Math 221 Homework 1

Atharv Sampath

Fall 2025

Problem 1

- **Question 1.1**: We have that

$$\det(A^\top A) = \det(A)\det(A^\top) = \det(A)^2 = 1.$$

  Hence, $\det(A) = \pm 1$. Without loss of generality, assume $\det(A) = +1$ and $\det(B) = -1$, then

$$\det(A + B) = \det(A^\top(A + B)) = \det(I + A^\top B).$$

  We can rewrite so that

$$\det(I + A^\top B) = \det((B^\top + A^\top)B) = -\det((A + B)^\top) = -\det(A + B).$$

  Thus, $\det(A + B) = 0$, so it is singular.

- **Question 1.3**: The columns of an orthogonal matrix must be an orthonormal set. Taking pairwise inner products of the columns of an upper triangular matrix shows that it is forced to be diagonal. Since the column vectors are orthonormal, they must also be unit vectors, which means that the entries of the diagonal are ± 1.

- **Question 1.5**: We have that

$$\|x + y\|_C = \|C(x + y)\| = \|Cx + Cy\| \leq \|Cx\| + \|Cy\| = \|x\|_C + \|y\|_C.$$

  Moreover, for k a constant,

$$\|kx\|_C = \|C(kx)\| = \|kCx\| = |k|\|Cx\| = |k|\|x\|_C.$$

  Finally, if

$$\|x\|_C = \|Cx\| = 0$$

  then we must have that $Cx = 0$. Then, since $\operatorname{rank}(C) = n$, we have that $\ker(C) = 0$, which means that $x = 0$. Thus $\|\cdot\|_C$ satisfies the required conditions for being a norm.

- **Question 1.9**: Let u be the unit roundoff and assume IEEE rounding to nearest. For any basic operation,

$$\mathrm{fl}(a \circ b) = (a \circ b)(1+\delta), \qquad |\delta| \le u.$$

Write $d = \mathrm{fl}(1+x) = (1+x)(1+\delta_1)$ with $|\delta_1| \le u$. The algorithm computes

$$\tilde{y}_1 = \mathrm{fl}\left(\frac{\log(d)}{x}\right) = \frac{\log(d)}{x}(1+\delta_2).$$

Since log is assumed exact,

$$\log(d) = \log(1+x) + \log(1+\delta_1) = \log(1+x) + \varepsilon, \qquad |\varepsilon| \le C_1 u.$$

Hence

$$\tilde{y}_1 = \frac{\log(1+x)}{x} + \frac{\varepsilon}{x} + O(u) = y(x) + \frac{\varepsilon}{x} + O(u).$$

The term $|\varepsilon|/|x| \gtrsim u/|x|$ makes the relative error unbounded as $x \to 0$. In particular, if $|x| < \frac{1}{2}\mathrm{ulp}(1) = u$, then $d = 1$, $\log(d) = 0$, and $\tilde{y}_1 = 0$ although $y(0) = 1$.

Let $d = \mathrm{fl}(1+x)$ as above. If $d = 1$ the algorithm returns 1, which equals $\lim_{x\to 0} \log(1+x)/x$. Otherwise it computes

$$\tilde{y}_2 = \mathrm{fl}\left(\frac{\log(d)}{d-1}\right) = \frac{\log(d)}{d-1}(1+\delta_3).$$

Set $g(z) = \log(z)/(z-1)$. For $|z-1| < \frac{1}{2}$ the series $\log z = (z-1) - \frac{1}{2}(z-1)^2 + \cdots$ gives g smooth and bounded, with $g(1) = 1$.

Now compare $g(d)$ with $y(x)$. Since $d = (1+x)(1+\delta_1)$,

$$d - 1 = x + \delta_1(1+x), \qquad \frac{x}{d-1} = 1 + O(u).$$

By the mean–value theorem,

$$\log(d) - \log(1+x) = \frac{d-(1+x)}{\xi} = \frac{\delta_1(1+x)}{\xi} \quad \text{for some } \xi \in (1+x, d),$$

so $\log(d)/\log(1+x) = 1 + O(u)$. Therefore

$$\frac{g(d)}{y(x)} = \frac{\log(d)}{\log(1+x)} \cdot \frac{x}{d-1} = (1+O(u))(1+O(u)) = 1 + O(u).$$

Including the last division gives

$$\frac{|\tilde{y}_2 - y(x)|}{|y(x)|} \le Cu,$$

with a modest constant C, uniformly for x near 0.

Thus the first algorithm suffers catastrophic cancellation (error $\sim u/|x|$), while the second is backward stable: it returns $g(d)$ with $d = \mathrm{fl}(1+x)$, and g is well conditioned.

- **Question 1.11**: Let $L \in \mathbb{R}^{n\times n}$ be lower triangular and assume IEEE rounding to nearest with unit roundoff ε, and no underflow or overflow. In floating point, $\mathrm{fl}(a \circ b) = (a \circ b)(1+\delta)$ with $|\delta| \le \varepsilon$, and the computed dot product satisfies the backward–error lemma: for any vectors u, v of length k,

  $$\mathrm{fl}(u^T v) = (u + \Delta u)^T v, \qquad |\Delta u| \le \gamma_k |u|, \qquad \gamma_k = \frac{k\varepsilon}{1-k\varepsilon} \le k\varepsilon + O(\varepsilon^2),$$

  with inequalities understood componentwise. Apply forward substitution. For each i, let $\widehat{s}_i = \mathrm{fl}\big(\sum_{j=1}^{i-1} l_{ij}\widehat{x}_j\big)$. By the lemma there exist perturbations Δl_{ij} for $j < i$ such that

  $$\widehat{s}_i = \sum_{j=1}^{i-1} (l_{ij} + \Delta l_{ij})\widehat{x}_j, \qquad |\Delta l_{ij}| \le \gamma_{i-1}|l_{ij}|.$$

  The next step computes $\widehat{x}_i = \mathrm{fl}\left((b_i - \widehat{s}_i)/l_{ii}\right) = ((b_i - \widehat{s}_i)/l_{ii})(1+\delta_i)$ with $|\delta_i| \le \gamma_2$. Multiply by l_{ii} and rearrange:

  $$b_i = (l_{ii} + \Delta l_{ii})\widehat{x}_i + \widehat{s}_i, \qquad \Delta l_{ii} := l_{ii}\big((1+\delta_i)^{-1} - 1\big), \qquad |\Delta l_{ii}| \le \gamma_2 |l_{ii}|.$$

  Substitute the expression for $\widehat{s}_i$ to obtain

  $$b_i = \sum_{j=1}^{i} (l_{ij} + \Delta l_{ij})\widehat{x}_j,$$

  which is the i-th equation of $(L+\Delta L)\widehat{x} = b$. Entrywise bounds follow from the previous estimates: for $j < i$, $|\Delta l_{ij}| \le \gamma_{i-1}|l_{ij}|$, and for $j = i$, $|\Delta l_{ii}| \le \gamma_2 |l_{ii}|$. Since $\gamma_k \le k\varepsilon/(1-k\varepsilon) \le n\varepsilon$ for $k \le n$ and ε sufficiently small, we have $|\Delta l_{ij}| \le n\varepsilon |l_{ij}|$ for all $j \le i$. Thus the computed solution $\widehat{x}$ satisfies $(L + \Delta L)\widehat{x} = b$ with $|\Delta L| \le n\varepsilon |L|$ componentwise, so forward substitution is backward stable. The same argument applied to the dot products that appear in backward substitution for an upper triangular system gives the identical bound, hence backward substitution is also backward stable.

- **Question 1.12**: Let $a = x + iy$ and $b = u + iv$ with IEEE rounding to nearest and unit roundoff ε; for real operations we use the standard model $\mathrm{fl}(\alpha \circ \beta) = (\alpha \circ \beta)(1+\delta)$, $|\delta| \le \varepsilon$, and the dot–product lemma $\mathrm{fl}(r^T s) = (r^T s)(1+\theta_k)$, $|\theta_k| \le \gamma_k := k\varepsilon/(1-k\varepsilon)$. For complex addition/subtraction the implementation rounds the two real components separately, hence $\widehat{z} = \mathrm{fl}(a \pm b) = ((x \pm u)(1+\delta_1)) + i((y \pm v)(1+\delta_2)) = (a \pm b) + e$ with $|e| \le \sqrt{2}\,\varepsilon\,|a \pm b|$; letting $\delta := e/(a \pm b)$ gives $\widehat{z} = (a \pm b)(1+\delta)$ and $|\delta| \le \sqrt{2}\varepsilon$. For complex multiplication write $\Re(ab) = [x, -y]\cdot[u, v]$ and $\Im(ab) = [x, y]\cdot[v, u]$; each is a length-2 dot product, so the computed parts satisfy $\widehat{r} = \Re(ab)(1+\theta_2)$, $\widehat{s} = \Im(ab)(1+\theta_2')$ with $|\theta_2|, |\theta_2'| \le \gamma_2$. Hence $\widehat{z} = \widehat{r} + i\widehat{s} = ab + e$ with $|e| \le \sqrt{2}\,\gamma_2\,|ab|$, so $\widehat{z} = ab(1+\delta)$ and $|\delta| \le \sqrt{2}\gamma_2 \le 3\varepsilon + O(\varepsilon^2)$. For division use Smith's scaled algorithm: if $|u| \ge |v|$, set $r = v/u$, $d = u + vr$, and compute $\widehat{z} = \mathrm{fl}\big((x + yr)/d\big) + i\,\mathrm{fl}\big((y - xr)/d\big)$; otherwise swap the roles of u and v. Then $|r| \le 1$ and $|d| \ge |u|$ (or $|d| \ge |v|$), so intermediates have size $O(|a| + |b|)$ and no squaring occurs; consequently the algorithm succeeds even when $|a|$ is very large (exceeding $\sqrt{\mathrm{OF}}$) or very small (below $\sqrt{\mathrm{UF}}$), in particular it computes $a/a \approx 1$. Each numerator is a length-3 dot product and the denominator is a length-2 dot product followed by one division, so by the dot–product model and quotient perturbation we obtain $\widehat{z} = (a/b)(1+\delta)$ with $|\delta| \le c\,\varepsilon$ for a modest constant c (e.g. $c \le 8$) provided no over/underflow occurs. Finally, it is **not** true that the real and imaginary parts of a complex product are always computed to high relative accuracy: if $xu \approx yv$ (or $xv \approx yu$) the corresponding component suffers catastrophic cancellation (e.g. $(1 + i\eta)(1 - i\eta)$ has exact imaginary part 0 but the computed one can be dominated by rounding), so only the complex relative error is uniformly small.

Problem 2

Assume IEEE rounding to nearest with unit roundoff ε and no under/overflow. For real scalars we use $\mathrm{fl}(a \circ b) = (a \circ b)(1+\delta)$, $|\delta| \le \varepsilon$, and for a sequential sum of m terms the standard bound $\mathrm{fl}(\sum_{k=1}^m t_k) = (\sum_{k=1}^m t_k)\,(1+\theta_m)$ with $|\theta_m| \le \gamma_m := m\varepsilon/(1-m\varepsilon)$. Consider the dot product computed in the usual way: $s_0 = 0$, $p_i = \mathrm{fl}(x_i y_i)$, $s_i = \mathrm{fl}(s_{i-1} + p_i)$ for $i = 1, \dots, d$. By the multiplication and summation models,

$$p_i = x_i y_i (1+\delta_i'), \quad |\delta_i'| \le \varepsilon, \qquad s_d = \Big(\sum_{i=1}^d p_i\Big)(1+\theta_{d-1}) = \Big(\sum_{i=1}^d x_i y_i (1+\delta_i')\Big)(1+\theta_{d-1}).$$

Hence $\mathrm{fl}\big(\sum_{i=1}^d x_i y_i\big) = \big(\sum_{i=1}^d x_i y_i\big)(1+\theta_{2d})$ with $|\theta_{2d}| \le \gamma_{2d}$. Writing $\delta_i \equiv \theta_{2d}$ for all i gives

$$\mathrm{fl}\Big(\sum_{i=1}^d x_i y_i\Big) = \sum_{i=1}^d x_i y_i (1+\delta_i), \qquad |\delta_i| \le \gamma_{2d} \le d\varepsilon + O(\varepsilon^2).$$

For matrix multiplication $C = \mathrm{fl}(AB)$ with $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n \times p}$, each entry is a length-n dot product:

$$\widehat{c}_{ij} = \mathrm{fl}\Big(\sum_{k=1}^n a_{ik} b_{kj}\Big) = \sum_{k=1}^n a_{ik} b_{kj} (1+\delta_k^{(ij)}), \qquad |\delta_k^{(ij)}| \le n\varepsilon + O(\varepsilon^2).$$

Therefore, componentwise,

$$|\widehat{c}_{ij} - c_{ij}| \le n\varepsilon \sum_{k=1}^n |a_{ik}||b_{kj}| = \big(n\varepsilon\,|A|\,|B|\big)_{ij},$$

hence $|\mathrm{fl}(AB) - AB| \le n\varepsilon\,|A|\,|B|$ to first order.

For the extension when the inputs are not exactly floating-point, let $\widetilde{A} = \mathrm{fl}(A)$ and $\widetilde{B} = \mathrm{fl}(B)$. Then $|\widetilde{A} - A| \le \varepsilon|A|$ and $|\widetilde{B} - B| \le \varepsilon|B|$ componentwise. Computing with $\widetilde{A}, \widetilde{B}$ yields

$$\mathrm{fl}(\widetilde{A}\widetilde{B}) = \widetilde{A}\widetilde{B} + \Delta_1, \qquad |\Delta_1| \le n\varepsilon\,|\widetilde{A}|\,|\widetilde{B}| \le n\varepsilon(1+\varepsilon)^2|A|\,|B| = n\varepsilon|A|\,|B| + O(n\varepsilon^2).$$

Moreover,

$$\widetilde{A}\widetilde{B} - AB = A(\widetilde{B} - B) + (\widetilde{A} - A)B + (\widetilde{A} - A)(\widetilde{B} - B),$$

so $|\widetilde{A}\widetilde{B} - AB| \le 2\varepsilon|A|\,|B| + O(\varepsilon^2)$. Combining,

$$|\mathrm{fl}(AB) - AB| \le (n+2)\varepsilon\,|A|\,|B| + O\big((n+1)\varepsilon^2\big),$$

which is the same first-order bound up to the unavoidable additional 2ε from the initial rounding of the data. The analogous extension for a scalar dot product gives $\mathrm{fl}\big(\sum_{i=1}^d x_i y_i\big) = \sum_{i=1}^d x_i y_i (1+\delta_i)$ with $|\delta_i| \le (d+2)\varepsilon + O(\varepsilon^2)$, reflecting one rounding for each input and the length-d accumulation.

Problem 3

Let $s = [s_1, s_2, \ldots, s_n]^\top$. Let $p \geq 1$ be a real number. Then, we have that $p \mapsto \|s\|_p$ is a differentiable function. Note that

$$\frac{\mathrm{d}}{\mathrm{d}\, p}\|s\|_p = \frac{1}{p\|s\|_p^{p-1}} \left(\sum_{i=0}^{\infty} |s_i|^p \ln(|s_i|) - \|s\|_p^p \ln\left(\|s\|_p\right) \right).$$

We see that p and $\|s\|_p^{p-1}$ must be positive, so let us consider

$$\sum_{i=0}^{\infty} |s_i|^p \ln(|s_i|) - \|s\|_p^p \ln\left(\|s\|_p\right) = \sum_{i=0}^{\infty} |x_i|^p \ln(|x_i|) - \frac{1}{p} \ln\left(\sum_{j=0}^{\infty} |x_j|^p\right) \sum_{i=0}^{\infty} |x_i|^p.$$

To determine, the sign of this, for an individual $i \in \mathbb{N}$, we simply need to consider

$$\ln(|x_i|) - \frac{1}{p} \ln\left(\sum_{j=0}^{\infty} |x_j|^p\right),$$

and taking exp of both terms to get

$$|x_i| - \left(\sum_{j=0}^{\infty} |x_j|^p\right)^{\frac{1}{p}} = (|x_i|^p)^{\frac{1}{p}} - \left(\sum_{j=0}^{\infty} |x_j|^p\right)^{\frac{1}{p}}.$$

And finally, taking both terms to the power of p, we have that the sign of the original expression is the same as the sign of

$$|x_i|^p - \sum_{j=0}^{\infty} |x_j|^p,$$

which is obviously negative. Thus, $\frac{\mathrm{d}}{\mathrm{d}\, p}\|s\|_p < 0$ for all $p \in [1, \infty)$, so $p \mapsto \|s\|_p$ is decreasing in p. Taking the s_i to be the singular values of a compact and bounded operator T between countable dimension Hilbert spaces (i.e. finite dimensional vector spaces or ℓ^2 spaces), we recover that $\|s\|_p$ is the Schatten p-norm of T. Thus, the Schatten norm is decreasing in p.

Problem 4

We have that

$$\|z+z'\|_* = \max_{\|x\|\leq 1}(z+z')^\top x = \max_{\|x\|\leq 1}(z^\top x + z'^\top x) \leq \max_{\|x\|\leq 1} z^\top x + \max_{\|x\|\leq 1} z'^\top x = \|z\|_* + \|z'\|_*.$$

Similarly,

$$\|kz\|_* = \max_{\|x\|\leq 1}(kz)^\top x = \max_{\|x\|\leq 1} k(z^\top x).$$

If k is positive, then we have that

$$\max_{\|x\|\leq 1} k(z^\top x) = k \max_{\|x\|\leq 1} z^\top x = |k|\|z\|_*.$$

Instead, if k is negative, suppose that

$$\hat{x} = \operatorname*{argmax}_{\|x\|\leq 1} z^\top x,$$

and note that

$$\max_{\|x\|\leq 1} k(z^\top x) = |k| \max_{\|x\|\leq 1} -z^\top x.$$

But if we pick $x = -\hat{x}$, then we have that

$$|k| \max_{\|x\|\leq 1} -z^\top x = |k| \max_{\|x\|\leq 1} z^\top x = |k|\|z\|_*$$

as desired. Finally, if

$$\|z\|_* = \max_{\|x\|\leq 1} z^\top x = 0 = \max_{\|x\|\leq 1} -z^\top x,$$

which implies that $z = 0$. Thus, $\|\cdot\|_*$ is indeed a norm.

Problem 5

First, substituting the expressions for Δx and Δa_i, we can write

$$f(x+\Delta x, a+\Delta a) \le \sum_{i=0}^{d}(a_i+\delta_a|a_i|)(x+\delta_x|x|)^i.$$

Subtracting $f(x,a)$, we have

$$f(x+\Delta x, a+\Delta a) - f(x,a) \le \sum_{i=0}^{d}(a_i+\delta_a|a_i|)(x+\delta_x|x|)^i - \sum_{i=0}^{d} a_i x^i.$$

Define

$$t := \frac{\Delta x}{x}, \qquad \alpha_i := \frac{\Delta a_i}{a_i},$$

so that $|t| \le \delta_x$ and $|\alpha_i| \le \delta_a$. Then

$$\begin{aligned} f(x+\Delta x, a+\Delta a) - f(x,a) &= \sum_{i=0}^{d}(a_i+\Delta a_i)(x+\Delta x)^i - \sum_{i=0}^{d} a_i x^i \\ &= \sum_{i=0}^{d} a_i \Big[(1+\alpha_i)\, x^i (1+t)^i - x^i\Big] \\ &= \sum_{i=0}^{d} a_i x^i \Big[(1+\alpha_i)(1+t)^i - 1\Big]. \end{aligned}$$

Taking absolute values and using the triangle inequality:

$$\Big|(1+\alpha_i)(1+t)^i - 1\Big| \le |\alpha_i|\,(1+|t|)^i + \Big|(1+t)^i - 1\Big|.$$

Now for $u \in [0,1)$ and any integer $i \ge 1$ such that $iu < 1$, we have

$$(1+u)^i \le e^{iu} \le \frac{1}{1-iu}$$

since

$$\sum_{k\ge 0} \frac{(iu)^k}{k!} \le \sum_{k\ge 0}(iu)^k.$$

Moreover,

$$(1+u)^i - 1 \le e^{iu} - 1 \le \frac{iu}{1-iu},$$

because

$$\sum_{k\ge 1} \frac{(iu)^k}{k!} \le \sum_{k\ge 1}(iu)^k.$$

With $u = |t| \le \delta_x$ (assuming $d\delta_x < 1$), these give

$$(1+|t|)^i \le \frac{1}{1-i\delta_x} \le \frac{1}{1-d\delta_x}, \qquad \Big|(1+t)^i - 1\Big| \le \frac{i\delta_x}{1-i\delta_x} \le \frac{d\delta_x}{1-d\delta_x}.$$

Therefore

$$\begin{aligned}
|f(x+\Delta x, a+\Delta a) - f(x,a)| &\leq \sum_{i=0}^{d} |a_i|\,|x|^i \Big(|\alpha_i|\,(1+|t|)^i + \big|(1+t)^i - 1\big| \Big) \\
&\leq \sum_{i=0}^{d} |a_i|\,|x|^i \left(\frac{\delta_a}{1-d\delta_x} + \frac{d\delta_x}{1-d\delta_x} \right) \\
&= \frac{\delta_a + d\delta_x}{1-d\delta_x} \sum_{i=0}^{d} |a_i|\,|x|^i \\
&= \frac{\delta_a + d\delta_x}{1-d\delta_x}\, f(|x|,|a|).
\end{aligned}$$

Which is the desired bound, assuming $d\delta_x < 1$ and $x \neq 0$.