Jacobians and Derivative Rules

Last modified: 2026-11-24 09:19:45 (PST)

Jacobians and derivative rules

Definition 0.1 (Row-vector derivative)**.** If $f(\beta)$ is a scalar-valued function of a $p \times 1$ vector β, such as $f(\beta) = \underset{\sim}{x} \cdot \beta$, then its **row-vector derivative** is the row vector of its partial derivatives (Definition 1 in Partial Derivatives and Gradients):

$$\frac{\partial f(\beta)}{\partial \underset{\sim}{\beta}^{\top}} = \begin{bmatrix} \frac{\partial}{\partial \beta_1} f(\underset{\sim}{\beta}) & \frac{\partial}{\partial \beta_2} f(\underset{\sim}{\beta}) & \cdots & \frac{\partial}{\partial \beta_p} f(\underset{\sim}{\beta}) \end{bmatrix} \tag{1}$$

When the derivative of a scalar-valued function is written as a row vector, it is also called the total derivative [1] of f.

Remark (Other notation for the row-vector derivative)

Remark 0.1 (Other notation for the row-vector derivative). Some sources write the same row vector as $\frac{\partial}{\partial \underset{\sim}{\beta}^{\top}} f(\underset{\sim}{\beta})$, with the operator on the left, read as each operator $\frac{\partial}{\partial \beta_i}$ applied to f rather than as a matrix product (see Remark 0.2).

Remark (Which side the operator goes on)

Remark 0.2 (Which side the operator goes on). Read $\frac{\partial}{\partial \underset{\sim}{\beta}}$ as a $p \times 1$ column vector of operators with entries $\frac{\partial}{\partial \beta_1}, \ldots, \frac{\partial}{\partial \beta_p}$, and $\frac{\partial}{\partial \underset{\sim}{\beta}^{\top}}$ as the $1 \times p$ row vector of operators with the same entries. Writing either one next to f can be read in two ways:

- **As a matrix product,** with f a 1×1 matrix, under the shape rule of matrix multiplication. The column-vector operator works on the left: $\frac{\partial}{\partial \underset{\sim}{\beta}} f$ is $(p \times 1)(1 \times 1)$, a $p \times 1$ column vector (Definition 2 in Partial Derivatives and Gradients). The row-vector operator does not: $\frac{\partial}{\partial \underset{\sim}{\beta}^{\top}} f$ is $(1 \times p)(1 \times 1)$, whose inner dimensions p and 1 do not match when $p > 1$. For the shape rule to give the $1 \times p$ row vector, the

operator would have to go on the right of f, as in $(1 \times 1)(1 \times p)$, which changes its meaning (see below).
- **As a scalar multiple,** with f a scalar, the way $c\underset{\sim}{x}$ is read in scalar multiplication, which has no shape rule: each entry of the operator vector is applied to f. Then both $\frac{\partial}{\partial \underset{\sim}{\beta}} f$ and $\frac{\partial}{\partial \underset{\sim}{\beta}^\top} f$ work with the operator on the left.

The fraction $\frac{\partial f}{\partial \underset{\sim}{\beta}^\top}$ of Equation 1 means the same thing under both readings: the transpose in the denominator says the result is a row vector, without making the derivative a product.
Treating the operators as vectors is sound for checking shapes, but "multiplying" an operator by a function means applying the operator to it, so two rules of ordinary matrix algebra do not carry over: scalars commute with matrices, and products can be regrouped.

- **Scalars do not commute with operators.** $\frac{\partial}{\partial \beta_i} f$ is a function, but $f \frac{\partial}{\partial \beta_i}$ is still an unapplied operator, waiting for something to act on. So, unlike $c\underset{\sim}{x} = \underset{\sim}{x}c$, the operator cannot move to the other side of f without changing the meaning. Putting the row-vector operator on the right of f gives the row-vector derivative only under the convention that it acts on the factor to its left.
- **Products cannot be regrouped.** An operator acts on the whole product to its right, so by the product rule, $\frac{\partial}{\partial \beta_i}(fg) \neq \left(\frac{\partial}{\partial \beta_i} f\right) g$ in general. For example, with $f(\underset{\sim}{\beta}) = \beta_1$, $\frac{\partial}{\partial \beta_1}(ff) = 2\beta_1$ but $\left(\frac{\partial}{\partial \beta_1} f\right) f = \beta_1$.

Theorem 0.1 (Row and column derivatives are transposes)**.**

$$\frac{\partial f(\beta)}{\partial \underset{\sim}{\beta}^\top} = \left(\frac{\partial f(\beta)}{\partial \underset{\sim}{\beta}}\right)^\top \tag{2}$$

$$\frac{\partial f(\beta)}{\partial \underset{\sim}{\beta}} = \left(\frac{\partial f(\beta)}{\partial \underset{\sim}{\beta}^\top}\right)^\top \tag{3}$$

Proof

Proof. By Definition 2 in Partial Derivatives and Gradients and Definition 0.1, entry j of both $\frac{\partial f(\beta)}{\partial \underset{\sim}{\beta}}$ and $\frac{\partial f(\beta)}{\partial \underset{\sim}{\beta}^\top}$ is $\frac{\partial}{\partial \beta_j} f(\underset{\sim}{\beta})$; the first is a $p \times 1$ column vector and the second a $1 \times p$ row vector with the same entries in the same order, so each is the transpose of the other. □

Example 0.1 (Row and column derivatives of a linear function)**.** For $f(\underset{\sim}{\beta}) = 3\beta_1 + 5\beta_2$:

$$\frac{\partial f(\underset{\sim}{\beta})}{\partial \underset{\sim}{\beta}} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}, \qquad \frac{\partial f(\underset{\sim}{\beta})}{\partial \underset{\sim}{\beta}^\top} = \begin{bmatrix} 3 & 5 \end{bmatrix},$$

and each is the transpose of the other.

Definition 0.2 (Derivative of a vector-valued function)**.** If

$$\begin{aligned} \underset{\sim}{y} &= \underset{\sim}{y}(\underset{\sim}{\beta}) \\ &= (y_1, \dots, y_q)^\top \end{aligned}$$

is a $q \times 1$ vector-valued function of the $p \times 1$ vector $\underset{\sim}{\beta}$, its **derivative with respect to** $\underset{\sim}{\beta}$ is the $p \times q$ matrix whose (i, j) entry is

$$\left[\frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{y}^\top\right]_{ij} \stackrel{\text{def}}{=} \frac{\partial}{\partial \beta_i} y_j, \qquad i = 1, \dots, p, \quad j = 1, \dots, q.$$

Both $\frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{y}$ and $\frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{y}^\top$ denote this $p \times q$ matrix.

Definition 0.3 (Denominator layout)**.** Writing derivatives as in Definition 0.2 is the **denominator layout** [2], which these notes use throughout: rows index the entries of β (the denominator) and columns index the entries of y (the numerator), so column j is the vector derivative $\frac{\partial}{\partial \underset{\sim}{\beta}} y_j$ (Definition 2 in Partial Derivatives and Gradients).

Example 0.2 (Differentiating a 3×1 function of a 2×1 vector)**.** Let $\underset{\sim}{\beta} = (\beta_1, \beta_2)^\top$ $(p = 2)$ and $\underset{\sim}{y}(\underset{\sim}{\beta}) = (\beta_1^2,\ \beta_1\beta_2,\ 3\beta_2)^\top$ $(q = 3)$. Then

$$\begin{aligned} \underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{y}^\top}_{2 \times 3} &= \begin{bmatrix} \frac{\partial}{\partial \beta_1}\beta_1^2 & \frac{\partial}{\partial \beta_1}\beta_1\beta_2 & \frac{\partial}{\partial \beta_1} 3\beta_2 \\ \frac{\partial}{\partial \beta_2}\beta_1^2 & \frac{\partial}{\partial \beta_2}\beta_1\beta_2 & \frac{\partial}{\partial \beta_2} 3\beta_2 \end{bmatrix} \\ &= \begin{bmatrix} 2\beta_1 & \beta_2 & 0 \\ 0 & \beta_1 & 3 \end{bmatrix} \end{aligned}$$

Definition 0.4 (Jacobian matrix)**.** Let

$$\begin{aligned} \underset{\sim}{y} &= \underset{\sim}{y}(\underset{\sim}{\beta}) \\ &= (y_1, \dots, y_q)^\top \end{aligned}$$

be a $q \times 1$ vector-valued function of the $p \times 1$ vector $\underset{\sim}{\beta}$. The **Jacobian matrix** [3] of $\underset{\sim}{y}$

is the $q \times p$ matrix whose (j, i) entry is $\frac{\partial}{\partial \beta_i} y_j$, for $j = 1, \ldots, q$ and $i = 1, \ldots, p$. It is the transpose of the derivative $\frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{y}^\top$ of Definition 0.2.

Definition 0.5 (Numerator layout)**.** Writing derivatives as in Definition 0.4, with rows indexing the entries of y (the numerator) and columns indexing the entries of $\underset{\sim}{\beta}$ (the denominator), is the **numerator layout** [2].

Example 0.3 (The Jacobian matrix of a 3×1 function of a 2×1 vector)**.** For $\underset{\sim}{y}(\underset{\sim}{\beta}) = (\beta_1^2,\ \beta_1\beta_2,\ 3\beta_2)^\top$ of Example 0.2, the Jacobian matrix is the 3×2 matrix

$$\begin{bmatrix} \frac{\partial}{\partial \beta_1}\beta_1^2 & \frac{\partial}{\partial \beta_2}\beta_1^2 \\ \frac{\partial}{\partial \beta_1}\beta_1\beta_2 & \frac{\partial}{\partial \beta_2}\beta_1\beta_2 \\ \frac{\partial}{\partial \beta_1}3\beta_2 & \frac{\partial}{\partial \beta_2}3\beta_2 \end{bmatrix} = \begin{bmatrix} 2\beta_1 & 0 \\ \beta_2 & \beta_1 \\ 0 & 3 \end{bmatrix},$$

the transpose of the 2×3 matrix found there.

Remark (Numerator layout)

Remark 0.3 (Numerator layout). Some sources use the numerator layout (Definition 0.5) throughout, so that their derivative of a vector-valued function is the Jacobian matrix, the transpose of the derivative used in these notes. In numerator layout, the derivative of a scalar-valued function is a row vector: the row-vector derivative (Definition 0.1). Check a source's layout before combining its formulas with these.

Definition 0.6 (Constant)**.** A $q \times 1$ vector $\underset{\sim}{x}$ is **constant with respect to** the $p \times 1$ vector $\underset{\sim}{\beta}$ if its derivative (Definition 0.2) is zero:

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{x}^\top}_{p \times q} = \underset{p \times q}{\underset{\sim}{\mathbf{0}}}$$

Example 0.4 (A constant vector)**.** Let $\underset{\sim}{\beta} = (\beta_1, \beta_2)^\top$ and $\underset{\sim}{x} = (3, 5)^\top$, so $x_1 = 3$ and $x_2 = 5$ do not depend on $\underset{\sim}{\beta}$. Expanding $\frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{x}^\top$ into its matrix of scalar partial derivatives (Definition 0.2) and evaluating each entry:

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{x}^\top}_{2\times 2} = \frac{\partial}{\partial \underset{\sim}{\beta}} \begin{bmatrix} x_1 & x_2 \end{bmatrix}$$
$$= \begin{bmatrix} \frac{\partial}{\partial \beta_1} x_1 & \frac{\partial}{\partial \beta_1} x_2 \\ \frac{\partial}{\partial \beta_2} x_1 & \frac{\partial}{\partial \beta_2} x_2 \end{bmatrix}$$
$$= \begin{bmatrix} \frac{\partial}{\partial \beta_1} 3 & \frac{\partial}{\partial \beta_1} 5 \\ \frac{\partial}{\partial \beta_2} 3 & \frac{\partial}{\partial \beta_2} 5 \end{bmatrix}$$
$$= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$
$$= \underset{2\times 2}{\underset{\sim}{\mathbf{0}}}$$

Every entry is the derivative of a constant, so $\frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{x}^\top = \underset{2\times 2}{\underset{\sim}{\mathbf{0}}}$ and $\underset{\sim}{x}$ is constant with respect to $\underset{\sim}{\beta}$ (Definition 0.6).

Example 0.5 (A vector that is not constant)**.** With the same $\underset{\sim}{\beta}$, let $\underset{\sim}{x} = (\beta_1, 3)^\top$. Then

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{x}^\top}_{2\times 2} = \begin{bmatrix} \frac{\partial}{\partial \beta_1} \beta_1 & \frac{\partial}{\partial \beta_1} 3 \\ \frac{\partial}{\partial \beta_2} \beta_1 & \frac{\partial}{\partial \beta_2} 3 \end{bmatrix}$$
$$= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

which is not $\underset{2\times 2}{\underset{\sim}{\mathbf{0}}}$, so this $\underset{\sim}{x}$ is not constant with respect to $\underset{\sim}{\beta}$: its first entry changes when β_1 does.

Exercise 0.1 (Partial derivative with respect to an unreferenced variable)**.** Adapted from Miller [4], Question 1.1.46.
Find $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ for

$$f(x, y, t) = 5t^4 - 4t^5 \cos(t \sin t)$$

Solution

Solution 0.1. Although the function is declared as a function of three variables (x, y, t), its expression depends solely on t.
Because neither x nor y appears in the formula, f is constant with respect to both x and y:

$$\frac{\partial f}{\partial x} = 0, \qquad \frac{\partial f}{\partial y} = 0$$

Remark: Always verify which variables actually appear in a formula before performing lengthy algebraic differentiation.

Theorem 0.2 (Derivative of a dot product)**.** *If $\underset{\sim}{x}$ is constant with respect to $\underset{\sim}{\beta}$, then:*

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{x} \cdot \underset{\sim}{\beta})}_{p\times 1} = \underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{\beta} \cdot \underset{\sim}{x})}_{p\times 1}$$
$$= \underset{p\times 1}{\underset{\sim}{x}}$$

Proof

Proof.

$$\frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{x} \cdot \underset{\sim}{\beta}) = \begin{bmatrix} \frac{\partial}{\partial \beta_1}(x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p) \\ \frac{\partial}{\partial \beta_2}(x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p) \\ \vdots \\ \frac{\partial}{\partial \beta_p}(x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p) \end{bmatrix}$$
$$= \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$
$$= \underset{\sim}{x}$$

□

Example 0.6 (Derivative of a dot product)**.** Let $\underset{\sim}{x} = (3, 5)^\top$ (constant with respect to $\underset{\sim}{\beta}$; see Example 0.4) and $\underset{\sim}{\beta} = (\beta_1, \beta_2)^\top$. Then $\underset{\sim}{x} \cdot \underset{\sim}{\beta} = 3\beta_1 + 5\beta_2$, and by Theorem 0.2:

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{x} \cdot \underset{\sim}{\beta})}_{2\times 1} = \underset{2\times 1}{\underset{\sim}{x}}$$
$$= \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

Verifying entry-wise:

$$\frac{\partial}{\partial \underset{\sim}{\beta}}(3\beta_1 + 5\beta_2) = \begin{pmatrix} \frac{\partial}{\partial \beta_1}(3\beta_1 + 5\beta_2) \\ \frac{\partial}{\partial \beta_2}(3\beta_1 + 5\beta_2) \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

Both methods agree.

Theorem 0.3 (Product rule for dot-products)**.** *If* $\underset{\sim}{a} = \underset{\sim}{a}(\underset{\sim}{x})$ *and* $\underset{\sim}{b} = \underset{\sim}{b}(\underset{\sim}{x})$ *are differentiable* $p \times 1$ *vector functions of* $\underset{\sim}{x}$*, then:*

$$\underset{p\times 1}{\frac{\partial}{\partial \underset{p\times 1}{\underset{\sim}{x}}}} \underset{p\times 1}{\underset{\sim}{a}} \cdot \underset{p\times 1}{\underset{\sim}{b}} = \left(\underset{p\times 1}{\frac{\partial}{\partial \underset{p\times 1}{\underset{\sim}{x}}}} \underset{1\times p}{\underset{\sim}{a}^\top} \right) \underset{p\times 1}{\underset{\sim}{b}} + \left(\underset{p\times 1}{\frac{\partial}{\partial \underset{p\times 1}{\underset{\sim}{x}}}} \underset{1\times p}{\underset{\sim}{b}^\top} \right) \underset{p\times 1}{\underset{\sim}{a}}$$

Proof

Proof. Entry-wise, for $i = 1, \ldots, p$:

$$\begin{aligned} \left[\frac{\partial}{\partial \underset{\sim}{x}}(\underset{\sim}{a} \cdot \underset{\sim}{b})\right]_i &= \frac{\partial}{\partial x_i} \sum_{k=1}^{p} a_k b_k \\ &= \sum_{k=1}^{p} \left(b_k \frac{\partial}{\partial x_i} a_k + a_k \frac{\partial}{\partial x_i} b_k \right) \\ &= \left[\left(\frac{\partial}{\partial \underset{\sim}{x}} \underset{\sim}{a}^\top \right) \underset{\sim}{b} \right]_i + \left[\left(\frac{\partial}{\partial \underset{\sim}{x}} \underset{\sim}{b}^\top \right) \underset{\sim}{a} \right]_i \end{aligned}$$

□

Example 0.7 (Example of the dot-product rule)**.** Apply Theorem 0.3 with the vector $\underset{\sim}{\beta} = (\beta_1, \beta_2)^\top$ in the role of $\underset{\sim}{x}$. Let $\underset{\sim}{a}(\underset{\sim}{\beta}) = (\beta_1, \beta_1\beta_2)^\top$ and $\underset{\sim}{b}(\underset{\sim}{\beta}) = (\beta_2, \beta_1)^\top$. Then:

$$\begin{aligned} \underset{\sim}{a} \cdot \underset{\sim}{b} &= \beta_1 \cdot \beta_2 + \beta_1\beta_2 \cdot \beta_1 \\ &= \beta_1\beta_2 + \beta_1^2\beta_2 \end{aligned}$$

By direct calculation:

$$\begin{aligned} \underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{a} \cdot \underset{\sim}{b})}_{2\times 1} &= \frac{\partial}{\partial \underset{\sim}{\beta}}(\beta_1\beta_2 + \beta_1^2\beta_2) \\ &= \begin{pmatrix} \beta_2 + 2\beta_1\beta_2 \\ \beta_1 + \beta_1^2 \end{pmatrix} \end{aligned}$$

By the product rule (Theorem 0.3), using $\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{a}^\top}_{2\times 2} = \begin{pmatrix} 1 & \beta_2 \\ 0 & \beta_1 \end{pmatrix}$ and $\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{b}^\top}_{2\times 2} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$:

$$\begin{aligned}
\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}} (\underset{\sim}{a} \cdot \underset{\sim}{b})}_{2\times 1} &= \underbrace{\begin{pmatrix} 1 & \beta_2 \\ 0 & \beta_1 \end{pmatrix}}_{2\times 2} \underbrace{\begin{pmatrix} \beta_2 \\ \beta_1 \end{pmatrix}}_{2\times 1} + \underbrace{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}}_{2\times 2} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_1\beta_2 \end{pmatrix}}_{2\times 1} \\
&= \begin{pmatrix} \beta_2 + \beta_1\beta_2 \\ \beta_1^2 \end{pmatrix} + \begin{pmatrix} \beta_1\beta_2 \\ \beta_1 \end{pmatrix} \\
&= \begin{pmatrix} \beta_2 + 2\beta_1\beta_2 \\ \beta_1^2 + \beta_1 \end{pmatrix}
\end{aligned}$$

Both methods agree.

Theorem 0.4 (Derivative of a linear map)**.** *If* $\mathbf{A}$ *is an* $m \times p$ *matrix that is constant with respect to* $\underset{\sim}{\beta}$*, then:*

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}} (\mathbf{A}\underset{\sim}{\beta})}_{p\times m} = \underbrace{\mathbf{A}^\top}_{p\times m}$$

Proof

Proof. For entry (i, j), where row i indexes the denominator $\underset{\sim}{\beta}$ (see Definition 0.2) and column j indexes the numerator $\mathbf{A}\underset{\sim}{\beta}$:

$$\begin{aligned}
\left[\frac{\partial}{\partial \underset{\sim}{\beta}} (\mathbf{A}\underset{\sim}{\beta})\right]_{ij} &= \frac{\partial}{\partial \beta_i} (\mathbf{A}\underset{\sim}{\beta})_j \\
&= \frac{\partial}{\partial \beta_i} \sum_{k=1}^{p} a_{jk}\beta_k \\
&= a_{ji} \\
&= [\mathbf{A}^\top]_{ij}
\end{aligned}$$

□

Example 0.8 (Derivative of a linear map)**.** Let $\mathbf{A} = \begin{pmatrix} 2 & 3 \end{pmatrix}$ (1×2) and $\underset{\sim}{\beta} = (\beta_1, \beta_2)^\top$. Then $\mathbf{A}\underset{\sim}{\beta} = 2\beta_1 + 3\beta_2$, and by Theorem 0.4:

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}}(\mathbf{A}\underset{\sim}{\beta})}_{2\times 1} = \underbrace{\mathbf{A}^\top}_{2\times 1}$$
$$= \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

Theorem 0.5 (Vector-derivative of a matrix-vector product)**.** *If* $\mathbf{A}$ *is an* $m \times q$ *matrix that is constant with respect to* $\underset{\sim}{\beta}$*, and* $\underset{\sim}{v} = \underset{\sim}{v}(\underset{\sim}{\beta})$ *is a* $q \times 1$ *vector that depends on the* $p \times 1$ *vector* $\underset{\sim}{\beta}$*, then:*

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}}(\mathbf{A}\underset{\sim}{v})}_{p\times m} = \underbrace{\left(\frac{\partial}{\partial \underset{\sim}{\beta}}\underset{\sim}{v}\right)}_{p\times q} \underbrace{\mathbf{A}^\top}_{q\times m}$$

Proof

Proof. For entry (i, j), where row i indexes the denominator $\underset{\sim}{\beta}$ and column j indexes the numerator $\mathbf{A}\underset{\sim}{v}$ (see Definition 0.2):

$$\begin{aligned}
\left[\frac{\partial}{\partial \underset{\sim}{\beta}}(\mathbf{A}\underset{\sim}{v})\right]_{ij} &= \frac{\partial}{\partial \beta_i}(\mathbf{A}\underset{\sim}{v})_j \\
&= \frac{\partial}{\partial \beta_i}\sum_{k=1}^{q} a_{jk}v_k \\
&= \sum_{k=1}^{q} a_{jk}\frac{\partial}{\partial \beta_i}v_k \\
&= \sum_{k=1}^{q}\left[\frac{\partial}{\partial \underset{\sim}{\beta}}\underset{\sim}{v}\right]_{ik}\left[\mathbf{A}^\top\right]_{kj} \\
&= \left[\left(\frac{\partial}{\partial \underset{\sim}{\beta}}\underset{\sim}{v}\right)\mathbf{A}^\top\right]_{ij}
\end{aligned}$$

□

Example 0.9 (Vector-derivative of a matrix-vector product)**.** Let $\mathbf{A} = \begin{pmatrix} 2 & 3 \end{pmatrix}$ (1×2, constant) and $\underset{\sim}{v}(\underset{\sim}{\beta}) = (\beta_1^2, \beta_2^2)^\top$. Then $\mathbf{A}\underset{\sim}{v} = 2\beta_1^2 + 3\beta_2^2$. By Theorem 0.5:

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}}(\mathbf{A}\underset{\sim}{v})}_{2\times 1} = \begin{pmatrix} 2\beta_1 & 0 \\ 0 & 2\beta_2 \end{pmatrix}\begin{pmatrix} 2 \\ 3 \end{pmatrix}$$
$$= \begin{pmatrix} 4\beta_1 \\ 6\beta_2 \end{pmatrix}$$

i Remark (The derivative of a linear map as a special case)

Remark 0.4 (The derivative of a linear map as a special case). This result generalizes Theorem 0.4, which is the special case $\underset{\sim}{v} = \underset{\sim}{\beta}$ (so that $q = p$, $\frac{\partial}{\partial \underset{\sim}{\beta}}\underset{\sim}{\beta} = \mathbf{I}_p$, and $\frac{\partial}{\partial \underset{\sim}{\beta}}(\mathbf{A}\underset{\sim}{\beta}) = \mathbf{I}_p\mathbf{A}^\top$, which is $\mathbf{A}^\top$). For example, Example 0.8 is the case $\mathbf{A} = \begin{pmatrix} 2 & 3 \end{pmatrix}$ and

$$\underset{\sim}{v} = \underset{\sim}{\beta}$$
$$= (\beta_1, \beta_2)^\top,$$

where this result gives $\mathbf{I}_2\mathbf{A}^\top = (2, 3)^\top$.

Theorem 0.6 (Vector-derivative of a product of matrices)**.** *If* $\mathbf{A}$ *$(k \times m)$ and* $\mathbf{B}$ *$(m \times q)$ are constant with respect to* $\underset{\sim}{\beta}$*, and* $\underset{\sim}{v} = \underset{\sim}{v}(\underset{\sim}{\beta})$ *is a* $q \times 1$ *vector that depends on the* $p \times 1$ *vector* $\underset{\sim}{\beta}$*, then:*

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}}(\mathbf{A}\mathbf{B}\underset{\sim}{v})}_{p\times k} = \underbrace{\left(\frac{\partial}{\partial \underset{\sim}{\beta}}\underset{\sim}{v}\right)}_{p\times q}\underbrace{\mathbf{B}^\top}_{q\times m}\underbrace{\mathbf{A}^\top}_{m\times k}$$

i Proof

Proof. Apply Theorem 0.5 with the constant $k \times q$ matrix $\mathbf{AB}$, then use the transpose of a product:

$$\frac{\partial}{\partial \underset{\sim}{\beta}}(\mathbf{A}\mathbf{B}\underset{\sim}{v}) = \left(\frac{\partial}{\partial \underset{\sim}{\beta}}\underset{\sim}{v}\right)(\mathbf{A}\mathbf{B})^\top \quad \text{(derivative of a matrix-vector product)}$$
$$= \left(\frac{\partial}{\partial \underset{\sim}{\beta}}\underset{\sim}{v}\right)\mathbf{B}^\top\mathbf{A}^\top \quad \text{(transpose of a product)}$$

□

Example 0.10. Let $\mathbf{A} = \begin{pmatrix} 1 & 0 \end{pmatrix}$ (1×2), $\mathbf{B} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$ (2×2), and $\underset{\sim}{v}(\underset{\sim}{\beta}) = \underset{\sim}{\beta}$ where $\underset{\sim}{\beta} = (\beta_1, \beta_2)^\top$. Then $\mathbf{AB}\underset{\sim}{v} = 2\beta_1$, and:

$$\underbrace{\frac{\partial}{\partial \beta}(\mathbf{AB}\underset{\sim}{v})}_{2\times 1} = \underbrace{\left(\frac{\partial}{\partial \beta}\underset{\sim}{\beta}\right)}_{2\times 2}\underset{2\times 2}{\mathbf{B}^\top}\underset{2\times 1}{\mathbf{A}^\top}$$
$$= \mathbf{I}_2\begin{pmatrix}2 & 0\\ 0 & 3\end{pmatrix}\begin{pmatrix}1\\ 0\end{pmatrix}$$
$$= \begin{pmatrix}2\\ 0\end{pmatrix}$$

Corollary 0.1 (Derivative of a dot product, transpose-product form)**.** *If $\underset{\sim}{x}$ is constant with respect to $\underset{\sim}{\beta}$, then:*

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{1\times p}{\underset{\sim}{x}^\top}\underset{p\times 1}{\underset{\sim}{\beta}})}_{p\times 1} = \underset{p\times 1}{\underset{\sim}{x}}$$

Proof

Proof. **Using Theorem 0.2:**
Since $\underset{\sim}{x}^\top\underset{\sim}{\beta} = \underset{\sim}{x}\cdot\underset{\sim}{\beta}$ (see dot product), and $\underset{\sim}{x}$ is constant with respect to $\underset{\sim}{\beta}$:

$$\frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{x}^\top\underset{\sim}{\beta}) = \frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{x}\cdot\underset{\sim}{\beta})$$
$$= \underset{\sim}{x}$$

by Theorem 0.2. □

Proof

Proof. **Using Theorem 0.5:**
Since $\underset{\sim}{x}$ is constant with respect to β, $\mathbf{A} = \underset{\sim}{x}^\top$ is a constant $1\times p$ matrix. Applying Theorem 0.5 with $\underset{\sim}{v} = \underset{\sim}{\beta}$ (so $\frac{\partial}{\partial \underset{\sim}{\beta}}\underset{\sim}{\beta} = \mathbf{I}$):

$$\frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{x}^\top\underset{\sim}{\beta}) = \left(\frac{\partial}{\partial \underset{\sim}{\beta}}\underset{\sim}{\beta}\right)(\underset{\sim}{x}^\top)^\top$$
$$= \mathbf{I}\cdot\underset{\sim}{x}$$
$$= \underset{\sim}{x}$$

□

Example 0.11 (Derivative of a transpose product)**.** Let $\underset{\sim}{x} = (3,5)^\top$ and $\underset{\sim}{\beta} = (\beta_1, \beta_2)^\top$. Then $\underset{\sim}{x}^\top \underset{\sim}{\beta} = 3\beta_1 + 5\beta_2$, and by Corollary 0.1:

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}} \left(\underset{\sim}{x}^\top \underset{\sim}{\beta} \right)}_{2\times 1} = \underset{2\times 1}{\underset{\sim}{x}}$$

$$= \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

i Remark (The coefficient gets transposed)

Remark 0.5 (The coefficient gets transposed). This vector derivative formula looks a lot like non-vector calculus, except that you have to transpose the coefficient: in scalar calculus $\frac{\partial}{\partial x}(cx) = c$, but here the coefficient $\underset{\sim}{x}^\top$ (a row vector) becomes $\underset{\sim}{x}$ (a column vector) in the result. For example, with $\underset{\sim}{x} = (2,-1)^\top$, $\underset{\sim}{x}^\top \underset{\sim}{\beta} = 2\beta_1 - \beta_2$, whose vector derivative is the column vector $(2,-1)^\top = \underset{\sim}{x}$, not the row vector $\underset{\sim}{x}^\top = (2,-1)$.

Theorem 0.7 (Derivative of a quadratic form)**.** *For a quadratic form (see quadratic form), if* $\mathbf{S}$ *is a* $p \times p$ *matrix that is constant with respect to* $\underset{\sim}{\beta}$*, then:*

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}} (\underset{\sim}{\beta}^\top \mathbf{S} \underset{\sim}{\beta})}_{p\times 1} = \underbrace{(\mathbf{S} + \mathbf{S}^\top)\underset{\sim}{\beta}}_{p\times 1}.$$

i Proof

Proof. Expanding entry-wise, $\underset{\sim}{\beta}^\top \mathbf{S} \underset{\sim}{\beta} = \sum_{j=1}^p \sum_{k=1}^p s_{jk}\beta_j\beta_k$. Differentiating component-wise with respect to β_i for $i = 1, \ldots, p$:

$$\begin{aligned}
\left[\frac{\partial}{\partial \underset{\sim}{\beta}} (\underset{\sim}{\beta}^\top \mathbf{S} \underset{\sim}{\beta}) \right]_i &= \frac{\partial}{\partial \beta_i} \sum_{j=1}^p \sum_{k=1}^p s_{jk}\beta_j\beta_k && \text{(expand quadratic form)} \\
&= \sum_{k=1}^p s_{ik}\beta_k + \sum_{j=1}^p s_{ji}\beta_j && \text{(product rule for } \beta_i\beta_k) \\
&= [\mathbf{S}\underset{\sim}{\beta}]_i + [\mathbf{S}^\top \underset{\sim}{\beta}]_i && \text{(matrix-vector multiplication definition)} \\
&= [(\mathbf{S} + \mathbf{S}^\top)\underset{\sim}{\beta}]_i && \text{(linearity of matrix multiplication)}
\end{aligned}$$

□

Corollary 0.2 (Derivative of a quadratic form with a symmetric matrix)**.** *If, in addition,* $\mathbf{S}$ *is symmetric (and, as in Theorem 0.7,* $p \times p$ *and constant with respect to* $\underset{\sim}{\beta}$*), then*

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{\beta}^\top \mathbf{S} \underset{\sim}{\beta})}_{p\times 1} = \underbrace{2\mathbf{S}\underset{\sim}{\beta}}_{p\times 1}.$$

Proof

Proof. If $\mathbf{S} = \mathbf{S}^\top$, then $\mathbf{S} + \mathbf{S}^\top = 2\mathbf{S}$, so Theorem 0.7 gives $\frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{\beta}^\top \mathbf{S} \underset{\sim}{\beta}) = 2\mathbf{S}\underset{\sim}{\beta}$. □

Remark (Like the derivative of cx^2)

Remark 0.6 (Like the derivative of cx^2). This operation is like taking the derivative of cx^2 with respect to x in non-vector calculus: $\frac{\partial}{\partial x}(cx^2) = 2cx$, and Corollary 0.2 says $\frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{\beta}^\top \mathbf{S} \underset{\sim}{\beta}) = 2\mathbf{S}\underset{\sim}{\beta}$. For example, with $p = 1$, $\mathbf{S} = (3)$, and $\underset{\sim}{\beta} = (\beta_1)$, $\underset{\sim}{\beta}^\top \mathbf{S} \underset{\sim}{\beta} = 3\beta_1^2$, and its derivative is $6\beta_1 = 2\mathbf{S}\underset{\sim}{\beta}$.

Example 0.12 (Derivative of a quadratic form)**.** Let $\mathbf{S} = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$ (2×2, symmetric and constant) and $\underset{\sim}{\beta} = (\beta_1, \beta_2)^\top$. Then $\underset{\sim}{\beta}^\top \mathbf{S} \underset{\sim}{\beta} = 3\beta_1^2 + 2\beta_1\beta_2 + 2\beta_2^2$. By Corollary 0.2:

$$\begin{aligned} \underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{\beta}^\top \mathbf{S} \underset{\sim}{\beta})}_{2\times 1} &= 2\mathbf{S}\underset{\sim}{\beta} \\ &= 2\begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \\ &= \begin{pmatrix} 6\beta_1 + 2\beta_2 \\ 2\beta_1 + 4\beta_2 \end{pmatrix} \end{aligned}$$

Differentiating component-wise directly:

$$\begin{pmatrix} \frac{\partial}{\partial \beta_1}(3\beta_1^2 + 2\beta_1\beta_2 + 2\beta_2^2) \\ \frac{\partial}{\partial \beta_2}(3\beta_1^2 + 2\beta_1\beta_2 + 2\beta_2^2) \end{pmatrix} = \begin{pmatrix} 6\beta_1 + 2\beta_2 \\ 2\beta_1 + 4\beta_2 \end{pmatrix}$$

Both methods agree.

For a matrix that is not symmetric, such as $\mathbf{S} = \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$, Theorem 0.7 gives

$$(\mathbf{S}+\mathbf{S}^\top)\underset{\sim}{\beta} = \begin{pmatrix} 2 & 2 \\ 2 & 6 \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 2\beta_1 + 2\beta_2 \\ 2\beta_1 + 6\beta_2 \end{pmatrix},$$

which matches differentiating $\underset{\sim}{\beta}^\top \mathbf{S}\underset{\sim}{\beta} = \beta_1^2 + 2\beta_1\beta_2 + 3\beta_2^2$ directly.

Corollary 0.3 (Derivative of a simple quadratic form)**.**

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{\beta}\cdot\underset{\sim}{\beta})}_{p\times 1} = \underset{p\times 1}{\underset{\sim}{2\beta}}$$

Proof

Proof. Applying Corollary 0.2 with $\mathbf{S} = \mathbf{I}_{p\times p}$ (which is symmetric and constant with respect to $\underset{\sim}{\beta}$):

$$\begin{aligned} \frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{\beta}\cdot\underset{\sim}{\beta}) = \frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{\beta}^\top \mathbf{I}_{p\times p}\underset{\sim}{\beta}) & \quad \text{(rewrite with identity matrix)} \\ = 2\underset{\sim}{\mathbf{I}_{p\times p}\beta} & \quad \text{(derivative of a quadratic form, with } \mathbf{S} = \mathbf{I}_{p\times p}\text{)} \\ = 2\underset{\sim}{\beta} & \quad \text{(identity matrix property)} \end{aligned}$$

□

Remark (Like the derivative of x^2)

Remark 0.7 (Like the derivative of x^2). This vector derivative is like taking the derivative of x^2: in scalar calculus $\frac{\partial}{\partial x}x^2 = 2x$, and Corollary 0.3 says $\frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{\beta}\cdot\underset{\sim}{\beta}) = 2\underset{\sim}{\beta}$. For example, with $p = 1$ and $\underset{\sim}{\beta} = (\beta_1)$, $\underset{\sim}{\beta}\cdot\underset{\sim}{\beta} = \beta_1^2$, and its derivative is $2\beta_1 = 2\underset{\sim}{\beta}$.

Example 0.13 (Derivative of a sum of squares)**.** Let $\underset{\sim}{\beta} = (\beta_1, \beta_2)^\top$, so $\underset{\sim}{\beta}\cdot\underset{\sim}{\beta} = \beta_1^2 + \beta_2^2$. By Corollary 0.3:

$$\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{\beta}\cdot\underset{\sim}{\beta})}_{2\times 1} = 2\underset{\sim}{\beta} = \begin{pmatrix} 2\beta_1 \\ 2\beta_2 \end{pmatrix}$$

Direct partial differentiation yields the same column vector.

Definition 0.7 (Residual vector and residuals)**.** Let $\mathbf{X}$ be an $n \times p$ matrix, $\underset{\sim}{y} \in \mathbb{R}^n$, and $\underset{\sim}{\beta} \in \mathbb{R}^p$. The **residual vector** of $\underset{\sim}{\beta}$ is $\underset{\sim}{\varepsilon}(\underset{\sim}{\beta}) \overset{\text{def}}{=} \underset{\sim}{y} - \mathbf{X}\underset{\sim}{\beta}$, and its entries $\varepsilon_i = y_i - (\mathbf{X}\underset{\sim}{\beta})_i$ are the **residuals** [5].

Definition 0.8 (Squared error)**.** In the setting of Definition 0.7, the square ε_i^2 is the ith **squared error**.

Definition 0.9 (Residual sum of squares)**.** In the setting of Definition 0.7, the sum of the squared errors (Definition 0.8)

$$\begin{aligned} \underset{\sim}{\varepsilon} \cdot \underset{\sim}{\varepsilon} &= \sum_{i=1}^{n} \underset{\sim}{\varepsilon}_i^2 \\ &= \|\underset{\sim}{y} - \mathbf{X}\underset{\sim}{\beta}\|^2 \end{aligned}$$

is the **residual sum of squares** [6].

Remark (Least squares solutions minimize the residual sum of squares)

Remark 0.8 (Least squares solutions minimize the residual sum of squares). A least squares solution of $\mathbf{X}\beta = y$ makes $\|y - \mathbf{X}\beta\|$ as small as possible, and so makes the residual sum of squares (Definition 0.9) as small as possible too.

Example 0.14 (Residuals of a constant fit)**.** Let $\mathbf{X} = (1,1)^\top$ ($n = 2$, $p = 1$) and $\underset{\sim}{y} = (1,3)^\top$. For $\underset{\sim}{\beta} = (2)$, $\mathbf{X}\underset{\sim}{\beta} = (2,2)^\top$, so

$$\begin{aligned} \underset{\sim}{\varepsilon}(\underset{\sim}{\beta}) &= (1,3)^\top - (2,2)^\top && \text{(definition of the residual vector)} \\ &= (-1,1)^\top, && \text{(subtract entry by entry)} \end{aligned}$$

the squared errors are $(-1)^2 = 1$ and $1^2 = 1$, and the residual sum of squares is $1+1 = 2$. For $\underset{\sim}{\beta} = (1)$ the residuals are 0 and 2, and the residual sum of squares is $0 + 4 = 4$.

Videos

Jon Krohn's "Calculus for Machine Learning" YouTube playlist has videos on gradients of sums and sample means of squared errors (Definition 0.9):

- The Gradient of Quadratic Cost
- The Gradient of Mean Squared Error

Theorem 0.8 (Vector chain rule)**.** *Let $\underset{\sim}{x}$ be a $p \times 1$ vector, let $\underset{\sim}{y} = \underset{\sim}{g}(\underset{\sim}{x})$ be a $q \times 1$ vector-valued function of $\underset{\sim}{x}$, and let $z = f(\underset{\sim}{y})$ be a scalar-valued function of $\underset{\sim}{y}$, where $\underset{\sim}{g}$ and f have continuous partial derivatives. Then $z = f(\underset{\sim}{g}(\underset{\sim}{x}))$, as a function of $\underset{\sim}{x}$, satisfies*

$$\underset{p\times 1}{\frac{\partial z}{\partial \underset{\sim}{x}}} = \underset{p\times q}{\frac{\partial \underset{\sim}{y}}{\partial \underset{\sim}{x}}} \underset{q\times 1}{\frac{\partial z}{\partial \underset{\sim}{y}}}$$

where $\frac{\partial \underset{\sim}{y}}{\partial \underset{\sim}{x}}$ is the derivative of Definition 0.2 and $\frac{\partial z}{\partial \underset{\sim}{x}}$ and $\frac{\partial z}{\partial \underset{\sim}{y}}$ are vector derivatives (Definition 2 in Partial Derivatives and Gradients).

Remark (The order of the factors matters)

Remark 0.9 (The order of the factors matters). The vector chain rule (Theorem 0.8) is like the univariate chain rule, but the order matters now: $\frac{\partial \underset{\sim}{y}}{\partial \underset{\sim}{x}}$ is $p \times q$ and $\frac{\partial z}{\partial \underset{\sim}{y}}$ is $q \times 1$, so the product $\frac{\partial z}{\partial \underset{\sim}{y}} \frac{\partial \underset{\sim}{y}}{\partial \underset{\sim}{x}}$ in the other order is not even defined unless $p = 1$.
The version presented here is for the gradient (Definition 2 in Partial Derivatives and Gradients), a column vector. The total derivative (Definition 0.1), a row vector, is the transpose of the gradient (Theorem 0.1), and transposing both sides gives $\left(\frac{\partial z}{\partial \underset{\sim}{x}}\right)^\top = \left(\frac{\partial z}{\partial \underset{\sim}{y}}\right)^\top \left(\frac{\partial \underset{\sim}{y}}{\partial \underset{\sim}{x}}\right)^\top$, with the factors in the reverse order; there $\left(\frac{\partial \underset{\sim}{y}}{\partial \underset{\sim}{x}}\right)^\top$ is the Jacobian matrix of $\underset{\sim}{y}$ as a function of $\underset{\sim}{x}$ (Definition 0.4).

Example 0.15 (Applying the vector chain rule)**.** Let $\underset{\sim}{x} = (x_1, x_2)^\top$,

$$\begin{aligned} \underset{\sim}{y} &= \underset{\sim}{g}(\underset{\sim}{x}) \\ &= (x_1 + x_2,\ x_1 x_2)^\top, \end{aligned}$$

and

$$\begin{aligned} z &= f(\underset{\sim}{y}) \\ &= y_1^2 + y_2. \end{aligned}$$

Then

$$
\begin{aligned}
\underbrace{\frac{\partial z}{\partial \underset{\sim}{x}}}_{2\times 1} &= \underbrace{\frac{\partial y}{\partial \underset{\sim}{x}}}_{2\times 2}\underbrace{\frac{\partial z}{\partial y}}_{2\times 1} && \text{(vector chain rule)} \\
&= \begin{bmatrix} 1 & x_2 \\ 1 & x_1 \end{bmatrix}\begin{bmatrix} 2y_1 \\ 1 \end{bmatrix} && \text{(differentiate } \underset{\sim}{g} \text{ and } f\text{)} \\
&= \begin{bmatrix} 2(x_1+x_2)+x_2 \\ 2(x_1+x_2)+x_1 \end{bmatrix} && \text{(multiply, and substitute } y_1 = x_1 + x_2\text{)} \\
&= \begin{bmatrix} 2x_1+3x_2 \\ 3x_1+2x_2 \end{bmatrix} && \text{(collect terms)}
\end{aligned}
$$

This matches differentiating z directly:

$$
\begin{aligned}
z &= (x_1+x_2)^2 + x_1x_2 \\
&= x_1^2 + 3x_1x_2 + x_2^2,
\end{aligned}
$$

so $\frac{\partial}{\partial x_1}z = 2x_1 + 3x_2$ and $\frac{\partial}{\partial x_2}z = 3x_1 + 2x_2$. The product in the other order, a 2×1 matrix times a 2×2 matrix, is not defined. The total derivative is the transpose, the row vector $(2x_1+3x_2,\ 3x_1+2x_2)$.

See Felippa [7], and Wikipedia contributors [8], section "Relationship with Frechet derivative".

Corollary 0.4 (Vector chain rule for quadratic forms)**.** *If $\underset{\sim}{\varepsilon} = \underset{\sim}{\varepsilon}(\underset{\sim}{\beta})$ is an $n\times 1$ vector-valued function of the $p\times 1$ vector $\underset{\sim}{\beta}$, with continuous partial derivatives, then*

$$
\underbrace{\frac{\partial}{\partial \underset{\sim}{\beta}}\left(\underset{\sim}{\varepsilon}(\underset{\sim}{\beta})\cdot\underset{\sim}{\varepsilon}(\underset{\sim}{\beta})\right)}_{p\times 1} = \underbrace{\left(\frac{\partial}{\partial \underset{\sim}{\beta}}\underset{\sim}{\varepsilon}(\underset{\sim}{\beta})\right)}_{p\times n}\underbrace{\left(2\underset{\sim}{\varepsilon}(\underset{\sim}{\beta})\right)}_{n\times 1}
$$

Proof

Proof. Apply Theorem 0.8 with $\underset{\sim}{x} = \underset{\sim}{\beta}$, $\underset{\sim}{y} = \underset{\sim}{\varepsilon}$, $q = n$, and

$$
\begin{aligned}
z &= f(\underset{\sim}{\varepsilon}) \\
&= \underset{\sim}{\varepsilon}\cdot\underset{\sim}{\varepsilon} \\
&= \underset{\sim}{\varepsilon}^{\top}\underset{\sim}{\varepsilon}:
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \underset{\sim}{\beta}}\left(\underset{\sim}{\varepsilon}\cdot\underset{\sim}{\varepsilon}\right) &= \left(\frac{\partial}{\partial \underset{\sim}{\beta}}\underset{\sim}{\varepsilon}\right)\frac{\partial}{\partial \underset{\sim}{\varepsilon}}\left(\underset{\sim}{\varepsilon}^{\top}\underset{\sim}{\varepsilon}\right) && \text{(vector chain rule)} \\
&= \left(\frac{\partial}{\partial \underset{\sim}{\beta}}\underset{\sim}{\varepsilon}\right)(2\underset{\sim}{\varepsilon}) && \text{(derivative of a simple quadratic form, in } \underset{\sim}{\varepsilon}\text{)}
\end{aligned}
$$

The second step is Corollary 0.3, applied to the $n\times 1$ vector $\underset{\sim}{\varepsilon}$ in place of $\underset{\sim}{\beta}$. □

Example 0.16 (Derivative of the residual sum of squares)**.** Let y $(n \times 1)$ and $\mathbf{X}$ $(n \times p)$ be constant with respect to β, and let $\underset{\sim}{\varepsilon}(\beta) = y - \mathbf{X}\beta$ be the residual vector (Definition 0.7). By Theorem 0.4, $\frac{\partial}{\partial \beta}(\mathbf{X}\beta) = \mathbf{X}^\top$, and $\frac{\partial}{\partial \beta} y = \mathbf{0}_{p \times n}$ because y is constant, so $\frac{\partial}{\partial \beta}\underset{\sim}{\varepsilon} = -\mathbf{X}^\top$ $(p \times n)$. By Corollary 0.4:

$$\begin{aligned} \frac{\partial}{\partial \underset{\sim}{\beta}}(\underset{\sim}{\varepsilon} \cdot \underset{\sim}{\varepsilon}) &= (-\mathbf{X}^\top)(2\underset{\sim}{\varepsilon}) && \text{(vector chain rule for quadratic forms)} \\ &= -2\,\mathbf{X}^\top(\underset{\sim}{y} - \mathbf{X}\underset{\sim}{\beta}) && \text{(substitute } \underset{\sim}{\varepsilon} = \underset{\sim}{y} - \mathbf{X}\underset{\sim}{\beta}) \end{aligned}$$

Setting this $p \times 1$ vector to $\underset{\sim}{0}$ gives the normal equations $\mathbf{X}^\top \mathbf{X}\underset{\sim}{\beta} = \mathbf{X}^\top \underset{\sim}{y}$ of least squares.

Videos

Jon Krohn's "Calculus for Machine Learning" YouTube playlist has videos on the chain rule for partial derivatives:

- The Chain Rule for Partial Derivatives
- Exercises on the Multivariate Chain Rule

Definition 0.10 (Matrix derivative (gradient))**.** For a scalar-valued function $f(\mathbf{X})$ of an $m \times n$ matrix $\mathbf{X}$, the **matrix derivative** [2] is the $m \times n$ matrix whose (i, j) entry is the partial derivative of f with respect to the (i, j) entry of $\mathbf{X}$:

$$\left[\frac{\partial}{\partial \mathbf{X}} f\right]_{ij} = \frac{\partial}{\partial X_{ij}} f$$

Like the vector derivative (Definition 2 in Partial Derivatives and Gradients), the matrix derivative is also called the **gradient** of f with respect to $\mathbf{X}$, written $\nabla_{\mathbf{X}} f$.

Example 0.17 (The matrix derivative of a trace)**.** Let $\mathbf{X}$ be a 2×2 matrix and

$$\begin{aligned} f(\mathbf{X}) &= \operatorname{tr}(\mathbf{X}) \\ &= X_{11} + X_{22} \end{aligned}$$

(see trace). Then $\frac{\partial}{\partial X_{ij}} f = 1$ if $i = j$ and 0 otherwise, so:

$$\frac{\partial}{\partial \mathbf{X}} f = \mathbf{I}_2$$

Theorem 0.9 (Matrix derivative of the trace of a matrix product)**.** *If* $\mathbf{A}$ *$(r \times m)$ and* $\mathbf{B}$ *$(n \times r)$ are constant with respect to the $m \times n$ matrix* $\mathbf{X}$*, then:*

$$\underbrace{\frac{\partial}{\partial \mathbf{X}} \operatorname{tr}(\mathbf{AXB})}_{m \times n} = \underbrace{\mathbf{A}^\top}_{m \times r} \underbrace{\mathbf{B}^\top}_{r \times n}$$

Proof

Proof. Write A_{kl}, X_{kl}, and B_{kl} for the entries of $\mathbf{A}$, $\mathbf{X}$, and $\mathbf{B}$. For entry (i, j):

$$\begin{aligned}
\left[\frac{\partial}{\partial \mathbf{X}} \operatorname{tr}(\mathbf{AXB})\right]_{ij} &= \frac{\partial}{\partial X_{ij}} \sum_{a=1}^{r} \sum_{b=1}^{m} \sum_{c=1}^{n} A_{ab} X_{bc} B_{ca} && \text{(trace of the } r \times r \text{ product } \mathbf{AXB}) \\
&= \sum_{a=1}^{r} A_{ai} B_{ja} && \text{(only the terms with } b = i,\ c = j \text{ depend on } X_{ij}) \\
&= \sum_{a=1}^{r} [\mathbf{A}^\top]_{ia} [\mathbf{B}^\top]_{aj} && \text{(definition of the transpose)} \\
&= [\mathbf{A}^\top \mathbf{B}^\top]_{ij} && \text{(definition of matrix multiplication)}
\end{aligned}$$

□

Remark (Why the theorem uses the trace)

Remark 0.10 (Why the theorem uses the trace). The trace makes $\operatorname{tr}(\mathbf{AXB})$ a scalar, so its matrix derivative (Definition 0.10) is again an $m \times n$ matrix. The matrix product $\mathbf{AXB}$ itself (without the trace) is an $r \times r$ matrix, and each of its r^2 entries has a partial derivative with respect to each of the mn entries of $\mathbf{X}$. Those r^2mn partial derivatives form a four-index array (a fourth-order tensor), not a matrix, which is why this result is stated for the scalar $\operatorname{tr}(\mathbf{AXB})$.
For example, with

$$\begin{aligned}
\mathbf{A} &= \mathbf{B} \\
&= \mathbf{I}_2,
\end{aligned}$$

the product $\mathbf{AXB} = \mathbf{X}$ has 4 entries, each with 4 partial derivatives, 16 in all, while its trace $X_{11} + X_{22}$ has the 4 partial derivatives that form the 2×2 matrix $\mathbf{I}_2$ of Example 0.17.

Example 0.18 (Differentiating a weighted trace)**.** Let $\mathbf{A} = \mathbf{I}_2$ (2×2) and $\mathbf{B} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$ (2×2). Then $\operatorname{tr}(\mathbf{AXB}) = 2X_{11} + 3X_{22}$, and:

$$\underbrace{\frac{\partial}{\partial \mathbf{X}} \operatorname{tr}(\mathbf{AXB})}_{2\times 2} = \underbrace{\mathbf{A}^\top}_{2\times 2}\,\underbrace{\mathbf{B}^\top}_{2\times 2}$$
$$= \mathbf{I}_2 \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$$
$$= \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$$

Exercise 0.2 (A gradient that does not mention its variable)**.** Take

$$A = \begin{bmatrix} 2 & -1 & 0 \\ 4 & 3 & -2 \end{bmatrix}$$

let $W \in \mathbb{R}^{2\times 3}$ vary, and define

$$g(W) = \sum_{i=1}^{2} \sum_{j=1}^{3} W_{ij} A_{ij}$$

which is the entry-by-entry version of the inner product for matrices. Compute $\nabla_W g(W)$, and say what is unusual about the answer.

Solution

Solution 0.2. Written out over the six positions,

$$g(W) = 2W_{11} - W_{12} + 0\,W_{13} + 4W_{21} + 3W_{22} - 2W_{23}$$

Each W_{ij} appears in exactly one term, multiplied by A_{ij} and by nothing else, so differentiating with respect to it leaves A_{ij} behind:

$$\frac{\partial g}{\partial W_{ij}} = A_{ij}$$

Collecting those partial derivatives into a matrix of the same shape as W, as Table 1 in Partial Derivatives and Gradients requires,

$$\nabla_W g(W) = \begin{bmatrix} 2 & -1 & 0 \\ 4 & 3 & -2 \end{bmatrix}$$
$$= A$$

What is unusual is that the gradient came out **constant**: it does not mention W at all. The constant answer is not a quirk of this particular A. g is a linear function of W, and the gradient of a linear function is constant everywhere, for the same reason the derivative of $f(w) = cw$ is c no matter where it is evaluated.

A constant gradient is the easy case, and it is not the case we usually face. Most functions we minimize are curved — the residual sum of squares of a linear model (Definition 0.9) is the standard example — so their gradient changes from point to point and the downhill direction has to be worked out afresh at every step.

Theorem 0.10 (Gradient of an entrywise inner product with a fixed matrix)**.** *Let A be a fixed $m \times n$ matrix, and for $W \in \mathbb{R}^{m \times n}$ define*

$$g(W) = \sum_{i=1}^{m} \sum_{j=1}^{n} W_{ij} A_{ij}$$

Then for every W, $\nabla_W g(W) = A$. In particular, the gradient is constant: it does not depend on W.

Proof

Proof. This is the solution to Exercise 0.2, worked in Solution 0.2, where $m = 2$, $n = 3$, and the entries of A are given. The same argument works for every A: each W_{ij} appears in exactly one term, multiplied by A_{ij}, so $\frac{\partial g}{\partial W_{ij}} = A_{ij}$. □

[1] Wikipedia contributors, "Total derivative — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Total_derivative

[2] Wikipedia contributors, "Matrix calculus — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Matrix_calculus

[3] Wikipedia contributors, "Jacobian matrix and determinant — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Jacobian_matrix_and_determinant

[4] S. J. Miller, *The probability lifesaver: Calculus review problems.* 2016. Available: https://web.williams.edu/Mathematics/sjmiller/public_html/probabilitylifesaver/index.htm#:~:text=http%3A//web.williams.edu/Mathematics/sjmiller/public_html/probabilitylifesaver/supplementalchap_calcreview.pdf

[5] Wikipedia contributors, "Errors and residuals — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Errors_and_residuals

[6] Wikipedia contributors, "Residual sum of squares — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Residual_sum_of_squares

[7] C. A. Felippa, "Matrix calculus." Appendix F of Introduction to Finite Element Methods, course notes, University of Colorado Boulder. Accessed: Oct. 06, 2026. [Online]. Available: https://quickfem.com/wp-content/uploads/IFEM.AppF_.pdf

[8] Wikipedia contributors, "Gradient — Wikipedia, the free encyclopedia." Available: https://en.wikipedia.org/wiki/Gradient