- **Linear Algebra** (15 points)

1. (5 points) Let

$$\boldsymbol{A} = \begin{bmatrix} \lambda & 1 \\ 1 & \lambda \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

For which values of λ does the equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ have

- a) No solutions?
- b) A unique solution? (Also, state the unique solution.)
- c) Infinitely many solutions? (Also, state the set of all solutions).

Solution.

Apply Gaussian elimination to the augmented matrix.

$$\left[\begin{array}{cc|c} \lambda & 1 & 1 \\ 1 & \lambda & 1 \end{array}\right]$$
$$\sim \left[\begin{array}{cc|c} 0 & 1-\lambda^2 & 1-\lambda \\ 1 & \lambda & 1 \end{array}\right] (R_1 := R_1 - \lambda R_2)$$
$$\sim \left[\begin{array}{cc|c} 1 & \lambda & 1 \\ 0 & 1-\lambda^2 & 1-\lambda \end{array}\right] (\text{Swap } R_1 \text{ and } R_2.)$$

Hence we obtain the following equations for the solution $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$.

$$x + \lambda y = 1$$

$$(1-\lambda^2)y = 1 - \lambda$$

Clearly, if $\lambda = \pm 1, 1 - \lambda^2 = 0$.

Proceed by cases,

(a) Assume $\lambda = -1$.
The above equation degenerates to $0y = 2$, hence no solutions exist.

(b) Assume $\lambda \neq \pm 1$.
We have

$$\begin{aligned} (1-\lambda^2)y &= 1 - \lambda \\ y &= \frac{1-\lambda}{1-\lambda^2} = \frac{1-\lambda}{(1-\lambda)(1+\lambda)} = \frac{1}{1+\lambda} \end{aligned}$$

Substituting into the other equation, we obtain

$$\begin{aligned} x + \lambda y &= 1 \\ x + \frac{\lambda}{1+\lambda} &= 1 \\ x &= 1 - \frac{\lambda}{1+\lambda} = \frac{1+\lambda}{1+\lambda} - \frac{\lambda}{1+\lambda} = \frac{1}{1+\lambda} \end{aligned}$$

Hence, the only unique solution is

$$\mathbf{x} = \begin{bmatrix} (1+\lambda)^{-1} \\ (1+\lambda)^{-1} \end{bmatrix}$$

(c) Assume $\lambda = 1$.
The above equation degenerates to $0y = 0$, and y is a free variable. The other equation gives us $x = 1 - y$, and thus the solution set is

$$\left\{ \begin{bmatrix} 1-\alpha \\ \alpha \end{bmatrix} : \alpha \in \mathbb{R} \right\}$$

and we have infinitely many solutions.

2. (4 points) Let

$$W = \left\{ \begin{bmatrix} r-s+t \\ 11s-t \\ 3r+5s \\ 7t \end{bmatrix} : \quad r,s,t \in \mathbb{R} \right\} \subseteq \mathbb{R}^4.$$

– a) Is W a vector subspace of $\mathbb{R}^4$? Explain
Solution. Yes, we verify the three subspace axioms.

(a) Zero vector.
Clearly, by choosing $r=0, s=0, t=0$, we obtain the zero vector.

(b) Closure under addition.
Let $\mathbf{w}_1, \mathbf{w}_2$ be vectors in W. Then there exists constants $r_1, s_1, t_1, r_2, s_2, t_2$ such that

$$\mathbf{w}_1 = \begin{bmatrix} r_1-s_1+t_1 \\ 11s_1-t_1 \\ 3r_1+5s_1 \\ 7t_1 \end{bmatrix} \text{ and } \mathbf{w}_2 = \begin{bmatrix} r_2-s_2+t_2 \\ 11s_2-t_2 \\ 3r_2+5s_2 \\ 7t_2 \end{bmatrix}$$

Then,

$$\begin{aligned} \mathbf{w}_1+\mathbf{w}_2 &= \begin{bmatrix} r_1-s_1+t_1 \\ 11s_1-t_1 \\ 3r_1+5s_1 \\ 7t_1 \end{bmatrix} + \begin{bmatrix} r_2-s_2+t_2 \\ 11s_2-t_2 \\ 3r_2+5s_2 \\ 7t_2 \end{bmatrix} \\ &= \begin{bmatrix} (r_1+r_2)-(s_1+s_2)+(t_1+t_2) \\ 11(s_1+s_2)-(t_1+t_2) \\ 3(r_1+r_2)+5(s_1+s_2) \\ 7(t_1+t_2) \end{bmatrix} \in W \end{aligned}$$

(c) Closure under scalar multiplication.
Let $\mathbf{w} \in W$. Then there exists scalars r, s, t such that

$$\mathbf{w} = \begin{bmatrix} r-s+t \\ 11s-t \\ 3r+5s \\ 7t \end{bmatrix}$$

Then,

$$\lambda\mathbf{w} = \lambda \begin{bmatrix} r-s+t \\ 11s-t \\ 3r+5s \\ 7t \end{bmatrix} = \begin{bmatrix} (\lambda r)-(\lambda s)+(\lambda t) \\ 11(\lambda s)-(\lambda t) \\ 3(\lambda r)+5(\lambda s) \\ 7(\lambda t) \end{bmatrix} \in W$$

Alternatively, one could simply verify that the constraints are all linear equations, and any subset of a vector space with linear constraints is a vector subspace.

– b) Calculate a basis of W. What is the dimension of W?
Solution. Decompose the vector down per variable.

$$\begin{bmatrix} r-s+t \\ 11s-t \\ 3r+5s \\ 7t \end{bmatrix} = r \begin{bmatrix} 1 \\ 0 \\ 3 \\ 0 \end{bmatrix} + s \begin{bmatrix} -1 \\ 11 \\ 5 \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ -1 \\ 0 \\ 7 \end{bmatrix}$$

We verify that these vectors are linearly independent via Gaussian elimination.

$$\begin{bmatrix} 1 & -1 & 1 \\ 0 & 11 & -1 \\ 3 & 5 & 0 \\ 0 & 0 & 7 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Hence a basis of W is given by

$$\left\{ \begin{bmatrix} 1 \\ 0 \\ 3 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 11 \\ 5 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 0 \\ 7 \end{bmatrix} \right\}$$

which is a 3 dimensional subspace.

3. (2 points) If $\boldsymbol{A} \in \mathbb{R}^{3\times 4}$, and the rows of $\boldsymbol{A}$ are linearly independent, compute $\text{rank}(\boldsymbol{A}^T)$.

   Solution. If the rows of $\mathbf{A}$ are linearly independent, then the columns of $\mathbf{A}^T \in \mathbb{R}^{4\times 3}$ are linearly independent, and hence the columns of $\mathbf{A}^T$ span a 3 dimensional space. Hence, $\text{rank}(\mathbf{A}^T) = 3$

4. (4 points) Let $(G, \otimes)$ be a group, and let $\phi : G \to G$ be a function satisfying the property that for all $a, b \in G$, $\phi(a \otimes b) = \phi(a) \otimes \phi(b)$. Prove that $\phi(e) = e$, where e is the neutral element in the group $(G, \otimes)$.

   Solution.

$$
\begin{aligned}
\phi(e) &= \phi(e \otimes e) && \text{By definition of neutral element} \\
\phi(e) &= \phi(e) \otimes \phi(e) && \text{By given property of } \phi. \\
\phi(e) \otimes \phi(e)^{-1} &= (\phi(e) \otimes \phi(e)) \otimes \phi(e)^{-1} && \text{Right multiply by inverse of } \phi(e). \\
e &= \phi(e) \otimes (\phi(e) \otimes \phi(e)^{-1}) && \text{Associtivity of } \otimes. \\
e &= \phi(e) \otimes e && \text{Definition of inverse} \\
e &= \phi(e) && \text{Definition of neutral element}
\end{aligned}
$$

- **Analytic Geometry** (10 points)

  Given a vector space V, we say that a norm $||\cdot||_\alpha : V \to \mathbb{R}$ is *equivalent* to another norm $||\cdot||_\beta : V \to \mathbb{R}$ if there exists constants $M, N > 0$ such that for any vector $\mathbf{x} \in V$, we have

  $$M||\boldsymbol{x}||_\alpha \leq ||\boldsymbol{x}||_\beta \leq N||\boldsymbol{x}||_\alpha$$

  We denote that $||\cdot||_\alpha$ is equivalent to $||\cdot||_\beta$ by writing $||\cdot||_\alpha \sim ||\cdot||_\beta$.

1. (2 points) Prove that equivalence is symmetric, that is,

   $$||\cdot||_\alpha \sim ||\cdot||_\beta \Rightarrow ||\cdot||_\beta \sim ||\cdot||_\alpha$$

   Solution.
   Assume that $\|\cdot\|_\alpha$ is equivalent to $\|\cdot\|_\beta$. There there exists $M, N > 0$ such that for all $\mathbf{x} \in V$, we have that

   $$M||\boldsymbol{x}||_\alpha \leq ||\boldsymbol{x}||_\beta \leq N||\boldsymbol{x}||_\alpha$$

   Note that since $M||\boldsymbol{x}||_\alpha \leq ||\boldsymbol{x}||_\beta$, we have $\|\mathbf{x}\|_\alpha \leq \frac{1}{M}\|\mathbf{x}\|_\beta$
   Also, $||\boldsymbol{x}||_\beta \leq N||\boldsymbol{x}||_\alpha$ implies $\frac{1}{N}||\boldsymbol{x}||_\beta \leq ||\boldsymbol{x}||_\alpha$. Combining, we have

   $$\frac{1}{N}||\boldsymbol{x}||_\beta \leq ||\boldsymbol{x}||_\alpha \leq \frac{1}{M}\|\mathbf{x}\|_\beta$$

   and thus $\|\cdot\|_\beta$ is equivalent to $\|\cdot\|_\alpha$.

2. In vector space $\mathbb{R}^2$, the *Euclidean* norm $||\cdot||_2$ and the *Manhattan* norm $||\cdot||_1$ are given by

$$\begin{aligned}||\boldsymbol{x}||_2 &:= \sqrt{x_1^2 + x_2^2}\\ ||\boldsymbol{x}||_1 &:= |x_1| + |x_2|.\end{aligned}$$

(5 points) Prove that for any vector $\boldsymbol{x} \in \mathbb{R}^2$,

$$||\boldsymbol{x}||_2 \leq ||\boldsymbol{x}||_1 \leq \sqrt{2}||\boldsymbol{x}||_2$$

Solution.

$$\begin{aligned}\|\mathbf{x}\|_2^2 &= x_1^2 + x_2^2\\ &= |x_1|^2 + |x_2|^2\\ &\leq |x_1|^2 + |x_2|^2 + 2|x_1||x_2|\\ &= (|x_1| + |x_2|)^2\\ &= \|\mathbf{x}\|_1^2\end{aligned}$$

Note that

$$\begin{aligned}(|x_1| - |x_2|)^2 &\geq 0\\ |x_1|^2 + |x_2|^2 - 2|x_1||x_2| &\geq 0\\ 2|x_1||x_2| &\leq |x_1|^2 + |x_2|^2\end{aligned}$$

Hence,

$$\begin{aligned}\|\mathbf{x}\|_1^2 &= (|x_1| + |x_2|)^2\\ &= |x_1|^2 + |x_2|^2 + 2|x_1||x_2|\\ &\leq |x_1|^2 + |x_2|^2 + |x_1|^2 + |x_2|^2\\ &= 2(|x_1|^2 + |x_2|^2)\\ &= 2(x_1^2 + x_2^2) = 2\|\mathbf{x}\|_2^2\end{aligned}$$

Combining the above statements, and square rooting,

$$\|\mathbf{x}\|_2^2 \leq \|\mathbf{x}\|_1^2 \leq 2\|\mathbf{x}\|_2^2$$

$$\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{2}\|\mathbf{x}\|_2$$

(0.5 point) Also, find a particular vector $\boldsymbol{y} \in \mathbb{R}^2$ such that

$$||\boldsymbol{y}||_1 = \sqrt{2}||\boldsymbol{y}||_2$$

Solution.
Choose $\mathbf{y} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

$$\|\mathbf{y}\|_1 = |1| + |1| = 2$$
$$\|\mathbf{y}\|_2 = \sqrt{1^2 + 1^2} = \sqrt{2}$$

Hence,

$$\sqrt{2}\|\mathbf{y}\|_2 = \sqrt{2}\sqrt{2} = 2 = \|\mathbf{y}\|_1$$

(0.5 point) and a particular vector $\boldsymbol{z} \in \mathbb{R}^2$ such that

$$||\boldsymbol{z}||_1 = ||\boldsymbol{z}||_2$$

Solution.
Choose $\mathbf{z} = \mathbf{0}$. Then

$$\|\mathbf{z}\|_2 = 0 = \|\mathbf{z}\|_1$$

by positive definiteness.

3. (2 points) Let $(V, \langle\cdot,\cdot\rangle)$ be an inner product space. Let

$$\mathrm{proj}_{\mathbf{u}}(\mathbf{v}) := \frac{\langle \mathbf{v}, \mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle}\mathbf{u}$$

denote the vector projection of $\mathbf{v}$ onto $\mathbf{u}$. Prove that $\mathbf{v} - \mathrm{proj}_{\mathbf{u}}(\mathbf{v})$ and $\mathbf{u}$ are orthogonal.
Solution.

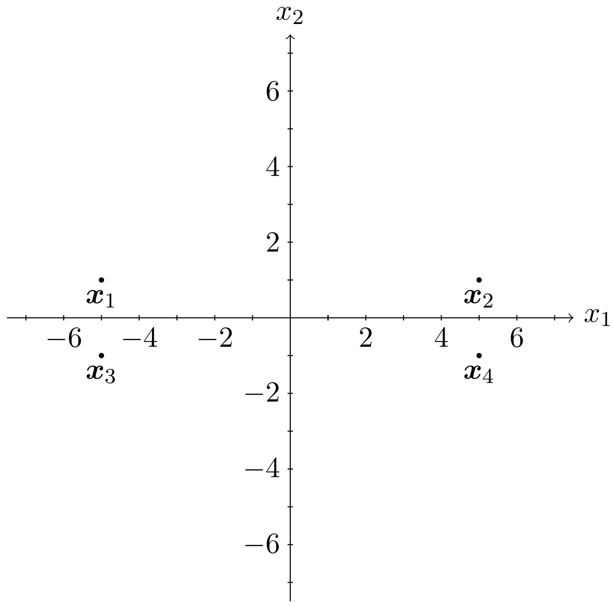
$$\begin{aligned}
\langle \mathbf{v} - \mathrm{proj}_{\mathbf{u}}(\mathbf{v}), \mathbf{u}\rangle &= \langle \mathbf{v} - \frac{\langle \mathbf{v}, \mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle}\mathbf{u}, \mathbf{u}\rangle \\
&= \langle \mathbf{v}, \mathbf{u}\rangle - \langle \frac{\langle \mathbf{v}, \mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle}\mathbf{u}, \mathbf{u}\rangle \\
&= \langle \mathbf{v}, \mathbf{u}\rangle - \frac{\langle \mathbf{v}, \mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle}\langle \mathbf{u}, \mathbf{u}\rangle \\
&= \langle \mathbf{v}, \mathbf{u}\rangle - \langle \mathbf{v}, \mathbf{u}\rangle = 0
\end{aligned}$$

- **Clustering** (8 points)

We look at conditions where K-means performs poorly. Consider a dataset of 4 points in $\mathbb{R}^2$,

$$\{\boldsymbol{x}_1 = (-5, 1), \boldsymbol{x}_2 = (5, 1), \boldsymbol{x}_3 = (-5, -1), \boldsymbol{x}_4 = (5, -1)\}$$

.



Assume that we are looking to group the points into two clusters ($K = 2$). As per usual, denote $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ as the representatives of the two clusters, and

$$r_{nk} = \begin{cases} 1 & \boldsymbol{x}_n \text{ assigned to cluster } k \\ 0 & \text{else} \end{cases}$$

1. (2 points) Describe the optimal choice of representatives r_{nk}, and assignments $\boldsymbol{\mu}_k$ of data points to clusters, such that the least squares error

$$L = \sum_{n=1}^{4} \sum_{k=1}^{2} r_{nk} ||\boldsymbol{x}_n - \boldsymbol{\mu}_k||_2^2$$

is minimised.

Solution.
Choose $r_{1,1} = r_{3,1} = 1$ and $r_{2,2} = r_{4,2} = 1$. All other r_{nk} terms are zero. (That is, assign $\mathbf{x}_1$ and $\mathbf{x}_3$ to cluster 1, and $\mathbf{x}_2$ and $\mathbf{x}_4$ to cluster 2.) (The loss function for this clustering evaluates to 4, as each data point is distance 1 away from it's representative.)

2. (4 points) There are two potential situations where K-means is guaranteed to NOT converge to the optimal configuration above in question 1. For each situation, find a set of initial starting values $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ that can create this situation.

   Situation 1: There is one empty cluster.

   Solution.
   Choose $\boldsymbol{\mu}_1 = (0, 0)$ and $\boldsymbol{\mu}_2 = (1000, 0)$. Clearly, all points are closer to $\boldsymbol{\mu}_1$ and are assigned to it. The other representative receives no points. $\boldsymbol{\mu}_1$ is already at the centroid of the four points (by symmetry) and hence doesn't move. $\boldsymbol{\mu}_2$ doesn't have any points to move towards. K-means terminates with all points in one cluster.

   Situation 2: There is no empty cluster, but the clustering result is not the optimal choice.

   Solution.
   Choose $\boldsymbol{\mu}_1 = (0, 1)$ and $\boldsymbol{\mu}_2 = (0, -1)$. $\mathbf{x}_1$ and $\mathbf{x}_2$ are assigned to $\boldsymbol{\mu}_1$, and $\mathbf{x}_3$ and $\mathbf{x}_4$ are assigned to $\boldsymbol{\mu}_2$. The representatives are already at their centroids, and do not move. Hence K-means terminates, and the loss function is

$$L = \|\mathbf{x}_1 - \boldsymbol{\mu}_1\|_2^2 + \|\mathbf{x}_2 - \boldsymbol{\mu}_1\|_2^2 + \|\mathbf{x}_3 - \boldsymbol{\mu}_2\|_2^2 + \|\mathbf{x}_4 - \boldsymbol{\mu}_2\|_2^2 = 4 \times 5^2 = 100 > 4$$

   which is suboptimal.

3. (2 points) In this specific example, show that using agglomerative clustering ($K = 2$) on this data set converges to an optimal clustering. Note: you can define a tie breaker. For example, given the choice between merging μ_i with μ_j, and merging μ_k with μ_l, let $N = \min(i, j, k, l)$ and then merge the cluster that contains μ_N.

   Solution.

   We merge either $\mathbf{x}_1$ with $\mathbf{x}_3$ or $\mathbf{x}_2$ with $\mathbf{x}_4$ first, as they are the closest sets of points. By the tie breaker strategy, merge $\mathbf{x}_1$ with $\mathbf{x}_3$. We then merge $\mathbf{x}_2$ with $\mathbf{x}_4$, as they are within distance 2, which is closer than the distance of 10+ from the centroid of the other cluster. We have two clusters, and terminate in the optimal clustering condition.

- **Vector Calculus** (17 points)

  Compute the derivatives of the following functions. Show intermediate steps, and the dimension of the result.

  1 (4 points)

  $f, g: \mathbb{R}^3 \to \mathbb{R}^3$. Let $f(x,y,z) = (u(x,y,z), v(x,y,z), w(x,y,x))$,
  $g(u,v,w) = (u-v, u+w, w+v)$, and let $h = g \circ f$.

  Calculate the gradient of h with respect to x, y, z

  Solution. *So the above question is very likely a typo, and should probably say $w(x,y,z)$ rather than $w(x,y,x)$. Answering either will receive full credit.*

$$\frac{d\mathbf{h}(x)}{dx} = \frac{d\mathbf{g}(\mathbf{f})}{d\mathbf{f}} \frac{d\mathbf{f}(x)}{dx}$$

$$\frac{d\mathbf{g}(\mathbf{f})}{d\mathbf{f}} = \begin{bmatrix} \frac{\partial g_1(\mathbf{f})}{\partial f_1} & \frac{\partial g_1(\mathbf{f})}{\partial f_2} & \frac{\partial g_1(\mathbf{f})}{\partial f_3} \\ \frac{\partial g_2(\mathbf{f})}{\partial f_1} & \frac{\partial g_2(\mathbf{f})}{\partial f_2} & \frac{\partial g_2(\mathbf{f})}{\partial f_3} \\ \frac{\partial g_3(\mathbf{f})}{\partial f_1} & \frac{\partial g_3(\mathbf{f})}{\partial f_2} & \frac{\partial g_3(\mathbf{f})}{\partial f_3} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial u}(u-v) & \frac{\partial}{\partial v}(u-v) & \frac{\partial}{\partial w}(u-v) \\ \frac{\partial}{\partial u}(u+w) & \frac{\partial}{\partial v}(u+w) & \frac{\partial}{\partial w}(u+w) \\ \frac{\partial}{\partial u}(w+v) & \frac{\partial}{\partial v}(w+v) & \frac{\partial}{\partial w}(w+v) \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

$$\frac{d\mathbf{f}}{dx} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x} \\ \frac{\partial f_2(x)}{\partial x} \\ \frac{\partial f_3(x)}{\partial x} \end{bmatrix} = \begin{bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial x} \\ \frac{\partial w}{\partial x} \end{bmatrix}$$

$$\frac{d\mathbf{h}(x)}{dx} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial x} \\ \frac{\partial w}{\partial x} \end{bmatrix} = \begin{bmatrix} \frac{\partial u}{\partial x} - \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial x} + \frac{\partial w}{\partial x} \\ \frac{\partial v}{\partial x} + \frac{\partial w}{\partial x} \end{bmatrix}$$

  If students answer the question using $w(x,y,x)$ rather than $w(x,y,z)$, then the answer is the same, but with $\frac{\partial w}{\partial x} = 0$. The gradient $\frac{dh}{dy}$ and $\frac{dh}{dz}$ is the same as $\frac{dh}{dz}$, but replace x with y and z respectively. So,

$$\nabla_{x,y,z}\mathbf{h} = \begin{bmatrix} \frac{\partial u}{\partial x} - \frac{\partial v}{\partial x} & \frac{\partial u}{\partial y} - \frac{\partial v}{\partial y} & \frac{\partial u}{\partial z} - \frac{\partial v}{\partial z} \\ \frac{\partial u}{\partial x} + \frac{\partial w}{\partial x} & \frac{\partial u}{\partial y} + \frac{\partial w}{\partial y} & \frac{\partial u}{\partial z} + \frac{\partial w}{\partial z} \\ \frac{\partial v}{\partial x} + \frac{\partial w}{\partial x} & \frac{\partial v}{\partial y} + \frac{\partial w}{\partial y} & \frac{\partial v}{\partial z} + \frac{\partial w}{\partial z} \end{bmatrix}$$

2 (5 points)

$$f, g: \mathbb{R}^n \to \mathbb{R}, \quad f(\boldsymbol{x}) = \boldsymbol{c}^T\boldsymbol{x}, \quad \boldsymbol{c} \in \mathbb{R}^n, \quad g(\boldsymbol{x}) = \sqrt{\boldsymbol{c}^T\boldsymbol{x} + \mu^2}, \quad \mu \in \mathbb{R}.$$

- a) (3 points) Prove $\frac{\mathrm{d}f(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} = \boldsymbol{c}^T$.
  Solution. Take the partial derivative with respect to one of the components.

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = \frac{\partial}{\partial x_i}\sum_{j=1}^{n} c_j x_j = \sum_{j=1}^{n} c_j \frac{\partial x_j}{\partial x_i} = c_i$$

  Hence,

$$\frac{df(\mathbf{x})}{d\mathbf{x}} = \mathbf{c}^T$$

- b) (2 points) Calculate $\frac{\mathrm{d}g}{\mathrm{d}\boldsymbol{x}}$.
  Solution. Use chain rule.

$$\frac{d\sqrt{\boldsymbol{c}^T\boldsymbol{x} + \mu^2}}{d\mathbf{x}} = \frac{1}{2\sqrt{\boldsymbol{c}^T\boldsymbol{x} + \mu^2}}\frac{d}{d\mathbf{x}}(\boldsymbol{c}^T\boldsymbol{x} + \mu^2) = \frac{1}{2\sqrt{\boldsymbol{c}^T\boldsymbol{x} + \mu^2}}\mathbf{c}^T$$

3 (3 points)

$f:\ \mathbb{R}^n \to \mathbb{R}, f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}, \quad \boldsymbol{B} \in \mathbb{R}^{n\times n}$

Prove $\frac{\mathrm{d}f(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} = \boldsymbol{x}^T(\boldsymbol{B}+\boldsymbol{B}^T)$

Solution. Compute each component of the derivative.

$$\begin{aligned}
\frac{\partial}{\partial x_p}(\mathbf{x}^T\mathbf{B}\mathbf{x}) &= \frac{\partial}{\partial x_p}\sum_k x_k(\mathbf{B}\mathbf{x})_k \\
&= \frac{\partial}{\partial x_p}\sum_k x_k \sum_j B_{kj}x_j \\
&= \sum_{j,k} B_{kj}\frac{\partial(x_k x_j)}{\partial x_p}
\end{aligned}$$

Note the following:

$$\frac{\partial(x_k x_j)}{\partial x_p} = \begin{cases} 2x_p & p = k = j \\ x_k & p = j \neq k \\ x_j & p = k \neq j \\ 0 & p \neq k, p \neq j \end{cases}$$

Hence, we can split the sum above,

$$\begin{aligned}
\frac{\partial}{\partial x_p}(\mathbf{x}^T\mathbf{B}\mathbf{x}) &= \sum_{\substack{j,k \\ p=k=j}} B_{kj}\frac{\partial(x_k x_j)}{\partial x_p} + \sum_{\substack{j,k \\ p=j\neq k}} B_{kj}\frac{\partial(x_k x_j)}{\partial x_p} \\
&+ \sum_{\substack{j,k \\ p=k\neq j}} B_{kj}\frac{\partial(x_k x_j)}{\partial x_p} + \sum_{\substack{j,k \\ p\neq k, p\neq j}} B_{kj}\frac{\partial(x_k x_j)}{\partial x_p} \\
&= B_{pp}2x_p + \sum_{\substack{k \\ p\neq k}} B_{kp}x_k + \sum_{\substack{j \\ p\neq j}} B_{pj}x_j
\end{aligned}$$

Add the $B_{pp}x_p$ terms back into each summation,

$$\begin{aligned}
&= \sum_k x_k B_{kp} + \sum_j x_j B_{pj} \\
&= (\mathbf{x}^T\mathbf{B})_p + \sum_j x_j(\mathbf{B}^T)_{jp} \\
&= (\mathbf{x}^T\mathbf{B})_p + (\mathbf{x}^T\mathbf{B}^T)_p \\
&= \mathbf{x}^T(\mathbf{B}+\mathbf{B}^T)_p
\end{aligned}$$

Hence,

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^T\mathbf{B}\mathbf{x}) = \mathbf{x}^T(\mathbf{B}+\mathbf{B}^T)$$

4 (5 points) Given a system of linear equations $\boldsymbol{Ax} = \boldsymbol{b}$, with $\boldsymbol{A} \in \mathbb{R}^{k\times n}, \boldsymbol{x} \in \mathbb{R}^{n\times 1}, \boldsymbol{b} \in \mathbb{R}^{k\times 1}$, sometimes there exists no solutions $\boldsymbol{x}$. So we'd like to find a approximate solution $\boldsymbol{Ax} \approx \boldsymbol{b}$. To achieve this, we formulate the following regularized least squares error

$$\ell(\boldsymbol{x}) = \|\boldsymbol{Ax} - \boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2 \,, \text{where } \lambda \in$$

Show that the gradient of the regularized least squares error above is given by

$$\frac{\mathrm{d}\ell(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} = 2(\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{b}^T\boldsymbol{A}) + 2\lambda\boldsymbol{x}^T$$

(Hint: you can directly use the conclusions from questions 2 and 3 above, together with the definition of the Euclidean norm.)

Solution.

$$\begin{aligned}
\ell(\boldsymbol{x}) &= \|\boldsymbol{Ax} - \boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2 \\
&= (\boldsymbol{Ax} - \boldsymbol{b})^T(\boldsymbol{Ax} - \boldsymbol{b}) + \lambda\boldsymbol{x}^T\boldsymbol{x} \\
&= (\boldsymbol{x}^T\boldsymbol{A}^T - \boldsymbol{b}^T)(\boldsymbol{Ax} - \boldsymbol{b}) + \lambda\boldsymbol{x}^T\boldsymbol{x} \\
&= \boldsymbol{x}^T(\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{b} - \boldsymbol{b}^T\boldsymbol{Ax} + \boldsymbol{b}^T\boldsymbol{b} + \lambda\boldsymbol{x}^T\boldsymbol{x} \\
&= \boldsymbol{x}^T(\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{x} - 2\boldsymbol{b}^T\boldsymbol{Ax} + \boldsymbol{b}^T\boldsymbol{b} + \lambda\boldsymbol{x}^T\boldsymbol{x}
\end{aligned}$$

So, by taking the derivative, and using the derivations above, together with the identity $\frac{d\mathbf{x}^T\mathbf{x}}{d\mathbf{x}} = 2\mathbf{x}^T$, and the fact that $\mathbf{b}^T\mathbf{b}$ has no dependence on $\mathbf{x}$, we obtain

$$\begin{aligned}
\frac{d\ell(\boldsymbol{x})}{d\boldsymbol{x}} &= \boldsymbol{x}^T(\boldsymbol{A}^T\boldsymbol{A} + (\boldsymbol{A}^T\boldsymbol{A})^T) - 2\boldsymbol{b}^T\boldsymbol{A} + 2\lambda\boldsymbol{x}^T \\
&= 2\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A} - 2\boldsymbol{b}^T\boldsymbol{A} + 2\lambda\boldsymbol{x}^T \\
&= 2(\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{b}^T\boldsymbol{A}) + 2\lambda\boldsymbol{x}^T
\end{aligned}$$

as required. Once could also use chain rule instead of expanding first.

——— End of the paper ———