Chapter 5

Rademacher Complexity Bounds for Concrete Models and Losses

In this chapter, we will instantiate Rademacher complexity for two important hypothesis classes: linear models and two-layer neural networks. In the process, we will develop margin theory and use it to bound the generalization gap for binary classifiers.

5.1 Margin theory for classification problems

5.1.1 Intuition

Assume that we are in the same setting as in the previous section. A fundamental problem we face in this setting is that we do not have a continuous loss: everything is discrete in the output space. We need to find a way to reason about the scale of the output. An example of this is logistic regression: the logistic regression model outputs a probability, and when we compare it to the outcome (0 or 1), its closeness to the true output gives us a measure of how confident we are in the prediction. $P(Y=1) = \frac{1}{1+e^{-(b_0+b_1X_{1i}+\cdots b_nX_{ni})}}$

Figure 5.1 gives similar intuition for linear classifiers. Intuitively, the black line is a "better" decision boundary than the red line because the minimum distance from any point to the black boundary is greater than the minimum distance from any point to the red boundary. In the next section, we will formalize this intuition by proving that the larger this margin is, the smaller the bound on the generalization gap is.

5.1.2 Formalizing margin theory

First, assume that the dataset $\mathcal{D} = ((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}))$ is *completely separable.* In other words, there exists some $h_\theta \in \mathcal{H}$ such that $y^{(i)} = \mathrm{sgn}(h_\theta(x^{(i)}))$ holds for all $(x^{(i)}, y^{(i)}) \in \mathcal{D}$. This is not a necessary condition for our final bound but will make the derivation cleaner.

Definition 5.1 ((Unnormalized) Margin)**.** Fix the hypothesis h_θ. The *(unnormalized) margin* for example (x, y) is defined as $\mathrm{margin}(x) = yh_\theta(x)$. Margin is only defined on examples where $\mathrm{sgn}(h_\theta(x)) = y$. (Note that $\mathrm{margin}(x) \geq 0$ because of our assumption of complete separability.)

Definition 5.2 (Minimum margin)**.** Given a dataset $\mathcal{D} = ((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}))$, the *minimum margin* over the dataset is defined as $\gamma_{\min} \triangleq \min_{i \in \{1,\ldots,|\mathcal{D}|\}} y^{(i)} h_\theta(x^{(i)})$.

$L(h) - \hat{L}(h$

Our final bound will have the form (generalization gap) $\leq f(\text{margin}, \text{parameter norm})$. This is very generic since there are many different bounds we could derive based on what margin we use. For this current setting we are using $\gamma_{\min}$, which is the minimum margin, but in other settings could use γ_{average}, which is the average margin of each point in the dataset.

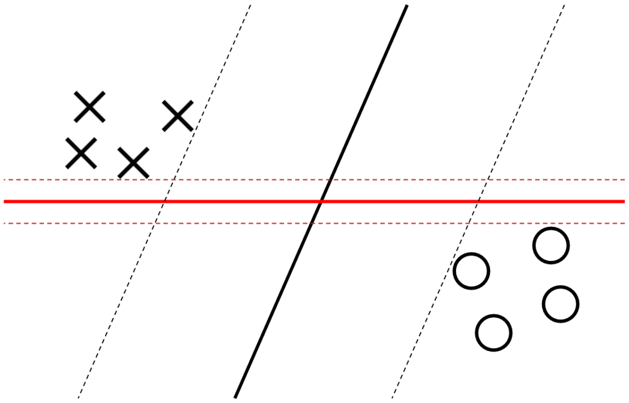

Figure 5.1: The red and black lines are two decision boundaries. The X's are positive examples and the O's are negative examples. The black line has a larger margin than the red line, and is intuitively a better classifier.

We will begin by introducing the idea of a *surrogate loss*, a loss function which approximates zero-one loss but takes the scale of the margin into account. The *margin loss* (also known as *ramp loss*) is defined as

$$\ell_\gamma(t) = \begin{cases} 0, & t \geq \gamma \\ 1, & t \leq 0 \\ 1 - t/\gamma, & 0 \leq t \leq \gamma \end{cases} \tag{5.1}$$

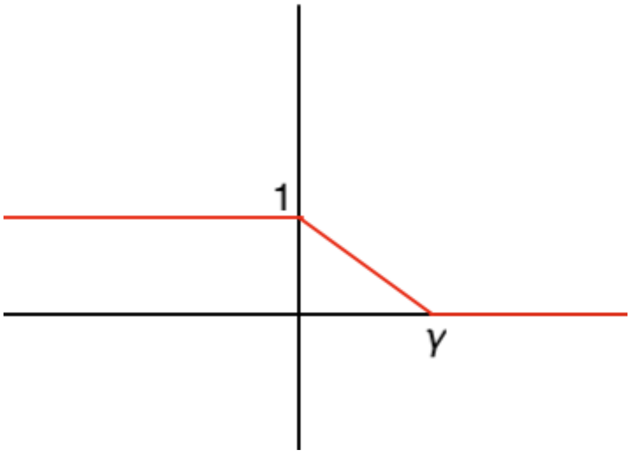

Figure 5.2: Plotted margin loss.

It is plotted in Figure 5.2. For convenience, define $\ell_\gamma((x,y),h) \triangleq \ell_\gamma(yh(x))$. We can view ℓ_γ as a continuous version of $\ell_{\text{0-1}}$ that is more sensitive to the scale of the margin on $[0,\gamma]$. Notice that $\ell_{\text{0-1}}$ is always less than or equal to the ℓ_γ when $\gamma \geq 0$, i.e.

$$\ell_{\text{0-1}}((x,y),h) = \mathbf{1}[yh(x) < 0] \leq \ell_\gamma(yh(x)) = \ell_\gamma((x,y),h) \tag{5.2}$$

holds for all $(x,y) \sim P$. Taking the expectation over (x,y) on both sides of this inequality, we see that

$$L(h) = \mathop{\mathbb{E}}_{(x,y)\sim P}[\ell_{\text{0-1}}((x,y),h)] \leq \mathop{\mathbb{E}}_{(x,y)\sim P}[\ell_\gamma((x,y),h)]\,. \tag{5.3}$$

Therefore, the population loss is bounded by the expectation of the margin loss, and so it is sufficient to bound the expectation of the margin loss in order to bound the population loss.

Define the population and empirical versions of the margin loss:

$$L_\gamma(h) = \mathop{\mathbb{E}}_{(x,y)\sim P}\left[\ell_\gamma((x,y),h)\right], \quad \hat{L}_\gamma(h) = \sum_{i=1}^{n}\left[\ell_\gamma((x^{(i)},y^{(i)}),h)\right]. \tag{5.4}$$

By Corollary 4.19, we see that with probability at least $1-\delta$,

$$L_\gamma(h) - \hat{L}_\gamma(h) \le 2R_S(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}, \tag{5.5}$$

where $\mathcal{F} = \{(x,y) \mapsto \ell_\gamma((x,y),h) \mid h \in \mathcal{H}\}$. Note that if we set $\gamma \le \gamma_{\min}$, then $\hat{L}_\gamma(h) = 0$. This follows because by definition of $\gamma_{\min}$, $y^{(i)}h(x^{(i)}) \ge \gamma_{\min}$ for any $(x^{(i)},y^{(i)}) \in \mathcal{D}$. As a result, $\ell_\gamma((x^{(i)},y^{(i)}),h) = \ell_\gamma(y^{(i)}h(x^{(i)})) = 0$ holds. Therefore, it suffices to bound $R_S(\mathcal{F})$.

We will now use *Talagrand's lemma* to bound $R_S(\mathcal{F})$ in terms of $R_S(\mathcal{H})$ to remove any dependence on the loss function from the upper bound.

$\|\phi(x) - \phi(x')\| \le k\|x-x'\|$

Lemma 5.3 (Talagrand's lemma). *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a κ-Lipschitz function. Then*

$$R_S(\phi \circ \mathcal{H}) \le \kappa R_S(\mathcal{H}), \tag{5.6}$$

where $\phi \circ \mathcal{H} = \{z \mapsto \phi(h(z)) \mid h \in \mathcal{H}\}$. (Lemma 4.29)

We can use Talagrand's lemma directly with $\phi(t) = \ell_\gamma(t)$, which is $\frac{1}{\gamma}$-Lipschitz. We can express $\mathcal{F}$ as $\mathcal{F} = \ell_\gamma \circ \mathcal{H}'$ where $\mathcal{H}' = \{(x,y) \to yh(x) \mid h \in \mathcal{H}\}$. Applying Talagrand's lemma, we see that

$$\begin{align}
R_S(\mathcal{F}) &\le \frac{1}{\gamma} R_S(\mathcal{H}') \tag{5.7}\\
&\qquad\qquad \downarrow \text{definition of } R_S \nonumber\\
&= \frac{1}{\gamma} \mathop{\mathbb{E}}_{\sigma_1,\dots,\sigma_n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i y^{(i)} h(x^{(i)})\right] \tag{5.8}\\
&\qquad\qquad \downarrow \text{by def of } \sigma_i \nonumber\\
&= \frac{1}{\gamma} \mathop{\mathbb{E}}_{\sigma_1,\dots,\sigma_n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(x^{(i)})\right] \tag{5.9}\\
&\qquad\qquad \downarrow \text{definition of } R_S \nonumber\\
&= \frac{1}{\gamma} R_S(\mathcal{H}). \tag{5.10}
\end{align}$$

Putting this all together, we have shown that for $\gamma = \gamma_{\min}$,

$$\begin{align}
L_{0\text{-}1}(h) \le L_\gamma(h) &\overset{(5.5),\,(5.10),\,\hat{L}_\gamma(h)=0}{\le} 0 + O\left(\frac{R_S(\mathcal{H})}{\gamma}\right) + \widetilde{O}\left(\sqrt{\frac{\log(2/\delta)}{2n}}\right) \tag{5.11}\\
&= O\left(\frac{R_S(\mathcal{H})}{\min_i y^{(i)}h(x^{(i)})}\right) + \widetilde{O}\left(\sqrt{\frac{\log(2/\delta)}{2n}}\right). \tag{5.12}
\end{align}$$

In other words, for training data of the form $S = \{(x^{(i)},y^{(i)})\}_{i=1}^{n} \subset \mathbb{R}^d \times \{-1,1\}$, a hypothesis class $\mathcal{H}$ and 0-1 loss, we can derive a bound of the form

$$\text{generalization loss} \le \frac{2R_S(\mathcal{H})}{\gamma_{\min}} + \text{low-order term}, \tag{5.13}$$

where $\gamma_{\min}$ is the minimum margin achievable on S over those hypotheses in $\mathcal{H}$ that separate the data, and $R_S(\mathcal{H})$ is the empirical Rademacher complexity of $\mathcal{H}$. Such bounds state that simpler models will generalize better beyond the training data, particularly for data that is strongly separable.

Remark 5.4. Note there is a subtlety here. If we think of the dataset as random, it follows that $\gamma_{\min}$ is a random variable. Consequently, the γ we choose to define the hypothesis class is random, which is not a valid choice when thinking about Rademacher complexity! Technically we cannot apply Talagrand's lemma with a random κ (which we took to be $1/\gamma$). Also, when we use concentration inequalities, we implicitly assume that the $\ell_\gamma((x^{(i)}, y^{(i)}), h)$ are independent of each other. That is not the case if γ is dependent on the data.

We sketch out how one might address this issue below. The main idea is to do another union bound over γ. Choose a family $\Gamma = \left\{2^k : k \in [-B, B]\right\}$ for some B. Then, for every fixed $\gamma \in \Gamma$, with probability greater than $1 - \delta$,

$$L_{0\text{-}1}(h) \le \widehat{L}_\gamma(h) + O\left(\frac{R_S(\mathcal{H})}{\gamma}\right) + \widetilde{O}\left(\sqrt{\frac{\log \frac{1}{\delta}}{n}}\right). \tag{5.14}$$

Taking a union bound over all $\gamma \in \Gamma$, it further holds that for all $\gamma \in (0, B)$,

Thm 4.8

$$L_{0\text{-}1}(h) \le \widehat{L}_\gamma(h) + O\left(\frac{R_S(\mathcal{H})}{\gamma}\right) + \widetilde{O}\left(\sqrt{\frac{\log \frac{1}{\delta}}{n}}\right) + \widetilde{O}\left(\sqrt{\frac{\log B}{n}}\right). \tag{5.15}$$

Last, choose the largest $\gamma \in \Gamma$ such that $\gamma \le \gamma_{\min}$. Then, for this value of γ, our desired bound directly follows from the bound in (5.15). Namely, we have that $\widehat{L}_\gamma(h) = 0$ and $O\left(\frac{R_S(\mathcal{H})}{\gamma}\right) = O\left(\frac{R_S(\mathcal{H})}{\gamma_{\min}}\right)$. The additional term, $\widetilde{O}\left(\sqrt{\frac{\log B}{n}}\right)$, is the price exacted by the uniform convergence argument required to correct the heuristic bound given in (5.13).

5.2 Linear models

5.2.1 Linear models with weights bounded in ℓ_2 norm

We begin with the Rademacher complexity of linear models using weights with bounded ℓ_2 norm.

Theorem 5.5. *Let $\mathcal{H} = \{x \mapsto \langle w, x\rangle \mid w \in \mathbb{R}^d, \|w\|_2 \le B\}$ for some constant $B > 0$. Moreover, assume $\mathbb{E}_{x\sim P}\left[\|x\|_2^2\right] \le C^2$, where P is some distribution and $C > 0$ is a constant. Then*

$$R_S(\mathcal{H}) \le \frac{B}{n}\sqrt{\sum_{i=1}^n \left\|x^{(i)}\right\|_2^2}, \tag{5.16}$$

and

$$R_n(\mathcal{H}) \le \frac{BC}{\sqrt{n}}. \tag{5.17}$$

Generally speaking, there are two methods with which we can bound the Rademacher complexity of a model. The first method, which we used in Chapter 4 consists of discretizing the space of possible outputs from our hypothesis class, then using a union bound or covering number argument to bound the Rademacher complexity of the model. While this method is powerful and generally applicable, it yields bounds that depend on the logarithm of the cardinality of this discretized output space, which in turn depends on the number of data points n. In the proof below, we will instead use a more elegant, albeit limited technique which does not rely on discretization of the output space.

Proof. We start with the proof of (5.16). By definition,

$$R_S(\mathcal{H}) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\|w\|_2\le B} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\left\langle w, x^{(i)}\right\rangle\right] \tag{5.18}$$

$$= \frac{1}{n}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\|w\|_2\le B}\left\langle w, \sum_{i=1}^{n}\sigma_i x^{(i)}\right\rangle\right] \tag{5.19}$$

$$= \frac{B}{n}\mathop{\mathbb{E}}_{\sigma}\left[\left\|\sum_{i=1}^{n}\sigma_i x^{(i)}\right\|_2\right] \qquad (\sup_{\|w\|_2\le B}\langle w, v\rangle = B\|v\|_2) \tag{5.20}$$

$\text{Pf. } (1) \leq B\|v\|_2$ (cauchy-schwarz ineq)

$(2) \geq B\|v\|_2$ $\left(w^* = B\frac{v}{\|v\|_2}\right)$

$$\le \frac{B}{n}\sqrt{\mathop{\mathbb{E}}_{\sigma}\left[\left\|\sum_{i=1}^{n}\sigma_i x^{(i)}\right\|_2^2\right]} \qquad (\text{Jensen's ineq. for } \alpha\mapsto\alpha^2) \tag{5.21}$$

$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$, $\phi(x) = x^2$

$$= \frac{B}{n}\sqrt{\mathop{\mathbb{E}}_{\sigma}\left[\sum_{i=1}^{n}\left(\underset{\searrow 1}{\sigma_i^2}\left\|x^{(i)}\right\|_2^2 + \left\langle \sigma_i x^{(i)}, \cancel{\sum_{j\neq i}^{n}\sigma_j x^{(j)}}\right\rangle\right)\right]} \tag{5.22}$$

$\mathbb{E}[\sigma_i\sigma_j] = 0$

$$= \frac{B}{n}\sqrt{\sum_{i=1}^{n}\left\|x^{(i)}\right\|_2^2}. \qquad (\sigma_i \text{ indep. and } \mathbb{E}[\sigma_i]=0) \tag{5.23}$$

This completes the proof of (5.16) for the empirical Rademacher complexity. The bound on the average Rademacher complexity in (5.17) follows from taking the expectation of both sides to get

$$R_n(\mathcal{H}) = \mathbb{E}\left[R_S(\mathcal{H})\right] \overset{\leq}{=} \frac{B}{n}\mathbb{E}\left[\sqrt{\sum_{i=1}^{n}\left\|x^{(i)}\right\|_2^2}\right]^{\sqrt{x}} \le \frac{B}{n}\sqrt{\underset{\sqrt{n}\,C}{\sum_{i=1}^{n}\mathbb{E}\left[\left\|x^{(i)}\right\|_2^2\right]}} \le \frac{BC}{\sqrt{n}}, \tag{5.24}$$

where the first inequality is another application of Jensen's inequality, and the second follows from the assumption $\mathbb{E}_{x\sim P}\left[\|x\|_2^2\right] \le C^2$.

□

We observe that both the empirical and average Rademacher complexities scale with the upper ℓ_2-norm bound $\|w\|_2 \le B$ on the parameters w, which motivates regularizing the model. However, smaller weights in the model may reduce the margin $\gamma_{\min}$, which in turn hurts generalization according to (5.13).

Remark 5.6. Note that if we scale the data by some multiplicative $\overset{x_i \to cx_i}{\text{factor}}$, the bound on empirical Rademacher complexity $R_S(\mathcal{H})$ will scale accordingly. However, at the same time, we expect the margin to scale by the same multiplicative factor, so the bound on the generalization gap in (5.13) does not change. This lines up with our intuition that the bound should not depend on the scaling of the data.

5.2.2 Linear models with weights bounded in ℓ_1 norm

Now, we consider linear models again, except we restrict the ℓ_1-norm of the parameters and assume an ℓ_∞-norm bound on the data.

Theorem 5.7. *Let $\mathcal{H} = \left\{x \mapsto \langle w, x\rangle \mid w\in\mathbb{R}^d, \|w\|_1 \le B\right\}$ for some constant $B > 0$. Moreover, assume $\left\|x^{(i)}\right\|_\infty \le C$ for some constant $C > 0$ and all points in $S = \{x^{(i)}\}_{i=1}^n \subset \mathbb{R}^d$. Then*

$$R_S(\mathcal{H}) \le BC\sqrt{\frac{2\log(2d)}{n}}. \tag{5.25}$$

To prove the theorem, we will need Massart's lemma, which provides a bound for the Rademacher complexity of a finite hypothesis class.

Lemma 5.8 (Massart's lemma). *Suppose $\mathcal{Q} \subset \mathbb{R}^n$ is finite and contained in the ℓ_2-norm ball of radius $M\sqrt{n}$ for some constant $M > 0$, i.e.,*

$$\mathcal{Q} \subset \{v \in \mathbb{R}^n \mid \|v\|_2 \leq M\sqrt{n}\}. \tag{5.26}$$

Then, for Rademacher variables $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n) \in \mathbb{R}^n$,

$$\mathop{\mathbb{E}}_{\sigma}\left[\sup_{v \in \mathcal{Q}} \frac{1}{n}\langle \sigma, v\rangle\right] \leq M\sqrt{\frac{2\log|\mathcal{Q}|}{n}}. \tag{5.27}$$

As a corollary, if $\mathcal{F}$ is a set of real-valued functions satisfying

$$\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(z^{(i)})^2 \leq M^2, \tag{5.28}$$

over some data $S = \{z^{(i)}\}_{i=1}^n$, then

$$R_S(\mathcal{F}) \leq M\sqrt{\frac{2\log|\mathcal{F}|}{n}}, \quad \textit{and} \quad R_n(\mathcal{F}) \leq M\sqrt{\frac{2\log|\mathcal{F}|}{n}}. \tag{5.29}$$

We will not prove Massart's lemma in detail. The intuition is to use concentration inequalities to bound $\frac{1}{n}\langle \sigma, v\rangle$ for fixed v, then to use a union bound over the elements $v \in \mathcal{Q}$.

We will now prove Theorem 5.7:

Proof of Theorem 5.7. By definition,

$$\begin{aligned} R_S(\mathcal{H}) &= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\|w\|_1 \leq B} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \left\langle w, x^{(i)}\right\rangle\right] && (5.30)\\ &= \frac{1}{n}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\|w\|_1 \leq B}\left\langle w, \sum_{i=1}^{n}\sigma_i x^{(i)}\right\rangle\right] && (5.31)\\ &= \frac{B}{n}\mathop{\mathbb{E}}_{\sigma}\left[\left\|\sum_{i=1}^{n}\sigma_i x^{(i)}\right\|_\infty\right], && (5.32)\end{aligned}$$

where the last equality is because $\sup_{\|w\|_1 \leq B}\langle w, v\rangle = B\|v\|_\infty$, i.e., the ℓ_∞-norm is the dual of the ℓ_1-norm, which is a consequence of Hölder's inequality. However, the ℓ_∞-norm is difficult to simplify further. Instead, we use the fact that $\sup_{\|w\|_1 \leq 1}\langle w, v\rangle$ for any $v \in \mathbb{R}^d$ is always attained at one of the vertices max element $\mathcal{W} = \bigcup_{i=1}^{d}\{-e_i, e_i\}$, where $e_i \in \mathbb{R}^d$ is the i-th coordinate unit vector. Defining the restricted hypothesis class $\bar{\mathcal{H}} = \{x \mapsto \langle w, x\rangle \mid w \in \mathcal{W}\} \subset \mathcal{H}$, this yields

$$\begin{aligned} R_S(\mathcal{H}) &= \frac{1}{n}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\|w\|_1 \leq B}\left\langle w, \sum_{i=1}^{n}\sigma_i x^{(i)}\right\rangle\right] && (5.33)\\ &= \frac{B}{n}\mathop{\mathbb{E}}_{\sigma}\left[\max_{w \in \mathcal{W}}\left\langle w, \sum_{i=1}^{n}\sigma_i x^{(i)}\right\rangle\right] && (5.34)\\ &= BR_S(\bar{\mathcal{H}}). && (5.35)\end{aligned}$$

In particular, the model class $\bar{\mathcal{H}}$ is bounded and finite with cardinality $|\bar{\mathcal{H}}| = 2d$. This suggests using Massart's lemma to complete the proof. To do so, we need to confirm that $\bar{\mathcal{H}}$ is bounded with respect to the ℓ_2-metric. Indeed, since the inner product of $x^{(i)}$ with a coordinate vector e_j just selects the j-th coordinate of $x^{(i)}$, for any $w \in \mathcal{W}$ we have

$$\frac{1}{n}\sum_{i=1}^{n}\left\langle w, x^{(i)}\right\rangle^2 \leq \frac{1}{n}\sum_{i=1}^{n}\left\|x^{(i)}\right\|_\infty^2 \leq \frac{1}{n}\sum_{i=1}^{n} C^2 = C^2, \tag{5.36}$$

better than not selected

where the last inequality uses the assumption $\|x_i\|_\infty \le C$. So $\bar{\mathcal{H}}$ is bounded in the ℓ_2-metric and finite, thus by Massart's Lemma we have

$$R_S(\mathcal{H}) = BR_S(\bar{\mathcal{H}}) \le BC\sqrt{\frac{2\log|\bar{\mathcal{H}}|}{n}} = BC\sqrt{\frac{2\log(2d)}{n}}, \tag{5.37}$$

which completes the proof. Proof on ℓ_n-norm? □

5.2.3 Comparing the bounds for different $\mathcal{H}$

First, we note that for this hypothesis class of linear models, it is possible to obtain an upper bound proportional to $\sqrt{d/n}$ using the VC dimension, which grows quickly with the data dimension d. Our bound is better since it does not have as strong of a dependence on d, and accounts for the norms of our model parameters and the data.

In the two subsections above, we considered two different hypothesis classes of linear models, each restricting different norms. In both cases, the bound on the average Rademacher complexity depended on the product of the norm bound on the parameters w and the norm bound on each data point x. To determine which choice of hypothesis class is better, consider the bounds

$$\|w\|_2\,\|x\|_2 \quad \text{vs.} \quad \|w\|_1\,\|x\|_\infty$$

and see how they compare in different settings. We consider 3 settings here:

- Suppose w and x are random variables with w_i and x_i close to the set of values $\{-1, 1\}$. Then we have

  $$\sqrt{d}\cdot\sqrt{d} \quad \text{vs.} \quad d\cdot 1.$$

  In this case, there is no difference in using either linear hypothesis class.

- If we additionally suppose w is sparse with at most k non-zero entries, then we have

  $$\sqrt{k}\cdot\sqrt{d} \quad \text{vs.} \quad k\cdot 1.$$

  So for $d \gg k$, we have $\sqrt{kd} \gg k$ and thus ℓ_1-norm regularization leads to a better complexity bound when w is suspected to be sparse. Indeed, $\sqrt{d}\,\|x\|_\infty \approx \|x\|_2$ when the entries of x are somewhat uniformly distributed, and so in the sparse case we have

  $$\|w\|_2\,\|x\|_2 \ge \sqrt{d}\,\|w\|_2\,\|x\|_\infty \ge \|w\|_1\,\|x\|_\infty. \tag{5.38}$$

- On the other hand, if w is dense in the sense that $\|w\|_2 \approx \sqrt{d}\,\|w\|_1$ (i.e., if all entries in w are close to each other in magnitude), then

  $$\|w\|_2\,\|x\|_2 \le \frac{1}{\sqrt{d}}\,\|w\|_1\cdot\sqrt{d}\,\|x\|_\infty \le \|w\|_1\,\|x\|_\infty. \tag{5.39}$$

  In this case, it makes sense to regularize the ℓ_2-norm instead.

In practice, other multiplicative factors enter the generalization bound, so regularizing both the ℓ_1- and ℓ_2-norms of the model parameters w is preferable.

Continuing with this rough style of analysis, for the hypothesis class with restricted ℓ_2-norm, we can write the bound on the generalization gap in (5.13) as

putting all-together

$$\text{generalization loss} \lesssim \frac{\|w\|_2\,\|x\|_2}{\sqrt{n}\gamma_{\min}} + \text{low-order term}. \tag{5.40}$$

The presence of $\|w\|_2/\gamma_{\min}$ motivates both the minimum norm and the maximum margin formulations of the Support Vector Machine (SVM) problem as good methods to improve generalization performance of binary classifiers.

5.3 Two-layer neural networks

We now compute a bound for the Rademacher complexity of two-layer neural networks. Throughout this section, we use the following notation:

- $\theta = (w, U)$ are the parameters of the model with $w \in \mathbb{R}^m$ and $U \in \mathbb{R}^{m \times d}$, where m denotes the number of hidden units. We use $u_i \in \mathbb{R}^d$ to denote the i-th row of U (written as a column vector).
- $\phi(z) = \max(z, 0)$ is the ReLU activation function applied element-wise.
- $f_\theta(x) = \langle w, \phi(Ux) \rangle = w^\top \phi(Ux)$ is the model.
- $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ is the training set, with $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \mathbb{R}$.

We start with a somewhat weak bound which introduces the technical tools we need to derive tighter bounds subsequently.

Theorem 5.9. *For some constants $B_w > 0$ and $B_u > 0$, let*

$$\mathcal{H} = \left\{ f_\theta \mid \|w\|_2 \le B_w,\ \|u_i\|_2 \le B_u,\ \forall i \in \{1, 2, \dots, m\} \right\}, \tag{5.41}$$

and suppose $\mathbb{E}\left[\|x\|_2^2\right] \le C^2$. Then

$$R_n(\mathcal{H}) \le 2 B_w B_u C \sqrt{\frac{m}{n}}. \tag{5.42}$$

This bound is not ideal as it depends on the number of neurons m. Empirically, it has been found that the generalization error does *not* increase monotonically with m. As more neurons are added to the model, thereby giving it more expressive power, studies have shown that generalization is improved [Belkin et al., 2019]. This contradicts the bound above, which states that more neurons leads to worse generalization. We also note that the theorem can be generalized straightforwardly to the setting where the w and U are jointly constrained in the sense that we set $\mathcal{H} = \{f_\theta \mid \|w\|_2 \cdot (\max_i \|u_i\|_2) \le B\}$ and obtain the generalization bound $R_n(\mathcal{H}) \le 2BC\sqrt{\frac{m}{n}}$. However, the $\sqrt{m}$ dependency still exists under this formulation of $\mathcal{H}$. Nevertheless, we now derive this bound.

Proof. By definition,

$$\begin{aligned}
R_S(\mathcal{H}) &= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\theta} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \left\langle w, \phi(Ux^{(i)})\right\rangle\right] && && (5.43)\\
&= \frac{1}{n}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{U:\|u_j\|_2 \le B_u}\ \sup_{\|w\|_2\le B_w} \left\langle w, \sum_{i=1}^{n}\sigma_i \phi(Ux^{(i)})\right\rangle\right] && && (5.44)\\
&= \frac{B_w}{n}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{U:\|u_j\|_2 \le B_u} \left\|\sum_{i=1}^{n}\sigma_i \phi(Ux^{(i)})\right\|_2\right] && (\sup_{\|w\|_2\le B}\langle w, v\rangle = B\|v\|_2) && (5.45)\\
&\le \frac{B_w\sqrt{m}}{n}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{U:\|u_j\|_2 \le B_u} \left\|\sum_{i=1}^{n}\sigma_i \phi(Ux^{(i)})\right\|_\infty\right] && (\|v\|_2 \le \sqrt{m}\|v\|_\infty) && (5.46)\\
&= \frac{B_w\sqrt{m}}{n}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{U:\|u_j\|_2 \le B_u}\ \max_{1\le j\le m} \left|\sum_{i=1}^{n}\sigma_i \phi(u_j^\top x^{(i)})\right|\right] && && (5.47)\\
&= \frac{B_w\sqrt{m}}{n}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\|u\|_2 \le B_u} \left|\sum_{i=1}^{n}\sigma_i \phi(u^\top x^{(i)})\right|\right] && && (5.48)\\
&\le \frac{2B_w\sqrt{m}}{n}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\|u\|_2 \le B_u} \sum_{i=1}^{n}\sigma_i \phi(u^\top x^{(i)})\right] && \text{(by Lemma 5.12)} && (5.49)\\
&\le \frac{2B_w\sqrt{m}}{n}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\|u\|_2 \le B_u} \sum_{i=1}^{n}\sigma_i u^\top x^{(i)}\right], && && (5.50)
\end{aligned}$$

where the last inequality follows by applying the contraction lemma (Talagrand's lemma) and observing that the ReLU function is 1-Lipschitz. (Observe that the expectation in (5.49) is the Rademacher complexity for $\{x \mapsto \phi(u^\top x) \mid \|u\|_2 \le B_u\}$: this is the family that we are applying the contraction lemma to.)

We now observe that the expectation in (5.50) is the Rademacher complexity of the family of linear models $\{x \mapsto \langle u, x\rangle \mid \|u\|_2 \le B_u\}$. Thus, applying Theorem ~~5.7~~ 5.5 yields

$$R_S(\mathcal{H}) \le \frac{2B_w\sqrt{m}}{n} B_u \sqrt{\sum_{i=1}^{n}\left\|x^{(i)}\right\|_2^2}. \tag{5.51}$$

Taking the expectation of both sides and using similar steps to those in the proof of Theorem 5.7 gives us

$$\begin{aligned}
R_n(\mathcal{H}) &= \mathbb{E}\left[R_S(\mathcal{H})\right] && (5.52)\\
&\le \frac{2B_wB_u\sqrt{m}}{n}\mathbb{E}\left[\sqrt{\sum_{i=1}^{n}\left\|x^{(i)}\right\|_2^2}\right] && (5.53)\\
&\le \frac{2B_wB_u\sqrt{m}}{n} C\sqrt{n} && (5.54)\\
&= 2B_wB_uC\sqrt{\frac{m}{n}}, && (5.55)
\end{aligned}$$

which completes the proof.

□

This upper bound is undesirable since it grows with the number of neurons m, contradicting empirical observations of the generalization error decreasing with m.

5.3.1 Refined bounds

Next, we look at a finer bound that results from defining a new complexity measure. A recurring theme in subsequent proofs will be the functional invariance of two-layer neural networks under a class of rescaling transformations. The key ingredient will be the *positive homogeneity* of the ReLU function, i.e.

$$\alpha\phi(x) = \phi(\alpha x) \qquad \forall \alpha > 0. \tag{5.56}$$

This implies that for any $\lambda_i > 0$ $(i = 1, \dots, m)$, the transformation $\theta = \{(w_i, u_i)\}_{1\le i\le m} \mapsto \theta' = \{(\lambda_i w_i, u_i/\lambda_i)\}_{1\le i\le m}$ has no net effect on the neural network's functionality (i.e. $f_\theta = f_{\theta'}$) since

$$w_i \cdot \phi\left(u_i^\top x^{(i)}\right) = (\lambda_i w_i) \cdot \phi\left(\left(\frac{u_i}{\lambda_i}\right)^\top x^{(i)}\right). \tag{5.57}$$

In light of this, we devise a new complexity measure $C(\theta)$ that is also invariant under such transformations and use it to prove a better bound for the Rademacher complexity. This positive homogeneity property is absent in the complexity measure used in the hypothesis class (5.41) of Theorem 5.9.

Theorem 5.10. Let $C(\theta) = \sum_{j=1}^m |w_j| \left\|u_j\right\|_2$, *and for some constant* $B > 0$ *consider the hypothesis class*

$$\mathcal{H} = \{f_\theta \mid C(\theta) \le B\}. \quad \text{joint bound.} \tag{5.58}$$

If $\left\|x^{(i)}\right\|_2 \le C$ *for all* $i \in \{1, \dots, n\}$, *then*

$$R_S(\mathcal{H}) \le \frac{2BC}{\sqrt{n}}. \tag{5.59}$$

Remark 5.11. Compared to Theorem 5.9, this bound does not explicitly depend on the number of neurons m. Thus, it is possible to use more neurons and still maintain a tight bound if the value of the new complexity measure $C(\theta)$ is reasonable. In contrast, the bound of Theorem 5.9 explicitly grows with the total number of neurons. In fact, Theorem 5.10 is strictly stronger than Theorem 5.9 as elaborated below. Note that

$$\begin{aligned} \sum |w_j| \|u_j\|_2 &\le \left(\sum |w_j|^2\right)^{1/2} \left(\sum \|u_j\|_2^2\right)^{1/2} && \text{(by Cauchy-Schwarz inequality)} \\ &\le \|w\|_2 \cdot \sqrt{m} \cdot \max_j \|u_j\|_2 && (5.60) \end{aligned}$$

Therefore, if we consider $\mathcal{H}^1 = \{f_\theta \mid \sum |w_j| \|u_j\|_2 \le B'\}$ and $\mathcal{H}^2 = \{f_\theta \mid \|w\|_2 \cdot \sqrt{m} \cdot \max_j \|u_j\|_2 \le B'\}$, then either Theorem 5.10 on $\mathcal{H}^1$ or Theorem 5.9 on $\mathcal{H}^2$ gives the same generalization bound $O(B'/\sqrt{n})$, but $\mathcal{H}^1 \supset \mathcal{H}^2$.

Moreover, Theorem 5.10 is stronger as we have more neurons—this is because the hypothesis class $\mathcal{H}$ as defined in (5.58) is bigger as m increases. Because of this, it's possible to obtain a generalization guarantee that decreases as m increases, as shown in Section 5.4.2.

Proof of Theorem 5.10. Due to the positive homogeneity of the ReLU function ϕ, it will be useful to define the ℓ_2-normalized weight vector $\bar{u}_j \triangleq u_j/\|u_j\|_2$ so that $\phi\left(u_j^\top x\right) = \|u_j\|_2 \cdot \phi(\bar{u}_j^\top x)$. The empirical Rademacher complexity satisfies

$$\begin{aligned} R_S(\mathcal{H}) &= \frac{1}{n} \mathop{\mathbb{E}}_{\sigma}\left[\sup_\theta \sum_{i=1}^n \sigma_i f_\theta\left(x^{(i)}\right)\right] && (5.61) \\ &= \frac{1}{n} \mathop{\mathbb{E}}_{\sigma}\left[\sup_\theta \sum_{i=1}^n \sigma_i \left[\sum_{j=1}^m w_j \phi\left(u_j^T x^{(i)}\right)\right]\right] && \text{(by dfn of } f_\theta\text{)} \quad (5.62) \\ &= \frac{1}{n} \mathop{\mathbb{E}}_{\sigma}\left[\sup_\theta \sum_{i=1}^n \sigma_i \left[\sum_{j=1}^m w_j \|u_j\|_2 \phi\left(\bar{u}_j^T x^{(i)}\right)\right]\right] && \text{(by positive homogeneity of } \phi\text{)} \quad (5.63) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\theta} \sum_{j=1}^{m} w_j \|u_j\|_2 \left[\sum_{i=1}^{n} \sigma_i \phi\left(\bar{u}_j^T x^{(i)}\right) \right] \right] && (5.64) \\
&\le \frac{1}{n} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\theta} \sum_{j=1}^{m} |w_j| \|u_j\|_2 \max_{k \in [n]} \left| \sum_{i=1}^{n} \sigma_i \phi\left(\bar{u}_k^T x^{(i)}\right) \right| \right] \qquad \left(\text{because } \sum_j \alpha_j \beta_j \le \sum_j |\alpha_j| \max_k |\beta_k| \right) && (5.65) \\
&\le \frac{B}{n} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\theta=(w,U)} \max_{k \in [n]} \left| \sum_{i=1}^{n} \sigma_i \phi\left(\bar{u}_k^T x^{(i)}\right) \right| \right] \qquad (\text{because } C(\theta) \le B) && (5.66) \\
&= \frac{B}{n} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\bar{u}:\|\bar{u}\|_2=1} \left| \sum_{i=1}^{n} \sigma_i \phi\left(\bar{u}^T x^{(i)}\right) \right| \right] && (5.67) \\
&\le \frac{B}{n} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\bar{u}:\|\bar{u}\|_2\le 1} \left| \sum_{i=1}^{n} \sigma_i \phi\left(\bar{u}^T x^{(i)}\right) \right| \right] && (5.68) \\
&\le \frac{2B}{n} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\bar{u}:\|\bar{u}\|_2\le 1} \sum_{i=1}^{n} \sigma_i \phi\left(\bar{u}^T x^{(i)}\right) \right] \qquad (\text{see Lemma 5.12}) && (5.69) \\
&= 2B R_S(\mathcal{H}'), && (5.70)
\end{aligned}
$$

where $\mathcal{H}' = \left\{x \mapsto \phi(\bar{u}^\top x) : \bar{u} \in \mathbb{R}^d, \|\bar{u}\|_2 \le 1\right\}$. By Talagrand's lemma, since ϕ is 1-Lipschitz, $R_S(\mathcal{H}') \le R_S(\mathcal{H}'')$ where $\mathcal{H}'' = \left\{x \mapsto \bar{u}^\top x : \bar{u} \in \mathbb{R}^d, \|\bar{u}\|_2 \le 1\right\}$ is a linear hypothesis space. Using $R_S(\mathcal{H}'') \le \frac{C}{\sqrt{n}}$ by Theorem 5.5 then concludes the proof.

□

We complete the proof by deriving the Lemma 5.12 used in the second-to-last inequality. Notably, the lemma's assumption holds in the current context, since

$$
\sup_{\theta} \langle \sigma, f_\theta(x) \rangle = \sup_{\bar{u}:\|\bar{u}\|_2 \le 1} \sum_{i=1}^{n} \sigma_i \phi\left(\bar{u}^\top x^{(i)}\right) \ge 0. \tag{5.71}
$$

since one can take $\bar{u} = 0$ for any $\sigma = (\sigma_1, \ldots, \sigma_n)$.

Lemma 5.12. *Let $\sigma = (\sigma_1, ..., \sigma_n)$ and $f_\theta(x) = \left(f_\theta\left(x^{(1)}\right), ..., f_\theta\left(x^{(n)}\right)\right)$. Suppose that for any $\sigma \in \{\pm 1\}^n$, $\sup_\theta \langle \sigma, f_\theta(x) \rangle \ge 0$. Then,*

$$
\mathbb{E}_\sigma \left[\sup_\theta |\langle \sigma, f_\theta(x) \rangle| \right] \le 2\mathbb{E}_\sigma \left[\sup_\theta \langle \sigma, f_\theta(x) \rangle \right]. \tag{5.72}
$$

Proof. Letting ϕ be the ReLU function, the lemma's assumption implies that $\sup_\theta \phi\left(\langle \sigma, f_\theta(x) \rangle\right) = \sup_\theta \langle \sigma, f_\theta(x) \rangle$ for any $\sigma \in \{\pm 1\}^n$. Observing that $|z| = \phi(z) + \phi(-z)$,

$$
\begin{aligned}
\sup_\theta |\langle \sigma, f_\theta(x) \rangle| &= \sup_\theta \left[\phi\left(\langle \sigma, f_\theta(x) \rangle\right) + \phi\left(\langle -\sigma, f_\theta(x) \rangle\right)\right] && (5.73) \\
&\le \sup_\theta \phi\left(\langle \sigma, f_\theta(x) \rangle\right) + \sup_\theta \phi\left(\langle -\sigma, f_\theta(x) \rangle\right) && (5.74) \\
&= \sup_\theta \left\langle \sigma, f_\theta(x) \right\rangle + \sup_\theta \left\langle -\sigma, f_\theta(x) \right\rangle. && (5.75)
\end{aligned}
$$

Taking the expectation over σ (and noting that $\sigma \overset{d}{=} -\sigma$), we get the desired conclusion. □

5.4 More implications and discussions on two-layer neural nets

In this section, we discuss practical implications of the refined neural network bound.