STAT 583 Final Project
On L_p-Metric Entropy of Convex Functions and Its Application to Convex Regression

Ming-Chun Wu
chunwu@uw.edu
Department of Statistics
University of Washington

Abstract

The Covering number/metric entropy is an idea to quantify the complexity of a statistical model and thus to provide general framework for theoretical analysis. To understand the entropy method, we consider an end-to-end analysis of convex regression as a motivated example. First, we follow Guntuboyina and Sen's work [4] to develop the upper bound of L_p-covering number of uniformly bounded convex functions defined on d-dimensional compact rectangles. Then, combined with Fano's method [12], Guntuboyina and Sen's work leads to an order $n^{-\frac{4}{4+d}}$ minimax lower bound with L_2-risk in a convex regression. The $n^{-\frac{4}{4+d}}$ rate is consistent with recent developments in more general settings [7].

1 Introduction

Metric entropy is important to analysis modern statistical procedures, to name a few, to derive oracle inequalities of empirical risk [8] or to provide minimax risk bound for nonparametric regressions [12]. The performance of such a statistical method usually depends on the complexity, or size, of the underlying statistical model. Metric entropy provides a clever way to turn the abstract concept of complexity into quantifiable theoretical guarantee.

Definition 1 (Covering number)**.** *Given a metric space $(\mathcal{F}, d)$, a ϵ-covering is a subset $A \subseteq \mathcal{F}$ such that for all $f \in \mathcal{F}$ there exists $\tilde{f} \in A$ such that $d(f, \tilde{f}) \leq \epsilon$. The covering number $N(\mathcal{F}, \epsilon, d)$ is defined as the smallest cardinality of all ϵ-coverings of $\mathcal{F}$ under metric d. And the metric entropy is defined as $\log N(\mathcal{F}, \epsilon, d)$.*

Definition 2 (Packing number)**.** *Given a metric space $(\mathcal{F}, d)$, a ϵ-packing is a subset $A \subseteq \mathcal{F}$ such that $d(f_1, f_2) > \epsilon$ for all $f_1, f_2 \in A, f_1 \neq f_2$. And the packing number $M(\mathcal{F}, \epsilon, d)$ is defined as the largest cardinality of all ϵ-packings of $\mathcal{F}$ under metric d.*

We usually cannot find the exact ϵ-covering/packing number unless we consider trivial scenarios. However, knowing the order of covering number with respect to ϵ is sufficient to do analysis in many cases. It is not difficult to see

$$M(\mathcal{F}, 2\epsilon, d) \leq N(\mathcal{F}, \epsilon, d) \leq M(\mathcal{F}, \epsilon, d) \tag{1}$$

, which says that covering and packing numbers are in the same order of ϵ. So, we focus ourselves on bounding covering number in this report.

As a good example, we review Guntuboyina and Sen's main contribution [4] on the study of covering number of uniformly bounded convex functions defined on d-dimensional compact rectangles. Specifically, they show that, for $L_p, 1 \leq p < \infty$, the metric entropy is of the order $\epsilon^{-d/2}$. We rigorously prove the upper bound in Section 2 by following their work.

Their work has some direct applications in statistics, and one of the most related is nonparametric convex regression where the regression function is assumed to be convex. In section 3, we show how to apply entropy method on convex regression. First, we review Fano's method as a general framework to determine minimax risk bound. Then, combine Guntuboyina and Sen's work and Fano's method, we get the $n^{-\frac{4}{4+d}}$ rate of L_2-minimax risk lower bound.

2 Covering Numbers of Convex Functions

Guntuboyina and Sen show that both the upper and lower bounds of ϵ-covering number are of order $\epsilon^{-2/d}$. We present a detailed proof of upper bound in this section. To upper bound the covering number of uniformly bounded

convex function defined on closed rectangles, we need a series of constructions starting from a more approachable function class. We first study the functions with additional Lipschitz constraints, then relax the constraints to get our desired result.

Definition 3 (Notations)**.** *We denote $\mathcal{C}(I,B)$ as the set of convex real-valued functions with domain $I \subset \mathbb{R}^d$ and uniformly bounded by B. We use $\mathcal{C}(I,B,\Gamma)$ to denote all the Γ-Lipschitz continuous functions in $\mathcal{C}(I,B)$. Also, $\mathcal{C}(I,B,\Gamma_{1:d})$ denote the subset of $\mathcal{C}(I,B)$ such that its member is Γ_i-Lipschitz with respect to the ith coordinate. That is, for all $f \in \mathcal{C}(I,B,\Gamma_{1:d})$ we have $|f(x_i,x_{-i}) - f(\tilde{x}_i,x_{-i})| \le \Gamma_i |x_i - \tilde{x}_i|, \forall i = 1,\dots d$.*

In this report, we focus on the cases when I is compact rectangle, $\prod_i [a_i,b_i] \subset \mathbb{R}^d$ and consider the L_p-covering number. Some useful relations are presented below.

Lemma 1 (Set of convex functions)**.**

1. $\mathcal{C}\left(\prod_{i=1}^d [a_i,b_i], B, \Gamma_{1:d}\right) \subseteq \mathcal{C}\left(\prod_{i=1}^d [a_i,b_i], B, \sqrt{\sum_{i=1}^d \Gamma_i^2}\right)$
2. *Scaling: Let* $\epsilon' = \epsilon \frac{[\prod_{i=1}^d (b_i-a_i)]^{-1/p}}{B}$
   (a) $N\left(\mathcal{C}(\prod_{i=1}^d [a_i,b_i], B, \Gamma_{1:d}), \epsilon, \|\cdot\|_p\right) = N\left(\mathcal{C}([0,1]^d, 1, \Gamma_1(b_1-a_1), \dots, \Gamma_1(b_d-a_d)), \epsilon', \|\cdot\|_p\right)$
   (b) $N\left(\mathcal{C}(\prod_{i=1}^d [a_i,b_i], B,), \epsilon, \|\cdot\|_p\right) = N\left(\mathcal{C}([0,1]^d, 1), \epsilon', \|\cdot\|_p\right)$
   (c) $N\left(\mathcal{C}\left(\prod_{i=1}^d [a_i,b_i], B, \Gamma_{1:d}\right), \epsilon, \|\cdot\|_\infty\right) = N\left(\mathcal{C}\left([0,1]^d, B, \Gamma_1(b_1-a_1), \dots, \Gamma_d(b_d-a_d)\right), \epsilon, \|\cdot\|_\infty\right)$
   (d) $\forall r > 0, N\left(\mathcal{C}\left(\prod_{i=1}^d [a_i,b_i], B, \Gamma_{1:d}\right), \epsilon, \|\cdot\|_\infty\right) = N\left(\mathcal{C}\left(\prod_{i=1}^d [a_i,b_i], B/r, \Gamma_1/r, \dots, \Gamma_d/r\right), \epsilon/r, \|\cdot\|_\infty\right)$

Proof. 1. For all $f \in \mathcal{C}(\prod_{i=1}^d [a_i,b_i], B, \Gamma_1, \dots, \Gamma_d)$

$$|f(x) - f(\tilde{x})| \le |f(x_{1:d}) - f(\tilde{x}_1, x_{2:d})| + |f(\tilde{x}_1, x_{2:d}) - f(\tilde{x}_{1:2}, x_{3:d})| + \dots |f(\tilde{x}_{1:d-1}, x_d) - f(\tilde{x}_{1:d})| \quad (2)$$

$$\le \sum_i^d \Gamma_i |x_i - \tilde{x}_i| \le \sqrt{\sum_{i=1}^d \Gamma_i^2}\, \|x - \tilde{x}\|_2 \quad \text{Cauchy-Schwarz} \quad (3)$$

2. (a) For every $f \in \mathcal{C}(\prod_{i=1}^d [a_i,b_i], B, \Gamma_i)$, define $\tilde{f}(x_1,\dots,x_d) = \frac{1}{B} f(a_1 + (b_1,a_1)x_1, \dots, a_d + (b_d - a_d)x_d)$. Since $\left|\tilde{f}(x) - \tilde{f}(x_i', x_{-i})\right| \le \Gamma_i |(b_i - a_i)(x_i - \tilde{x}_i)|$, $\tilde{f} \in \mathcal{C}([0,1]^d, 1, \Gamma_i(b_i - a_i))$. For any $f, g \in \mathcal{C}(\prod_{i=1}^d [a_i,b_i], B, \Gamma_i)$ and their counterpart $\tilde{f}, \tilde{g} \in \mathcal{C}([0,1]^d, 1, \Gamma_i(b_i - a_i))$, by change of variable we have

$$\left\|\tilde{f} - \tilde{g}\right\|_p = \frac{(\prod_{i=1}^d (b_i - a_i))^{-1/p}}{B} \|f - g\|_p \quad (4)$$

Hence, if we find a ϵ'-covering $\left\{\tilde{f}_i\right\}$ of $\mathcal{C}([0,1]^d, 1, \Gamma_i(b_i - a_i))$, the counterparts $\{f_i\}$ is a ϵ-covering of $\mathcal{C}(\prod_{i=1}^d [a_i,b_i], B, \Gamma_i)$. Then, we finish the proof.

(b) Letting $\Gamma_i = \infty$ in (a).

(c) Similar to part (a), for every $f \in \mathcal{C}\left(\prod_{i=1}^d [a_i,b_i], B, \Gamma_{1:d}\right)$, define

$$\tilde{f}(x_1,\dots,x_d) = f(a_1 + (b_1,a_1)x_1, \dots, a_d + (b_d - a_d)x_d) \quad (5)$$

and the remaining is obvious.

(d) It is obvious if we define $\tilde{f} = f/r$ for all $f \in \mathcal{C}\left(\prod_{i=1}^d [a_i,b_i], B, \Gamma_{1:d}\right)$.

□

2.1 Covering number of $\mathcal{C}(I, B, \Gamma_{1:d})$

The story starts from the extension of Bronshtein's theorem on the covering number of convex sets.

Definition 4 (Hausdorff Metric)**.** *Given a metric space $(\mathcal{F}, d)$, for all $U, V \subseteq \mathcal{F}$, the Hausdorff distance is defined as*

$$d_H(U, V) := \max\left\{\sup_{a\in U}\inf_{b\in V} d(a, b), \sup_{b\in V}\inf_{a\in U} d(a, b)\right\} \tag{6}$$

Theorem 2 (Bronshtein [3])**.** *Let $\mathcal{K}^{d+1}(r)$ be the set of all the nonempty closed convex subsets of the open ball with radius r and centered at origin, there exist positive constants c, ϵ_0 depending only on d such that*

$$\log N(\mathcal{K}^{d+1}(r), \epsilon, d_H) \le c(r/\epsilon)^{d/2}, \forall \epsilon < r\epsilon_0 \tag{7}$$

Note that Bronshtein's result has the desired order $\epsilon^{-d/2}$, so we should try to use it. However, at the first glance, one may wonder how could we use it for convex functions. First of all, our target is convex functions, and Bronshtein's theorem is on convex sets. Secondly, it is the Hausdorff metric that are used in Bronshtein's theorem instead of the L_p-norm we are interested in. Therefore, we need to address these two obstacles to use (7).

There are two key ideas to overcome the challenges. The first is to observe that the size of a function class is highly related to the size of its graphs. An example is the universality theorem in empirical process theory [6], which says that we can upper bound the packing number of a function class using the VC dimension of the corresponding VC-subgraph class. This suggests us work on some types of graphs of the convex functions. Since it is well known that the epigraph of a convex function is a convex set, working on epigraph allows us to apply Bronshtein's theorem.

Definition 5 (Epigraph)**.** *Given function that is bounded above $f \le B$, its epigraph is the set defined as*

$$V_f := \{(x, t) : f(x) \le t \le B\} \tag{8}$$

The second idea bridges the gap between Hausdorff metric and L_∞ norm through the lemma below.

Lemma 3. *For any $f, g \in \mathcal{C}([0,1]^d, B, \Gamma)$, we have*

$$\|f - g\|_\infty \le d_H(V_f, V_g)\sqrt{1 + \Gamma^2} \tag{9}$$

, which implies

$$N\left(\mathcal{C}([0,1]^d, B, \Gamma_{1:d}), \epsilon, \|\cdot\|_\infty\right) \le N\left(\left\{V_f, f \in \mathcal{C}([0,1]^d, B, \Gamma_{1:d})\right\}, \frac{\epsilon}{\sqrt{1 + \sum_{i=1}^d \Gamma_i^2}}, d_H\right) \tag{10}$$

Proof. The inequality (9) is useless unless we assume that $\Gamma < \infty$ for all i. The strategy is to show that for all $x \in [0,1]^d$, $|f(x) - g(x)| \le d_H(V_f, V_g)\sqrt{1 + \Gamma^2}$ for all $f, g \in \mathcal{C}([0,1]^d, B, \Gamma_{1:d})$. Now, fix a x, because the inequality holds if $f(x) = g(x)$, WLOG, we assume $f(x) < g(x)$.

Since the domain of f, g is a subset in $\mathbb{R}^d$, their epigraphs lie in $\mathbb{R}^{d+1}$. We consider the metric space $(\mathbb{R}^{d+1}, \|\cdot\|_2)$ in Hausdorff metric. By the definition of Hausdorff metric, we have

$$\sup_{a\in V_f}\inf_{b\in V_g} \|a - b\|_2 \le d_H(V_f, V_g), \quad \inf_{b\in V_g} \|a - b\|_2 \le d_H(V_f, V_g), \forall a \in V_f \tag{11}$$

Let $a = (x, f(x)) \in V_f$, there must exist $b := (x', t') \in V_g$ such that $\|(x, f(x)) - (x', t')\|_2 \le d_H(V_f, V_g)$. Since $(x', t') \in V_g$ implies $g(x') \le t'$, we have

$$\|(x, f(x)) - (x', g(x'))\|_2 \le \|(x, f(x)) - (x', t')\|_2 \le d_H(V_f, V_g). \tag{12}$$

So, just pick $b = (x', g(x'))$. Then,

$$\begin{aligned}
|g(x) - f(x)| &\le |g(x) - g(x')| + |g(x') - f(x)| && && (13)\\
&\le \Gamma \|x - x'\|_2 + |g(x') - f(x)| && \text{Lipschitz} && (14)\\
&= (\Gamma, 1) \cdot (\|x - x'\|_2, |g(x) - g(x'|) && && (15)\\
&\le \sqrt{1 + \Gamma^2}\, \|(x, f(x)) - (x', g(x'))\|_2 && \text{Cauchy-Schwarz} && (16)\\
&\le d_H(V_g, V_f)\sqrt{1 + \Gamma^2} && (12) && (17)
\end{aligned}$$

Then, apply lemma 1.1 we then have (10). □

Now, we are ready to upper bound the L_∞-covering number of $\mathcal{C}(\prod_i[a_i,b_i],B,\Gamma_{1:d})$.

Theorem 4 (Theorem 3.2 in Guntuboyina and Sen[4])**.** *There exist positive constants c,ϵ_0 depending only on d such that*

$$\log N\left(\mathcal{C}\left(\prod_{i=1}^d[a_i,b_i],B,\Gamma_{1:d}\right),\epsilon,\|\cdot\|_\infty\right)\le c\left(\frac{B+\sum_i\Gamma_i(b_i-a_i)}{\epsilon}\right)^{d/2} \tag{18}$$

for all $\epsilon<\epsilon_0\left[B+\sum_i\Gamma_i(b_a-a_i)\right]$.

Proof. For any $r>0$, let $\tilde\Gamma_i=\Gamma_i(b_i-a_i)/r$, then

$$\log N\left(\mathcal{C}\left(\prod_i[a_i,b_i],B,\Gamma_{1:d}\right),\epsilon,\|\cdot\|_\infty\right) \tag{19}$$

$$=\log N\left(\mathcal{C}\left([0,1]^d,B,\tilde\Gamma_{1:d}\right),\epsilon,\|\cdot\|_\infty\right) \qquad \text{lemma 1.1.(c)} \tag{20}$$

$$=\log N\left(\mathcal{C}\left([0,1]^d,B/r,\tilde\Gamma_{1:d}\right),\epsilon/r,\|\cdot\|_\infty\right) \qquad \text{lemma 1.1.(d)} \tag{21}$$

$$\le\log N\left(\left\{V_f,f\in\mathcal{C}\left([0,1]^d,B/r,\tilde\Gamma_{1:d}\right)\right\},\frac{\epsilon/r}{\sqrt{1+\sum_{i=1}^d\tilde\Gamma_i^2}},d_H\right) \qquad (10) \tag{22}$$

$$\le\log N\left(\mathcal{K}^{d+1}(\sqrt{d+(B/r)^2}),\frac{\epsilon/r}{\sqrt{1+\sum_{i=1}^d\tilde\Gamma_i^2}},d_H\right) \qquad \{V_f\}\subset\mathcal{K}^{d+1}\left(\sqrt{d+(\frac{B}{r})^2}\right) \tag{23}$$

$$\le c\left(\frac{\sqrt{(r^2d+B^2)(1+\sum_i(b_i-a_i)^2(\Gamma_i/r)^2)}}{\epsilon}\right)^{d/2} \qquad (7) \tag{24}$$

$$=c\left(\frac{B+\sqrt{d\sum_i\Gamma_i^2(b_i-a_i)^2}}{\epsilon}\right)^{d/2} \qquad r^4=\frac{B^2\sum_i\Gamma_i^2(b_i-a_i)^2}{d} \tag{25}$$

$$\le c\left(\frac{\sqrt{d}B+\sqrt{d\sum_i\Gamma_i^2(b_i-a_i)^2}}{\epsilon}\right)^{d/2} \tag{26}$$

$$\le c\left(\sqrt{d}\frac{B+\sum_i\Gamma_i(b_i-a_i)}{\epsilon}\right)^{d/2}=c'\left(\frac{B+\sum_i\Gamma_i(b_i-a_i)}{\epsilon}\right)^{d/2} \tag{27}$$

for all $\epsilon\le\epsilon_0\left(B+\sum_i\Gamma_i(b_i-a_i)\right)$ where c',ϵ_0 are constants depending only on d. The last inequality is because $\Gamma_i(b_i-a_i)>0$, we have $\sum_i\Gamma_i^2(b_i-a_i)^2\le\left(\sum_i\Gamma_i(b_i-a_i)\right)^2$. □

Corollary 5. *There exist positive constant c,ϵ_0 depending only on d,p such that for all $\epsilon<\epsilon_0\frac{B+\sum_i\Gamma_i(b_a-a_i)}{[\prod_i(b_i-a_i)]^{1/p}}$,*

$$\log N\left(\mathcal{C}\left(\prod_{i=1}^d[a_i,b_i],B,\Gamma_{1:d}\right),\epsilon,\|\cdot\|_p\right)\le c\left(\frac{B+\sum_i\Gamma_i(b_a-a_i)}{\epsilon[\prod_i(b_i-a_i)]^{1/p}}\right)^{d/2} \tag{28}$$

Proof. Note that we implicitly use the Lebesque measure μ for the L_p norm and $|f-g|\le\|f-g\|_\infty\le 2B<\infty$, we then have $\|f-g\|_p\le\|f-g\|_\infty[\mu(\prod_i[a_i,b_i])]^{1/p}=\|f-g\|_\infty[\prod_i(b_i-a_i)]^{1/p}$. So,

$$N\left(\mathcal{C}\left(\prod_{i=1}^d[a_i,b_i],B,\Gamma_{1:d}\right),\epsilon,\|\cdot\|_p\right)\le N\left(\mathcal{C}\left(\prod_{i=1}^d[a_i,b_i],B,\Gamma_{1:d}\right),\epsilon[\prod_i(b_i-a_i)]^{1/p},\|\cdot\|_\infty\right) \tag{29}$$

, which finishes the proof. □

2.2 Covering number of $\mathcal{C}(I,B)$

To get the L_p-covering number of $\mathcal{C}(I,B)$, we need to relax the Lipschitz constants in $\mathcal{C}(I,B,\Gamma_{1:d})$. Specifically, we want to let $\Gamma_i=\infty$. The strategy is to apply the mathematical induction to relax one Lipschitz constant Γ_i at a time. The challenge is how to use the result on Lipschitz functions when the Lipschitz assumption is no longer valid. Hence, here come two two important ideas. First, use the fact that convex function is locally Lipschitz.

Lemma 6 (Locally Lipschitz). *Suppose f is a convex function on (a,b), then it is Lipschitz on $[c,d] \subset (a,b)$ with Lipschitz constant at most* $\max\left\{\left|\frac{f(c)-f(a)}{c-a}\right|, \left|\frac{f(b)-f(d)}{b-d}\right|\right\}$.

Proof. Let $a < t_1 \le c \le x \le y \le d \le t_2 < b$, by the definition of convex function we have

$$\frac{f(c)-f(t_1)}{c-t_1} \le \frac{f(y)-f(x)}{y-x} \le \frac{f(t_2)-f(d)}{t_2-d} \tag{30}$$

Take $t_1 \to a, t_2 \to b$ and absolute value, we get $\left|\frac{f(y)-f(x)}{y-x}\right| \le \max\left\{\left|\frac{f(c)-f(a)}{c-a}\right|, \left|\frac{f(b)-f(d)}{b-d}\right|\right\}, \forall x, y \in [c,d]$. □

The second idea is the partition argument. Roughly speaking, we partition the domain to get several locally Lipschitz function classes. We know how to find the covering number of these locally Lipschitz convex function classes. Then, we can argue that the local covering numbers can be used to upper bound the global covering number. The detail is summarized in the following lemma.

Lemma 7 (Partition). *Given a set of functions $\mathcal{F}$ with domain $I = \cup_{m=1}^{M} I_m$, WLOG, assume I_m's are disjoint. Define the restriction of $\mathcal{F}$ on I_m as $\mathcal{F}_{I_m} := \{f\mathbb{1}_{I_m} : f \in \mathcal{F}\}$, then*

$$N(\mathcal{F}, \epsilon, \|\cdot\|_p) \le \prod_m N(\mathcal{F}_{I_m}, \epsilon_m, \|\cdot\|_p) \tag{31}$$

where $\epsilon = 2(\sum_m \epsilon_m^p)^{1/p}$.

Proof. Let V_m be a ϵ_m-covering of $\mathcal{F}_{I_m} = \{f\mathbb{1}_{I_m}\}$. Then, for any $f \in \mathcal{F}$ and $m = 1, \dots, M$, there is a $f_m \in V_m$ such that $\|(f-f_m)\mathbb{1}_{I_m}\|_p \le \epsilon_m$. For a sequence $(f_m)_1^M, f_m \in V_m$, if possible, pick a $\tilde{f} \in \mathcal{F}$ such that $\left\|(\tilde{f}-f_m)\mathbb{1}_{I_m}\right\|_p \le \epsilon_m, \forall m$, we then collect such $\tilde{f}$ to form a covering $V = \left\{\tilde{f}\right\}$. Since there are at most $\prod |V_m|$ different sequences, the cardinality of V is at most $\prod |V_m|$. For any $f \in \mathcal{F}$, there must be a sequence $(f_m)_1^M, f_m \in V_m$ such that $\|(f-f_m)\mathbb{1}_{I_m}\|_p \le \epsilon_m, \forall m$. Because of the construction of V, there is a $\tilde{f} \in V$ such that

$$\left\|f-\tilde{f}\right\|_p \le \left\|f - \sum_m f_m \mathbb{1}_{I_m}\right\|_p + \left\|\tilde{f} - \sum_m f_m \mathbb{1}_{I_m}\right\|_p \tag{32}$$

$$= \left(\sum_m \|(f-f_m)\mathbb{1}_{I_m}\|_p^p\right)^{1/p} + \left(\sum_m \left\|(\tilde{f}-f_m)\mathbb{1}_{I_m}\right\|_p^p\right)^{1/p} \le 2(\sum_m \epsilon_m^p)^{1/p} := \epsilon \tag{33}$$

Therefore, we have construct a ϵ-covering V such that $|V| \le \prod_m |V_m|$, which finishes the proof. □

We are ready to remove the Lipschitz constants.

Theorem 8. *For $1 \le p < \infty$, there exist positive constant ϵ_0, c depending only on d, p, such that*

$$\log N\left(\mathcal{C}\left([0,1]^d, 1, \Gamma_{1:d}\right), \epsilon, \|\cdot\|_p\right) \le c\left(\frac{2+\sum_i \Gamma_{i=1}^d \mathbb{1}(\Gamma_i < \infty)}{\epsilon}\right)^{d/2} \tag{34}$$

for all $\epsilon \le \epsilon_0$.

Proof. We use induction to prove the theorem. Let k be the number of $\Gamma_i = \infty$, we want to show that (34) holds for $k = d$. Corollary 5 implies that (34) holds when $k = 0$. Suppose (34) holds for $k-1$, WLOG, let $\Gamma_2, \dots, \Gamma_k = \infty$, there exist c, ϵ_0 depending only on d, p such that

$$\log N\left(\mathcal{C}\left([0,1]^d, 1, \Gamma_1, \infty, \dots, \infty, \Gamma_{k+1:d}\right), \epsilon, \|\cdot\|_p\right) \le c\left(\frac{2+\Gamma_1+\sum_{i=k+1}^d \Gamma_i}{\epsilon}\right)^{d/2}, \epsilon \le \epsilon_0. \tag{35}$$

Then, we want to show that it is also true for k, say, we want to extend to $\Gamma_1 = \infty$. Given a $u < 1/2$, partition the first coordinate into three intervals $[0,1] = [0,u] \cup [u, 1-u] \cup [1-u, 1]$. We want to find coverings of subintervals and then combine them using partition argument to proceed from $k-1$ to k.

Part 1: We aim to find a covering of the restrictions on $[u, 1-u]$. Using lemma 6 with $a = 0, c = u, d = 1-u, b = 1$ and $|f| < 1$, the Lipschitz constant is at most $2/u$, that is,

$$\mathcal{F}_{[u,1-u]} := \left\{ f \mathbb{1}_{[u,1-u]}(x_1) : f \in \mathcal{C}\left([0,1]^d, 1, \infty, \ldots, \Gamma_{k+1:d}\right) \right\} \tag{36}$$

$$\subseteq \mathcal{C}\left([u, 1-u] \times [0,1]^{d-1}, 1, \frac{2}{u}, \infty, \ldots, \Gamma_{k+1:d}\right) \tag{37}$$

Note that $u < \frac{1}{2}$, we have

$$N(\mathcal{F}_{[u,1-u]}, \eta, \|\cdot\|_p) \tag{38}$$

$$\leq N\left(\mathcal{C}\left([u, 1-u] \times [0,1]^{d-1}, 1, \frac{2}{u}, \infty, \ldots, \Gamma_{k+1:d}\right), \eta, \|\cdot\|_p\right) \tag{39}$$

$$\leq N\left(\mathcal{C}\left([u, 1-u] \times [0,1]^{d-1}, 1, \frac{2}{u}, \infty, \ldots, \Gamma_{k+1:d}\right), \eta(1-2u)^{1/p}, \|\cdot\|_p\right) \qquad \eta(1-2u) < \eta \tag{40}$$

$$= N\left(\mathcal{C}\left([0,1]^d, 1, \frac{2(1-2u)}{u}, \infty, \ldots, \Gamma_{k+1:d}\right), \eta, \|\cdot\|_p\right) \qquad \text{lemma 1.1.(b)} \tag{41}$$

$$\leq N\left(\mathcal{C}\left([0,1]^d, 1, \frac{2}{u}, \infty, \ldots, \Gamma_{k+1:d}\right), \eta, \|\cdot\|_p\right) \qquad \frac{2(1-2u)}{u} \leq \frac{2}{u} \tag{42}$$

$$\leq \exp\left(c\left(\frac{2 + 2/u + \sum_{i>k}\Gamma_i}{\eta}\right)^{d/2}\right), \eta \leq \epsilon_0 \qquad (35) \tag{43}$$

$$\leq \exp\left(c(2/u)^{d/2}\left(\frac{2 + \sum_{i>k}\Gamma_i}{\eta}\right)^{d/2}\right), \eta \leq \epsilon_0 \qquad u < 1/2 \tag{44}$$

Part 2: For $[0, u]$, suppose we have a finer partition, $0 < \delta_1 < \delta_2, \ldots, \delta_A < u < \delta_{A+1}$, which will be be determined later. For each interval $[\delta_m, \delta_{m+1}], m = 1, \ldots, A-1$, since $\delta_m < \delta_{m+1} < u \leq 1/2$, we have $\frac{2}{\delta_m} > \frac{2}{1-\delta_{m+1}}$. By lemma 6, the restrictions on $[\delta_m, \delta_{m+1}]$ is $2/\delta$-Lipschitz, so give $0 < a_m < \epsilon_0$

$$N(\mathcal{F}_{[\delta_m,\delta_{m+1}]}, a_m(\delta_{m+1} - \delta_m)^{1/p}, \|\cdot\|_p) \tag{45}$$

$$\leq N\left(\mathcal{C}\left([\delta_m, \delta_{m+1}] \times [0,1]^{d-1}, 1, \frac{2}{\delta_m}, \infty, \ldots, \Gamma_{k+1:d}\right), a_m(\delta_{m+1} - \delta_m)^{1/p}, \|\cdot\|_p\right) \tag{46}$$

$$= N\left(\mathcal{C}\left([0,1]^d, 1, \frac{2(\delta_{m+1} - \delta_m)}{\delta_m}, \infty, \ldots, \Gamma_{k+1:d}\right), a_m, \|\cdot\|_p\right) \qquad \text{lemma 1.1.(b)} \tag{47}$$

$$\leq \exp\left(c\left(\frac{2 + \frac{2(\delta_{m+1}-\delta_m)}{\delta_m} + \sum_{i>k}\Gamma_i}{a_m}\right)^{d/2}\right), a_m \leq \epsilon_0 \qquad (35) \tag{48}$$

$$\leq \exp\left(c\left(\frac{\delta_{m+1}}{a_m\delta_m}\right)^{d/2}\left(2 + \sum_{i>k}\Gamma_i\right)^{d/2}\right), a_m \leq \epsilon_0 \qquad \frac{\delta_{m+1}}{\delta_m} > 1 \tag{49}$$

We need to be careful about the intervals $[0, \delta_1]$ and $[\delta_A, u]$. For $[\delta_A, u]$, since $\delta_A < u < 1/2$, the Lipschitz constant in lemma 6 is at most $2/\delta_A$. Then

$$N(\mathcal{F}_{[\delta_A,u]}, a_A(u - \delta_A)^{1/p}, \|\cdot\|_p) \tag{50}$$

$$\leq N\left(\mathcal{C}\left([\delta_A, u] \times [0,1]^{d-1}, 1, \frac{2}{\delta_A}, \infty, \ldots, \Gamma_{k+1:d}\right), a_A(u - \delta_A)^{1/p}, \|\cdot\|_p\right) \tag{51}$$

$$= N\left(\mathcal{C}\left([0,1]^d, 1, \frac{2(u - \delta_A)}{\delta_A}, \infty, \ldots, \Gamma_{k+1:d}\right), a_A, \|\cdot\|_p\right) \qquad \text{lemma 1.1.(b)} \tag{52}$$

$$\leq N\left(\mathcal{C}\left([0,1]^d, 1, \frac{2(\delta_{A+1} - \delta_A)}{\delta_A}, \infty, \ldots, \Gamma_{k+1:d}\right), a_A, \|\cdot\|_p\right) \qquad \frac{2(\delta_{A+1} - \delta_A)}{\delta_A} > \frac{2(u - \delta_A)}{\delta_A} \tag{53}$$

$$\leq \exp\left(c\left(\frac{\delta_{A+1}}{a_A\delta_A}\right)^{d/2}\left(2 + \sum_{i>k}\Gamma_i\right)^{d/2}\right), a_A \leq \epsilon_0 \qquad m = A, (49) \tag{54}$$

For $[0, \delta_1]$, it is not difficult to see that zero function is a $\delta_1^{1/p}$-covering with cardinality 1.
Part 3: For $[1-u, 1]$, by symmetry, we can get the same result as in part 2.

Now, apply lemma 6 to combine the coverings of all the subintervals in part 1 to part 3, we have a s-covering in L_p-norm with cardinality N if $\eta, a_m < \epsilon_0$ where

$$s = 2\left(2\left(\delta_1 + \sum_{m=1}^{A} a_m^p(\delta_{m+1} - \delta_m)\right) + \eta^p\right)^{1/p} \tag{55}$$

$$\log N = c\left[\left(\frac{2}{u}\right)^{d/2} + 2\sum_{m=1}^{A}\left(\frac{\eta\delta_{m+1}}{\delta_m a_m}\right)^{d/2}\right]\left(\frac{2+\sum_{i=k+1}^{d}\Gamma_i}{\eta}\right)^{d/2} \tag{56}$$

The remaining is to choose the parameters u, δ_m, a_m, η to get the desired result. Let

$$\eta \le \epsilon_0 \tag{57}$$

$$u = \exp\left(-2(p+1)^2(p+2)\log 2\right) \le \frac{1}{2} \tag{58}$$

$$\delta_m = \exp\left(p\left(\frac{p+1}{p+2}\right)^{m-1}\log\eta\right) \tag{59}$$

$$a_m = \eta\exp\left(-p\frac{(p+1)^{m-2}}{(p+2)^{m-1}}\log\eta\right) \le \eta \tag{60}$$

and after some tedious algebra, see [4] for details, we can show

$$\left(\delta_1 + \sum_{m=1}^{A} a_m^p(\delta_{m+1} - \delta_m)\right) \le \frac{7}{3}\eta^p, \quad \sum_{m=1}^{A}\left(\frac{\eta\delta_{m+1}}{\delta_m a_m}\right)^{d/2} \le \frac{2^d}{2^d - 1} := c_d \tag{61}$$

Hence,

$$s \le 2\left(\frac{17}{3}\right)^{1/p}\eta, \quad \log N \le c(2c_d + (2/u)^{d/2})\left(\frac{2+\sum_{i>k}\Gamma_i}{\eta}\right)^{d/2} \tag{62}$$

Let $\epsilon := 2\left(\frac{17}{3}\right)^{1/p}\eta$, we can find a ϵ-covering in L_p-norm with cardinality at most

$$c\left(2\left(\frac{17}{3}\right)^{1/p}\right)^{2/d}\left(2c_d + \left(\frac{2}{u}\right)^{d/2}\right)\left(\frac{2+\sum_{i>k}\Gamma_i}{\eta}\right)^{d/2}, \forall\epsilon < 2\left(\frac{17}{3}\right)^{1/p}\epsilon_0 \tag{63}$$

Since u only depends on p and c_d only depends on d, we show that (34) holds for k and finish the proof by induction. □

Corollary 9 (Theorem 3.1 in Guntuboyina and Sen [4])**.** *For $1 \le p < \infty$, there exist positive constant ϵ_0, c depending only on d, p, such that*

$$\log N\left(\mathcal{C}\left(\prod_i [a_i, b_i], B\right), \epsilon, \|\cdot\|_p\right) \le c\left(\frac{B(\prod_i [a_i, b_i])^{1/p}}{\epsilon}\right)^{d/2} \tag{64}$$

for all $\epsilon \le \epsilon_0 B(\prod_i [a_i, b_i])^{1/p}$.

Proof. Apply scaling property, lemma 1.1.(b), it is equivalent to show that $\log N(\mathcal{C}([0,1]^d, 1), \epsilon, \|\cdot\|_p) \le c(1/\epsilon)^{d/2}$. Apply theorem 8 with $\Gamma_i = \infty$ to finish the proof. □

2.3 Summary of the proof

After deriving the final result, Corollary 9, we brief summarize the key steps of the entire proof.

1. Start from Bronshtein's theorem, which shows the order $\epsilon^{-d/2}$ upper bound of covering number of convex sets in Hausdorff metric.

2. Because of convexity of the epigraph of convex function, we can move from convex set to convex function.
3. With Lipschitz condition, we can move from Hausdorff metric to L_∞-norm by lemma 3 to get Theorem 4 (Theorem 3.2 in Guntuboyina and Sen[4]).
4. To move from L_∞ to L_p, it requires uniform boundedness and compact domain (proof of Corollary 5).
5. Use the locally Lipschitz property of convex function and partition argument (lemma 7) to relax the Lipschitz constants and get the final result Corollary 9 (Theorem 3.1 in Guntuboyina and Sen[4]).

3 Minimax Lower Bound of Convex Regression

An important application of metric entropy is to derive minimax risk bound of nonparametric regression. A general framework is via the celebrated Fano's inequality in information theory [2, 11]. To apply Fano's inequality, we need to reduce a estimation problem to testing multiple hypotheses and this is where entropy method comes in. Then, Fano's inequality provides the lower bound of the testing error and leads to the minimax lower bound of the estimation problem. In this section, we first review Fano's approach on determining minimax lower bound. Then, combined with the result of covering number derived in last section, we have $n^{-\frac{4}{4+d}}$ rate of L_2-minimax lower bound of an example of convex regression.

We specifically consider the following nonparametric estimation problem. Given a model $\{p(\cdot|f), f \in \mathcal{F}\}$ indexed by a function class $\mathcal{F}$, observations $x_i|f \overset{i.i.d.}{\sim} p(\cdot|f)$ and estimator $\phi : x_1^n \to \hat{f} \in \mathcal{F}$, the minimax risk with respect to a metric d is defined as

$$r_n(\mathcal{F}) := \inf_{\hat{f}\in\mathcal{F}} \sup_{f\in\mathcal{F}} \mathbb{E}_{x_1^n}[d^2(\hat{f}, f)|f] \tag{65}$$

Our main goal here is to derive the lower bound of the minimax risk.

3.1 Minimax Lower Bound via Fano's Method

We first introduce basic concepts that are necessary to derive the lower bound.

Definition 6 (Mutual Information)**.** *Define KL-divergence of two distribution functions P, Q as $D_{KL}(dP||dQ) = \mathbb{E}_P \log \frac{dP}{dQ}$ and the mutual information of two random object X, Y is defined as $I(X;Y) = D_{KL}(p_{x,y}||p_x p_y)$, which is the KL divergence of their joint distribution $P_{x,y}$ and the product of their marginal distributions $P_x P_y$.*

Theorem 10 (Fano's inequality [2])**.** *Suppose a random object X is uniformly distributed over a finite set $\mathcal{X}$. For any Markov chain $X \to Y \to \hat{X}$, we have*

$$\mathbb{P}\left(X \neq \hat{X}\right) \geq 1 - \frac{I(X;Y) + \log 2}{\log|\mathcal{X}|} \tag{66}$$

where $I(X;Y)$ is the mutual information.

The inequality says the lower bound of the testing error in a multiple hypotheses testing is determined by the mutual information of X, Y, where Y is interpreted as the observation. Moreover, we should keep in mind that the lower bound is independent to the choice of the estimator/testing rule $\hat{X}$.

Theorem 11 (Barron and Yang [12])**.** *Let $d_{KL}(f_1, f_2) := D_{KL}(p(x|f_1)||p(x|f_2))$, suppose*

$$\log N(\mathcal{F}, \epsilon, d_{KL}) \approx n\epsilon^2, \quad \log N(\mathcal{F}, 2\epsilon', d) \approx 4n\epsilon^2 + 2\log 2 \tag{67}$$

, we have

$$r_n(\mathcal{F}) := \inf_{\hat{f}\in\mathcal{F}} \sup_{f\in\mathcal{F}} \mathbb{E}_{x_1^n}[d^2(\hat{f}, f)|f] \geq \frac{1}{2}\epsilon'^2 \tag{68}$$

Proof.
Step 1: It suffices to find a uniform lower bound of $\sup_{f\in\mathcal{F}} \mathbb{E}_{x_1^n}[d^2(\hat{f}, f)|f]$ over all $\hat{f}$. When it comes to maximizing

over a uncountable set, metric entropy provides a way to discretize the problem. Let $\mathcal{F}_{2\epsilon'} := \{f_1 \dots, f_M\} \subset \mathcal{F}$ be a $2\epsilon'$-packing set of $\mathcal{F}$ in metric d, we then have

$$\sup_{f\in\mathcal{F}} \mathbb{E}_{x_1^n}[d^2(\hat{f}, f)|f] \geq \sup_{f\in\mathcal{F}_{2\epsilon'}} \mathbb{E}_{x_1^n}[d^2(\hat{f}, f)|f] \geq \sum_{m=1}^{M} \frac{1}{M}\mathbb{E}_{x_1^n}[d^2(\hat{f}, f_m)|f_m] \tag{69}$$

$$\geq \sum_{m=1}^{M} \frac{1}{M}\epsilon^2 \mathbb{P}\left(d^2(\hat{f}, f_m) > \epsilon^2 | f_m\right) \qquad \text{Markov inequality} \tag{70}$$

Let m be uniformly distributed on $\{1, \dots, M\}$ and define $\hat{m}(x_1^n) := \arg\min_i d(\phi(x_1^n), f_i) = \arg\min_i d(\hat{f}, f_i)$ as a valid testing rule, we then have the valid testing problem and the corresponding Markov chain

$$m \overset{p(\cdot|f_m)}{\to} x_1^n \overset{\hat{m}(\phi(x_1^n))}{\to} \hat{m} \tag{71}$$

By the construction of $\mathcal{F}_{2\epsilon'}$, $d(\hat{f}, f_m) > \epsilon$ if and only if $\hat{m}(\hat{f}) \neq m$. We have

$$(70) = \epsilon^2 \mathbb{P}\left(d^2(\hat{f}, f_m) > \epsilon^2\right) \qquad \text{uniformly distributed } m \tag{72}$$

$$\geq \epsilon^2 \left(1 - \frac{I(m, x_1^n) + \log 2}{\log |\mathcal{F}_{2\epsilon'}|}\right) \qquad \text{Fano} \tag{73}$$

The Markov chain (71) holds no matter what estimator ϕ is used, so

$$\inf_{\hat{f}\in\mathcal{F}} \sup_{f\in\mathcal{F}} \mathbb{E}_{x_1^n}[d^2(\hat{f}, f)|f] \geq \epsilon^2 \left(1 - \frac{I(m, x_1^n) + \log 2}{\log |\mathcal{F}_{2\epsilon'}|}\right) \tag{74}$$

Next, we want to remove the dependence on the mutual information, which can be done by finding a upper bound of $I(m, x_1^n)$.

Step 2: Find a upper bound of $I(m; x_1^n)$. By the definition of mutual information, we have

$$I(m; x_1^n) = D_{KL}\left(p(m)p(x_1^n|f_m)||p(m)\sum p(m)p(x_1^n|f_m)\right) \tag{75}$$

$$= \mathbb{E}_m\left[D_{KL}\left(p(x_1^n|f_m)||\sum p(m)p(x_1^n|f_m)\right)\right] \tag{76}$$

$$= \mathbb{E}_m \mathbb{E}_{p(x_1^n|f_m)}\left[\log\frac{p(x_1^n|f_m)}{q} + \log\frac{q}{\sum p(m)p(x_1^n|f_m)}\right], \forall q \tag{77}$$

$$= \mathbb{E}_m D_{KL}\left(p(x_1^n|f_m)||q\right) - \mathbb{E}_{p(x_1^n, m)} \log\frac{\sum p(m)p(x_1^n|f_m)}{q} \tag{78}$$

$$\leq \mathbb{E}_m D_{KL}\left(p(x_1^n|f_m)||q\right) \qquad \text{Jensen} \tag{79}$$

$$\leq \max_m D_{KL}\left(p(x_1^n|f_m)||q\right), \forall q \tag{80}$$

Let V_ϵ be a ϵ-covering of $\mathcal{F}$ with respect to $d_{KL}^2(f_1, f_2) := D_{KL}(p(x|f_1)||p(x|f_2))$. That is, $f \in \mathcal{F}$ there exist a $f_i' \in V_\epsilon$ such that $D_{KL}(p(x|f)||p(x|f_i')) \leq \epsilon^2$. Let $q(x_1^n) = \sum_{f_i'\in V_\epsilon} \frac{1}{|V_\epsilon|} p(x_1^n|f_i')$, then

$$D_{KL}\left(p(x_1^n|f_m)||q\right) = \mathbb{E}_{p(x_1^n|f_m)} \log\frac{p(x_1^n|f_m)}{\sum \frac{1}{|V_\epsilon|} p(x_1^n|f_i')} \tag{81}$$

$$\leq \log|V_\epsilon| + \mathbb{E}_{p(x_1^n|f_m)} \log\frac{p(x_1^n|f_m)}{p(x_1^n|f_i')}, \forall i \qquad \sum p(x_1^n|f_i') \geq p(x_1^n|f_i') \tag{82}$$

$$= \log|V_\epsilon| + \sum_{j=1}^{n} \mathbb{E}_{p(x_j|f_m)} \log\frac{p(x_j|f_m)}{p(x_j|f_i')}, \forall i \qquad x_1^n \text{ are i.i.d.} \tag{83}$$

$$\leq \log|V_\epsilon| + \min_i n D_{KL}(p(x|f_m)||p(x|f_i')) \leq \log|V_\epsilon| + n\epsilon^2 \qquad \epsilon\text{-covering} \tag{84}$$

This is true for all ϵ and any ϵ-covering in d_{KL}, so

$$I(m; x_1^n) \leq \inf_{\epsilon>0} \left\{n\epsilon^2 + \log N(\mathcal{F}, \epsilon, d_{KL})\right\} \tag{85}$$

Step 3: Control the bound (73). Let

$$\log N(\mathcal{F}, \epsilon, d_{KL}) \approx n\epsilon^2 \tag{86}$$

$$\log |\mathcal{F}_{2\epsilon'}| \overset{(1)}{\geq} \log N(\mathcal{F}, 2\epsilon', d) \approx 4n\epsilon^2 + 2\log 2 \tag{87}$$

Finally, under this construction (73) $\geq \frac{1}{2}\epsilon'^2$.

□

3.2 Convex Regression

Consider the regression problem $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \sim N(0,1), x_i \sim U[0,1]^d$ and ϵ_i, x_i are all independent. Suppose $f \in \mathcal{F} = \mathcal{C}([0,1]^d, 1)$, we can use theorem 11 to derive the minimax lower bound under L_2-norm. Note that

$$d_{KL}^2(f_1, f_2) = D_{KL}(p(x,y|f_1)||p(x,y|f_2)) \tag{88}$$

$$= \int\int p(x,y|f_1,x)p(x)\log\frac{p(x,y|f_1,x)p(x)}{p(x,y|f_2,x)p(x)}dxdy \tag{89}$$

$$= \int\int \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(y - f_1(x)\right)^2\right]p(x)\left[-\frac{f_1^2(x) - 2yf_1(x) - f_2^2(x) + 2yf_2(x)}{2}\right]dydx \tag{90}$$

$$= \frac{1}{2}\int (f_1 - f_2)^2 p(x)dx = \frac{1}{2}\left\|f_1 - f_2\right\|_2^2 \tag{91}$$

Use the result from Guntuboyina and Sen, which says $\log N(\mathcal{F}, \epsilon, \|\cdot\|_2) \approx \epsilon^{-d/2}$, and solve

$$\log N(\mathcal{F}, \epsilon, d_{KL}) = \log N(\mathcal{F}, \frac{1}{\sqrt{2}}\epsilon, \|\cdot\|_2) \approx n\epsilon^2 \tag{92}$$

$$\log N(\mathcal{F}, 2\epsilon', \|\cdot\|_2) \approx 4n\epsilon^2 \tag{93}$$

We get $\epsilon'^2 \approx c(1/n)^{4/(4+d)}$ and the minimax lower bound has rate $(1/n)^{4/(4+d)}$.

4 Concluding Remarks

The whole report serves as an self-contained example of entropy method in statistics. Start from deriving the L_p-covering number of uniform bounded convex function defined on compact rectangles, we end up with a $n^{-4/(4+d)}$ rate minimax lower bound of L_2-risk in a convex regression.

Briefly speaking, for nonparametric regression, L_2-minimax rate ϵ_n^2 can be determined by solving the *metric entropy master equation* [1]

$$\log N(\mathcal{F}, 2\epsilon_n, \|\cdot\|_2) \approx n\epsilon_n^2 \tag{94}$$

, which is an implication of the aforementioned Fano's method. As we can see, sharper control of the metric entropy, either on covering or packing number, leads to more accurate minimax risk bound. Besides Fano's method, there are other entropy approaches in determining the minimax lower bound such as Assouad's method [2, 11].

For convex regression, [9] provides a brief overview. Some recent advances of metric entropy of convex functions lead to better understanding of convex regression. For example, Guntuboyina [5] himself extends the result by relaxing the uniform boundedness constraint. Moreover, [7] shows that the performance of convex regression is highly related to the support. Specifically, it shows that with polytope support in $\mathbb{R}^d$, the minimax rate is $n^{-4/(4+d)}$, which is a more general result of the example in section 3.

We should emphasize that discretization is the essence of entropy method. In deriving minimax bound, we reduce estimation problem with uncountable parameter space to a testing problem indexed by finite parameters. In other word, we divide the original parameter space into finite local neighborhoods. We analysis the performance on each local neighborhood, then combine the local results by taking the advantage that there are only finite neighborhood. The main cost is the discretization error in each local neighborhood, which is controlled by carefully constructing a covering or packing set. The same idea also appears in Dudley's entropy integral [10], which is based on discretization and chaining argument. To sum up, such discretization property of entropy method helps us control the performance of a statistical procedure even with infinite dimensional parameter space.

References

[1] Emmanuel Abbe and Martin Wainwright. Tutorial part i: Information theory meets machine learning. URL: http://www.princeton.edu/~eabbe/publications/tuto_slides_part1.pdf.

[2] John Duchi. Lecture notes: Information theory and statistics. URL: https://stanford.edu/class/stats311/Lectures/full_notes.pdf.

[3] R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1999. doi:10.1017/CBO9780511665622.

[4] A. Guntuboyina and B. Sen. Covering numbers for convex functions. *IEEE Transactions on Information Theory*, 59(4):1957–1965, April 2013.

[5] Adityanand Guntuboyina. Covering numbers of l_p-balls of convex functions and sets. *Constructive Approximation*, 43(1):135–151, Feb 2016. URL: https://doi.org/10.1007/s00365-015-9279-1.

[6] Fang Han. Stat 583 lecture note: Uniform entropy. URL: https://www.stat.washington.edu/~fanghan/teaching/STAT583/lec2.pdf.

[7] Qiyang Han and Jon A. Wellner. Multivariate convex regression: global risk bounds and adaptation. URL: https://arxiv.org/pdf/1601.06844.pdf.

[8] Philippe Rigollet. Lecture notes: Mathematics of machine learning. URL: https://ocw.mit.edu/courses/mathematics/18-657-mathematics-of-machine-learning-fall-2015/lecture-notes/MIT18_657F15_L6.pdf.

[9] Bodhisattva Sen. Convex regression. URL: http://www.stat.columbia.edu/~bodhi/Talks/NPCvxReg.pdf.

[10] Martin Wainwright. Metric entropy and its uses. URL: https://www.stat.berkeley.edu/~wainwrig/nachdiplom/Chap5_Sep10_2015.pdf.

[11] Yihong Wu. Lecture notes: Information-theoretic methods in high-dimensional statistics. URL: http://www.stat.yale.edu/~yw562/teaching/it-stats.pdf.

[12] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 10 1999. URL: https://doi.org/10.1214/aos/1017939142, doi:10.1214/aos/1017939142.