Elements of Statistical Learning Theory

Ju Sun[*]

December 10, 2024

Overview In practice, we evaluate the performance of any machine learning model on a test set, that is disjoint from but related to the training set. We hope that good performance on the training set can continue on to the test set. This is called *generalization*. Why is generalization possible? How to ensure a good generalization? In particular, how do factors, such as the size of the training set and the size of the hypothesis class, affect generalization? Here, we study these topics rigorously, drawing tools mostly from probability and statistics.

Contents

1 Warmup: finite $\mathcal{H}$, realizable case

1.1 A first learning setup

We will start with binary classification with input space $\mathcal{X} \subset \mathbb{R}^d$, output space $\mathcal{Y} = \{+1, -1\}$, and a finite hypothesis class $\mathcal{H}$, i.e., $|\mathcal{H}| < \infty$, that consists of functions of the form $h : \mathbb{R}^d \to \{+1, -1\}$.

[*] Department of Computer Science and Engineering, University of Minnesota at Twin Cities. Email: jusun@umn.edu.

For generalization to be possible at all, we need to make assumptions about data generation—if there is absolutely no relationship between the training and the test sets, it is hard to imagine generalization. As a first trial, we make the following assumptions about data generation.

- **Assumption A-I**: There is a *fixed but unknown* distribution $\mathcal{D}_{\mathcal{X}}$ on $\mathcal{X}$;
- **Assumption A-II**: There is a *fixed but unknown* labeling function $f_* : \mathcal{X} \to \{+1, -1\}$, so that $y = f_*(\boldsymbol{x})$ for any $\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}$;
- **Assumption A-III**: The training set $S \doteq \{(\boldsymbol{x}_i, y_i)\}_{i \in [N]}$ is generated as follows: $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ are randomly drawn, in an iid manner, from $\mathcal{D}_{\mathcal{X}}$, i.e., $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \sim_{iid} \mathcal{D}_{\mathcal{X}}$, and $y_i = f_*(\boldsymbol{x}_i) \; \forall \, i \in [N]$.

To quantify generalization, we need appropriate metrics to measure prediction performance, and also it is natural to assume that the test set follows the same (or similar) generation process as that of the training set. These considerations lead to the following quantities:

$$\widehat{\mathsf{R}}_S(f) \doteq \frac{1}{N} \sum\nolimits_{i \in [N]} \mathbb{1}\left\{f(\boldsymbol{x}_i) \neq f_*(\boldsymbol{x}_i)\right\} \qquad (\textbf{empirical risk, or training error rate}), \tag{1.1}$$

$$\mathsf{R}(f) \doteq \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}} \mathbb{1}\left\{f(\boldsymbol{x}) \neq f_*(\boldsymbol{x})\right\} \qquad (\textbf{risk, or error rate}). \tag{1.2}$$

Here, the predictor error is measured by $0/1$ loss of the form $\mathbb{1}\left\{f(\boldsymbol{x}_i) = f_*(\boldsymbol{x}_i)\right\}$, and the risk is taken with respect to the data distribution $\mathcal{D}_{\mathcal{X}}$—during test, we typically use a finite number of test points to approximate the risk; the same data distribution $\mathcal{D}_{\mathcal{X}}$ and the same labeling function f_* connect the training and the test sets.

For training, we focus on the popular *empirical risk minimization* (ERM) framework:

$$h_S \in \arg\min\nolimits_{h \in \mathcal{H}} \widehat{\mathsf{R}}_S(h), \quad (\textbf{empirical risk minimization, or ERM}) \tag{1.3}$$

where we use $\in$ but not $=$, as the minimization problem might not have a unique global solution, and in this case we allow an arbitrary global solution.

1.2 The generalization question

After the above setup effort, we are tempted to ask the natural question: is $\mathsf{R}(h_S)$ small? In particular, is it close to the best we can do, perhaps in the sense that

$$\mathsf{R}(h_S) \leq \inf\nolimits_{h \in \mathcal{H}} \mathsf{R}(h) + \varepsilon \text{ for a sufficiently small } \varepsilon > 0? \tag{1.4}$$

Another consideration is that h_S is random, as it depends on the random set S. Although we draw the elements of $S = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ iid from $\mathcal{D}_{\mathcal{X}}$, there is always a small but nonzero probability that **the training points are not sufficiently representative**. For example, suppose that $\mathcal{X}$ is the unit sphere in $\mathbb{R}^d$, and $\mathcal{D}_{\mathcal{X}}$ is the uniform distribution on the unit sphere. Then,

$$\mathbb{P}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}}[\boldsymbol{x} \text{ from the upper hemisphere}] = \mathbb{P}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}}[\boldsymbol{x} \text{ from the upper hemisphere}] = 1/2 \tag{1.5}$$

$$\Longrightarrow \mathbb{P}_{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \sim_{iid} \mathcal{D}_{\mathcal{X}}}[\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \text{ all come from the lower hemisphere}] = 1/2^N. \tag{1.6}$$

When the training set is not representative, we expect h_S returned by the ERM rule to have a high risk.

So, a sensible technical question to ask here is: how many training points do we need to guarantee $\mathsf{R}(h_S) \leq \inf_{h \in \mathcal{H}} \mathsf{R}(h) + \varepsilon$ for a sufficiently small ε with high probability. In other words, we hope to see that

$$\mathbb{P}_S[\mathsf{R}(h_S) \leq \inf\nolimits_{h \in \mathcal{H}} \mathsf{R}(h) + \varepsilon] \geq 1 - \delta, \tag{1.7}$$

for a sufficiently small $\varepsilon \in (0, 1)$ and a sufficiently small $\delta \in (0, 1)$, provided that N is sufficiently large.

1.3 A first generalization result for finite $\mathcal{H}$ with realizability assumption

To state our first result in this direction, we make one more assumption:

- **Assumption A-IV—realizability**: There is a certain $h_* \in \mathcal{H}$, so $\mathsf{R}(h_*) = 0$. This implies that $\widehat{\mathsf{R}}_S(h_*) = 0$ for any training set S of any size $N \geq 1$ (with probability one).

Note that this assumption does not imply $h_* = f_*$; it only requires $h_*(\boldsymbol{x}) = f_*(\boldsymbol{x})$ for "almost all" $\boldsymbol{x} \in \mathcal{X}$[1]—but they do not need to agree with each other for $\boldsymbol{x} \in \mathcal{X}^c$. Also, under this assumption,

$$\inf_{h \in \mathcal{H}} \mathsf{R}(h) = 0. \tag{1.8}$$

We are now to state our first generalization result.

Theorem 1.1. *For any $\varepsilon \in (0,1)$, any $\delta \in (0,1)$, any labeling function $f_* : \mathcal{X} \to \{+1,-1\}$, and any distribution $\mathcal{D}_\mathcal{X}$ on $\mathcal{X}$, if the realizability assumption holds, i.e., there exists an $h_* \in \mathcal{H}$ so that $\mathsf{R}(h_*) = 0$, every predictor h_S returned by the ERM rule over a training set $S = \{(\boldsymbol{x}_i, f_*(\boldsymbol{x}_i)\}_{i \in [N]}$, where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \sim_{iid} \mathcal{D}_\mathcal{X}$, satisfies all of the following three equivalent statements:*

i) $\mathbb{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}[\mathsf{R}(h_S) \leq \varepsilon] \geq 1 - \delta$, *provided that* $N \geq \varepsilon^{-1}\log(|\mathcal{H}|/\delta)$;

ii) $\mathbb{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left[\mathsf{R}(h_S) \leq \frac{1}{N}\log\frac{|\mathcal{H}|}{\delta}\right] \geq 1 - \delta$;

iii) $\mathbb{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}[\mathsf{R}(h_S) \leq \varepsilon] \geq 1 - |\mathcal{H}|e^{-N\varepsilon}$.

Proof. We first prove (iii), and then show the equivalence of the three statements.

We start by considering $\mathbb{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}[\mathsf{R}(h_S) \geq \varepsilon]$. Since h_S is returned by the ERM rule, we have $\widehat{\mathsf{R}}(h_S) = 0$ by the realizability assumption. Moreover, define the "bad" subset of $\mathcal{H}$ as $\mathcal{H}_B \doteq \{h \in \mathcal{H} : \mathsf{R}(h) \geq \varepsilon\}$. Then, $h_S \in \mathcal{H}_B$. So,

$$\mathsf{R}(h_S) \geq \varepsilon \Longrightarrow \widehat{\mathsf{R}}_S(h_S) = 0 \text{ and } h_S \in \mathcal{H}_B, \tag{1.9}$$

implying that

$$\begin{aligned} \mathbb{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}[\mathsf{R}(h_S) \geq \varepsilon] &\leq \mathbb{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left[\widehat{\mathsf{R}}_S(h_S) = 0 \text{ and } h_S \in \mathcal{H}_B\right] & (1.10)\\ &\leq \mathbb{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left[\exists h \in \mathcal{H}_B \text{ with } \widehat{\mathsf{R}}_S(h) = 0\right] & (1.11)\\ &= \mathbb{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left[\widehat{\mathsf{R}}_S(h_1) = 0 \text{ or } \ldots \text{ or } \widehat{\mathsf{R}}_S(h_{|\mathcal{H}_B|}) = 0\right], & (1.12) \end{aligned}$$

where $h_1, \ldots, h_{|\mathcal{H}_B|}$ are the predictors in $\mathcal{H}_B$. Now we invoke the famous union bound, which says that for any two events A_1 and A_2, $\mathbb{P}[A_1 \cup A_2] \leq \mathbb{P}[A_1] + \mathbb{P}[A_2]$, and obtain that

$$\mathbb{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left[\widehat{\mathsf{R}}_S(h_1) = 0 \text{ or } \ldots \text{ or } \widehat{\mathsf{R}}_S(h_{|\mathcal{H}_B|}) = 0\right] \leq \sum\nolimits_{b \in [|\mathcal{H}_B|]} \mathbb{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left[\widehat{\mathsf{R}}_S(h_b) = 0\right]. \tag{1.13}$$

Now, for any $b \in [|\mathcal{H}_B|]$, $\mathsf{R}(h_b) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}} \mathbb{1}\{h_b(\boldsymbol{x}) \neq f_*(\boldsymbol{x})\} = \mathbb{P}_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}}[h_b(\boldsymbol{x}) \neq f_*(\boldsymbol{x})] \geq \varepsilon$, so $\mathbb{P}_{\boldsymbol{x}\sim\mathcal{D}_\mathcal{X}}[h_b(\boldsymbol{x}) = f_*(\boldsymbol{x})] \leq 1 - \varepsilon$. For N iid points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$,

$$\mathbb{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left[\widehat{\mathsf{R}}_S(h_b) = 0\right] = \mathbb{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}[h_b(\boldsymbol{x}_i) = f_*(\boldsymbol{x}_i)\ \forall\, i \in [N]] \leq (1-\varepsilon)^N. \tag{1.14}$$

Thus,

$$\sum\nolimits_{b \in [|\mathcal{H}_B|]} \mathbb{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left[\widehat{\mathsf{R}}_S(h_b) = 0\right] \leq |\mathcal{H}_B|(1-\varepsilon)^N \leq |\mathcal{H}_B|e^{-\varepsilon N}. \quad (\text{as } 1 - z \leq e^{-z}\ \forall\, z \in \mathbb{R}) \tag{1.15}$$

[1] In probability theory, exception can hold on a zero-measure set.

Piecing together the above results, we conclude that $\mathbb{P}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}[\mathsf{R}(h_S) \ge \varepsilon] \le |\mathcal{H}_B| e^{-\varepsilon N} \le |\mathcal{H}| e^{-\varepsilon N}$, which is equivalent to iii).

Now, set $|\mathcal{H}| e^{-\varepsilon N} = \delta$, we get $N = \varepsilon^{-1} \log(|\mathcal{H}|/\delta)$, which implies i), or $\varepsilon = N^{-1} \log(|\mathcal{H}|/\delta)$, which implies ii), completing the proof. ■

Remark 1.2. *We can relax the realizability assumption as follows: for any training set S of any size $N \ge 1$, there exists an $h \in \mathcal{H}$ so that $\widehat{\mathsf{R}}_S(h) = 0$. The results in Theorem 1.1 still hold under this weaker assumption.*

Given any fixed hypothesis class $\mathcal{H}$ and hence its size $|\mathcal{H}|$, Theorem 1.1 describes the quantitative relationship between N, ε, and δ—the three equivalent statements i), ii), iii) each fixes two of the three quantities, and states the other as a function of the fixed two. This is the hallmark of statistical learning theory.

Here, ε is called the *accuracy parameter*, as it measures prediction performance. δ is called the *confidence parameter*, which measures the uncertainty of performance over the randomness of the training set.

Henceforth, when we state similar generalization results, we do not state all forms, expecting the reader to make appropriate translation when needed.

2 General learning setting with finite $\mathcal{H}$

Our previous learning setup is restrictive in many senses. The most critical one is assuming the existence of a labeling function. For example, suppose that we try to predict the gender of a person (y) from their height (x). There is no single function that maps x *deterministically* to y, as people of the same height can be of any gender, albeit with different probabilities. In this case, it is more appropriate to think of $\mathbb{P}[y|x]$, i.e., considering stochastic prediction.

2.1 A stochastic setup for binary classification

This motivates us to generalize the previous learning setup as follows. Again, consider an input space $\mathcal{X} \subset \mathbb{R}^d$, an output space $\mathcal{Y} = \{+1, -1\}$, and a hypothesis class $\mathcal{H}$, which may or may not be finite, that consists of functions of the form $h : \mathbb{R}^d \to \{+1, -1\}$.

- **Assumption B-I**: There is a *fixed but unknown* distribution $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$ on $\mathcal{X} \times \mathcal{Y}$. This joint distribution $\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$ can be decomposed as $\mathcal{D}_{\mathcal{X}}$—for generation of $\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}$, and $\mathcal{D}_{\mathcal{Y}|\mathcal{X}}$—for conditional generation of $y \sim \mathcal{D}_{\mathcal{Y}|\mathcal{X}}$;
- **Assumption B-II**: The training set $S \doteq \{(\boldsymbol{x}_i, y_i)\}_{i\in[N]}$ is generated as $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_N, y_N) \sim_{iid} \mathcal{D}_{\mathcal{X}\times\mathcal{Y}}$.

Then, we can define the empirical risk and the risk as follows, natural generalizations of those considered in Section 1.1:

$$\widehat{\mathsf{R}}_S(f) \doteq \frac{1}{N} \sum\nolimits_{i\in[N]} \mathbb{1}\{f(\boldsymbol{x}_i) \ne y_i\} \qquad (\textbf{empirical risk, or training error rate}), \tag{2.1}$$

$$\mathsf{R}(f) \doteq \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\mathcal{X}\times\mathcal{Y}}} \mathbb{1}\{f(\boldsymbol{x}) \ne y\} \qquad (\textbf{risk, or error rate}), \tag{2.2}$$

and ERM rule as

$$h_S \in \arg\min\nolimits_{h\in\mathcal{H}} \widehat{\mathsf{R}}_S(h) \quad (\textbf{empirical risk minimization, or ERM}). \tag{2.3}$$

A reasonable goal for generalization is:

$$\mathsf{R}(h_S) \le \inf\nolimits_{h\in\mathcal{H}} \mathsf{R}(h) + \varepsilon \text{ for a sufficiently small } \varepsilon > 0. \tag{2.4}$$

Note that unlike the realizable case discussed in Section 1.3 where $\inf_{h \in \mathcal{H}} \mathsf{R}(h) = 0$, here $\inf_{h \in \mathcal{H}} \mathsf{R}(h) > 0$ in general. To see this, we can consider Bayes optimal classifiers, defined as

$$h_* \in \arg\min_{h \text{ ``reasonable''}} \mathsf{R}(h). \tag{2.5}$$

Here, "reasonable" is a minimal condition that one wants to put on the function h, e.g., measurable[2],to ensure that $\mathsf{R}(h)$ is well defined. Any classifier that achieves the minimal risk is called a Bayes optimal classifier. It can be shown that decision rule

$$\arg\max_{y \in \{+1,-1\}} \mathbb{P}[y|\boldsymbol{x}] \tag{2.6}$$

is a Bayes optimal classifier, although in practice we typically cannot implement it as we do not know $\mathbb{P}[y|\boldsymbol{x}]$. The associated risk is

$$\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}} \min(\mathbb{P}[1|\boldsymbol{x}], \mathbb{P}[-1|\boldsymbol{x}]) \in [0, 1/2], \tag{2.7}$$

which can be interpreted as the *intrinsic noise* in the data.

The implication: if the intrinsic noise is positive, we cannot expect to obtain 100% prediction accuracy, no matter what $\mathcal{H}$ we choose and how large N is. **We cannot expect perfect classifiers in general**. Similar claim can be made for regression problems (see, e.g., Exercise 2.14 of [Moh18]).

2.2 Further generalization for general learning problems

Our learning setup in Section 2.1 is still restrictive: it is only for binary classification. We have numerous other machine learning frameworks, e.g., multiclass classification, regression, unsupervised learning, self-supervised learning—a core learning framework for training foundation models and large language models in the current frontier of deep learning. Can we unify them in our setup?

It is relatively easy to generalize the setup in Section 2.1 to general supervised learning, including multiclass classification and regression, by considering the general input-output space $\mathcal{X} \times \mathcal{Y}$, and any distribution on it $\mathcal{D}_{\mathcal{X} \times \mathcal{Y}}$. But to generalize further, we make the following abstraction:

$$\mathcal{X} \times \mathcal{Y} \text{ abstracted into } \mathcal{Z} \quad (\textbf{data space}), \tag{2.8}$$

$$\mathcal{D}_{\mathcal{X} \times \mathcal{Y}} \text{ abstracted into } \mathcal{D}_{\mathcal{Z}} \quad (\textbf{distribution on the data space}), \tag{2.9}$$

$$\mathbb{1}\{f(\boldsymbol{x}) \neq y\} \text{ abstracted into } \ell(f, \boldsymbol{z}). \tag{2.10}$$

Now, our assumptions on data generation are

- **Assumption C-I**: There is a *fixed but unknown* distribution $\mathcal{D}_{\mathcal{Z}}$ on the data space $\mathcal{Z}$;
- **Assumption C-II**: The training set $S \doteq \{\boldsymbol{z}_i\}_{i \in [N]}$ is generated as $\boldsymbol{z}_1, \dots, \boldsymbol{z}_N \sim_{iid} \mathcal{D}_{\mathcal{Z}}$.

Moreover, the empirical risk and the risk are defined naturally as

$$\widehat{\mathsf{R}}_S(f) \doteq \frac{1}{N} \sum_{i \in [N]} \ell(f, \boldsymbol{z}_i) \qquad (\textbf{empirical risk, or training error rate}), \tag{2.11}$$

$$\mathsf{R}(f) \doteq \mathbb{E}_{\boldsymbol{z} \sim \mathcal{D}_{\mathcal{Z}}} \ell(f, \boldsymbol{z}) \qquad (\textbf{risk, or error rate}). \tag{2.12}$$

ERM rule looks exactly like before:

$$h_S \in \arg\min_{h \in \mathcal{H}} \widehat{\mathsf{R}}_S(h) \quad (\textbf{empirical risk minimization, or ERM}). \tag{2.13}$$

Now, let us look at several examples to appreciate how general our setup is.

[2] https://en.wikipedia.org/wiki/Measurable_function

- **Multiclass classification**: $\mathcal{Z} = \mathbb{R}^d \times \{1, \dots, K\}$ (assuming a K-class problem; i.e., $\boldsymbol{z} = (\boldsymbol{x}, y)$), and $\ell(f, \boldsymbol{z}) = \mathbb{1}\{f(\boldsymbol{x}) \neq y\}$—one can choose other losses as well, such as hinge loss (as in SVM), exponential loss (as in Adaboost);
- **Regression**: $\mathcal{Z} = \mathbb{R}^d \times \mathbb{R}$ (i.e., $\boldsymbol{z} = (\boldsymbol{x}, y)$) and $\ell(f, \boldsymbol{z}) = (f(\boldsymbol{x}) - y)^2$;
- **Metric learning**: the goal is to learn a transformation f on any input $\boldsymbol{x}$, so that for pair $(\boldsymbol{x}, \boldsymbol{x}')$, $\|f(\boldsymbol{x}) - f(\boldsymbol{x}')\|_2$ matches a predefined nominal distance $d_{\boldsymbol{x},\boldsymbol{x}'}$. Here, $\mathcal{Z} = \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$, as each data point takes the form $(\boldsymbol{x}, \boldsymbol{x}', d_{\boldsymbol{x},\boldsymbol{x}'})$ and $\ell(f, \boldsymbol{z}) = \left| \|f(\boldsymbol{x}) - f(\boldsymbol{x}')\|_2 - d_{\boldsymbol{x},\boldsymbol{x}'} \right|$.

As before, a reasonable learning goal here is to ensure that

$$\mathsf{R}(\widehat{h}_S) \le \inf_{h \in \mathcal{H}} \mathsf{R}(h) + \varepsilon \tag{2.14}$$

for a sufficiently small $\varepsilon > 0$, with a sufficiently high probability.

2.3 Generalization via uniform convergence

To obtain generalization results for the general learning setup in Section 2.2, we distill a sufficient condition called *uniform convergence*.

Definition 2.1 (ε-uniform convergence)**.** *A training set S is said to ensure ε-uniform convergence with respect to $\mathcal{Z}, \mathcal{D}_{\mathcal{Z}}, \ell$, and $\mathcal{H}$ if*

$$\forall\, h \in \mathcal{H}, \; \left| \widehat{\mathsf{R}}_S(h) - \mathsf{R}(h) \right| \le \varepsilon. \tag{2.15}$$

Theorem 2.2 (Uniform convergence ensures generalization)**.** *Assume that a training set S ensures $\varepsilon/2$-uniform convergence with respect to $\mathcal{Z}, \mathcal{D}_{\mathcal{Z}}, \ell$, and $\mathcal{H}$. Then, any h_S returned by the ERM rule, i.e., $h_S \in \arg\min_{h \in \mathcal{H}} \widehat{\mathsf{R}}_S(h)$, satisfies*

$$\mathsf{R}(h_S) \le \inf_{h \in \mathcal{H}} \mathsf{R}(h) + \varepsilon. \tag{2.16}$$

Proof. We have the following chain of inequalities

$$\begin{aligned} \mathsf{R}(h_S) &\le \widehat{\mathsf{R}}_S(h_S) + \varepsilon/2 \quad (\varepsilon/2 \text{ uniform convergence}) && (2.17)\\ &\le \widehat{\mathsf{R}}_S(h_*) + \varepsilon/2 \quad (h_* \in \arg\min_{h \in \mathcal{H}} \mathsf{R}(h);\, h_S \in \arg\min_{h \in \mathcal{H}} \widehat{\mathsf{R}}_S(h)) && (2.18)\\ &\le \mathsf{R}(h_*) + \varepsilon/2 + \varepsilon/2 = \mathsf{R}(h_*) + \varepsilon \quad (\varepsilon/2 \text{ uniform convergence}), && (2.19) \end{aligned}$$

completing the proof. ■

Since the translation from uniform convergence to generalization results is mechanical as described in Theorem 2.2, people often stop once establishing uniform convergence, expecting the reader to complete the trivial translation step.

Next, we establish a general uniform convergence result for learning problems with bounded loss and finite $\mathcal{H}$. Toward this, we need a famous concentration bound called *Hoeffding's inequality*.

Recall that the law of large numbers (LLN) says that under mild conditions, sample average converges to the expectation as the size of sample grows. Specifically, suppose that $\theta_1, \dots, \theta_m$ are sampled iid from a distribution with mean μ,

$$\frac{1}{m} \sum_{j \in [m]} \theta_j \to \mu \text{ as } m \to \infty. \tag{2.20}$$

This result is asymptotic, as it only describes the behavior of $\frac{1}{m}\sum_{j\in[m]}\theta_j$ in the large m limit. Concentration inequalities, in contrast, describe nonasymptotic behaviors for finite m and take the generic form

$$\mathbb{P}\left[|\frac{1}{m}\sum\nolimits_{j\in[m]}\theta_j - \mu| \geq \varepsilon\right] \leq \delta(\varepsilon, m), \tag{2.21}$$

for a small ε and small $\delta(\varepsilon, m)$ which is a function of ε and m. So, concentration inequalities are typically more informative than the LLN, or they are quantitative versions of the LLN.

Hoeffding's inequality provides a concentration inequality for bounded random variables.

Theorem 2.3 (Hoeffding's inequality)**.** *Let $\theta_1, \dots, \theta_m$ be iid versions of a scalar random variable θ with $\mathbb{E}\theta = \mu$ and $\theta \in [a, b]$. Then for any $\varepsilon > 0$,*

$$\mathbb{P}\left[|\frac{1}{m}\sum\nolimits_{j\in[m]}\theta_j - \mu| > \varepsilon\right] \leq 2\exp\left(-\frac{2m\varepsilon^2}{(a-b)^2}\right). \tag{2.22}$$

With Hoeffding's inequality, we can easily obtain the following uniform convergence result for learning with bounded loss and finite $\mathcal{H}$.

Theorem 2.4. *Consider a data space $\mathcal{Z}$ and a finite hypothesis class $\mathcal{H}$. Assume that the loss ℓ is bounded in $[0, 1]$. For any $\varepsilon \in (0, 1)$, any $\delta \in (0, 1)$, and any distribution $\mathcal{D}_{\mathcal{Z}}$ on $\mathcal{Z}$, we have the following uniform convergence result:*

$$\mathbb{P}_{\boldsymbol{z}_1,\dots,\boldsymbol{z}_N}\left[|\mathsf{R}(h) - \widehat{\mathsf{R}}_S(h)| \leq \varepsilon \ \forall h \in \mathcal{H}\right] \geq 1 - 2|\mathcal{H}|e^{-2N\varepsilon^2}, \tag{2.23}$$

over the random drawing of $\{\boldsymbol{z}_i\}_{i\in[N]}$ iid from $\mathcal{D}_{\mathcal{Z}}$.

Proof. Since the loss ℓ is bounded in $[0, 1]$ and $z_1, \dots, z_N$ are iid, for each fixed $h \in \mathcal{H}$, the summation terms inside the empirical risk $\widehat{\mathsf{R}}_S(h) = 1/N\sum_{i\in[N]}\ell(h, \boldsymbol{z}_i)$, i.e., $\ell(h, \boldsymbol{z}_i)$'s, are iid scalar random variables bounded in $[0, 1]$. Applying Hoeffding's inequality, we have

$$\mathbb{P}_{\boldsymbol{z}_1,\dots,\boldsymbol{z}_N}\left[|\mathsf{R}(h) - \widehat{\mathsf{R}}_S(h)| \geq \varepsilon \text{ for a fixed } h\right] \leq 2e^{-2N\varepsilon^2}. \tag{2.24}$$

Applying the union bound on all $h \in \mathcal{H}$, we have

$$\mathbb{P}_{\boldsymbol{z}_1,\dots,\boldsymbol{z}_N}\left[\exists h \in \mathcal{H} \text{ with } |\mathsf{R}(h) - \widehat{\mathsf{R}}_S(h)| \geq \varepsilon\right] \leq 2|\mathcal{H}|e^{-2N\varepsilon^2}, \tag{2.25}$$

leading to the claimed result. ■

Note that for binary classification with loss $0/1$, this implies that we need to make $N \geq 2\varepsilon^{-2}\log(2|\mathcal{H}|/\delta)$ to ensure that $\mathbb{P}[\mathsf{R}(h_S) \leq \inf_{h\in\mathcal{H}}\mathsf{R}(h) + \varepsilon] \geq 1 - \delta$, vs. $N \geq \varepsilon^{-1}\log(|\mathcal{H}|/\delta)$ obtained in Theorem 1.1 for the realizable case. The gap between ε^{-2} and ε^{-1} is the main price to pay for making weaker assumptions about the data generation process.

3 General learning setting with infinite $\mathcal{H}$

Hypothesis classes that we work with in practice are often infinite in size. For example, the set of hyperplane classifiers in $\mathbb{R}^d$: $\mathcal{H}_{HC} \doteq \left\{\boldsymbol{x} \mapsto \text{sign}\left(\boldsymbol{w}^\intercal\boldsymbol{x} + b\right) : \boldsymbol{w} \in \mathbb{R}^d, n \in \mathbb{R}\right\}$ used in SVMs. How do we deal with these infinite classes? Is generalization possible?

Before a rigorous treatment, we sketch an informal argument for why generalization is possible with infinite hypothesis classes. For simplicity, let us focus on $\mathcal{H}_{HC}$, and the informal argument goes as follows: although $\boldsymbol{w}$ and b contain real-valued elements, on finite-precision computers, we still have only finitely many of them that we can encode. Assuming 64-bit (i.e., double float) precision, we have in total no more than 2^{64d} possible $\boldsymbol{w}$' and no more than 2^{64} possible b's. Together, we have no more than $O(2^{64d})$ hyperplane classifiers, i.e., $|\mathcal{H}_{HC}| \in O(2^{64d})$. Then, if the loss ℓ is also bounded, we can invoke Theorem 2.4 to obtain generalization bounds. Although simple, such informal arguments often lead to very reasonable results, at least in terms of the order of dependency on key quantities.

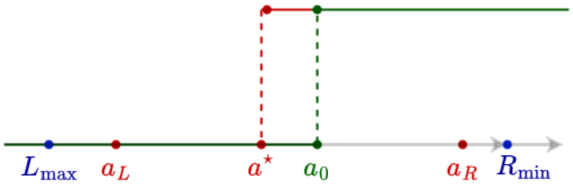


Figure 1: Illustration for Example 3.1 and key quantities involved.

Another question is whether we could have a good generalization at all with infinite hypothesis classes. We answer it using an example.

Example 3.1 (Learning threshold functions; adapted from Example 6.1 of [SSS14])**.** *Consider an unknown labeling function $f_\star(a) = \mathbb{1}\{a \geq a^\star\}$, and the hypothesis class $\mathcal{H}_{TF} \doteq \{a \mapsto \mathbb{1}\{a \geq a_0\} : a_0 \in \mathbb{R}\}$, i.e., the set of one-dimensional threshold functions. Obviously, $\mathcal{H}_{TF}$ is infinite in size, and $h_* \in \mathcal{H}_{TF}$.*

Now consider any unknown distribution $\mathcal{D}_{\mathbb{R}}$, $\{x_i\}_{i\in[N]}$ drawn iid from $\mathcal{D}_{\mathbb{R}}$, and the resulting training set $S = \{(x_i, f_\star(x_i))\}_{i\in[N]}$. Invoking the ERM rule with the $0/1$ loss, we have

$$\widehat{\mathsf{R}}_S(h_{a_0}) = \frac{1}{N}\sum\nolimits_{i\in[N]} \mathbb{1}\{h_{a_0}(x_i) \neq f_\star(x_i)\} = \frac{1}{N}\sum\nolimits_{i\in[N]} \mathbb{1}\{x_i \text{ is between } a^\star \text{ and } a_0\}. \tag{3.1}$$

Let $L_{\max} \doteq \max\{x_i : f_\star(x_i) = 0\}$ ($L_{\max} = -\infty$ if no point is labeled 0) and $R_{\min} \doteq \min\{x_i : f_\star(x_i) = 1\}$ ($R_{\min} = +\infty$ if no point is labeled 1). Clearly, any $a_0 \in (L_{\max}, R_{\min})$ leads to an h_0 with zero empirical risk, and thus the ERM rule can return any $a_0 \in (L_{\max}, R_{\min})$.

Now for any $\varepsilon \in (0,1)$, locate $a_L < a^\star$ and $a_R > a^\star$ so that $\mathbb{P}_x[x \in (a_L, a^\star)] = \varepsilon$ (if no such a_L exists, set $a_L = -\infty$) and $\mathbb{P}_x[x \in (a^\star, a_R)] = \varepsilon$ (if no such a_R exists, set $a_R = +\infty$). Now

$$\mathsf{R}(h_{a_0}) = \mathbb{E}_{x\sim\mathcal{D}_{\mathbb{R}}} \mathbb{1}\{x \text{ is between } a^\star \text{ and } a_0\} = \mathbb{P}_{x\sim\mathcal{D}_{\mathbb{R}}}[x \text{ is between } a^\star \text{ and } a_0]. \tag{3.2}$$

To proceed, we have three cases:

- *If $L_{\max} = -\infty$ (i.e., $L_{\min} < +\infty$), there is no point labeled as 0, implying that $x_i \geq a^\star\ \forall i$. So,*

$$\mathbb{P}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}[\mathsf{R}(h_{a_0}) \geq \varepsilon] \leq \mathbb{P}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}[a_R \leq R_{\min}] \leq (1-\varepsilon)^N \leq e^{-\varepsilon N}. \tag{3.3}$$

- *If $L_{\min} = +\infty$ (i.e., $L_{\max} > -\infty$), there is no point labeled as 1, implying that $x_i \leq a^\star\ \forall i$. So,*

$$\mathbb{P}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}[\mathsf{R}(h_{a_0}) \geq \varepsilon] \leq \mathbb{P}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}[a_L \geq L_{\max}] \leq (1-\varepsilon)^N \leq e^{-\varepsilon N}. \tag{3.4}$$

- *If $L_{\max} > -\infty$ and $L_{\min} < +\infty$, we have that*

$$a_L \leq L_{\max} \text{ and } a_R \geq R_{\min} \Longrightarrow \mathsf{R}(h_{a_0}) \leq \varepsilon \Longleftrightarrow \mathsf{R}(h_{a_0}) \geq \varepsilon \Longrightarrow a_L \geq L_{\max} \text{ or } a_R \leq R_{\min}.$$

  Now,

  - *If $a_L = -\infty$, $a_L \geq L_{\max}$ can never happen. We have*

$$\begin{aligned} \mathbb{P}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}[\mathsf{R}(h_{a_0}) \geq \varepsilon] &\leq \mathbb{P}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}[a_L \geq L_{\max} \text{ or } a_R \leq R_{\min}] \\ &= \mathbb{P}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}[a_R \leq R_{\min}] \leq (1-\varepsilon)^N \leq e^{-\varepsilon N}. \end{aligned} \tag{3.5}$$

- *If $a_R = +\infty$, $a_R \le R_{\min}$ can never happen. We have*

$$\begin{aligned}\mathbb{P}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}[\mathsf{R}(h_{a_0}) \ge \varepsilon] &\le \mathbb{P}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}[a_L \ge L_{\max} \text{ or } a_R \le R_{\min}] \\ &= \mathbb{P}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}[a_L \ge L_{\max}] \le (1-\varepsilon)^N \le e^{-\varepsilon N}. \quad (3.6)\end{aligned}$$

- *If $a_L > -\infty$ and $a_R < +\infty$,*

$$\begin{aligned}\mathbb{P}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}[\mathsf{R}(h_{a_0}) \ge \varepsilon] &\le \mathbb{P}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}[a_L \ge L_{\max} \text{ or } a_R \le R_{\min}] && (3.7)\\ &\le \mathbb{P}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}[a_L \ge L_{\max}] + \mathbb{P}[a_R \le R_{\min}] \quad \textit{(union bound)} && (3.8)\\ &\le (1-\varepsilon)^N + (1-\varepsilon)^N \le 2e^{-\varepsilon N}. && (3.9)\end{aligned}$$

In summary, $\mathbb{P}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N}[\mathsf{R}(h_{a_0}) \ge \varepsilon] \le 2e^{-\varepsilon N}$, or we need $N \ge \varepsilon^{-1}\log(2/\delta)$ to ensure at least $1-\delta$ level confidence.

So, the next question is what kind of infinite hypothesis classes can be learned?

3.1 Complexity of sets and an idea for measuring complexity of infinite $\mathcal{H}$

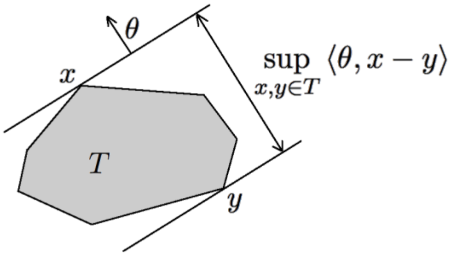


Figure 2: Geometric meaning of spherical width of a set—spherical width is the average width of the set over all directions; figure adapted from Figure 7.3 of [Ver18]

To answer the previous question, we need to measure the *effective complexity* of $\mathcal{H}$: although $\mathcal{H}$ has infinitely many elements, it may have only very limited power.

Example 3.2. *Consider the hyperthesis class: $\mathcal{H} \doteq \{x \mapsto \mathrm{sign}(x^p) : p \in \mathbb{N}\}$. It clearly contains infinitely many functions, but there is too much redundancy: for all $x \ge 0$, the sign is always $+1$. For $x < 0$, all $x \mapsto \mathrm{sign}(x^p)$ with even p's predict $+1$, and those with odd p's predict -1. In this sense, there are only two effective functions inside, $x \mapsto \mathrm{sign}(x)$ and $x \mapsto \mathrm{sign}(x^2)$. Intuitively, the high level of redundancy substantially reduces the effective complexity of $\mathcal{H}$, despite its superficial infinite capacity. Moreover, if we draw N random points, the two effective functions can generate at most two labeling patterns on them.* ***This suggests that we can measure the effective complexity of $\mathcal{H}$ by checking the number/size of the output patterns when passing the training set through all functions within $\mathcal{H}$, as we do below.***

3.1.1 Measuring the size of a set

How do we measure the size of a set $A \subset \mathbb{R}^n$ in general? If A is finite, we typically use its cardinality $|A|$. But cardinality is powerless when describing infinite sets—when we talk about their size, we are intuitively thinking of volume-like quantities. This motivates us to consider the *directional width*. For any fixed direction $\boldsymbol{u}$ (with $\|\boldsymbol{u}\|_2 = 1$), the width of the set A in direction $\boldsymbol{u}$ is the smallest distance between two hyperplanes that are orthogonal to $\boldsymbol{u}$ and contain A in between (see Fig. 2), analytically expressed as

$$\sup\nolimits_{\boldsymbol{a},\boldsymbol{a}'\in A} \langle \boldsymbol{u}, \boldsymbol{a}-\boldsymbol{a}' \rangle . \quad (3.10)$$

Now, we can measure the overall width of the set by averaging over all directions:

$$\mathbb{E}_{\boldsymbol{u}\sim \mathrm{Unif}(\mathbb{S}^{n-1})} \sup\nolimits_{\boldsymbol{a},\boldsymbol{a}'\in A} \langle \boldsymbol{u}, \boldsymbol{a}-\boldsymbol{a}' \rangle = \frac{1}{\mathrm{area}(\mathbb{S}^{n-1})} \int_{\boldsymbol{u}\in\mathbb{S}^{n-1}} \sup\nolimits_{\boldsymbol{a},\boldsymbol{a}'\in A} \langle \boldsymbol{u}, \boldsymbol{a}-\boldsymbol{a}' \rangle \; d\boldsymbol{u}, \quad (3.11)$$

where $\mathbb{S}^{n-1} \subset \mathbb{R}^n$ is the unit sphere in $\mathbb{R}^n$ and $\mathrm{Unif}(\mathbb{S}^{n-1})$ means the uniform distribution on the sphere[3]. This motivates the following definition of *spherical width* (also called *mean width*):

$$(\textbf{spherical width}) \quad \mathsf{w}_s(A) \doteq \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{n-1})} \sup\nolimits_{\boldsymbol{a} \in A} \langle \boldsymbol{u}, \boldsymbol{a} \rangle . \tag{3.12}$$

Note that the quantity defined in Eq. (3.11) is $\mathsf{w}_s(A - A)$ (here, $A - A \doteq \{\boldsymbol{a} - \boldsymbol{a}' : \boldsymbol{a}, \boldsymbol{a}' \in A\}$), and $\mathsf{w}_s(A - A) = 2\mathsf{w}_s(A)$, as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{n-1})} \sup\nolimits_{\boldsymbol{a}, \boldsymbol{a}' \in A} \langle \boldsymbol{u}, \boldsymbol{a} - \boldsymbol{a}' \rangle &= \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{n-1})} \sup\nolimits_{\boldsymbol{a}, \boldsymbol{a}' \in A} \langle \boldsymbol{u}, \boldsymbol{a} \rangle + \langle \boldsymbol{u}, -\boldsymbol{a}' \rangle \\ &= \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{n-1})} \sup\nolimits_{\boldsymbol{a} \in A} \langle \boldsymbol{u}, \boldsymbol{a} \rangle + \sup\nolimits_{\boldsymbol{a}' \in A} \langle \boldsymbol{u}, -\boldsymbol{a}' \rangle \\ &= \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{n-1})} \sup\nolimits_{\boldsymbol{a} \in A} \langle \boldsymbol{u}, \boldsymbol{a} \rangle + \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{n-1})} \sup\nolimits_{\boldsymbol{a}' \in A} \langle -\boldsymbol{u}, \boldsymbol{a}' \rangle \\ &= \mathsf{w}_s(A) + \mathsf{w}_s(A) = 2\mathsf{w}_s(A), \end{aligned} \tag{3.13}$$

where we note that

$$\mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{n-1})} \sup\nolimits_{\boldsymbol{a}' \in A} \langle -\boldsymbol{u}, \boldsymbol{a}' \rangle = \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{n-1})} \sup\nolimits_{\boldsymbol{a}' \in A} \langle \boldsymbol{u}, \boldsymbol{a}' \rangle = \mathsf{w}_s(A), \tag{3.14}$$

as $\mathbb{E}_{\boldsymbol{v}} f(\boldsymbol{v}) = \mathbb{E}_{\boldsymbol{v}} f(-\boldsymbol{v})$ for any symmetric random vector $\boldsymbol{v}$ (i.e., $\boldsymbol{v}$ and $-\boldsymbol{v}$ have the same distribution) and any integrable function f.

The spherical width is not always easy to compute or estimate—we typically only need a reasonable upper bound of it. A convenient alternative is the *Gaussian width*:

$$(\textbf{Gaussian width/complexity}) \quad \mathsf{w}_g(A) \doteq \mathbb{E}_{\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})} \sup\nolimits_{\boldsymbol{a} \in A} \langle \boldsymbol{g}, \boldsymbol{a} \rangle , \tag{3.15}$$

where we basically replace the $\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{n-1})$ in spherical width by a Gaussian random vector $\boldsymbol{g}$. Similarly, we can replace the Gaussian vector $\boldsymbol{g}$ by a Rademacher vector $\boldsymbol{r} \sim_{iid} \mathrm{Rad}$, where a Rademacher random variable takes $+1$ and -1 with $1/2$ probability each, yielding the *Rademacher width*:

$$(\textbf{Rademacher width/complexity}) \quad \mathsf{w}_r(A) \doteq \mathbb{E}_{\boldsymbol{r} \sim_{iid} \mathrm{Rad}} \sup\nolimits_{\boldsymbol{a} \in A} \langle \boldsymbol{r}, \boldsymbol{a} \rangle . \tag{3.16}$$

We are going to use Rademacher complexity later for measuring the complexity of infinite hypothesis classes, so we create a special notation to denote its normalized version:

$$(\textbf{(Normalized) Rademacher complexity}) \quad \mathsf{RC}(A) \doteq \frac{1}{n} \mathbb{E}_{\boldsymbol{r} \sim_{iid} \mathrm{Rad}} \sup_{\boldsymbol{a} \in A} \langle \boldsymbol{r}, \boldsymbol{a} \rangle , \tag{3.17}$$

where we normalize by $1/n$, where n is the ambient dimension of the set A.

3.1.2 Measuring the complexity of an infinite hypothesis class

We are not going to measure the size of the set $\mathcal{H}$ directly, which is in a functional space. As suggested in Example 3.2, we will measure the size of the image set of S passed through the hypothesis class $\mathcal{H}$. To proceed, with $S = \{\boldsymbol{z}_i\}_{i \in [N]}$ as the training set, we define

$$(\ell \circ h)[S] \doteq [\ell(h, \boldsymbol{z}_1), \dots, \ell(h, \boldsymbol{z}_N)] \in \mathbb{R}^N, \ \forall h \in \mathcal{H} \tag{3.18}$$

$$(\ell \circ \mathcal{H})[S] \doteq \{(\ell \circ h)[S] : h \in \mathcal{H}\} . \tag{3.19}$$

[3] The unit sphere in $\mathbb{R}^n$ is written as $\mathbb{S}^{n-1}$ because its effective (manifold) dimension (i.e., degrees of freedom) is only $n - 1$.

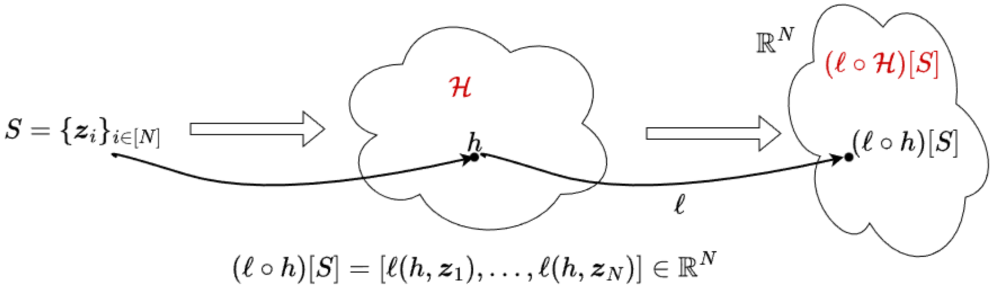

$S = \{\boldsymbol{z}_i\}_{i\in[N]}$
$\mathcal{H}$
h
$\mathbb{R}^N$
$(\ell \circ \mathcal{H})[S]$
$(\ell \circ h)[S]$
ℓ
$(\ell \circ h)[S] = [\ell(h, \boldsymbol{z}_1), \dots, \ell(h, \boldsymbol{z}_N)] \in \mathbb{R}^N$

Figure 3: The key idea of measuring the complexity of an infinite hypothesis class $\mathcal{H}$—we measure the size of image set $(\ell \circ \mathcal{H})[S] \subset \mathbb{R}^N$, produced by passing a training set S through all $h \in \mathcal{H}$.

The key idea is to measure the size of $(\ell \circ \mathcal{H})[S]$, using any reasonable complexity measure discussed in Section 3.1.1 or alike; see Fig. 3.

Below, we implement this idea in two different settings: the first pertains to the general learning setting, and we measure the size of $(\ell \circ \mathcal{H})[S]$ via Rademacher complexity (Section 3.2); the second is about binary classification, and we take ℓ as the $0/1$ loss, and measure the size of the discrete set $(\ell \circ \mathcal{H})[S]$ simply by its cardinality, resulting in the growth function (Section 3.3). The notion of VC dimension in Section 3.4 is closely related to growth function.

3.2 Rademacher complexity

For the general learning setting, the *empirical Rademacher complexity* of $\ell \circ \mathcal{H}$ is defined as

$$\widehat{\mathsf{RC}}_S(\ell \circ \mathcal{H}) \doteq \mathsf{RC}((\ell \circ \mathcal{H})[S]) = \frac{1}{N}\mathbb{E}_{\boldsymbol{r}\sim_{iid}\mathrm{Rad}} \sup_{h\in\mathcal{H}} \langle \boldsymbol{r}, (\ell \circ h)[S]\rangle \,, \tag{3.20}$$

i.e., the Rademacher complexity of the set $(\ell \circ \mathcal{H})[S]$. The word "empirical" here stresses that the complexity measure depends on the training set S, and is hence random. To get rid of the randomness, we can take the expectation with respect to S, resulting in the *Rademacher complexity* of $\ell \circ \mathcal{H}$:

$$\mathsf{RC}_N(\ell \circ \mathcal{H}) \doteq \mathbb{E}_{S\sim\mathcal{D}_{\mathcal{Z}}^N} \widehat{\mathsf{RC}}_S(\ell \circ \mathcal{H}). \tag{3.21}$$

We can also define the (empirical) Gaussian complexity in an analogous way, as we explore in the homework.

We have the following uniform convergence results for infinite $\mathcal{H}$ with bounded ℓ, where the (empirical) Rademacher complexity dominates the convergence gap.

Theorem 3.3 (Uniform convergence via Rademacher complexity)**.** *Consider a data space $\mathcal{Z}$ and a hypothesis class $\mathcal{H}$. Assume the loss ℓ is bounded by $c > 0$ for all $\boldsymbol{z} \in \mathcal{Z}$ and all $h \in \mathcal{H}$, i.e., $|\ell(h, \boldsymbol{z})| \le c$. Then, for any $\delta \in (0,1)$ and any $\mathcal{D}_{\mathcal{Z}}$, with probability at least $1-\delta$ over the draw of an iid training set S of size N, the following uniform convergence results hold*

$$\mathsf{R}(h) \le \widehat{\mathsf{R}}_S(h) + 2\widehat{\mathsf{RC}}_S(\ell \circ \mathcal{H}) + C_1 c\sqrt{\log(1/\delta)/N} \quad \forall h \in \mathcal{H} \tag{3.22}$$

$$\mathsf{R}(h) \le \widehat{\mathsf{R}}_S(h) + 2\mathsf{RC}(\ell \circ \mathcal{H}) + C_1 c\sqrt{\log(1/\delta)/N} \quad \forall h \in \mathcal{H}, \tag{3.23}$$

where $C_1, C_2 > 0$ are two universal constants.

With this master theorem at hand, we typically only need to focus on estimating $\widehat{\mathsf{RC}}_S(\ell \circ \mathcal{H})$ or $\mathsf{RC}(\ell \circ \mathcal{H})$, and plug them into the bounds.

For this purpose, the following properties of the Rademacher complexity of sets can be useful, as, remember, we define the Rademacher complexity of a hypothesis class by the Rademacher complexity of its image set.

Theorem 3.4 (Properties of Rademacher complexity of sets). *Assume all sets below are subsets of $\mathbb{R}^n$.*

1. ***nonnegativity*** *For any set A, $\mathsf{RC}(A) \ge 0$*
2. ***monotonicity*** *If $A \subset A'$, $\mathsf{RC}(A) \le \mathsf{RC}(A')$*
3. ***summation*** *For any two sets A, B, $\mathsf{RC}(A + B) = \mathsf{RC}(A) + \mathsf{RC}(B)$*
4. ***affine transform*** *For any set A and any real numbers a, b*

$$\mathsf{RC}(aA + b) = |a|\mathsf{RC}(A) \tag{3.24}$$

5. ***convexity*** *For any A and its convex hull*

$$\mathrm{conv}(A) \doteq \left\{ \sum_{j \in [N]} \alpha_j \boldsymbol{x}_i : N \in \mathbb{N}, \boldsymbol{x}_j \in A\ \forall j, \boldsymbol{\alpha} \ge \mathbf{0}, \mathbf{1}^\intercal \boldsymbol{\alpha} = 1 \right\}, \tag{3.25}$$

$\mathsf{RC}(A) = \mathsf{RC}(\mathrm{conv}(A))$.

6. ***Talagrand's contraction lemma***[4] *For $j \in [n]$, consider L-Lipschitz functions $\phi_j : \mathbb{R} \to \mathbb{R}$, i.e., $|\phi_j(x) - \phi_j(x')| \le L|x - x'|$ for all $x, x' \in \mathbb{R}$. Write $\boldsymbol{\phi} \doteq [\phi_1, \dots, \phi_n]$. For any A, write $\boldsymbol{\phi} \circ \boldsymbol{a} \doteq [\phi_1(a_1), \dots, \phi_n(a_n)]$ and $\boldsymbol{\phi} \circ A \doteq \{\boldsymbol{\phi} \circ \boldsymbol{a} : \boldsymbol{a} \in A\}$. It holds that*

$$\mathsf{RC}(\boldsymbol{\phi} \circ A) \le L \cdot \mathsf{RC}(A). \tag{3.26}$$

The contraction lemma can often help simplify the estimation of Rademacher complexity. For example, for supervised learning problems, we can obtain the following.

Corollary 3.5 (Contraction lemma for supervised learning). *Consider $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} \subset \mathbb{R}$, and a hypothesis class $\mathcal{H}$ consisting of predictors mapping from $\mathcal{X}$ to $\mathcal{Y}$. Assume that $\forall y \in \mathcal{Y}$, the mapping $a \mapsto \ell(a, y)$ is L-Lipschitz. Then, for any $\mathcal{D}_{\mathcal{X} \times \mathcal{Y}}$ and any $S = \{(\boldsymbol{x}_i, y_i)\}_{i \in [N]}$, write $S_{\boldsymbol{x}} = \{\boldsymbol{x}_i\}_{i \in [N]}$, we have*

$$\widehat{\mathsf{RC}}_S(\ell \circ \mathcal{H}) \le L \cdot \widehat{\mathsf{RC}}_{S_{\boldsymbol{x}}}(\mathcal{H}). \tag{3.27}$$

Proof. Consider $A = \mathcal{H}[S_{\boldsymbol{x}}] = \{h[S_{\boldsymbol{x}}] : h \in \mathcal{H}\}$ where $\boldsymbol{h}[S_{\boldsymbol{x}}] \doteq [h(\boldsymbol{x}_1), \dots, h(\boldsymbol{x}_N)] \subset \mathbb{R}^N$, and $\boldsymbol{\phi} \circ A = (\ell \circ \mathcal{H})[S] \doteq \{(\ell \circ h)[S] : h \in \mathcal{H}\}$ where $(\ell \circ h)[S] \doteq [\ell(h(\boldsymbol{x}_1), y_1), \dots, \ell(h(\boldsymbol{x}_N), y_N)] \subset \mathbb{R}^N$. So, here $i \in [N]$, $\phi_i : h(\boldsymbol{x}_i) \mapsto \ell(h(\boldsymbol{x}_i), y_i)$, which is L-Lipschitz by our assumption. Invoking the contraction lemma, we have

$$\mathsf{RC}((\ell \circ \mathcal{H})[S]) \le L \cdot \mathsf{RC}(\mathcal{H}[S_{\boldsymbol{x}}]), \tag{3.28}$$

implying the claimed result. ■

Moreover, for finite sets, we have the following estimates.

Lemma 3.6 (Massart lemma). *Let $A = \{\boldsymbol{a}_1, \dots, \boldsymbol{a}_m\}$ be a finite set of vectors in $\mathbb{R}^n$. Definite $\overline{\boldsymbol{a}} \doteq \frac{1}{m} \sum_{j \in [m]} \boldsymbol{x}_j$, i.e., center of the set. Then,*

$$\mathsf{RC}(A) \le \max_{\boldsymbol{a} \in A} \|\boldsymbol{a} - \overline{\boldsymbol{a}}\| \sqrt{2 \log m}/n. \tag{3.29}$$

[4] Due to the famous mathematician and Abel Prize laureate Michel Talagrand (https://en.wikipedia.org/wiki/Michel_Talagrand).

Next, we work out a couple of examples for supervised learning. For simplicity, we assume that the loss ℓ is 1-Lipschitz for any y, so we focus on estimating $\widehat{\mathsf{RC}}_{S_{\boldsymbol{x}}}(\mathcal{H})$ due to the contraction lemma.

Lemma 3.7. *Consider $\mathcal{H}_1 \doteq \left\{\boldsymbol{x} \mapsto \boldsymbol{w}^\intercal \boldsymbol{x} : \|\boldsymbol{w}\|_1 \le 1, \boldsymbol{x} \in \mathbb{R}^d\right\}$ and any input set $S_{\boldsymbol{x}} = \{\boldsymbol{x}_i\}_{i\in[N]}$. We have*

$$\widehat{\mathsf{RC}}_{S_{\boldsymbol{x}}}(\mathcal{H}_1) \le \max_i \|\boldsymbol{x}_i\|_\infty \sqrt{\frac{2\log(2d)}{N}}. \tag{3.30}$$

Proof. We have

$$\begin{aligned}
\widehat{\mathsf{RC}}_{S_{\boldsymbol{x}}}(\mathcal{H}_1) &= \frac{1}{N}\mathbb{E}_{\boldsymbol{r}\sim_{iid}\text{Rad}} \sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_1\le 1} \langle \boldsymbol{r}, \boldsymbol{X}\boldsymbol{w}\rangle \quad (\boldsymbol{X} \text{ collects all } \boldsymbol{x}_i\text{'s as its rows}) && (3.31)\\
&= \frac{1}{N}\mathbb{E}_{\boldsymbol{r}\sim_{iid}\text{Rad}} \sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_1\le 1} \langle \boldsymbol{X}^\intercal\boldsymbol{r}, \boldsymbol{w}\rangle && (3.32)\\
&\le \frac{1}{N}\mathbb{E}_{\boldsymbol{r}\sim_{iid}\text{Rad}} \|\boldsymbol{X}^\intercal\boldsymbol{r}\|_\infty, && (3.33)
\end{aligned}$$

where to obtain the last line we use the Hölder's inequality, which says that for any two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, $|\langle \boldsymbol{u}, \boldsymbol{v}\rangle| \le \|\boldsymbol{u}\|_p \|\boldsymbol{v}\|_q$ for any p, q with $1/p + 1/q = 1$, i.e., $\|\cdot\|_p$ and $\|\cdot\|_q$ are the dual norm to each other. Now

$$\mathbb{E}_{\boldsymbol{r}\sim_{iid}\text{Rad}} \|\boldsymbol{X}^\intercal\boldsymbol{r}\|_\infty = \mathbb{E}_{\boldsymbol{r}\sim_{iid}\text{Rad}} \sup_{j\in[d]} \left|\boldsymbol{e}_j^\intercal \boldsymbol{X}^\intercal \boldsymbol{r}\right| = \mathbb{E}_{\boldsymbol{r}\sim_{iid}\text{Rad}} \sup_{s=\pm 1, j\in[d]} \langle s\boldsymbol{X}\boldsymbol{e}_j, \boldsymbol{r}\rangle = N\mathsf{RC}(A), \tag{3.34}$$

where $A \doteq \{s\boldsymbol{X}\boldsymbol{e}_j : s = \pm 1, j \in [d]\}$. Clearly, $|A| = 2d$, and $\|s\boldsymbol{X}\boldsymbol{e}_j\|_2 \le \sqrt{N}\max_i \|\boldsymbol{x}_i\|_\infty$. By the Massart lemma, we have

$$\mathsf{RC}(A) \le \max_i \|\boldsymbol{x}_i\|_\infty \sqrt{\frac{2\log(2d)}{N}}. \tag{3.35}$$

Combining the above chain of results completes the proof. ■

Lemma 3.8. *Consider $\mathcal{H}_2 \doteq \{\boldsymbol{x} \mapsto \boldsymbol{w}^\intercal \boldsymbol{x} : \|\boldsymbol{w}\|_2 \le 1\}$ and any input set $S_{\boldsymbol{x}} = \{\boldsymbol{x}_i\}_{i\in[N]}$. We have*

$$\widehat{\mathsf{RC}}_{S_{\boldsymbol{x}}}(\mathcal{H}_2) \le \max_i \|\boldsymbol{x}_i\|_2^2/\sqrt{N}. \tag{3.36}$$

Proof. We have

$$\begin{aligned}
\widehat{\mathsf{RC}}_{S_{\boldsymbol{x}}}(\mathcal{H}_1) &= \frac{1}{N}\mathbb{E}_{\boldsymbol{r}\sim_{iid}\text{Rad}} \sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\le 1} \langle \boldsymbol{r}, \boldsymbol{X}\boldsymbol{w}\rangle \quad (\boldsymbol{X} \text{ collects all } \boldsymbol{x}_i\text{'s as its rows}) && (3.37)\\
&= \frac{1}{N}\mathbb{E}_{\boldsymbol{r}\sim_{iid}\text{Rad}} \sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\le 1} \langle \boldsymbol{X}^\intercal\boldsymbol{r}, \boldsymbol{w}\rangle && (3.38)\\
&\le \frac{1}{N}\mathbb{E}_{\boldsymbol{r}\sim_{iid}\text{Rad}} \|\boldsymbol{X}^\intercal\boldsymbol{r}\|_2, && (3.39)
\end{aligned}$$

where again we use the Hölder's inequality to obtain the last line. Now Jensen's inequality says that for convex (resp. concave) function $\phi : \mathbb{R}^n \to \mathbb{R}$, $\mathbb{E}_{\boldsymbol{z}}\phi(\boldsymbol{z}) \ge \phi(\mathbb{E}_{\boldsymbol{z}}\boldsymbol{z})$ (resp. $\mathbb{E}_{\boldsymbol{z}}\phi(\boldsymbol{z}) \le \phi(\mathbb{E}_{\boldsymbol{z}}\boldsymbol{z})$). Now

$$\mathbb{E}_{\boldsymbol{r}\sim_{iid}\text{Rad}} \|\boldsymbol{X}^\intercal\boldsymbol{r}\|_2 = \mathbb{E}_{\boldsymbol{r}\sim_{iid}\text{Rad}} \sqrt{\|\boldsymbol{X}^\intercal\boldsymbol{r}\|_2^2} \le \sqrt{\mathbb{E}_{\boldsymbol{r}\sim_{iid}\text{Rad}}\|\boldsymbol{X}^\intercal\boldsymbol{r}\|_2^2}, \tag{3.40}$$

where we apply Jensen's inequality to the concave function $z \mapsto \sqrt{z}$ for $z \geq 0$. Now, we estimate $\mathbb{E}_{\boldsymbol{r} \sim_{iid} \mathrm{Rad}} \|\boldsymbol{X}^\intercal \boldsymbol{r}\|_2^2$:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{r} \sim_{iid} \mathrm{Rad}} \|\boldsymbol{X}^\intercal \boldsymbol{r}\|_2^2 = \mathbb{E}_{\boldsymbol{r} \sim_{iid} \mathrm{Rad}} \langle \boldsymbol{X}^\intercal \boldsymbol{r}, \boldsymbol{X}^\intercal \boldsymbol{r} \rangle = \mathbb{E}_{\boldsymbol{r} \sim_{iid} \mathrm{Rad}} \langle \boldsymbol{X}\boldsymbol{X}^\intercal, \boldsymbol{r}\boldsymbol{r}^\intercal \rangle = \langle \boldsymbol{X}\boldsymbol{X}^\intercal, \mathbb{E}_{\boldsymbol{r} \sim_{iid} \mathrm{Rad}} \boldsymbol{r}\boldsymbol{r}^\intercal \rangle \\ = \langle \boldsymbol{X}\boldsymbol{X}^\intercal, \boldsymbol{I} \rangle = \mathrm{tr}\left(\boldsymbol{X}\boldsymbol{X}^\intercal\right) = \|\boldsymbol{X}\|_F^2 \leq N \max_i \|\boldsymbol{x}_i\|_2^2. \quad (3.41)\end{aligned}$$

Combining the above chain of results completes the proof. ■

3.3 Growth function

For binary classification problems, the growth function and the VC dimension are useful alternative complexity measures that can be easier to estimate than the Rademacher complexity.

To motivate the growth function, we refer back to the general strategy we describe in Section 3.1.2 (see also Fig. 3). Now $Z = \mathcal{X} \times \{+1, -1\}$, $S = \{(\boldsymbol{x}_i, y_i)\}_{i \in [N]}$, and $\ell(h, \boldsymbol{z}_i) = \mathbb{1}\{h(\boldsymbol{x}_i) \neq y_i\}$ $\forall i$, i.e., we take the $0/1$ loss. It is easy to see that no matter what $\mathcal{H}$ is, $(\ell \circ \mathcal{H})[S]$ must be finite, as there are at most 2^N binary vectors there. So we will measure the complexity of $(\ell \circ \mathcal{H})[S]$ by its cardinality. Now, since y_i's in S are fixed and binary-valued,

$$|(\ell \circ \mathcal{H})[S]| = |\mathcal{H}[S_{\boldsymbol{x}}]|. \quad (3.42)$$

Therefore, we define the growth function as follows.

Definition 3.9 (Growth function). *Consider $S_{\boldsymbol{x}} = \{\boldsymbol{x}_i\}_{i \in [N]}$, $h(S_{\boldsymbol{x}}) \doteq [h(\boldsymbol{x}_1), \dots, h(\boldsymbol{x}_N)]$, and $\mathcal{H}(S_{\boldsymbol{x}}) \doteq \{h(S) : h \in \mathcal{H}\}$. The growth function (also called the shattering coefficient) of $\mathcal{H}$, as a function of N (for any integer $N \geq 1$), is defined as*

$$\Pi_{\mathcal{H}}(N) = \max_{S_{\boldsymbol{x}} : |S_{\boldsymbol{x}}| = N} |\mathcal{H}(S_{\boldsymbol{x}})|. \quad (3.43)$$

In other words, if we call each distinct binary labeling $h(S)$ of S a *dichotomy*, the growth function counts the number of distinct dichotomies realizable by $\mathcal{H}$. Note that inside the definition, $\max$ is taken over all possible $S_{\boldsymbol{x}}$, as this definition is only combinatorial in nature—if we take a very bad, non-representative set of N points (e.g., all of them are almost identical to each other), we may well expect to see far fewer dichotomies. Also, $\Pi_{\mathcal{H}}(N) \leq 2^N$, as there is a total of 2^N dichotomies on any S with $|S| = N$.

For binary classification, one can connect the Rademacher complexity and the growth function as follows.

Lemma 3.10. *For binary classification with ℓ as the $0/1$ loss and any hypothesis class $\mathcal{H}$, we have*

$$\widehat{\mathsf{RC}}_S(\ell \circ \mathcal{H}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(N)}{N}}, \quad (3.44)$$

which implies that

$$\mathsf{RC}_N(\ell \circ \mathcal{H}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(N)}{N}}. \quad (3.45)$$

One can easily combine the upper bounds here with the uniform convergence results in Theorem 3.3 to obtain uniform convergence results via the growth function.

3.4 VC dimensions

The growth function is a function of N; in principle, we need to estimate it for all N. The VC (i.e., Vapnik-Chervonenkis) dimension, to be discussed below, offers a single-number summary of the growth function and tends to be easy to estimate.

Definition 3.11 (Shattering)**.** *A set of points S is* schattered *by a hypothesis class $\mathcal{H}$, if each dichotomy on S can be realized by a certain $h \in \mathcal{H}$, that is, if $\mathcal{H}$ realizes all possible dichotomies of S.*

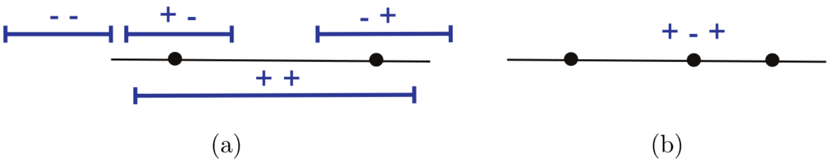


Figure 4: Illustration of shattering on interval. Figure adapted from Fig. 3.1 of [Moh18].

For example, consider the interval hypothesis class

$$\mathcal{H}_{itvl} \doteq \{x \mapsto 2 \cdot \mathbb{1}\{x \in (a,b)\} - 1 : a < b, a, b \in \mathbb{R}\}. \tag{3.46}$$

As Fig. 4 shows, any two distinct points on the real line can be shattered by $\mathcal{H}_{itvl}$, but not three points. Note that the combinatorial nature of shattering means that we often need to draw (mental) pictures to figure this out.

Definition 3.12 (VC dimension)**.** *The VC dimension of $\mathcal{H}$, denoted as* $\mathsf{VCdim}(\mathcal{H})$*, is the size of the largest set that can be shattered by $\mathcal{H}$, or equivalently,*

$$\mathsf{VCdim}(\mathcal{H}) = \max\left\{N : \Pi_{\mathcal{H}}(N) = 2^N\right\}. \tag{3.47}$$

Obviously, the definition implies that $\Pi_{\mathcal{H}}(N) < 2^N$ for any $N > \mathsf{VCdim}(\mathcal{H})$. Hence, to prove $\mathsf{VCdim}(\mathcal{H}) = d$, one needs to show that

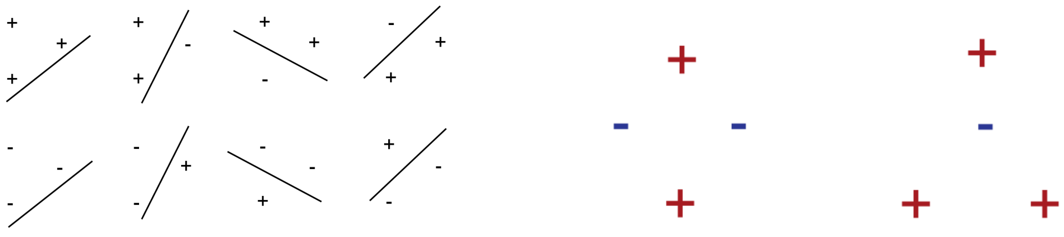
- there exists a set of size d that can be shattered by $\mathcal{H}$, **and**
- no set of cardinality $d+1$ can be shattered by $\mathcal{H}$.

We present several quick examples.

Example 3.13 (Examples of VC dimension estimation)**.** *Consider the following examples.*

- $\mathcal{H}_{itvl}$. *The shattering argument above implies that* $\mathsf{VCdim}(\mathcal{H}_{itvl}) = 2$.
- ***Hyperplane classifiers in*** $\mathbb{R}^d$ *Consider* $\mathcal{H}^d_{HC} \doteq \left\{\boldsymbol{x} \mapsto \mathrm{sign}(\boldsymbol{w}^\intercal \boldsymbol{x} + b) : \boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}\right\}$. *We first look at the $d = 2$ case, which is easy to visualize. Any 3 points in general positions (i.e., non-colinear) in $\mathbb{R}^2$ can be shattered, as shown in Fig. 5(a). Now, for any 4 points, clearly we cannot shatter them if any 3 of them are colinear (as the $+-+$ pattern on 3 colinear points cannot be realized). Assume that no 3 points of the 4 points are colinear. Then, either all 4 points are the vertices of their convex hull, or the 1 point lies in the interior of their convex hull, with the remaining 3 points being the vertices; see Fig. 5(b). For either case, we can easily construct a dichotomy that cannot be realized, as shown in Fig. 5(b). So,* $\mathsf{VCdim}(\mathcal{H}^2_{HC}) = 3$.

  For general $d \ge 3$, we can extend the above argument, in a more systematic way.
  - $\mathcal{H}^d_{HC}$ *can shatter $d+1$ points. Consider $\boldsymbol{x}_0 = \boldsymbol{0}$ and $\boldsymbol{x}_i = \boldsymbol{e}_i$ for $i \in [d]$. For any binary labeling $\boldsymbol{y} \in \{+1,-1\}^{d+1}$ on this, the hyperplane classifier* $\mathrm{sign}\left(\boldsymbol{y}^\intercal \boldsymbol{x}_i + y_0/2\right) = y_i$ *for all $i = 0, 1, \ldots, d$. So, these $d+1$ points are shattered by $\mathcal{H}^d_{HC}$.*



(a) (b)

Figure 5: (a) Illustration of shattering patterns of hyperplane classifiers in $\mathbb{R}^2$ on three points in general positions; (b) Generic configurations of four points that cannot be shattered by hyperplane classifiers.

- *$\mathcal{H}_{HC}^d$ cannot shatter $d+2$ points.* *To prove this, we need the famous* ***Radon's theorem*** *from convex analysis: any set A of $d+2$ points in $\mathbb{R}^d$ can be partitioned into two subsets A_1 and A_2 so that the convex hulls of A_1 and A_2 intersect. Now, let us consider such a set A with $|A| = d+2$ and the associated A_1 and A_2 so that $\mathrm{conv}(A_1) \cap \mathrm{conv}(A_2) \neq \emptyset$. Assign $+1$ to all points in A_1 and -1 to all points in A_2. Now, if there is an hyperplane that separates A_1 and A_2, it also separates their convex hulls, so that the convex hulls are disjoint, violating Radon's theorem. So, no $d+2$ points can be shattered by $\mathcal{H}_{HC}^d$.*

Together, the above arguments imply that $\mathsf{VCdim}(\mathcal{H}_{HC}) = d+1$.

- ***Sine functions*** *Consider $\mathcal{H}_{\sin} \doteq \{x \mapsto \mathrm{sign}\,(\sin(\omega x)) : \omega \in \mathbb{R}\}$. For any $N \in \mathbb{N}$, consider N points $2^1, \ldots, 2^N$ and $\omega = -\pi \sum_{i\in[N]} 2^{-i}\mathbb{1}\{y_i = +1\} - \pi 2^{-N}$. Also, note that* $\mathrm{sign}(\sin(z\pi)) = (-1)^{\lfloor z \rfloor}$. *Now*

$$\mathrm{sign}\,(\omega x_j) = \mathrm{sign}\left(\sin\left(-\pi \sum\nolimits_{i\in[N]} 2^{j-i}\mathbb{1}\{y_i = +1\} - \pi 2^{-N}\right)\right). \tag{3.48}$$

  - *For any $y_j = +1$,*

$$-\sum\nolimits_{i\in[N]} 2^{j-i}\mathbb{1}\{y_i=+1\} - 2^{-N} = -\underbrace{\sum\nolimits_{i<j} 2^{j-i}\mathbb{1}\{y_i=+1\}}_{\textit{even number}} -1 - \underbrace{(\sum\nolimits_{i>j} 2^{j-i}\mathbb{1}\{y_i=+1\} + 2^{-N})}_{\in(0,1)},$$

  So, $\lfloor -\sum_{i\in[N]} 2^{j-i}\mathbb{1}\{y_i=+1\} - 2^{-N}\rfloor = -M$ for an even $M \geq 0$, and $(-1)^M = 1$;

  - *For any $y_i = -1$,*

$$-\sum\nolimits_{i\in[N]} 2^{j-i}\mathbb{1}\{y_i=+1\} - 2^{-N} = -\underbrace{\sum\nolimits_{i<j} 2^{j-i}\mathbb{1}\{y_i=+1\}}_{\textit{even number}} - \underbrace{(\sum\nolimits_{i>j} 2^{j-i}\mathbb{1}\{y_i=+1\} + 2^{-N})}_{\in(0,1)},$$

  So, $\lfloor -\sum_{i\in[N]} 2^{j-i}\mathbb{1}\{y_i=+1\} - 2^{-N}\rfloor = -M$ for an odd $M \geq 0$, and $(-1)^M = -1$;

The above argument for any $N \in \mathbb{N}$, implying that $\mathsf{VCdim}(\mathcal{H}_{\sin}) = \infty$.

The following lemma describes the quantitative relationship between the growth function and the VC dimension.

Lemma 3.14 (Sauer's lemma)**.** *For a hypothesis class $\mathcal{H}$ with $\mathsf{VCdim}(\mathcal{H}) = d$. The following inequalities hold:*

$$\Pi_{\mathcal{H}}(N) \le \sum_{i=0}^{d} \binom{N}{i} \text{ for any integer } N \ge 1, \tag{3.49}$$

$$\text{and } \Pi_{\mathcal{H}}(N) \le (eN/d)^d \in O(N^d) \text{ for any integer } N \ge d. \tag{3.50}$$

Proof. The proofs are those for Theorem 3.17 and Corollary 3.18 of [Moh18]. ■

Theorem 3.15 (VC dimension uniform convergence bounds)**.** *Consider a data space $\mathcal{Z} = \mathcal{X} \times \{+1, -1\}$ and any hypothesis class $\mathcal{H}$ consisting of binary predictions taking values in $\{+1, -1\}$. For any $\delta \in (0, 1)$ and any $\mathcal{D}_{\mathcal{Z}}$, with probability at least $1 - \delta$ over the draw of an iid training set S of size N, the following holds*

$$\mathsf{R}(h) \le \widehat{\mathsf{R}}_S(h) + C\sqrt{\frac{\mathsf{VCdim}(\mathcal{H}) + \log(1/\delta)}{N}} \quad \forall h \in \mathcal{H}, \tag{3.51}$$

where $C > 0$ is a universal constant.

Further reading

Main references are Chapters 2–4 of [Moh18] and Chapters 2–8, 11 of [SSS14]. The basics of probability and concentration inequalities are reviewed in Appendices C and D of [Moh18] and Appendix B of [SSS14]. Classic references for statistical and computational learning theory are [Kea94, Vap00], and [BBL04, BBL05] give excellent summary of recent developments. [Ver18] is a masterpiece on modern high-dimensional probability and concentration inequalities for non-experts.

Disclaimer

This set of notes is preliminary and has not been thoroughly proofread. Typos and factual errors are well expected, and hence use it with caution. Bug reports are very welcome and should be sent to Prof. Ju Sun via `jusun@umn.edu`.

References

[BBL04] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi, *Introduction to statistical learning theory*, Advanced Lectures on Machine Learning, Springer Berlin Heidelberg, 2004, pp. 169–207.

[BBL05] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi, *Theory of classification: a survey of some recent advances*, ESAIM: Probability and Statistics **9** (2005), 323–375.

[Kea94] Michael Kearns, *An introduction to computational learning theory*, MIT Press, Cambridge, Mass, 1994.

[Moh18] Mehryar Mohri, *Foundations of machine learning*, 2 ed., The MIT Press, Cambridge, Massachusetts, 2018.

[SSS14] Shai Ben-David Shai Shalev-Shwartz, *Understanding machine learning*, Cambridge University Press, 2014.

[Vap00] Vladimir Vapnik, *The nature of statistical learning theory*, Springer-Verlag New York, 2000.

[Ver18] Roman Vershynin, *High-dimensional probability*, Cambridge University Pr., 2018.