CSCI 145 Problem 9: Training Error Lies, in a Predictable Direction

July 25, 2026

Submission Instructions

Please upload your work to Gradescope by **11:59pm on DATE**.

- Turn in a single PDF of *handwritten* solutions. Typed (including LaTeX) solutions will not be accepted.
- Problems are grouped into three-week **cycles**. Once per cycle, you will meet with the grader and present the solution to *one* problem from that cycle, selected at random when you arrive — so bring your work and be ready to explain any of them.

Problem 9: Training Error Lies, in a Predictable Direction

In the class demo, the winning polynomial scored a mean squared error of 24.1 on the 15 days it was fit to and 32.8 on 60 fresh days. Neither number is a mistake, and the direction of the gap was predictable before either one was computed. Lecture quantified that gap exactly for a single fitted constant. This problem shows the same effect in general: first for the simplest quantity a model can estimate, the noise level itself, and then through the bias–variance decomposition that lecture stated without proof.

Throughout, the n training points are generated by

$$y^{(i)} = f^*(x^{(i)}) + \eta^{(i)}, \qquad \eta^{(i)} \sim \mathcal{N}(0, \sigma^2) \text{ independent},$$

exactly as in lecture, and $r^{(i)} = y^{(i)} - f(x^{(i)})$ is the residual of a hypothesis f on point i.

Part A: Even Estimating σ^2 Is Optimistic

Lecture treated the noise level σ^2 as known. Here we estimate it, using the same maximum likelihood recipe that produced squared error in the first place.

Holding the hypothesis f fixed, the Gaussian log-likelihood of the dataset is

$$\log L(\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(r^{(i)})^2.$$

Differentiate with respect to σ^2 (treat σ^2 itself as the variable, not σ), set the derivative to zero, and conclude that the maximum likelihood estimate of the noise level is the mean squared residual,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(r^{(i)})^2.$$

Now ask what that estimate is worth, in two situations. First, suppose someone hands us the *true* f^*, so that $r^{(i)} = \eta^{(i)}$ and the estimate is $\hat{\sigma}^2_{\text{true}} = \frac{1}{n}\sum_i(\eta^{(i)})^2$. Using $\mathbb{E}[(\eta^{(i)})^2] = \sigma^2$ and linearity of expectation, show that $\hat{\sigma}^2_{\text{true}}$ is unbiased. Second, suppose we have only the fitted $\hat{f}$ that empirical risk minimization returned on these same n points, giving $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_i(y^{(i)} - \hat{f}(x^{(i)}))^2$, which is exactly the training error. Assume the model class is rich enough to contain f^*, so f^* was one of the candidates $\hat{f}$ was chosen over. Explain in one or two sentences, with no computation, why this forces $\hat{\sigma}^2_{\text{MLE}} \le \hat{\sigma}^2_{\text{true}}$ on *every* dataset, and conclude that $\hat{\sigma}^2_{\text{MLE}}$ is biased low: fitting absorbs some of the noise into $\hat{f}$, leaving residuals that understate the true noise level.

Part B: The Decomposition

Fix a test point x with label $y = f^*(x) + \eta$, where $\eta \sim \mathcal{N}(0, \sigma^2)$ is independent of the training set. Let $\hat{f}$ be fit on a training set drawn at random from the data distribution, so $\hat{f}(x)$ is itself a random variable: it varies with which training set we happened to draw. Write $\bar{f}(x) = \mathbb{E}[\hat{f}(x)]$ for the average prediction over training sets. Starting from $\mathbb{E}[(y - \hat{f}(x))^2]$, add and subtract *both* $f^*(x)$ and $\bar{f}(x)$ inside the square:

$$y - \hat{f}(x) = \big(y - f^*(x)\big) + \big(f^*(x) - \bar{f}(x)\big) + \big(\bar{f}(x) - \hat{f}(x)\big).$$

Expand the square of this three-term sum and show that all three cross terms vanish in expectation. (In each one, either a factor is a constant multiplying something with expectation zero, or the two factors are independent and one has expectation zero, as in the Regression lecture's floor proof.) Conclude that

$$\mathbb{E}\big[(y - \hat{f}(x))^2\big] = \underbrace{\big(f^*(x) - \bar{f}(x)\big)^2}_{\text{bias}^2} + \underbrace{\mathrm{Var}\big(\hat{f}(x)\big)}_{\text{variance}} + \underbrace{\sigma^2}_{\text{noise}},$$

where the variance term is the $\mathbb{E}[(\bar{f}(x) - \hat{f}(x))^2]$ your expansion produced, by the definition of variance.

Part C: Reading the Curve

Use the decomposition to answer three questions, one or two sentences each. (i) A constant model predicts the same number regardless of x: which of the three terms is large, and which is exactly zero? (ii) A model that interpolates every training point (lecture's degree $= n_{\text{train}} - 1$ polynomial) does the opposite: which term is large now, and why does interpolating the labels make it so? (iii) Suppose we keep the model class fixed and double the number of training points. Which of the three terms shrink, which do not, and what does that say about the two curves in the demo?

Finally, state in one sentence the lesson of the decomposition: name the two distinct ways a model can be wrong, and say which one you make worse by adding capacity.

Where this goes next. This decomposition is the lens the rest of the course uses to talk about model capacity: every time a model gets bigger (more layers, more parameters, more heads), we are trading bias against variance. Near the end of the semester, the Generalization lecture pushes capacity far past the interpolation threshold from today's demo, where the variance term does something the classical U-curve does not predict.