Chapter 11

Generalization

This chapter is a very brief introduction to statistical learning theory. This is not my area of expertise, so I'm just giving a tour of some key results. For more details, please see the notes by Ma (2022) which I'm heavily drawing from.

11.1 Introduction

Assume the following setup for supervised binary classification:

- An input space $\mathcal{X}$ and output space $\mathcal{Y} = \{-1, +1\}$. For example, $\mathcal{X}$ is the set of all possible images and an output of +1 means "cat" and −1 means "not a cat."

- A distribution $\pi(X, Y)$ where $X \in \mathcal{X}$, $Y \in \mathcal{Y}$. Anything with "population" in its name is related to π.

- Training data $\mathcal{D} = (x^{(0)}, y^{(0)}), \ldots, (x^{(N-1)}, y^{(N-1)}) \in \mathcal{X} \times \mathcal{Y}$. Anything with "sample" or "empirical" in its name is related to the training data.

- A hypothesis space $\mathcal{H}$ containing functions $\mathcal{X} \to \mathbb{R}$. For example, $\mathcal{H}$ contains all possible parameter settings of a neural network. If $h \in \mathcal{H}$ and $x \in \mathcal{X}$, then $h(x) > 0$ is a positive classification and $h(x) < 0$ is a negative classification.

- A loss function $\ell \colon \mathbb{R} \to \mathbb{R}^{\geq 0}$. If $h \in \mathcal{H}$, then $\ell(yh(x))$ measures how bad h's prediction on x is when the correct answer is y. For example, the 0–1 loss is
$$\ell_{0-1}(s) = \begin{cases} 0 & \text{if } s > 0 \\ 1 & \text{if } s < 0. \end{cases}$$

Adjusting the setup for multi-class classification or so that different kinds of mistakes incur different losses is possible but messy.

Given the training data $\mathcal{D}$, the learning problem is to choose the "best" hypothesis from $\mathcal{H}$. To pinpoint what "best" means, we need a few more definitions, where $h \in \mathcal{H}$:

- Population risk: $L(h) = E_{(x,y)\sim\pi}[\ell(yh(x))]$
- Empirical risk: $\hat{L}(h) = \frac{1}{N}\sum_{i=1}^{N} \ell(yh(x))$.
- Excess risk: $L(\hat{h}) - L(h^*)$ where $\hat{h}$ minimizes empirical risk and h^* minimizes population risk.

Ideally, we want to find an h that minimizes $L(h)$. But in practice, we don't have access to π, but just the training data $\mathcal{D}$. So the best we can do is to find an h that minimizes $\hat{L}(h)$, known as *empirical risk minimization (ERM).* Does ERM also minimize $L(h)$?

We use the stronger criterion of *uniform convergence*, which is to bound the difference $|L(h) - \hat{L}(h)|$ for *all* $h \in \mathcal{H}$. This criterion is related to excess risk as follows. Assume that $\hat{h}$ minimizes the empirical risk $\hat{L}$. Then

$$L(\hat{h}) - L(h^*) = \underbrace{L(\hat{h}) - \hat{L}(\hat{h})}_{\text{generalization}} + \underbrace{\hat{L}(\hat{h}) - \hat{L}(h^*)}_{\text{optimization}} + \underbrace{\hat{L}(h^*) - L(h^*)}_{\text{concentration}}. \tag{11.1}$$

The first term is generalization: minimizing this term means that the model learned from the training data will do well on unseen data. The second term is optimization (though not exactly the way we worded it in Section 1.1): minimizing this term means that the model does well on training data. By assumption, $\hat{h}$ does minimize this term, so it is less than or equal to zero. The third term is also kind of related to generalization, but h^* is not affected by training, so there's nothing here to minimize. For now, we can combine the first and third terms to get

$$L(\hat{h}) - L(h^*) \le 2 \sup_h |L(h) - \hat{L}(h)|.$$

That is, if, for *all* models, empirical risk tracks population risk, then we know that ERM generalizes well.

11.2 Finite Sets of Functions

Let's start with the simplest case where $\mathcal{H}$ is a *finite* set of functions. For example, you have 100 randomly initialized neural networks, and for some reason, you don't want to train them, but just choose the one that best fits the training data.

Theorem 11.1. *Let $\mathcal{H}$ be finite and $\ell(yh(x)) \in [0,1]$. Then for any δ, with probability at least $(1-\delta)$, we have for all $h \in \mathcal{H}$,*

$$|L(h) - \hat{L}(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2N}}. \tag{11.2}$$

The empirical risk $\hat{L}$ depends on the training data $\mathcal{D}$, which is chosen at random. We can't bound $|L(h) - \hat{L}(h)|$ for certain, but only with some probability $(1-\delta)$. We get to choose δ, but the more certainty we need, the looser the bound will be.

Proof. If $X_1, \ldots, X_N$ are bounded independent random variables, *Hoeffding's inequality* bounds the probability of their sum falling more than ϵ away from its expected value. For the special case where all the X_i are in $[0,1]$, it says

$$P(|\hat{L}(h) - L(h)| \geq \epsilon) \leq 2\exp(-2N\epsilon^2).$$

That's the probability of a single hypothesis having $|L(h) - \hat{L}(h)| \geq \epsilon$. We have $|\mathcal{H}|$ hypotheses, so the probability that any one of them has $|L(h) - \hat{L}(h)| \geq \epsilon$ is (this is called the *union bound*):

$$P(\exists h \in \mathcal{H}. |\hat{L}(h) - L(h)| \geq \epsilon) \leq 2|\mathcal{H}|\exp(-2N\epsilon^2).$$

Let this be δ and solve for ϵ:

$$\epsilon = \sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2N}}.$$

The probability that all the hypotheses in $\mathcal{H}$ have $|L(h) - \hat{L}(h)|$ less than this is $(1-\delta)$. □

11.3 Rademacher Complexity

Just about any interesting hypothesis space $\mathcal{H}$ has infinitely many functions, so the above approach won't work. Instead, we need a measure of the "complexity" of $\mathcal{H}$. Perhaps the best-known such measure is VC dimension, but Rademacher complexity is more useful for our purposes.

Definition 11.2. Let $\mathcal{F}$ be a family of functions $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and let $\mathcal{D} = ((x^{(i)}, y^{(i)}))_{i \in [N]}$. The *empirical Rademacher complexity* of $\mathcal{F}$ and $\mathcal{D}$ is

$$\mathrm{Rad}_{\mathcal{D}}(\mathcal{F}) = E_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i \in [N]} \sigma_i f(x^{(i)}, y^{(i)})\right] \tag{11.3}$$

where each σ_i is a random variable which is either +1 or −1 with equal probability.

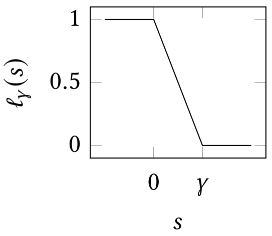
Theorem 11.3. *Let $\mathcal{F}$ be a family of functions $\mathcal{X}\times\mathcal{Y} \to \mathbb{R}$, such that for all $f \in \mathcal{F}$ and $(x, y) \in \mathcal{X}\times\mathcal{Y}$, $f(x, y) \in [0, 1]$. If $\mathcal{D} = ((x^{(i)}, y^{(i)}))_{i\in[N]}$ is drawn at random from a distribution π, then, with probability at least $(1-\delta)$, for any $f \in \mathcal{F}$,*

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i, y_i) - E_\pi[f(x, y)] \le 2\,\mathrm{Rad}_{\mathcal{D}}(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Definition 11.4. A function $f\colon \mathbb{R} \to \mathbb{R}$ is *ρ-Lipschitz continuous* (or simply *ρ-Lipschitz*) if, for all $x, y \in \mathbb{R}$, $|f(x) - f(y)| \le \rho|x - y|$.

For example, the *ramp loss* is $1/\gamma$-Lipschitz:

$$\begin{aligned} \ell_\gamma &\colon \mathbb{R} \to [0,1] \\ s &\mapsto \begin{cases} 1 & \text{if } s \le 0 \\ 1 - s/\gamma & \text{if } 0 \le s \le \gamma \\ 1 & \text{if } s \ge \gamma. \end{cases} \end{aligned}$$



Lemma 11.5 (Contraction)**.** *Let $\mathcal{F}$ be a family of functions $\mathcal{X}\times\mathcal{Y} \to \mathbb{R}$, and let $f\colon \mathbb{R} \to \mathbb{R}$ be ρ-Lipschitz continuous. For any $\mathcal{D} = ((x^{(i)}, y^{(i)}))_{i\in[N]}$,*

$$\mathrm{Rad}_{\mathcal{D}}(f \circ \mathcal{F}) \le \rho\,\mathrm{Rad}_{\mathcal{D}}(\mathcal{F}).$$

Corollary 11.6. *Let $\mathcal{H}$ be a hypothesis space, let ℓ be a ρ-Lipschitz continuous loss function such that $\ell(yh(x)) \in [0, 1]$, and let π be a distribution over $\mathcal{X}\times\mathcal{Y}$. Then, with probability at least $(1-\delta)$, for any $h \in \mathcal{H}$,*

$$\hat{L}(h) - L(h) \le 2\rho\,\mathrm{Rad}_{\mathcal{D}}(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Proof. Use Theorem 11.3 and Lemma 11.5. □

Lemma 11.5 is useful for "peeling" a function apart. Here are some other properties (though we won't use all of them):

Lemma 11.7. *Let $\mathcal{F}$ and $\mathcal{F}'$ be a families of functions $\mathcal{X}\times\mathcal{Y} \to \mathbb{R}$. For any $\mathcal{D} = ((x^{(i)}, y^{(i)}))_{i\in[N]}$,*

(a) $\mathrm{Rad}_{\mathcal{D}}(c\mathcal{F}) = |c|\,\mathrm{Rad}_{\mathcal{D}}(\mathcal{F})$

(b) $\mathrm{Rad}_{\mathcal{D}}(\mathcal{F} + \mathcal{F}') = \mathrm{Rad}_{\mathcal{D}}(\mathcal{F}) + \mathrm{Rad}_{\mathcal{D}}(\mathcal{F}')$

(c) $\mathrm{Rad}_{\mathcal{D}}(|\mathcal{F}|) = 2\,\mathrm{Rad}_{\mathcal{D}}(\mathcal{F})$

where any operation applied to $\mathcal{F}$ means to apply the operation to every function in $\mathcal{F}$.

11.4 Linear Models

We have the following bound for a hypothesis space of linear models with bounded norm (Ma, 2022).

Theorem 11.8. *For any $B > 0$, let $\mathcal{H}$ be the set of linear models, $\mathcal{H} = \{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} \mid \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\| \leq B\}$, and let $\mathcal{D} = (\mathbf{x}^{(i)})_{i \in [N]}$ and $R = \max_{i \in [N]} \|\mathbf{x}^{(i)}\|$. Then*

$$\mathrm{Rad}_{\mathcal{D}}(\mathcal{H}) \leq \frac{BR}{\sqrt{N}}.$$

Lemma 11.9.

$$E_\sigma \left[\sup_{\substack{\|\mathbf{w}\| \leq B \\ f \in \mathcal{F}}} \frac{1}{N} \sum_{i \in [N]} \sigma_i (\mathbf{w} \cdot f(\mathbf{x})) \right] \leq B \cdot E_\sigma \left[\sup_{f \in \mathcal{F}} \left\| \frac{1}{N} \sum_{i \in [N]} \sigma_i f(\mathbf{x}) \right\| \right].$$

Proof.

$$\begin{aligned}
&E_\sigma \left[\sup_{\substack{\|\mathbf{w}\| \leq B \\ f \in \mathcal{F}}} \frac{1}{N} \sum_{i \in [N]} \sigma_i (\mathbf{w} \cdot f(\mathbf{x})) \right] \\
&\quad = E_\sigma \left[\sup_{\substack{\|\mathbf{w}\| \leq B \\ f \in \mathcal{F}}} \mathbf{w} \cdot \left(\frac{1}{N} \sum_{i \in [N]} \sigma_i f(\mathbf{x}) \right) \right] \\
&\quad \leq E_\sigma \left[\sup_{\substack{\|\mathbf{w}\| \leq B \\ f \in \mathcal{F}}} \|\mathbf{w}\| \left\| \frac{1}{N} \sum_{i \in [N]} \sigma_i f(\mathbf{x}) \right\| \right] && \text{Cauchy-Schwarz} \\
&\quad = B \cdot E_\sigma \left[\sup_{f \in \mathcal{F}} \left\| \frac{1}{N} \sum_{i \in [N]} \sigma_i f(\mathbf{x}) \right\| \right].
\end{aligned}$$

□

Lemma 11.10. *Let $\mathcal{D} = (\mathbf{x}^{(i)})_{i \in [N]}$ and $R = \max_{i \in [N]} \|\mathbf{x}^{(i)}\|$. Then*

$$E_\sigma \left[\left\| \frac{1}{N} \sum_{i \in [N]} \sigma_i \mathbf{x}^{(i)} \right\| \right] \leq \frac{R}{\sqrt{N}}.$$

Proof.

$$
\begin{aligned}
&E_\sigma\left[\left\|\frac{1}{N}\sum_{i\in[N]}\sigma_i\mathbf{x}^{(i)}\right\|\right] \\
&\quad= \frac{1}{N}\sqrt{E_\sigma\left[\left\|\sum_{i\in[N]}\sigma_i\mathbf{x}^{(i)}\right\|^2\right]} && \text{Jensen's inequality} \\
&\quad= \frac{1}{N}\sqrt{E_\sigma\left[\left(\sum_{i\in[N]}\sigma_i\mathbf{x}^{(i)}\right)\cdot\left(\sum_{j\in[N]}\sigma_j\mathbf{x}^{(j)}\right)\right]} \\
&\quad= \frac{1}{N}\sqrt{E_\sigma\left[\sum_{i\in[N]}\sum_{j\in[N]}\sigma_i\mathbf{x}^{(i)}\cdot\sigma_j\mathbf{x}^{(j)}\right]} \\
&\quad= \frac{1}{N}\sqrt{E_\sigma\left[\sum_{i\in[N]}\|\mathbf{x}^{(i)}\|^2+\sum_{\substack{i,j\in[N]\\ i\neq j}}\sigma_i\mathbf{x}^{(i)}\cdot\sigma_j\mathbf{x}^{(j)}\right]} \\
&\quad= \frac{1}{N}\sqrt{\sum_{i\in[N]}\|\mathbf{x}^{(i)}\|^2} \\
&\quad\le \frac{R}{\sqrt{N}}.
\end{aligned}
$$

□

Proof of Theorem 11.8.

$$
\begin{aligned}
\mathrm{Rad}_{\mathcal{D}}(\mathcal{F}) &= E_\sigma\left[\sup_{\|\mathbf{w}\|\le B}\frac{1}{N}\sum_{i\in[N]}\sigma_i(\mathbf{w}\cdot\mathbf{x}^{(i)})\right] && \text{def. of Rad} \\
&\le B\cdot E_\sigma\left[\left\|\frac{1}{N}\sum_{i\in[N]}\sigma_i\mathbf{x}^{(i)}\right\|\right] && \text{Lemma 11.9} \\
&\le \frac{BR}{\sqrt{N}}. && \text{Lemma 11.10}
\end{aligned}
$$

□

Putting this together with Corollary 11.6, we obtain the generalization bound that, with probability $(1-\delta)$, for all $h \in \mathcal{H}$:

$$\hat{L}(h) - L(h) \leq \frac{2BR}{\gamma\sqrt{N}} + 3\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Recall that for the perceptron, we defined γ relative to $\|\mathbf{w}\|$; equivalently, it could have been defined as the minimum margin such that $\|\mathbf{w}\| = 1$. Accordingly, if we set $B = 1$ in the generalization bound above, it is similar to the perceptron (Theorem 2.8), except for a square root.

11.5 Two-Layer Networks

There are lots of different generalization bounds for FFNNs; here's a relatively easy one (Golowich et al., 2018; Bartlett, 2019).

If $\mathbf{W}$ is a matrix, we write $\|\mathbf{W}\|_F$ for the *Frobenius norm* of $\mathbf{W}$, which is just the usual 2-norm of $\mathbf{W}$ flattened into a vector: $\|\mathbf{W}\|_F = \sqrt{\sum_{i,j} \mathbf{W}[i,j]^2}$.

Theorem 11.11. *Let $\mathcal{F}_{L,d_{\text{out}}}$ be the class of L-layer ReLU FFNNs with output size d_{out}, in which each weight matrix has Frobenius norm at most B, and let $\mathcal{D} = (\mathbf{x}^{(i)})_{i\in[N]}$ and $R = \max_{i\in[N]} \|\mathbf{x}^{(i)}\|$. Then*

$$\operatorname{Rad}_{\mathcal{D}}(\mathcal{F}_{L,1}) \leq \frac{(2B)^L R}{\sqrt{N}}.$$

This theorem works for any activation that is *positively 1-homogeneous*, which means that for any $c > 0$, we have $\operatorname{ReLU}(cx) = c \operatorname{ReLU}(x)$.

Lemma 11.12.

$$E_\sigma\left[\sup_{\substack{f\in\mathcal{F}\\ \|\mathbf{W}\|_F\leq B}} \left\| \frac{1}{N}\sum_{i\in[N]} \sigma_i \operatorname{ReLU}(\mathbf{W}f(\mathbf{x}^{(i)}))\right]\right\| \leq 2B \cdot E_\sigma\left[\sup_{f\in\mathcal{F}} \left\| \frac{1}{N}\sum_{i\in[N]} \sigma_i f(\mathbf{x}^{(i)})\right\|\right].$$

Proof.

$$
\begin{aligned}
&E_\sigma\left[\sup_{\substack{f\in\mathcal{F}\\ \|\mathbf{W}\|_F\le B}}\left\|\frac{1}{N}\sum_{i\in[N]}\sigma_i\,\mathrm{ReLU}(\mathbf{W}f(\mathbf{x}^{(i)}))\right\|\right]\\
&\quad= E_\sigma\left[\sup_{\substack{f\in\mathcal{F}\\ \|\mathbf{W}\|_F\le B}}\frac{1}{N}\sqrt{\sum_{j\in[d]}\left(\sum_{i\in[N]}\sigma_i\,\mathrm{ReLU}(\mathbf{W}[j]\cdot f(\mathbf{x}^{(i)}))\right)^2}\right]\\
&\quad= E_\sigma\left[\sup_{\substack{f\in\mathcal{F}\\ \|\mathbf{W}\|_F\le B}}\frac{1}{N}\sqrt{\sum_{j\in[d]}\|\mathbf{W}[j]\|^2\left(\sum_{i\in[N]}\sigma_i\,\mathrm{ReLU}\left(\frac{\mathbf{W}[j]}{\|\mathbf{W}[j]\|}\cdot f(\mathbf{x}^{(i)})\right)\right)^2}\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{by positive homogeneity}\\
&\quad= E_\sigma\left[\sup_{\substack{f\in\mathcal{F}\\ \sum_j\|\mathbf{W}[j]\|^2\le B^2}}\frac{1}{N}\sqrt{\sum_{j\in[d]}\|\mathbf{W}[j]\|^2\left(\sum_{i\in[N]}\sigma_i\,\mathrm{ReLU}\left(\frac{\mathbf{W}[j]}{\|\mathbf{W}[j]\|}\cdot f(\mathbf{x}^{(i)})\right)\right)^2}\right]\\
&\quad\le B\cdot E_\sigma\left[\sup_{\substack{f\in\mathcal{F}\\ \|\mathbf{w}\|=1}}\left|\frac{1}{N}\sum_{i\in[N]}\sigma_i\,\mathrm{ReLU}\left(\mathbf{w}\cdot f(\mathbf{x}^{(i)})\right)\right|\right]\\
&\quad\le 2B\cdot E_\sigma\left[\sup_{\substack{f\in\mathcal{F}\\ \|\mathbf{w}\|=1}}\frac{1}{N}\sum_{i\in[N]}\sigma_i\,\mathrm{ReLU}\left(\mathbf{w}\cdot f(\mathbf{x}^{(i)})\right)\right] &&\text{Lemma 11.7c}\\
&\quad\le 2B\cdot E_\sigma\left[\sup_{\substack{f\in\mathcal{F}\\ \|\mathbf{w}\|=1}}\frac{1}{N}\sum_{i\in[N]}\sigma_i(\mathbf{w}\cdot f(\mathbf{x}^{(i)}))\right] &&\text{Lemma 11.5}\\
&\quad\le 2B\cdot E_\sigma\left[\sup_{f\in\mathcal{F}}\left\|\frac{1}{N}\sum_{i\in[N]}\sigma_i f(\mathbf{x}^{(i)})\right\|\right]. &&\text{Lemma 11.9}
\end{aligned}
$$

□

Proof of Theorem 11.11. Each function in $\mathcal{F}_{L,1}$ is of the form $\mathbf{w}\cdot\mathrm{ReLU}(f(\mathbf{x}))$ where

$f \in \mathcal{F}_{L-1,d}$, $\mathbf{w} \in \mathbb{R}^d$, and $\|\mathbf{w}\| \leq B$. Then

$$\begin{aligned}
\mathrm{Rad}_{\mathcal{D}}(\mathcal{F}_{L,1}) &= E_\sigma \left[\sup_{\substack{f \in \mathcal{F}_{L-1,d} \\ \|\mathbf{w}\| \leq B}} \frac{1}{N} \sum_{i \in [N]} \sigma_i \left(\mathbf{w} \cdot f(\mathbf{x}) \right) \right] && \text{def. of Rad} \\
&\leq B \cdot E_\sigma \left[\sup_{f \in \mathcal{F}_{L-1,d}} \left\| \frac{1}{N} \sum_{i \in [N]} \sigma_i f(\mathbf{x}) \right\| \right] && \text{Lemma 11.9}
\end{aligned}$$

Then we prove the rest by induction: the base case is Lemma 11.10, and the inductive step is Lemma 11.12. □

Bibliography

Bartlett, Peter (2019). Generalization in deep networks II. Tutorial at YES workshop on "Understanding Deep Learning: Generalization, Approximation and Optimization".

Golowich, Noah, Alexander Rakhlin, and Ohad Shamir (2018). Size-independent sample complexity of neural networks. In: *Proceedings of the 31st Conference On Learning Theory (COLT)*. Vol. 75. Proceedings of Machine Learning Research. full version available as arXiv:1712.06541, pp. 297–299.

Ma, Tengyu (2022). Lecture notes for machine learning theory (CS229M/STATS214).