CSCI 145 Problem 7 Solution: Can Units Cure Poor Conditioning?

Solution

Divide the columns by their lengths a and b. This equalizes their lengths and gives the smallest possible contraction factor, $|\gamma|$. AM–GM shows why equal column lengths are optimal.

1. The Optimal Scaling

Positive rescaling preserves the cosine similarity γ. If the rescaled columns have lengths $u, v > 0$, the MSE Hessian is

$$\mathbf{H} = \frac{2}{n}\begin{bmatrix} u^2 & uv\gamma \\ uv\gamma & v^2 \end{bmatrix}.$$

Let $\lambda_{\max} \geq \lambda_{\min} > 0$ be the eigenvalues of $\mathbf{H}$. The contraction formula in the hint gives

$$q = \frac{\kappa - 1}{\kappa + 1} = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}, \qquad \kappa = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

Use $(A - B)^2 = (A + B)^2 - 4AB$, then identify the eigenvalue product and sum as the determinant and trace:

$$\begin{aligned} q^2 &= \frac{(\lambda_{\max} - \lambda_{\min})^2}{(\lambda_{\max} + \lambda_{\min})^2} = 1 - \frac{4\lambda_{\max}\lambda_{\min}}{(\lambda_{\max} + \lambda_{\min})^2} \\ &= 1 - \frac{4\det(\mathbf{H})}{\operatorname{tr}(\mathbf{H})^2} = 1 - \frac{4u^2v^2(1-\gamma^2)}{(u^2+v^2)^2}. \end{aligned}$$

The common factor $(2/n)^2$ cancels in the last step. By AM–GM, $u^2 + v^2 \geq 2uv$, with equality exactly when $u = v$. Since $1 - \gamma^2 > 0$, this gives

$$q^2 \geq 1 - (1 - \gamma^2) = \gamma^2,$$

with equality exactly when the rescaled columns have equal lengths. Thus the smallest contraction factor is $|\gamma|$. Dividing by a, b makes both lengths one. Any common further scaling is equally optimal.

2. The Remaining Contraction Factor

The proposed scaling preserves every prediction by rescaling the weights in the opposite way:

$$\frac{\mathbf{x}_1}{a}(aw_1) + \frac{\mathbf{x}_2}{b}(bw_2) = \mathbf{x}_1 w_1 + \mathbf{x}_2 w_2.$$

Since $a, b > 0$, this correspondence works in both directions.

For orthogonal columns, $\gamma = 0$ and the best contraction factor is zero. For nearly parallel or antiparallel columns, $|\gamma|$ is close to one, so even the best scaling gives slow worst-direction convergence. Changing units can equalize lengths, but it cannot change the angle between the features.