CPSC 532D — 4. INFINITE $\mathcal{H}$ WITH COVERING NUMBERS

Danica J. Sutherland

University of British Columbia, Vancouver

Fall 2025

In Chapter 2, we showed that ERM probably gets approximately the right answer on finite hypothesis classes. As we discussed in Section 2.3.1, even though everything we do on a computer is ultimately finite, it's not really satisfying to analyze things that way: we need to incorporate somehow that "similar" hypotheses probably have similar error. So, we might as well go to a case where $|\mathcal{H}| = \infty$.

In *logistic regression*, our data is in a subset of $\mathbb{R}^d$, our labels are in $\mathcal{Y} = \{-1, 1\}$, and we try to predict with a confidence score in $\widehat{\mathcal{Y}} = \mathbb{R}$. Our predictors are linear functions of the form $h_w(x) = w \cdot x$, and the logistic loss is given by

This is more convenient than $\mathcal{Y} = \{0, 1\}$ here...

You usually want an intercept term, $w \cdot x + w_0$, but you can achieve that by padding x with an always-one dimension.

$$\ell_{log}(h, (x, y)) = l_y^{log}(h(x)) = \log(1 + \exp(-h(x)y)). \tag{4.1}$$

For the probabilistically-minded among you, this corresponds to maximizing the likelihood of a model that takes $\hat{p}(y \mid x) = 1/(1 + \exp(-h(x)))$.

We'll use the hypothesis class $\mathcal{H} = \{h_w = x \mapsto w \cdot x : w \in \mathbb{R}^d, \|w\| \le B\}$ for some constant B; this avoids overfitting by using really-really complex w, and is basically equivalent to doing L_2-regularized logistic regression (we'll talk about this more later). This $\mathcal{H}$ is still infinite, but it has finite volume.

Now, our analysis is going to be based on the idea that if w and v are similar predictors, i.e. $h_w(x) \approx h_v(x)$ for all x, then they'll behave similarly: $L_\mathcal{D}(h_w) \approx L_\mathcal{D}(h_v)$ and $L_S(h_w) \approx L_S(h_v)$. Thus we don't have to do a totally separate concentration bound on their empirical risks; we can exploit that they're similar.

The fundamental idea is going to be one of a "set cover," or an "ε-net." To handle an infinite $\mathcal{H}$ that's nonetheless bounded, we're going to choose some *finite* set $\mathcal{H}_0$ such that everything in $\mathcal{H}$ is close to something in $\mathcal{H}_0$, use Proposition 2.2 to say that $L_\mathcal{D}(h) - L_S(h)$ isn't too big for anything in $\mathcal{H}_0$, and then argue that since $L_\mathcal{D}(h) - L_S(h)$ is smooth, this means it can't be too big for anything in $\mathcal{H}$ at all.

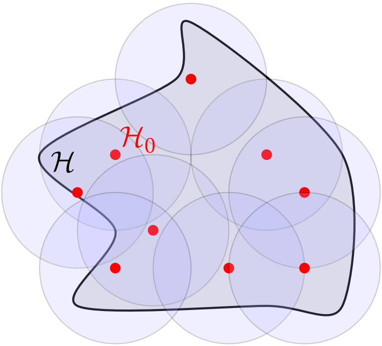

Figure 4.1: A (non-minimal) set cover.

For more, visit https://cs.ubc.ca/~dsuth/532D/25w1/.

4.1 SMOOTHNESS: LIPSCHITZ FUNCTIONS

To formalize the idea that similar weight vectors give similar loss, we'll want a bound like

$$|L_{\mathcal{D}}(h) - L_{\mathcal{D}}(g)| \le M\, \rho_{\mathcal{H}}(h, g),$$

for some notion of a distance metric on $\mathcal{H}$.

Definition 4.1. A *metric* on a set $\mathcal{X}$ is a function $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for all $x, y, z \in \mathcal{X}$: (i) $\rho(x, x) = 0$; (ii) if $x \neq y$, then $\rho(x, y) > 0$; (iii) $\rho(x, y) = \rho(y, x)$; (iv) $\rho(x, z) \le \rho(x, y) + \rho(y, z)$.

Definition 4.2. A function $f : \mathcal{X} \to \mathcal{Y}$ is M-*Lipschitz* with respect to metrics $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$ if for all $x, x' \in \mathcal{X}$, $\rho_{\mathcal{Y}}(f(x), f(x')) \le M\, \rho_{\mathcal{X}}(x, x')$. The smallest M for which this inequality holds is *the Lipschitz constant*, denoted $\|f\|_{\mathrm{Lip}}$.

If $\mathcal{X}$ and/or $\mathcal{Y}$ are subsets of $\mathbb{R}^d$, we assume that the corresponding ρ is Euclidean distance unless we say otherwise.

So, for example, $x \mapsto |x|$ is a 1-Lipschitz function, since $\big||x| - |y|\big| \le |x - y|$.

The notation $\|f\|_{\mathrm{Lip}}$ is justified by the following result, which uses notions of function spaces described in Appendix B.

Lemma 4.3. *Let $\mathcal{F}$ be a vector space of functions $\mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is a normed space, such that $f + g \in \mathcal{F}$ is the function $x \mapsto f(x) + g(x)$ and $af \in \mathcal{F}$ is the function $x \mapsto af(x)$ for any $a \in \mathbb{R}$. $\|\cdot\|_{\mathrm{Lip}}$ with respect to the metric $\rho_{\mathcal{Y}}(y, y') = \|y - y'\|_{\mathcal{Y}}$ is a seminorm on $\mathcal{F}$: it satisfies $\|f + g\|_{\mathrm{Lip}} \le \|f\|_{\mathrm{Lip}} + \|g\|_{\mathrm{Lip}}$ and $\|af\| = |a|\,\|f\|_{\mathrm{Lip}}$.*

Proof. Both properties can be shown fairly directly:

$$\begin{aligned}
\|f + g\|_{\mathrm{Lip}} &= \sup_{x \neq x'} \frac{\|f(x) + g(x) - f(x') - g(x')\|}{\rho_{\mathcal{X}}(x, x')} \\
&\le \sup_{x \neq x'} \frac{\|f(x) - f(x')\|}{\rho_{\mathcal{X}}(x, x')} + \frac{\|g(x) - g(x')\|}{\rho_{\mathcal{X}}(x, x')} \le \|f\|_{\mathrm{Lip}} + \|g\|_{\mathrm{Lip}} \\
\|af\|_{\mathrm{Lip}} &= \sup_{x \neq x'} \frac{\|af(x) - af(x')\|}{\rho_{\mathcal{X}}(x, x')} = \sup_{x \neq x'} \frac{|a|\,\|f(x) - f(x')\|}{\rho_{\mathcal{X}}(x, x')} = |a|\,\|f\|_{\mathrm{Lip}}.
\end{aligned}$$

□

It's only a seminorm, not a proper norm, because there are nonzero functions with zero Lipschitz constant: for instance, $x \mapsto a$ for any $a \in \mathbb{R}$.

So, what is $\|L_{\mathcal{D}}\|_{\mathrm{Lip}}$? Well, to start, the properties above imply that

$$|L_{\mathcal{D}}(h) - L_{\mathcal{D}}(g)| = \left| \mathop{\mathbb{E}}_{z \sim \mathcal{D}} \ell(h, z) - \mathop{\mathbb{E}}_{z \sim \mathcal{D}} \ell(g, z) \right| \le \mathop{\mathbb{E}}_{z \sim \mathcal{D}} |\ell(h, z) - \ell(g, z)|,$$

which means that $\|L_{\mathcal{D}}\|_{\mathrm{Lip}} \le \mathbb{E}_{z \sim \mathcal{D}} \|h \mapsto \ell(h, z)\|_{\mathrm{Lip}}$, where the Lipschitz constant of the function $h \mapsto \ell(h, z)$, which we can also more compactly write as $\ell(\cdot, z)$, is with respect to some metric on $\mathcal{H}$.

For the same reason, we have that $\|L_S\|_{\mathrm{Lip}} \le \frac{1}{m} \sum_{i=1}^{m} \|\ell(\cdot, z_i)\|_{\mathrm{Lip}}$; in fact, this is a special case of the result for $L_{\mathcal{D}}$, noting that $L_S = L_{\hat{\mathcal{D}}}$ where $\hat{\mathcal{D}}$ is the *empirical distribution*, the discrete distribution that puts $1/m$ probability at each of the points in S.

To make this a little more concrete, let's think about $z = (x, y)$ and $\ell(h, (x, y)) = l_y(h(x))$, along with $\mathcal{H}$ being linear predictors, $h_w = (x \mapsto w \cdot x)$. Then we have that

$$\begin{aligned} |\ell(h_w, (x, y)) - \ell(h_v, (x, y))| &= |l_y(h_w(x)) - l_y(h_v(x))| \\ &= |l_y(w \cdot x) - l_y(v \cdot x)| \\ &\le \|l_y\|_{\mathrm{Lip}} |w \cdot x - v \cdot x| \\ &\le \|l_y\|_{\mathrm{Lip}} \, \|x\| \, \|w - v\|, \end{aligned}$$

and so if we use the metric on $\mathcal{H}$ of $\rho(h_w, h_v) = \|w - v\|$, then we've shown that $\|\ell(\cdot, (x, y))\|_{\mathrm{Lip}} \le \|l_y\|_{\mathrm{Lip}} \, \|x\|$. This implies from the properties of Lemma 4.3 that

$$\|\mathrm{L}_{\mathcal{D}} - \mathrm{L}_{\mathrm{S}}\|_{\mathrm{Lip}} \le \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \|x\| \, \|l_y\|_{\mathrm{Lip}} + \frac{1}{m} \sum_{i=1}^{m} \|x_i\| \, \|l_{y_i}\|_{\mathrm{Lip}}. \tag{4.2}$$

If we assume for simplicity that the distribution is bounded, $\Pr_{(x,y)\sim\mathcal{D}}(\|x\| \le \mathrm{C}) = 1$, and that $\|l_y\|_{\mathrm{Lip}} \le \mathrm{M}$ for each y, then $\mathrm{L}_{\mathcal{D}} - \mathrm{L}_{\mathrm{S}}$ is guaranteed to be $(2\mathrm{CM})$-Lipschitz.

4.1.1 *Lipschitz constants of scalar losses*

For logistic regression (and other problems), we can compute the Lipschitz constant of $l_y^{log} : \mathbb{R} \to \mathbb{R}$ with a little calculus:

Lemma 4.4. *Let $\mathcal{X} \subseteq \mathbb{R}$ be a connected set. If a function $f : \mathcal{X} \to \mathbb{R}$ is differentiable everywhere on the interior of $\mathcal{X}$, $\|f\|_{\mathrm{Lip}} = \sup_{x\in\mathcal{X}} |f'(x)|$.*

Proof. We apply the fundamental theorem of calculus: assuming without loss of generality that $x' \ge x$,

$$|f(x') - f(x)| = \left| \int_x^{x'} f'(t) \, \mathrm{d}t \right| \le \int_x^{x'} |f'(t)| \, \mathrm{d}t \le \int_x^{x'} \left(\sup_{s\in\mathcal{X}} |f'(s)| \right) \mathrm{d}t = \left(\sup_{s\in\mathcal{X}} |f'(s)| \right) |x' - x|,$$

showing $\|f\|_{\mathrm{Lip}} \le \sup_{x\in\mathcal{X}} |f'(x)|$. Equality follows by considering an x with $|f'(x)|$ arbitrarily close to the supremum and looking at $|f(x + \varepsilon) - f(x)|$. □

We won't need this now, but it's worth noting that if $\mathcal{X} \subseteq \mathbb{R}^d$ is convex and $f : \mathcal{X} \to \mathbb{R}$ is everywhere differentiable, the exact same proof with a directional derivative from x to x' gives that $\|f\|_{\mathrm{Lip}} = \sup_{x\in\mathcal{X}} \|\nabla f(x)\|$.

Lemma 4.5. *For any $|y| \le 1$, $\|l_y^{log}\|_{\mathrm{Lip}} \le 1$.*

Proof. l_y^{log} is differentiable everywhere on $\mathbb{R}$, and so using Lemma 4.4,

$$\begin{aligned} \left| \frac{\mathrm{d}}{\mathrm{d}\hat{y}} l_y^{\log}(\hat{y}) \right| &= \left| \frac{\mathrm{d}}{\mathrm{d}\hat{y}} \log(1 + \exp(-y\hat{y})) \right| = \left| \frac{1}{1 + \exp(-y\hat{y})} \exp(-y\hat{y}) \, (-y) \right| \\ &= \left| \frac{\exp(-y\hat{y})}{1 + \exp(-y\hat{y})} \cdot \frac{\exp(y\hat{y})}{\exp(y\hat{y})} \right| |-y| = \left| \frac{1}{1 + \exp(y\hat{y})} \right| |y| \le 1. \end{aligned}$$ □

4.2 SET COVERING

The goal of all of this was to say that $L_{\mathcal{D}} - L_S$ is smooth and so, if everything in $\mathcal{H}$ is near something in $\mathcal{H}_0$, then $L_{\mathcal{D}} - L_S$ can't be too much bigger on $\mathcal{H}$ than it is on $\mathcal{H}_0$. Now the question is: how big does $\mathcal{H}_0$ have to be?

Lemma 4.6. *Let $\mathcal{X}$ be a normed vector space with finite dimension d, and ρ the metric induced by its norm. Let $U \subseteq \mathcal{X}$ be such that there is some $o \in \mathcal{X}$ with $\sup_{u \in U} \rho(o, u) \le R$, and let $\eta \in (0, R)$. Then there exists a $T \subseteq U$ with $|T| \le (3R/\eta)^d$ such that for all $u \in U$, there is a $t \in T$ with $\rho(t, u) \le \eta$.*

This is proved in Section 4.4 based on comparing volumes. The $\mathbb{R}^d$ case is a relatively straightforward result that you can definitely directly understand, it just doesn't really have anything to do with the rest of the course so we won't go through it; the general case uses the exact same structure, but might require taking a bit of measure theory on faith.

4.3 PUTTING IT ALL TOGETHER

We now have all the tools we need for the following result about linear models with bounded Lipschitz losses.

Proposition 4.7. *Let $h_w(x) = w \cdot x$ and $\mathcal{H} \subseteq \{h_w : \|w\| \le B\}$ for some $B > 0$. Consider a loss $\ell(h, (x, y)) = l_y(h(x))$ for functions $l_y : \mathbb{R} \to \mathbb{R}$ which each have Lipschitz constant at most M and are bounded in $[a, b]$. Assume that $\|x\| \le C$ almost surely under $\mathcal{D}$. Then, with probability at least $1 - \delta$,*

$$\sup_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - L_S(h) \le \frac{1}{\sqrt{2m}} \left[BCM + (b - a)\sqrt{\log \frac{1}{\delta} + \frac{d}{2} \log(72m)} \right].$$

This implies that for an ERM $\hat{h}_S$,

$$L_{\mathcal{D}}(\hat{h}_S) - \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \le \frac{1}{\sqrt{2m}} \left[BCM + 2(b - a)\sqrt{\log \frac{2}{\delta}} + (b - a)\sqrt{\frac{d}{2} \log(72m)} \right].$$

Proof. We'll first choose an η-cover $\mathcal{H}_0 = \{h_1, \dots, h_{N_\eta}\} \subseteq \mathcal{H}$, where η is a parameter to be set later. Then, for any $h \in \mathcal{H}$, let $\mathrm{nn}_{\mathcal{H}_0}(h) \in \arg\min_{h' \in \mathcal{H}_0} \rho(h, h')$, where again $\rho(h_w, h_v) = \|w - v\|$. Define the function $\Delta(h) := L_{\mathcal{D}}(h) - L_S(h)$ for brevity. Then

$$\begin{aligned} \sup_{h \in \mathcal{H}} \Delta(h) &= \sup_{h \in \mathcal{H}} \Delta(h) - \Delta(\mathrm{nn}_{\mathcal{H}_0}(h)) + \Delta(\mathrm{nn}_{\mathcal{H}_0}(h)) \\ &\le \sup_{h \in \mathcal{H}} \left[\Delta(h) - \Delta(\mathrm{nn}_{\mathcal{H}_0}(h)) \right] + \sup_{h \in \mathcal{H}} \Delta(\mathrm{nn}_{\mathcal{H}_0}(h)) \\ &\le \eta \|\Delta\|_{\mathrm{Lip}} + \max_{h' \in \mathcal{H}_0} \Delta(h'). \end{aligned}$$

By (4.2), $\|\Delta\|_{\mathrm{Lip}} \le 2CM$. The other term is uniform convergence over a finite hypothesis class $\mathcal{H}_0$, as in Proposition 2.2. We have $|\mathcal{H}_0| = N(\mathcal{H}, \eta)$, and so we can use Hoeffding for each element of $\mathcal{H}_0$ with a failure probability of $\delta/N(\mathcal{H}, \eta)$.

Applying Proposition 4.11, this gives that with probability at least $1 - \delta$,

$$\begin{aligned}\sup_{h \in \mathcal{H}} \Delta(h) &\le 2\mathrm{CM}\eta + (b-a)\sqrt{\frac{1}{2m}\log\frac{\mathrm{N}(\mathcal{H},\eta)}{\delta}} \\ &\le 2\mathrm{CM}\eta + (b-a)\sqrt{\frac{1}{2m}\left[\log\frac{1}{\delta} + d\log\frac{3\mathrm{B}}{\eta}\right]}.\end{aligned}$$

Now, we could try to exactly optimize the value of η, but I think we won't be able to do that analytically. Instead, let's notice that if η is $o(1/\sqrt{m})$, the first term being smaller doesn't really help in rate since the other term is $1/\sqrt{m}$ anyway – but choosing a smaller η makes the $\log\frac{1}{\eta}$ worse. Also, the dependence on η there is only in a log term, so it's probably okay-ish to choose $\eta = \alpha/\sqrt{m}$ for some $\alpha > 0$, giving us

$$\sup_{h \in \mathcal{H}} [\mathrm{L}_{\mathcal{D}}(h) - \mathrm{L}_{\mathrm{S}}(h)] \le \frac{1}{\sqrt{m}}\left[2\mathrm{CM}\alpha + \frac{b-a}{\sqrt{2}}\sqrt{\log\frac{1}{\delta} + d\log\frac{3\mathrm{B}\sqrt{m}}{\alpha}}\right].$$

Picking $\alpha = \mathrm{B}/(2\sqrt{2})$, the first result follows by $\log(6\sqrt{2m}) = \frac{1}{2}\log(36 \cdot 2m)$.

To show the ERM bound, recall from (1.5) that for any fixed $h^* \in \mathcal{H}$,

$$\mathrm{L}_{\mathcal{D}}(\hat{h}_{\mathrm{S}}) - \mathrm{L}_{\mathcal{D}}(h^*) \le \left(\mathrm{L}_{\mathcal{D}}(\hat{h}_{\mathrm{S}}) - \mathrm{L}_{\mathrm{S}}(\hat{h}_{\mathrm{S}})\right) + (\mathrm{L}_{\mathrm{S}}(h^*) - \mathrm{L}_{\mathcal{D}}(h^*)).$$

Bounding the first term by the previous result, with a failure probability of $\delta/2$, and the latter term by Hoeffding (Proposition 2.1) with the same failure probability:

$$\mathrm{L}_{\mathcal{D}}(\hat{h}_{\mathrm{S}}) - \mathrm{L}_{\mathcal{D}}(h^*) \le \frac{1}{\sqrt{2m}}\left[\mathrm{BCM} + \frac{b-a}{\sqrt{2}}\sqrt{\log\frac{2}{\delta} + \frac{d}{2}\log(72m)}\right] + (b-a)\sqrt{\frac{1}{2m}\log\frac{2}{\delta}}.$$

To make it look a little nicer, we can slightly loosen the bound with $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. The fact that the left-hand side holds for each h^*, and the right-hand side does not depend on the choice of h^*, gives the desired result. □

4.3.1 *Logistic regression in particular*

For our motivating problem of logistic regression, $\mathrm{M} = 1$, but there's one catch: we can use $a = 0$ but there isn't an "inherent" upper bound for b. Given that we know $\|x\| \le \mathrm{C}$ and $\|w\| \le \mathrm{B}$, though, we have that $|h(x)| = |w \cdot x| \le \mathrm{BC}$. Thus

$$\begin{aligned}\ell(h,(x,y)) &= \log(1 + \exp(-yh(x)) \le \log(1 + \exp(\mathrm{BC})) =: b \\ \ell(h,(x,y)) &= \log(1 + \exp(-yh(x)) \ge \log(1 + \exp(-\mathrm{BC})) =: a \\ b - a &= \log(1 + \exp(\mathrm{BC})) - \log(1 + \exp(-\mathrm{BC})) \\ &= \log\left(\frac{1 + \exp(\mathrm{BC})}{1 + \exp(-\mathrm{BC})} \times \frac{\exp(\mathrm{BC})}{\exp(\mathrm{BC})}\right) \\ &= \log\left(\frac{1 + \exp(\mathrm{BC})}{\exp(\mathrm{BC}) + 1} \times \exp(\mathrm{BC})\right) = \log\exp(\mathrm{BC}) = \mathrm{BC}. \qquad (4.3)\end{aligned}$$

Plugging into Proposition 4.7 gives us that with probability at least $1-\delta$, logistic regression with bounded-norm weights on bounded-norm data satisfies

$$\sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h) \le \frac{BC}{\sqrt{2m}}\left[1+\sqrt{\log\frac{1}{\delta}+\frac{d}{2}\log(72m)}\right] = \mathcal{O}_p\left(BC\sqrt{\frac{d\log m}{m}}\right), \tag{4.4}$$

and similarly ERM satisfies

$$L_{\mathcal{D}}(\hat{h}_S) - \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) \le \frac{BC}{\sqrt{2m}}\left[1+2\sqrt{\log\frac{2}{\delta}}+\sqrt{\frac{d}{2}\log(72m)}\right] = \mathcal{O}_p\left(BC\sqrt{\frac{d\log m}{m}}\right). \tag{4.5}$$

This machinery is called "chaining"; we probably won't cover it in class, but Wainwright [Wai19, Section 5.3.3] has a reasonable overview.

That $\sqrt{\log m}$ factor is actually unnecessary, but getting rid of it with covering number-type arguments requires some more advanced machinery. Instead, soon we'll see a simpler way to show a $\mathcal{O}_p(BC/\sqrt{m})$ rate, which drops the $\sqrt{\log m}$ factor but far more importantly also removes the dependence on d – that will also be very generally applicable.

More general versions We used the following properties about the problem:

- A bounded loss, to apply Hoeffding. This could be weakened in various ways, e.g. another kind of subgaussianity, or other ways to show concentration for a finite number of points.
- A Lipschitz loss. Some form of this is definitely necessary. You could potentially use a locally Lipschitz loss (where the constant varies through space), but then you have to be more careful in bounding (4.2) or similar.
- A covering number bound for $\mathcal{H}$. We did this for linear models, but you could anything, as long as it's in the same metric as you used for Lipschitzness in the previous part. This generality is often useful, e.g. for nonparametric $\mathcal{H}$.

4.4 Aside: proof of Lemma 4.6

To show our bound on how big $\mathcal{H}_0$ has to be, we'll use the following concepts. Here $\mathcal{X}$ is any set with metric ρ; denote its *closed balls* by $\mathcal{B}_\eta(x) = \{y \in \mathcal{X} : \rho(x,y) \le \eta\}$.

This is sometimes called an "internal cover," as opposed to an "external cover" which allows centres outside of U. *Some sources use open balls, in which case we'd allow* $\ge \eta$ *for* η*-separated sets; this doesn't really matter, but closed balls are a slightly closer fit to what we're using them for.* N^{sep} *is sometimes called the "packing number," but that also gets used for a different concept that's equivalent in normed spaces but slightly different in general metric spaces.*

Definition 4.8. An η-*cover* of a set $U \subseteq \mathcal{X}$ is a set $T \subseteq U$ such that $U \subseteq \cup_{t\in T}\mathcal{B}_\eta(t)$. The *covering number* $N^{\text{cov}}_\eta(U)$ is the size of the smallest η-cover of U.

An η-*separated set* is one where $\rho(u_i,u_j) > \eta$ for all $i \ne j$ – note that this inequality is strict. Let $N^{\text{sep}}_\eta(U)$ denote the size of the largest η-separated subset of U.

Proposition 4.9. *Any maximially sized* η*-separated subset of* U *is also an* η*-cover of* U*; hence* $N^{\text{cov}}_\eta(U) \le N^{\text{sep}}_\eta(U)$.

Proof. Let T be a maximally-sized η-separated subset of U. Suppose there were some $u \in U$ such that $\rho(u,t) > \eta$ for all $t \in T$. Then $T \cup \{u\}$ would be a larger η-separated subset of U, contradicting that T was of maximal size. Thus T is an η-cover of U. □

Proposition 4.10. *For any* $U \subseteq V$*, we have that* $N^{\text{sep}}_\eta(U) \le N^{\text{sep}}_\eta(V)$*. The same does* not *necessarily hold for* N^{cov}.

Proof. Any η-separated subset of U is also a subset of V, and so the largest η-separated subset of V can only be larger than that for U.

However, an η-cover of V may not be be a subset of U. For example, in $\mathbb{R}$, $\{0\}$ is an η-cover of the set $[-\eta,\eta]$, but there is no size-one η-cover of the set $[-\eta,0)\cup(0,\eta]$. □

These concepts are well-defined for any metric space. For our main bound of interest, we'll further need things to be in a d-dimensional *normed space*: see Definition B.8 for details, but for now, it's enough to say that both $\mathbb{R}^d$ and $\{h_w = (x \mapsto w \cdot x) : w \in \mathbb{R}^d\}$, with the metric $\rho(h_w, h_v) = \|w - v\|$, satisfy this assumption.

Proposition 4.11. *Let $\mathcal{X}$ be a normed space with dimension d, and $o \in \mathcal{X}$. If $\eta \geq \mathrm{R}$, trivially $\mathrm{N}^{\mathrm{cov}}_\eta(\mathcal{B}_\mathrm{R}(o)) = \mathrm{N}^{\mathrm{sep}}_\eta(\mathcal{B}_\mathrm{R}(o)) = 1$. Otherwise, for $\eta \in (0,\mathrm{R}]$ and finite d, we have*

$$\left(\frac{\mathrm{R}}{\eta}\right)^d \leq \mathrm{N}^{\mathrm{cov}}_\eta(\mathcal{B}_\mathrm{R}(o)) \leq \mathrm{N}^{\mathrm{sep}}_\eta(\mathcal{B}_\mathrm{R}(o)) \leq \left(\frac{2\mathrm{R}}{\eta}+1\right)^d \leq \left(\frac{3\mathrm{R}}{\eta}\right)^d.$$

We won't actually *use* the lower bound here for anything other than vague reassurance that at least this upper bound isn't too loose, but it's not much extra work.

Proof. We'll get both of the meaningful inequalities here by comparison of volumes. Formally, let vol denote the Haar measure for $\mathcal{X}$ under addition; this exists for any finite-dimensional normed space. For $\mathbb{R}^d$, this is just Lebesgue measure, which is (for "reasonable" sets) exactly the intuitive notion of volume. Any other finite-dimensional normed space is isomorphic to $\mathbb{R}^d$ and so its vol satisfies the same properties as Lebesgue measure.

For the lower bound, let T be a minimal η-cover of $\mathcal{B}_\mathrm{R}(o)$. Then we have that

$$\mathrm{vol}\left(\mathcal{B}_\mathrm{R}(o)\right) \leq \mathrm{vol}\left(\bigcup_{t\in\mathrm{T}} \mathcal{B}_\eta(t)\right) \leq \sum_{t\in\mathrm{T}} \mathrm{vol}\left(\mathcal{B}_\eta(t)\right) = \mathrm{N}^{\mathrm{cov}}_\eta(\mathcal{B}_\mathrm{R}(o))\ \mathrm{vol}\left(\mathcal{B}_\eta(o)\right),$$

where the final equality follows because volume is translation-invariant and any two balls of the same size are translations of each other. But, using the notation $a\mathrm{U} = \{au : u \in \mathrm{U}\}$ for "dilations," we have that $\mathrm{B}_\eta(o) = \frac{\mathrm{R}}{\eta}\mathrm{B}_\eta(o)$, and moreover it is well-known that d-dimensional volumes satisfy $\mathrm{vol}(a\mathrm{U}) = a^d\,\mathrm{vol}(\mathrm{U})$. Thus we've shown

$$\mathrm{N}^{\mathrm{cov}}_\eta(\mathcal{B}_\mathrm{R}(o)) \geq \frac{\mathrm{vol}(\mathcal{B}_\mathrm{R}(o))}{\mathrm{vol}(\mathcal{B}_\eta(o))} = \left(\frac{\mathrm{R}}{\eta}\right)^d.$$

For the upper bound, let T be a maximal η-separated subset of $\mathcal{B}_\mathrm{R}(o)$. This implies that the $\mathcal{B}_{\eta/2}(t)$ for each $t \in \mathrm{T}$ are disjoint, since if $\mathcal{B}_{\eta/2}(t)$ and $\mathcal{B}_{\eta/2}(t')$ contained a common point x, we would have $\rho(t,t') \leq \rho(t,x) + \rho(x,t') \leq \eta/2 + \eta/2 = \eta$. We also have, since $\mathrm{T} \subseteq \mathcal{B}_\mathrm{R}(o)$, that $\mathcal{B}_{\eta/2}(t) \subseteq \mathcal{B}_{\mathrm{R}+\eta/2}(o)$. Thus

$$\sum_{t\in\mathrm{T}} \mathrm{vol}\left(\mathcal{B}_{\eta/2}(t)\right) = \mathrm{vol}\left(\bigcup_{t\in\mathrm{T}} \mathcal{B}_{\eta/2}(t)\right) \leq \mathrm{vol}\left(\mathcal{B}_{\mathrm{R}+\eta/2}(o)\right).$$

Since the left-hand side here is equal to $\mathrm{N}^{\mathrm{sep}}_\eta(\mathcal{B}_\mathrm{R}(o))\,\mathrm{vol}\left(\mathcal{B}_{\eta/2}(o)\right)$, we have

$$\mathrm{N}^{\mathrm{sep}}_\eta(\mathcal{B}_\mathrm{R}(o)) \leq \frac{\mathrm{vol}\left(\mathcal{B}_{\mathrm{R}+\eta/2}(o)\right)}{\mathrm{vol}\left(\mathcal{B}_{\eta/2}(o)\right)} = \left(\frac{\mathrm{R}+\eta/2}{\eta/2}\right)^d = \left(\frac{2\mathrm{R}}{\eta}+1\right)^d.$$

The final inequality follows from $\eta \leq \mathrm{R}$. □

Unfortunately, for infinite-dimensional Banach spaces a Haar measure doesn't exist, and indeed the covering number is infinite [Isr15]. So, this kind of "hypothesis-covering" approach cannot work there.

We now have the bound we wanted:

Lemma 4.6. *Let $\mathcal{X}$ be a normed vector space with finite dimension d, and ρ the metric induced by its norm. Let $U \subseteq \mathcal{X}$ be such that there is some $o \in \mathcal{X}$ with $\sup_{u \in U} \rho(o, u) \leq R$, and let $\eta \in (0, R)$. Then there exists a $T \subseteq U$ with $|T| \leq (3R/\eta)^d$ such that for all $u \in U$, there is a $t \in T$ with $\rho(t, u) \leq \eta$.*

Proof. We can write our assumption about o as $\mathcal{H} \subseteq \mathcal{B}_R(o)$. Applying Propositions 4.9 to 4.11, we find

$$N_\eta^{\text{cov}}(\mathcal{H}) \leq N_\eta^{\text{sep}}(\mathcal{H}) \leq N_\eta^{\text{sep}}(\mathcal{B}_R(o)) \leq (3R/\eta)^d. \qquad \square$$

References

[Isr15] Robert Israel. *Can the ball $B(0, r_0)$ be covered with a finite number of balls of radius $< r_0$*. Mathematics Stack Exchange. April 1, 2015.

[Wai19] Martin Wainwright. *High-dimensional statistics: a non-asymptotic viewpoint*. Cambridge University Press, 2019.