1 Notation

- Scalars are written as lower case letters.
- Vectors are written as lower case bold letters, such as $\mathbf{x}$, and can be either row (dimensions $1 \times n$) or column (dimensions $n \times 1$). Column vectors are the default choice, unless otherwise mentioned. Individual elements are indexed by subscripts, such as $x_i \ (i \in \{1, \cdots, n\})$.
- Matrices are written as upper case bold letters, such as $\mathbf{X}$, and have dimensions $m \times n$ corresponding to $m$ rows and $n$ columns. Individual elements are indexed by double subscripts for row and column, such as $X_{ij} \ (i \in \{1, \cdots, m\}, \ j \in \{1, \cdots, n\})$.
- Occasionally higher order tensors occur, such as 3rd order with dimensions $m \times n \times p$, etc.

Note that a matrix is a 2nd order tensor. A row vector is a matrix with 1 row, and a column vector is a matrix with 1 column. A scalar is a matrix with 1 row and 1 column. Essentially, scalars and vectors are special cases of matrices.

The derivative of $f$ with respect to $x$ is $\frac{\partial f}{\partial x}$. Both $x$ and $f$ can be a scalar, vector, or matrix, leading to 9 types of derivatives. The gradient of $f$ w.r.t $x$ is $\nabla_x f = \left( \frac{\partial f}{\partial x} \right)^T$, i.e. gradient is transpose of derivative. The gradient at any point $x_0$ in the domain has a physical interpretation, its direction is the direction of maximum increase of the function $f$ at the point $x_0$, and its magnitude is the rate of increase in that direction. We do not generally deal with the gradient when $x$ is a scalar.

2 Basic Rules

This document follows numerator layout convention. There is an alternative denominator layout convention, where several results are transposed. Do not mix different layout conventions.

We’ll first state the most general matrix-matrix derivative type. All other types are simplifications since scalars and vectors are special cases of matrices. Consider a function $F(\cdot)$ which maps $m \times n$ matrices to $p \times q$ matrices, i.e. domain $\subset \mathbb{R}^{m\times n}$ and range $\subset \mathbb{R}^{p\times q}$. So, $F(\cdot) : \mathbf{X} \rightarrow F(\mathbf{X})$. Its derivative $\frac{\partial F}{\partial \mathbf{X}}$ is a 4th order tensor of dimensions $p \times q \times n \times m$. This is an outer matrix of dimensions $n \times m$ (transposed dimensions of the denominator $\mathbf{X}$), with
each element being a $p \times q$ inner matrix (same dimensions as the numerator $F$). It is given as:

$$
\frac{\partial F}{\partial X} = \begin{bmatrix}
\frac{\partial F}{\partial X_{1,1}} & \cdots & \frac{\partial F}{\partial X_{1,m}} \\
\vdots & \ddots & \vdots \\
\frac{\partial F}{\partial X_{1,n}} & \cdots & \frac{\partial F}{\partial X_{m,n}} 
\end{bmatrix} \tag{1a}
$$

which has $n$ rows and $m$ columns, and the $(i,j)$th element is given as:

$$
\frac{\partial F}{\partial X_{i,j}} = \begin{bmatrix}
\frac{\partial F_{1,1}}{\partial X_{i,j}} & \cdots & \frac{\partial F_{1,q}}{\partial X_{i,j}} \\
\vdots & \ddots & \vdots \\
\frac{\partial F_{p,1}}{\partial X_{i,j}} & \cdots & \frac{\partial F_{p,q}}{\partial X_{i,j}} 
\end{bmatrix} \tag{1b}
$$

which has $p$ rows and $q$ columns.

Whew! Now that that’s out of the way, let’s get to some general rules (for the following, $x$ and $y$ can represent scalar, vector or matrix):

- The derivative $\frac{\partial y}{\partial x}$ always has outer matrix dimensions = transposed dimensions of denominator $x$, and each individual element (inner matrix) has dimensions = same dimensions of numerator $y$. If you do a calculation and the dimension doesn’t come out right, the answer is not correct.
- Derivatives usually obey the chain rule, i.e. $\frac{\partial f\left(g(x)\right)}{\partial x} = \frac{\partial f}{\partial g(x)} \frac{\partial g(x)}{\partial x}$.
- Derivatives usually obey the product rule, i.e. $\frac{\partial f(x)g(x)}{\partial x} = f(x)\frac{\partial g(x)}{\partial x} + g(x)\frac{\partial f(x)}{\partial x}$.

3 Types of derivatives

3.1 Scalar by scalar

Nothing special here. The derivative is a scalar, and can also be written as $f'(x)$. For example, if $f(x) = \sin x$, then $f'(x) = \cos x$.

3.2 Scalar by vector

$f(\cdot) : \mathbb{R}^m \rightarrow \mathbb{R}^n$. For this, the derivative is a $1 \times m$ row vector:

$$
\frac{\partial f}{\partial x} = \begin{bmatrix}
\frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_m}
\end{bmatrix} \tag{2}
$$

The gradient $\nabla_x f$ is its transposed column vector.
3.3 Vector by scalar

\( f(\cdot): x_{1 \times 1} \rightarrow f(x)_{n \times 1} \). For this, the derivative is a \( n \times 1 \) column vector:

\[
\frac{\partial f}{\partial x} = \begin{bmatrix}
\frac{\partial f_1}{\partial x} \\
\frac{\partial f_2}{\partial x} \\
\vdots \\
\frac{\partial f_n}{\partial x}
\end{bmatrix}
\]  

(3)

3.4 Vector by vector

\( f(\cdot): x_{m \times 1} \rightarrow f(x)_{n \times 1} \). Derivative, also known as the Jacobian, is a matrix of dimensions \( n \times m \). Its \((i, j)\)th element is the scalar derivative of the \( i \)th output component w.r.t the \( j \)th input component, i.e.:

\[
\frac{\partial f}{\partial x} = \begin{bmatrix}
\frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} \\
\vdots & \ddots & \vdots \\
\frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_m}
\end{bmatrix}
\]  

(4)

3.4.1 Special case – Vectorized scalar function

This is a scalar-scalar function applied element-wise to a vector, and is denoted by \( f(\cdot): x_{m \times 1} \rightarrow f(x)_{m \times 1} \). For example:

\[
f \left( \begin{bmatrix}
x_1 \\
x_2 \\
\vdots \\
x_m
\end{bmatrix} \right) = \begin{bmatrix}
f(x_1) \\
f(x_2) \\
\vdots \\
f(x_m)
\end{bmatrix}
\]  

(5)

In this case, both the derivative and gradient are the same \( m \times m \) diagonal matrix, given as:

\[
\nabla_x f = \frac{\partial f}{\partial x} = \begin{bmatrix}
f'(x_1) & 0 \\
f'(x_2) & \ddots \\
0 & \ddots & 0 \\
0 & \cdots & f'(x_m)
\end{bmatrix}
\]  

(6)

where \( f'(x_i) = \frac{\partial f(x_i)}{\partial x_i} \).
Note: Some texts take the derivative of a vectorized scalar function by taking element-wise derivatives to get a $m \times 1$ vector. To avoid confusion with (6), we will refer to this as $f'(x)$.

$$f'(x) = \begin{bmatrix} f'(x_1) \\ f'(x_2) \\ \vdots \\ f'(x_m) \end{bmatrix} \quad (7)$$

To realize the effect of this, let’s say we want to multiply the gradient from (6) with some $m$-dimensional vector $a$. This would result in:

$$\left( \nabla_x f \right) a = \begin{bmatrix} f'(x_1) a_1 \\ f'(x_2) a_2 \\ \vdots \\ f'(x_m) a_m \end{bmatrix} \quad (8)$$

Achieving the same result with $f'(x)$ from (7) would require the Hadamard product $\circ$, defined as element-wise multiplication of 2 vectors:

$$f'(x) \circ a = \begin{bmatrix} f'(x_1) a_1 \\ f'(x_2) a_2 \\ \vdots \\ f'(x_m) a_m \end{bmatrix} \quad (9)$$

### 3.4.2 Special Case – Hessian

Consider the type of function in Sec. 3.2 i.e. $f(\cdot) : \mathbb{R}^{m \times 1} \rightarrow \mathbb{R}^{1 \times 1}$. Its gradient is a vector-to-vector function given as $\nabla_x f(\cdot) : \mathbb{R}^{m \times 1} \rightarrow \mathbb{R}^{m \times 1}$. The transpose of its derivative is the Hessian:

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_m \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_m^2} \end{bmatrix} \quad (10)$$

i.e. $H = \left( \frac{\partial \nabla_x f}{\partial x} \right)^T$. If derivatives are continuous, then $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$, so the Hessian is symmetric.
3.5 Scalar by matrix

\( f(\cdot) : X_{m \times n} \rightarrow f(X) \). In this case, the derivative is a \( n \times m \) matrix:

\[
\frac{\partial f}{\partial X} = \begin{bmatrix}
\frac{\partial f}{\partial X_{1,1}} & \cdots & \frac{\partial f}{\partial X_{m,1}} \\
\vdots & \ddots & \vdots \\
\frac{\partial f}{\partial X_{1,n}} & \cdots & \frac{\partial f}{\partial X_{m,n}}
\end{bmatrix}
\] (11)

The gradient has the same dimensions as the input matrix, i.e. \( m \times n \).

3.6 Matrix by scalar

\( f(\cdot) : x_{1 \times 1} \rightarrow F(x) \). In this case, the derivative is a \( p \times q \) matrix:

\[
\frac{\partial F}{\partial x} = \begin{bmatrix}
\frac{\partial F_{1,1}}{\partial x} & \cdots & \frac{\partial F_{1,q}}{\partial x} \\
\vdots & \ddots & \vdots \\
\frac{\partial F_{p,1}}{\partial x} & \cdots & \frac{\partial F_{p,q}}{\partial x}
\end{bmatrix}
\] (12)

3.7 Vector by matrix

\( f(\cdot) : X_{m \times n} \rightarrow f(X) \). In this case, the derivative is a 3rd-order tensor with dimensions \( p \times n \times m \).

This is the same \( n \times m \) matrix in (11), but with \( f \) replaced by the \( p \)-dimensional vector \( f \), i.e.:

\[
\frac{\partial f}{\partial X} = \begin{bmatrix}
\frac{\partial f}{\partial X_{1,1}} & \cdots & \frac{\partial f}{\partial X_{m,1}} \\
\vdots & \ddots & \vdots \\
\frac{\partial f}{\partial X_{1,n}} & \cdots & \frac{\partial f}{\partial X_{m,n}}
\end{bmatrix}
\] (13)

3.8 Matrix by vector

\( F(\cdot) : x_{m \times 1} \rightarrow F(x) \). In this case, the derivative is a 3rd-order tensor with dimensions \( p \times q \times m \).

This is the same \( m \times 1 \) row vector in (2), but with \( f \) replaced by the \( p \times q \) matrix \( F \), i.e.:

\[
\frac{\partial F}{\partial x} = \begin{bmatrix}
\frac{\partial F}{\partial x_1} & \frac{\partial F}{\partial x_2} & \cdots & \frac{\partial F}{\partial x_m}
\end{bmatrix}
\] (14)
4 Operations and Examples

4.1 Commutation

If things normally don’t commute (such as for matrices, $AB \neq BA$), then order should be maintained when taking derivatives. If things normally commute (such as for vector inner product, $a \cdot b = b \cdot a$), their order can be switched when taking derivatives. **Output dimensions must always come out right.**

For example, let $f(x) = (a^T x)^T b$. The derivative $\frac{\partial f}{\partial x}$ should be an $n \times m$ matrix. Keeping order fixed, we get $\frac{\partial f}{\partial x} = a^T \frac{\partial x}{\partial x} b = a^T b = a^T b$. This is a scalar, which is wrong! The solution? Note that $(a^T x)$ is a scalar, which can sit either to the right or the left of vector $b$, i.e. ordering doesn’t really matter. Rewriting $f(x) = b(a^T x)$, we get $\frac{\partial f}{\partial x} = b a^T \frac{\partial x}{\partial x} = b a^T I = b a^T$, which is the correct $n \times m$ matrix.

If this seems confusing, it might be useful to take a simple example with low values for $m$ and $n$, and write out the full derivative in matrix form as shown in (4). The resulting matrix will be $b a^T$.

4.2 Derivative of a transposed vector

The derivative of a transposed vector w.r.t itself is the identity matrix, but the transpose gets applied to everything after. For example, let $f(w) = (y - w^T x)^2 = y^2 - (w^T x) y - y (w^T x) + (w^T x) (w^T x)$, where $y$ and $x$ are not a function of $w$. Taking derivative of the terms individually:

- $\frac{\partial y^2}{\partial w} = 0^T$, i.e. a row vector of all 0s.
- $\frac{\partial (w^T x) y}{\partial w} = \frac{\partial w^T}{\partial w} x y = (x y)^T = y^T x^T$. Since $y$ is a scalar, this is simply $y x^T$.
- $\frac{\partial y (w^T x)}{\partial w} = y \frac{\partial w^T}{\partial w} x = y x^T$
- $\frac{\partial (w^T x) (w^T x)}{\partial w} = \frac{\partial w^T}{\partial w} x (w^T x) + (w^T x) \frac{\partial w^T}{\partial w} x = (x^T w) x^T + (w^T x) x^T$. Since vector inner products commute, this is $2 (w^T x) x^T$.

So $\frac{\partial f}{\partial w} = -2 y x^T + 2 (w^T x) x^T$.

4.3 Dealing with tensors

A tensor of dimensions $p \times q \times n \times m$ (such as given in (4)) can be pre- and post-multiplied by vectors just like an ordinary matrix. **These vectors must be compatible with the inner matrices
of dimensions $p \times q$, i.e. for each inner matrix, pre-multiply with a $1 \times p$ row vector and post-multiply with a $q \times 1$ column vector to get a scalar. This gives a final matrix of dimensions $n \times m$.

Example: $f(W) = a^T W b$. This is a scalar, so $\frac{\partial f}{\partial W}$ should be a matrix which has transposed dimensions as $W$, i.e. $n \times m$. Now, $\frac{\partial f}{\partial W} = a^T \frac{\partial W}{\partial W} b$, where $\frac{\partial W}{\partial W}$ has dimensions $m \times n \times n \times m$. For example if $m = 3$, $n = 2$, then:

$$
\frac{\partial W}{\partial W} = \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
\end{bmatrix}
$$

Note that the $(i, j)$th inner matrix has a 1 in its $(j, i)$th position. Pre- and post-multiplying the $(i, j)$th inner matrix with $a^T$ and $b$ gives $a_j b_i$, where $i \in \{1, 2\}$ and $j \in \{1, 2, 3\}$. So:

$$
a^T \frac{\partial W}{\partial W} b = \begin{bmatrix} a_1 b_1 & a_2 b_1 & a_3 b_1 \\
a_1 b_2 & a_2 b_2 & a_3 b_2 \end{bmatrix}
$$

Thus, $\frac{\partial f}{\partial W} = b a^T$.

### 4.4 Gradient Example: L2 Norm

Problem: Given $f(x) = \|x - a\|_2$, find $\nabla_x f$.

Note that $\|x - a\|_2 = \sqrt{(x - a)^T (x - a)}$, which is a scalar. So the derivative will be a row vector and gradient will be a column vector of the same dimension as $x$. Let’s use the chain rule:

$$
\frac{\partial f}{\partial x} = \frac{\partial \sqrt{(x - a)^T(x - a)}}{\partial (x - a)^T(x - a)} \times \frac{\partial (x - a)^T(x - a)}{\partial x}
$$

The first term is a scalar-scalar derivative equal to $\frac{1}{2\sqrt{(x - a)^T(x - a)}}$. The second term is:

$$
\frac{\partial (x - a)^T(x - a)}{\partial x} = \frac{\partial (x^T x - a^T x - x^T a + a^T a)}{\partial x} = (x^T + x^T) - a^T - a^T + 0^T = 2(x^T - a^T)
$$

So $\frac{\partial f}{\partial x} = \frac{a^T - a^T}{\sqrt{(x - a)^T(x - a)}}$.

So $\nabla_x f = \frac{x - a}{\|x - a\|_2}$, which is basically the unit displacement vector from $a$ to $x$. This means that to get maximum increase in $f(x)$, one should move away from $a$ along the straight line joining $a$ and $x$. Alternatively, to get maximum decrease in $f(x)$, one should move from $x$ directly towards $a$, which makes sense geometrically.
5 Notes and Further Reading

The chain rule and product rule do not always hold when dealing with matrices. However, some modified forms can hold when using the \emph{Trace}(\cdot) function. For a full list of derivatives, the reader should consult a textbook or websites such as [Wikipedia’s page on Matrix calculus](https://en.wikipedia.org/wiki/Matrix_calculus). Keep in mind that some texts may use denominator layout convention, where results will look different.