THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

# SDS 384 11: Theoretical Statistics

## Lecture 14: Uniform Law of Large Numbers- Covering number

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

# Definitions

- Recall that a metric space $(\mathcal{T}, \rho)$ consists of a nonempty set $\mathcal{T}$ and a mapping $\rho : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ that satisfies:
  - Non-negative: $\rho(\theta, \theta') \geq 0$ for all $(\theta, \theta')$ with equality iff $\theta = \theta'$.
  - Symmetric: $\rho(\theta, \theta') = \rho(\theta', \theta)$ for all pairs $(\theta', \theta)$, and
  - Triangle ineq holds: $\rho(\theta, \theta') + \rho(\theta', \theta'') \geq \rho(\theta, \theta'')$

- Examples:
  - $\mathcal{T} = \mathbb{R}^d$, $\rho(\theta, \theta') = \|\theta - \theta'\|_2$
  - $\mathcal{T} = \{0, 1\}^d$ with $\rho(\theta, \theta') = \frac{1}{d} \sum_i 1(\theta_i \neq \theta'_i)$

# Covering numbers

**Definition**

A $\delta$ cover of a set $\mathcal{T}$ w.r.t to a metric $\rho$ is a set $\{\theta^1, \ldots, \theta^N\} \subset \mathcal{T}$ such that for every $\theta \in \mathcal{T}$, $\exists i \in [N]$, s.t. $\rho(\theta, \theta^i) \leq \delta$. The $\delta$ covering number $N(\delta; \mathcal{T}, \rho)$ is the cardinality of the smallest $\delta$ cover.

- We will consider metric spaces which are totally bounded, i.e. $N(\delta; \mathcal{T}, \rho) < \infty$ for all $\delta > 0$.
- The covering number is non-increasing in $\delta$, i.e. $N(\delta) \geq N(\delta')$ for all $\delta < \delta'$
- We are interested in something called Metric entropy, which is the logarithm of the covering number.

# Picture

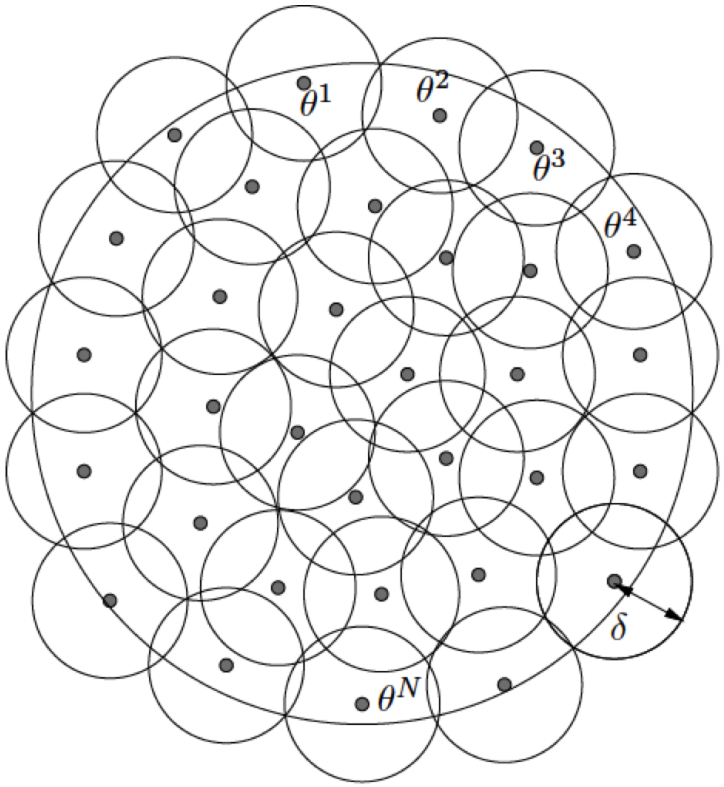


**Figure 1:** [courtesy: Martin Wainwright's book]

- A $\delta$ covering can be thought of as a union of balls with radius $\delta$.

# Covering number of a unit cube

**Example**

Consider the interval $[-1, 1]$ with $\rho(\theta, \theta') = |\theta - \theta'|$. We have $N(\delta; [-1, 1], |.|) \leq \frac{1}{\delta} + 1$

- Divide the interval into $L$ sub-intervals centered at $\theta^i := -1 + (2i - 1)\delta$ for $i \in [L]$ and each of length at most $2\delta$.
- By construction this is a $\delta$ covering.
- So $L \leq 1 + 1/\delta$

# Covering the binary hypercube

**Example**

Consider a $d$ dimensional binary hypercube $\mathcal{T} = \{0,1\}^d$ with the Hamming metric defined before.

$$\frac{\log N(\delta; \mathcal{T}, \rho)}{\log 2} \leq \lceil d(1-\delta) \rceil$$

- Let $S = \{1, 2, \ldots, \lceil \delta d \rceil\}$
- Consider the set of binary vectors $\mathcal{S}(\delta) := \{\theta \in \mathcal{T} : \theta_j = 0, j \in \mathcal{S}(\delta)\}$.
- By construction, for every binary vector $\theta' \in \mathcal{T}$, we can find a vector $\theta \in \mathcal{S}(\delta)$ such that $\rho(\theta, \theta') \leq \delta$
- $N(\delta; \mathcal{T}, \rho) \leq |\mathcal{S}(\delta)| = 2^{\lceil d(1-\delta) \rceil}$

# Lower bound on Covering number of the binary hypercube

- Let $\delta \in (0, 1/2)$
- If $\{\theta^1, \ldots, \theta^N\}$ is a $\delta$ covering, then the (unrescaled) Hamming balls of radius $s = \delta d$ around each $\theta^\ell$ must contain all $2^d$ vectors.
- Let $s = \lfloor \delta d \rfloor$
- For each $\theta^i$ there are exactly $\sum_{j=0}^{s} \binom{d}{j}$ vectors within $\delta d$ distance.
- So $N \sum_{j=0}^{s} \binom{d}{j} \geq 2^d$

# Lower bound on Covering number of the binary hypercube

- Let $\delta \in (0, 1/2)$
- So $N \sum_{j=0}^{s} \binom{d}{j} \geq 2^d$
- Now take a Binomial $(d, 1/2)$ random variable $X$.
- $P(X \leq \delta d) = \sum_{j=0}^{s} \binom{d}{j} / 2^d$
- So $N \geq \frac{1}{P(X \leq \delta d)}$
- Using the Hoeffding bound gives: $N \geq \exp(\frac{d}{2}(1/2 - \delta)^2)$
- Using the refined version in your homework gives: $N \geq \exp(d \times KL(\delta || 1/2))$

# Packing numbers

**Definition**

An $\delta-$packing of $\mathcal{T}$ w.r.t a metric $\rho$ is a set $\{\theta^1, \ldots, \theta^M\} \subset \mathcal{T}$ such that $\rho(\theta^i, \theta^j) > \delta$ for every distinct pair $i, j \in [M]$. The $\delta$ packing number $M(\delta; \mathcal{T}, \rho)$ is the cardinality of the largest $\delta$ packing.

# Picture

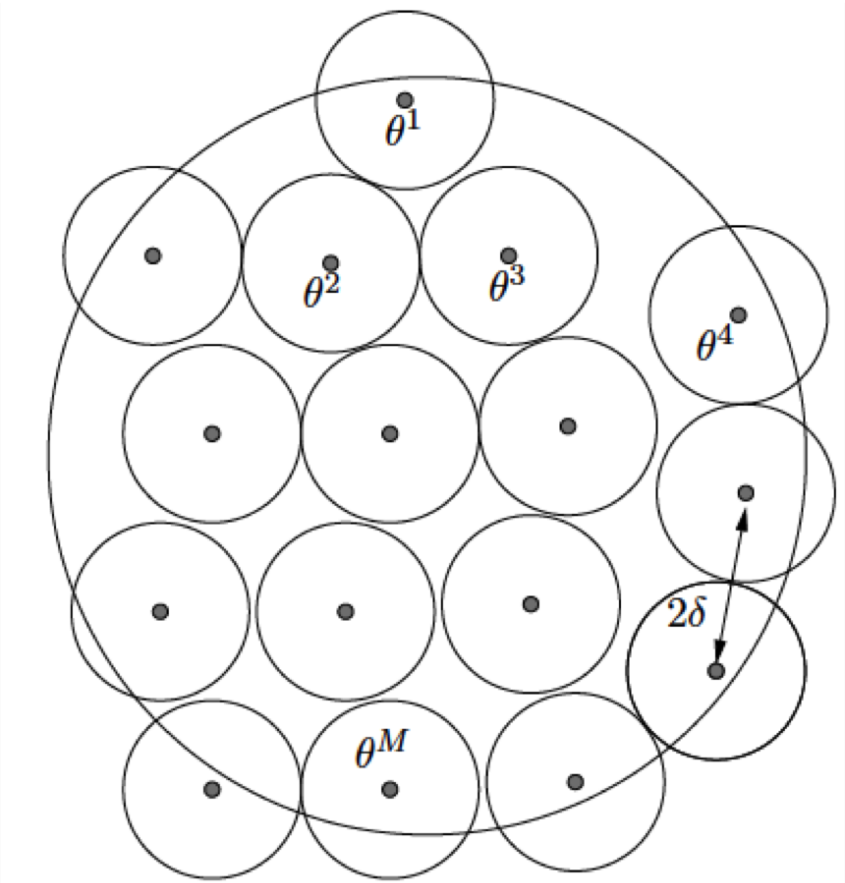


**Figure 2:** [courtesy: Martin Wainwright's book]

- A $2\delta$ packing can be thought of as a union of balls with radius $\delta$ such that no two balls touch.

# Relationship between packing and covering numbers

**Theorem**

*For all* $\delta > 0$,

$$M(2\delta; \mathcal{T}, \rho) \leq N(\delta; \mathcal{T}, \rho) \leq M(\delta; \mathcal{T}, \rho)$$

- This is saying that packing and covering numbers exhibit the same scaling behavior as $\delta \to 0$.

# Proof

- **Upper bound:** Let $V = \{x_1, \ldots, x_N\}$ be a $\delta$ packing of $\mathcal{T}$. So for each $y \in T \setminus V$, $\exists i, \|y - x_i\| \leq \delta$. Otherwise we could have added this point and increased the packing number. So, $V$ is also a $\delta$ cover. But since the covering number is the size of the smallest $\delta$ covering, the lower bound holds.
- **Lower bound:** Say there is a $2\delta$ packing $\{y_1, \ldots, y_M\}$ and a $\delta$ covering $\{v_1, \ldots, v_n\}$ with $M > n$. Now by pigeonhole, there must be two $y_i, y_j$ who both are in the $\delta$ ball around some $v_k$. But using triangle, we will have $|y_i - y_j| \leq 2\delta$, which is a contradiction. So we must have $m \leq n$.

# Covering and Packing numbers-example

**Theorem**

*Let $\rho$ be the Euclidean norm on $\mathbb{R}^d$. Let $B_1(0)$ be the unit ball centered at the origin (WLOG).*

$$\frac{1}{\epsilon^d} \leq N(\epsilon, B_1, \rho) \leq (1 + 2/\epsilon)^d$$

- Consider an $\epsilon$ cover $\{\theta^1, \ldots, \theta^N\}$. Now,

$$B_1 \subseteq \bigcup_{i=1}^{N} B_\epsilon(\theta^i)$$

$$\text{vol}(B_1) \leq N\text{vol}(B_\epsilon(\theta^i)) = N\epsilon^d \text{vol}(B_1)$$

$$N \geq 1/\epsilon^d$$

# Proof-upper bound

- Consider a $\epsilon$ packing $\{\theta^1, \ldots, \theta^M\}$
- This is an union of disjoint balls of radius $\epsilon/2$

$$\bigcup_i B_{\epsilon/2}(\theta^i) \subseteq B_{1+\epsilon/2}$$

$$M\text{vol}(B_{\epsilon/2}(\theta^i)) \leq (1+\epsilon/2)\text{vol}(B_{1+\epsilon/2})$$

$$M(\epsilon/2)^d\text{vol}(B_1) \leq (1+\epsilon/2)^d\text{vol}(B_1)$$

$$M \leq (1+2/\epsilon)^d$$

# Suprema over an infinite space

**Theorem**

*Consider a $d$ dimensional vector of independent subG($\sigma^2$) random variables. Let $B_d$ be the unit ball in $\|.\|_2$ norm. Then the following holds:*

$$E[\sup_{\theta \in B_d} \theta^T X] \leq 4\sigma\sqrt{d}$$

*Also, for $\delta \in (0,1)$, with probability $1-\delta$,*

$$\sup_{\theta \in B_d} \theta^T X \leq 4\sigma\sqrt{d} + \sqrt{2\sigma \log(1/\delta)}.$$

# [Recall] Size of a function class $\mathcal{F}$

- Let $\mathcal{F}(X) = \{(f(X_1), \ldots, f(X_n)) : f \in \mathcal{F}\}$
- $\mathcal{R}_{\mathcal{F}} = E \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \epsilon_i f(X_i) \right| = E \left[ E \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_i \epsilon_i f(X_i)| \middle| X_1, \ldots, X_n \right]$
- In the next slide we will bound this using the cardinality of $\mathcal{F}(X)$

**Theorem**

*Let* $A \subseteq \mathbb{R}^n$, $R = \max_{a \in A} \|a\|$,

$$E \sup_{a \in A} \langle \epsilon, a \rangle \leq \sqrt{2R^2 \log |A|}.$$

*And,*

$$E \sup_{a \in A} |\langle \epsilon, a \rangle| \leq \sqrt{2R^2 \log |2A|}.$$

# Proof of first half

- Let $\mathcal{N}_{1/2}$ be a half covering of $B_1$. So $N(1/2, B_1, \|\|_2) \leq 5^d$
- So for each $\theta \in B_d$, $\exists z_\theta \in \mathcal{N}_{1/2}$ such that

$$\theta = z_\theta + x, \qquad \|x\| \leq 1/2$$

- So,

$$Y := \sup_{\theta \in B_1} \theta^T X \leq \max_{z_\theta \in \mathcal{N}_{1/2}} z_\theta^T X + \underbrace{\sup_{x \in B_{1/2}} x^T X}_{Y/2}$$

- Thus, we have:

$$EY \leq 2E\left[\max_{z_\theta \in \mathcal{N}_{1/2}} z_\theta^T X\right] \leq 2\sigma\sqrt{2\log|\mathcal{N}_{1/2}|} \leq \sigma\sqrt{8d\log 5} \leq 4\sigma\sqrt{d}$$

- We used the same result as last time.

# Proof of part 2

$$
\begin{aligned}
P(Y \geq t) &\leq P(\max_{z \in \mathcal{N}_{1/2}} z^T X \geq t/2) \\
&\leq |\mathcal{N}_{1/2}| P(z^T X \geq t/2) \\
&\leq 5^d \exp(-t^2/8\sigma^2 \|z\|^2) \leq 5^d \exp(-t^2/4\sigma^2) = \delta
\end{aligned}
$$

Solving for $t$ gives, $\sqrt{\log 5 + \log(1/\delta)} = t/2\sigma$. In fact, we can get an upper bound on $t$ as follows.

$$
t = 2\sigma\sqrt{d \log 5 + \log(1/\delta)} \leq 2\sigma\sqrt{d \log 5} + 2\sigma\sqrt{\log(1/\delta)} =: t_0
$$

Thus, $P(Y \geq t_0) \leq P(Y \geq t) \leq \delta$

# Example-smoothly parametrized problems

- Consider the following function class parametrized by $\theta \in \Theta$.

$$\mathcal{F} := \{f_\theta(.) : \theta \in \Theta\}$$

- Let $\|.\|_\Theta$ be the norm for $\theta$ and $\|.\|_\mathcal{F}$ be the norm for $\mathcal{F}$.
- Say $\|f_\theta(.) - f_{\theta'}(.)\|_\mathcal{F} \leq L\|\theta - \theta'\|_\Theta$
- Then $N(\epsilon; \mathcal{F}, \|.\|_F) \leq N(\epsilon/L; \Theta, \|.\|_\Theta)$

# Example-smoothly parametrized problems

- A Lipschtiz parametrization allows us to go from cover of the $\Theta$ space to cover of the $f_\theta$ space with a loss of $L$.
- If $\mathcal{F}$ is parametrized by a compact set of $d$ parameters then $N(\epsilon, \mathcal{F}) = O(1/\epsilon^d)$

# A parametric class of Lipschitz continuous functions

**Example**

For any fixed $\theta$, define the real-valued function $f_\theta(x) := \exp(-\theta x)$, and consider the function class

$$\mathcal{F} = \{f_\theta : [0,1] \to \mathbb{R} | \theta \in [0,1]\}$$

Using the uniform norm as a metric, i.e.

$\|f - g\|_\infty := \sup_{x \in [0,1]} |f(x) - g(x)|$. Prove that

$$\lfloor \frac{1 - 1/e}{2\delta} \rfloor + 1 \leq N(\delta; \mathcal{F}, \|.\|_\infty) \leq \frac{1}{2\delta} + 2.$$

# Proof-upper bound

- First note that $\|f_\theta - f_{\theta'}\|_\infty \leq |\theta - \theta'|$
- For any $\delta \in (0,1)$, let $T = \lfloor \frac{1}{2\delta} \rfloor$
- Consider $S = \{\theta^0, \ldots, \theta^{T+1}\}$ where $\theta^i = 2\delta i$ for $i \leq T$ and $\theta^{T+1} = 1$.
- $\{f_{\theta^i} : \theta^i \in S\}$ is a $\delta$ cover for $\mathcal{F}$.
- For any $\theta \in [0,1]$ we can find $\theta^i \in S$ such that $|\theta^i - \theta| \leq \delta$
- Indeed we have,

$$\begin{aligned} \|f_{\theta^i} - f_\theta\|_\infty &= \sup_{x \in [0,1]} |\exp(-\theta^i x) - \exp(-\theta x)| \\ &\leq |\theta^i - \theta| \leq \delta \end{aligned}$$

So $N(\delta; \mathcal{F}, \|.\|_\infty) \leq 2 + T \leq 2 + \frac{1}{2\delta}$

# Proof-lower bound

- We will do a $\delta$ packing.
- Let $\theta^i = -\log(1 - i\delta)$ for $i = 0, \ldots, T$
- $-\log(1 - T\delta) = 1$, and so the largest integral value is $T = \lfloor \frac{1 - 1/e}{\delta} \rfloor$
- So $M(\delta; \mathcal{F}, \|.\|_\infty) \geq 1 + \lfloor \frac{1 - 1/e}{\delta} \rfloor$
- $N(\delta; \mathcal{F}, \|.\|_\infty) \geq M(2\delta; \mathcal{F}, \|.\|_\infty) \geq 1 + \lfloor \frac{1 - 1/e}{2\delta} \rfloor$

# Proof-lower bound

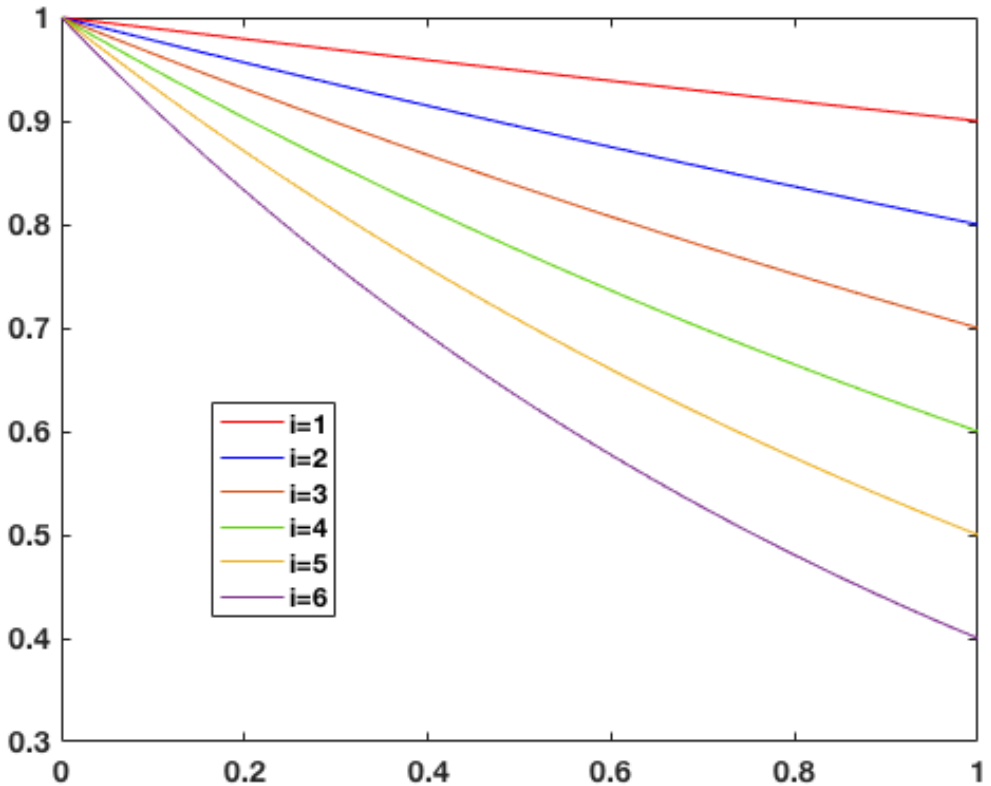


**Figure 3:** $\exp(-\theta^i x)$ where $\theta^i = -\log(1 - i\delta)$

# Example-Lipschitz functions on the unit interval

**Example**

$$\mathcal{F}_L = \{g : [0,1] \to \mathbb{R} | g(0) = 0, |g(x) - g(y)| \leq L|x - x'|, \forall x, x' \in [0,1]\}$$

Metric entropy scales as $\log N(\delta; \mathcal{F}_L, \|.\|_\infty) \asymp L/\delta$ for small enough $\delta > 0$.

# Proof

- Its sufficient to consider a sufficiently large packing of $\mathcal{F}_L$
- For a given $\epsilon$ define $M = \lfloor \frac{1}{\epsilon} \rfloor$
- Let $x_i = (i-1)\epsilon$ for $i = 1, \ldots, M+1$
-

$$\phi(x) := \begin{cases} 0 & x < 0 \\ x & x \in [0,1] \\ 1 & x > 1 \end{cases} \tag{1}$$

- Define $f_\beta(x) = \sum_{i=1} \beta_i L\epsilon\phi\left(\frac{x - x_i}{\epsilon}\right)$ for $\beta \in \{-1, 1\}^M$

# Picture

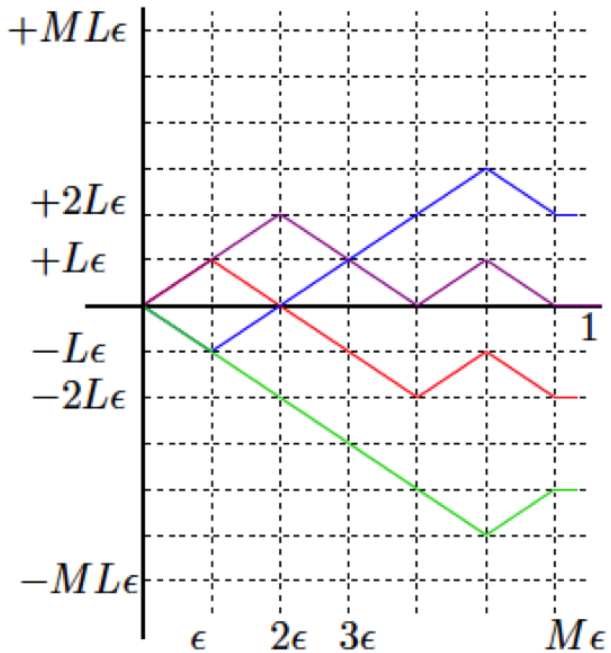


**Figure 5-2.** The function class $\{f_\beta, \beta \in \{-1,+1\}^M\}$ used to construct a packing of the Lipschitz class $\mathscr{F}_L$. Each function is piecewise linear over the intervals $[0,\epsilon]$, $[\epsilon,2\epsilon]$, ..., $[(M-1)\epsilon, M\epsilon]$ with slope either $+L$ or $-L$. There are $2^M$ functions in total, where $M = \lfloor 1/\epsilon \rfloor$.

# example

- Note that the functions in the packing are all continuous, although they are not differentiable.
- Lipschitz continuous functions are continuous, but they do not have to be necessarily differentiable.
- Lipschitz continuous functions are differentiable *almost everywhere*.
- Rademacher's theorem basically says that the number of discontinuities of a Lipschitz continuous function are of measure zero.

# example

- For any pair $\beta \neq \beta' \in \{-1, 1\}^M$ there is at least one interval where they have the same starting point.
- So $\|f_\beta(x) - f'_\beta(x)\|_\infty \geq 2L\epsilon$
- $f_\beta \in \mathcal{F}_L$ for all $\beta \in \{-1, 1\}^M$
- So $f_\beta$ forms a $2L\epsilon$ packing.
- Making $\epsilon L = \delta$ we see

$$N(\delta; \mathcal{F}_L, \|.\|_\infty) \geq M(2L\epsilon; \mathcal{F}_L, \|.\|_\infty) = 2^{\lfloor \frac{1}{\epsilon} \rfloor} = 2^{\lfloor \frac{L}{\delta} \rfloor}$$

- Also the set $f_\beta$ also form a suitable covering of the original functions, and this gives the upper bound.

# example

- The last example can be extended to Lipschitz functions on the Unit cube in higher dimensions, i.e.

$$|f(x) - f(y)| \leq \|x - y\|_\infty \qquad \text{for all } x, y \in [0,1]^d$$

- The same method can be used to show that the metric entropy for this class is the same order as $(L/\delta)^d$

# Make a comparison

- Recall that for a $L$ Lipschitz continuous functions supported on $[0,1]$ with $f(0) = 0$, the metric entropy was $L/\delta$
- Also recall that for a $L$ Lipschitz continuous functions supported on $[0,1]^d$ with $f(0) = 0$, the metric entropy was $(L/\delta)^d$
- However for a given function class like the last one the metric entropy is $\log(1/\delta)$
- Recall that for Unit hypercubes in $d$ dimensions the metric entropy is $d \log(1 + 1/\delta)$
- Note that for Lipschitz continuous functions the dependence on $d$ is exponential. This is a much richer class of functions, so the size is considerably larger and scales poorly with $d$.

# Acknowledgment

This lecture was very much based on Martin Wainwright's book.