1. The second derivative test

In one variable calculus, the mean value theorem relates the first derivative of a function to the nearby values of the function. The analogue for second (and higher order) derivatives is known as 'Taylor's Theorem (with remainder)'. Here I state it only for second order derivatives.

Theorem 1.1 (Second order Taylor's Theorem with remainder)**.** *Let $I \subset \mathbf{R}$ be an open interval and $f : I \to \mathbf{R}$ be twice differentiable. Then for any $a, t \in I$, there exists a number c between a and t such that*

$$f(t) = f(a) + f'(a)(t-a) + \frac{1}{2}f''(c)(t-a)^2$$

This theorem generalizes to scalar-valued functions of more than one variable as follows.

Corollary 1.2. *Suppose that $f : \mathbf{R}^n \to \mathbf{R}$ is twice differentiable on an open ball $B(\mathbf{a}, r) \subset \mathbf{R}^n$. Then for any displacement $\mathbf{h} \in \mathbf{R}^n$ with magnitude $\|\mathbf{h}\| < r$, there exists $c \in (0,1)$ such that*

$$f(\mathbf{a}+\mathbf{h}) = f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot \mathbf{h} + \frac{1}{2}\mathbf{h}^T Hf(\mathbf{a}+c\mathbf{h})\mathbf{h}.$$

Proof. Let $\gamma(t) = \mathbf{a} + t\mathbf{h}$ parametrize the line through $\mathbf{a}$ in direction $\mathbf{h}$ and $g : \mathbf{R} \to \mathbf{R}$ be the composite function

$$g(t) := f \circ \gamma(t) = f(\mathbf{a} + t\mathbf{h}).$$

Then $g(t)$ is defined for all t in an open interval containing $t = 0$ and $t = 1$. Moreover, the Chain rule tells us that g is twice differentiable on this interval and allows us to compute the derivatives of g in terms of derivatives of f:

$$g'(t) = Df(\gamma(t))\gamma'(t) = \nabla f(\mathbf{a}+t\mathbf{h}) \cdot \mathbf{h} = \sum_{j=1}^n \frac{\partial f}{\partial x_j}(\mathbf{a}+t\mathbf{h})h_j,$$

and

$$g''(t) = \sum_{j=1}^n \sum_{k=1}^n \frac{\partial^2 f}{\partial x_k\, \partial x_j}(\mathbf{a}+t\mathbf{h})h_k h_j = \mathbf{h}^T\, Hf(\mathbf{a}+t\mathbf{h})\, \mathbf{h}.$$

Applying Taylor's Theorem (above) to g with $a = 0$ and $x = 1$, I obtain $c \in (0,1)$ such that

$$g(1) = g(0) + g'(0) + \frac{1}{2}g''(c).$$

In light of our computations above, this can be rewritten in terms of f as

$$f(\mathbf{a}+\mathbf{h}) = f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot \mathbf{h} + \frac{1}{2}\mathbf{h}^T Hf(\mathbf{a}+c\mathbf{h})\mathbf{h},$$

which is what I wanted to show. □

By itself Corollary 1.2 is not very useful, because we don't know much about the number c, or more specifically, about the relationship between $Hf(\mathbf{a}+c\mathbf{h})$ and $Hf(\mathbf{a})$. However, if we assume that f is C^2 at $\mathbf{a}$ (i.e. that all entries of Hf are continuous at $\mathbf{a}$), then it follows from Proposition 1.10 in my notes about limits that

$$\lim_{\mathbf{x}\to\mathbf{a}} Hf(\mathbf{x}) = Hf(\mathbf{a}).$$

Here I am thinking of $Hf : \mathbf{R}^n \to \mathbf{R}^{n^2}$ as a vector-valued function with one component for each second order partial derivative $\frac{\partial^2 f}{\partial x_j\, \partial x_k}$ of f. Hence I can restate Corollary 1.2 in the following less precise but ultimately more useful fashion.

Theorem 1.3. *Suppose that $f : \mathbf{R}^n \to \mathbf{R}$ is C^2 at $\mathbf{a}$. Then*

$$f(\mathbf{a}+\mathbf{h}) = f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot \mathbf{h} + \frac{1}{2}\mathbf{h}^T Hf(\mathbf{a})\mathbf{h} + E_2(\mathbf{h}),$$

where $\lim_{\mathbf{h}\to\mathbf{0}} \frac{E_2(\mathbf{h})}{\|\mathbf{h}\|^2} = \mathbf{0}$.

Proof. Given $\mathbf{a} \in \mathbf{R}^n$, choose $r > 0$ such that $f(\mathbf{x})$ is defined for all $\mathbf{x} \in B(\mathbf{a}, r)$. Then for any $\mathbf{h}$ with magnitude $\|\mathbf{h}\| < r$, Corollary 1.2 tells me that

$$E_2(\mathbf{h}) = \frac{1}{2}(\mathbf{h}^T Hf(\mathbf{a})\mathbf{h} - \mathbf{h}^T Hf(\mathbf{a} + c\mathbf{h})\mathbf{h}) = \frac{1}{2}\mathbf{h}^T(Hf(\mathbf{a})\mathbf{h} - Hf(\mathbf{a} + c\mathbf{h}))\mathbf{h}$$

for some $c \in (0, 1)$. By continuity of Hf at $\mathbf{a}$, this expression tends to 0 as $\mathbf{h} \to \mathbf{0}$. I must show that it does so faster than $\|\mathbf{h}\|^2$, and for this I resort to the definition of limit.

Let $\epsilon > 0$ be given. Since f is C^2 at $\mathbf{a}$ there exists $\delta > 0$ such that $\|\mathbf{h}\| < \delta$ implies that

$$\|Hf(\mathbf{a} + \mathbf{h}) - Hf(\mathbf{a})\| < 2\epsilon.$$

Note that if $\|\mathbf{h}\| < \delta$, then for any $c \in (0, 1)$, I have $\|c\mathbf{h}\| < \delta$, too. Hence $0 < \|\mathbf{h}\| < \delta$ implies that

$$\frac{|E_2(\mathbf{h})|}{\|\mathbf{h}\|^2} = \frac{|\frac{1}{2}\mathbf{h}^T(Hf(\mathbf{a} + c\mathbf{h}) - Hf(\mathbf{a}))\mathbf{h}|}{\|\mathbf{h}\|^2} \leq \frac{1}{2}\|Hf(\mathbf{a} + c\mathbf{h}) - Hf(\mathbf{a})\| < \epsilon.$$

The '$\leq$' comes from the Cauchy-Schwarz inequality for matrices. This proves that $\frac{|E_2(\mathbf{h})|}{\|\mathbf{h}\|^2} \to 0$ as $\mathbf{h} \to \mathbf{0}$. □

In order to pass from this theorem to the second derivative test, I must introduce a bit of terminology associated to symmetric matrices. The *quadratic form* associated to a symmetric $n \times n$ matrix A is the function $Q : \mathbf{R}^n \to \mathbf{R}$ given by

$$Q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}.$$

Note that $Q(c\mathbf{x}) = c^2 Q(\mathbf{x})$ for any scalar $c \in \mathbf{R}$.

Definition 1.4. *Let $A \in \mathcal{M}_{n\times n}$ be a symmetric square matrix and $Q : \mathbf{R}^n \to \mathbf{R}$ be the associated quadratic form. We say that*

- *A is* positive definite *if $Q(\mathbf{x}) > 0$ for all non-zero $\mathbf{x} \in \mathbf{R}^n$;*
- *A is* negative definite *if $Q(\mathbf{x}) < 0$ for all non-zero $\mathbf{x} \in \mathbf{R}^n$;*
- *A is* indefinite *if there exist $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$ such that $Q(\mathbf{x}) < 0 < Q(\mathbf{y})$.*

A symmetric matrix can satisfy at most one of these three conditions, but it's hard to tell just by looking which, if any, holds for a given matrix. For 2×2 matrices, there is a fairly convenient condition one can apply.

Theorem 1.5. *A 2×2 symmetric matrix $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ is*

- *positive definite if and only if $a > 0$ and $ac > b^2$;*
- *negative definite if and only if $a < 0$ and $ac > b^2$;*
- *indefinite if and only if $ac < b^2$.*

Proof. See Shifrin. □

For larger matrices, there is another criterion one can use, but it depends on the notion of an 'eigenvalue'. Though I do not do so here, it can be proved using the method of Lagrange multipliers.

Theorem 1.6. *An $n \times n$ symmetric matrix A is*

- *positive definite if and only if all (real) eigenvalues of A are positive;*
- *negative definite if and only if all (real) eigenvalues of A are negative;*
- *indefinite if and only if A has both positive and negative eigenvalues.*

Jones gives another useful criterion for definiteness (see 'The Definiteness Test' in Jones chapter 4E) that I do not record here. Instead, I move on to

Theorem 1.7 (Second derivative test)**.** *Suppose that $f : \mathbf{R}^n \to \mathbf{R}$ is C^2 at some critical point $\mathbf{a} \in \mathbf{R}^n$ for f. Then f has*

- *a local minimum at $\mathbf{a}$ if $Hf(\mathbf{a})$ is positive definite;*
- *a local maximum at $\mathbf{a}$ if $Hf(\mathbf{a})$ is negative definite;*
- *neither a local maximum nor a local minimum if $Hf(\mathbf{a})$ is indefinite.*

Proof. Let $Q(\mathbf{h}) = \mathbf{h}^T Hf(\mathbf{a})\mathbf{h}$ be the quadratic form associated to the symmetric matrix $Hf(\mathbf{a})$. Since Q is continuous, and since the unit sphere $\{\mathbf{h} \in \mathbf{R}^n : \|\mathbf{h}\| = 1\}$ is a compact subset of $\mathbf{R}^n$, the Extreme Value Theorem gives me unit vectors $\mathbf{h}_{max}, \mathbf{h}_{\min}$ such that

$$Q(\mathbf{h}_{max}) \geq Q(\mathbf{h}) \geq Q(\mathbf{h}_{\min}) \tag{1}$$

for all unit vectors $\mathbf{h} \in \mathbf{R}^n$. As I noted above, $Q(c\mathbf{h}) = c^2 Q(\mathbf{h})$, so I can extend this inequality to *any* vector $\mathbf{h} \in \mathbf{R}^n$, regardless of length:

$$\|\mathbf{h}\|^2 Q(\mathbf{h}_{max}) \geq Q(\mathbf{h}) \geq \|\mathbf{h}\|^2 Q(\mathbf{h}_{min}).$$

In particular, Q is positive definite if and only if $Q(\mathbf{h}_{min}) > 0$, negative definite if and only if $Q(\mathbf{h}_{max}) < 0$ and indefinite if and only if $Q(\mathbf{h}_{min}) < 0 < Q(\mathbf{h}_{max})$.

Suppose now that $Hf(\mathbf{a})$ is positive definite. Since $\mathbf{a}$ is a critical point of f, Theorem 1.3 tells me that for any small displacement $\mathbf{h}$,

$$f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) = Q(\mathbf{h}) + E_2(\mathbf{h}) \geq Q(\mathbf{h}_{min}) \|\mathbf{h}\|^2 + E_2(\mathbf{h}) = \|\mathbf{h}\|^2 \left(Q(\mathbf{h}_{min}) + \frac{E_2(\mathbf{h})}{\|\mathbf{h}\|^2} \right).$$

Since $Hf(\mathbf{a})$ is positive definite, I know that $\frac{1}{3}Q(\mathbf{h}_{min}) > 0$. So Theorem 1.3 tells me further that there exists $\delta > 0$ such that $\|\mathbf{h}\| < \delta$ implies that $\frac{E_2(\mathbf{h})}{\|\mathbf{h}\|^2} < \frac{1}{3}Q(\mathbf{h}_{min})$. Hence $\|\mathbf{h}\| < \delta$ implies that

$$f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) \geq \frac{2}{3} \|\mathbf{h}\|^2 Q(\mathbf{h}_{min}) > 0.$$

That is, f has a local minimum at $\mathbf{a}$. The case where $Hf(\mathbf{a})$ is negative definite is proved in the same fashion.

It remains to deal with the case where $Hf(\mathbf{a})$ is indefinite. This time I take $\mathbf{h} = t\mathbf{h}_{max}$, and obtain that

$$f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) = Q(\mathbf{h}) + E_2(\mathbf{h}) = t^2 \left(Q(\mathbf{h}_{max}) + \frac{E_2(t\mathbf{h}_{max})}{t^2} \right).$$

Since $Hf(\mathbf{a})$ is indefinite, $Q(\mathbf{h}_{max})$ is positive, and I obtain $\delta_1 > 0$ such that $\|\mathbf{h}\| < \delta$ implies that $\left|\frac{E_2(\mathbf{h})}{\|\mathbf{h}\|^2}\right| < \frac{1}{3}Q(\mathbf{h}_{max})$. This holds in particular, if $\mathbf{h} = t\mathbf{h}_{max}$ for any $|t| < \delta_1$. So $|t| < \delta_1$ implies that

$$f(\mathbf{a}+t\mathbf{h}_{max}) - f(\mathbf{a}) > \frac{2}{3}t^2 Q(\mathbf{h}_{max}) > 0.$$

A similar argument shows that there exists δ_2 such that $|t| < \delta_2$

$$f(\mathbf{a}+t\mathbf{h}_{max}) - f(\mathbf{a}) < \frac{2}{3}t^2 Q(\mathbf{h}_{min}) < 0.$$

So for any $\delta > 0$ I may set $t = \frac{1}{2}\min\{\delta, \delta_1, \delta_2\}$ and obtain displacements $t\mathbf{h}_{min}, t\mathbf{h}_{max} \in B(0,\delta)$ such that

$$f(\mathbf{a}+t\mathbf{h}_{min}) - f(\mathbf{a}) < 0 < f(\mathbf{a}+t\mathbf{h}_{max}) - f(\mathbf{a}).$$

That is, f does not have a local maximum or a local minimum at $\mathbf{a}$. $\square$