CSCI 145 Problem 9: The Exact Price of Reusing the Data

August 27, 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: The Exact Price of Reusing the Data 2 points

Fix a full-column-rank design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$. Suppose the linear model is correct and the training labels are

$$\mathbf{y} = \mathbf{X}\mathbf{w}^\star + \boldsymbol{\eta}, \qquad \mathbb{E}[\boldsymbol{\eta}] = \mathbf{0}, \qquad \mathrm{Cov}(\boldsymbol{\eta}) = \sigma^2\mathbf{I}.$$

Least squares fits $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, where $\mathbf{H}$ is the projection matrix from Problem 7. To measure performance on the same feature vectors without reusing their noise, an independent replicate has labels $\mathbf{y}' = \mathbf{X}\mathbf{w}^\star + \boldsymbol{\eta}'$, with $\boldsymbol{\eta}'$ distributed like $\boldsymbol{\eta}$ and independent of it.

Quantify exactly how optimistic the training mean squared error is. Derive the expected training error $n^{-1}\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$ and the expected replicate error $n^{-1}\|\mathbf{y}' - \hat{\mathbf{y}}\|_2^2$ as functions of n, d, and σ^2. Use your formulas to audit a report claiming that a nearly zero training error near $d = n$ proves that the noise level is nearly zero. Your explanation should identify where the missing error went and connect the resulting gap to the bias–variance account in the reading.

Results you may use. The projection satisfies $\mathbf{H}^\top = \mathbf{H}$, $\mathbf{H}^2 = \mathbf{H}$, $\mathrm{tr}(\mathbf{H}) = d$, and $\mathbf{H}\mathbf{X} = \mathbf{X}$. If a mean-zero random vector has covariance $\sigma^2\mathbf{I}$, then

$$\mathbb{E}\|\mathbf{A}\boldsymbol{\eta}\|_2^2 = \sigma^2\,\mathrm{tr}(\mathbf{A}^\top\mathbf{A}).$$