# Between Chapters 1 and 2

Chapter 2 of the textbook starts discussing the statistcal properties of learning algorithms. Unfortunately, the textbook material jumps in difficulty a LOT between chapters 1 and 2. The purpose of these notes is to help "fill the gap." In particular, the textbook introduces the idea of *generalization bounds* using infinite hypothesis classes in chapter 2. Infinite hypothesis classes are fairly abstract and can be a bit tricky to understand; so we'll start by looking at finite hypothesis classes, which are more concrete and easier to understand.

**Problem 1.** For each hypothesis class below, draw a picture of what a "typical" hypothesis looks like and write the number of elements in the hypothesis class.

$$
\begin{aligned}
\mathcal{H}_{\text{binary}} &= \Big\{ \mathbf{x} \mapsto +1, \mathbf{x} \mapsto -1 \Big\} \\
\mathcal{H}_{\text{axis}} &= \Big\{ \mathbf{x} \mapsto \operatorname{sign}(x_i) : i \in [d] \Big\} \\
\mathcal{H}_{\text{axis2}} &= \Big\{ \mathbf{x} \mapsto \sigma \operatorname{sign}(x_i) : \sigma \in \{+1, -1\}, i \in [d] \Big\} \\
\mathcal{H}_{\text{multiaxis2}} &= \Big\{ \mathbf{x} \mapsto \sum_{j=1}^{d} \sigma_j \operatorname{sign}(x_j) : \sigma_i \in \{+1, -1\}, i \in [d] \Big\} \\
\mathcal{H}_{\text{multiaxis3}} &= \Big\{ \mathbf{x} \mapsto \sum_{j=1}^{d} \sigma_j \operatorname{sign}(x_j) : \sigma_i \in \{+1, 0, -1\}, i \in [d] \Big\}
\end{aligned}
$$

**Problem 2.** One simple idea for selecting a hypothesis in the hypothesis class is to select the best hypothesis on the training data. That is, select

$$g = \arg\min_{h \in \mathcal{H}} E_{\text{in}}(h). \tag{1}$$

The *Try Everything Algorithm* (TEA) is a simple algorithm for computing Formula 1 above. The pseudo-code is shown below:

```
g = None
g_score = -infinity
for h in H:
    h_score = E_in(h)
    if h_score > g_score:
        g = h
        g_score = h_score
```

What is the runtime of the TEA algorithm?

Page 40 in the textbook defines the following terms and equations.

**Definition 1.** The *generalization error* of a hypothesis $g$ is defined to be $|E_{\text{in}}(g) - E_{\text{out}}(g)|$.

**Theorem 1** (Finite Hypothesis Class Generalization)**.** *Let $\mathcal{H}$ be a hypothesis class of size $M$, let $g$ be an arbitrary hypothesis in $\mathcal{H}$ (in particular, $g$ is allowed to be the result of the TEA algorithm), and let $N$ be the size of the dataset. Then we have that for all $\epsilon > 0$,*

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| \geq \epsilon] \leq 2M \exp(-2\epsilon^2 N).$$

*This implies that with probability at least $1 - \delta$,*

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{1}{2N} \log \frac{2M}{\delta}}.$$

*Proof on pages 23-24 of the textbook.*

**Problem 3.** Graph the in-sample error, out-of-sample error, and the generalization error as a function of the hypothesis class size $M$.

This should be done informally/at an “intuition level”; the next two problems explore these trade-offs more formally. This is very similar to Figure 2.3 on page 59 of the textbook.

**Problem 4.** You have a dataset with 100 dimensions and 500 data points. Due to the small amount of training data, you have decided not to split the dataset into training and testing datasets, and instead you will use the entire dataset for training. Your particular application requires that your final model have a generalization error less than 0.1 with probability at least 0.99.

1. Which of the finite hypothesis classes above can you select?

2. Of the hypothesis classes that you CAN select, which one SHOULD you select?

**Problem 5.** Given two finite hypothesis classes $\mathcal{H}_1$ and $\mathcal{H}_2$, we can construct a third hypothesis class

$$\mathcal{H}_3 = \mathcal{H}_1 \cup \mathcal{H}_2. \tag{2}$$

1. Let $g_i$ be the result of the TEA algorithm on hypothesis class $\mathcal{H}_\mathrm{i}$. How does the in-sample error of $g_3$ compare to the in-sample errors of $g_1$ and $g_2$?

2. How does the bound on the generalization error from the FHCG theorem for $\mathcal{H}_3$ compare to the generalization error of $\mathcal{H}_1$ and $\mathcal{H}_2$?