EE4-66
EE9-SO28

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2017

MSc and EEE PART IV: MEng and ACGI

Corrected copy

TOPICS IN LARGE DIMENSIONAL DATA PROCESSING

Tuesday, 9 May 10:00 am

Time allowed: 3:00 hours

There are FOUR questions on this paper.

Answer ALL questions.

All questions carry equal marks

Any special instructions for invigilators and information for candidates are on page 1.

Examiners responsible First Marker(s) : W. Dai
Second Marker(s) : C. Ling

EE4-66 Topics in Large Dimensional Data Processing

Instructions for Candidates

Answer all questions. Each question carries 20 marks.

1. (Linear Algebra and Restricted Isometry Property)

(a) Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be a matrix with rank $r < \min(m, n)$. Let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be its singular value decomposition (SVD) and $\boldsymbol{A} = \boldsymbol{U}_r\boldsymbol{\Sigma}_r\boldsymbol{V}_r^T$ be its ***compact*** SVD, where $\boldsymbol{\Sigma}_r$ is a diagonal matrix containing the nonzero singular values.

i Specify the dimension of the matrices $\boldsymbol{U}_r$, $\boldsymbol{\Sigma}_r$, and $\boldsymbol{V}_r$. [2]

ii Give the singular value decomposition of $\boldsymbol{A}^T\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^T$ using $\boldsymbol{U}_r$, $\boldsymbol{\Sigma}_r$, and $\boldsymbol{V}_r$. [2]

iii Let $\boldsymbol{q} \in \mathbb{R}^n$ be an arbitrary vector with $\|\boldsymbol{q}\|_2 = 1$. Prove that $\left\|\boldsymbol{V}^T\boldsymbol{q}\right\|_2 = 1$ and $\left\|\boldsymbol{V}_r^T\boldsymbol{q}\right\|_2 \leq 1$. [2]

iv Let $\boldsymbol{q} \in \mathbb{R}^n$ be an arbitrary vector with $\|\boldsymbol{q}\|_2 = 1$. Find upper and lower bounds on $\|\boldsymbol{A}\boldsymbol{q}\|_2$. Prove your results and explain the choices of $\boldsymbol{q}$ to achieve the upper and lower bounds respectively. [4]

v For any given vector $\boldsymbol{x} \in \mathbb{R}^m$, the projection of $\boldsymbol{x}$ on the subspace $\mathrm{span}(\boldsymbol{A})$ is given by $\mathrm{proj}(\boldsymbol{x}, \boldsymbol{A}) = \boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{x}$ where $\boldsymbol{A}^\dagger$ is the pseudoinverse of $\boldsymbol{A}$. Decompose the matrices $\boldsymbol{A}^\dagger$ and $\boldsymbol{A}\boldsymbol{A}^\dagger$ using the (compact) SVD of $\boldsymbol{A}$. [2]

vi Prove that $\boldsymbol{x}_{\mathrm{res}} = \mathrm{resid}(\boldsymbol{x}, \boldsymbol{A}) := \boldsymbol{x} - \mathrm{proj}(\boldsymbol{x}, \boldsymbol{A})$ is orthogonal to $\boldsymbol{A}$. [2]

(b) Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be a flat matrix, i.e., $m < n$. Let S be a positive integer such that $2S < m$. Suppose that $\boldsymbol{A}$ satisfies Restricted Isometry Property (RIP) with Restricted Isometry Constant (RIC) δ_{2S}.

i Give the definition of RIP (with RIC δ_{2S}) using singular values of submatrices. [2]

ii Suppose that $\mathcal{I}, \mathcal{J} \subset [n]$, $|\mathcal{I}| = |\mathcal{J}| = S$, and $\mathcal{I} \bigcap \mathcal{J} = \phi$. Let $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^S$ be arbitrary vectors with $\|\boldsymbol{p}\|_2 = \|\boldsymbol{q}\|_2 = 1$. Find lower and upper bounds on $\|\boldsymbol{A}_{\mathcal{I}}\boldsymbol{p} + \boldsymbol{A}_{\mathcal{J}}\boldsymbol{q}\|_2$ and $\|\boldsymbol{A}_{\mathcal{I}}\boldsymbol{p} - \boldsymbol{A}_{\mathcal{J}}\boldsymbol{q}\|_2$. [2]

iii Based on the above bounds, show that $\sigma_{\max}\left(\boldsymbol{A}_{\mathcal{I}}^T\boldsymbol{A}_{\mathcal{J}}\right) \leq \delta_{2S}$ where $\sigma_{\max}$ denotes the largest singular value. [2]

In other words, RIP implies that 'disjoint' submatrices are near orthogonal.

2. (SVM)

(a) A hyperplane in $\mathbb{R}^n$ can be defined as

$$\mathcal{P} = \left\{ \boldsymbol{x} : \boldsymbol{\beta}^T \boldsymbol{x} = b \right\} \subset \mathbb{R}^n,$$

for some $\boldsymbol{\beta} \in \mathbb{R}^n$ with $\boldsymbol{\beta} \neq \mathbf{0}$. Show how the distance between an $\boldsymbol{x} \in \mathbb{R}^n$ and $\mathcal{P}$ can be computed? Proof is not required. [2]

(b) Let $\{\boldsymbol{x}_i, y_i\}$, $i = 1, \cdots, m$, be the training dataset where $y_i \in \{-1, 1\}$. Suppose that this training dataset is linearly separable. Give the mathematical definition of linearly separable. [2]

(c) Define the primal optimization problem for Support Vector Machine (SVM). [2]

(d) Find the corresponding Lagrangian $L(\boldsymbol{\beta}, b, \boldsymbol{\lambda})$ where $\boldsymbol{\lambda} = [\cdots, \lambda_i, \cdots]^T$ is the vector of Lagrange multipliers. Specify the constraints of λ_is. [2]

(e) Show that the Lagrange dual function $L_D(\boldsymbol{\lambda}) := \min_{\boldsymbol{\beta}, b} L(\boldsymbol{\beta}, b, \boldsymbol{\lambda}) = -\frac{1}{2}\boldsymbol{\lambda}^T \boldsymbol{K} \boldsymbol{\lambda} + \mathbf{1}^T \boldsymbol{\lambda}$ where $K_{i,j} = y_i \boldsymbol{x}_i^T \boldsymbol{x}_j y_j$. [4]

(f) Give the dual optimization problem. [2]

(g) Give the KKT conditions for the optimal solutions $\boldsymbol{\beta}$, b, and $\boldsymbol{\lambda}$. [4]

(h) Discuss how to find the supporting vectors for the resulting SVM using KKT conditions. [2]

3. (ℓ_1-minimization)

(a) (Iterative Shrinkage-Thresholding Algorithm)

i Let $f(x) = |x|$. Find the subdifferential of $f(x)$. [2]

ii Define $f: \mathbb{R}^n \to \mathbb{R}$ by $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2 + \lambda\|\boldsymbol{x}\|_1$ where $\boldsymbol{z} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^+$ are given. Find the optimal $\boldsymbol{x}^*$ to minimize $f(\boldsymbol{x})$. Express your answer using the famous soft thresholding function η. [2]

iii Let $g(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$. Find its gradient $\nabla g(\boldsymbol{x})$. (No proof is needed.) [2]

iv Let $g(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$. For any given $\boldsymbol{x}^k$, define $\tilde{g}(\boldsymbol{x}) = g(\boldsymbol{x}^k) + \langle \boldsymbol{x} - \boldsymbol{x}^k, \nabla g(\boldsymbol{x}^k)\rangle + \frac{1}{2t_k}\|\boldsymbol{x} - \boldsymbol{x}^k\|_2^2$ for some given constant $t_k \in \mathbb{R}^+$. Simplify $\tilde{g}(\boldsymbol{x})$ into the form $\tilde{g}(\boldsymbol{x}) = \frac{1}{c_1}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2 + c_2$. Find c_1 and $\boldsymbol{z}$. [3]

v Note that in above $\tilde{g}$ is a local approximation of g at $\boldsymbol{x}^k$. Based on $\tilde{g}$, derive the so called Iterative Shrinkage-Thresholding Algorithm to find the global minimum of $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1$. [3]

(b) Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{w}$ where $\boldsymbol{x}_0$ is an S-sparse signal supported at $\mathcal{T}_0 \subset [n]$ and the noise satisfies $\|\boldsymbol{w}\|_2 \le \epsilon$. Consider the constrained optimization problem

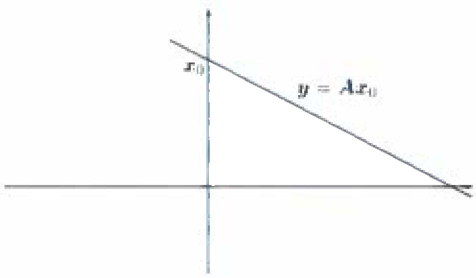
$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad \text{subject to } \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \le \epsilon.$$

Let $\boldsymbol{x}^{\#}$ be the output of this optimization program. Define the estimate error $\boldsymbol{h} := \boldsymbol{x}^{\#} - \boldsymbol{x}_0$.

i Prove the tube constraint $\|\boldsymbol{A}\boldsymbol{h}\|_2 \le 2\epsilon$. [2]

ii Prove the cone constraint $\|\boldsymbol{h}_{\mathcal{T}_0^c}\|_1 \le \|\boldsymbol{h}_{\mathcal{T}_0}\|_1$. [3]

iii Based on the 2-dimensional illustration below, illustrate the regions corresponding to $\|\boldsymbol{A}\boldsymbol{h}\|_2 \le 2\epsilon$ and $\|\boldsymbol{h}_{\mathcal{T}_0^c}\|_1 \le \|\boldsymbol{h}_{\mathcal{T}_0}\|_1$ and use your illustration to explain the intuition why $\boldsymbol{h}$ should be small (under some mathematical assumptions for example RIP). [3]



4. (a) Consider a matrix $\boldsymbol{X}$ with rank 2. The partial observations of $\boldsymbol{X}$ are given as follows:

$$\begin{bmatrix} 1 & 1 & -1 & -1 & 1 & 1 & ? & ? \\ ? & ? & -1 & -1 & 1 & 1 & -1 & -1 \\ 2 & 2 & 0 & 0 & ? & ? & 0 & 0 \\ 1 & 1 & 1 & ? & 1 & 1 & 1 & ? \end{bmatrix}$$

Find the complete $\boldsymbol{X}$ and explain your steps briefly. [5]

(b) (Gaussian Random Vectors) The Gaussian Conditioning Lemma is as follows. Suppose that $\boldsymbol{X} = \left[\boldsymbol{X}_{\mathcal{A}}^T, \boldsymbol{X}_{\mathcal{B}}^T\right]^T \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and

$$\boldsymbol{K} := \Sigma^{-1} = \begin{bmatrix} \boldsymbol{K}_{\mathcal{A}\mathcal{A}} & \boldsymbol{K}_{\mathcal{A}\mathcal{B}} \\ \boldsymbol{K}_{\mathcal{B}\mathcal{A}} & \boldsymbol{K}_{\mathcal{B}\mathcal{B}} \end{bmatrix}.$$

Then $\boldsymbol{X}_{\mathcal{A}}|\boldsymbol{X}_{\mathcal{B}} \sim \mathcal{N}\left(-\boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{K}_{\mathcal{A}\mathcal{B}}\boldsymbol{X}_{\mathcal{B}},\ \boldsymbol{K}_{\mathcal{A}\mathcal{A}}^{-1}\right)$.

i Consider a linear system $\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{W}$ where $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \Sigma_x)$ and $\boldsymbol{W} \sim \mathcal{N}(\mathbf{0}, \Sigma_w)$. What is the distribution of $\boldsymbol{Y}$? What is the joint distribution of $\left[\boldsymbol{X}^T, \boldsymbol{Y}^T\right]^T$? In many signal processing applications, one is asked to estimate $\boldsymbol{X}$ from the given observation $\boldsymbol{Y} = \boldsymbol{y}$. Describe how to use Gaussian conditioning lemma to estimate $\boldsymbol{X}$ (no detailed computations are required). [5]

ii In the Gaussian Graphical model, under what condition will nodes a and b have a connection? Explain your answer using Gaussian conditioning lemma. [5]

iii Now given data points $\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(m)}$, we would like to use these data points to estimate the underlying sparse graph. In particular, we assume that $\mathrm{E}\left[X_a|\boldsymbol{X}_{\sim\{a\}}\right] = \sum_{b\neq a}\theta_{ba}X_b$ and want to find θs to minimize $\mathrm{E}\left[\sum_a\left(X_a - \sum_{b\neq a}\theta_{ba}X_b\right)^2\right]$. Let $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_{(1)}^T \\ \vdots \\ \boldsymbol{x}_{(m)}^T \end{bmatrix}$ be the data matrix. Give the optimization problem for the sparse graph estimation. [5]