Matrix calculus conventions

Ten problems on layout, shapes and the gradients of Ax, $x^T Ax$, $\|Ax - b\|^2$, traces and Frobenius norms, with full worked solutions and the mistakes that come from mixing conventions.

BEFORE YOU START

Most wrong answers in matrix calculus are not calculus errors. They are layout errors: a transpose in the wrong place, or a product whose shapes do not multiply. These ten problems fix one convention and use it throughout; every later page on this site uses the same one.

- Vectors are columns. $x \in \mathbb{R}^n$ is $n \times 1$.
- Numerator layout: for $y \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$, the derivative $\partial y / \partial x$ is the $m \times n$ Jacobian, with $(\partial y / \partial x)_{ij} = \partial y_i / \partial x_j$. Rows follow the output, columns follow the input.
- For a scalar f, $\partial f / \partial x$ is therefore the $1 \times n$ row. The gradient is its transpose, $\nabla_x f = (\partial f / \partial x)^\top$, the $n \times 1$ column. Differentials read off the row: if $df = g^\top dx$ then $\nabla_x f = g$.
- For a scalar f of a matrix $X \in \mathbb{R}^{m \times n}$, $\nabla_X f$ has the shape of X, with $(\nabla_X f)_{ij} = \partial f / \partial X_{ij}$. In differential form: if $df = \operatorname{tr}(G^\top dX)$ then $\nabla_X f = G$.
- $\|\cdot\|$ with no subscript means the Euclidean norm $\|\cdot\|_2$.

If a textbook you use writes $\partial (Ax)/\partial x = A^\top$, it is using denominator layout. Every Jacobian and row derivative here transposes under that convention; the gradients, which take the shape of the input, do not. Neither convention is wrong; mixing them within one derivation is.

Problems

1. Shapes. For $y = f(x)$ with $y \in \mathbb{R}^m$, $x \in \mathbb{R}^n$, give the shape of $\partial y / \partial x$. For scalar $f(x)$, give the shapes of $\partial f / \partial x$ and $\nabla_x f$. For scalar $f(X)$ with $X \in \mathbb{R}^{m \times n}$, give the shape of $\nabla_X f$ and say what its (i, j) entry is. difficulty ·

2. Let $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$. Compute $\partial (Ax)/\partial x$ and state its shape. difficulty ·

3. Let $a, x \in \mathbb{R}^n$. Compute $\nabla_x (a^\top x)$. difficulty ·

4. Let $A \in \mathbb{R}^{n \times n}$ (not necessarily symmetric), $x \in \mathbb{R}^n$. Compute $\nabla_x (x^\top A x)$. What does it become when A is symmetric? difficulty ··

5. Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. Compute $\nabla_x \|Ax - b\|_2^2$. difficulty ··

6. Set the gradient from Problem 5 to zero and derive the normal equations. What condition on A makes the solution unique? difficulty ··

7. Let $A \in \mathbb{R}^{m \times n}$, $X \in \mathbb{R}^{n \times m}$. Compute $\nabla_X \operatorname{tr}(AX)$. difficulty ··

8. Let $A \in \mathbb{R}^{n \times n}$, $X \in \mathbb{R}^{n \times m}$. Compute $\nabla_X \operatorname{tr}(X^\top A X)$. difficulty ···

9. For $x \neq 0$, compute $\nabla_x \|x\|_2$ (the norm itself, not its square). difficulty ··

10. Let $X \in \mathbb{R}^{N \times d}$, $W \in \mathbb{R}^{d \times k}$, $Y \in \mathbb{R}^{N \times k}$. Compute $\nabla_X \|X\|_F^2$ and $\nabla_W \|XW - Y\|_F^2$. difficulty ···

Answers

1. $\partial y/\partial x$: $m \times n$; $\partial f/\partial x$: $1 \times n$; $\nabla_x f$: $n \times 1$; $\nabla_X f$: $m \times n$ with $(\nabla_X f)_{ij} = \partial f / \partial X_{ij}$.
2. $\partial (Ax)/\partial x = A$, shape $m \times n$.
3. $\nabla_x (a^\top x) = a$
4. $\nabla_x (x^\top A x) = (A + A^\top)x$, and $2Ax$ when $A = A^\top$
5. $\nabla_x \|Ax - b\|_2^2 = 2A^\top (Ax - b)$
6. $A^\top A\, x = A^\top b$; unique iff A has full column rank ($\operatorname{rank} A = n$), since then $A^\top A$ is invertible.
7. $\nabla_X \operatorname{tr}(AX) = A^\top$
8. $\nabla_X \operatorname{tr}(X^\top A X) = (A + A^\top)X$
9. $\nabla_x \|x\|_2 = x / \|x\|_2$
10. $\nabla_X \|X\|_F^2 = 2X$; $\nabla_W \|XW - Y\|_F^2 = 2X^\top (XW - Y)$

Worked solutions

Problem 1

> **Shapes.** For $y = f(x)$ with $y \in \mathbb{R}^m$, $x \in \mathbb{R}^n$, give the shape of $\partial y / \partial x$. For scalar $f(x)$, give the shapes of $\partial f / \partial x$ and $\nabla_x f$. For scalar $f(X)$ with $X \in \mathbb{R}^{m \times n}$, give the shape of $\nabla_X f$ and say what its (i, j) entry is.

$\partial y / \partial x$ has one row per y_i and one column per x_j: $m \times n$.	Numerator layout: the output indexes the rows, the input the columns.
A scalar f is the case $m = 1$, so $\partial f / \partial x$ is $1 \times n$.	Same rule with a one-row output.
$\nabla_x f = (\partial f / \partial x)^\top$ is $n \times 1$.	The gradient is defined as the transpose of the row, so it has the shape of x and can be added to x in a gradient step.
$\nabla_X f$ is $m \times n$, entry (i, j) is $\partial f / \partial X_{ij}$.	A matrix gradient is arranged in the shape of its input, entry by entry, for the same reason: $X - \eta \nabla_X f$ must make sense.
$\partial y / \partial x$: $m \times n$; $\partial f / \partial x$: $1 \times n$; $\nabla_x f$: $n \times 1$; $\nabla_X f$: $m \times n$ **with** $(\nabla_X f)_{ij} = \partial f / \partial X_{ij}$.	

Problem 2

> Let $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$. Compute $\partial (Ax) / \partial x$ and state its shape.

$y = Ax$, so $y_i = \sum_j A_{ij} x_j$.	Index form turns the matrix derivative into ordinary partial derivatives.
$\partial y_i / \partial x_j = A_{ij}$.	Each x_j appears in exactly one term of the sum for y_i, with coefficient A_{ij}.
$(\partial y / \partial x)_{ij} = A_{ij}$.	In numerator layout entry (i, j) of the Jacobian is $\partial y_i / \partial x_j$.

$\partial(Ax)/\partial x = A$, **shape** $m \times n$.

Another route: with differentials

$dy = A\,dx$.	A is constant, so the differential passes through it.
$dy = J\,dx$ defines the Jacobian J.	The Jacobian is the matrix that maps a small input change to the output change.
$\partial(Ax)/\partial x = A$, **shape** $m \times n$.	

Problem 3

> Let $a, x \in \mathbb{R}^n$. Compute $\nabla_x(a^\top x)$.

$f = a^\top x = \sum_k a_k x_k$.	Writing the inner product out as a sum shows which term contains each x_j.
$\partial f/\partial x_j = a_j$.	Only the $k = j$ term depends on x_j.
$\partial f/\partial x = a^\top$, a $1 \times n$ row.	This is Problem 2 with $A = a^\top$ and $m = 1$.
$\nabla_x(a^\top x) = a$	The gradient is the transpose of the row derivative, so it is the column a, the same shape as x.

Problem 4

> Let $A \in \mathbb{R}^{n \times n}$ (not necessarily symmetric), $x \in \mathbb{R}^n$. Compute $\nabla_x(x^\top A x)$. What does it become when A is symmetric?

$f = x^\top A x = \sum_{i,j} A_{ij} x_i x_j$.	Index form of the quadratic form; x appears twice, once on each side of A.

$\dfrac{\partial f}{\partial x_k} = \sum_j A_{kj} x_j + \sum_i A_{ik} x_i.$	Product rule on $x_i x_j$: the first sum collects the terms with $i = k$, the second the terms with $j = k$. The $i = j = k$ term appears in both, which is correct since it is $A_{kk} x_k^2$ with derivative $2A_{kk} x_k$.
$\dfrac{\partial f}{\partial x_k} = (Ax)_k + (A^\top x)_k.$	$\sum_j A_{kj} x_j$ is row k of A times x; $\sum_i A_{ik} x_i$ is column k of A times x, which is row k of $A^\top$ times x.
Stacking over k: $\nabla_x f = Ax + A^\top x.$	The gradient is the column of these partials.
$\nabla_x (x^\top A x) = (A + A^\top)x$, **and** $2Ax$ **when** $A = A^\top$	For symmetric A the two terms are equal.

Another route: with differentials

$f = x^\top A x$, so $df = dx^\top A x + x^\top A \, dx.$	Product rule; A is constant.
$dx^\top A x = x^\top A^\top dx.$	It is a scalar, so it equals its own transpose.
$df = x^\top (A^\top + A) \, dx.$	Collect the two terms so dx is on the right.
$\partial f / \partial x = x^\top (A + A^\top).$	The row derivative is whatever multiplies dx.
$\nabla_x (x^\top A x) = (A + A^\top)x$, **and** $2Ax$ **when** $A = A^\top$	Transpose the row: $(x^\top (A + A^\top))^\top = (A + A^\top)x$, since $A + A^\top$ is symmetric.

Problem 5

> Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. Compute $\nabla_x \|Ax - b\|_2^2$.

$f = (Ax - b)^\top (Ax - b).$	The squared 2-norm of a vector is its inner product with itself.

Let $r = Ax - b$; then $dr = A\,dx$.	b is constant and $d(Ax) = A\,dx$ from Problem 2.
$df = dr^\top r + r^\top dr = 2r^\top dr.$	Product rule; $dr^\top r$ is a scalar, so it equals $r^\top dr$.
$df = 2(Ax - b)^\top A\,dx.$	Substitute dr. Check the shapes: $(1 \times m)(m \times n)$ is the $1 \times n$ row derivative.
$\nabla_x \lVert Ax - b \rVert_2^2 = 2A^\top (Ax - b)$	The gradient is the transpose of the row: $(2(Ax - b)^\top A)^\top = 2A^\top (Ax - b)$, which is $n \times 1$ like x.

Problem 6

> Set the gradient from Problem 5 to zero and derive the normal equations. What condition on A makes the solution unique?

$2A^\top (Ax - b) = 0.$	$f(x) = \lVert Ax - b \rVert_2^2$ is a convex quadratic, so a point where its gradient vanishes is a global minimizer.
$A^\top A\,x - A^\top b = 0.$	Divide by 2 and distribute $A^\top$.
$A^\top A$ is $n \times n$.	The system is square, so the solution is unique exactly when $A^\top A$ is invertible, that is, when its null space is $\{0\}$.
$A^\top A v = 0 \iff \lVert Av \rVert^2 = 0 \iff Av = 0.$	Multiplying $A^\top A v = 0$ on the left by $v^\top$ gives $v^\top A^\top A v = \lVert Av \rVert^2 = 0$; the other direction is immediate. So $A^\top A$ and A have the same null space.
The normal equations always have a solution.	$A^\top b$ lies in the column space of $A^\top$, the orthogonal complement of the null space of A. $A^\top A$ is symmetric, so its column space is the orthogonal complement of its own null space, which is the same one (step 4). So the only question is uniqueness.

$A^\top A\,x = A^\top b$; **unique iff A has full column rank (**$\operatorname{rank} A = n$**), since then $A^\top A$ is invertible.**	By steps 3 and 4, $A^\top A$ is invertible iff A has a trivial null space, which is iff its columns are independent. That needs $m \geq n$, at least as many rows as columns.

Problem 7

> Let $A \in \mathbb{R}^{m\times n}$, $X \in \mathbb{R}^{n\times m}$. Compute $\nabla_X \operatorname{tr}(AX)$.

$f = \operatorname{tr}(AX)$; $(AX)_{ii} = \sum_j A_{ij}X_{ji}$, so $f = \sum_{i,j} A_{ij}X_{ji}$.	The trace sums the diagonal; write each diagonal entry as a row of A times a column of X.
$\partial f / \partial X_{ji} = A_{ij}$.	X_{ji} appears in exactly one term of the double sum, with coefficient A_{ij}.
$(\nabla_X f)_{ji} = A_{ij} = (A^\top)_{ji}$.	Entry (j,i) of the gradient is the partial with respect to X_{ji}; the indices on A come out swapped.
$\nabla_X \operatorname{tr}(AX) = A^\top$	Shape check: $A^\top$ is $n \times m$, the shape of X.

Problem 8

> Let $A \in \mathbb{R}^{n\times n}$, $X \in \mathbb{R}^{n\times m}$. Compute $\nabla_X \operatorname{tr}(X^\top AX)$.

$f = \operatorname{tr}(X^\top AX)$, so $df = \operatorname{tr}(dX^\top AX) + \operatorname{tr}(X^\top A\,dX)$.	Product rule inside the trace; the trace is linear, so it passes through d.
$\operatorname{tr}(dX^\top AX) = \operatorname{tr}(X^\top A^\top dX)$.	$\operatorname{tr}(M^\top) = \operatorname{tr}(M)$ with $M = dX^\top AX$.
$df = \operatorname{tr}((AX)^\top dX) + \operatorname{tr}((A^\top X)^\top dX)$.	Rewrite $X^\top A^\top = (AX)^\top$ and $X^\top A = (A^\top X)^\top$ to get both terms in the form $\operatorname{tr}(G^\top dX)$.
$df = \operatorname{tr}(((A + A^\top)X)^\top dX)$.	Linearity of the trace and of the transpose.

$\nabla_X \operatorname{tr}(X^\top A X) = (A + A^\top) X$	From $df = \operatorname{tr}(G^\top dX)$ the gradient is G, because $\operatorname{tr}(G^\top dX) = \sum_{ij} G_{ij}\, dX_{ij}$. With $m = 1$ this is Problem 4.

Problem 9

For $x \neq 0$, compute $\nabla_x \|x\|_2$ (the norm itself, not its square).

$\|x\|_2 = (x^\top x)^{1/2}.$	Write the norm as a scalar function of the squared norm, whose gradient is easy.
$\nabla_x (x^\top x) = 2x.$	Problem 4 with $A = I$.
$\nabla_x \|x\|_2 = \frac{1}{2}(x^\top x)^{-1/2} \cdot 2x.$	Chain rule for $g(u) = u^{1/2}$ with $u = x^\top x$: the scalar $g'(u)$ multiplies the gradient of u.
$\nabla_x \|x\|_2 = x / \|x\|_2$	Simplify; $(x^\top x)^{1/2} = \|x\|_2$. This is a unit vector along x. At $x = 0$ the norm has a corner and no gradient, which is why the problem excludes it.

Problem 10

Let $X \in \mathbb{R}^{N \times d}$, $W \in \mathbb{R}^{d \times k}$, $Y \in \mathbb{R}^{N \times k}$. Compute $\nabla_X \|X\|_F^2$ and $\nabla_W \|XW - Y\|_F^2$.

$\|X\|_F^2 = \sum_{ij} X_{ij}^2 = \operatorname{tr}(X^\top X).$	The Frobenius norm squares every entry; the diagonal of $X^\top X$ holds the squared column norms.
$\partial \|X\|_F^2 / \partial X_{ij} = 2X_{ij}$, so $\nabla_X \|X\|_F^2 = 2X.$	Each entry appears in one term of the sum. Equivalently, Problem 8 with $A = I$.
Let $R = XW - Y$ and $f = \|R\|_F^2 = \operatorname{tr}(R^\top R)$; then $dR = X\, dW$.	X and Y are constant here; only W varies.

$df = \operatorname{tr}(dR^\top R) + \operatorname{tr}(R^\top dR) = 2\operatorname{tr}(R^\top dR)$.	Product rule, then $\operatorname{tr}(dR^\top R) = \operatorname{tr}(R^\top dR)$ because a matrix and its transpose have the same trace.
$df = 2\operatorname{tr}((XW - Y)^\top X\, dW)$.	Substitute $dR = X\, dW$.
$df = \operatorname{tr}(G^\top dW)$ with $G = 2X^\top(XW - Y)$.	$(XW - Y)^\top X = (X^\top(XW - Y))^\top$; writing the factor in front of dW as $G^\top$ puts it in the form that identifies the gradient.
$\nabla_X \lVert X\rVert_F^2 = 2X$; $\nabla_W \lVert XW - Y\rVert_F^2 = 2X^\top(XW - Y)$	Read G off $df = \operatorname{tr}(G^\top dW)$. Shape check: $(d \times N)(N \times k) = d \times k$, the shape of W. With $k = 1$ this is Problem 5.

Where this goes wrong

1. Transposing the Jacobian by habit

The gradient of $a^\top x$ is a, a column, and it is easy to carry that transpose reflex over to Ax.

$y = Ax$, $y_i = \sum_j A_{ij} x_j$	Right so far: this is the index form, and it already determines every entry of the Jacobian.
$\nabla_x(a^\top x) = a$, so "the derivative of a linear map is its coefficient, transposed"	The analogy that causes the mistake. The transpose in $\nabla_x(a^\top x) = a$ belongs to the gradient, not to the linear map.
$\partial(Ax)/\partial x = A^\top$ ✗	The habit from $\nabla_x(a^\top x) = a$, where the transpose comes from turning a row derivative into a gradient. A Jacobian is not a gradient: entry (i, j) is $\partial y_i/\partial x_j = A_{ij}$, so it is A. $A^\top$ is $n \times m$, the wrong shape unless $m = n$, and even then wrong unless A is symmetric.

2. Forgetting the transpose term in $x^T Ax$

The scalar rule $\frac{d}{dx}(ax^2) = 2ax$ suggests an answer that is only half right.

$f = x^\top Ax$	Right so far: the function is stated correctly, and nothing has been differentiated yet.
"This is $x^\top Ax$, the matrix version of ax^2."	The analogy that causes the mistake.
$\nabla_x(x^\top Ax) = 2Ax$ ✗	Only right when A is symmetric. The two occurrences of x give Ax and $A^\top x$ (Problem 4), and they coincide only if $A = A^\top$. Most textbook examples use a covariance or Hessian, which are symmetric, so the shortcut sticks.

3. Chain rule with the shapes in the wrong order

The scalar chain rule gives $\frac{d}{dx}(ax - b)^2 = 2(ax - b)a$, and copying that order to vectors produces a product that does not exist.

$f = \|Ax - b\|^2$, outer derivative $2(Ax - b)$, inner derivative A	Both pieces are right; the error only appears when they are multiplied.
$\nabla_x\|Ax - b\|^2 = 2(Ax - b)A$ ✗	$(m \times 1)(m \times n)$ does not multiply. Scalars commute, so the scalar chain rule never forced an order. The fix is not to transpose factors until it fits but to write the row derivative $2(Ax - b)^\top A$ and transpose the whole thing, giving $2A^\top(Ax - b)$.

4. Treating the trace as a constant

Linearity makes $\mathrm{tr}(AX)$ look like "A times X", and the derivative of aX is a.

$f = \mathrm{tr}(AX)$ is linear in X	True, and useful: the gradient is constant. Linearity says the gradient does not depend on X; it does not say which constant it is.

$\nabla_X \operatorname{tr}(AX) = A$ ✗	Borrowed from $\frac{d}{dx}(ax) = a$. Index it out: $\operatorname{tr}(AX) = \sum_{i,j} A_{ij} X_{ji}$, so the entry that multiplies X_{ji} is A_{ij}, which puts $A^\top$ in position. The shape gives it away too: A is $m \times n$ and X is $n \times m$.

5. Differentiating the norm as if it were the squared norm

Least squares uses $\|x\|^2$ so often that $2x$ becomes the reflex for anything with a norm in it.

$f = \|x\|_2$	Right so far: the function is the norm, not its square, which is what the next line forgets.
$\nabla_x \|x\|_2 = 2x$ ✗	That is the gradient of $\|x\|^2$. The norm is $(x^\top x)^{1/2}$, and the chain rule adds the factor $\frac{1}{2}(x^\top x)^{-1/2}$, giving $x/\|x\|_2$. A quick sanity check: $\|x\|_2$ grows at rate 1 along x, so its gradient must have length 1, not $2\|x\|$.

More at eulerpad.app. This is one of 82 free practice sets with worked solutions at eulerpad.app/practice, alongside guides to the topics behind them.

EulerPad, the app, reads your handwritten math on iPad, iPhone and Android and checks it line by line. It's free to start.

This set is also Chapter 2 of *Deep Learning Math by Hand*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/matrix-calculus-conventions