Matrix Calculus

Massimiliano Tomassoli

23/7/2013

Matrix Calculus is a set of techniques which let us differentiate functions of matrices without computing the single partial derivatives by hand. What this means will be clear in a moment.

1 The derivative

Let's talk about notation a little bit. If $f(X)$ is a scalar function of an $m \times n$ matrix X, then

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

The definition above is valid even if $m = 1$ or $n = 1$, that is if f is a function of a row vector, a column vector or a scalar, in which case the result is a row vector, a column vector or a scalar, respectively.

If $f(X)$ is an $m \times n$ matrix function of a matrix, then

$$\frac{\partial f(X)}{\partial X} = \begin{bmatrix} \frac{\partial f_{11}(X)}{\partial X} & \cdots & \frac{\partial f_{1n}(X)}{\partial X} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_{m1}(X)}{\partial X} & \cdots & \frac{\partial f_{mn}(X)}{\partial X} \end{bmatrix}$$

where the matrix above is a *block matrix*. The definition above is valid even when $m = 1$ or $n = 1$, that is when f is a row vector function, a column vector function or a scalar function, in which case the block matrix is a row of blocks, a column of blocks or just a single block, respectively.

If f is

- a *scalar* function of a *scalar*, *vector* or *matrix*, or
- a *vector* function of a *scalar* or *vector*, or
- a *matrix* function of a *scalar*,

then the *derivative*, also called *Jacobian matrix*, of f is

$$Df(x) = \frac{\partial f(x)}{\partial x^T}$$

For instance, if $f(x)$ is a vector function of a vector, then

$$Df(x) := \frac{\partial f(x)}{\partial x^T} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x^T} \\ \vdots \\ \frac{\partial f_m(x)}{\partial x^T} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix}.$$

As another example, if $f(X)$ is a scalar function of a matrix, then

$$Df(X) := \frac{\partial f(X)}{\partial X^T} = \begin{bmatrix} \frac{\partial f(X)}{\partial x_{11}} & \cdots & \frac{\partial f(X)}{\partial x_{m1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(X)}{\partial x_{1n}} & \cdots & \frac{\partial f(X)}{\partial x_{mn}} \end{bmatrix}.$$

Some authors prefer to find the *gradient* of a function, which is defined as the *transpose* of the Jacobian matrix.

Now, how do we define the derivative of an $m \times n$ matrix function f of a $p \times q$ matrix? It's clear that we have $mnpq$ partial derivatives. We could just define the derivative of f as we did above for the other cases, but we'll follow Magnus's way [1, 2] and give the following definition:

$$Df(X) = \frac{\partial \operatorname{vec} f(X)}{\partial (\operatorname{vec} X)^T}$$

The result is an $mn \times pq$ matrix. We'll talk about the *vec* operation in a moment. For now let's just say that it *vectorizes* a matrix by stacking its columns on top of one another. More formally,

$$\operatorname{vec}(A) = \begin{bmatrix} a_{11} & \cdots & a_{m1} & a_{12} & \cdots & a_{m2} & \cdots\cdots & a_{1n} & \cdots & a_{mn} \end{bmatrix}^T$$

This means that $\operatorname{vec} f(X)$ is a vector function and $\operatorname{vec} X$ is a vector.

So what about a scalar function of a matrix? According to this last definition, the derivative should be a row vector, while according to our first definition of derivative it should be a matrix. Both are viable options and we'll see how easy it is to go from one to the other.

The second derivative in the multidimensional case is the *Hessian matrix* (or just *Hessian*), a square matrix of second-order partial derivatives of a scalar function. More precisely, if f is a scalar function of a vector, then the Hessian is defined as

$$Hf(x) = \frac{\partial^2 f(x)}{\partial x \partial x^T} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

Note that the Hessian is the Jacobian of the gradient (i.e. the transpose of the Jacobian) of f:

$$Hf(x) = \frac{\partial^2 f(x)}{\partial x \partial x^T} = \frac{\partial}{\partial x^T} \frac{\partial f(x)}{\partial x} = J(\nabla f)(x)$$

If f is a scalar function of a matrix (rather than of a vector), we vectorize the matrix:

$$Hf(X) = \frac{\partial^2 f(X)}{\partial (\operatorname{vec} X) \partial (\operatorname{vec} X)^T}$$

In this article we assume that the second-order partial derivatives are *continuous* so the order of differentiation is irrelevant and the Hessian is always symmetric.

2 The differential

The method described here is based on *differentials*, therefore let's recall what a differential is.

In the one-dimensional case, the derivative of f at x is defined as

$$\lim_{u \to 0} \frac{f(x+u) - f(x)}{u} = f'(x)$$

This can be rewritten as

$$f(x+u) = f(x) + f'(x)u + r_x(u)$$

where $r_x(u)$ is $o(u)$, i.e. $r_x(u)/u \to 0$ as $u \to 0$. The differential of f at x with increment u is $df(x;u) = f'(x)u$.

In the vector case, we have

$$f(x+u) = f(x) + (Df(x))u + r_x(u)$$

and the differential of f at x with increment u is $df(x; u) = (Df(x))u$. As we can see, the differential is the best linear approximation of $f(x + u) - f(x)$ at x. In practice, we write dx instead of u, so, for instance, $df(x; u) = f'(x)dx$. We'll justify this notation in a moment, but first let's introduce the so-called *identification* results. They are needed to get the derivatives from the differentials.

The first result says that, if f is a vector function of a vector,

$$df(x) = A(x)dx \iff Df(x) = A(x).$$

More generally, if f is a matrix function of a matrix,

$$d \operatorname{vec} f(X) = A(X) d \operatorname{vec} X \iff Df(x) = A(X).$$

The second result is about the second differential and says that if f is a scalar function of a vector, then

$$d^2 f(x) = (dx)^T B(x) dx \iff Hf(x) = \frac{1}{2}(B(x) + B(x)^T)$$

where $Hf(x)$ denotes the *Hessian matrix*.

Another important result is *Cauchy's rule of invariance* which says that if $h(x) = g(f(x))$, then

$$dh(x; u) = dg(f(x); df(x; u))$$

This is related to the *chain rule* for the derivatives; in fact, it can be proved by making use of the chain rule:

$$\begin{aligned} dh(x; u) &= Dh(x)u \\ &= D(g \circ f)(x)u \\ &= Dg(f(x))Df(x)u \\ &= Dg(f(x))df(x; u) \\ &= dg(f(x); df(x; u)) \end{aligned}$$

Now we can justify the abbreviated notation dx. If $y = f(x)$, then we write

$$dy = df(x; dx)$$

where x and y are *variables*. Basically, we name the differential after the variables rather than after the functions. But now suppose that $x = g(t)$ and $dx = dg(t; dt)$. We now have $y = f(g(t)) = h(t)$ and, therefore,

$$dy = dh(t; dt).$$

For our abbreviated notation to be consistent, it must be the case that $df(x; dx) = dh(t; dt)$. Fortunately, thanks to Cauchy's rule of invariance, we can see that it is so:

$$dh(t; dt) = d(f \circ g)(t; dt) = df(g(t); dg(t; dt)) = df(x; dx).$$

3 Two important operators

Before proceeding with the actual computation of differentials and derivatives, we need to introduce two important operators: the *Kronecker product* and the *vec* operator. We've already talked a little about the vec operator but here we'll see and prove some useful results for manipulating expressions involving these two operators.

As we said before, the vec operator is defined as follows:

$$\operatorname{vec}(A) = \begin{bmatrix} a_{11} & \cdots & a_{m1} & a_{12} & \cdots & a_{m2} & \cdots\cdots & a_{1n} & \cdots & a_{mn} \end{bmatrix}^T$$

For instance,

$$\operatorname{vec} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 4 & 2 & 5 & 3 & 6 \end{bmatrix}^T$$

The Kronecker product between two matrix A and B is defined as follows:

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}$$

For instance,

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 & 2 \\ 1 & 1 & 2 & 2 \\ 3 & 3 & 4 & 4 \\ 3 & 3 & 4 & 4 \end{bmatrix}$$

Here's a list of properties of the Kronecker product and the vec operator:

1. $A \otimes (B \otimes C) = (A \otimes B) \otimes C \quad$ (associativity)
2. $A \otimes (B + C) = (A \otimes B) + (A \otimes C)$
   $(A + B) \otimes C = (A \otimes C) + (B \otimes C) \quad$ (distributivity)
3. $\forall a \in \mathbb{R}, \quad a \otimes A = A \otimes a = aA$
4. $\forall a, b \in \mathbb{R}, \quad aA \otimes bB = ab(A \otimes B)$
5. For conforming matrices, $\quad (A \otimes B)(C \otimes D) = AC \otimes BD$
6. $(A \otimes B)^T = A^T \otimes B^T, \quad (A \otimes B)^H = A^H \otimes B^H$
7. For all vectors a and b, $\quad a^T \otimes b = ba^T = b \otimes a^T$
8. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$
9. The vec operator is linear.
10. For all vectors a and b, $\quad \mathrm{vec}(ab^T) = b \otimes a$
11. $\mathrm{vec}(AXC) = (C^T \otimes A)\,\mathrm{vec}(X)$
12. $\mathrm{tr}(AB) = \mathrm{vec}(A^T)^T\,\mathrm{vec}(B)$

We'll denote with $\{f(i,j)\}$ the matrix whose generic (i,j) element is $f(i,j)$. Note that if A is a block matrix with blocks B_{ij}, then

$$A \otimes C = \begin{bmatrix} B_{11} \otimes C & \cdots & B_{1n} \otimes C \\ \vdots & \ddots & \vdots \\ B_{m1} \otimes C & \cdots & B_{mn} \otimes C \end{bmatrix}$$

Point (3) follows directly from the definition, while point (4) can be proved as follows:

$$xA \otimes yB = \{xa_{ij}(yB)\} = \{xya_{ij}B\} = xy\{a_{ij}B\} = xy(A \otimes B)$$

Now let's prove the other points one by one.

1. $A \otimes (B \otimes C) = \{a_{ij}(B \otimes C)\} = \{a_{ij}B \otimes C\} = \{a_{ij}B\} \otimes C = (A \otimes B) \otimes C$
2. $A \otimes (B + C) = \{a_{ij}(B + C)\} = \{a_{ij}B + a_{ij}C\} = \{a_{ij}B\} + \{a_{ij}C\} = A \otimes B + A \otimes C$
   The other case is analogous.
3. See above.
4. See above.

5. $(A \otimes B)(C \otimes D) = \{a_{ij}B\}\{c_{ij}D\} = \{\sum_k a_{ik}Bc_{kj}D\} = \{(\sum_k a_{ik}c_{kj})\, BD\} = \{\sum_k a_{ik}c_{kj}\} \otimes BD = AC \otimes BD$

6. $(A \otimes B)^T = \{a_{ij}B\}^T = \{a_{ji}B^T\} = A^T \otimes B^T$
   The other case is analogous.

7. This is very easy.

8. By (5), $\quad (A \otimes B)(A^{-1} \otimes B^{-1}) = AA^{-1} \otimes BB^{-1} = I \otimes I = I$

9. This is very easy.

10. $\mathrm{vec}(ab^T) = \mathrm{vec}\begin{bmatrix} a_1b_1 & \cdots & a_1b_n \\ \vdots & \ddots & \vdots \\ a_mb_1 & \cdots & a_mb_n \end{bmatrix} = \mathrm{vec}\begin{bmatrix} b_1a & \cdots & b_na \end{bmatrix} = \begin{bmatrix} b_1a \\ \vdots \\ b_na \end{bmatrix} = b \otimes a$

11. Let $x_1, \ldots, x_n$ be the columns of the matrix X and $e_1, \ldots, e_n$ the columns of the identity matrix of order n. You should convince yourself that $X = \sum_k x_k e_k^T$. Here we go:

$$
\begin{aligned}
\mathrm{vec}(AXC) &= \mathrm{vec}\left(A\left(\sum_k x_k e_k^T\right)C\right) \\
&= \mathrm{vec}\left(\sum_k (Ax_k)(e_k^T C)\right) \\
&= \sum_k \mathrm{vec}((Ax_k)(e_k^T C)) && \text{(by (9))} \\
&= \sum_k ((e_k^T C)^T \otimes (Ax_k)) && \text{(by (10))} \\
&= \sum_k ((C^T e_k) \otimes (Ax_k)) \\
&= \sum_k ((C^T \otimes A)(e_k \otimes x_k)) && \text{(by (5))} \\
&= (C^T \otimes A)\sum_k (e_k \otimes x_k) \\
&= (C^T \otimes A)\sum_k \mathrm{vec}(x_k e_k^T) && \text{(by (10))} \\
&= (C^T \otimes A)\,\mathrm{vec}\sum_k (x_k e_k^T) && \text{(by (9))} \\
&= (C^T \otimes A)\,\mathrm{vec}(X)
\end{aligned}
$$

12. The trace of the product AB is just the sum of all the products $a_{ij}b_{ij}$:

$$\mathrm{tr}(AB) = \sum_i [AB]_{ii} = \sum_i \sum_k a_{ik}b_{ki} = \sum_i \sum_k [A^T]_{ki}b_{ki} = \mathrm{vec}(A^T)^T\,\mathrm{vec}(B)$$

Since we'll be using the *trace* operator quite a bit, recall that:

1. tr is linear.

2. $\mathrm{tr}(A) = \mathrm{tr}(A^T)$

3. $\mathrm{tr}(AB) = \mathrm{tr}(BA)$

The first two properties are trivial (but still very useful). Let's see why the last one is also true:

$$\operatorname{tr}(AB) = \sum_i [AB]_{ii} = \sum_i \sum_k a_{ik} b_{ki} = \sum_k \sum_i b_{ki} a_{ik} = \sum_k [BA]_{kk} = \operatorname{tr}(BA)$$

Also note that the trace of a scalar is the scalar itself. This means that when we have a scalar function we can add a trace. This simple observation will be very useful.

4 Basic rules of differentiation

In order to be able to differentiate expressions, we'll need a set of simple rules:

1. $dA = 0, \qquad$ where A is constant
2. $d(\alpha X) = \alpha dX, \qquad$ where α is a scalar
3. $d(X + Y) = dX + dY$
4. $d(\operatorname{tr}(X)) = \operatorname{tr}(dX)$
5. $d(XY) = (dX)Y + XdY$
6. $d(X \otimes Y) = (dX) \otimes Y + X \otimes dY$
7. $d(X^{-1}) = -X^{-1}(dX)X^{-1}$
8. $d|X| = |X| \operatorname{tr}(X^{-1}dX)$
9. $d \log |X| = \operatorname{tr}(X^{-1}dX)$
10. $d(X^\star) = (dX)^\star, \qquad$ where * is any operator which rearranges elements such as *transpose* and *vec*

A matrix is really a matrix of scalar functions and the differential of a matrix is the matrix of the differentials of the single scalar functions. More formally, $[dX]_{ij} = d(X_{ij})$. Remember that if f is a scalar function of a vector, then $df(x; u) = \sum_i D_i f(x) u_i$, where the $D_i f(x)$ are the partial derivatives of f at x. If f is, instead, a function of a matrix, we just generalize the previous relation: $df(x; u) = \sum_{ij} D_{ij} f(x) u_{ij}$. That said, many of the rules above can be readily proved.

As an example, let's prove the *product rule* (5). Let f and g be two scalar functions of a matrix. Then

$$\begin{aligned}
d(fg)(x; u) &= \sum_{i,j} D_{ij}(fg)(x) u_{ij} \\
&= \sum_{i,j} ((D_{ij} f(x)) g(x) + f(x) D_{ij} g(x)) u_{ij} \\
&= \sum_{i,j} (D_{ij} f(x)) u_{ij} g(x) + \sum_{i,j} f(x) D_{ij} g(x) u_{ij} \\
&= df(x; u) g(x) + f(x) dg(x; u)
\end{aligned}$$

where we used the usual product rule for derivatives. Now we can use this result to prove the general one about matrices of scalar functions:

$$
\begin{aligned}
[d(XY)]_{ij} &= d[XY]_{ij} \\
&= d\left(\sum_k x_{ik}y_{kj}\right) \\
&= \sum_k d(x_{ik}y_{kj}) \\
&= \sum_k (dx_{ik}y_{kj} + \sum_k x_{ik}dy_{kj} \\
&= \sum_k [dX]_{ik}y_{kj} + \sum_k x_{ik}[dY]_{kj} \\
&= [(dX)Y]_{ij} + [XdY]_{ij} \\
&= [(dX)Y + XdY]_{ij}
\end{aligned}
$$

Now we can prove (7) by using the product rule:

$$0 = dI = d(XX^{-1}) = (dX)X^{-1} + Xd(X^{-1}) \implies dX^{-1} = -X^{-1}(dX)X^{-1}$$

To prove (8) we observe that, for any $i = 1, \ldots, n$, we have $|X| = \sum_j x_{ij}C_{ij}$, where C_{ij} is the cofactor of the element x_{ij}, i.e. $(-1)^{i+j}$ times the determinant of the matrix obtained by removing from X the i-th row and the j-th column. Because C_{ij} doesn't depend on x_{ij}, then

$$\frac{\partial |X|}{\partial x_{ij}} = \sum_j \frac{\partial x_{ij}C_{ij}}{\partial x_{ij}} = C_{ij}.$$

Now note that C is the matrix of the cofactors and recall that $X^{-1} = \frac{1}{|X|}C^T$ and thus $C^T = |X|X^{-1}$. This last result will be used in the following derivation:

$$d|X| = d|\cdot|(X;dX) = \sum_{i,j} C_{ij}[dX]_{ij} = \sum_{i,j}[C^T]_{ji}[dX]_{ij} = \sum_j [C^T dX]_{jj} = \operatorname{tr}(C^T dX) = |X| \operatorname{tr}(X^{-1}dX)$$

Note that (9) follows directly from (8).

5 The special form $\operatorname{tr}(AdX)$

At the beginning of this article we gave two definitions of the derivative of a scalar function of a matrix. The first is

$$Df(X) = \frac{\partial f(X)}{\partial X^T}$$

and the second is

$$Df(X) = \frac{\partial f(X)}{\partial (\operatorname{vec} X)^T}.$$

In the first case the result is a matrix whereas in the second the result is a row vector. Magnus suggests to use the second definition, but we'll opt for the first one.

Let's consider the differential $\operatorname{tr}(AdX)$. We can find the derivative according to the second definition above by using property (12) in the section "Two important operators". We'll also use rule of differentiation (10) with $\star = \operatorname{vec}$. We get

$$\operatorname{tr}(AdX) = \operatorname{vec}(A^T)^T \operatorname{vec} dX = \operatorname{vec}(A^T)^T d \operatorname{vec} X$$

and, therefore,

$$\frac{\partial \operatorname{tr}(AdX)}{\partial (\operatorname{vec} X)^T} = \operatorname{vec}(A^T)^T$$

To get the result in matrix form (first definition) we need to *unvectorize* the result above. We'll proceed step by step:

$$\frac{\partial \operatorname{tr}(AdX)}{\partial(\operatorname{vec} X)^T} = \operatorname{vec}(A^T)^T \implies \frac{\partial \operatorname{tr}(AdX)}{\partial \operatorname{vec} X} = \operatorname{vec}(A^T) \implies \frac{\partial \operatorname{tr}(AdX)}{\partial X} = A^T \implies \frac{\partial \operatorname{tr}(AdX)}{\partial X^T} = A.$$

So, the result is simply A. As we saw in the previous section,

$$d|X| = |X| \operatorname{tr}(X^{-1}dX) = \operatorname{tr}(|X|X^{-1}dX).$$

This means that the derivative is $|X|X^{-1} = C^T$ which agree with the result

$$\frac{\partial |X|}{\partial x_{ij}} = C_{ij}$$

that we derived in the previous section.

6 Examples

In practice, the derivative of an expression involving matrices can be computed by using the rules of differentiation to get a result of the form $\phi(X)dX$, $\operatorname{tr}(\phi(X)dX)$ or $\phi(X)d\operatorname{vec} X$. After having done that, we can read off the derivative which is simply $\phi(X)$.

To find the Hessian we must differentiate a second time, that is we must differentiate $f(x)$ and find $df(x; dx)$, and then differentiate $df(x; dx)$ itself by keeping in mind that dx is a constant increment and thus, for instance, $d(a^T dx) = 0$ and $d(x^T Adx) = dx^T Adx$.

According to the second identification result, we must get a result of the form $dx^T\phi(x)dx$ or of the form $(d\operatorname{vec} X)^T\phi(X)d\operatorname{vec} X$ and the Hessian is $\frac{1}{2}(\phi(x) + \phi(x)^T)$. Note that if $\phi(x)$ is symmetric then the Hessian is just $\phi(x)$.

From time to time we'll need to use the *commutation matrix* K_{mn}, which is the permutation matrix satisfying

$$K_{mn} \operatorname{vec} X = \operatorname{vec}(X^T)$$

where X is $m \times n$. It can be shown that

$$K_{mn}^T = K_{mn}^{-1} = K_{nm}$$

In particular, $K_{nn} = K_n$ is symmetric. Another useful fact is the following. If A is an $m \times n$ matrix, B a $p \times q$ matrix and X a $q \times n$ matrix, then

$$(B \otimes A)K_{qn} = K_{pm}(A \otimes B)$$

Let's prove that:

$$\begin{aligned}
K_{pm}(A \otimes B) \operatorname{vec} X &= K_{pm} \operatorname{vec}(BXA^T) \\
&= \operatorname{vec}((BXA^T)^T) \\
&= \operatorname{vec}(AX^TB^T) \\
&= (B \otimes A) \operatorname{vec}(X^T) \\
&= (B \otimes A)K_{qn} \operatorname{vec} X
\end{aligned}$$

which is true for all $\operatorname{vec} X \in \mathbb{R}^{qn\times 1}$ and hence the proof is complete.

In this section we'll see some examples of computation of the derivative, of the Hessian, and then an example of *maximum likelihood* estimation with a multivariate Gaussian distribution,

Matrices will be written in uppercase and vectors in lowercase.

Let's get started!

$$f(x) = a^T x \qquad \begin{aligned} d(a^T x) &= a^T dx \\ &\implies Df(x) = a^T \end{aligned}$$

$$f(x) = x^T A x \qquad \begin{aligned} d(x^T A x) &= d(x^T) A x + x^T d(Ax) \\ &= (dx)^T A x + x^T A dx \\ &= x^T A^T dx + x^T A dx \\ &= x^T (A^T + A) dx \\ &\implies Df(x) = x^T (A^T + A) \end{aligned}$$

$$f(X) = a^T X b \qquad \begin{aligned} d(a^T X b) &= d\,\mathrm{tr}(a^T X b) \\ &= \mathrm{tr}(a^T d(X) b) \\ &= \mathrm{tr}(b a^T dX) \\ &\implies Df(X) = b a^T \end{aligned}$$

$$f(X) = a^T X X^T a \qquad \begin{aligned} d(a^T X X^T a) &= \mathrm{tr}(a^T d(X X^T) a) \\ &= \mathrm{tr}(a a^T d(X X^T)) \\ &= \mathrm{tr}(a a^T ((dX) X^T + X dX^T)) \\ &= \mathrm{tr}(a a^T (dX) X^T) + \mathrm{tr}((dX) X^T a a^T) \\ &= \mathrm{tr}(X^T a a^T dX) + \mathrm{tr}(X^T a a^T dX) \\ &= 2\,\mathrm{tr}(X^T a a^T dX) \\ &\implies Df(X) = 2 X^T a a^T \end{aligned}$$

$$f(X) = \mathrm{tr}(A X^T B X C) \qquad \begin{aligned} d\,\mathrm{tr}(A X^T B X C) &= \mathrm{tr}(A d(X^T) B X C) + \mathrm{tr}(A X^T B (dX) C) \\ &= \mathrm{tr}(C^T X^T B^T (dX) A^T) + \mathrm{tr}(C A X^T B dX) \\ &= \mathrm{tr}((A^T C^T X^T B^T + C A X^T B) dX) \\ &\implies Df(X) = A^T C^T X^T B^T + C A X^T B \end{aligned}$$

$$f(X) = \mathrm{tr}(A X^{-1} B) \qquad \begin{aligned} d\,\mathrm{tr}(A X^{-1} B) &= \mathrm{tr}(B A d(X^{-1})) \\ &= -\,\mathrm{tr}(B A X^{-1} (dX) X^{-1}) \\ &= -\,\mathrm{tr}(X^{-1} B A X^{-1} dX) \\ &\implies Df(X) = -X^{-1} B A X^{-1} \end{aligned}$$

$$f(X) = |X^T X| \qquad \begin{aligned} d|X^T X| &= |X^T X|\,\mathrm{tr}((X^T X)^{-1} d(X^T X)) \\ &= |X^T X|\,\mathrm{tr}((X^T X)^{-1} (d(X^T) X + X^T dX)) \\ &= |X^T X| \left[\mathrm{tr}((X^T X)^{-1} d(X^T) X) + \mathrm{tr}((X^T X)^{-1} X^T dX)\right] \\ &= |X^T X| \left[\mathrm{tr}(X^T (dX) (X^T X)^{-1}) + \mathrm{tr}((X^T X)^{-1} X^T dX)\right] \\ &= 2|X^T X|\,\mathrm{tr}((X^T X)^{-1} X^T dX) \\ &\implies Df(X) = 2|X^T X| (X^T X)^{-1} X^T \end{aligned}$$

$$f(X) = \operatorname{tr}(X^p) \qquad \begin{aligned} d\operatorname{tr}(X^p) &= \operatorname{tr}((dX)X^{p-1} + X(dX)X^{p-2} + \cdots + X^{p-1}dX) \\ &= \operatorname{tr}(X^{p-1}dX + X^{p-1}dX + \cdots + X^{p-1}dX) \\ &= p\operatorname{tr}(X^{p-1}dX) \\ &\implies Df(X) = pX^{p-1} \end{aligned}$$

$$f(X) = Xa \qquad \begin{aligned} d(Xa) &= (dX)a \\ &= \operatorname{vec}((dX)a) \qquad \text{(the vec of a vector is the vector itself)} \\ &= \operatorname{vec}(I_n(dX)a) \\ &= (a^T \otimes I_n)d\operatorname{vec} X \\ &\implies Df(X) = a^T \otimes I_n \end{aligned}$$

To differentiate a matrix we need to vectorize it:

$$f(x) = xx^T \qquad \begin{aligned} d\operatorname{vec}(xx^T) &= \operatorname{vec}((dx)x^T + xdx^T) \\ &= \operatorname{vec}(I_n(dx)x^T) + \operatorname{vec}(x(dx)^T I_n) \\ &= (x \otimes I_n)d\operatorname{vec} x + (I_n \otimes x)d\operatorname{vec}(x^T) \\ &= (x \otimes I_n)d\operatorname{vec} x + (I_n \otimes x)d\operatorname{vec} x \\ &\implies Df(x) = x \otimes I_n + I_n \otimes x \end{aligned}$$

$$f(X) = X^2 \qquad \begin{aligned} d\operatorname{vec}(X^2) &= \operatorname{vec}((dX)X + XdX) \\ &= \operatorname{vec}(I_n(dX)X + X(dX)I_n) \\ &= (X^T \otimes I_n + I_n \otimes X)d\operatorname{vec} X \\ &\implies Df(X) = X^T \otimes I_n + I_n \otimes X \end{aligned}$$

Now let's compute some Hessians.

$$f(x) = a^T x \qquad \begin{aligned} d(a^T x) &= a^T dx \\ d(a^T dx) &= 0 \\ &\implies Hf(x) = 0 \end{aligned}$$

$$f(X) = \operatorname{tr}(AXB) \qquad \begin{aligned} d\operatorname{tr}(AXB) &= \operatorname{tr}(a(dX)B) \\ &= \operatorname{tr}(BAdX) \\ d\operatorname{tr}(BAdX) &= 0 \\ &\implies Hf(x) = 0 \end{aligned}$$

$$f(x) = x^T Ax \qquad \begin{aligned} d(x^T Ax) &= dx^T Ax + x^T Adx \\ &= x^T A^T dx + x^T Adx \\ &= x^T(A^T + A)dx \\ d(x^T(A^T + A)dx) &= dx^T(A^T + A)dx \\ &\implies Hf(x) = A^T + A \end{aligned}$$

$f(X) = \operatorname{tr}(X^T X)$

$$\begin{aligned} d\operatorname{tr}(X^T X) &= \operatorname{tr}(dX^T X + X^T dX) \\ &= \operatorname{tr}(X^T dX + X^T dX) \\ &= 2\operatorname{tr}(X^T dX) \\ d(2\operatorname{tr}(X^T dX)) &= 2\operatorname{tr}(dX^T dX) \\ &= 2(d\operatorname{vec} X)^T d\operatorname{vec} X \\ &\implies Hf(X) = 2I_{mn} \qquad (X \text{ is } m \times n) \end{aligned}$$

$f(X) = \operatorname{tr}(AX^T BX)$

$$\begin{aligned} d\operatorname{tr}(AX^T BX) &= \operatorname{tr}(AdX^T BX + AX^T BdX) \\ &= \operatorname{tr}(X^T B^T dXA^T + AX^T BdX) \\ &= \operatorname{tr}(A^T X^T B^T dX + AX^T BdX) \\ d(\operatorname{tr}(A^T X^T B^T dX + AX^T BdX)) &= \operatorname{tr}(A^T dX^T B^T dX + AdX^T BdX) \\ &= \operatorname{tr}(dX^T BdXA + AdX^T BdX) \\ &= 2\operatorname{tr}(AdX^T BdX) \\ &= 2\operatorname{tr}(dX^T B(dX)A) \\ &= 2(d\operatorname{vec} X)^T \operatorname{vec}(B(dX)A) \\ &= 2(d\operatorname{vec} X)^T (A^T \otimes B) d\operatorname{vec} X \\ &\implies Hf(X) = \frac{1}{2}(2A^T \otimes B + 2A \otimes B^T) \\ &\implies Hf(X) = A^T \otimes B + A \otimes B^T \end{aligned}$$

Here we'll make use of the commutation matrix K_{mn}. Remember that $K_{nn} = K_n$ is symmetric.

$f(X) = \operatorname{tr}(X^2)$

$$\begin{aligned} d\operatorname{tr}(X^2) &= \operatorname{tr}((dX)X + XdX) \qquad (X \text{ is } n \times n) \\ &= 2\operatorname{tr}(XdX) \\ d(2\operatorname{tr}(XdX)) &= 2\operatorname{tr}(dXdX) \\ &= 2(\operatorname{vec}(dX^T))^T d\operatorname{vec} X \\ &= 2(K_n \operatorname{vec}(dX))^T d\operatorname{vec} X \\ &= 2(d\operatorname{vec} X)^T K_n d\operatorname{vec} X \\ &\implies Hf(X) = 2K_n \end{aligned}$$

$f(X) = \operatorname{tr}(AXBX)$

$$\begin{aligned} d\operatorname{tr}(AXBX) &= \operatorname{tr}(A(dX)BX + AXBdX) \qquad (X \text{ is } m \times n) \\ &= \operatorname{tr}(BXAdX + AXBdX) \\ d\operatorname{tr}(BXAdX + AXBdX) &= \operatorname{tr}(B(dX)AdX + A(dX)BdX) \\ &= 2\operatorname{tr}(A(dX)BdX) \\ &= 2\operatorname{tr}((dX)B(dX)A) \\ &= 2(\operatorname{vec}(dX^T))^T \operatorname{vec}(B(dX)A) \\ &= 2(\operatorname{vec}(dX^T))^T (A^T \otimes B) d\operatorname{vec} X \\ &= 2(K_{mn} \operatorname{vec}(dX))^T (A^T \otimes B) d\operatorname{vec} X \\ &= 2(d\operatorname{vec} X)^T K_{nm}(A^T \otimes B) d\operatorname{vec} X \\ &\implies Hf(X) = K_{nm}(A^T \otimes B) + (A \otimes B^T)K_{mn} \\ &\implies Hf(X) = K_{nm}(A^T \otimes B + B^T \otimes A) \end{aligned}$$

In the last step above we used the fact that $(A \otimes B^T)K_{mn} = K_{nm}(B^T \otimes A)$.

$$
\begin{aligned}
f(X) = a^T X X^T a \qquad\qquad d(a^T X X^T a) &= a^T(dX)X^T a + a^T X(dX)^T a \\
&= a^T(dX)X^T a + a^T(dX)X^T a \\
&= 2a^T(dX)X^T a \\
d(2a^T(dX)X^T a) &= 2a^T(dX)(dX)^T a \\
&= 2\operatorname{tr}(a^T(dX)(dX)^T a) \\
&= 2\operatorname{tr}((dX)^T a a^T dX) \\
&= 2(d\operatorname{vec} X)^T \operatorname{vec}(a a^T dX) \\
&= 2(d\operatorname{vec} X)^T \operatorname{vec}(a a^T (dX) I) \\
&= 2(d\operatorname{vec} X)^T (I \otimes a a^T) d\operatorname{vec} X \\
&\implies Hf(X) = 2(I \otimes a a^T)
\end{aligned}
$$

$$
\begin{aligned}
f(X) = \operatorname{tr}(X^{-1}) \qquad\qquad d\operatorname{tr}(X^{-1}) &= \operatorname{tr}(d(X^{-1})) \\
&= -\operatorname{tr}(X^{-1}(dX)X^{-1}) \\
d(-\operatorname{tr}(X^{-1}(dX)X^{-1})) &= -\operatorname{tr}(d(X^{-1})(dX)X^{-1} + X^{-1}(dX)d(X^{-1})) \\
&= -\operatorname{tr}(-X^{-1}(dX)X^{-1}(dX)X^{-1} - X^{-1}(dX)X^{-1}(dX)X^{-1}) \\
&= 2\operatorname{tr}((dX)X^{-1}(dX)X^{-2}) \\
&= 2(\operatorname{vec}(dX^T))^T \operatorname{vec}(X^{-1}(dX)X^{-2}) \\
&= 2(K_n \operatorname{vec}(dX))^T \operatorname{vec}(X^{-1}(dX)X^{-2}) \\
&= 2(d\operatorname{vec} X)^T K_n (X^{-2T} \otimes X^{-1}) d\operatorname{vec} X \\
&\implies Hf(X) = K_n(X^{-2T} \otimes X^{-1}) + (X^{-2} \otimes X^{-T})K_n \\
&\implies Hf(X) = K_n(X^{-2T} \otimes X^{-1} + X^{-T} \otimes X^{-2})
\end{aligned}
$$

$$
\begin{aligned}
f(X) = |X| \qquad\qquad d|X| &= |X| \operatorname{tr}(X^{-1} dX) \\
d(|X| \operatorname{tr}(X^{-1} dX)) &= d|X| \operatorname{tr}(X^{-1} dX) + |X| \operatorname{tr}(d(X^{-1}) dX) \\
&= |X| (\operatorname{tr}(X^{-1} dX))^2 - |X| \operatorname{tr}(X^{-1}(dX)X^{-1} dX) \\
&= |X| \operatorname{tr}((dX)^T X^{-T}) \operatorname{tr}(X^{-1} dX) - |X| \operatorname{tr}((dX)X^{-1}(dX)X^{-1}) \\
&= |X| (d\operatorname{vec} X)^T \operatorname{vec}(X^{-T})(\operatorname{vec}(X^{-T}))^T d\operatorname{vec} X \\
&\quad - |X| (\operatorname{vec}(dX^T))^T (X^{-T} \otimes X^{-1}) d\operatorname{vec} X \\
&= |X| (d\operatorname{vec} X)^T \left(\operatorname{vec}(X^{-T})(\operatorname{vec}(X^{-T}))^T - K_n(X^{-T} \otimes X^{-1})\right) d\operatorname{vec} X \\
&\implies Hf(X) = |X| \left(\operatorname{vec}(X^{-T})(\operatorname{vec}(X^{-T}))^T - K_n(X^{-T} \otimes X^{-1})\right)
\end{aligned}
$$

Note that the Hessian above (like all the others) is symmetric; in fact,

$$
(K_n(X^{-T} \otimes X^{-1}))^T = (X^{-1} \otimes X^{-T})K_n = K_n(X^{-T} \otimes X^{-1})
$$

We conclude this section with an example about MLE.

Given a set of vectors $x_1, \ldots, x_N$ drawn independently from a multivariate Gaussian distribution, we want to estimate the parameters of the distribution by maximum likelihood. Note that the covariance matrix Σ, and thus Σ^{-1}, is symmetric. The log likelihood function is

$$
\ln p(x_1, \ldots, x_N | \mu, \Sigma) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^T \Sigma^{-1} (x_n - \mu).
$$

We first find the derivative with respect to μ:

$$
\begin{aligned}
d\left(-\frac{1}{2}\sum_{i=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)\right) &= -\frac{1}{2}\sum_{i=1}^{N}\left(d(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)+(x_n-\mu)^T\Sigma^{-1}d(x_n-\mu)\right)\\
&= -\frac{1}{2}\sum_{i=1}^{N}\left(-d\mu^T\Sigma^{-1}(x_n-\mu)-(x_n-\mu)^T\Sigma^{-1}d\mu\right)\\
&= \frac{1}{2}\sum_{i=1}^{N}\left((x_n-\mu)^T\Sigma^{-1}d\mu+(x_n-\mu)^T\Sigma^{-1}d\mu\right)\\
&= \left(\sum_{i=1}^{N}(x_n-\mu)^T\Sigma^{-1}\right)d\mu\\
\implies \frac{\partial \ln p}{\partial \mu^T} &= \sum_{i=1}^{N}(x_n-\mu)^T\Sigma^{-1}
\end{aligned}
$$

Therefore,

$$
\sum_{i=1}^{N}(x_n-\mu)^T\Sigma^{-1}=0 \iff \sum_{n=1}^{N}x_n^T=N\mu^T \iff \hat{\mu}=\frac{1}{N}\sum_{i=1}^{N}x_n
$$

Finally, we find the derivative with respect to Σ:

$$
\begin{aligned}
d\left(-\frac{N}{2}\ln|\Sigma|-\frac{1}{2}\sum_{i=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)\right) &= -\frac{N}{2}\operatorname{tr}(\Sigma^{-1}d\Sigma)-\frac{1}{2}\sum_{n=1}^{N}\operatorname{tr}\left((x_n-\mu)^Td(\Sigma^{-1})(x_n-\mu)\right)\\
&= -\frac{N}{2}\operatorname{tr}(\Sigma^{-1}d\Sigma)+\frac{1}{2}\sum_{n=1}^{N}\operatorname{tr}\left((x_n-\mu)^T\Sigma^{-1}(d\Sigma)\Sigma^{-1}(x_n-\mu)\right)\\
&= -\frac{N}{2}\operatorname{tr}(\Sigma^{-1}d\Sigma)+\frac{1}{2}\sum_{n=1}^{N}\operatorname{tr}\left(\Sigma^{-1}(x_n-\mu)(x_n-\mu)^T\Sigma^{-1}d\Sigma\right)\\
\implies \frac{\partial \ln p}{\partial \Sigma^T} &= -\frac{N}{2}\Sigma^{-1}+\frac{1}{2}\sum_{i=1}^{N}\Sigma^{-1}(x_n-\mu)(x_n-\mu)^T\Sigma^{-1}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
-\frac{N}{2}\Sigma^{-1}+\frac{1}{2}\sum_{i=1}^{N}\Sigma^{-1}(x_n-\mu)(x_n-\mu)^T\Sigma^{-1}=0 &\iff \frac{N}{2}\Sigma^{-1}=\frac{1}{2}\sum_{i=1}^{N}\Sigma^{-1}(x_n-\mu)(x_n-\mu)^T\Sigma^{-1}\\
&\iff N=\Sigma^{-1}\sum_{i=1}^{N}(x_n-\mu)(x_n-\mu)^T\\
&\iff \hat{\Sigma}=\frac{1}{N}\sum_{i=1}^{N}(x_n-\hat{\mu})(x_n-\hat{\mu})^T
\end{aligned}
$$

References

[1] Abadir, K. M., & Magnus, J. R. (2005). *Matrix algebra*. Cambridge, UK: Cambridge University Press.

[2] Magnus, J. R. & Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and economics*. New York, NY: John Wiley & Sons

[3] Steven W. Nydick (2012). *A Different(ial) Way: Matrix Derivatives Again.*

[4] Thomas Minka (2000). *Old and New Matrix Algebra Useful for Statistics.*