Statistical Learning Theory

SDS 391P.6, Spring 2026

Pratik Patil

These notes are a work in progress and are provided as-is for instructional purposes only. They are not (yet) at the level of a scholarly document. In particular, the notes draw from various sources and do not (yet) have sufficient references to the original sources. Additionally, almost surely the notes have errors and they are only probably approximately correct. The notes will be updated regularly as the course progresses. Last updated: 2026-04-05.

1 Motivation

In the last two lectures, we developed empirical process tools for controlling quantities of the form

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right|.$$

First, we used Dudley's inequality and metric entropy to prove uniform laws of large numbers for structured function classes such as Lipschitz functions. Then we introduced symmetrization, Rademacher complexity, and VC dimension, which allow us to replace metric complexity by combinatorial complexity when the class is Boolean.

The natural next step is to use these tools in statistical learning. At a high level, a learning algorithm tries to pick a function h from a hypothesis class $\mathcal{H}$ using training data, with the goal that h performs well on new unseen samples. The central question is: How much worse is the predictor chosen from the data than the best predictor in the hypothesis class?

This question leads to the notion of *generalization error*, or more precisely *excess risk*. The key point is that excess risk is controlled by a uniform empirical process. So the abstract tools we developed earlier become concrete learning-theoretic bounds.

In this lecture, we begin with the general supervised learning framework and the empirical risk minimization (ERM) principle. We then prove two basic generalization results:

- a finite-class bound, where the complexity is $\log |\mathcal{H}|$;
- a VC-based bound for Boolean classification, where the complexity is $\mathrm{VC}(\mathcal{H})$.

We end with a nonparametric regression example that connects back to the Lipschitz-function empirical process bounds from the previous lecture.

2 Supervised learning and empirical risk minimization

Let (X, Y) be a random pair taking values in $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the feature space and $\mathcal{Y}$ is the label space. We observe training data

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

which are i.i.d. copies of (X, Y).

A *hypothesis* is a measurable function $h : \mathcal{X} \to \mathcal{Y}$ (or into some prediction space). We do not search over all possible functions, but only over a prescribed hypothesis class $\mathcal{H}$.

To measure performance, fix a loss function

$$\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty).$$

For a hypothesis $h \in \mathcal{H}$, define the *risk*

$$R(h) := \mathbb{E}[\ell(h(X), Y)].$$

This is the population prediction error. Since the distribution of (X, Y) is unknown, $R(h)$ is not directly computable.

The observable proxy is the *empirical risk*

$$R_n(h) := \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i).$$

The population-optimal hypothesis in the class is

$$h^\star \in \arg\min_{h \in \mathcal{H}} R(h),$$

while the empirical risk minimizer is

$$\hat{h}_n \in \arg\min_{h \in \mathcal{H}} R_n(h).$$

This is the *empirical risk minimization principle*: replace the unknown population risk by the empirical one, and minimize that instead.

Measures of errors. The quantity $R(h^\star)$ is the best risk achievable within the class $\mathcal{H}$. It reflects the *approximation* quality of the class. The additional error

$$R(\hat{h}_n) - R(h^\star)$$

is the *excess risk*: the price we pay for learning from finitely many samples.

Thus

$$R(\hat{h}_n) = R(h^\star) + \big(R(\hat{h}_n) - R(h^\star)\big),$$

which separates the total prediction error into an approximation term and a statistical term.

3 Excess risk control

The main deterministic observation is that ERM can only be bad if empirical risks fail to approximate true risks uniformly over the hypothesis class.

Proposition 3.1 (Excess risk bound)**.** *For any hypothesis class $\mathcal{H}$,*

$$R(\hat{h}_n) - R(h^\star) \leqslant 2 \sup_{h \in \mathcal{H}} |R_n(h) - R(h)|.$$

Proof. Let

$$\Delta := \sup_{h \in \mathcal{H}} |R_n(h) - R(h)|.$$

Then for every $h \in \mathcal{H}$,

$$R(h) \leqslant R_n(h) + \Delta \qquad \text{and} \qquad R_n(h) \leqslant R(h) + \Delta.$$

Applying these inequalities first to $\hat{h}_n$, then using that $\hat{h}_n$ minimizes R_n, and finally applying the second inequality to $h^\star$, we get

$$R(\hat{h}_n) \leqslant R_n(\hat{h}_n) + \Delta \leqslant R_n(h^\star) + \Delta \leqslant R(h^\star) + 2\Delta.$$

Subtract $R(h^\star)$ from both sides. □

So the whole learning problem reduces to bounding

$$\sup_{h \in \mathcal{H}} |R_n(h) - R(h)|.$$

This is exactly the supremum of an empirical process indexed by the loss class.

Indeed, define

$$\mathcal{L} := \{(x, y) \mapsto \ell(h(x), y) : \ h \in \mathcal{H}\}.$$

Then

$$\sup_{h \in \mathcal{H}} |R_n(h) - R(h)| = \sup_{g \in \mathcal{L}} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i, Y_i) - \mathbb{E}g(X, Y) \right|.$$

This is the bridge between empirical process theory and learning theory.

4 Warm-up: finite hypothesis classes

Before using VC dimension, it is helpful to see the simplest case where $\mathcal{H}$ is finite.

Assume throughout this section that the loss is bounded:

$$0 \leqslant \ell(h(x), y) \leqslant 1 \qquad \text{for all } h, x, y.$$

This includes the 0-1 loss used in classification.

Fix $h \in \mathcal{H}$. Then the random variables

$$Z_i(h) := \ell(h(X_i), Y_i)$$

are i.i.d. in $[0, 1]$, so Hoeffding's inequality gives, for every $t > 0$,

$$\mathbb{P}\{|R_n(h) - R(h)| \geqslant t\} \leqslant 2e^{-2nt^2}.$$

Applying the union bound over $h \in \mathcal{H}$, we obtain

$$\mathbb{P}\left\{ \sup_{h \in \mathcal{H}} |R_n(h) - R(h)| \geqslant t \right\} \leqslant 2|\mathcal{H}|\, e^{-2nt^2}.$$

Combining this with Proposition 3.1 yields the following.

Theorem 4.1 (Finite-class generalization bound)**.** *Assume* $|\mathcal{H}| < \infty$ *and* $0 \leqslant \ell \leqslant 1$*. Then for every* $\delta \in (0,1)$*, with probability at least* $1-\delta$*,*

$$R(\hat{h}_n) - R(h^\star) \leqslant 2\sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2n}}.$$

In particular, the complexity of a finite hypothesis class enters only through $\log|\mathcal{H}|$*.*

This already shows an important point: even a very large finite class can be learnable, as long as $\log|\mathcal{H}|$ is moderate compared with n.

Of course, many natural hypothesis classes are infinite. To go beyond finite classes, we need a more structural notion of complexity.

5 Binary classification and VC dimension

We now specialize to Boolean classification. Assume

$$\mathcal{Y} = \{0,1\}, \qquad \mathcal{H} \subseteq \{h : \mathcal{X} \to \{0,1\}\},$$

and use the 0-1 loss

$$\ell(h(x), y) := \mathbf{1}_{\{h(x)\neq y\}}.$$

Then the risk is simply the misclassification probability

$$R(h) = \mathbb{P}\{h(X) \neq Y\},$$

and the empirical risk is the training misclassification rate

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{h(X_i)\neq Y_i\}}.$$

To apply the empirical process bounds from the previous lecture, define the loss class

$$\mathcal{L} := \left\{(x,y) \mapsto \mathbf{1}_{\{h(x)\neq y\}} : \ h \in \mathcal{H}\right\}.$$

The key observation is that $\mathcal{L}$ has the same combinatorial complexity as $\mathcal{H}$.

Lemma 5.1 (Loss class has the same trace complexity)**.** *For every sample* $(x_1,y_1),\dots,(x_n,y_n)$*,*

$$|\mathcal{L}|_{(x_i,y_i)_{i=1}^n}| = |\mathcal{H}|_{x_{1:n}}|.$$

Consequently,

$$\Pi_{\mathcal{L}}(n) \leqslant \Pi_{\mathcal{H}}(n) \qquad \textit{and} \qquad \mathrm{VC}(\mathcal{L}) \leqslant \mathrm{VC}(\mathcal{H}).$$

Proof. For any $h \in \mathcal{H}$, the trace of the corresponding loss function on the labeled sample is

$$\big(\mathbf{1}_{\{h(x_1)\neq y_1\}}, \dots, \mathbf{1}_{\{h(x_n)\neq y_n\}}\big).$$

This vector is obtained from the label vector

$$(h(x_1), \dots, h(x_n))$$

by coordinatewise XOR with the fixed vector $(y_1,\dots,y_n)$. Since XOR with a fixed binary vector is a bijection on $\{0,1\}^n$, the number of distinct traces is preserved. □

Now we can invoke the VC-based empirical process bound from the previous lecture.

Theorem 5.2 (VC generalization bound)**.** *Let $\mathcal{H}$ be a class of Boolean hypotheses with finite VC dimension*

$$d := \mathrm{VC}(\mathcal{H}) < \infty.$$

Assume $n \geqslant d$. Then

$$\mathbb{E}\big[R(\hat{h}_n) - R(h^\star)\big] \leqslant C\sqrt{\frac{d\log(en/d)}{n}},$$

where $C > 0$ is an absolute constant. In particular,

$$\mathbb{E}\big[R(\hat{h}_n) - R(h^\star)\big] \lesssim \sqrt{\frac{d\log n}{n}}.$$

Proof. By Proposition 3.1,

$$R(\hat{h}_n) - R(h^\star) \leqslant 2\sup_{g\in\mathcal{L}}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i) - \mathbb{E}g(X, Y)\right|.$$

Take expectations. The preliminary VC bound for empirical processes from the previous lecture gives

$$\mathbb{E}\sup_{g\in\mathcal{L}}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i) - \mathbb{E}g(X, Y)\right| \leqslant C\sqrt{\frac{\mathrm{VC}(\mathcal{L})\log(en/\mathrm{VC}(\mathcal{L}))}{n}}.$$

Using Lemma 5.1, we have $\mathrm{VC}(\mathcal{L}) \leqslant \mathrm{VC}(\mathcal{H}) = d$. Absorb the factor 2 into the constant. □

This theorem says that the sample complexity of Boolean classification is controlled by the VC dimension of the hypothesis class.

Interpretation. If $\mathcal{H}$ is too rich, then ERM can overfit the training data and fail to generalize. Finite VC dimension rules out excessive combinatorial richness. The theorem quantifies this by saying that roughly

$$n \gg d$$

samples are needed, up to the logarithmic factor coming from our current empirical process bound.

Remark 5.3 (Sharper bounds)**.** Using the sharper VC entropy estimate discussed after the previous lecture, one can improve the theorem to

$$\mathbb{E}\big[R(\hat{h}_n) - R(h^\star)\big] \lesssim \sqrt{\frac{d}{n}}.$$

We do not pursue that refinement here.

6 A parametric classification example: linear classification

Consider the class of half-space classifiers in $\mathbb{R}^d$:

$$\mathcal{H} = \left\{ x \mapsto \mathbf{1}_{\{\langle w,x\rangle + b \geqslant 0\}} : \ w \in \mathbb{R}^d, \ b \in \mathbb{R} \right\}.$$

As we discussed earlier, this class has VC dimension

$$\mathrm{VC}(\mathcal{H}) = d + 1.$$

Therefore Theorem 5.2 yields

$$\mathbb{E}\big[R(\hat{h}_n) - R(h^\star)\big] \leqslant C\sqrt{\frac{(d+1)\log(en/(d+1))}{n}}.$$

So, at least at the level of sample complexity, linear classification in $\mathbb{R}^d$ needs on the order of d labeled examples, up to logarithmic factors.

This matches the intuition that linear classifiers have about d effective degrees of freedom.

7 A nonparametric regression example: Lipschitz regression

To see that the same philosophy also applies beyond Boolean classes, let us return to the Lipschitz-function class from the previous lecture.

Assume now that $\Omega = [0,1]$, that the label $Y \in [0,1]$, and that the hypothesis class is

$$\mathcal{H}_L := \{h : [0,1] \to [0,1] : \|h\|_{\mathrm{Lip}} \leqslant L\}.$$

We use squared loss

$$\ell(h(x), y) := (h(x) - y)^2.$$

For each $h \in \mathcal{H}_L$, define the loss function

$$g_h(x,y) := (h(x) - y)^2.$$

If $h_1, h_2 \in \mathcal{H}_L$, then pointwise

$$|g_{h_1}(x,y) - g_{h_2}(x,y)| = |(h_1(x) - y)^2 - (h_2(x) - y)^2| \leqslant 2|h_1(x) - h_2(x)|.$$

So the map $h \mapsto g_h$ is 2-Lipschitz in the uniform norm on function classes:

$$\|g_{h_1} - g_{h_2}\|_\infty \leqslant 2\|h_1 - h_2\|_\infty.$$

Therefore

$$\mathcal{N}(\mathcal{L}, \|\cdot\|_\infty, \varepsilon) \leqslant \mathcal{N}(\mathcal{H}_L, \|\cdot\|_\infty, \varepsilon/2),$$

where $\mathcal{L} = \{g_h : h \in \mathcal{H}_L\}$ is the loss class.

From the previous lecture, we know that

$$\log \mathcal{N}(\mathcal{H}_L, \|\cdot\|_\infty, \varepsilon) \lesssim \frac{L}{\varepsilon} \qquad (0 < \varepsilon \leqslant 1).$$

Applying the empirical-process bound in the sup norm therefore gives

$$\mathbb{E} \sup_{h \in \mathcal{H}_L} |R_n(h) - R(h)| \lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\frac{L}{\varepsilon}}\, d\varepsilon \lesssim \sqrt{\frac{L}{n}}.$$

By Proposition 3.1, we conclude

$$\mathbb{E}\big[R(\hat{h}_n) - R(h^\star)\big] \lesssim \sqrt{\frac{L}{n}}.$$

So in one dimension, the excess risk of ERM over a Lipschitz hypothesis class decays at the same $n^{-1/2}$ scale as the uniform empirical process bound.

Remark 7.1 (Extension to higher dimensions)**.** For Lipschitz functions on $[0,1]^d$, the metric entropy is much larger: roughly

$$\log \mathcal{N}(\mathcal{H}_L, \|\cdot\|_\infty, \varepsilon) \asymp \varepsilon^{-d}.$$

Then the Dudley integral diverges at 0 once $d \geqslant 2$, and the simple one-scale argument breaks down. This is one manifestation of the curse of dimensionality in nonparametric learning.

8 Discussion

This lecture shows how empirical process theory enters statistical learning: the excess risk of ERM is controlled by a uniform empirical process over the loss class. This provides us a concrete mathematical expression of the classical fit-versus-complexity tradeoff.

If the hypothesis class $\mathcal{H}$ is too small, then $R(h^\star)$ may be large: the class cannot approximate the true relationship well. If the class is too large, then the uniform empirical error

$$\sup_{h \in \mathcal{H}} |R_n(h) - R(h)|$$

becomes large, so ERM may overfit.

Thus learning theory is largely about finding the right notion of complexity for $\mathcal{H}$, and then proving that the empirical risk is a good proxy for the true risk at the corresponding sample size.

For finite Boolean classes, the relevant complexity is $\log|\mathcal{H}|$. For infinite Boolean classes, VC dimension $\mathrm{VC}(\mathcal{H})$ provides a robust combinatorial replacement. For richer real-valued classes, one often uses metric entropy, Rademacher complexity, covering numbers, or more refined chaining functionals.

Source material

Parts of this lecture are based on references: Vershynin (2018); Tropp (2023), in addition to the author's accumulated experience working on related topics.

References

Tropp, J. A. (2023). Probability in high dimensions. Caltech CMS Lecture Notes 2021-01.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.