Introduction and learning objectives

This tutorial builds on Lecture 5's analysis of validation-based model selection and Lecture 6's finite-class PAC bounds. The goal is to understand how performance on a sampled dataset supports guarantees on new examples, and how those guarantees change between the realizable and agnostic settings.

By the end of the tutorial, you should be able to:

- Interpret Lecture 6's PAC guarantees: distinguish training error, true error and excess risk, and explain ε and δ.
- Apply the lectures' Hoeffding and union-bound argument to every hypothesis, including one selected using the data.
- Explain Lecture 6's realizable-case argument and compare its $1/\varepsilon$ sample-size dependence with the general $1/\varepsilon^2$ bound.
- Use the lecture bounds to calculate sufficient sample sizes and excess-risk guarantees, explaining the role of repeated inputs.

1 PAC Bound: Realizable Case

Problem 1.1. Each of the 8 inputs can be labelled 0 or 1 independently, so $|\mathcal{H}| = 2^8 = 256$.

Problem 1.2. If $N \geq \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}$, then with probability at least $1 - \delta$ over the i.i.d. draw of $\mathcal{D}$, the learner returns $\hat{h}$ with $L_p(\hat{h}) \leq \varepsilon$. Realizability and empirical risk minimization give $L_{\mathcal{D}}(\hat{h}) = 0$ with probability one, so the gap between training and true error is exactly $L_p(\hat{h})$, and the guarantee bounds this gap by ε.

To see where the bound comes from, call h *bad* if $L_p(h) > \varepsilon$. For any fixed bad h, the N examples are independent, so

$$\Pr_{\mathcal{D}}\big(L_{\mathcal{D}}(h) = 0\big) = (1 - L_p(h))^N \leq (1 - \varepsilon)^N \leq e^{-N\varepsilon},$$

using $1 - \varepsilon \leq e^{-\varepsilon}$. The learner can return a bad hypothesis only if some bad hypothesis has zero training error. A union bound over at most $|\mathcal{H}|$ bad hypotheses gives

$$\Pr_{\mathcal{D}}\big(L_p(\hat{h}) > \varepsilon\big) \leq |\mathcal{H}| e^{-N\varepsilon} \leq \delta \quad \text{when} \quad N \geq \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}.$$

This is the lecture's realizable-case argument. The general bound $N \geq \frac{1}{2\varepsilon^2} \ln \frac{2|\mathcal{H}|}{\delta}$ also applies here, but it does not use the fact that the training error is zero.

Problem 1.3. $\ln \frac{256}{0.05} = \ln 5120 \approx 8.5409$. Then

$$N \geq \frac{1}{0.1} \ln 5120 \approx 85.41, \qquad \boxed{N_{\min} = \lceil 10 \ln 5120 \rceil = 86.}$$

So 86 i.i.d. draws suffice for true error at most 0.1 with probability at least 0.95. This is a sufficient sample size from the bound; it does not say that fewer samples cannot work.

Comparison with the general bound. With the same ε and δ, the general bound requires

$$N \geq \frac{1}{2(0.1)^2} \ln \frac{2 \cdot 256}{0.05} = 50 \ln 10240 \approx 461.7, \qquad \text{so } N \geq 462.$$

The realizable bound grows as $1/\varepsilon$ and the general bound as $1/\varepsilon^2$; halving ε doubles the first and quadruples the second.

Why the bound can exceed 8. The bound counts draws, and draws may repeat inputs. Eight i.i.d. draws need not cover all eight inputs, and the inputs need not be equally likely; a rarely drawn input may not appear at all. If all eight distinct inputs were observed, the noiseless labelling would be identified exactly, but i.i.d. sampling does not guarantee this coverage. The bound holds for every distribution satisfying realizability, including very uneven ones, and the union bound over all 256 hypotheses adds the $\ln |\mathcal{H}|$ term. Both make the bound larger than the number of distinct inputs.

2 Agnostic (Non-Realizable) Case

Recall the notation: $h^* \in \arg\min_{h\in\mathcal{H}} L_p(h)$ is the best hypothesis in the class, $\hat{h} \in \arg\min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$ is the learner's output, and the excess risk is $L_p(\hat{h}) - L_p(h^*)$.

Problem 2.1. With noisy labels, h^* need not have zero true error even though $\mathcal{H}$ contains every deterministic labelling. Fix a hypothesis h (chosen before seeing the data). Its zero-one losses on the N i.i.d. samples are independent, lie in $[0,1]$, and have mean $L_p(h)$. Hoeffding's inequality gives

$$\Pr\left(|L_{\mathcal{D}}(h) - L_p(h)| > \varepsilon\right) \le 2e^{-2N\varepsilon^2}.$$

A union bound over the $|\mathcal{H}|$ hypotheses gives

$$\Pr\left(\exists h \in \mathcal{H} : |L_{\mathcal{D}}(h) - L_p(h)| > \varepsilon\right) \le 2|\mathcal{H}|e^{-2N\varepsilon^2}.$$

Set the right-hand side equal to δ and solve for ε:

$$2|\mathcal{H}|e^{-2N\varepsilon^2} = \delta \iff 2N\varepsilon^2 = \ln\frac{2|\mathcal{H}|}{\delta} \iff \varepsilon = \sqrt{\frac{1}{2N}\ln\frac{2|\mathcal{H}|}{\delta}}.$$

With this ε, the probability that some hypothesis deviates by more than ε is at most δ, so with probability at least $1-\delta$ over $\mathcal{D}$, $|L_{\mathcal{D}}(h) - L_p(h)| \le \varepsilon$ for every $h \in \mathcal{H}$.

Problem 2.2. Hoeffding's inequality cannot be applied to $\hat{h}$ directly: $\hat{h}$ is chosen using $\mathcal{D}$, so its training error is not an average of losses of a hypothesis fixed in advance. The uniform bound from Problem 2.1 avoids this, because it holds for every $h \in \mathcal{H}$ at once, including whichever one the learner selects. On that event,

$$L_p(\hat{h}) \;\le\; L_{\mathcal{D}}(\hat{h}) + \varepsilon \;\le\; L_{\mathcal{D}}(h^*) + \varepsilon \;\le\; L_p(h^*) + 2\varepsilon.$$

The first and third inequalities are the uniform bound applied to $\hat{h}$ and to h^*. The middle inequality holds because $\hat{h}$ minimises the training error. Hence $\boxed{L_p(\hat{h}) \le L_p(h^*) + 2\varepsilon}$ with probability at least $1-\delta$: the excess risk is at most 2ε, where ε is the uniform-deviation tolerance from Problem 2.1.

Problem 2.3.

$$\varepsilon = \sqrt{\frac{1}{400}\ln\frac{2\cdot 256}{0.05}} = \sqrt{\frac{\ln 10240}{400}} \approx 0.1519, \qquad 2\varepsilon \approx 0.3039.$$

So, rounding the excess-risk bound upward, $\boxed{L_p(\hat{h}) \le L_p(h^*) + 0.304}$ with probability at least 0.95. With 200 samples, the bound only guarantees a true error within about 0.30 of the best in the class.