conventions for tensor calculus

a vector $\vec{x} \in \mathbb{R}^n$ is a column vector of size n:

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

a matrix $A \in \mathbb{R}^{n \times m}$ is of shape $n \times m$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & & & \\ \vdots & & & \\ a_{n1} & \cdots & \cdots & a_{nm} \end{bmatrix}$$

note that a vector $\vec{x} \in \mathbb{R}^n$ is also an $n \times 1$ matrix.

use the following conventions for derivatives involving matrices and vectors:

1) if $a \in \mathbb{R}$ and $\vec{x} \in \mathbb{R}^n$, then $\frac{\partial a}{\partial \vec{x}} = \left[\frac{\partial a}{\partial x_1}, \frac{\partial a}{\partial x_2}, \ldots, \frac{\partial a}{\partial x_n} \right]$

so. $\frac{\partial a}{\partial \vec{x}}$ is a row vector, i.e $\frac{\partial a}{\partial \vec{x}} \in \mathbb{R}^{1 \times n}$

2) if $\vec{y} \in \mathbb{R}^m$ and $\vec{x} \in \mathbb{R}^n$ then

$\left[\frac{\partial \vec{y}}{\partial \vec{x}} \right]_{ij} = \frac{\partial y_i}{\partial x_j}$, so $\frac{\partial \vec{y}}{\partial \vec{x}} \in \mathbb{R}^{m \times n}$ (shape $m \times n$)

3.) if $a \in \mathbb{R}$ and $A \in \mathbb{R}^{n \times m}$, then

$$\left[\frac{\partial a}{\partial A}\right]_{ij} = \frac{\partial a}{\partial A_{ji}}, \text{ so } \frac{\partial a}{\partial A} \in \mathbb{R}^{m \times n}$$

note: the book "mathematics for machine learning" uses the opposite convention: $\left[\frac{\partial a}{\partial A}\right]_{ij} = \frac{\partial a}{\partial A_{ij}}$ (*)

this is an inconvenient choice as it doesn't have (1) as a special case.
Meaning: if I consider a vector $\vec{x} \in \mathbb{R}^{n \times 1}$ as an $n \times 1$ matrix, and insert this into (*), I do not get back (1).

4). $$\nabla_x f(x) = \frac{\partial f(x)}{\partial x}$$

so the shape of the gradient is the same as the shape of the derivative.
Note: officially this is not true, for instance $\nabla_x f(x)$ should be a column if $f: \mathbb{R}^n \to \mathbb{R}$ and $x \in \mathbb{R}^n$, and $\frac{\partial f(x)}{\partial x}$ a row vector. To avoid confusion we will not make this distinction in this class.
so for ML1 $\nabla_x f(x)$ is also a row vector in the above example.