EE4-66 Topics in Large Dimensional Data Processing

Instructions for Candidates

Answer all three questions. Each question carries 25 marks.

1. (Matrix Analysis)

Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be a matrix.

(a) State the definition of the mutual coherence constance $\mu(\boldsymbol{A})$ of the matrix $\boldsymbol{A}$. [2]

(b)

i State the definitions of Restricted Isometry Property (RIP) and Restricted Isometry Constant (RIC) respectively using squared ℓ_2-norm. [3]

ii State the equivalent definitions of RIP and RIC using the singular values of the relevant matrices. [3]

(c) RIP implies the near orthogonality of two disjoint submatrices of $\boldsymbol{A}$. Specifically, assume that the matrix $\boldsymbol{A}$ satisfies the RIP. Let $\mathcal{I}, \mathcal{J} \subset \{1, \cdots, n\}$. Assume that $|\mathcal{I}| = |\mathcal{J}| = k$ and $\mathcal{I} \bigcap \mathcal{J} = \phi$. RIP implies that for all $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^k$,

$$|\langle \boldsymbol{A}_{\mathcal{I}}\boldsymbol{a}, \boldsymbol{A}_{\mathcal{J}}\boldsymbol{b}\rangle| \leq c \|\boldsymbol{a}\|_2 \|\boldsymbol{b}\|_2, \tag{1.1}$$

for some constant c. Answer the following sub-questions in order to find out the value of the constant c (in terms of RIC).

i Define $\boldsymbol{a}' = \boldsymbol{a}/\|\boldsymbol{a}\|_2$ and $\boldsymbol{b}' = \boldsymbol{b}/\|\boldsymbol{b}\|_2$. Compute the squared ℓ_2-norm of the vectors $\begin{bmatrix} \boldsymbol{a}' \\ \boldsymbol{b}' \end{bmatrix}$ and $\begin{bmatrix} \boldsymbol{a}' \\ -\boldsymbol{b}' \end{bmatrix}$. [2]

ii Define $\boldsymbol{x}' = \boldsymbol{A}_{\mathcal{I}}\boldsymbol{a}'$ and $\boldsymbol{y}' = \boldsymbol{A}_{\mathcal{J}}\boldsymbol{b}'$. Find the lower and upper bounds of $\|\boldsymbol{x}' + \boldsymbol{y}'\|_2^2$ and $\|\boldsymbol{x}' - \boldsymbol{y}'\|_2^2$ using RIC of the matrix $\boldsymbol{A}$. [3]

iii Using the results of the previous sub-question to derive the constant c in (1.1) in terms of RIC. [3]

iv Based on (1.1), it holds that

$$\left\|\boldsymbol{A}_{\mathcal{I}}^T \boldsymbol{A}_{\mathcal{J}}\boldsymbol{b}\right\|_2 \leq c \|\boldsymbol{b}\|_2 \tag{1.2}$$

with the same constant c. Prove this result. [2]

(d) Let $\boldsymbol{x} \in \mathbb{R}^n$ be a k-sparse vector with support set $\operatorname{supp}(\boldsymbol{x}) = \mathcal{I}$. Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$.
Remark: For simplicity you can directly use the fact that $\delta_k \leq \delta_{2k} < 1$.

i Establish an upper bound on $\left\|\boldsymbol{A}_{\mathcal{J}}^T \boldsymbol{y}\right\|_2$ using RIC δ_{2k}. [2]

ii Establish a lower bound on $\left\|\boldsymbol{A}_{\mathcal{I}}^T \boldsymbol{y}\right\|_2$ using RIC δ_{2k}. [2]

iii Find the specific range of δ_{2k} so that $\left\|\boldsymbol{A}_{\mathcal{I}}^{T}\boldsymbol{y}\right\|_2 \geq \left\|\boldsymbol{A}_{\mathcal{J}}^{T}\boldsymbol{y}\right\|_2$. [1]

iv Let $\boldsymbol{A}_{\mathcal{J}}^{\dagger}$ and $\boldsymbol{A}_{\mathcal{I}}^{\dagger}$ be the pseudo-inverse of the matrices $\boldsymbol{A}_{\mathcal{J}}$ and $\boldsymbol{A}_{\mathcal{I}}$, respectively. Establish an upper bound on $\left\|\boldsymbol{A}_{\mathcal{J}}^{\dagger}\boldsymbol{y}\right\|_2$ using RIC. Find the specific range of RIC δ_{2k} so that $\left\|\boldsymbol{A}_{\mathcal{I}}^{\dagger}\boldsymbol{y}\right\|_2 \geq \left\|\boldsymbol{A}_{\mathcal{J}}^{\dagger}\boldsymbol{y}\right\|_2$. [2]

(Total marks: 25)

2. (Convex Optimisation Basics)

(a)

i State the definition of a convex set $\mathcal{S} \subset \mathbb{R}^n$. [2]

ii State the definition of a convex function $f: \mathbb{R}^n \to \mathbb{R}$. [2]

(b)

i Consider a differentiable function $f: \mathbb{R}^n \to \mathbb{R}$. Denote its gradient at $\boldsymbol{x} \in \mathbb{R}^n$ by $\nabla f(\boldsymbol{x})$. For given $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the directional gradient of f along the direction $\boldsymbol{y} - \boldsymbol{x}$ is defined as

$$\nabla_{\boldsymbol{y}-\boldsymbol{x}} f(\boldsymbol{x}) = \lim_{\lambda \to 0} \frac{f(\boldsymbol{x} + \lambda(\boldsymbol{y} - \boldsymbol{x})) - f(\boldsymbol{x})}{\lambda}.$$

State the dimension of $\nabla f(\boldsymbol{x})$ and $\nabla_{\boldsymbol{y}-\boldsymbol{x}} f(\boldsymbol{x})$. State how to compute the directional gradient $\nabla_{\boldsymbol{y}-\boldsymbol{x}} f(\boldsymbol{x})$ using the gradient $\nabla f(\boldsymbol{x})$. [2]

ii Assume that a convex function $f: \mathbb{R}^n \to \mathbb{R}$ is differentiable. Prove that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, it holds that

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}). \quad (2.3)$$

[2]

Hint: Apply the definition of convex function to $f((1-\lambda)\boldsymbol{x} + \lambda \boldsymbol{y})$, $\lambda \in [0, 1]$.

iii Assume that a convex function $f: \mathbb{R}^n \to \mathbb{R}$ is *not* differentiable at a point $\boldsymbol{x} \in \mathbb{R}^n$. State the definition of the subgradient at the point $\boldsymbol{x}$. Make the dimension of the subgradient explicit in your answer. [2]

iv The set of subgradients at $\boldsymbol{x}$ is called the subdifferential at $\boldsymbol{x}$ and is denoted by $\partial f(\boldsymbol{x})$. Find the subdifferential of $f(x) = |x|$ for all $x \in \mathbb{R}$. [3]

(c)

i Assume that a convex function $f: \mathbb{R}^n \to \mathbb{R}$ is differentiable. Prove that $\boldsymbol{x}$ is a global minimiser of f if $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$. [2]
Hint: You are allowed to use the result in (2.3) directly.

ii Assume that a convex function $f: \mathbb{R}^n \to \mathbb{R}$ is *not* differentiable at a point $\boldsymbol{x} \in \mathbb{R}^n$. Prove that $\boldsymbol{x}$ is a global minimiser of f if $\boldsymbol{0} \in \partial f(\boldsymbol{x})$. [2]

iii The soft thresholding function $\eta(\cdot)$ is designed to give a global minimiser of the simplified Lasso problem

$$\min_{\boldsymbol{x}} \frac{1}{2} \left\| \boldsymbol{x} - \boldsymbol{z} \right\|_2^2 + \lambda \left\| \boldsymbol{x} \right\|_1 . \tag{2.4}$$

State the form of the soft thresholding function $\eta(\cdot)$. (Derivations are not required.) [2]

iv The famous Lasso formulation for sparse recovery is given by

$$\min_{\boldsymbol{x}} \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \right\|_2^2 + \lambda \left\| \boldsymbol{x} \right\|_1 , \tag{2.5}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is a given matrix and $\lambda > 0$ is a given parameter. State the Iterative Shrinkage Thresholding (IST) algorithm to solve the Lasso problem (2.5). (Derivations are not required.) [3]

v Consider the low-rank matrix recovery problem $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X})$, where $\mathcal{A} : \mathbb{R}^{n_r \times n_c} \to \mathbb{R}^m$ is a linear operator. State the counterpart of the IST algorithm designed to solve the low-rank matrix recovery problem. Give the definition of corresponding soft thresholding function used in your algorithm. [3]

(Total marks: 25)

3. (Convex Optimisation)

(a)

i State the standard form of a convex optimisation problem (with equality and inequality constraints). [2]

ii Let u_i be the Lagrange multipliers of the inequality constraints and v_j be the Lagrange multipliers of the equality constraints, respectively. State the corresponding Lagrangian. [2]

iii State the corresponding Lagrange dual function and Lagrange dual problem. [2]

iv State the Karush-Kuhn-Tucker (KKT) conditions for a global optimum. [4]

(b) Alternating direction method of multipliers (ADMM) solves optimisation problems in the form

$$\begin{aligned} &\min_{\boldsymbol{x},\boldsymbol{z}} f(\boldsymbol{x}) + g(\boldsymbol{z}) \\ &\text{subject to } \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} = \boldsymbol{c}. \end{aligned} \tag{3.6}$$

Consider the equivalent problem

$$\begin{aligned} &\min_{\boldsymbol{x},\boldsymbol{z}} f(\boldsymbol{x}) + g(\boldsymbol{z}) + \frac{\rho}{2}\left\|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c}\right\|_2^2 \\ &\text{subject to } \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} = \boldsymbol{c}, \end{aligned} \tag{3.7}$$

where $\rho > 0$ is a constant.

i State the corresponding Lagrangian of (3.7), denoted by $L_\rho(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v})$. [1]

ii ADMM is an iterative algorithm with three steps in each iteration:

- Update $\boldsymbol{x}$ to obtain $\boldsymbol{x}^{k+1}$.
- Update $\boldsymbol{z}$ to obtain $\boldsymbol{z}^{k+1}$.
- Update $\boldsymbol{v}$ to obtain $\boldsymbol{v}^{k+1}$ via

$$\boldsymbol{v}^{k+1} = \boldsymbol{v}^k + \rho\left(\boldsymbol{A}\boldsymbol{x}^{k+1} + \boldsymbol{B}\boldsymbol{z}^{k+1} - \boldsymbol{c}\right).$$

State the details of the first two steps. [4]

iii The form of ADMM iterations can be highly simplified by introducing $\boldsymbol{w} = \frac{1}{\rho}\boldsymbol{v}$ (i.e. $\boldsymbol{v} = \rho\boldsymbol{w}$). Rewrite the details of the three steps in each iteration by replacing $\boldsymbol{v}$ with $\boldsymbol{w}$. [2]

In the literature, the simplified form is called the scaled form of ADMM.

(c) In the following, we are applying the scaled form of ADMM to solve two non-smooth convex optimisation problems.

i To solve the Lasso problem

$$\min_{\boldsymbol{x}} \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \right\|_2^2 + \lambda \left\| \boldsymbol{x} \right\|_1 ,$$

state the standard ADMM problem formulation in the form of (3.6) and the scaled form of ADMM iterations. [3]

ii To solve the constrained Lasso problem

$$\min_{\boldsymbol{x}} \frac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \right\|_2^2 + \lambda \left\| \boldsymbol{x} \right\|_1$$
$$\text{subject to } \boldsymbol{B}\boldsymbol{x} \leq \boldsymbol{0},$$

state the standard ADMM problem formulation in the form of (3.6) and the scaled form of ADMM iterations. [5]

(Total marks: 25)

Solution to Problem 1.

(a) The mutual coherence is defined as

$$\mu(\boldsymbol{A}) = \max_{i \neq j} \frac{|\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle|}{\|\boldsymbol{a}_i\|_2 \|\boldsymbol{a}_j\|_2} = \max_{i \neq j} \frac{\left|\boldsymbol{a}_i^T \boldsymbol{a}_j\right|}{\|\boldsymbol{a}_i\|_2 \|\boldsymbol{a}_j\|_2}.$$

If the columns of the matrix $\boldsymbol{A}$ are normalised (in ℓ_2-norm), then the above definition is simplified to

$$\mu(\boldsymbol{A}) = \max_{i \neq j} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle| = \max_{i \neq j} \left|\boldsymbol{a}_i^T \boldsymbol{a}_j\right|.$$

Henceforth, we assume a column-normalised matrix $\boldsymbol{A}$ below for composition simplicity. [2]

(b)

i The matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is said to satisfy the RIP with parameters (K, δ), if for all $\mathcal{T} \subset \{1, \cdots, n\}$ such that $|\mathcal{T}| \leq K$ and for all $\boldsymbol{q} \in \mathbb{R}^{|\mathcal{T}|}$, it holds that

$$(1-\delta)\|\boldsymbol{q}\|_2^2 \leq \|\boldsymbol{A}_{\mathcal{T}} \boldsymbol{q}\|_2^2 \leq (1+\delta)\|\boldsymbol{q}\|_2^2.$$

The RIC δ_K is defined as the smallest constant δ for which the (K, δ)-RIP holds, i.e.,

$$\begin{aligned}\delta_K = \inf \{ \delta : \ & (1-\delta)\|\boldsymbol{q}\|_2^2 \leq \|\boldsymbol{A}_{\mathcal{T}} \boldsymbol{q}\|_2^2 \leq (1+\delta)\|\boldsymbol{q}\|_2^2 \\ & \forall |\mathcal{T}| \leq K, \ \forall \boldsymbol{q} \in \mathbb{R}^{|\mathcal{T}|} \}.\end{aligned}$$

[3]

Remark. Some students forgot to mention the properties in the definition should hold for all $|\mathcal{T}| \leq K$ and all $\boldsymbol{q} \in \mathbb{R}^{|\mathcal{T}|}$.

ii The matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is said to satisfy the RIP with RIC $\delta_K \in (0,1)$, if for all $\mathcal{T} \subset \{1, \cdots, n\}$ such that $|\mathcal{T}| \leq K$, it holds that

$$\sqrt{1-\delta_K} \leq \sigma_{\min}(\boldsymbol{A}_{\mathcal{T}}) \leq \sigma_{\max}(\boldsymbol{A}_{\mathcal{T}}) \leq \sqrt{1+\delta_K}.$$

[3]

(c)

i It is clear that

$$\left\| \begin{bmatrix} \boldsymbol{a}' \\ \boldsymbol{b}' \end{bmatrix} \right\|_2^2 = \sum_i (a_i')^2 + \sum_j \left(b_j'\right)^2$$
$$= \|\boldsymbol{a}'\|_2^2 + \|\boldsymbol{b}'\|_2^2 = 2.$$

Similarly,

$$\left\| \begin{bmatrix} \boldsymbol{a}' \\ -\boldsymbol{b}' \end{bmatrix} \right\|_2^2 = 2.$$

[2]

ii As

$$\boldsymbol{x}' + \boldsymbol{y}' = [\boldsymbol{A}_{\mathcal{I}} \boldsymbol{A}_{\mathcal{J}}] \begin{bmatrix} \boldsymbol{a}' \\ \boldsymbol{b}' \end{bmatrix},$$

RIP implies

$$2\,(1 - \delta_{2k}) \le \|\boldsymbol{x}' + \boldsymbol{y}'\|_2^2 \le 2\,(1 + \delta_{2k})\,.$$

Similarly, it is also true that

$$2\,(1 - \delta_{2k}) \le \|\boldsymbol{x}' - \boldsymbol{y}'\|_2^2 \le 2\,(1 + \delta_{2k})\,.$$

[3]

iii It holds that

$$\langle \boldsymbol{x}', \boldsymbol{y}' \rangle \le \frac{\|\boldsymbol{x}' + \boldsymbol{y}'\|_2^2 - \|\boldsymbol{x}' - \boldsymbol{y}'\|_2^2}{4} \le \delta_{2k},$$

and

$$-\langle \boldsymbol{x}', \boldsymbol{y}' \rangle \le \frac{\|\boldsymbol{x}' - \boldsymbol{y}'\|_2^2 - \|\boldsymbol{x}' + \boldsymbol{y}'\|_2^2}{4} \le \delta_{2k}.$$

As a consequence,

$$|\langle \boldsymbol{x}', \boldsymbol{y}' \rangle| \le \delta_{2k},$$

and

$$|\langle \boldsymbol{A}_{\mathcal{I}} \boldsymbol{a}, \boldsymbol{A}_{\mathcal{J}} \boldsymbol{b} \rangle| \le \delta_{2k} \|\boldsymbol{a}\|_2 \|\boldsymbol{b}\|_2\,.$$

It concludes that

$$c = \delta_{2k}.$$

[3]

iv

$$
\begin{aligned}
\left\| A_{\mathcal{I}}^T A_{\mathcal{J}} b \right\|_2 &= \sup_{a:\, \|a\|_2 = 1} \left| a^T \left(A_{\mathcal{I}}^T A_{\mathcal{J}} b \right) \right| \\
&= \sup_{a:\, \|a\|_2 = 1} \left| \langle A_{\mathcal{I}} a, A_{\mathcal{J}} b \rangle \right| \\
&\le \delta_{2k} \|b\|_2 .
\end{aligned}
$$

[2]

Remark. Some students do not realise that the sup operator in the first line is necessary.

(d)

i

$$
\left\| A_{\mathcal{J}}^T y \right\|_2 = \left\| A_{\mathcal{J}}^T A_{\mathcal{I}} x_{\mathcal{I}} \right\|_2 \le \delta_{2k} \|x\|_2 .
$$

[2]

ii

$$
\begin{aligned}
\left\| A_{\mathcal{I}}^T y \right\|_2 &= \left\| A_{\mathcal{I}}^T A_{\mathcal{I}} x_{\mathcal{I}} \right\|_2 \\
&\ge (1 - \delta_k) \|x\|_2 \ge (1 - \delta_{2k}) \|x\|_2 .
\end{aligned}
$$

[2]

iii In order to make $\left\| A_{\mathcal{I}}^T y \right\|_2 \ge \left\| A_{\mathcal{J}}^T y \right\|_2$, it is sufficient to have $1 - \delta_{2k} \ge \delta_{2k}$, i.e., $\delta_{2k} \le \frac{1}{2}$. [1]

iv

$$
\begin{aligned}
\left\| A_{\mathcal{J}}^{\dagger} y \right\|_2 &= \left\| \left(A_{\mathcal{J}}^T A_{\mathcal{J}} \right)^{-1} A_{\mathcal{J}}^T A_{\mathcal{I}} x_{\mathcal{I}} \right\|_2 \\
&\le \frac{1}{1 - \delta_{2k}} \left\| A_{\mathcal{J}}^T A_{\mathcal{I}} x_{\mathcal{I}} \right\|_2 \\
&\le \frac{\delta_{2k}}{1 - \delta_{2k}} \|x\|_2 .
\end{aligned}
$$

At the same time,

$$
\begin{aligned}
\left\| A_{\mathcal{I}}^{\dagger} y \right\|_2 &= \left\| \left(A_{\mathcal{I}}^T A_{\mathcal{I}} \right)^{-1} A_{\mathcal{I}}^T A_{\mathcal{I}} x_{\mathcal{I}} \right\|_2 \\
&= \|x_{\mathcal{I}}\|_2 = \|x\|_2 .
\end{aligned}
$$

In order to make $\left\| A_{\mathcal{I}}^{\dagger} y \right\|_2 \ge \left\| A_{\mathcal{J}}^{\dagger} y \right\|_2$, it is sufficient to have $1 \ge \frac{\delta_{2k}}{1 - \delta_{2k}}$, i.e., $\delta_{2k} \leq \frac{1}{2}$. [2]

Remark. This sub-question is testing an application of RIP property. Such an application has been used in some of the proofs in the lecture notes but was not emphasised in the lectures. As a consequence, this sub-question appears difficult for many students.

Solution to Problem 2.

(a)

i A set $\mathcal{S} \subset \mathbb{R}^n$ is convex if

$$\lambda \boldsymbol{x} + (1-\lambda)\, \boldsymbol{y} \in \mathcal{S},$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ and all $\lambda \in [0,1]$. [2]

ii A function $f:\ \mathbb{R}^n \to \mathbb{R}$ is convex if

$$f\left(\lambda \boldsymbol{x} + (1-\lambda)\, \boldsymbol{y}\right) \le \lambda f\left(\boldsymbol{x}\right) + (1-\lambda)\, f\left(\boldsymbol{y}\right)$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and all $\lambda \in [0,1]$. [2]

(b)

i It is clear that $\nabla f\left(\boldsymbol{x}\right) \in \mathbb{R}^n$, $\nabla_{\boldsymbol{y}-\boldsymbol{x}} f\left(\boldsymbol{x}\right) \in \mathbb{R}$, and

$$\nabla_{\boldsymbol{y}-\boldsymbol{x}} f\left(\boldsymbol{x}\right) = \nabla f\left(\boldsymbol{x}\right)^T \left(\boldsymbol{y} - \boldsymbol{x}\right).$$

[2]

ii By definition of convexity, for all $\lambda \in [0,1]$,

$$f\left((1-\lambda)\, \boldsymbol{x} + \lambda \boldsymbol{y}\right) \le (1-\lambda)\, f\left(\boldsymbol{x}\right) + \lambda f\left(\boldsymbol{y}\right).$$

This implies

$$\lambda f\left(\boldsymbol{y}\right) \ge \lambda f\left(\boldsymbol{x}\right) + f\left(\boldsymbol{x} + \lambda\left(\boldsymbol{y} - \boldsymbol{x}\right)\right) - f\left(\boldsymbol{x}\right),$$

or equivalently,

$$f\left(\boldsymbol{y}\right) \ge f\left(\boldsymbol{x}\right) + \frac{f\left(\boldsymbol{x} + \lambda\left(\boldsymbol{y} - \boldsymbol{x}\right)\right) - f\left(\boldsymbol{x}\right)}{\lambda}.$$

Take the limit $\lambda \to 0$. One has

$$f\left(\boldsymbol{y}\right) \ge f\left(\boldsymbol{x}\right) + \nabla f\left(\boldsymbol{x}\right)^T \left(\boldsymbol{y} - \boldsymbol{x}\right).$$

[2]

> *Remark.* Some students failed to answer this sub-question correctly. It appears to me that they may not realise the connection between gradient and the terms in the convexity definition.

iii A vector $\boldsymbol{v} \in \mathbb{R}^n$ is a subgradient of $f(\boldsymbol{x})$ at a point $\boldsymbol{x}$ if

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{v}^T (\boldsymbol{y} - \boldsymbol{x}) \text{ for all } \boldsymbol{y} \in \mathbb{R}^n.$$

[2]

iv For $f(x) = |x|$, its subdifferential is given by

$$\partial f = \begin{cases} 1 & \text{if } x > 0, \\ [-1, 1] & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

[3]

(c)

i As proved in (2.3), it holds that for all $\boldsymbol{y} \in \mathbb{R}^n$,

$$\begin{aligned} f(\boldsymbol{y}) &\geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) \\ &= f(\boldsymbol{x}), \end{aligned}$$

where the equality follows from the fact that $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$. Hence $\boldsymbol{x}$ is a global minimiser of f. [2]

ii By the definition of subgradient, it holds that for all $\boldsymbol{y} \in \mathbb{R}^n$,

$$\begin{aligned} f(\boldsymbol{y}) &\geq f(\boldsymbol{x}) + \boldsymbol{0}^T (\boldsymbol{y} - \boldsymbol{x}) \\ &= f(\boldsymbol{x}). \end{aligned}$$

Hence $\boldsymbol{x}$ is a global minimiser of f. [2]

iii The soft-thresholding function is of the form

$$x^* = \eta(z; \lambda) = \begin{cases} z - \lambda & \text{if } z \geq \lambda, \\ 0 & \text{if } -\lambda < z < \lambda, \\ z + \lambda & \text{if } z \leq -\lambda. \end{cases}$$

[2]

iv The IST algorithm is an iterative algorithm where in the k^{th} iteration the variable $\boldsymbol{x}^k$ is updated by

$$\boldsymbol{x}^k = \eta\left(\boldsymbol{x}^{k-1} + t_k \boldsymbol{A}^T\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{k-1}\right); \lambda t_k\right),$$

and $t_k > 0$ is an appropriately chosen step size. [3]

v Define the soft thresholding function $\eta_\sigma(\boldsymbol{Z};\lambda)$ as

$$\boldsymbol{X} = \eta_\sigma(\boldsymbol{Z};\lambda) = \sum_i \boldsymbol{u}_i \eta(\sigma_i;\lambda)\boldsymbol{v}_i^T,$$

where σ_i is the i^{th} singular value of $\boldsymbol{Z}$, $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are the corresponding singular vectors. The IST algorithm to solve the low-rank matrix recovery problem is an iterative algorithm where in the k^{th} iteration the matrix $\boldsymbol{X}^k$ is updated by

$$\boldsymbol{X}^k = \eta_\sigma\left(\boldsymbol{X}^{k-1} + t_k \mathcal{A}^*\left(\boldsymbol{y} - \mathcal{A}\left(\boldsymbol{X}^{k-1}\right)\right); \lambda t_k\right),$$

where $\mathcal{A}^*$ is the adjoint operator (transpose) of $\mathcal{A}$ and $t_k > 0$ is an appropriately chosen step size. [3]

Solution to Problem 3.

(a)

i The standard form of a convex optimisation problem is

$$\begin{aligned} &\min_{\boldsymbol{x}} f(\boldsymbol{x}) \\ &\text{subject to } h_i(\boldsymbol{x}) \le 0, \quad i \in \{1, \cdots, m\} \\ &\qquad\qquad \ell_j(\boldsymbol{x}) = 0, \quad j \in \{1, \cdots, p\} \end{aligned}$$

where $f(\boldsymbol{x})$, $h_i(\boldsymbol{x})$ are convex and ℓ_j is affine. [2]

Remark. Many students forgot to mention that $f(\boldsymbol{x})$, $h_i(\boldsymbol{x})$ are convex and ℓ_j is affine, and hence lost 1 point that could be easily save.

ii The corresponding Lagrangian is given by

$$L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} u_i h_i(\boldsymbol{x}) + \sum_{j=1}^{p} v_j \ell_j(\boldsymbol{x})$$

where $\boldsymbol{u} \in \mathbb{R}^m$, $\boldsymbol{v} \in \mathbb{R}^p$, and $\boldsymbol{u} \ge \boldsymbol{0}$. [2]

Remark. Again, many students forgot to mention that $\boldsymbol{u} \ge 0$, losing 1 point.

iii The Lagrange dual function is given as

$$g(\boldsymbol{u}, \boldsymbol{v}) = \min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}).$$

And the Lagrange dual problem is stated as

$$\begin{aligned} &\max_{\boldsymbol{u} \in \mathbb{R}^m,\, \boldsymbol{v} \in \mathbb{R}^p} g(\boldsymbol{u}, \boldsymbol{v}) \\ &\text{subject to } \boldsymbol{u} \ge \boldsymbol{0}. \end{aligned}$$

[2]

iv The KKT conditions are given as

- $\boldsymbol{0} \in \partial f(\boldsymbol{x}) + \sum_{i=1}^{m} u_i \partial h_i(\boldsymbol{x}) + \sum_{j=1}^{p} v_j \partial \ell_j(\boldsymbol{x})$ [1/4]
- $u_i h_i(\boldsymbol{x}) = 0$ for all $i \in \{1, \cdots, m\}$ [1/4]
- $h_i(\boldsymbol{x}) \le 0$, $\ell_j(\boldsymbol{x}) = 0$, and $u_i \ge 0$ for all $i \in \{1, \cdots, m\}$ and $j \in \{1, \cdots, p\}$. [2/4]

(b)

i

$$L_\rho(\boldsymbol{x},\boldsymbol{z},\boldsymbol{v}) = f(\boldsymbol{x}) + g(\boldsymbol{z}) + \frac{\rho}{2}\|\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{z}-\boldsymbol{c}\|_2^2 \\ + \boldsymbol{v}^T(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{z}-\boldsymbol{c}).$$

[1]

ii

- The variable $\boldsymbol{x}$ is updated via

$$\begin{aligned}\boldsymbol{x}^{k+1} &= \arg\min_{\boldsymbol{x}} L_\rho\left(\boldsymbol{x},\boldsymbol{z}^k,\boldsymbol{v}^k\right) \\ &= \arg\min_{\boldsymbol{x}} f(\boldsymbol{x}) + \frac{\rho}{2}\left\|\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{z}^k-\boldsymbol{c}\right\|_2^2 \\ &\quad + \left(\boldsymbol{v}^k\right)^T\left(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{z}^k-\boldsymbol{c}\right).\end{aligned}$$

[2/4]

- The variable $\boldsymbol{z}$ is updated via

$$\begin{aligned}\boldsymbol{z}^{k+1} &= \arg\min_{\boldsymbol{z}} L_\rho\left(\boldsymbol{x}^{k+1},\boldsymbol{z},\boldsymbol{v}^k\right) \\ &= \arg\min_{\boldsymbol{z}} g(\boldsymbol{z}) + \frac{\rho}{2}\left\|\boldsymbol{A}\boldsymbol{x}^{k+1}+\boldsymbol{B}\boldsymbol{z}-\boldsymbol{c}\right\|_2^2 \\ &\quad + \left(\boldsymbol{v}^k\right)^T\left(\boldsymbol{A}\boldsymbol{x}^{k+1}+\boldsymbol{B}\boldsymbol{z}-\boldsymbol{c}\right).\end{aligned}$$

[2/4]

iii By replacing $\boldsymbol{v}$ with $\boldsymbol{w} = \frac{1}{\rho}\boldsymbol{v}$, the three steps in each iteration of ADMM can be simplified to

$$\begin{aligned}\boldsymbol{x}^{k+1} &= \arg\min_{\boldsymbol{x}} f(\boldsymbol{x}) + \frac{\rho}{2}\left\|\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{z}^k-\boldsymbol{c}+\boldsymbol{w}^k\right\|_2^2 \\ \boldsymbol{z}^{k+1} &= \arg\min_{\boldsymbol{z}} g(\boldsymbol{z}) + \frac{\rho}{2}\left\|\boldsymbol{A}\boldsymbol{x}^{k+1}+\boldsymbol{B}\boldsymbol{z}-\boldsymbol{c}+\boldsymbol{w}^k\right\|_2^2 \\ \boldsymbol{w}^{k+1} &= \boldsymbol{w}^k + \boldsymbol{A}\boldsymbol{x}^{k+1}+\boldsymbol{B}\boldsymbol{z}^{k+1}-\boldsymbol{c}.\end{aligned}$$

[2]

(c)

i The ADMM formulation is given by

$$\min_{\boldsymbol{x},\boldsymbol{z}} \frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\right\|_2^2+\lambda\left\|\boldsymbol{z}\right\|_1$$
$$\text{subject to } \boldsymbol{x}-\boldsymbol{z}=\boldsymbol{0}.$$

[2/3]

The scaled form of ADMM iterations are given by

$$\begin{aligned}
\boldsymbol{x}^{k+1} &= \arg\min_{\boldsymbol{x}} \frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\right\|_2^2+\frac{\rho}{2}\left\|\boldsymbol{x}-\boldsymbol{z}^k+\boldsymbol{w}^k\right\|_2^2\\
\boldsymbol{z}^{k+1} &= \arg\min_{\boldsymbol{z}} \lambda\left\|\boldsymbol{z}\right\|_1+\frac{\rho}{2}\left\|\boldsymbol{x}^{k+1}-\boldsymbol{z}+\boldsymbol{w}^k\right\|_2^2\\
\boldsymbol{w}^{k+1} &= \boldsymbol{w}^k+\boldsymbol{x}^{k+1}-\boldsymbol{z}^{k+1}.
\end{aligned}$$

[1/3]

ii The ADMM formulation is given by

$$\min_{\boldsymbol{x},\boldsymbol{z}} \frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\right\|_2^2+g\left(\boldsymbol{z}\right)$$
$$\text{subject to } \underbrace{\begin{bmatrix}\boldsymbol{I}\\ \boldsymbol{B}\end{bmatrix}}_{\boldsymbol{B}'}\boldsymbol{x}+\underbrace{\begin{bmatrix}-\boldsymbol{I} & \\ & \boldsymbol{I}\end{bmatrix}}_{\boldsymbol{D}'}\underbrace{\begin{bmatrix}\boldsymbol{z}_1\\ \boldsymbol{z}_2\end{bmatrix}}_{\boldsymbol{z}}=\boldsymbol{0},$$

where

$$\begin{aligned}
g\left(\boldsymbol{z}\right) &= \lambda\left\|\boldsymbol{z}_1\right\|_1+\mathbb{1}_{\geq 0}\left(\boldsymbol{z}_2\right)\\
\mathbb{1}_{\geq 0}\left(z\right) &= \begin{cases}\infty & \text{if } z<0\\ 0 & \text{if } z\geq 0\end{cases}.
\end{aligned}$$

[3/5]

The scaled form of ADMM iterations are given by

$$\begin{aligned}
\boldsymbol{x}^{k+1} &= \arg\min_{\boldsymbol{x}} \frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\right\|_2^2+\frac{\rho}{2}\left\|\boldsymbol{B}'\boldsymbol{x}+\boldsymbol{D}'\boldsymbol{z}^k+\boldsymbol{w}^k\right\|_2^2\\
\boldsymbol{z}^{k+1} &= \arg\min_{\boldsymbol{z}_1,\boldsymbol{z}_2} \lambda\left\|\boldsymbol{z}_1\right\|_1+\frac{\rho}{2}\left\|\boldsymbol{x}^{k+1}-\boldsymbol{z}_1+\boldsymbol{w}_1^k\right\|_2^2\\
&\qquad\qquad +\mathbb{1}_{\geq 0}\left(\boldsymbol{z}_2\right)+\frac{\rho}{2}\left\|\boldsymbol{B}\boldsymbol{x}^{k+1}+\boldsymbol{z}_2+\boldsymbol{w}_2^k\right\|_2^2\\
\boldsymbol{w}^{k+1} &= \boldsymbol{w}^k+\boldsymbol{B}'\boldsymbol{x}^{k+1}+\boldsymbol{D}'\boldsymbol{z}^{k+1}.
\end{aligned}$$

[2/5]

Remark. This sub-question is perhaps the most difficult one. It requires an auxiliary variable $\boldsymbol{z}_2$ so that the inequality constraint can be easily handled. I am glad that a few students got the idea.