CSCI 145 Problem 8: The Slowest Direction Sets the Pace

July 20, 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 8: The Slowest Direction Sets the Pace

In the demo, gradient descent spent 20,000 steps on a dataset with two features and still had one of the two weights sitting at 0.009 instead of 2.007; rescaling those same two features found the minimum in six steps. The number of parameters never changed, only the shape of the bowl. This problem derives the rate that explains the gap, starting from the simplest possible bowl.

Part A: One Dimension

Let $\mathcal{L}(w) = \frac{\lambda}{2} w^2$ for a fixed curvature $\lambda > 0$, so $\nabla \mathcal{L}(w) = \lambda w$ and the minimum sits at $w^\star = 0$. Write out one gradient descent step $w^{(t+1)} = w^{(t)} - \alpha \lambda w^{(t)}$ with learning rate $\alpha > 0$, and show by induction that

$$w^{(t)} = (1 - \alpha\lambda)^t \, w^{(0)}.$$

For what range of α (in terms of λ) does $w^{(t)} \to 0$ as $t \to \infty$? The number $|1 - \alpha\lambda|$ is the **contraction factor**, the fraction of the distance to the minimum that survives one step, and everything below is built out of it.

Part B: Every Direction at Once

Lecture showed that mean squared error, measured from its minimizer, is a sum of one-dimensional bowls, one per right singular vector of the design matrix:

$$\mathcal{L}(\mathbf{w}^\star + \mathbf{e}) - \mathcal{L}(\mathbf{w}^\star) = \frac{1}{2} \sum_{i=1}^{d} \lambda_i \, (\mathbf{v}_i^\top \mathbf{e})^2, \qquad \lambda_i = \frac{2\sigma_i^2}{n},$$

where $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^d$ are the orthonormal right singular vectors of $\mathbf{X} \in \mathbb{R}^{n \times d}$ and σ_i are its singular values. So with $\mathbf{A} = \sum_i \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$ and error $\mathbf{e}^{(t)} = \mathbf{w}^{(t)} - \mathbf{w}^\star$, gradient descent on $\mathcal{L}$ is gradient descent on $\frac{1}{2} \mathbf{e}^\top \mathbf{A} \mathbf{e}$.

Track the error one direction at a time: let $c_i^{(t)} = \mathbf{v}_i^\top \mathbf{e}^{(t)}$ be its component along $\mathbf{v}_i$. Using $\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i$ and orthonormality of the $\mathbf{v}_i$, show that gradient descent updates each component on its own, exactly as in Part A:

$$c_i^{(t+1)} = (1 - \alpha\lambda_i) \, c_i^{(t)}, \qquad \text{hence} \qquad c_i^{(t)} = (1 - \alpha\lambda_i)^t \, c_i^{(0)}.$$

There are d contraction factors and only one α to satisfy them all. Argue that the steepest direction sets the largest usable step size, $\alpha < 2/\lambda_{\max}$, and that at any such α the shallowest direction's factor $|1 - \alpha\lambda_{\min}|$ sits just below 1. That is lecture's zig-zag, in coordinates.

Part C: The Rate

Write $\kappa = \lambda_{\max}/\lambda_{\min}$ for the bowl's aspect ratio; lecture showed that for mean squared error this equals $\kappa(\mathbf{X})^2$, the squared condition number of the design matrix. The slowest surviving component is what we wait on, so the best step size is the one minimizing $\max_i |1 - \alpha\lambda_i|$. Substitute $\alpha^\star$ below into the two extreme contraction factors, check that they come out as negatives of each other, and argue in a sentence that no other α does better (raising α inflates one, lowering it inflates the other):

$$\alpha^\star = \frac{2}{\lambda_{\max} + \lambda_{\min}}, \qquad \text{at which} \qquad |1 - \alpha^\star \lambda_{\max}| = |1 - \alpha^\star \lambda_{\min}| = \frac{\kappa - 1}{\kappa + 1}.$$

Now put numbers on that rate. Evaluate $\frac{\kappa-1}{\kappa+1}$ at $\kappa = 1$, at $\kappa = 15$ (the stretched bowl in lecture's figure), and at $\kappa \approx 9.5 \times 10^5$ (the demo's raw features, where $\kappa(\mathbf{X}) \approx 977$), and in each case estimate how many steps it takes to shrink the error by a factor of 10. Compare the last number against the 20,000 steps the demo actually ran. Close with one sentence saying what really controls the number of iterations, and why a problem with many parameters is not automatically a slow one.

Where this goes next. The rate $\frac{\kappa-1}{\kappa+1}$ decides how long training takes on any loss that is quadratic near its minimum, and for mean squared error the curvatures $\lambda_i = 2\sigma_i^2/n$ come straight from the singular values of $\mathbf{X}$: the SVD from Unit 1 sets the training time. It is also why gradient descent gets a companion in the Gradient Descent lecture, momentum, which improves the dependence from κ to $\sqrt{\kappa}$ by remembering which way it has been heading instead of reacting to each gradient alone.