IN5400
Theoretical Exercises

Mohamed Ismail

April 2, 2020

These exercises can give you a hint about how exercises for the written exam can be.

1 Week 3

1.1 Linear Algebra

$$a = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad b = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$$
$$P = \begin{pmatrix} 3 & 6 \\ 2 & 4 \end{pmatrix}, \quad Q = \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix}$$

Compute x in the following cases (if it is not possible, state why).

a)

$$\begin{aligned} x &= a^T b && (1)\\ &= (1 \quad 2) \cdot \begin{pmatrix} 4 \\ 2 \end{pmatrix} && (2)\\ &= 8 && (3) \end{aligned}$$

b)

$$\begin{aligned} x &= Pa && (4)\\ &= \begin{pmatrix} 3 & 6 \\ 2 & 4 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 2 \end{pmatrix} && (5)\\ &= \begin{pmatrix} 15 \\ 10 \end{pmatrix} && (6) \end{aligned}$$

c)

$$x = PQ \tag{7}$$
$$= \begin{pmatrix} 3 & 6 \\ 2 & 4 \end{pmatrix} \cdot \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix} \tag{8}$$
$$= \begin{pmatrix} 18 & 30 \\ 12 & 20 \end{pmatrix} \tag{9}$$

d)

$$Px = a \tag{10}$$
$$P^{-1}Px = P^{-1}a \tag{11}$$
$$Ix = P^{-1}a \tag{12}$$

This only possible if the matrix P has an inverse matrix such that $P^{-1}P = I$. One possible way of checking if such a matrix exists is to calculate the determinant of P and verify that $\det(P) \neq 0$, then a P^{-1} exists.

$$det(P) = \left| \begin{pmatrix} 3 & 6 \\ 2 & 4 \end{pmatrix} \right| \tag{13}$$
$$= 3 \times 4 - 6 \times 2 \tag{14}$$
$$= 12 - 12 \tag{15}$$
$$= 0. \tag{16}$$

Since $\det(P) = 0$, then the equation $Px = a$ is not solvable problem.

e)

$$Qx = b \tag{17}$$
$$Q^{-1}Qx = Q^{-1}b \tag{18}$$
$$Ix = Q^{-1}b \tag{19}$$
$$x = Q^{-1}b \tag{20}$$
$$= \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 4 \\ 2 \end{pmatrix} \tag{21}$$
$$= \begin{pmatrix} 3 \\ -1 \end{pmatrix} \tag{22}$$

1.2 Derivatives in higher dimensions

The gradient of a scalar-valued, multi-variable function $f : \mathbb{R}^n \to \mathbb{R}$ is given by

$$\nabla_x f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} \tag{23}$$

For the same function, we can state the Hessian matrix of f w.r.t. x as

$$H_x(f(x)) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 x_1} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial x_2 x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \frac{\partial^2 f}{\partial x_n x_2} & \cdots & \frac{\partial^2 f}{\partial x_n x_n} \end{pmatrix} \tag{24}$$

For a vector-valued, multi-variable function $g : \mathbb{R}^n \to \mathbb{R}^m$ the Jacobian matrix of g w.r.t. x is given by

$$J_x(g(x)) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_2}{\partial x_1} & \cdots & \frac{\partial g_m}{\partial x_1} \\ \frac{\partial g_1}{\partial x_2} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_m}{\partial x_2} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial g_1}{\partial x_n} & \frac{\partial g_2}{\partial x_n} & \cdots & \frac{\partial g_m}{\partial x_n} \end{pmatrix} \tag{25}$$

a) Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by

$$f(x) = x^T Ax + b^T x + c \tag{26}$$

where $A \in \mathbb{R}^{n \times n}, b, x \in \mathbb{R}^n$. Give the expression of the gradient of f w.r.t. $x, \nabla_x f(x)$.

$$f(x) = x^T Ax + b^T x + c \tag{27}$$

$$= [x_1, \ldots, x_n] \begin{pmatrix} \sum_{i=1}^n a_{1i} x_i \\ \sum_{i=1}^n a_{2i} x_i \\ \vdots \\ \sum_{i=1}^n a_{ni} x_i \end{pmatrix} + \sum_{i=1}^n b_i x_i + c \tag{28}$$

$$= \Big(x_1 \sum_{i=1}^n a_{1i} x_i + x_2 \sum_{i=1}^n a_{2i} x_i + \cdots + x_n \sum_{i=1}^n a_{ni} x_i \Big) + \sum_{i=1}^n b_i x_i + c \tag{29}$$

$$\nabla_x f(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix} \left[\Big(x_1 \sum_{i=1}^n a_{1i} x_i + x_2 \sum_{i=1}^n a_{2i} x_i + \cdots + x_n \sum_{i=1}^n a_{ni} x_i \Big) + \sum_{i=1}^n b_i x_i + c \right] \tag{30}$$

$$= \begin{pmatrix} \sum_{i=1}^n a_{1i} x_i \\ \sum_{i=1}^n a_{2i} x_i \\ \vdots \\ \sum_{i=1}^n a_{ni} x_i \end{pmatrix} + \begin{pmatrix} \sum_{i=1}^n a_{i1} x_i \\ \sum_{i=1}^n a_{i2} x_i \\ \vdots \\ \sum_{i=1}^n a_{in} x_i \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{31}$$

$$= Ax + A^T x + b \tag{32}$$

$$= \Big(A + A^T \Big) x + b \tag{33}$$

b) Compute the Hessian matrix of f w.r.t. $x, H_x(f(x))$.

$$f = \left(x_1 \sum_{i=1}^{n} a_{1i}x_i + x_2 \sum_{i=1}^{n} a_{2i}x_i + \cdots + x_n \sum_{i=1}^{n} a_{ni}x_i\right) + \sum_{i=1}^{n} b_i x_i + c \tag{34}$$

$$\nabla_x f(x) = \begin{pmatrix} \sum_{i=1}^{n} a_{1i}x_i \\ \sum_{i=1}^{n} a_{2i}x_i \\ \vdots \\ \sum_{i=1}^{n} a_{ni}x_i \end{pmatrix} + \begin{pmatrix} \sum_{i=1}^{n} a_{i1}x_i \\ \sum_{i=1}^{n} a_{i2}x_i \\ \vdots \\ \sum_{i=1}^{n} a_{in}x_i \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{35}$$

$$H_x(f(x)) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} + \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \cdots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{pmatrix} \tag{36}$$

$$= \begin{pmatrix} 2a_{11} & a_{12} + a_{21} & \cdots & a_{1n} + a_{n1} \\ a_{21} + a_{12} & 2a_{22} & \cdots & a_{1n} + a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} + a_{1n} & a_{n2} + a_{2n} & \cdots & 2a_{nn} \end{pmatrix} \tag{37}$$

$$= A + A^T \tag{38}$$

c) Compute the Jacobian matrix of the gradient of f w.r.t. $x, J_x(\nabla_x f(x))$.

$$\nabla_x f(x) = \begin{pmatrix} \sum_{i=1}^{n} a_{1i}x_i \\ \sum_{i=1}^{n} a_{2i}x_i \\ \vdots \\ \sum_{i=1}^{n} a_{ni}x_i \end{pmatrix} + \begin{pmatrix} \sum_{i=1}^{n} a_{i1}x_i \\ \sum_{i=1}^{n} a_{i2}x_i \\ \vdots \\ \sum_{i=1}^{n} a_{in}x_i \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{39}$$

$$J_x(f(x)) = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \cdots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{pmatrix} + \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \tag{40}$$

$$= A^T + A \tag{41}$$

d) Show how, in general, the Hessian matrix relates to the Jacobian matrix.

We see that from the previous exercises that

$$H_x(f(x)) = J_x(\nabla_x f(x)) = A + A^T \tag{42}$$

1.3 Chain rule

For a single variable, scalar-valued functions $f, g : \mathbb{R} \to \mathbb{R}$, the derivative of the composition $(f \circ g)(x) = f(g(x))$ w.r.t. x is given by the so-called chain rule of differentiation

$$\frac{\partial}{\partial x} f(g(x)) = \frac{\partial f}{\partial g}\frac{\partial g}{\partial x} \tag{43}$$

Compute the derivative $\frac{\partial f}{\partial x}$ on the following expressions.

a)

$$f(x) = \sin(x^2) \tag{44}$$
$$g(x) = x^2 \tag{45}$$
$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial g}\frac{\partial g}{\partial x} \tag{46}$$
$$= \cos(x^2)2x \tag{47}$$

b)

$$f(x) = e^{\sin(x^2)} \tag{48}$$
$$g(x) = \sin x^2 \tag{49}$$
$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial g}\frac{\partial g}{\partial x} \tag{50}$$
$$= e^{\sin x^2} 2x\cos(x^2) \tag{51}$$

c) In the case where $f : \mathbb{R}^m \to \mathbb{R}, g : \mathbb{R}^n \to \mathbb{R}^m$, and $x \in R^n$, the derivative of f

$$f(g(x)) = f(g_1(x), \ldots, g_m(x)) \tag{52}$$
$$= f(g_1(x_1, \ldots, x_n), \ldots, g_m(x_1, \ldots, x_n)) \tag{53}$$

w.r.t. one of the components of x, can be given by a generalisation of the above chain rule

$$\frac{\partial f}{\partial x_i} = \sum_{j=1}^{m} \frac{\partial f}{\partial g_j}\frac{\partial g_j}{\partial x_i} \tag{54}$$

Compute the derivatives of $\frac{\partial f}{\partial x_1}$ and $\frac{\partial f}{\partial x_2}$ when

$$\begin{cases} f &= \sin g_1 + g_2^2 \\ g_1 &= x_1 e^{x_2} \\ g_2 &= x_1 + x_2^2 \end{cases} \tag{55}$$

$$f = \sin g_1 + g_2^2, \quad g_1 = x_1 e^{x_2}, \quad g_2 = x_1 + x_2^2 \tag{56}$$
$$\frac{\partial f}{\partial x_1} = \frac{\partial f}{\partial g_1}\frac{\partial g_1}{\partial x_1} + \frac{\partial f}{\partial g_2}\frac{\partial g_2}{\partial x_1} \quad \& \quad \frac{\partial f}{\partial x_2} = \frac{\partial f}{\partial g_1}\frac{\partial g_1}{\partial x_2} + \frac{\partial f}{\partial g_2}\frac{\partial g_2}{\partial x_2} \tag{57}$$
$$= \cos(x_1 e^{x_2})e^{x_2} + 2(x_1 + x_2^2) \quad \& \quad = \cos(x_1 e^{x_2})x_1 e^{x_2} + 2(x_1 + x_2^2)2x_2 \tag{58}$$

1.4 Forward Propagation

Suppose we have a small dense neural network. The input vector is

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix} \tag{59}$$

In the first layer we have the following weigth and bias parameters

$$W_1 = \begin{pmatrix} w_{11}^1 & w_{12}^1 & w_{13}^1 \\ w_{21}^2 & w_{22}^2 & w_{23}^2 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 3 \\ 2 & -1 & 1 \end{pmatrix}, \quad \boldsymbol{b}_1 = \begin{pmatrix} b_1^1 \\ b_2^1 \\ b_3^1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \tag{60}$$

In the second layer we have the following weight and bias parameters

$$W_2 = \begin{pmatrix} w_{11}^2 \\ w_{21}^2 \\ w_{31}^2 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}, \quad \boldsymbol{b}_2 = \begin{pmatrix} b_1^2 \end{pmatrix} = \begin{pmatrix} 1 \end{pmatrix}. \tag{61}$$

a) Compute the value of the activation in the second layer, $\hat{y}$, when the activation function in the first and second are identity functions.

First we calculate the input for the activation function in the first layer, z_1, then we set that into the activation function $a_1 = g(z_1)$ and use that as the input for the second layer.

$$z_1 = W_1^T \boldsymbol{x} + \boldsymbol{b_1} = \begin{pmatrix} 9 \\ -2 \\ 5 \end{pmatrix} \tag{62}$$

$$a_1 = g(z_1) = z_1 \tag{63}$$

$$z_2 = W_2^T a_1 + \boldsymbol{b_2} = 36 \tag{64}$$

$$a_2 = g(z_2) = z_2 \tag{65}$$

b) Compute the value of the activation in the second layer, $\hat{y}$, when the activation functions in the first layer are ReLU functions, and in the second layer is the identity function.

First we calculate the input for the activation function in the first layer, z_1, then we set that into the activation function $a_1 = g(z_1)$ and use that as the input for the second layer.

$$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases} \tag{66}$$

$$z_1 = W_1^T \boldsymbol{x} + \boldsymbol{b_1} = \begin{pmatrix} 9 \\ -2 \\ 5 \end{pmatrix} \tag{67}$$

$$a_1 = g(z_1) = \begin{pmatrix} 9 \\ 0 \\ 5 \end{pmatrix} \tag{68}$$

$$z_2 = W_2^T a_1 + \boldsymbol{b_2} = 38 \tag{69}$$

$$a_2 = g(z_2) = z_2 \tag{70}$$

1.5 Cost functions and optimization

Let $\theta^k = [1, 3]^T$ be the value of some parameter $\theta = [\theta_1, \theta_2]^T$ at iteration k of a gradient descent method. Let the loss function be

$$L(\theta) = 2\big(\theta_1 - 2\big)^2 + \theta_2 \tag{71}$$

With a learning rate of $\lambda = 2$, find the value of θ^{k+1} when it has been updated with the gradient descent method.

$$L(\theta_1, \theta_2) = 2\big(\theta_1 - 2\big)^2 + \theta_2 \tag{72}$$

$$\nabla_1 L(\theta_1, \theta_2) = 4(\theta_1 - 2) \quad \& \quad \nabla_2 L(\theta_1, \theta_2) = 1 \tag{73}$$

$$\theta_1^{k+1} = \theta_1^k - \lambda \nabla_1 L(\theta_1^k, \theta_2^k) = 1 - 2 * 4(1 - 2) = 9 \tag{74}$$

$$\theta_2^{k+1} = \theta_2^k - \lambda \nabla_2 L(\theta_1^k, \theta_2^k) = 3 - 2 * (1) = 1 \tag{75}$$

$$\theta^{k+1} = \begin{pmatrix} 9 \\ 1 \end{pmatrix} \tag{76}$$

1.6 Optimizing a convex objective function

Let the loss function L be convex and quadratic

$$L(\theta) = \frac{1}{2}\theta^T Q \theta - b^T \theta \tag{77}$$

where $Q \in \mathbb{R}^{n \times n}$ is a symmetric and positive definite matrix, $b \in \mathbb{R}^n$ is constant vector, and $\theta \in \mathbb{R}^n$ is a vector of parameters.

a) Find an expression for the unique minimizer θ^* of L.

Since Q is a symmetric, then that means

$$Q = Q^T \tag{78}$$

$$L(\theta) = \frac{1}{2}\theta^T Q\theta - b^T\theta \tag{79}$$

$$\nabla L(\theta) = \frac{1}{2}\Big(Q + Q^T\Big)\theta - b = Q\theta - b \tag{80}$$

The derivation of $\nabla L(\theta)$ can be found in an earlier exercise.

To find the expression for the unique minimizer θ^* we must find the θ that makes $\nabla L(\theta) = 0$.

$$\nabla L(\theta) = Q\theta - b \tag{81}$$

$$\nabla L(\theta^*) = Q\theta^* - b = 0 \tag{82}$$

$$Q\theta^* = b \tag{83}$$

$$\theta^* = Q^{-1}b \tag{84}$$

b) Instead of solving the optimization problem analytically, we want to take an iterative approach using gradient descent. Let ∇L_k be the gradient of L w.r.t. θ evaluated at θ_k. Show that the optimal step length at this iteration is given by

$$\lambda_k = \frac{\nabla L_k^T \nabla L_k}{\nabla L_k^T Q \nabla L_k}. \tag{85}$$

By optimal we mean the step length that yields the smallest value of L at step $k+1$.

$$\theta_{k+1} = \theta_k - \lambda_k \nabla L_k(\theta_k) \tag{86}$$

$$L(\theta_{k+1}) = \frac{1}{2}\theta_{k+1}^T Q\theta_{k+1} - b^T\theta_{k+1} \tag{87}$$

$$\frac{\partial L(\theta_{k+1})}{\partial \lambda_k} = \Big(Q\theta_{k+1} - b\Big) \times -\nabla L_k \tag{88}$$

$$\Big(\theta_k^T - \lambda_k \nabla L_k^T\Big) Q \nabla L_k = b^T \nabla L_k \tag{89}$$

$$\theta_k^T Q \nabla L_k - b^T \nabla L_k = \lambda_k \nabla L_k^T Q \nabla L_k \tag{90}$$

$$\lambda_k \nabla L_k^T Q \nabla L_k = \Big(\theta_k^T Q - b^T\Big)\nabla L_k \tag{91}$$

$$\lambda_k = \frac{\nabla L_k^T \nabla L_k}{\nabla L_k^T Q \nabla L_k} \tag{92}$$