# Covering and packing numbers

Jonathan Ma
johnma@udel.edu

July 2, 2025

## 1 Covering numbers

Let's start by defining something familiar.

**Definition 1.1.** Let $S$ be a set. We call $d : S \times S \to \mathbb{R}^+$ a *pseudometric* if it is symmetric, satisfies $d(x, x) = 0$, $\forall x \in S$, and satisfies the triangle inequality. We call $(S, d)$ a *pseudometric space.*

**Note.** Unlike a metric, we don't insist $d(x, y) = 0 \implies x = y$.

**Definition 1.2.** An *$\epsilon$-cover* of a subset $T$ of a pseudometric space $(S, d)$ is a set $\widehat{T} \subset T$ such that for each $t \in T$, there is a $\hat{t} \in \widehat{T}$ such that $d(t, \hat{t}) \leq \varepsilon$. [1]

**Note.** I've seen some definitions drop the requirement that $\widehat{T}$ be a subset of $T$. I do not know how that changes the analysis we will do. I've seen more resources use our given definition, and for the topic of dimensionality reduction, our definition probably makes more sense to use.

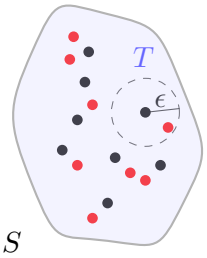


Figure 1: An *$\epsilon$-cover* (red) of a subset $T$ of $S$.

**Definition 1.3.** The *$\epsilon$-covering number* of $T \subset S$ is

$$N(\epsilon, T, d) = \min \left\{ |\widehat{T}| \mid \widehat{T} \text{ is a } \epsilon\text{-cover of T} \right\}.$$

**Definition 1.4.** A set $T$ is *totally bounded* if for all $\epsilon > 0$, $N(\epsilon, T, d) < \infty$.

When $d$ is known, we may drop it from the notation $N(\epsilon, T, d)$, or even the set $T$.

**Definition 1.5.** The function $\epsilon \mapsto \log N(\epsilon, T, d)$ is called the *metric entropy* of $T$.

**Definition 1.6.** If $\lim_{\epsilon \to 0} \frac{\log N(\epsilon, T)}{\log(1/\epsilon)}$ exists, it is called the *metric dimension* of $T$.

**Example 1.7.** Let $d \in \mathbb{Z}^+$. Consider $([0, 1]^d, l^\infty)$, the unit $d$-cube endowed the with metric induced by the $l^\infty$ norm. We claim that $N(\epsilon, [0, 1]^d, l^\infty) = \Theta\left(\frac{1}{\epsilon^d}\right)$.

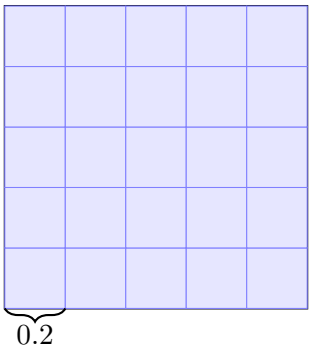


Figure 2: A .2-covering of the unit square $[0,1]^2$ under the $l^\infty$ metric.

See Fig. 2. We can see that when $\epsilon = .2$, a minimal $\epsilon$-cover of $[0,1]^2$ is given by tiling the square. (It helps that $.2 \mid 1$, in this case.) Generalizing to arbitrary $\epsilon > 0$, we can show that "tiling" the square provides an $\epsilon$-cover with $\lceil \frac{1}{\epsilon} \rceil^d = \Theta\left(\frac{1}{\epsilon^d}\right)$ balls. So intuitively, *we would expect $d$-dimensional sets in general to have metric dimension $d$*, or that $N(\epsilon) = \Theta(1/\epsilon^d)$.

## 1.1 Packing numbers

**Definition 1.8.** An *$\epsilon$-packing* of a subset $T$ of a pseudometric space $(S, d)$ is a subset $\widehat{T} \subset T$ such that each pair $s, t \in \widehat{T}$ satisfies $d(s,t) > \epsilon$. The *$\epsilon$-packing number* of $T$ is

$$M(\epsilon, T, d) = \max\left\{ |\widehat{T}| \mid \widehat{T} \text{ is an } \epsilon\text{-packing of } T \right\}.$$

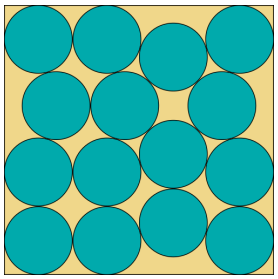

Figure 3: (Not my image.) Let's say this is the unit square, and the metric is induced by the $l^2$ norm. Set our $\epsilon$-packing to be the centers of each ball, and say that each ball has radius $\epsilon$.

Circle packing (in $\mathbb{R}^d$) is a classical problem of packing circles into a set (typically connected, with boundary). Famously it was studied by Lagrange, in his calculus of variations (among other topics). Here we can think of this definition as a generalization of this idea. The notion of covering numbers involves a minimization problem, while the notion of packing numbers involves a maximization problem. It turns out that these two quantities are closely related.

**Theorem 1.9.** *For all $\epsilon > 0$, $M(2\epsilon) \leq N(\epsilon) \leq M(\epsilon)$.*

Thus the scaling of covering and packing numbers is the same.

*Proof.* For the first inequality, consider a minimal $\epsilon$-cover $\widehat{T}$ of $T$. Any two elements of a $2\epsilon$-packing of $T$ cannot be within $\epsilon$ of the same element of $\widehat{T}$, otherwise the triangle inequality shows that they are within

[1] Fun fact. This same notion of covering sets is alternatively used in the Neural Operator JMLR paper, under the language of "discretization invariance."

$2\epsilon$ of each other. Thus there can be no more than one element of a $2\epsilon$ packing for each of the $N(\epsilon)$ elements of $\widehat{T}$.

For the second inequality, consider an $\epsilon$-packing $\widehat{T}$ of size $M(\epsilon)$. Since it is maximal, no other point $s \in T$ can be added for which some $t \in \widehat{T}$ has $d(s,t) > \epsilon$. Thus $\widehat{T}$ is an $\epsilon$-cover, so the minimal $\epsilon$-cover has size $N(\epsilon) \leq M(\epsilon)$. □

**Example 1.10.** Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$, and let $B$ be the unit ball under that norm. Then we claim

$$\frac{1}{\epsilon^d} \leq N(\epsilon, B, \|\cdot\|) \leq \left(\frac{2}{\epsilon} + 1\right)^d.$$

*Proof.* For the lower bound, let $\{x_1, \ldots, x_N\}$ be an $\epsilon$-cover of size $N = N(\epsilon, B)$, and consider that

$$B \subseteq \bigcup_{i=1}^{N} (x_i + \epsilon B).$$

By countable subadditivity of Lebesgue measure,

$$\mu(B) \leq N(\epsilon, B)\mu(\epsilon B) = N(\epsilon, B)\epsilon^d \mu(B),$$

hence $N(\epsilon, B) \geq 1/\epsilon^d$. For the upper bound, consider a maximal $\epsilon$-packing $\{x_1, \ldots, x_M\}$ of size $M = M(\epsilon, B)$. Since it is a packing, the balls $x_i + (\epsilon/2)B$ are disjoint. Each of these balls is contained in $(1+\epsilon/2)B$. Thus

$$\bigcup_{i=1}^{M} (x_i + \frac{\epsilon}{2}B) \subseteq (1 + \epsilon/2)B.$$

Therefore by additivity of Lebesgue measure,

$$\begin{aligned} M\mu((\epsilon/2)B) &\leq \mu((1+\epsilon/2)B) \\ \implies \left(\frac{\epsilon}{2}\right)^d \mu(B) &\leq \left(1 + \frac{\epsilon}{2}\right)^d \mu(B). \end{aligned}$$

Hence, $N(\epsilon, B) \leq M(\epsilon, B) \leq (2/\epsilon + 1)^d$. □

# 2 Lipschitz mappings

**Theorem 2.1.** *Let $F$ be a parametrized class of functions*

$$F = \{f(\theta, \cdot) \mid \theta \in \Theta\}.$$

*Let $\|\cdot\|_\Theta$ be a norm on $\Theta$, and let $\|\cdot\|_F$ be a norm on $F$. Suppose that the mapping $\theta \mapsto f(\theta, \cdot)$ is $L$-Lipschitz, that is*

$$\left\|f(\theta, \cdot) - f(\theta', \cdot)\right\|_F \leq L \left\|\theta - \theta'\right\|_\Theta.$$

*Then $N(\epsilon, F, \|\cdot\|_F) \leq N(\epsilon/L, \Theta, \|\cdot\|_\Theta)$.*

*Proof.* Let's try to prove this live to fill time? Or look it up in the morning. □

We care about Lipschitz parametrization as they allow us to translate covers of the parameter space into covers of the function space. For example, if $F$ is smoothly parametrized by a compact set of $d$ parameters, then $N(\epsilon, F) = \mathcal{O}(1/\epsilon^d)$.

**Example 2.2** (1-dimensional Lipschitz functions)**.** Let $F$ be the set of $L$-Lipschitz functions mapping from $[0,1] \to [0,1]$. Then in the infinity norm $\|f\|_\infty = \sup_{x \in [0,1]} |f(x)|$,

$$\log N(\epsilon, F, \|\cdot\|_\infty) = \Theta(L/\epsilon).$$

*Proof idea.* Form an $\epsilon$ grid of the $y$-axis, and an $\epsilon/L$ grid of the $x$-axis, and consider all functions that are piecewise linear on this grid, where all pieces have slope $L$ or $L$. There are $1/\epsilon$ starting points, and for each starting point, there are $2^{L/\epsilon}$ slope choices. It remains to show that this set is an $\mathcal{O}(\epsilon)$ packing and an $\mathcal{O}(\epsilon)$ cover. □

**Example 2.3.** Let $F_d$ be the set of $L$-Lipschitz functions wrt $\|\cdot\|_\infty$, this time mapping from $[0,1]^d$ to $[0,1]$. Then

$$\log N(\epsilon, F_d, \|\cdot\|_\infty) = \Theta((L/\epsilon)^d).$$

Note the exponential dependence on the dimension.

## 3 Johnson-Lindenstrauss lemma and its optimality

Many famous people have worked on Johnson-Lindenstrauss, including Noga Alon, and Terrence Tao. Alon brought a combinatorial approach towards JL, while it seems Tao focuses on JL's applications to signal reconstruction in "*compressed sensing*."

**Theorem 3.1** (Johnson-Lindenstrauss lemma)**.** *Let $X \subset \mathbb{R}^d$ be any set of size $n$, and let $\epsilon \in (0, 1/2)$ be arbitrary. Then there exists a map $f : X \to \mathbb{R}^m$ for some $m = \mathcal{O}(\epsilon^{-2} \log n)$ such that*

$$\forall x, y \in X, \quad (1-\epsilon)\|x-y\|_2^2 \le \|f(x)-f(y)\|_2^2 \le (1+\epsilon)\|x-y\|_2^2. \tag{1}$$

JL lemma is a so-called bi-Lipschitz embedding, in a finite setting. An elegant and short proof of this theorem uses the probabilistic method of combinatorics, made famous by Alon and Erdös. In the original JL paper, it was proven for any $\epsilon < 1/2$, there exists $n$ point sets $X \subset \mathbb{R}^n$ for which any embedding $f : X \to \mathbb{R}^m$ satisfying (1) must have $m = \Omega(\log n)$. Alon and Levenshtein later showed the existence of an $n$ point set $X \subset \mathbb{R}^n$ such that any $f$ satisfying (1) must have $m = \Omega(\min\{n, \epsilon^{-2}\log n/\log(1/\epsilon)\})$. Larson and Nelson[2] prove the following result settling the optimality of the JL lemma, for almost the full range of $\epsilon$.

**Theorem 3.2.** *For any integers $n, d \ge 2$, and $\epsilon \in (\log^{0.5001} n/\sqrt{\min\{n,d\}}, 1)$, there exists a set of points $X \subset \mathbb{R}^d$ of size $n$ such that any map $f : X \to \mathbb{R}^m$ providing (1) must have*

$$m = \Omega(\epsilon^{-2}\log \epsilon^2 n).$$

Their proof uses covering numbers, but it is too long to cover in this presentation. This should conclude this presentation.

## References


[1] Kasper Green Larsen and Jelani Nelson. Optimality of the Johnson-Lindenstrauss Lemma, 2017.

[2] Patrick Rebeschini. VC Dimension, Covering and Packing Numbers. https://www.stats.ox.ac.uk/~rebeschi/teaching/AFoL/22/material/lecture04.pdf, 2022. Accessed: 2025-07-02.

[3] Peter L. Bartlett. Theoretical Statistics. Lecture 12, Metric Entropy. https://www.stat.berkeley.edu/~bartlett/courses/2013spring-stat210b/notes/12notes.pdf, 2013. Accessed: 2025-07-02.

[4] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the Restricted Isometry Property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[5] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Contemporary Mathematics*, volume 26, pages 189–206. American Mathematical Society, 1984.


[2] Nelson is chair at UC Berkeley CS now.