Math 221 Homework 2

Atharv Sampath

Fall 2025

Problem 1

- **Question 2.2**: Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. Using Gaussian elimination with pivoting costs $\frac{2}{3}n^3+O(n^2)$ flops to compute an LU factorization, and each forward/back substitution to solve $Ly = Pb$, $Ux = y$ costs $O(n^2)$ flops, so solving one right-hand side after the factorization is $O(n^2)$ and m right-hand sides cost about $2n^2m$ flops in total. Hence Algorithm 1 (factor once, then solve) costs

  $$\boxed{\tfrac{2}{3}n^3 + 2n^2m + O(n^2)}.$$

  To "compute A^{-1} and multiply," observe that forming A^{-1} by solving $AX = I$ with the same LU takes the factorization cost $\frac{2}{3}n^3$ plus n solves, each $\approx 2n^2$ flops (two triangular substitutions), for about $2n^3$ more; thus inverting costs

  $$\tfrac{2}{3}n^3 + 2n^3 = \tfrac{8}{3}n^3 + O(n^2).$$

  Afterward, multiplying $X = A^{-1}B$ (dense GEMM) costs about $2n^2m$ flops. Therefore Algorithm 2 (invert then multiply) costs

  $$\boxed{\tfrac{8}{3}n^3 + 2n^2m + O(n^2)}.$$

  (The $2n^2m$ figure follows from the $2n^3$ cost for square GEMM in Table 2.1, scaling to an $n \times n$ by $n \times m$ product.) Comparing the leading terms shows Algorithm 2 exceeds Algorithm 1 by $2n^3$ flops, so the factor-and-solve approach is strictly cheaper.

- **Question 2.3**: Let $\|\cdot\|$ be the spectral (two-) norm. Write $\delta x = \hat{x} - x = A^{-1}(-\delta A\,\hat{x} + \delta b)$, so the bound

  $$\|\delta x\| \le \|A^{-1}\|\,(\|\delta A\|\,\|\hat{x}\| + \|\delta b\|)$$

  comes from two applications of norm inequalities: $\|A^{-1}w\| \le \|A^{-1}\|\,\|w\|$ with $w = -\delta A\,\hat{x} + \delta b$, and $\|u + v\| \le \|u\| + \|v\|$ with $u = -\delta A\,\hat{x}$, $v = \delta b$. To attain equality in both for the 2-norm, choose $v \in \mathbb{R}^n$ to be a right singular vector of A^{-1} associated with $\|A^{-1}\| = \sigma_{\max}(A^{-1})$; then $\|A^{-1}v\| = \|A^{-1}\|\,\|v\|$. Pick any $s, t > 0$ and define

  $$\delta A = -\frac{s}{\|\hat{x}\|^2}\,v\,\hat{x}^T, \qquad \delta b = t\,v.$$

Then $\delta A\,\hat{x} = -\frac{s}{\|\hat{x}\|^2}\,v\,\hat{x}^T\hat{x} = -s\,v$, so $-\delta A\,\hat{x}$ and δb are nonnegative multiples of the same vector v; hence $\|-\delta A\,\hat{x}+\delta b\| = \|-\delta A\,\hat{x}\| + \|\delta b\| = s+t$. Moreover $\|\delta A\| = \frac{s}{\|\hat{x}\|}$ (rank-one with unit left/right singular vectors v and $\hat{x}/\|\hat{x}\|$), so $\|\delta A\|\,\|\hat{x}\|+\|\delta b\| = s+t$. Therefore

$$\|\delta x\| = \|A^{-1}(-\delta A\,\hat{x}+\delta b)\| = \|A^{-1}\|\,\|-\delta A\,\hat{x}+\delta b\| = \|A^{-1}\|\big(\|\delta A\|\,\|\hat{x}\|+\|\delta b\|\big),$$

i.e., (2.2) holds with equality. Because s,t may be taken arbitrarily small, this works for sufficiently small $\|\delta A\|$ with $\delta A \neq 0$ and $\delta b \neq 0$. This shows the constant $\|A^{-1}\|$ in (2.2) is sharp for the 2-norm, and after rewriting (2.2) as a relative bound the optimal multiplicative factor is $\kappa(A) = \|A^{-1}\|\,\|A\|$, justifying its name as the condition number.

- **Question 2.4**: Let x solve $Ax=b$ and let $\hat{x}$ be any approximation. Write $\delta x = \hat{x}-x = A^{-1}(-\delta A\,\hat{x}+\delta b)$. The notes show that for any absolute vector norm (so $\|\,|z|\,\| = \|z\|$) and componentwise relative perturbations $|\delta A| \le \varepsilon|A|$, $|\delta b| \le \varepsilon|b|$, one has

  $$\|\delta x\| \le \varepsilon\,\big\|\,|A^{-1}|\big(|A|\,|\hat{x}|+|b|\big)\big\| \qquad \text{and, if } \delta b = 0,\quad \frac{\|\delta x\|}{\|x\|} \le \varepsilon\,\big\|\,|A^{-1}|\,|A|\,\big\|,$$

  i.e. (2.7) and (2.8). To show both bounds are attainable, set $g := |A|\,|\hat{x}|+|b| \ge 0$ and pick a dual vector u with $\|u\|_* = 1$ attaining the norm of the nonnegative vector $|A^{-1}|g$, so $u^T|A^{-1}|g = \big\||A^{-1}|g\big\|$ (for absolute norms such a maximizer can be chosen with nonnegative entries). Define the sign vector $s := \operatorname{sign}\big((A^{-1})^Tu\big) \in \{\pm 1\}^n$ and set

  $$\delta A = -\varepsilon\ \mathrm{Diag}(s)\,|A|\ \mathrm{Diag}(\operatorname{sign}(\hat{x})), \qquad \delta b = \varepsilon\ \mathrm{Diag}(s)\,|b|.$$

  Then $|\delta A| = \varepsilon|A|$, $|\delta b| = \varepsilon|b|$, and

  $$-\delta A\,\hat{x}+\delta b = \varepsilon\ \mathrm{Diag}(s)\big(|A|\,|\hat{x}|+|b|\big) = \varepsilon\ \mathrm{Diag}(s)\,g =: y.$$

  Consequently

  $$\|\delta x\| = \big\|A^{-1}y\big\| \ \ge\ u^TA^{-1}y = \varepsilon\,u^TA^{-1}\,\mathrm{Diag}(s)\,g = \varepsilon\,u^T|A^{-1}|\,g = \varepsilon\,\big\||A^{-1}|g\big\|.$$

  Since the derivation of (2.7) gave $\|\delta x\| \le \varepsilon\big\||A^{-1}|g\big\|$, we have equality. Thus (2.7) is sharp and attained by the explicit nonzero $\delta A, \delta b$ above. For (2.8), take $\delta b = 0$, so $y = \varepsilon\ \mathrm{Diag}(s)\,|A|\,|\hat{x}|$ and the same argument yields

  $$\|\delta x\| = \varepsilon\,\big\||A^{-1}|\,|A|\,|\hat{x}|\big\|.$$

  If $\hat{x}$ is chosen to attain the operator norm of $|A^{-1}|\,|A|$ (i.e. $\big\||A^{-1}|\,|A|\,|\hat{x}|\big\| = \big\||A^{-1}|\,|A|\big\|\,\|\hat{x}\|$), the weakened bound (2.8) also holds with equality. Therefore both (2.7) and (2.8) are attainable.

- **Question 2.7**: Let A be nonsingular and symmetric and suppose $A = LDM^T$ with L, M unit lower triangular and D diagonal. Since $A^T = A$, we also have $A^T = (LDM^T)^T = MDL^T$, hence $A = MDL^T$. Set $P := M^{-1}L$. From $A = LDM^T$ we get $M^{-1}AM^{-T} = PD$; from $A = MDL^T$ we get $M^{-1}AM^{-T} = DP^T$. Therefore $PD = DP^T$. Writing this entrywise and using that D is diagonal and nonsingular, for $i > j$ we have $(PD)_{ij} = p_{ij}d_{jj}$ and $(DP^T)_{ij} = d_{ii}p_{ji} = 0$ because P is lower triangular, so $p_{ij}d_{jj} = 0$ and hence $p_{ij} = 0$. Thus P has no strictly lower entries; since P is also lower triangular with unit diagonal, $P = I$. So, $M^{-1}L = I$ and $L = M$ as desired.

- **Question 2.8**: Let $A = \begin{bmatrix} 1 & 1 \\ 1 & 1+\varepsilon \end{bmatrix}$ with $0 < \varepsilon \ll 1$ and let $b \approx [a_{12}, a_{22}]^T = [1, 1+\varepsilon]^T$. In exact arithmetic the solution is $x = (0,1)^T$. Work in fixed-precision decimal floating point with four digits after the point (unit roundoff $u = \frac{1}{2} \cdot 10^{-4} = 5 \times 10^{-5}$), and perform every arithmetic operation in this arithmetic. Take $\varepsilon = 3 \times 10^{-5} < u$ and $b = \begin{bmatrix} 1 \\ 1+\varepsilon \end{bmatrix}$ (both components are exactly representable in this format). Cramer's rule forms

$$\widehat{\det} = \mathrm{fl}(a_{11}a_{22}) - \mathrm{fl}(a_{12}a_{21}), \qquad \widehat{x}_1 = \frac{\mathrm{fl}(a_{22}b_1) - \mathrm{fl}(a_{12}b_2)}{\widehat{\det}}, \qquad \widehat{x}_2 = \frac{-\,\mathrm{fl}(a_{21}b_1) + \mathrm{fl}(a_{11}b_2)}{\widehat{\det}}.$$

  Because $\varepsilon < u$, $\mathrm{fl}(a_{22}) = \mathrm{fl}(1+\varepsilon) = 1.0000$. Hence $\mathrm{fl}(a_{11}a_{22}) = \mathrm{fl}(1.0000 \cdot 1.0000) = 1.0000$ and $\mathrm{fl}(a_{12}a_{21}) = \mathrm{fl}(1.0000 \cdot 1.0000) = 1.0000$, so $\widehat{\det} = 0.0000$. Thus Cramer's rule divides by zero (or a denormal rounded to zero), i.e. it fails catastrophically on a problem whose exact determinant is $\det A = \varepsilon = 3 \times 10^{-5} \neq 0$. In particular, there is no $(\delta A, \delta b)$ with $\|\delta A\|/\|A\|,\ \|\delta b\|/\|b\| = O(u)$ such that Cramer's output $\widehat{x}$ satisfies $(A + \delta A)\widehat{x} = b + \delta b$; the algorithm has not produced any $\widehat{x}$ with a small backward error, because it has not produced a finite $\widehat{x}$ at all. By contrast, Gaussian elimination with partial pivoting (GEPP) applied in the same arithmetic performs one elimination step with pivot 1.0000 and returns $\widehat{x} = (0,1)^T$ with residual $r = b - A\widehat{x} = 0$, which is consistent with backward stability (residual of size $O(u) \cdot (\|A\|\|\widehat{x}\| + \|b\|)$). Therefore a concrete floating-point example shows that Cramer's rule is not backward stable.

Problem 2

Let $\|\cdot\|$ be any vector norm on $\mathbb{R}^n$ with dual norm $\|\cdot\|_* \equiv \max_{\|x\| \le 1} x^\top(\cdot)$. For the rank-one matrix $A = uv^\top$ and any $y \neq 0$,

$$Ay = u(v^\top y) \quad \Rightarrow \quad \|Ay\| = \|u\| \cdot |v^\top y| \le \|u\|\, \|v\|_*\, \|y\|,$$

where the inequality is the defining Hölder–dual inequality $|v^\top y| \le \|v\|_* \|y\|$. Taking the supremum over $y \neq 0$ in the induced norm gives $\|uv^\top\| \le \|u\|\,\|v\|_*$ (using the induced-norm definition $\|A\| = \max_{y \neq 0} \|Ay\|/\|y\|$). For the reverse inequality, compactness of the unit ball and continuity of $y \mapsto |v^\top y|$ guarantee a maximizer y_* with $\|y_*\| = 1$ and $|v^\top y_*| = \|v\|_*$; substituting y_* above yields $\|uv^\top y_*\| = \|u\|\|v\|_*$, hence $\|uv^\top\| \ge \|u\|\|v\|_*$. Combining both directions proves $\boxed{\|uv^\top\| = \|u\|\,\|v\|_*}$. (We use the standard induced-norm maximization over y as in the notes' operator-norm definitions/uses for general vector norms.)