Can Units Cure Poor Conditioning?

Due: Wednesday, September 30 at 11:59pm

Instructions: Please submit your *handwritten* solutions as a single PDF on Gradescope.

A calibration model has two nonzero feature columns $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ with lengths $a, b > 0$ and cosine similarity

$$\gamma = \frac{\mathbf{x}_1^\top \mathbf{x}_2}{ab}, \qquad |\gamma| < 1.$$

An engineer may divide each column by any positive scale before minimizing MSE with fixed-step gradient descent. For each scaling, the learning rate is chosen to minimize the worst-direction contraction factor.

The scaling must preserve the set of attainable predictions.

1. The column scales determine the Hessian's condition number. Find the optimal positive diagonal scaling, with proof.
2. The features now use the optimal scaling from part 1. Determine the smallest worst-direction contraction factor.
3. Nearly redundant sensors have $|\gamma|$ close to one. Can changing units guarantee fast convergence in this case?

Hint: For a two-dimensional symmetric matrix, the eigenvalue sum is its trace and their product is its determinant.

Analyze q^2, where

$$q = \frac{\kappa - 1}{\kappa + 1}$$

is the best fixed-step contraction factor and κ is the Hessian condition number. Use the arithmetic–geometric mean (AM–GM) inequality.

Takeaway: *Rescaling the two features to have equal lengths gives the fastest convergence among these choices of units. It cannot fix features that are nearly parallel. Good scaling helps gradient descent, but redundant features can still make it slow.*