CSCI 145 Problem 14: More Parameters, Less Error?

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 14: More Parameters, Less Error?

In the class demo, a model with $d = n = 100$ features had test error 94.6, while the same model handed 130 features had test error 9.0. Both fit the training data perfectly, so thirty surplus parameters made the model ten times more accurate on data it had never seen. Lecture blamed the spike on singular values shrinking toward zero at $d = n$; this problem makes that exact and gives the double-descent curve a formula.

Throughout, $\mathbf{X} \in \mathbb{R}^{n \times d}$ is a design matrix with $d > n$ and full row rank n, with singular value decomposition

$$\mathbf{X} = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top, \qquad \sigma_1 \geq \cdots \geq \sigma_n > 0,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_n \in \mathbb{R}^n$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^d$ are orthonormal. The labels are $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\eta}$ for a true weight vector $\mathbf{w}^* \in \mathbb{R}^d$ and a noise vector $\boldsymbol{\eta} \in \mathbb{R}^n$ with independent mean-zero entries of variance σ_η^2 (subscripted to keep it apart from the singular values). As in lecture, $\hat{\mathbf{w}} = \mathbf{X}^+\mathbf{y}$ is the minimum-norm fit.

Part A: One Fit Out of Infinitely Many

Lecture dropped a perpendicular to find the minimum-norm fit of one equation in two unknowns; here is the same picture in d dimensions. Starting from

$$\hat{\mathbf{w}} = \mathbf{X}^+\mathbf{y} = \sum_{i=1}^{n} \frac{1}{\sigma_i} \mathbf{v}_i \, (\mathbf{u}_i^\top \mathbf{y}),$$

substitute the SVD of $\mathbf{X}$, use orthonormality of the $\mathbf{v}_i$'s to collapse the double sum, and then use that $\mathbf{u}_1, \ldots, \mathbf{u}_n$ is an orthonormal basis of $\mathbb{R}^n$ to verify that $\mathbf{X}\hat{\mathbf{w}} = \mathbf{y}$. Next, show that $\hat{\mathbf{w}} + \mathbf{z}$ fits the data exactly as well, for every $\mathbf{z}$ in the null space $\{\mathbf{z} \in \mathbb{R}^d : \mathbf{X}\mathbf{z} = \mathbf{0}\}$, and say what the dimension of that null space is. Finally, argue that $\mathbf{X}\mathbf{z} = \mathbf{0}$ forces $\mathbf{v}_i^\top \mathbf{z} = 0$ for every i, and conclude that $\|\hat{\mathbf{w}} + \mathbf{z}\|_2^2 = \|\hat{\mathbf{w}}\|_2^2 + \|\mathbf{z}\|_2^2$, so $\hat{\mathbf{w}}$ is the smallest of all the exact fits.

Part B: Where the Small Singular Values Do the Damage

Substituting $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\eta}$ splits the fit into a signal term and a noise term:

$$\hat{\mathbf{w}} = \underbrace{\sum_{i=1}^{n} \mathbf{v}_i \mathbf{v}_i^\top \mathbf{w}^*}_{\text{signal, projected}} + \underbrace{\sum_{i=1}^{n} \frac{1}{\sigma_i} \mathbf{v}_i \, (\mathbf{u}_i^\top \boldsymbol{\eta})}_{\text{noise, amplified by } 1/\sigma_i}.$$

(You may take the signal term's simplification $\mathbf{X}^+\mathbf{X}\mathbf{w}^* = \sum_i \mathbf{v}_i \mathbf{v}_i^\top \mathbf{w}^*$ as given.) Using orthonormality of the $\mathbf{v}_i$'s and the fact that $\mathbf{u}_i^\top \boldsymbol{\eta}$ has mean zero and variance σ_η^2, show that the noise term's expected squared length is $\sigma_\eta^2 \sum_{i=1}^{n} 1/\sigma_i^2$. Then, in one sentence, say what the signal term misses: it is a projection of $\mathbf{w}^*$ onto a subspace of $\mathbb{R}^d$ of what dimension, and how does that dimension compare to d?

Part C: Why the Peak Sits Exactly at $d = n$

Part B measured the damage in singular values; now measure it in n and d. Write $\mathrm{tr}(\mathbf{M})$ for the sum of the diagonal entries of a square matrix $\mathbf{M}$, so that $\mathrm{tr}(\mathbf{u}\mathbf{u}^\top) = \|\mathbf{u}\|_2^2$ for any vector $\mathbf{u}$. Substitute the SVD into $\mathbf{X}\mathbf{X}^\top$, invert it term by term, and take the trace to show

$$\sum_{i=1}^{n} \frac{1}{\sigma_i^2} = \mathrm{tr}\big((\mathbf{X}\mathbf{X}^\top)^{-1}\big).$$

Now suppose, as in the demo, that the entries of $\mathbf{X}$ are independent standard Gaussians, and use without proof that $\mathbb{E}\big[(\mathbf{X}\mathbf{X}^\top)^{-1}\big] = \frac{1}{d-n-1}\mathbf{I}$ whenever $d > n+1$. Averaging over the draw of $\mathbf{X}$ as well as the noise, combine this with Part B to write the noise term's expected squared length as a function of n, d, and σ_η^2 alone. Read your formula back: what happens as d decreases toward $n+1$, and how fast does it fall once $d \gg n$? With your answer to Part B's last question, conclude in one sentence what the double-descent peak is.

Where this goes next. Your $\sum_i 1/\sigma_i^2$ is the condition number story from the Optimization lecture applied to *variance* rather than to precision: small singular values are where a design matrix hurts. The next lecture turns the same spectrum around and discards the smallest singular values instead of dividing by them.