1. Question: Derivatives of functions taking scalars as inputs (*elementary*)

1.1. Calculate the gradient of the following two functions

(i) $F : \mathbb{R} \longrightarrow \mathbb{R}^2$

$$F(x) = \begin{pmatrix} x^3 \\ 2e^x \end{pmatrix}.$$

(ii) $G : \mathbb{R} \longrightarrow \mathbb{R}^3$

$$G(x) = \begin{pmatrix} 0 \\ x^3 + 2x^2 \\ \cos(x) \end{pmatrix}.$$

1.2. Calculate the gradient of the following two functions

(i) $F : \mathbb{R} \longrightarrow \mathbb{R}^{2\times 3}$

$$F(x) = \begin{pmatrix} x^2 & 2e^x & 0 \\ 0 & x & \ln(x) \end{pmatrix}.$$

(ii) $G : \mathbb{R} \longrightarrow \mathbb{R}^{3\times 2}$

$$G(x) = \begin{pmatrix} 5x & \sin(x) \\ 2 & x^3 + 2x^2 \\ x^2 + 3x & 1 \end{pmatrix}.$$

1.3. Consider two functions $\boldsymbol{f} : \mathbb{R} \longrightarrow \mathbb{R}^n$ and $\boldsymbol{g} : \mathbb{R} \longrightarrow \mathbb{R}^n$. Verify the general **sum rule** and **product rule** for these two functions.

Solution:

1.1. (i) The gradient of F is

$$\frac{\partial F(x)}{\partial x} = \begin{pmatrix} 2x \\ 2e^x \end{pmatrix}.$$

(ii) The gradient of G is

$$\frac{\partial G(x)}{\partial x} = \begin{pmatrix} 0 \\ 3x^2 + 4x \\ -\sin(x) \end{pmatrix}.$$

1.2. (i) The gradient of F is

$$\frac{\partial F(x)}{\partial x} = \begin{pmatrix} 2x & 2e^x & 0 \\ 0 & 1 & 1/x \end{pmatrix}.$$

(ii) The gradient of G is

$$\frac{\partial G(x)}{\partial x} = \begin{pmatrix} 5 & \cos(x) \\ 0 & 3x^2 + 4x \\ 2x + 3 & 0 \end{pmatrix}.$$

1.3. We start by writing

$$\boldsymbol{f}(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix} \quad \text{and} \quad \boldsymbol{g}(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_n(x) \end{pmatrix}.$$

The general sum rule is easily proven:

$$\frac{\partial}{\partial x}(\boldsymbol{f}(x) + \boldsymbol{g}(x)) = \frac{\partial}{\partial x}\begin{pmatrix} f_1(x) + g_1(x) \\ \vdots \\ f_n(x) + g_n(x) \end{pmatrix} \overset{\text{Univariate sum rule}}{=} \begin{pmatrix} \frac{\partial}{\partial x} f_1(x) + \frac{\partial}{\partial x} g_1(x) \\ \vdots \\ \frac{\partial}{\partial x} f_n(x) + \frac{\partial}{\partial x} g_n(x) \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial}{\partial x} f_1(x) \\ \vdots \\ \frac{\partial}{\partial x} f_n(x) \end{pmatrix} + \begin{pmatrix} \frac{\partial}{\partial x} g_1(x) \\ \vdots \\ \frac{\partial}{\partial x} g_n(x) \end{pmatrix} = \frac{\partial \boldsymbol{f}}{\partial x} + \frac{\partial \boldsymbol{g}}{\partial x}.$$

And the product rule does not take much more:

$$\begin{aligned}
\frac{\partial}{\partial x}(\boldsymbol{f}(x)\boldsymbol{g}(x)) &= \frac{\mathrm{d}}{\partial x}\left(\sum_{i=1}^{n} f_i(x)g_i(x)\right) = \sum_{i=1}^{n} \frac{\partial}{\partial x}\left(f_i(x)g_i(x)\right) \\
&= \sum_{i=1}^{n}\left(\frac{\partial}{\partial x}\left(f_i(x)\right)g_i(x) + f_i(x)\frac{\partial}{\partial x}\left(g_i(x)\right)\right) \\
&= \sum_{i=1}^{n}\frac{\partial}{\partial x}\left(f_i(x)\right)g_i(x) + \sum_{i=1}^{n} f_i(x)\frac{\partial}{\partial x}\left(g_i(x)\right) \\
&= \frac{\partial \boldsymbol{f}}{\partial x}\boldsymbol{g}(x) + \boldsymbol{f}(x)\frac{\partial \boldsymbol{g}}{\partial x}.
\end{aligned}$$

2. Question: Derivatives of functions taking vectors as inputs *(elementary)*

2.1. Calculate the Jacobian matrix of the following two functions

(i) $F : \mathbb{R}^2 \to \mathbb{R}^3$ where:

$$F(x, y) = \begin{bmatrix} x^2 + \sin(x) \\ x(y-2) \\ y^2 - 3xy \end{bmatrix}$$

(ii) $G : \mathbb{R}^3 \to \mathbb{R}^2$ where:

$$G(x, y, z) = \begin{bmatrix} x^2 - y^2 \\ 3xyz - 5 \end{bmatrix}$$

2.2. Determine the gradient $\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}}$ of the following function, where $M, N \in \mathbb{N}_{>0}$

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}, \quad \boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^M, \quad \boldsymbol{A} \in \mathbb{R}^{M\times N}, \quad \boldsymbol{x} \in \mathbb{R}^N.$$

2.3. Consider the function $h : \mathbb{R} \to \mathbb{R}, h(t) = (f \circ g)(t)$ with

$$\begin{aligned}
f &: \mathbb{R}^2 \to \mathbb{R} \\
g &: \mathbb{R} \to \mathbb{R}^2 \\
f(\boldsymbol{x}) &= \exp\left(x_1 x_2^2\right), \\
\boldsymbol{x} &= \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = g(t) = \begin{bmatrix} t\cos t \\ t\sin t \end{bmatrix}
\end{aligned}$$

and compute the gradient of h with respect to t.

2.4. Use the chain rule, **both according to Proposition 7.1 and according to Remark 7.1**, to find the gradient of

$$F : \mathbb{R}^3 \longrightarrow \mathbb{R}, \quad (x, y, z) \mapsto f \circ \varphi\,(x, y, z)$$

for

$$\varphi : \mathbb{R}^3 \longrightarrow \mathbb{R}^3, \quad (x, y, z) \mapsto (h(x), g(x, y), z)$$

and scalar-valued functions f, g, and h defined as $f(x, y, z) := x^2 + yz$, $g(x, y) := y^3 + xy$, and $h(x) := \sin x$.

Solution:

2.1. (i) The Jacobian matrix is:

$$\boldsymbol{J}_F(x,y) = \begin{bmatrix} 2x+\cos(x) & 0 \\ y-2 & x \\ -3y & 2y-3x \end{bmatrix}$$

(ii) The Jacobian is:

$$\boldsymbol{J}_G(x,y,z) = \begin{bmatrix} 2x & -2y & 0 \\ 3yz & 3xz & 3xy \end{bmatrix}$$

2.2. To compute the gradient $\mathrm{d}\boldsymbol{f}/\mathrm{d}\boldsymbol{x}$ we first determine the dimension of $\mathrm{d}\boldsymbol{f}/\mathrm{d}\boldsymbol{x}$: Since $\boldsymbol{f}:\mathbb{R}^N \to \mathbb{R}^M$, it follows that $\mathrm{d}\boldsymbol{f}/\mathrm{d}\boldsymbol{x} \in \mathbb{R}^{M\times N}$. Second, to compute the gradient we determine the partial derivatives of f with respect to every x_j :

$$f_i(\boldsymbol{x}) = \sum_{j=1}^{N} A_{ij}x_j \Longrightarrow \frac{\partial f_i}{\partial x_j} = A_{ij}$$

We collect the partial derivatives in the Jacobian and obtain the gradient

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & & \vdots \\ A_{M1} & \cdots & A_{MN} \end{bmatrix} = \boldsymbol{A} \in \mathbb{R}^{M\times N}.$$

2.3. Since $f:\mathbb{R}^2 \to \mathbb{R}$ and $g:\mathbb{R}\to\mathbb{R}^2$ we note that

$$\frac{\partial f}{\partial \boldsymbol{x}} \in \mathbb{R}^{1\times 2}, \quad \frac{\partial g}{\partial t} \in \mathbb{R}^{2\times 1}$$

The desired gradient is computed by applying the chain rule:

$$\begin{aligned} \frac{\mathrm{d}h}{\mathrm{d}t} &= \frac{\partial f}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial t} \end{bmatrix} \\ &= \begin{bmatrix} \exp\left(x_1x_2^2\right)x_2^2 & 2\exp\left(x_1x_2^2\right)x_1x_2 \end{bmatrix} \begin{bmatrix} \cos t - t\sin t \\ \sin t + t\cos t \end{bmatrix} \\ &= \exp\left(x_1x_2^2\right)\left(x_2^2(\cos t - t\sin t) + 2x_1x_2(\sin t + t\cos t)\right), \end{aligned}$$

where $x_1 = t\cos t$ and $x_2 = t\sin t$.

2.4. Here $f\circ\varphi(x,y,z) = f(h(x),g(x,y),z) = h(x)^2 + g(x,y)z$. The chain rule gives

$$\begin{aligned} \frac{\partial F}{\partial x} &= \frac{\partial f}{\partial h}\frac{\partial h}{\partial x} + \frac{\partial f}{\partial g}\frac{\partial g}{\partial x} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial x} = 2\sin x\cos x + zy + 0 \\ \frac{\partial F}{\partial y} &= \frac{\partial f}{\partial h}\frac{\partial h}{\partial y} + \frac{\partial f}{\partial g}\frac{\partial g}{\partial y} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial y} = 0 + z\left(3y^2+x\right) + 0 \\ \frac{\partial F}{\partial z} &= \frac{\partial f}{\partial h}\frac{\partial h}{\partial z} + \frac{\partial f}{\partial g}\frac{\partial g}{\partial z} + \frac{\partial f}{\partial z}\frac{\partial z}{\partial z} = 0 + 0 + \left(y^3 + xy\right) \end{aligned}$$

Therefore $\boldsymbol{J}_F(x,y,z) = \begin{pmatrix} 2\sin x\cos x + zy & xz + 3y^2z & y^3 + xy \end{pmatrix}$. Alternatively, we can use Jacobean matrices: $\boldsymbol{J}_F(\boldsymbol{x}) = \boldsymbol{J}_{f\circ\varphi}(\boldsymbol{x}) = \boldsymbol{J}_f(\varphi(\boldsymbol{x}))\circ\boldsymbol{J}_\varphi(\boldsymbol{x})$. In this case

$$\boldsymbol{J}_F(x,y,z) = \begin{pmatrix} 2h(x) & z & g(x,y) \end{pmatrix} \cdot \begin{pmatrix} \cos x & 0 & 0 \\ y & 3y^2+x & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and we get the same answer as before.

3. Question: Derivatives of functions taking matrices as inputs (*elementary*)

Note that the Booklet only contains instructions on taking the derivative of scalar-valued functions taking matrices as inputs (matrix norms being a common case). If you are interested in the derivation of vector and matrix valued functions taking matrices as indices, see Examples 5.12 *and* 5.13 *of* Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). Mathematics for Machine Learning

3.1. For a matrix $A \in \mathbb{R}^{m \times n}$, the *Frobenius norm* is defined as $\|\boldsymbol{X}\|_F := \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}$.

Calculate the gradient of the squared Frobenius norm, i.e. the function

$$f : \mathbb{R}^{m \times n} \longrightarrow \mathbb{R}, \quad \boldsymbol{X} \mapsto \|\boldsymbol{X}\|_F^2.$$

3.2. Prove the following identities

(i) $\nabla_{A^T} f(A) = (\nabla_A f(A))^T$, for a differentiable function $f : \mathbb{R}^{m \times n} \longrightarrow \mathbb{R}$, $m, n \in \mathbb{N}_{>0}$.

(ii) $\nabla_A \operatorname{tr}(AB) = B^T$.

Solution:

3.1. Since matrix norms are scalar valued functions, we must simply compute the matrix of partial derivatives from definition 7.4.
Since $f(\boldsymbol{X}) = \left(\sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}\right)^2 = \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2$, it immediately follows that $\dfrac{\partial f(\boldsymbol{X})}{\partial x_{ij}} = 2x_{ij} \ \forall i \in \{1, \ldots, m\}, j \in \{1, \ldots, n\}$. Therefore, thre gradient of the squared Frobenius norm is

$$\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}} = \begin{pmatrix} \dfrac{\partial f(\boldsymbol{X})}{\partial x_{11}} & \cdots & \dfrac{\partial f(\boldsymbol{X})}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f(\boldsymbol{X})}{\partial x_{m1}} & \cdots & \dfrac{\partial f(\boldsymbol{X})}{\partial x_{mn}} \end{pmatrix} = \begin{pmatrix} 2x_{11} & \cdots & 2x_{1n} \\ \vdots & \ddots & \vdots \\ 2x_{m1} & \cdots & 2x_{mn} \end{pmatrix} = 2\boldsymbol{X}.$$

3.2. (i)

$$\nabla_{A^T} f(A) = \begin{bmatrix} \dfrac{\partial f(A)}{\partial a_1} & \dfrac{\partial f(A)}{\partial a_1} & \cdots & \dfrac{\partial f(A)}{\partial a_{n1}} \\ \dfrac{\partial f(A)}{\partial a_{12}} & \dfrac{\partial f(A)}{\partial a_{22}} & \cdots & \dfrac{\partial f(A)}{\partial a_{n2}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f(A)}{\partial a_{1n}} & \dfrac{\partial f(A)}{\partial a_{2n}} & \cdots & \dfrac{\partial f(A)}{\partial a_{nn}} \end{bmatrix} = (\nabla_A f(A))^T.$$

(ii)

$$\operatorname{tr}(AB) = \operatorname{tr} \begin{bmatrix} \longleftarrow \overrightarrow{a_1} \longrightarrow \\ \longleftarrow \overrightarrow{a_2} \longrightarrow \\ \vdots \\ \longleftarrow \overrightarrow{a_n} \longrightarrow \end{bmatrix} \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \overrightarrow{b_1} & \overrightarrow{b_2} & \cdots & \overrightarrow{b_n} \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} = \sum_{i=1}^m a_{1i} b_{i1} + \sum_{i=1}^m a_{2i} b_{i2} + \ldots + \sum_{i=1}^m a_{ni} b_{in}$$

$$\Rightarrow \frac{\partial \operatorname{tr}(AB)}{\partial a_{ij}} = b_{ji}$$

$$\Rightarrow \nabla_A \operatorname{tr}(AB) = B^T.$$

4. Question: Directional derivative *(a bit more advanced)*

Evaluating partial derivatives only gives us the slope of a function in the direction of one of the inputs, or, equivalently, the direction of the corresponding canonical vector. (A canonical vector is a vector each of whose components are all zero, except one that equals 1.)
If we are interested in the slope of a function in the direction of a non-canonical vector, i.e. when changing several inputs at once, we can use the ***directional derivative****. The directional derivative of function f at $\mathbf{x}$ along $\mathbf{u}$ is defined as*

$$D_{\mathbf{u}} f(\mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{u}) - f(\mathbf{x})}{h}.$$

For differentiable functions f and unit vector $\mathbf{u}$, i.e. $\|\mathbf{u}\| = 1$, the directional derivative is simply computed as $D_{\mathbf{u}} f(\mathbf{x}) = \nabla f(\mathbf{x})\mathbf{u}$.

4.1. Evaluate the directional derivative $D_{\mathbf{u}} f(\mathbf{x})$ for the following

(i) $f(x, y) = e^x \cos(\pi y)$, $\mathbf{x} = (0, -1)^\top$ and $\mathbf{u} = \left(-\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}\right)^\top$.

(ii) $f(x, y) = xy^2 + x^3 y$, $\mathbf{x} = (4, -2)^\top$ and $\mathbf{u} = \left(\frac{1}{\sqrt{10}}, \frac{3}{\sqrt{10}}\right)^\top$

4.2. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called homogeneous of degree m if $f(tx) = t^m f(x)$ for all $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$. If f is differentiable, show that for $x \in \mathbb{R}^n$,

$$\nabla f(x) x = m f(x), \text{ that is, } \sum_{i=1}^{n} x_i \frac{\partial f}{\partial x_i} = m f(x).$$

Show that maps multilinear in k variables, which are characterized by the following property

$$\begin{aligned} &L\left(x_1, \ldots, x_{i-1}, \alpha u + \beta w, x_{i+1}, \ldots, x_n\right) \\ =&\alpha L\left(x_1, \ldots, x_{i-1}, u, x_{i+1}, \ldots, x_n\right) + \beta\left(x_1, \ldots, x_{i-1}, w, x_{i+1}, \ldots, x_n\right) \end{aligned}$$

give rise to homogeneous functions of degree k. Give other examples.

Solution:

4.1.

(i) We have

$$(\nabla f)(x, y) = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) = \left(e^x \cos(\pi y), -\pi e^x \sin(\pi y)\right)$$

and thus if we evaluate at $(0, -1)$ we find

$$(\nabla f)(0, -1) = (-1, 0)$$

Since $\mathbf{u}$ is a unit vector and f differentiable, the directional derivative in general is $(\nabla f)(x_1, x_2) \cdot \mathbf{u}$, so for this problem the answer is

$$(-1, 0) \cdot \left(-\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}\right) = \frac{1}{\sqrt{5}}.$$

(ii) We have

$$(\nabla f)(x, y) = \left(y^2 + 3x^2 y, 2xy + x^3\right)$$

and thus if we evaluate at $(4, -2)$ we find

$$(\nabla f)(4, -2) = (-92, 48)$$

Again, since $\mathbf{u}$ is a unit vector and f differentiable, the directional derivative in general is $(\nabla f)(x_1, x_2) \cdot \mathbf{u}$, so for this problem the answer is

$$(-92, 48) \cdot \left(\frac{1}{\sqrt{10}}, \frac{3}{\sqrt{10}}\right) = \frac{52}{\sqrt{10}}.$$

4.2. By definition of the directional derivative,

$$\nabla f(x)x = \lim_{h\to 0} \frac{f(x+hx) - f(x)}{h} = \lim_{h\to 0} \frac{f((1+h)x) - f(x)}{h}$$

Using the fact that f is homogeneous of degree m, we get

$$\begin{aligned}
\nabla f(x)x &= \lim_{h\to 0} \frac{(1+h)^m f(x) - f(x)}{h} = \lim_{h\to 0} \left(\frac{(1+h)^m - 1}{h}\right) f(x) \\
&= \lim_{h\to 0} \left(\frac{1^m + \binom{m}{1}h + \binom{m}{2}h^2 + \cdots + \binom{m}{m}h^m - 1}{h}\right) f(x) \\
&= \lim_{h\to 0} \left(m + \binom{m}{2}h + \cdots + \binom{m}{m}h^{m-1}\right) f(x) = mf(x)
\end{aligned}$$

as desired. k-linear maps are characterized by the property

$$\begin{aligned}
&L\left(x_1, \ldots, x_{i-1}, \alpha u + \beta w, x_{i+1}, \ldots, x_n\right) \\
=&\alpha L\left(x_1, \ldots, x_{i-1}, u, x_{i+1}, \ldots, x_n\right) + \beta\left(x_1, \ldots, x_{i-1}, w, x_{i+1}, \ldots, x_n\right)
\end{aligned}$$

If we define $g(x) = L(\underbrace{x, \ldots, x}_{k \text{ times}})$, then it follows that

$$g(tx) = L(tx, \ldots, tx) = t^k L(x, \ldots, x) = t^k g(x)$$

Therefore, maps multilinear in k variables give rise to homogeneous functions of degree k. An example of a non-linear homogeneous function is $f(x, y) = x^2 + y^2$. This is homogeneous of degree 2 since $f(kx, ky) = k^2\left(x^2 + y^2\right) = k^2 f(x, y)$.

If you have any questions or feedback, please feel free to contact me via E-mail at hannah.kuempel@stat.uni-muenchen.de!!

Also, thank you to the authors of the books *Mathematics for Machine Learning* as well as Steven J. Miller and Anthony Varilly whose exercises this sheet was inspired by.