Chapter 6

Data-dependent Generalization Bounds for Deep Nets

In Theorem 5.20, we proved the following bound on the Rademacher complexity of deep neural networks:

$$R_S(\mathcal{F}) \le \prod_{i=1}^{r} \|W_i\|_{\text{op}} \cdot \text{poly}(\|W_1\|, \ldots, \|W_r\|). \tag{6.1}$$

This bound, however, suffers from multiple deficiencies. In particular, it grows exponentially in the depth, r, of the network and $\|W_i\|_{\text{op}}$ measures the worst-case Lipschitz-ness of the network layers over the input space.

In this section, we obtain a tighter generalization bound that depends upon the realized Lipschitz-ness of the model on the training data. To further motivate this approach, we also note that stochastic gradient descent, i.e. the typical optimization method typically used to fit deep neural networks, prefers models that are more Lipschitz (see Chapter (TBD) for further discussion) . This preference must be realized by the model *on empirical data*, however, as no learning algorithm has access to the model's properties over the entire data space.

Ultimately, we aim to prove a tighter bound on the population loss that grows polynomially in the Lipschitz-ness of f on the empirical data. Namely, given that f is parameterized by some θ, we hope to derive a bound on the population loss at θ that is a *polynomial* function of the Lipschitz-ness of f on $x^{(1)}, \ldots, x^{(n)}$ as well as the norm of θ.

Uniform convergence with a data-dependent hypothesis class. So far in this course, given some complexity measure that we denote as $\text{comp}(\cdot)$, our uniform convergence results always appear in one of the two following forms (which are essentially equivalent). Namely, with high probability,

$$\forall f \in \mathcal{F}, \quad L(f) \le \frac{\text{comp}(\mathcal{F})}{\sqrt{n}} \tag{I} \tag{6.2}$$

$$\forall f, \quad L(f) \le \frac{\text{comp}(f)}{\sqrt{n}} \tag{II} \tag{6.3}$$

Remark 6.1. Most of the results we have obtained so far are of type I, e.g. with $\text{comp}(\mathcal{F})/\sqrt{n} = R_n(\mathcal{F})$. We obtain results of type II by considering a restricted set of functions $\mathcal{F}_C = \{f : \text{comp}(f) \le C\}$. We then apply a type I bound to $\mathcal{F}_C$ and take a union bound over all C. Therefore, these two type of bounds are essentially equivalent (up to a small additive factor difference due to the additional union bound over the choices of C.)

Note, however, that neither of these approaches produce bounds that depend upon the data. By contrast, in the sequel, we will derive a new *data-dependent* generalization bound. These bounds state that with high

probability over the choice of the empirical data and, for all functions $f \in \mathcal{F}$,

$$L(f) \leq \text{comp}\left(f, \{(x^{(i)}, y^{(i)})\}_{i=1}^n\right) \tag{6.4}$$

Even though the complexity measure depends on the training data, and is thus a random variable by itself, it can be used as a regularizer which can be added to the original training loss.

Remark 6.2. Although there is no universal consensus on the type of generalization bound we should derive, we can argue that there is no way to leverage more information in a generalization bound beyond the empirical data. For example, one might try to use the input distribution P to define the complexity measure, but if we allowed ourselves access to P, we could just define $\text{comp}(f, P) = \mathbb{E}_P[f(X)]$. In some sense, defining a generalization bound using the true distribution amounts to cheating, and the dependence on the empirical data seems to be proper because the bound can still be used as a regularizer.

In this new paradigm, we can no longer take the previous approach of obtaining type I bounds and then derive a type II bound via a reduction. To see why, suppose that we have the hypothesis class

$$\mathcal{F}_C = \{f : \text{comp}(f, \{(x^{(i)}, y^{(i)})\}_{i=1}^n) \leq C)\} \tag{6.5}$$

If our complexity measure depends on the empirical data, then so does our hypothesis class $\mathcal{F}_C$, which makes $\mathcal{F}_C$ itself a random variable. However, our theorems regarding Rademacher complexity require that the hypothesis class be fixed before we ever see the empirical data.

We may hope to get around this by changing the way we think about uniform convergence. Consider the simplified case where our new complexity measure is separable, i.e.

$$\text{comp}(f, \{(x^{(i)}, y^{(i)})\}_{i=1}^n) = \sum_{i=1}^n h(f, x^{(i)}), \tag{6.6}$$

for some function g. Then we can consider an *augmented loss*:

$$\tilde{\ell}(f) = \ell(f)\mathbf{1}[h(f, x^{(i)}) \leq C)] \tag{6.7}$$

Suppose we have a region of low complexity in our existing loss function as depicted in Figure 6.1. Because this region is random, so we cannot selectively apply uniform convergence. However, we can use our new surrogate loss function $\tilde{\ell}$ in that region. By modifying the loss function in this way, we can still fix the hypothesis class ahead of time, allowing us to apply existing tools to $\tilde{\ell}(f)$. The surrogate loss was used in [Wei and Ma, 2019a] to obtain a data-dependent generalization bound, though there are possibly various other ways to define surrogate losses and apply existing uniform convergence guarantees. In the sequel, we introduce a particular surrogate "margin" that allows us to cleanly apply our previous results to a (implicitly) data-dependent hypothesis class [Wei and Ma, 2019a].

6.1 All-layer margin

We next introduce a new surrogate loss called the *all-layer margin* that can also be thought of as a surrogate margin. This loss will essentially zero out high-complexity regions so that we may focus on low-complexity regions for which we can expect small generalization gap. Note that the all-layer margin we analyze will not explicitly zero-out high-complexity regions using an indicator function, but instead implicitly takes into account some data-dependent characteristics of the model. Once we adopt this new loss function, we will be able to apply some of our earlier methods.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a classification model. Recall that the standard margin is defined as $yf(x)$, with y in $\{-1, 1\}$. We will say that $g_f(x, y)$ is a *generalized margin* if it satisfies

$$g_f(x, y) = \begin{cases} 0, & \text{if } f(x)y \leq 0 \text{ (an incorrect classification)} \\ > 0, & \text{if } f(x)y > 0 \text{ (a correct classification)} \end{cases} \tag{6.8}$$

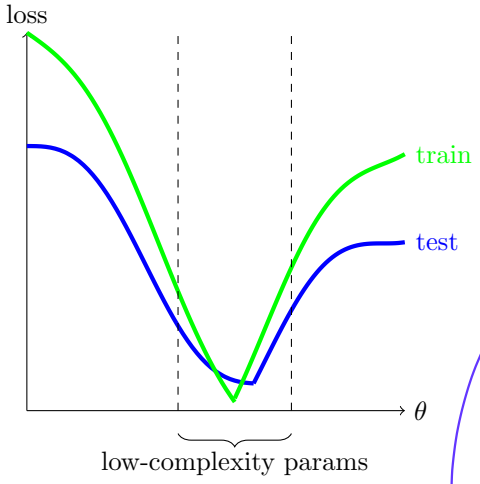

Figure 6.1: These curves depict a "low-complexity" region in parameter space. The blue curve is the unobserved test loss we aim to bound, while the green curve denotes the empirical training loss we observe. Observe that in the region of θ that we identify as being "low-complexity," the gap between the train and test losses is smaller than in the high-complexity regions.

To simplify the exposition of the machinery below, we also introduce the ∞-*covering number* $N_\infty(\epsilon, \mathcal{F})$ as the minimum cover size with respect to the metric ρ defined as the infinity-norm distance on an input domain $\mathcal{X}$:

$$\rho(f, f') \triangleq \sup_{x \in \mathcal{X}} |f(x) - f'(x)| \triangleq \|f - f'\|_\infty.^1 \tag{6.9}$$

Remark 6.3. Notice that $N_\infty(\epsilon, \mathcal{F}) \geq N(\epsilon, \mathcal{F}, L_2(P_n))$. This is because the $\rho = L_\infty(\mathcal{X})$ is a more demanding measure of error: f and f' must be close on *every* input, not just the empirical data. That is,

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - f'(x_i))^2} \leq \sup_{x \in \mathcal{X}} |f(x) - f'(x)|. \tag{6.10}$$

Lemma 6.4. *Suppose g_f is a generalized margin. Let $\mathcal{G} = \{g_f : f \in \mathcal{F}\}$. Suppose that for some R, $\log N_\infty(\epsilon, \mathcal{G}) \leq \lfloor \frac{R^2}{\epsilon^2} \rfloor$ for all $\epsilon > 0$. Then, with high probability over the randomness in the training data, for every f in $\mathcal{F}$ that correctly predicts all the training examples,*

$$L_{01} \leq \widetilde{O}\left(\frac{1}{\sqrt{n}} \cdot \frac{R}{\min_{i \in [n]} g_f(x^{(i)}, y^{(i)})}\right) + \widetilde{O}\left(\frac{1}{\sqrt{n}}\right). \tag{6.11}$$

Proof. The high-level idea of our proof is to replace $\mathcal{F}$ with $\mathcal{G}$ before repeating the standard margin theory argument from Section 5.1.2.

Let ℓ_γ be the ramp loss given in (5.1), which is 1 for negative values, 0 for values greater than γ, and a linear interpolation between 1 and 0 for values between 0 and γ. We define the surrogate loss as $\hat{L}_\gamma(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_\gamma(g_{f_\theta}(x^{(i)}, y^{(i)}))$, and the surrogate population loss as $L_\gamma(\theta) = \mathbb{E}[\ell_\gamma(g_{f_\theta}(x, y))]$. Applying Corollary 4.19, where we used the Rademacher complexity to control the generalization error, we conclude that

$$L_\gamma(\theta) - \hat{L}_\gamma(\theta) \leq R_S(\ell_\gamma \circ \mathcal{G}) + \widetilde{O}\left(\frac{1}{\sqrt{n}}\right). \tag{6.12}$$

[1] If f maps $\mathcal{X}$ to multi-dimensional outputs, we will define $\rho(f, f) \triangleq \sup_{x \in \mathcal{X}} \|f(x) - f'(x)\| \triangleq \|f - f'\|_\infty$ where the norm in $\|f(x) - f'(x)\|$ is a norm in the output space of f (which will be the Euclidean norm in this rest of this section).

[2] Recall that this is the worst dependency on ϵ that we can tolerate when converting covering number bounds to Rademacher complexity.

Next we observe that

$$\log N(\epsilon, \ell_\gamma \circ \mathcal{G}, L_2(P_n)) \le \log N(\epsilon\gamma, \mathcal{G}, L_2(P_n)) \quad \text{(Lemma 4.29)} \tag{6.13}$$

$$\le \log N_\infty(\epsilon\gamma, \mathcal{G}) \tag{6.14}$$

$$\le \left\lfloor \frac{R^2}{\epsilon^2\gamma^2} \right\rfloor = \left\lfloor \frac{C_0^2(\mathcal{G})}{\epsilon^2\gamma^2} \right\rfloor \quad \text{(by our assumption)}. \tag{6.15}$$

Then, using our results relating the log of the covering number to a bound on the Rademacher complexity (recall (4.152) and Theorem 4.28), we conclude that $R_S(\ell_\gamma \circ \mathcal{G}) \le \widetilde{O}\left(\frac{R}{\gamma\sqrt{n}}\right)$. Take $\gamma = \gamma_{\min} = \min_i g_\gamma(x^{(i)}, y^{(i)})$.[3] Using Corollary 4.19, we conclude that $\hat{L}_{\gamma_{\min}}(\theta) \le 0 + \widetilde{O}\left(\frac{R}{\sqrt{n}\cdot\gamma_{\min}}\right) + \widetilde{O}\left(\frac{1}{\sqrt{n}}\right)$, as desired. □

For which g_f can we bound the covering number? If we take $g_f(x, y) = yf(x)$, then the covering number depends on the product $\prod_i \|W_i\|_{\text{op}}$, but we originally set out to do better than this. If we have a linear model $f : w^\top x$, the normalized margin, $\frac{y \cdot w^\top x}{\|w\|}$, governs the generalization performance. But how do we normalize for more general models?

For a deep neural net, a potential normalizer is the product of the Lipschitz constants of the layers. However, we do not want to normalize by a constant that depends only on the function class, so we take a different approach. We interpret the normalized margin as the solution to the following optimization problem:

$$\begin{aligned} \min_\delta \quad & \|\delta\|_2 \\ \text{s.t.} \quad & w^\top(x+\delta)y \le 0 \end{aligned} \tag{6.16}$$

In plain English, this problem searches for the minimum perturbation that gets our data point across the boundary.

This perturbation view of the standard margin can be extended naturally to multiple layers. For the math to work, it turns out that we need to perturb all the layers. We define the *all-layer margin* as below. We will consider perturbed models $\delta = (\delta_1, \ldots, \delta_r)$, where each δ_i is a perturbation *vector* associated with the i-th layer (and it has the same dimensionality as the i-th layer activation). We incorporate these perturbations into our model in the following way (so that we can handle the scaling in a clean way):

$$h_1(x, \delta) = W_1 x + \delta_1 \cdot \|x\|_2 \tag{6.17}$$

$$h_2(x, \delta) = \sigma(W_2 h_1(x, \delta)) + \delta_2 \cdot \|h_1(x, \delta)\|_2 \tag{6.18}$$

$$\vdots$$

$$f(x, \delta) = h_r(x, \delta) = \sigma(W_r h_{r-1}(x, \delta)) + \delta_r \cdot \|h_{r-1}(x, \delta)\|_2. \tag{6.19}$$

We can then ask: what was the smallest perturbation that changed our decision? That is, let

$$m_f(x, y) \triangleq \min_{\delta_1, \ldots, \delta_r} \sqrt{\sum_{i=1}^r \|\delta_i\|_2^2} \quad \text{s.t.} \quad f(x, \delta)y \le 0, \tag{6.20}$$

i.e. the smallest perturbation that yields incorrect predictions.

Informally, $m_f(x, y)$ is a measure of how hard it is to perturb the model f. f can be hard to perturb for two reasons: f is Lipschitz (in its intermediate layers) and/or $yf(x)$ is large. In other words, the all-layer margin is a normalized version of the standard margin, normalized by the Lipschitzness of the model at the particular data point (x, y).

We now introduce our main result regarding the all-layer margin.

[3] A caveat: because γ is a random variable, proving this result rigorously requires taking a union bound over a discretized γ. We sketched out this argument more thoroughly in Remark 5.4.

Theorem 6.5. *With high probability, for all f with training error 0,*

$$L_{01}(f) \le \widetilde{O}\left(\frac{1}{\sqrt{n}} \cdot \frac{\sum_{i=1}^{r} \|W_i\|_{1,1}}{\min_{i\in[n]} m_f(x^{(i)}, y^{(i)})}\right) + \widetilde{O}\left(\frac{r}{\sqrt{n}}\right), \tag{6.21}$$

where $\|W\|_{1,1}$ is the sum of the absolute values of the entries of W.

In summary, robustness to perturbations in intermediate layers implies good generalization. We will interpret the bound, compare the bounds with previous works, and discuss further extensions in the remarks following the proofs of the theorem. (E.g, in Remark 6.8, we will argue that this bound is strictly better than the one we obtained in Theorem 5.20; in the worst case, we still have that $\frac{1}{m_f(x,y)} \le \frac{\prod \|W_i\|_{\mathrm{op}}}{f(x)}$.)

To prove this theorem, it suffices to bound $N_\infty(\epsilon, \mathcal{G})$ by $O(\frac{\sum \|W_i\|_{1,1}}{\epsilon^2})$ and apply Lemma 6.4. Towards this goal, let $\mathcal{F}_i = \{z \mapsto \sigma(W_i z) : \|W_i\|_{1,1} \le \beta_i\}$. Then, $\mathcal{F} = \mathcal{F}_r \circ \mathcal{F}_{r-1} \circ \cdots \circ \mathcal{F}_1$.

Lemma 6.6 (Decomposition Lemma). *Let $m \circ \mathcal{F}$ denote $\{m_f : f \in \mathcal{F}\}$. Then,*

$$\log N_\infty\left(\sqrt{\sum_{i=1}^{r} \epsilon_i^2}, m \circ \mathcal{F}\right) \le \sum_{i=1}^{r} \log N_\infty(\epsilon_i, \mathcal{F}_i), \tag{6.22}$$

where $N_\infty(\epsilon_i, \mathcal{F}_i)$ is defined with respect to the input domain $\mathcal{X} = \{x : \|x\|_2 \le 1\}$.

That is, we only have to find the covering number for each layer, and then we have the covering number for the (all-layer margin of the) composed function class. Notice that we bounded the covering number of $m \circ \mathcal{F}$ in the above lemma, not $\mathcal{F}$.

Then, the desired result follows directly from the preceding decomposition lemma.

Corollary 6.7. Assume that $\log N_\infty(\epsilon_i, \mathcal{F}_i) \le \left\lfloor \frac{c_i^2}{\epsilon_i^2} \right\rfloor$ for every $\mathcal{F}_i$, i.e. the function class corresponding to the i-th layer of f in Theorem 6.5. Then, by taking $\epsilon_i = \epsilon \cdot \frac{c_i}{\sqrt{\sum_i c_i^2}}$, we have that

$$\log N_\infty(\epsilon, m \circ \mathcal{F}) \le \frac{\sum_i c_i^2}{\epsilon^2}. \tag{6.23}$$

This result gives the complexity of the composed model in terms of the complexity of the layers, with each c_i given by $\|W_i\|_{1,1}$. For linear models, we can show $N_\infty(\epsilon_i, \mathcal{F}_i) \le \widetilde{O}\left(\frac{\beta_i^2}{\epsilon^2}\right)$ (where β_i is a bound on $\|W_i\|_{1,1}$), and this implies Theorem 6.5[4]. Finally, we are only left with the proof of Lemma 6.6.

Proof of Lemma 6.6. Now we will prove a limited form of the decomposition lemma for affine models: $\mathcal{F}_i = \{z \mapsto \sigma(W_i z) : \|W_i\|_{1,1} \le \beta_i\}$. There are two crucial steps to this problem. First, we will prove that $m_f(x, y)$ is 1-Lipschitz in f. That is, for all $\mathcal{F} = \mathcal{F}_r \circ \mathcal{F}_{r-1} \circ \cdots \circ \mathcal{F}_1$ and $\mathcal{F}' = \mathcal{F}'_r \circ \mathcal{F}'_{r-1} \circ \cdots \circ \mathcal{F}'_1$,

$$|m_f(x, y) - m_{f'}(x, y)| \le \sqrt{\sum_{i=1}^{r} \max_{\|x\|_2 \le 1} \|f_i(x) - f'_i(x)\|_2^2}. \tag{6.24}$$

Notice that now we are working with a clean sum of differences, with no multipliers!

Second, we construct a cover: Let $U_1, \ldots, U_r$ be $\epsilon_1, \ldots, \epsilon_r$-covers of $\mathcal{F}_1, \ldots, \mathcal{F}_r$, respectively, such that $|U_i| = N_\infty(\epsilon_i, \mathcal{F}_i)$. By definition, for all f_i in $\mathcal{F}_i$, there exists a $u_i \in U_i$ such that $\max_{\|x\|\le 1} \|f_i(x) - u_i(x)\|_2 \le \epsilon_i$. Take $U = U_r \circ U_{r-1} \circ \cdots \circ U_1 = \{u_r \circ u_{r-1} \circ \cdots \circ u_1\}$ as the cover for $m \circ \mathcal{F}$. Suppose we were given

[4] Technically, we also need to union bound over the choices of β_i, which can also be achieved following Remark 5.4.

Pf of Lemma 6.6

$f = f_r \circ \cdots \circ f_1 \in \mathcal{F}$. Let $u_r, \ldots, u_1$ be the nearest neighbors of $f_r, \ldots, f_1$. Then

$$|m_f(x,y) - m_u(x,y)| \le \sqrt{\sum_{i=1}^{r} \max_{||x|| \le 1} ||f_i(x) - u_i(x)||_2^2} \tag{6.25}$$

$$\le \sqrt{\sum_{i=1}^{r} \epsilon_i^2} \qquad \text{(by construction)}. \tag{6.26}$$

Having established the validity of our cover, we now return to our claim of 1-Lipschitz-ness stated in (6.24). By symmetry, it is sufficient to prove an upper bound for $m_{f'}(x,y) - m_f(x,y)$.

Let $\delta_1^*, \ldots, \delta_r^*$ be the optimal choices of δ in defining $m_f(x,y)$. Our goal is to turn these into a feasible solution of $m_{f'}(x,y)$, which we denote by $\hat{\delta}_1, \ldots, \hat{\delta}_r$. If this solution is feasible, we obtain the bound $m_{f'}(x,y) \le \sqrt{\sum \|\hat{\delta}_i\|_2^2}$.

Intuitively, we want to define a perturbation for f' that does the same thing as $\delta_1^*, \ldots, \delta_r^*$ for f. In plain English, $(f', \hat{\delta}_1, \ldots, \hat{\delta}_r)$ should do the same thing as $(f, \delta_1^*, \ldots, \delta_r^*)$. Recall that f has parameters $W_1, \ldots, W_r$ and f' has parameters $W_1', \ldots, W_r'$. Then, under the optimal perturbation,

$$h_1 = W_1 x + \delta_1^* \|x\|_2 \tag{6.27}$$

$$h_2 = \sigma(W_2 h_1) + \delta_2^* \|h_1\|_2 \tag{6.28}$$

$$\vdots$$

$$h_r = \sigma(W_r h_{r-1}) + \delta_r^* \|h_{r-1}\|_2 \tag{6.29}$$

We want to imitate this by perturbing f' in some way. In particular, let

$$h_1 = W_1' x + \underbrace{\delta_1^* \|x\|_2 + (W_1 - W_1')x}_{\triangleq\ \hat{\delta}_1 \|x\|_2}, \tag{6.30}$$

where the last term serves to compensate for the difference between W_1 and W_1'. Thus, $\hat{\delta}_1 \triangleq \delta_1^* + \frac{(W_1 - W_1')x}{\|x\|_2}$. We repeat this argument for every layer. Using the second layer as an example,

$$h_2 = \sigma(W_2' h_1) + \underbrace{\delta_2^* \|h_1\| + \sigma(W_2 h_1) - \sigma(W_2' h_1)}_{\triangleq\ \hat{\delta}_2 \|h_1\|_2}. \tag{6.31}$$

So, $\hat{\delta}_2 = \delta_2^* + \frac{\sigma(W_2 h_1) - \sigma(W_2' h_1)}{\|h_1\|_2}$. In general,

$$\hat{\delta}_i \triangleq \delta_i^* + \frac{\sigma(W_i h_{i-1}) - \sigma(W_i' h_{i-1})}{\|h_{i-1}\|_2} \tag{6.32}$$

Then $\hat{\delta}_1, \ldots, \hat{\delta}_r$ on f' are making the same predictions as $\delta_1, \ldots, \delta_r$ on f. Last, observe that

$$m_{f'}(x,y) \le \sqrt{\sum \|\hat{\delta}_i\|_2^2} \tag{6.33}$$

$$\le \sqrt{\sum \|\delta_i^*\|_2^2} + \sqrt{\sum_{i=1}^{r} \left(\frac{\sigma(W_i h_{i-1}) - \sigma(W_i' h_{i-1})}{\|h_{i-1}\|_2} \right)^2} \qquad \text{(Minkowski's Ineq.)}^5 \tag{6.34}$$

$$\le m_f(x,y) + \sqrt{\sum_{i=1}^{r} \max_{\|x\|_2 \le 1} (\sigma(W_i x) - \sigma(W_i' x))^2} \tag{6.35}$$

$$= m_f(x,y) + \sqrt{\sum_{i=1}^{r} \max_{\|x\|_2 \le 1} (f_i(x) - f_i'(x))^2} \tag{6.36}$$

obtained an upper bnd for $m_{f'}(x,y) - m_f(x,y)$!

Claim 2.5. *For any two compositions $F = f_k \circ \cdots \circ f_1$ and $\widehat{F} = \widehat{f}_k \circ \cdots \circ \widehat{f}_1$ and any (x,y), we have $|m_F(x,y) - m_{\widehat{F}}(x,y)| \le \sqrt{\sum_{i=1}^{k} \|f_i - \widehat{f}_i\|_{\text{op}}^2}$.*

Proof sketch. Let δ^* be the optimal choice of δ in the definition of $m_F(x,y)$. We will construct $\hat{\delta}$ such that $\|\hat{\delta}\|_2 \le \|\delta^*\|_2 + \sqrt{\sum_i \|f_i - \widehat{f}_i\|_{\text{op}}^2}$ and $\gamma(\widehat{F}(x,\hat{\delta}), y) = 0$ as follows: define $\hat{\delta}_i \triangleq \delta_i^* + \Delta_i$ for $\Delta_i \triangleq \frac{f_i(h_{i-1}(x,\delta^*)) - \widehat{f}_i(h_{i-1}(x,\delta^*))}{\|h_{i-1}(x,\delta^*)\|_2}$, where h is defined as in (2.1) with respect to the classifier F. Note that by our definition of $\|\cdot\|_{\text{op}}$, we have $\|\Delta_i\|_2 \le \|f_i - \widehat{f}_i\|_{\text{op}}$. Now it is possible to check inductively that $\widehat{F}(x,\hat{\delta}) = F(x,\delta^*)$. In particular, $\hat{\delta}$ is satisfies the misclassification constraint in the all-layer margin objective for $\widehat{F}$. Thus, it follows that $m_{\widehat{F}}(x,y) \le \|\hat{\delta}\|_2 \le \|\delta^*\|_2 + \|\Delta\|_2 \le m_F(x,y) + \sqrt{\sum_i \|f_i - \widehat{f}_i\|_{\text{op}}^2}$, where the last inequality followed from $\|\Delta_i\|_2 \le \|f_i - \widehat{f}_i\|_{\text{op}}$. With the same reasoning,we obtain $m_F(x,y) \le m_{\widehat{F}}(x,y) + \sqrt{\sum_i \|f_i - \widehat{f}_i\|_{\text{op}}^2}$, so $|m_F(x,y) - m_{\widehat{F}}(x,y)| \le \sqrt{\sum_i \|f_i - \widehat{f}_i\|_{\text{op}}^2}$. □

Note that in (6.35), constraining $\|x\|_2 \leq 1$ is equivalent to dividing by the ℓ_2-norm of x. $\square$

Remark 6.8. We can compare the above with Theorem 5.20 proven in [Bartlett et al., 2017].

$$\begin{aligned} f(x, \delta) - f(x) \leq\ & \|\delta_r\|_2 \cdot \|W_{r-1}\|_{\text{op}} \cdots \|W_1\|_{\text{op}} \\ & + \|W_r\|_{\text{op}} \cdot \|\delta_{r-1}\|_2 \cdot \|W_{r-2}\|_{\text{op}} \cdots \|W_1\|_{\text{op}} \\ & + \cdots \\ & + \|W_r\|_{\text{op}} \cdots \|W_2\|_{\text{op}} \cdot \|\delta_1\|_2. \end{aligned} \tag{6.37}$$

Ignoring minor details (e.g. dependency on r), we suppose that $y = 1$. Then, if $f(x) > 0$ and $f(x + \delta) \leq 0$, it must be the case that $\|\delta\|_2 \lesssim \frac{|f(x)|}{\prod_{i=1}^r \|W_i\|_{\text{op}}}$. This further implies that

$$\frac{m_f(x, y)}{y f(x)} \gtrsim \frac{1}{\prod_{i=1}^r \|W_i\|_{\text{op}}}. \tag{6.38}$$

Rearranging, we conclude that we have obtained a tighter bound since the inverse margin $\frac{1}{m_f(x,y)} \lesssim \frac{1}{y f(x)} \cdot \prod_{i=1}^r \|W_i\|_{\text{op}}$.

Remark 6.9. Later, we will show that SGD prefers Lipschitz solutions and Lipschitzness on data points. Implicitly, SGD seems to be maximizing the all-layer margin. Since the algorithm is (in a sense) minimizing Lipschitzness on a data point, this likely accounts for the empirically observed gap between the two bounds.

Remark 6.10. The approach we have described here is also similar to other methods in the deep learning literature. Other authors have introduced a method known as SAM (a form of sharpness-aware regularization); this method applies a perturbation to the parameter θ itself (rather than on the intermediate hidden parameters h_i.) However, these two methods are related! If we consider the (single-example) loss $\frac{\partial \ell}{\partial W_i}$, it equals $\frac{\partial \ell}{\partial h_{i+1}} \cdot h_i^\top$. Note that the norm of the term on the left is bounded by the product of the norms of the two terms of the right; this observation relates the model's Lipschitzness (with respect to the parameters) to its Lipschitzness (with respect to the hidden layer outputs.)

Remark 6.11. Finally, we can prove a more general version of this result in which we do not need to study the minimum margin of the entire dataset, and instead consider the average margin. Using this approach, we can show that the test error is bounded above by $\frac{1}{n}\sqrt{\frac{1}{n}\sum_{i=1}^n \frac{1}{m_f(x^{(i)}, y^{(i)})^2}}$ times the sum of complexities of each layer, plus a low-order term.

[5] Minkowski's inequality, which states that $\sqrt{\sum \|a_i + b_i\|_2^2} \leq \sqrt{\sum \|a_i\|_2^2} + \sqrt{\sum \|b_i\|_2^2}$. In this setting, this inequality can also be proved using Cauchy-Schwarz.