**CIS3990-002: Mathematics of Machine Learning** Fall 2025

Lecture: Application: Generalization Bound

*Date: September 24th, 2025* *Author: Surbhi Goel*

# 1 The Generalization Problem

In the first lecture, we introduced the core problem of generalization in machine learning. We learn a model by minimizing the empirical risk, $\hat{R}(f)$, on our training data, but our true goal is for the model to perform well on new, unseen data, which is measured by the true risk, $R(f)$. Recall their definitions:

$$\hat{R}(f) = \frac{1}{N}\sum_{i=1}^{N} \ell\big(f(x_i), y_i\big) \quad \text{and} \quad R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell\big(f(x), y\big)\big],$$

where $\ell(\cdot,\cdot)$ is the loss function.

In this lecture, we will prove our first formal guarantee that minimizing the former helps us do well on the latter. Our analysis will center on two key functions: the one we learn from data, $\hat{f} \in \mathcal{F}$, and the best possible one, $f^* \in \mathcal{F}$. Formally, $\hat{f}$ is the function that minimizes the empirical risk (what we can measure), and $f^*$ is the function that minimizes the true risk (what we wish we could measure):

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \hat{R}(f) \quad \text{and} \quad f^* = \arg\min_{f\in\mathcal{F}} R(f)$$

Our goal is to formally bound the excess risk, $R(\hat{f}) - R(f^*)$, which measures how much worse our learned model's true performance is compared to the best possible true performance we could hope for from our hypothesis class.

# 2 From Generalization to Uniform Convergence

The key to bounding the excess risk is to relate it to a quantity we can control with our concentration tools. We can do this with a decomposition that adds and subtracts the empirical risks:

$$R(\hat{f}) - R(f^*) = \underbrace{[R(\hat{f}) - \hat{R}(\hat{f})]}_{\text{Gap for } \hat{f}} + \underbrace{[\hat{R}(\hat{f}) - \hat{R}(f^*)]}_{\text{Empirical gap } (\leq 0)} + \underbrace{[\hat{R}(f^*) - R(f^*)]}_{\text{Gap for } f^*}$$

The second term, the empirical gap, is less than or equal to zero by the definition of $\hat{f}$ as the empirical risk minimizer. The other two terms are the generalization gaps for our learned model and the best model in the class, respectively. By dropping the middle term and taking absolute values, we can bound the excess risk:

$$R(\hat{f}) - R(f^*) \leq |R(\hat{f}) - \hat{R}(\hat{f})| + |\hat{R}(f^*) - R(f^*)|$$

Both terms on the right are individual instances of the deviation between true and empirical risk. Therefore, both are bounded by the worst-case, or supremum, deviation over the entire function class. This gives us:

$$R(\hat{f}) - R(f^*) \leq \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| + \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)|$$

This simplifies to the crucial inequality:

$$R(\hat{f}) - R(f^*) \leq 2 \cdot \sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)|$$

This derivation is powerful: it reduces the problem of bounding the excess risk to a new problem of bounding the worst-case deviation between the empirical and true risks. The concentration inequalities we developed in the last lectures guarantee that for any single function $f$, its empirical risk $\hat{R}(f)$ will be close to its true risk $R(f)$ with high probability. However, our new goal requires something stronger: we must show that all empirical risks for all functions in the class $\mathcal{F}$ are simultaneously close to their true risks. This stronger property is known as uniform convergence.

## 3 A Generalization Bound for a Finite Hypothesis Class

**Theorem 1** (Uniform Convergence for Finite Classes)**.** *Let $\mathcal{F}$ be a finite hypothesis class. Assume the loss function $\ell$ is bounded such that $0 \leq \ell(f(x), y) \leq B$ for all $f \in \mathcal{F}$ and all $(x, y)$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of the $N$ training examples, the following holds:*

$$\sup_{f \in \mathcal{F}} \left| \hat{R}(f) - R(f) \right| \leq B \sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2N}}$$

This theorem provides the uniform convergence guarantee we need. It says that the worst-case deviation between the empirical and true risks shrinks as we get more data ($N$). It also depends on the complexity of our model ($|\mathcal{F}|$) and our desired confidence ($1 - \delta$). Combining this theorem with our excess risk inequality gives our final result.

**Corollary 2** (Excess Risk Bound)**.** *Under the same assumptions, with probability at least $1 - \delta$:*

$$R(\hat{f}) - R(f^*) \leq B \sqrt{\frac{2 \log(2|\mathcal{F}|/\delta)}{N}}$$

*Proof of Theorem 1.* The proof follows a two-step argument. First we apply Hoeffding's inequality to a single function, then we use the union bound to extend the guarantee to all functions in the class.

**Step 1: Hoeffding's Inequality.** Let's fix a single function $f \in \mathcal{F}$. The empirical risk is defined as the average loss on the training set:

$$\hat{R}(f) = \frac{1}{N} \sum_{i=1}^{N} \ell(f(x_i), y_i)$$

Each term $\ell(f(x_i), y_i)$ in the sum is a random variable, because it depends on the randomly drawn data point $(x_i, y_i)$. Let's call these random variables $Z_i = \ell(f(x_i), y_i)$. Since each data point is drawn i.i.d., the variables $Z_1, \ldots, Z_N$ are also i.i.d.

By our assumption, the loss is bounded between 0 and $B$, so each $Z_i \in [0, B]$. Furthermore, the expectation of each $Z_i$ is the true risk of $f$:

$$\mathbb{E}[Z_i] = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)] = R(f)$$

So, for a fixed $f$, the empirical risk $\hat{R}(f)$ is a sample mean of $N$ i.i.d. bounded random variables. This is exactly the setup for Hoeffding's Inequality. Applying it gives us a direct bound on the probability of $\hat{R}(f)$ deviating from its true mean $R(f)$ by more than some $\epsilon > 0$:

$$\mathbb{P}\left(\left|\hat{R}(f) - R(f)\right| > \epsilon\right) \leq 2\exp\left(-\frac{2N\epsilon^2}{B^2}\right)$$

**Step 2: Union Bound.** Now, we need to extend this from a guarantee for a single function to a guarantee for all functions in $\mathcal{F}$ simultaneously. We want to bound the probability that the worst-case deviation is large. The union bound is the tool for this. The event that the supremum is large is the union of the events that any individual deviation is large:

$$\begin{aligned}
\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\hat{R}(f) - R(f)\right| > \epsilon\right) &= \mathbb{P}\left(\bigcup_{f\in\mathcal{F}}\left\{\left|\hat{R}(f) - R(f)\right| > \epsilon\right\}\right) \\
&\leq \sum_{f\in\mathcal{F}} \mathbb{P}\left(\left|\hat{R}(f) - R(f)\right| > \epsilon\right) \\
&\leq \sum_{f\in\mathcal{F}} 2\exp\left(-\frac{2N\epsilon^2}{B^2}\right) \\
&= |\mathcal{F}| \cdot 2\exp\left(-\frac{2N\epsilon^2}{B^2}\right)
\end{aligned}$$

This final expression bounds the "probability of failure." We want this probability to be at most $\delta$. So, we set this expression equal to $\delta$ and solve for $\epsilon$:

$$\delta = 2|\mathcal{F}|\exp\left(-\frac{2N\epsilon^2}{B^2}\right) \implies \epsilon = B\sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2N}}$$

This tells us that with probability at least $1 - \delta$, the worst-case deviation for any function in our class is bounded by this value of $\epsilon$. This completes the proof. □

## 4 Interpreting the Bound

Let's make this bound concrete by revisiting the spam detection example from Lecture 1. Suppose our goal is to learn a simple rule to classify emails.

**Setup.** Our input $x$ is an email. We process it by checking for the presence of words from a dictionary of $d = 5000$ common English words. Our hypothesis class $\mathcal{F}$ consists of simple single keyword rules. A function $f_w \in \mathcal{F}$ is defined by a single keyword $w$ from our dictionary. The rule is:

$$f_w(x) = \begin{cases} \text{spam} & \text{if keyword } w \text{ is in the email} \\ \text{ham} & \text{otherwise} \end{cases}$$

This is a finite hypothesis class where each function corresponds to one keyword. The size of the class is therefore $|\mathcal{F}| = d = 5000$. Our learning algorithm will test all 5000 keyword rules on the training data and pick the one with the lowest empirical error.

**Applying the Bound.** How much data do we need to be confident that our chosen rule will generalize? Let's use the theorem to find out. We will use the 0-1 loss for classification, so our loss is bounded by $B = 1$. Let's say we want to be 95% confident (so $\delta = 0.05$) that the true error of our chosen rule is no more than 5% worse than its empirical error on the training set (so we are bounding $|\hat{R}(f) - R(f)|$ by $\epsilon = 0.05$ for all $f$).

We use the bound from Theorem 1 and solve for $N$:

$$\epsilon \leq B\sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2N}} \implies N \geq \frac{B^2}{2\epsilon^2}\log\left(\frac{2|\mathcal{F}|}{\delta}\right)$$

Plugging in our values:

$$N \geq \frac{1^2}{2\cdot(0.05)^2}\log\left(\frac{2\cdot 5000}{0.05}\right) \geq \frac{1}{0.005}\log\left(\frac{10000}{0.05}\right) \geq 200\cdot\log(200000) \approx 200\cdot 12.2 = 2440.$$

This calculation tells us that we need about 2,440 labeled emails to guarantee, with 95% confidence, that the empirical error of every single-keyword rule is within 5% of its true error. This ensures that the rule we pick, $\hat{f}$, which has the lowest empirical error, will also have a true error close to that value, preventing a situation where we overfit to a keyword that just happened to work well on our specific training sample.

**Takeaways.** This example illustrates the trade-offs in generalization. If we had used a more complex hypothesis class, such as rules involving pairs of keywords, the size of the class would explode to $|\mathcal{F}| = \binom{5000}{2} \approx 12.5$ million. How much more data would we need to achieve the same $(\epsilon, \delta)$ guarantee?

$$N \geq \frac{1}{2\cdot(0.05)^2}\log\left(\frac{2\cdot 12.5\times 10^6}{0.05}\right) \geq 200\cdot\log(5\times 10^8) \approx 200\cdot 20.0 = 4000.$$

So, while the model complexity increased by a factor of 2,500 (from 5,000 to 12.5 million), the sample size required only increased from about 2,440 to 4,000. This is a direct consequence of the $\log(|\mathcal{F}|)$ term in the bound: the data requirement scales only logarithmically with the hypothesis class size.