

CMPUT 267: Machine Learning I · Fall 2026

Evaluating Models (in Practice)

Xi Ye

2026-10-15

Choices in Learning a Predictor

We already have several choices that can change the predictor we learn.

Choice	What changes?
Polynomial degree	How flexible the predictor can be
Learning rate	The size of each optimization update
Batch size	How many examples contribute to each update
Number of training steps	How long we optimize

❓ Which choices will give good predictions on new examples?

Recall: The Two Steps of ERM

Empirical loss is the average loss on the training data. ERM uses it to choose a predictor.

Step 1: Estimation

$$L(f) \rightsquigarrow \hat{L}_D(f)$$

Use the dataset to estimate true loss with empirical loss.

Step 2: Optimization

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

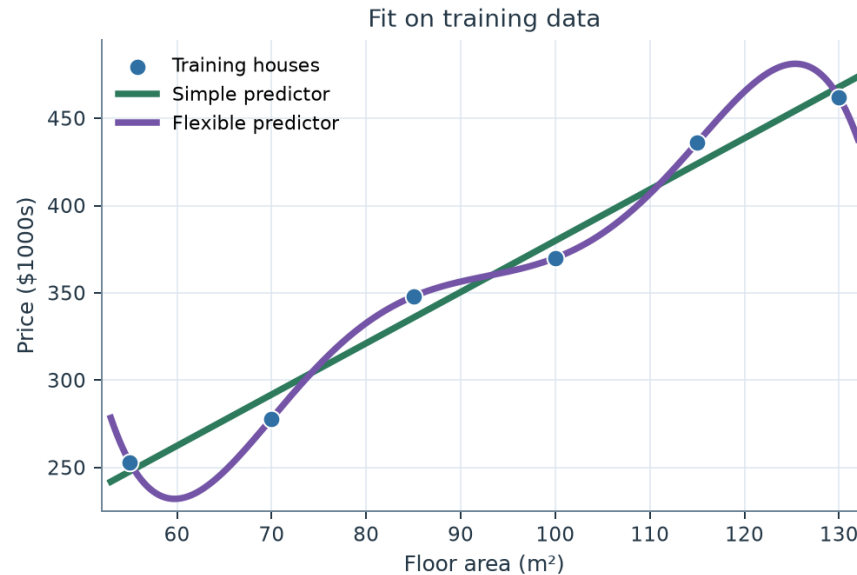
Choose the predictor with the lowest training loss.



We choose $\hat{f}_D$ **based on its training loss**. Using that same loss to evaluate it can be optimistic: the chosen predictor may fit noise in the training data.

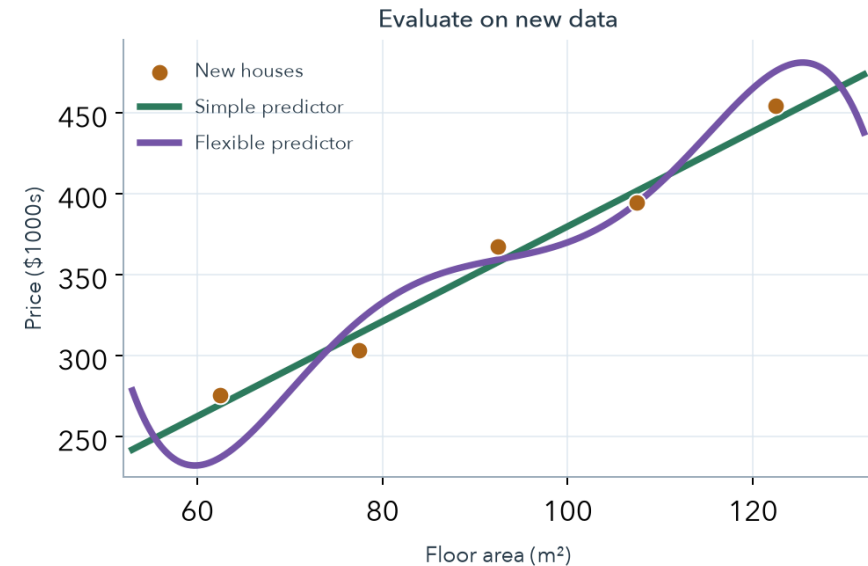
Training Loss vs New-Data Loss

Fit on training houses (blue).



Training loss (MSE): simple 107.5, flexible 0.0

Evaluate on new houses (orange).



Loss on new houses (MSE): simple 72.1, flexible 469.3



The lowest training loss does not necessarily identify the predictor that works best on new examples.

This Lecture: Evaluating and Selecting Predictors

We can fit many predictors, but training loss alone does not tell us which will work best on new examples.

① How should we **evaluate** a predictor and **select** among the available choices?

② This lecture gives a **practical** view first. Next two lectures will dive into theories.

1 · Model Complexity

How much freedom should the predictor have?

2 · Training-Data Size

How does more training data change loss?

3 · Training Steps

Does more optimization improve predictions?

4 · Evaluation in Practice

How do we compare models when true risk is unknown?

Recall: Empirical Risk Minimization

We want low **true loss**, but only observe a finite training dataset.

True Loss

Average loss over the population.

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

Also called **expected loss** or **risk**.

Empirical Loss

Average loss on the training dataset D .

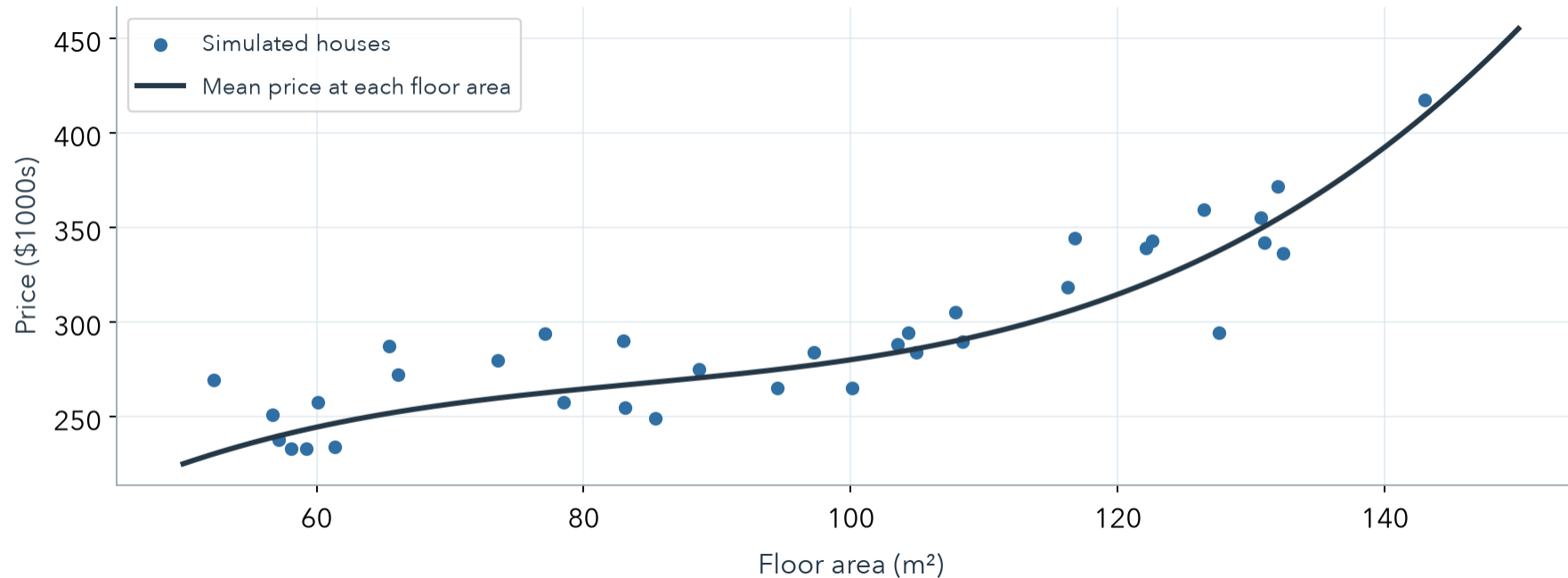
$$\hat{L}_D(f) = \frac{1}{n} \sum_{i=1}^n \ell(f(\mathbf{X}_i), Y_i).$$

Also called **training loss** or **empirical risk**.

i **ERM:** use empirical loss to estimate true loss, then minimize it over a chosen predictor class $\mathcal{F}$.

Example: House-Price Regression

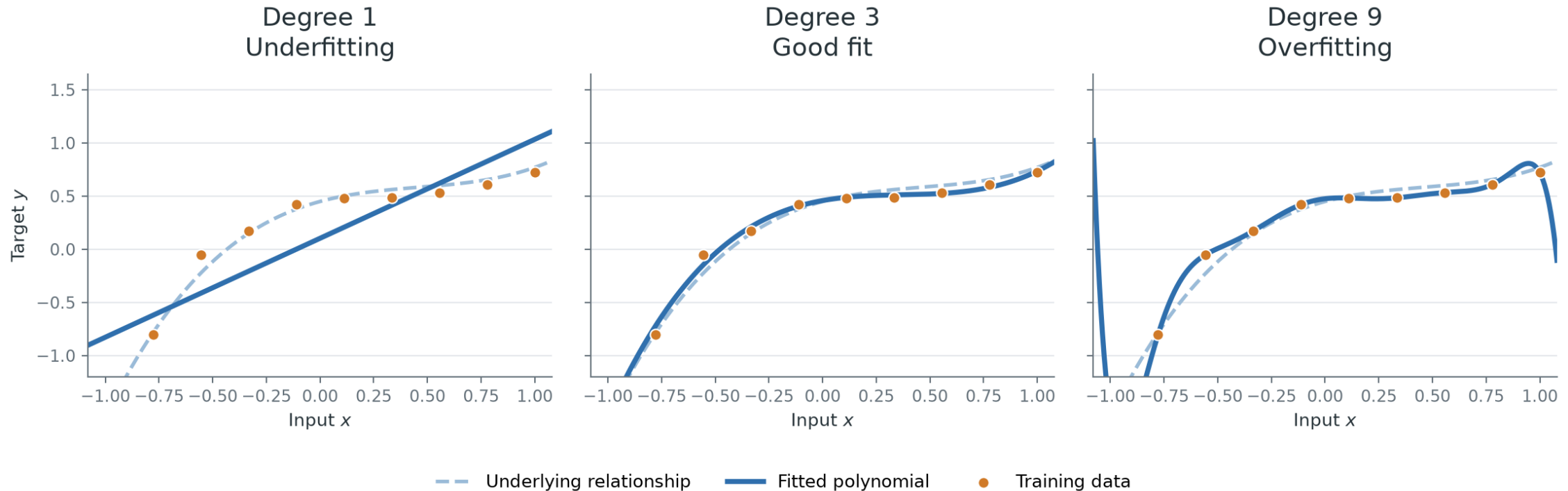
In this illustrative task, we predict house price Y (in \$1000s) using **only floor area** X (in m^2).



1 • Model Complexity

Polynomial Degree

Higher degree gives the polynomial more freedom to fit the data.



❓ Which degree should we choose to predict new examples well?

Polynomial Model Complexity

Let $\mathcal{F}_p$ contain all polynomials of degree **at most** p .

A degree- p predictor is also in $\mathcal{F}_{p+1}$: set its new coefficient to zero.

$$\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots \subseteq \mathcal{F}_p.$$

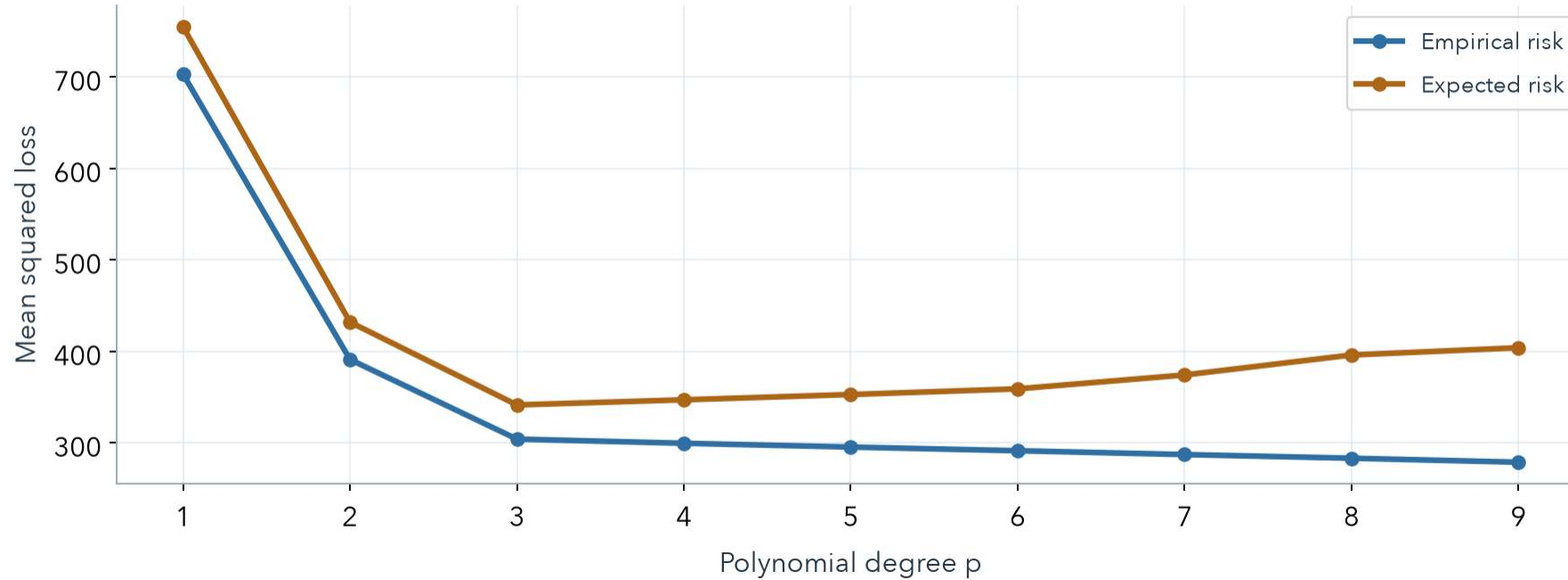
$$\min_{f \in \mathcal{F}_{p+1}} \hat{L}_D(f) \leq \min_{f \in \mathcal{F}_p} \hat{L}_D(f).$$



For the same dataset and exact minimization, the best training loss cannot increase as the class expands.

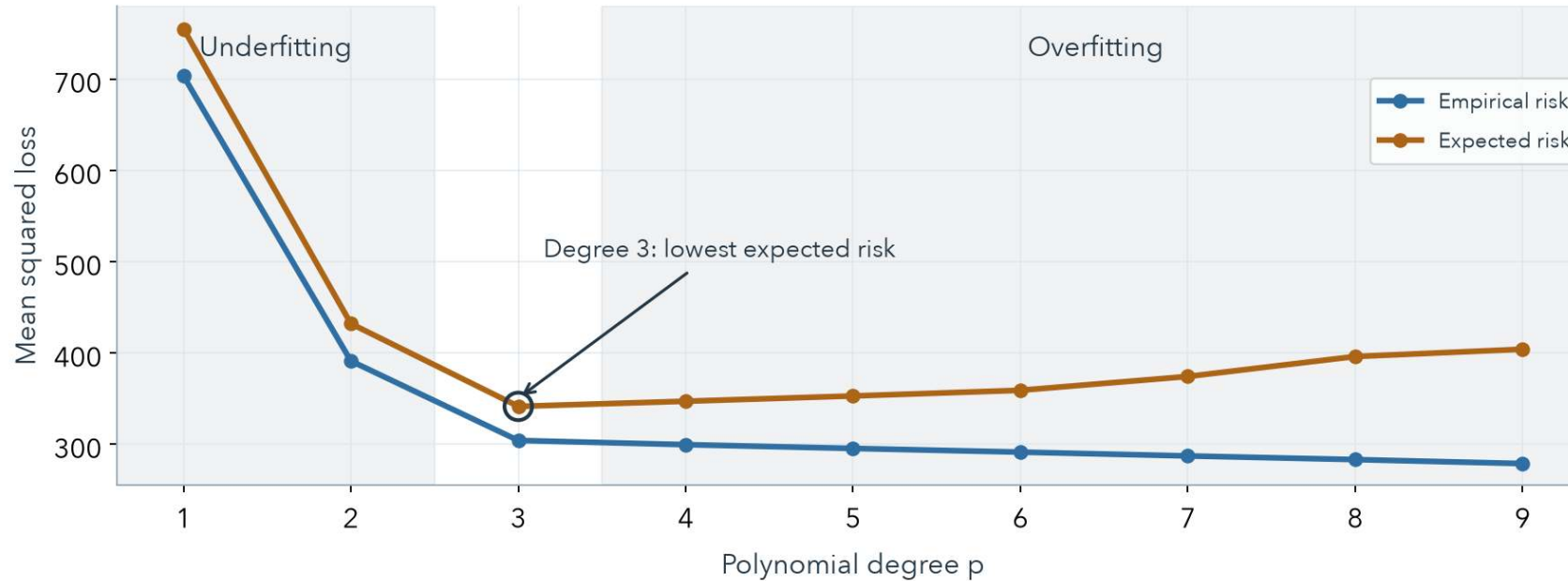
Loss vs. Model Complexity

For now, **assume true risk is known**. Keep the training-set size fixed.



i A lower training loss can accompany a higher true risk.

Underfitting and Overfitting



Degrees 1-2: too little flexibility to capture the curved relationship.

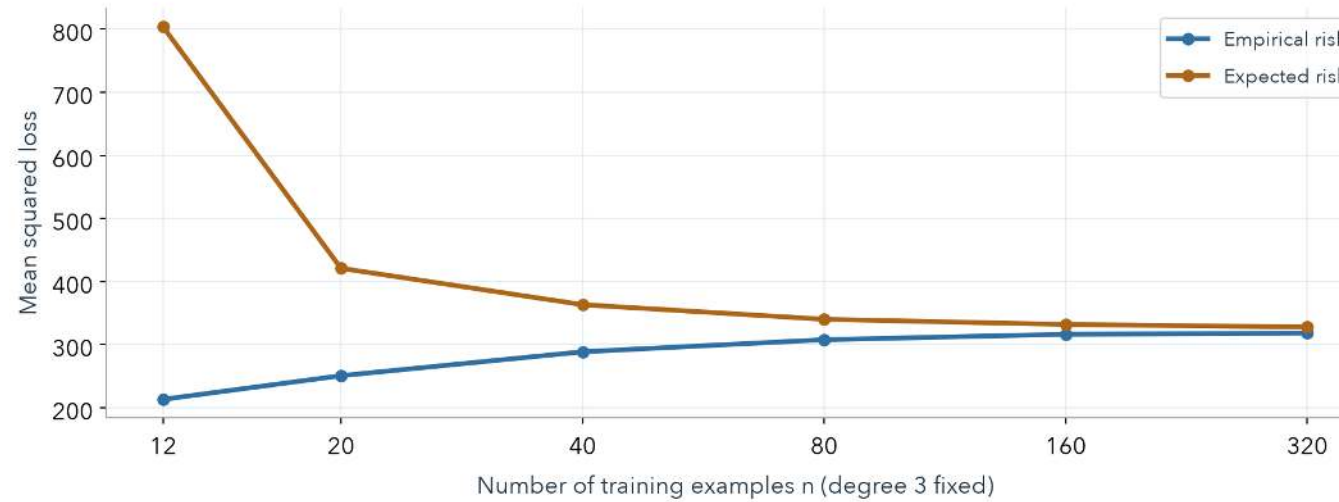
Degrees 4-9: fitting noise lowers training loss but raises true loss.

i Degree 3 is best here: it has the lowest true loss, averaged over training datasets.

2 · Training-Data Size

Loss vs. Dataset Size

Keep the polynomial degree fixed at 3 and vary the number of training examples.



Training loss can rise: a larger sample is harder to fit exactly, so the training score becomes less optimistic.

True risk can fall: more data makes the fitted predictor less sensitive to noise in a particular sample.



With fixed model complexity, more data **typically narrows the gap**. These are average trends, not guarantees for every dataset; neither loss must approach zero.

Check: More Data

Keep degree 3 fixed and increase the number of training examples. Which statements must be true?

1. The optimized training loss decreases.
2. The true risk becomes zero.
3. The training-true-risk gap shrinks for every new sample.

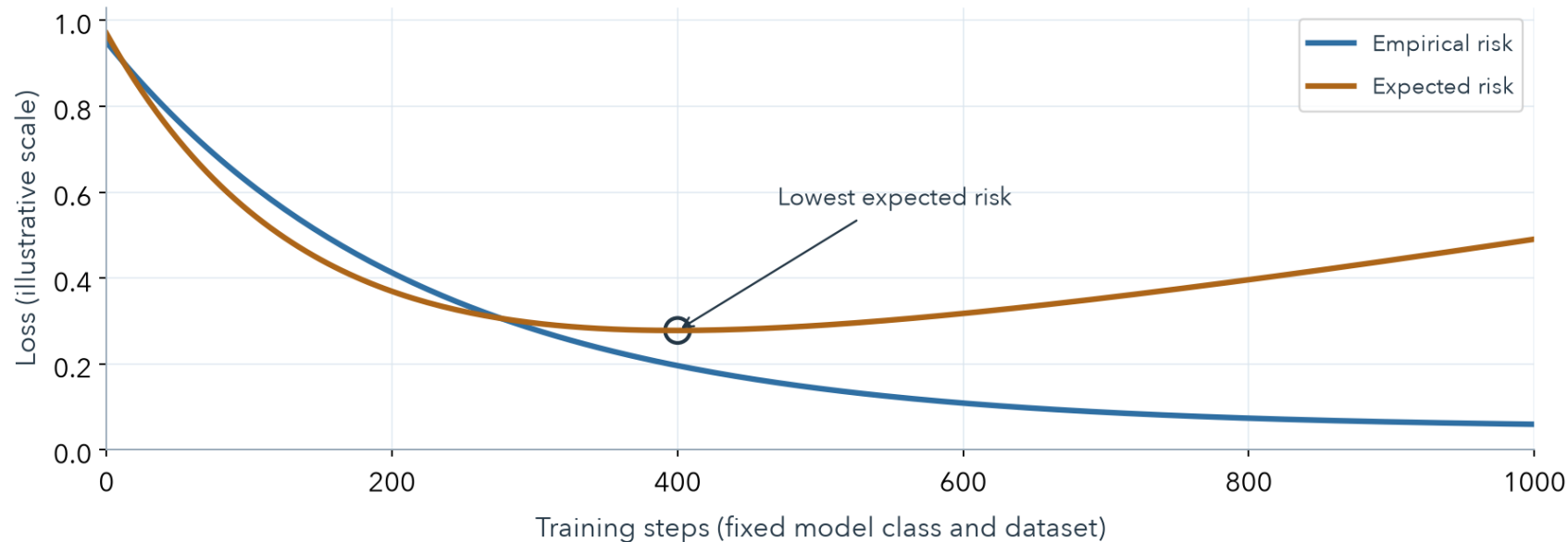


None is guaranteed. Training loss can rise, unpredictable variation can keep true risk positive, and an individual sample can depart from the average trend.

3 · Training Steps

Loss vs. Training Steps: Overfitting

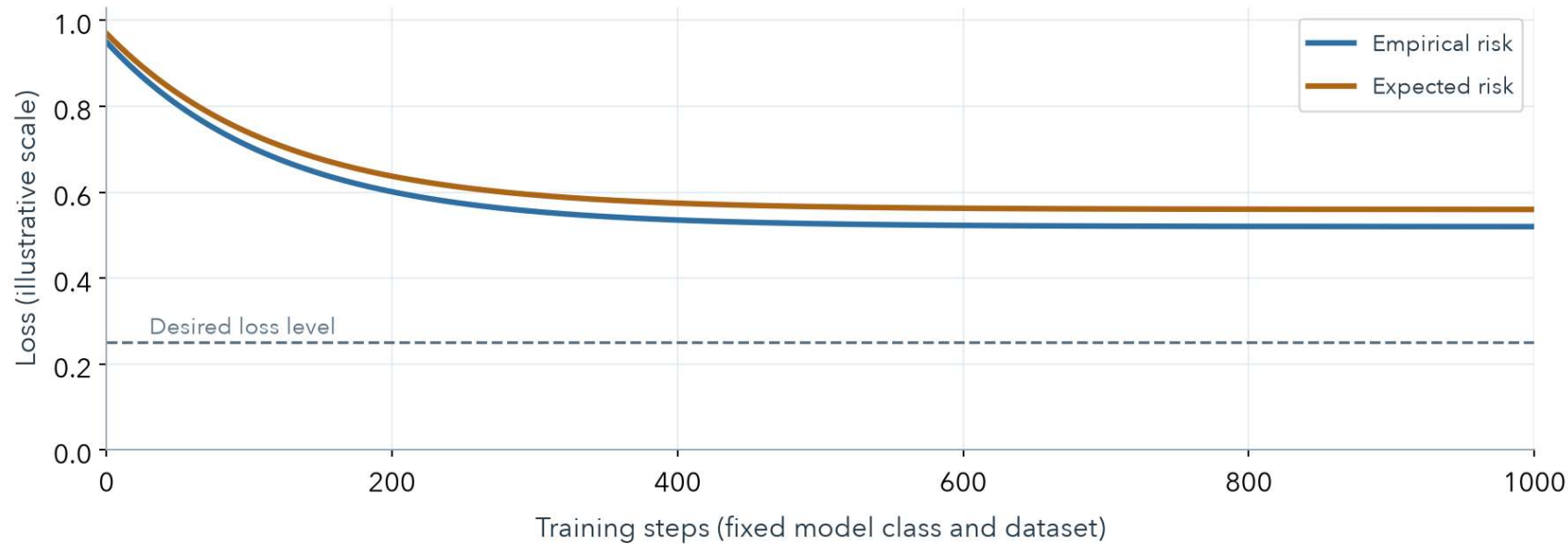
Keep the model class and dataset fixed; **assume we can see true risk** as training proceeds.



i Later updates lower training loss but raise true risk. Stopping earlier can give better predictions.

Loss vs. Training Steps: Underfitting

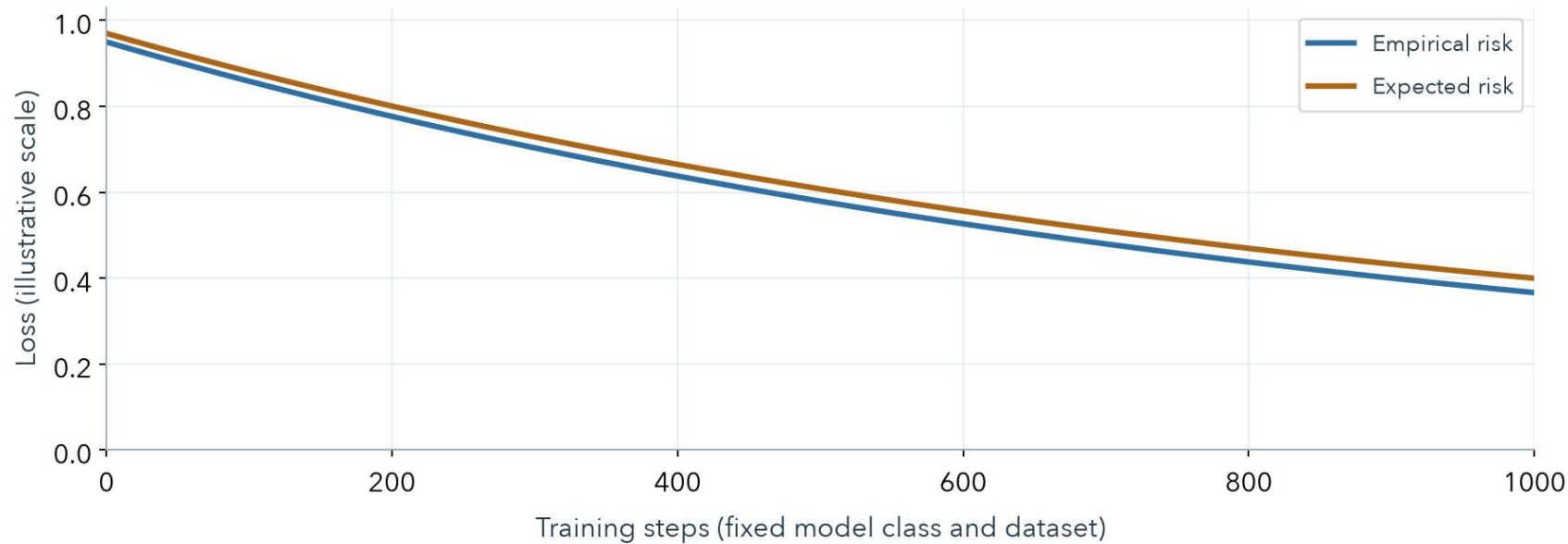
Both losses plateau above the desired level, even after optimization has settled.



More updates may not help. Investigate better features or a more flexible model class.

Loss vs. Training Steps: Optimization

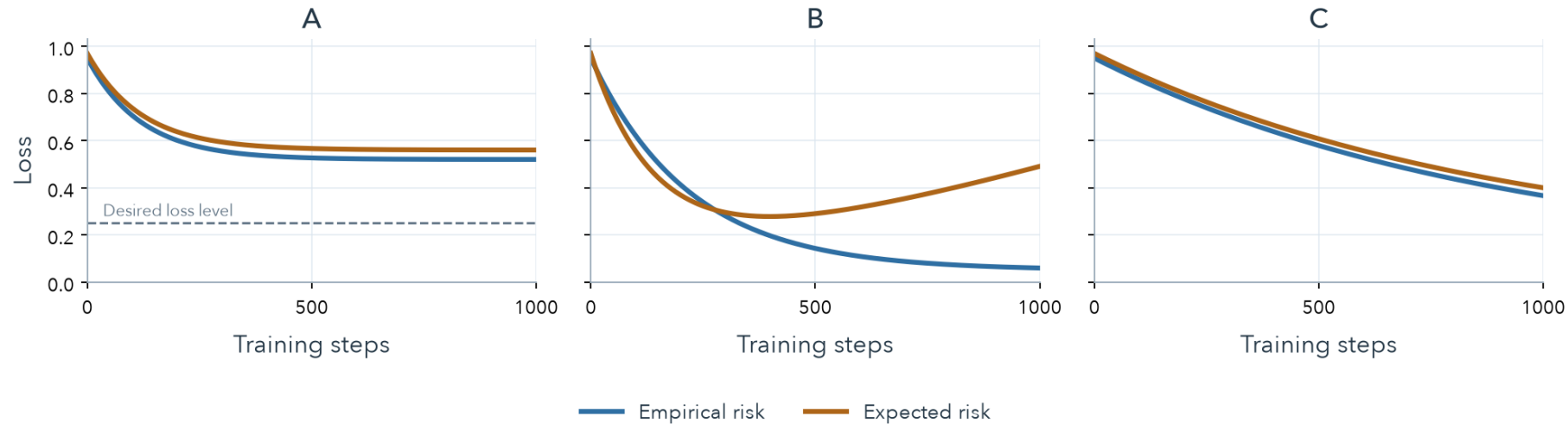
Training loss and true risk are still falling when the current training run ends.



i Optimization may be incomplete. Try more updates and check the step size; monitor whether true risk keeps improving.

Check: Reading Training Curves

For each curve, what would you try next (e.g., more training steps)?



A: Try a more complex model. Both losses plateau high; extra flexibility may capture the missing pattern.

B: Try a less complex model—or stop earlier. Further fitting lowers training loss but raises true risk.

C: Try more training steps. Both losses are still falling; continue while monitoring performance.

4 · Evaluation in Practice

Fitting and Generalizing

We have shown that there could be discrepancy between empirical risk and expected risk (true risk that we cannot access).

How to estimate true risk in practice?



Use fresh new data

Why Training Loss Is Optimistic

For each training dataset D , ERM chooses $\hat{f}_D$: the predictor in $\mathcal{F}$ with the **lowest training loss** on D .

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

Repeat this process with new training datasets, refitting each time. On average:

$$\underbrace{\mathbb{E}_D[\hat{L}_D(\hat{f}_D)]}_{\text{loss on the data used to choose the predictor}} \leq \underbrace{\mathbb{E}_D[L(\hat{f}_D)]}_{\text{true loss of that same predictor}}.$$

 Selection can reward fitting training noise.

Why Training Loss Is Optimistic: Proof

Let f^* minimize **true loss** in $\mathcal{F}$, independently of D .

1. ERM minimizes training loss.

$$\begin{aligned}\hat{L}_D(\hat{f}_D) &\leq \hat{L}_D(f^*) \\ \mathbb{E}_D[\hat{L}_D(\hat{f}_D)] &\leq L(f^*).\end{aligned}$$

For fixed f^* , **linearity** gives $\mathbb{E}_D[\hat{L}_D(f^*)] = L(f^*)$:
sample averages average to true loss.

2. f^* minimizes true loss.

$$\begin{aligned}L(f^*) &\leq L(\hat{f}_D) \\ L(f^*) &\leq \mathbb{E}_D[L(\hat{f}_D)].\end{aligned}$$

Every fitted predictor $\hat{f}_D$ belongs to $\mathcal{F}$, so it cannot
have lower true loss than f^* .

$$\boxed{\mathbb{E}_D[\hat{L}_D(\hat{f}_D)] \leq L(f^*) \leq \mathbb{E}_D[L(\hat{f}_D)].}$$

Assume iid training examples, a fixed class, exact minimization, and finite expected losses.

Training and Validation Sets

True risk is usually unknown. Split the available data to estimate it using examples that were not used for fitting.

 Use validation loss in place of the unknown true-risk curve to make model choices.

Training

Fit coefficients

Validation

Choose degree or stopping point

Parameters and Hyperparameters

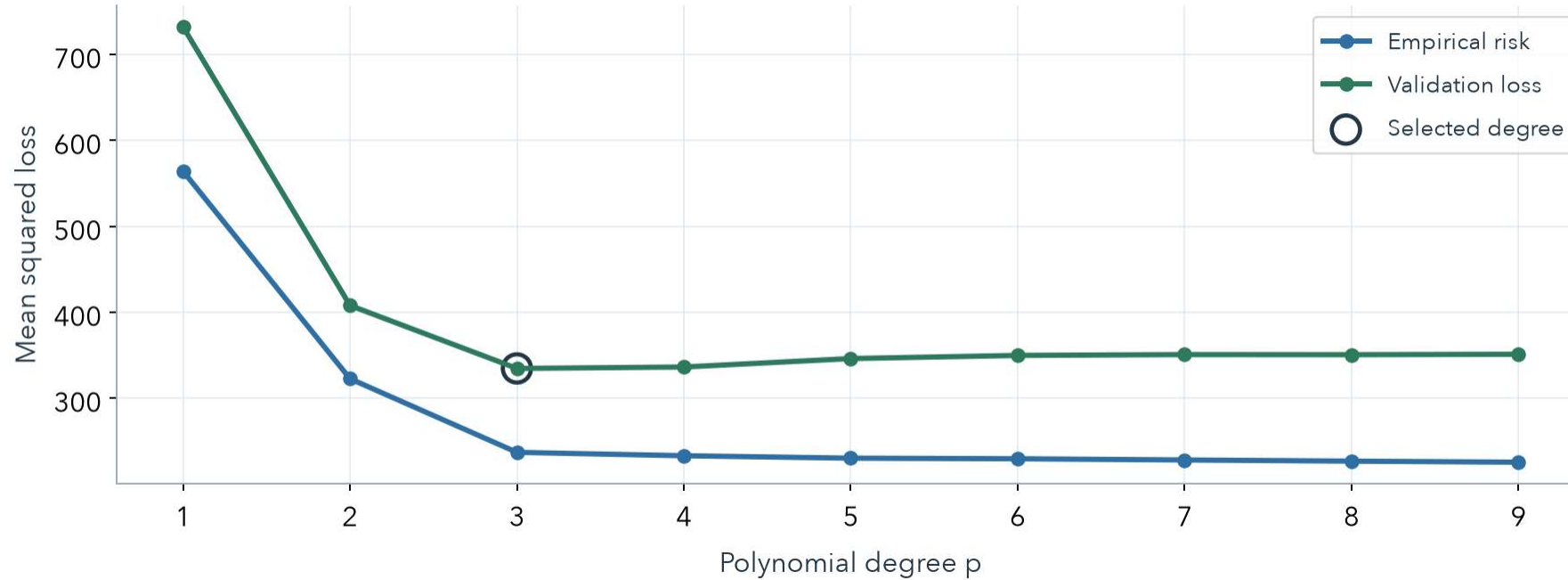
Parameters are fitted by the training procedure. **Hyperparameters** specify choices about that procedure or model.

Parameters	Hyperparameters
Intercept w_0	Polynomial degree p
Feature weights $w_1, \dots, w_d$	Learning rate η
Fitted polynomial coefficients	Number of epochs; batch size

 Changing a hyperparameter generally means fitting a new predictor.

Example: Choosing Polynomial Degree

Fit each degree on the **training houses**. Compare on **validation houses**.



Check: When to Stop Training?

Suppose one training run gives these losses:

Training step	Training loss	Validation loss
200	0.30	0.35
500	0.15	0.25
1000	0.05	0.40

Which saved predictor would you choose? Which data determined that choice?



Step 500, chosen using validation data. Its validation loss is lowest, even though training loss continues to fall.

Validation Loss Can Be Optimistic Too

We use validation loss to choose hyperparameters, such as polynomial degree or when to stop training.

 After choosing the lowest validation loss, is that same score a fair evaluation of the selected model?

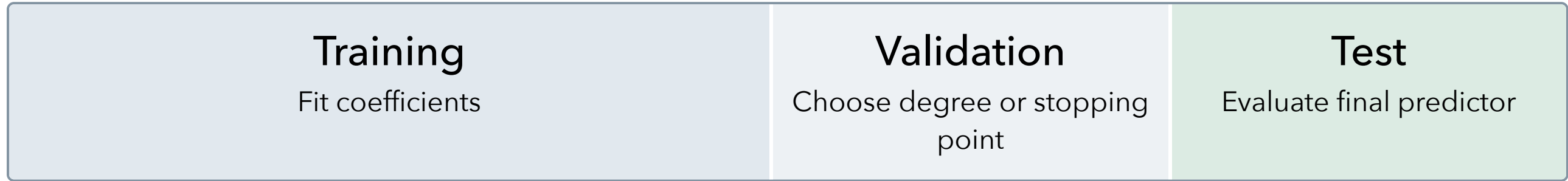
The choice depends on the validation data. A candidate can win partly because it fits accidental details of that particular validation set.

 We are selecting and evaluating with the same data again. The winning validation score can be optimistic.

Training, Validation, and Test Sets

Reserve a third portion, the **test set**, to evaluate the final choice.

All available data



Final evaluation uses data involved in **neither fitting parameters nor choosing models**.

Model-Selection Procedure

1. Split data into training, validation, and test subsets.
2. For each candidate degree p , fit $\hat{f}_p$ on the training set.
3. Choose $p^* = \arg \min_p \hat{L}_{\text{val}}(\hat{f}_p)$.
4. Freeze the chosen predictor and report its test loss.

 The test score estimates performance of this final predictor on new houses.

Information Leakage

Leakage occurs when information from held-out data influences training or model selection in a way that invalidates the intended evaluation.

Procedure	Appropriate?
Fit coefficients using training houses	Yes
Choose degree using validation loss	Yes
Repeatedly change degree after checking test loss	No
Estimate preprocessing statistics from all three sets	No

Check: Is This Still a Final Evaluation?

A student chooses degree 3 using validation loss and evaluates it on the test set. After seeing the test result, they try degrees 4 and 5 and report whichever has the lowest test loss.

❓ Is the reported score still an independent final evaluation? Why?

i **No.** The test data now influence model selection. Final evaluation needs fresh data that played no role in those choices.

Check: Which Dataset?

Decision or calculation	Which subset?
Estimate the slope of a fitted line	?
Choose degree 3 instead of degree 8	?
Report the final model's predictive loss	?

Training; validation; test.

Wrap-up

- Training loss can be optimistic because the predictor was chosen using those data.
- Model complexity, training-data size, and training steps affect generalization.
- Both training and validation scores can be optimistic after selection.
- Evaluate the final choice on data used for neither fitting nor model selection.

 Next lecture: theoretical analysis of training loss.