**Introduction**

These are handy results needed from mathematics. These will be applied; more fundamental building blocks (mostly from analysis, algebra and calculus) are found in [1]; standard texts should also be consulted.

**Differentiation in $\mathbb{R}^n$**

We define an element $x \in \mathbb{R}^n$ as a standard $nx1$ column vector $(x_1, \ldots, x_n)$. We define our matrices as $nxk$ where $n$ is the number of rows (observations) and $k$ (or variously $m, p, d$) the number of columns (components/variables); we believe this represents the majority usage in engineering and statistics. (Unfortunately there are users who define $n$ variables as columns with $m$ row observations.

Let a function $g: \mathbb{R}^d \rightarrow \mathbb{R}^k$. We say the derivative of $g$ at $x \in \mathbb{R}^d$ is defined to be the linear map:

$$T: h \mapsto g(x + h) - g(x) + o(||h||).$$

Other parameterizations are possible; the interpretation of the derivative as a linear map is not optional. Usually $\nabla$ is the matrix associated with this transformation, and we often denote it variously as $\dot{g}, \nabla g, d_g$, or $\frac{\partial g}{\partial x}$. The second derivative at $x$ is defined similarly and is denoted $\ddot{g}, \nabla^2 g, D_g$, or $\frac{\partial^2 g}{\partial x \partial x^T}$.

**Definitional Notation**

$$\begin{align*}
g &: \mathbb{R} \rightarrow \mathbb{R} \\
\dot{g} &: \mathbb{R} \rightarrow \mathbb{R} \quad \dot{g} = g(x) \\
\ddot{g} &: \mathbb{R} \rightarrow \mathbb{R} \\
g(x) &= a \sin(x), \dot{g}(x) = a \cos(x), \text{ and } \ddot{g}(x) = -g(x).
\end{align*}$$

$$\begin{align*}
g &: \mathbb{R} \rightarrow \mathbb{R}^{d_xk} = \mathbb{R}^k \\
\dot{g} &: \mathbb{R}^k \rightarrow \mathbb{R}^{d_xk} = \mathbb{R}^k \\
\ddot{g} &: \mathbb{R}^k \rightarrow \mathbb{R}^k \\
g(x) &= xx^T \in \mathbb{R}^{d_d}; \text{ we have } \nabla g = \dot{g}(x): \mathbb{R}^d \rightarrow \mathbb{R}^d, \text{ with } \dot{g}(x) = \nabla xx^T = \nabla xx + x \nabla x = 2x.
\end{align*}$$

$$\begin{align*}
g &: \mathbb{R}^d \rightarrow \mathbb{R} \\
\dot{g} &: \mathbb{R}^d \rightarrow \mathbb{R}^{d_xk} = \mathbb{R}^d \\
\ddot{g} &: \mathbb{R}^{d_xk} = \mathbb{R}^d \rightarrow \mathbb{R}^{d_xk \times d} = \mathbb{R}^{d_xd} \\
g(x_1, \ldots, x_d) &= \dot{g} = \nabla g = \\
\ddot{g}(x_1, \ldots, x_d) &= \frac{\partial^2 g}{\partial x \partial x^T}.
\end{align*}$$
Helpful Mathematical Relations, Functions, and Shortcuts

Useful Matrix Derivatives

\[ \frac{\partial^2 g}{\partial x_i \partial x_j} = \begin{bmatrix} \frac{\partial^2 g}{\partial x_1 \partial x_1} & \frac{\partial^2 g}{\partial x_1 \partial x_2} & \ldots & \frac{\partial^2 g}{\partial x_1 \partial x_d} \\ \frac{\partial^2 g}{\partial x_2 \partial x_1} & \frac{\partial^2 g}{\partial x_2 \partial x_2} & \ldots & \frac{\partial^2 g}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 g}{\partial x_d \partial x_1} & \frac{\partial^2 g}{\partial x_d \partial x_2} & \ldots & \frac{\partial^2 g}{\partial x_d \partial x_d} \end{bmatrix} \]

For example, suppose \( g : R^d \rightarrow R \) as \( g(x) = x^T x \). \( \dot{g} : R^d \rightarrow R^d \) as \( \nabla x^T x = 2x \). Or, consider \( g : R^{d \times d} \rightarrow R, \ g(X) = \det(X) \). It can be shown that \( \dot{g} : R^{d \times d} \rightarrow R^{d \times d} \) as \( \nabla |X| \in R^{d \times d} \).

\[
\begin{align*}
g & : R^d \rightarrow R^k \\
x \dot{g} & : R^d \rightarrow R^{d \times k} \\
\ddot{g} & : R^{d \times k} \rightarrow R^{d \times k \times d} \\
g(x) & = \begin{bmatrix} g_1(x_1, \ldots, x_d) \\ \vdots \\ g_k(x_1, \ldots, x_d) \end{bmatrix} \\
\dot{g} & = \nabla g^T = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_2}{\partial x_1} & \ldots & \frac{\partial g_k}{\partial x_1} \\ \frac{\partial g_1}{\partial x_2} & \frac{\partial g_2}{\partial x_2} & \ldots & \frac{\partial g_k}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial x_d} & \frac{\partial g_2}{\partial x_d} & \ldots & \frac{\partial g_k}{\partial x_d} \end{bmatrix} \\
\ddot{g} & = \nabla^2 g = \nabla [?] \end{align*}
\]

Useful Matrix Derivatives

**Chain Rules**

- If \( f : R^d \rightarrow R^e, g : R^e \rightarrow R^k \), and \( h = g(f(x)) : R^d \rightarrow R^k \) then \( \dot{h}(x) = \dot{g}(f(x)) \dot{f}(x) \)
- If both \( f, g : R^d \rightarrow R^k \), and \( h \in R^{d \times k} = f^T(x)g(x) \) then \( \dot{h}(x) = g(x)^T \dot{f}(x) + f(x)^T \dot{g}(x) \)  
  (NOTE need to check this…)

**Remarks**

- It is best to have a complete guide to differentiation of scalars, vectors and matrices with respect to scalars, vectors and matrices; Gentle [4] provides a good summary. Just the first derivatives for these 9 combinations can result in tensors of rank higher than 2.
- Note that a non-negative measure of variation \( h(\dot{f}) \), such as \( \left| \frac{df}{d\theta} \right| \) or \( \left( \frac{df}{d\theta} \right)^2 \), may be accumulated by summation/integration to give an overall variation as \( \int h(\mu) \). For \( \dot{f} \in L^r \) we define our \( L^r \) norm

---

1. This would be a 3\textsuperscript{rd} order array. See Dr. Genevera Allen re. 3\textsuperscript{rd} Order tensor operations
2. Note these are convex!
as \( \|f\|_p = \left( \int |f|^p d\mu \right)^{1/p} \). Considering the norm squared, we have \( \|f\|_2 = \int f^2 d\mu \). For \( h : \mathbb{R}^k \to \mathbb{R} \) we might use \( \int \nabla f \nabla f^T d\mu \).

- Note that in the case of the log likelihood, \( \hat{l}(\theta | x) = \frac{d}{d\theta} \int f(x | \theta) d\theta \) is the RELATIVE variation w.r.t. \( \theta \); using \( h = \left( \frac{d}{d\theta} \right)^2 \), we have \( \int \hat{l}(\theta | x) d\mu = \int \hat{l}(\theta | x) f(x | \theta) dx = E(U^2) = I(\theta) \), where \( U(\theta | x) = \nabla \ell \) is the score function (statistic).

- Note that \( \nabla \ell \nabla \ell^T \) is not equal to \( -\nabla^2 \ell = -\nabla \nabla^T = -H_i(\theta) \), although under regularity their expectations are. E.g., \( f = x_1^2 + x_1 x_2 \Rightarrow \nabla f = \begin{pmatrix} x_1 + x_2 \\ x_1 \end{pmatrix} \); but \( \nabla f \nabla f^T = \begin{bmatrix} (x_1 + x_2)^2 & x_2 (x_1 + x_2) \\
 x_1 (x_1 + x_2) & x_1 x_2 \end{bmatrix} \) is not equal to \( \nabla \nabla f^T = \begin{bmatrix} 1 & 1 \\
 1 & 0 \end{bmatrix} \).

**References**

3. Various vector space and applied analysis books, esp. w.r.t \( \mathbb{R}^n \).