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

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

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]

Rearranging the Equation 1

$$f(y + a(x-y)) \leq f(y) + a\,(f(x) - f(y)).$$

Or

$$\frac{f(y + a(x-y)) - f(y)}{a} \leq f(x) - f(y).$$

Letting $a = \epsilon/(x-y)$ **then**

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

Taking the limit $\epsilon \to 0$ **then using**

$$\lim_{\epsilon \to 0} \frac{f(y+\epsilon) - f(y)}{\epsilon} = f'(y)$$

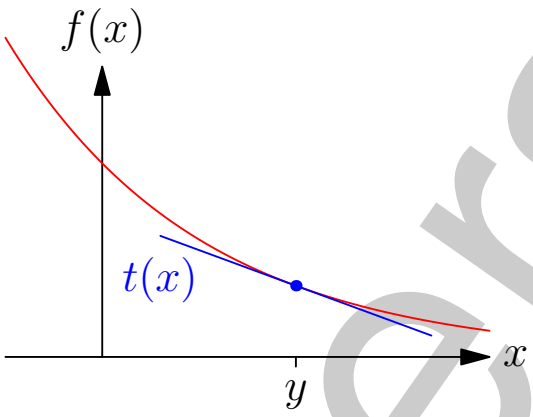
So that

$$(x-y)f'(y) \leq f(x) - f(y)$$

or

$$f(x) \geq f(y) + (x-y)f'(y) = t(x).$$

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



(c) Starting from the inequality for a convex function

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

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 for a convex function $f''(x) \geq 0$. [4 marks]

We use the expansions

$$f(x+\epsilon) = f(x) + \epsilon f'(x) + \frac{\epsilon^2}{2} f''(x) + O(\epsilon^3)$$
$$f'(x+\epsilon) = f'(x) + \epsilon f''(x) + O(\epsilon^2)$$

Substituting into the Equation (2)

$$f(x) \geq f(x) + \epsilon f'(x) + \frac{\epsilon^2}{2} f''(x) + O(\epsilon^3) - \epsilon \left(f'(x) + \epsilon f''(x) + O(\epsilon^2) \right)$$

Or subtraction $f(x)$ on both sides

$$0 \geq -\frac{\epsilon^2}{2} f''(x) + O(\epsilon^3)$$

Since this has to be true for all $\epsilon > 0$ this requires $f''(x) \geq 0$.

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

$$\frac{\mathrm{d}^2 x^4}{\mathrm{d}x^2} = 12x^2 \geq 0$$

End of question 1

2

(a) Show by writing out in component for that $\mathrm{tr}\,\mathbf{AB} = \mathrm{tr}\,\mathbf{BA}$ where $\mathrm{tr}\,\mathbf{M} = \sum_i M_{ii}$ (i.e. the trace of a matrix is equal to the sum of terms down the diagonal). [2 marks]

$$\mathrm{tr}\,\mathbf{AB} = \sum_{i,j} A_{ij} B_{ji} \qquad \mathrm{tr}\,\mathbf{BA} = \sum_{k,l} B_{kl} A_{lk}$$

These are identical (which is clearer if we let $l = i$ and $k = j$).

(b) Using the fact that we can write a symmetric matrix $\mathbf{M}$ as $\mathbf{M} = \mathbf{V\Lambda V}^\mathsf{T}$ where $\mathbf{V}$ is an orthogonal matrix and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots)$ (i.e. a diagonal matrix with $\Lambda_{ii} = \lambda_i$). Show that $\mathrm{tr}\,\mathbf{M} = \sum_i \lambda_i$. [2 marks]

$$\mathrm{tr}\,\mathbf{M} = \mathrm{tr}\left(\mathbf{V\Lambda V^\mathsf{T}}\right) = \mathrm{tr}\left(\mathbf{V^\mathsf{T} V\Lambda}\right) = \mathrm{tr}\,\mathbf{\Lambda} = \sum_i \lambda_i$$

where we have used (1) $\mathrm{tr}\,\mathbf{AB} = \mathrm{tr}\,\mathbf{BA}$ with $\mathbf{A} = \mathbf{V\Lambda}$ and $\mathbf{B} = \mathbf{V}^\mathsf{T}$ and (2) $\mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$.

(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 $\mathrm{tr}\,\mathbf{X}^\mathsf{T}\mathbf{X}$ [2 marks]

$$\mathrm{tr}\,\mathbf{X}^\mathsf{T}\mathbf{X} = \mathrm{tr}\begin{pmatrix} \boldsymbol{x}_1^\mathsf{T}\boldsymbol{x}_1 & \boldsymbol{x}_1^\mathsf{T}\boldsymbol{x}_2 & \ldots & \boldsymbol{x}_1^\mathsf{T}\boldsymbol{x}_n \\ \boldsymbol{x}_2^\mathsf{T}\boldsymbol{x}_1 & \boldsymbol{x}_2^\mathsf{T}\boldsymbol{x}_2 & \ldots & \boldsymbol{x}_2^\mathsf{T}\boldsymbol{x}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_n^\mathsf{T}\boldsymbol{x}_1 & \boldsymbol{x}_n^\mathsf{T}\boldsymbol{x}_2 & \ldots & \boldsymbol{x}_n^\mathsf{T}\boldsymbol{x}_n \end{pmatrix} = \sum_{i=1}^{n} \boldsymbol{x}_i^\mathsf{T}\boldsymbol{x}_i = \sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2$$

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

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

using the previous result show that $\|\mathbf{X}\|_F^2 = \mathrm{tr}\,\mathbf{X}^\mathsf{T}\mathbf{X}$ [2 marks]

$$\|\mathbf{X}\|_F^2 = \sum_{i,j} X_{ij}^2 = \sum_j \left(\sum_i X_{ij}^2 \right) = \sum_j \|\boldsymbol{x}_j\|^2 = \mathrm{tr}\,\mathbf{X}^\mathsf{T}\mathbf{X}$$

where we have used the fact that $\sum_i X_{ij}^2$ is a sum over all elements in the j^{th} column but the j^{th} column is equal to the vector $\boldsymbol{x}_j$.

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

Now

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \left(\mathbf{VS}^\mathsf{T}\mathbf{U}^\mathsf{T}\right)\left(\mathbf{USV}^\mathsf{T}\right) = \mathbf{VS}^\mathsf{T}\mathbf{SV}^\mathsf{T}$$

And so using

$$\mathrm{tr}\,\mathbf{X}^\mathsf{T}\mathbf{X} = \mathrm{tr}\,\mathbf{S}^\mathsf{T}\mathbf{S}$$

but $\mathbf{S}^\mathsf{T}\mathbf{S}$ is a diagonal matrix with elements s_i^2 so $\mathrm{tr}\,\mathbf{X}^\mathsf{T}\mathbf{X} = \sum_i s_i^2$. But we have shown that $\|\mathbf{X}\|_F^2 = \mathrm{tr}\,\mathbf{X}^\mathsf{T}\mathbf{X}$, so $\|\mathbf{X}\|_F^2 = \sum_i s_i^2$.

End of question 2

3 The p-norm of a matrix $\mathbf{M}$ 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 $\|\mathbf{UA}\|_2 = \|\mathbf{A}\|_2$. [2 marks]

We note that

$$\|\mathbf{U}\boldsymbol{v}\|_2^2 = \boldsymbol{v}^\mathsf{T}\mathbf{U}^\mathsf{T}\mathbf{U}\boldsymbol{v} = \boldsymbol{v}^\mathsf{T}\boldsymbol{v} = \|\boldsymbol{v}\|_2^2$$

since norms are positive $\|\mathbf{U}\boldsymbol{v}\|_2 = \|\boldsymbol{v}\|_2$. We note that $\|\mathbf{UA}\boldsymbol{x}\|_2 = \|\mathbf{A}\boldsymbol{x}\|_2$, thus the unit vector $\boldsymbol{x}$ that maximises $\|\mathbf{A}\boldsymbol{x}\|_2$ will also maximise $\|\mathbf{UA}\boldsymbol{x}\|_2$. By Equation (4) this implies $\|\mathbf{UA}\|_2 = \|\mathbf{A}\|_2$.

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

We consider a unit-vector $\boldsymbol{x}$ that maximises $\|\mathbf{A}\boldsymbol{x}\|_2$ then by definition (4) $\|\mathbf{A}\boldsymbol{x}\|_2 = \|\mathbf{A}\|_2$. But we note that $\boldsymbol{y} = \mathbf{V}\boldsymbol{x}$ will be a unit vector that maximises $\|\mathbf{AV}^\mathsf{T}\boldsymbol{y}\|_2$ since

$$\|\mathbf{AV}^\mathsf{T}\boldsymbol{y}\|_2 = \|\mathbf{AV}^\mathsf{T}\mathbf{V}\boldsymbol{x}\|_2 = \|\mathbf{A}\boldsymbol{x}\|_2$$

using $,\mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$. But as $\boldsymbol{x}$ is a unit vector that maximises $\|\mathbf{A}\boldsymbol{x}\|_2$ then $\boldsymbol{y}$ will be a unit vector that maximise $\|\mathbf{AV}^\mathsf{T}\boldsymbol{y}\|_2$ and $\|\mathbf{AV}^\mathsf{T}\| = \|\mathbf{A}\|$.

(c) Using the SVD $\mathbf{M} = \mathbf{USV}^\mathsf{T}$ and the results of part (a) and part (b) show that $\|\mathbf{M}\|_2 = \|\mathbf{S}\|_2$. [1 marks]

We note that $\|\mathbf{M}\|_2 = \|\mathbf{USV}^\mathsf{T}\|_2$. But by part (a) $\|\mathbf{USV}^\mathsf{T}\|_2 = \|\mathbf{SV}^\mathsf{T}\|_2$ and by part (b) $\|\mathbf{SV}^\mathsf{T}\|_2 = \|\mathbf{S}\|_2$.

(d) Compute $\|\mathbf{S}\boldsymbol{x}\|_2^2$ where $\mathbf{S} = \mathrm{diag}(s_1,s_2,\ldots,s_n)$ is the diagonal matrix of singular values [1 marks]

Since $\mathbf{S} = \mathrm{diag}(s_1,s_2,\ldots,s_n)$ then

$$\|\mathbf{S}\boldsymbol{x}\|_2^2 = \sum_{i=1}^{n} s_i^2 x_i^2$$

(e) Write down the Lagrangian 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, i such that $s_i = s_\alpha$. Using the extrumum 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. [4 marks]

The Lagrangian is equal to

$$L = \sum_{i=1}^{n} s_i^2 x_i^2 - \lambda\left(\sum_{i=1}^{n} x_i^2 - 1\right)$$

The extremum conditions are

$$\frac{\partial L}{\partial x_i} = x_i\left(s_i^2 - \lambda\right) = 0$$

(and the constraint). The extremum conditions are thus either $x_i = 0$ or $s_i^2 = \lambda$. Thus the value of Lagrange multipliers are $\lambda = s_\alpha$ where $x_i = 0$ if $i \notin I_\alpha$. Thus

$$\|\mathbf{S}\boldsymbol{x}\|_2^2 = s_\alpha^2 \sum_{i\in I_\alpha} x_i^2 = s_\alpha^2$$

where we have used $\sum_i x_i^2 = 1$. The maximum condition is just given by the case s_{max} (the largest maximum eigenvalue). Thus using Equation (4) we have $\|\mathbf{S}\|_2 = s_{max}$. But since $\|\mathbf{M}\|_2 = \|\mathbf{S}\|_2$ it follows that $\|\mathbf{M}\|_2 = s_{max}$.

End of question 3

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 $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\epsilon}$ so that under the mapping $\boldsymbol{y}' = \mathbf{M}\boldsymbol{x}'$. Compute an upper bound on $\|\boldsymbol{y}' - \boldsymbol{y}\|_2$ in terms of $\|\epsilon\|$ and s_{max}. [2 marks]

$$\|\boldsymbol{y}' - \boldsymbol{y}\|_2 = \|\mathbf{M}\boldsymbol{\epsilon}\|_2 \leq \|\mathbf{M}\|_2\|\epsilon\|_2 = s_{max}\|\epsilon\|_2$$

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

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

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

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

[3 marks]

$$\|\mathbf{X}\boldsymbol{x}\|_2 = \|\mathbf{U}\mathbf{S}\mathbf{V}^\mathsf{T}\boldsymbol{x}\|_2 = \|\mathbf{S}\mathbf{V}^\mathsf{T}\boldsymbol{x}\|_2$$

but using $\boldsymbol{a} = \mathbf{V}^\mathsf{T}\boldsymbol{x}/\|\boldsymbol{x}\|_2$

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

where $\|a\|_2 = 1$**. But**

$$\|\mathbf{S}\boldsymbol{a}\|_2^2 = \sum_i a_i^2 s_i^2 \geq s_{min}^2 \sum_i a_i^2 = s_{min}^2 = s_{min}^2\|\boldsymbol{a}\|^2 = s_{min}^2.$$

Thus $\|\mathbf{S}\boldsymbol{a}\|_2 \geq s_{min}$ **and**

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

(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 $\|\boldsymbol{x}\|$. [1 marks]

$$\frac{\|\boldsymbol{y}' - \boldsymbol{y}\|_2}{\|\boldsymbol{y}\|_2} \leq \frac{s_{max}}{s_{min}}\frac{\|\boldsymbol{\epsilon}\|_2}{\|\boldsymbol{x}\|_2}$$

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

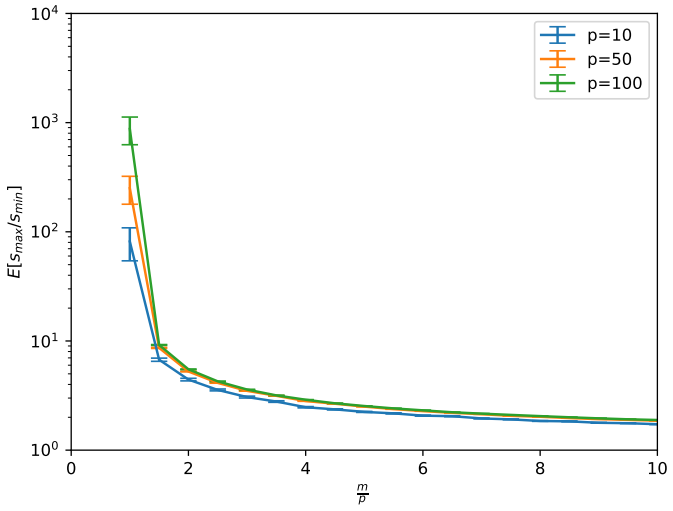
We observe that $\|\mathbf{A}\|_2 = s_{max}$ **and** $\|\mathbf{A}^{-1}\|_2 = s_{min}^{-1}$ **so that**

$$\kappa_2(\mathbf{M}) = \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2 = \frac{s_{max}}{s_{min}}$$

(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 the maximum and minimum no-zero singular value of the design matrix. Consider randomly drawn feature vectors

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

Using python generate the $m \times p$ dimensional design matrix **X** with rows $\boldsymbol{x}_i^\mathsf{T}$. By computing the singular values for **X** for $m = i \times p$ where $i = 1,2,\ldots,10$ find s_{max}/s_{min}. Repeat this 10 times to obtain an estimate of $\mathbb{E}\left[s_{max}/s_{min}\right]$. Plot a graph of you 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]



End of question 4

END OF PAPER