CSCI 145 Problem 18: The Polar Express

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 18: The Polar Express

Lecture ran two Newton–Schulz steps from a single starting value and watched it grow by a factor of about $3/2$ each time. A real gradient matrix hands the iteration a whole range of singular values at once, and a badly conditioned one starts very low: a matrix whose rescaled singular values reach down to 0.01 needs thirteen steps before the smallest of them clears 0.9. This problem explains where that thirteen comes from, and proves that repeating one fixed cubic is already enough to drive every singular value to 1.

Throughout, $p(x) = \frac{3x-x^3}{2}$ is lecture's Newton–Schulz cubic, and the matrix has been rescaled (by its Frobenius norm, as in lecture) so that every nonzero singular value lies in an interval $[\ell, 1]$ with $0 < \ell < 1$. Since orthogonalization succeeds exactly when every singular value reaches 1, the entire state of the algorithm is one number: how far the *worst* singular value still has to go.

Part A: One Step Tightens the Interval

Compute $p'(x)$ and check its sign to show that p is increasing on $[0, 1]$. Conclude that p maps $[\ell, 1]$ onto $[p(\ell), p(1)] = [p(\ell), 1]$, so the upper end stays pinned at the fixed point 1 and only the lower end can move. Then compute $p(\ell) - \ell$, factor it, and show it is positive for every $\ell \in (0, 1)$. One Newton–Schulz step therefore leaves the singular values inside a strictly smaller interval $[p(\ell), 1]$. Write $\ell_0 = \ell$ and $\ell_{t+1} = p(\ell_t)$ for that interval's low end after t steps.

Part B: Two Speeds

The iteration is slow at one end of $[0, 1]$ and fast at the other, which Part A says nothing about. First, factor $p(x) = \frac{3}{2}x\left(1 - \frac{x^2}{3}\right)$ and conclude that while ℓ_t is small the low end grows by a factor of nearly $\frac{3}{2}$ per step, so $\ell_t \approx \ell_0(3/2)^t$. Second, show that the *remaining* error obeys

$$1 - p(x) = \frac{(1-x)^2(2+x)}{2} \leq \frac{3}{2}(1-x)^2$$

for $x \in [0, 1]$ (expand and factor the cubic $x^3 - 3x + 2$), so once ℓ_t is close to 1 the distance to 1 is squared at every step.

Now put the two rates together on the hook's numbers. Starting from $\ell_0 = 0.01$, use the first rate to estimate the smallest t with $\ell_t > 0.9$ by solving $0.01 \cdot (1.5)^t = 0.9$, and compare it to the true answer of $t = 13$. Then use the second rate to bound how many further steps take the low end from 0.9 to within 10^{-6} of 1. Say in one sentence which of the two phases the running time of orthogonalization is actually spent in.

Part C: Where the Iteration Ends Up

Part A makes $(\ell_t)_{t=0}^{\infty}$ a strictly increasing sequence bounded above by 1. Recall (or take as given) that a bounded, monotonically increasing sequence of real numbers converges to some limit L. Using the continuity of p, argue that this limit satisfies $L = p(L)$, so L is one of the fixed points found in class (-1, 0, and 1). Since $\ell_0 > 0$ and the sequence increases, explain why L can be neither 0 nor -1, and conclude $L = 1$: repeating the *same* cubic, forever, sends every singular value in $(0, 1]$ to exactly 1, which is the orthogonalized gradient lecture asked for.

Finally, state in one sentence what orthogonalizing a matrix has turned into, and which single property of the gradient decides how long it takes.

Where this goes next. The Polar Express of Amsel, Persson, Musco, and Gower shortens Part B's slow climb by choosing a *different* odd polynomial at each step, the one lifting ℓ_t as high as any polynomial of that degree can. The state of the algorithm is unchanged: one worst-case interval, the same one-number summary the condition number gave us in the Optimization lecture.