

STA 35C: Statistical Data Science III

Lecture 12: The Bias-variance Tradeoff

Dogyoon Song

Fall 2026, UC Davis

Announcements

Office hours

- Instructor office hours today (Wed, Oct 21) are **canceled**; please post questions on Piazza

Midterm grading policy

- Midterm 1 exam and solutions are available on the course webpage
- Only the **higher of your two midterm scores** will count; the lower score will be dropped
- Prepare for Midterm 2: attend lectures and discussion regularly, attempt problems before reading solutions, and ask questions early

Coming up

- The lecture on Fri, Oct 23 will also be held on Zoom; see Canvas for the link
- Homework 4 will be released Fri, Oct 23 and is due Thu, Oct 29, 11:59 PM PT

Agenda

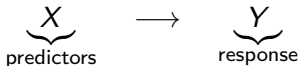
Brief recap of pre-Midterm 1 topics: core concepts and quick checks

Today: Bias-variance tradeoff

- Model accuracy: MSE and classification error
- Training error vs. test error
- Bias, variance, and irreducible error
- Why overly simple and overly flexible models might both fail

Roadmap ahead: estimating test error, controlling model flexibility, and multiple testing

Recap: Supervised learning



Goal: learn a rule or function $\hat{f}$ that uses X to explain or predict Y

- **Prediction:** use $\hat{f}(x_{\text{new}})$ to predict a future response
- **Inference:** understand how Y is related to the predictors

Two major problem types

- **Regression:** Y is quantitative
- **Classification:** Y is categorical

Key point: We fit models on training data, but want good performance on new test data

Quick check: Predict an exam score or pass/fail: which is regression?

Recap: Probability and random variables

Probability language

- Outcome ω , sample space Ω , event $A \subseteq \Omega$
- Conditional probability:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0$$

- Bayes' theorem updates probabilities after observing evidence (posterior)

Random variables

- A random variable is a function $X : \Omega \rightarrow \mathbb{R}$
- PMF/PDF/CDF describe the distribution of X

Expectation and variance summarize center and spread

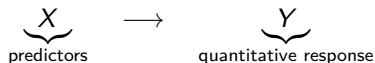
$$\mathbb{E}[X] = \sum_x x p_X(x) \quad \rightarrow \quad \mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$$

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] \quad \rightarrow \quad \text{Var}(aX + bY + c) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y)$$

Quick check: Let X count heads in two independent fair coin tosses. Find $P(X = 1)$ and $\mathbb{E}[X]$.

Recap: Regression

Regression setup



- **Target:** The mean quantitative response $\mathbb{E}[Y \mid X = x]$.

Linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \quad \mathbb{E}[\varepsilon \mid X] = 0$$

- Coefficients are estimated by least squares

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- β_j is interpreted as the average change in the mean response per unit increase in X_j , holding other predictors fixed
- Standard errors, confidence intervals, and t -tests quantify uncertainty
- R^2 , adjusted R^2 , and RSE summarize model fit

Quick check: If $\hat{y} = 4 + 2x$, what is predicted at $x = 3$? Must $Y = 10$?

Recap: Classification

Classification setup



- **Target:** A class label, often obtained from estimated class probabilities.

Logistic regression (discriminative)

$$\log \left(\frac{p(x)}{1 - p(x)} \right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \quad p(x) = P(Y = 1 \mid X = x)$$

- Predict class 1 if $\hat{p}(x) \geq p^*$
- Changing p^* changes the false positive / false negative tradeoff

Linear discriminant analysis (generative)

$$P(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_j \pi_j f_j(x)}$$

- LDA assumes Gaussian class-conditional distributions with common covariance

Recap: Classification errors and threshold selection

Class 1 is positive. Consider these illustrative counts:

	$\hat{y} = 1$	$\hat{y} = 0$
$Y = 1$	$TP = 10$	$FN = 10$
$Y = 0$	$FP = 4$	$TN = 76$

$$\text{Error rate} = \frac{FP + FN}{TP + TN + FP + FN},$$

$$\text{FPR} = \frac{FP}{TN + FP} \quad (\text{actual negatives}),$$

$$\text{FNR} = \frac{FN}{TP + FN} \quad (\text{actual positives}).$$

Quick check: What are the three rates for this example?

Answer: Error rate = 14%, FPR = 5%, FNR = 50%.

Changing the threshold p^* :

- With the same fitted model and observations, lowering p^* cannot decrease FPR or increase FNR
- Choose p^* to balance false-positive and false-negative costs; 0.5 is not always appropriate

Model accuracy: Regression vs. classification

Regression: Mean squared error (MSE)

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{f}(x_i) \right)^2$$

- Smaller MSE means predictions are closer to observed responses
- Least squares minimizes training MSE, equivalently RSS

Classification: Error rate (=rate of misclassification)

$$\text{Error rate} = \frac{1}{n} \sum_{i=1}^n I(\hat{y}_i \neq y_i)$$

- Smaller error rate means fewer misclassified observations
- Other metrics may matter when false positives and false negatives have different costs

Training error vs. test error

Training error is computed on the same data used to fit the model:

$$\text{MSE}_{\text{train}} = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{f}(x_i) \right)^2$$

Test error measures performance on new, unseen data:

$$\text{MSE}_{\text{test}} = \frac{1}{n_{\text{test}}} \sum_{j=1}^{n_{\text{test}}} \left(y_j^{\text{test}} - \hat{f}(x_j^{\text{test}}) \right)^2$$

Key distinction

- Training error measures how well the model fits the data it saw
- Test error measures how well the model generalizes to new data
- We care most about test error, but it is usually unknown when choosing a model

The challenge in practice

Reality: We usually do not know test error when choosing a model before deployment

- We fit models using training data
- But the model with the smallest training error may not have the smallest test error

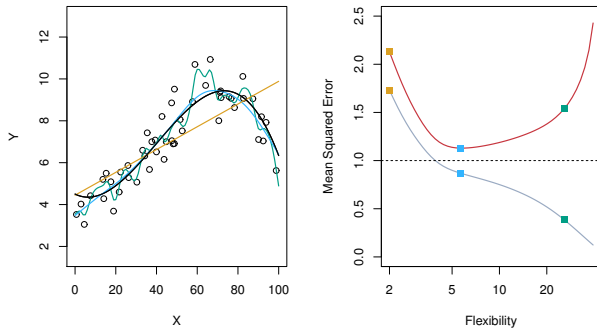


Figure: As model flexibility grows, training error usually decreases, while test error can eventually increase; Linear, Degree-5, Degree-25 regression [JWHT21, Figure 2.9].

Bias-variance tradeoff

At a fixed test point x_0 , $\hat{f}(x_0)$ varies across training samples. A new response satisfies

$$Y_0 = f(x_0) + \varepsilon_0, \quad \mathbb{E}[\varepsilon_0] = 0, \quad \text{Var}(\varepsilon_0) = \sigma^2,$$

where ε_0 is independent of the training sample.

The expected test error decomposes as

$$\mathbb{E} \left[\left(Y_0 - \hat{f}(x_0) \right)^2 \right] = \underbrace{\left(f(x_0) - \mathbb{E}[\hat{f}(x_0)] \right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E} \left[\left(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)] \right)^2 \right]}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Irreducible error}}$$

where the expectation is over both the training sample used to fit $\hat{f}$ and the new noise ε_0 .

Interpretation

- **Bias:** difference between $\mathbb{E}[\hat{f}(x_0)]$ and $f(x_0)$; systematic error from the model in use
- **Variance:** sensitivity of $\hat{f}$ to the particular training sample
- **Irreducible error:** new-response noise not predictable from the available predictors

Bias and variance: A numerical example

Example

At fixed x_0 , suppose $f(x_0) = 10$ and independent mean-zero noise has variance 1.

Across training samples, each model gives either listed prediction with probability 1/2:

Model	Possible $\hat{f}(x_0)$	$\mathbb{E}[\hat{f}(x_0)]$
A	8 or 10	9
B	6 or 14	10

Quick check: Model B is unbiased. Does it have lower expected squared test error?

Model	Bias ²	Variance	Expected squared test error
A	$(9 - 10)^2 = 1$	1	$1 + 1 + 1 = 3$
B	$(10 - 10)^2 = 0$	16	$0 + 16 + 1 = 17$

Takeaway: Unbiased does not mean accurate in every sample; variance also matters.

Interpreting bias and variance

High bias

- Can arise when the model is too simple to capture the true relationship
- Example: fitting a line when the true relationship is strongly curved
- Leads to underfitting: the model misses real structure in the X - Y relationship

High variance

- Can arise when the model is too sensitive to small changes in the training data
- Example: fitting an overly flexible curve through noisy observations
- Leads to overfitting: the model fits noise rather than signal

Goal: Find the sweet spot

- Choose enough flexibility to reduce bias
- But not so much flexibility that variance becomes too large

Model flexibility and test error

As model flexibility increases:

- Training error usually decreases
- Test error may decrease at first, then increase
 - Bias tends to decrease
 - Variance tends to increase

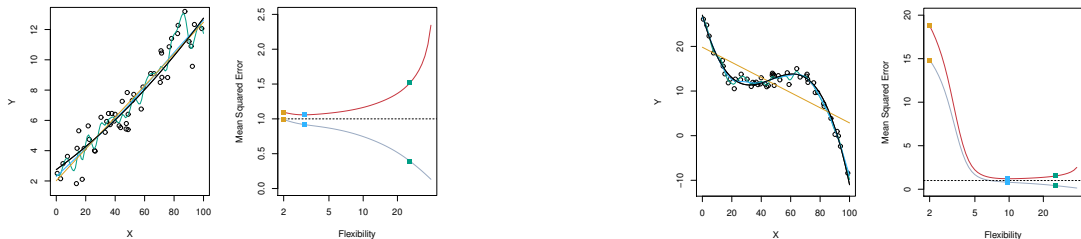


Figure: The same bias-variance idea appears across many settings [JWHT21, Figures 2.10 & 2.11].

$$\text{Expected squared test error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible error}$$

Pop-up quiz: Training vs. test error

Three regression models are fit to the same training data and evaluated on a separate test set:

Model	Flexibility	Training MSE	Test MSE
A	Low	30	32
B	Moderate	18	20
C	High	6	28

Question: Which statement is most appropriate?

- A) Choose Model C because it has the smallest training MSE.
- B) Model A has the highest variance because it is the simplest model.
- C) Test MSE always decreases as flexibility increases.
- D) Model B has the smallest observed error on this test set.

Answer: D. Model C fits the training data best but likely overfits; Model B gives the best performance on new data among the three.

Summary of the bias-variance tradeoff

Overly simple models might underfit the data

- High bias
- Low variance

Overly flexible models might overfit the training data

- Low bias
- High variance

Best predictive model often lies between these extremes

- Balances bias and variance to minimize test error
- Should be chosen using evidence about generalization, not training error alone

Looking ahead: Roadmap until Midterm 2

Main theme: Building models that fit signal, not noise, thereby generalizing well

The bias-variance tradeoff raises several practical questions.

1. **How can we estimate test performance using training data only?**
 - Cross-validation
2. **How can we quantify uncertainty when formulas are difficult?**
 - Bootstrap
3. **How can we allow flexibility but avoid overfitting?**
 - Model selection and regularization
4. **When many predictors or tests are considered, which findings are reliable?**
 - Multiple hypotheses testing and false discovery control

Unifying question: How do we learn useful patterns without mistaking noise for signal?

Wrap-up

- Before Midterm 1, we focused on probability, regression, classification, and interpreting fitted models.
- After Midterm 1, the central question becomes: *How well will a model perform on new data?*
- Training error measures fit to observed data; test error measures performance on new data.
- More flexibility can reduce bias, but too much flexibility can increase variance and cause overfitting.
- The bias-variance tradeoff explains why the model with the smallest training error is not necessarily the best predictive model.
- Upcoming lectures cover tools—cross-validation, bootstrap, regularization, and multiple-testing adjustments—that help us estimate performance, control overfitting, and quantify uncertainty.

Suggested reading: [JWHT21, Ch. 2.2]

Next lecture: Validation set approach and leave-one-out cross-validation

References



Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.

An Introduction to Statistical Learning: with Applications in R, volume 112 of *Springer Texts in Statistics*.

Springer, New York, NY, 2nd edition, 2021.