## Rules for the differentials[1]

Let $\alpha$, $a$, $A$ be constants and $\phi$, $\psi$, $u$, $v$, $x$, $f$, $U$, $V$, $F$ be functions.

$$\mathsf{d}\alpha = 0 \qquad \mathsf{d}a = 0_n \qquad \mathsf{d}A = 0_{mn} \tag{1}$$

$$\mathsf{d}(\alpha\phi) = \alpha\,\mathsf{d}\phi \qquad \mathsf{d}(\alpha u) = \alpha\,\mathsf{d}u \qquad \mathsf{d}(\alpha U) = \alpha\,\mathsf{d}U \tag{2}$$

$$\mathsf{d}(\phi+\psi) = \mathsf{d}\phi + \mathsf{d}\psi \qquad \mathsf{d}(u+v) = \mathsf{d}u + \mathsf{d}v \qquad \mathsf{d}(U+V) = \mathsf{d}U + \mathsf{d}V \tag{3}$$

$$\mathsf{d}(\phi\,\psi) = (\mathsf{d}\phi)\psi + \phi\,\mathsf{d}\psi \qquad \mathsf{d}(u^\mathsf{T} v) = (\mathsf{d}u)^\mathsf{T} v + u^\mathsf{T}\mathsf{d}v \qquad \mathsf{d}(UV) = (\mathsf{d}U)V + U\mathsf{d}V \tag{4}$$

$$\mathsf{d}(\phi/\psi) = ((\mathsf{d}\phi)\psi - \phi\,\mathsf{d}\psi)\,/\,\psi^2 \qquad \mathsf{d}(u^\mathsf{T}) = (\mathsf{d}u)^\mathsf{T} \qquad \mathsf{d}(U^\mathsf{T}) = (\mathsf{d}U)^\mathsf{T} \tag{5}$$

$$\mathsf{d}\operatorname{vec} U = \operatorname{vec}\mathsf{d}U \qquad \mathsf{d}\operatorname{tr} U = \operatorname{tr}\mathsf{d}U \tag{6}$$

$$\mathsf{d}(U\otimes V) = (\mathsf{d}U)\otimes V + U\otimes\mathsf{d}V \qquad \mathsf{d}(U\odot V) = (\mathsf{d}U)\odot V + U\odot\mathsf{d}V \tag{7}$$

$$\mathsf{d}(\phi^\alpha) = \alpha\phi^{\alpha-1}\mathsf{d}\phi \qquad \mathsf{d}(U^{-1}) = -U^{-1}(\mathsf{d}U)\,U^{-1} \tag{8}$$

$$\mathsf{d}\det U = \det(U)\operatorname{tr}(U^{-1}\mathsf{d}U) \qquad \mathsf{d}\log(\det U) = \operatorname{tr}(U^{-1}\mathsf{d}U) \tag{9}$$

$$\mathsf{d}\exp\phi = \exp(\phi)\,\mathsf{d}\phi \qquad \operatorname{tr}(\mathsf{d}\exp U) = \operatorname{tr}(\exp(U)\,\mathsf{d}U) \tag{10}$$

## Identification table

Note that the differential has always the same shape as the function.

| | function | differential | derivative | shape of derivative | |
|---|---|---|---|---|---|
| $\phi(\xi)$ | $\mathbb{R}\to\mathbb{R}$ | $\mathsf{d}\phi = \alpha(\xi)\mathsf{d}\xi$ | $\mathsf{D}\phi(\xi) = \alpha(\xi)$ | $1\times 1$ | (11) |
| $\phi(x)$ | $\mathbb{R}^n\to\mathbb{R}$ | $\mathsf{d}\phi = a(x)^\mathsf{T}\mathsf{d}x$ | $\mathsf{D}\phi(x) = a(x)^\mathsf{T}$ | $1\times n$ | (12) |
| $\phi(X)$ | $\mathbb{R}^{n\times q}\to\mathbb{R}$ | $\mathsf{d}\phi = \operatorname{tr} A(X)^\mathsf{T}\mathsf{d}X$ | $\mathsf{D}\phi(X) = (\operatorname{vec} A(X))^\mathsf{T}$ | $1\times nq$ | (13) |
| $f(\xi)$ | $\mathbb{R}\to\mathbb{R}^m$ | $\mathsf{d}f = a(\xi)\,\mathsf{d}\xi$ | $\mathsf{D}f(\xi) = a(\xi)$ | $m\times 1$ | (14) |
| $f(x)$ | $\mathbb{R}^n\to\mathbb{R}^m$ | $\mathsf{d}f = A(x)\,\mathsf{d}x$ | $\mathsf{D}f(x) = A(x)$ | $m\times n$ | (15) |
| $f(X)$ | $\mathbb{R}^{n\times q}\to\mathbb{R}^m$ | $\mathsf{d}f = A(X)\ \mathsf{d}\operatorname{vec} X$ | $\mathsf{D}f(X) = A(X)$ | $m\times nq$ | (16) |
| $F(\xi)$ | $\mathbb{R}\to\mathbb{R}^{m\times p}$ | $\mathsf{d}F = A(\xi)\,\mathsf{d}\xi$ | $\mathsf{D}F(\xi) = \operatorname{vec} A(\xi)$ | $mp\times 1$ | (17) |
| $F(x)$ | $\mathbb{R}^n\to\mathbb{R}^{m\times p}$ | $\mathsf{d}\operatorname{vec} F = A(x)\,\mathsf{d}x$ | $\mathsf{D}F(x) = A(x)$ | $mp\times n$ | (18) |
| $F(X)$ | $\mathbb{R}^{n\times q}\to\mathbb{R}^{m\times p}$ | $\mathsf{d}\operatorname{vec} F = A(X)\,\mathsf{d}\operatorname{vec} X$ | $\mathsf{D}F(X) = A(X)$ | $mp\times nq$ | (19) |

## Notation

| | | |
|---|---|---|
| $\alpha, \beta, \phi, \xi, \ldots$ | lower case Greek letters are scalars | (20) |
| $a, b, c, f, x, \ldots$ | lower case Latin letters are (column) vectors | (21) |
| $A, B, C, F, X, \ldots$ | capital Latin letters are matrices | (22) |
| $\mathsf{d}\phi$ | differential of $\phi$ | (23) |
| $\mathsf{D}\phi(x)$ | derivative of $\phi$ at $x$ | (24) |
| $0_n, 0_{mn}$ | $n$ vector of zeros, $m\times n$ matrix of zeros | (25) |
| $1_n, 1_{mn}$ | $n$ vector of ones, $m\times n$ matrix of ones | (26) |
| $I_n, I_{mn}$ | $n\times n$ identity matrix, $m\times n$ identity matrix | (27) |
| $A^\mathsf{T}, \operatorname{tr} A, \det A$ | transpose of $A$, trace of $A$, determinant of $A$ | (28) |
| $\operatorname{vec} A$ | vector containing the stacked columns of $A$ | (29) |
| $\operatorname{diag} A$ | vector containing the diagonal of $A$ | (30) |
| $\operatorname{Diag} a$ | diagonal matrix with $a$ along the diagonal | (31) |
| $\exp\alpha$ | scalar exponential function | (32) |
| $\exp A$ | matrix exponential function, $\det(\exp(A)) = \exp(\operatorname{tr}(A))$ | (33) |
| $A\odot B$ | Hadamard product (component-wise product) | (34) |
| $A\oslash B$ | component-wise division | (35) |
| $A\otimes B$ | Kronecker product, $\operatorname{vec} ab^\mathsf{T} = b\otimes a$ | (36) |

[1] Most of the material of this document is from [2], available at at the author's webpage `http://www.janmagnus.nl/misc/mdc2007-3rdedition.pdf`. Another good resource for matrix wisdom is [1].

## More rules

$$\mathrm{tr}(AB) = \mathrm{tr}(BA) \qquad \mathrm{tr}\,A = 1_m^\top (A \odot I_{mn})\, 1_n = 1^T \,\mathrm{diag}(A) = \mathrm{tr}\,A^\top \tag{37}$$

$$\mathrm{diag}(UV^\top) = (U \odot V)1_n \qquad \mathrm{tr}(U^\top (V \odot C)) = \mathrm{tr}((U^\top \odot V^\top)C) \tag{38}$$

$$A \otimes 1_l = (I_m \otimes 1_l)A \qquad 1_l \otimes A = (1_l \otimes I_m)A \tag{39}$$

$$\mathrm{Diag}\, a = a1_n^T \odot I_n \qquad \mathrm{diag}\, A = \mathrm{vec}(A \odot I_n) = (A \odot I_n)1_n \tag{40}$$

$$\mathrm{Diag}(\mathrm{diag}\, A) = I_n \odot A \qquad \|U\|^2_{\mathrm{Fro}} = \mathrm{tr}(U^\top U) = \mathrm{vec}(U)^\top \mathrm{vec}(U) \tag{41}$$

$$\mathrm{vec}(a) = \mathrm{vec}(a^\top) = a \qquad \mathrm{vec}(ABC) = (C^\top \otimes A)\,\mathrm{vec}(B) \tag{42}$$

$$\mathrm{tr}(u^\top v) = \mathrm{tr}(v^\top u) = v^\top u \qquad ABc = (c^\top \otimes A)\,\mathrm{vec}\, B = (A \otimes c^\top)\,\mathrm{vec}\, B^\top = \mathrm{vec}(c^\top B^\top A^\top) \tag{43}$$

## Interlude: finite differencing

You should always check your derivatives with finite differencing which is an alternative way to calculate a derivative! Here is some matlab code:

```
function df = finitediff(fun, x, d, varargin)
%FINITEDIFF estimates a gradient by finite-differencing method.
% (c) Stefan Harmeling, 2012-07-10.
sx = size(x);
nx = numel(x);
df = zeros(sx);
dx = zeros(sx);
for i = 1:nx
  dx(i) = d;
  df(i) = (fun(x+dx, varargin{:})-fun(x-dx, varargin{:}))/(2*d);
  dx(i) = 0;
end
```

## Some pros and cons of matrix differential calculus

+ clean notation
+ vectorized function leads to vectorized derivative (good for coding)
+ powerful: [2] shows how to take the derivative of eigenvalues and eigenvectors
− complicated formulas
− requires tricks and practice

## General recipe and examples

(i) write the letter d in front of the expression
(ii) identity the constants and variables
(ii) transform the expression
(iv) read off the derivative using the identification table

1. Find the derivative of $\phi(\xi) = \xi^2$.

$$\mathsf{d}\phi = \mathsf{d}\xi^2 = 2\xi \mathsf{d}\xi \;\text{ thus }\; \mathsf{D}\phi(\xi) = 2\xi \tag{44}$$

2. Find the derivative of $\phi(x) = x^\top Ax$.

$$\mathsf{d}\phi = (\mathsf{d}x)^\top Ax + x^\top A\mathsf{d}x = x^\top A^\top \mathsf{d}x + x^\top A\,\mathsf{d}x = x^\top (A + A^\top)\,\mathsf{d}x \;\text{ thus }\; \mathsf{D}\phi(x) = x^\top (A + A^\top) \tag{45}$$

3. Find the derivative of $\phi(X) = \mathrm{tr}(X^\top X)$.

$$\mathsf{d}\phi = \mathsf{d}\,\mathrm{tr}(X^\top X) = \mathrm{tr}((\mathsf{d}X)^\top X + X^\top \mathsf{d}X) = \mathrm{tr}(2X^\top \mathsf{d}X) \;\text{ thus }\; \mathsf{D}\phi(X) = 2(\mathrm{vec}\, X)^\top \tag{46}$$

4. Find the derivative of $\phi(x) = (y - Ax)^2$.

$$\mathsf{d}\phi = \mathsf{d}((y - Ax)^\mathsf{T}(y - Ax)) = -2(y - Ax)^T A\,\mathsf{d}x \text{ thus } \mathsf{D}\phi(x) = -2A^\mathsf{T}(y - Ax) \tag{47}$$

5. Find the derivative of $f(X) = (X^\mathsf{T} X)^{-1} X^\mathsf{T} y$. We write $A = (X^\mathsf{T} X)^{-1}$.

$$\mathsf{d}f = \mathsf{d}(X^\mathsf{T} X)^{-1} X^\mathsf{T} y \tag{48}$$
$$= -A((\mathsf{d}X)^T X + X^\mathsf{T}(\mathsf{d}X))AX^\mathsf{T} y \tag{49}$$
$$= -A(\mathsf{d}X)^T XAX^\mathsf{T} y - AX^\mathsf{T}(\mathsf{d}X)AX^\mathsf{T} y - A(\mathsf{d}X)^\mathsf{T} y \tag{50}$$
$$= -(A \otimes (XAX^\mathsf{T} y)^\mathsf{T})\,\mathsf{d}\mathrm{vec}\,X - ((XAX^\mathsf{T} y)^\mathsf{T} \otimes A)\,\mathsf{d}\mathrm{vec}\,X - (A \otimes y^\mathsf{T})\,\mathsf{d}\mathrm{vec}\,X \tag{51}$$
$$= \underbrace{-(A \otimes (XAX^\mathsf{T} y)^\mathsf{T} + (XAX^\mathsf{T} y)^\mathsf{T} \otimes A + A \otimes y^\mathsf{T})}_{\mathsf{D}f(X)}\,\mathsf{d}\mathrm{vec}\,X \tag{52}$$

6. Often it is easier to find the differential of a scalar function $\phi(X) = c^\mathsf{T}(X^\mathsf{T} X)^{-1} X^\mathsf{T} y$.

$$\mathsf{d}\phi = -c^\mathsf{T} A(\mathsf{d}X)^T XAX^\mathsf{T} y - c^\mathsf{T} AX^\mathsf{T}(\mathsf{d}X)AX^\mathsf{T} y - c^\mathsf{T} A(\mathsf{d}X)^\mathsf{T} y \tag{53}$$
$$= -\operatorname{tr}(c^\mathsf{T} A(\mathsf{d}X)^T XAX^\mathsf{T} y) - \operatorname{tr}(c^\mathsf{T} AX^\mathsf{T}(\mathsf{d}X)AX^\mathsf{T} y) - \operatorname{tr}(c^\mathsf{T} A(\mathsf{d}X)^\mathsf{T} y) \tag{54}$$
$$= -\operatorname{tr}(yXA^\mathsf{T} X^\mathsf{T}(\mathsf{d}X)A^\mathsf{T} c) - \operatorname{tr}(c^\mathsf{T} AX^\mathsf{T}(\mathsf{d}X)AX^\mathsf{T} y) - \operatorname{tr}(y^\mathsf{T}(\mathsf{d}X)A^\mathsf{T} c) \tag{55}$$
$$= -\operatorname{tr}(A^\mathsf{T} c\,yXA^\mathsf{T} X^\mathsf{T}\mathsf{d}X) - \operatorname{tr}(AX^\mathsf{T} yc^\mathsf{T} AX^\mathsf{T}\mathsf{d}X) - \operatorname{tr}(A^\mathsf{T} cy^\mathsf{T}\mathsf{d}X) \tag{56}$$
$$= -\operatorname{tr}((A^\mathsf{T} c\,yXA^\mathsf{T} X^\mathsf{T} + AX^\mathsf{T} yc^\mathsf{T} AX^\mathsf{T} + A^\mathsf{T} cy^\mathsf{T})\mathsf{d}X) \tag{57}$$

7. Sometimes it is good to rewrite with indices:

$$\mathsf{d}\operatorname{tr}(A\,\mathrm{Diag}\,v) = \mathsf{d}\operatorname{tr}(A(I_n \odot v1_n^\mathsf{T})) = \mathsf{d}\operatorname{tr}((A \otimes I)v1_n^\mathsf{T} = \mathsf{d}1_n^\mathsf{T}\,\mathrm{Diag}(A)v = \mathsf{d}\,\mathrm{diag}(A)^\mathsf{T} v = \mathrm{diag}(A)^\mathsf{T}\mathsf{d}v \tag{58}$$
$$\mathsf{d}\operatorname{tr}(A\,\mathrm{Diag}\,v) = \mathsf{d}\sum_i A_{ii}v_i = \mathsf{d}\operatorname{tr}\mathrm{diag}(A)^\mathsf{T} v = \operatorname{tr}\mathrm{diag}(A)^\mathsf{T}\mathsf{d}v \tag{59}$$

8. Find the derivative of Rayleigh coefficient $\phi(x) = x^\mathsf{T} Ax/(x^\mathsf{T} x)$ for symmetric $A$.

$$\mathsf{d}\phi = \frac{2x^\mathsf{T} A(\mathsf{d}x)(x^\mathsf{T} x) - 2x^\mathsf{T} Axx^\mathsf{T}\mathsf{d}x}{(x^\mathsf{T} x)^2} \tag{60}$$
$$= \frac{2(x^\mathsf{T} x)x^\mathsf{T} A(\mathsf{d}x) - 2x^\mathsf{T} Axx^\mathsf{T}\mathsf{d}x}{(x^\mathsf{T} x)^2} \tag{61}$$
$$= \frac{2(x^\mathsf{T} x)x^\mathsf{T} A - 2x^\mathsf{T} Axx^\mathsf{T}}{(x^\mathsf{T} x)^2}\mathsf{d}x \tag{62}$$
$$= \frac{2}{(x^\mathsf{T} x)^2}\, x^\mathsf{T}(xx^\mathsf{T} A - Axx^\mathsf{T})\,\mathsf{d}x \tag{63}$$
$$\tag{64}$$

Thus the derivative is:

$$\mathsf{D}\phi(x) = \frac{2}{(x^\mathsf{T} x)^2}\,(Axx^\mathsf{T} - xx^\mathsf{T} A)x \tag{65}$$
$$= 2\frac{Ax}{x^T x} - 2\frac{x^\mathsf{T} Ax}{(x^T x)^2}x \tag{66}$$

**More difficult examples**

9. Consider a steepest descent algorithm for minimizing the previous function:

$$x^{(k+1)} = x^{(k)} - \xi A^\mathsf{T}(y - Ax^{(k)}) \tag{67}$$

(a) Find the derivative of $x^{(1)}(x^{(0)})$ and of $x^{(k+1)}(x^{(k)})$.

$$\mathsf{d}x^{(1)} = \mathsf{d}x^{(0)} + \xi A^\mathsf{T} A\,\mathsf{d}x^{(0)} = (I_n + \xi A^\mathsf{T} A)\,\mathsf{d}x^{(0)} \tag{68}$$
$$\mathsf{d}x^{(k+1)} = (I_n + \xi A^\mathsf{T} A)\,\mathsf{d}x^{(k)} \tag{69}$$

(b) Find the derivative of $x^{(2)}(x^{(0)})$.

$$\mathsf{d}x^{(2)} = (I_n + \xi A^\mathsf{T} A)\,\mathsf{d}x^{(1)} = (I_n + \xi A^\mathsf{T} A)\,(I_n + \xi A^\mathsf{T} A)\,\mathsf{d}x^{(0)} \tag{70}$$

(c) Find the derivative of $x^{(k)}(x^{(0)})$.

$$\mathsf{d}x^{(k)} = (I_n + \xi A^\mathsf{T} A)^k\,\mathsf{d}x^{(0)} \tag{71}$$

(d) Find the derivative of $x^{(1)}(\xi)$.

$$\mathsf{d}x^{(1)} = -A^\mathsf{T}(y - Ax^{(0)})\,\mathsf{d}\xi \tag{72}$$

(e) Find the derivative of $x^{(2)}(\xi)$.

$$\begin{aligned}
\mathsf{d}x^{(2)} &= \mathsf{d}(-\xi A^\mathsf{T}(y - Ax^{(1)}) + x^{(1)}) && (73)\\
&= -A^\mathsf{T}(y - Ax^{(1)})\,\mathsf{d}\xi + \xi A^\mathsf{T} A\,\mathsf{d}x^{(1)} + \mathsf{d}x^{(1)} && (74)\\
&= -A^\mathsf{T}(y - Ax^{(1)})\,\mathsf{d}\xi + (\xi A^\mathsf{T} A + I_n)\,\mathsf{d}x^{(1)} && (75)\\
&= (-A^\mathsf{T}(y - Ax^{(1)}) - (\xi A^\mathsf{T} A + I_n)A^\mathsf{T}(y - Ax^{(0)}))\,\mathsf{d}\xi && (76)\\
& && (77)
\end{aligned}$$

(f) Find the derivative of $x^{(3)}(\xi)$.

$$\begin{aligned}
\mathsf{d}x^{(3)} &= \mathsf{d}(-\xi A^\mathsf{T}(y - Ax^{(2)}) + x^{(2)}) && (78)\\
&= -A^\mathsf{T}(y - Ax^{(2)})\,\mathsf{d}\xi + (\xi A^\mathsf{T} A + I_n)\,\mathsf{d}x^{(2)} && (79)\\
&= -A^\mathsf{T}(y - Ax^{(2)})\,\mathsf{d}\xi - (\xi A^\mathsf{T} A + I_n)A^\mathsf{T}(y - Ax^{(1)}) - (\xi A^\mathsf{T} A + I_n)^2 A^\mathsf{T}(y - Ax^{(0)})\,\mathsf{d}\xi && (80)\\
&= -\left(\sum_{i=0}^{2}(\xi A^\mathsf{T} A + I_n)^i A^\mathsf{T}(y - Ax^{(2-i)})\right)\mathsf{d}\xi && (81)
\end{aligned}$$

(g) Find the derivative of $x^{(k+1)}(\xi)$.

$$\mathsf{d}x^{(k+1)} = -\left(\sum_{i=0}^{k}(\xi A^\mathsf{T} A + I_n)^i A^\mathsf{T}(y - Ax^{(k-i)})\right)\mathsf{d}\xi \tag{82}$$

(h) Find the derivative of $x^{(1)}(A)$.

$$\begin{aligned}
\mathsf{d}x^{(1)} &= \mathsf{d}(x^{(0)} - \xi A^\mathsf{T}(y - Ax^{(0)})) && (83)\\
&= -\xi(\mathsf{d}A)^\mathsf{T}(y - Ax^{(0)}) + \xi A^\mathsf{T}(\mathsf{d}A)x^{(0)} && (84)\\
&= -\xi(I_n \otimes (y - Ax^{(0)})^\mathsf{T})\,\mathsf{d}\mathrm{vec}\,A + \xi((x^{(0)})^\mathsf{T} \otimes A^\mathsf{T})\,\mathsf{d}\mathrm{vec}\,A && (85)\\
&= -\xi(I_n \otimes (y - Ax^{(0)})^\mathsf{T} + (x^{(0)})^\mathsf{T} \otimes A^\mathsf{T})\,\mathsf{d}\mathrm{vec}\,A && (86)
\end{aligned}$$

10. Consider the norm $\phi = r^2 = r^\mathsf{T} r$ of the residual $r = y - Ax$. Plug into $\phi$ the minimizer $x = (A^T A)^{-1} A^\mathsf{T} y$, we obtain

$$\phi = (y - A(A^T A)^{-1} A^\mathsf{T} y)^2 = ((I_n - A(A^T A)^{-1} A^\mathsf{T})y)^2 \tag{87}$$

Additionally we have an expression for $A$:

$$A = C^\mathsf{T}\,\mathrm{Diag}(Za) B \tag{88}$$

So we can view $\phi$ as a function of $a$. Find the derivative of $\phi(a)$. Left as an exercise ;)

## References

[1] H. Lütkepohl. *Handbook of matrices.* Wiley, 1996.

[2] J.R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley, Chichester, 1999.