

## COMP3670/6670: Introduction to Machine Learning

These exercises will concentrate on vector calculus, and how to compute derivatives of functions that live in higher dimensions.

### Preliminaries

The formal definition of the derivative of a function  $f : \mathbb{R} \rightarrow \mathbb{R}$  is given by

$$\frac{df}{dx} := \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a function of a vector  $\mathbf{x}$ . The derivative of  $f(\mathbf{x})$  with respect to  $\mathbf{x}$  is defined as

$$\nabla_{\mathbf{x}} f = \text{grad} f = \frac{df}{d\mathbf{x}} := \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} & \frac{\partial f(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \in (\mathbb{R}^n \rightarrow \mathbb{R})^{1 \times n}$$

Note that  $\frac{df}{d\mathbf{x}}$  is a row vector, where each element is a function of the form  $\mathbb{R}^n \rightarrow \mathbb{R}$ . We write  $\nabla_{\mathbf{x}} f \in (\mathbb{R}^n \rightarrow \mathbb{R})^{1 \times n}$ . Some authors write  $\nabla_{\mathbf{x}} f \in \mathbb{R}^{1 \times n}$  as an abuse of notation for the sake of brevity, and ease of matching dimensions. Keep in mind that each element of the row vector isn't a real number, but itself a function.

Let  $\mathbf{g} : \mathbb{R} \rightarrow \mathbb{R}^n$  be a function of a scalar  $t$ . The derivative of  $\mathbf{g}(t)$  with respect to  $t$  is defined as

$$\frac{d\mathbf{g}}{dt} := \begin{bmatrix} \frac{dg_1(t)}{dt} \\ \frac{dg_2(t)}{dt} \\ \vdots \\ \frac{dg_n(t)}{dt} \end{bmatrix} \in (\mathbb{R} \rightarrow \mathbb{R})^{n \times 1}$$

Note that  $\frac{d\mathbf{g}}{dt}$  is a column vector, where each element is itself a function of the form  $\mathbb{R} \rightarrow \mathbb{R}$ . As before, we notate this using an abuse of notation as  $\frac{d\mathbf{g}}{dt} \in \mathbb{R}^{n \times 1}$ ,

The reason why the derivatives are defined this way, is so that the dimensions match when we define the chain rule.

Given  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  and  $\mathbf{g} : \mathbb{R} \rightarrow \mathbb{R}^n$ , we can define two new functions

$$h : \mathbb{R} \rightarrow \mathbb{R}, \quad h(t) = f(\mathbf{g}(t))$$

$$\mathbf{k} : \mathbb{R}^n \rightarrow \mathbb{R}^n \quad \mathbf{k}(\mathbf{x}) = \mathbf{g}(f(\mathbf{x}))$$

and we can define their derivatives as

$$\frac{dh}{dt} = \frac{df}{d\mathbf{g}} \frac{d\mathbf{g}}{dt} = \begin{bmatrix} \frac{\partial f(\mathbf{g})}{\partial g_1} & \dots & \frac{\partial f(\mathbf{g})}{\partial g_n} \end{bmatrix} \begin{bmatrix} \frac{\partial g_1}{\partial t} \\ \vdots \\ \frac{\partial g_n}{\partial t} \end{bmatrix} = \sum_{i=1}^n \frac{\partial f(\mathbf{g})}{\partial g_i} \frac{\partial g_i}{\partial t}$$

and

$$\frac{d\mathbf{k}}{d\mathbf{x}} = \frac{d\mathbf{g}}{df} \frac{df}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial g_1}{\partial f} \\ \vdots \\ \frac{\partial g_n}{\partial f} \end{bmatrix} \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1}{\partial f} \frac{\partial f(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial g_1}{\partial f} \frac{\partial f(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial f} \frac{\partial f(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial g_n}{\partial f} \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} = \mathbf{A}$$

where  $\mathbf{A}_{ij} = \frac{\partial g_i}{\partial f} \frac{\partial f(\mathbf{x})}{\partial x_j}$ .

(Here, the term  $\frac{\partial f(\mathbf{g})}{\partial g_i}$  means to substitute each output component of  $\mathbf{g}$  into the inputs for  $f$ , and take the partial derivative with respect to the  $g_i$ , the  $i^{\text{th}}$  component of  $\mathbf{g}$ .)

For a vector valued function  $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , we define the matrix of all first order derivatives as the *Jacobian*, which is given by

$$\mathbf{J} = \nabla_{\mathbf{x}} \mathbf{f} = \frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}, \quad \mathbf{J}_{ij} = \frac{\partial f_i(\mathbf{x})}{\partial x_j}$$

.

You may also need the definition of matrix multiplication.

If  $\mathbf{A} \in \mathbb{R}^{n \times m}$  and  $\mathbf{B} \in \mathbb{R}^{m \times p}$ , the product  $\mathbf{C} = \mathbf{AB}$  is a matrix in  $\mathbb{R}^{n \times p}$  satisfying

$$C_{ij} = \sum_{k=1}^m A_{ik} B_{kj}$$

If  $\mathbf{A} \in \mathbb{R}^{n \times m}$ ,  $\mathbf{b} \in \mathbb{R}^{m \times 1}$  and  $\mathbf{c} \in \mathbb{R}^{n \times 1}$  then the matrix vector products  $\mathbf{Ab}$  and  $\mathbf{c}^T \mathbf{A}$  satisfy the properties

$$(\mathbf{Ab})_k = \sum_{j=1}^m A_{kj} b_j$$

and

$$(\mathbf{c}^T \mathbf{A})_k = \sum_{i=1}^n A_{ik} c_i$$

For  $\mathbf{x} \in \mathbb{R}^n$ , the Euclidean norm  $\|\cdot\|_2$  is given by

$$\|\mathbf{x}\|_2 := \sqrt{\mathbf{x}^T \mathbf{x}}$$

For all problems below, state the dimension of the answer where appropriate.

**Question 1**

**Formal definition of derivative**

Compute the derivative of  $f : \mathbb{R} \rightarrow \mathbb{R}, f(x) = x^2$  from the formal limit definition of the derivative.

**Solution.**

$$\begin{aligned}\frac{d}{dx}x^2 &= \lim_{h \rightarrow 0} \frac{(x+h)^2 - x^2}{h} \\ &= \lim_{h \rightarrow 0} \frac{x^2 + 2xh + h^2 - x^2}{h} \\ &= \lim_{h \rightarrow 0} \frac{2xh + h^2}{h} \\ &= \lim_{h \rightarrow 0} 2x + h \\ &= 2x \in \mathbb{R}^{1 \times 1}\end{aligned}$$

**Question 2**

**Vector Derivative of Scalar Function**

Given  $f : \mathbb{R}^2 \rightarrow \mathbb{R}, f(\mathbf{x}) = 2x_1x_2 + x_1 + 3x_2 + 5$ , compute  $\frac{df}{d\mathbf{x}}$ .

**Solution.**

$$\begin{aligned}\frac{\partial f}{\partial x_1} &= 2x_2 + 1 \\ \frac{\partial f}{\partial x_2} &= 2x_1 + 3\end{aligned}$$

Hence,

$$\frac{df}{d\mathbf{x}} = [2x_2 + 1, \quad 2x_1 + 3] \in \mathbb{R}^{1 \times 2}$$

**Question 3**

**Scalar Derivative of Vector Function**

Given  $\mathbf{g}(t) : \mathbb{R} \rightarrow \mathbb{R}^2, \mathbf{g}(t) = \begin{bmatrix} t^2 \\ e^t \end{bmatrix}$  compute  $\frac{d\mathbf{g}}{dt}$ .

**Solution.**

$$\frac{d\mathbf{g}}{dt} = \begin{bmatrix} \frac{\partial}{\partial t} t^2 \\ \frac{\partial}{\partial t} e^t \end{bmatrix} = \begin{bmatrix} 2t \\ e^t \end{bmatrix} \in \mathbb{R}^{2 \times 1}$$

**Question 4**

**Derivative of the L2 Norm**

Let  $\mathbf{x} \in \mathbb{R}^n$ , and define  $k : \mathbb{R}^n \rightarrow \mathbb{R}, k(\mathbf{x}) = \|\mathbf{x}\|_2^2 := \mathbf{x}^T \mathbf{x}$ . Compute  $\frac{dk}{d\mathbf{x}}$ .

**Solution.** We proceed by computing one of the partial derivatives.

$$\frac{\partial}{\partial x_i} \mathbf{x}^T \mathbf{x} = \frac{\partial}{\partial x_i} \sum_{j=1}^n x_j^2 = \frac{\partial}{\partial x_i} x_i^2 = 2x_i$$

Hence,

$$\frac{df}{d\mathbf{x}} = [2x_1, \dots, 2x_n] = 2\mathbf{x}^T \in \mathbb{R}^{1 \times n}$$

**Question 5**

**Chain Rule, Scalar Derivative**

Let  $h : \mathbb{R} \rightarrow \mathbb{R}, h(t) = f(\mathbf{g}(t))$ , where  $f$  and  $\mathbf{g}$  are defined in Question 2 and Question 3 respectively.

1. Compute  $\frac{dh}{dt}$  by using the chain rule.

**Solution.** The chain rule here is given by

$$\frac{dh}{dt} = \frac{df}{d\mathbf{g}} \frac{d\mathbf{g}}{dt}$$

Using the previous exercises to help us, and treating  $f$  as a function of  $\mathbf{g}$ ,

$$\frac{df}{d\mathbf{g}} = [2g_2 + 1, 2g_1 + 3]$$

$$\frac{d\mathbf{g}}{dt} = \begin{bmatrix} 2t \\ e^t \end{bmatrix}$$

Hence,

$$\frac{dh}{dt} = [2g_2 + 1, 2g_1 + 3] \begin{bmatrix} 2t \\ e^t \end{bmatrix} = (2g_2 + 1)2t + (2g_1 + 3)e^t$$

Substituting  $g_1 = t^2$  and  $g_2 = e^t$ , we obtain

$$\begin{aligned} &= (2e^t + 1)2t + (2t^2 + 3)e^t \\ &= 2t^2e^t + 4te^t + 2t + 3e^t \in \mathbb{R} \end{aligned}$$

2. Compute  $\frac{dh}{dt}$  by evaluating  $f(\mathbf{g}(t))$  first, and then differentiating the entire expression by  $t$ . Compare your answer to the above and check that they match.

**Solution.**

$$\begin{aligned} \frac{dh}{dt} &= \frac{d}{dt} f(\mathbf{g}(t)) \\ &= \frac{d}{dt} (2t^2e^t + t^2 + 3e^t + 5) \\ &= 2(2te^t + t^2e^t) + 2t + 3e^t \\ &= 2t^2e^t + 4te^t + 2t + 3e^t \in \mathbb{R} \end{aligned}$$

which matches the answer above.

## Question 6

## Chain Rule, Vector Derivative

Let  $\mathbf{k} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $\mathbf{k}(\mathbf{x}) = \mathbf{g}(f(\mathbf{x}))$ , where  $f$  and  $\mathbf{g}$  are defined in Question 2 and Question 3 respectively.

1. Compute  $\frac{d\mathbf{k}}{d\mathbf{x}}$  using the chain rule.

**Solution.** Using the chain rule, we have

$$\frac{d\mathbf{k}}{d\mathbf{x}} = \frac{d\mathbf{g}}{df} \frac{df}{d\mathbf{x}}$$

$$\frac{d\mathbf{g}}{df} = \begin{bmatrix} 2f \\ e^f \end{bmatrix} \quad \frac{df}{d\mathbf{x}} = [2x_2 + 1, 2x_1 + 3]$$

Hence,

$$\frac{d\mathbf{k}}{d\mathbf{x}} = \begin{bmatrix} 2f \\ e^f \end{bmatrix} [2x_2 + 1, 2x_1 + 3] = \begin{bmatrix} 2f(2x_2 + 1) & 2f(2x_1 + 3) \\ e^f(2x_2 + 1) & e^f(2x_1 + 3) \end{bmatrix}$$

Substituting  $f = 2x_1x_2 + x_1 + 3x_2 + 5$ , we obtain

$$\frac{d\mathbf{k}}{d\mathbf{x}} = \begin{bmatrix} 2(2x_1x_2 + x_1 + 3x_2 + 5)(2x_2 + 1) & 2(2x_1x_2 + x_1 + 3x_2 + 5)(2x_1 + 3) \\ (2x_2 + 1)e^{2x_1x_2 + x_1 + 3x_2 + 5} & (2x_1 + 3)e^{2x_1x_2 + x_1 + 3x_2 + 5} \end{bmatrix} \in \mathbb{R}^{2 \times 2}$$

2. Compute  $\frac{d\mathbf{k}}{d\mathbf{x}}$  directly by using the Jacobian to differentiate  $\mathbf{g}(f(\mathbf{x}))$ . Check your answer matches the above using chain rule.

**Solution.** Computing directly,

$$\begin{aligned}\mathbf{k}(\mathbf{x}) &= \mathbf{g}(f(\mathbf{x})) \\ &= \mathbf{g}(2x_1x_2 + x_1 + 3x_2 + 5) \\ &= \begin{bmatrix} (2x_1x_2 + x_1 + 3x_2 + 5)^2 \\ e^{2x_1x_2 + x_1 + 3x_2 + 5} \end{bmatrix}\end{aligned}$$

Hence, we can compute the derivative using the Jacobian:

$$\begin{aligned}\frac{d\mathbf{k}}{d\mathbf{x}} &= \begin{bmatrix} \frac{\partial k_1(\mathbf{x})}{\partial x_1} & \frac{\partial k_1(\mathbf{x})}{\partial x_2} \\ \frac{\partial k_2(\mathbf{x})}{\partial x_1} & \frac{\partial k_2(\mathbf{x})}{\partial x_2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial}{\partial x_1}(2x_1x_2 + x_1 + 3x_2 + 5)^2 & \frac{\partial}{\partial x_2}(2x_1x_2 + x_1 + 3x_2 + 5)^2 \\ \frac{\partial}{\partial x_1}e^{2x_1x_2 + x_1 + 3x_2 + 5} & \frac{\partial}{\partial x_2}e^{2x_1x_2 + x_1 + 3x_2 + 5} \end{bmatrix} \\ &= \begin{bmatrix} 2(2x_1x_2 + x_1 + 3x_2 + 5)(2x_2 + 1) & 2(2x_1x_2 + x_1 + 3x_2 + 5)(2x_1 + 3) \\ (2x_2 + 1)e^{2x_1x_2 + x_1 + 3x_2 + 5} & (2x_1 + 3)e^{2x_1x_2 + x_1 + 3x_2 + 5} \end{bmatrix} \in \mathbb{R}^{2 \times 2}\end{aligned}$$

which matches the above.

## Question 7

## More Derivatives

1. Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}, f(\mathbf{x}) = (\mathbf{x}^T \mathbf{x} + 1)^2$ .  
Compute  $\frac{d}{d\mathbf{x}} f(\mathbf{x})$  using the chain rule. (You can use the previous questions to help you.)

**Solution.** Let  $g : \mathbb{R} \rightarrow \mathbb{R}, g(u) = (u + 1)^2$  and  $h : \mathbb{R}^n \rightarrow \mathbb{R}, h(\mathbf{x}) = \mathbf{x}^T \mathbf{x}$ . Note that  $f(\mathbf{x}) = g(h(\mathbf{x}))$ . Hence, we can apply the chain rule.

$$\begin{aligned}\frac{df}{d\mathbf{x}} &= \frac{dg}{dh} \frac{dh}{d\mathbf{x}} \\ \frac{dg}{dh} &= \frac{dg(h)}{dh} = \frac{d}{dh}(h + 1)^2 = 2(h + 1) \\ \frac{dh}{d\mathbf{x}} &= \frac{d}{d\mathbf{x}} \mathbf{x}^T \mathbf{x} = 2\mathbf{x}^T \text{ (by .)}\end{aligned}$$

Hence,

$$\frac{df}{d\mathbf{x}} = 2(h + 1)2\mathbf{x}^T = 4(\mathbf{x}^T \mathbf{x} + 1)\mathbf{x}^T \in \mathbb{R}^{1 \times n}$$

2. Directly compute  $\frac{d}{d\mathbf{x}} f(\mathbf{x})$  by expanding out  $(\mathbf{x}^T \mathbf{x} + 1)^2$  first. Your result should match the above.

**Solution.**

$$\begin{aligned}\frac{d}{d\mathbf{x}} f(\mathbf{x}) &= \frac{d}{d\mathbf{x}} (\mathbf{x}^T \mathbf{x} + 1)^2 \\ &= \frac{d}{d\mathbf{x}} ((\mathbf{x}^T \mathbf{x})^2 + 2\mathbf{x}^T \mathbf{x} + 1) \\ &= 2(\mathbf{x}^T \mathbf{x}) \left( \frac{d}{d\mathbf{x}} \mathbf{x}^T \mathbf{x} \right) + 2 \left( \frac{d}{d\mathbf{x}} \mathbf{x}^T \mathbf{x} \right) \\ &= 2(\mathbf{x}^T \mathbf{x})(2\mathbf{x}^T) + 2(2\mathbf{x}^T) \\ &= 4(\mathbf{x}^T \mathbf{x})\mathbf{x}^T + 4\mathbf{x}^T \\ &= 4(\mathbf{x}^T \mathbf{x} + 1)\mathbf{x}^T \in \mathbb{R}^{1 \times n}\end{aligned}$$

which matches 1.

**Question 8****Derivative of a Matrix-Vector product**

Let  $\mathbf{A} \in \mathbb{R}^{m \times n}$  and  $\mathbf{x} \in \mathbb{R}^{n \times 1}$ . Show that  $\frac{d}{d\mathbf{x}}(\mathbf{Ax}) = \mathbf{A}$ .

**Solution.** Note that the vector derivative of a vector valued function will be a matrix. We take the partial derivative of each component, with respect to each element of  $\mathbf{x}$ .

$$\frac{\partial}{\partial x_p}(\mathbf{Ax})_q = \frac{\partial}{\partial x_p} \sum_j A_{qj} x_j = \sum_j A_{qj} \frac{\partial x_j}{\partial x_p} = A_{qp} = (\mathbf{A})_{qp}$$

Hence,

$$\frac{d}{d\mathbf{x}}(\mathbf{Ax}) = \mathbf{A}$$

**Question 9****Linear Regression**

Let  $\Phi \in \mathbb{R}^{n \times m}$ ,  $\mathbf{w} \in \mathbb{R}^{n \times 1}$ ,  $\mathbf{t} \in \mathbb{R}^{m \times 1}$ .

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f(\mathbf{w}) = \|(\mathbf{w}^T \Phi)^T - \mathbf{t}\|_2^2$

1. Verify that  $f$  is well defined (the dimensions of all the components match up).

**Solution.** We have  $\mathbf{w} \in \mathbb{R}^{n \times 1}$ . So  $\mathbf{w}^T \in \mathbb{R}^{1 \times n}$ . So  $\mathbf{w}^T \Phi \in \mathbb{R}^{1 \times m}$ . Transposing the result,  $(\mathbf{w}^T \Phi)^T \in \mathbb{R}^{m \times 1}$ . The vector  $\mathbf{t} \in \mathbb{R}^{m \times 1}$ , and subtraction is defined for vectors of the same size. So  $(\mathbf{w}^T \Phi)^T - \mathbf{t} \in \mathbb{R}^{m \times 1}$ . Norms are only defined for column vectors, which  $(\mathbf{w}^T \Phi)^T - \mathbf{t}$  is. So  $\|(\mathbf{w}^T \Phi)^T - \mathbf{t}\|_2 \in \mathbb{R}$ . Real numbers are closed under squaring and halving, so  $f(\mathbf{w}) \in \mathbb{R}$ .

2. Compute  $\frac{d}{d\mathbf{w}} f(\mathbf{w})$ .

**Solution.** Note that  $(\mathbf{w}^T \Phi)^T = \Phi^T \mathbf{w}$ .

Let  $g : \mathbb{R}^{m \times 1} \rightarrow \mathbb{R}$ ,  $g(\mathbf{x}) = \|\mathbf{x}\|_2^2 := \mathbf{x}^T \mathbf{x}$  and

$\mathbf{h} : \mathbb{R}^{n \times 1} \rightarrow \mathbb{R}^{m \times 1}$ ,  $\mathbf{h}(\mathbf{w}) = (\Phi^T \mathbf{w}) - \mathbf{t}$ . Then  $f(\mathbf{w}) = g(\mathbf{h}(\mathbf{w}))$ . Apply chain rule.

$$\frac{df}{d\mathbf{w}} = \frac{dg}{d\mathbf{h}} \frac{d\mathbf{h}}{d\mathbf{w}}$$

From , we have

$$\frac{dg}{d\mathbf{h}} = 2\mathbf{h}^T$$

From , together with the property that  $\mathbf{t}$  has no dependence on  $\mathbf{w}$ , we have

$$\frac{d\mathbf{h}}{d\mathbf{w}} = \frac{d}{d\mathbf{w}} (\Phi^T \mathbf{w} - \mathbf{t}) = \frac{d}{d\mathbf{w}} (\Phi^T \mathbf{w}) = \Phi^T$$

Hence,

$$\frac{df}{d\mathbf{w}} = 2\mathbf{h}^T \Phi^T = 2(\Phi^T \mathbf{w} - \mathbf{t})^T \Phi^T = 2(\mathbf{w}^T \Phi - \mathbf{t}^T) \Phi^T \in \mathbb{R}^{1 \times n}$$

**Question 10****Matrix Gradient**

Given  $\mathbf{X} \in \mathbb{R}^{n \times m}$  and some vectors  $\mathbf{a} \in \mathbb{R}^{? \times ?}$ ,  $\mathbf{b} \in \mathbb{R}^{? \times ?}$ .

1. What are the dimensions of  $\mathbf{a}$  and  $\mathbf{b}$  such that  $\mathbf{a}^T \mathbf{X} \mathbf{b}$  is well defined?<sup>1</sup> What is the dimension of the result?

**Solution.** Since  $\mathbf{a}, \mathbf{b}$  are vectors, one of the dimensions must be 1. We have that  $\mathbf{a}^T \in \mathbb{R}^{? \times ?}$  and the inner dimensions must match to right multiply by  $\mathbf{X}$ , hence  $\mathbf{a}^T \in \mathbb{R}^{? \times n}$ . The missing

<sup>1</sup>Note that if  $\mathbf{X}$  is square, symmetric and positive definite, then defining  $\langle \mathbf{a}, \mathbf{b} \rangle := \mathbf{a}^T \mathbf{X} \mathbf{b}$  gives an inner product.

dimension is one, so  $\mathbf{a}^T \in \mathbb{R}^{1 \times n}$ , which implies  $\mathbf{a} \in \mathbb{R}^{n \times 1}$ .

Similarly, we are left multiplying  $\mathbf{X} \in \mathbb{R}^{n \times m}$  by  $\mathbf{b}$ , hence  $\mathbf{b} \in \mathbb{R}^{m \times 1}$ .

The dimension of the result is a scalar, as the outer dimensions of  $\mathbf{a}^T$  and  $\mathbf{b}$  are both 1.

2. Compute the matrix gradient  $\frac{d}{d\mathbf{X}} \mathbf{a}^T \mathbf{X} \mathbf{b}$ .

**Solution.** We write out the definition of  $\mathbf{a}^T \mathbf{X} \mathbf{b}$ .

$$\begin{aligned} \mathbf{a}^T \mathbf{X} \mathbf{b} &= \mathbf{a}^T (\mathbf{X} \mathbf{b}) \\ &= \sum_j a_j (\mathbf{X} \mathbf{b})_j \\ &= \sum_j a_j \left( \sum_k X_{jk} b_k \right) \\ &= \sum_j \left( \sum_k a_j X_{jk} b_k \right) \end{aligned}$$

Now, to take the matrix gradient, we take the partial with respect to each element of the matrix.

$$\frac{\partial}{\partial X_{pq}} (\mathbf{a}^T \mathbf{X} \mathbf{b}) = \frac{\partial}{\partial X_{pq}} \left( \sum_j \left( \sum_k a_j X_{jk} b_k \right) \right) = \sum_{j,k} a_j b_k \frac{\partial X_{jk}}{\partial X_{pq}} = a_p b_q = (\mathbf{a} \mathbf{b}^T)_{pq}$$

as clearly,  $\frac{\partial X_{jk}}{\partial X_{pq}}$  is 1 if  $j = p$ ,  $k = q$ , and 0 otherwise. Hence,

$$\frac{d}{d\mathbf{X}} (\mathbf{a}^T \mathbf{X} \mathbf{b}) = \mathbf{a} \mathbf{b}^T \in \mathbb{R}^{n \times m}$$