f_{j_\gamma}]$, where $[0,v] := \{\lambda v : \lambda \in [0,1]\}$. By Lemma 4.3, $P_{i,j} \subset \sum_{\alpha=1}^{k}[0, \gamma_\alpha e_\alpha]$ where γ_α are given by the lemma. Thus by (53), for some $x \in G_{i,j}$,

$$G_{i,j} \subset x + \sum_{\alpha=1}^{k}[0, \gamma_\alpha e_\alpha] + \sum_{\alpha=k+1}^{d}\left[-2L_{k,1}\rho_{j,\alpha}e_\alpha, 2L_{k,1}\rho_{j,\alpha}e_\alpha\right]. \tag{21}$$

Now, using (19), we apply Theorem 2.1 to see

$$\begin{aligned}&\log N_{[\,]}\left(\epsilon \operatorname{Vol}_d(G_{i,j})^{1/p}, \mathcal{C}(D,1)\,|_{G_{i,j}}, L_p\right)\\ &\le c_d \epsilon^{-d/2}\left(1 + \sum_{\alpha=1}^{k}\frac{2\gamma_\alpha}{d(G_{i,j}, \partial D, e_\alpha)} + \sum_{\alpha=k+1}^{d}\frac{8L_{k,1}\rho_{j,\alpha}}{u}\right)^{d/2}\end{aligned} \tag{22}$$

Now by applying (55), (61), and (62) with $v = e_\alpha$, we see that

$$\frac{2\gamma_\alpha}{d(G_{i,j}, \partial D, e_\alpha)} \le \frac{2d \operatorname{diam}(G_{i,j}, e_\alpha)}{d(G_{i,j}, \partial D, e_\alpha)} \le \frac{2d \min_{\beta=1,\ldots,k}\frac{\delta_{i_\beta+1}}{\left|\langle e_\alpha, v_{j_\beta}\rangle\right|}}{\max_{\beta=1,\ldots,k}\frac{\delta_{i_\beta}}{\left|\langle e_\alpha, v_{j_\beta}\rangle\right|}} \le \frac{2d}{L_{j,4}}\max_{\beta=1,\ldots,k}\frac{\delta_{i_\beta+1}}{\delta_{i_\beta}} \tag{23}$$

where

$$L_{j,4} := \min_{e_1,\ldots,e_d}\ \min_{v_{j_\beta} : \langle v_{j_\beta}, e_\alpha\rangle > 0}\left|\langle e_\alpha v_{j_\beta}\rangle\right|. \tag{24}$$

(We can restrict to v_{j_β} such that $\langle v_{j_\beta}, e_\alpha\rangle > 0$ in the definition of $L_{j,4}$ because the numerator in (23) is finite.) Thus (22) is bounded above by

$$c_d\epsilon^{-d/2}\left(1 + \frac{2d^2}{L_{j,4}}\max_{\beta=1,\ldots,k}\frac{\delta_{i_\beta+1}}{\delta_{i_\beta}} + \sum_{\alpha=k+1}^{d}\frac{8L_{k,1}\rho_{j,\alpha}}{u}\right)^{d/2}. \quad \square$$

Now we present our main theorem. It gives a bracketing entropy of order $\epsilon^{-d/2}$ when D is a fixed simple polytope. Its proof relies on embedding $G_{i,j}$ in a set $R_{i,j}$ (defined in (52)) which is a set-sum of a parallelotope and a hyperrectangle with axes given by Proposition 3.2. We need to control the distance of $G_{i,j}$ to ∂D, and we need to control the size of $R_{i,j}$ in terms of the widths along its axes. Then we can use the results of Section 2 on $R_{i,j}$ and thus on $G_{i,j}$. We defer some statements and proofs of needed facts about $G_{i,j}$ and $R_{i,j}$ until Section 4.

Theorem 3.5. *Let Assumption 1 hold for a convex polytope $D \subseteq \prod_{i=1}^{d}[a_i, b_i]$. Fix $p \ge 1$. Then for all $\epsilon > 0$,*

$$\log N_{[\,]}\left(\epsilon, \mathcal{C}(D,B), L_p\right) \le S\epsilon^{-d/2}\left(B\left(\prod_{i=1}^{d}(b_i - a_i)\right)^{1/p}\right)^{d/2}, \tag{25}$$

where S is a constant depending only on d and D.

The form of the constant S is given in the proof of the theorem.

Proof. Fix $\epsilon > 0$. First, we will reduce to the case where $D \subset [0,1]^d$ and $B = 1$ by a scaling argument. Let C be an affine map from $\prod_{i=1}^{d}[a_i, b_i]$ to $[0,1]$, where $\tilde{D}$ is the image of D, and assume we have a bracketing cover $[\tilde{l}_1, \tilde{u}_1], \ldots, [\tilde{l}_N, \tilde{u}_N]$ of $\mathcal{C}\left(\tilde{D}, 1\right)$. Let $l_i := B\tilde{l}_i \circ C$ and similarly for u_i, so that $[l_1, u_1], \ldots, [l_N, u_N]$ form brackets for $\mathcal{C}(D, B)$. Their L_p^p size is

$$\int_D (u_i(x) - l_i(x))^p \, dx = B^p \int_{\tilde{D}} (\tilde{u}_i(x) - \tilde{l}_i(x))^p \prod^d (b_i - a_i) dx.$$

Thus,

$$N_{[\,]}\left(\epsilon B \left(\prod^d b_i - a_i\right)^{1/p}, \mathcal{C}(D, B), L_p\right) \leq N_{[\,]}\left(\epsilon, \mathcal{C}\left(\tilde{D}, 1\right), L_p\right),$$

so apply the theorem with $\eta = \epsilon / B\left(\prod^d b_i - a_i\right)^{1/p}$ for ϵ. Note that the constant S depends on $\tilde{D}$, the version of D normalized to lie in $[0, 1]^d$.

We now assume $D \subset [0, 1]^d$ and $B = 1$. We specify the sequence in (9) and $a_{\boldsymbol{i},k} \equiv a_{\boldsymbol{i}} > 0$, which will govern the L_p-sizes of our brackets on $G_{\boldsymbol{i},\boldsymbol{j}}$, as follows. Let

$$\delta_i := \exp\left\{ p \left(\frac{p+1}{p+2}\right)^{i-1} \log \epsilon \right\} \quad \text{for } i = 1, \ldots, A, \quad \text{and} \quad \delta_0 = 0. \tag{26}$$

Note that this implicitly defines A, by (9) and (15). For $k \in \{1, \ldots, d\}$ and $\boldsymbol{i} \in I_k$, we will let $a_{(i_1,\ldots,i_k)} = 2$ if $i_\alpha = 0$ for any $\alpha \in \{1, \ldots, k\}$, and otherwise we let

$$a_{(i_1,\ldots,i_k)} := \prod_{\beta=1}^k a_{i_\beta} := \prod_{\beta=1}^k \epsilon^{1/k} \exp\left\{ -p \frac{(p+1)^{i_\beta - 2}}{(p+2)^{i_\beta - 1}} \log \epsilon \right\}.$$

For the $k = 0$ case, let $a_A := \epsilon / u$. Let

$$a = \left(\sum_{k=0}^d \sum_{\boldsymbol{j} \in J_k, \boldsymbol{i} \in I_k} a_{\boldsymbol{i}}^p \operatorname{Vol}_d(G_{\boldsymbol{i},\boldsymbol{j}}) \right)^{1/p}. \tag{27}$$

Then since $D \subset \cup_{k=0}^d \cup_{\boldsymbol{j} \in J_k, \boldsymbol{i} \in I_k} G_{\boldsymbol{i},\boldsymbol{j}}$ by Lemma 3.3, as in (5),

$$\log N_{[\,]}\left(a, \mathcal{C}(D, 1), L_p\right) \leq \sum_{k=0}^d \sum_{\boldsymbol{j} \in J_k} \sum_{\boldsymbol{i} \in I_k} \log N_{[\,]}\left(a_{\boldsymbol{i}} \operatorname{Vol}_d(G_{\boldsymbol{i},\boldsymbol{j}})^{1/p}, \mathcal{C}(D, 1)\,|_{G_{\boldsymbol{i},\boldsymbol{j}}}, L_p\right). \tag{28}$$

First, consider the case $k \in \{1, \ldots, d\}$ and compute the sum over I_k for a fixed $\boldsymbol{j} \in J_k$. We use the trivial bracket $[-1, 1]$ for any $G_{\boldsymbol{i},\boldsymbol{j}}$ where $i_\alpha = 0$ for any $\alpha \in \{1, \ldots, k\}$. Otherwise apply Lemma 3.4 which shows us that the sum over the remaining terms in (28) is bounded by

$$\sum_{i_1=1}^A \cdots \sum_{i_k=1}^A c_d a_{\boldsymbol{i}}^{-d/2} \left(1 + \frac{2d^2}{L_{\boldsymbol{j},4}} \max_{\alpha=1,\ldots,k} \frac{\delta_{i_\alpha+1}}{\delta_{i_\alpha}} + \sum_{\alpha=k+1}^d \frac{8 L_{k,1} \rho_{\boldsymbol{j},\alpha}}{u} \right)^{d/2}. \tag{29}$$

Since $L_{k,1} \geq 1$ and $u \leq \rho_{\boldsymbol{j},\alpha}$ by (15) for all $k, \boldsymbol{i}, \boldsymbol{j}$ and $\alpha = k+1, \ldots, d$, we have $\sum_{\alpha=k+1}^d \frac{8\rho_{\boldsymbol{j},\alpha} L_{k,1}}{u} = 4L_{k,1} \sum_{\alpha=k+1}^d \frac{2\rho_{\boldsymbol{j},\alpha}}{u} \leq 4L_{k,1} \prod_{\alpha=k+1}^d \frac{2\rho_{\boldsymbol{j},\alpha}}{u}$ (using the fact that for $a, b \geq 2$, $ab \geq a + b$). We also bound $\max_{\alpha=1,\ldots,k} 2\delta_{i_\alpha+1}/\delta_{i_\alpha} \leq \prod_{\alpha=1}^k 2\delta_{i_\alpha+1}/\delta_{i_\alpha}$ since $2\delta_{i_\alpha+1}/\delta_{i_\alpha} > 2$. Thus (29) is bounded above by

$$c_d d^2 L_{\boldsymbol{j},4}^{-1} \left(1 + 2^{d-k+2} L_{k,1} \prod_{\alpha=k+1}^d \frac{\rho_{\boldsymbol{j},\alpha}}{u} \right)^{d/2} \sum_{i_1=1}^A \cdots \sum_{i_k=1}^A a_{\boldsymbol{i}}^{-d/2} \prod_{\alpha=1}^k \left(\frac{2\delta_{i_\alpha+1}}{\delta_{i_\alpha}} \right)^{d/2}, \tag{30}$$

which is

$$c_d d^2 L_{\boldsymbol{j},4}^{-1}\left(1+2^{d-k+2}L_{k,1}\prod_{\alpha=k+1}^{d}\frac{\rho_{\boldsymbol{j}\alpha}}{u}\right)^{d/2}\sum_{i_1=1}^{A}\cdots\sum_{i_k=1}^{A}\prod_{\beta=1}^{k}\left(\frac{2\delta_{i_\beta+1}}{\delta_{i_\beta}a_{i_\beta}}\right)^{d/2}. \tag{31}$$

Note that when $k=d$ we take the product over an empty set to be 1. For $i=1,\ldots,A$, let

$$\zeta_i \equiv \zeta_{i,k} := \sqrt{\epsilon^{1/k}\delta_{i+1}/(\delta_i a_i)}, \tag{32}$$

so that $\sum_{i_1=1}^{A}\cdots\sum_{i_k=1}^{A}\prod_{\beta=1}^{k}\left(\frac{2\delta_{i_\beta+1}}{\delta_{i_\beta}a_{i_\beta}}\right)^{d/2}$ equals

$$\begin{aligned}\sum_{i_1=1}^{A}\cdots\sum_{i_k=1}^{A}2^{kd/2}\epsilon^{-d/2}\prod_{\beta=1}^{k}\zeta_{i_\beta}^{d} &= 2^{kd/2}\epsilon^{-d/2}\sum_{i_1=1}^{A}\zeta_{i_1}^{d}\sum_{i_2=1}^{A}\zeta_{i_2}^{d}\cdots\sum_{i_k=1}^{A}\zeta_{i_k}^{d}\\ &= \epsilon^{-d/2}2^{kd/2}B_u^k\end{aligned}$$

where, for $0<\epsilon\le 1$

$$B_u := \sum_{i=1}^{A}\zeta_i^d \le 2u^{d/(2(p+1)(p+2))}, \tag{33}$$

by Lemma 3.6.

Next, we will relate the term $\left(1+2^{d-k+2}L_{k,1}\prod_{\alpha=k+1}^{d}\frac{\rho_{\boldsymbol{j}\alpha}}{u}\right)^{d/2}$ to $\mathrm{Vol}_{d-k}(G_{\boldsymbol{j}})$. Recall that $A_{\boldsymbol{j}}$ is the ellipsoid defined in (8) which has diameter in the e_α direction given by $\gamma_{\boldsymbol{j},\alpha}$. By (8), $\rho_{\boldsymbol{j},\alpha}\le d\gamma_{\boldsymbol{j},\alpha}$. The volume of $A_{\boldsymbol{j}}$ is $\mathrm{Vol}_{d-k}(A_{\boldsymbol{j}})=\left(\prod_{\alpha=k+1}^{d}\gamma_{\boldsymbol{j},\alpha}/2\right)\pi^{(d-k)/2}/\Gamma((d-k)/2+1)$. Thus, letting $C_d := \frac{(2d)^{d-k}\Gamma((d-k)/2+1)}{\pi^{(d-k)/2}}$, we have

$$\prod_{\alpha=k+1}^{d}\rho_{\boldsymbol{j},\alpha}\le C_d\,\mathrm{Vol}_{d-k}(A_{\boldsymbol{j}})\le C_d\,\mathrm{Vol}_{d-k}(G_{\boldsymbol{j}}).$$

Thus we have shown that (31) is bounded above by

$$c_d d^2 L_{\boldsymbol{j},4}^{-1}2^{kd/2}\left(1+2^{d-k+2}L_{k,1}u^{-(d-k)}C_d\,\mathrm{Vol}_{d-k}(G_{\boldsymbol{j}})\right)^{d/2}B_u^k\cdot\epsilon^{-d/2}. \tag{34}$$

Therefore, letting $\tilde{c}_{d,k}:=c_d d^2 2^k 2^{kd/2}$, we have shown that

$$\begin{aligned}&\sum_{\boldsymbol{i}\in I_k}\log N_{[\,]}\left(a_{\boldsymbol{i}}\,\mathrm{Vol}_d(G_{\boldsymbol{i},\boldsymbol{j}})^{1/p},\mathcal{C}(D,1)|_{G_{\boldsymbol{i},\boldsymbol{j}}},L_p\right)\\ &\quad\le L_{\boldsymbol{j},4}^{-1}\tilde{c}_{d,k}u^{kd/2(p+1)(p+2)}\left(1+2^{d-k+2}L_{k,1}u^{-(d-k)}C_d\,\mathrm{Vol}_{d-k}(G_{\boldsymbol{j}})\right)^{d/2}\epsilon^{-d/2}.\end{aligned} \tag{35}$$

Display (35) holds for $k\in\{1,\ldots,d\}$. When $k=0$, recalling $a_A=\epsilon/u$, we have

$$\log N_{[\,]}\left(a_A\,\mathrm{Vol}_d(G_{A,1})^{1/p},\mathcal{C}(D,1)|_{G_{A,1}},L_p\right)\le c_d\,(u+2d)^{d/2}\epsilon^{-d/2} \tag{36}$$

by Theorem 2.1 since $\mathcal{C}(D,1)|_{G_{A,1}}\subset\mathcal{C}\left(G_{A,1},1,\frac{2}{u}\mathbf{1}\right)$ where $\mathbf{1}\in\mathbb{R}^d$ is the vector of all 1's. Then, combining (35) and (36), the cardinality of the collection of brackets covering the entire domain D is given by summing over $\boldsymbol{j}\in J_k$ and $k\in\{0,\ldots,d\}$.

We have computed the cardinality of the brackets. Now we bound their size. Let I_k^0 be the subset of $\boldsymbol{i} \in I_k$ such that some i_α is 0, and let $I_k^+ := I_k \setminus I_k^0$. We have

$$\begin{aligned} a^p &\le a_A^p \operatorname{Vol}_d(D) + \sum_{k=0}^{d} \sum_{\boldsymbol{j} \in J_k, \boldsymbol{i} \in I_k^0} 2^p \operatorname{Vol}_d(G_{\boldsymbol{i},\boldsymbol{j}}) \\ &\quad + \sum_{k=1}^{d} (2L_{k,1})^{d-k} \sum_{\boldsymbol{j} \in J_k} \operatorname{Vol}_{d-k}(G_{\boldsymbol{j}}) \sum_{\boldsymbol{i} \in I_k^+} a_{\boldsymbol{i}}^p \prod_{\alpha=1}^{k} \frac{\delta_{i_\alpha+1} - \delta_{i_\alpha}}{\left\langle \tilde{f}_\alpha, v_{j_\alpha} \right\rangle} \end{aligned} \tag{37}$$

by Proposition 4.2 with $\tilde{f}_\alpha \equiv \tilde{f}_{\boldsymbol{j},\alpha}$ defined there. Recalling $\delta_1 = \epsilon^p$, note that

$$\sum_{k=0}^{d} \sum_{\boldsymbol{j} \in J_k, \boldsymbol{i} \in I_k^0} 2^p \operatorname{Vol}_d(G_{\boldsymbol{i},\boldsymbol{j}}) \le 2^p \epsilon^p \operatorname{Vol}_{d-1}(\partial D). \tag{38}$$

Fixing $k \in \{1, \ldots, d\}$, we have

$$\begin{aligned} \sum_{\boldsymbol{j} \in J_k} \operatorname{Vol}_{d-k}(G_{\boldsymbol{j}}) \sum_{\boldsymbol{i} \in I_k^+} a_{\boldsymbol{i}}^p \prod_{\alpha=1}^{k} \frac{\delta_{i_\alpha+1} - \delta_{i_\alpha}}{\left\langle \tilde{f}_\alpha, v_{j_\alpha} \right\rangle} &\le \sum_{\boldsymbol{j} \in J_k} \operatorname{Vol}_{d-k}(G_{\boldsymbol{j}}) L_{\boldsymbol{j},3}^k \sum_{i_1=1}^{A} \cdots \sum_{i_k=1}^{A} \prod_{\alpha=1}^{k} a_{i_\alpha}^p \delta_{i_\alpha+1} \\ &\le \sum_{\boldsymbol{j} \in J_k} \operatorname{Vol}_{d-k}(G_{\boldsymbol{j}}) L_{\boldsymbol{j},3}^k \sum_{i_1=1}^{A} a_{i_1}^p \delta_{i_1+1} \cdots \sum_{i_k=1}^{A} a_{i_k}^p \delta_{i_k+1}. \end{aligned}$$

where $L_{\boldsymbol{j},3} := \max_{\alpha \in \{1,\ldots,k\}} 1/\left\langle \tilde{f}_{\boldsymbol{j},\alpha}, v_{j_\alpha} \right\rangle$. We have

$$\sum_{\alpha=1}^{A} a_\alpha^p \delta_{\alpha+1} = \epsilon^{p/k} \sum_{\alpha=1}^{A} \frac{\epsilon^{1/k} \delta_{\alpha+1}}{\delta_\alpha a_\alpha} = \epsilon^{p/k} \sum_{\alpha=1}^{A} \zeta_\alpha^2 =: \epsilon^{p/k} A_u, \tag{39}$$

where $A_u \le 2u^{1/(p+1)^2}$ by Lemma 3.6. Thus

$$\sum_{\boldsymbol{j} \in J_k} \operatorname{Vol}_{d-k}(G_{\boldsymbol{j}}) L_{\boldsymbol{j},3}^k \left(\sum_{i_1=0}^{A} a_{i_1}^p \delta_{i_1+1} \right) \cdots \left(\sum_{i_k=0}^{A} a_{i_k}^p \delta_{i_k+1} \right) \le \epsilon^p A_u^k \sum_{\boldsymbol{j} \in J_k} \operatorname{Vol}_{d-k}(G_{\boldsymbol{j}}) L_{\boldsymbol{j},3}^k,$$

so by (37) $a \le S_{D,s}^{2/d} \epsilon$ where

$$S_{D,s}^{2/d} := \Big(\frac{\operatorname{Vol}_d(D)}{u^p} + 2^p \operatorname{Vol}_{d-1}(\partial D) + \sum_{k=1}^{d} (2L_{k,1})^{d-k} A_u^k \sum_{\boldsymbol{j} \in J_k} \operatorname{Vol}_{d-k}(G_{\boldsymbol{j}}) L_{\boldsymbol{j},3}^k \Big)^{1/p}. \tag{40}$$

We have thus bounded the bracketing entropy when $D \subset [0,1]^d$ and $B = 1$. Thus, by the scaling at the beginning of the proof, for any convex polytope $D \subset \prod_{i=1}^d [a_i, b_i]$ and any $B > 0$, we have shown for $0 < \epsilon \le B \left(\prod_{i=1}^d b_i - a_i \right)^{1/p}$ that

$$\log N_{[]} \left(\epsilon S_{\tilde{D},s}^{2/d}, \mathcal{C}(D, B), L_p \right) \le S_{\tilde{D},c} \epsilon^{-d/2} \Bigg(B \left(\prod_{i=1}^{d} (b_i - a_i) \right)^{1/p} \Bigg)^{d/2}$$

where $S_{\tilde{D},c}$ equals

$$c_d(u+2d)^{d/2}+\sum_{k=1}^{d}\sum_{j\in J_k} L_{j,4}^{-1}\tilde{c}_{d,k}u^{kd/2(p+1)(p+2)}\left(1+2^{d-k+2}L_{k,1}u^{-(d-k)}C_d \operatorname{Vol}_{d-k}(G_j)\right)^{d/2}. \tag{41}$$

Letting $\delta := S_{\tilde{D},s}^{2/d}\epsilon$, we have shown that

$$\log N_{[\,]}\left(\delta, \mathcal{C}(D,B), L_p\right) \le S_{\tilde{D},c}S_{\tilde{D},s}\delta^{-d/2}\left(B\left(\prod_{i=1}^{d}(b_i-a_i)\right)^{1/p}\right)^{d/2} \tag{42}$$

for $0<\delta\le S_{\tilde{D},s}^{2/d}B\prod_{i=1}^{d}(b_i-a_i)^{1/p}$.

Finally, we can extend from requiring $\delta \le S_{D,s}B\prod_{i=1}^{d}(b_i-a_i)^{1/p}$ to allowing any $\delta>0$, just as in the proof of Theorem 2.1, in [13] (at the slight cost of increasing the constant on the right hand side of (42)). □

Note that the constants $S_{\tilde{D},s}$ and $S_{\tilde{D},c}$ should be calculated using the rescaling of D that lies in $[0,1]^d$, $\tilde{D}$. The following lemma was used above.

Lemma 3.6. *For any $\gamma\ge 1$, $0<\epsilon\le 1$, with ζ_i given in (32), and with A and u given by (9) and (26), we have*

$$\sum_{\alpha=1}^{A}\zeta_\alpha^\gamma \le 2u^{\gamma/(2(p+1)^2)}.$$

Proof. Straightforward algebra shows

$$\zeta_\alpha^2 = \exp\left\{p\frac{(p+1)^{\alpha-2}}{(p+2)^\alpha}\log\epsilon\right\}. \tag{43}$$

We have, for $\alpha=1,\ldots,A-1$,

$$\frac{\zeta_\alpha}{\zeta_{\alpha+1}} = \exp\left\{\frac{p\log\epsilon}{2(p+1)^2(p+2)}\left(\frac{p+1}{p+2}\right)^\alpha\right\},$$

which is bounded above by

$$\exp\left\{\frac{p\log\epsilon}{2(p+1)^2(p+2)}\left(\frac{p+1}{p+2}\right)^{A-1}\right\} \le \exp\left\{\frac{\log u}{2(p+1)^2(p+2)}\right\} =: R^{-1}.$$

Then, $\zeta_\alpha^\gamma(R^\gamma-1)\le \zeta_\alpha^\gamma R^\gamma-(R\zeta_{\alpha-1})^\gamma$ so $\zeta_\alpha^\gamma \le (R^\gamma/(R^\gamma-1))(\zeta_\alpha^\gamma-\zeta_{\alpha-1}^\gamma)$ and thus

$$\sum_{\alpha=1}^{A}\zeta_\alpha^\gamma \le \zeta_1^\gamma + \frac{R^\gamma}{R^\gamma-1}\sum_{\alpha=2}^{A}(\zeta_\alpha^\gamma-\zeta_{\alpha-1}^\gamma) = \zeta_1^\gamma+\frac{R^\gamma}{R^\gamma-1}(\zeta_A^\gamma-\zeta_1^\gamma)\le \frac{R^\gamma}{R^\gamma-1}\zeta_A^\gamma \tag{44}$$

and $\zeta_A^\gamma \le u^{\gamma/(2(p+1)(p+2))}$. Since $u\le \exp\left(-2(p+1)^2(p+2)\log 2\right)$ by its definition (15), $R\ge 2$ so $R^\gamma/(R^\gamma-1)\le 2$ for any $\gamma\ge 1$. □

For any convex D and convex subset $\tilde{D}\subset D$, note that $\mathcal{C}(D,1)|_{\tilde{D}} \subset \mathcal{C}\left(\tilde{D},1\right)$. Thus by covering any convex polytope D by simple polytopes $D_i\subset D$, we can bound

$N_{[\,]}\left(\epsilon, \mathcal{C}(D, B), L_p\right)$ by applying Theorem 3.5 repeatedly to $\mathcal{C}(D_i, 1)$ and using (5). A cover of D can be attained by, for instance, subdividing D into simple polytopes [35], such as simplices. The constant in the bound then depends on the subdivision of D.

Corollary 3.7. *Fix $d \geq 1$ and $p \geq 1$. Let $D \subseteq \prod_{i=1}^{d}[a_i, b_i]$ be any convex polytope. Then for $\epsilon > 0$,*

$$\log N_{[\,]}\left(\epsilon, \mathcal{C}(D, B), L_p\right) \leq C_{d,D}\epsilon^{-d/2}\left(B\left(\prod_{i=1}^{d}(b_i - a_i)\right)^{1/p}\right)^{d/2}.$$

Proof. By the same scaling argument as in the proof of Theorem 3.5 we may assume $[a_i, b_i] = [0, 1]$ and $B = 1$. The $d = 1$ case is given by [18]. Any convex polytope D can be triangulated into d-dimensional simplices (see e.g. [11,38]). We are done by applying Theorem 3.5 to each of those simplices, by (5). □

4. Properties of $G_{i,j}$

In this section we show how to embed the domains $G_{i,j}$, which partition D, into hyperrectangles. We used this in the proof of Theorem 3.5 so we could apply Theorem 2.1. Theorem 2.1 says that the bracketing entropy of convex functions on domain D with Lipschitz constraints along directions $e_1, \ldots, e_k$ depends on $w(D, e_i)$ (since that gives the maximum "rise" in "rise over run"). In our proof of Theorem 3.5 we partitioned D into sets related to parallelotopes. Thus we will study these parallelotopes. We know the width of $G_{i,j}$ in the directions v_{j_α}, which are $\delta_{i_\alpha+1} - \delta_{i_\alpha}$, by definition.

A polytope P is a d-parallelotope if $P = \sum_{i=1}^{d}[a_i, b_i]$ for vectors $a_i, b_i \in \mathbb{R}^d$, where for all i, $[a_i, b_i]$ is not parallel to the affine hull of $[a_j, b_j]$ for any $j \neq i$ ([28] page 56). We will rely on the following representation for a k-dimensional parallelotope.

Lemma 4.1. *Let k be a positive integer and let $P := \cap_{\beta=1}^{k}\tilde{E}_\beta$ be a parallelotope where $\tilde{E}_\beta := \left\{x \in \mathbb{R}^k : 0 \leq \langle x, v_\beta\rangle \leq d_\beta\right\}$ for k linearly independent normal unit vectors v_β. Let $H_\beta^0 := \left\{x \in \mathbb{R}^k : \langle x, v_\beta\rangle = 0\right\}$. Let $\tilde{f}_\beta$ be the unit vector lying in $\cap_{\gamma=1,\gamma\neq\beta}^{k}\tilde{H}_\beta^0$ with $\left\langle \tilde{f}_\beta, v_\beta\right\rangle > 0$, for $\beta = 1, \ldots, k$. Then 0 is a vertex of P and we can write*

$$P = \sum_{\beta=1}^{k}[0, f_\beta]$$

where $f_\beta := d_\beta \tilde{f}_\beta / \left\langle \tilde{f}_\beta, v_\beta\right\rangle$, $[0, f_\beta] = \left\{\lambda f_\beta : \lambda \in [0, 1]\right\}$.

Proof. Since the vectors v_β are unique, $\cap_{\beta=1}^{k}H_\beta^0 = 0$ and the intersection of any $k - 1$ of the hyperplanes H_β^0 gives a 1-dimensional space, span $\left\{\tilde{f}_\beta\right\}$. A k-dimensional parallelotope can be written as the set-sum of the k intervals emanating from the vertex, each given by the intersection of $k - 1$ of the hyperplanes H_β^0. See page 56 of [28]. Note that f_β satisfy $\langle f_\beta, v_\beta\rangle = d_\beta$ so that $f_\beta \in \tilde{H}_\beta^+ := \left\{x \in \mathbb{R}^k : \langle x, v_\beta\rangle = d_\beta\right\}$; thus the k intervals are given by $[0, f_\beta]$, $\beta = 1, \ldots, k$. □

Note the vector $\tilde{f}_\beta$ can be written as $(I - Q)v_\beta$ where I is the identity projection in $\mathbb{R}^k$ and Q is the projection onto span $\left\{v_1, \ldots, v_{\beta-1}, v_{\beta+1}, \ldots, v_k\right\}$. The next proposition uses Lemma 4.1

to bound the widths of $G_{\boldsymbol{i},\boldsymbol{j}}$, in certain directions, in terms of the width of $G_{\boldsymbol{j}}$ in those directions. We will need the following constant (depending on D). For $k \in \{1, \ldots, d-1\}$, let

$$L_{k,1} := 1 \vee \max_{\boldsymbol{j} \in J_k} \max_{\substack{\|e\|=1 \\ e \in E_{\boldsymbol{j}}}} \max_{\substack{j \in \{1,\ldots,N\} \setminus \boldsymbol{j};\ \langle e, v_j \rangle < 0 \\ \langle v_i, v_j \rangle > 0,\ \text{some } i \in \boldsymbol{j}}} \langle -e, v_j \rangle^{-1}, \tag{45}$$

where $E_{\boldsymbol{j}} := \operatorname{span}\{e_{\boldsymbol{j},k+1}, \ldots, e_{\boldsymbol{j},d}\}$ from Proposition 3.2, and we abuse notation as convenient to treat $\boldsymbol{j}$ as if it were a set rather than a vector. We also (arbitrarily) define $L_{d,1} := 1$, for ease of presentation later on.

Proposition 4.2. *For each $k \in \{1, \ldots, d-1\}$, $\boldsymbol{i} \in I_k$, $\boldsymbol{j} \in J_k$, and each $G_{\boldsymbol{i},\boldsymbol{j}}$, and the basis $\boldsymbol{e} \equiv \boldsymbol{e}_{\boldsymbol{i},\boldsymbol{j}}$ from Proposition 3.2, for $\alpha \in \{k+1, \ldots, d\}$, we have*

$$w(G_{\boldsymbol{i},\boldsymbol{j}}, e_\alpha) \le 2L_{k,1} w(G_{\boldsymbol{j}}, e_\alpha). \tag{46}$$

Then for $k \in \{1, \ldots, d\}$, let $\tilde{f}_\alpha \equiv \tilde{f}_{\boldsymbol{j},\alpha}$ be the unit vector with $\left\langle \tilde{f}_\alpha, v_{j_\alpha} \right\rangle > 0$ lying in $\operatorname{span}\{v_{j_1}, \ldots, v_{j_k}\} \cap \left(\cap_{\gamma=1,\gamma\neq\alpha}^k H_{j_\gamma}^0\right)$, $\alpha = 1, \ldots, k$, where $H_{j_\gamma}^0 := \{y \in \mathbb{R}^d : \langle y, v_{j_\gamma} \rangle = 0\}$. Then for $k \in \{1, \ldots, d\}$, we have

$$\operatorname{Vol}_d(G_{\boldsymbol{i},\boldsymbol{j}}) \le (2L_{k,1})^{d-k} \operatorname{Vol}_{d-k}\left(G_{\boldsymbol{j}}\right) \cdot \prod_{\alpha=1}^k \frac{\delta_{i_\alpha+1} - \delta_{i_\alpha}}{\left\langle \tilde{f}_\alpha, v_{j_\alpha} \right\rangle} \tag{47}$$

where $L_{k,1}$ is given by (45) for $k \in \{1, \ldots, d-1\}$ (and we set $L_{d,1} := 1$ arbitrarily).

Proof. Fix $k \in \{1, \ldots, d-1\}$. Let $x \equiv x_{\boldsymbol{j}} \in G_{\boldsymbol{j}}$ (from (8)). Let f_{j_γ} be as given in (20). Let $P_{\boldsymbol{i},\boldsymbol{j}} := \sum_{\gamma=1}^k [0, f_{j_\gamma}]$. We will show that $G_{\boldsymbol{i},\boldsymbol{j}}$ is contained in the set-sum of a hyperrectangle and $P_{\boldsymbol{i},\boldsymbol{j}}$. To begin with let $G_{\boldsymbol{i},\boldsymbol{j}} \ni z = x + \sum_{\gamma=1}^k f_{j_\gamma}^*$ where $f_{j_\gamma}^* = d_{j_\gamma} \tilde{f}_{j_\gamma}$ where

$$0 \le d_{j_\gamma} \le (\delta_{i_\gamma+1} - \delta_{i_\gamma}) / \left\langle \tilde{f}_{j_\gamma}, v_{j_\gamma} \right\rangle \le u / \left\langle \tilde{f}_{j_\gamma}, v_{j_\gamma} \right\rangle. \tag{48}$$

Take an arbitrary $e \in \operatorname{span}\{e_{k+1}, \ldots, e_d\}$ with $\|e\| = 1$. Let $\lambda_{z,e} := d^+(z, \partial G_{\boldsymbol{i},\boldsymbol{j}}, e)$ and let j give the corresponding facet of $G_{\boldsymbol{i},\boldsymbol{j}}$ that $x + \lambda_{z,e} e$ hits, so that $\langle z + \lambda_{z,e} e, v_j \rangle = p_j + u$ for some $j \notin \boldsymbol{j}$ (abusing notation to treat $\boldsymbol{j}$ as if were a set rather than a vector). Note that this means

$$\langle e, v_j \rangle < 0. \tag{49}$$

If $\left\langle \sum_{\gamma=1}^k f_{j_\gamma}^*, v_j \right\rangle \le 0$ then

$$\lambda_{z,e} \le d^+(x, \partial G_{\boldsymbol{i},\boldsymbol{j}}, e). \tag{50}$$

Thus if (50) does not hold then $\left\langle f_{j_\gamma}^*, v_j \right\rangle > 0$ for some $\gamma \in \{1, \ldots, k\}$, so $\langle v_{j_\alpha}, v_j \rangle > 0$ for some $\alpha \in \{1, \ldots, k\}$. Now, since $\langle z + \lambda_{z,e} e, v_j \rangle = p_j + u$, we have

$$\lambda_{z,e} = \frac{p_j + u - \langle z, v_j \rangle}{\langle e, v_j \rangle} \le \frac{\langle x, v_j \rangle - p_j + u \sum_{\gamma=1}^k \frac{\langle \tilde{f}_\gamma, v_j \rangle}{\langle \tilde{f}_\gamma, v_{j_\gamma} \rangle}}{\langle -e, v_j \rangle}$$

Now

$$\langle x, v_j \rangle - p_j \le d(x, H_j) \le d^+(x, \partial_r G_{\boldsymbol{j}}, e)$$

since H_j is the closest hyperplane to x in the direction e. Recall the definition of $L_{k,1}$ in (45). Now, by (15) and the definition of $L_{k,2}$ (16), we have shown

$$\lambda_{z,e} \le 2L_{k,1} d^+(x, \partial_r G_{\boldsymbol{j}}, e), \tag{51}$$

by (49) and (50). This means that

$$(G_{\boldsymbol{i},\boldsymbol{j}} - z) \cap \operatorname{span}\{e_{k+1}, \dots, e_d\} \subset 2L_{k,1}\left(G_{\boldsymbol{j}} - x\right)$$

so we can conclude that $w(G_{\boldsymbol{i},\boldsymbol{j}} - z, e_\alpha) \le 2L_{k,1} w(G_{\boldsymbol{j}}, e_\alpha)$ and $w(G_{\boldsymbol{i},\boldsymbol{j}}, e_\alpha) \le 2L_{k,1} w(G_{\boldsymbol{j}}, e_\alpha)$ since $\langle z, e_\alpha \rangle = 0$ for all d_{j_γ} given by the range (48), $\alpha = k+1, \dots, d$, for $k = 1, \dots, d-1$.

Let $\rho_{\boldsymbol{j},\alpha} := w(G_{\boldsymbol{j}}, e_\alpha)$. Then let

$$R_{\boldsymbol{i},\boldsymbol{j}} := P_{\boldsymbol{i},\boldsymbol{j}} + \sum_{\alpha=k+1}^{d} \left[-2L_{k,1}\rho_{\boldsymbol{j},\alpha} e_\alpha, 2L_{k,1}\rho_{\boldsymbol{j},\alpha} e_\alpha\right]. \tag{52}$$

Then for any $x \in G_{\boldsymbol{i},\boldsymbol{j}}$ such that $\langle x, v_{j_\alpha} \rangle = p_{j_\alpha} + \delta_{i_\alpha}$ for $\alpha \in \{1, \dots, k\}$, we have shown

$$G_{\boldsymbol{i},\boldsymbol{j}} \subset x + R_{\boldsymbol{i},\boldsymbol{j}}. \tag{53}$$

It then also follows that

$$\operatorname{Vol}_d(G_{\boldsymbol{i},\boldsymbol{j}}) \le (2L_{k,1})^{d-k} \operatorname{Vol}_{d-k}\left(G_{\boldsymbol{j}}\right) \cdot \operatorname{Vol}_k\left(\sum_{\alpha=1}^{k} \left[0, f_{j_\alpha}\right]\right). \tag{54}$$

Since of parallelotopes with given axis lengths, the one with largest volume is the hyperrectangle, $\operatorname{Vol}_k\left(\sum_{\alpha=1}^{k}\left[0, f_{j_\alpha}\right]\right) \le \prod_{\alpha=1}^{k} \frac{\delta_{i_\alpha+1} - \delta_{i_\alpha}}{\left\langle \tilde{f}_{j_\alpha}, v_{j_\alpha}\right\rangle}$, and so we have shown (47) (with this bound on $\operatorname{Vol}_k\left(\sum_{\alpha=1}^{k}\left[0, f_{j_\alpha}\right]\right)$ being all that is needed in the $k = d$ case). $\square$

The previous proposition controls the width and volume of $G_{\boldsymbol{i},\boldsymbol{j}}$ in directions lying in span$\left\{G_{\boldsymbol{j}}\right\}$. Next we control width, volume, and also distance to ∂D in directions perpendicular to span$\left\{G_{\boldsymbol{j}}\right\}$.

Lemma 4.3. *Let $P := \sum_{\alpha=1}^{k}[0, f_\alpha]$ be a parallelotope in $\mathbb{R}^k$ where $f_1, \dots, f_k$ are k linearly independent vectors. Then there exists an orthonormal basis of $\mathbb{R}^k$, $e_1, \dots, e_k \in \mathbb{R}^k$ and $\gamma_1, \dots, \gamma_k \in \mathbb{R}$, such that*

$$P \subset \sum_{\alpha=1}^{k}[0, \gamma_\alpha e_\alpha] \qquad \text{where} \qquad |\gamma_\alpha| \le k \operatorname{diam}(P, e_\alpha). \tag{55}$$

Proof. We will construct a permutation π of $\{1, \dots, k\}$ and inductively define $e_1, \dots, e_k$ based on the sequence $f_{\pi(1)}, \dots, f_{\pi(k)}$. Let $e_1 := f_{\pi(1)}/\|f_{\pi(1)}\|$ where $\|f_{\pi(1)}\|$ is maximal over $\{\|f_\alpha\|\}_{\alpha=1}^{k}$. Now let Q_{j-1} be the projection of $\mathbb{R}^k$ onto span$\left\{e_1, \dots, e_{j-1}\right\}$ and let $Q_{j-1}^{\perp}$ be the projection onto span$\left\{e_1, \dots, e_{j-1}\right\}^{\perp}$. Then let $e_j := Q_{j-1}^{\perp} f_{\pi(j)}/\|Q_{j-1}^{\perp} f_{\pi(j)}\|$ where $\pi(j) \in \{1, \dots, k\} \setminus \{\pi(1), \dots, \pi(j-1)\}$ is defined so that $\|Q_{j-1}^{\perp} f_{\pi(j)}\|$ is maximal.

Let $P_j := \sum_{\alpha=1}^{j}[0, f_{\pi(j)}]$. Now, diam$(P_j, e_\alpha)$ is given by the value of $\langle x - y, e_\alpha \rangle$ such that $x, y \in P_j$ and $\langle x - y, e_\alpha \rangle$ is maximal. Since $f_{\pi(j)} \notin \operatorname{span}\left\{f_{\pi(i)}\right\}_{i \ne j}$, we also have that

$e_j \notin \operatorname{span}\left\{f_{\pi(i)}\right\}_{i\neq j}$. Thus for $i \geq j$, $\operatorname{diam}(P_i, e_j) \leq \left\langle f_{\pi(j)}, e_j\right\rangle$ and so in fact $\operatorname{diam}(P_i, e_j) = \operatorname{diam}(P, e_j) = \left\langle f_{\pi(j)}, e_j\right\rangle$.

Now we prove by induction that

$$P_j \subset \sum_{\alpha=1}^{j} [0, \gamma_{j,\alpha} e_\alpha] \tag{56}$$

where $0 \leq \gamma_{j,\alpha} \leq j \operatorname{diam}(P_j, e_\alpha) = j \operatorname{diam}(P, e_\alpha)$. The statement is immediate for $j = 1$. Thus let $1 < j \leq k$ and assume the induction hypothesis holds for $j - 1$. Then for $1 < i \leq j$

$$\left|\left\langle e_i, f_{\pi(j)}\right\rangle\right| \leq \|Q_i^{\perp} f_{\pi(j)}\| \leq \|Q_i^{\perp} f_{\pi(i)}\| = \left|\left\langle e_i, f_{\pi(i)}\right\rangle\right| = \operatorname{diam}(P, e_i) \tag{57}$$

where the first inequality is because $e_i \in \operatorname{span}\{e_1, \ldots, e_{i-1}\}^{\perp}$, and the next inequality and equality are by the definition of e_i. Also, (57) is immediately verifiable for $i = 1$.

Now, we can write

$$f_{\pi(j)} = \lambda_{j,1} e_1 + \cdots + \lambda_{j,j} e_j \tag{58}$$

where $|\lambda_{j,i}| \leq \operatorname{diam}(P, e_i)$ by (57). For any $x \in P_j = P_{j-1} + [0, f_{\pi(j)}]$, we can write

$$x = \sum_{\alpha=1}^{j-1} \eta_{j-1,\alpha} e_\alpha + \eta f_{\pi(j)} \tag{59}$$

where $0 \leq \eta \leq 1$ and $|\eta_{j-1,\alpha}| \leq (j-1) \operatorname{diam}(P, e_\alpha)$ by the induction hypothesis. Thus (59) equals

$$\sum_{\alpha=1}^{j-1} \left(\eta_{j-1,\alpha} + \eta\lambda_{j,\alpha}\right) e_\alpha + \eta\lambda_{j,j} e_j,$$

and $|\eta_{j-1,\alpha} + \eta\lambda_{j,\alpha}| \leq (j-1) \operatorname{diam}(P, e_\alpha) + \operatorname{diam}(P, e_\alpha)$ for $\alpha \leq j-1$ and $|\lambda_{j,j}| = \operatorname{diam}(P, e_j)$, so the induction hypothesis is shown. □

To state the next lemma we make the following definitions. For a set $D \subset \mathbb{R}^d$ and a unit vector v, let

$$\operatorname{diam}(D, v) := \sup_{\substack{x,y\in D \\ x-y\in \operatorname{span}\{v\}}} \|x - y\|. \tag{60}$$

Lemma 4.4. *Let Assumption 1 hold. Let $k \in \{1, \ldots, d\}$, $\boldsymbol{i} \in I_k$, and $\boldsymbol{j} \in J_k$. Then for any unit length $v \in \operatorname{span}\left\{v_{j_1}, \ldots, v_{j_k}\right\}$,*

$$\operatorname{diam}(G_{\boldsymbol{i},\boldsymbol{j}}, v) \leq \min_{\alpha\in\{1,\ldots,k\}} \frac{\delta_{i_\alpha+1}}{\left|\left\langle v, v_{j_\alpha}\right\rangle\right|}, \quad \textit{and} \tag{61}$$

$$d(G_{\boldsymbol{i},\boldsymbol{j}}, \partial D, v) \geq \max_{\alpha\in\{1,\ldots,k\}} \frac{\delta_{i_\alpha}}{\left|\left\langle -v, v_{j_\alpha}\right\rangle\right|}. \tag{62}$$

Proof. Fix $k \in \{1, \ldots, d\}$, $\boldsymbol{i} \in I_k$, $\boldsymbol{j} \in J_k$. Fix $v \in \operatorname{span}\left\{v_{j_1}, \ldots, v_{j_k}\right\}$ with $\|v\| = 1$, fix $\alpha \in \{1, \ldots, k\}$. Since $\operatorname{diam}(G_{\boldsymbol{i},\boldsymbol{j}}, v) = \operatorname{diam}(G_{\boldsymbol{i},\boldsymbol{j}}, -v)$, we restrict attention to v such that

$$\left\langle -v, v_{j_\alpha}\right\rangle \geq 0. \tag{63}$$

We will upper bound $\operatorname{diam}(G_{\boldsymbol{i},\boldsymbol{j}}, v)$. Consider $x, y \in G_{\boldsymbol{i},\boldsymbol{j}}$ such that $x - y \in \operatorname{span}\{v\}$. In particular, assume without loss of generality that $x - y = \lambda v$ for $\lambda \geq 0$. Since $x, y \in G_{\boldsymbol{i},\boldsymbol{j}}$, $\langle y, v_{j_\alpha} \rangle \leq p_{j_\alpha} + \delta_{i_\alpha+1}$ and $p_{j_\alpha} + \delta_{i_\alpha} \leq \langle x, v_{j_\alpha} \rangle$; thus $\delta_{i_\alpha} - \delta_{i_\alpha+1} \leq p_{j_\alpha} + \delta_{i_\alpha} - \langle y, v_{j_\alpha} \rangle \leq \lambda \langle v, v_{j_\alpha} \rangle$. Since $\langle -v, v_{j_\alpha} \rangle \geq 0$, we have $\lambda \leq (\delta_{i_\alpha+1} - \delta_{i_\alpha}) / \langle -v, v_{j_\alpha} \rangle$. Thus we see $\operatorname{diam}(G_{\boldsymbol{i},\boldsymbol{j}}, -v) = \operatorname{diam}(G_{\boldsymbol{i},\boldsymbol{j}}, v) \leq (\delta_{i_\alpha+1} - \delta_{i_\alpha}) / |\langle v, v_{j_\alpha} \rangle|$. This holds for all $\alpha \in \{1, \ldots, k\}$, so for any $\tilde{v} \in \operatorname{span}\{v_{j_1}, \ldots, v_{j_k}\}$ (where we do not assume $\langle \tilde{v}, v_{j_\alpha} \rangle \geq 0$)

$$\operatorname{diam}(G_{\boldsymbol{i},\boldsymbol{j}}, \tilde{v}) \leq \min_{\alpha \in \{1,\ldots,k\}} \frac{\delta_{i_\alpha+1} - \delta_{i_\alpha}}{|\langle \tilde{v}, v_{j_\alpha} \rangle|}. \tag{64}$$

Next we take v as above and now lower bound $d(G_{\boldsymbol{i},\boldsymbol{j}}, \partial D, v)$. Fix $\alpha \in \{1, \ldots, k\}$. We begin by considering $d(G_{\boldsymbol{i},\boldsymbol{j}}, H_{j_\alpha}, v)$. Again, since $d(G_{\boldsymbol{i},\boldsymbol{j}}, \partial D, v) = d(G_{\boldsymbol{i},\boldsymbol{j}}, \partial D, -v)$, we can and do assume (63) holds. Fix $x \in G_{\boldsymbol{i},\boldsymbol{j}}$. Consider $\lambda > 0$ such that $x + \lambda v \in H_{j_\alpha}$. Then $\lambda \langle v, v_{j_\alpha} \rangle = p_{j_\alpha} - \langle x, v_{j_\alpha} \rangle \leq -\delta_{i_\alpha}$ since $\langle x, v_{j_\alpha} \rangle \geq \delta_{i_\alpha} + p_{j_\alpha}$, and so $\lambda \geq \delta_{i_\alpha} / \langle -v, v_{j_\alpha} \rangle$. This shows for any $\beta \in \{1, \ldots, k\}$ that

$$d(G_{\boldsymbol{i},\boldsymbol{j}}, \cup_{\alpha=1}^{k} F_{j_\alpha}, v) \geq \min_{\alpha \in \{1,\ldots,k\}} \frac{\delta_{i_\alpha}}{|\langle v, v_{j_\alpha} \rangle|}. \tag{65}$$

To complete the proof, note for $j \in \{1, \ldots, N\} \setminus \boldsymbol{j}$, that

$$d(x, F_j, v) \geq u = u \min_{\alpha=1,\ldots,k} |\langle v, v_{j_\alpha} \rangle|^{-1}$$

which is larger than the right hand side of (65). □

5. Further applications

We now consider further entropy bounds that rely on the above ideas, results, or their proofs. In Section 5.1 we consider so-called univariate and multivariate *m-monotone functions*. In Section 5.2 we briefly consider estimation of level sets of convex functions and the question of adaptation to polytopal level sets. Further discussion is given at the beginning of the two subsections.

5.1. Bracketing entropy of m-monotone function classes

The shape constraint of *m-monotonicity*, for $m \in \{0, 1, 2, \ldots\}$, is useful because it serves, roughly, as a higher order convexity restriction (when $m > 2$). An m-monotone function f satisfies further convexity restrictions besides simply convexity of f, and so in many settings is even nicer to work with than convex functions are. When $d = 1$, m-monotonicity is defined as follows (by, e.g., [43,44]).

Definition 5.1. A function $f : [0, \infty) \to \mathbb{R}$ is 0-monotone if it is nonnegative, 1-monotone if it is nonnegative and nonincreasing, and 2-monotone if it is nonnegative, nonincreasing, and convex; f is m-monotone for $m \geq 2$ if $(-1)^l f^{(l)}$ exists and is nonnegative, nonincreasing, and convex for $l = 0, 1, \ldots, m - 2$.

(Here $f^{(l)}$ is the lth derivative, with $f^{(0)} \equiv f$.) When $m = 1$ or 2, a large body of statistical work and results exists (some of which also allows the case where $d > 1$), some of which is referenced in the introduction of this paper. Statistical properties of two (nonparametric) estimators of a (univariate) m-monotone density, for general m (and $d = 1$), were introduced and studied in [1,2]; see also [26]. For instance, in statistical settings, if a function being nonparametrically estimated is known to be m-monotone, then it can be estimated at a faster rate of convergence than if it were just convex [1,2,26]. As discussed in Section 1, in the univariate setting m-monotone functions have been studied, but we do not even know of a formal definition of m-monotonicity in the multivariate case. In fact, as discussed below, there are several possible definitions one could use for m-monotonicity that generalize the univariate definition. We present one definition which has the benefit of being amenable to finding bracketing upper bounds. We then show that our proof for the bracketing upper bound for convex functions, Theorem 3.5 (and Corollary 3.7), applies to yield a bracketing bound for classes of m-monotone functions. This is the main result of Section 5.1 and is given in Theorem 5.16. Recall that the proof of Theorem 3.5 relies on Theorem 2.1. There is no known or immediate analog for Theorem 2.1 in the general m-monotone case. Thus we prove an analog, Theorem 5.11 (under a certain technical restriction, given below in Definition 5.2 Part A), which we use to prove Theorem 5.16.

As mentioned above, there are many possible methods for defining a class of m-monotone functions in the multivariate setting. This is perhaps illustrated by the fact that there are many competing definitions of monotonicity (i.e., 1-monotonicity) in dimension $d \geq 2$. One can define a function f to be multivariate monotone (or unimodal) via *star monotonicity*, meaning that along all rays emanating from a special fixed point, f is monotone. This, for instance, is a suggested definition used in the related context of hyperbolic monotonicity by [9, p. 600]. [25] consider entropy bounds for *block-decreasing* densities. Even "block decreasing" can be defined in multiple ways: [25] and [12] differ in their definitions of this term. Very recent statistical work has considered *entire monotonicity* in the regression setting [22]. See [12, chapter 2] for several other possible definitions of unimodality (they focus on unimodality rather than monotonicity, but the two settings are very similar). Many of the above definitions are not amenable to accurate entropy computations, at least with the tools we are aware of at present. In Sub Section 5.1.1 we present a definition of multivariate m-monotonicity that is amenable to entropy calculations; the results we get suggest that the entropies are of the "right" order of magnitude ($\epsilon^{-d/m}$ as $\epsilon \searrow 0$) that we might expect a priori. This suggests that our definition is indeed a reasonable one. In Sub Section 5.1.2 we return briefly to the particular $d = 1$ case.

5.1.1. Multivariate m-monotone functions

Fix the dimension $d \geq 1$. We will use so-called d-dimensional multi-index notation: a vector of nonnegative integers $\boldsymbol{i} = (i_1, \ldots, i_d)$ is a multi-index. We let $|\boldsymbol{i}| := i_1 + \cdots + i_d$. Let $\mathbb{I}_m$ be the set of multi-indices $\boldsymbol{i}$ with $|\boldsymbol{i}| = m$. For two vectors $K = (K_1, \ldots, K_j)$, $L = (L_1, \ldots, L_j) \in \mathbb{R}^j$ with $L_i > 0$, we let $L^K := L_1^{K_1} \cdots L_j^{K_j}$. For any function f, we let $f^{(\boldsymbol{i})}$ be $\frac{\partial^{|\boldsymbol{i}|}}{\partial x_{i_1} \cdots \partial x_{i_d}} f$, whenever this is well-defined. We let $\frac{\partial}{\partial e_i} f(x)$ denote $\frac{d}{dt}|_{t=0} f(x + t e_i)$, and for an orthonormal basis $\boldsymbol{e} := \{e_1, \ldots, e_d\}$ and $\boldsymbol{j} \in \mathbb{I}_i$, we let $f_{\boldsymbol{e}}^{(\boldsymbol{j})} := \frac{\partial^{|\boldsymbol{j}|}}{\partial e_1^{j_1} \cdots \partial e_d^{j_d}} f$.

Our m-monotone classes are based on any subclass $\mathcal{C}^*$ of convex functions having a certain needed property. The idea of multivariate m-monotonicity involves convexity of partial derivatives in certain directions; since such convexity is not preserved by rotation (see Remark 5.4), we will define m-monotonicity to be *relative* to a domain D_0. In the case where D_0 is a hyperrectangle, the definition simplifies (see Remark 5.4).

Definition 5.2.

A. For a convex set $G \subset \prod_{i=1}^d [a_i, b_i]$, let $\mathcal{C}^*(G)$ be any subclass of $\mathcal{C}(G)$ such that for all B, $\boldsymbol{\Gamma}$, $\mathcal{C}^*(G, B, \boldsymbol{\Gamma}) := \mathcal{C}^*(G) \cap \mathcal{C}(G, B, \boldsymbol{\Gamma})$ satisfies the following L_∞ cover property. For all $\epsilon > 0$, there exists a L_∞-ϵ-cover of log cardinality no larger than $c\epsilon^{-d/2}(B+\sum_{i=1}^d \Gamma_i(b_i-a_i))^{d/2}$. The cover must satisfy the following. For any $f \in \mathcal{C}^*(G, B, \boldsymbol{\Gamma})$, any $x, y \in G$, and any rotation matrix $A \in \mathbb{R}^{d\times d}$ ($\det A = 1$ and $A' = A^{-1}$), we have

$$\frac{\partial}{\partial x_i}\int_0^1 |h(A(x+t(y-x))) - f(A(x+t(y-x)))|dt \le \epsilon \tag{66}$$

for all $i \in \{1,\dots,d\}$, where h is the L_∞-closest element of the cover to f.

B. Let $D_0 \subset [0,\infty)^d$ be a convex polytope, let $0 < B < \infty$, and let $m \ge 2$ be an integer. We define the class of *m-monotone functions relative to* D_0, denoted $\mathcal{C}^m(D_0, B)$, to be the set of all $f \in \mathcal{C}(D_0, B)$ satisfying the following. For each vertex of D_0, for all ($d!$ possible) orthornomal bases $\boldsymbol{e}$ given by Proposition 3.2, either $f_{\boldsymbol{e}}^{(\boldsymbol{i})}$ or $-f_{\boldsymbol{e}}^{(\boldsymbol{i})}$ lies in $\mathcal{C}^*(D_0)$ for all $\boldsymbol{i} \in \mathbb{I}_j$, $0 \le j \le m-2$.

Remark 5.3. The fundamental idea of m-monotonicity is given by Part B of Definition 5.2. The technical requirement (66) of Part A is needed for the proof of our bracketing bound. It is not clear at this point if it can be removed or not. When $y = x + x_j$ and A is the identity, for continuously differentiable h and f, the property holds automatically by the Fundamental Theorem of Calculus. Ideally we would like to replace $\mathcal{C}^*$ by the full class $\mathcal{C}$ (which is possible when $d = 1$, see Remark 5.13 and Sub Section 5.1.2). We leave an investigation of whether this is possible for future work.

Remark 5.4. The property of m-monotonicity is preserved by translation and rescaling. However, while rotations of convex functions are still convex, if $f^{(\boldsymbol{j})}$ is convex, $|\boldsymbol{j}| > 2$, then after rotation (by, say, a matrix A), $g^{(\boldsymbol{j})} := (f(A\cdot))^{(\boldsymbol{j})}$ is not necessarily convex. This is because a mixed partial derivative of a rotation is a linear combination of mixed partial derivatives; if some of the linear coefficients are negative, then the resulting function may no longer be convex. This is why our definition is relative to the domain, D_0 (so an m-monotone function after a rotation will be m-monotone relative to the rotated domain). Note that when D_0 is a hyperrectangle, for f to be m-monotone relative to D_0, it is sufficient that $f^{(\boldsymbol{i})}$ be convex for all $\boldsymbol{i} \in \mathbb{I}_i$, $0 \le i \le m-2$.

Remark 5.5. Our definition of multivariate m-monotonicity captures a higher order type of convexity. It does not enforce the alternating sign condition "$(-1)^j f^j \ge 0$" that is generally required in the univariate case. That is, we allow $f_{\boldsymbol{e}}^{(\boldsymbol{i})}$ to be either convex or to be concave. In the univariate case, there is only one direction in which one is computing a derivative. In the multivariate case, since we consider many different bases $\boldsymbol{e}$, and may have instances where a vector e_i and its opposite $-e_i$ are contained in two different bases, thus potentially switching the sign of $f_{\boldsymbol{e}}^{(\boldsymbol{i})}$, we must allow $f_{\boldsymbol{e}}^{(\boldsymbol{i})}$ to be either convex or concave. Further restrictions to our definition could be enforced if needed in a specific application; our entropy bounds would of course still apply.

Example 5.6. Let $a \in \mathbb{R}^d$ have nonnegative components and let $b > 0$; let $(\cdot)_+ := \max(\cdot, 0)$. The function $f(x) := b\,(1 - \langle a, x\rangle)_+^{m-1}\,\mathbb{1}_{[0,\infty)^d}(x)$ is a primary example of an m-monotone function (i.e., satisfies Part B of Definition 5.2). For any $m \ge 1$, the function $f(x) := e^{-b\langle a,x\rangle}\mathbb{1}_{[0,\infty)^d}(x)$ is m-monotone. Both are m-monotone relative to any hyperrectangle. Further examples of m-monotone functions can be generated by taking linear combinations.

Example 5.7. The functions in Example 5.6 are also m-monotone relative to polytopes beyond hyperrectangles. For simplicity, let $f(x) := (1-(x_1+x_2))_+^{m-1}\mathbb{1}_{[0,\infty)^2}(x)$. Note that if we let $g(x) = (1-(a_1x_1+a_2x_2))_+^{m-1}\mathbb{1}_{[0,\infty)^2}(x)$, then for $k \le m-1$,

$$\frac{\partial^k}{\partial x_1^j \partial x_2^{k-j}} g(x) = \frac{(m-1)!}{(m-1-k)!}(-1)^k a_1^j a_2^{k-j}(1-a_1x_1-a_2x_2)_+^{(m-1-k)} \tag{67}$$

Thus $(-1)^k \frac{\partial^k}{\partial x_1^j \partial x_2^{k-j}} g(x)$ is convex if $a_i > 0$, $i = 1, 2$, and $k \le m-2$.

Let $e_1 = (1, 0)'$, $e_2 = (0, 1)'$ (where $'$ denotes transpose), and let the basis $\boldsymbol{d} := \{d_1, d_2\}$ be defined by $d_i := Ae_i$ where

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

is the matrix giving rotation by angle θ. If $g(y) := f(Ay')$ then $g^{(j)} = f_{\boldsymbol{d}}^{(j)}$. Thus, by (67), f is m-monotone relative to any polytope since the partial derivatives are always either convex or concave.

Furthermore, as long as $\cos\theta + \sin\theta \ge 0$ and $\cos\theta - \sin\theta \ge 0$, i.e., as long as $-\pi/4 \le \theta \le \pi/4$, $f_{\boldsymbol{d}}^{(\boldsymbol{j})}$ is convex (for $\boldsymbol{j} \in \mathbb{I}_j$, $0 \le j \le m-2$). Thus, for instance, if we take $-\pi/4 \le \theta \le 0$, and let $H_\theta := \{x : \langle d_2, x\rangle \le \cos\theta\}$ (the rotation of the line $\{x_2 = 1\}$ by angle θ about the point $(0, 1)$), then if we let $D_0 := [0,1]^2 \cap H_\theta$, then $f_{\boldsymbol{d}}^{(\boldsymbol{j})}$ is convex where $\boldsymbol{d}$ is one basis given by Proposition 3.2 at the upper right vertex of D_0. (The other basis given by Proposition 3.2 is the standard basis $\boldsymbol{e} = \{e_1, e_2\}$.)

We now define classes of Lipschitz bounded m-monotone functions, which are needed for us to generalize Theorem 2.1.

Definition 5.8. Let $D_0 \subset [0,\infty)^d$ be a convex polytope, let $0 < B < \infty$, and let $m \ge 2$ be an integer. For all vertices v of D_0, for all orthonormal bases $\boldsymbol{e}$ given by Proposition 3.2, and all $\boldsymbol{i} \in \mathbb{I}_{m-1}$, let $0 < \Gamma_{\boldsymbol{e},\boldsymbol{i}} < \infty$, and let $\boldsymbol{\Gamma}$ be the set of all such $\Gamma_{\boldsymbol{e},\boldsymbol{i}}$. Let $\mathcal{C}_d^m(D_0, B, \boldsymbol{\Gamma})$ be the class of functions $f \in \mathcal{C}_d^m(D_0, B)$ such that for all $\boldsymbol{i} \in \mathbb{I}_{m-2}$ and orthonormal bases $\boldsymbol{e}$ (given by Proposition 3.2 for any vertex of D_0) the function $f_{\boldsymbol{e}}^{(\boldsymbol{i})}$ is Lipschitz in the following sense. For each $\boldsymbol{i} \in \mathbb{I}_{m-1}$ and $\boldsymbol{j} \in \mathbb{I}_{m-2}$, $\boldsymbol{i} - \boldsymbol{j}$ is 1 in a single coordinate, which we denote $\alpha_{\boldsymbol{i},\boldsymbol{j}}$. Then, for *any* $\boldsymbol{j} \in \mathbb{I}_{m-2}$, assume for all $x, x + \lambda e_{\alpha_{\boldsymbol{i},\boldsymbol{j}}} \in D_0$ that $|f^{(\boldsymbol{j})}(x + \lambda e_{\alpha_{\boldsymbol{i},\boldsymbol{j}}}) - f^{(\boldsymbol{j})}(x)| \le \Gamma_{\boldsymbol{e},\boldsymbol{i}}|\lambda|$.

Remark 5.9. Let $\mathcal{C}_d^{m,\circ}(D_0, B, \boldsymbol{\Gamma})$ denote the subset of $f \in \mathcal{C}_d^m(D_0, B, \boldsymbol{\Gamma})$ that are $(m-1)$-times continuously differentiable. Note that for such f, we have $|f^{(\boldsymbol{i})}| \le \Gamma_{\boldsymbol{i}}$. Let C^{m-1} denote the class of $(m-1)$-times continuously differentiable functions. Then $C^1 \cap \mathcal{C}_d(D_0, B, \boldsymbol{\Gamma})$ is L_∞-dense in $\mathcal{C}_d(D_0, B, \boldsymbol{\Gamma})$, so $\mathcal{C}_d^{m,\circ}(D_0, B, \boldsymbol{\Gamma}) := C^{m-1} \cap \mathcal{C}_d^m(D_0, B, \boldsymbol{\Gamma})$ is dense in $\mathcal{C}_d^m(D_0, B, \boldsymbol{\Gamma})$ [10]; see also Lemma 1.1 of [24]. This means that any L_∞ (bracketing or metric) entropy bound for $\mathcal{C}_d^{m,\circ}(D_0, B, \boldsymbol{\Gamma})$ implies the same bound on $\mathcal{C}_d^m(D_0, B, \boldsymbol{\Gamma})$. We will use this in our proofs.

The following lemma provides uniform bounds on the smoothness of the functions in $\mathcal{C}_d^m(D, 1)|_{G_{\boldsymbol{i},\boldsymbol{j}}}$, which, together with Theorem 5.11, allows us to later prove Lemma 5.15 and thus prove the main Theorem 5.16.

Lemma 5.10. *Let D_0 be a convex polytope, let $\mathcal{C}_d^{m,\circ}(D_0, 1)$ be as defined in Remark 5.9 for $m \ge 2$, and let $f \in \mathcal{C}_d^{m,\circ}(D_0, 1)$.*

A. *If x is interior to D_0 such that $B(x, r_0) \subset D_0$, $r_0 > 0$, then for any $\boldsymbol{j} \in \mathbb{I}_j$, $0 \le j \le m-1$, we have $|f^{(\boldsymbol{j})}(x)| \le K_j / r_0^j$ for a constant $0 < K_j$.*
B. *Let $\boldsymbol{e}$ be any orthonormal basis of $\mathbb{R}^d$ such that $f_{\boldsymbol{e}}^{(\boldsymbol{i})}$ is convex for all $\boldsymbol{i} \in \mathbb{I}_j$, $0 \le j \le m-2$. Assume we have $d(x, \partial D_0, e_i) \ge \delta_i$, where $\delta_i > 0$ for $i = 1, \ldots, d$. Let $\delta = (\delta_1, \ldots, \delta_d)$. Then for any $\boldsymbol{j} \in \mathbb{I}_j$, $0 \le j \le m-1$, we have $|f_{\boldsymbol{e}}^{(\boldsymbol{j})}(x)| \le K_j / \delta^{\boldsymbol{j}}$.*

Proof. We first show part A. We will show, by induction on $l < m-1$, that for $\boldsymbol{i} \in \mathbb{I}_l$, we have that $(f^{(\boldsymbol{i})})|_{B(x, r_0/2^l)} \in \mathcal{C}(B(x, \frac{r_0}{2^l}), \frac{2^{l(l+3)/2}}{r_0^l})$. When $l = m-1$, the statement $(|f^{(\boldsymbol{i})}|)|_{B(x, r_0/2^l)} \le \frac{2^{l(l+3)/2}}{r_0^l}$ holds. The base case of $l = 0$ is satisfied trivially by assumption, since $f \in \mathcal{C}_d^{m,\circ}(D_0, 1)$. Now we show the induction hypothesis holds for a general $l \le m-1$ by assuming it holds for the $l-1$ case. Take $\boldsymbol{i} \in \mathbb{I}_l$. Write (non-uniquely) $\boldsymbol{i} = \boldsymbol{i}_1 + \boldsymbol{i}_2$ for $\boldsymbol{i}_1 \in \mathbb{I}_{l-1}$ and $\boldsymbol{i}_2 \in \mathbb{I}_1$. Since $l-1 \le m-2$, $f^{(\boldsymbol{i}_1)}$ is convex by assumption, so by the induction hypothesis $(f^{(\boldsymbol{i}_1)})|_{B(x, r_0/2^{l-1})} \in \mathcal{C}(B(x, \frac{r_0}{2^{l-1}}), \frac{2^{(l-1)(l+2)/2}}{r_0^{l-1}})$. Note for $z \in B\left(x, \frac{r_0}{2^l}\right)$, that $d(x, \partial B\left(x, \frac{r_0}{2^{l-1}}\right), e_i) \ge r_0/2^l$ for any i. Since $f^{(\boldsymbol{i})} = (f^{(\boldsymbol{i}_1)})^{(\boldsymbol{i}_2)}$, Lemma 3.1 implies for any $z \in B\left(x, \frac{r_0}{2^l}\right)$ that $|f^{(\boldsymbol{i})}(z)| \le \frac{2}{r_0/2^l} \frac{2^{(l-1)(l+2)/2}}{r_0^{l-1}} = \frac{2^{l(l+3)}}{r_0^l}$. Thus part A has been shown. Part B follows from part A by a simple scaling argument: let A be the diagonal matrix of δ, so that $g(y) := f(Ay)$ is defined on a hyperrectangle E where $d(Ax, \partial E, e_i) \ge 1$. Note $\delta^{\boldsymbol{i}} f^{(\boldsymbol{i})} = g^{(\boldsymbol{i})}$, and then apply part A. □

Theorem 5.11. *Let $m \ge 2$. Let $D = \prod_{i=1}^d [a_i, b_i]$ be a hyperrectangle, with $-\infty < a_i < b_i < \infty$. For all $\boldsymbol{i} \in \mathbb{I}_{m-1}$, let $0 < \Gamma_{\boldsymbol{i}} < \infty$ and let $\boldsymbol{\Gamma} := \{\Gamma_{\boldsymbol{i}} : \boldsymbol{i} \in \mathbb{I}_{m-1}\}$. Let $0 < B \le \max_{\boldsymbol{i} \in \mathbb{I}_{m-1}} \Gamma_{\boldsymbol{i}} (b-a)^{\boldsymbol{i}}$. Then there exists $c \equiv c_{m,d}$ such that for all $\epsilon > 0$,*

$$N_{[\,]}\left(\epsilon, \mathcal{C}_d^m(D, B, \boldsymbol{\Gamma}), L_\infty\right) \le \exp\left\{ c \left(\frac{\max_{\boldsymbol{i} \in \mathbb{I}_{m-1}} \Gamma_{\boldsymbol{i}} (b-a)^{\boldsymbol{i}}}{\epsilon} \right)^{d/m} \right\}, \tag{68}$$

where $a := (a_1, \ldots, a_d)$ and $b := (b_1, \ldots, b_d)$. Note in (68), $\mathcal{C}_d^m(D, B, \boldsymbol{\Gamma})$ may trivially be replaced by $\mathcal{C}_d^m(D, B, \boldsymbol{\Gamma})|_G$ for any $G \subset D$.

The proof proceeds via several lemmas. The following lemma was inspired in part by Lemma 1 in [26].

Lemma 5.12. *Let $\mathcal{F}$ be a class of functions on $\prod_{i=1}^d [0, L_i]$, $0 < L_i < \infty$, let $x \in [0,1]^d$, and let*

$$\mathcal{G} := \left\{ y \mapsto \int_0^1 f(x + t(y-x))dt : f \in \mathcal{F} \right\}.$$

Assume $\log N_{[\,]}(\epsilon, \mathcal{F}, L_\infty) \le \phi(\epsilon) < \infty$ for some function ϕ and all $\epsilon > 0$, and assume further that the ϵ-bracketing cover of $\mathcal{F}$ can be taken to satisfy (66) with A the identity (and where h is replaced by the lower and upper bracket of f). Then there exists $0 < C < \infty$ such that

$$\log N_{[\,]}\left(\epsilon/\phi(\epsilon)^{1/d}, \mathcal{G}, L_\infty\right) \le C\phi(\epsilon).$$

Proof. By (4), we will bound the metric covering number rather than the bracketing number, just for ease of notation. Without loss of generality, assume $\phi(\epsilon)$ takes on integer values and take $x = 0$. Let $\{f_i\}_{i=1}^{e^{\phi(\epsilon)}}$ be an ϵ-L_∞-net for $\mathcal{F}$. For $f \in \mathcal{F}$ write $g(y) := \int_0^1 f(ty)dt = \int_0^1 (f(ty) - f_i(ty))dt + g_i(y)$ where $g_i(y) := \int_0^1 f_i(ty)dt$ and $L_\infty(f - f_i) \le \epsilon$. Define

$$\mathcal{G}_i := \left\{ g(y) = \int_0^1 (f(ty) - f_i(ty))dt : y \in [0,1]^d,\, f \in \mathcal{F},\, L_\infty(f - f_i) \le \epsilon \right\}.$$

Thus $\mathcal{G} \subseteq \cup_i (\mathcal{G}_i + g_i)$.

Now, for each i, $\mathcal{G}_i$ consists of functions g satisfying $L_\infty(g) \le \epsilon$, and also (by (66)) satisfying

$$L_\infty\left(\frac{\partial}{\partial x_j} g\right) \le \epsilon \text{ for } j \in \{1, \dots, d\}.$$

Thus, by Theorem 5.23 (in the Appendix), we see that $\log N(\delta, \mathcal{G}_i, L_\infty) \le C(2\epsilon/\delta)^d$ for a constant C and any $\delta > 0$. Take $\delta = \epsilon/\phi(\epsilon)^{1/d}$ and see

$$\log N(\epsilon/\phi(\epsilon)^{1/d}, \mathcal{G}_i, L_\infty) \le 2C\phi(\epsilon),$$

and let g_{ij}, $1 \le j \le e^{C\phi(\epsilon)}$, denote a corresponding cover. Then $\left\{g_i + g_{ij}\right\}_{i,j}$ is an L_∞-cover of $\mathcal{G}$ with size $(\epsilon/\phi^{1/d}(\epsilon))$ and with cardinality no larger than $e^{(2C+1)\phi(\epsilon)}$, so we are done. □

Remark 5.13. The above lemma depends on (66). Note that in the $d = 1$ case, this property is satisfied for the entire class $\mathcal{C}(L, B, \boldsymbol{\Gamma})$: see Lemma 5.18.

Let $D \subset \prod_{i=1}^d [0, L_i]$ be convex, and for simplicity assume $0 \in D$. Let $0 < B$ and $\boldsymbol{\Gamma} := (\Gamma_1, \dots, \Gamma_d)$. For $m \ge 3$, let

$$\mathcal{G}_d^m(D, B, \boldsymbol{\Gamma}) := \left\{ x \mapsto \int_0^1 \int_0^{z_1} \cdots \int_0^{z_{m-2}} f(sx)\, ds dz_{m-2} \cdots dz_1 : f \in \mathcal{C}_d(D, B, \boldsymbol{\Gamma}) \right\},$$

be a class of functions, where $x \in D$. Note that the functions in $\mathcal{G}_d^m(D, B, \boldsymbol{\Gamma})$ are normalized so their size does not increase with the size of D.

Lemma 5.14. *Fix $D \subset \prod_{i=1}^d [0, L_i]$ be convex with $0 \in D$. Let $\boldsymbol{\Gamma} := (\Gamma_1, \dots, \Gamma_d) \in (0, \infty)^d$ and $L := (L_1, \dots, L_d) \in (0, \infty)^d$. Let $0 < B \le \sum_{i=1}^d \Gamma_i L_i$. Let $m \ge 2$ be an integer and $p \ge 1$. Then, abbreviating $\mathcal{G}_d^m \equiv \mathcal{G}_d^m(D, B, \boldsymbol{\Gamma})$, we have*

$$\log N_{[\,]}(\epsilon \operatorname{Vol}_d(D)^{1/p}, \mathcal{G}_d^m, L_p) \le \log N_{[\,]}\left(\epsilon, \mathcal{G}_d^m, L_\infty\right) \le c_m \left(\frac{\sum_{i=1}^d \Gamma_i L_i}{\epsilon}\right)^{d/m}. \tag{69}$$

Proof. The first inequality of (69) is immediate. The proof of the second inequality is by induction. We can start with the base case of $m = 2$ by identifying $\mathcal{G}_d^2(D, B, \boldsymbol{\Gamma})$ with $\mathcal{C}_d(D, B, \boldsymbol{\Gamma})$ and then the result is by Theorem 2.1. Now we assume the $m - 1$ case holds, i.e.,

$$\log N_{[\,]}\left(\epsilon, \mathcal{G}_d^{m-1}(D, B, \boldsymbol{\Gamma}), L_\infty\right) \le c_{m-1} \left(\frac{\sum_{i=1}^d \Gamma_i L_i}{\epsilon}\right)^{d/(m-1)}, \tag{70}$$

and show (69) holds. By (70) and Lemma 5.12, we have

$$\log N_{[\,]}\left(\frac{\epsilon}{\left(\frac{\sum_{i=1}^d \Gamma_i L_i}{\epsilon}\right)^{\frac{1}{m-1}}}, \mathcal{G}_d^m(D, B, \boldsymbol{\Gamma}), L_\infty\right) \le C_m \left(\frac{\sum_{i=1}^d \Gamma_i L_i}{\epsilon}\right)^{\frac{d}{m-1}} \tag{71}$$

which is equivalent to (69). □

Proof of Theorem 5.11. We consider (only) $f \in \mathcal{C}_d^{m,\circ}(D, B, \boldsymbol{\Gamma})$, by Remark 5.9. Since D has nonempty interior there exists an open ball contained in D, which, by translation, we take to be $B(0, r_0)$ without loss of generality, for $r_0 > 0$. Now, by iterated application of the Fundamental Theorem of Calculus, for any $(m-1)$-times continuously differentiable $h\colon \mathbb{R} \to \mathbb{R}$, we can write

$$h(x) = h(0) + \cdots + \frac{h^{(m-2)}(0)}{(m-2)!} x^{m-2} + \int_0^x \int_0^{z_1} \cdots \int_0^{z_{m-2}} h^{(m-1)}(s) ds dz_{m-2} \cdots dz_1. \tag{72}$$

By applying (72) to $t \mapsto f(ty)$, for any $y \in D$ we can write

$$f(y) = \sum_{i=0}^{m-3} \sum_{\boldsymbol{j} \in \mathbb{I}_i} \frac{1}{\boldsymbol{j}!} f^{(\boldsymbol{j})}(0) y^{\boldsymbol{j}} + \sum_{\boldsymbol{j} \in \mathbb{I}_{m-2}} \binom{m-2}{\boldsymbol{j}} y^{\boldsymbol{j}} I_{m-2}(f^{(\boldsymbol{j})}, y) \tag{73}$$

where

$$I_{m-2}(f^{(\boldsymbol{j})}, y) := \int_0^1 \int_0^{z_1} \cdots \int_0^{z_{m-3}} f^{(\boldsymbol{j})}(sy)\, ds dz_{m-3} \cdots dz_1.$$

Let

$$\mathcal{P}^m := \left\{ y \mapsto \sum_{i=0}^{m-3} \sum_{\boldsymbol{j} \in \mathbb{I}_i} a_{\boldsymbol{j}} y^{\boldsymbol{j}} : 0 \le a_{\boldsymbol{j}} \le c_{\boldsymbol{j}} \right\}$$

where $c_{\boldsymbol{j}} := K_m / r_0^m \boldsymbol{j}!$ and where $K_m := \max_j K_j$ comes from Lemma 5.10.

Now, for $\boldsymbol{i} \in \mathbb{I}_{m-2}$, let $\boldsymbol{j}_\alpha(\boldsymbol{i}) := \boldsymbol{i} + (0, \ldots, 0, 1, 0, \ldots, 0)$, where the 1 is in the α index. Let $\boldsymbol{\Gamma}_{\boldsymbol{i}} := (\boldsymbol{\Gamma}_{\boldsymbol{j}_1(\boldsymbol{i})}, \ldots, \boldsymbol{\Gamma}_{\boldsymbol{j}_d(\boldsymbol{i})})$. This is the vector of Lipschitz constraints for $f^{(\boldsymbol{i})}$. Let

$$\mathcal{F}^m := \left\{ y \mapsto \sum_{\boldsymbol{j} \in \mathbb{I}_{m-2}} \binom{m-2}{\boldsymbol{j}} y^{\boldsymbol{j}} I_{m-2}(g_{\boldsymbol{j}}, y) : g_{\boldsymbol{j}} \in \mathcal{C}_d(D, B, \boldsymbol{\Gamma}_{\boldsymbol{j}}) \right\}. \tag{74}$$

Then $\mathcal{C}_d^m(D, B, \boldsymbol{\Gamma}) \subset \mathcal{P}^m + \mathcal{F}^m$ so

$$N(\epsilon, \mathcal{C}_d^m(D, B, \boldsymbol{\Gamma}), L_\infty) \le N(\epsilon/2, \mathcal{P}^m, L_\infty) N(\epsilon/2, \mathcal{F}^m, L_\infty). \tag{75}$$

Recall by (4), L_∞-ϵ-bracketing numbers equal L_∞-$(\epsilon/2)$-covering numbers, and so simply for ease of notation and without any loss of generality, we form a L_∞ (metric) cover rather than L_∞ bracketing cover.

First we form a cover for $\mathcal{P}^m$. For an integer $N \ge 1$, we can construct a grid to cover $\mathcal{P}^m$ by taking $a_{\boldsymbol{j}} \in \{c_{\boldsymbol{j}}/N, \ldots, N c_{\boldsymbol{j}}/N\}$. Since $\boldsymbol{j} \in \mathbb{I}_i$, $0 \le i \le m$ takes on no more than m^d values, the cover has cardinality N^{m^d}. The L_∞ size is $m^d C L^\mu / N$ where $C := \max_{\boldsymbol{j}} c_{\boldsymbol{j}}$, $L := b - a$, and $\mu := (m-3, \ldots, m-3)$. Take N to be $\lceil \epsilon^{-1} m^d C L^\mu \rceil$. Then we have formed a cover for $\mathcal{P}^m$ in the L_∞ norm with distances no larger than ϵ and log cardinality bounded above by $m^d \log(1 + m^d C L^\mu \epsilon^{-1})$.

Now consider forming an L_∞-cover for $\mathcal{F}^m$. By Lemma 5.14, for a fixed $\boldsymbol{j} \in \mathbb{I}_{m-2}$, we can form an ϵ-L_∞-cover, $h_{\boldsymbol{j},i}$, for $i = 1, \ldots, N_{\boldsymbol{j}}$, for the functions $I_{m-2}(g_{\boldsymbol{j}})$, where $\log N_{\boldsymbol{j}} \le c_m \left(\sum_{i=1}^d \Gamma_{\boldsymbol{j},i} L_i/\epsilon\right)^{d/m}$ and $L_i := b_i - a_i$. Let $f_{\boldsymbol{j},i}(y) := \binom{m-2}{\boldsymbol{j}} y^{\boldsymbol{j}} h_{\boldsymbol{j},i}(y)$, $i = 1, \ldots, N_{\boldsymbol{j}}$. Let $L = (L_1, \ldots, L_d)$. Then for a function $f(y) = \binom{m-2}{\boldsymbol{j}} y^{\boldsymbol{j}} I_{m-2}(g_{\boldsymbol{j}})$, with $L_\infty(f_{\boldsymbol{j},i} - f) \le \epsilon$, we have $L_\infty(f_{\boldsymbol{j},i} - f) \le m^m L^{\boldsymbol{j}} \epsilon$. Equivalently, we can cover the same function class with size ϵ and log cardinality $c_{m,2} \left(\sum_{i=1}^d \Gamma_{\boldsymbol{j},i} L_i L^{\boldsymbol{j}}/\epsilon\right)^{d/m}$. Let $g_{\boldsymbol{j},i}$ denote such a cover. Then the class $\mathcal{F}^m$ is covered by the set of functions $\sum_{\boldsymbol{j}\in\mathbb{I}_{m-2}} g_{\boldsymbol{j},i}$ which have L_∞ distance bounded above by $\sum_{\boldsymbol{j}\in\mathbb{I}_{m-2}} \epsilon$ and log cardinality bounded above by $c_{m,2} \sum_{\boldsymbol{j}\in\mathbb{I}_{m-2}} \left(\sum_{i=1}^d \Gamma_{\boldsymbol{j},i} L_i L^{\boldsymbol{j}}/\epsilon\right)^{d/m}$. Equivalently, $\mathcal{F}^m$ can be covered by a class with $\epsilon/2$ L_∞-distance and log cardinality bounded above by $c_{m,3}\epsilon^{-d/m}(\max_{\boldsymbol{j}\in\mathbb{I}_{m-1}} \Gamma_{\boldsymbol{j}} L^{\boldsymbol{j}})^{d/m}$. Then, by (75), we have

$$\log N(\epsilon, \mathcal{C}_d^m(D, B, \boldsymbol{\Gamma}), L_\infty) \le m^d \log(1 + m^d 2 C L^\mu \epsilon^{-1}) + c_{m,3}\epsilon^{-d/m} \left(\max_{\boldsymbol{j}\in\mathbb{I}_{m-1}} \Gamma_{\boldsymbol{j}} L^{\boldsymbol{j}}\right)^{d/m}.$$

This completes the proof. □

We can now prove the following bound on bracketing entropy of m-monotone function classes, using the same approach used to prove Theorem 3.5. We use the same $G_{\boldsymbol{i},\boldsymbol{j}}$ partition construction as in the proof of Theorem 3.5, except that we modify the δ_i's. First we have the following m-monotone version of Lemma 3.4.

Lemma 5.15. *Let Assumption 1 hold. Fix $k \in \{1, \ldots, d\}$, $\boldsymbol{i} \in I_k$, $\boldsymbol{j} \in J_k$. Then for any $p \ge 1$ and $\epsilon > 0$,*

$$\log N_{[\,]} \left(\epsilon \operatorname{Vol}_d(G_{\boldsymbol{i},\boldsymbol{j}})^{1/p}, \mathcal{C}_d^m(D_0, 1)|_{G_{\boldsymbol{i},\boldsymbol{j}}}, L_p\right) \le c_{D,d} \left(\frac{1}{\epsilon} \max_{\alpha=1,\ldots,k} \frac{\delta_{i_\alpha+1}^{m-1}}{\delta_{i_\alpha}^{m-1}}\right)^{d/m}.$$

Proof. Let $e_{\boldsymbol{i},\boldsymbol{j},\alpha} \equiv e_\alpha$, $\alpha = 1, \ldots, d$ be given by Proposition 3.2, so that $d(G_{\boldsymbol{i},\boldsymbol{j}}, \partial D_0, e_\alpha) \ge \max_{\beta\in 1,\ldots,k} \delta_{i_\beta}/|\langle e_\beta, v_{j_\beta}\rangle|$ for $\alpha = 1, \ldots, k$, and $d(G_{\boldsymbol{i},\boldsymbol{j}}, \partial D_0, e_\alpha) \ge 2/u$ for $\alpha = k+1, \ldots, d$. Recall (21) from the proof of Lemma 3.4, and recall that $\gamma_\alpha \le d L_{\boldsymbol{j},4}^{-1} \max_{\beta\in\{1,\ldots,k\}} \delta_{i_\beta+1}$. Let $\eta_0 := (\delta_{i_{\beta_1}}, \ldots, \delta_{i_{\beta_k}}, 2/u, \ldots, 2/u)$ and $\eta_1 := (\delta_{i_{\beta_1}+1}, \ldots, \delta_{i_{\beta_k}+1}, 4L_{k,1}\rho_{\boldsymbol{j},k+1}, \ldots, 4L_{k,1}\rho_{\boldsymbol{j},d})$. For $\boldsymbol{l} \in \mathbb{I}_{m-1}$, let $\Gamma_{\boldsymbol{l}} := \sup_{x\in G_{\boldsymbol{i},\boldsymbol{j}}, f\in\mathcal{C}_d^m(D_0,1)} |f^{(\boldsymbol{l})}(x)|$. Then by Lemma 5.10 we have

$$\Gamma_{\boldsymbol{l}} \eta_1^{\boldsymbol{l}} \le c \frac{\eta_1^{\boldsymbol{l}}}{\eta_0^{\boldsymbol{l}}} \le c \max_{\beta\in\{1,\ldots,k\}} \frac{\delta_{i_\beta+1}^{m-1}}{\delta_{i_\beta}^{m-1}}.$$

Now we can apply Theorem 5.11. We use the fact that $G_{\boldsymbol{i},\boldsymbol{j}}$ is embedded in a hyperrectangle H with axes given by the orthonormal basis $\boldsymbol{e} := \{e_\alpha\}_{\alpha=1}^d$ specified by Proposition 3.2. Thus $\mathcal{C}_d^m(D, 1)|_{G_{\boldsymbol{i},\boldsymbol{j}}}$ is contained in $\mathcal{C}_d^m(H, 1)|_{G_{\boldsymbol{i},\boldsymbol{j}}}$. Thus, letting $\boldsymbol{\Gamma} := \{\boldsymbol{\Gamma}_{\boldsymbol{l}} : \boldsymbol{l} \in \mathbb{I}_{m-1}\}$, we may apply Theorem 5.11 to $\mathcal{C}^m(H, 1, \boldsymbol{\Gamma})|_{G_{\boldsymbol{i},\boldsymbol{j}}}$ (after applying a rotation, see Remark 5.4). By (68) (and the logic leading to (6)), the proof is complete. □

We are now in a position to prove the following m-monotone version of Theorem 3.5.

Theorem 5.16. *Assume the setup and conclusions of (68) hold. Fix $p \geq 1$. Then for all $\epsilon > 0$,*

$$\log N_{[\,]}\left(\epsilon, \mathcal{C}_d^m(D,B), L_p\right) \leq C\epsilon^{-d/m}\left(B\prod_{i=1}^d (b_i - a_i)^{1/p}\right)^{-d/m}. \tag{76}$$

Proof. The same scaling argument as given in (the beginning of) the proof of Theorem 3.5 applies here, since rescalings of m-monotone functions are still m-monotone. Thus assume $D \subset [0,1]^d$ and $B = 1$. Now we define

$$u \equiv u_D := r_D/2 \wedge 2^{-m(1+p(m-1))^2(2+p(m-1))} \wedge \min_{k\in\{1,\ldots,d-1\}} \min_{\boldsymbol{j}\in J_k, e\in E_{\boldsymbol{j}}} \frac{d^+(x_{\boldsymbol{j}}, \partial_r G_{\boldsymbol{j}}, e)}{L_{k,2}} \tag{77}$$

(where $L_{k,2}$ is still given by (16)). We define A and $\{\delta_i\}_{i=1}^A$ as in (9), by

$$\delta_i := \exp\left\{p\left(\frac{p-1/(m-1)}{p+2/(m-1)}\right)^{i-1}\log\epsilon\right\} \quad \text{for} \quad i = 1,\ldots,A, \text{ and } \delta_0 = 0. \tag{78}$$

For $k \in \{1,\ldots,d\}$, $\boldsymbol{i} \in I_k$ we let $a_{(i_1\ldots i_k)} = 2$ if $i_\alpha = 0$ for any $\alpha \in \{1,\ldots,k\}$, and otherwise let

$$a_{(i_1,\ldots,i_k)} := \prod_{\beta=1}^k a_{i_\beta} := \prod_{\beta=1}^k \epsilon^{1/k}\exp\left\{-p\frac{(p+1/(m-1))^{i_\beta-2}}{(p+2/(m-1))^{i_\beta-1}}\log\epsilon\right\}. \tag{79}$$

When $k = 0$, let $a_A := \epsilon/u$. Now define a by (27), as before, and then

$$\log N_{[\,]}\left(a, \mathcal{C}_d^m(D,1), L_p\right) \leq \sum_{k=0}^d \sum_{\boldsymbol{j}\in J_k}\sum_{\boldsymbol{i}\in I_k} \log N_{[\,]}\left(a_{\boldsymbol{i}}\,\mathrm{Vol}_d(G_{\boldsymbol{i},\boldsymbol{j}})^{1/p}, \mathcal{C}_d^m(D,1)|_{G_{\boldsymbol{i},\boldsymbol{j}}}, L_p\right). \tag{80}$$

holds. We consider the case where $k \in \{1,\ldots,d\}$ (i.e., $k \neq 0$), and compute the sum above over I_k for a fixed $\boldsymbol{j} \in J_k$. We again use the trivial bracket $[-1,1]$ for any $G_{\boldsymbol{i},\boldsymbol{j}}$ where $i_\alpha = 0$ for any $\alpha \in \{1,\ldots,k\}$. By Lemma 5.15, the sum over the remaining terms is bounded above by

$$\sum_{i_1=1}^A \cdots \sum_{i_k=1}^A c_1 a_{\boldsymbol{i}}^{-d/m}\left(\max_{\alpha=1,\ldots,k}\frac{\delta_{i_\alpha+1}^{m-1}}{\delta_{i_\alpha}^{m-1}}\right)^{d/m} \tag{81}$$

which (using $\max_{\alpha\in\{1,\ldots,k\}} 2\delta_{i_\alpha+1}/\delta_{i_\alpha} \leq \prod_{\alpha=1}^k 2\delta_{i_\alpha+1}/\delta_{i_\alpha}$ as in the proof of Theorem 3.5) is bounded above by

$$c_2\sum_{i_1=1}^A \cdots \sum_{i_k=1}^A \prod_{\beta=1}^k \left(\frac{\delta_{i_\beta+1}^{m-1}}{\delta_{i_\beta}^{m-1} a_{i_\beta}}\right)^{d/m}. \tag{82}$$

The constants c_1, c_2 depend on D and d.

We now let

$$\zeta_i \equiv \zeta_{i,k,m} := \left(\epsilon^{1/k}\delta_{i+1}^{m-1}/(\delta_i^{m-1}a_i)\right)^{1/m}. \tag{83}$$

By Lemma 5.17, $\sum_{i=1}^{A} \zeta_i^d \le K_d < \infty$ for a constant K_d. Now,

$$\prod_{\alpha=k+1}^{d} \rho_{j,\alpha}^{m-1} \le C_d \operatorname{Vol}_{d-k}(A_j)^{m-1} \le C_d \operatorname{Vol}_{d-k}(G_j)^{m-1}. \tag{84}$$

Thus we have shown that (82) is bounded above by $K\epsilon^{-d/m}$. where K depends only on D, d, m, and p, but not on ϵ. Therefore, for a different constant K, we have shown that the right side of (80) is bounded above by $K\epsilon^{-d/m}$ for all $\epsilon > 0$.

We now bound the size of the brackets, a. Define I_k^+, I_k^0 as in the proof of Theorem 3.5. Just as in Theorem 3.5, (37) holds (using the current definitions of δ_{i_α} and a_{i_α}). For the middle term on the right side of (37), recall (38), and for the first term recall $a_A := \epsilon/u$. It remains only to bound the last term. We can check that $a_i^p \delta_{i+1} \le \epsilon^{p/k} \zeta_i^m$ (equality holding when $m = 2$). Thus, arguing as in the proof of Theorem 3.5, we can see we need only bound $\epsilon^{p/k} \sum_{\alpha=1}^{A} \zeta_\alpha^m$ which by Lemma 5.17 is bounded above by $\epsilon^{p/k} A_u$. Thus $a \le C\epsilon$ for a constant C not depending on ϵ. This completes the proof. □

In the m-monotone case, we did not relate the constants involved in the bound to the volumes of the faces of D as explicitly as we did in the convex case. The following lemma was used to bound both the cardinality and the size of the brackets in Theorem 5.16.

Lemma 5.17. *Define A, u, δ_i, and a_i by (9), (77), (78), and (79), respectively. Assume $0 < \epsilon \le 1$. Let ζ_i be defined by (83). Then for any $\gamma \ge 1$,*

$$\sum_{i=1}^{A} \zeta_i^\gamma \le 2^\gamma/(2^\gamma - 1).$$

Proof. Straightforward algebra shows

$$\zeta_i = \exp\left\{ \frac{(m-1)p \left(\frac{p+\frac{1}{m-1}}{p+\frac{2}{m-1}} \right)^i}{m(1+(m-1)p)^2} \log\epsilon \right\}.$$

Thus, for $\alpha = 1, \dots, A-1$, further algebra shows

$$\frac{\zeta_\alpha^m}{\zeta_{\alpha+1}^m} = \exp\left\{ (m-1)p \frac{(1+p(m-1))^{\alpha-2}}{(2+p(m-1))^{\alpha+1}} \log\epsilon \right\},$$

which (since $\alpha \le A-1$) is bounded above by

$$\exp\left\{ \frac{(m-1)p}{(1+p(m-1))^2(2+p(m-1))} \frac{(1+p(m-1))^{A-1}}{(2+p(m-1))^{A-1}} \log\epsilon \right\}$$
$$\le \exp\left\{ \frac{\log u}{(1+p(m-1))^2(2+p(m-1))} \right\} =: R^{-m}.$$

Now, note that $\zeta_A \le 1$ (since $\epsilon \le 1$), and by its definition (77), we see $R \ge 2$. The rest of the proof follows in fashion similar to the proof of Lemma 3.6. □

5.1.2. Univariate m-monotonicity

In this subsection, we prove that when $d = 1$, (66) holds, and so the conclusion of Theorem 5.11 holds with $\mathcal{C}^*$ replaced by the full class $\mathcal{C}$. We point out that [26] provide bracketing entropy for classes of bounded (univariate) m-monotone functions on a compact interval in Hellinger distance, but their result does not immediately give a bound for L_∞-bracketing entropy. The methods of the previous section do give such a bound though. Recall λ is Lebesgue measure.

Lemma 5.18. *Let $\mathcal{F}$ be a class of functions on $[0, L]$ and let $\mathcal{G} := \left\{x \mapsto \int_0^x f d\lambda : f \in \mathcal{F}\right\}$ be the class of primitives of $\mathcal{F}$ on $[0, L]$. Assume $\log N_{[\,]}(\epsilon, \mathcal{F}, L_\infty) \le \phi(\epsilon) < \infty$ for a function ϕ and $\epsilon > 0$. Then there exists $0 < C < \infty$ such that*

$$\log N_{[\,]}\left(\epsilon/\phi(\epsilon), L^{-1}\mathcal{G}, L_\infty\right) \le C\phi(\epsilon). \tag{85}$$

Proof. By (4), we will bound the metric covering number rather than the bracketing number, just for ease of notation. We take $L = 1$, by rescaling: if $\mathcal{F}$ and $\mathcal{G}$ are classes of functions defined on $[0, 1]$, let $\tilde{\mathcal{F}}$ be $\{x \mapsto f(xL) : f \in \mathcal{F}\}$ and define $\tilde{\mathcal{G}} := \{x \mapsto Lg(xL) : g \in \mathcal{G}\}$. Then $N(\epsilon, \mathcal{F}, L_\infty) = N(\epsilon, \tilde{\mathcal{F}}, L_\infty)$ and $\tilde{\mathcal{G}}$ is the class of primitives of $\tilde{\mathcal{F}}$. We see that $N(\epsilon, \mathcal{G}, L_\infty) = N\left(\epsilon, L^{-1}\tilde{\mathcal{G}}, L_\infty\right)$ so we can take $L = 1$. Now, by the Fundamental Theorem of Calculus, (66) holds, and so we can apply Lemma 5.12. This completes the proof. □

Now let $\mathcal{G}^1 \equiv \mathcal{G}^1(L, B)$ be the class of non-decreasing functions f on $[0, L]$ satisfying $0 \le f \le B$, and let

$$\mathcal{G}^k \equiv \mathcal{G}^k(L, B) := \left\{ g(x) = \int_0^x \int_0^{z_1} \cdots \int_0^{z_{k-2}} f(s) ds dz_{k-2} \cdots dz_1 : f \in \mathcal{G}^1 \right\} \tag{86}$$

where g is defined on $[0, L]$. Note, when f is continuous, then $0 \le g^{(k-1)} \le B$.

Lemma 5.19. *Fix $L, B > 0$ and define $\mathcal{G}^k(L, B)$ by (86). Let $k \ge 2$ be an integer and $p \ge 1$. We have*

$$\log N(\epsilon L^{1/p}, \mathcal{G}^k(L, B), L_p) \le \log N_{[\,]}\left(\epsilon, \mathcal{G}^k(L, B), L_\infty\right) \le c_k \left(\frac{BL^{k-1}}{\epsilon}\right)^{1/k}. \tag{87}$$

Proof. This follows from using Lemma 5.18 in the proof of Lemma 5.14 (i.e., from the fact that (66) is satisfied when $d = 1$). □

For $L, B, \Gamma > 0$, let $\mathcal{C}^m(L, B, \Gamma)$ be the class of m-monotone functions (per Definition 5.1) f on $[0, L]$ satisfying $0 \le f \le B$ and $f^{(m-2)}$ is Lipschitz with constant Γ.

Theorem 5.20. *Let $B, L, \Gamma > 0$. Let $m \ge 2$ be an integer. Assume $B < \Gamma L^{m-1}$. Then there exists a constant $c_m > 0$ (not depending on B, L, Γ, or ϵ) such that for all $\epsilon > 0$,*

$$\log N_{[\,]}\left(\epsilon, \mathcal{C}^m(L, B, \Gamma), L_\infty\right) \le c_m \left(\frac{\Gamma L^{m-1}}{\epsilon}\right)^{1/m}. \tag{88}$$

Proof. This follows from Theorem 5.11 together with the fact that (66) is satisfied when $d = 1$. □

5.2. Entropy of classes related to level set estimation

Now, we consider the entropy of certain classes of functions related to estimating the level sets of convex functions; we may consider, for instance, estimating the level set of a convex or concave regression function, or of a so-called log- or s-concave density.[3] We refer to [14] for the definition of log- or s-concavity. We are specifically concerned with the case when the level set is a polytope.

In the present paper, we are concerned with bracketing entropy bounds, rather than statistical methodological developments. We provide here an extremely brief discussion of methodology to motivate the classes we are developing bounds for. The methodology is to first pick a bandwidth $h > 0$, then to minimize an objective function Q, based on i.i.d. data points, over functions of the form $f|_{\mathcal{L}(f)+B(0,h)}$ for convex f. The bandwidth h will converge to 0 as the sample size increases. The model will need to satisfy some regularity conditions for M-estimation theory to apply (it must be such that if $H(p_f, p_{f_0}) \le \delta$ then $l_H(\mathcal{L}(f), D_0) \le C\delta$ where $H^2(p_f, p_{f_0}) := \int (\sqrt{p_f} - \sqrt{p_{f_0}})^2 d\mu$ is Hellinger distance for some dominating measure μ on the sample space, and p_f is the data generating density corresponding to convex function f).

We now proceed to find bracketing entropy bounds for a function class that will govern rates of convergence for the above procedure, specifically when the level set is a polytope. We operate under the following basic setup or assumption.

Assumption 2. Let C_0 be a closed, bounded convex set in $\mathbb{R}^d$ with nonempty interior. Let $f_0 \in \mathcal{C}(C_0, B)$ satisfy $\inf_{x\in C_0} f_0(x) < \inf_{x\in\partial C_0} f_0(x)$. Let $\lambda \in \mathbb{R}$ satisfy $\inf_{x\in C_0} f_0(x) < \lambda < \sup_{x\in C_0} f_0(x)$. Assume further that $D_0 := \mathcal{L}(f_0)$ is a polytope.

The assumption restricts attention to functions which attain their minimum on the interior of C_0 and are strictly larger everywhere on their boundary than the minimum. (This is somewhat analogous to the assumption of so-called "coercivity", except that we are restricting attention to a compact domain C_0.) For a function f, define $\mathcal{L}(f) \equiv \mathcal{L}_\lambda(f) := \left\{x \in \mathbb{R}^d : f(x) = \lambda\right\}$. For two sets $C, D \subset \mathbb{R}^d$, define the Hausdorff distance between them by

$$l_H(C, D) := \max\left(\sup_{x\in D} \inf_{y\in C} \|x - y\|, \sup_{y\in C} \inf_{x\in D} \|x - y\| \right).$$

Let $\mathcal{S}_\delta := \{D : l_H(D, D_0) \le \delta\}$. Define set addition $A_1 + A_2 := \{a_1 + a_2 : a_1 \in A_1, a_2 \in A_2\}$ and recall that $f|_{A_1}$ is the restriction of a function f to the set A_1. For $h > 0$, the class of functions we consider is

$$\mathcal{C}_{\delta,h}(C_0, B) := \left\{ f|_{\mathcal{L}(f)+B(0,h)} : f \in \mathcal{C}(C_0, B), \mathcal{L}(f) \in S_\delta \right\}; \tag{89}$$

this is a class of bounded convex functions on C_0 restricted to a neighborhood about their λ-level set (which is generally not a convex set).

Let F_j, $j = 1, \ldots, N$, be the facets of D_0. Let $T_j := F_j + B(0, \delta + h)$. Note that T_j is a convex set. Thus for any $f \in \mathcal{C}(C_0, B)$ with $l_H(\mathcal{L}(f), D_0) \le \delta$, we have $\mathcal{L}(f) + B(0, h) \subseteq \cup_{j=1}^N T_j$. Thus for $\epsilon > 0$, restating (5), we have

$$N_{[\,]}\left(N^{1/p}\epsilon, \mathcal{C}_{\delta,h}(C_0, B), L_p\right) \le \prod_{j=1}^N N_{[\,]}\left(\epsilon, \mathcal{C}_{\delta,h}(C_0, B)|_{T_j}, L_p\right). \tag{90}$$

To bound the terms on the right side of the above display we will use Theorem 2.1. To do so, we need to compute $\mathrm{Vol}_d(T_j)$, we need to find a hyperrectangle containing T_j, and we need

[3] As shown by [14] in the univariate log- or s-concave cases and [32] in the multivariate log-concave case, bracketing entropies of log- or s-concave density classes are related to those of bounded concave function classes.

to show that $\mathcal{C}_{\delta,h}(C_0, B)|_{T_j}$ is Lipschitz, which we will do by using an idea from the proof of Theorem 3.5.

By Assumption 2, $0 < l_H(D_0, C_0)$. Let $\eta_0 := l_H(D_0, C_0)$. Thus for any $f \in \mathcal{C}_{\delta,h}(C_0, B)$, for δ, h small enough, for any $j \in \{1, \ldots, N\}$, $l_H(T_j, C_0) \geq \eta_0 - \delta - h > \eta_0/2$. Thus by Lemma 3.1,

$$\mathcal{C}_{\delta,h}(C_0, B)|_{T_j} \subset \mathcal{C}(T_j, B, \boldsymbol{\Gamma}) \tag{91}$$

where $\boldsymbol{\Gamma} := (4B/\eta_0, \ldots, 4B/\eta_0)$. Now, let $V_{0,j} := \mathrm{Vol}_{d-1}(F_j)$. Then

$$\mathrm{Vol}_d(T_j) \leq 2V_{0,j} \cdot 2(\delta + h). \tag{92}$$

Next, note each facet F_j is compact so can be embedded in a hyperrectangle. Let $\prod_{i=1}^{d-1}[a_{j,i}, b_{j,i}]$ be a hyperrectangle of minimum volume containing F_j (after an orthogonal rotation). Then by its definition, T_j is (after rotation) contained in $\left(\prod_{i=1}^{d-1}[a_{j,i} - \delta - h, b_{j,i} + \delta + h]\right) \times [-\delta - h, \delta + h]$, for $\delta, h > 0$ small enough. Thus by Theorem 2.1, for $1 \leq p < \infty$, for any $j \in \{1, \ldots, N\}$, and for δ, h small enough, $\log N_{[\,]}\left(\epsilon, \mathcal{C}(T_j, B, \boldsymbol{\Gamma}), L_p\right)$ is bounded above by

$$\begin{aligned} & c\left(\frac{\mathrm{Vol}_d(T_j)^{1/p}}{\epsilon}\right)^{d/2}\left(B + \frac{4B}{\eta_0}\left(2(\delta+h) + \sum_{i=1}^{d-1}(2(\delta+h) + b_{j,i} - a_{j,i})\right)\right)^{d/2} \\ & \leq c_2\left(\frac{V_{0,j}^{1/p}(\delta+h)}{\epsilon}\right)^{d/2}\left(B + \frac{4B}{\eta_0}\left(\sum_{i=1}^{d-1} 2(b_{j,i} - a_{j,i})\right)\right)^{d/2}. \end{aligned}$$

Let $U_{0,j} := \sum_{i=1}^{d-1}(b_{j,i} - a_{j,i})$. Then the right side of the above display is bounded above by $c_3\left(V_{0,j}^{1/p}(\delta+h)/\epsilon\right)^{d/2}\left(B + BU_{0,j}/\eta_0\right)^{d/2}$. Thus the log of the right of (90) is bounded above by $\sum_{j=1}^{N} c_3\left(V_{0,j}^{1/p}(\delta+h)/\epsilon\right)^{d/2}\left(B + BU_{0,j}/\eta_0\right)^{d/2}$. Thus we have shown the following.

Theorem 5.21. *Let Assumption 2 hold. Fix $\delta > 0$, $h > 0$, and let $\mathcal{C}_{\delta,h}(C_0, B)$ be defined as in (89). Then*

$$\log N_{[\,]}\left(\epsilon, \mathcal{C}_{\delta,h}(C_0, B), L_p\right) \leq S_0\left(\frac{\delta+h}{\epsilon}\right)^{d/2},$$

where S_0 is a constant depending on B, f_0, and C_0.

This result suggests that fast rates of convergence may be possible when estimating polytopal level sets of convex functions. As mentioned above, we do not here develop a full estimation procedure. To do so will require studying the optimal choice of the bandwidth h. If we can take $h = \delta$ for instance, then the bound is of order $(\delta/\epsilon)^{d/2}$. When $d = 2$ or 3, one can then compute the entropy integral (see, e.g., [21,41,42]), $\int_0^\delta \sqrt{\log N_{[\,]}\left(\epsilon, \mathcal{C}_{\delta,h}(C_0, B), L_2\right)} d\epsilon = \delta^{d/4}\int_0^\delta \epsilon^{-d/4} d\epsilon$ and see that it is of order $\delta^{d/4}\delta^{1-d/4} = \delta$. This corresponds, at least heuristically, to a $\sqrt{n}$ rate of convergence (the rate $\sqrt{n}$ arises from combining, e.g., Lemma 3.4.2 (p. 324) and Theorem 3.2.5 (p. 289) of [42]). These calculations are only suggestive in nature (and indeed we have not formally proposed an estimator in a specific model!). They are presented to explain potential repercussions of Theorem 5.21.

Appendix

Theorem 5.22 (*John's Theorem, [31]; Theorem 13.4.1 [36]*)**.** *Let $K \subset \mathbb{R}^d$ be a bounded closed convex body with nonempty interior. Then there exists an ellipsoid E of maximal volume such that $E \subseteq K \subseteq nE$.*

Theorem 5.23 (*Theorem 2.7.1 of [42]*)**.** *Let $\mathcal{L}$ be a class of functions on $\prod_{i=1}^d [0, L_i]$, $0 < L_i < \infty$, such that for all $f \in \mathcal{L}$, we have $L_\infty(f) \le B < \infty$ and f has Lipschitz constant in the direction x_i given by $\Gamma_i < \infty$. Then*

$$\log N(\epsilon, \mathcal{L}, L_\infty) \le K \left(\frac{B + \sum_{i=1}^d \Gamma_i L_i}{\epsilon} \right)^d. \tag{93}$$

Proof. The theorem is given by [42] when the domain is $[0, 1]^d$ and the sup and Lipschitz bounds are all 1. A scaling argument gives the general form. □

References

[1] F. Balabdaoui, J.A. Wellner, Estimation of a k-monotone density: limit distribution theory and the spline connection, Ann. Statist. 35 (6) (2007) 2536–2564.

[2] F. Balabdaoui, J.A. Wellner, Estimation of a k-monotone density: characterizations, consistency and minimax lower bounds, Statist. Neerlandica 64 (1) (2010) 45–70.

[3] K. Ball, Ellipsoids of maximal volume in convex bodies, Geom. Dedicata 41 (2) (1992) 241–250.

[4] K. Ball, in: S. Levy (Ed.), An Elementary Introduction to Modern Convex Geometry, in: Flavors of Geometry, vol. 31, Cambridge University Press, 1997, pp. 1–58.

[5] L. Birgé, P. Massart, Rates of convergence for minimum contrast estimators, Probab. Theory Related Fields 97 (1–2) (1993) 113–150.

[6] A. Brøndsted, An introduction to convex polytopes, in: Graduate Texts in Mathematics, vol. 90, Springer-Verlag, New York-Berlin, ISBN: 0-387-90722-X, 1983, p. viii+160, 683612.

[7] E. Bronshtein, ϵ-Entropy of convex sets and functions, Sib. Math. J. 17 (1976) 393–398.

[8] V.-E. Brunel, Methods for estimation of convex sets, Statist. Sci. 33 (4) (2018) 615–632.

[9] M. Cule, R. Samworth, M. Stewart, Maximum likelihood estimation of a multi-dimensional log-concave density, J. R. Stat. Soc. Ser. B Stat. Methodol. 72 (5) (2010) 545–607.

[10] M.-O. Czarnecki, L. Rifford, Approximation and regularization of Lipschitz functions: convergence of the gradients, Trans. Amer. Math. Soc. 358 (10) (2006) 4467–4520.

[11] T. Dey, J. Pach, Extremal problems for geometric hypergraphs, Discrete Comput. Geom. 19 (4) (1998) 473–484.

[12] S. Dharmadhikari, K. Joag-Dev, Unimodality, convexity, and applications, in: Probability and Mathematical Statistics, Academic Press Inc., Boston, MA, 1988.

[13] C.R. Doss, Bracketing numbers of convex functions on polytopes, 2017, arXiv.org arXiv:1506.00034v2.

[14] C.R. Doss, J.A. Wellner, Global rates of convergence of the MLEs of log-concave and s-concave densities, Ann. Statist. 44 (3) (2016) 954–981.

[15] C.R. Doss, J.A. Wellner, Inference for the mode of a log-concave density, 2016, arXiv.org, arXiv:1611.1034 8.

[16] C.R. Doss, J.A. Wellner, Mode-constrained estimation of a log-concave density, 2016, arXiv.org, myehosthttp://arxiv.org/abs/1611.10335v2[arXiv:1611.10335v2].

[17] C.R. Doss, G. Weng, Bandwidth selection for kernel density estimators of multivariate level sets and highest density regions, Electron. J. Stat. 12 (2) (2018) 4313–4376.

[18] D. Dryanov, Kolmogorov Entropy for classes of convex functions, Constr. Approx. (ISSN: 0176-4276) 30 (1) (2009) 137–153.

[19] R. Dudley, Central limit theorems for empirical measures, Ann. Probab. 6 (6) (1978) 899–929.

[20] R.M. Dudley, A course on empirical processes, in: école D'été de Probabilités de Saint-FLour, XII—1982, in: Lecture Notes in Math., vol. 1097, Springer, Berlin, 1984, pp. 1–142.

[21] R. Dudley, Uniform central limit theorems, in: Cambridge Studies in Advanced Mathematics, vol. 63, Cambridge University Press, Cambridge, 1999.

[22] B. Fang, A. Guntuboyina, B. Sen, Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy-Krause variation, 2019, arXiv.org arXiv:1903.01395v1.

[23] H. Federer, Geometric measure theory, in: Die Grundlehren der mathematischen Wissenschaften, Band vol. 153, Springer-Verlag New York Inc., New York, 1969.

[24] F. Gao, Entropy estimate for k-monotone functions via small ball probability of integrated brownian motion, Electron Commun. Probab. (2008).

[25] F. Gao, J.A. Wellner, Entropy estimate for high-dimensional monotonic functions, J. Multivariate Anal. 98 (9) (2007) 1751–1764.

[26] F. Gao, J.A. Wellner, On the rate of convergence of the maximum likelihood estimator of a k-monotone density, Sci. China Ser. A-Math. 52 (7) (2009) 1525–1538.

[27] F. Gao, J.A. Wellner, Entropy of convex functions on $\mathbb{R}^d$, Constr. Approx. 65 (2) (2017) 181–28.

[28] B. Grünbaum, Convex polytopes, in: With the cooperation of Victor Klee, M. A. Perles and G. C. Shephard. Pure and Applied Mathematics, Vol. 16, Interscience Publishers John Wiley & Sons, Inc., New York, 1967.

[29] A. Guntuboyina, B. Sen, Covering numbers for convex functions, IEEE Trans. Inform. Theory 59 (4) (2013) 1957–1965.

[30] A. Guntuboyina, B. Sen, Global risk bounds and adaptation in univariate convex regression, Probab. Theory Related Fields (ISSN: 0178-8051) 163 (1–2) (2015) 379–411, 3405621.

[31] F. John, Extremum problems with inequalities as subsidiary conditions, in: Studies and Essays Presented To R. Courant on His 60th Birthday, January 8, 1948, Interscience Publishers, Inc., New York, N. Y., 1948, pp. 187–204.

[32] A.K. Kim, R.J. Samworth, Global rates of convergence in log-concave density estimation, Ann. Statist. 44 (6) (2016) 2756–2779.

[33] R. Koenker, I. Mizera, Quasi-concave density estimation, Ann. Statist. (ISSN: 0090-5364) 38 (5) (2010) 2998–3027.

[34] A. Kolmogorov, V. Tihomirov, ε-Entropy and ε-capacity of sets in function spaces, AMS Transl.: Ser. 2 17 (1961) 277–364.

[35] C.W. Lee, Subdivisions and triangulations of polytopes, in: Handbook of Discrete and Computational Geometry, CRC, Boca Raton, FL, 1997, pp. 271–290.

[36] J. Matoušek, Lectures on discrete geometry, in: Graduate Texts in Mathematics, vol. 212, Springer-Verlag, New York, New York, NY, 2002.

[37] R. Rockafellar, Convex Analysis, in: Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ, 1970.

[38] B. Rothschild, E. Straus, On triangulations of the convex hull of n points, Combinatorica 5 (2) (1985) 167–179.

[39] E. Seijo, B. Sen, Nonparametric least squares estimation of a multivariate convex regression function, Ann. Statist. 39 (3) (2011) 1633–1657.

[40] A. Seregin, J.A. Wellner, Nonparametric estimation of multivariate convex-transformed densities, Ann. Statist. (ISSN: 0090-5364) 38 (6) (2010) 3751–3781, With supplementary material available online.

[41] S.A. van de Geer, Empirical Processes in M-Estimation, Cambridge University Press, 2000.

[42] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes, in: Springer Series in Statistics, Springer-Verlag, New York, 1996.

[43] R.E. Williamson, On multiply monotone functions and their laplace transforms, (Ph.D. thesis), ProQuest LLC, Ann Arbor, MI, University of Pennsylvania, 1955.

[44] R. Williamson, Multiply monotone functions and their Laplace transforms, Duke Math. J. 23 (1956) 189–207.

[45] M. Xu, M. Chen, J. Lafferty, Faithful variable screening for high-dimensional convex regression, Ann. Statist. 44 (6) (2016) 2624–2660.