L₁ Covering Numbers for Uniformly Bounded Convex Functions

Adityanand Guntuboyina  
Department of Statistics  
University of California, Berkeley  
367 Evans Hall Berkeley, CA 94720-3860

Bodhisattva Sen  
Statistical Laboratory  
University of Cambridge  
Wilberforce Road, Cambridge  
CB3 0WB, England

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 \( M(C([a,b]^d, B), \epsilon; L_1) \) in terms of the relevant constants, where \( d \geq 1 \), \( a < b \in \mathbb{R} \), \( B > 0 \), and \( C([a,b]^d, B) \) denotes the set of all convex functions on \([a,b]^d\) that are uniformly bounded by \( B \). In fact, our lower bound result is valid for the \( L_p \) metric, \( 1 \leq p \leq \infty \). 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, \( L_p \) 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 \((X, \rho)\), the \( \epsilon \)-covering number \( M(\mathcal{F}, \epsilon; \rho) \) is defined as the smallest number of balls of radius \( \epsilon \) 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 \( \epsilon \)-covering number \( M(C([a,b]^d, B), \epsilon; L_1) \) in terms of the relevant constants, where \( d \geq 1 \), \( a, b \in \mathbb{R} \), \( B > 0 \), and \( C([a,b]^d, B) \) 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). However, the proof in that paper is quite complicated and

© 2012 A. Guntuboyina & B. Sen.
does not seem to be easily generalizable to higher dimensions. 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 $\epsilon$ 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 $\epsilon$) multiple of $\epsilon^{-d/2}$. Note that the $L_\infty$ 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 $\Gamma$ centered at the origin. In Theorem 3 (and Remark 1) of Bronshtein (1976), he proved that there exist positive constants $c$ and $\epsilon_0$ depending only on $d$ such that

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

(1)
where $\ell_H$ denotes the Hausdorff distance defined by

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

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 $L_\infty$ 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 $\epsilon_0$, depending only on the dimension $d$, such that, for every $B, L > 0$ and $b > a$, we have, for every $\epsilon \leq \epsilon_0(B + L(b - a))$,

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

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 $L_\infty$ metric are infinite. In other words, the space $\mathcal{C}([a,b]^d, B)$ is not totally bounded under the $L_\infty$ metric. This can be seen, for example, by noting that the functions

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

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

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

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 $L_p$ metric. We recall that under the $L_p$ metric, $1 \leq p < \infty$, the distance between two functions $f$ and $g$ on $[a,b]^d$ is defined as

$$\|f - g\|_p := \left( \int_{x \in [a,b]^d} |f(x) - g(x)|^p dx \right)^{1/p}.$$

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

3
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 $M(C([a, b], B), \epsilon; L_p)$ is based on the application of Bronshtein’s bound for covering numbers of $C([c, d], B, L)$ for suitable subintervals $[c, d] \subset (a, b)$ and for suitable values of $L$. Unfortunately, his selection of these subintervals is rather complicated and does not seem to generalize easily to higher dimensions where one needs to work with subrectangles as opposed to intervals. 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 $C([a, b]^d, B)$

In this section, we prove upper and lower bounds for the $\epsilon$-covering number of $C([a, b]^d, B)$ 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 $f \in C([a, b]^d, B)$, let us define $\tilde{f}$ on $[0, 1]^d$ by $\tilde{f}(x) := f(a1 + (b-a)x)/B$, where $1 = (1, \ldots, 1) \in \mathbb{R}^d$. Clearly $\tilde{f} \in C([0, 1]^d, 1)$ and, for $1 \leq p < \infty$,

$$B^p \int_{x \in [0, 1]^d} \left| \tilde{f}(x) - g(x) \right|^p dx = (b - a)^{-d} \int_{y \in [a, b]^d} \left| f(y) - Bg \left( \frac{y-a1}{b-a} \right) \right|^p dy.$$  

It follows that covering $f$ to within $\epsilon$ 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 $1 \leq p < \infty$,

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

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

**Theorem 2** There exist positive constants $c$ and $\epsilon_0$, depending only on the dimension $d$, such that, for every $B > 0$ and $b > a$, we have,

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

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

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

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

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

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 $C([0,1]^d, 1)$, we have
\[ ||f - g||_1 \leq (1 + 20d)\ell_H(V_f, V_g), \] (4)
where $V_f$ and $V_g$ are defined as in (3).

Inequality (4) implies that
\[ M \left( C([0,1]^d, 1), \epsilon; L_1 \right) \leq M \left( K^{d+1}(\sqrt{d}+1), \frac{\epsilon}{2(1+20d)}; \ell_H \right). \]

Thus from (1), we deduce the existence of two positive constants $c$ and $\epsilon_0$, depending only on $d$, such that
\[ \log M \left( C([0,1]^d, 1), \epsilon; L_1 \right) \leq c\epsilon^{-d/2} \quad \text{whenever } \epsilon \leq \epsilon_0, \]
which completes the proof of the theorem.

Proof [Proof of Lemma 3] For $f \in C([0,1]^d, 1)$ and $x \in (0,1)^d$, let $m_f(x)$ denote any subgradient of the convex function $f$ at $x$. Let $\ell_H(V_f, V_g) = \rho > 0$. Our first step is to observe that
\[ |f(x) - g(x)| \leq \rho (1 + ||m_f(x)|| + ||m_g(x)||) \quad \text{for every } x \in (0,1)^d, \]
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
\[ g(x') \geq g(x) + \langle m_g(x), x' - x \rangle. \]
Therefore,
\[ 0 \leq g(x) - f(x) = g(x) - g(x') + g(x') - f(x) \leq \langle m_g(x), x - x' \rangle + |g(x') - f(x)| \leq ||m_g(x)|| ||x - x'|| + |g(x') - f(x)| \leq \sqrt{||m_g(x)||^2 + 1} ||(x, f(x)) - (x', y')|| \leq \rho \sqrt{||m_g(x)||^2 + 1} \leq \rho(1 + ||m_g(x)||). \]

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
\[
||f - g||_1 = \int_{[0,1]^d \setminus [\rho, 1 - \rho]^d} |f(x) - g(x)| \, dx + \int_{[\rho, 1 - \rho]^d} |f(x) - g(x)| \, dx
\]
\[
\leq 2 \left(1 - (1 - 2\rho)^d \right) + \rho \left(1 + \int_{[\rho, 1 - \rho]^d} ||m_f(x)|| \, dx + \int_{[\rho, 1 - \rho]^d} ||m_g(x)|| \, dx\right)
\]
\[
\leq \rho \left(1 + 4d + \int_{[\rho, 1 - \rho]^d} ||m_f(x)|| \, dx + \int_{[\rho, 1 - \rho]^d} ||m_g(x)|| \, dx\right),
\]
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 \(f \in C([0,1]^d, 1)\). 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 \(i\)th 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
\[
|m_f(x)(i)| \leq |f'(x; e_i)| + |f'(x; -e_i)| \quad \text{for } i = 1, \ldots, d.
\]

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

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, \(v'_r(z)\) 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
\[
v(y_2) - v(y_1) = \int_{y_1}^{y_2} v'_r(z) \, dz \quad \text{for } 0 < y_1 < y_2 < 1.
\]
Consequently,
\[
\int_\rho^{1-\rho} |v'_\rho(z)|dz \leq \sup_{\rho \leq \rho \leq 1-\rho} \left( - \int_\rho^{c} v'_\rho(z)dz + \int_c^{1-\rho} v'_\rho(z)dz \right) = \sup_{\rho \leq \rho \leq 1-\rho} (v(\rho) + v(1 - \rho) - 2v(c)).
\]

The function \( v(z) = f(z, x_2, \ldots, x_d) \) clearly satisfies \( |v(z)| \leq 1 \) because \( f \in C([0,1]^d, 1) \). This implies that \( \int_\rho^{1-\rho} |v'_\rho(z)|dz \leq 4(1 - 2\rho) \leq 4 \). The inequality (7) therefore gives
\[
\int_{[\rho, 1-\rho]^d} |f'(x; e_1)|dx \leq \int_{(x_2, \ldots, x_d) \in [\rho, 1-\rho]^{d-1}} \left( \int_\rho^{1-\rho} |v'_\rho(z)|dz \right) dx_2 \ldots dx_d \leq 4.
\]

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. \( \blacksquare \)

Remark 4 The proof of Theorem 2 is crucially based on Lemma 3 which bounds the \( L_1 \) distance between two functions in \( C([0,1]^d, 1) \) 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 \( 1 \leq p < \infty \),
\[
||f_\alpha - g||_p := \frac{\alpha^{1/p}}{(1 + p)^{1/p}} \quad \text{and} \quad \ell_H(V_{f_\alpha}, V_g) := \frac{\alpha}{\sqrt{1 + \alpha^2}}.
\]

As \( \alpha \) 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 \( p > 1 \). As already mentioned in Section 2, Dryanov (2009) proved the one-dimensional version (i.e., \( d = 1 \)) of Theorem 2 for all \( 1 \leq p < \infty \). It is however unclear if his intricate proof can be extended easily to higher dimensions. Nevertheless, Dryanov’s result suggests that Theorem 2 should hold for the \( L_p \) metric for all \( 1 \leq p < \infty \).

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

Theorem 5 There exist positive constants \( c \) and \( \epsilon_0 \), depending only on the dimension \( d \), such that for every \( B > 0 \) and \( b > a \), we have
\[
\log M \left( C([a, b]^d, B), \epsilon; L_1 \right) \geq c \left( \frac{\epsilon}{B(b-a)^d} \right)^{-d/2},
\]

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 $\epsilon$ sufficiently small, there exists an $\epsilon$-packing subset of $\mathcal{C}([0, 1]^d, 1)$ of 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

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

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

1. $0 \leq u(1) < v(1) \leq u(2) < v(2) \leq \cdots \leq u(k) < v(k) \leq 1$,
2. $v(i) - u(i) = \sqrt{\eta}$, for $i = 1, \ldots, k$,
3. $u(i + 1) - v(i) = \frac{1}{2} \sqrt{\eta(d - 1)}$ for $i = 1, \ldots, k - 1$.

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

For each $S \in S$ 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

$$h_S(x) = h_S(x_1, \ldots, x_d) := \frac{1}{d} \sum_{j=1}^{d} [u^2(i_j) + \{v(i_j) + u(i_j)\}\{x_j - u(i_j)\}] = f_0(x) + \frac{1}{d} \sum_{j=1}^{d}\{x_j - u(i_j)\}\{v(i_j) - x_j\}, \quad (9)$$

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

1. $h_S$ is affine and hence convex.
2. For every $x \in [0, 1]^d$, we have $h_S(x) \leq h_S(1, \ldots, 1) \leq 1$.
3. For every $x \in S$, we have $h_S(x) \geq f_0(x)$. This is because whenever $x \in S$, we have $u(i_j) \leq x_j \leq v(i_j)$ for each $j$, which implies $\{x_j - u(i_j)\}\{v(i_j) - x_j\} \geq 0$.
4. Let $S, S' \in S$ with $S \neq S'$. For every $x \in S'$, we have $h_S(x) \leq f_0(x)$. 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 $1 \leq j \leq d$. If $I(i_j) = I'(i'_j)$, then $x_j \in I(i_j) = [u(i_j), v(i_j)]$ and hence

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

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

$$\{x_j - u(i_j)\}\{v(i_j) - x_j\} \leq -\{u(i_j) - v(i'_j)\}^2 = -\frac{d - 1}{4}\eta.$$
The same above bound holds if \( u(i_j) < v(i_j) < u(i'_j) < v(i'_j) \). Because \( S \neq S' \), at least one of \( i_j \) and \( i'_j \) will be different. Consequently,

\[
h_S(x) = f_0(x) + \sum_j \{x_j - u(i_j)\}\{v(i_j) - x_j\} \\
\leq f_0(x) + \sum_{j:i_j = i'_j} \eta 4 - \sum_{j:i_j \neq i'_j} (d - 1) \eta 4 \leq f_0(x).
\]

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

For each \( \theta \in \{0, 1\}^S \), let

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

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

\[
g_{\theta}(x) = h_S(x)\theta(S) + f_0(x)(1 - \theta(S)) \quad \text{for } x \in S.
\]

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

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

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 \( S \in S \). We have thus shown that

\[
\|g_{\theta} - g_{\theta'}\|_1 \geq \zeta \mathcal{Y}(\theta, \theta') \quad \text{for all } \theta, \theta' \in \{0, 1\}^S,
\]

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

The quantity \( \zeta \) 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

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

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 \( \zeta = \eta^{d/2} \gamma_d \) where

\[
\gamma_d := \int_{[0,1]^d} \frac{1}{d} \sum_{j=1}^d y_j (1 - y_j) dy.
\]

Note that \( \gamma_d \) is a constant that depends on the dimension \( d \) alone. Thus, from (10), we deduce

\[
||g_{\theta} - g_{\theta'}||_1 \geq \gamma_d \eta^{d/2} \eta \Upsilon(\theta, \theta') \quad \text{for all } \theta, \theta' \in \{0,1\}^S. \tag{11}
\]

We now use the Varshamov-Gilbert lemma (see e.g., Massart (2007, Lemma 4.7)) which asserts the existence of a subset \( W \) of \( \{0,1\}^S \) with cardinality, \( |W| \geq \exp(|S|/8) \) such that \( \Upsilon(\tau, \tau') \geq |S|/4 \) 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 \( \tau \neq \tau' \),

\[
||g_{\theta} - g_{\theta'}||_1 \geq \gamma_d \eta^{d/2} \eta^{k_d} \geq c_1 \eta
\]

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

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

where \( c \) depends only on the dimension \( d \). This completes the proof. \( \blacksquare \)

**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 \( \epsilon_0 \), depending only on the dimension \( d \), such that for every \( p \geq 1 \), \( B > 0 \) and \( b > a \), we have

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

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

In this paper we have studied the covering numbers of $C([a, b]^d, B)$, the class of all uniformly bounded convex functions, defined on the hypercube $[a, b]^d$, under the $L_p$ 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 $\epsilon$-covering numbers grow at the same order $\epsilon^{-d/2}$, under the $L_1$ metric. Specifically, we prove that the $\epsilon$-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 $a, b, B, L$.

A natural question that now arises is whether we can find such optimal upper and lower bounds when using the $L_p$ metric, $1 < p < \infty$. As described in Remark 4, our proof of Theorem 2 does not directly generalize to the $L_p$ metric as Lemma 3 no longer holds under the $L_p$ metric for $p > 1$. Thus, this remains an open question for $d > 1$ and will be a topic of future research. The lower bound, on the other hand, works for all $p \geq 1$ and is of the same order $\epsilon^{-d/2}$ as explained in Remark 7. We expect the upper bound for the $L_p$ metric, for $p > 1$, to be of the order $\epsilon^{-d/2}$ as well.

References


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 \( \hat{f} \) on \([0,1]^d\) by \( \hat{f}(x) := f(a1 + (b-a)x) \), where \( 1 = (1,1,\ldots,1) \in \mathbb{R}^d \). Clearly \( \hat{f} \in \mathcal{C}([0,1]^d, B, L(b-a)) \) and covering \( \hat{f} \) to within \( \epsilon \) in the \( L_\infty \) metric is equivalent to covering \( f \). Thus,

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

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

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

\[
x_1^2 + \cdots + x_d^2 + x_{d+1}^2 \leq 1 + \cdots + 1 + B^2 = d + B^2,
\]
which implies that \( V_f \subseteq K^{d+1}(\sqrt{d + B^2}) \). We now show that

\[
\|f - g\|_\infty \leq (\sqrt{1 + L^2})\ell_H(V_f, V_g),
\]

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 \( \| (x, f(x)) - (x', y') \| \leq \rho \). 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,

\[
0 \leq g(x) - f(x) = g(x) - g(x') + g(x') - f(x) \\
\leq L \| x - x' \| + |g(x') - f(x)| \\
\leq \sqrt{L^2 + 1} \sqrt{\| x - x' \|^2 + |g(x') - f(x)|^2} \\
= \sqrt{L^2 + 1} \| (x, f(x)) - (x', y') \| \leq (\sqrt{L^2 + 1})\rho,
\]

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

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

Thus from (1), we deduce the existence of two positive constants \( c \) and \( \epsilon_0 \), depending only on \( d \), such that

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

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

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

if \( \epsilon \leq \epsilon_0 \sqrt{(d + B^2)(1 + L^2(b-a)^2)} \). 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,

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

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

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

13
we deduce finally that, for $\epsilon \leq \epsilon_0 \left( B + L(b - a)\sqrt{d} \right)$,

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

The $\sqrt{d}$ term above can be absorbed in the constants $c$ and $\epsilon_0$.