Name:	Student ID:

PROBLEM SHEET 2 FOR ADVANCED MACHINE LEARNING (COMP6208)

1

(a) To find the minimum of a 1-d function, we can do an iterative update

$$x^{(t+1)} = x^{(t)} - r\, f'(x^{(t)})$$

where r is a learning rate and $f'(t)$ is the derivative of the function, $f(x)$ we are minimising. Supposing that

$$f(x) = \frac{c}{2}(x - x^*)^2$$

where $c > 0$. Write down a recursion formula for $x^{(t+1)}$ and $x^{(t)}$. [2 marks]

$\overline{2}$

(b) Show by induction that $x^{(t)} = F(t) = x^* + (x^{(0)} - x^*)(1 - cr)^t$ is a solution to the recursion relation. Hence find a condition on the value of r to ensure convergence. [5 marks]

$\overline{5}$

(c) Now consider the case when $\boldsymbol{x} \in \mathbb{R}^n$. We assume that

$$g(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^{\mathsf{T}} \mathbf{Q} (\boldsymbol{x} - \boldsymbol{x}^*)$$

where **Q** is a symmetric, positive-definite matrix. The Hessian, **H**, of $g(\boldsymbol{x})$ is a matrix with components

$$H_{ij} = \frac{\partial^2 g(\boldsymbol{x})}{\partial x_i \partial x_j}.$$

By writing out $g(\boldsymbol{x})$ as a double sum over the components compute the Hessian [2 marks]

$\overline{2}$

(d) Gradient descent in $\mathbb{R}^n$ is given by

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - r \boldsymbol{\nabla} g(\boldsymbol{x}).$$

Using the definition of $g(\boldsymbol{x})$ write down a recursion relation between $\boldsymbol{x}^{(t+1)}$ and $\boldsymbol{x}^{(t)}$. [2 marks]

$\overline{2}$

(e) Defining $\boldsymbol{\Delta}^{(t)} = (\boldsymbol{x}^{(t)} - \boldsymbol{x}^*)$ obtain a recursion relation between $\boldsymbol{\Delta}^{(t+1)}$ and $\boldsymbol{\Delta}^{(t)}$. (This is easy if you subtract $\boldsymbol{x}^*$ from both sides of the recursion equation for $\boldsymbol{x}^{(t+1)}$.) [2 marks]

$\overline{2}$

(f) Using the eigenvalue decomposition $\mathbf{Q} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathsf{T}}$ and defining $\boldsymbol{z}^{(t)} = \mathbf{V}^{\mathsf{T}}\boldsymbol{\Delta}^{t}$ write out a recursion relation between $\boldsymbol{z}^{(t+1)}$ and $\boldsymbol{z}^{(t)}$. (This is helped by multiplying the recursion relation on the left by $\mathbf{V}^{\mathsf{T}}$ and using the fact that $\mathbf{V}$ is an orthogonal matrix.) [3 marks]

$\overline{3}$

(g) Solve the recursion relation to obtain a formula for $\boldsymbol{x}^{(t)}$ in terms of the initial state $\boldsymbol{x}^{(0)}$. Express this formula for the i^{th} component of $\boldsymbol{x}^{(t)}$ and hence find a condition on the learning rate r to ensure convergence. [4 marks]

$\overline{4}$

End of question 1

(a) $\frac{}{2}$ (b) $\frac{}{5}$ (c) $\frac{}{2}$ (d) $\frac{}{2}$ (e) $\frac{}{2}$ (f) $\frac{}{3}$ (g) $\frac{}{4}$ Total $\frac{}{20}$

2

(a) Consider the non-quadratic minimum at x^* given by

$$f(x) = \frac{c}{2}(x - x^*)^2 + \frac{d}{6}(x - x^*)^3$$

where we use Newton's method

$$x^{(t+1)} = x^{(t)} - \frac{f'(x^{(t)})}{f''(x^{(t)})}.$$

by computing the derivatives and expanding for small $x^{(t)} - x^*$ show that $x^{(t+1)} - x^* = O\left(\left(x^{(t)} - x^*\right)^2\right)$.

(To do this we need to expand a term with the structure

$$\frac{r + s\epsilon}{u + v\epsilon}.$$

Note that we can use the geometric series expansion to write

$$\frac{1}{u + v\epsilon} = \frac{1}{u}\frac{1}{1 + \frac{v}{u}\epsilon} = \frac{1}{u}\left(1 - \frac{v}{u}\epsilon + \left(\frac{v}{u}\epsilon\right)^2 - \cdots\right)$$

which is convergent provided $|\frac{v}{u}\epsilon| < 1$.) [5 marks]

$\overline{5}$

(b) Consider the function

$$h(x) = -x\log(x)$$

defined for $0 < x \le 1$. By computing $h'(x)$ and setting $h'(x) = 0$ compute the value of x^* that maximises $h(x)$. [2 marks]

$\overline{2}$

(c) Compute $h''(x)$ and thus compute the Newton update function

$$n(x) = x - \frac{h'(x)}{h''(x)}$$

(the answer is rather surprising and in no way general). [3 marks]

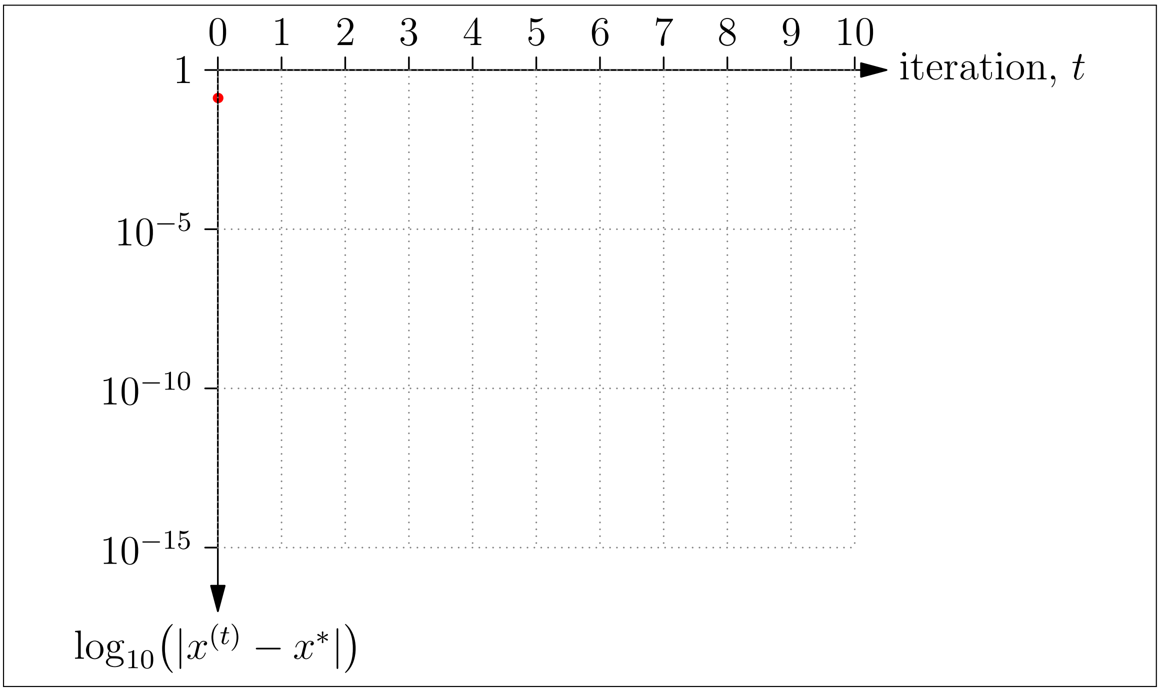
$\overline{3}$

(d) On the axes given below plot $x^{(t)}$ for $t = 0,1,2,\ldots,10$, starting from $x^{(0)} = 0.5$ where we use the gradient ascent updates

$$x^{(t+1)} = x^{(t)} + r\,h'(x^t)$$

for $r = \{0.3, 0.4, 0.5\}$ (that is you should plot three curves).

Also plot $x^{(t)}$ where $x^{(t+1)} = n(x^{(t)})$ (that is using Newton's update formula) for $t = 0,1,2,3$ and 4—note that to machine precision $x^{(5)} = x^*$. [10 marks]



$\overline{}$ / 10

End of question 2

(a) $\frac{}{5}$ (b) $\frac{}{2}$ (c) $\frac{}{3}$ (d) $\frac{}{10}$ Total $\frac{}{20}$

3

(a) Show that for $p_i > 0$ the function

$$h(\boldsymbol{p}) = -\sum_i p_i \log(p_i)$$

is strongly convex-down. Hint: show that the Hessian matrix is negative-definite.

[3 marks]

$\overline{3}$

(b) Write down the Lagrangian, $\mathcal{L}$, for the problem of maximising $h(\boldsymbol{p})$ subject to the constraints

$$\sum_i p_i = 1 \qquad\qquad \sum_i p_i E_i = U.$$

Then explain why there is a unique solution to this constrained optimisation problem.

[2 marks]

$\overline{2}$

(c) By setting $\partial\mathcal{L}/\partial p_i = 0$ find the value of p_i that maximises $\mathcal{L}$ in terms of E_i and the Lagrange multipliers. Use the constraint $\sum_i p_i = 1$ to eliminate the Lagrange multiplier that enforces this constraint. [5 marks]

$\overline{5}$

End of question 3

(a) $\frac{}{3}$	(b) $\frac{}{2}$	(c) $\frac{}{5}$	Total $\frac{}{10}$