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 R^n

We define an element $x \in R^n$ as a standard *n*x*1* column vector $(x_1, \cdots, x_n)$. We define our matrices as $n \times k$ 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 $g: R^d \to R^k$. We say the derivative of g at $x \in R^d$ is defined to be the linear map: $T \cdot \|h\| = g(x+h) - g(x) + o(\|h\|)$. Other parameterizations are possible; the interpretation of the derivative as a linear map is not optional. Usually ∇_x is the matrix associated with this transformation, and we often denote it variously as $\dot{g}, \nabla g$, $\mathrm{d}g$, or $\frac{\partial g}{\partial x}$. The second derivative at *x* is defined similarly and is denoted $\ddot{g}$, $\nabla^2 g$, $\mathrm{D}g$, or $\frac{\partial^2 g}{\partial x \partial x^T}$.

Definitional Notation

$$\left.\begin{array}{l} g: R \to R \\ \dot{g}: R \to R \\ \ddot{g}: R \to R \end{array}\right\} g = g(x) \qquad \dot{g} = \nabla g = \frac{dg}{dx} \qquad \ddot{g} = \nabla^2 g = \frac{d^2 g}{dx^2}.$$ For example,

$g(x) = a\sin(x), \dot{g}(x) = a\cos(x)$, and $\ddot{g}(x) = -g(x)$.

$$\left.\begin{array}{l} g: R \to R^{(d \times k)^T} = R^k \\ \dot{g}: R^k \to R^{(d \times k)^T} = R^k \\ \ddot{g}: R^k \to R^k \end{array}\right\} g = \begin{pmatrix} g_1(x) \\ \vdots \\ g_k(x) \end{pmatrix} \qquad \dot{g} = \nabla g = \begin{bmatrix} \frac{dg_1(x)}{dx} \\ \vdots \\ \frac{dg_k(x)}{dx} \end{bmatrix} \qquad \ddot{g} = \nabla^2 g = \begin{bmatrix} \frac{d^2 g_1(x)}{dx^2} \\ \vdots \\ \frac{d^2 g_k(x)}{dx^2} \end{bmatrix}.$$ For example, let

$g(x) = xx^T \in R^{dxd}$; we have $\nabla g = \dot{g}(x): R^d \to R^d$, with $\dot{g}(x) = \nabla xx^T = \nabla xx + x\nabla x = 2x$.

$$\left.\begin{array}{l} g: R^d \to R \\ \dot{g}: R^d \to R^{d \times k} = R^d \\ \ddot{g}: R^{d \times k} = R^d \to R^{d \times k \times d} = R^{d \times d} \end{array}\right\} g = g(x_1, \cdots, x_d) \qquad \dot{g} = \nabla g = \begin{bmatrix} \frac{\partial g(x_1, \cdots, x_d)}{\partial x_1} \\ \frac{\partial g(x_1, \cdots, x_d)}{\partial x_2} \\ \vdots \\ \frac{\partial g(x_1, \cdots, x_d)}{\partial x_d} \end{bmatrix}$$

$$\ddot{g} = \nabla^2 g = \nabla\nabla g^T = \begin{bmatrix} \frac{\partial^2 g}{\partial x_1 \partial x_1} & \frac{\partial^2 g}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 g}{\partial x_1 \partial x_d} \\ \frac{\partial^2 g}{\partial x_2 \partial x_1} & \cdots & \cdots & \frac{\partial^2 g}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 g}{\partial x_d \partial x_1} & \frac{\partial^2 g}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 g}{\partial x_d \partial x_d} \end{bmatrix}$$

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

$$\left.\begin{array}{l} g : R^d \to R^k \\ \dot{g} : R^d \to R^{dxk} \\ \ddot{g} : R^{d \times k} \to R^{dxk\,x\,?} \end{array}\right\} g = \begin{pmatrix} g_1(x_1,\cdots,x_d) \\ \vdots \\ g_k(x_1,\cdots,x_d) \end{pmatrix} \quad \dot{g} = \nabla g^T = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_2}{\partial x_1} & \cdots & \frac{\partial g_k}{\partial x_1} \\ \frac{\partial g_1}{\partial x_2} & \cdots & \cdots & \frac{\partial g_k}{\partial x_2} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial g_1}{\partial x_d} & \cdots & \cdots & \frac{\partial g_k}{\partial x_d} \end{bmatrix} \quad \ddot{g} = \nabla^2 g = \nabla[?]$$ [1]

Useful Matrix Derivatives

Chain Rules

- If $f : R^d \to R^s, g : R^s \to R^k$, and $h = g(f(x)) : R^d \to R^k$ then $\dot{h}(x) = \dot{g}(f(x))\dot{f}(x)$
- If both $f, g : R^d \to R^k$, and $h \in R^{k \times k} = f^T(x) g(x)$ 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[2] $\left|\frac{df}{d\theta}\right|$ or $\left(\frac{df}{d\theta}\right)^2$, may be accumulated by summation/integration to give an overall variation as $\int h\mu$. For $\dot{f} \in L^p$ we define our L^p norm

[1] This would be a 3rd order array. See Dr. Genevera Allen re. 3rd Order tensor operations

[2] Note these are convex!

as $\left\|\dot{f}\right\|_p = \left(\int \left|\dot{f}\right|^p d\mu\right)^{\frac{1}{p}}$. Considering the norm squared, we have $\left\|\dot{f}\right\|_2^2 = \int \dot{f}^2 d\mu$. For $h: R^k \to R$ we might use $\int \nabla f \nabla f^T d\mu$.

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

- Note that $\nabla \ell \nabla \ell^T$ is not equal to $-\nabla^2 \ell = -\nabla \nabla \ell^T = -H_\ell(\theta)$, 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

1. Dobelman, J.A., *Helpful Results from Analysis useful for Statistics*, Working Paper (2012), Rice University
2. Cox, D. (2004), *The Theory of Statistics and Its Applications*, Working Edition, Rice University
3. Various vector space and applied analysis books, esp. w.r.t R^n.
4. Gentle, J.E. (200x) A Companion for Mathematical Statstics.