Appendix D

Matrix calculus

From too much study, and from extreme passion, cometh madnesse.

−Isaac Newton [168, §5]

D.1 Directional derivative, Taylor series

D.1.1 Gradients

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

$$\nabla f(x) \triangleq \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_K} \end{bmatrix} \in \mathbb{R}^K \tag{1860}$$

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

$$\nabla^2 f(x) \triangleq \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_K} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_K \partial x_1} & \frac{\partial^2 f(x)}{\partial x_K \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_K^2} \end{bmatrix} \in \mathbb{S}^K \tag{1861}$$

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

$$\nabla v(x) \triangleq \begin{bmatrix} \frac{\partial v_1(x)}{\partial x} & \frac{\partial v_2(x)}{\partial x} & \cdots & \frac{\partial v_N(x)}{\partial x} \end{bmatrix} \in \mathbb{R}^N \tag{1862}$$

while the second-order gradient is

$$\nabla^2 v(x) \triangleq \begin{bmatrix} \frac{\partial^2 v_1(x)}{\partial x^2} & \frac{\partial^2 v_2(x)}{\partial x^2} & \cdots & \frac{\partial^2 v_N(x)}{\partial x^2} \end{bmatrix} \in \mathbb{R}^N \tag{1863}$$

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

$$\begin{aligned} \nabla h(x) &\triangleq \begin{bmatrix} \frac{\partial h_1(x)}{\partial x_1} & \frac{\partial h_2(x)}{\partial x_1} & \cdots & \frac{\partial h_N(x)}{\partial x_1} \\ \frac{\partial h_1(x)}{\partial x_2} & \frac{\partial h_2(x)}{\partial x_2} & \cdots & \frac{\partial h_N(x)}{\partial x_2} \\ \vdots & \vdots & & \vdots \\ \frac{\partial h_1(x)}{\partial x_K} & \frac{\partial h_2(x)}{\partial x_K} & \cdots & \frac{\partial h_N(x)}{\partial x_K} \end{bmatrix} \\ &= [\, \nabla h_1(x) \;\; \nabla h_2(x) \;\; \cdots \;\; \nabla h_N(x) \,] \in \mathbb{R}^{K \times N} \end{aligned} \tag{1864}$$

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

$$\begin{aligned} \nabla^2 h(x) &\triangleq \begin{bmatrix} \nabla \frac{\partial h_1(x)}{\partial x_1} & \nabla \frac{\partial h_2(x)}{\partial x_1} & \cdots & \nabla \frac{\partial h_N(x)}{\partial x_1} \\ \nabla \frac{\partial h_1(x)}{\partial x_2} & \nabla \frac{\partial h_2(x)}{\partial x_2} & \cdots & \nabla \frac{\partial h_N(x)}{\partial x_2} \\ \vdots & \vdots & & \vdots \\ \nabla \frac{\partial h_1(x)}{\partial x_K} & \nabla \frac{\partial h_2(x)}{\partial x_K} & \cdots & \nabla \frac{\partial h_N(x)}{\partial x_K} \end{bmatrix} \\ &= \left[\, \nabla^2 h_1(x) \;\; \nabla^2 h_2(x) \;\; \cdots \;\; \nabla^2 h_N(x) \,\right] \in \mathbb{R}^{K \times N \times K} \end{aligned} \tag{1865}$$

where the gradient of each real entry is with respect to vector x as in (1860).

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

$$\begin{aligned} \nabla g(X) &\triangleq \begin{bmatrix} \frac{\partial g(X)}{\partial X_{11}} & \frac{\partial g(X)}{\partial X_{12}} & \cdots & \frac{\partial g(X)}{\partial X_{1L}} \\ \frac{\partial g(X)}{\partial X_{21}} & \frac{\partial g(X)}{\partial X_{22}} & \cdots & \frac{\partial g(X)}{\partial X_{2L}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial g(X)}{\partial X_{K1}} & \frac{\partial g(X)}{\partial X_{K2}} & \cdots & \frac{\partial g(X)}{\partial X_{KL}} \end{bmatrix} \in \mathbb{R}^{K \times L} \\ &= \begin{matrix} \left[\, \nabla_{X(:,1)}\, g(X) \right. & & & \\ & \nabla_{X(:,2)}\, g(X) & & \\ & & \ddots & \\ & & & \left. \nabla_{X(:,L)}\, g(X) \,\right] \end{matrix} \in \mathbb{R}^{K \times 1 \times L} \end{aligned} \tag{1866}$$

where gradient $\nabla_{X(:,i)}$ is with respect to the i^{th} column of X. The strange appearance of (1866) in $\mathbb{R}^{K \times 1 \times L}$ is meant to suggest a third dimension perpendicular to the page (not

[D.1] The word *matrix* comes from the Latin for *womb*; related to the prefix *matri-* derived from *mater* meaning *mother*.

a diagonal matrix). The second-order gradient has representation

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

$$= \left[\begin{array}{llll} \nabla\nabla_{X(:,1)}\, g(X) & & & \\ & \nabla\nabla_{X(:,2)}\, g(X) & & \\ & & \ddots & \\ & & & \nabla\nabla_{X(:,L)}\, g(X) \end{array}\right] \in \mathbb{R}^{K\times 1\times L\times K\times L} \tag{1867}$$

where the gradient ∇ is with respect to matrix X.

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

$$\nabla g(X) \triangleq \left[\begin{array}{llll} \nabla_{X(:,1)}\, g_1(X) & \nabla_{X(:,1)}\, g_2(X) & \cdots & \nabla_{X(:,1)}\, g_N(X) \\ \quad\nabla_{X(:,2)}\, g_1(X) & \quad\nabla_{X(:,2)}\, g_2(X) & \cdots & \quad\nabla_{X(:,2)}\, g_N(X) \\ \qquad\ddots & \qquad\ddots & & \qquad\ddots \\ \qquad\nabla_{X(:,L)}\, g_1(X) & \qquad\nabla_{X(:,L)}\, g_2(X) & \cdots & \qquad\nabla_{X(:,L)}\, g_N(X) \end{array}\right] \tag{1868}$$

$$= [\, \nabla g_1(X) \;\; \nabla g_2(X) \;\; \cdots \;\; \nabla g_N(X)\,] \in \mathbb{R}^{K\times N\times L}$$

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

$$\nabla^2 g(X) \triangleq \left[\begin{array}{llll} \nabla\nabla_{X(:,1)}\, g_1(X) & \nabla\nabla_{X(:,1)}\, g_2(X) & \cdots & \nabla\nabla_{X(:,1)}\, g_N(X) \\ \quad\nabla\nabla_{X(:,2)}\, g_1(X) & \quad\nabla\nabla_{X(:,2)}\, g_2(X) & \cdots & \quad\nabla\nabla_{X(:,2)}\, g_N(X) \\ \qquad\ddots & \qquad\ddots & & \qquad\ddots \\ \qquad\nabla\nabla_{X(:,L)}\, g_1(X) & \qquad\nabla\nabla_{X(:,L)}\, g_2(X) & \cdots & \qquad\nabla\nabla_{X(:,L)}\, g_N(X) \end{array}\right] \tag{1869}$$

$$= \left[\, \nabla^2 g_1(X) \;\; \nabla^2 g_2(X) \;\; \cdots \;\; \nabla^2 g_N(X)\,\right] \in \mathbb{R}^{K\times N\times L\times K\times L}$$

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

$$\nabla g(X) \triangleq \begin{bmatrix} \nabla g_{11}(X) & \nabla g_{12}(X) & \cdots & \nabla g_{1N}(X) \\ \nabla g_{21}(X) & \nabla g_{22}(X) & \cdots & \nabla g_{2N}(X) \\ \vdots & \vdots & & \vdots \\ \nabla g_{M1}(X) & \nabla g_{M2}(X) & \cdots & \nabla g_{MN}(X) \end{bmatrix} \in \mathbb{R}^{M\times N\times K\times L} \tag{1870}$$

while the second-order gradient has a six-dimensional representation

$$\nabla^2 g(X) \triangleq \begin{bmatrix} \nabla^2 g_{11}(X) & \nabla^2 g_{12}(X) & \cdots & \nabla^2 g_{1N}(X) \\ \nabla^2 g_{21}(X) & \nabla^2 g_{22}(X) & \cdots & \nabla^2 g_{2N}(X) \\ \vdots & \vdots & & \vdots \\ \nabla^2 g_{M1}(X) & \nabla^2 g_{M2}(X) & \cdots & \nabla^2 g_{MN}(X) \end{bmatrix} \in \mathbb{R}^{M\times N\times K\times L\times K\times L} \tag{1871}$$

and so on.

D.1.2 Product rules for matrix-functions

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

$$\nabla_X \big(f(X)^{\mathrm T} g(X)\big) = \nabla_X(f)\, g \,+\, \nabla_X(g)\, f \tag{1872}$$

while [54, §8.3] [333]

$$\nabla_X \operatorname{tr}\big(f(X)^{\mathrm T} g(X)\big) \,=\, \nabla_X \Big(\operatorname{tr}\big(f(X)^{\mathrm T} g(Z)\big) \,+\, \operatorname{tr}\big(g(X)\, f(Z)^{\mathrm T}\big)\Big)\Big|_{Z\leftarrow X} \tag{1873}$$

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

$$\nabla_X \big(f(X)^{\mathrm T} g(X)\big) = \nabla_X\, a^{\mathrm T} X^2 b \tag{1874}$$

using the product rule. Formula (1872) calls for

$$\nabla_X\, a^{\mathrm T} X^2 b = \nabla_X (X^{\mathrm T} a)\, Xb \,+\, \nabla_X (Xb)\, X^{\mathrm T} a \tag{1875}$$

Consider the first of the two terms:

$$\begin{aligned} \nabla_X(f)\, g &= \nabla_X (X^{\mathrm T} a)\, Xb \\ &= \big[\, \nabla(X^{\mathrm T} a)_1 \quad \nabla(X^{\mathrm T} a)_2 \,\big]\, Xb \end{aligned} \tag{1876}$$

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

$$\nabla_X (X^{\mathrm T} a)\, Xb \,=\, \left[\begin{array}{cccc} \frac{\partial (X^{\mathrm T}a)_1}{\partial X_{11}} & & \frac{\partial (X^{\mathrm T}a)_2}{\partial X_{11}} & \\ & \frac{\partial (X^{\mathrm T}a)_1}{\partial X_{12}} & & \frac{\partial (X^{\mathrm T}a)_2}{\partial X_{12}} \\ \frac{\partial (X^{\mathrm T}a)_1}{\partial X_{21}} & & \frac{\partial (X^{\mathrm T}a)_2}{\partial X_{21}} & \\ & \frac{\partial (X^{\mathrm T}a)_1}{\partial X_{22}} & & \frac{\partial (X^{\mathrm T}a)_2}{\partial X_{22}} \end{array}\right] \begin{bmatrix} (Xb)_1 \\ \\ (Xb)_2 \end{bmatrix} \in \mathbb{R}^{\mathbf{2\times 1\times 2}} \tag{1877}$$

Because gradient of the product (1874) 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 that second dimension of the cubix indicated by dotted line segments;

$$\begin{aligned}\nabla_X(X^{\mathrm{T}}a)\,Xb &= \begin{bmatrix} a_1 & 0 & \\ & 0 & a_1 \\ a_2 & 0 & \\ & 0 & a_2 \end{bmatrix}\begin{bmatrix} b_1X_{11}+b_2X_{12} \\ b_1X_{21}+b_2X_{22}\end{bmatrix} \in \mathbb{R}^{\mathbf{2}\times 1\times\mathbf{2}} \\ &= \begin{bmatrix} a_1(b_1X_{11}+b_2X_{12}) & a_1(b_1X_{21}+b_2X_{22}) \\ a_2(b_1X_{11}+b_2X_{12}) & a_2(b_1X_{21}+b_2X_{22})\end{bmatrix} \in \mathbb{R}^{\mathbf{2}\times\mathbf{2}} \\ &= ab^{\mathrm{T}}X^{\mathrm{T}}\end{aligned} \tag{1878}$$

where the cubix appears as a complete $2\times2\times2$ matrix. In like manner for the second term $\nabla_X(g)\,f$

$$\begin{aligned}\nabla_X(Xb)\,X^{\mathrm{T}}a &= \begin{bmatrix} b_1 & 0 & \\ & b_2 & 0 \\ 0 & b_1 & \\ & 0 & b_2 \end{bmatrix}\begin{bmatrix} X_{11}a_1+X_{21}a_2 \\ X_{12}a_1+X_{22}a_2\end{bmatrix} \in \mathbb{R}^{\mathbf{2}\times 1\times\mathbf{2}} \\ &= X^{\mathrm{T}}ab^{\mathrm{T}} \in \mathbb{R}^{\mathbf{2}\times\mathbf{2}}\end{aligned} \tag{1879}$$

The solution

$$\nabla_X\, a^{\mathrm{T}}X^2b = ab^{\mathrm{T}}X^{\mathrm{T}} + X^{\mathrm{T}}ab^{\mathrm{T}} \tag{1880}$$

can be found from Table **D.2.1** or verified using (1873). □

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,` *tensor product.* Although it sees reversal in the literature, [344, §2.1] we adopt the definition: for $A\in\mathbb{R}^{m\times n}$ and $B\in\mathbb{R}^{p\times q}$

$$B\otimes A \triangleq \begin{bmatrix} B_{11}A & B_{12}A & \cdots & B_{1q}A \\ B_{21}A & B_{22}A & \cdots & B_{2q}A \\ \vdots & \vdots & & \vdots \\ B_{p1}A & B_{p2}A & \cdots & B_{pq}A \end{bmatrix} \in \mathbb{R}^{pm\times qn} \tag{1881}$$

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

One advantage to vectorization is existence of the 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 $A\,,B\in\mathbb{R}^{n\times n}$ [182, §5.2] [13, §3]

$$\nabla^2_{\mathrm{vec}\,X}\,\mathrm{tr}(AXBX^{\mathrm{T}}) \;=\; \nabla^2_{\mathrm{vec}\,X}\,\mathrm{vec}(X)^{\mathrm{T}}(B^{\mathrm{T}}\otimes A)\,\mathrm{vec}\,X \;=\; B\otimes A^{\mathrm{T}} + B^{\mathrm{T}}\otimes A \in \mathbb{R}^{n^{\mathbf{2}}\times n^{\mathbf{2}}} \tag{1882}$$

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

$$W \triangleq \begin{bmatrix} w_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0}^{\mathrm{T}} & & w_N \end{bmatrix} \in \mathbb{S}^N \tag{1883}$$

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

$$SW = S\begin{bmatrix} w_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0}^{\mathrm{T}} & & w_N \end{bmatrix} = \begin{bmatrix} S(:,1)w_1 & \cdots & S(:,N)w_N \end{bmatrix} \in \mathbb{R}^{M\times N} \tag{1884}$$

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

$$SW = \left(\delta(W)^{\mathrm{T}} \otimes I\right)\begin{bmatrix} S(:,1) & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & S(:,2) & \ddots & \\ & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & & \mathbf{0} & S(:,N) \end{bmatrix} \in \mathbb{R}^{M\times N} \tag{1885}$$

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

$$WS = \begin{bmatrix} S(1,:) & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & S(2,:) & \ddots & \\ & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & & \mathbf{0} & S(M,:) \end{bmatrix}\left(\delta(W) \otimes I\right) \in \mathbb{R}^{M\times N} \tag{1886}$$

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

$$S\circ Y = \begin{bmatrix} \delta(Y(:,1)) & \cdots & \delta(Y(:,N)) \end{bmatrix}\begin{bmatrix} S(:,1) & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & S(:,2) & \ddots & \\ & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & & \mathbf{0} & S(:,N) \end{bmatrix} \in \mathbb{R}^{M\times N} \tag{1887}$$

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

$$s\circ y = \delta(s)y \tag{1888}$$

$$\begin{aligned} s^{\mathrm{T}}\otimes y &= ys^{\mathrm{T}} \\ s\otimes y^{\mathrm{T}} &= sy^{\mathrm{T}} \end{aligned} \tag{1889}$$

D.1.3 Chain rules for composite matrix-functions

Given dimensionally compatible matrix-valued functions of matrix variable $f(X)$ and $g(X)$ [235, §15.7]

$$\nabla_X\, g\big(f(X)^{\mathrm{T}}\big) = \nabla_X f^{\mathrm{T}}\, \nabla_f\, g \tag{1890}$$

$$\nabla_X^2\, g\big(f(X)^{\mathrm{T}}\big) \;=\; \nabla_X\big(\nabla_X f^{\mathrm{T}}\, \nabla_f\, g\big) \;=\; \nabla_X^2 f\, \nabla_f\, g \;+\; \nabla_X f^{\mathrm{T}}\, \nabla_f^2 g\, \nabla_X f \tag{1891}$$

D.1.3.1 Two arguments

$$\nabla_X\, g\big(f(X)^{\mathrm{T}},\, h(X)^{\mathrm{T}}\big) = \nabla_X f^{\mathrm{T}}\, \nabla_f\, g \;+\; \nabla_X h^{\mathrm{T}}\, \nabla_h\, g \tag{1892}$$

D.1.3.1.1 Example. *Chain rule for two arguments.* [43, §1.1]

$$g\big(f(x)^{\mathrm{T}},\, h(x)^{\mathrm{T}}\big) = \big(f(x) + h(x)\big)^{\mathrm{T}} A\big(f(x) + h(x)\big) \tag{1893}$$

$$f(x) = \begin{bmatrix} x_1 \\ \varepsilon x_2 \end{bmatrix}, \qquad h(x) = \begin{bmatrix} \varepsilon x_1 \\ x_2 \end{bmatrix} \tag{1894}$$

$$\nabla_x\, g\big(f(x)^{\mathrm{T}},\, h(x)^{\mathrm{T}}\big) = \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \end{bmatrix}(A + A^{\mathrm{T}})(f + h) + \begin{bmatrix} \varepsilon & 0 \\ 0 & 1 \end{bmatrix}(A + A^{\mathrm{T}})(f + h) \tag{1895}$$

$$\nabla_x\, g\big(f(x)^{\mathrm{T}},\, h(x)^{\mathrm{T}}\big) = \begin{bmatrix} 1+\varepsilon & 0 \\ 0 & 1+\varepsilon \end{bmatrix}(A + A^{\mathrm{T}})\left(\begin{bmatrix} x_1 \\ \varepsilon x_2 \end{bmatrix} + \begin{bmatrix} \varepsilon x_1 \\ x_2 \end{bmatrix}\right) \tag{1896}$$

$$\lim_{\varepsilon \to 0} \nabla_x\, g\big(f(x)^{\mathrm{T}},\, h(x)^{\mathrm{T}}\big) = (A + A^{\mathrm{T}})x \tag{1897}$$

from Table **D.2.1**. □

These foregoing formulae remain correct when gradient produces hyperdimensional representation:

D.1.4 First directional derivative

Assume that a differentiable function $g(X) : \mathbb{R}^{K\times L} \to \mathbb{R}^{M\times N}$ has continuous first- and second-order gradients ∇g and $\nabla^2 g$ over $\operatorname{dom} g$ 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, $\overset{\to Y}{dg} \in \mathbb{R}^{M\times N}$ and $\overset{\to Y}{dg^2} \in \mathbb{R}^{M\times N}$.

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

$$\frac{\partial g_{mn}(X)}{\partial X_{kl}} \;=\; \lim_{\Delta t \to 0} \frac{g_{mn}(X + \Delta t\, e_k e_l^{\mathrm{T}}) - g_{mn}(X)}{\Delta t} \in \mathbb{R} \tag{1898}$$

where e_k is the k^{th} standard basis vector in $\mathbb{R}^K$ while e_l is the l^{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^{th} entry of the gradient is

$$\nabla g_{mn}(X) = \begin{bmatrix} \frac{\partial g_{mn}(X)}{\partial X_{11}} & \frac{\partial g_{mn}(X)}{\partial X_{12}} & \cdots & \frac{\partial g_{mn}(X)}{\partial X_{1L}} \\ \frac{\partial g_{mn}(X)}{\partial X_{21}} & \frac{\partial g_{mn}(X)}{\partial X_{22}} & \cdots & \frac{\partial g_{mn}(X)}{\partial X_{2L}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial g_{mn}(X)}{\partial X_{K1}} & \frac{\partial g_{mn}(X)}{\partial X_{K2}} & \cdots & \frac{\partial g_{mn}(X)}{\partial X_{KL}} \end{bmatrix} \in \mathbb{R}^{K\times L} \tag{1899}$$

while the gradient is a quartix

$$\nabla g(X) = \begin{bmatrix} \nabla g_{11}(X) & \nabla g_{12}(X) & \cdots & \nabla g_{1N}(X) \\ \nabla g_{21}(X) & \nabla g_{22}(X) & \cdots & \nabla g_{2N}(X) \\ \vdots & \vdots & & \vdots \\ \nabla g_{M1}(X) & \nabla g_{M2}(X) & \cdots & \nabla g_{MN}(X) \end{bmatrix} \in \mathbb{R}^{M\times N\times K\times L} \tag{1900}$$

By simply rotating our perspective of a four-dimensional representation of gradient matrix, we find one of three useful transpositions of this quartix (connoted $^{\mathrm{T}_1}$):

$$\nabla g(X)^{\mathrm{T}_1} = \begin{bmatrix} \frac{\partial g(X)}{\partial X_{11}} & \frac{\partial g(X)}{\partial X_{12}} & \cdots & \frac{\partial g(X)}{\partial X_{1L}} \\ \frac{\partial g(X)}{\partial X_{21}} & \frac{\partial g(X)}{\partial X_{22}} & \cdots & \frac{\partial g(X)}{\partial X_{2L}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial g(X)}{\partial X_{K1}} & \frac{\partial g(X)}{\partial X_{K2}} & \cdots & \frac{\partial g(X)}{\partial X_{KL}} \end{bmatrix} \in \mathbb{R}^{K\times L\times M\times N} \tag{1901}$$

When the limit for $\Delta t\in\mathbb{R}$ exists, it is easy to show by substitution of variables in (1898)

$$\frac{\partial g_{mn}(X)}{\partial X_{kl}} Y_{kl} = \lim_{\Delta t\to 0} \frac{g_{mn}(X + \Delta t\, Y_{kl}\, e_k e_l^{\mathrm{T}}) - g_{mn}(X)}{\Delta t} \in \mathbb{R} \tag{1902}$$

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^{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^{th} entry of the directional derivative is the corresponding total differential [235, §15.8]

$$\begin{aligned} dg_{mn}(X)|_{dX\to Y} &= \sum_{k,l} \frac{\partial g_{mn}(X)}{\partial X_{kl}} Y_{kl} = \operatorname{tr}\big(\nabla g_{mn}(X)^{\mathrm{T}} Y\big) & (1903)\\ &= \sum_{k,l} \lim_{\Delta t\to 0} \frac{g_{mn}(X + \Delta t\, Y_{kl}\, e_k e_l^{\mathrm{T}}) - g_{mn}(X)}{\Delta t} & (1904)\\ &= \lim_{\Delta t\to 0} \frac{g_{mn}(X + \Delta t\, Y) - g_{mn}(X)}{\Delta t} & (1905)\\ &= \left.\frac{d}{dt}\right|_{t=0} g_{mn}(X + t\,Y) & (1906) \end{aligned}$$

where $t\in\mathbb{R}$. Assuming finite Y, equation (1905) is called the *Gâteaux differential* [42, App.A.5] [215, §D.2.1] [377, §5.28] whose existence is implied by existence of the *Fréchet differential* (the sum in (1903)). [266, §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.[D.2] Hence the directional derivative,

$$\begin{aligned}
\overset{\to Y}{dg}(X) &\triangleq \left.\begin{bmatrix} dg_{11}(X) & dg_{12}(X) & \cdots & dg_{1N}(X) \\ dg_{21}(X) & dg_{22}(X) & \cdots & dg_{2N}(X) \\ \vdots & \vdots & & \vdots \\ dg_{M1}(X) & dg_{M2}(X) & \cdots & dg_{MN}(X) \end{bmatrix}\right|_{dX\to Y} \in \mathbb{R}^{M\times N} \\
&= \begin{bmatrix} \operatorname{tr}\big(\nabla g_{11}(X)^{\mathrm T} Y\big) & \operatorname{tr}\big(\nabla g_{12}(X)^{\mathrm T} Y\big) & \cdots & \operatorname{tr}\big(\nabla g_{1N}(X)^{\mathrm T} Y\big) \\ \operatorname{tr}\big(\nabla g_{21}(X)^{\mathrm T} Y\big) & \operatorname{tr}\big(\nabla g_{22}(X)^{\mathrm T} Y\big) & \cdots & \operatorname{tr}\big(\nabla g_{2N}(X)^{\mathrm T} Y\big) \\ \vdots & \vdots & & \vdots \\ \operatorname{tr}\big(\nabla g_{M1}(X)^{\mathrm T} Y\big) & \operatorname{tr}\big(\nabla g_{M2}(X)^{\mathrm T} Y\big) & \cdots & \operatorname{tr}\big(\nabla g_{MN}(X)^{\mathrm T} Y\big) \end{bmatrix} \qquad (1907) \\
&= \begin{bmatrix} \sum\limits_{k,l} \frac{\partial g_{11}(X)}{\partial X_{kl}} Y_{kl} & \sum\limits_{k,l} \frac{\partial g_{12}(X)}{\partial X_{kl}} Y_{kl} & \cdots & \sum\limits_{k,l} \frac{\partial g_{1N}(X)}{\partial X_{kl}} Y_{kl} \\ \sum\limits_{k,l} \frac{\partial g_{21}(X)}{\partial X_{kl}} Y_{kl} & \sum\limits_{k,l} \frac{\partial g_{22}(X)}{\partial X_{kl}} Y_{kl} & \cdots & \sum\limits_{k,l} \frac{\partial g_{2N}(X)}{\partial X_{kl}} Y_{kl} \\ \vdots & \vdots & & \vdots \\ \sum\limits_{k,l} \frac{\partial g_{M1}(X)}{\partial X_{kl}} Y_{kl} & \sum\limits_{k,l} \frac{\partial g_{M2}(X)}{\partial X_{kl}} Y_{kl} & \cdots & \sum\limits_{k,l} \frac{\partial g_{MN}(X)}{\partial X_{kl}} Y_{kl} \end{bmatrix}
\end{aligned}$$

from which it follows

$$\overset{\to Y}{dg}(X) = \sum_{k,l} \frac{\partial g(X)}{\partial X_{kl}} Y_{kl} \qquad (1908)$$

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

$$g(X+t\,Y) = g(X) \;+\; t\,\overset{\to Y}{dg}(X) \;+\; o(t^2) \qquad (1909)$$

which is the first-order Taylor series expansion about X. [235, §18.4] [166, §2.3.4] Differentiation with respect to t and subsequent t-zeroing isolates the second term of expansion. Thus differentiating and zeroing $g(X+t\,Y)$ in t is an operation equivalent to individually differentiating and zeroing every entry $g_{mn}(X+t\,Y)$ as in (1906). So the directional derivative of $g(X):\mathbb{R}^{K\times L}\to\mathbb{R}^{M\times N}$ in any direction $Y\in\mathbb{R}^{K\times L}$ evaluated at $X\in\operatorname{dom} g$ becomes

$$\overset{\to Y}{dg}(X) = \left.\frac{d}{dt}\right|_{t=0} g(X+t\,Y) \in \mathbb{R}^{M\times N} \qquad (1910)$$

[294, §2.1, §5.4.5] [35, §6.3.1] which is simplest. In case of a real function $g(X):\mathbb{R}^{K\times L}\to\mathbb{R}$

$$\overset{\to Y}{dg}(X) = \operatorname{tr}\big(\nabla g(X)^{\mathrm T} Y\big) \qquad (1932)$$

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

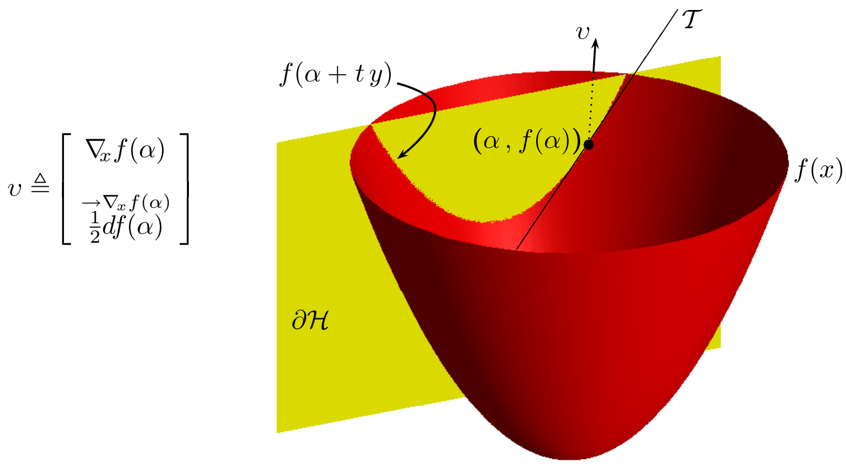

Figure 174: Strictly convex quadratic bowl in $\mathbb{R}^{\mathbf{2}}\times\mathbb{R}$; $f(x)=x^{\mathrm{T}}x : \mathbb{R}^{\mathbf{2}}\to\mathbb{R}$ *versus* x on some open disc in $\mathbb{R}^{\mathbf{2}}$. Plane slice $\partial\mathcal{H}$ is perpendicular to function domain. Slice intersection with domain connotes bidirectional vector y. Slope of tangent line $\mathcal{T}$ at point $(\alpha\,,\,f(\alpha))$ is value of directional derivative $\nabla_x f(\alpha)^{\mathrm{T}}y$ (1935) at α in slice direction y. Negative gradient $-\nabla_x f(x)\in\mathbb{R}^{\mathbf{2}}$ is direction of *steepest descent.* [414] [235, §15.6] [166] When vector $\upsilon\in\mathbb{R}^{\mathbf{3}}$ entry υ_3 is half directional derivative in gradient direction at α and when $\begin{bmatrix} \upsilon_1 \\ \upsilon_2 \end{bmatrix} = \nabla_x f(\alpha)$, then $-\upsilon$ points directly toward bowl bottom.

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

$$\overset{\to Y}{dg}(X) = \nabla g(X)^{\mathrm{T}}Y \tag{1935}$$

Unlike gradient, directional derivative does not expand dimension; directional derivative (1910) 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, $\overset{\to Y}{dg}(X) = g(Y)$. [266, §7.2]

D.1.4.1 Interpretation of directional derivative

In the case of any differentiable real function $g(X) : \mathbb{R}^{K\times L}\to\mathbb{R}$, the directional derivative of $g(X)$ at X in any direction Y yields the slope of g along the line $\{X+t\,Y \mid t\in\mathbb{R}\}$ through its domain evaluated at $t=0$. For higher-dimensional functions, by (1907), this slope interpretation can be applied to each entry of the directional derivative.

Figure **174**, for example, shows a plane slice of a real convex bowl-shaped function $f(x)$ along a line $\{\alpha + t\,y \mid t\in\mathbb{R}\}$ through its domain. The slice reveals a one-dimensional real function of t ; $f(\alpha + t\,y)$. The directional derivative at $x=\alpha$ in direction y is the slope of $f(\alpha + t\,y)$ with respect to t at $t=0$. In the case of a real function having vector argument $h(X) : \mathbb{R}^K\to\mathbb{R}$, its directional derivative in the normalized direction of its gradient is the gradient magnitude. (1935) 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* §3.6)

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 $e_k e_l^{\mathrm{T}}$ (60), then a single partial derivative $\partial g(X)/\partial X_{kl}$ is obtained from directional derivative (1908); such is each entry of gradient $\nabla g(X)$ in equalities (1932) and (1935), for example.

D.1.4.1.1 Theorem. *Directional derivative optimality condition.* [266, §7.4]
Suppose $f(X):\mathbb{R}^{K\times L}\to\mathbb{R}$ is minimized on convex set $\mathcal{C}\subseteq\mathbb{R}^{K\times L}$ by $X^\star$, and the directional derivative of f exists there. Then for all $X\in\mathcal{C}$

$$\overset{\to X-X^\star}{df(X)} \geq 0 \tag{1911}$$

◇

D.1.4.1.2 Example. *Simple bowl.*
Bowl function (Figure **174**)

$$f(x):\mathbb{R}^K\to\mathbb{R} \triangleq (x-a)^{\mathrm{T}}(x-a)-b \tag{1912}$$

has function offset $-b\in\mathbb{R}$, axis of revolution at $x=a$, and positive definite Hessian (1861) 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 $-\upsilon$ based anywhere in $\operatorname{dom} f\times\mathbb{R}$ pointing toward the unique bowl-bottom is specified:

$$\upsilon \propto \begin{bmatrix} x-a \\ f(x)+b \end{bmatrix} \in \mathbb{R}^K\times\mathbb{R} \tag{1913}$$

Such a vector is

$$\upsilon = \begin{bmatrix} \nabla_x f(x) \\ \\ \frac{1}{2}\overset{\to\nabla_x f(x)}{df(x)} \end{bmatrix} \tag{1914}$$

since the gradient is

$$\nabla_x f(x) = 2(x-a) \tag{1915}$$

and the directional derivative in direction of the gradient is (1935)

$$\overset{\to\nabla_x f(x)}{df(x)} = \nabla_x f(x)^{\mathrm{T}}\,\nabla_x f(x) = 4(x-a)^{\mathrm{T}}(x-a) = 4(f(x)+b) \tag{1916}$$

□

D.1.5 Second directional derivative

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

$$\nabla \frac{\partial g_{mn}(X)}{\partial X_{kl}} = \frac{\partial \nabla g_{mn}(X)}{\partial X_{kl}} = \begin{bmatrix} \frac{\partial^2 g_{mn}(X)}{\partial X_{kl}\partial X_{11}} & \frac{\partial^2 g_{mn}(X)}{\partial X_{kl}\partial X_{12}} & \cdots & \frac{\partial^2 g_{mn}(X)}{\partial X_{kl}\partial X_{1L}} \\ \frac{\partial^2 g_{mn}(X)}{\partial X_{kl}\partial X_{21}} & \frac{\partial^2 g_{mn}(X)}{\partial X_{kl}\partial X_{22}} & \cdots & \frac{\partial^2 g_{mn}(X)}{\partial X_{kl}\partial X_{2L}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 g_{mn}(X)}{\partial X_{kl}\partial X_{K1}} & \frac{\partial^2 g_{mn}(X)}{\partial X_{kl}\partial X_{K2}} & \cdots & \frac{\partial^2 g_{mn}(X)}{\partial X_{kl}\partial X_{KL}} \end{bmatrix} \in \mathbb{R}^{K\times L} \tag{1917}$$

$$\begin{aligned} \nabla^2 g_{mn}(X) &= \begin{bmatrix} \nabla \frac{\partial g_{mn}(X)}{\partial X_{11}} & \nabla \frac{\partial g_{mn}(X)}{\partial X_{12}} & \cdots & \nabla \frac{\partial g_{mn}(X)}{\partial X_{1L}} \\ \nabla \frac{\partial g_{mn}(X)}{\partial X_{21}} & \nabla \frac{\partial g_{mn}(X)}{\partial X_{22}} & \cdots & \nabla \frac{\partial g_{mn}(X)}{\partial X_{2L}} \\ \vdots & \vdots & & \vdots \\ \nabla \frac{\partial g_{mn}(X)}{\partial X_{K1}} & \nabla \frac{\partial g_{mn}(X)}{\partial X_{K2}} & \cdots & \nabla \frac{\partial g_{mn}(X)}{\partial X_{KL}} \end{bmatrix} \in \mathbb{R}^{K\times L\times K\times L} \\ &= \begin{bmatrix} \frac{\partial \nabla g_{mn}(X)}{\partial X_{11}} & \frac{\partial \nabla g_{mn}(X)}{\partial X_{12}} & \cdots & \frac{\partial \nabla g_{mn}(X)}{\partial X_{1L}} \\ \frac{\partial \nabla g_{mn}(X)}{\partial X_{21}} & \frac{\partial \nabla g_{mn}(X)}{\partial X_{22}} & \cdots & \frac{\partial \nabla g_{mn}(X)}{\partial X_{2L}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial \nabla g_{mn}(X)}{\partial X_{K1}} & \frac{\partial \nabla g_{mn}(X)}{\partial X_{K2}} & \cdots & \frac{\partial \nabla g_{mn}(X)}{\partial X_{KL}} \end{bmatrix} \end{aligned} \tag{1918}$$

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

$$\nabla^2 g(X) = \begin{bmatrix} \nabla^2 g_{11}(X) & \nabla^2 g_{12}(X) & \cdots & \nabla^2 g_{1N}(X) \\ \nabla^2 g_{21}(X) & \nabla^2 g_{22}(X) & \cdots & \nabla^2 g_{2N}(X) \\ \vdots & \vdots & & \vdots \\ \nabla^2 g_{M1}(X) & \nabla^2 g_{M2}(X) & \cdots & \nabla^2 g_{MN}(X) \end{bmatrix} \in \mathbb{R}^{M\times N\times K\times L\times K\times L} \tag{1919}$$

$$\nabla^2 g(X)^{\mathrm{T}_1} = \begin{bmatrix} \nabla \frac{\partial g(X)}{\partial X_{11}} & \nabla \frac{\partial g(X)}{\partial X_{12}} & \cdots & \nabla \frac{\partial g(X)}{\partial X_{1L}} \\ \nabla \frac{\partial g(X)}{\partial X_{21}} & \nabla \frac{\partial g(X)}{\partial X_{22}} & \cdots & \nabla \frac{\partial g(X)}{\partial X_{2L}} \\ \vdots & \vdots & & \vdots \\ \nabla \frac{\partial g(X)}{\partial X_{K1}} & \nabla \frac{\partial g(X)}{\partial X_{K2}} & \cdots & \nabla \frac{\partial g(X)}{\partial X_{KL}} \end{bmatrix} \in \mathbb{R}^{K\times L\times M\times N\times K\times L} \tag{1920}$$

$$\nabla^2 g(X)^{\mathrm{T}_2} = \begin{bmatrix} \frac{\partial \nabla g(X)}{\partial X_{11}} & \frac{\partial \nabla g(X)}{\partial X_{12}} & \cdots & \frac{\partial \nabla g(X)}{\partial X_{1L}} \\ \frac{\partial \nabla g(X)}{\partial X_{21}} & \frac{\partial \nabla g(X)}{\partial X_{22}} & \cdots & \frac{\partial \nabla g(X)}{\partial X_{2L}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial \nabla g(X)}{\partial X_{K1}} & \frac{\partial \nabla g(X)}{\partial X_{K2}} & \cdots & \frac{\partial \nabla g(X)}{\partial X_{KL}} \end{bmatrix} \in \mathbb{R}^{K\times L\times K\times L\times M\times N} \tag{1921}$$

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

$$\frac{\partial^2 g_{mn}(X)}{\partial X_{kl}\,\partial X_{ij}} = \lim_{\Delta\tau,\Delta t\to 0} \frac{g_{mn}(X+\Delta t\, e_k e_l^{\mathrm{T}}+\Delta\tau\, e_i e_j^{\mathrm{T}})-g_{mn}(X+\Delta t\, e_k e_l^{\mathrm{T}})-\big(g_{mn}(X+\Delta\tau\, e_i e_j^{\mathrm{T}})-g_{mn}(X)\big)}{\Delta\tau\,\Delta t} \tag{1922}$$

Differentiating (1902) and then scaling by Y_{ij}

$$\begin{aligned}
\frac{\partial^2 g_{mn}(X)}{\partial X_{kl}\,\partial X_{ij}} Y_{kl} Y_{ij} &= \lim_{\Delta t\to 0} \frac{\partial g_{mn}(X+\Delta t\, Y_{kl}\, e_k e_l^{\mathrm{T}})-\partial g_{mn}(X)}{\partial X_{ij}\,\Delta t} Y_{ij} &(1923)\\
= \lim_{\Delta\tau,\Delta t\to 0} &\frac{g_{mn}(X+\Delta t\, Y_{kl}\, e_k e_l^{\mathrm{T}}+\Delta\tau\, Y_{ij}\, e_i e_j^{\mathrm{T}})-g_{mn}(X+\Delta t\, Y_{kl}\, e_k e_l^{\mathrm{T}})-\big(g_{mn}(X+\Delta\tau\, Y_{ij}\, e_i e_j^{\mathrm{T}})-g_{mn}(X)\big)}{\Delta\tau\,\Delta t}
\end{aligned}$$

which can be proved by substitution of variables in (1922). The mn^{th} second-order total differential due to any $Y\in\mathbb{R}^{K\times L}$ is

$$\begin{aligned}
d^2 g_{mn}(X)|_{dX\to Y} &= \sum_{i,j}\sum_{k,l} \frac{\partial^2 g_{mn}(X)}{\partial X_{kl}\,\partial X_{ij}} Y_{kl} Y_{ij} = \operatorname{tr}\Big(\nabla_X \operatorname{tr}\big(\nabla g_{mn}(X)^{\mathrm{T}} Y\big)^{\mathrm{T}} Y\Big) &(1924)\\
&= \sum_{i,j} \lim_{\Delta t\to 0} \frac{\partial g_{mn}(X+\Delta t\, Y)-\partial g_{mn}(X)}{\partial X_{ij}\,\Delta t} Y_{ij} &(1925)\\
&= \lim_{\Delta t\to 0} \frac{g_{mn}(X+2\Delta t\, Y)-2g_{mn}(X+\Delta t\, Y)+g_{mn}(X)}{\Delta t^2} &(1926)\\
&= \left.\frac{d^2}{dt^2}\right|_{t=0} g_{mn}(X+t\,Y) &(1927)
\end{aligned}$$

Hence the second directional derivative,

$$\overset{\to Y}{dg^2}(X) \triangleq \left.\begin{bmatrix} d^2 g_{11}(X) & d^2 g_{12}(X) & \cdots & d^2 g_{1N}(X)\\ d^2 g_{21}(X) & d^2 g_{22}(X) & \cdots & d^2 g_{2N}(X)\\ \vdots & \vdots & & \vdots\\ d^2 g_{M1}(X) & d^2 g_{M2}(X) & \cdots & d^2 g_{MN}(X)\end{bmatrix}\right|_{dX\to Y} \in \mathbb{R}^{M\times N}$$

$$= \begin{bmatrix} \operatorname{tr}\Big(\nabla \operatorname{tr}\big(\nabla g_{11}(X)^{\mathrm{T}} Y\big)^{\mathrm{T}} Y\Big) & \operatorname{tr}\Big(\nabla \operatorname{tr}\big(\nabla g_{12}(X)^{\mathrm{T}} Y\big)^{\mathrm{T}} Y\Big) & \cdots & \operatorname{tr}\Big(\nabla \operatorname{tr}\big(\nabla g_{1N}(X)^{\mathrm{T}} Y\big)^{\mathrm{T}} Y\Big)\\ \operatorname{tr}\Big(\nabla \operatorname{tr}\big(\nabla g_{21}(X)^{\mathrm{T}} Y\big)^{\mathrm{T}} Y\Big) & \operatorname{tr}\Big(\nabla \operatorname{tr}\big(\nabla g_{22}(X)^{\mathrm{T}} Y\big)^{\mathrm{T}} Y\Big) & \cdots & \operatorname{tr}\Big(\nabla \operatorname{tr}\big(\nabla g_{2N}(X)^{\mathrm{T}} Y\big)^{\mathrm{T}} Y\Big)\\ \vdots & \vdots & & \vdots\\ \operatorname{tr}\Big(\nabla \operatorname{tr}\big(\nabla g_{M1}(X)^{\mathrm{T}} Y\big)^{\mathrm{T}} Y\Big) & \operatorname{tr}\Big(\nabla \operatorname{tr}\big(\nabla g_{M2}(X)^{\mathrm{T}} Y\big)^{\mathrm{T}} Y\Big) & \cdots & \operatorname{tr}\Big(\nabla \operatorname{tr}\big(\nabla g_{MN}(X)^{\mathrm{T}} Y\big)^{\mathrm{T}} Y\Big)\end{bmatrix}$$

$$= \begin{bmatrix} \sum\limits_{i,j}\sum\limits_{k,l} \frac{\partial^2 g_{11}(X)}{\partial X_{kl}\,\partial X_{ij}} Y_{kl} Y_{ij} & \sum\limits_{i,j}\sum\limits_{k,l} \frac{\partial^2 g_{12}(X)}{\partial X_{kl}\,\partial X_{ij}} Y_{kl} Y_{ij} & \cdots & \sum\limits_{i,j}\sum\limits_{k,l} \frac{\partial^2 g_{1N}(X)}{\partial X_{kl}\,\partial X_{ij}} Y_{kl} Y_{ij}\\ \sum\limits_{i,j}\sum\limits_{k,l} \frac{\partial^2 g_{21}(X)}{\partial X_{kl}\,\partial X_{ij}} Y_{kl} Y_{ij} & \sum\limits_{i,j}\sum\limits_{k,l} \frac{\partial^2 g_{22}(X)}{\partial X_{kl}\,\partial X_{ij}} Y_{kl} Y_{ij} & \cdots & \sum\limits_{i,j}\sum\limits_{k,l} \frac{\partial^2 g_{2N}(X)}{\partial X_{kl}\,\partial X_{ij}} Y_{kl} Y_{ij}\\ \vdots & \vdots & & \vdots\\ \sum\limits_{i,j}\sum\limits_{k,l} \frac{\partial^2 g_{M1}(X)}{\partial X_{kl}\,\partial X_{ij}} Y_{kl} Y_{ij} & \sum\limits_{i,j}\sum\limits_{k,l} \frac{\partial^2 g_{M2}(X)}{\partial X_{kl}\,\partial X_{ij}} Y_{kl} Y_{ij} & \cdots & \sum\limits_{i,j}\sum\limits_{k,l} \frac{\partial^2 g_{MN}(X)}{\partial X_{kl}\,\partial X_{ij}} Y_{kl} Y_{ij}\end{bmatrix} \tag{1928}$$

from which it follows

$$\overset{\to Y}{dg^2}(X) = \sum_{i,j}\sum_{k,l}\frac{\partial^2 g(X)}{\partial X_{kl}\,\partial X_{ij}}Y_{kl}Y_{ij} = \sum_{i,j}\frac{\partial}{\partial X_{ij}}\overset{\to Y}{dg}(X)\,Y_{ij} \tag{1929}$$

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

$$g(X+t\,Y) = g(X) \;+\; t\,\overset{\to Y}{dg}(X) \;+\; \frac{1}{2!}t^2\,\overset{\to Y}{dg^2}(X) \;+\; o(t^3) \tag{1930}$$

which is the second-order Taylor series expansion about X. [235, §18.4] [166, §2.3.4] Differentiating twice with respect to t and subsequent t-zeroing isolates the third term of the expansion. Thus differentiating and zeroing $g(X+t\,Y)$ in t is an operation equivalent to individually differentiating and zeroing every entry $g_{mn}(X+t\,Y)$ as in (1927). So the second directional derivative of $g(X):\mathbb{R}^{K\times L}\to\mathbb{R}^{M\times N}$ becomes [294, §2.1, §5.4.5] [35, §6.3.1]

$$\overset{\to Y}{dg^2}(X) = \left.\frac{d^2}{dt^2}\right|_{t=0} g(X+t\,Y) \in \mathbb{R}^{M\times N} \tag{1931}$$

which is again simplest. (*confer* (1910)) Directional derivative retains the dimensions of g.

D.1.6 directional derivative expressions

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

$$\overset{\to Y}{dg}(X) = \operatorname{tr}\big(\nabla g(X)^{\mathrm T} Y\big) \tag{1932}$$

$$\overset{\to Y}{dg^2}(X) = \operatorname{tr}\Big(\nabla_X \operatorname{tr}\big(\nabla g(X)^{\mathrm T} Y\big)^{\mathrm T} Y\Big) = \operatorname{tr}\bigg(\nabla_X\,\overset{\to Y}{dg}(X)^{\mathrm T} Y\bigg) \tag{1933}$$

$$\overset{\to Y}{dg^3}(X) = \operatorname{tr}\bigg(\nabla_X \operatorname{tr}\Big(\nabla_X \operatorname{tr}\big(\nabla g(X)^{\mathrm T} Y\big)^{\mathrm T} Y\Big)^{\mathrm T} Y\bigg) = \operatorname{tr}\bigg(\nabla_X\,\overset{\to Y}{dg^2}(X)^{\mathrm T} Y\bigg) \tag{1934}$$

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

$$\overset{\to Y}{dg}(X) = \nabla g(X)^{\mathrm T} Y \tag{1935}$$

$$\overset{\to Y}{dg^2}(X) = Y^{\mathrm T}\nabla^2 g(X) Y \tag{1936}$$

$$\overset{\to Y}{dg^3}(X) = \nabla_X\big(Y^{\mathrm T}\nabla^2 g(X) Y\big)^{\mathrm T} Y \tag{1937}$$

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 (1909) and (1930). Assuming $g(X)$ has continuous first-, second-, and third-order gradients over the open set $\operatorname{dom} g$, then for $X \in \operatorname{dom} g$ and any $Y \in \mathbb{R}^{K\times L}$ the complete Taylor series is expressed on some open interval of $\mu \in \mathbb{R}$

$$g(X+\mu Y) = g(X) \;+\; \mu \overset{\rightarrow Y}{dg}(X) \;+\; \frac{1}{2!}\mu^2 \overset{\rightarrow Y}{dg^2}(X) \;+\; \frac{1}{3!}\mu^3 \overset{\rightarrow Y}{dg^3}(X) \;+\; o(\mu^4) \tag{1938}$$

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

$$g(Y) = g(X) \;+\; \overset{\rightarrow Y-X}{dg}(X) \;+\; \frac{1}{2!}\overset{\rightarrow Y-X}{dg^2}(X) \;+\; \frac{1}{3!}\overset{\rightarrow Y-X}{dg^3}(X) \;+\; o(\|Y\|^4) \tag{1939}$$

which are third-order expansions about X. The *mean value theorem* from calculus is what insures finite order of the series. [235] [43, §1.1] [42, App.A.5] [215, §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 596,

$$\overset{\rightarrow Y}{dg}(X) = \left.\frac{d}{dt}\right|_{t=0} g(X+t\,Y) \;=\; -X^{-1}YX^{-1} \tag{1940}$$

$$\overset{\rightarrow Y}{dg^2}(X) = \left.\frac{d^2}{dt^2}\right|_{t=0} g(X+t\,Y) \;=\; 2X^{-1}YX^{-1}YX^{-1} \tag{1941}$$

$$\overset{\rightarrow Y}{dg^3}(X) = \left.\frac{d^3}{dt^3}\right|_{t=0} g(X+t\,Y) \;=\; -6X^{-1}YX^{-1}YX^{-1}YX^{-1} \tag{1942}$$

Let's find the Taylor series expansion of g about $X = I$: Since $g(I) = I$, for $\|Y\|_2 < 1$ ($\mu = 1$ in (1938))

$$g(I+Y) \;=\; (I+Y)^{-1} \;=\; I - Y + Y^2 - Y^3 + \ldots \tag{1943}$$

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

$$g(X+Y) \;=\; (X+Y)^{-1} \;=\; X^{-1} - \; X^{-1}YX^{-1} \;+\; 2X^{-1}YX^{-1}YX^{-1} \;-\; \ldots \tag{1944}$$

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

D.1.7.0.2 Exercise. *log det*. (*confer* [63, 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.3] Had we instead set $g(Y) = (I+Y)^{-1}$, then the equivalent expansion would have been about $X = \mathbf{0}$.

D.1.8 Correspondence of gradient to derivative

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

D.1.8.1 first-order

Removing evaluation at $t=0$ from (1910),[D.4] we find an expression for the directional derivative of $g(X)$ in direction Y evaluated anywhere along a line $\{X+t\,Y \mid t\in\mathbb{R}\}$ intersecting $\operatorname{dom} g$

$$\overset{\rightarrow Y}{dg}(X+t\,Y)=\frac{d}{dt}g(X+t\,Y) \tag{1945}$$

In the general case $g(X):\mathbb{R}^{K\times L}\to\mathbb{R}^{M\times N}$, from (1903) and (1906) we find

$$\operatorname{tr}\big(\nabla_X\, g_{mn}(X+t\,Y)^{\mathrm T}Y\big)=\frac{d}{dt}g_{mn}(X+t\,Y) \tag{1946}$$

which is valid at $t=0$, of course, when $X\in\operatorname{dom} g$. In the important case of a real function $g(X):\mathbb{R}^{K\times L}\to\mathbb{R}$, from (1932) we have simply

$$\operatorname{tr}\big(\nabla_X\, g(X+t\,Y)^{\mathrm T}Y\big)=\frac{d}{dt}g(X+t\,Y) \tag{1947}$$

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

$$\nabla_X\, g(X+t\,Y)^{\mathrm T}Y=\frac{d}{dt}g(X+t\,Y) \tag{1948}$$

D.1.8.1.1 Example. *Gradient.*
$g(X)=w^{\mathrm T}X^{\mathrm T}Xw$, $X\in\mathbb{R}^{K\times L}$, $w\in\mathbb{R}^{L}$. Using the tables in §D.2,

$$\begin{aligned}\operatorname{tr}\big(\nabla_X\, g(X+t\,Y)^{\mathrm T}Y\big) &= \operatorname{tr}\big(2ww^{\mathrm T}(X^{\mathrm T}+t\,Y^{\mathrm T})Y\big) &(1949)\\ &= 2w^{\mathrm T}(X^{\mathrm T}Y+t\,Y^{\mathrm T}Y)w &(1950)\end{aligned}$$

Applying equivalence (1947),

$$\begin{aligned}\frac{d}{dt}g(X+t\,Y) &= \frac{d}{dt}w^{\mathrm T}(X+t\,Y)^{\mathrm T}(X+t\,Y)w &(1951)\\ &= w^{\mathrm T}\big(X^{\mathrm T}Y+Y^{\mathrm T}X+2t\,Y^{\mathrm T}Y\big)\,w &(1952)\\ &= 2w^{\mathrm T}(X^{\mathrm T}Y+t\,Y^{\mathrm T}Y)w &(1953)\end{aligned}$$

which is the same as (1950). Hence, the equivalence is demonstrated.

[D.4] Justified by replacing X with $X+t\,Y$ in (1903)-(1905); beginning,

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

It is easy to extract $\nabla g(X)$ from (1953) knowing only (1947):

$$\begin{array}{rcl}\operatorname{tr}\big(\nabla_X\, g(X+t\,Y)^{\mathrm T} Y\big) &=& 2w^{\mathrm T}(X^{\mathrm T} Y + t\,Y^{\mathrm T} Y)w\\ &=& 2\operatorname{tr}\big(ww^{\mathrm T}(X^{\mathrm T}+t\,Y^{\mathrm T})Y\big)\\ \operatorname{tr}\big(\nabla_X\, g(X)^{\mathrm T} Y\big) &=& 2\operatorname{tr}\big(ww^{\mathrm T} X^{\mathrm T} Y\big)\\ &\Leftrightarrow& \\ \nabla_X\, g(X) &=& 2Xww^{\mathrm T}\end{array} \tag{1954}$$

$\square$

D.1.8.2 second-order

Likewise removing the evaluation at $t=0$ from (1931),

$$\overset{\to Y}{dg^2}(X+t\,Y) = \frac{d^2}{dt^2} g(X+t\,Y) \tag{1955}$$

we can find a similar relationship between second-order gradient and second derivative: In the general case $g(X) : \mathbb{R}^{K\times L}\to\mathbb{R}^{M\times N}$ from (1924) and (1927),

$$\operatorname{tr}\Big(\nabla_X \operatorname{tr}\big(\nabla_X\, g_{mn}(X+t\,Y)^{\mathrm T} Y\big)^{\mathrm T} Y\Big) = \frac{d^2}{dt^2} g_{mn}(X+t\,Y) \tag{1956}$$

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

$$\operatorname{tr}\Big(\nabla_X \operatorname{tr}\big(\nabla_X\, g(X+t\,Y)^{\mathrm T} Y\big)^{\mathrm T} Y\Big) = \frac{d^2}{dt^2} g(X+t\,Y) \tag{1957}$$

From (1936), the simpler case, where real function $g(X) : \mathbb{R}^{K}\to\mathbb{R}$ has vector argument,

$$Y^{\mathrm T} \nabla_X^2 g(X+t\,Y) Y = \frac{d^2}{dt^2} g(X+t\,Y) \tag{1958}$$

D.1.8.2.1 Example. *Second-order gradient.*
We want to find $\nabla^2 g(X) \in \mathbb{R}^{K\times K\times K\times K}$ given real function $g(X) = \log\det X$ having domain $\operatorname{int}\mathbb{S}_+^K$. From the tables in §D.2,

$$h(X) \triangleq \nabla g(X) = X^{-1} \in \operatorname{int}\mathbb{S}_+^K \tag{1959}$$

so $\nabla^2 g(X) = \nabla h(X)$. By (1946) and (1909), for $Y\in\mathbb{S}^K$

$$\operatorname{tr}\big(\nabla h_{mn}(X)^{\mathrm T} Y\big) = \left.\frac{d}{dt}\right|_{t=0} h_{mn}(X+t\,Y) \tag{1960}$$

$$= \left(\left.\frac{d}{dt}\right|_{t=0} h(X+t\,Y)\right)_{mn} \tag{1961}$$

$$= \left(\left.\frac{d}{dt}\right|_{t=0} (X+t\,Y)^{-1}\right)_{mn} \tag{1962}$$

$$= -\big(X^{-1}\,Y X^{-1}\big)_{mn} \tag{1963}$$

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

$$\nabla^2 g(X)_{mnkl} \;=\; \operatorname{tr}\big(\nabla h_{mn}(X)^{\mathrm T} e_k e_l^{\mathrm T}\big) \;=\; \nabla h_{mn}(X)_{kl} \;=\; -\big(X^{-1} e_k e_l^{\mathrm T} X^{-1}\big)_{mn} \tag{1964}$$

$$\nabla^2 g(X)_{kl} \;=\; \nabla h(X)_{kl} \;=\; -\big(X^{-1} e_k e_l^{\mathrm T} X^{-1}\big) \in \mathbb{R}^{K\times K} \tag{1965}$$

□

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. [115] When proving results for symmetric matrices algebraically, it is critical to take gradients ignoring symmetry and to then substitute symmetric entries afterward. [182] [67]

- $a\,,b\in\mathbb{R}^n$, $x\,,y\in\mathbb{R}^k$, $A\,,B\in\mathbb{R}^{m\times n}$, $X,Y\in\mathbb{R}^{K\times L}$, $t\,,\mu\in\mathbb{R}$, $i\,,j\,,k\,,\ell\,,K\,,L\,,m\,,n\,,M\,,N$ are integers, unless otherwise noted.

- x^μ means $\delta(\delta(x)^\mu)$ for $\mu\in\mathbb{R}$; *id est*, entrywise vector exponentiation. δ is the main-diagonal linear operator (1504). $x^0 \triangleq \mathbf{1}$, $X^0 \triangleq I$ 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), $\sqrt[\circ]{x}$ (entrywise square root), *etcetera*, are maps $f:\mathbb{R}^k\to\mathbb{R}^k$ that maintain dimension; *e.g*, (§A.1.1)

$$\frac{d}{dx}x^{-1} \;\triangleq\; \nabla_x\, \mathbf{1}^{\mathrm T}\delta(x)^{-1}\mathbf{1} \tag{1966}$$

- For A a scalar or square matrix, we have the Taylor series [80, §3.6]

$$e^A \triangleq \sum_{k=0}^{\infty}\frac{1}{k!}A^k \tag{1967}$$

  Further, [348, §5.4]

$$e^A \succ 0 \qquad \forall\, A\in\mathbb{S}^m \tag{1968}$$

- For all square A and integer k

$$\det^k A = \det A^k \tag{1969}$$

D.2.1 algebraic

$\nabla_x\, x = \nabla_x\, x^{\mathrm T} = I \in \mathbb{R}^{k\times k}$	$\nabla_X X = \nabla_X X^{\mathrm T} \triangleq I \in \mathbb{R}^{K\times L\times K\times L}$ (Identity)
$\nabla_x(Ax-b) = A^{\mathrm T}$	
$\nabla_x\big(x^{\mathrm T}A - b^{\mathrm T}\big) = A$	
$\nabla_x(Ax-b)^{\mathrm T}(Ax-b) = 2A^{\mathrm T}(Ax-b)$ $\nabla_x^2(Ax-b)^{\mathrm T}(Ax-b) = 2A^{\mathrm T}A$	
$\nabla_x\|Ax-b\| = A^{\mathrm T}(Ax-b)/\|Ax-b\|$	
$\nabla_x z^{\mathrm T}\lvert Ax-b\rvert = A^{\mathrm T}\delta(z)\,\mathrm{sgn}(Ax-b)\,,\;\; z_i\neq 0 \Rightarrow (Ax-b)_i\neq 0$ $\nabla_x \mathbf{1}^{\mathrm T} f(\lvert Ax-b\rvert) = A^{\mathrm T}\delta\!\left(\left.\frac{df(y)}{dy}\right\rvert_{y=\lvert Ax-b\rvert}\right)\mathrm{sgn}(Ax-b)$	
$\nabla_x\big(x^{\mathrm T}Ax + 2x^{\mathrm T}By + y^{\mathrm T}Cy\big) = \big(A + A^{\mathrm T}\big)x + 2By$ $\nabla_x(x+y)^{\mathrm T}A(x+y) = (A+A^{\mathrm T})(x+y)$ $\nabla_x^2\big(x^{\mathrm T}Ax + 2x^{\mathrm T}By + y^{\mathrm T}Cy\big) = A + A^{\mathrm T}$	
	$\nabla_X\, a^{\mathrm T}Xb \,=\, \nabla_X\, b^{\mathrm T}X^{\mathrm T}a \,=\, ab^{\mathrm T}$
	$\nabla_X\, a^{\mathrm T}X^2 b = X^{\mathrm T}ab^{\mathrm T} + ab^{\mathrm T}X^{\mathrm T}$
	$\nabla_X\, a^{\mathrm T}X^{-1}b = -X^{-\mathrm T}ab^{\mathrm T}X^{-\mathrm T}$
	$\nabla_X(X^{-1})_{kl} = \dfrac{\partial X^{-1}}{\partial X_{kl}} = -X^{-1}e_k e_l^{\mathrm T}X^{-1},$ *confer* (1901) (1965)
$\nabla_x\, a^{\mathrm T}x^{\mathrm T}xb = 2xa^{\mathrm T}b$	$\nabla_X\, a^{\mathrm T}X^{\mathrm T}Xb = X(ab^{\mathrm T} + ba^{\mathrm T})$
$\nabla_x\, a^{\mathrm T}xx^{\mathrm T}b = (ab^{\mathrm T} + ba^{\mathrm T})x$	$\nabla_X\, a^{\mathrm T}XX^{\mathrm T}b = (ab^{\mathrm T} + ba^{\mathrm T})X$
$\nabla_x\, a^{\mathrm T}x^{\mathrm T}xa = 2xa^{\mathrm T}a$	$\nabla_X\, a^{\mathrm T}X^{\mathrm T}Xa = 2Xaa^{\mathrm T}$
$\nabla_x\, a^{\mathrm T}xx^{\mathrm T}a = 2aa^{\mathrm T}x$	$\nabla_X\, a^{\mathrm T}XX^{\mathrm T}a = 2aa^{\mathrm T}X$
$\nabla_x\, a^{\mathrm T}yx^{\mathrm T}b = ba^{\mathrm T}y$	$\nabla_X\, a^{\mathrm T}YX^{\mathrm T}b = ba^{\mathrm T}Y$
$\nabla_x\, a^{\mathrm T}y^{\mathrm T}xb = yb^{\mathrm T}a$	$\nabla_X\, a^{\mathrm T}Y^{\mathrm T}Xb = Yab^{\mathrm T}$
$\nabla_x\, a^{\mathrm T}xy^{\mathrm T}b = ab^{\mathrm T}y$	$\nabla_X\, a^{\mathrm T}XY^{\mathrm T}b = ab^{\mathrm T}Y$
$\nabla_x\, a^{\mathrm T}x^{\mathrm T}yb = ya^{\mathrm T}b$	$\nabla_X\, a^{\mathrm T}X^{\mathrm T}Yb = Yba^{\mathrm T}$

algebraic continued

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

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

$$\frac{d^2}{dt^2}B^{\mathrm T}(X+t\,Y)^{-1}A = \quad 2B^{\mathrm T}(X+t\,Y)^{-1}\,Y(X+t\,Y)^{-1}\,Y(X+t\,Y)^{-1}A$$
$$\frac{d^3}{dt^3}B^{\mathrm T}(X+t\,Y)^{-1}A = -6B^{\mathrm T}(X+t\,Y)^{-1}\,Y(X+t\,Y)^{-1}\,Y(X+t\,Y)^{-1}Y(X+t\,Y)^{-1}A$$

$$\frac{d}{dt}\big((X+t\,Y)^{\mathrm T}A(X+t\,Y)\big) = Y^{\mathrm T}AX + X^{\mathrm T}AY + 2\,t\,Y^{\mathrm T}AY$$
$$\frac{d^2}{dt^2}\big((X+t\,Y)^{\mathrm T}A(X+t\,Y)\big) = 2\,Y^{\mathrm T}AY$$
$$\frac{d}{dt}\big((X+t\,Y)^{\mathrm T}A(X+t\,Y)\big)^{-1}$$
$$= -\big((X+t\,Y)^{\mathrm T}A(X+t\,Y)\big)^{-1}(Y^{\mathrm T}AX + X^{\mathrm T}AY + 2\,t\,Y^{\mathrm T}AY)\big((X+t\,Y)^{\mathrm T}A(X+t\,Y)\big)^{-1}$$
$$\frac{d}{dt}((X+t\,Y)A(X+t\,Y)) = YAX + XAY + 2\,t\,YAY$$
$$\frac{d^2}{dt^2}((X+t\,Y)A(X+t\,Y)) = 2\,YAY$$

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

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

$$\nabla^2_{\mathrm{vec}\,X}\,\mathrm{tr}(AXBX^{\mathrm T}) = \nabla^2_{\mathrm{vec}\,X}\,\mathrm{vec}(X)^{\mathrm T}(B^{\mathrm T}\otimes A)\,\mathrm{vec}\,X = B\otimes A^{\mathrm T} + B^{\mathrm T}\otimes A$$

D.2.3 trace

$\nabla_x \, \mu \, x = \mu I$	$\nabla_X \operatorname{tr} \mu X = \nabla_X \, \mu \operatorname{tr} X = \mu I$
$\nabla_x \, \mathbf{1}^{\mathrm{T}} \delta(x)^{-1} \mathbf{1} = \frac{d}{dx} x^{-1} = -x^{-2}$ $\nabla_x \, \mathbf{1}^{\mathrm{T}} \delta(x)^{-1} y = -\delta(x)^{-2} y$	$\nabla_X \operatorname{tr} X^{-1} = -X^{-2\mathrm{T}}$ $\nabla_X \operatorname{tr}(X^{-1} Y) = \nabla_X \operatorname{tr}(Y X^{-1}) = -X^{-\mathrm{T}} Y^{\mathrm{T}} X^{-\mathrm{T}}$
$\frac{d}{dx} x^{\mu} = \mu x^{\mu-1}$	$\nabla_X \operatorname{tr} X^{\mu} = \mu X^{\mu-1}\,, \qquad X \in \mathbb{S}^M$
	$\nabla_X \operatorname{tr} X^j = j X^{(j-1)\mathrm{T}}$
$\nabla_x (b - a^{\mathrm{T}} x)^{-1} = (b - a^{\mathrm{T}} x)^{-2} a$	$\nabla_X \operatorname{tr}\big((B - AX)^{-1}\big) = \big((B - AX)^{-2} A\big)^{\mathrm{T}}$
$\nabla_x (b - a^{\mathrm{T}} x)^{\mu} = -\mu (b - a^{\mathrm{T}} x)^{\mu-1} a$	
$\nabla_x \, x^{\mathrm{T}} y = \nabla_x \, y^{\mathrm{T}} x = y$	$\nabla_X \operatorname{tr}(X^{\mathrm{T}} Y) = \nabla_X \operatorname{tr}(Y X^{\mathrm{T}}) = \nabla_X \operatorname{tr}(Y^{\mathrm{T}} X) = \nabla_X \operatorname{tr}(X Y^{\mathrm{T}}) = Y$
	$\nabla_X \operatorname{tr}(AXBX^{\mathrm{T}}) = \nabla_X \operatorname{tr}(XBX^{\mathrm{T}}A) = A^{\mathrm{T}} X B^{\mathrm{T}} + AXB$ $\nabla_X \operatorname{tr}(AXBX) = \nabla_X \operatorname{tr}(XBXA) = A^{\mathrm{T}} X^{\mathrm{T}} B^{\mathrm{T}} + B^{\mathrm{T}} X^{\mathrm{T}} A^{\mathrm{T}}$
	$\nabla_X \operatorname{tr}(AXAXAX) = \nabla_X \operatorname{tr}(XAXAXA) = 3(AXAXA)^{\mathrm{T}}$
	$\nabla_X \operatorname{tr}(Y X^k) = \nabla_X \operatorname{tr}(X^k Y) = \sum\limits_{i=0}^{k-1} \big(X^i Y X^{k-1-i}\big)^{\mathrm{T}}$
	$\nabla_X \operatorname{tr}(Y^{\mathrm{T}} X X^{\mathrm{T}} Y) = \nabla_X \operatorname{tr}(X^{\mathrm{T}} Y Y^{\mathrm{T}} X) = 2\, Y Y^{\mathrm{T}} X$ $\nabla_X \operatorname{tr}(Y^{\mathrm{T}} X^{\mathrm{T}} X Y) = \nabla_X \operatorname{tr}(X Y Y^{\mathrm{T}} X^{\mathrm{T}}) = 2 X Y Y^{\mathrm{T}}$
	$\nabla_X \operatorname{tr}\big((X+Y)^{\mathrm{T}}(X+Y)\big) = 2(X+Y) = \nabla_X \\|X+Y\\|_{\mathrm{F}}^2$ $\nabla_X \operatorname{tr}((X+Y)(X+Y)) = 2(X+Y)^{\mathrm{T}}$
	$\nabla_X \operatorname{tr}(A^{\mathrm{T}} X B) = \nabla_X \operatorname{tr}(X^{\mathrm{T}} A B^{\mathrm{T}}) = AB^{\mathrm{T}}$ $\nabla_X \operatorname{tr}(A^{\mathrm{T}} X^{-1} B) = \nabla_X \operatorname{tr}(X^{-\mathrm{T}} A B^{\mathrm{T}}) = -X^{-\mathrm{T}} A B^{\mathrm{T}} X^{-\mathrm{T}}$
	$\nabla_X \, a^{\mathrm{T}} X b = \nabla_X \operatorname{tr}(b a^{\mathrm{T}} X) = \nabla_X \operatorname{tr}(X b a^{\mathrm{T}}) = a b^{\mathrm{T}}$ $\nabla_X \, b^{\mathrm{T}} X^{\mathrm{T}} a = \nabla_X \operatorname{tr}(X^{\mathrm{T}} a b^{\mathrm{T}}) = \nabla_X \operatorname{tr}(a b^{\mathrm{T}} X^{\mathrm{T}}) = a b^{\mathrm{T}}$ $\nabla_X \, a^{\mathrm{T}} X^{-1} b = \nabla_X \operatorname{tr}(X^{-\mathrm{T}} a b^{\mathrm{T}}) = -X^{-\mathrm{T}} a b^{\mathrm{T}} X^{-\mathrm{T}}$ $\nabla_X \, a^{\mathrm{T}} X^{\mu} b = \ldots$

trace continued

$$\frac{d}{dt}\operatorname{tr} g(X+tY) = \operatorname{tr}\frac{d}{dt}g(X+tY) \qquad [219, \text{p.491}]$$

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

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

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

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

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

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

$$\begin{aligned}
\frac{d}{dt}\operatorname{tr}\big((X+tY)^{-1}Y\big) &= -\operatorname{tr}\big((X+tY)^{-1}Y(X+tY)^{-1}Y\big)\\
\frac{d}{dt}\operatorname{tr}\big(B^{\mathrm T}(X+tY)^{-1}A\big) &= -\operatorname{tr}\big(B^{\mathrm T}(X+tY)^{-1}Y(X+tY)^{-1}A\big)\\
\frac{d}{dt}\operatorname{tr}\big(B^{\mathrm T}(X+tY)^{-\mathrm T}A\big) &= -\operatorname{tr}\big(B^{\mathrm T}(X+tY)^{-\mathrm T}\,Y^{\mathrm T}(X+tY)^{-\mathrm T}A\big)\\
\frac{d}{dt}\operatorname{tr}\big(B^{\mathrm T}(X+tY)^{-k}A\big) &= \ldots\,, \quad k>0\\
\frac{d}{dt}\operatorname{tr}\big(B^{\mathrm T}(X+tY)^{\mu}A\big) &= \ldots\,, \quad -1\le\mu\le1,\quad X\,,Y\in\mathbb{S}_+^M
\end{aligned}$$

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

$$\begin{aligned}
&\frac{d}{dt}\operatorname{tr}\big((X+tY)^{\mathrm T}A(X+tY)\big) = \operatorname{tr}\big(Y^{\mathrm T}AX + X^{\mathrm T}AY + 2\,t\,Y^{\mathrm T}AY\big)\\
&\frac{d^2}{dt^2}\operatorname{tr}\big((X+tY)^{\mathrm T}A(X+tY)\big) = 2\operatorname{tr}\big(Y^{\mathrm T}AY\big)\\
&\frac{d}{dt}\operatorname{tr}\Big(\big((X+tY)^{\mathrm T}A(X+tY)\big)^{-1}\Big)\\
&\quad= -\operatorname{tr}\Big(\big((X+tY)^{\mathrm T}A(X+tY)\big)^{-1}(Y^{\mathrm T}AX + X^{\mathrm T}AY + 2\,t\,Y^{\mathrm T}AY)\big((X+tY)^{\mathrm T}A(X+tY)\big)^{-1}\Big)\\
&\frac{d}{dt}\operatorname{tr}\big((X+tY)A(X+tY)\big) = \operatorname{tr}(YAX + XAY + 2\,t\,YAY)\\
&\frac{d^2}{dt^2}\operatorname{tr}\big((X+tY)A(X+tY)\big) = 2\operatorname{tr}(YAY)
\end{aligned}$$

D.2.4 logarithmic determinant

$x \succ 0$, $\det X > 0$ on some neighborhood of X, and $\det(X + tY) > 0$ on some open interval of t; otherwise, $\log(\,)$ would be discontinuous. [86, p.75]

$\frac{d}{dx}\log x = x^{-1}$	$\nabla_X \log\det X = X^{-\mathrm{T}}$
	$\nabla_X^2 \log\det(X)_{kl} = \dfrac{\partial X^{-\mathrm{T}}}{\partial X_{kl}} = -\left(X^{-1}e_k e_l^{\mathrm{T}} X^{-1}\right)^{\mathrm{T}}$, *confer* (1918)(1965)
$\frac{d}{dx}\log x^{-1} = -x^{-1}$	$\nabla_X \log\det X^{-1} = -X^{-\mathrm{T}}$
$\frac{d}{dx}\log x^{\mu} = \mu x^{-1}$	$\nabla_X \log\det^{\mu} X = \mu X^{-\mathrm{T}}$
	$\nabla_X \log\det X^{\mu} = \mu X^{-\mathrm{T}}$
	$\nabla_X \log\det X^{k} = \nabla_X \log\det^{k} X = kX^{-\mathrm{T}}$
	$\nabla_X \log\det^{\mu}(X + tY) = \mu(X + tY)^{-\mathrm{T}}$
$\nabla_x \log(a^{\mathrm{T}}x + b) = a\frac{1}{a^{\mathrm{T}}x+b}$	$\nabla_X \log\det(AX + B) = A^{\mathrm{T}}(AX + B)^{-\mathrm{T}}$
	$\nabla_X \log\det(I \pm A^{\mathrm{T}}XA) = \pm A(I \pm A^{\mathrm{T}}XA)^{-\mathrm{T}}A^{\mathrm{T}}$
	$\nabla_X \log\det(X + tY)^{k} = \nabla_X \log\det^{k}(X + tY) = k(X + tY)^{-\mathrm{T}}$
	$\frac{d}{dt}\log\det(X + tY) = \operatorname{tr}\left((X + tY)^{-1}Y\right)$
	$\frac{d^2}{dt^2}\log\det(X + tY) = -\operatorname{tr}\left((X + tY)^{-1}Y(X + tY)^{-1}Y\right)$
	$\frac{d}{dt}\log\det(X + tY)^{-1} = -\operatorname{tr}\left((X + tY)^{-1}Y\right)$
	$\frac{d^2}{dt^2}\log\det(X + tY)^{-1} = \operatorname{tr}\left((X + tY)^{-1}Y(X + tY)^{-1}Y\right)$
	$\frac{d}{dt}\log\det\left(\delta(A(x + ty) + a)^2 + \mu I\right)$ $= \operatorname{tr}\Big(\left(\delta(A(x + ty) + a)^2 + \mu I\right)^{-1} 2\delta\left(A(x + ty) + a\right)\delta(Ay)\Big)$

D.2.5 determinant

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

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

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

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

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

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

$$\nabla_X \det{}^{\mu}(X+t\,Y) = \mu \det{}^{\mu}(X+t\,Y)(X+t\,Y)^{-\mathrm{T}}$$

$$\nabla_X \det(X+t\,Y)^k = \nabla_X \det{}^k(X+t\,Y) = k \det{}^k(X+t\,Y)(X+t\,Y)^{-\mathrm{T}}$$

$$\tfrac{d}{dt} \det(X+t\,Y) = \det(X+t\,Y)\operatorname{tr}((X+t\,Y)^{-1}\,Y)$$

$$\tfrac{d^2}{dt^2} \det(X+t\,Y) = \det(X+t\,Y)\big(\operatorname{tr}^2\big((X+t\,Y)^{-1}\,Y\big) - \operatorname{tr}((X+t\,Y)^{-1}\,Y(X+t\,Y)^{-1}\,Y)\big)$$

$$\tfrac{d}{dt} \det(X+t\,Y)^{-1} = -\det(X+t\,Y)^{-1}\operatorname{tr}((X+t\,Y)^{-1}\,Y)$$

$$\tfrac{d^2}{dt^2} \det(X+t\,Y)^{-1} = \det(X+t\,Y)^{-1}\big(\operatorname{tr}^2((X+t\,Y)^{-1}\,Y) + \operatorname{tr}((X+t\,Y)^{-1}\,Y(X+t\,Y)^{-1}\,Y)\big)$$

$$\tfrac{d}{dt} \det{}^{\mu}(X+t\,Y) = \mu \det{}^{\mu}(X+t\,Y)\operatorname{tr}((X+t\,Y)^{-1}\,Y)$$

D.2.6 logarithmic

Matrix logarithm.

$$\tfrac{d}{dt}\log(X+t\,Y)^{\mu} = \mu Y(X+t\,Y)^{-1} = \mu(X+t\,Y)^{-1}Y\,, \qquad XY = YX$$

$$\tfrac{d}{dt}\log(I-t\,Y)^{\mu} = -\mu Y(I-t\,Y)^{-1} = -\mu(I-t\,Y)^{-1}Y \qquad [219, p.493]$$

D.2.7 exponential

Matrix exponential. [80, §3.6, §4.5] [348, §5.4]

$$\nabla_X e^{\operatorname{tr}(Y^{\mathrm T}X)} = \nabla_X \det e^{Y^{\mathrm T}X} = e^{\operatorname{tr}(Y^{\mathrm T}X)}Y \qquad (\forall\, X, Y)$$

$$\nabla_X \operatorname{tr} e^{YX} = e^{Y^{\mathrm T}X^{\mathrm T}}Y^{\mathrm T} = Y^{\mathrm T}e^{X^{\mathrm T}Y^{\mathrm T}}$$

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

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

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

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

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

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

$$\tfrac{d}{dt}e^{tY} = e^{tY}Y = Ye^{tY}$$

$$\tfrac{d}{dt}e^{X+tY} = e^{X+tY}Y = Ye^{X+tY}, \qquad XY = YX$$

$$\tfrac{d^2}{dt^2}e^{X+tY} = e^{X+tY}Y^2 = Ye^{X+tY}Y = Y^2e^{X+tY}, \qquad XY = YX$$

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