JMLR: Workshop and Conference Proceedings vol (2012) 1–13 25th Annual Conference on Learning Theory

L_1 Covering Numbers for Uniformly Bounded Convex Functions

author names withheld

Editor: Under Review for COLT 2012

Abstract

In this paper we study the covering numbers of the space of convex and uniformly bounded functions in multi-dimension. We find optimal upper and lower bounds for the ϵ-covering number $M(\mathcal{C}([a,b]^d, B), \epsilon; L_1)$ in terms of the relevant constants, where $d \geq 1$, $a < b \in \mathbb{R}$, $B > 0$, and $\mathcal{C}([a,b]^d, B)$ denotes the set of all convex functions on $[a,b]^d$ that are uniformly bounded by B. We summarize previously known results on covering numbers for convex functions and also provide alternate proofs of some known results. Our results have direct implications in the study of rates of convergence of empirical minimization procedures as well as optimal convergence rates in the numerous convexity constrained function estimation problems.

Keywords: convexity constrained function estimation, empirical risk minimization, Hausdorff distance, Kolmogorov entropy, L_1 metric, metric entropy, packing numbers.

1. Introduction

Ever since the work of Kolmogorov and Tihomirov (1961), covering numbers (and their logarithms, known as metric entropy numbers) have been studied extensively in a variety of disciplines. For a subset $\mathcal{F}$ of a metric space $(\mathcal{X}, \rho)$, the *ϵ-covering number* $M(\mathcal{F}, \epsilon; \rho)$ is defined as the smallest number of balls of radius ϵ whose union contains $\mathcal{F}$. Covering numbers capture the *size* of the underlying metric space and play a central role in a number of areas in information theory and statistics, including nonparametric function estimation, density estimation, empirical processes and machine learning.

In this paper we study the covering numbers of the space of convex and uniformly bounded functions in multi-dimension. Specifically, we find optimal upper and lower bounds for the ϵ-covering number $M(\mathcal{C}([a,b]^d, B), \epsilon; L_1)$ in terms of the relevant constants, where $d \geq 1$, $a, b \in \mathbb{R}$, $B > 0$, and $\mathcal{C}([a,b]^d, B)$ denotes the set of all convex functions on $[a,b]^d$ that are uniformly bounded by B. We also summarize previously known results on covering numbers for convex functions. The special case of the problem when $d = 1$ has been recently addressed in Dryanov (2009). Prior to Dryanov (2009), the only other result on the covering numbers of convex functions is due to Bronshtein (1976) (also see (Dudley, 1999, Chapter 8)) who considered convex functions that are uniformly bounded and uniformly Lipschitz with a *known* Lipschitz constant.

In recent years there has been an upsurge of interest in nonparametric function estimation under convexity based constraints, especially in multi-dimension. In general function estimation, it is well-known (see e.g., Birgé (1983); Le Cam (1973); Yang and Barron (1999); Guntuboyina (2011b)) that the covering numbers of the underlying function space can be used to characterize optimal rates of convergence. They are also useful for studying the rates of convergence of empirical minimization procedures (see e.g., Van de Geer (2000); Birgé and Massart (1993)). Our results have direct implications in this regard in the context of understanding the rates of convergence of the numerous convexity constrained function estimators, e.g., the nonparametric least squares estimator of a convex regression function studied in Seijo and Sen (2011); Hannah and Dunson (2011); the maximum likelihood estimator of a log-concave density in multi-dimension studied in Seregin and Wellner (2010); Cule et al. (2010); Dümbgen et al. (2011). Also, similar problems that crucially use convexity/concavity constraints to estimate sets have also received recent attention in the statistical and machine learning literature, see e.g., Guntuboyina (2011a); Gardner et al. (2006), and our results can be applied in such settings.

The paper is organized as follows. In Section 2, we set up notation, describe the previous work on covering numbers of convex functions and provide motivation for our main result, which is proved in Section 3. We conclude in Section 4 with a brief summary of the paper and some open questions that remain. The appendix contains the proof of an auxiliary result.

2. Motivation

The first result on covering numbers for convex functions was proved by Bronshtein (1976), who considered convex functions defined on a cube in $\mathbb{R}^d$ that are uniformly bounded and uniformly Lipschitz. Specifically, let $\mathcal{C}([a,b]^d, B, L)$ denote the class of real-valued convex functions defined on $[a,b]^d$ that are uniformly bounded in absolute value by B and uniformly Lipschitz with constant L. In Theorem 6 of Bronshtein (1976), he proved that for ϵ sufficiently small, the logarithm of $M(\mathcal{C}([a,b]^d, B, L), \epsilon; L_\infty)$ can be bounded from above and below by a positive constant (not depending on ϵ) multiple of $\epsilon^{-d/2}$. Note that the L_∞ distance between two functions f and g on $[a,b]^d$ is defined as $||f-g||_\infty := \sup_{x \in [a,b]^d} |f(x) - g(x)|$.

Bronshtein's proof of the upper bound on $M(\mathcal{C}([a,b]^d, B, L), \epsilon; L_\infty)$ is based on the following result on covering numbers of convex sets proved in the same paper. For $\Gamma > 0$, let $\mathcal{K}^{d+1}(\Gamma)$ denote the set of all compact, convex subsets of the ball in $\mathbb{R}^{d+1}$ of radius Γ centered at the origin. In Theorem 3 (and Remark 1) of Bronshtein (1976), he proved that there exist positive constants c and ϵ_0 depending only on d such that

$$\log M(\mathcal{K}^{d+1}(\Gamma), \epsilon; \ell_H) \leq c \left(\frac{\Gamma}{\epsilon}\right)^{d/2} \qquad \text{for } \epsilon \leq \Gamma \epsilon_0, \tag{1}$$

where ℓ_H denotes the Hausdorff distance defined by

$$\ell_H(B, C) := \max\left(\sup_{x \in B} \inf_{y \in C} |x - y|, \sup_{x \in C} \inf_{y \in B} |x - y|\right) \qquad \text{for } B, C \in \mathcal{K}^{d+1}(\Gamma).$$

A more detailed account of Bronshtein's proof of (1) can be found in Section 8.4 of Dudley (1999).

Bronshtein proved the upper bound on $M(\mathcal{C}([a,b]^d, B, L), \epsilon; L_\infty)$ by relating the L_∞ distance between two functions in $\mathcal{C}([a,b]^d, B, L)$ to the Hausdorff distance between their epigraphs, which allowed him to use (1). However, he did not state the dependence of the upper bound on the constants a, b, B and L. We state Bronshtein's upper bound result below showing the explicit dependence on the constants a, b, B and L. The proof of the result can be found in the Appendix.

Theorem 1 *There exist positive constants c and ϵ_0, depending only on the dimension d, such that, for every $B, L > 0$ and $b > a$, we have, for every $\epsilon \leq \epsilon_0(B + L(b-a))$,*

$$\log M\left(\mathcal{C}([a,b]^d, B, L), \epsilon; L_\infty\right) \leq c\left(\frac{\epsilon}{B + L(b-a)}\right)^{-d/2}.$$

Note that Bronshtein worked with the class $\mathcal{C}([a,b]^d, B, L)$ where the functions are uniformly Lipschitz. However, in convexity-based function estimation problems, one usually does not have a known uniform Lipschitz bound on the unknown function class. This leads to difficulties in the analysis of empirical minimization procedures via Bronshtein's result. To the best of our knowledge, there does not exist any other result on the covering numbers of convex functions that deals with all $d \geq 1$ and does not require the Lipschitz constraint.

In the absence of the uniformly Lipschitz constraint (i.e., if one works with the class $\mathcal{C}([a,b]^d, B)$ instead of $\mathcal{C}([a,b]^d, B, L)$), the covering numbers under the L_∞ metric are infinite. In other words, the space $\mathcal{C}([a,b]^d, B)$ is not totally bounded under the L_∞ metric. This can be seen, for example, by noting that the functions

$$f_j(t) := \max\left(0, 1 - 2^j t\right), \qquad \text{for } t \in [0,1],$$

are in $\mathcal{C}([0,1], 1)$, for all $j \geq 1$, and satisfy

$$||f_j - f_k||_\infty \geq |f_j(2^{-k}) - f_k(2^{-k})| = 1 - 2^{j-k} \geq 1/2,$$

for all $j < k$.

This motivated us to study the covering numbers of the class $\mathcal{C}([a,b]^d, B)$ under a different metric, namely the L_1 metric. We recall that under the L_1 metric, the distance between two functions f and g on $[a,b]^d$ is defined as

$$||f - g||_1 := \int_{x \in [a,b]^d} |f(x) - g(x)| dx.$$

Our main result in this paper shows that if one works with the L_1 metric as opposed to L_∞, then the covering numbers of $\mathcal{C}([a,b]^d, B)$ are finite. Moreover, their logarithms are bounded from above and below by constant multiples of $\epsilon^{-d/2}$ for sufficiently small ϵ.

The special case of our main result for $d = 1$ has been recently established by Dryanov (2009) who actually proved it for every L_p metric with $1 \leq p < \infty$. Dryanov's proof of the upper bound for $M(\mathcal{C}([a,b], B), \epsilon; L_p)$ is based on the application of Bronshtein's bound for covering numbers of $\mathcal{C}([c,d], B, L)$ for suitable subintervals $[c,d] \subset (a,b)$ and for suitable values of L. Unfortunately, his selection of these subintervals is rather complicated. In contrast, our proofs for both the upper and lower bounds work for all $d \geq 1$ and are much simpler than Dryanov's. The disadvantage with our approach, however, is that our proof of the upper bound result only works for the L_1 metric and does not generalize to the L_p metric, $1 < p < \infty$. Our lower bound argument, on the other hand, is valid for all $1 \leq p < \infty$.

3. L_1 - covering number bounds for $\mathcal{C}([a,b]^d, B)$

In this section, we prove upper and lower bounds for the ϵ-covering number of $\mathcal{C}([a,b]^d, B)$ under the L_1 metric. Let us start by noting a simple scaling identity that allows us to take $a = 0, b = 1$ and $B = 1$ without loss of generality. For each $f \in \mathcal{C}([a,b]^d, B)$, let us define $\tilde{f}$ on $[0,1]^d$ by $\tilde{f}(x) := f(a\mathbf{1} + (b-a)x)/B$, where $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^d$. Clearly $\tilde{f} \in \mathcal{C}([0,1]^d, 1)$ and, for $1 \le p < \infty$,

$$B^p \int_{x \in [0,1]^d} \left| \tilde{f}(x) - g(x) \right|^p dx = (b-a)^{-d} \int_{y \in [a,b]^d} \left| f(y) - Bg\left(\frac{y - a\mathbf{1}}{b-a}\right) \right|^p dy.$$

It follows that covering f to within ϵ in the L_p metric on $[a,b]^d$ is equivalent to covering $\tilde{f}$ to within $(b-a)^{-d/p}\epsilon/B$ in the L_p metric on $[0,1]^d$. Therefore, for $1 \le p < \infty$,

$$M(\mathcal{C}([a,b]^d, B), \epsilon; L_p) = M(\mathcal{C}([0,1]^d, 1), (b-a)^{-d/p}\epsilon/B, L_p). \tag{2}$$

3.1. Upper Bound for $M(\mathcal{C}([a,b]^d, B), \epsilon; L_1)$

Theorem 2 *There exist positive constants c and ϵ_0, depending only on the dimension d, such that, for every $B > 0$ and $b > a$, we have,*

$$\log M\left(\mathcal{C}([a,b]^d, B), \epsilon; L_1\right) \le c \left(\frac{\epsilon}{B(b-a)^d}\right)^{-d/2},$$

for every $\epsilon \le \epsilon_0 B(b-a)^d$.

Proof [Proof of Theorem 2] The scaling identity (2) lets us take $a = 0$, $b = 1$ and $B = 1$. For $f \in \mathcal{C}([0,1]^d, 1)$, we define its (bounded) epigraph $V_f \subseteq \mathbb{R}^{d+1}$ to be the compact, convex set defined by

$$V_f = \left\{(x_1, \ldots, x_d, x_{d+1}) : (x_1, \ldots, x_d) \in [0,1]^d \text{ and } f(x_1, \ldots, x_d) \le x_{d+1} \le 1\right\}. \tag{3}$$

For every $(x_1, \ldots, x_{d+1}) \in V_f$, we clearly have $x_1^2 + \cdots + x_{d+1}^2 \le d + 1$. As a result, $V_f \in \mathcal{K}^{d+1}(\sqrt{d+1})$.

In the following lemma, we relate the L_1 distance between the functions f and g to the Hausdorff distance between V_f and V_g. The proof of the lemma is provided at the end of this proof.

Lemma 3 *For every pair of functions f and g in $\mathcal{C}([0,1]^d, 1)$, we have*

$$||f - g||_1 \le (1 + 20d)\ell_H(V_f, V_g), \tag{4}$$

where V_f and V_g are defined as in (3).

Inequality (4), along with a simple relationship between covering numbers and packing numbers, see e.g., Theorem 1.2.1 of Dudley (1999), implies that

$$M\left(\mathcal{C}([0,1]^d, 1), \epsilon; L_1\right) \le M\left(\mathcal{K}^{d+1}(\sqrt{d+1}), \frac{\epsilon}{2(1+20d)}; \ell_H\right).$$

Thus from (1), we deduce the existence of two positive constants c and ϵ_0, depending only on d, such that

$$\log M\left(\mathcal{C}([0,1]^d, 1), \epsilon; L_1\right) \leq c\epsilon^{-d/2} \qquad \text{whenever } \epsilon \leq \epsilon_0,$$

which completes the proof of the theorem. ■

Proof [Proof of Lemma 3] For $f \in \mathcal{C}([0,1]^d, 1)$ and $x \in (0,1)^d$, let $m_f(x)$ denote any subgradient of the convex function f at x. Fix two functions f and g in $\mathcal{C}([0,1]^d, 1)$ with $\ell_H(V_f, V_g) = \rho > 0$. Our first step is to observe that

$$|f(x) - g(x)| \leq \rho\,(1 + ||m_f(x)|| + ||m_g(x)||) \qquad \text{for every } x \in (0,1)^d, \tag{5}$$

where $||m_f(x)||$ denotes the Euclidean norm of the subgradient vector $m_f(x) \in \mathbb{R}^d$. To see this, fix $x \in (0,1)^d$ with $f(x) \neq g(x)$. We assume, without loss of generality, that $f(x) < g(x)$. Clearly $(x, f(x)) \in V_f$ and because $\ell_H(V_f, V_g) = \rho$, there exists $(x', y') \in V_g$ with $||(x, f(x)) - (x', y')|| \leq \rho$. Since $f(x) < g(x)$, the point $(x, f(x))$ lies outside the convex set V_g and we can thus take $y' = g(x')$. By the definition of the subgradient, we have

$$g(x') \geq g(x) + \left\langle m_g(x), x' - x \right\rangle.$$

Therefore,

$$\begin{aligned} 0 \leq g(x) - f(x) &= g(x) - g(x') + g(x') - f(x) \\ &\leq \left\langle m_g(x), x - x' \right\rangle + |g(x') - f(x)| \\ &\leq ||m_g(x)||||x - x'|| + |g(x') - f(x)| \\ &\leq \sqrt{||m_g(x)||^2 + 1}||(x, f(x)) - (x', y')|| \\ &\leq \rho\sqrt{||m_g(x)||^2 + 1} \leq \rho(1 + ||m_g(x)||). \end{aligned}$$

Note that the Cauchy-Schwarz inequality has been used twice in the above chain of inequalities. We have thus shown that $g(x) - f(x) \leq \rho(1 + ||m_g(x)||)$ in the case when $f(x) < g(x)$. One would have a similar inequality in the case when $f(x) > g(x)$. Combining these two, we obtain (5).

As a consequence of (5), we get

$$\begin{aligned} ||f - g||_1 &= \int_{[0,1]^d \setminus [\rho, 1-\rho]^d} |f(x) - g(x)| dx + \int_{[\rho, 1-\rho]^d} |f(x) - g(x)| dx \\ &\leq 2\left(1 - (1 - 2\rho)^d\right) + \rho\left(1 + \int_{[\rho,1-\rho]^d} ||m_f(x)|| dx + \int_{[\rho,1-\rho]^d} ||m_g(x)|| dx\right) \\ &\leq \rho\left(1 + 4d + \int_{[\rho,1-\rho]^d} ||m_f(x)|| dx + \int_{[\rho,1-\rho]^d} ||m_g(x)|| dx\right), \end{aligned}$$

where we have used the inequality $(1 - 2\rho)^d \geq 1 - 2d\rho$.

To complete the proof of (4), we show that $\int_{[\rho,1-\rho]^d} ||m_f(x)||dx \leq 8d$ for every $f \in \mathcal{C}([0,1]^d, 1)$. We write $m_f(x) = (m_f(x)(1), \ldots, m_f(x)(d)) \in \mathbb{R}^d$ and use the definition of the subgradient to note that for every $x \in [\rho, 1-\rho]^d$ and $1 \leq i \leq d$,

$$f(x + te_i) - f(x) \geq t\, m_f(x)(i) \tag{6}$$

for $t > 0$ sufficiently small, where e_i is the unit vector in the ith coordinate direction i.e., $e_i(j) := 1$ if $i = j$ and 0 otherwise. Dividing both sides by t and letting $t \downarrow 0$, we would get $m_f(x)(i) \leq f'(x; e_i)$ (we use $f'(x;v)$ to denote the directional derivative of f in the direction v; directional derivatives exist as f is convex). Using (6) for $t < 0$, we get $m_f(x)(i) \geq -f'(x; -e_i)$. Combining these two inequalities, we get

$$|m_f(x)(i)| \leq |f'(x; e_i)| + |f'(x; -e_i)| \qquad \text{for } i = 1, \ldots, d.$$

As a result,

$$\begin{aligned} \int_{[\rho,1-\rho]^d} ||m_f(x)||dx &\leq \sum_{i=1}^{d} \int_{[\rho,1-\rho]^d} |m_f(x)(i)|dx \\ &\leq \sum_{i=1}^{d} \left(\int_{[\rho,1-\rho]^d} |f'(x;e_i)|dx + \int_{[\rho,1-\rho]^d} |f'(x;-e_i)|dx \right). \end{aligned}$$

We now show that for each i, both the integrals $\int_{[\rho,1-\rho]^d} |f'(x; e_i)|$ and $\int_{[\rho,1-\rho]^d} |f'(x; -e_i)|$ are bounded from above by 4. Assume, without loss of generality, that $i = 1$ and notice

$$\int_{[\rho,1-\rho]^d} |f'(x;e_1)|dx \leq \int_{(x_2,\ldots,x_d)\in[\rho,1-\rho]^{d-1}} \left(\int_{\rho}^{1-\rho} |f'(x;e_1)|dx_1 \right) dx_2 \ldots dx_d. \tag{7}$$

We fix $(x_2, \ldots, x_d) \in [\rho, 1-\rho]^{d-1}$ and focus on the inner integral. Let $v(z) := f(z, x_2, \ldots, x_d)$ for $z \in [0,1]$. Clearly v is a convex function on $[0,1]$ and its right derivative, $v_r'(x_1)$ at the point $z = x_1 \in (0,1)$ equals $f'(x; e_1)$ where $x = (x_1, \ldots, x_d)$. The inner integral thus equals $\int_{\rho}^{1-\rho} |v_r'(z)|dz$. Because of the convexity of v, its right derivative $v_r'(z)$ is non-decreasing and satisfies

$$v(y_2) - v(y_1) = \int_{y_1}^{y_2} v_r'(z)dz \qquad \text{for } 0 < y_1 < y_2 < 1.$$

Consequently,

$$\begin{aligned} \int_{\rho}^{1-\rho} |v_r'(z)|dz &\leq \sup_{\rho \leq c \leq 1-\rho} \left(-\int_{\rho}^{c} v_r'(z)dz + \int_{c}^{1-\rho} v_r'(z)dz \right) \\ &= \sup_{\rho \leq c \leq 1-\rho} \left(v(\rho) + v(1-\rho) - 2v(c) \right). \end{aligned}$$

The function $v(z) = f(z, x_2, \ldots, x_d)$ clearly satisfies $|v(z)| \leq 1$ because $f \in \mathcal{C}([0,1]^d, 1)$. This implies that $\int_{\rho}^{1-\rho} |v_r'(z)|dz \leq 4(1-2\rho) \leq 4$. The inequality (7) therefore gives

$$\int_{[\rho,1-\rho]^d} |f'(x;e_1)|dx \leq \int_{(x_2,\ldots,x_d)\in[\rho,1-\rho]^{d-1}} \left(\int_{\rho}^{1-\rho} |v_r'(z)|dz \right) dx_2 \ldots dx_d \leq 4.$$

Similarly, by working with left derivatives as opposed to right, we can prove that

$$\int_{[\rho,1-\rho]^d} |f'(x;-e_1)|dx \leq 4.$$

Therefore,

$$\int_{[\rho,1-\rho]^d} ||m_f(x)||dx \leq \sum_{i=1}^{d} \left(\int_{[\rho,1-\rho]^d} |f'(x;e_i)|dx + \int_{[\rho,1-\rho]^d} |f'(x;-e_i)|dx \right) \leq 8d,$$

thereby completing the proof of Lemma 3. ■

Remark 4 *The proof of Theorem 2 is crucially based on Lemma 3 which bounds the L_1 distance between two functions in $\mathcal{C}([0,1]^d, 1)$ by a constant multiple of the Hausdorff distance between their epigraphs. This is not true if L_1 is replaced by L_p for $p > 1$. Indeed, if $d = 1$ and $f_\alpha(x) := \max(0, 1 - (x/\alpha))$ for $0 < \alpha \leq 1$ and $g(x) := 0$ for all $x \in [0,1]$, then it can be easily checked that for $1 \leq p < \infty$,*

$$||f_\alpha - g||_p := \frac{\alpha^{1/p}}{(1+p)^{1/p}} \quad \textit{and} \quad \ell_H(V_{f_\alpha}, V_g) := \frac{\alpha}{\sqrt{1+\alpha^2}}.$$

As α can be arbitrarily close to zero, this clearly rules out any inequality of the form (4) with the L_1 metric replaced by L_p for $1 < p \leq \infty$. Therefore, our proof of Theorem 2 will break down for the L_p metric with $p > 1$. However, Theorem 2 does indeed hold for all $1 \leq p < \infty$. The proof requires different techniques and can be found in Guntuboyina and Sen (2012).

3.2. Lower bound for $M(\mathcal{C}([a,b]^d, B), \epsilon; L_1)$

Theorem 5 *There exist positive constants c and ϵ_0, depending only on the dimension d, such that for every $B > 0$ and $b > a$, we have*

$$\log M\left(\mathcal{C}([a,b]^d, B), \epsilon; L_1\right) \geq c \left(\frac{\epsilon}{B(b-a)^d} \right)^{-d/2},$$

for $\epsilon \leq \epsilon_0 B(b-a)^d$.

Proof As before, by the scaling identity (2), we take $a = 0$, $b = 1$ and $B = 1$. We prove that for ϵ sufficiently small, there exists an ϵ-packing subset of $\mathcal{C}([0,1]^d, 1)$ of log-cardinality larger than a constant multiple of $\epsilon^{-d/2}$. By a packing subset of $\mathcal{C}([0,1]^d, 1)$, we mean a subset F satisfying $||f - g||_1 \geq \epsilon$ whenever $f, g \in F$ with $f \neq g$.

Fix $0 < \eta \leq 4(2 + \sqrt{d-1})^{-2}$ and let $k := k(\eta)$ be the positive integer satisfying

$$k \leq \frac{2\eta^{-1/2}}{2 + \sqrt{d-1}} < k + 1 \leq 2k. \tag{8}$$

Consider the intervals $I(i) = [u(i), v(i)]$ for $i = 1, \ldots, k$, such that

1. $0 \le u(1) < v(1) \le u(2) < v(2) \le \cdots \le u(k) < v(k) \le 1$,

2. $v(i) - u(i) = \sqrt{\eta}$, for $i = 1, \dots, k$,

3. $u(i+1) - v(i) = \frac{1}{2}\sqrt{\eta(d-1)}$ for $i = 1, \dots, k-1$.

Let $\mathcal{S}$ denote the set of all d-dimensional cubes of the form $I(i_1) \times \cdots \times I(i_d)$ where $i_1, \dots, i_d \in \{1, \dots, k\}$. The cardinality of $\mathcal{S}$, denoted by $|\mathcal{S}|$, is clearly k^d.

For each $S \in \mathcal{S}$ with $S = I(i_1) \times \cdots \times I(i_d)$ where $I(i_j) = [u(i_j), v(i_j)]$, let us define the function $h_S : [0,1]^d \to \mathbb{R}$ as

$$\begin{aligned} h_S(x) = h_S(x_1, \dots, x_d) &:= \frac{1}{d}\sum_{j=1}^{d}\left[u^2(i_j) + \{v(i_j) + u(i_j)\}\{x_j - u(i_j)\}\right] \\ &= f_0(x) + \frac{1}{d}\sum_{j=1}^{d}\{x_j - u(i_j)\}\{v(i_j) - x_j\}, \qquad (9) \end{aligned}$$

where $f_0(x) := \frac{1}{d}\left(x_1^2 + \cdots + x_d^2\right)$, for $x \in [0,1]^d$. The functions $h_S, S \in \mathcal{S}$ have the following four key properties:

1. h_S is affine and hence convex.

2. For every $x \in [0,1]^d$, we have $h_S(x) \le h_S(1, \dots, 1) \le 1$.

3. For every $x \in S$, we have $h_S(x) \ge f_0(x)$. This is because whenever $x \in S$, we have $u(i_j) \le x_j \le v(i_j)$ for each j, which implies $\{x_j - u(i_j)\}\{v(i_j) - x_j\} \ge 0$.

4. Let $S, S' \in \mathcal{S}$ with $S \ne S'$. For every $x \in S'$, we have $h_S(x) \le f_0(x)$. To see this, let $S' = I(i'_1) \times \cdots \times I(i'_d)$ with $I(i'_j) = [u(i'_j), v(i'_j)]$. Let $x \in S'$ and fix $1 \le j \le d$. If $I(i_j) = I(i'_j)$, then $x_j \in I(i_j) = [u(i_j), v(i_j)]$ and hence

$$\{x_j - u(i_j)\}\{v(i_j) - x_j\} \le \frac{\{v(i_j) - u(i_j)\}^2}{4} = \frac{\eta}{4}.$$

If $I(i_j) \ne I(i'_j)$ and $u(i'_j) < v(i'_j) < u(i_j) < v(i_j)$, then

$$\{x_j - u(i_j)\}\{v(i_j) - x_j\} \le -\{u(i_j) - v(i'_j)\}^2 \le -\frac{d-1}{4}\eta.$$

The same above bound holds if $u(i_j) < v(i_j) < u(i'_j) < v(i'_j)$. Because $S \ne S'$, at least one of i_j and i'_j will be different. Consequently,

$$\begin{aligned} h_S(x) &= f_0(x) + \sum_j \{x_j - u(i_j)\}\{v(i_j) - x_j\} \\ &\le f_0(x) + \sum_{j: i_j = i'_j} \frac{\eta}{4} - \sum_{j: i_j \ne i'_j} (d-1)\frac{\eta}{4} \le f_0(x). \end{aligned}$$

Let $\{0,1\}^{\mathcal{S}}$ denote the collection of all $\{0,1\}$-valued functions on $\mathcal{S}$. The cardinality of $\{0,1\}^{\mathcal{S}}$ clearly equals $2^{|\mathcal{S}|}$ (recall that $|\mathcal{S}| = k^d$).

For each $\theta \in \{0,1\}^{\mathcal{S}}$, let

$$g_\theta(x) := \max\left(\max_{S\in\mathcal{S}:\theta(S)=1} h_S(x), f_0(x)\right).$$

The first two properties of $h_S, S \in \mathcal{S}$ ensure that $g_\theta \in \mathcal{C}([0,1]^d, 1)$. The last two properties imply that

$$g_\theta(x) = h_S(x)\theta(S) + f_0(x)(1-\theta(S)) \qquad \text{for } x \in S.$$

We now bound from below the L_1 distance between g_θ and $g_{\theta'}$ for $\theta, \theta' \in \{0,1\}^{\mathcal{S}}$. Because the interiors of the cubes in $\mathcal{S}$ are all disjoint, we can write

$$||g_\theta - g_{\theta'}||_1 \geq \sum_{S\in\mathcal{S}} \int_{x\in S} |g_\theta(x) - g_{\theta'}(x)|\, dx = \sum_{S\in\mathcal{S}} \left\{\theta(S) \neq \theta'(S)\right\} \int_{x\in S} |h_S(x) - f_0(x)| dx.$$

Note that from (9) and by symmetry, that the value of integral

$$\zeta := \int_{x\in S} |h_S(x) - f_0(x)| dx$$

is the same for all $S \in \mathcal{S}$. We have thus shown that

$$||g_\theta - g_{\theta'}||_1 \geq \zeta \Upsilon(\theta, \theta') \qquad \text{for all } \theta, \theta' \in \{0,1\}^{\mathcal{S}}, \tag{10}$$

where $\Upsilon(\theta,\theta') := \sum_{S\in\mathcal{S}} \{\theta(S) \neq \theta'(S)\}$ denotes the Hamming distance.

The quantity ζ can be computed in the following way. Let $S = I(i_1) \times \cdots \times I(i_d)$ where $I(i_j) = [u(i_j), v(i_j)]$. We write

$$\zeta = \int_{u(i_1)}^{v(i_1)} \cdots \int_{u(i_d)}^{v(i_d)} \frac{1}{d} \sum_{j=1}^{d} \{x_j - u(i_j)\}\{v(i_j) - x_j\} dx_d \ldots dx_1.$$

By the change of variable $y_j = \{x_j - u(i_j)\}/\{v(i_j) - u(i_j)\}$ for $j = 1, \ldots, d$, we get

$$\zeta = \prod_{j=1}^{d} \{v(i_j) - u(i_j)\} \int_{[0,1]^d} \frac{1}{d} \sum_{j=1}^{d} \{v(i_j) - u(i_j)\}^2 y_j(1-y_j) dy.$$

Recalling that $v(i) - u(i) = \sqrt{\eta}$ for all $i = 1, \ldots, k$, we get $\zeta = \eta^{d/2}\eta/6$. Thus, from (10), we deduce

$$||g_\theta - g_{\theta'}||_1 \geq \eta^{d/2}\eta\Upsilon(\theta,\theta')/6 \qquad \text{for all } \theta, \theta' \in \{0,1\}^{\mathcal{S}}. \tag{11}$$

We now use the Varshamov-Gilbert lemma (see e.g., Massart (2007, Lemma 4.7)) which asserts the existence of a subset W of $\{0,1\}^{\mathcal{S}}$ with cardinality, $|W| \geq \exp(|\mathcal{S}|/8)$ such that $\Upsilon(\tau, \tau') \geq |\mathcal{S}|/4$ for all $\tau, \tau' \in W$ with $\tau \neq \tau'$. Thus, from (11) and (8), we get that for every $\tau, \tau' \in W$ with $\tau \neq \tau'$,

$$||g_\theta - g_{\theta'}||_1 \geq \eta^{d/2}\eta\frac{|\mathcal{S}|}{24} = \frac{1}{24}\eta^{d/2}\eta k^d \geq c_1\eta$$

where $c_1 := (2+\sqrt{d-1})^{-d}/24$. Taking $\epsilon := c_1\eta$, we have obtained for $\epsilon \leq \epsilon_0 := 4c_1(2+\sqrt{d-1})^{-2}$, an ϵ-packing subset of $\mathcal{C}([0,1]^d, 1)$ of size $M := |W|$ where

$$\log M \geq \frac{|\mathcal{S}|}{8} = \frac{k^d}{8} \geq \frac{(2+\sqrt{d-1})^{-d}}{8}\eta^{-d/2} = \frac{c_1^{d/2}}{8(2+\sqrt{d-1})^d}\epsilon^{-d/2} = c\epsilon^{-d/2},$$

where c depends only on the dimension d. This completes the proof. ■

Remark 6 *The explicit packing subset constructed in the above proof consists of functions that can be viewed as perturbations of the quadratic function f_0. Previous lower bounds on the covering numbers of convex functions in (Bronshtein, 1976, Proof of Theorem 6) and (Dryanov, 2009, Section 2) (for $d = 1$) are based on perturbations of a function whose graph is a subset of a sphere; a more complicated convex function than f_0. The perturbations of f_0 in the above proof can also be used to simplify the lower bound arguments in those papers.*

Remark 7 *For functions defined on $[0,1]^d$, the L_p metric, $p > 1$, is larger than L_1. Thus, when $a = 0, b = 1$, the conclusion of Theorem 5 also holds for the L_p metric with $p > 1$. The scaling identity (2) then gives the following inequality for arbitrary $a < b$: There exist positive constants c and ϵ_0, depending only on the dimension d, such that for every $p \geq 1$, $B > 0$ and $b > a$, we have*

$$\log M\left(\mathcal{C}([a,b]^d, B), \epsilon; L_p\right) \geq c\left(\frac{\epsilon}{B(b-a)^{d/p}}\right)^{-d/2},$$

for $\epsilon \leq \epsilon_0 B(b-a)^{d/p}$.

4. Concluding remarks

In this paper we have studied the covering numbers of $\mathcal{C}([a,b]^d, B)$, the class of all uniformly bounded convex functions, defined on the hypercube $[a,b]^d$, under the L_1 metric, $1 \leq p \leq \infty$. Our main result shows that we can forgo the assumption of a uniform Lipschitz norm for the underlying class of convex functions (as was assumed in Bronshtein (1976)) and still show that the logarithm of the ϵ-covering number grows at the same order $\epsilon^{-d/2}$, under the L_1 metric. Specifically, we prove that the logarithm of the ϵ-covering number under the L_1 metric is bounded from both above and below by a constant multiple of $\epsilon^{-d/2}$. Our proof of the upper bound in Theorem 2 is based on Lemma 3 which bounds the L_1 distance between two convex functions by a constant multiple of the Hausdorff distance between their epigraphs. Our proof of the lower bound in Theorem 5 is based on an explicit construction of a finite packing subset of the space of uniformly bounded convex functions. In the Appendix, we provide a slightly improved proof of the known upper bound result (Bronshtein, 1976, Theorem 6) for the class of all uniformly bounded (by B) convex functions with a uniform Lipschitz norm L that explicitly shows the dependence of the covering numbers on a, b, B, L.

After the submission of this paper, we managed to extend the results to the case of the L_p metric, for all $1 \leq p < \infty$. These results, which required more involved arguments, can be found in Guntuboyina and Sen (2012).

References

L. Birgé. Approximation dans les espaces metriques et theorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65:181–237, 1983.

L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150, 1993.

E. M. Bronshtein. ϵ-entropy of convex sets and functions. *Siberian Mathematical Journal*, 17:393–398, 1976.

M. L. Cule, R. J. Samworth, and M. I. Stewart. Maximum likelihood estimation of a multi-dimensional log-concave density (with discussion). *Journal of the Royal Statistical Society, Series B*, 72:545–600, 2010.

D. Dryanov. Kolmogorov entropy for classes of convex functions. *Constructive Approximation*, 30:137–153, 2009.

R. M. Dudley. *Uniform Central Limit Theorems.* Cambridge University Press, 1999.

L. Dümbgen, R. J. Samworth, and D. Schuhmacher. Approximation by log-concave distributions with applications to regression. *Annals of Statistics*, 39:702–730, 2011.

R. J. Gardner, M. Kiderlen, and P. Milanfar. Convergence of algorithms for reconstructing convex bodies and directional measures. *Annals of Statistics*, 34:1331–1374, 2006.

A. Guntuboyina. Optimal rates of convergence for the estimation of reconstruction of convex bodies from noisy support function measurements. *Annals of Statistics*, 2011a. to appear.

A. Guntuboyina. Lower bounds for the minimax risk using f divergences, and applications. *IEEE Transactions on Information Theory*, 57:2386–2399, 2011b.

A. Guntuboyina and B. Sen. Covering numbers for convex functions. Available at http://arxiv.org/abs/1204.0147, 2012.

L. A. Hannah and D. Dunson. Bayesian nonparametric multivariate convex regression. Submitted, 2011.

A. N. Kolmogorov and V. M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in function spaces. *Amer. Math. Soc. Transl. (2)*, 17:277–364, 1961.

L. Le Cam. Convergence of estimates under dimensionality restrictions. *Annals of Statistics*, 1:38–53, 1973.

P. Massart. *Concentration inequalities and model selection. Lecture notes in Mathematics*, volume 1896. Springer, Berlin, 2007.

E. Seijo and B. Sen. Nonparametric least squares estimation of a multivariate convex regression function. *Annals of Statistics*, 39:1633–1657, 2011.

A. Seregin and J. A. Wellner. Nonparametric estimation of multivariate convex-transformed densities. *Annals of Statistics*, 38:3751–3781, 2010.

S. Van de Geer. *Applications of Empirical Process Theory*. Cambridge University Press, 2000.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27:1564–1599, 1999.

Appendix A. Proof of the Theorem 1

We mostly follow the proof of Theorem 6 in Bronshtein (1976) but are more careful and use a scaling argument in the end so that the dependence on the various constants involved is maintained. For each $f \in \mathcal{C}([a,b]^d, B, L)$, let us define $\tilde{f}$ on $[0,1]^d$ by $\tilde{f}(x) := f(a\mathbf{1}+(b-a)x)$, where $\mathbf{1} = (1,1,\dots,1) \in \mathbb{R}^d$. Clearly $\tilde{f} \in \mathcal{C}([0,1]^d, B, L(b-a))$ and covering $\tilde{f}$ to within ϵ in the L_∞ metric is equivalent to covering f. Thus,

$$M\left(\mathcal{C}([a,b]^d, B, L), \epsilon; L_\infty\right) = M\left(\mathcal{C}([0,1]^d, B, L(b-a)), \epsilon; L_\infty\right). \tag{12}$$

We can thus take, without loss of generality, $a = 0$ and $b = 1$. Note that, unlike the proof of Theorem 2, we may not take $B = 1$ or $L = 1$ here. For every $f \in \mathcal{C}([0,1]^d, B, L)$, we define the compact, convex set $V_f \subseteq \mathbb{R}^{d+1}$ by

$$V_f := \left\{(x_1,\dots,x_d,x_{d+1}) : (x_1,\dots,x_d) \in [0,1]^d \text{ and } f(x_1,\dots,x_d) \le x_{d+1} \le B\right\}.$$

For every $(x_1,\dots,x_{d+1}) \in V_f$, we have

$$x_1^2 + \cdots + x_d^2 + x_{d+1}^2 \le 1 + \dots 1 + B^2 = d + B^2,$$

which implies that $V_f \in \mathcal{K}^{d+1}(\sqrt{d+B^2})$. We now show that

$$||f-g||_\infty \le (\sqrt{1+L^2})\ell_H(V_f, V_g), \tag{13}$$

for all $f, g \in \mathcal{C}([0,1]^d, B, L)$. To see this, fix $f, g \in \mathcal{C}([0,1]^d, B, L)$ and let $\ell_H(V_f, V_g) = \rho$. Fix $x \in [0,1]^d$ with $f(x) \neq g(x)$. Suppose, without loss of generality, that $f(x) < g(x)$. Now $(x, f(x)) \in V_f$ and because $\ell_H(V_f, V_g) = \rho$, there exists $(x', y') \in V_g$ with $||(x, f(x)) - (x', y')|| \le \rho$. As $f(x) < g(x)$, the point $(x, f(x))$ lies outside the convex set V_g which lets us take $y' = g(x')$. Therefore,

$$\begin{aligned} 0 \le g(x) - f(x) &= g(x) - g(x') + g(x') - f(x) \\ &\le L||x - x'|| + |g(x') - f(x)| \\ &\le \sqrt{L^2+1}\sqrt{||x-x'||^2 + |g(x') - f(x)|^2} \\ &= \sqrt{L^2+1}||(x, f(x)) - (x', y')|| \le (\sqrt{L^2+1})\rho, \end{aligned} \tag{14}$$

where (14) follows from Cauchy-Schwarz inequality. Therefore (13) follows as $x \in [0,1]^d$ is arbitrary in the above argument.

We now use (13) to deduce that

$$M\left(\mathcal{C}([0,1]^d, B, L), \epsilon; L_\infty\right) \le M\left(\mathcal{K}^{d+1}(\sqrt{d+B^2}), \frac{\epsilon}{2\sqrt{1+L^2}}; \ell_H\right).$$

Thus from (1), we deduce the existence of two positive constants c and ϵ_0, depending only on d, such that

$$\log M\left(\mathcal{C}([0,1]^d, B, L), \epsilon; L_\infty\right) \le c\left(\frac{\sqrt{(d+B^2)(1+L^2)}}{\epsilon}\right)^{d/2},$$

if $\epsilon \le \epsilon_0\sqrt{(d+B^2)(1+L^2)}$. By the scaling identity (12), we obtain

$$\log M\left(\mathcal{C}([a,b]^d, B, L), \epsilon; L_\infty\right) \le c\left(\frac{\sqrt{(d+B^2)(1+L^2(b-a)^2)}}{\epsilon}\right)^{d/2}$$

if $\epsilon \le \epsilon_0\sqrt{(d+B^2)(1+L^2(b-a)^2)}$. By another scaling argument, it follows that, for every $\Gamma > 0$,

$$M\left(\mathcal{C}([a,b]^d, B, L), \epsilon; L_\infty\right) = M\left(\mathcal{C}([a,b]^d, B/\Gamma, L/\Gamma), \epsilon/\Gamma; L_\infty\right)$$

and, as a consequence, we get,

$$\log M\left(\mathcal{C}([a,b]^d, B, L), \epsilon; L_\infty\right) \le c\left(\frac{\sqrt{(d\Gamma^2+B^2)(1+L^2(b-a)^2/\Gamma^2)}}{\epsilon}\right)^{d/2}.$$

if $\epsilon \le \epsilon_0\sqrt{(d\Gamma^2+B^2)(1+L^2(b-a)^2/\Gamma^2)}$. Choosing (by differentiation)

$$\Gamma^4 = \frac{B^2L^2(b-a)^2}{d},$$

we deduce finally that, for $\epsilon \le \epsilon_0\left(B + L(b-a)\sqrt{d}\right)$,

$$\log M\left(\mathcal{C}([a,b]^d, B, L), \epsilon; L_\infty\right) \le c\left(\frac{B + L(b-a)\sqrt{d}}{\epsilon}\right)^{d/2}.$$

The $\sqrt{d}$ term above can be absorbed in the constants c and ϵ_0.