CSCI 145 Problem 18: How Close Is a Finite Muon Step?

October 27, 2026

Submission Instructions

Please upload your work to Gradescope by **11:59pm on Wednesday, November 18**.

- Turn in a single PDF of *handwritten* solutions. Typed (including LaTeX) solutions will not be accepted.
- A complete, good-faith attempt earns submission credit. The grader may select any recent assigned problem for a brief conversation, so keep your work and be prepared to explain each solution and answer a related question. Scheduling details will be announced in class.

Problem 18: How Close Is a Finite Muon Step?

Muon uses a few matrix multiplications to approach a matrix gradient's polar factor rather than computing an SVD at every optimizer step. Let a normalized nonzero matrix gradient of active rank $r \geq 2$ have compact SVD

$$\mathbf{X}^{(0)} = \mathbf{U}\,\mathrm{diag}(x_1, \ldots, x_r)\mathbf{V}^\top, \qquad 1 \geq x_1 \geq \cdots \geq x_r = \ell > 0.$$

Starting from $\mathbf{X}^{(0)}$, the accelerator performs only an integer $T \geq 0$ steps of the Newton–Schulz iteration from the reading. The ideal unit-spectral-norm descent direction would instead use the polar factor $\mathbf{U}\mathbf{V}^\top$ immediately.

1. In terms of $p^{\circ T}$, what is $\mathbf{X}^{(T)}$, and what ℓ-dependent guarantee follows for the fraction of the ideal polar step's first-order loss decrease achieved by $-\alpha\mathbf{X}^{(T)}$?
2. What is the spectral-norm distance between $\mathbf{X}^{(T)}$ and the polar factor, and can any fixed finite T bound it uniformly over normalized gradients without a positive lower bound on ℓ?

Hint: The Newton–Schulz update is

$$\mathbf{X}^{(t+1)} = \frac{3}{2}\mathbf{X}^{(t)} - \frac{1}{2}\mathbf{X}^{(t)}(\mathbf{X}^{(t)})^\top\mathbf{X}^{(t)},$$

and it preserves singular vectors while sending each singular value through

$$p(x) = \frac{3x - x^3}{2}.$$

On $[0, 1]$, p is nondecreasing, maps the interval into itself, and its positive iterates converge to 1, while $p^{\circ T}(0) = 0$. You may also use trace cyclicity and the first-order change $\mathcal{L}(\mathbf{W} + \mathbf{\Delta}) - \mathcal{L}(\mathbf{W}) \approx \langle \mathbf{X}^{(0)}, \mathbf{\Delta} \rangle_F$. Rescaling the gradient by a positive constant does not change the displayed fraction.

Takeaway: *A few Newton–Schulz steps can leave very small singular values far from their target of one. A fixed iteration budget therefore cannot guarantee a close approximation to the polar factor for every gradient.*