This summary contains a account of the *basic* facts concerning derivatives of multivariate functions, up to the second derivative test for local extrema, with an emphasis on honesty, coherence and the extent to which the multivariate case reflects the univariate case. Recall that entities having several components are usually denoted by bold faced letters (*e.g.*, $\boldsymbol{f}$ for functions, $\mathbb{R}^n \to \mathbb{R}^m$, $\boldsymbol{x}$ for points in $\mathbb{R}^n$ or vectors in $V_n = V_n(\mathbb{R})$, $\hat{\boldsymbol{u}}$ for unit vectors), that $\boldsymbol{h} = |\boldsymbol{h}|\hat{\boldsymbol{h}}$ if $\boldsymbol{h}$ is a non-zero vector, and that for sets A and B, $A \subset B$ means that A is a *subset* of B, *i.e.*, every element of A is an element of B,

Notation. If $U \subset \mathbb{R}^n$ and $\boldsymbol{f} \colon U \to \mathbb{R}^m$, the components of $\boldsymbol{f}$ will be denoted by $f^i \colon \mathbb{R}^n \to \mathbb{R}$, for $1 \leqslant i \leqslant n$; one may write $\boldsymbol{f} = \left(f^1, \dots, f^m\right)$. The reason for using superscripts as indices is that subscripts have been (and will be) used to denote partial derivatives. Superscripts will also be used to index the coordinates of a point, the entries of a vector and the arguments of a multivariate function; *e.g.*, in the notation used here $f^i_j = \partial f^i / \partial x^j$ is the partial derivative of the i^{th} component of $\boldsymbol{f}$, f^i, with respect to its j^{th} argument, x^j, for $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$. The notation should not cause any confusion because exponents will be used rarely, and this use should be clear from the context; *e.g.*, in (14), where the base is a complex expression enclosed in braces.

The standard basis vectors for V_n will be denoted by $\hat{\boldsymbol{e}}_1, \dots, \hat{\boldsymbol{e}}_n$, where $\hat{\boldsymbol{e}}_j$ has a 1 in position j and a 0 in every other position. Remember that the angle bracket notation for vectors is just a way of expressing (column) vectors using horizontal notation: $\left\langle \alpha^1, \dots, \alpha^n \right\rangle = \begin{pmatrix}\alpha^1 & \dots & \alpha^n\end{pmatrix}^t \in V_n$.

An element $\boldsymbol{x} \in \mathbb{R}^n$ belongs to the *interior* of $U \subset \mathbb{R}^n$, in symbols $\boldsymbol{x} \in U^\circ$, if there is a positive number ε such that $\{\, \boldsymbol{y} \in \mathbb{R}^n \colon |\boldsymbol{y} - \boldsymbol{x}| < \varepsilon \,\} \subset U$. This means that a point in $\mathbb{R}^n$ belongs to U provided it is sufficiently close to $\boldsymbol{x}$. This situation is also expressed by saying that U is a *neighbourhood* of $\boldsymbol{x}$. Notice that any function continuous at $\boldsymbol{x}$ is defined on some neighbourhood of $\boldsymbol{x}$.

The derivative. Suppose that $U \subset \mathbb{R}^n$, $\boldsymbol{f} \colon U \to \mathbb{R}^m$, and that $\boldsymbol{x}$ is a point in the interior of U. $\boldsymbol{f}$ is *differentiable* at $\boldsymbol{x}$ if there is a linear transformation $\boldsymbol{f}'(\boldsymbol{x}) \colon V_n \to V_m$, called the *derivative of* $\boldsymbol{f}$ *at* $\boldsymbol{x}$, such that

$$\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{h}) + |\boldsymbol{h}|\,\boldsymbol{\delta}, \tag{1}$$

and $\boldsymbol{\delta} \to \boldsymbol{0}$ as $\boldsymbol{h} \to \boldsymbol{0}$. $\boldsymbol{\delta}$ is a function of $\boldsymbol{x}$ and $\boldsymbol{h}$, and is defined whenever $\boldsymbol{x} + \boldsymbol{h} \in U$; if it is necessary to record the dependence of $\boldsymbol{\delta}$ on $\boldsymbol{x}$ and $\boldsymbol{h}$, the notation $\boldsymbol{\delta}_{\boldsymbol{x}}(\boldsymbol{h})$ will be used. The derivative of $\boldsymbol{f}$ at $\boldsymbol{x}$ is sometimes written $D\boldsymbol{f}(\boldsymbol{x})$. It was proved in class that there is at most one linear transformation satisfying the condition defining $\boldsymbol{f}'(\boldsymbol{x})$, *i.e.*, the derivative of $\boldsymbol{f}$ at $\boldsymbol{x}$, if it exists, is unique. The reason for this will be recalled. Suppose that a linear transformation $T(\boldsymbol{x}) \colon V_n \to V_m$ satisfies the condition defining $\boldsymbol{f}'(\boldsymbol{x})$, with $\boldsymbol{\varepsilon}$ in place of $\boldsymbol{\delta}$. It is clear that $T(\boldsymbol{x})(\boldsymbol{0}) = \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{0}) = \boldsymbol{0}$, and for any $\boldsymbol{v} \neq \boldsymbol{0}$ in V_n and $t \neq 0$ in $\mathbb{R}$,

$$\begin{aligned}\frac{|T(\boldsymbol{x})(\boldsymbol{v}) - \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{v})|}{|\boldsymbol{v}|} &= \frac{|T(\boldsymbol{x})(t\boldsymbol{v}) - \boldsymbol{f}'(\boldsymbol{x})(t\boldsymbol{v})|}{|t\boldsymbol{v}|} \\ &= |\boldsymbol{\delta}_{\boldsymbol{x}}(t\boldsymbol{v}) - \boldsymbol{\varepsilon}_{\boldsymbol{x}}(t\boldsymbol{v})| \to \boldsymbol{0}\end{aligned}$$

as $t \to 0$, so $T(\boldsymbol{x})(\boldsymbol{v}) = \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{v})$, since the left hand side of the equation is independent of t.

Notice that $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$ if, and only if, f^i is differentiable for $1 \leqslant i \leqslant m$. This is because (1) is true if, and only if, it is true for each of the (m) coordinates of the points involved, and because $\boldsymbol{\delta} = \left\langle \delta^1, \dots, \delta^m \right\rangle \to \boldsymbol{0}$ if, and only if, $\delta^i \to 0$ for $1 \leqslant i \leqslant m$. When f is real-valued (*i.e.*, $f \colon U \to \mathbb{R}$) it is sometimes customary to replace the last term in (1), which in this case is of the form $|\boldsymbol{h}|\delta$, by a sum of terms of the form $\eta^j \delta^j$, where $\boldsymbol{h} = \left\langle \eta^1, \dots, \eta^n \right\rangle$, and require that each $\delta^j \to 0$ as $\boldsymbol{h} \to 0$. The notions are clearly equivalent: given δ let $\delta^j = \eta^j \delta / |\boldsymbol{h}|$; given δ^j, for $1 \leqslant j \leqslant n$, let $\delta = \sum_{j=1}^n \eta^j \delta^j / |\boldsymbol{h}|$ (for $\boldsymbol{h} \neq \boldsymbol{0}$).

It is left as a (straightforward) exercise to show that differentiation is linear in $\boldsymbol{f}$, just as for functions of a single real variable: If $\boldsymbol{f}$ and $\boldsymbol{g}$ are differentiable at $\boldsymbol{x}$ and $\alpha, \beta \in \mathbb{R}$, then $\alpha\boldsymbol{f} + \beta\boldsymbol{g}$ is differentiable at $\boldsymbol{x}$ and

$$(\alpha\boldsymbol{f} + \beta\boldsymbol{g})'(\boldsymbol{x}) = \alpha\boldsymbol{f}'(\boldsymbol{x}) + \beta\boldsymbol{g}'(\boldsymbol{x}).$$

The following is really an easy result of basic linear algebra.

2**. Lemma.** *If $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$ then there is a real number M such that $|\boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{v})| \leqslant M|\boldsymbol{v}|$ for all $\boldsymbol{v} \in V_n$. In particular, $\boldsymbol{f}'(\boldsymbol{x})$ is continuous on V_n.*

Proof. Let $M = nK$, where K is the maximum of $|\boldsymbol{f}'(\boldsymbol{x})(\hat{\boldsymbol{e}}_j)|$, for $1 \leqslant j \leqslant n$; then for $\boldsymbol{v} = \left\langle \beta^1, \dots, \beta^n \right\rangle \in V_n$,

$$|\boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{v})| = \Big|\sum_{j=1}^n \beta^j \boldsymbol{f}'(\boldsymbol{x})(\hat{\boldsymbol{e}}_j)\Big| \leqslant \sum_{j=1}^n |\beta^j \boldsymbol{f}'(\boldsymbol{x})(\hat{\boldsymbol{e}}_j)| \leqslant \sum_{j=1}^n K|\boldsymbol{v}| = M|\boldsymbol{v}|.$$

□

With the real notion of the derivative of a multivariate function in hand, a fundamental connection between differentiability and continuity for univariate functions—which turned out to be false for partial derivatives—can be recovered.

3**. Theorem.** *If $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$ then $\boldsymbol{f}$ is continuous at $\boldsymbol{x}$.*

Proof. It is required to show that $\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{h}) \to \boldsymbol{f}(\boldsymbol{x})$ as $\boldsymbol{h} \to \boldsymbol{0}$ if $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$. Referring to (1), $\boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{h}) \to \boldsymbol{0}$ as $\boldsymbol{h} \to \boldsymbol{0}$ by lemma 2, and $|\boldsymbol{h}|\,\boldsymbol{\delta} \to \boldsymbol{0}$ as $\boldsymbol{h} \to \boldsymbol{0}$ by definition (since $\boldsymbol{\delta} \to \boldsymbol{0}$). Therefore, the limit of the right hand side of (1) as $\boldsymbol{h} \to \boldsymbol{0}$ is $\boldsymbol{f}(\boldsymbol{x})$, as required. □

Directional derivatives. There are concepts intermediate between the derivative of $\boldsymbol{f}$ and its partial derivatives, which turn out to be of some independent interest. If $\boldsymbol{v} \in V_n$ and $\boldsymbol{f}(\boldsymbol{x} + t\boldsymbol{v})$ is defined for all t in an open interval containing 0 then

$$\boldsymbol{f}_{\boldsymbol{v}}(\boldsymbol{x}) = \lim_{t \to 0} \frac{\boldsymbol{f}(\boldsymbol{x} + t\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{x})}{t} = \frac{d}{dt}\boldsymbol{f}(\boldsymbol{x} + t\boldsymbol{v})\Big|_{t=0} \tag{4}$$

If $\hat{\boldsymbol{u}}$ is a unit vector in V_n then $\boldsymbol{f}_{\hat{\boldsymbol{u}}}(\boldsymbol{x})$ is denoted by $D_{\hat{\boldsymbol{u}}}\boldsymbol{f}(\boldsymbol{x})$ and is called the *directional derivative of* $\boldsymbol{f}$ *in the direction* of $\hat{\boldsymbol{u}}$. (4) is called the directional derivative of $\boldsymbol{f}$ *along* $\boldsymbol{v}$ or, of $\boldsymbol{f}$ *with respect to* $\boldsymbol{v}$ but, as such usage is uncommon in children's books, that will not be done here. It follows directly from (4) and the definition of partial derivatives that, for $1 \leqslant j \leqslant m$,

$$\boldsymbol{f}_{\hat{e}_j} = \boldsymbol{f}_j = \left\langle\, f^1_j, f^2_j, \dots, f^m_j \,\right\rangle = \left\langle \frac{\partial f^1}{\partial x^j}, \frac{\partial f^2}{\partial x^j}, \dots, \frac{\partial f^m}{\partial x^j} \right\rangle. \tag{5}$$

Where $\boldsymbol{f}$ is differentiable, the $\boldsymbol{f}_{\boldsymbol{v}}$ are precisely the values of the derivative of $\boldsymbol{f}$; *i.e.*,

$$\boldsymbol{f}_{\boldsymbol{v}}(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{v}). \tag{6}$$

for $\boldsymbol{v} \in V_n$. This is seen by using the definition (4) and putting $t\boldsymbol{v}$ for $\boldsymbol{h}$ in (1):

$$\begin{aligned}\boldsymbol{f}_{\boldsymbol{v}}(\boldsymbol{x}) &= \lim_{t \to 0} \frac{\boldsymbol{f}'(\boldsymbol{x})(t\boldsymbol{v}) + |tv|\boldsymbol{\delta}}{t} = \lim_{t \to 0} \frac{t\boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{v}) + |tv|\boldsymbol{\delta}}{t} \\ &= \lim_{t \to 0} \left\{ \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{v}) + \frac{|t\boldsymbol{v}|}{t}\,\boldsymbol{\delta} \right\} = \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{v}),\end{aligned}$$

since $||t\boldsymbol{v}|/t| = |\boldsymbol{v}|$ (for $t \neq 0$) and $\boldsymbol{\delta} \to \boldsymbol{0}$ as $t \to 0$.

$\boldsymbol{f}'$, when it exists, can be expressed in terms the partial derivatives of $\boldsymbol{f}$.

7**. Theorem.** *If $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$ and $\boldsymbol{v} = \left\langle \beta^1, \dots, \beta^n \right\rangle$ then*

$$\boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{v}) = \sum_{j=1}^n \beta^j \boldsymbol{f}_j(\boldsymbol{x}).$$

Proof. The proof is a routine calculation using the linearity of $\boldsymbol{f}'(\boldsymbol{x})$, (5) and (6):

$$\boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{v}) = \sum_{j=1}^n \beta^j \boldsymbol{f}'(\boldsymbol{x})(\hat{\boldsymbol{e}}_j) = \sum_{j=1}^n \beta^j \boldsymbol{f}_{\hat{\boldsymbol{e}}_j}(\boldsymbol{x}) = \sum_{j=1}^n \beta^j \boldsymbol{f}_j(\boldsymbol{x})$$ □

The last theorem implies that, where $\boldsymbol{f}$ is differentiable, the derivative of $\boldsymbol{f}$ can be represented, with respect to the standard bases on V_n and V_m, by the matrix $\begin{pmatrix}\boldsymbol{f}_1 & \boldsymbol{f}_2 & \dots & \boldsymbol{f}_m\end{pmatrix}$, *i.e.*,

$$\begin{pmatrix} f^1_1 & f^1_2 & \dots & f^1_n \\ f^2_1 & f^2_2 & \dots & f^2_n \\ \vdots & \vdots & \ddots & \vdots \\ f^m_1 & f^m_2 & \dots & f^m_n \end{pmatrix} = \begin{pmatrix} \dfrac{\partial f^1}{\partial x^1} & \dfrac{\partial f^1}{\partial x^2} & \dots & \dfrac{\partial f^1}{\partial x^n} \\ \dfrac{\partial f^2}{\partial x^1} & \dfrac{\partial f^2}{\partial x^2} & \dots & \dfrac{\partial f^2}{\partial x^n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f^m}{\partial x^1} & \dfrac{\partial f^m}{\partial x^2} & \dots & \dfrac{\partial f^m}{\partial x^n} \end{pmatrix}$$

(where the dependence of this matrix on $\boldsymbol{x}$ has been suppressed). This matrix is called the *Jacobian matrix* of $\boldsymbol{f}$ (at $\boldsymbol{x}$); there is no universally standard notation for it, and it is sometimes denoted by one of the symbols used for the derivative of $\boldsymbol{f}$: $\boldsymbol{f}'$ or $D\boldsymbol{f}$.

In particular, if $f\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\boldsymbol{x}$ and $\boldsymbol{v} = \langle \beta^1, \ldots, \beta^n \rangle$, then

$$f'(\boldsymbol{x})(\boldsymbol{v}) = \nabla f(\boldsymbol{x}) \cdot \boldsymbol{v} = \sum_{j=1}^{n} \beta^j \frac{\partial f}{\partial x^j}(\boldsymbol{x}), \tag{8}$$

where $\nabla f(\boldsymbol{x}) = \langle f_1(\boldsymbol{x}), \ldots, f_n(\boldsymbol{x}) \rangle = \langle (\partial f/\partial x_1)(\boldsymbol{x}), \ldots, (\partial f/\partial x_n)(\boldsymbol{x}) \rangle$ is called the *gradient* (vector) of f at $\boldsymbol{x}$. From (8) it is obvious that $\nabla f(\boldsymbol{x})$ points in the direction of greatest increase of f (starting at $\boldsymbol{x}$); *i.e.*, the largest value of $D_{\hat{\boldsymbol{u}}} f(\boldsymbol{x})$ is $|\nabla f(\boldsymbol{x})|$, which occurs when $\hat{\boldsymbol{u}} = \nabla f(\boldsymbol{x})/|\nabla f(\boldsymbol{x})|$ provided $\nabla f(\boldsymbol{x}) \neq \boldsymbol{0}$ (otherwise, all directional derivatives of f are zero).

When the Jacobian matrix of $\boldsymbol{f}$ is square, *i.e.*, when $m = n$, its determinant is called the *Jacobian determinant* of $\boldsymbol{f}$, or just the *Jacobian* of $\boldsymbol{f}$, and is denoted by $J(\boldsymbol{f})$ or

$$\frac{\partial(f^1, f^2, \ldots, f^n)}{\partial(x^1, x^2, \ldots, x^n)}.$$

The Jacobian determinant of $\boldsymbol{f}$ will turn out to be of fundamental importance for the calculus of several functions of several variables.

The differentiability of $\boldsymbol{f}$ is often most easily checked using the following

9. **Theorem.** *If $\boldsymbol{f}_j$ is continuous at $\boldsymbol{x}$, for $1 \leqslant j \leqslant n$, then $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$.*

Proof. By the remark in the first paragraph following the definition of the derivative, is sufficient to consider a function $f\colon U \to \mathbb{R}$, where $U \subset \mathbb{R}^n$ is a neighbourhood of $\boldsymbol{x}$. Let $\boldsymbol{h} = \langle \eta^1, \ldots, \eta^n \rangle \neq \boldsymbol{0}$ be such that f_j, $1 \leqslant j \leqslant n$, is defined on $\{\, \boldsymbol{y} \in \mathbb{R}^n \colon |\boldsymbol{y} - \boldsymbol{x}| < |\boldsymbol{h}| \,\}$, and let $\boldsymbol{h}_j = \sum_{i=1}^{j} \eta^i \hat{\boldsymbol{e}}_i$; note that $\boldsymbol{h}_0 = \boldsymbol{0}$, $\boldsymbol{h}_n = \boldsymbol{h}$, and $\boldsymbol{h}_j - \boldsymbol{h}_{j-1} = \eta^j \hat{\boldsymbol{e}}_j$ for $1 \leqslant j \leqslant n$. The mean value theorem, applied to $f(\boldsymbol{x} + \boldsymbol{h}_{j-1} + t\eta^j \hat{\boldsymbol{e}}_j)$ for $t \in [\,0, 1\,]$, yields real numbers ϑ_j such that $0 < \vartheta_j < 1$, and $f(\boldsymbol{x} + \boldsymbol{h}_j) - f(\boldsymbol{x} + \boldsymbol{h}_{j-1}) = \eta^j f_j(\boldsymbol{x} + \boldsymbol{h}_{j-1} + \vartheta_j \eta^j \hat{\boldsymbol{e}}_j)$, for $1 \leqslant j \leqslant n$. Hence,

$$\begin{aligned} f(\boldsymbol{x} + \boldsymbol{h}) &= f(\boldsymbol{x}) + \sum_{j=1}^{n} \bigl\{ f(\boldsymbol{x} + \boldsymbol{h}_j) - f(\boldsymbol{x} + \boldsymbol{h}_{j-1}) \bigr\} \\ &= f(\boldsymbol{x}) + \sum_{j=1}^{n} \eta^j f_j(\boldsymbol{x} + \boldsymbol{h}_{j-1} + \vartheta_j \eta^j \hat{\boldsymbol{e}}_j) \\ &= f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}) \cdot \boldsymbol{h} + |\boldsymbol{h}| \sum_{j=1}^{n} \delta_j, \end{aligned}$$

where $\delta_j = \bigl(\eta^j/|\boldsymbol{h}|\bigr) \bigl\{ f_j(\boldsymbol{x} + \boldsymbol{h}_{j-1} + \vartheta_j \eta^j \hat{\boldsymbol{e}}_j) - f(\boldsymbol{x}) \bigr\}$ for $1 \leqslant j \leqslant n$. Since $|\eta^j/|\boldsymbol{h}|| \leqslant 1$ and f_j is continuous at $\boldsymbol{x}$, $\delta_j \to 0$ as $\boldsymbol{h} \to \boldsymbol{0}$. Therefore, $\delta = \sum_{j=1}^{n} \delta_j \to 0$ as $\boldsymbol{h} \to \boldsymbol{0}$, so f is differentiable at $\boldsymbol{x}$ (with $f'(\boldsymbol{x})$ given by the dot product with $\nabla f(\boldsymbol{x})$, as expected). $\square$

The chain rule for multivariate functions takes on a more conceptual form (the derivative of a composite is the composite of the derivatives) and plays a more fundamental role (*e.g.*, multiplication is a differentiable function of two variables, so the product rule is a consequence of the chain rule) than its single variable counterpart.

10. **Theorem** (The chain rule). *Suppose that $f\colon U \to \mathbb{R}^m$ is differentiable at $\boldsymbol{x}$ and $g\colon V \to \mathbb{R}^p$ is differentiable at $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$, where $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$; then $\boldsymbol{g} \circ \boldsymbol{f}$ is differentiable at $\boldsymbol{x}$ and*

$$(\boldsymbol{g} \circ \boldsymbol{f})'(\boldsymbol{x}) = \boldsymbol{g}'(\boldsymbol{y}) \circ \boldsymbol{f}'(\boldsymbol{x}).$$

Proof. Referring to the right hand side of (1), with $\boldsymbol{\delta} = \boldsymbol{\delta}_{\boldsymbol{x}}(\boldsymbol{h})$, let

$$\Delta \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{h}) = \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{h}) + |\boldsymbol{h}|\, \boldsymbol{\delta} \quad \text{and} \quad \Gamma \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{h}) = \boldsymbol{f}'(\boldsymbol{x})(\hat{\boldsymbol{h}}) + \boldsymbol{\delta},$$

so that $\Delta \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{h}) = |\boldsymbol{h}|\, \Gamma \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{h})$. The the differentiability assumptions yield (where $\boldsymbol{\varepsilon} \to \boldsymbol{0}$ as $\Delta \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{h}) \to \boldsymbol{0}$ by the differentiability of $\boldsymbol{g}$ at $\boldsymbol{y}$),

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{h})) &= \boldsymbol{g}(\boldsymbol{y} + \Delta \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{h})) \\ &= \boldsymbol{g}(\boldsymbol{y}) + \boldsymbol{g}'(\boldsymbol{y})(\Delta \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{h})) + |\Delta \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{h})|\, \boldsymbol{\varepsilon} \\ &= \boldsymbol{g}(\boldsymbol{y}) + \boldsymbol{g}'(\boldsymbol{y})(\boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{h})) + \boldsymbol{g}'(\boldsymbol{y})(|\boldsymbol{h}|\, \boldsymbol{\delta}) + |\Delta \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{h})|\, \boldsymbol{\varepsilon} \\ &= \boldsymbol{g}(\boldsymbol{y}) + \boldsymbol{g}'(\boldsymbol{y}) \circ \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{h}) + |\boldsymbol{h}|\, \{ \boldsymbol{g}'(\boldsymbol{y})(\boldsymbol{\delta}) + |\Gamma \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{h})|\, \boldsymbol{\varepsilon} \} \end{aligned}$$

Now as $\boldsymbol{h} \to 0$, $\boldsymbol{\delta} \to \boldsymbol{0}$ by the differentiability of $\boldsymbol{f}$ at $\boldsymbol{x}$, both $\boldsymbol{g}'(\boldsymbol{y})(\boldsymbol{\delta}) \to \boldsymbol{0}$ and $\Delta \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{h}) \to \boldsymbol{0}$ by lemma 2, and thus $\boldsymbol{\varepsilon} \to \boldsymbol{0}$ by the differentiability of $\boldsymbol{g}$ at $\boldsymbol{y}$. Again by lemma 2, there is a real number M such that $|\Gamma \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{h})| \leqslant M + |\boldsymbol{\delta}|$, from which it follows that $\boldsymbol{g}'(\boldsymbol{y})(\boldsymbol{\delta}) + |\Gamma \boldsymbol{f}(\boldsymbol{x})(\boldsymbol{h})|\, \boldsymbol{\varepsilon} \to \boldsymbol{0}$ as $\boldsymbol{h} \to \boldsymbol{0}$, and the theorem is proved. $\square$

Recall the following instance of the Taylor development, with the Lagrange form of the remainder, of a univariate real-valued function φ which is twice differentiable on an interval I containing x and $x + h$:

$$\varphi(x + h) = \varphi(x) + \varphi'(x)h + \tfrac{1}{2}\varphi''(x + \vartheta h)h^2, \tag{11}$$

for some real number ϑ such that $0 < \vartheta < 1$. A version of this development, for multivariate functions, will now be given.

Theorem. *If $U \subset \mathbb{R}^n$, and $f\colon U \to \mathbb{R}$ and its partial derivatives f_i, for $1 \leqslant i \leqslant n$, are differentiable on the segment $\{\, \boldsymbol{x} + t\boldsymbol{h} \colon 0 \leqslant t \leqslant 1 \,\}$, where $\boldsymbol{h} = \langle \eta^1, \ldots, \eta^n \rangle \in V_n$, then there is a real number ϑ such that $0 < \vartheta < 1$ and*

$$f(\boldsymbol{x} + \boldsymbol{h}) = f(\boldsymbol{x}) + f'(\boldsymbol{x})(\boldsymbol{h}) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \eta^i \eta^j f_{ij}(\boldsymbol{x} + \vartheta \eta). \tag{12}$$

Proof. Let $\varphi(t) = f(\boldsymbol{x} + t\boldsymbol{h})$, $0 \leqslant t \leqslant 1$, so that (11) yields

$$f(\boldsymbol{x} + \boldsymbol{h}) = f(\boldsymbol{x}) + \varphi'(0) + \tfrac{1}{2}\varphi''(\vartheta),$$

for some real number ϑ such that $0 < \vartheta < 1$. By the chain rule,

$$\varphi'(t) = f'(\boldsymbol{x} + t\boldsymbol{h})(\boldsymbol{h}) = \sum_{i=1}^{n} \eta^i f_i(\boldsymbol{x} + t\boldsymbol{h}),$$

so that $\varphi'(0) = f'(\boldsymbol{x})(\boldsymbol{h})$ is the second term of the right hand side of (12), and

$$\begin{aligned} \varphi''(t) &= \sum_{i=1}^{n} \frac{d}{dt}\bigl(\eta^i f_i(\boldsymbol{x} + t\boldsymbol{h})\bigr) = \sum_{i=1}^{n} \eta^i f_i'(\boldsymbol{x} + t\boldsymbol{h})(\boldsymbol{h}) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \eta^i \eta^j f_{ij}(\boldsymbol{x} + t\boldsymbol{h}), \end{aligned}$$

so that $\frac{1}{2}\varphi''(\vartheta)$ is the last term on the right hand side of (12). $\square$

The previous theorem can be generalized to give a full Taylor development for multivariate functions—if you want to do a module assignment on this let, me know—although (12) suffices for the present ends. Two applications of (12) will be given. The first is the linear approximation of a differentiable function, together with an error bound. The second is a second derivative test for local extrema of multivariate real-valued functions.

The *linearization* of f near $\boldsymbol{x}$ is given by

$$L_f(\boldsymbol{x} + \boldsymbol{h}) = f(\boldsymbol{x}) + f'(\boldsymbol{x})(\boldsymbol{h}). \tag{13}$$

It follows from (12) that if $|f_{ij}(\boldsymbol{y})| \leqslant M$ for $1 \leqslant i, j \leqslant n$, and $|y^i - x^i| \leqslant |\eta^i|$ for $1 \leqslant i \leqslant n$, then the (absolute value of the) error when approximating $f(\boldsymbol{y})$ by $L_f(\boldsymbol{y}) = f(\boldsymbol{x}) + f'(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})$ is at most

$$\tfrac{1}{2} M \biggl\{ \sum_{i=1}^{n} |\eta^i| \biggr\}^2. \tag{14}$$

It is not uncommon to use the so-called *total differential* of f,

$$df = \sum_{i=1}^{n} f_i(\boldsymbol{x}) dx^i = \frac{\partial f}{\partial x^1} dx^1 + \frac{\partial f}{\partial x^2} dx^2 + \cdots + \frac{\partial f}{\partial x^n} dx^n,$$

as a version of $\boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{h})$, *i.e.*as an approximation to the change in $f(\boldsymbol{x})$ corresponding to a change in $\boldsymbol{x}$. When estimating an actual change in the value of f, an actual increment $\boldsymbol{h} = \langle \eta^1, \ldots, \eta^n \rangle$ is used in place of $\langle dx^1, \ldots, dx^n \rangle$, as in the rightmost term of (13), in which case the error involved in such an estimate is bounded by (14) under the given hypotheses. The significance of (14) lies with the obvious fact that an estimate is essentially worthless unless it is accompanied by a bound on the error involved.

If the real-valued function f has all second order partial derivatives at $\boldsymbol{x} \in \mathbb{R}^n$, then the $n \times n$ matrix of second order partial derivatives

$$H_f(\boldsymbol{x}) = \begin{pmatrix} f_{11}(\boldsymbol{x}) & f_{12}(\boldsymbol{x}) & \ldots & f_{1n}(\boldsymbol{x}) \\ f_{21}(\boldsymbol{x}) & f_{22}(\boldsymbol{x}) & \ldots & f_{2n}(\boldsymbol{x}) \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1}(\boldsymbol{x}) & f_{n2}(\boldsymbol{x}) & \ldots & f_{nn}(\boldsymbol{x}) \end{pmatrix}$$

is called the *Hessian* matrix of f (at $\boldsymbol{x}$), and the quadratic form $\mathscr{H}_f(\boldsymbol{x})$ defined by

$$\mathscr{H}_f(\boldsymbol{x})(\boldsymbol{h}) = \boldsymbol{h}^t H_f(\boldsymbol{x})\, \boldsymbol{h} = \sum_{i=1}^{n} \sum_{j=1}^{n} \eta^i \eta^j f_{ij}(\boldsymbol{x})$$

is called the *Hessian* (form) of f (at $\boldsymbol{x}$). Note that (12) can be written as

$$f(\boldsymbol{x} + \boldsymbol{h}) = f(\boldsymbol{x}) + f'(\boldsymbol{x})(\boldsymbol{h}) + \tfrac{1}{2}\mathscr{H}(\boldsymbol{x} + \vartheta \boldsymbol{h})(\boldsymbol{h}). \tag{15}$$

15. **(In)definiteness.** The next application of (12), or (15), will require some basic linear algebra. The *quadratic form* $\mathscr{A}$ associated to an $n \times n$ matrix $A = (\alpha_{ij})$ is given by

$$\mathscr{A}(\boldsymbol{v}) = \boldsymbol{v}^t A\, \boldsymbol{v} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij} \beta^i \beta^j$$

for $\boldsymbol{v} = \langle \beta^1, \ldots, \beta^n \rangle \in V_n$ (it is called a quadratic form because each term of the rightmost expression is a quadratic–*i.e.*, degree 2—function of the entries of $\boldsymbol{v}$). Notice that replacing the entries α_{ij} and α_{ji} of A by $\frac{1}{2}(\alpha_{ij}+\alpha_{ji})$ defines the same quadratic form $\mathscr{A}$, so there is no loss of generality in assuming that A is *symmetric* (*i.e.*, $A^t = A$), at least as far as the study of $\mathscr{A}$ is concerned. $\mathscr{A}$ is *positive definite* if $\mathscr{A}(\boldsymbol{v}) > 0$ whenever $\boldsymbol{v} \neq \boldsymbol{0}$, *negative definite* if $\mathscr{A}(\boldsymbol{v}) < 0$ whenever $\boldsymbol{v} \neq \boldsymbol{0}$, and *indefinite* if there are $\boldsymbol{u}, \boldsymbol{v}$ such that $\mathscr{A}(\boldsymbol{u}) < 0 < \mathscr{A}(\boldsymbol{v})$. (It should be clear that these notions do not exhaust the possible behaviour of a quadratic form. There are manifest notions of positive, and negative, semidefinite quadratic forms which, though unnecessary to formulate and prove the second derivative test, would play an evident role in a systematic treatment of local extrema of multivariate real-valued functions.) The terms positive definite, negative definite and indefinite are applied to the matrix A, with the same meaning. Notice that A is negative definite if, and only if, $-A$ is positive definite.

For an $n \times n$ matrix M, and $1 \leqslant p \leqslant n$, M_p denotes the matrix obtained from M by deleting the bottom $n-p$ rows and the rightmost $n-p$ columns, and $\mu_p(M)$ denotes the determinant of M_p; so, *e.g.*, $\mu_1(M)$ is the top left entry of M and $\mu_n(M) = \det(M)$. The numbers $\mu_p(M)$ are called the leading principal minors of M. In general, a *principal minor* of M is the determinant of a square matrix obtained by deleting at least one, and fewer than n, same-indexed rows and columns of M (*i.e.*, row i is deleted if, and only if, column i is deleted, for $1 \leqslant i \leqslant n$), and the *order* of a principal minor is the number of rows (equivalently, columns) of the matrix of which it is the determinant. Observe that for A to be (positive or negative) definite, it is necessary that A_p be nonsingular, (*i.e.*, $\mu_p(A) \neq 0$) for $1 \leqslant p \leqslant n$.

Suppose that A is symmetric and that $\mu_p(A) \neq 0$, for $1 \leqslant p \leqslant n$. It is easy to verify that there is a lower triangular $n \times n$ matrix L, each of whose diagonal entries is a 1, such that $LAL^t = D$ is an invertible diagonal matrix, and $L_p A_p {L_p}^t = D_p$ for $1 \leqslant p \leqslant n$ (L is the book-keeping matrix for the Gaussian elimination process that puts A into upper triangular form). Now, A is positive definite if, and only if, D is positive definite, because L is invertible, and D is positive definite if, and only if, each of its diagonal entries is positive. Since $\mu_p(D)$ is the product of the first p diagonal entries of D, and $\mu_p(L) = 1$, it follows that A is positive definite if, and only if, each of its leading principal minors is positive (*i.e.*, $\mu_p(A) > 0$ for $1 \leqslant p \leqslant n$). Also, A is negative definite if, and only if, its leading principal minors are alternately negative and positive (*i.e.*, $(-1)^p \mu_p(A) > 0$ for $1 \leqslant p \leqslant n$). This is a consequence of the fact that A is negative definite if, and only if, $-A$ is positive definite, and that $\mu_p(A)$ is a sum of products of p entries of A (of A_p, actually, but that extra information is not needed). Finally, if the leading principal minors do not fall into one of these patterns (*i.e.*, either all positive, or alternately negative and positive), then A is indefinite.

The results of the previous paragraph do not characterize indefinite quadratic forms as required by the second derivative test, although handling the additional cases is but a slight headache. If some of the leading principal minors of A are zero, then it may be necessary to rearrange the rows and (corresponding) columns of A before carrying out the (row and column) elimination recorded by L, and D may have zeros on its diagonal. The additional complication, then, is the need to consider all principal minors, and not just the leading ones. Incorporating these additional considerations yields the following criterion. A is indefinite if, and only if, there is a negative principal minor of A of even order, or there is a negative and a positive principal minor of A, both of odd order.

Local extrema. In preparation for the second derivative test, some definitions and a basic result will be given. Suppose that $f\colon U \to \mathbb{R}$, where $U \subset \mathbb{R}^n$ is a neighbourhood of $\boldsymbol{x}$. f has a *local minimum at* $\boldsymbol{x}$ if there is a neighbourhood $W \subset U$ of $\boldsymbol{x}$ such that $f(\boldsymbol{x}) \leqslant f(\boldsymbol{y})$ for $\boldsymbol{y} \in W$. The notion of local maximum is defined similarly: f has a *local maximum at* $\boldsymbol{x}$ if there is a neighbourhood $W \subset U$ of $\boldsymbol{x}$ such that $f(\boldsymbol{y}) \leqslant f(\boldsymbol{x})$ for $\boldsymbol{y} \in W$. Remember that an *extremum* is a maximum or a minimum, and that a local extremum is *strict* if the inequality can be made strict for every $\boldsymbol{y} \neq \boldsymbol{x}$ in some neighbourhood of $\boldsymbol{x}$. Finally, f has a *saddle point at* $\boldsymbol{x}$ if, for every neighbourhood $W \subset U$ of $\boldsymbol{x}$, there are $\boldsymbol{y}, \boldsymbol{z} \in W$ such that $f(\boldsymbol{y}) < f(\boldsymbol{x}) < f(\boldsymbol{z})$.

The following is an analogue of the Fermat criterion for local extrema from single variable calculus:

16**. Theorem.** *Suppose that f is continuous at $\boldsymbol{x}$, and has a local extremum at $\boldsymbol{x}$. Then $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$ or $f_i(\boldsymbol{x})$ is undefined for some i, $1 \leqslant i \leqslant n$.*

Proof. Define $\varphi(t) = f(\boldsymbol{x} + t\hat{\mathbf{e}}_i)$, so that φ has a local extremum at $t = 0$. By the Fermat criterion for local extrema of univariate functions, $f_i(\boldsymbol{x}) = \varphi'(0) = 0$, provided $f_i(\boldsymbol{x})$ exists. □

The Fermat criterion motivates the notions of critical and stationary points of a multivariate function. $\boldsymbol{x}$ is a *stationary point* of f if $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$, and $\boldsymbol{x}$ is a *critical point* of f if $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$ or $f_i(\boldsymbol{x})$ is undefined for some i, $1 \leqslant i \leqslant n$. As in single variable calculus, the Fermat criterion insures that where f is continuous its local extrema occur at critical points. Also as in single variable calculus, f can have a critical point at $\boldsymbol{x}$ without having a local extremum at $\boldsymbol{x}$. However, multivariate functions can exhibit behaviour that univariate functions do not, as illustrated by the function $f(x, y) = xy$: the intersection of the graph of f and the plane $y = x$ has a local minimum at the origin, and the intersection of the graph of f and the plane $y = -x$ has a local maximum at the origin. Sketching the graph of $z = xy$ reveals why the origin is called a saddle point of f.

Suppose now that $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$ and that all second order partial derivatives of f are continuous at $\boldsymbol{x}$. If $\mathscr{H}_f(\boldsymbol{x})$ is positive definite then, since the entries of $H_f(\boldsymbol{x})$ are continuous at $\boldsymbol{x}$, there is a positive real number ε such that $\mu_p(H_f(\boldsymbol{y})) > 0$ for $1 \leqslant p \leqslant n$, and hence $\mathscr{H}_f(\boldsymbol{y})$ is positive definite, provided $|\boldsymbol{y} - \boldsymbol{x}| < \varepsilon$. Therefore, by (12), there is a real number ϑ such that $0 < \vartheta < 1$ and

$$f(\boldsymbol{x} + \boldsymbol{h}) = f(\boldsymbol{x}) + \tfrac{1}{2}\mathscr{H}_f(\boldsymbol{x} + \vartheta\boldsymbol{h}) > f(\boldsymbol{x})$$

if $0 < |\boldsymbol{h}| < \varepsilon$, so f has a (strict) local minimum at $\boldsymbol{x}$. Similarly, if $\mathscr{H}_f(\boldsymbol{x})$ is negative definite, f has a (strict) local maximum at $\boldsymbol{x}$, and if $\mathscr{H}_f(\boldsymbol{x})$ is indefinite then f has a saddle point at $\boldsymbol{x}$. Together with the characterizations of definite and indefinite (symmetric) matrices given in 15, this proves the following

17**. Theorem** (The second derivative test)**.** *Suppose that $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$, and that all second order partial derivatives of f are continuous at $\boldsymbol{x}$; then*

- *(i) f has a local minimum at $\boldsymbol{x}$ if every leading principal minor of $H_f(\boldsymbol{x})$ is positive.*
- *(ii) f has a local maximum at $\boldsymbol{x}$ if the leading principal minors of $H_f(\boldsymbol{x})$ are alternately negative and positive (i.e., $(-1)^p \mu_p(H_f(\boldsymbol{x})) > 0$ for $1 \leqslant p \leqslant n$).*
- *(iii) f has a saddle point at $\boldsymbol{x}$ if $H_f(\boldsymbol{x})$ has a negative principal minor of even order, or negative and positive principal minors of odd order.*

In any other case, further investigation is required. The second derivative test will now be spelled out for real-valued functions of two and three real variables. Suppose that f_{11}, f_{12}, f_{21} and f_{22} are continuous at $\boldsymbol{x} \in \mathbb{R}^2$, and that $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$. Let

$$\mu_1 = f_{11}(\boldsymbol{x}) \quad \text{and} \quad \mu_2 = \begin{vmatrix} f_{11}(\boldsymbol{x}) & f_{12}(\boldsymbol{x}) \\ f_{21}(\boldsymbol{x}) & f_{22}(\boldsymbol{x}) \end{vmatrix};$$

then

- f has a local minimum at $\boldsymbol{x}$ if $\mu_1 > 0$ and $\mu_2 > 0$,
- f has a local maximum at $\boldsymbol{x}$ if $\mu_1 < 0$ and $\mu_2 > 0$,
- f has a saddle point at $\boldsymbol{x}$ if $\mu_2 < 0$;

and the remaining cases require further investigation. Next, suppose that f_{ij} is continuous at $\boldsymbol{x} \in \mathbb{R}^3$ for $1 \leqslant i, j \leqslant 3$, and that $\nabla f(\boldsymbol{x}) = \boldsymbol{0}$. Let f_{ij} denote $f_{ij}(\boldsymbol{x})$, for $1 \leqslant i, j \leqslant 3$, and consider

$$\Delta_1 = \begin{vmatrix} f_{22} & f_{23} \\ f_{32} & f_{33} \end{vmatrix}, \quad \Delta_2 = \begin{vmatrix} f_{11} & f_{13} \\ f_{31} & f_{33} \end{vmatrix}, \quad \Delta_3 = \begin{vmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{vmatrix},$$

$$\delta_1 = f_{11}, \quad \delta_2 = f_{22}, \quad \delta_3 = f_{33}, \quad \text{and} \quad \Delta = \begin{vmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{vmatrix};$$

then

- f has a local minimum at $\boldsymbol{x}$ if $\delta_1, \Delta_3, \Delta > 0$,
- f has a local maximum at $\boldsymbol{x}$ if $\delta_1, \Delta < 0$ and $\Delta_3 > 0$,
- f has a saddle point at $\boldsymbol{x}$ if at least one of $\Delta_1, \Delta_2, \Delta_3$ is negative, or there are both positive and negative numbers among $\delta_1, \delta_2, \delta_3, \Delta$,

and the remaining cases require further investigation.