Name:	Student ID:

PROBLEM SHEET 2 FOR ADVANCED MACHINE LEARNING (COMP6208)

This problem sheet asks you to prove some well known results. Although the algebra is easy the proofs are not entirely straightforward. There are marks assigned to the readability of the solution and also how well laid out and explained the steps you make are. (A good proof needs to be easy to follow: you need not comment on trivial algebra, but there should not be steps that are difficult to follow).

This looks very mathematical, but it helps to develop the tools and language that is used to describe machine learning.

1

(a) Starting from the definition of a convex function where, for $a \in [0,1]$,

$$f(ax + (1-a)y) \leq a f(x) + (1-a) f(y) \tag{1}$$

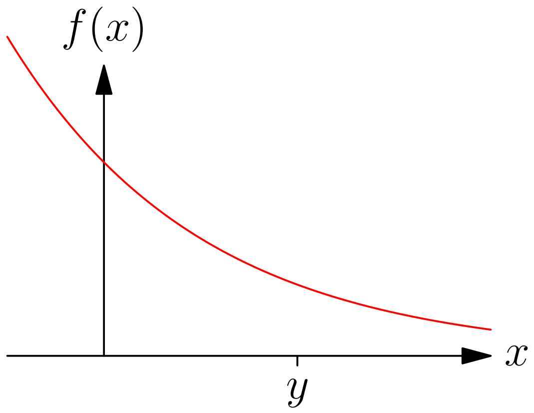
Let $a = \epsilon/(x-y)$ and rearrange the inequality to give

$$(x-y)\left(\frac{f(y+\epsilon) - f(y)}{\epsilon}\right)$$

on the left-hand side. Taking the limit $\epsilon \to 0$ show that the function $f(x)$ lies above the tangent line $t(x) = f(y) + (x-y)f'(y)$ going through the point y. [4 marks]

$\overline{4}$

(b) Sketch the tangent line, $t(x)$, at the point y in the graph shown below. [1 mark]



$\frac{}{1}$

(c) Starting from the inequality for a convex function, f,

$$f(x) \geq f(y) + (x - y) f'(y) \tag{2}$$

consider the case $y = x + \epsilon$, then by Taylor expanding $f(x + \epsilon)$ and $f'(x + \epsilon)$ around x and keeping all terms up to order ϵ^2, show that $f''(x) \geq 0$. [4 marks]

$\frac{}{4}$

(d) Prove that x^4 is convex. [1 mark]

$\frac{}{1}$

End of question 1

(a) $\frac{}{4}$ (b) $\frac{}{1}$ (c) $\frac{}{4}$ (d) $\frac{}{1}$ Total $\frac{}{10}$

2

(a) Show by writing out in component for that $\operatorname{tr}\mathbf{AB} = \operatorname{tr}\mathbf{BA}$ where $\operatorname{tr}\mathbf{M} = \sum_i M_{ii}$ (i.e. the trace of a matrix is equal to the sum of terms down the diagonal).

[2 marks]

$\overline{2}$

(b) Using the fact that we can write a symmetric matrix $\mathbf{M}$ as $\mathbf{M} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}}$ where $\mathbf{V}$ is an orthogonal matrix and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots)$ (i.e. a diagonal matrix with $\Lambda_{ii} = \lambda_i$). Show that $\operatorname{tr}\mathbf{M} = \sum_i \lambda_i$.

[2 marks]

$\overline{2}$

(c) Consider the matrix $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ where the i^{th} column of $\mathbf{X}$ is the vector $\boldsymbol{x}_i$. Compute $\operatorname{tr}\mathbf{X}^{\mathsf{T}}\mathbf{X}$

[2 marks]

$\overline{2}$

(d) The Frobenius norm, $\|\mathbf{X}\|_F$ for a matrix $\mathbf{X}$ is given by

$$\|\mathbf{X}\|_F = \sqrt{\sum_{i,j} X_{ij}^2},$$

where X_{ij} is the (i,j) component of $\mathbf{X}$. Using the previous result, show that $\|\mathbf{X}\|_F^2 = \operatorname{tr} \mathbf{X}^\mathsf{T}\mathbf{X}$ [2 marks]

$\overline{2}$

(e) By using the SVD $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\mathsf{T}$ where $\mathbf{S} = \operatorname{diag}(s_1, s_2, \ldots, s_n)$ (i.e. a diagonal matrix where $S_{ii} = s_i$—the i^{th} singular value) and using the previous results, show that $\|\mathbf{X}\|_F^2 = \sum_i s_i^2$. [2 marks]

$\overline{2}$

End of question 2

(a) $\frac{}{2}$ (b) $\frac{}{2}$ (c) $\frac{}{2}$ (d) $\frac{}{2}$ (e) $\frac{}{2}$ Total $\frac{}{10}$

3 The p-norm of a matrix $\mathbf{M}$, for $p \geq 1$ is defined to satisfy

$$\|\mathbf{M}\|_p = \max_{\boldsymbol{x} \neq \mathbf{0}} \frac{\|\mathbf{M}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p} \tag{3}$$

$$= \max_{\boldsymbol{x}:\|\boldsymbol{x}\|_p = 1} \|\mathbf{M}\boldsymbol{x}\|_p \tag{4}$$

where $\|\boldsymbol{x}\|_p$ is the p norm of a vector defined by

$$\|\boldsymbol{x}\|_p = \left(\sum_i |x_i|^p \right)^{1/p}.$$

Note that with this definition $\|\mathbf{M}\boldsymbol{x}\|_p \leq \|\mathbf{M}\|_p \|\boldsymbol{x}\|_p$ (where the inequality is tight, i.e. there exists a vector where the inequality becomes an equality).

(a) If $\mathbf{U}$ is an orthogonal matrix show that for any vector $\boldsymbol{v}$ that $\|\mathbf{U}\boldsymbol{v}\|_2 = \|\boldsymbol{v}\|_2$. Use this to show $\|\mathbf{U}\mathbf{A}\|_2 = \|\mathbf{A}\|_2$. [2 marks]

$\overline{2}$

(b) If $\boldsymbol{V}$ is an orthogonal matrix show that $\|\mathbf{A}\mathbf{V}^{\mathsf{T}}\|_2 = \|\mathbf{A}\|_2$. [2 marks]

$\overline{2}$

(c) Use the SVD $\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}}$ and the results of part (a) and part (b) to show that $\|\mathbf{M}\|_2 = \|\mathbf{S}\|_2$. [1 mark]

$\overline{1}$

(d) Compute $\|\mathbf{S}\boldsymbol{x}\|_2^2$ where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{S} = \mathrm{diag}(s_1, s_2, \ldots, s_n)$ is the diagonal matrix of singular values, s_i.

[1 mark]

$\overline{1}$

(e) Write down the Lagrangian, L, to maximise $\|\mathbf{S}\boldsymbol{x}\|_2^2$ subject to $\|\boldsymbol{x}\|_2^2 = 1$. Compute the extrumum conditions given by $\partial L / \partial x_i = 0$. Let $(s_\alpha | \alpha = 1, 2, \ldots)$ be the set of unique singular values and I_α the set of indices such that $s_i = s_\alpha$ if $I \in I_\alpha$. Using the extremum condition and the constraint, write down the set of extremum values for $\|\mathbf{S}\boldsymbol{x}\|$ and hence show that $\|\mathbf{M}\|_2 = s_{max}$ where s_{max} is the maximum singular value and $\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}^\mathsf{T}$.

[4 marks]

$\overline{4}$

End of question 3

(a) $\frac{}{2}$ (b) $\frac{}{2}$ (c) $\frac{}{1}$ (d) $\frac{}{1}$ (e) $\frac{}{4}$ Total $\frac{}{10}$

4

(a) We consider the mapping $\boldsymbol{y} = \mathbf{M}\boldsymbol{x}$ where $\mathbf{M}$ is an $n \times n$ matrix. Suppose there is some noise in $\boldsymbol{x}$ so that $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\epsilon}$ and under the mapping $\boldsymbol{y}' = \mathbf{M}\boldsymbol{x}'$. Compute an upper bound for $\|\boldsymbol{y}' - \boldsymbol{y}\|_2$ in terms of $\|\boldsymbol{\epsilon}\|$ and s_{max}, where s_{max} follows the same definition as in Q3(e). [2 marks]

$\overline{2}$

(b) For a matrix $\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}}$ show that

$$\|\mathbf{M}\boldsymbol{x}\|_2 = \|\mathbf{S}\boldsymbol{a}\|_2 \|\boldsymbol{x}\|_2$$

where $\boldsymbol{a} = \mathbf{V}^{\mathsf{T}}\boldsymbol{x}/\|\boldsymbol{x}\|_2$ so that $\|\boldsymbol{a}\|_2 = 1$. Show that we can lower bound $\|\mathbf{S}\boldsymbol{a}\|_2^2$ by s_{min}^2 and hence prove

$$\|\mathbf{M}\boldsymbol{x}\|_2 \geq s_{min}\|\boldsymbol{x}\|_2.$$

where s_{min} is the minimum non-zero singular value analogous to the definition of s_{max}. [3 marks]

$\overline{3}$

(c) Using the previous results, obtain an upper bound for the relative error

$$\frac{\|\boldsymbol{y}' - \boldsymbol{y}\|_2}{\|\boldsymbol{y}\|_2}$$

in terms of s_{max}, s_{min}, $\|\boldsymbol{\epsilon}\|_2$ and $\|\boldsymbol{x}\|_2$. [1 mark]

$\overline{1}$

TURN OVER

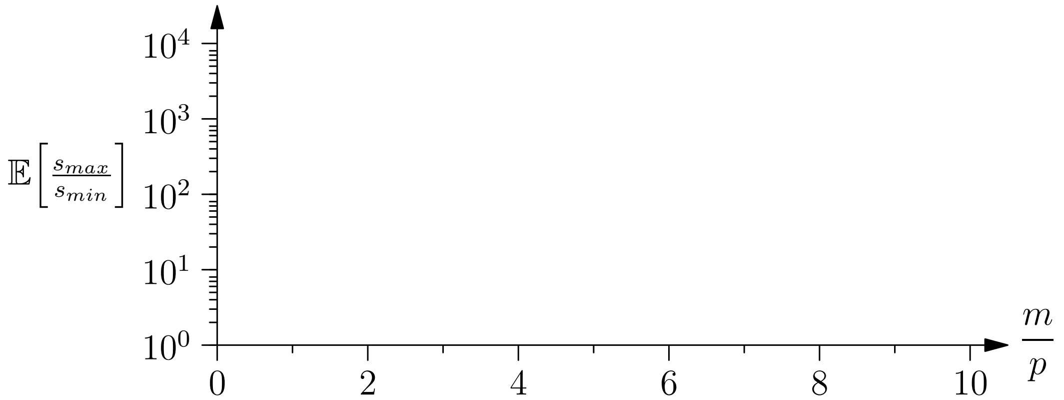
(d) The condition number for an invertible square matrix $\mathbf{M}$ is given by $\kappa_2(\mathbf{M}) = \|\mathbf{M}\|_2 \|\mathbf{M}^{-1}\|_2$ (there are different condition numbers for different norms.) Write down the condition number of $\mathbf{M}$ in terms of s_{max} and s_{min}. [1 mark]

$\frac{\quad}{1}$

(e) In linear regression we make predictions $\hat{y} = \boldsymbol{x}^\mathsf{T} \boldsymbol{w}$ given an input $\boldsymbol{x}$ where $\boldsymbol{w} = \mathbf{X}^+ \boldsymbol{y}$, where $\mathbf{X}^+ = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$ is the pseudo inverse of the design matrix $\mathbf{X}$ and $\boldsymbol{y}$ is a vector of training examples. There are bounds on the accuracy of linear regression depending on $\mathbb{E}\left[s_{max}/s_{min}\right]$ where s_{max} and s_{min} are respectively the maximum and minimum no-zero singular values of the design matrix. Consider randomly drawn feature vectors

$$\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Use python to generate the $m \times p$ dimensional design matrix $\mathbf{X}$ with rows $\boldsymbol{x}_i^\mathsf{T}$. By computing the singular values for $\mathbf{X}$ with $m = i \times p$ where $i \in \{1, 2, \ldots, 10\}$, find s_{max}/s_{min}. Repeat this 10 times for each i to obtain an estimate of $\mathbb{E}\left[s_{max}/s_{min}\right]$. Plot a graph of your estimate for $\mathbb{E}\left[s_{max}/s_{min}\right]$ (on a log-axis) versus m/p for $p = 10, 50$ and 100. [3 marks]



$\frac{\quad}{3}$

End of question 4

(a) $\frac{\quad}{2}$ (b) $\frac{\quad}{3}$ (c) $\frac{\quad}{1}$ (d) $\frac{\quad}{1}$ (e) $\frac{\quad}{3}$ Total $\frac{\quad}{10}$

