

---

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]

---

$$\begin{aligned} x^{(t+1)} &= x^{(t)} - cr(x^{(t)} - x^*) \\ &= (1 - cr)x^{(t)} + crx^*. \end{aligned}$$

---

- (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]
- 

**In the base case, if we take  $t = 0$  then**

$$\begin{aligned} F(0) &= x^* + (x^{(0)} - x^*)(1 - cr)^0 \\ &= x^* + (x^{(0)} - x^*) = x^{(0)}. \end{aligned}$$

**Thus the formula is true for  $t = 0$ . Assuming the formula is true for  $x^{(t)}$  then substituting this into the recursion relation**

$$\begin{aligned} x^{(t+1)} &= (1 - cr)F(t) + crx^* \\ &= (1 - cr)\left(x^* + (x^{(0)} - x^*)(1 - cr)^t\right) + crx^* \\ &= x^* + (x^{(0)} - x^*)(1 - cr)^{(t+1)} = F(t + 1). \end{aligned}$$

**Thus, we shown that assuming  $x^{(t)} = F(t)$  then  $x^{(t+1)} = F(t + 1)$ , but as it is true for  $t = 0$  the formula will be true for all non-negative integers.**

**The condition for convergence is that  $0 < cr < 2$ . Assuming  $c > 0$  then  $0 < r < 2/c$ .**

---

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

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

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

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

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

$$H_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} \frac{1}{2} \sum_{k\ell} (x_k - x_k^*) Q_{k\ell} (x_\ell - x_\ell^*) = Q_{ij}$$

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

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

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

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - r \mathbf{Q}(\mathbf{x}^{(t)} - \mathbf{x}^*)$$

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

$$\Delta^{(t+1)} = (\mathbf{I} - r \mathbf{Q}) \Delta^{(t)}.$$

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

$$\mathbf{z}^{(t+1)} = (\mathbf{I} - r \mathbf{\Lambda}) \mathbf{z}^{(t)}$$

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

**The solution to the recursion equation is trivially.**

$$\mathbf{z}^{(t)} = (\mathbf{I} - r \mathbf{\Lambda})^t \mathbf{z}^{(0)}.$$

As  $\Lambda$  is diagonal

$$z_i^{(t)} = (1 - r\lambda_i)^t z_i^{(0)}.$$

Thus the condition for  $z_i^{(t)}$  to converge is that  $0 < r\lambda_i < 2$ . For all  $z^{(t)}$  to converge we require  $0 < r < 2/\lambda_{\max}$ .

---

End of question 1

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  $x^{(t+1)} - x^* = O\left((x^{(t)} - x^*)^2\right)$ .

(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

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

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

[5 marks]

---

Using

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

so that

$$x^{(t+1)} = x^{(t)} - \frac{c(x^{(t)} - x^*) + \frac{d}{2}(x^{(t)} - x^*)^2}{c + d(x^{(t)} - x^*)}$$

Subtracting  $x^*$  from both sides of this equation and writing  $\epsilon^{(n)} = x^{(t)} - x^*$ .

We find

$$\begin{aligned}\epsilon^{(t+1)} &= \epsilon^{(t)} - \frac{c\epsilon^{(t)} + \frac{d}{2}(\epsilon^{(t)})^2}{c + d\epsilon^{(t)}} \\ &= \epsilon^{(t)} - \left(c\epsilon^{(t)} + \frac{d}{2}(\epsilon^{(t)})^2\right) \frac{1}{c} \left(1 - \frac{d}{c}\epsilon^{(t)} + \frac{d^2}{c^2}(\epsilon^{(t)})^2 + \dots\right) \\ &= \frac{d}{2c}(\epsilon^{(t)})^2 + O\left((\epsilon^{(t)})^3\right) = O\left((\epsilon^{(t)})^2\right)\end{aligned}$$

that is  $x^{(t+1)} - x^* = O\left((x^{(t)} - x^*)^2\right)$ .

---

(b) Consider the function

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

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

---

$$h'(x) = -\log(x) - 1$$

Thus  $h'(x) = 0$  has a solution at  $x^* = e^{-1}$ .

---

(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]

---

$$h''(x) = \frac{-1}{x} \text{ so that}$$

$$n(x) = x - x(x \log(x) + 1) = -x \log(x).$$

It is a coincidence that  $n(x) = h(x)$ .

---

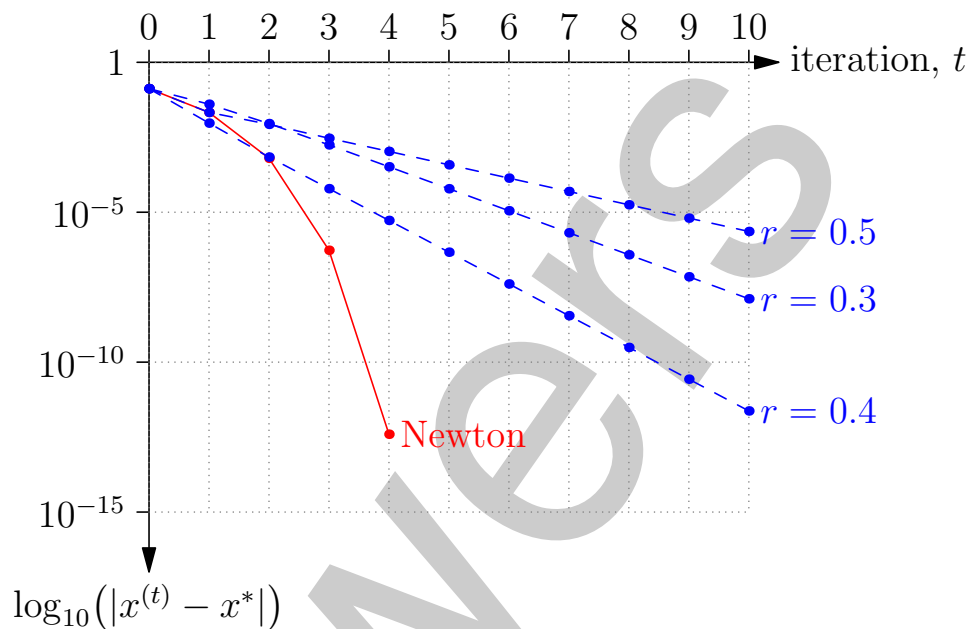
(d) On the axes given below plot  $x^{(t)}$  for  $t = 0, 1, 2, \dots, 10$ , starting from  $x^{(0)} = 0.5$  where we use the gradient ascent updates

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

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]

---



End of question 2

3

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

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

is strongly convex-down. Hint: show that the Hessian matrix is negative-definite. [3 marks]

**We note that the Hessian H has entries**

$$H_{ij} = \frac{\partial^2 h(\mathbf{p})}{\partial p_i \partial p_j} = 0 \quad H_{ii} = \frac{\partial^2 h(\mathbf{p})}{\partial p_i^2} = -\frac{1}{p_i}$$

**Thus the Hessian is a diagonal matrix with  $H_{ii} = -1/p_i < 0$  (since  $p_i > 0$ ). The eigenvalues of a diagonal matrix are equal to the diagonal elements so that  $\lambda_i < 0$  for all  $i$ . This is a necessary and sufficient condition for H to be negative-definite and hence  $h(\mathbf{p})$  is strongly convex-down.**

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

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

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

[2 marks]

TURN OVER  
Page 5 of 6

---

The Lagrangian is given by

$$\mathcal{L} = h(\mathbf{p}) - \alpha \left( \sum_i p_i - 1 \right) - \beta \left( \sum_i p_i E_i - U \right).$$

The problem has a unique solution because the Lagrangian is strongly convex-down. This follows because  $h(\mathbf{p})$  is strongly convex-down and the constraints are linear (both convex-up and convex-down). The sum of a strongly convex function and another convex functions is strongly convex.

---

- (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]
- 

We note that

$$\frac{\partial h(\mathbf{p})}{\partial p_i} = -\log(p_i) - 1 - \alpha - \beta E_i$$

so that

$$p_i = e^{-\alpha-1-\beta E_i}$$

Using the constraint  $\sum_i p_i = 1$  then

$$\sum_i e^{-\alpha-1-\beta E_i} = 1$$

Dividing through by this we find

$$p_i = \frac{e^{-\beta E_i}}{\sum_j e^{-\beta E_j}}.$$

This is the famous Boltzmann distribution.

---

End of question 3

**END OF PAPER**