The Australian National University Semester 2, 2024
School of Computing Tutorial 5

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 $f : \mathbb{R} \to \mathbb{R}$ is given by

$$\frac{\mathrm{d}f}{\mathrm{d}x} := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

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

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

Note that $\frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}}$ is a row vector, where each element is a function of the form $\mathbb{R}^n \to \mathbb{R}$. We write $\nabla_{\mathbf{x}} f \in (\mathbb{R}^n \to \mathbb{R})^{1 \times n}$. 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 $\mathbf{g} : \mathbb{R} \to \mathbb{R}^n$ be a function of a scalar t. The derivative of $\mathbf{g}(t)$ with respect to t is defined as

$$\frac{\mathrm{d}\mathbf{g}}{\mathrm{d}t} := \begin{bmatrix} \frac{\mathrm{d}g_1(t)}{\mathrm{d}t} \\ \frac{\mathrm{d}g_2(t)}{\mathrm{d}t} \\ \vdots \\ \frac{\mathrm{d}g_n(t)}{\mathrm{d}t} \end{bmatrix} \in (\mathbb{R} \to \mathbb{R})^{n \times 1}$$

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

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

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

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

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

and we can define their derivatives as

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

and

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

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

(Here, the term $\frac{\partial f(\mathbf{g})}{\partial g_1}$ 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^{th} component of $\mathbf{g}$.)

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

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

.

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.

Ensure you are comfortable with the preliminary problems, some of which are being provided with solutions. The tutorial will primarily focus on the advanced questions.

Question 1 Formal definition of derivative

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

Solution.

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

Question 2 Vector Derivative of Scalar Function

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

Solution.

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

Hence,

$$\frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}} = \begin{bmatrix} 2x_2 + 1, & 2x_1 + 3 \end{bmatrix} \in \mathbb{R}^{1\times 2}$$

Question 3 Scalar Derivative of Vector Function

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

Solution.

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

Question 4 Derivative of the L2 Norm

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

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,

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

Question 5 Chain Rule, Scalar Derivative

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

1. Compute $\frac{\mathrm{d}h}{\mathrm{d}t}$ by using the chain rule.

   Solution. The chain rule here is given by

$$\frac{\mathrm{d}h}{\mathrm{d}t} = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{g}}\frac{\mathrm{d}\mathbf{g}}{\mathrm{d}t}$$

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

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

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

   Hence,

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

   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 $\frac{\mathrm{d}h}{\mathrm{d}t}$ 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.

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

   which matches the answer above.

Question 6 **Chain Rule, Vector Derivative**

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

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

   Solution. Using the chain rule, we have

$$\frac{\mathrm{d}\mathbf{k}}{\mathrm{d}\mathbf{x}} = \frac{\mathrm{d}\mathbf{g}}{\mathrm{d}f}\frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}}$$

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

   Hence,

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

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

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

2. Compute $\frac{\mathrm{d}\mathbf{k}}{\mathrm{d}\mathbf{x}}$ 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:

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

   which matches the above.

Question 7 **More Derivatives**

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

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

$$\frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}} = \frac{\mathrm{d}g}{\mathrm{d}h}\frac{\mathrm{d}h}{\mathrm{d}\mathbf{x}}$$

$$\frac{\mathrm{d}g}{\mathrm{d}h} = \frac{\mathrm{d}g(h)}{\mathrm{d}h} = \frac{\mathrm{d}}{\mathrm{d}h}(h+1)^2 = 2(h+1)$$

$$\frac{\mathrm{d}h}{\mathrm{d}\mathbf{x}} = \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\mathbf{x}^T\mathbf{x} = 2\mathbf{x}^T \text{ (by .)}$$

Hence,

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

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.

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

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

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 $\mathbf{x}$.

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

Hence,

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

Question 9 Linear Regression

Let $\mathbf{\Phi} \in \mathbb{R}^{n\times m}, \mathbf{w} \in \mathbb{R}^{n\times 1}, \mathbf{t} \in \mathbb{R}^{m\times 1}$.
Let $f : \mathbb{R}^n \to \mathbb{R}, f(\mathbf{w}) = \left\|((\mathbf{w}^T\mathbf{\Phi})^T - \mathbf{t})\right\|_2^2$

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

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

   Solution. Note that $(\mathbf{w}^T\mathbf{\Phi})^T = \mathbf{\Phi}^T\mathbf{w}$.
   Let $g : \mathbb{R}^{m\times 1} \to \mathbb{R}, g(\mathbf{x}) = \|\mathbf{x}\|_2^2 := \mathbf{x}^T\mathbf{x}$ and
   $\mathbf{h} : \mathbb{R}^{n\times 1} \to \mathbb{R}^{m\times 1}, \mathbf{h}(\mathbf{w}) = (\mathbf{\Phi}^T\mathbf{w}) - \mathbf{t}$. Then $f(\mathbf{w}) = g(\mathbf{h}(\mathbf{w}))$. Apply chain rule.

$$\frac{\mathrm{d}f}{\mathrm{d}\mathbf{w}} = \frac{\mathrm{d}g}{\mathrm{d}\mathbf{h}}\frac{\mathrm{d}\mathbf{h}}{\mathrm{d}\mathbf{w}}$$

From , we have

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

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

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

Hence,

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

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?[1] 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 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 $\frac{\mathrm{d}}{\mathrm{d}\mathbf{X}}\mathbf{a}^T\mathbf{X}\mathbf{b}$.

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

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

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

$$\frac{\partial}{\partial X_{pq}}(\mathbf{a}^T\mathbf{X}\mathbf{b}) = \frac{\partial}{\partial X_{pq}}(\sum_j(\sum_k a_jX_{jk}b_k)) = \sum_{j,k} a_jb_k\frac{\partial X_{jk}}{\partial X_{pq}} = a_pb_q = (\mathbf{a}\mathbf{b}^T)_{pq}$$

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

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

[1] 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.