

CMPUT 267: Machine Learning I · Fall 2026

Understanding Prediction Error

Xi Ye

2026-10-20

Notation Reference

D : n training examples; $\mathcal{F}$: candidate class; ℓ : loss on one example.

Symbol / definition	Meaning
$L(f) = \mathbb{E}_{\mathbf{X}, Y}[\ell(f(\mathbf{X}), Y)]$	Expected risk: average loss over the population.
$\hat{L}_D(f) = \frac{1}{n} \sum_{i=1}^n \ell(f(\mathbf{x}_i), y_i)$	Empirical risk: average loss on training data D .
$f_{\text{Bayes}} \in \arg \min_f L(f)$	Bayes predictor: best expected risk over all predictors of the available inputs.
$f^* \in \arg \min_{f \in \mathcal{F}} L(f)$	Best in class: minimizes expected risk within $\mathcal{F}$.
$\hat{f}_D \in \arg \min_{f \in \mathcal{F}} \hat{L}_D(f)$	Learned predictor: minimizes empirical risk within $\mathcal{F}$ using D .

Recall: Evaluating Models in Practice

Last lecture, we studied practical choices that affect predictions on new examples.

1 · Model Complexity

How flexible should the predictor be?

2 · Training-Data Size

How does more training data change loss?

3 · Training Steps

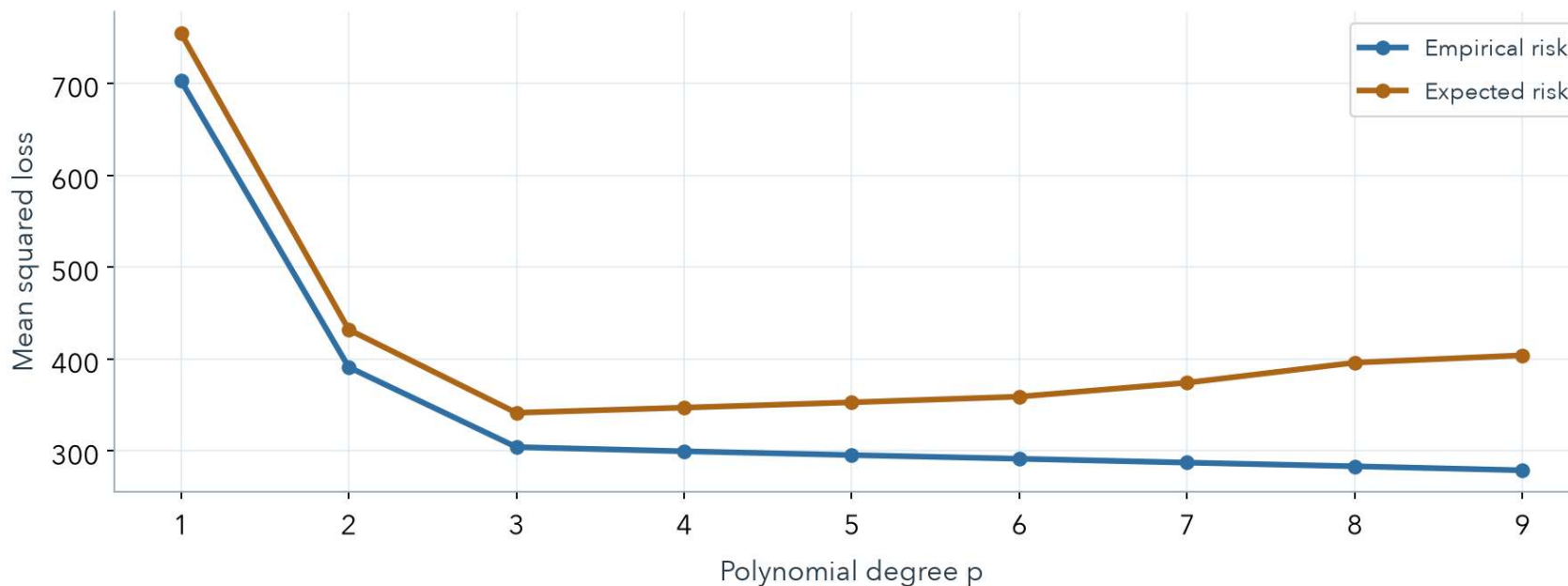
Does more optimization improve predictions?

4 · Evaluation in Practice

Split data for training, validation, and testing.

Recall: Model Complexity

For the same training-set size, increasing polynomial degree lowers the best training loss.



? Why does expected risk first fall, then rise as the model becomes more flexible?

This Lecture: Analyzing Prediction Error

Our goal is to make **expected risk** on new examples small.

Expected Risk

Average loss over the population:

$$L(f) = \mathbb{E}_{\mathbf{X}, Y}[\ell(f(\mathbf{X}), Y)].$$

Empirical Risk

Average loss on the training dataset:

$$\hat{L}_D(f) = \frac{1}{n} \sum_{i=1}^n \ell(f(\mathbf{x}_i), y_i).$$

i We fit $\hat{f}_D$ by minimizing empirical risk. We analyze its expected risk, averaged over training datasets: $\mathbb{E}_D[L(\hat{f}_D)]$.

Three Sources of Prediction Error

Expected risk = Estimation error + Approximation error + Irreducible error.

Estimation error: learning from a finite sample instead of knowing the true distribution.

Approximation error: restricting the predictor to a chosen function class.

Irreducible error: loss that remains even with the true distribution and the best predictor among all functions of the available inputs.

Decomposing Expected Risk

$$\begin{aligned}\mathbb{E}_D[L(\hat{f}_D)] &= \underbrace{\mathbb{E}_D[L(\hat{f}_D)] - L(f^*)}_{\text{Estimation error}} \\ &\quad + \underbrace{L(f^*) - L(f_{\text{Bayes}})}_{\text{Approximation error}} + \underbrace{L(f_{\text{Bayes}})}_{\text{Irreducible error}} .\end{aligned}$$

- f_{Bayes} : the predictor with the smallest expected risk among **all functions** of the available inputs.
- f^* : the predictor with the smallest expected risk **within our chosen function class** $\mathcal{F}$.

Irreducible Error

$$\text{Irreducible error} = L(f_{\text{Bayes}}).$$

The **Bayes optimal predictor** achieves the smallest expected risk among all functions of the available inputs:

$$f_{\text{Bayes}} \in \arg \min_{f: \mathcal{X} \rightarrow \mathcal{Y}} L(f).$$

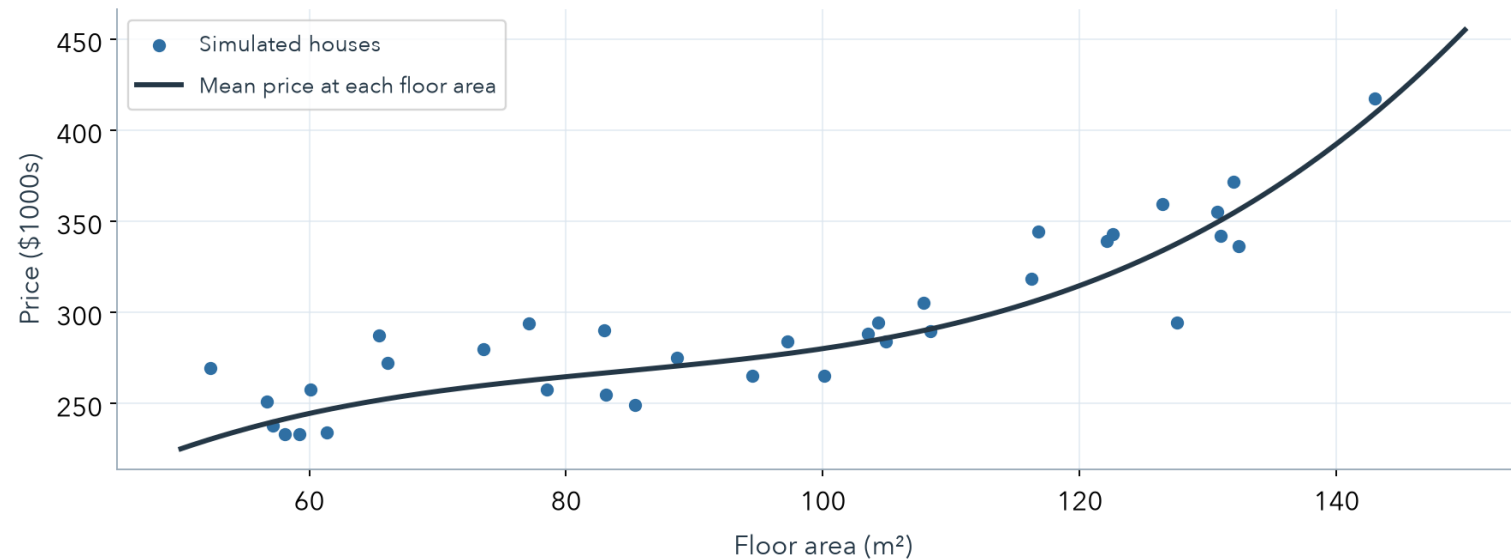


This is a **hypothetical reference**: assume we know the full population, or the joint distribution of $\mathbf{X}$ and Y .

Example: The Bayes Predictor for House Prices

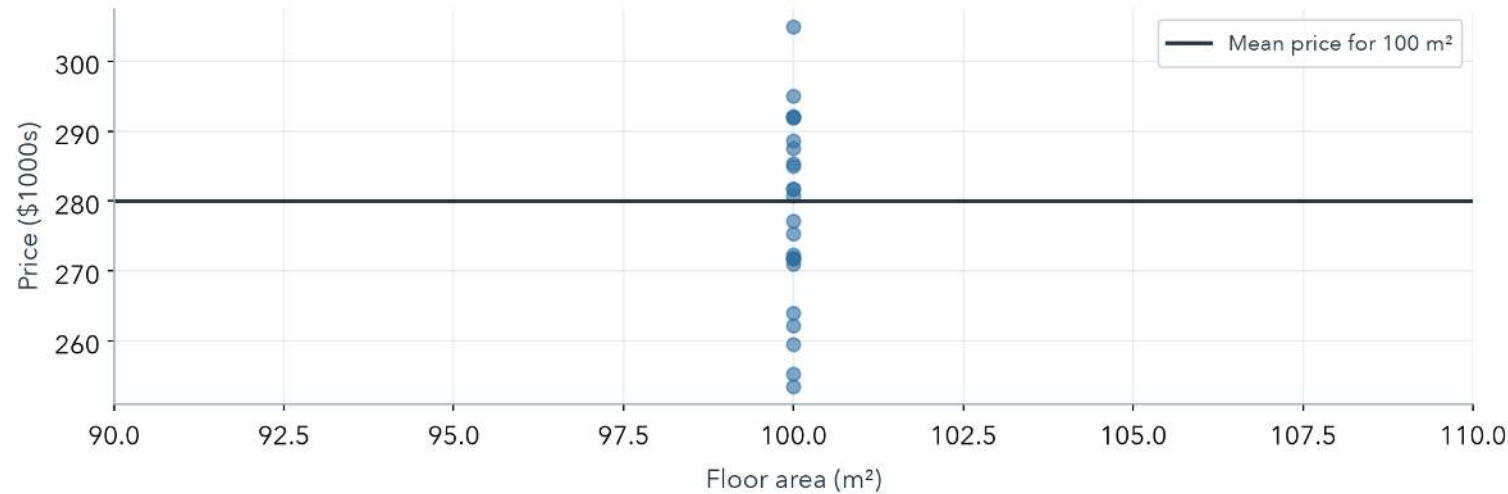
Use the same houses as last lecture: predict **price** from **floor area**.

For squared loss, the curve is $f_{\text{Bayes}}(x) = \mathbb{E}[Y \mid X = x]$.



Why the Error Is Irreducible

Even for **one fixed input**, $x = 100 \text{ m}^2$, there can be **different prices** y .



i Any predictor must give one value at this x . Even $f_{\text{Bayes}}(100) = 280$ cannot match every price: randomness remains in Y given X .

In this case, $L(f_{\text{Bayes}}) = \mathbb{E}_X[\text{Var}(Y \mid X)]$

More Informative Features

For real houses, age and location may explain price differences that floor area alone cannot.

$$\mathbf{X} = \text{Floor area} \longrightarrow \mathbf{X}' = (\text{Floor area}, \text{Age}, \text{Location}).$$



More informative inputs can reduce irreducible error. “Irreducible” is relative to the information available to the predictor.

Three Sources of Prediction Error

Expected risk = Estimation error + **Approximation error** + Irreducible error.

Estimation error: learning from a finite sample instead of knowing the true distribution.

Approximation error: restricting the predictor to a chosen function class.

Irreducible error: loss that remains even with the true distribution and the best predictor among all functions of the available inputs.

 Next: **approximation error**, the cost of restricting our function class.

Approximation Error

$$\text{Approximation error} = L(f^*) - L(f_{\text{Bayes}}).$$

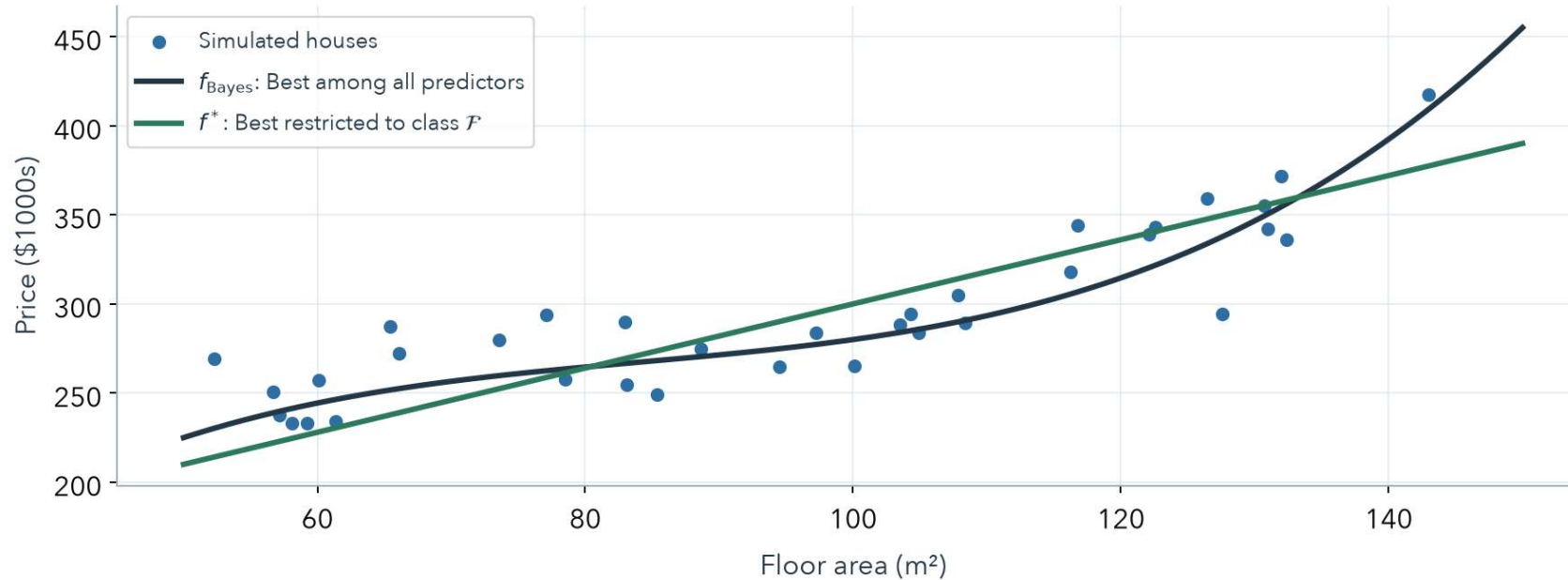
The **best predictor in the class**, f^* , minimizes expected risk within our chosen function class $\mathcal{F}$:

$$f^* \in \arg \min_{f \in \mathcal{F}} L(f).$$

 Again assume we know the full population. This time, we must choose a predictor from $\mathcal{F}$.

Example: The Best Line

Keep the same house-price population, but allow **only straight-line predictors**.



i f_{Bayes} is best among all predictors. f^* is best when restricted to $\mathcal{F}$. In this example, we choose from the class of straight lines.

Why There Is Approximation Error

We restrict our candidates to a function class $\mathcal{F}$.

The best predictor **within that class** may still be worse than the best predictor **among all possible functions**.

 In our example, the best straight line cannot match the curved mean price.

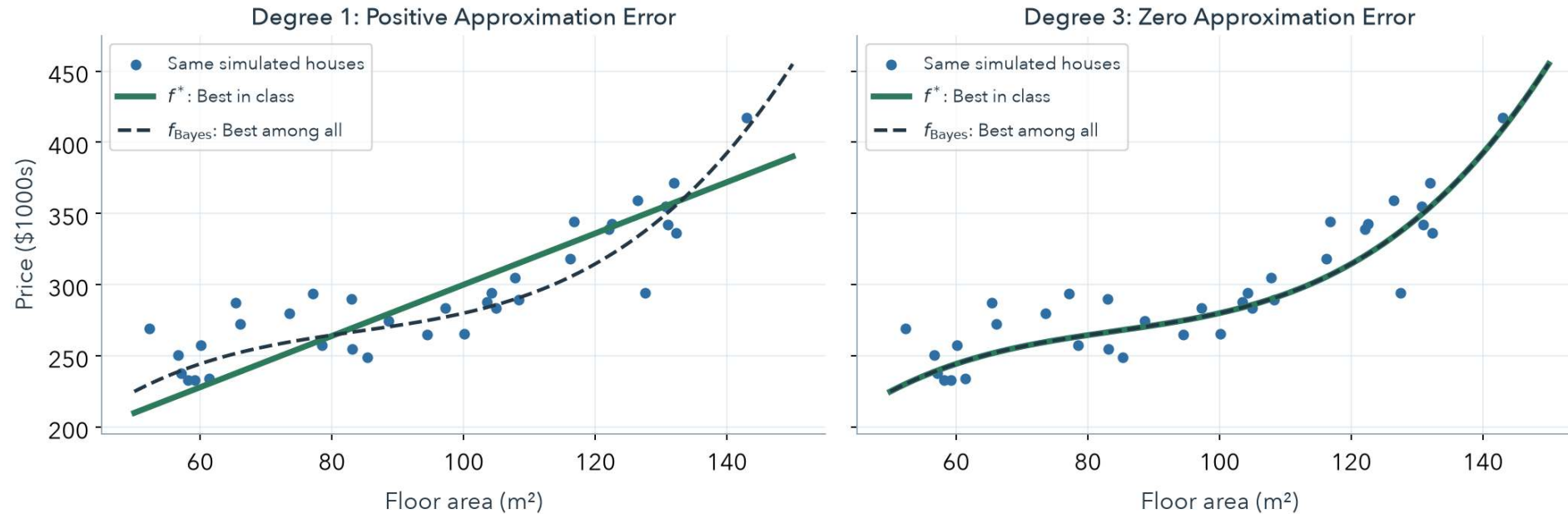
 How can we reduce approximation error?

Use a **richer function class**: allow more candidate predictors.

A Richer Function Class

Allowing more functions can reduce approximation error. If $\mathcal{F} \subseteq \mathcal{G}$:

$$\min_{f \in \mathcal{G}} L(f) \leq \min_{f \in \mathcal{F}} L(f).$$



Three Sources of Prediction Error

Expected risk = **Estimation error** + Approximation error + Irreducible error.

Estimation error: learning from a finite sample instead of knowing the true distribution.

Approximation error: restricting the predictor to a chosen function class.

Irreducible error: loss that remains even with the true distribution and the best predictor among all functions of the available inputs.

 Next: **estimation error**, the cost of learning from a finite sample.

Estimation Error

$$\text{EE} = \mathbb{E}_D[L(\hat{f}_D)] - L(f^*).$$

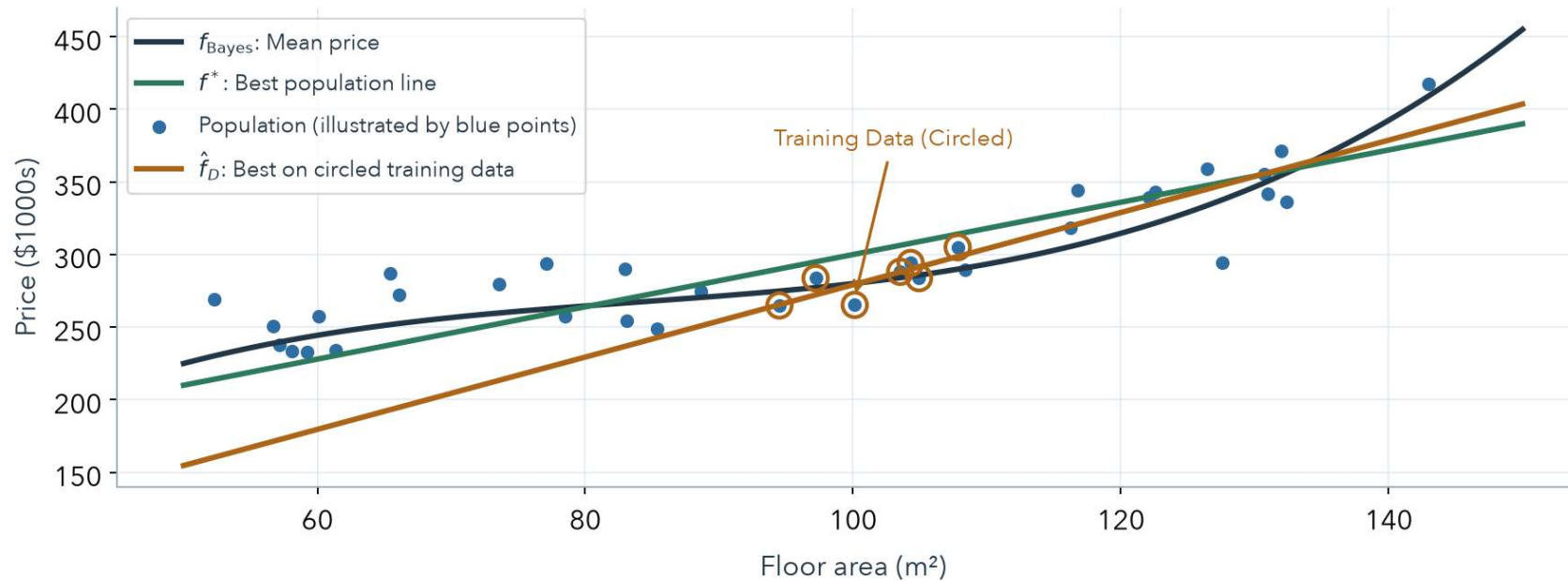
The learned predictor minimizes **empirical risk** within the same class:

$$\hat{f}_D \in \arg \min_{f \in \mathcal{F}} \hat{L}_D(f).$$

i We now have a finite training dataset, rather than the full population. Its best predictor can differ from f^* .

Example: A Fitted Line

All blue points illustrate the population. The circled middle points are the **training data**.



Empirical risk uses only the circled points. **Expected risk** averages over the whole population: training gives us only a partial view.

Why There Is Estimation Error

Training loss is an estimate of expected risk. Minimizing that estimate can select a different predictor from minimizing expected risk.

$$\hat{L}_D(\hat{f}_D) \leq \hat{L}_D(f^*) \quad \text{but} \quad L(f^*) \leq L(\hat{f}_D).$$

❓ What helps the training loss identify a good population predictor?

More training examples, or fewer candidate functions to distinguish.

Estimation Error and Training-Data Size

Keep the function class $\mathcal{F}$ fixed.

$$\text{EE} = \underbrace{\mathbb{E}_D[L(\hat{f}_D)]}_{\text{Learned from } n \text{ examples}} - \underbrace{L(f^*)}_{\text{Fixed best-in-class risk}} .$$

② What happens to estimation error if we increase the training-data size?

Estimation error typically decreases.

More houses give a more representative view of the population. Accidental patterns in a small sample have less influence on which predictor we choose.

Estimation Error and Function-Class Size

Keep the training-data size fixed.

$$\text{EE} = \mathbb{E}_D[L(\hat{f}_D)] - L(f^*).$$

❓ What happens to estimation error if we enlarge the function class $\mathcal{F}$?

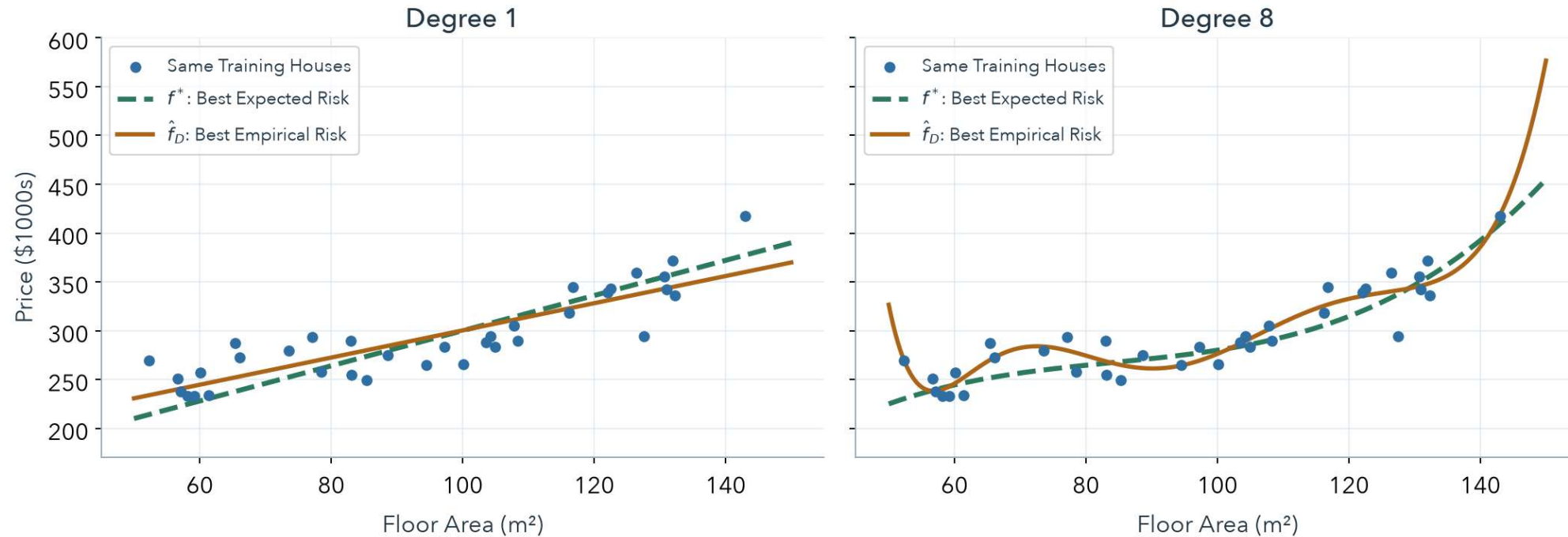
Estimation error often increases—but not always.

f^* minimizes expected risk, so $L(f^*)$ cannot increase. But $\hat{f}_D$ minimizes empirical risk. EE grows if:

- Its expected risk **rises**, for example by fitting sample-specific noise.
- Its expected risk falls, but **less than** $L(f^*)$ **falls**.

Example: Fitting with a Richer Class

Fit the **same training houses** with degree 1 and degree 8 polynomials.



The richer class improves f^* , but its fitted predictor can follow sample noise and move away from f^* .

Approximation-Estimation Tradeoff

With training-data size fixed, enlarging $\mathcal{F}$ has a benefit and a cost.

Approximation error

More candidates in $\mathcal{F}$ can represent the population pattern better.

Can decrease: the best predictor in the class gets closer to the Bayes risk.

Estimation error

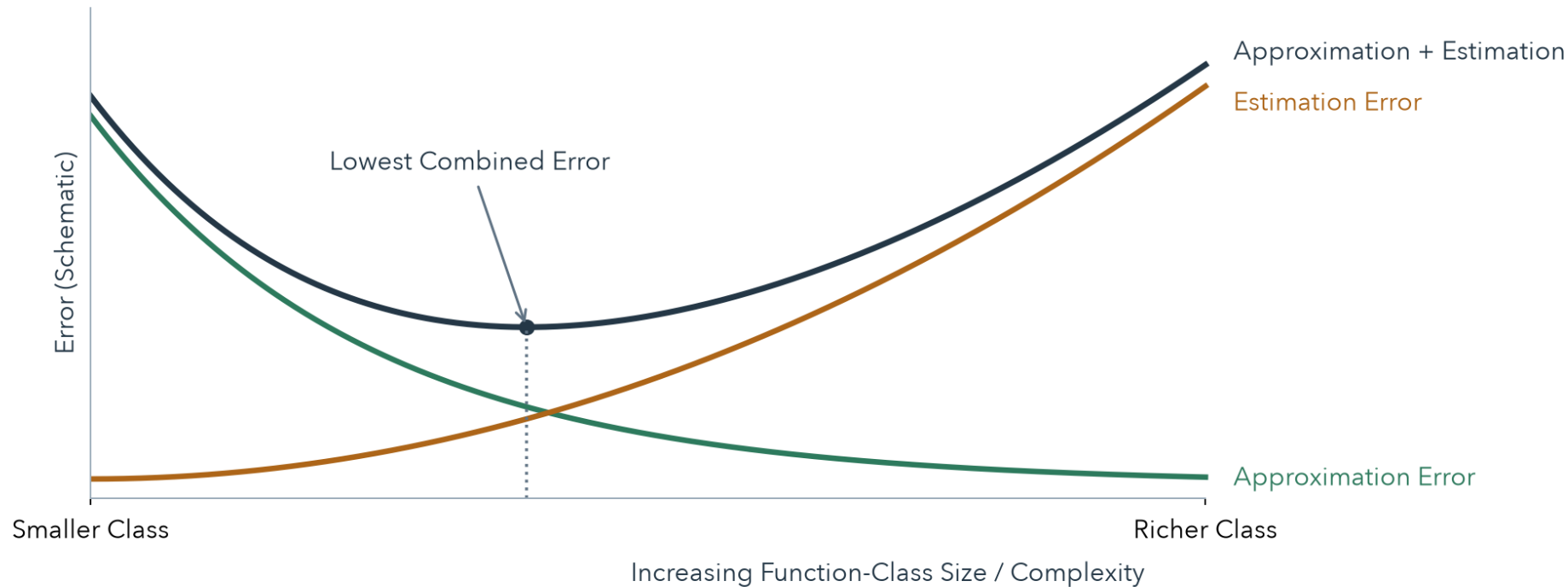
More candidates in $\mathcal{F}$ can also fit accidental patterns in the sample.

Often increases: selecting a predictor close to the best in the class becomes harder.

$$\mathbb{E}_D[L(\hat{f}_D)] = \underbrace{\text{EE} + \text{AE}}_{\text{Balance when choosing } \mathcal{F}} + \underbrace{L(f_{\text{Bayes}})}_{\text{Unchanged}}.$$

Visualizing the Tradeoff

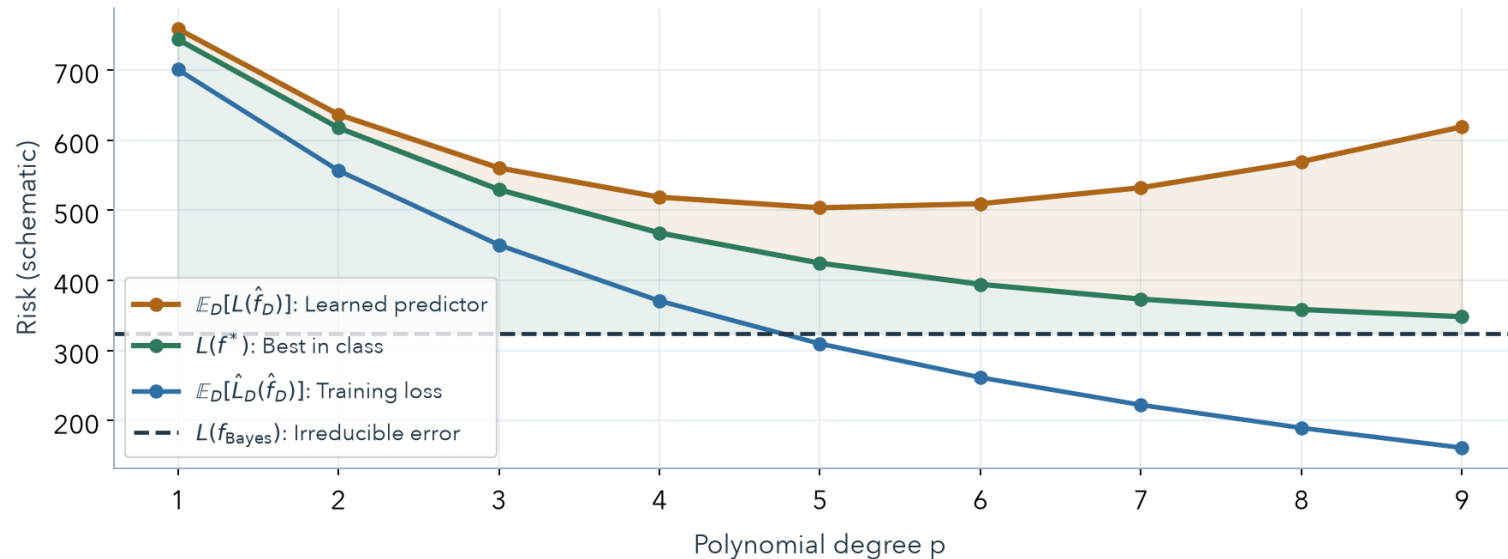
Keep the training-data size fixed while increasing function-class complexity.



Irreducible error stays fixed, so minimizing **approximation + estimation error** also minimizes expected risk.

Error and Polynomial Degree

Fixed training-data size; increasing polynomial degree (schematic).



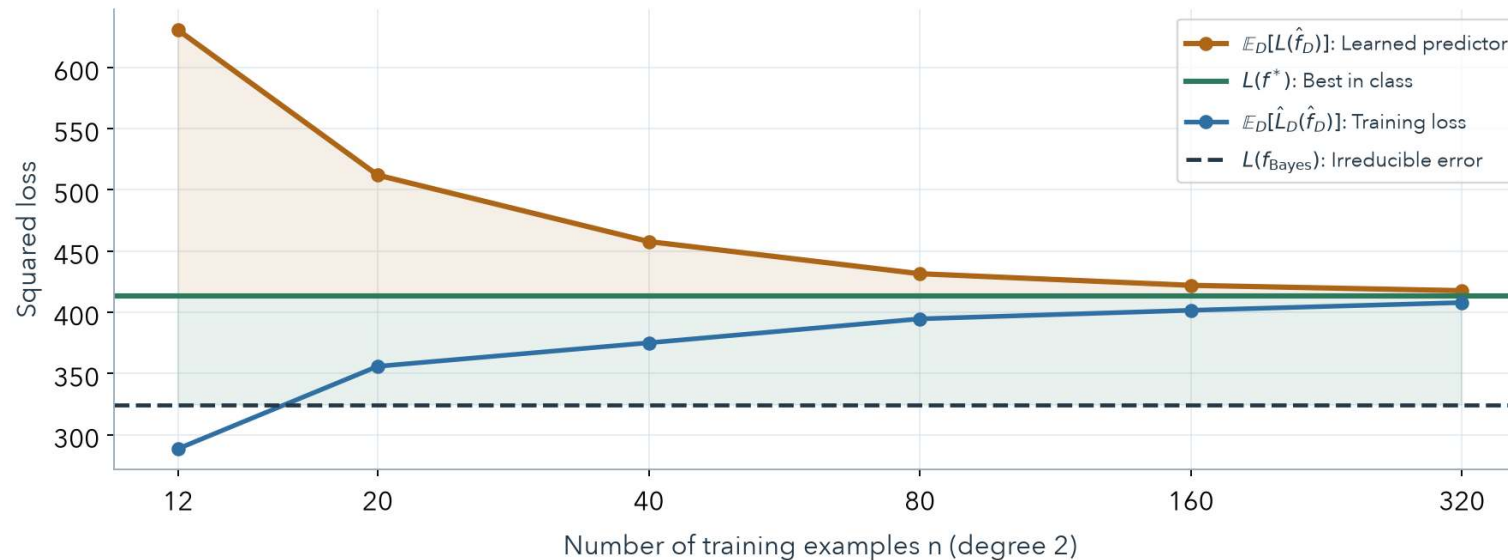
$$\underbrace{\mathbb{E}\mathbb{E} = \mathbb{E}_D[L(\hat{f}_D)] - L(f^*)}_{\text{Estimation: orange minus green}}$$

$$\underbrace{\text{AE} = L(f^*) - L(f_{\text{Bayes}})}_{\text{Approximation: green minus dashed}}.$$

A richer class shrinks the approximation gap, but can widen the estimation gap.

Error and Training-Data Size

Keep **degree** fixed and increase the number of training examples.



$$\underbrace{\mathbb{E}E = \mathbb{E}_D[L(\hat{f}_D)] - L(f^*)}_{\text{Estimation: orange minus green}}$$

$$\underbrace{AE = L(f^*) - L(f_{\text{Bayes}})}_{\text{Approximation: green minus dashed}}$$

More training data shrinks the estimation gap; the approximation gap stays fixed.

Summary: Class Size and Data Size

Keep the population, inputs, and loss fixed. Predict each effect before revealing the answer.

Error component	Higher polynomial degree	More training examples
Irreducible error	Unchanged	Unchanged
Approximation error	Cannot increase	Unchanged
Estimation error	Often increases	Usually decreases



Choosing a class balances approximation and estimation error. More data can make a richer class useful.

Interpreting Underfitting and Overfitting

Expected risk = Estimation error + Approximation error + Irreducible error.

Underfitting

Large approximation error: even the best predictor in the class misses important structure.

Overfitting

Large estimation error: the learned predictor fits the sample but is far from the best population predictor in its class.

 Both errors can be substantial. These are useful diagnoses, not mutually exclusive laws.

Check: What Would You Change?

Situation	First idea to investigate
Even the best predictor in $\mathcal{F}$ misses important patterns	Richer $\mathcal{F} \rightarrow$ approximation error
Fitted predictors change substantially across training samples	More training data? $\rightarrow$ estimation error
Houses with the same recorded features differ because of location	Include location in $\mathbf{X} \rightarrow$ irreducible error may decrease

Check: Can More Data Remove Every Error?

Keep $\mathcal{F}$, the input features, and the population fixed.

- ② We collect much more training data. Must the learned predictor's expected risk approach the irreducible error?

No. Approximation error can remain.

$$\mathbb{E}_D[L(\hat{f}_D)] = \text{EE} + \underbrace{\text{AE} + L(f_{\text{Bayes}})}_{L(f^*): \text{unchanged by more data}} .$$

Even if estimation error approaches zero, the class may still be unable to represent the Bayes predictor.

Check: Does a Richer Class Always Help?

Keep the training-data size fixed and add candidates to $\mathcal{F}$.

② Approximation error decreases. Must the learned predictor's average expected risk also decrease?

No. Estimation error can increase by more.

Check: Training Loss and Irreducible Error

❓ Can a fitted predictor's training loss be lower than the irreducible error?

Yes. Empirical risk can satisfy $\hat{L}_D(\hat{f}_D) < L(f_{\text{Bayes}})$.

❓ Does this contradict the definition of irreducible error?

No. Irreducible error is the minimum **expected risk**, not the minimum training loss.

A predictor can fit the observed sample—even perfectly—while still making errors on new examples from the population.

Wrap-up

Expected risk = Estimation error + Approximation error + Irreducible error.

- Better features can reduce uncertainty in the label.
- A suitable class represents the useful pattern.
- More data can improve the fitted predictor within that class.
- A richer class trades possible approximation gains against estimation difficulty.

 Next lecture: bias, variance, and regularization as a way to control the fitted model.

Optional: Bias-Variance Preview and Estimation Bounds

Preview: Bias and Variance

Across repeated training datasets, define the average predictor:

$$\bar{f}(\mathbf{x}) = \mathbb{E}_D[\hat{f}_D(\mathbf{x})].$$

Bias

How far is the **average prediction** from the best prediction?

$$\bar{f}(\mathbf{x}) - f_{\text{Bayes}}(\mathbf{x})$$

Example: consistently missing curvature in house prices.

Variance

How much does the **fitted prediction change** across training datasets?

$$\text{Var}_D[\hat{f}_D(\mathbf{x})]$$

Example: a fitted curve changes with the sampled houses.

Preview: Decomposing Squared Loss

For squared loss, the Bayes predictor is the conditional mean:

$$f_{\text{Bayes}}(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}].$$

$$\begin{aligned} \mathbb{E}_D[L(\hat{f}_D)] &= \underbrace{\mathbb{E}_{\mathbf{X}}[(\bar{f}(\mathbf{X}) - f_{\text{Bayes}}(\mathbf{X}))^2]}_{\text{Squared bias, averaged over inputs}} \\ &\quad + \underbrace{\mathbb{E}_{\mathbf{X}}[\text{Var}_D(\hat{f}_D(\mathbf{X}))]}_{\text{Variance, averaged over inputs}} + \underbrace{\mathbb{E}_{\mathbf{X}}[\text{Var}(Y \mid \mathbf{X})]}_{\text{Irreducible error}}. \end{aligned}$$

i Individual terms need not equal AE and EE, but their sums agree:
Squared bias + Variance = AE + EE.

Estimation Error: Learning from a Sample

With limited training data, the predictor with the lowest empirical risk may not have the lowest expected risk in the class.

$$\hat{f}_D \in \arg \min_{f \in \mathcal{F}} \hat{L}_D(f), \quad f^* \in \arg \min_{f \in \mathcal{F}} L(f). \\ \text{EE}(n, \mathcal{F}) = \mathbb{E}_D[L(\hat{f}_D)] - L(f^*).$$

For house prices, we fit the observed houses; f^* is the best predictor in the class over the whole distribution.

 How do the number of training examples n and the function class $\mathcal{F}$ affect this error?

Estimation Error and the Generalization Gap

$$\begin{aligned}\text{EE} &= \mathbb{E}_D[L(\hat{f}_D) - \hat{L}_D(\hat{f}_D)] \\ &\quad + \underbrace{\mathbb{E}_D[\hat{L}_D(\hat{f}_D) - \hat{L}_D(f^*)]}_{\leq 0 \text{ by ERM}} \\ &\quad + \underbrace{\mathbb{E}_D[\hat{L}_D(f^*)] - L(f^*)}_{=0: f^* \text{ is independent of } D} \\ &\leq \mathbb{E}_D[L(\hat{f}_D) - \hat{L}_D(\hat{f}_D)].\end{aligned}$$

- i Estimation error is bounded by the average gap between expected risk and empirical risk of the fitted predictor.

More Data: More Precise Empirical Risks

For a fixed predictor f , iid data, and finite loss variance:

$$\hat{L}_D(f) = \frac{1}{n} \sum_{i=1}^n \ell(f(\mathbf{X}_i), Y_i).$$

$$\mathbb{E}_D[\hat{L}_D(f)] = L(f), \quad \text{Var}_D[\hat{L}_D(f)] = \frac{\text{Var}[\ell(f(\mathbf{X}), Y)]}{n}.$$

More houses give a more precise estimate of a fixed predictor's expected risk.

i But $\hat{f}_D$ is selected using the data. To control its error, we need empirical risks to be accurate across the candidate class.

Controlling Error Across the Class

Let $\Delta_D(\mathcal{F})$ be the largest discrepancy among candidate predictors:

$$\begin{aligned}\Delta_D(\mathcal{F}) &= \sup_{f \in \mathcal{F}} |L(f) - \hat{L}_D(f)|. \\ L(\hat{f}_D) &\leq \hat{L}_D(\hat{f}_D) + \Delta_D(\mathcal{F}) \\ &\leq \hat{L}_D(f^*) + \Delta_D(\mathcal{F}) \quad (\text{ERM}) \\ &\leq L(f^*) + 2\Delta_D(\mathcal{F}).\end{aligned}$$

Subtract $L(f^*)$ and average over training datasets:

$$\boxed{\mathbb{E}E(n, \mathcal{F}) \leq 2\mathbb{E}_D[\Delta_D(\mathcal{F})].}$$

Training-Data Size and Estimation Error

Fix M candidate predictors before observing the iid training data. Assume the loss lies in $[0, 1]$.

$$\mathbb{E}_D[\Delta_D(\mathcal{F})] \leq \sqrt{\frac{\log(2M)}{2n}} \implies \mathbb{EE}(n, \mathcal{F}) \leq \sqrt{\frac{2 \log(2M)}{n}}.$$

With the class fixed, increasing n makes the bound smaller. Four times as much data halves this bound.

i Estimation error typically decreases with more training data. Here, convergence to zero is guaranteed; a decrease at every successive n is not.

Function-Class Size and Estimation Error

Keep n fixed and enlarge the class from $\mathcal{F}$ to $\mathcal{G}$:

$$\begin{aligned}\mathcal{F} \subseteq \mathcal{G} &\implies \Delta_D(\mathcal{F}) \leq \Delta_D(\mathcal{G}). \\ \mathbb{E}\mathbb{E}(n, \mathcal{F}) &\leq 2\mathbb{E}_D[\Delta_D(\mathcal{F})].\end{aligned}$$

More candidates give ERM more opportunities to select a lucky underestimate of expected risk. For house prices, extra polynomial terms can fit sample-specific fluctuations.

 A richer class typically increases estimation error. The bound above cannot decrease as the class grows; actual estimation error need not increase in every case.

Data Size and Function-Class Size

Change	Approximation error	Estimation error
More training data; fixed class	Unchanged	Typically decreases
Richer class; fixed data size	Cannot increase	Typically increases

$$\text{EE}(n, \mathcal{F}) \leq \sqrt{\frac{2 \log(2M)}{n}} \quad (\text{fixed } M \text{ candidates, loss in } [0, 1], \text{ iid data}).$$

 A richer class can represent more patterns, but choosing among more candidates requires more data for the same guarantee.