Week 3: Transductive, i.i.d., and PAC learning

The fixed-pool model, population guarantees, and the language of learnability

Tianhao Wang
tianhaowang@ucsd.edu

UCSD · Spring 2026
DSC 190/291 Topics: Learning Theory

Contents

Where We Left Off

Week 2. Online transductive learning:

- Fixed pool of n points.

Where We Left Off

Week 2. Online transductive learning:

- Fixed pool of n points.
- Labels revealed in an *adversarial* order.

Where We Left Off

Week 2. Online transductive learning:

- Fixed pool of n points.
- Labels revealed in an *adversarial* order.
- We counted *mistakes*. Halving achieved $\log_2 \Gamma_{\mathcal{H}}(n)$.

Where We Left Off

Week 2. Online transductive learning:

- Fixed pool of n points.
- Labels revealed in an *adversarial* order.
- We counted *mistakes*. Halving achieved $\log_2 \Gamma_{\mathcal{H}}(n)$.

Recall: $\Gamma_{\mathcal{H}}(n)$ is the *growth function*, the maximum number of distinct label patterns $\mathcal{H}$ can realize on any n points:

$$\Gamma_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n} |\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}|.$$

Where We Left Off

Week 2. Online transductive learning:

- Fixed pool of n points.
- Labels revealed in an *adversarial* order.
- We counted *mistakes*. Halving achieved $\log_2 \Gamma_{\mathcal{H}}(n)$.

Recall: $\Gamma_{\mathcal{H}}(n)$ is the *growth function*, the maximum number of distinct label patterns $\mathcal{H}$ can realize on any n points:

$$\Gamma_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n} |\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}|.$$

The Halving bound is worst-case and pessimistic by design.

From Adversarial to Random Splits

Natural question. If the split is *random* instead of adversarial, can we give an *average-error* guarantee rather than a worst-case mistake count?

From Adversarial to Random Splits

Natural question. If the split is *random* instead of adversarial, can we give an *average-error* guarantee rather than a worst-case mistake count?

This week's pivot. Same fixed pool as Week 2, but

- the train/test split is drawn *uniformly at random*, and

From Adversarial to Random Splits

Natural question. If the split is *random* instead of adversarial, can we give an *average-error* guarantee rather than a worst-case mistake count?

This week's pivot. Same fixed pool as Week 2, but

- the train/test split is drawn *uniformly at random*, and
- we measure *error rate* on the held-out portion, not mistake count.

From Adversarial to Random Splits

Natural question. If the split is *random* instead of adversarial, can we give an *average-error* guarantee rather than a worst-case mistake count?

This week's pivot. Same fixed pool as Week 2, but

- the train/test split is drawn *uniformly at random*, and
- we measure *error rate* on the held-out portion, not mistake count.

The classical (Vapnik) transductive setting: no distributional assumption on the pool, and the cleanest place to meet the combinatorial core.

Transductive Learning

Classical Transductive Learning Setup

Protocol (notation: $[k] = \{1, ..., k\}$)

- Fix a full labeled sample

  $S_{n+u} = ((x_1, y_1), ..., (x_{n+u}, y_{n+u}))$.

Classical Transductive Learning Setup

Protocol (notation: $[k] = \{1, ..., k\}$)

- Fix a full labeled sample

  $$S_{n+u} = ((x_1, y_1), ..., (x_{n+u}, y_{n+u})).$$

- Reveal only the unlabeled pool

  $$x_1, ..., x_{n+u}.$$

Classical Transductive Learning Setup

Protocol (notation: $[k] = \{1, ..., k\}$)

- Fix a full labeled sample

$$S_{n+u} = ((x_1, y_1), ..., (x_{n+u}, y_{n+u})).$$

- Reveal only the unlabeled pool

$$x_1, ..., x_{n+u}.$$

- Draw a training index set $T \subseteq [n+u]$ uniformly among all size-n subsets, and reveal y_i for $i \in T$.

Classical Transductive Learning Setup

Protocol (notation: $[k] = \{1, ..., k\}$)

- Fix a full labeled sample

$$S_{n+u} = ((x_1, y_1), ..., (x_{n+u}, y_{n+u})).$$

- Reveal only the unlabeled pool

$$x_1, ..., x_{n+u}.$$

- Draw a training index set $T \subseteq [n+u]$ uniformly among all size-n subsets, and reveal y_i for $i \in T$.
- Predict y_i for each $i \in U := [n+u] \setminus T$ (the test indices, with $|U| = u$).

Classical Transductive Learning Setup

Protocol (notation: $[k] = \{1, ..., k\}$)

- Fix a full labeled sample

$$S_{n+u} = ((x_1, y_1), ..., (x_{n+u}, y_{n+u})).$$

- Reveal only the unlabeled pool

$$x_1, ..., x_{n+u}.$$

- Draw a training index set $T \subseteq [n+u]$ uniformly among all size-n subsets, and reveal y_i for $i \in T$.
- Predict y_i for each $i \in U := [n+u] \setminus T$ (the test indices, with $|U| = u$).

What is random?

The pool itself is fixed.

The randomness is only in the uniform train/test split T.

The Transductive Learning Goal

The learner outputs a hypothesis $\hat{h}$, chosen based on the revealed examples $\{(x_i, y_i) : i \in T\}$.

The Transductive Learning Goal

The learner outputs a hypothesis $\hat{h}$, chosen based on the revealed examples $\{(x_i, y_i) : i \in T\}$.

Goal. Small error on the test points:

$$L_U\left(\hat{h}\right) = \frac{1}{u} \sum_{i \in U} \mathbf{1}\left[\hat{h}(x_i) \neq y_i\right].$$

The Transductive Learning Goal

The learner outputs a hypothesis $\hat{h}$, chosen based on the revealed examples $\{(x_i, y_i) : i \in T\}$.

Goal. Small error on the test points:

$$L_U\big(\hat{h}\big) = \frac{1}{u} \sum_{i \in U} \mathbf{1}\big[\hat{h}(x_i) \neq y_i\big].$$

Obstacle. We never see the labels on U, so $L_U(h)$ is unknown for every h.

The Transductive Learning Goal

The learner outputs a hypothesis $\hat{h}$, chosen based on the revealed examples $\{(x_i, y_i) : i \in T\}$.

Goal. Small error on the test points:

$$L_U\left(\hat{h}\right) = \frac{1}{u} \sum_{i \in U} \mathbf{1}\left[\hat{h}(x_i) \neq y_i\right].$$

Obstacle. We never see the labels on U, so $L_U(h)$ is unknown for every h.

Strategy. Use the training error as a proxy:

$$L_T(h) = \frac{1}{n} \sum_{i \in T} \mathbf{1}[h(x_i) \neq y_i].$$

Pick $\hat{h}$ that does well on the labels we see, and hope it also does well on the labels we don't.

Why the Strategy Could Work

For a fixed h, both $L_T(h)$ and $L_U(h)$ are random (they depend on the split). They average to the deterministic **pool error**

$$L(h) = \frac{1}{n+u} \sum_{i=1}^{n+u} \mathbf{1}[h(x_i) \neq y_i],$$

which depends only on the fixed pool.

Why the Strategy Could Work

For a fixed h, both $L_T(h)$ and $L_U(h)$ are random (they depend on the split). They average to the deterministic **pool error**

$$L(h) = \frac{1}{n+u} \sum_{i=1}^{n+u} \mathbf{1}[h(x_i) \neq y_i],$$

which depends only on the fixed pool.

Since T and U partition $[n+u]$,

$$L(h) = \frac{n}{n+u} L_T(h) + \frac{u}{n+u} L_U(h)$$

Why the Strategy Could Work

For a fixed h, both $L_T(h)$ and $L_U(h)$ are random (they depend on the split). They average to the deterministic **pool error**

$$L(h) = \frac{1}{n+u} \sum_{i=1}^{n+u} \mathbf{1}[h(x_i) \neq y_i],$$

which depends only on the fixed pool.

Since T and U partition $[n+u]$,

$$L(h) = \frac{n}{n+u} L_T(h) + \frac{u}{n+u} L_U(h)$$

and rearranging,

$$L_U(h) - L_T(h) = \frac{n+u}{u} (L(h) - L_T(h)).$$

Why the Strategy Could Work

For a fixed h, both $L_T(h)$ and $L_U(h)$ are random (they depend on the split). They average to the deterministic **pool error**

$$L(h) = \frac{1}{n+u} \sum_{i=1}^{n+u} \mathbf{1}[h(x_i) \neq y_i],$$

which depends only on the fixed pool.

Since T and U partition $[n+u]$,

$$L(h) = \frac{n}{n+u} L_T(h) + \frac{u}{n+u} L_U(h)$$

and rearranging,

$$L_U(h) - L_T(h) = \frac{n+u}{u} (L(h) - L_T(h)).$$

The right-hand side involves only one random quantity, $L_T(h)$, compared against the fixed $L(h)$. Make it small, and the left-hand side shrinks with it.

Concentration for a Fixed Hypothesis: Setup

Fix any hypothesis h, chosen **before** the random split T is drawn.

Concentration for a Fixed Hypothesis: Setup

Fix any hypothesis h, chosen **before** the random split T is drawn.

Define the error indicators on the pool:

$$z_i(h) := \mathbf{1}[h(x_i) \neq y_i] \in \{0, 1\}, \quad i \in [n + u].$$

Since the pool is fixed, these are $n + u$ fixed numbers.

Concentration for a Fixed Hypothesis: Setup

Fix any hypothesis h, chosen **before** the random split T is drawn.

Define the error indicators on the pool:

$$z_i(h) := \mathbf{1}[h(x_i) \neq y_i] \in \{0, 1\}, \quad i \in [n + u].$$

Since the pool is fixed, these are $n + u$ fixed numbers.

In these terms:

- $L(h) = \frac{1}{n+u} \sum_{i=1}^{n+u} z_i(h)$ is the average over the whole population of $n + u$ indicators.

Concentration for a Fixed Hypothesis: Setup

Fix any hypothesis h, chosen **before** the random split T is drawn.

Define the error indicators on the pool:

$$z_i(h) := \mathbf{1}[h(x_i) \neq y_i] \in \{0, 1\}, \quad i \in [n + u].$$

Since the pool is fixed, these are $n + u$ fixed numbers.

In these terms:

- $L(h) = \frac{1}{n+u} \sum_{i=1}^{n+u} z_i(h)$ is the average over the whole population of $n + u$ indicators.
- $L_T(h) = \frac{1}{n} \sum_{i \in T} z_i(h)$ is the average over a uniformly random size-n subset of that population.

Concentration for a Fixed Hypothesis: Setup

Fix any hypothesis h, chosen **before** the random split T is drawn.

Define the error indicators on the pool:

$$z_i(h) := \mathbf{1}[h(x_i) \neq y_i] \in \{0, 1\}, \quad i \in [n + u].$$

Since the pool is fixed, these are $n + u$ fixed numbers.

In these terms:

- $L(h) = \frac{1}{n+u} \sum_{i=1}^{n+u} z_i(h)$ is the average over the whole population of $n + u$ indicators.
- $L_T(h) = \frac{1}{n} \sum_{i \in T} z_i(h)$ is the average over a uniformly random size-n subset of that population.

Controlling $|L_T(h) - L(h)|$ is a classical "sample mean vs. population mean" question, with the twist that the sample is drawn *without replacement*.

Hoeffding Inequality for Sampling Without Replacement

The probability fact we need, stated abstractly (no learning yet):

Hoeffding Inequality for Sampling Without Replacement

The probability fact we need, stated abstractly (no learning yet):

Hoeffding inequality for sampling without replacement

Let $z_1, \ldots, z_N \in [0,1]$ be fixed, and let T be drawn uniformly at random from the size-n subsets of $[N]$. Then for every $\varepsilon > 0$,

$$\mathbb{P}_T\left(\left|\frac{1}{n}\sum_{i\in T} z_i - \frac{1}{N}\sum_{i=1}^{N} z_i\right| > \varepsilon\right) \leq 2\exp\left(-2n\varepsilon^2\right).$$

Hoeffding Inequality for Sampling Without Replacement

The probability fact we need, stated abstractly (no learning yet):

Hoeffding inequality for sampling without replacement

Let $z_1, \ldots, z_N \in [0,1]$ be fixed, and let T be drawn uniformly at random from the size-n subsets of $[N]$. Then for every $\varepsilon > 0$,

$$\mathbb{P}_T\left(\left|\frac{1}{n}\sum_{i\in T} z_i - \frac{1}{N}\sum_{i=1}^{N} z_i\right| > \varepsilon\right) \le 2\exp\left(-2n\varepsilon^2\right).$$

Equivalently, setting $2\exp\left(-2n\varepsilon^2\right) = \delta$ and solving for ε: with probability at least $1-\delta$ over T,

$$\left|\frac{1}{n}\sum_{i\in T} z_i - \frac{1}{N}\sum_{i=1}^{N} z_i\right| \le \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}.$$

Hoeffding Inequality for Sampling Without Replacement

The probability fact we need, stated abstractly (no learning yet):

Hoeffding inequality for sampling without replacement

Let $z_1, \ldots, z_N \in [0,1]$ be fixed, and let T be drawn uniformly at random from the size-n subsets of $[N]$. Then for every $\varepsilon > 0$,

$$\mathbb{P}_T\left(\left|\frac{1}{n}\sum_{i\in T} z_i - \frac{1}{N}\sum_{i=1}^{N} z_i\right| > \varepsilon\right) \le 2\exp(-2n\varepsilon^2).$$

Equivalently, setting $2\exp(-2n\varepsilon^2) = \delta$ and solving for ε: with probability at least $1-\delta$ over T,

$$\left|\frac{1}{n}\sum_{i\in T} z_i - \frac{1}{N}\sum_{i=1}^{N} z_i\right| \le \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}.$$

Same rate as the classical (with-replacement) Hoeffding inequality: the population size N does not appear.

Concentration for a Fixed Hypothesis: The Bound

Concentration for a fixed hypothesis

Applying the probability fact with $z_i = z_{i(h)}$ and $N = n + u$: for any h fixed before the split, with probability at least $1 - \delta$,

$$|L_T(h) - L(h)| \leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}.$$

Concentration for a Fixed Hypothesis: The Bound

Concentration for a fixed hypothesis

Applying the probability fact with $z_i = z_{i(h)}$ and $N = n + u$: for any h fixed before the split, with probability at least $1 - \delta$,

$$|L_T(h) - L(h)| \leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}.$$

Combining with the previous identity:

$$|L_T(h) - L_U(h)| = \frac{n+u}{u}|L_T(h) - L(h)| \leq \frac{n+u}{u}\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}.$$

Concentration for a Fixed Hypothesis: The Bound

Concentration for a fixed hypothesis

Applying the probability fact with $z_i = z_{i(h)}$ and $N = n + u$: for any h fixed before the split, with probability at least $1 - \delta$,

$$|L_T(h) - L(h)| \leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}.$$

Combining with the previous identity:

$$|L_T(h) - L_U(h)| = \frac{n+u}{u}|L_T(h) - L(h)| \leq \frac{n+u}{u}\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}.$$

A training average over n points tracks the **pool average** $L(h)$, not any **individual test point**:

Concentration for a Fixed Hypothesis: The Bound

Concentration for a fixed hypothesis

Applying the probability fact with $z_i = z_{i(h)}$ and $N = n + u$: for any h fixed before the split, with probability at least $1 - \delta$,

$$|L_T(h) - L(h)| \leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}.$$

Combining with the previous identity:

$$|L_T(h) - L_U(h)| = \frac{n+u}{u}|L_T(h) - L(h)| \leq \frac{n+u}{u}\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}.$$

A training average over n points tracks the **pool average** $L(h)$, not any **individual test point**:

- If $u = 1$, $L_U(h)$ is the loss on a single point and has high variance; $L_T(h)$ need not match it.

Concentration for a Fixed Hypothesis: The Bound

Concentration for a fixed hypothesis

Applying the probability fact with $z_i = z_{i(h)}$ and $N = n + u$: for any h fixed before the split, with probability at least $1 - \delta$,

$$|L_T(h) - L(h)| \leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}.$$

Combining with the previous identity:

$$|L_T(h) - L_U(h)| = \frac{n+u}{u}|L_T(h) - L(h)| \leq \frac{n+u}{u}\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}.$$

A training average over n points tracks the **pool average** $L(h)$, not any **individual test point**:

- If $u = 1$, $L_U(h)$ is the loss on a single point and has high variance; $L_T(h)$ need not match it.
- The bound is meaningful for $u \gtrsim n$, where it recovers the standard $\tilde{O}\left(\sqrt{1/n}\right)$ rate.

But $\hat{h}$ Is Not Fixed

The issue. The previous bound requires h to be fixed *before* the split T is drawn.

But $\hat{h}$ Is Not Fixed

The issue. The previous bound requires h to be fixed *before* the split T is drawn.

But the learner's $\hat{h}$ is a function of $\{(x_i, y_i) \mid i \in T\}$, so $\hat{h}$ *depends on* T. The fixed-h bound does not apply to $\hat{h}$.

But $\hat{h}$ Is Not Fixed

The issue. The previous bound requires h to be fixed *before* the split T is drawn.

But the learner's $\hat{h}$ is a function of $\{(x_i, y_i) \mid i \in T\}$, so $\hat{h}$ *depends on* T. The fixed-h bound does not apply to $\hat{h}$.

What we need. With probability at least $1 - \delta$ over T, simultaneously for every $h \in \mathcal{H}$,

$$|L_T(h) - L(h)| \leq \varepsilon.$$

But $\hat{h}$ Is Not Fixed

The issue. The previous bound requires h to be fixed *before* the split T is drawn.

But the learner's $\hat{h}$ is a function of $\{(x_i, y_i) \mid i \in T\}$, so $\hat{h}$ *depends on* T. The fixed-h bound does not apply to $\hat{h}$.

What we need. With probability at least $1 - \delta$ over T, simultaneously for every $h \in \mathcal{H}$,

$$|L_T(h) - L(h)| \leq \varepsilon.$$

How we get it. Apply the fixed-h bound across all of $\mathcal{H}$ via a union bound. The result is called a *uniform convergence* bound.

Only Label Patterns Matter

Write $C = \{x_1, ..., x_{n+u}\}$ for the pool of unlabeled points.

Only Label Patterns Matter

Write $C = \{x_1, \ldots, x_{n+u}\}$ for the pool of unlabeled points.

The fixed-h bound only depends on h through its **error indicators**

$$z_i(h) = \mathbf{1}[h(x_i) \neq y_i], \quad i \in [n+u].$$

Only Label Patterns Matter

Write $C = \{x_1, \ldots, x_{n+u}\}$ for the pool of unlabeled points.

The fixed-h bound only depends on h through its **error indicators**

$$z_i(h) = \mathbf{1}[h(x_i) \neq y_i], \quad i \in [n+u].$$

If two hypotheses h, h' agree on every $x_i \in C$, they produce **identical error indicators**, hence identical events in the concentration bound.

Only Label Patterns Matter

Write $C = \{x_1, \ldots, x_{n+u}\}$ for the pool of unlabeled points.

The fixed-h bound only depends on h through its **error indicators**

$$z_i(h) = \mathbf{1}[h(x_i) \neq y_i], \quad i \in [n+u].$$

If two hypotheses h, h' agree on every $x_i \in C$, they produce **identical error indicators**, hence identical events in the concentration bound.

So for the union bound, only **distinct label patterns on C** matter. Denote this *restriction set* by

$$\mathcal{H}|_C := \{(h(x_1), \ldots, h(x_{n+u})) \mid h \in \mathcal{H}\} \subseteq \{0,1\}^{n+u}.$$

Only Label Patterns Matter

Write $C = \{x_1, \ldots, x_{n+u}\}$ for the pool of unlabeled points.

The fixed-h bound only depends on h through its **error indicators**

$$z_i(h) = \mathbf{1}[h(x_i) \neq y_i], \quad i \in [n+u].$$

If two hypotheses h, h' agree on every $x_i \in C$, they produce **identical error indicators**, hence identical events in the concentration bound.

So for the union bound, only **distinct label patterns on C** matter. Denote this *restriction set* by

$$\mathcal{H}|_C := \{(h(x_1), \ldots, h(x_{n+u})) \mid h \in \mathcal{H}\} \subseteq \{0,1\}^{n+u}.$$

We will show that the error is governed by the size of the restriction set $\mathcal{H}|_C$, which is controlled by the *growth function* $\Gamma_{\mathcal{H}}(n+u)$ from Week 2.

Uniform Convergence on a Fixed Pool

Recall the restriction set $\mathcal{H}|_C$ from the previous slide, and let N be its size: the number of distinct label patterns that $\mathcal{H}$ realizes on the pool C.

Uniform Convergence on a Fixed Pool

Recall the restriction set $\mathcal{H}|_C$ from the previous slide, and let N be its size: the number of distinct label patterns that $\mathcal{H}$ realizes on the pool C.

Finite-pool uniform convergence

For every $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random split T, every $h \in \mathcal{H}$ satisfies

$$|L_U(h) - L_T(h)| \leq \frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}}.$$

Uniform Convergence on a Fixed Pool

Recall the restriction set $\mathcal{H}|_C$ from the previous slide, and let N be its size: the number of distinct label patterns that $\mathcal{H}$ realizes on the pool C.

Finite-pool uniform convergence

For every $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random split T, every $h \in \mathcal{H}$ satisfies

$$|L_U(h) - L_T(h)| \leq \frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}}.$$

By the growth function from Week 2, $N \leq \Gamma_{\mathcal{H}}(n+u)$, so the bound above can always be replaced with $\log \Gamma_{\mathcal{H}}(n+u)$ in the numerator.

Proof of the Pool-Wise Uniform Convergence Bound

Proof.

- Hypotheses with the same label pattern on C produce identical L_T and L_U, so only N distinct deviation events need to be controlled.

Proof of the Pool-Wise Uniform Convergence Bound

Proof.

- Hypotheses with the same label pattern on C produce identical L_T and L_U, so only N distinct deviation events need to be controlled.
- Pick one representative $h_1, \ldots, h_N$, one per pattern in $\mathcal{H}|_C$.

Proof of the Pool-Wise Uniform Convergence Bound

Proof.

- Hypotheses with the same label pattern on C produce identical L_T and L_U, so only N distinct deviation events need to be controlled.
- Pick one representative $h_1, \ldots, h_N$, one per pattern in $\mathcal{H}|_C$.
- Apply the fixed-hypothesis bound to each h_j at failing probability δ/N, then union bound: with probability at least $1 - \delta$, every h_j satisfies

$$\left|L_T(h_j) - L_U(h_j)\right| \leq \frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}}.$$

Proof of the Pool-Wise Uniform Convergence Bound

Proof.

- Hypotheses with the same label pattern on C produce identical L_T and L_U, so only N distinct deviation events need to be controlled.
- Pick one representative $h_1, \ldots, h_N$, one per pattern in $\mathcal{H}|_C$.
- Apply the fixed-hypothesis bound to each h_j at failing probability δ/N, then union bound: with probability at least $1 - \delta$, every h_j satisfies

$$\left|L_T(h_j) - L_U(h_j)\right| \leq \frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}}.$$

- Every $h \in \mathcal{H}$ shares its L_T and L_U values with some h_j, so the bound extends to all of $\mathcal{H}$. □

We want small $L_U\left(\hat{h}\right)$. By the triangle inequality, for any choice of $\hat{h}$,

$$L_U\left(\hat{h}\right) \leq L_T\left(\hat{h}\right) + \underbrace{\left|L_U\left(\hat{h}\right) - L_T\left(\hat{h}\right)\right|}_{\leq \tilde{O}\left(\frac{1}{\sqrt{n}}\right) \text{ by uniform convergence}}.$$

We want small $L_U\left(\hat{h}\right)$. By the triangle inequality, for any choice of $\hat{h}$,

$$L_U\left(\hat{h}\right) \leq L_T\left(\hat{h}\right) + \underbrace{\left|L_U\left(\hat{h}\right) - L_T\left(\hat{h}\right)\right|}_{\leq \tilde{O}\left(\frac{1}{\sqrt{n}}\right) \text{ by uniform convergence}}.$$

Since the second term is already small for *every* $h \in \mathcal{H}$, it suffices to choose $\hat{h}$ that minimizes the training error:

$$\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_T(h).$$

We want small $L_U\left(\hat{h}\right)$. By the triangle inequality, for any choice of $\hat{h}$,

$$L_U\left(\hat{h}\right) \leq L_T\left(\hat{h}\right) + \underbrace{\left|L_U\left(\hat{h}\right) - L_T\left(\hat{h}\right)\right|}_{\leq \tilde{O}\left(\frac{1}{\sqrt{n}}\right) \text{ by uniform convergence}}.$$

Since the second term is already small for *every* $h \in \mathcal{H}$, it suffices to choose $\hat{h}$ that minimizes the training error:

$$\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_T(h).$$

This is *empirical risk minimization* (ERM): minimize what you can see, and let uniform convergence handle the rest.

ERM on the Fixed Pool

Transductive ERM guarantee

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_T(h)$. With probability at least $1 - \delta$ over the random split T,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}},$$

where N is the size of $\mathcal{H}|_C$.

ERM on the Fixed Pool

Transductive ERM guarantee

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_T(h)$. With probability at least $1 - \delta$ over the random split T,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}},$$

where N is the size of $\mathcal{H}|_C$.

Reading the pieces of the bound:

ERM on the Fixed Pool

Transductive ERM guarantee

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_T(h)$. With probability at least $1 - \delta$ over the random split T,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}},$$

where N is the size of $\mathcal{H}|_C$.

Reading the pieces of the bound:

- $\log N$ is a complexity term, coming from the union bound over label patterns.

ERM on the Fixed Pool

Transductive ERM guarantee

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_T(h)$. With probability at least $1 - \delta$ over the random split T,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}},$$

where N is the size of $\mathcal{H}|_C$.

Reading the pieces of the bound:

- $\log N$ is a complexity term, coming from the union bound over label patterns.
- $\log(2/\delta)$ is a failing-probability term, coming from the fixed-hypothesis concentration bound.

ERM on the Fixed Pool

Transductive ERM guarantee

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_T(h)$. With probability at least $1 - \delta$ over the random split T,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}},$$

where N is the size of $\mathcal{H}|_C$.

Reading the pieces of the bound:

- $\log N$ is a complexity term, coming from the union bound over label patterns.
- $\log(2/\delta)$ is a failing-probability term, coming from the fixed-hypothesis concentration bound.
- Dividing by $2n$ says more training points sharpen both terms, giving the $\sqrt{1/n}$ rate from Hoeffding.

ERM on the Fixed Pool

Transductive ERM guarantee

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_T(h)$. With probability at least $1 - \delta$ over the random split T,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}},$$

where N is the size of $\mathcal{H}|_C$.

Reading the pieces of the bound:

- $\log N$ is a complexity term, coming from the union bound over label patterns.
- $\log(2/\delta)$ is a failing-probability term, coming from the fixed-hypothesis concentration bound.
- Dividing by $2n$ says more training points sharpen both terms, giving the $\sqrt{1/n}$ rate from Hoeffding.
- $(n+u)/u$ is the amplification from converting the pool-error concentration into a training-vs-test gap; it blows up when $u \ll n$ (few test points to generalize to).

ERM on the Fixed Pool

Transductive ERM guarantee

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_T(h)$. With probability at least $1 - \delta$ over the random split T,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}},$$

where N is the size of $\mathcal{H}|_C$.

Reading the pieces of the bound:

- $\log N$ is a complexity term, coming from the union bound over label patterns.
- $\log(2/\delta)$ is a failing-probability term, coming from the fixed-hypothesis concentration bound.
- Dividing by $2n$ says more training points sharpen both terms, giving the $\sqrt{1/n}$ rate from Hoeffding.
- $(n+u)/u$ is the amplification from converting the pool-error concentration into a training-vs-test gap; it blows up when $u \ll n$ (few test points to generalize to).
- N is pool-dependent. To get a pool-independent guarantee we need to control it uniformly over C.

Proof of the Transductive ERM Guarantee

Write $\varepsilon = \frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}}$ for the bound in the theorem, and let $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} L_U(h)$ be the best hypothesis in $\mathcal{H}$.

Proof of the Transductive ERM Guarantee

Write $\varepsilon = \frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}}$ for the bound in the theorem, and let $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} L_U(h)$ be the best hypothesis in $\mathcal{H}$.

On the uniform convergence event, every $h \in \mathcal{H}$ satisfies $|L_T(h) - L_U(h)| \leq \varepsilon$. Chain three inequalities:

$$\begin{aligned} L_U\left(\hat{h}\right) &\leq L_T\left(\hat{h}\right) + \varepsilon \quad \text{(uniform convergence)} \\ &\leq L_T(h^*) + \varepsilon \quad \text{(definition of } \hat{h}\text{)} \\ &\leq L_U(h^*) + 2\varepsilon \quad \text{(uniform convergence).} \end{aligned}$$

Proof of the Transductive ERM Guarantee

Write $\varepsilon = \frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}}$ for the bound in the theorem, and let $h^* \in \operatorname{argmin}_{h\in\mathcal{H}} L_U(h)$ be the best hypothesis in $\mathcal{H}$.

On the uniform convergence event, every $h \in \mathcal{H}$ satisfies $|L_T(h) - L_U(h)| \le \varepsilon$. Chain three inequalities:

$$\begin{aligned} L_U\left(\hat{h}\right) &\le L_T\left(\hat{h}\right) + \varepsilon \quad \text{(uniform convergence)} \\ &\le L_T(h^*) + \varepsilon \quad \text{(definition of } \hat{h}\text{)} \\ &\le L_U(h^*) + 2\varepsilon \quad \text{(uniform convergence).} \end{aligned}$$

Since $L_U(h^*) = \inf_{h\in\mathcal{H}} L_U(h)$, this is exactly the theorem. □

From the Pool-Dependent Bound to a VC Bound

Recall the transductive ERM guarantee: with probability at least $1 - \delta$,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}},$$

where N is the size of $\mathcal{H}|_C$.

From the Pool-Dependent Bound to a VC Bound

Recall the transductive ERM guarantee: with probability at least $1 - \delta$,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}},$$

where N is the size of $\mathcal{H}|_C$.

To get a pool-independent guarantee, recall from Week 2:

From the Pool-Dependent Bound to a VC Bound

Recall the transductive ERM guarantee: with probability at least $1 - \delta$,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}},$$

where N is the size of $\mathcal{H}|_C$.

To get a pool-independent guarantee, recall from Week 2:

- **Growth function.** $N \leq \Gamma_{\mathcal{H}}(n+u)$ for every pool of size $n+u$.

From the Pool-Dependent Bound to a VC Bound

Recall the transductive ERM guarantee: with probability at least $1 - \delta$,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}},$$

where N is the size of $\mathcal{H}|_C$.

To get a pool-independent guarantee, recall from Week 2:

- **Growth function.** $N \leq \Gamma_{\mathcal{H}}(n+u)$ for every pool of size $n+u$.
- **Sauer–Shelah.** If $d = \mathrm{VCdim}(\mathcal{H})$ and $n+u \geq d \geq 1$, then $\Gamma_{\mathcal{H}}(n+u) \leq \left(e\frac{n+u}{d}\right)^d$.

From the Pool-Dependent Bound to a VC Bound

Recall the transductive ERM guarantee: with probability at least $1 - \delta$,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{\log N + \log\left(\frac{2}{\delta}\right)}{2n}},$$

where N is the size of $\mathcal{H}|_C$.

To get a pool-independent guarantee, recall from Week 2:

- **Growth function.** $N \leq \Gamma_{\mathcal{H}}(n+u)$ for every pool of size $n+u$.
- **Sauer–Shelah.** If $d = \mathrm{VCdim}(\mathcal{H})$ and $n+u \geq d \geq 1$, then $\Gamma_{\mathcal{H}}(n+u) \leq \left(e\frac{n+u}{d}\right)^d$.

Plugging $\log N \leq d \log(e(n+u)/d)$ into the previous theorem yields a pool-independent bound.

VC corollary for transductive ERM

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_T(h)$. With probability at least $1 - \delta$ over the random split T,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{d\log\left(e\frac{n+u}{d}\right) + \log\left(\frac{2}{\delta}\right)}{2n}}.$$

VC corollary for transductive ERM

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_T(h)$. With probability at least $1 - \delta$ over the random split T,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{d\log\left(e\frac{n+u}{d}\right) + \log\left(\frac{2}{\delta}\right)}{2n}}.$$

VC dimension governs the learning difficulty in *both* transductive settings:

VC Corollary for Transductive ERM

VC corollary for transductive ERM

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_T(h)$. With probability at least $1 - \delta$ over the random split T,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{d\log\left(e\frac{n+u}{d}\right) + \log\left(\frac{2}{\delta}\right)}{2n}}.$$

VC dimension governs the learning difficulty in *both* transductive settings:

- **Online** (Week 2): Halving mistake bound $O(d\log(n/d))$.

VC Corollary for Transductive ERM

VC corollary for transductive ERM

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_T(h)$. With probability at least $1 - \delta$ over the random split T,

$$L_U\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_U(h) + 2\frac{n+u}{u}\sqrt{\frac{d\log\left(e\frac{n+u}{d}\right) + \log\left(\frac{2}{\delta}\right)}{2n}}.$$

VC dimension governs the learning difficulty in *both* transductive settings:

- **Online** (Week 2): Halving mistake bound $O(d\log(n/d))$.
- **Batch** (here): ERM error rate $\tilde{O}\left(\sqrt{d/n}\right)$ when $u \gtrsim n$.

The i.i.d. Model

From Fixed Pools to Random Draws

The transductive setting rested on three choices:

From Fixed Pools to Random Draws

The transductive setting rested on three choices:

- the pool is fixed and known up front;

From Fixed Pools to Random Draws

The transductive setting rested on three choices:

- the pool is fixed and known up front;
- the train/test split is the only randomness;

From Fixed Pools to Random Draws

The transductive setting rested on three choices:

- the pool is fixed and known up front;
- the train/test split is the only randomness;
- error is measured on the held-out points of the pool.

From Fixed Pools to Random Draws

The transductive setting rested on three choices:

- the pool is fixed and known up front;
- the train/test split is the only randomness;
- error is measured on the held-out points of the pool.

This fits problems where the prediction target is a concrete batch.

From Fixed Pools to Random Draws

The transductive setting rested on three choices:

- the pool is fixed and known up front;
- the train/test split is the only randomness;
- error is measured on the held-out points of the pool.

This fits problems where the prediction target is a concrete batch.

Now we turn to a different question:

From Fixed Pools to Random Draws

The transductive setting rested on three choices:

- the pool is fixed and known up front;
- the train/test split is the only randomness;
- error is measured on the held-out points of the pool.

This fits problems where the prediction target is a concrete batch.

Now we turn to a different question:

What if the data themselves are random draws from an unknown distribution?

From Fixed Pools to Random Draws

The transductive setting rested on three choices:

- the pool is fixed and known up front;
- the train/test split is the only randomness;
- error is measured on the held-out points of the pool.

This fits problems where the prediction target is a concrete batch.

Now we turn to a different question:

What if the data themselves are random draws from an unknown distribution?

i.i.d. setting: (x_i, y_i) are drawn i.i.d. from an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$.

From Fixed Pools to Random Draws

The transductive setting rested on three choices:

- the pool is fixed and known up front;
- the train/test split is the only randomness;
- error is measured on the held-out points of the pool.

This fits problems where the prediction target is a concrete batch.

Now we turn to a different question:

What if the data themselves are random draws from an unknown distribution?

i.i.d. setting: (x_i, y_i) are drawn i.i.d. from an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$.

The combinatorics carry over unchanged; only the concentration step is different.

Data-generating assumption

Unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. Training sample $(x_1, y_1), ..., (x_n, y_n)$ drawn i.i.d. from $\mathcal{D}$.

Data-generating assumption

Unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. Training sample $(x_1, y_1), \ldots, (x_n, y_n)$ drawn i.i.d. from $\mathcal{D}$.

Population risk, the quantity we care about:

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}(h(x) \neq y).$$

Data-generating assumption

Unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. Training sample $(x_1, y_1), \ldots, (x_n, y_n)$ drawn i.i.d. from $\mathcal{D}$.

Population risk, the quantity we care about:

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}(h(x) \neq y).$$

Empirical risk, the quantity the learner sees:

$$L_n(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[h(x_i) \neq y_i].$$

Data-generating assumption

Unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. Training sample $(x_1, y_1), \ldots, (x_n, y_n)$ drawn i.i.d. from $\mathcal{D}$.

Population risk, the quantity we care about:

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}(h(x) \neq y).$$

Empirical risk, the quantity the learner sees:

$$L_n(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[h(x_i) \neq y_i].$$

Goal. Output $\hat{h}$ with small population risk $L_{\mathcal{D}}\left(\hat{h}\right)$, using only the training sample.

Example: Thresholds on the Real Line

Let $\mathcal{H}$ be the class of thresholds $h_\theta(x) = \mathbf{1}[x \leq \theta]$ on $\mathbb{R}$.

Example: Thresholds on the Real Line

Let $\mathcal{H}$ be the class of thresholds $h_\theta(x) = \mathbf{1}[x \leq \theta]$ on $\mathbb{R}$.

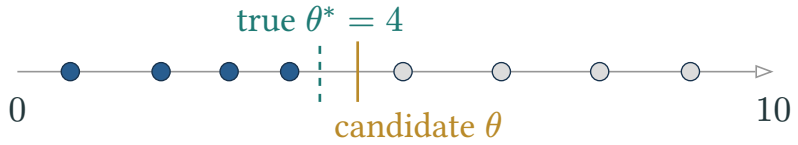
Underlying distribution $\mathcal{D}$: x is drawn uniformly from $[0, 10]$, and $y = \mathbf{1}[x \leq 4]$.

Example: Thresholds on the Real Line

Let $\mathcal{H}$ be the class of thresholds $h_\theta(x) = \mathbf{1}[x \le \theta]$ on $\mathbb{R}$.

Underlying distribution $\mathcal{D}$: x is drawn uniformly from $[0, 10]$, and $y = \mathbf{1}[x \le 4]$.

A sample of $n = 8$ points with their labels (blue $= 1$, gray $= 0$):

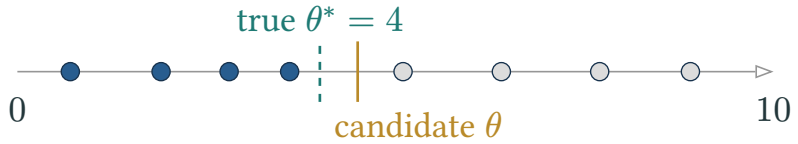


Example: Thresholds on the Real Line

Let $\mathcal{H}$ be the class of thresholds $h_\theta(x) = \mathbf{1}[x \le \theta]$ on $\mathbb{R}$.

Underlying distribution $\mathcal{D}$: x is drawn uniformly from $[0, 10]$, and $y = \mathbf{1}[x \le 4]$.

A sample of $n = 8$ points with their labels (blue $= 1$, gray $= 0$):



The candidate $\theta = 4.5$ agrees with every labeled point, so $L_n(h_\theta) = 0$. Its population risk is

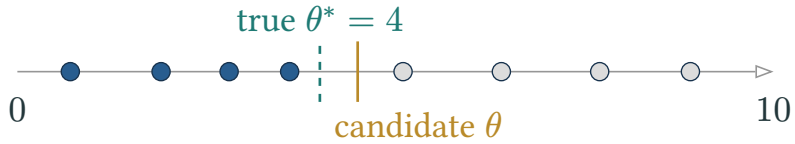
$$L_{\mathcal{D}}(h_\theta) = \mathbb{P}_x(4 < x \le 4.5) = 0.05.$$

Example: Thresholds on the Real Line

Let $\mathcal{H}$ be the class of thresholds $h_\theta(x) = \mathbf{1}[x \le \theta]$ on $\mathbb{R}$.

Underlying distribution $\mathcal{D}$: x is drawn uniformly from $[0, 10]$, and $y = \mathbf{1}[x \le 4]$.

A sample of $n = 8$ points with their labels (blue $= 1$, gray $= 0$):



The candidate $\theta = 4.5$ agrees with every labeled point, so $L_n(h_\theta) = 0$. Its population risk is

$$L_{\mathcal{D}}(h_\theta) = \mathbb{P}_x(4 < x \le 4.5) = 0.05.$$

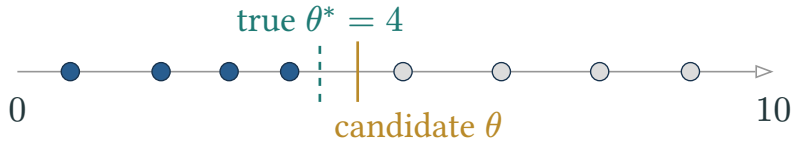
$L_n(h)$ is what the learner sees; $L_{\mathcal{D}}(h)$ is what the learner actually wants to minimize.

Example: Thresholds on the Real Line

Let $\mathcal{H}$ be the class of thresholds $h_\theta(x) = \mathbf{1}[x \le \theta]$ on $\mathbb{R}$.

Underlying distribution $\mathcal{D}$: x is drawn uniformly from $[0, 10]$, and $y = \mathbf{1}[x \le 4]$.

A sample of $n = 8$ points with their labels (blue $= 1$, gray $= 0$):



The candidate $\theta = 4.5$ agrees with every labeled point, so $L_n(h_\theta) = 0$. Its population risk is

$$L_{\mathcal{D}}(h_\theta) = \mathbb{P}_x(4 < x \le 4.5) = 0.05.$$

$L_n(h)$ is what the learner sees; $L_{\mathcal{D}}(h)$ is what the learner actually wants to minimize.

In general, how fast does $|L_n(h) - L_{\mathcal{D}}(h)|$ shrink as n grows?

Concentration for One Fixed Predictor

Fix a predictor h before seeing S, and define the mistake indicator $Z_i = \mathbf{1}[h(x_i) \neq y_i]$.

Concentration for One Fixed Predictor

Fix a predictor h before seeing S, and define the mistake indicator $Z_i = \mathbf{1}[h(x_i) \neq y_i]$.

Since (x_i, y_i) are i.i.d. from $\mathcal{D}$, the Z_i's are i.i.d. Bernoulli with mean

$$\mathbb{E}[Z_i] = \mathbb{P}_{(x,y)\sim\mathcal{D}}(h(x) \neq y) = L_{\mathcal{D}}(h),$$

and $L_n(h) = \frac{1}{n}\sum_{i=1}^{n} Z_i$ is their empirical average.

Concentration for One Fixed Predictor

Fix a predictor h before seeing S, and define the mistake indicator $Z_i = \mathbf{1}[h(x_i) \neq y_i]$.

Since (x_i, y_i) are i.i.d. from $\mathcal{D}$, the Z_i's are i.i.d. Bernoulli with mean

$$\mathbb{E}[Z_i] = \mathbb{P}_{(x,y)\sim\mathcal{D}}(h(x) \neq y) = L_{\mathcal{D}}(h),$$

and $L_n(h) = \frac{1}{n}\sum_{i=1}^{n} Z_i$ is their empirical average.

Applying the classical Hoeffding inequality:

Concentration for One Fixed Predictor

Fix a predictor h before seeing S, and define the mistake indicator $Z_i = \mathbf{1}[h(x_i) \neq y_i]$.

Since (x_i, y_i) are i.i.d. from $\mathcal{D}$, the Z_i's are i.i.d. Bernoulli with mean

$$\mathbb{E}[Z_i] = \mathbb{P}_{(x,y)\sim\mathcal{D}}(h(x) \neq y) = L_{\mathcal{D}}(h),$$

and $L_n(h) = \frac{1}{n}\sum_{i=1}^{n} Z_i$ is their empirical average.

Applying the classical Hoeffding inequality:

i.i.d. concentration for one fixed predictor

For any h fixed before S, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$,

$$|L_n(h) - L_{\mathcal{D}}(h)| \leq \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Concentration for One Fixed Predictor

Fix a predictor h before seeing S, and define the mistake indicator $Z_i = \mathbf{1}[h(x_i) \neq y_i]$.

Since (x_i, y_i) are i.i.d. from $\mathcal{D}$, the Z_i's are i.i.d. Bernoulli with mean

$$\mathbb{E}[Z_i] = \mathbb{P}_{(x,y)\sim\mathcal{D}}(h(x) \neq y) = L_{\mathcal{D}}(h),$$

and $L_n(h) = \frac{1}{n}\sum_{i=1}^{n} Z_i$ is their empirical average.

Applying the classical Hoeffding inequality:

i.i.d. concentration for one fixed predictor

For any h fixed before S, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$,

$$|L_n(h) - L_{\mathcal{D}}(h)| \leq \sqrt{\frac{\log(2/\delta)}{2n}}.$$

For a *fixed* h, empirical and population risks agree up to $\tilde{O}\left(\sqrt{1/n}\right)$.

But $\hat{h}$ Is Not Fixed

The issue. The previous bound requires h to be fixed *before* S is drawn.

But $\hat{h}$ Is Not Fixed

The issue. The previous bound requires h to be fixed *before* S is drawn.

But the learner's $\hat{h}$ is a function of S, so $\hat{h}$ *depends on* S. The fixed-h bound does not apply to $\hat{h}$.

But $\hat{h}$ Is Not Fixed

The issue. The previous bound requires h to be fixed *before* S is drawn.

But the learner's $\hat{h}$ is a function of S, so $\hat{h}$ *depends on* S. The fixed-h bound does not apply to $\hat{h}$.

What we need. With probability at least $1 - \delta$ over S, simultaneously for every $h \in \mathcal{H}$,

$$|L_n(h) - L_{\mathcal{D}}(h)| \leq \varepsilon.$$

But $\hat{h}$ Is Not Fixed

The issue. The previous bound requires h to be fixed *before* S is drawn.

But the learner's $\hat{h}$ is a function of S, so $\hat{h}$ *depends on* S. The fixed-h bound does not apply to $\hat{h}$.

What we need. With probability at least $1 - \delta$ over S, simultaneously for every $h \in \mathcal{H}$,

$$|L_n(h) - L_{\mathcal{D}}(h)| \leq \varepsilon.$$

How we get it. Apply the fixed-h bound across all of $\mathcal{H}$ via a union bound; this is *uniform convergence*.

But $\hat{h}$ Is Not Fixed

The issue. The previous bound requires h to be fixed *before* S is drawn.

But the learner's $\hat{h}$ is a function of S, so $\hat{h}$ *depends on* S. The fixed-h bound does not apply to $\hat{h}$.

What we need. With probability at least $1 - \delta$ over S, simultaneously for every $h \in \mathcal{H}$,

$$|L_n(h) - L_{\mathcal{D}}(h)| \leq \varepsilon.$$

How we get it. Apply the fixed-h bound across all of $\mathcal{H}$ via a union bound; this is *uniform convergence*.

This is exactly the same strategy as in the transductive setting.

But $\hat{h}$ Is Not Fixed

The issue. The previous bound requires h to be fixed *before* S is drawn.

But the learner's $\hat{h}$ is a function of S, so $\hat{h}$ *depends on* S. The fixed-h bound does not apply to $\hat{h}$.

What we need. With probability at least $1 - \delta$ over S, simultaneously for every $h \in \mathcal{H}$,

$$|L_n(h) - L_{\mathcal{D}}(h)| \leq \varepsilon.$$

How we get it. Apply the fixed-h bound across all of $\mathcal{H}$ via a union bound; this is *uniform convergence*.

This is exactly the same strategy as in the transductive setting.

We start with the simplest case: *finite* $\mathcal{H}$, where the union bound applies directly.

Finite-class uniform convergence

If $|\mathcal{H}| < \infty$, then for every $\delta \in (0,1)$, with probability at least $1-\delta$ over $S \sim \mathcal{D}^n$, every $h \in \mathcal{H}$ satisfies

$$|L_n(h) - L_{\mathcal{D}}(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2n}}. \quad (\star)$$

Finite-class uniform convergence

If $|\mathcal{H}| < \infty$, then for every $\delta \in (0,1)$, with probability at least $1-\delta$ over $S \sim \mathcal{D}^n$, every $h \in \mathcal{H}$ satisfies

$$|L_n(h) - L_{\mathcal{D}}(h)| \le \sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2n}}. \quad (\star)$$

Proof. Apply the fixed-hypothesis bound with failing probability $\delta/|\mathcal{H}|$ to each $h \in \mathcal{H}$; then $(\star)$ holds with probability at least $1 - \delta/|\mathcal{H}|$.

Finite-class uniform convergence

If $|\mathcal{H}| < \infty$, then for every $\delta \in (0,1)$, with probability at least $1-\delta$ over $S \sim \mathcal{D}^n$, every $h \in \mathcal{H}$ satisfies

$$|L_n(h) - L_{\mathcal{D}}(h)| \le \sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2n}}. \quad (\star)$$

Proof. Apply the fixed-hypothesis bound with failing probability $\delta/|\mathcal{H}|$ to each $h \in \mathcal{H}$; then $(\star)$ holds with probability at least $1 - \delta/|\mathcal{H}|$.

Let B_h be the "bad event" that $(\star)$ is violated for h, so $\mathbb{P}(B_h) \le \delta/|\mathcal{H}|$. By the union bound,

$$\mathbb{P}\left(\bigcup_{h\in\mathcal{H}} B_h\right) \le \sum_{h\in\mathcal{H}} \mathbb{P}(B_h) \le |\mathcal{H}| \cdot \frac{\delta}{|\mathcal{H}|} = \delta.$$

Finite-class uniform convergence

If $|\mathcal{H}| < \infty$, then for every $\delta \in (0,1)$, with probability at least $1-\delta$ over $S \sim \mathcal{D}^n$, every $h \in \mathcal{H}$ satisfies

$$|L_n(h) - L_{\mathcal{D}}(h)| \le \sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2n}}. \quad (\star)$$

Proof. Apply the fixed-hypothesis bound with failing probability $\delta/|\mathcal{H}|$ to each $h \in \mathcal{H}$; then $(\star)$ holds with probability at least $1 - \delta/|\mathcal{H}|$.

Let B_h be the "bad event" that $(\star)$ is violated for h, so $\mathbb{P}(B_h) \le \delta/|\mathcal{H}|$. By the union bound,

$$\mathbb{P}\left(\bigcup_{h\in\mathcal{H}} B_h\right) \le \sum_{h\in\mathcal{H}} \mathbb{P}(B_h) \le |\mathcal{H}| \cdot \frac{\delta}{|\mathcal{H}|} = \delta.$$

Thus with probability at least $1-\delta$, $(\star)$ holds for every $h \in \mathcal{H}$ simultaneously. □

Same triangle inequality as before, now comparing **population risk** with **empirical risk**:

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq L_n\left(\hat{h}\right) + \underbrace{\left|L_{\mathcal{D}}\left(\hat{h}\right) - L_n\left(\hat{h}\right)\right|}_{\leq \tilde{O}\left(\frac{1}{\sqrt{n}}\right) \text{ by uniform convergence}} .$$

Same triangle inequality as before, now comparing **population risk** with **empirical risk**:

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq L_n\left(\hat{h}\right) + \underbrace{\left|L_{\mathcal{D}}\left(\hat{h}\right) - L_n\left(\hat{h}\right)\right|}_{\leq \tilde{O}\left(\frac{1}{\sqrt{n}}\right) \text{ by uniform convergence}} .$$

The quantity $|L_{\mathcal{D}}(h) - L_n(h)|$ is called the *generalization gap*: how much the empirical risk can deviate from the population risk.

Same triangle inequality as before, now comparing **population risk** with **empirical risk**:

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq L_n\left(\hat{h}\right) + \underbrace{\left|L_{\mathcal{D}}\left(\hat{h}\right) - L_n\left(\hat{h}\right)\right|}_{\leq \tilde{O}\left(\frac{1}{\sqrt{n}}\right) \text{ by uniform convergence}} .$$

The quantity $|L_{\mathcal{D}}(h) - L_n(h)|$ is called the *generalization gap*: how much the empirical risk can deviate from the population risk.

By the same reasoning, it suffices to choose $\hat{h}$ that minimizes the empirical risk:

$$\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_n(h).$$

Same triangle inequality as before, now comparing **population risk** with **empirical risk**:

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq L_n\left(\hat{h}\right) + \underbrace{\left|L_{\mathcal{D}}\left(\hat{h}\right) - L_n\left(\hat{h}\right)\right|}_{\leq \tilde{O}\left(\frac{1}{\sqrt{n}}\right) \text{ by uniform convergence}} .$$

The quantity $|L_{\mathcal{D}}(h) - L_n(h)|$ is called the *generalization gap*: how much the empirical risk can deviate from the population risk.

By the same reasoning, it suffices to choose $\hat{h}$ that minimizes the empirical risk:

$$\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_n(h).$$

Same *empirical risk minimization* (ERM) rule; what changed is the quantity we are trying to control: population risk $L_{\mathcal{D}}(h)$ instead of test error $L_U(h)$ on the held-out subset.

Finite-class ERM guarantee

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_n(h)$. With probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$,

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + 2\sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2n}}.$$

Finite-class ERM guarantee

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_n(h)$. With probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$,

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + 2\sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2n}}.$$

An *excess-risk* bound: ERM is nearly as good as the best hypothesis in $\mathcal{H}$, up to an $\tilde{O}\left(\sqrt{\log|\mathcal{H}|/n}\right)$ gap.

Finite-class ERM guarantee

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_n(h)$. With probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$,

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + 2\sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2n}}.$$

An *excess-risk* bound: ERM is nearly as good as the best hypothesis in $\mathcal{H}$, up to an $\tilde{O}\left(\sqrt{\log|\mathcal{H}|/n}\right)$ gap.

Sample complexity form. To guarantee excess risk at most ε, it suffices to take

$$n \geq \frac{2(\log|\mathcal{H}| + \log(2/\delta))}{\varepsilon^2}.$$

Finite-class ERM guarantee

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_n(h)$. With probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$,

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + 2\sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2n}}.$$

An *excess-risk* bound: ERM is nearly as good as the best hypothesis in $\mathcal{H}$, up to an $\tilde{O}\left(\sqrt{\log|\mathcal{H}|/n}\right)$ gap.

Sample complexity form. To guarantee excess risk at most ε, it suffices to take

$$n \geq \frac{2(\log|\mathcal{H}| + \log(2/\delta))}{\varepsilon^2}.$$

Under *realizability* ($\exists h^* \in \mathcal{H}$ with $L_{\mathcal{D}}(h^*) = 0$), the rate improves from $1/\varepsilon^2$ to $1/\varepsilon$, a "fast rate" we return to when discussing PAC learning.

Proof of the Finite-Class ERM Guarantee

Write $\varepsilon = \sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2n}}$ and let $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$.

Proof of the Finite-Class ERM Guarantee

Write $\varepsilon = \sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2n}}$ and let $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$.

Same three-line chain as the transductive ERM proof:

$$\begin{aligned} L_{\mathcal{D}}\left(\hat{h}\right) &\leq L_n\left(\hat{h}\right) + \varepsilon \quad && \text{(uniform convergence)} \\ &\leq L_n(h^*) + \varepsilon \quad && (\ \hat{h}\ \text{ minimizes }\ L_n) \\ &\leq L_{\mathcal{D}}(h^*) + 2\varepsilon. && \text{(uniform convergence)} \end{aligned}$$

Since $L_{\mathcal{D}}(h^*) = \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$, this is the theorem. $\square$

Proof of the Finite-Class ERM Guarantee

Write $\varepsilon = \sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2n}}$ and let $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$.

Same three-line chain as the transductive ERM proof:

$$\begin{aligned} L_{\mathcal{D}}\left(\hat{h}\right) &\leq L_n\left(\hat{h}\right) + \varepsilon \quad && \text{(uniform convergence)} \\ &\leq L_n(h^*) + \varepsilon \quad && (\ \hat{h}\ \text{ minimizes }\ L_n) \\ &\leq L_{\mathcal{D}}(h^*) + 2\varepsilon. && \text{(uniform convergence)} \end{aligned}$$

Since $L_{\mathcal{D}}(h^*) = \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$, this is the theorem. □

Approximation vs. estimation. The bound cleanly decomposes excess risk:

Proof of the Finite-Class ERM Guarantee

Write $\varepsilon = \sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2n}}$ and let $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$.

Same three-line chain as the transductive ERM proof:

$$\begin{aligned} L_{\mathcal{D}}\left(\hat{h}\right) &\leq L_n\left(\hat{h}\right) + \varepsilon \quad \text{(uniform convergence)} \\ &\leq L_n(h^*) + \varepsilon \quad (\ \hat{h}\ \text{ minimizes }\ L_n) \\ &\leq L_{\mathcal{D}}(h^*) + 2\varepsilon. \ \text{(uniform convergence)} \end{aligned}$$

Since $L_{\mathcal{D}}(h^*) = \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$, this is the theorem. □

Approximation vs. estimation. The bound cleanly decomposes excess risk:

- $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$: **approximation error**, set by the richness of $\mathcal{H}$.

Proof of the Finite-Class ERM Guarantee

Write $\varepsilon = \sqrt{\frac{\log|\mathcal{H}| + \log(2/\delta)}{2n}}$ and let $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$.

Same three-line chain as the transductive ERM proof:

$$\begin{aligned} L_{\mathcal{D}}\left(\hat{h}\right) &\leq L_n\left(\hat{h}\right) + \varepsilon \quad \text{(uniform convergence)} \\ &\leq L_n(h^*) + \varepsilon \quad (\ \hat{h}\ \text{ minimizes }\ L_n) \\ &\leq L_{\mathcal{D}}(h^*) + 2\varepsilon. \ \text{(uniform convergence)} \end{aligned}$$

Since $L_{\mathcal{D}}(h^*) = \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$, this is the theorem. □

Approximation vs. estimation. The bound cleanly decomposes excess risk:

- $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$: **approximation error**, set by the richness of $\mathcal{H}$.
- $2\varepsilon = \tilde{O}\left(\sqrt{\log|\mathcal{H}|/n}\right)$: **estimation error**, shrinking with n.

Sample-Complexity Scaling

Plotting $n \geq 2(\log|\mathcal{H}| + \log(2/\delta))/\varepsilon^2$ (with $\delta = 0.05$):

Sample-Complexity Scaling

Plotting $n \geq 2(\log|\mathcal{H}| + \log(2/\delta))/\varepsilon^2$ (with $\delta = 0.05$):

$\varepsilon \setminus |\mathcal{H}|$	$\mathbf{10}$	$\mathbf{10^3}$	$\mathbf{10^6}$	$\mathbf{10^{\{10\}}}$
0.10	1.2k	2.1k	3.5k	5.3k
0.05	4.8k	8.5k	14.0k	21.4k
0.01	120k	212k	350k	535k

Sample-Complexity Scaling

Plotting $n \geq 2(\log|\mathcal{H}| + \log(2/\delta))/\varepsilon^2$ (with $\delta = 0.05$):

$\varepsilon \setminus |\mathcal{H}|$	$\mathbf{10}$	$\mathbf{10^3}$	$\mathbf{10^6}$	$\mathbf{10^{\{10\}}}$
0.10	1.2k	2.1k	3.5k	5.3k
0.05	4.8k	8.5k	14.0k	21.4k
0.01	120k	212k	350k	535k

- Halving ε **quadruples** n (every row down).

Sample-Complexity Scaling

Plotting $n \geq 2(\log|\mathcal{H}| + \log(2/\delta))/\varepsilon^2$ (with $\delta = 0.05$):

$\varepsilon \setminus |\mathcal{H}|$	$\mathbf{10}$	$\mathbf{10^3}$	$\mathbf{10^6}$	$\mathbf{10^{\{10\}}}$
0.10	1.2k	2.1k	3.5k	5.3k
0.05	4.8k	8.5k	14.0k	21.4k
0.01	120k	212k	350k	535k

- Halving ε **quadruples** n (every row down).
- Blowing up $|\mathcal{H}|$ by 10^9 only **quintuples** n (leftmost to rightmost), showing *logarithmic* dependence.

Sample-Complexity Scaling

Plotting $n \geq 2(\log|\mathcal{H}| + \log(2/\delta))/\varepsilon^2$ (with $\delta = 0.05$):

$\varepsilon \setminus |\mathcal{H}|$	$\mathbf{10}$	$\mathbf{10^3}$	$\mathbf{10^6}$	$\mathbf{10^{\{10\}}}$
0.10	1.2k	2.1k	3.5k	5.3k
0.05	4.8k	8.5k	14.0k	21.4k
0.01	120k	212k	350k	535k

- Halving ε **quadruples** n (every row down).
- Blowing up $|\mathcal{H}|$ by 10^9 only **quintuples** n (leftmost to rightmost), showing *logarithmic* dependence.

This bound applies only to *finite* $\mathcal{H}$. For infinite classes we need a different combinatorial quantity: the **growth function**.

From $|\mathcal{H}|$ to the Growth Function

Problem. If $\mathcal{H}$ is infinite, $\log|\mathcal{H}| = \infty$ and the finite-class bound is vacuous.

From $|\mathcal{H}|$ to the Growth Function

Problem. If $\mathcal{H}$ is infinite, $\log|\mathcal{H}| = \infty$ and the finite-class bound is vacuous.

Recall the **growth function** from Week 2:

$$\Gamma_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} |\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}|.$$

From $|\mathcal{H}|$ to the Growth Function

Problem. If $\mathcal{H}$ is infinite, $\log|\mathcal{H}| = \infty$ and the finite-class bound is vacuous.

Recall the **growth function** from Week 2:

$$\Gamma_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} |\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}|.$$

Example. Thresholds on $\mathbb{R}$ have $|\mathcal{H}| = \infty$, yet $\Gamma_{\mathcal{H}}(n) = n + 1$.

From $|\mathcal{H}|$ to the Growth Function

Problem. If $\mathcal{H}$ is infinite, $\log|\mathcal{H}| = \infty$ and the finite-class bound is vacuous.

Recall the **growth function** from Week 2:

$$\Gamma_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} |\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}|.$$

Example. Thresholds on $\mathbb{R}$ have $|\mathcal{H}| = \infty$, yet $\Gamma_{\mathcal{H}}(n) = n + 1$.

The union bound was wasteful. In the finite-class proof, $\log|\mathcal{H}|$ came from a union bound over $|\mathcal{H}|$ "bad events", one per hypothesis. But:

From $|\mathcal{H}|$ to the Growth Function

Problem. If $\mathcal{H}$ is infinite, $\log|\mathcal{H}| = \infty$ and the finite-class bound is vacuous.

Recall the **growth function** from Week 2:

$$\Gamma_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} |\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}|.$$

Example. Thresholds on $\mathbb{R}$ have $|\mathcal{H}| = \infty$, yet $\Gamma_{\mathcal{H}}(n) = n + 1$.

The union bound was wasteful. In the finite-class proof, $\log|\mathcal{H}|$ came from a union bound over $|\mathcal{H}|$ "bad events", one per hypothesis. But:

- Two hypotheses with the same labels on S produce the same L_n.

From $|\mathcal{H}|$ to the Growth Function

Problem. If $\mathcal{H}$ is infinite, $\log|\mathcal{H}| = \infty$ and the finite-class bound is vacuous.

Recall the **growth function** from Week 2:

$$\Gamma_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} |\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}|.$$

Example. Thresholds on $\mathbb{R}$ have $|\mathcal{H}| = \infty$, yet $\Gamma_{\mathcal{H}}(n) = n + 1$.

The union bound was wasteful. In the finite-class proof, $\log|\mathcal{H}|$ came from a union bound over $|\mathcal{H}|$ "bad events", one per hypothesis. But:

- Two hypotheses with the same labels on S produce the same L_n.
- So they give the same "bad event", and should only be counted once.

From $|\mathcal{H}|$ to the Growth Function

Problem. If $\mathcal{H}$ is infinite, $\log|\mathcal{H}| = \infty$ and the finite-class bound is vacuous.

Recall the **growth function** from Week 2:

$$\Gamma_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} |\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}|.$$

Example. Thresholds on $\mathbb{R}$ have $|\mathcal{H}| = \infty$, yet $\Gamma_{\mathcal{H}}(n) = n + 1$.

The union bound was wasteful. In the finite-class proof, $\log|\mathcal{H}|$ came from a union bound over $|\mathcal{H}|$ "bad events", one per hypothesis. But:

- Two hypotheses with the same labels on S produce the same L_n.
- So they give the same "bad event", and should only be counted once.
- Only **distinct label patterns on S** contribute, and there are at most $\Gamma_{\mathcal{H}}(n)$ of them.

From $|\mathcal{H}|$ to the Growth Function

Problem. If $\mathcal{H}$ is infinite, $\log|\mathcal{H}| = \infty$ and the finite-class bound is vacuous.

Recall the **growth function** from Week 2:

$$\Gamma_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} |\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}|.$$

Example. Thresholds on $\mathbb{R}$ have $|\mathcal{H}| = \infty$, yet $\Gamma_{\mathcal{H}}(n) = n + 1$.

The union bound was wasteful. In the finite-class proof, $\log|\mathcal{H}|$ came from a union bound over $|\mathcal{H}|$ "bad events", one per hypothesis. But:

- Two hypotheses with the same labels on S produce the same L_n.
- So they give the same "bad event", and should only be counted once.
- Only **distinct label patterns on S** contribute, and there are at most $\Gamma_{\mathcal{H}}(n)$ of them.

Making this rigorous requires *symmetrization*, since S itself is random.

Obstacle. Consider the deviation we want to bound:

$$\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_n(h)).$$

Symmetrization: Obstacle and Idea

Obstacle. Consider the deviation we want to bound:

$$\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_n(h)).$$

- $L_{\mathcal{D}}(h)$ depends on h's values *everywhere*, not just on S.

Symmetrization: Obstacle and Idea

Obstacle. Consider the deviation we want to bound:

$$\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_n(h)).$$

- $L_{\mathcal{D}}(h)$ depends on h's values *everywhere*, not just on S.
- Two hypotheses can agree on all of S but have different $L_{\mathcal{D}}$.

Symmetrization: Obstacle and Idea

Obstacle. Consider the deviation we want to bound:

$$\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_n(h)).$$

- $L_{\mathcal{D}}(h)$ depends on h's values *everywhere*, not just on S.
- Two hypotheses can agree on all of S but have different $L_{\mathcal{D}}$.
- So the previous counting argument breaks.

Symmetrization: Obstacle and Idea

Obstacle. Consider the deviation we want to bound:

$$\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_n(h)).$$

- $L_{\mathcal{D}}(h)$ depends on h's values *everywhere*, not just on S.
- Two hypotheses can agree on all of S but have different $L_{\mathcal{D}}$.
- So the previous counting argument breaks.

Idea. Introduce a *ghost sample* $S' \sim \mathcal{D}^n$ independent of S, and let $L'_n(h)$ be the empirical risk on S'.

Symmetrization: Obstacle and Idea

Obstacle. Consider the deviation we want to bound:

$$\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_n(h)).$$

- $L_{\mathcal{D}}(h)$ depends on h's values *everywhere*, not just on S.
- Two hypotheses can agree on all of S but have different $L_{\mathcal{D}}$.
- So the previous counting argument breaks.

Idea. Introduce a *ghost sample* $S' \sim \mathcal{D}^n$ independent of S, and let $L'_n(h)$ be the empirical risk on S'.

For each fixed h, $\mathbb{E}_{S'}[L'_n(h)] = L_{\mathcal{D}}(h)$, so the population risk is itself an expected empirical risk.

Symmetrization: Obstacle and Idea

Obstacle. Consider the deviation we want to bound:

$$\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_n(h)).$$

- $L_{\mathcal{D}}(h)$ depends on h's values *everywhere*, not just on S.
- Two hypotheses can agree on all of S but have different $L_{\mathcal{D}}$.
- So the previous counting argument breaks.

Idea. Introduce a *ghost sample* $S' \sim \mathcal{D}^n$ independent of S, and let $L'_n(h)$ be the empirical risk on S'.

For each fixed h, $\mathbb{E}_{S'}[L'_n(h)] = L_{\mathcal{D}}(h)$, so the population risk is itself an expected empirical risk.

Replacing $L_{\mathcal{D}}$ with L'_n turns the deviation into a difference of two empirical averages, so we can **count label patterns** on the $2n$ combined points of $S \cup S'$.

Symmetrization: Obstacle and Idea

Obstacle. Consider the deviation we want to bound:

$$\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_n(h)).$$

- $L_{\mathcal{D}}(h)$ depends on h's values *everywhere*, not just on S.
- Two hypotheses can agree on all of S but have different $L_{\mathcal{D}}$.
- So the previous counting argument breaks.

Idea. Introduce a *ghost sample* $S' \sim \mathcal{D}^n$ independent of S, and let $L'_n(h)$ be the empirical risk on S'.

For each fixed h, $\mathbb{E}_{S'}[L'_n(h)] = L_{\mathcal{D}}(h)$, so the population risk is itself an expected empirical risk.

Replacing $L_{\mathcal{D}}$ with L'_n turns the deviation into a difference of two empirical averages, so we can **count label patterns** on the $2n$ combined points of $S \cup S'$.

Why $|S'| = n$?

Symmetrization: Obstacle and Idea

Obstacle. Consider the deviation we want to bound:

$$\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_n(h)).$$

- $L_{\mathcal{D}}(h)$ depends on h's values *everywhere*, not just on S.
- Two hypotheses can agree on all of S but have different $L_{\mathcal{D}}$.
- So the previous counting argument breaks.

Idea. Introduce a *ghost sample* $S' \sim \mathcal{D}^n$ independent of S, and let $L'_n(h)$ be the empirical risk on S'.

For each fixed h, $\mathbb{E}_{S'}[L'_n(h)] = L_{\mathcal{D}}(h)$, so the population risk is itself an expected empirical risk.

Replacing $L_{\mathcal{D}}$ with L'_n turns the deviation into a difference of two empirical averages, so we can **count label patterns** on the $2n$ combined points of $S \cup S'$.

Why $|S'| = n$?

- Smaller ghost: L' fluctuates at $O(1)$, bound fails to shrink with n.

Symmetrization: Obstacle and Idea

Obstacle. Consider the deviation we want to bound:

$$\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_n(h)).$$

- $L_{\mathcal{D}}(h)$ depends on h's values *everywhere*, not just on S.
- Two hypotheses can agree on all of S but have different $L_{\mathcal{D}}$.
- So the previous counting argument breaks.

Idea. Introduce a *ghost sample* $S' \sim \mathcal{D}^n$ independent of S, and let $L'_n(h)$ be the empirical risk on S'.

For each fixed h, $\mathbb{E}_{S'}[L'_n(h)] = L_{\mathcal{D}}(h)$, so the population risk is itself an expected empirical risk.

Replacing $L_{\mathcal{D}}$ with L'_n turns the deviation into a difference of two empirical averages, so we can **count label patterns** on the $2n$ combined points of $S \cup S'$.

Why $|S'| = n$?

- Smaller ghost: L' fluctuates at $O(1)$, bound fails to shrink with n.
- Larger ghost: concentration is still bottlenecked by $|S| = n$, but we pay a larger $\Gamma_{\mathcal{H}}(|S| + |S'|)$.

Symmetrization inequality

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\left(L_{\mathcal{D}}(h)-L_n(h)\right)\right]\leq\mathbb{E}_{(S,S')}\left[\sup_{h\in\mathcal{H}}\left(L'_n(h)-L_n(h)\right)\right].$$

Symmetrization inequality

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\left(L_{\mathcal{D}}(h)-L_n(h)\right)\right]\leq\mathbb{E}_{(S,S')}\left[\sup_{h\in\mathcal{H}}\left(L'_n(h)-L_n(h)\right)\right].$$

Proof.

Symmetrization inequality

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{H}}(L_{\mathcal{D}}(h)-L_n(h))\right]\leq\mathbb{E}_{(S,S')}\left[\sup_{h\in\mathcal{H}}(L'_n(h)-L_n(h))\right].$$

Proof.

- For each fixed h: $L_{\mathcal{D}}(h)-L_n(h)=\mathbb{E}_{S'}[L'_n(h)-L_n(h)]$.

Symmetrization inequality

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{H}}(L_{\mathcal{D}}(h)-L_n(h))\right]\leq\mathbb{E}_{(S,S')}\left[\sup_{h\in\mathcal{H}}(L'_n(h)-L_n(h))\right].$$

Proof.

- For each fixed h: $L_{\mathcal{D}}(h)-L_n(h)=\mathbb{E}_{S'}[L'_n(h)-L_n(h)]$.
- Apply inside the supremum:

$$\sup_{h\in\mathcal{H}}(L_{\mathcal{D}}(h)-L_n(h))=\sup_{h\in\mathcal{H}}\mathbb{E}_{S'}[L'_n(h)-L_n(h)].$$

Symmetrization inequality

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{H}}(L_{\mathcal{D}}(h)-L_n(h))\right]\leq\mathbb{E}_{(S,S')}\left[\sup_{h\in\mathcal{H}}(L'_n(h)-L_n(h))\right].$$

Proof.

- For each fixed h: $L_{\mathcal{D}}(h)-L_n(h)=\mathbb{E}_{S'}[L'_n(h)-L_n(h)]$.
- Apply inside the supremum:

$$\sup_{h\in\mathcal{H}}(L_{\mathcal{D}}(h)-L_n(h))=\sup_{h\in\mathcal{H}}\mathbb{E}_{S'}[L'_n(h)-L_n(h)].$$

- Jensen's inequality (sup of expectation $\leq$ expectation of sup):

$$\sup_{h\in\mathcal{H}}\mathbb{E}_{S'}[L'_n(h)-L_n(h)]\leq\mathbb{E}_{S'}\left[\sup_{h\in\mathcal{H}}(L'_n(h)-L_n(h))\right].$$

Symmetrization inequality

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{H}}(L_{\mathcal{D}}(h)-L_n(h))\right]\leq\mathbb{E}_{(S,S')}\left[\sup_{h\in\mathcal{H}}(L'_n(h)-L_n(h))\right].$$

Proof.

- For each fixed h: $L_{\mathcal{D}}(h)-L_n(h)=\mathbb{E}_{S'}[L'_n(h)-L_n(h)]$.
- Apply inside the supremum:

$$\sup_{h\in\mathcal{H}}(L_{\mathcal{D}}(h)-L_n(h))=\sup_{h\in\mathcal{H}}\mathbb{E}_{S'}[L'_n(h)-L_n(h)].$$

- Jensen's inequality (sup of expectation $\leq$ expectation of sup):

$$\sup_{h\in\mathcal{H}}\mathbb{E}_{S'}[L'_n(h)-L_n(h)]\leq\mathbb{E}_{S'}\left[\sup_{h\in\mathcal{H}}(L'_n(h)-L_n(h))\right].$$

- Take $\mathbb{E}_S$ on both sides. □

Proof Sketch: Setup and Symmetrization

Step 1: Decomposition of the uniform deviation. Let $\Phi(S) = \sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_n(h))$. Write

$$\Phi(S) = \mathbb{E}[\Phi(S)] + \underbrace{(\Phi(S) - \mathbb{E}[\Phi(S)])}_{\leq \tilde{O}\left(\sqrt{\log(1/\delta)/n}\right)}.$$

The deviation term is bounded by concentration (McDiarmid); **details deferred to HW**.

Proof Sketch: Setup and Symmetrization

Step 1: Decomposition of the uniform deviation. Let $\Phi(S) = \sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_n(h))$. Write

$$\Phi(S) = \mathbb{E}[\Phi(S)] + \underbrace{(\Phi(S) - \mathbb{E}[\Phi(S)])}_{\leq \tilde{O}\left(\sqrt{\log(1/\delta)/n}\right)}.$$

The deviation term is bounded by concentration (McDiarmid); **details deferred to HW**.

Step 2: Symmetrization, conditioning, tail integration.

$$\begin{aligned}
\mathbb{E}[\Phi(S)] &\leq \mathbb{E}\left[\sup_{h \in \mathcal{H}} (L'_n(h) - L_n(h))\right] && \text{(symmetrization)} \\
&= \mathbb{E}\left[\mathbb{E}\left[\sup_{h \in \mathcal{H}} (L'_n(h) - L_n(h)) \mid S \cup S'\right]\right] && \text{(tower rule)} \\
&= \mathbb{E}\left[\int_0^\infty \mathbb{P}\left[\sup_{h \in \mathcal{H}} (L'_n(h) - L_n(h)) > t \mid S \cup S'\right] dt\right] && \text{(tail integration).}
\end{aligned}$$

The last step uses the identity $\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > t]\, dt$ for any nonnegative X.

Proof Sketch: Reducing to the Transductive Setting

After conditioning on $S \cup S'$:

Proof Sketch: Reducing to the Transductive Setting

After conditioning on $S \cup S'$:

- the combined sample is a fixed pool of $2n$ points;

Proof Sketch: Reducing to the Transductive Setting

After conditioning on $S \cup S'$:

- the combined sample is a fixed pool of $2n$ points;
- S is a uniformly random size-n subset, and S' is its complement.

Proof Sketch: Reducing to the Transductive Setting

After conditioning on $S \cup S'$:

- the combined sample is a fixed pool of $2n$ points;
- S is a uniformly random size-n subset, and S' is its complement.

This is **exactly the transductive setting** from part 1, with training size n, test size n, and pool $S \cup S'$.

Proof Sketch: Reducing to the Transductive Setting

After conditioning on $S \cup S'$:

- the combined sample is a fixed pool of $2n$ points;
- S is a uniformly random size-n subset, and S' is its complement.

This is **exactly the transductive setting** from part 1, with training size n, test size n, and pool $S \cup S'$.

The transductive uniform convergence bound, applied to this pool, gives

$$\mathbb{P}\left[\sup_{h \in \mathcal{H}} (L'_n(h) - L_n(h)) > t \mid S \cup S'\right] \leq \Gamma_{\mathcal{H}}(2n) \cdot \exp(-nt^2/2).$$

Proof Sketch: Reducing to the Transductive Setting

After conditioning on $S \cup S'$:

- the combined sample is a fixed pool of $2n$ points;
- S is a uniformly random size-n subset, and S' is its complement.

This is **exactly the transductive setting** from part 1, with training size n, test size n, and pool $S \cup S'$.

The transductive uniform convergence bound, applied to this pool, gives

$$\mathbb{P}\left[\sup_{h \in \mathcal{H}} (L'_n(h) - L_n(h)) > t \mid S \cup S'\right] \leq \Gamma_{\mathcal{H}}(2n) \cdot \exp(-nt^2/2).$$

Two final pieces:

Proof Sketch: Reducing to the Transductive Setting

After conditioning on $S \cup S'$:

- the combined sample is a fixed pool of $2n$ points;
- S is a uniformly random size-n subset, and S' is its complement.

This is **exactly the transductive setting** from part 1, with training size n, test size n, and pool $S \cup S'$.

The transductive uniform convergence bound, applied to this pool, gives

$$\mathbb{P}\left[\sup_{h \in \mathcal{H}} (L'_n(h) - L_n(h)) > t \mid S \cup S'\right] \leq \Gamma_{\mathcal{H}}(2n) \cdot \exp(-nt^2/2).$$

Two final pieces:

1. Integrate in t and plug back into step 2's tail-integration chain to bound $\Phi(S)$ with high probability.

Proof Sketch: Reducing to the Transductive Setting

After conditioning on $S \cup S'$:

- the combined sample is a fixed pool of $2n$ points;
- S is a uniformly random size-n subset, and S' is its complement.

This is **exactly the transductive setting** from part 1, with training size n, test size n, and pool $S \cup S'$.

The transductive uniform convergence bound, applied to this pool, gives

$$\mathbb{P}\left[\sup_{h \in \mathcal{H}} (L'_n(h) - L_n(h)) > t \mid S \cup S'\right] \leq \Gamma_{\mathcal{H}}(2n) \cdot \exp(-nt^2/2).$$

Two final pieces:

1. Integrate in t and plug back into step 2's tail-integration chain to bound $\Phi(S)$ with high probability.
2. Apply the ERM three-line chain (as in the finite-class proof) to turn the bound on $\Phi(S)$ into a bound on $L_{\mathcal{D}}\left(\hat{h}\right)$.

Growth-function ERM guarantee

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_n(h)$. With probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$,

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + O\left(\sqrt{\frac{\log \Gamma_{\mathcal{H}}(2n) + \log(1/\delta)}{n}}\right).$$

Growth-function ERM guarantee

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_n(h)$. With probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$,

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + O\left(\sqrt{\frac{\log \Gamma_{\mathcal{H}}(2n) + \log(1/\delta)}{n}}\right).$$

Recall Sauer–Shelah: for $d = \mathrm{VCdim}(\mathcal{H})$, $\log \Gamma_{\mathcal{H}}(2n) \leq d \log(2en/d)$. Plugging in:

Growth-function ERM guarantee

Let $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} L_n(h)$. With probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$,

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + O\left(\sqrt{\frac{\log \Gamma_{\mathcal{H}}(2n) + \log(1/\delta)}{n}}\right).$$

Recall Sauer–Shelah: for $d = \mathrm{VCdim}(\mathcal{H})$, $\log \Gamma_{\mathcal{H}}(2n) \leq d \log(2en/d)$. Plugging in:

VC corollary

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + O\left(\sqrt{\frac{d \log(2en/d) + \log(1/\delta)}{n}}\right).$$

Transductive and i.i.d.: Same Combinatorics, Different Target

Question	Transductive	i.i.d.
What is random?	The split on a fixed pool	The sample itself
Quantity to predict	Error on held-out points of a fixed pool	Expected error under the unknown distribution
Basic concentration	Sampling without replacement	Independent sampling
Uniform step	Count label patterns on the fixed pool	After symmetrization, count label patterns on a fixed $2n$-point pool
ERM step	Same three-line comparison to the best benchmark	Same three-line comparison to the best benchmark

Both models are controlled by counting *finite-sample behaviors*, not raw hypotheses.

VC Dimension Governs All Three Settings

One number, $d = \mathrm{VCdim}(\mathcal{H})$, controls all three settings:

VC Dimension Governs All Three Settings

One number, $d = \mathrm{VCdim}(\mathcal{H})$, controls all three settings:

- **Online transductive** (Week 2): Halving mistake bound $O(d \log(n/d))$.

VC Dimension Governs All Three Settings

One number, $d = \mathrm{VCdim}(\mathcal{H})$, controls all three settings:

- **Online transductive** (Week 2): Halving mistake bound $O(d \log(n/d))$.
- **Batch transductive** (part 1): ERM excess risk $\tilde{O}\left(\sqrt{d/n}\right)$ when $u \gtrsim n$.

VC Dimension Governs All Three Settings

One number, $d = \mathrm{VCdim}(\mathcal{H})$, controls all three settings:

- **Online transductive** (Week 2): Halving mistake bound $O(d \log(n/d))$.
- **Batch transductive** (part 1): ERM excess risk $\tilde{O}\left(\sqrt{d/n}\right)$ when $u \gtrsim n$.
- **i.i.d.** (part 2): ERM excess risk $\tilde{O}\left(\sqrt{d/n}\right)$.

VC Dimension Governs All Three Settings

One number, $d = \mathrm{VCdim}(\mathcal{H})$, controls all three settings:

- **Online transductive** (Week 2): Halving mistake bound $O(d \log(n/d))$.
- **Batch transductive** (part 1): ERM excess risk $\tilde{O}\left(\sqrt{d/n}\right)$ when $u \gtrsim n$.
- **i.i.d.** (part 2): ERM excess risk $\tilde{O}\left(\sqrt{d/n}\right)$.

VC dimension is the same complexity measure governing learnability in all three settings.

PAC Learning

So far our i.i.d. theorems have the form

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon(n, \delta).$$

So far our i.i.d. theorems have the form

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon(n, \delta).$$

We now repackage this guarantee as a question of *learnability*: when is a class learnable, and how many samples does it take?

So far our i.i.d. theorems have the form

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon(n, \delta).$$

We now repackage this guarantee as a question of *learnability*: when is a class learnable, and how many samples does it take?

PAC stands for **Probably Approximately Correct** (Valiant, 1984).

So far our i.i.d. theorems have the form

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon(n, \delta).$$

We now repackage this guarantee as a question of *learnability*: when is a class learnable, and how many samples does it take?

PAC stands for **Probably Approximately Correct** (Valiant, 1984).

- **Approximately correct:** $\hat{h}$ has error at most ε above the best in $\mathcal{H}$.

So far our i.i.d. theorems have the form

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon(n, \delta).$$

We now repackage this guarantee as a question of *learnability*: when is a class learnable, and how many samples does it take?

PAC stands for **Probably Approximately Correct** (Valiant, 1984).

- **Approximately correct:** $\hat{h}$ has error at most ε above the best in $\mathcal{H}$.
- **Probably:** this holds with probability at least $1 - \delta$ over S.

So far our i.i.d. theorems have the form

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon(n, \delta).$$

We now repackage this guarantee as a question of *learnability*: when is a class learnable, and how many samples does it take?

PAC stands for **Probably Approximately Correct** (Valiant, 1984).

- **Approximately correct:** $\hat{h}$ has error at most ε above the best in $\mathcal{H}$.
- **Probably:** this holds with probability at least $1 - \delta$ over S.
- **Sample complexity:** the smallest n, as a function of ε and δ, for which the above holds uniformly over all distributions $\mathcal{D}$.

So far our i.i.d. theorems have the form

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon(n, \delta).$$

We now repackage this guarantee as a question of *learnability*: when is a class learnable, and how many samples does it take?

PAC stands for **Probably Approximately Correct** (Valiant, 1984).

- **Approximately correct:** $\hat{h}$ has error at most ε above the best in $\mathcal{H}$.
- **Probably:** this holds with probability at least $1 - \delta$ over S.
- **Sample complexity:** the smallest n, as a function of ε and δ, for which the above holds uniformly over all distributions $\mathcal{D}$.

PAC is the vocabulary for *learnability* and *sample complexity*.

Realizable PAC learning

A class $\mathcal{H}$ is PAC learnable if there is a learning rule $A : \bigcup_{n \geq 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ such that for every $\varepsilon, \delta > 0$, there exists a sample size $n_{\mathcal{H}}(\varepsilon, \delta)$ with the following property.

For every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ that is realizable by $\mathcal{H}$, meaning that

$$\exists h^* \in \mathcal{H} \text{ with } L_{\mathcal{D}}(h^*) = 0,$$

if $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$ and $S \sim \mathcal{D}^n$, then

$$\mathbb{P}(L_{\mathcal{D}}(A(S)) \leq \varepsilon) \geq 1 - \delta.$$

Realizable PAC learning

A class $\mathcal{H}$ is PAC learnable if there is a learning rule $A : \bigcup_{n \geq 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ such that for every $\varepsilon, \delta > 0$, there exists a sample size $n_{\mathcal{H}}(\varepsilon, \delta)$ with the following property.

For every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ that is realizable by $\mathcal{H}$, meaning that

$$\exists h^* \in \mathcal{H} \text{ with } L_{\mathcal{D}}(h^*) = 0,$$

if $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$ and $S \sim \mathcal{D}^n$, then

$$\mathbb{P}(L_{\mathcal{D}}(A(S)) \leq \varepsilon) \geq 1 - \delta.$$

Realizability is a modeling assumption about the world: nature lies in $\mathcal{H}$. The benchmark is absolute error, and the goal is to drive it below ε.

Realizable PAC learning

A class $\mathcal{H}$ is PAC learnable if there is a learning rule $A : \bigcup_{n\geq 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ such that for every $\varepsilon, \delta > 0$, there exists a sample size $n_{\mathcal{H}}(\varepsilon, \delta)$ with the following property.

For every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ that is realizable by $\mathcal{H}$, meaning that

$$\exists h^* \in \mathcal{H} \text{ with } L_{\mathcal{D}}(h^*) = 0,$$

if $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$ and $S \sim \mathcal{D}^n$, then

$$\mathbb{P}(L_{\mathcal{D}}(A(S)) \leq \varepsilon) \geq 1 - \delta.$$

Realizability is a modeling assumption about the world: nature lies in $\mathcal{H}$. The benchmark is absolute error, and the goal is to drive it below ε.

Example. Thresholds on $\mathbb{R}$ with labels generated by an unknown true threshold $h_{\theta^*}(x) = \mathbf{1}[x \leq \theta^*]$. Then $L_{\mathcal{D}}(h_{\theta^*}) = 0$.

Agnostic PAC learning

A class $\mathcal{H}$ is agnostically PAC learnable if there is a learning rule $A : \bigcup_{n \geq 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ such that for every $\varepsilon, \delta > 0$, there exists $n_{\mathcal{H}}(\varepsilon, \delta)$ with the following property.

For every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, if $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$ and $S \sim \mathcal{D}^n$, then

$$\mathbb{P}\Big(L_{\mathcal{D}}(A(S)) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon \Big) \geq 1 - \delta.$$

Agnostic PAC learning

A class $\mathcal{H}$ is agnostically PAC learnable if there is a learning rule $A : \bigcup_{n \geq 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ such that for every $\varepsilon, \delta > 0$, there exists $n_{\mathcal{H}}(\varepsilon, \delta)$ with the following property.

For every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, if $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$ and $S \sim \mathcal{D}^n$, then

$$\mathbb{P}\Big(L_{\mathcal{D}}(A(S)) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon \Big) \geq 1 - \delta.$$

No realizability assumption; the benchmark is the *best in class*, $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$.

Agnostic PAC learning

A class $\mathcal{H}$ is agnostically PAC learnable if there is a learning rule $A : \bigcup_{n \geq 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ such that for every $\varepsilon, \delta > 0$, there exists $n_{\mathcal{H}}(\varepsilon, \delta)$ with the following property.

For every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, if $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$ and $S \sim \mathcal{D}^n$, then

$$\mathbb{P}\Big(L_{\mathcal{D}}(A(S)) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon \Big) \geq 1 - \delta.$$

No realizability assumption; the benchmark is the *best in class*, $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$.

Agnostic subsumes realizable: plugging $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = 0$ recovers the realizable guarantee.

Agnostic PAC learning

A class $\mathcal{H}$ is agnostically PAC learnable if there is a learning rule $A : \bigcup_{n \geq 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ such that for every $\varepsilon, \delta > 0$, there exists $n_{\mathcal{H}}(\varepsilon, \delta)$ with the following property.

For every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, if $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$ and $S \sim \mathcal{D}^n$, then

$$\mathbb{P}\Big(L_{\mathcal{D}}(A(S)) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon \Big) \geq 1 - \delta.$$

No realizability assumption; the benchmark is the *best in class*, $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$.

Agnostic subsumes realizable: plugging $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = 0$ recovers the realizable guarantee.

Example. $\mathcal{H}$ is thresholds on $\mathbb{R}$, x uniform on $[0, 3]$, and $y = \mathbf{1}[x \in [1, 2]]$. No threshold agrees with this labeling on all of $[0, 3]$, so $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) > 0$.

Recall the finite-class ERM theorem:

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + 2\sqrt{\frac{\log|\mathcal{H}| + \log\left(\frac{2}{\delta}\right)}{2n}}.$$

Recall the finite-class ERM theorem:

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + 2\sqrt{\frac{\log|\mathcal{H}| + \log\left(\frac{2}{\delta}\right)}{2n}}.$$

Setting the deviation term $\leq \varepsilon$ and solving for n:

$$n_{\mathcal{H}}(\varepsilon, \delta) = 2\frac{\log|\mathcal{H}| + \log\left(\frac{2}{\delta}\right)}{\varepsilon^2}.$$

Recall the finite-class ERM theorem:

$$L_{\mathcal{D}}\big(\hat{h}\big) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + 2\sqrt{\frac{\log|\mathcal{H}| + \log\big(\frac{2}{\delta}\big)}{2n}}.$$

Setting the deviation term $\leq \varepsilon$ and solving for n:

$$n_{\mathcal{H}}(\varepsilon, \delta) = 2\frac{\log|\mathcal{H}| + \log\big(\frac{2}{\delta}\big)}{\varepsilon^2}.$$

- Hoeffding's $1/\sqrt{n}$ rate becomes $1/\varepsilon^2$ in sample complexity: *halving ε quadruples the samples.*

Recall the finite-class ERM theorem:

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + 2\sqrt{\frac{\log|\mathcal{H}| + \log\left(\frac{2}{\delta}\right)}{2n}}.$$

Setting the deviation term $\leq \varepsilon$ and solving for n:

$$n_{\mathcal{H}}(\varepsilon, \delta) = 2\frac{\log|\mathcal{H}| + \log\left(\frac{2}{\delta}\right)}{\varepsilon^2}.$$

- Hoeffding's $1/\sqrt{n}$ rate becomes $1/\varepsilon^2$ in sample complexity: *halving ε quadruples the samples.*
- Dependence on $|\mathcal{H}|$ is *logarithmic*, so doubling the class only adds a constant.

Recall the finite-class ERM theorem:

$$L_{\mathcal{D}}\left(\hat{h}\right) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + 2\sqrt{\frac{\log|\mathcal{H}| + \log\left(\frac{2}{\delta}\right)}{2n}}.$$

Setting the deviation term $\leq \varepsilon$ and solving for n:

$$n_{\mathcal{H}}(\varepsilon, \delta) = 2\frac{\log|\mathcal{H}| + \log\left(\frac{2}{\delta}\right)}{\varepsilon^2}.$$

- Hoeffding's $1/\sqrt{n}$ rate becomes $1/\varepsilon^2$ in sample complexity: *halving ε quadruples the samples.*
- Dependence on $|\mathcal{H}|$ is *logarithmic*, so doubling the class only adds a constant.

Every finite class is agnostically PAC learnable by ERM. The same inversion converts the growth-function and VC bounds into their own PAC learnability statements.

Realizable Thresholds: Key Observation

$\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}\}$, $h_\theta(x) = \mathbf{1}[x \le \theta]$, and realizability: some θ^* has $y = \mathbf{1}[x \le \theta^*]$ almost surely.

Realizable Thresholds: Key Observation

$\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}\}$, $h_\theta(x) = \mathbf{1}[x \le \theta]$, and realizability: some θ^* has $y = \mathbf{1}[x \le \theta^*]$ almost surely.

ERM. Any $A(S)$ consistent with S. Under realizability it has empirical risk 0, hence is an ERM.

Realizable Thresholds: Key Observation

$\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}\}$, $h_\theta(x) = \mathbf{1}[x \le \theta]$, and realizability: some θ^* has $y = \mathbf{1}[x \le \theta^*]$ almost surely.

ERM. Any $A(S)$ consistent with S. Under realizability it has empirical risk 0, hence is an ERM.

Key observation. For any θ, the population error equals the probability mass of the disagreement interval between θ and θ^*. So we just need $A(S)$ close to θ^* *in mass*, not in distance.

Realizable Thresholds: Key Observation

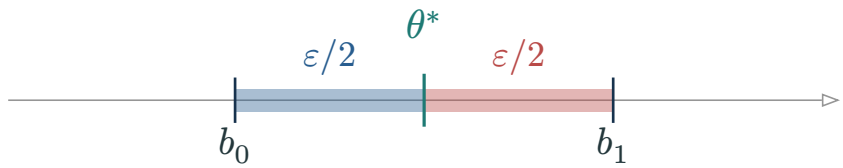
$\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}\}$, $h_\theta(x) = \mathbf{1}[x \le \theta]$, and realizability: some θ^* has $y = \mathbf{1}[x \le \theta^*]$ almost surely.

ERM. Any $A(S)$ consistent with S. Under realizability it has empirical risk 0, hence is an ERM.

Key observation. For any θ, the population error equals the probability mass of the disagreement interval between θ and θ^*. So we just need $A(S)$ close to θ^* *in mass*, not in distance.

Landmarks. Assume x has a density. Fix $b_0 < \theta^* < b_1$ (before seeing S) such that

$$\mathbb{P}_x(b_0 < x \le \theta^*) = \frac{\varepsilon}{2} \quad \text{and} \quad \mathbb{P}_x(\theta^* < x \le b_1) = \frac{\varepsilon}{2}.$$



Realizable Thresholds: Sandwich Argument

Claim. If S contains a positive point in $(b_0, \theta^*]$ *and* a negative point in $(\theta^*, b_1]$, then every consistent $A(S)$ has $L_{\mathcal{D}}\left(h_{A(S)}\right) \leq \varepsilon$.

Realizable Thresholds: Sandwich Argument

Claim. If S contains a positive point in $(b_0, \theta^*]$ *and* a negative point in $(\theta^*, b_1]$, then every consistent $A(S)$ has $L_{\mathcal{D}}\left(h_{A(S)}\right) \leq \varepsilon$.

> **Proof.** A positive sample $x^+ \in (b_0, \theta^*]$ forces $A(S) \geq x^+ > b_0$. A negative sample $x^- \in (\theta^*, b_1]$ forces $A(S) < x^- \leq b_1$. So $A(S) \in (b_0, b_1)$, and its disagreement interval is contained in $(b_0, b_1]$, whose total mass is $\leq \varepsilon$.

Realizable Thresholds: Sandwich Argument

Claim. If S contains a positive point in $(b_0, \theta^*]$ *and* a negative point in $(\theta^*, b_1]$, then every consistent $A(S)$ has $L_{\mathcal{D}}\left(h_{A(S)}\right) \leq \varepsilon$.

> **Proof.** A positive sample $x^+ \in (b_0, \theta^*]$ forces $A(S) \geq x^+ > b_0$. A negative sample $x^- \in (\theta^*, b_1]$ forces $A(S) < x^- \leq b_1$. So $A(S) \in (b_0, b_1)$, and its disagreement interval is contained in $(b_0, b_1]$, whose total mass is $\leq \varepsilon$.

Sample complexity. The bad event $L_{\mathcal{D}}\left(h_{A(S)}\right) > \varepsilon$ requires at least one of the two intervals to be empty of samples. By union bound:

$$\mathbb{P}\left(L_{\mathcal{D}}\left(h_{A(S)}\right) > \varepsilon\right) \leq 2\left(1 - \frac{\varepsilon}{2}\right)^n \leq 2e^{-\varepsilon\frac{n}{2}}.$$

Realizable Thresholds: Sandwich Argument

Claim. If S contains a positive point in $(b_0, \theta^*]$ *and* a negative point in $(\theta^*, b_1]$, then every consistent $A(S)$ has $L_{\mathcal{D}}\left(h_{A(S)}\right) \leq \varepsilon$.

> **Proof.** A positive sample $x^+ \in (b_0, \theta^*]$ forces $A(S) \geq x^+ > b_0$. A negative sample $x^- \in (\theta^*, b_1]$ forces $A(S) < x^- \leq b_1$. So $A(S) \in (b_0, b_1)$, and its disagreement interval is contained in $(b_0, b_1]$, whose total mass is $\leq \varepsilon$.

Sample complexity. The bad event $L_{\mathcal{D}}\left(h_{A(S)}\right) > \varepsilon$ requires at least one of the two intervals to be empty of samples. By union bound:

$$\mathbb{P}\left(L_{\mathcal{D}}\left(h_{A(S)}\right) > \varepsilon\right) \leq 2\left(1 - \frac{\varepsilon}{2}\right)^n \leq 2e^{-\varepsilon\frac{n}{2}}.$$

Setting this $\leq \delta$ and solving:

$$n_{\mathcal{H}}(\varepsilon, \delta) = \frac{2}{\varepsilon}\log\left(\frac{2}{\delta}\right).$$

Realizable Thresholds: Sandwich Argument

Claim. If S contains a positive point in $(b_0, \theta^*]$ *and* a negative point in $(\theta^*, b_1]$, then every consistent $A(S)$ has $L_{\mathcal{D}}\left(h_{A(S)}\right) \leq \varepsilon$.

> **Proof.** A positive sample $x^+ \in (b_0, \theta^*]$ forces $A(S) \geq x^+ > b_0$. A negative sample $x^- \in (\theta^*, b_1]$ forces $A(S) < x^- \leq b_1$. So $A(S) \in (b_0, b_1)$, and its disagreement interval is contained in $(b_0, b_1]$, whose total mass is $\leq \varepsilon$.

Sample complexity. The bad event $L_{\mathcal{D}}\left(h_{A(S)}\right) > \varepsilon$ requires at least one of the two intervals to be empty of samples. By union bound:

$$\mathbb{P}\left(L_{\mathcal{D}}\left(h_{A(S)}\right) > \varepsilon\right) \leq 2\left(1 - \frac{\varepsilon}{2}\right)^n \leq 2e^{-\varepsilon\frac{n}{2}}.$$

Setting this $\leq \delta$ and solving:

$$n_{\mathcal{H}}(\varepsilon, \delta) = \frac{2}{\varepsilon}\log\left(\frac{2}{\delta}\right).$$

Thresholds are realizably PAC learnable at a $1/\varepsilon$ rate instead of $1/\varepsilon^2$.

Recall: the Uniform Convergence Property

$\mathcal{H}$ has the **uniform convergence property** if for every $\varepsilon, \delta > 0$, there is $n_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta)$ such that for every distribution $\mathcal{D}$ and every $n \geq n_{\mathcal{H}}^{\mathrm{UC}}$,

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_n(h)| \leq \varepsilon \right) \geq 1 - \delta.$$

Recall: the Uniform Convergence Property

$\mathcal{H}$ has the **uniform convergence property** if for every $\varepsilon, \delta > 0$, there is $n_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta)$ such that for every distribution $\mathcal{D}$ and every $n \geq n_{\mathcal{H}}^{\mathrm{UC}}$,

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_n(h)| \leq \varepsilon \right) \geq 1 - \delta.$$

For finite classes, this is exactly what we proved. For finite-VC classes, the proof sketch above gives the bound up to the usual logarithmic refinements, and a sharper analysis yields:

Recall: the Uniform Convergence Property

$\mathcal{H}$ has the **uniform convergence property** if for every $\varepsilon, \delta > 0$, there is $n_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta)$ such that for every distribution $\mathcal{D}$ and every $n \geq n_{\mathcal{H}}^{\mathrm{UC}}$,

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_n(h)| \leq \varepsilon \right) \geq 1 - \delta.$$

For finite classes, this is exactly what we proved. For finite-VC classes, the proof sketch above gives the bound up to the usual logarithmic refinements, and a sharper analysis yields:

- **Finite classes:** $n_{\mathcal{H}}^{\mathrm{UC}} = O((\log|\mathcal{H}| + \log(1/\delta))/\varepsilon^2)$ via union bound.

Recall: the Uniform Convergence Property

$\mathcal{H}$ has the **uniform convergence property** if for every $\varepsilon, \delta > 0$, there is $n_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta)$ such that for every distribution $\mathcal{D}$ and every $n \geq n_{\mathcal{H}}^{\mathrm{UC}}$,

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_n(h)| \leq \varepsilon \right) \geq 1 - \delta.$$

For finite classes, this is exactly what we proved. For finite-VC classes, the proof sketch above gives the bound up to the usual logarithmic refinements, and a sharper analysis yields:

- **Finite classes:** $n_{\mathcal{H}}^{\mathrm{UC}} = O((\log|\mathcal{H}| + \log(1/\delta))/\varepsilon^2)$ via union bound.
- **Finite-VC classes:** $n_{\mathcal{H}}^{\mathrm{UC}} = O((d + \log(1/\delta))/\varepsilon^2)$ with a sharper VC analysis.

Fundamental theorem of PAC learning

Let $\mathcal{H}$ be a binary hypothesis class with $d = \mathrm{VCdim}(\mathcal{H})$. The following are equivalent:

1. $\mathcal{H}$ has the uniform convergence property.
2. ERM is a successful agnostic PAC learner for $\mathcal{H}$.
3. $\mathcal{H}$ is agnostically PAC learnable.
4. $\mathcal{H}$ is (realizably) PAC learnable.
5. $\mathcal{H}$ has finite VC dimension ($d < \infty$).

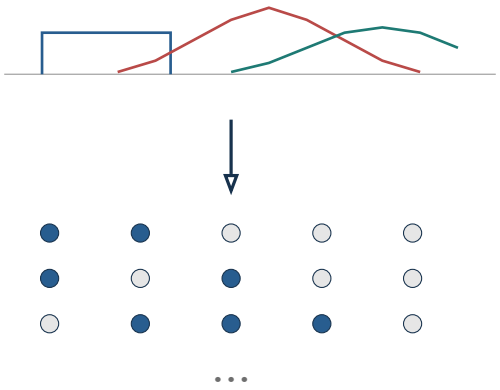
Moreover, when $d < \infty$, there exist absolute constants $C_1, C_2 > 0$ such that

$$C_1 \frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2} \leq n_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2} \quad \text{(uniform convergence, agnostic)},$$

$$C_1 \frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon} \leq n_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon} \quad \text{(realizable)}.$$

Interpreting the Fundamental Theorem

The theorem is striking: it turns an infinite-dimensional question into a finite combinatorial one.

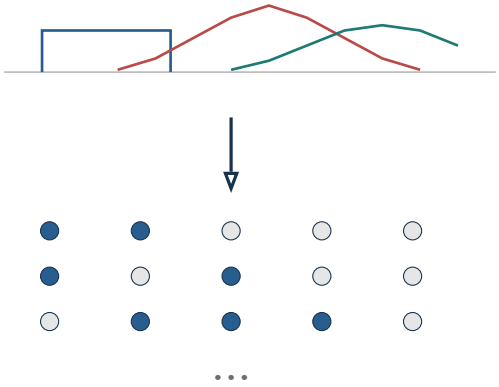


… *"Can we learn from every $\mathcal{D}$?"*

"How many labelings on n points?"

Interpreting the Fundamental Theorem

The theorem is striking: it turns an infinite-dimensional question into a finite combinatorial one.



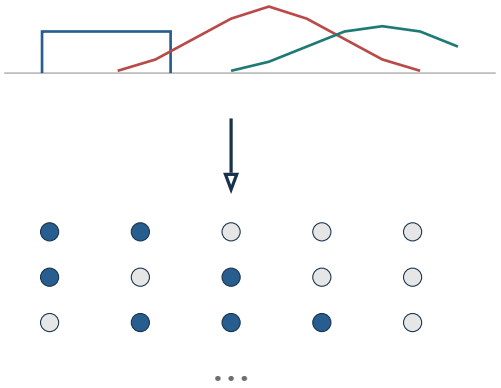
… *"Can we learn from every $\mathcal{D}$?"*

"How many labelings on n points?"

- VC dimension is *necessary and sufficient* for PAC learnability.

Interpreting the Fundamental Theorem

The theorem is striking: it turns an infinite-dimensional question into a finite combinatorial one.



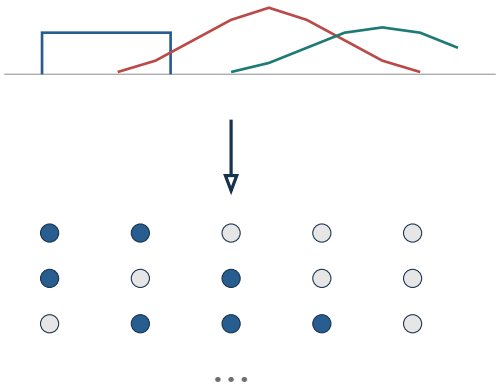
… *"Can we learn from every $\mathcal{D}$?"*

"How many labelings on n points?"

- VC dimension is *necessary and sufficient* for PAC learnability.
- ERM is *sufficient* whenever PAC learning is possible: if any rule PAC-learns $\mathcal{H}$, then ERM does too.

Interpreting the Fundamental Theorem

The theorem is striking: it turns an infinite-dimensional question into a finite combinatorial one.



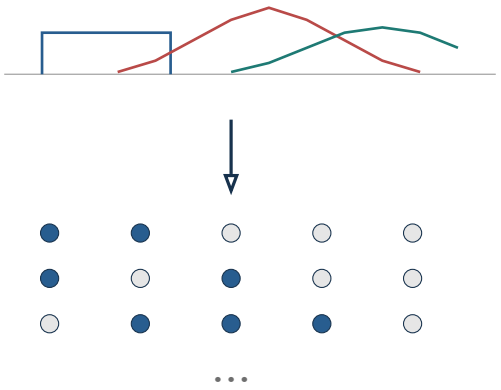
… *"Can we learn from every $\mathcal{D}$?"*

"How many labelings on n points?"

- VC dimension is *necessary and sufficient* for PAC learnability.
- ERM is *sufficient* whenever PAC learning is possible: if any rule PAC-learns $\mathcal{H}$, then ERM does too.
- Uniform convergence *exactly characterizes* PAC learnability.

Interpreting the Fundamental Theorem

The theorem is striking: it turns an infinite-dimensional question into a finite combinatorial one.



… *"Can we learn from every $\mathcal{D}$?"*

"How many labelings on n points?"

- VC dimension is *necessary and sufficient* for PAC learnability.
- ERM is *sufficient* whenever PAC learning is possible: if any rule PAC-learns $\mathcal{H}$, then ERM does too.
- Uniform convergence *exactly characterizes* PAC learnability.

PAC learnability is *combinatorial*: a class is learnable exactly when its finite-sample behavior is controlled.

Summary

How the Pieces Fit

- **Transductive learning:** fixed pool, random split, predict the test points.

How the Pieces Fit

- **Transductive learning:** fixed pool, random split, predict the test points.
- **i.i.d. learning:** random sample from an unknown distribution, predict future draws.

How the Pieces Fit

- **Transductive learning:** fixed pool, random split, predict the test points.
- **i.i.d. learning:** random sample from an unknown distribution, predict future draws.
- **PAC learning:** language for summarizing the i.i.d. guarantees as learnability and sample complexity.

How the Pieces Fit

- **Transductive learning:** fixed pool, random split, predict the test points.
- **i.i.d. learning:** random sample from an unknown distribution, predict future draws.
- **PAC learning:** language for summarizing the i.i.d. guarantees as learnability and sample complexity.

The combinatorial core is the same in both the transductive and i.i.d. settings:

$$\text{finite-sample behaviors} \rightarrow \text{generalization guarantees.}$$