

# Appendix D

## Matrix calculus

*From too much study, and from extreme passion, cometh madness.*

– Isaac Newton [150, §5]

### D.1 Directional derivative, Taylor series

#### D.1.1 Gradients

*Gradient* of a differentiable real function  $f(x) : \mathbb{R}^K \rightarrow \mathbb{R}$  with respect to its vector argument is defined in terms of partial derivatives

$$\nabla f(x) \triangleq \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_K} \end{bmatrix} \in \mathbb{R}^K \quad (1719)$$

while the second-order gradient of the twice differentiable real function with respect to its vector argument is traditionally called the *Hessian*;

$$\nabla^2 f(x) \triangleq \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_K} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_K \partial x_1} & \frac{\partial^2 f(x)}{\partial x_K \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_K^2} \end{bmatrix} \in \mathbb{S}^K \quad (1720)$$

The gradient of vector-valued function  $v(x) : \mathbb{R} \rightarrow \mathbb{R}^N$  on real domain is a row-vector

$$\nabla v(x) \triangleq \left[ \frac{\partial v_1(x)}{\partial x} \quad \frac{\partial v_2(x)}{\partial x} \quad \dots \quad \frac{\partial v_N(x)}{\partial x} \right] \in \mathbb{R}^N \quad (1721)$$

while the second-order gradient is

$$\nabla^2 v(x) \triangleq \left[ \frac{\partial^2 v_1(x)}{\partial x^2} \quad \frac{\partial^2 v_2(x)}{\partial x^2} \quad \dots \quad \frac{\partial^2 v_N(x)}{\partial x^2} \right] \in \mathbb{R}^N \quad (1722)$$

Gradient of vector-valued function  $h(x) : \mathbb{R}^K \rightarrow \mathbb{R}^N$  on vector domain is

$$\begin{aligned} \nabla h(x) &\triangleq \begin{bmatrix} \frac{\partial h_1(x)}{\partial x_1} & \frac{\partial h_2(x)}{\partial x_1} & \dots & \frac{\partial h_N(x)}{\partial x_1} \\ \frac{\partial h_1(x)}{\partial x_2} & \frac{\partial h_2(x)}{\partial x_2} & \dots & \frac{\partial h_N(x)}{\partial x_2} \\ \vdots & \vdots & & \vdots \\ \frac{\partial h_1(x)}{\partial x_K} & \frac{\partial h_2(x)}{\partial x_K} & \dots & \frac{\partial h_N(x)}{\partial x_K} \end{bmatrix} \\ &= [\nabla h_1(x) \quad \nabla h_2(x) \quad \dots \quad \nabla h_N(x)] \in \mathbb{R}^{K \times N} \end{aligned} \quad (1723)$$

while the second-order gradient has a three-dimensional representation dubbed *cubix*; [D.1](#)

$$\begin{aligned} \nabla^2 h(x) &\triangleq \begin{bmatrix} \nabla \frac{\partial h_1(x)}{\partial x_1} & \nabla \frac{\partial h_2(x)}{\partial x_1} & \dots & \nabla \frac{\partial h_N(x)}{\partial x_1} \\ \nabla \frac{\partial h_1(x)}{\partial x_2} & \nabla \frac{\partial h_2(x)}{\partial x_2} & \dots & \nabla \frac{\partial h_N(x)}{\partial x_2} \\ \vdots & \vdots & & \vdots \\ \nabla \frac{\partial h_1(x)}{\partial x_K} & \nabla \frac{\partial h_2(x)}{\partial x_K} & \dots & \nabla \frac{\partial h_N(x)}{\partial x_K} \end{bmatrix} \\ &= [\nabla^2 h_1(x) \quad \nabla^2 h_2(x) \quad \dots \quad \nabla^2 h_N(x)] \in \mathbb{R}^{K \times N \times K} \end{aligned} \quad (1724)$$

where the gradient of each real entry is with respect to vector  $x$  as in (1719).

---

<sup>D.1</sup>The word *matrix* comes from the Latin for *womb*; related to the prefix *matri-* derived from *mater* meaning *mother*.

The gradient of real function  $g(X) : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}$  on matrix domain is

$$\begin{aligned} \nabla g(X) &\triangleq \begin{bmatrix} \frac{\partial g(X)}{\partial X_{11}} & \frac{\partial g(X)}{\partial X_{12}} & \dots & \frac{\partial g(X)}{\partial X_{1L}} \\ \frac{\partial g(X)}{\partial X_{21}} & \frac{\partial g(X)}{\partial X_{22}} & \dots & \frac{\partial g(X)}{\partial X_{2L}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial g(X)}{\partial X_{K1}} & \frac{\partial g(X)}{\partial X_{K2}} & \dots & \frac{\partial g(X)}{\partial X_{KL}} \end{bmatrix} \in \mathbb{R}^{K \times L} \\ &= \begin{bmatrix} \nabla_{X(:,1)} g(X) \\ \nabla_{X(:,2)} g(X) \\ \vdots \\ \nabla_{X(:,L)} g(X) \end{bmatrix} \in \mathbb{R}^{K \times 1 \times L} \end{aligned} \tag{1725}$$

where the gradient  $\nabla_{X(:,i)}$  is with respect to the  $i^{\text{th}}$  column of  $X$ . The strange appearance of (1725) in  $\mathbb{R}^{K \times 1 \times L}$  is meant to suggest a third dimension perpendicular to the page (not a diagonal matrix). The second-order gradient has representation

$$\begin{aligned} \nabla^2 g(X) &\triangleq \begin{bmatrix} \nabla \frac{\partial g(X)}{\partial X_{11}} & \nabla \frac{\partial g(X)}{\partial X_{12}} & \dots & \nabla \frac{\partial g(X)}{\partial X_{1L}} \\ \nabla \frac{\partial g(X)}{\partial X_{21}} & \nabla \frac{\partial g(X)}{\partial X_{22}} & \dots & \nabla \frac{\partial g(X)}{\partial X_{2L}} \\ \vdots & \vdots & & \vdots \\ \nabla \frac{\partial g(X)}{\partial X_{K1}} & \nabla \frac{\partial g(X)}{\partial X_{K2}} & \dots & \nabla \frac{\partial g(X)}{\partial X_{KL}} \end{bmatrix} \in \mathbb{R}^{K \times L \times K \times L} \\ &= \begin{bmatrix} \nabla \nabla_{X(:,1)} g(X) \\ \nabla \nabla_{X(:,2)} g(X) \\ \vdots \\ \nabla \nabla_{X(:,L)} g(X) \end{bmatrix} \in \mathbb{R}^{K \times 1 \times L \times K \times L} \end{aligned} \tag{1726}$$

where the gradient  $\nabla$  is with respect to matrix  $X$ .

Gradient of vector-valued function  $g(X) : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}^N$  on matrix domain is a cubix

$$\begin{aligned} \nabla g(X) &\triangleq \begin{bmatrix} \nabla_{X(:,1)} g_1(X) & \nabla_{X(:,1)} g_2(X) & \cdots & \nabla_{X(:,1)} g_N(X) \\ \nabla_{X(:,2)} g_1(X) & \nabla_{X(:,2)} g_2(X) & \cdots & \nabla_{X(:,2)} g_N(X) \\ \vdots & \vdots & \ddots & \vdots \\ \nabla_{X(:,L)} g_1(X) & \nabla_{X(:,L)} g_2(X) & \cdots & \nabla_{X(:,L)} g_N(X) \end{bmatrix} \\ &= [\nabla g_1(X) \quad \nabla g_2(X) \quad \cdots \quad \nabla g_N(X)] \in \mathbb{R}^{K \times N \times L} \end{aligned} \quad (1727)$$

while the second-order gradient has a five-dimensional representation;

$$\begin{aligned} \nabla^2 g(X) &\triangleq \begin{bmatrix} \nabla \nabla_{X(:,1)} g_1(X) & \nabla \nabla_{X(:,1)} g_2(X) & \cdots & \nabla \nabla_{X(:,1)} g_N(X) \\ \nabla \nabla_{X(:,2)} g_1(X) & \nabla \nabla_{X(:,2)} g_2(X) & \cdots & \nabla \nabla_{X(:,2)} g_N(X) \\ \vdots & \vdots & \ddots & \vdots \\ \nabla \nabla_{X(:,L)} g_1(X) & \nabla \nabla_{X(:,L)} g_2(X) & \cdots & \nabla \nabla_{X(:,L)} g_N(X) \end{bmatrix} \\ &= [\nabla^2 g_1(X) \quad \nabla^2 g_2(X) \quad \cdots \quad \nabla^2 g_N(X)] \in \mathbb{R}^{K \times N \times L \times K \times L} \end{aligned} \quad (1728)$$

The gradient of matrix-valued function  $g(X) : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}^{M \times N}$  on matrix domain has a four-dimensional representation called *quartix* (*fourth-order tensor*)

$$\nabla g(X) \triangleq \begin{bmatrix} \nabla g_{11}(X) & \nabla g_{12}(X) & \cdots & \nabla g_{1N}(X) \\ \nabla g_{21}(X) & \nabla g_{22}(X) & \cdots & \nabla g_{2N}(X) \\ \vdots & \vdots & \ddots & \vdots \\ \nabla g_{M1}(X) & \nabla g_{M2}(X) & \cdots & \nabla g_{MN}(X) \end{bmatrix} \in \mathbb{R}^{M \times N \times K \times L} \quad (1729)$$

while the second-order gradient has six-dimensional representation

$$\nabla^2 g(X) \triangleq \begin{bmatrix} \nabla^2 g_{11}(X) & \nabla^2 g_{12}(X) & \cdots & \nabla^2 g_{1N}(X) \\ \nabla^2 g_{21}(X) & \nabla^2 g_{22}(X) & \cdots & \nabla^2 g_{2N}(X) \\ \vdots & \vdots & \ddots & \vdots \\ \nabla^2 g_{M1}(X) & \nabla^2 g_{M2}(X) & \cdots & \nabla^2 g_{MN}(X) \end{bmatrix} \in \mathbb{R}^{M \times N \times K \times L \times K \times L} \quad (1730)$$

and so on.

### D.1.2 Product rules for matrix-functions

Given dimensionally compatible matrix-valued functions of matrix variable  $f(X)$  and  $g(X)$

$$\nabla_X (f(X)^T g(X)) = \nabla_X (f) g + \nabla_X (g) f \tag{1731}$$

while [51, §8.3] [309]

$$\nabla_X \operatorname{tr}(f(X)^T g(X)) = \nabla_X \left( \operatorname{tr}(f(X)^T g(Z)) + \operatorname{tr}(g(X) f(Z)^T) \right) \Big|_{Z=X} \tag{1732}$$

These expressions implicitly apply as well to scalar-, vector-, or matrix-valued functions of scalar, vector, or matrix arguments.

#### D.1.2.0.1 Example. Cubix.

Suppose  $f(X) : \mathbb{R}^{2 \times 2} \rightarrow \mathbb{R}^2 = X^T a$  and  $g(X) : \mathbb{R}^{2 \times 2} \rightarrow \mathbb{R}^2 = X b$ . We wish to find

$$\nabla_X (f(X)^T g(X)) = \nabla_X a^T X^2 b \tag{1733}$$

using the product rule. Formula (1731) calls for

$$\nabla_X a^T X^2 b = \nabla_X (X^T a) X b + \nabla_X (X b) X^T a \tag{1734}$$

Consider the first of the two terms:

$$\begin{aligned} \nabla_X (f) g &= \nabla_X (X^T a) X b \\ &= \left[ \nabla (X^T a)_1 \quad \nabla (X^T a)_2 \right] X b \end{aligned} \tag{1735}$$

The gradient of  $X^T a$  forms a cubix in  $\mathbb{R}^{2 \times 2 \times 2}$ ; a.k.a, *third-order tensor*.

$$\nabla_X (X^T a) X b = \left[ \begin{array}{cc} \frac{\partial (X^T a)_1}{\partial X_{11}} & \dots & \frac{\partial (X^T a)_2}{\partial X_{11}} \\ \vdots & \frac{\partial (X^T a)_1}{\partial X_{12}} & \dots & \frac{\partial (X^T a)_2}{\partial X_{12}} \\ \frac{\partial (X^T a)_1}{\partial X_{21}} & \dots & \frac{\partial (X^T a)_2}{\partial X_{21}} \\ \vdots & \frac{\partial (X^T a)_1}{\partial X_{22}} & \dots & \frac{\partial (X^T a)_2}{\partial X_{22}} \end{array} \right] \begin{bmatrix} (X b)_1 \\ (X b)_2 \end{bmatrix} \in \mathbb{R}^{2 \times 1 \times 2} \tag{1736}$$

Because gradient of the product (1733) requires total change with respect to change in each entry of matrix  $X$ , the  $Xb$  vector must make an inner product with each vector in the second dimension of the cubix (indicated by dotted line segments);

$$\begin{aligned} \nabla_X(X^T a) X b &= \begin{bmatrix} a_1 & 0 & & \\ & 0 & a_1 & \\ a_2 & 0 & & \\ & 0 & a_2 & \end{bmatrix} \begin{bmatrix} b_1 X_{11} + b_2 X_{12} \\ b_1 X_{21} + b_2 X_{22} \end{bmatrix} \in \mathbb{R}^{2 \times 1 \times 2} \\ &= \begin{bmatrix} a_1(b_1 X_{11} + b_2 X_{12}) & a_1(b_1 X_{21} + b_2 X_{22}) \\ a_2(b_1 X_{11} + b_2 X_{12}) & a_2(b_1 X_{21} + b_2 X_{22}) \end{bmatrix} \in \mathbb{R}^{2 \times 2} \\ &= ab^T X^T \end{aligned} \quad (1737)$$

where the cubix appears as a complete  $2 \times 2 \times 2$  matrix. In like manner for the second term  $\nabla_X(g) f$

$$\begin{aligned} \nabla_X(Xb) X^T a &= \begin{bmatrix} b_1 & 0 & & \\ & b_2 & 0 & \\ 0 & & b_1 & \\ & 0 & & b_2 \end{bmatrix} \begin{bmatrix} X_{11} a_1 + X_{21} a_2 \\ X_{12} a_1 + X_{22} a_2 \end{bmatrix} \in \mathbb{R}^{2 \times 1 \times 2} \\ &= X^T ab^T \in \mathbb{R}^{2 \times 2} \end{aligned} \quad (1738)$$

The solution

$$\nabla_X a^T X^2 b = ab^T X^T + X^T ab^T \quad (1739)$$

can be found from Table D.2.1 or verified using (1732).  $\square$

### D.1.2.1 Kronecker product

A partial remedy for venturing into *hyperdimensional* matrix representations, such as the cubix or quartix, is to first vectorize matrices as in (37). This device gives rise to the Kronecker product of matrices  $\otimes$ ; a.k.a, *direct product* or *tensor product*. Although it sees reversal in the literature, [321, §2.1] we adopt the definition: for  $A \in \mathbb{R}^{m \times n}$  and  $B \in \mathbb{R}^{p \times q}$

$$B \otimes A \triangleq \begin{bmatrix} B_{11}A & B_{12}A & \cdots & B_{1q}A \\ B_{21}A & B_{22}A & \cdots & B_{2q}A \\ \vdots & \vdots & \ddots & \vdots \\ B_{p1}A & B_{p2}A & \cdots & B_{pq}A \end{bmatrix} \in \mathbb{R}^{pm \times qn} \quad (1740)$$

for which  $A \otimes 1 = 1 \otimes A = A$  (real unity acts like identity).

One advantage to vectorization is existence of a traditional two-dimensional matrix representation (*second-order tensor*) for the second-order gradient of a real function with respect to a vectorized matrix. For example, from §A.1.1 no.33 (§D.2.1) for square  $A, B \in \mathbb{R}^{n \times n}$  [162, §5.2] [12, §3]

$$\nabla_{\text{vec } X}^2 \text{tr}(AXBX^T) = \nabla_{\text{vec } X}^2 \text{vec}(X)^T (B^T \otimes A) \text{vec } X = B \otimes A^T + B^T \otimes A \in \mathbb{R}^{n^2 \times n^2} \quad (1741)$$

To disadvantage is a large new but known set of algebraic rules (§A.1.1) and the fact that its mere use does not generally guarantee two-dimensional matrix representation of gradients.

Another application of the Kronecker product is to reverse order of appearance in a matrix product: Suppose we wish to weight the columns of a matrix  $S \in \mathbb{R}^{M \times N}$ , for example, by respective entries  $w_i$  from the main diagonal in

$$W \triangleq \begin{bmatrix} w_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0}^T & & w_N \end{bmatrix} \in \mathbb{S}^N \quad (1742)$$

A conventional means for accomplishing column weighting is to multiply  $S$  by diagonal matrix  $W$  on the right-hand side:

$$SW = S \begin{bmatrix} w_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0}^T & & w_N \end{bmatrix} = [S(:, 1)w_1 \quad \cdots \quad S(:, N)w_N] \in \mathbb{R}^{M \times N} \quad (1743)$$

To reverse product order such that diagonal matrix  $W$  instead appears to the left of  $S$ : for  $I \in \mathbb{S}^M$  (Sze Wan)

$$SW = (\delta(W)^T \otimes I) \begin{bmatrix} S(:, 1) & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & S(:, 2) & \ddots & \\ & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & & \mathbf{0} & S(:, N) \end{bmatrix} \in \mathbb{R}^{M \times N} \quad (1744)$$

To instead weight the rows of  $S$  via diagonal matrix  $W \in \mathbb{S}^M$ , for  $I \in \mathbb{S}^N$

$$WS = \begin{bmatrix} S(1, :) & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & S(2, :) & \ddots & \\ & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & & \mathbf{0} & S(M, :) \end{bmatrix} (\delta(W) \otimes I) \in \mathbb{R}^{M \times N} \quad (1745)$$

For any matrices of like size,  $S, Y \in \mathbb{R}^{M \times N}$

$$S \circ Y = [\delta(Y(:, 1)) \cdots \delta(Y(:, N))] \begin{bmatrix} S(:, 1) & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & S(:, 2) & \cdots & \\ & & \ddots & \mathbf{0} \\ \mathbf{0} & & & S(:, N) \end{bmatrix} \in \mathbb{R}^{M \times N} \quad (1746)$$

which converts a Hadamard product into a standard matrix product. In the special case that  $S = s$  and  $Y = y$  are vectors in  $\mathbb{R}^M$

$$s \circ y = \delta(s)y \quad (1747)$$

$$\begin{aligned} s^T \otimes y &= ys^T \\ s \otimes y^T &= sy^T \end{aligned} \quad (1748)$$

### D.1.3 Chain rules for composite matrix-functions

Given dimensionally compatible matrix-valued functions of matrix variable  $f(X)$  and  $g(X)$  [214, §15.7]

$$\nabla_X g(f(X)^T) = \nabla_X f^T \nabla_f g \quad (1749)$$

$$\nabla_X^2 g(f(X)^T) = \nabla_X (\nabla_X f^T \nabla_f g) = \nabla_X^2 f \nabla_f g + \nabla_X f^T \nabla_f^2 g \nabla_X f \quad (1750)$$

#### D.1.3.1 Two arguments

$$\nabla_X g(f(X)^T, h(X)^T) = \nabla_X f^T \nabla_f g + \nabla_X h^T \nabla_h g \quad (1751)$$

**D.1.3.1.1 Example.** *Chain rule for two arguments.* [41, §1.1]

$$g(f(x)^T, h(x)^T) = (f(x) + h(x))^T A (f(x) + h(x)) \quad (1752)$$

$$f(x) = \begin{bmatrix} x_1 \\ \varepsilon x_2 \end{bmatrix}, \quad h(x) = \begin{bmatrix} \varepsilon x_1 \\ x_2 \end{bmatrix} \quad (1753)$$

$$\nabla_x g(f(x)^T, h(x)^T) = \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \end{bmatrix} (A + A^T)(f + h) + \begin{bmatrix} \varepsilon & 0 \\ 0 & 1 \end{bmatrix} (A + A^T)(f + h) \quad (1754)$$

$$\nabla_x g(f(x)^T, h(x)^T) = \begin{bmatrix} 1 + \varepsilon & 0 \\ 0 & 1 + \varepsilon \end{bmatrix} (A + A^T) \left( \begin{bmatrix} x_1 \\ \varepsilon x_2 \end{bmatrix} + \begin{bmatrix} \varepsilon x_1 \\ x_2 \end{bmatrix} \right) \quad (1755)$$

$$\lim_{\varepsilon \rightarrow 0} \nabla_x g(f(x)^T, h(x)^T) = (A + A^T)x \quad (1756)$$

from Table [D.2.1](#).  $\square$

These foregoing formulae remain correct when gradient produces hyperdimensional representation:

#### D.1.4 First directional derivative

Assume that a differentiable function  $g(X) : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}^{M \times N}$  has continuous first- and second-order gradients  $\nabla g$  and  $\nabla^2 g$  over  $\text{dom } g$  which is an open set. We seek simple expressions for the first and second directional derivatives in direction  $Y \in \mathbb{R}^{K \times L}$ : respectively,  $\overset{\rightarrow Y}{dg} \in \mathbb{R}^{M \times N}$  and  $\overset{\rightarrow Y}{dg^2} \in \mathbb{R}^{M \times N}$ .

Assuming that the limit exists, we may state the partial derivative of the  $mn^{\text{th}}$  entry of  $g$  with respect to the  $kl^{\text{th}}$  entry of  $X$ ;

$$\frac{\partial g_{mn}(X)}{\partial X_{kl}} = \lim_{\Delta t \rightarrow 0} \frac{g_{mn}(X + \Delta t e_k e_l^T) - g_{mn}(X)}{\Delta t} \in \mathbb{R} \quad (1757)$$

where  $e_k$  is the  $k^{\text{th}}$  standard basis vector in  $\mathbb{R}^K$  while  $e_l$  is the  $l^{\text{th}}$  standard basis vector in  $\mathbb{R}^L$ . The total number of partial derivatives equals  $KLMN$  while the gradient is defined in their terms; the  $mn^{\text{th}}$  entry of the gradient is

$$\nabla g_{mn}(X) = \begin{bmatrix} \frac{\partial g_{mn}(X)}{\partial X_{11}} & \frac{\partial g_{mn}(X)}{\partial X_{12}} & \cdots & \frac{\partial g_{mn}(X)}{\partial X_{1L}} \\ \frac{\partial g_{mn}(X)}{\partial X_{21}} & \frac{\partial g_{mn}(X)}{\partial X_{22}} & \cdots & \frac{\partial g_{mn}(X)}{\partial X_{2L}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial g_{mn}(X)}{\partial X_{K1}} & \frac{\partial g_{mn}(X)}{\partial X_{K2}} & \cdots & \frac{\partial g_{mn}(X)}{\partial X_{KL}} \end{bmatrix} \in \mathbb{R}^{K \times L} \quad (1758)$$

while the gradient is a quartix

$$\nabla g(X) = \begin{bmatrix} \nabla g_{11}(X) & \nabla g_{12}(X) & \cdots & \nabla g_{1N}(X) \\ \nabla g_{21}(X) & \nabla g_{22}(X) & \cdots & \nabla g_{2N}(X) \\ \vdots & \vdots & & \vdots \\ \nabla g_{M1}(X) & \nabla g_{M2}(X) & \cdots & \nabla g_{MN}(X) \end{bmatrix} \in \mathbb{R}^{M \times N \times K \times L} \quad (1759)$$

By simply rotating our perspective of the four-dimensional representation of gradient matrix, we find one of three useful transpositions of this quartix (connoted  $T_1$ ):

$$\nabla g(X)^{T_1} = \begin{bmatrix} \frac{\partial g(X)}{\partial X_{11}} & \frac{\partial g(X)}{\partial X_{12}} & \dots & \frac{\partial g(X)}{\partial X_{1L}} \\ \frac{\partial g(X)}{\partial X_{21}} & \frac{\partial g(X)}{\partial X_{22}} & \dots & \frac{\partial g(X)}{\partial X_{2L}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial g(X)}{\partial X_{K1}} & \frac{\partial g(X)}{\partial X_{K2}} & \dots & \frac{\partial g(X)}{\partial X_{KL}} \end{bmatrix} \in \mathbb{R}^{K \times L \times M \times N} \quad (1760)$$

When the limit for  $\Delta t \in \mathbb{R}$  exists, it is easy to show by substitution of variables in (1757)

$$\frac{\partial g_{mn}(X)}{\partial X_{kl}} Y_{kl} = \lim_{\Delta t \rightarrow 0} \frac{g_{mn}(X + \Delta t Y_{kl} e_k e_l^T) - g_{mn}(X)}{\Delta t} \in \mathbb{R} \quad (1761)$$

which may be interpreted as the change in  $g_{mn}$  at  $X$  when the change in  $X_{kl}$  is equal to  $Y_{kl}$ , the  $kl^{\text{th}}$  entry of any  $Y \in \mathbb{R}^{K \times L}$ . Because the total change in  $g_{mn}(X)$  due to  $Y$  is the sum of change with respect to each and every  $X_{kl}$ , the  $mn^{\text{th}}$  entry of the directional derivative is the corresponding total differential [214, §15.8]

$$dg_{mn}(X)|_{dX \rightarrow Y} = \sum_{k,l} \frac{\partial g_{mn}(X)}{\partial X_{kl}} Y_{kl} = \text{tr}(\nabla g_{mn}(X)^T Y) \quad (1762)$$

$$= \sum_{k,l} \lim_{\Delta t \rightarrow 0} \frac{g_{mn}(X + \Delta t Y_{kl} e_k e_l^T) - g_{mn}(X)}{\Delta t} \quad (1763)$$

$$= \lim_{\Delta t \rightarrow 0} \frac{g_{mn}(X + \Delta t Y) - g_{mn}(X)}{\Delta t} \quad (1764)$$

$$= \left. \frac{d}{dt} \right|_{t=0} g_{mn}(X + tY) \quad (1765)$$

where  $t \in \mathbb{R}$ . Assuming finite  $Y$ , equation (1764) is called the *Gâteaux differential* [40, App.A.5] [195, §D.2.1] [351, §5.28] whose existence is implied by existence of the *Fréchet differential* (the sum in (1762)). [244, §7.2] Each may be understood as the change in  $g_{mn}$  at  $X$  when the change in  $X$  is equal

in magnitude and direction to  $Y$ .<sup>D.2</sup> Hence the directional derivative,

$$\begin{aligned}
\overset{\rightarrow Y}{dg}(X) &\triangleq \left[ \begin{array}{cccc} dg_{11}(X) & dg_{12}(X) & \cdots & dg_{1N}(X) \\ dg_{21}(X) & dg_{22}(X) & \cdots & dg_{2N}(X) \\ \vdots & \vdots & & \vdots \\ dg_{M1}(X) & dg_{M2}(X) & \cdots & dg_{MN}(X) \end{array} \right] \Bigg|_{dX \rightarrow Y} \in \mathbb{R}^{M \times N} \\
&= \left[ \begin{array}{cccc} \text{tr}(\nabla g_{11}(X)^T Y) & \text{tr}(\nabla g_{12}(X)^T Y) & \cdots & \text{tr}(\nabla g_{1N}(X)^T Y) \\ \text{tr}(\nabla g_{21}(X)^T Y) & \text{tr}(\nabla g_{22}(X)^T Y) & \cdots & \text{tr}(\nabla g_{2N}(X)^T Y) \\ \vdots & \vdots & & \vdots \\ \text{tr}(\nabla g_{M1}(X)^T Y) & \text{tr}(\nabla g_{M2}(X)^T Y) & \cdots & \text{tr}(\nabla g_{MN}(X)^T Y) \end{array} \right] \\
&= \left[ \begin{array}{cccc} \sum_{k,l} \frac{\partial g_{11}(X)}{\partial X_{kl}} Y_{kl} & \sum_{k,l} \frac{\partial g_{12}(X)}{\partial X_{kl}} Y_{kl} & \cdots & \sum_{k,l} \frac{\partial g_{1N}(X)}{\partial X_{kl}} Y_{kl} \\ \sum_{k,l} \frac{\partial g_{21}(X)}{\partial X_{kl}} Y_{kl} & \sum_{k,l} \frac{\partial g_{22}(X)}{\partial X_{kl}} Y_{kl} & \cdots & \sum_{k,l} \frac{\partial g_{2N}(X)}{\partial X_{kl}} Y_{kl} \\ \vdots & \vdots & & \vdots \\ \sum_{k,l} \frac{\partial g_{M1}(X)}{\partial X_{kl}} Y_{kl} & \sum_{k,l} \frac{\partial g_{M2}(X)}{\partial X_{kl}} Y_{kl} & \cdots & \sum_{k,l} \frac{\partial g_{MN}(X)}{\partial X_{kl}} Y_{kl} \end{array} \right] \quad (1766)
\end{aligned}$$

from which it follows

$$\overset{\rightarrow Y}{dg}(X) = \sum_{k,l} \frac{\partial g(X)}{\partial X_{kl}} Y_{kl} \quad (1767)$$

Yet for all  $X \in \text{dom } g$ , any  $Y \in \mathbb{R}^{K \times L}$ , and some open interval of  $t \in \mathbb{R}$

$$g(X + tY) = g(X) + t \overset{\rightarrow Y}{dg}(X) + o(t^2) \quad (1768)$$

which is the first-order Taylor series expansion about  $X$ . [214, §18.4] [149, §2.3.4] Differentiation with respect to  $t$  and subsequent  $t$ -zeroing isolates the second term of expansion. Thus differentiating and zeroing  $g(X + tY)$  in  $t$  is an operation equivalent to individually differentiating and zeroing every entry  $g_{mn}(X + tY)$  as in (1765). So the directional derivative of  $g(X) : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}^{M \times N}$  in any direction  $Y \in \mathbb{R}^{K \times L}$  evaluated at  $X \in \text{dom } g$  becomes

$$\overset{\rightarrow Y}{dg}(X) = \frac{d}{dt} \Bigg|_{t=0} g(X + tY) \in \mathbb{R}^{M \times N} \quad (1769)$$

<sup>D.2</sup> Although  $Y$  is a matrix, we may regard it as a vector in  $\mathbb{R}^{KL}$ .

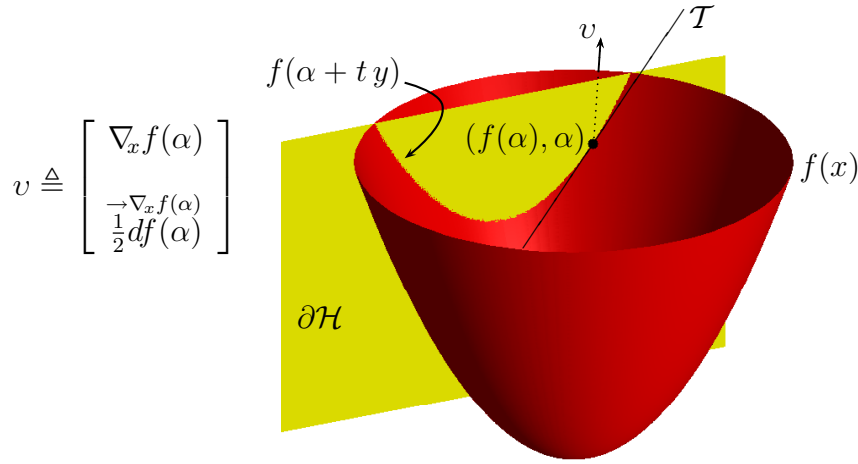


Figure 157: Strictly convex quadratic bowl in  $\mathbb{R}^2 \times \mathbb{R}$ ;  $f(x) = x^T x : \mathbb{R}^2 \rightarrow \mathbb{R}$  versus  $x$  on some open disc in  $\mathbb{R}^2$ . Plane slice  $\partial\mathcal{H}$  is perpendicular to function domain. Slice intersection with domain connotes bidirectional vector  $y$ . Slope of tangent line  $\mathcal{T}$  at point  $(\alpha, f(\alpha))$  is value of  $\nabla_x f(\alpha)^T y$  directional derivative (1794) at  $\alpha$  in slice direction  $y$ . Negative gradient  $-\nabla_x f(x) \in \mathbb{R}^2$  is direction of *steepest descent*. [368] [214, §15.6] [149] When vector  $v \in \mathbb{R}^3$  entry  $v_3$  is half directional derivative in gradient direction at  $\alpha$  and when  $\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \nabla_x f(\alpha)$ , then  $-v$  points directly toward bowl bottom.

[271, §2.1, §5.4.5] [33, §6.3.1] which is simplest. In case of a real function  $g(X) : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}$

$$\overrightarrow{dg}(X) = \text{tr}(\nabla g(X)^T Y) \quad (1791)$$

In case  $g(X) : \mathbb{R}^K \rightarrow \mathbb{R}$

$$\overrightarrow{dg}(X) = \nabla g(X)^T Y \quad (1794)$$

Unlike gradient, directional derivative does not expand dimension; directional derivative (1769) retains the dimensions of  $g$ . The derivative with respect to  $t$  makes the directional derivative resemble ordinary calculus (§D.2); e.g., when  $g(X)$  is linear,  $\overrightarrow{dg}(X) = g(Y)$ . [244, §7.2]

### D.1.4.1 Interpretation of directional derivative

In the case of any differentiable real function  $g(X) : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}$ , the directional derivative of  $g(X)$  at  $X$  in any direction  $Y$  yields the slope of  $g$  along the line  $\{X + tY \mid t \in \mathbb{R}\}$  through its domain evaluated at  $t = 0$ . For higher-dimensional functions, by (1766), this slope interpretation can be applied to each entry of the directional derivative.

Figure 157, for example, shows a plane slice of a real convex bowl-shaped function  $f(x)$  along a line  $\{\alpha + ty \mid t \in \mathbb{R}\}$  through its domain. The slice reveals a one-dimensional real function of  $t$ ;  $f(\alpha + ty)$ . The directional derivative at  $x = \alpha$  in direction  $y$  is the slope of  $f(\alpha + ty)$  with respect to  $t$  at  $t = 0$ . In the case of a real function having vector argument  $h(X) : \mathbb{R}^K \rightarrow \mathbb{R}$ , its directional derivative in the normalized direction of its gradient is the gradient magnitude. (1794) For a real function of real variable, the directional derivative evaluated at any point in the function domain is just the slope of that function there scaled by the real direction. (*confer* §3.7)

Directional derivative generalizes our one-dimensional notion of derivative to a multidimensional domain. When direction  $Y$  coincides with a member of the standard Cartesian basis  $e_k e_l^T$  (60), then a single partial derivative  $\partial g(X) / \partial X_{kl}$  is obtained from directional derivative (1767); such is each entry of gradient  $\nabla g(X)$  in equalities (1791) and (1794), for example.

#### D.1.4.1.1 Theorem. Directional derivative optimality condition.

[244, §7.4] Suppose  $f(X) : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}$  is minimized on convex set  $\mathcal{C} \subseteq \mathbb{R}^{p \times k}$  by  $X^*$ , and the directional derivative of  $f$  exists there. Then for all  $X \in \mathcal{C}$

$$df(X) \stackrel{\rightarrow X - X^*}{\geq} 0 \quad (1770)$$

◇

#### D.1.4.1.2 Example. Simple bowl.

Bowl function (Figure 157)

$$f(x) : \mathbb{R}^K \rightarrow \mathbb{R} \triangleq (x - a)^T(x - a) - b \quad (1771)$$

has function offset  $-b \in \mathbb{R}$ , axis of revolution at  $x = a$ , and positive definite Hessian (1720) everywhere in its domain (an open *hyperdisc* in  $\mathbb{R}^K$ ); *id est*, strictly convex quadratic  $f(x)$  has unique global minimum equal to  $-b$  at

$x = a$ . A vector  $-v$  based anywhere in  $\text{dom } f \times \mathbb{R}$  pointing toward the unique bowl-bottom is specified:

$$v \propto \begin{bmatrix} x - a \\ f(x) + b \end{bmatrix} \in \mathbb{R}^K \times \mathbb{R} \quad (1772)$$

Such a vector is

$$v = \begin{bmatrix} \nabla_x f(x) \\ -\nabla_x f(x) \\ \frac{1}{2} df(x) \end{bmatrix} \quad (1773)$$

since the gradient is

$$\nabla_x f(x) = 2(x - a) \quad (1774)$$

and the directional derivative in the direction of the gradient is (1794)

$$\begin{aligned} \begin{matrix} \rightarrow \\ \nabla_x f(x) \end{matrix} df(x) &= \nabla_x f(x)^T \nabla_x f(x) = 4(x - a)^T (x - a) = 4(f(x) + b) \quad (1775) \\ &\quad \square \end{aligned}$$

### D.1.5 Second directional derivative

By similar argument, it so happens: the second directional derivative is equally simple. Given  $g(X) : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}^{M \times N}$  on open domain,

$$\nabla \frac{\partial g_{mn}(X)}{\partial X_{kl}} = \frac{\partial \nabla g_{mn}(X)}{\partial X_{kl}} = \begin{bmatrix} \frac{\partial^2 g_{mn}(X)}{\partial X_{ki} \partial X_{11}} & \frac{\partial^2 g_{mn}(X)}{\partial X_{ki} \partial X_{12}} & \cdots & \frac{\partial^2 g_{mn}(X)}{\partial X_{ki} \partial X_{1L}} \\ \frac{\partial^2 g_{mn}(X)}{\partial X_{ki} \partial X_{21}} & \frac{\partial^2 g_{mn}(X)}{\partial X_{ki} \partial X_{22}} & \cdots & \frac{\partial^2 g_{mn}(X)}{\partial X_{ki} \partial X_{2L}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 g_{mn}(X)}{\partial X_{ki} \partial X_{K1}} & \frac{\partial^2 g_{mn}(X)}{\partial X_{ki} \partial X_{K2}} & \cdots & \frac{\partial^2 g_{mn}(X)}{\partial X_{ki} \partial X_{KL}} \end{bmatrix} \in \mathbb{R}^{K \times L} \quad (1776)$$

$$\nabla^2 g_{mn}(X) = \begin{bmatrix} \nabla \frac{\partial g_{mn}(X)}{\partial X_{11}} & \nabla \frac{\partial g_{mn}(X)}{\partial X_{12}} & \cdots & \nabla \frac{\partial g_{mn}(X)}{\partial X_{1L}} \\ \nabla \frac{\partial g_{mn}(X)}{\partial X_{21}} & \nabla \frac{\partial g_{mn}(X)}{\partial X_{22}} & \cdots & \nabla \frac{\partial g_{mn}(X)}{\partial X_{2L}} \\ \vdots & \vdots & & \vdots \\ \nabla \frac{\partial g_{mn}(X)}{\partial X_{K1}} & \nabla \frac{\partial g_{mn}(X)}{\partial X_{K2}} & \cdots & \nabla \frac{\partial g_{mn}(X)}{\partial X_{KL}} \end{bmatrix} \in \mathbb{R}^{K \times L \times K \times L} \quad (1777)$$

$$= \begin{bmatrix} \frac{\partial \nabla g_{mn}(X)}{\partial X_{11}} & \frac{\partial \nabla g_{mn}(X)}{\partial X_{12}} & \cdots & \frac{\partial \nabla g_{mn}(X)}{\partial X_{1L}} \\ \frac{\partial \nabla g_{mn}(X)}{\partial X_{21}} & \frac{\partial \nabla g_{mn}(X)}{\partial X_{22}} & \cdots & \frac{\partial \nabla g_{mn}(X)}{\partial X_{2L}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial \nabla g_{mn}(X)}{\partial X_{K1}} & \frac{\partial \nabla g_{mn}(X)}{\partial X_{K2}} & \cdots & \frac{\partial \nabla g_{mn}(X)}{\partial X_{KL}} \end{bmatrix}$$

Rotating our perspective, we get several views of the second-order gradient:

$$\nabla^2 g(X) = \begin{bmatrix} \nabla^2 g_{11}(X) & \nabla^2 g_{12}(X) & \cdots & \nabla^2 g_{1N}(X) \\ \nabla^2 g_{21}(X) & \nabla^2 g_{22}(X) & \cdots & \nabla^2 g_{2N}(X) \\ \vdots & \vdots & \ddots & \vdots \\ \nabla^2 g_{M1}(X) & \nabla^2 g_{M2}(X) & \cdots & \nabla^2 g_{MN}(X) \end{bmatrix} \in \mathbb{R}^{M \times N \times K \times L \times K \times L} \quad (1778)$$

$$\nabla^2 g(X)^{T_1} = \begin{bmatrix} \nabla \frac{\partial g(X)}{\partial X_{11}} & \nabla \frac{\partial g(X)}{\partial X_{12}} & \cdots & \nabla \frac{\partial g(X)}{\partial X_{1L}} \\ \nabla \frac{\partial g(X)}{\partial X_{21}} & \nabla \frac{\partial g(X)}{\partial X_{22}} & \cdots & \nabla \frac{\partial g(X)}{\partial X_{2L}} \\ \vdots & \vdots & \ddots & \vdots \\ \nabla \frac{\partial g(X)}{\partial X_{K1}} & \nabla \frac{\partial g(X)}{\partial X_{K2}} & \cdots & \nabla \frac{\partial g(X)}{\partial X_{KL}} \end{bmatrix} \in \mathbb{R}^{K \times L \times M \times N \times K \times L} \quad (1779)$$

$$\nabla^2 g(X)^{T_2} = \begin{bmatrix} \frac{\partial \nabla g(X)}{\partial X_{11}} & \frac{\partial \nabla g(X)}{\partial X_{12}} & \cdots & \frac{\partial \nabla g(X)}{\partial X_{1L}} \\ \frac{\partial \nabla g(X)}{\partial X_{21}} & \frac{\partial \nabla g(X)}{\partial X_{22}} & \cdots & \frac{\partial \nabla g(X)}{\partial X_{2L}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \nabla g(X)}{\partial X_{K1}} & \frac{\partial \nabla g(X)}{\partial X_{K2}} & \cdots & \frac{\partial \nabla g(X)}{\partial X_{KL}} \end{bmatrix} \in \mathbb{R}^{K \times L \times K \times L \times M \times N} \quad (1780)$$

Assuming the limits exist, we may state the partial derivative of the  $mn^{\text{th}}$  entry of  $g$  with respect to the  $kl^{\text{th}}$  and  $ij^{\text{th}}$  entries of  $X$ ;

$$\frac{\partial^2 g_{mn}(X)}{\partial X_{kl} \partial X_{ij}} = \lim_{\Delta\tau, \Delta t \rightarrow 0} \frac{g_{mn}(X + \Delta t e_k e_l^T + \Delta\tau e_i e_j^T) - g_{mn}(X + \Delta t e_k e_l^T) - (g_{mn}(X + \Delta\tau e_i e_j^T) - g_{mn}(X))}{\Delta\tau \Delta t} \quad (1781)$$

Differentiating (1761) and then scaling by  $Y_{ij}$

$$\begin{aligned} \frac{\partial^2 g_{mn}(X)}{\partial X_{kl} \partial X_{ij}} Y_{kl} Y_{ij} &= \lim_{\Delta t \rightarrow 0} \frac{\partial g_{mn}(X + \Delta t Y_{kl} e_k e_l^T) - \partial g_{mn}(X)}{\partial X_{ij} \Delta t} Y_{ij} \\ &= \lim_{\Delta\tau, \Delta t \rightarrow 0} \frac{g_{mn}(X + \Delta t Y_{kl} e_k e_l^T + \Delta\tau Y_{ij} e_i e_j^T) - g_{mn}(X + \Delta t Y_{kl} e_k e_l^T) - (g_{mn}(X + \Delta\tau Y_{ij} e_i e_j^T) - g_{mn}(X))}{\Delta\tau \Delta t} \end{aligned} \quad (1782)$$

which can be proved by substitution of variables in (1781). The  $mn^{\text{th}}$  second-order total differential due to any  $Y \in \mathbb{R}^{K \times L}$  is

$$d^2g_{mn}(X)|_{dX \rightarrow Y} = \sum_{i,j} \sum_{k,l} \frac{\partial^2 g_{mn}(X)}{\partial X_{kl} \partial X_{ij}} Y_{kl} Y_{ij} = \text{tr} \left( \nabla_X \text{tr} (\nabla g_{mn}(X)^T Y)^T Y \right) \quad (1783)$$

$$= \sum_{i,j} \lim_{\Delta t \rightarrow 0} \frac{\partial g_{mn}(X + \Delta t Y) - \partial g_{mn}(X)}{\partial X_{ij} \Delta t} Y_{ij} \quad (1784)$$

$$= \lim_{\Delta t \rightarrow 0} \frac{g_{mn}(X + 2\Delta t Y) - 2g_{mn}(X + \Delta t Y) + g_{mn}(X)}{\Delta t^2} \quad (1785)$$

$$= \left. \frac{d^2}{dt^2} \right|_{t=0} g_{mn}(X + t Y) \quad (1786)$$

Hence the second directional derivative,

$$\begin{aligned} \overset{-Y}{dg^2}(X) &\triangleq \left[ \begin{array}{cccc} d^2g_{11}(X) & d^2g_{12}(X) & \cdots & d^2g_{1N}(X) \\ d^2g_{21}(X) & d^2g_{22}(X) & \cdots & d^2g_{2N}(X) \\ \vdots & \vdots & & \vdots \\ d^2g_{M1}(X) & d^2g_{M2}(X) & \cdots & d^2g_{MN}(X) \end{array} \right] \Bigg|_{dX \rightarrow Y} \in \mathbb{R}^{M \times N} \\ &= \left[ \begin{array}{cccc} \text{tr} \left( \nabla \text{tr} (\nabla g_{11}(X)^T Y)^T Y \right) & \text{tr} \left( \nabla \text{tr} (\nabla g_{12}(X)^T Y)^T Y \right) & \cdots & \text{tr} \left( \nabla \text{tr} (\nabla g_{1N}(X)^T Y)^T Y \right) \\ \text{tr} \left( \nabla \text{tr} (\nabla g_{21}(X)^T Y)^T Y \right) & \text{tr} \left( \nabla \text{tr} (\nabla g_{22}(X)^T Y)^T Y \right) & \cdots & \text{tr} \left( \nabla \text{tr} (\nabla g_{2N}(X)^T Y)^T Y \right) \\ \vdots & \vdots & & \vdots \\ \text{tr} \left( \nabla \text{tr} (\nabla g_{M1}(X)^T Y)^T Y \right) & \text{tr} \left( \nabla \text{tr} (\nabla g_{M2}(X)^T Y)^T Y \right) & \cdots & \text{tr} \left( \nabla \text{tr} (\nabla g_{MN}(X)^T Y)^T Y \right) \end{array} \right] \\ &= \left[ \begin{array}{cccc} \sum_{i,j} \sum_{k,l} \frac{\partial^2 g_{11}(X)}{\partial X_{kl} \partial X_{ij}} Y_{kl} Y_{ij} & \sum_{i,j} \sum_{k,l} \frac{\partial^2 g_{12}(X)}{\partial X_{kl} \partial X_{ij}} Y_{kl} Y_{ij} & \cdots & \sum_{i,j} \sum_{k,l} \frac{\partial^2 g_{1N}(X)}{\partial X_{kl} \partial X_{ij}} Y_{kl} Y_{ij} \\ \sum_{i,j} \sum_{k,l} \frac{\partial^2 g_{21}(X)}{\partial X_{kl} \partial X_{ij}} Y_{kl} Y_{ij} & \sum_{i,j} \sum_{k,l} \frac{\partial^2 g_{22}(X)}{\partial X_{kl} \partial X_{ij}} Y_{kl} Y_{ij} & \cdots & \sum_{i,j} \sum_{k,l} \frac{\partial^2 g_{2N}(X)}{\partial X_{kl} \partial X_{ij}} Y_{kl} Y_{ij} \\ \vdots & \vdots & & \vdots \\ \sum_{i,j} \sum_{k,l} \frac{\partial^2 g_{M1}(X)}{\partial X_{kl} \partial X_{ij}} Y_{kl} Y_{ij} & \sum_{i,j} \sum_{k,l} \frac{\partial^2 g_{M2}(X)}{\partial X_{kl} \partial X_{ij}} Y_{kl} Y_{ij} & \cdots & \sum_{i,j} \sum_{k,l} \frac{\partial^2 g_{MN}(X)}{\partial X_{kl} \partial X_{ij}} Y_{kl} Y_{ij} \end{array} \right] \quad (1787) \end{aligned}$$

from which it follows

$$\overset{\rightarrow Y}{dg^2}(X) = \sum_{i,j} \sum_{k,l} \frac{\partial^2 g(X)}{\partial X_{kl} \partial X_{ij}} Y_{kl} Y_{ij} = \sum_{i,j} \frac{\partial}{\partial X_{ij}} \overset{\rightarrow Y}{dg}(X) Y_{ij} \quad (1788)$$

Yet for all  $X \in \text{dom } g$ , any  $Y \in \mathbb{R}^{K \times L}$ , and some open interval of  $t \in \mathbb{R}$

$$g(X + tY) = g(X) + t \overset{\rightarrow Y}{dg}(X) + \frac{1}{2!} t^2 \overset{\rightarrow Y}{dg^2}(X) + o(t^3) \quad (1789)$$

which is the second-order Taylor series expansion about  $X$ . [214, §18.4] [149, §2.3.4] Differentiating twice with respect to  $t$  and subsequent  $t$ -zeroing isolates the third term of the expansion. Thus differentiating and zeroing  $g(X + tY)$  in  $t$  is an operation equivalent to individually differentiating and zeroing every entry  $g_{mn}(X + tY)$  as in (1786). So the second directional derivative of  $g(X) : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}^{M \times N}$  becomes [271, §2.1, §5.4.5] [33, §6.3.1]

$$\overset{\rightarrow Y}{dg^2}(X) = \left. \frac{d^2}{dt^2} \right|_{t=0} g(X + tY) \in \mathbb{R}^{M \times N} \quad (1790)$$

which is again simplest. (*confer*(1769)) Directional derivative retains the dimensions of  $g$ .

### D.1.6 directional derivative expressions

In the case of a real function  $g(X) : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}$ , all its directional derivatives are in  $\mathbb{R}$ :

$$\overset{\rightarrow Y}{dg}(X) = \text{tr}(\nabla g(X)^T Y) \quad (1791)$$

$$\overset{\rightarrow Y}{dg^2}(X) = \text{tr}\left(\nabla_X \text{tr}(\nabla g(X)^T Y)^T Y\right) = \text{tr}\left(\nabla_X \overset{\rightarrow Y}{dg}(X)^T Y\right) \quad (1792)$$

$$\overset{\rightarrow Y}{dg^3}(X) = \text{tr}\left(\nabla_X \text{tr}\left(\nabla_X \text{tr}(\nabla g(X)^T Y)^T Y\right)^T Y\right) = \text{tr}\left(\nabla_X \overset{\rightarrow Y}{dg^2}(X)^T Y\right) \quad (1793)$$

In the case  $g(X) : \mathbb{R}^K \rightarrow \mathbb{R}$  has vector argument, they further simplify:

$$\overset{\rightarrow Y}{dg}(X) = \nabla g(X)^T Y \quad (1794)$$

$$\overset{\rightarrow Y}{dg^2}(X) = Y^T \nabla^2 g(X) Y \quad (1795)$$

$$\overset{\rightarrow Y}{dg^3}(X) = \nabla_X (Y^T \nabla^2 g(X) Y)^T Y \quad (1796)$$

and so on.

### D.1.7 Taylor series

Series expansions of the differentiable matrix-valued function  $g(X)$ , of matrix argument, were given earlier in (1768) and (1789). Assuming  $g(X)$  has continuous first-, second-, and third-order gradients over the open set  $\text{dom } g$ , then for  $X \in \text{dom } g$  and any  $Y \in \mathbb{R}^{K \times L}$  the complete Taylor series is expressed on some open interval of  $\mu \in \mathbb{R}$

$$g(X + \mu Y) = g(X) + \mu \overset{\rightarrow Y}{dg}(X) + \frac{1}{2!} \mu^2 \overset{\rightarrow Y}{dg^2}(X) + \frac{1}{3!} \mu^3 \overset{\rightarrow Y}{dg^3}(X) + o(\mu^4) \quad (1797)$$

or on some open interval of  $\|Y\|_2$

$$g(Y) = g(X) + \overset{\rightarrow Y-X}{dg}(X) + \frac{1}{2!} \overset{\rightarrow Y-X}{dg^2}(X) + \frac{1}{3!} \overset{\rightarrow Y-X}{dg^3}(X) + o(\|Y\|^4) \quad (1798)$$

which are third-order expansions about  $X$ . The *mean value theorem* from calculus is what insures finite order of the series. [214] [41, §1.1] [40, App.A.5] [195, §0.4] These somewhat unbelievable formulae imply that a function can be determined over the whole of its domain by knowing its value and all its directional derivatives at a single point  $X$ .

#### D.1.7.0.1 Example. Inverse matrix function.

Say  $g(Y) = Y^{-1}$ . From the table on page 680,

$$\overset{\rightarrow Y}{dg}(X) = \left. \frac{d}{dt} \right|_{t=0} g(X + tY) = -X^{-1} Y X^{-1} \quad (1799)$$

$$\overset{\rightarrow Y}{dg^2}(X) = \left. \frac{d^2}{dt^2} \right|_{t=0} g(X + tY) = 2X^{-1} Y X^{-1} Y X^{-1} \quad (1800)$$

$$\overset{\rightarrow Y}{dg^3}(X) = \left. \frac{d^3}{dt^3} \right|_{t=0} g(X + tY) = -6X^{-1} Y X^{-1} Y X^{-1} Y X^{-1} \quad (1801)$$

Let's find the Taylor series expansion of  $g$  about  $X = I$ : Since  $g(I) = I$ , for  $\|Y\|_2 < 1$  ( $\mu = 1$  in (1797))

$$g(I + Y) = (I + Y)^{-1} = I - Y + Y^2 - Y^3 + \dots \quad (1802)$$

If  $Y$  is small,  $(I + Y)^{-1} \approx I - Y$ . <sup>D.3</sup> Now we find Taylor series expansion about  $X$ :

$$g(X + Y) = (X + Y)^{-1} = X^{-1} - X^{-1}YX^{-1} + 2X^{-1}YX^{-1}YX^{-1} - \dots \quad (1803)$$

If  $Y$  is small,  $(X + Y)^{-1} \approx X^{-1} - X^{-1}YX^{-1}$ . □

**D.1.7.0.2 Exercise.** *log det.* (confer [59, p.644])  
 Find the first three terms of a Taylor series expansion for  $\log \det Y$ . Specify an open interval over which the expansion holds in vicinity of  $X$ . ▼

### D.1.8 Correspondence of gradient to derivative

From the foregoing expressions for directional derivative, we derive a relationship between the gradient with respect to matrix  $X$  and the derivative with respect to real variable  $t$ :

#### D.1.8.1 first-order

Removing evaluation at  $t = 0$  from (1769), <sup>D.4</sup> we find an expression for the directional derivative of  $g(X)$  in direction  $Y$  evaluated anywhere along a line  $\{X + tY \mid t \in \mathbb{R}\}$  intersecting  $\text{dom } g$

$$\overset{\rightarrow Y}{dg}(X + tY) = \frac{d}{dt}g(X + tY) \quad (1804)$$

In the general case  $g(X) : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}^{M \times N}$ , from (1762) and (1765) we find

$$\text{tr}(\nabla_X g_{mn}(X + tY)^T Y) = \frac{d}{dt}g_{mn}(X + tY) \quad (1805)$$

---

<sup>D.3</sup>Had we instead set  $g(Y) = (I + Y)^{-1}$ , then the equivalent expansion would have been about  $X = \mathbf{0}$ .

<sup>D.4</sup>Justified by replacing  $X$  with  $X + tY$  in (1762)-(1764); beginning,

$$dg_{mn}(X + tY)|_{dX \rightarrow Y} = \sum_{k,l} \frac{\partial g_{mn}(X + tY)}{\partial X_{kl}} Y_{kl}$$

which is valid at  $t = 0$ , of course, when  $X \in \text{dom } g$ . In the important case of a real function  $g(X) : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}$ , from (1791) we have simply

$$\text{tr}(\nabla_X g(X + tY)^T Y) = \frac{d}{dt} g(X + tY) \quad (1806)$$

When, additionally,  $g(X) : \mathbb{R}^K \rightarrow \mathbb{R}$  has vector argument,

$$\nabla_X g(X + tY)^T Y = \frac{d}{dt} g(X + tY) \quad (1807)$$

**D.1.8.1.1 Example.** *Gradient.*

$g(X) = w^T X^T X w$ ,  $X \in \mathbb{R}^{K \times L}$ ,  $w \in \mathbb{R}^L$ . Using the tables in §D.2,

$$\text{tr}(\nabla_X g(X + tY)^T Y) = \text{tr}(2ww^T(X^T + tY^T)Y) \quad (1808)$$

$$= 2w^T(X^T Y + tY^T Y)w \quad (1809)$$

Applying the equivalence (1806),

$$\frac{d}{dt} g(X + tY) = \frac{d}{dt} w^T (X + tY)^T (X + tY) w \quad (1810)$$

$$= w^T (X^T Y + Y^T X + 2tY^T Y) w \quad (1811)$$

$$= 2w^T (X^T Y + tY^T Y) w \quad (1812)$$

which is the same as (1809); hence, equivalence is demonstrated.

It is easy to extract  $\nabla g(X)$  from (1812) knowing only (1806):

$$\begin{aligned} \text{tr}(\nabla_X g(X + tY)^T Y) &= 2w^T (X^T Y + tY^T Y) w \\ &= 2 \text{tr}(ww^T (X^T + tY^T) Y) \\ \text{tr}(\nabla_X g(X)^T Y) &= 2 \text{tr}(ww^T X^T Y) \\ &\Leftrightarrow \\ \nabla_X g(X) &= 2Xww^T \end{aligned} \quad (1813)$$

□

**D.1.8.2 second-order**

Likewise removing the evaluation at  $t = 0$  from (1790),

$$\frac{\rightarrow Y}{dg^2}(X + tY) = \frac{d^2}{dt^2} g(X + tY) \quad (1814)$$

we can find a similar relationship between the second-order gradient and the second derivative: In the general case  $g(X) : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}^{M \times N}$  from (1783) and (1786),

$$\operatorname{tr}\left(\nabla_X \operatorname{tr}\left(\nabla_X g_{mn}(X+tY)^T Y\right)^T Y\right) = \frac{d^2}{dt^2} g_{mn}(X+tY) \quad (1815)$$

In the case of a real function  $g(X) : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}$  we have, of course,

$$\operatorname{tr}\left(\nabla_X \operatorname{tr}\left(\nabla_X g(X+tY)^T Y\right)^T Y\right) = \frac{d^2}{dt^2} g(X+tY) \quad (1816)$$

From (1795), the simpler case, where the real function  $g(X) : \mathbb{R}^K \rightarrow \mathbb{R}$  has vector argument,

$$Y^T \nabla_X^2 g(X+tY) Y = \frac{d^2}{dt^2} g(X+tY) \quad (1817)$$

**D.1.8.2.1 Example.** *Second-order gradient.*

Given real function  $g(X) = \log \det X$  having domain  $\operatorname{int} \mathbb{S}_+^K$ , we want to find  $\nabla^2 g(X) \in \mathbb{R}^{K \times K \times K \times K}$ . From the tables in §D.2,

$$h(X) \triangleq \nabla g(X) = X^{-1} \in \operatorname{int} \mathbb{S}_+^K \quad (1818)$$

so  $\nabla^2 g(X) = \nabla h(X)$ . By (1805) and (1768), for  $Y \in \mathbb{S}^K$

$$\operatorname{tr}(\nabla h_{mn}(X)^T Y) = \left. \frac{d}{dt} \right|_{t=0} h_{mn}(X+tY) \quad (1819)$$

$$= \left( \left. \frac{d}{dt} \right|_{t=0} h(X+tY) \right)_{mn} \quad (1820)$$

$$= \left( \left. \frac{d}{dt} \right|_{t=0} (X+tY)^{-1} \right)_{mn} \quad (1821)$$

$$= -(X^{-1} Y X^{-1})_{mn} \quad (1822)$$

Setting  $Y$  to a member of  $\{e_k e_l^T \in \mathbb{R}^{K \times K} \mid k, l = 1 \dots K\}$ , and employing a property (39) of the trace function we find

$$\nabla^2 g(X)_{mnkl} = \operatorname{tr}(\nabla h_{mn}(X)^T e_k e_l^T) = \nabla h_{mn}(X)_{kl} = -(X^{-1} e_k e_l^T X^{-1})_{mn} \quad (1823)$$

$$\nabla^2 g(X)_{kl} = \nabla h(X)_{kl} = -(X^{-1} e_k e_l^T X^{-1}) \in \mathbb{R}^{K \times K} \quad (1824)$$

□

From all these first- and second-order expressions, we may generate new ones by evaluating both sides at arbitrary  $t$  (in some open interval) but only after the differentiation.

## D.2 Tables of gradients and derivatives

- Results may be numerically proven by Romberg extrapolation. [105]  
When proving results for symmetric matrices algebraically, it is critical to take gradients ignoring symmetry and to then substitute symmetric entries afterward. [162] [63]
- $a, b \in \mathbb{R}^n$ ,  $x, y \in \mathbb{R}^k$ ,  $A, B \in \mathbb{R}^{m \times n}$ ,  $X, Y \in \mathbb{R}^{K \times L}$ ,  $t, \mu \in \mathbb{R}$ ,  $i, j, k, \ell, K, L, m, n, M, N$  are integers, unless otherwise noted.
- $x^\mu$  means  $\delta(\delta(x)^\mu)$  for  $\mu \in \mathbb{R}$ ; *id est*, entrywise vector exponentiation.  $\delta$  is the main-diagonal linear operator (1391).  $x^0 \triangleq \mathbf{1}$ ,  $X^0 \triangleq I$  if square.
- $\frac{d}{dx} \triangleq \begin{bmatrix} \frac{d}{dx_1} \\ \vdots \\ \frac{d}{dx_k} \end{bmatrix}$ ,  $\overset{\rightarrow y}{dg}(x)$ ,  $\overset{\rightarrow y}{dg^2}(x)$  (directional derivatives §D.1),  $\log x$ ,  $e^x$ ,  $|x|$ ,  $\operatorname{sgn} x$ ,  $x/y$  (Hadamard quotient),  $\sqrt{x}$  (entrywise square root), *etcetera*, are maps  $f : \mathbb{R}^k \rightarrow \mathbb{R}^k$  that maintain dimension; *e.g.*, (§A.1.1)

$$\frac{d}{dx} x^{-1} \triangleq \nabla_x \mathbf{1}^T \delta(x)^{-1} \mathbf{1} \quad (1825)$$

- For  $A$  a scalar or square matrix, we have the Taylor series [75, §3.6]

$$e^A \triangleq \sum_{k=0}^{\infty} \frac{1}{k!} A^k \quad (1826)$$

Further, [325, §5.4]

$$e^A \succ 0 \quad \forall A \in \mathbb{S}^m \quad (1827)$$

- For all square  $A$  and integer  $k$

$$\det^k A = \det A^k \quad (1828)$$



**algebraic** continued

$$\frac{d}{dt}(X + tY) = Y$$

$$\frac{d}{dt}B^T(X + tY)^{-1}A = -B^T(X + tY)^{-1}Y(X + tY)^{-1}A$$

$$\frac{d}{dt}B^T(X + tY)^{-T}A = -B^T(X + tY)^{-T}Y^T(X + tY)^{-T}A$$

$$\frac{d}{dt}B^T(X + tY)^\mu A = \dots, \quad -1 \leq \mu \leq 1, \quad X, Y \in \mathbb{S}_+^M$$

$$\frac{d^2}{dt^2}B^T(X + tY)^{-1}A = -2B^T(X + tY)^{-1}Y(X + tY)^{-1}Y(X + tY)^{-1}A$$

$$\frac{d^2}{dt^2}B^T(X + tY)^{-1}A = -6B^T(X + tY)^{-1}Y(X + tY)^{-1}Y(X + tY)^{-1}Y(X + tY)^{-1}A$$

$$\frac{d}{dt}((X + tY)^T A (X + tY)) = Y^T A X + X^T A Y + 2tY^T A Y$$

$$\frac{d^2}{dt^2}((X + tY)^T A (X + tY)) = 2Y^T A Y$$

$$\frac{d}{dt}((X + tY) A (X + tY)) = Y A X + X A Y + 2tY A Y$$

$$\frac{d^2}{dt^2}((X + tY) A (X + tY)) = 2Y A Y$$

**D.2.1.0.1 Exercise.** *Expand these tables.*

Provide unfinished table entries indicated by ... throughout §D.2. ▼

**D.2.1.0.2 Exercise.** *log.* (§D.1.7)

Find the first four terms of the Taylor series expansion for  $\log x$  about  $x = 1$ .

Prove that  $\log x \leq x - 1$ ; alternatively, plot the supporting hyperplane to

the hypograph of  $\log x$  at  $\begin{bmatrix} x \\ \log x \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ . ▼

## D.2.2 trace Kronecker

$$\nabla_{\text{vec } X} \text{tr}(A X B X^T) = \nabla_{\text{vec } X} \text{vec}(X)^T (B^T \otimes A) \text{vec } X = (B \otimes A^T + B^T \otimes A) \text{vec } X$$

$$\nabla_{\text{vec } X}^2 \text{tr}(A X B X^T) = \nabla_{\text{vec } X}^2 \text{vec}(X)^T (B^T \otimes A) \text{vec } X = B \otimes A^T + B^T \otimes A$$

## D.2.3 trace

$\nabla_x \mu x = \mu I$	$\nabla_X \operatorname{tr} \mu X = \nabla_X \mu \operatorname{tr} X = \mu I$
$\nabla_x \mathbf{1}^T \delta(x)^{-1} \mathbf{1} = \frac{d}{dx} x^{-1} = -x^{-2}$	$\nabla_X \operatorname{tr} X^{-1} = -X^{-2T}$
$\nabla_x \mathbf{1}^T \delta(x)^{-1} y = -\delta(x)^{-2} y$	$\nabla_X \operatorname{tr}(X^{-1} Y) = \nabla_X \operatorname{tr}(Y X^{-1}) = -X^{-T} Y^T X^{-T}$
$\frac{d}{dx} x^\mu = \mu x^{\mu-1}$	$\nabla_X \operatorname{tr} X^\mu = \mu X^{\mu-1}, \quad X \in \mathbb{S}^M$
	$\nabla_X \operatorname{tr} X^j = j X^{(j-1)T}$
$\nabla_x (b - a^T x)^{-1} = (b - a^T x)^{-2} a$	$\nabla_X \operatorname{tr}((B - AX)^{-1}) = ((B - AX)^{-2} A)^T$
$\nabla_x (b - a^T x)^\mu = -\mu (b - a^T x)^{\mu-1} a$	
$\nabla_x x^T y = \nabla_x y^T x = y$	$\nabla_X \operatorname{tr}(X^T Y) = \nabla_X \operatorname{tr}(Y X^T) = \nabla_X \operatorname{tr}(Y^T X) = \nabla_X \operatorname{tr}(X Y^T) = Y$
	$\nabla_X \operatorname{tr}(A X B X^T) = \nabla_X \operatorname{tr}(X B X^T A) = A^T X B^T + A X B$
	$\nabla_X \operatorname{tr}(A X B X) = \nabla_X \operatorname{tr}(X B X A) = A^T X^T B^T + B^T X^T A^T$
	$\nabla_X \operatorname{tr}(A X A X A X) = \nabla_X \operatorname{tr}(X A X A X A) = 3(A X A X A)^T$
	$\nabla_X \operatorname{tr}(Y X^k) = \nabla_X \operatorname{tr}(X^k Y) = \sum_{i=0}^{k-1} (X^i Y X^{k-1-i})^T$
	$\nabla_X \operatorname{tr}(Y^T X X^T Y) = \nabla_X \operatorname{tr}(X^T Y Y^T X) = 2 Y Y^T X$
	$\nabla_X \operatorname{tr}(Y^T X^T X Y) = \nabla_X \operatorname{tr}(X Y Y^T X^T) = 2 X Y Y^T$
	$\nabla_X \operatorname{tr}((X + Y)^T (X + Y)) = 2(X + Y) = \nabla_X \ X + Y\ _F^2$
	$\nabla_X \operatorname{tr}((X + Y)(X + Y)) = 2(X + Y)^T$
	$\nabla_X \operatorname{tr}(A^T X B) = \nabla_X \operatorname{tr}(X^T A B^T) = A B^T$
	$\nabla_X \operatorname{tr}(A^T X^{-1} B) = \nabla_X \operatorname{tr}(X^{-T} A B^T) = -X^{-T} A B^T X^{-T}$
	$\nabla_X a^T X b = \nabla_X \operatorname{tr}(b a^T X) = \nabla_X \operatorname{tr}(X b a^T) = a b^T$
	$\nabla_X b^T X^T a = \nabla_X \operatorname{tr}(X^T a b^T) = \nabla_X \operatorname{tr}(a b^T X^T) = a b^T$
	$\nabla_X a^T X^{-1} b = \nabla_X \operatorname{tr}(X^{-T} a b^T) = -X^{-T} a b^T X^{-T}$
	$\nabla_X a^T X^\mu b = \dots$

**trace** continued

$$\frac{d}{dt} \operatorname{tr} g(X+tY) = \operatorname{tr} \frac{d}{dt} g(X+tY) \quad [199, \text{p.491}]$$

$$\frac{d}{dt} \operatorname{tr}(X+tY) = \operatorname{tr} Y$$

$$\frac{d}{dt} \operatorname{tr}^j(X+tY) = j \operatorname{tr}^{j-1}(X+tY) \operatorname{tr} Y$$

$$\frac{d}{dt} \operatorname{tr}(X+tY)^j = j \operatorname{tr}((X+tY)^{j-1} Y) \quad (\forall j)$$

$$\frac{d}{dt} \operatorname{tr}((X+tY)Y) = \operatorname{tr} Y^2$$

$$\frac{d}{dt} \operatorname{tr}((X+tY)^k Y) = \frac{d}{dt} \operatorname{tr}(Y(X+tY)^k) = k \operatorname{tr}((X+tY)^{k-1} Y^2), \quad k \in \{0, 1, 2\}$$

$$\frac{d}{dt} \operatorname{tr}((X+tY)^k Y) = \frac{d}{dt} \operatorname{tr}(Y(X+tY)^k) = \operatorname{tr} \sum_{i=0}^{k-1} (X+tY)^i Y (X+tY)^{k-1-i} Y$$

$$\frac{d}{dt} \operatorname{tr}((X+tY)^{-1} Y) = -\operatorname{tr}((X+tY)^{-1} Y (X+tY)^{-1} Y)$$

$$\frac{d}{dt} \operatorname{tr}(B^T(X+tY)^{-1} A) = -\operatorname{tr}(B^T(X+tY)^{-1} Y (X+tY)^{-1} A)$$

$$\frac{d}{dt} \operatorname{tr}(B^T(X+tY)^{-T} A) = -\operatorname{tr}(B^T(X+tY)^{-T} Y^T (X+tY)^{-T} A)$$

$$\frac{d}{dt} \operatorname{tr}(B^T(X+tY)^{-k} A) = \dots, \quad k > 0$$

$$\frac{d}{dt} \operatorname{tr}(B^T(X+tY)^\mu A) = \dots, \quad -1 \leq \mu \leq 1, \quad X, Y \in \mathbb{S}_+^M$$

$$\frac{d^2}{dt^2} \operatorname{tr}(B^T(X+tY)^{-1} A) = 2 \operatorname{tr}(B^T(X+tY)^{-1} Y (X+tY)^{-1} Y (X+tY)^{-1} A)$$

$$\frac{d}{dt} \operatorname{tr}((X+tY)^T A (X+tY)) = \operatorname{tr}(Y^T A X + X^T A Y + 2t Y^T A Y)$$

$$\frac{d^2}{dt^2} \operatorname{tr}((X+tY)^T A (X+tY)) = 2 \operatorname{tr}(Y^T A Y)$$

$$\frac{d}{dt} \operatorname{tr}((X+tY) A (X+tY)) = \operatorname{tr}(Y A X + X A Y + 2t Y A Y)$$

$$\frac{d^2}{dt^2} \operatorname{tr}((X+tY) A (X+tY)) = 2 \operatorname{tr}(Y A Y)$$

### D.2.4 logarithmic determinant

$x > 0$ ,  $\det X > 0$  on some neighborhood of  $X$ , and  $\det(X + tY) > 0$  on some open interval of  $t$ ; otherwise,  $\log(\cdot)$  would be discontinuous. [80, p.75]

$\frac{d}{dx} \log x = x^{-1}$	$\nabla_X \log \det X = X^{-T}$
	$\nabla_X^2 \log \det(X)_{kl} = \frac{\partial X^{-T}}{\partial X_{kl}} = -(X^{-1} e_k e_l^T X^{-1})^T$ , confer (1777)(1824)
$\frac{d}{dx} \log x^{-1} = -x^{-1}$	$\nabla_X \log \det X^{-1} = -X^{-T}$
$\frac{d}{dx} \log x^\mu = \mu x^{-1}$	$\nabla_X \log \det^\mu X = \mu X^{-T}$
	$\nabla_X \log \det X^\mu = \mu X^{-T}$
	$\nabla_X \log \det X^k = \nabla_X \log \det^k X = kX^{-T}$
	$\nabla_X \log \det^\mu(X + tY) = \mu(X + tY)^{-T}$
$\nabla_x \log(a^T x + b) = a \frac{1}{a^T x + b}$	$\nabla_X \log \det(AX + B) = A^T(AX + B)^{-T}$
	$\nabla_X \log \det(I \pm A^T X A) = \pm A(I \pm A^T X A)^{-T} A^T$
	$\nabla_X \log \det(X + tY)^k = \nabla_X \log \det^k(X + tY) = k(X + tY)^{-T}$
	$\frac{d}{dt} \log \det(X + tY) = \text{tr}((X + tY)^{-1} Y)$
	$\frac{d^2}{dt^2} \log \det(X + tY) = -\text{tr}((X + tY)^{-1} Y (X + tY)^{-1} Y)$
	$\frac{d}{dt} \log \det(X + tY)^{-1} = -\text{tr}((X + tY)^{-1} Y)$
	$\frac{d^2}{dt^2} \log \det(X + tY)^{-1} = \text{tr}((X + tY)^{-1} Y (X + tY)^{-1} Y)$
	$\frac{d}{dt} \log \det(\delta(A(x + ty) + a)^2 + \mu I)$ $= \text{tr}((\delta(A(x + ty) + a)^2 + \mu I)^{-1} 2\delta(A(x + ty) + a)\delta(Ay))$

**D.2.5 determinant**

$$\nabla_X \det X = \nabla_X \det X^T = \det(X)X^{-T}$$

$$\nabla_X \det X^{-1} = -\det(X^{-1})X^{-T} = -\det(X)^{-1}X^{-T}$$

$$\nabla_X \det^\mu X = \mu \det^\mu(X)X^{-T}$$

$$\nabla_X \det X^\mu = \mu \det(X^\mu)X^{-T}$$

$$\nabla_X \det X^k = k \det^{k-1}(X)(\operatorname{tr}(X)I - X^T), \quad X \in \mathbb{R}^{2 \times 2}$$

$$\nabla_X \det X^k = \nabla_X \det^k X = k \det(X^k)X^{-T} = k \det^k(X)X^{-T}$$

$$\nabla_X \det^\mu(X + tY) = \mu \det^\mu(X + tY)(X + tY)^{-T}$$

$$\nabla_X \det(X + tY)^k = \nabla_X \det^k(X + tY) = k \det^k(X + tY)(X + tY)^{-T}$$

$$\frac{d}{dt} \det(X + tY) = \det(X + tY) \operatorname{tr}((X + tY)^{-1}Y)$$

$$\frac{d^2}{dt^2} \det(X + tY) = \det(X + tY)(\operatorname{tr}^2((X + tY)^{-1}Y) - \operatorname{tr}((X + tY)^{-1}Y(X + tY)^{-1}Y))$$

$$\frac{d}{dt} \det(X + tY)^{-1} = -\det(X + tY)^{-1} \operatorname{tr}((X + tY)^{-1}Y)$$

$$\frac{d^2}{dt^2} \det(X + tY)^{-1} = \det(X + tY)^{-1}(\operatorname{tr}^2((X + tY)^{-1}Y) + \operatorname{tr}((X + tY)^{-1}Y(X + tY)^{-1}Y))$$

$$\frac{d}{dt} \det^\mu(X + tY) = \mu \det^\mu(X + tY) \operatorname{tr}((X + tY)^{-1}Y)$$

**D.2.6 logarithmic**

Matrix logarithm.

$$\frac{d}{dt} \log(X + tY)^\mu = \mu Y(X + tY)^{-1} = \mu(X + tY)^{-1}Y, \quad XY = YX$$

$$\frac{d}{dt} \log(I - tY)^\mu = -\mu Y(I - tY)^{-1} = -\mu(I - tY)^{-1}Y \quad [199, \text{p.493}]$$

**D.2.7 exponential**

Matrix exponential. [75, §3.6, §4.5] [325, §5.4]

$$\nabla_X e^{\text{tr}(Y^T X)} = \nabla_X \det e^{Y^T X} = e^{\text{tr}(Y^T X)} Y \quad (\forall X, Y)$$

$$\nabla_X \text{tr} e^{Y X} = e^{Y^T X^T} Y^T = Y^T e^{X^T Y^T}$$

$$\nabla_x \mathbf{1}^T e^{Ax} = A^T e^{Ax}$$

$$\nabla_x \mathbf{1}^T e^{|Ax|} = A^T \delta(\text{sgn}(Ax)) e^{|Ax|} \quad (Ax)_i \neq 0$$

$$\nabla_x \log(\mathbf{1}^T e^x) = \frac{1}{\mathbf{1}^T e^x} e^x$$

$$\nabla_x^2 \log(\mathbf{1}^T e^x) = \frac{1}{\mathbf{1}^T e^x} \left( \delta(e^x) - \frac{1}{\mathbf{1}^T e^x} e^x e^{x^T} \right)$$

$$\nabla_x \prod_{i=1}^k x_i^{\frac{1}{k}} = \frac{1}{k} \left( \prod_{i=1}^k x_i^{\frac{1}{k}} \right) \mathbf{1}/x$$

$$\nabla_x^2 \prod_{i=1}^k x_i^{\frac{1}{k}} = -\frac{1}{k} \left( \prod_{i=1}^k x_i^{\frac{1}{k}} \right) \left( \delta(x)^{-2} - \frac{1}{k} (\mathbf{1}/x)(\mathbf{1}/x)^T \right)$$

$$\frac{d}{dt} e^{tY} = e^{tY} Y = Y e^{tY}$$

$$\frac{d}{dt} e^{X+tY} = e^{X+tY} Y = Y e^{X+tY}, \quad XY = YX$$

$$\frac{d^2}{dt^2} e^{X+tY} = e^{X+tY} Y^2 = Y e^{X+tY} Y = Y^2 e^{X+tY}, \quad XY = YX$$

$$\frac{d^j}{dt^j} e^{\text{tr}(X+tY)} = e^{\text{tr}(X+tY)} \text{tr}^j(Y)$$