Definitions

**Definition 1** Let \( f : I \to \mathbb{R} \), where \( I \subseteq \mathbb{R} \) is an open interval. \( f \) is **differentiable** at \( x \in I \) if

\[
\lim_{h \to 0} \frac{f(x + h) - f(x)}{h} = a
\]

for some \( a \in \mathbb{R} \).

This is equivalent to \( \exists a \in \mathbb{R} \) such that:

\[
\lim_{h \to 0} \frac{f(x + h) - (f(x) + ah)}{h} = 0
\]

\[
\Leftrightarrow \forall \varepsilon > 0 \ \exists \delta > 0 \text{ s.t. } 0 < |h| < \delta \Rightarrow \frac{|f(x + h) - (f(x) + ah)|}{|h|} < \varepsilon
\]

Recall that the limit considers \( h \) near zero, but not \( h = 0 \).

**Definition 2** If \( X \subseteq \mathbb{R}^n \) is open, \( f : X \to \mathbb{R}^m \) is **differentiable** at \( x \in X \) if

\[
\exists T_x \in L(\mathbb{R}^n, \mathbb{R}^m) \text{ s.t. } \lim_{h \to 0, h \in \mathbb{R}^n} \frac{|f(x + h) - (f(x) + T_x(h))|}{|h|} = 0
\]

(1)

\( f \) is **differentiable** if it is differentiable at all \( x \in X \).

Note that \( T_x \) is uniquely determined by Equation (1). \( h \) is a small, nonzero element of \( \mathbb{R}^n \); \( h \to 0 \) from any direction, from above, below, along a spiral, etc. The definition requires that one linear operator \( T_x \) works no matter how \( h \) approaches zero. In this case, \( f(x) + T_x(h) \) is the best linear approximation to \( f(x + h) \) for small \( h \).

**Notation:**

- \( y = O(|h|^n) \) as \( h \to 0 \) – read “\( y \) is big-Oh of \( |h|^n \)” – means

\[
\exists K, \delta > 0 \text{ s.t. } |h| < \delta \Rightarrow |y| \leq K|h|^n
\]

\(^1\)Recall \( | \cdot | \) denotes the Euclidean distance.
• \( y = o(|h|^n) \) as \( h \to 0 \) – read “\( y \) is little-oh of \( |h|^n \)” – means
  \[
  \lim_{h \to 0} \frac{|y|}{|h|^n} = 0
  \]

Note that the statement \( y = O(|h|^{n+1}) \) as \( h \to 0 \) implies \( y = o(|h|^n) \) as \( h \to 0 \).

Also note that if \( y \) is either \( O(|h|^{n}) \) or \( o(|h|^n) \), then \( y \to 0 \) as \( h \to 0 \); the difference in whether \( y \) is “big-Oh” or “little-oh” tells us something about the rate at which \( y \to 0 \).

Using this notation, note that \( f \) is differentiable at \( x \Leftrightarrow \exists T_x \in L(\mathbb{R}^n, \mathbb{R}^m) \) such that
  \[
  f(x + h) = f(x) + T_x(h) + o(h) \text{ as } h \to 0
  \]

Notation:

• \( df_x \) is the linear transformation \( T_x \)

• \( Df(x) \) is the matrix of \( df_x \) with respect to the standard basis.

  This is called the *Jacobian* or *Jacobian matrix* of \( f \) at \( x \)

• \( E_f(h) = f(x + h) - (f(x) + df_x(h)) \) is the *error term*

Using this notation,
  \[
  f \text{ is differentiable at } x \Leftrightarrow E_f(h) = o(h) \text{ as } h \to 0
  \]

Now compute \( Df(x) = (a_{ij}) \). Let \( \{e_1, \ldots, e_n\} \) be the standard basis of \( \mathbb{R}^n \). Look in direction \( e_j \) (note that \( |\gamma e_j| = |\gamma| \)).

\[
\begin{align*}
o(\gamma) &= f(x + \gamma e_j) - (f(x) + T_x(\gamma e_j)) \\
&= f(x + \gamma e_j) - f(x) - \left( \begin{array}{ccc}
a_{11} & \cdots & a_{1j} \\
\vdots & \ddots & \vdots \\
a_m1 & \cdots & a_{mj}
\end{array} \right) \left( \begin{array}{c}
\gamma \\
0 \\
\vdots \\
0
\end{array} \right) \\
&= f(x + \gamma e_j) - f(x) - \left( \begin{array}{c}
\gamma a_{1j} \\
\vdots \\
\gamma a_{mj}
\end{array} \right)
\end{align*}
\]
For \( i = 1, \ldots, m \), let \( f^i \) denote the \( i \)th component of the function \( f \):

\[
f^i(x + \gamma e_j) - f^i(x) = o(\gamma)
\]

so \( a_{ij} = \frac{\partial f^i}{\partial x_j}(x) \)

**Theorem 3 (Thm. 3.3)** Suppose \( X \subseteq \mathbb{R}^n \) is open and \( f : X \to \mathbb{R}^m \) is differentiable at \( x \in X \). Then \( \frac{\partial f^i}{\partial x_j} \) exists at \( x \) for \( 1 \leq i \leq m, 1 \leq j \leq n \), and

\[
Df(x) = \begin{pmatrix}
\frac{\partial f^1}{\partial x_1}(x) & \cdots & \frac{\partial f^1}{\partial x_n}(x) \\
\vdots & \ddots & \vdots \\
\frac{\partial f^m}{\partial x_1}(x) & \cdots & \frac{\partial f^m}{\partial x_n}(x)
\end{pmatrix}
\]

i.e. the Jacobian at \( x \) is the matrix of partial derivatives at \( x \).

**Remark:** If \( f \) is differentiable at \( x \), then all first-order partial derivatives \( \frac{\partial f^i}{\partial x_j} \) exist at \( x \). However, the converse is false: existence of all the first-order partial derivatives does not imply that \( f \) is differentiable. The missing piece is continuity of the partial derivatives:

**Theorem 4 (Thm. 3.4)** If all the first-order partial derivatives \( \frac{\partial f^i}{\partial x_j} \) (\( 1 \leq i \leq m, 1 \leq j \leq n \)) exist and are continuous at \( x \), then \( f \) is differentiable at \( x \).

**Directional Derivatives:**

Suppose \( X \subseteq \mathbb{R}^n \) open, \( f : X \to \mathbb{R}^m \) is differentiable at \( x \), and \( |u| = 1 \).

\[
f(x + \gamma u) - (f(x) + T_x(\gamma u)) = o(\gamma) \text{ as } \gamma \to 0
\]

\[
\Rightarrow \quad f(x + \gamma u) - (f(x) + \gamma T_x(u)) = o(\gamma) \text{ as } \gamma \to 0
\]

\[
\Rightarrow \quad \lim_{\gamma \to 0} \frac{f(x + \gamma u) - f(x)}{\gamma} = T_x(u) = Df(x)u
\]

i.e. the directional derivative in the direction \( u \) (with \(|u| = 1\)) is

\[
Df(x)u \in \mathbb{R}^m
\]

**Theorem 5 (Thm. 3.5, Chain Rule)** Let \( X \subseteq \mathbb{R}^n, Y \subseteq \mathbb{R}^m \) be open, \( f : X \to Y \), \( g : Y \to \mathbb{R}^p \). Let \( x_0 \in X \) and \( F = g \circ f \). If \( f \) is differentiable at \( x_0 \) and \( g \) is differentiable at \( f(x_0) \), then \( F = g \circ f \) is differentiable at \( x_0 \) and

\[
dF_{x_0} = dg_{f(x_0)} \circ df_{x_0}
\]

(composition of linear transformations)

\[
DF(x_0) = Dg(f(x_0))DF(x_0)
\]

(matrix multiplication)
Remark: The statement is exactly the same as in the univariate case, except we replace the univariate derivative by a linear transformation. The proof is more or less the same, with a bit of linear algebra added.

Theorem 6 (Thm. 1.7, Mean Value Theorem, Univariate Case) Let \( a, b \in \mathbb{R} \). Suppose \( f : [a, b] \to \mathbb{R} \) is continuous on \([a, b]\) and differentiable on \((a, b)\). Then there exists \( c \in (a, b) \) such that

\[
\frac{f(b) - f(a)}{b-a} = f'(c)
\]

that is, such that

\[
f(b) - f(a) = f'(c)(b - a)
\]

Proof: Consider the function

\[
g(x) = f(x) - f(a) - \frac{f(b) - f(a)}{b-a}(x - a)
\]

Then \( g(a) = 0 = g(b) \). See Figure 1. Note that for \( x \in (a, b) \),

\[
g'(x) = f'(x) - \frac{f(b) - f(a)}{b-a}
\]

so it suffices to find \( c \in (a, b) \) such that \( g'(c) = 0 \).

Case I: If \( g(x) = 0 \) for all \( x \in [a, b] \), choose an arbitrary \( c \in (a, b) \), and note that \( g'(c) = 0 \), so we are done.

Case II: Suppose \( g(x) > 0 \) for some \( x \in [a, b] \). Since \( g \) is continuous on \([a, b]\), it attains its maximum at some point \( c \in (a, b) \). Since \( g \) is differentiable at \( c \) and \( c \) is an interior point of the domain of \( g \), we have \( g'(c) = 0 \), and we are done.

Case III: If \( g(x) < 0 \) for some \( x \in [a, b] \), the argument is similar to that in Case II. ■

Remark: The Mean Value Theorem is useful for estimating bounds on functions and error terms in approximation of functions.

Notation:

\[
\ell(x, y) = \{ \alpha x + (1 - \alpha)y : \alpha \in [0, 1] \}
\]

is the line segment from \( x \) to \( y \).

Theorem 7 (Mean Value Theorem) Suppose \( f : \mathbb{R}^n \to \mathbb{R} \) is differentiable on an open set \( X \subseteq \mathbb{R}^n \), \( x, y \in X \) and \( \ell(x, y) \subseteq X \). Then there exists \( z \in \ell(x, y) \) such that

\[
f(y) - f(x) = Df(z)(y - x)
\]
Remark: This statement is different from Theorem 3.7 in de la Fuente. Notice that the statement is exactly the same as in the univariate case. For $f : \mathbb{R}^n \to \mathbb{R}^m$, we can apply the Mean Value Theorem to each component, to obtain $z_1, \ldots, z_m \in \ell(x, y)$ such that

$$f^i(y) - f^i(x) = Df^i(z_i)(y - x)$$

However, we cannot find a single $z$ which works for every component. Note that each $z_i \in \ell(x, y) \subset \mathbb{R}^n$; there are $m$ of them, one for each component in the range.

The following result plays the same role in estimating function values and error terms for functions taking values in $\mathbb{R}^m$ as the Mean Value Theorem plays for functions from $\mathbb{R}$ to $\mathbb{R}$.

**Theorem 8** Suppose $X \subset \mathbb{R}^n$ is open and $f : X \to \mathbb{R}^m$ is differentiable. If $x, y \in X$ and $\ell(x, y) \subseteq X$, then there exists $z \in \ell(x, y)$ such that

$$|f(y) - f(x)| \leq |df_z(y - x)|$$

$$\leq \|df_z\||y - x|$$

Remark: To understand why we don’t get equality, consider $f : [0, 1] \to \mathbb{R}^2$ defined by

$$f(t) = (\cos 2\pi t, \sin 2\pi t)$$

$f$ maps $[0, 1]$ to the unit circle in $\mathbb{R}^2$. Note that $f(0) = f(1) = (1, 0)$, so $|f(1) - f(0)| = 0$. However, for any $z \in [0, 1]$,

$$|df_z(1 - 0)| = |2\pi(- \sin 2\pi z, \cos 2\pi z)|$$

$$= 2\pi \sqrt{\sin^2 2\pi z + \cos^2 2\pi z}$$

$$= 2\pi$$

**Section 4.4. Taylor’s Theorem**

**Theorem 9 (Thm. 1.9, Taylor’s Theorem in $\mathbb{R}^1$)** Let $f : I \to \mathbb{R}$ be $n$-times differentiable, where $I \subseteq \mathbb{R}$ is an open interval. If $x, x + h \in I$, then

$$f(x + h) = f(x) + \sum_{k=1}^{n-1} \frac{f^{(k)}(x)h^k}{k!} + E_n$$

where $f^{(k)}$ is the $k$th derivative of $f$ and

$$E_n = \frac{f^{(n)}(x + \lambda h)h^n}{n!} \text{ for some } \lambda \in (0, 1)$$
Motivation: Let

\[ T_n(h) = f(x) + \sum_{k=1}^{n} \frac{f^{(k)}(x)h^k}{k!} \]

\[ T_n(0) = f(x) \]

\[ T_n'(h) = f'(x) + f''(x)h + \cdots + \frac{f^{(n)}(x)h^{n-1}}{(n-1)!} \]

\[ T_n'(0) = f'(x) \]

\[ T_n''(h) = f''(x) + \cdots + \frac{f^{(n)}(x)h^{n-2}}{(n-2)!} \]

\[ T_n''(0) = f''(x) \]

\[ \vdots \]

\[ T_n^{(n)}(0) = f^{(n)}(x) \]

so \( T_n(h) \) is the unique \( n^{th} \) degree polynomial such that

\[ T_n(0) = f(x) \]

\[ T_n'(0) = f'(x) \]

\[ \vdots \]

\[ T_n^{(n)}(0) = f^{(n)}(x) \]

The proof of the formula for the remainder \( E_n \) is essentially the Mean Value Theorem; the problem in applying it is that the point \( x + \lambda h \) is not known in advance.

Theorem 10 (Alternate Taylor's Theorem in \( \mathbb{R}^1 \)) Let \( f : I \to \mathbb{R} \) be \( n \) times differentiable, where \( I \subseteq \mathbb{R} \) is an open interval and \( x \in I \). Then

\[ f(x + h) = f(x) + \sum_{k=1}^{n} \frac{f^{(k)}(x)h^k}{k!} + o(h^n) \text{ as } h \to 0 \]

If \( f \) is \((n + 1)\) times continuously differentiable (i.e. all derivatives up to order \( n + 1 \) exist and are continuous), then

\[ f(x + h) = f(x) + \sum_{k=1}^{n} \frac{f^{(k)}(x)h^k}{k!} + O(h^{n+1}) \text{ as } h \to 0 \]

Remark: The first equation in the statement of the theorem is essentially a restatement of the definition of the \( n^{th} \) derivative. The second statement is proven from Theorem 1.9, and the continuity of the derivative, hence the boundedness of the derivative on a neighborhood of \( x \).
Definition 11 Let $X \subseteq \mathbb{R}^n$ be open. A function $f : X \to \mathbb{R}^m$ is continuously differentiable on $X$ if

- $f$ is differentiable on $X$ and
- $df_x$ is a continuous function of $x$ from $X$ to $L(\mathbb{R}^n, \mathbb{R}^m)$, with operator norm $\|df_x\|

$f$ is $C^k$ if all partial derivatives of order less than or equal to $k$ exist and are continuous in $X$.

Theorem 12 (Thm. 4.3) Suppose $X \subseteq \mathbb{R}^n$ is open and $f : X \to \mathbb{R}^m$. Then $f$ is continuously differentiable on $X$ if and only if $f$ is $C^1$.

Remark: The notation in Taylor’s Theorem is difficult. If $f : \mathbb{R}^n \to \mathbb{R}^m$, the quadratic terms are not hard for $m = 1$; for $m > 1$, we handle each component separately. For cubic and higher order terms, the notation is a mess.

Linear Terms:

Theorem 13 Suppose $X \subseteq \mathbb{R}^n$ is open and $x \in X$. If $f : X \to \mathbb{R}^m$ is differentiable, then

$$f(x + h) = f(x) + Df(x)h + o(h) \text{ as } h \to 0$$

The previous theorem is essentially a restatement of the definition of differentiability.

Theorem 14 (Corollary of 4.4) Suppose $X \subseteq \mathbb{R}^n$ is open and $x \in X$. If $f : X \to \mathbb{R}^m$ is $C^2$, then

$$f(x + h) = f(x) + Df(x)h + O(|h|^2) \text{ as } h \to 0$$

Quadratic Terms:

We treat each component of the function separately, so consider $f : X \to \mathbb{R}, X \subseteq \mathbb{R}^n$ an open set. Let

$$D^2 f(x) = \begin{pmatrix}
\frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\
\frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \cdots & \cdots & \frac{\partial^2 f}{\partial x_n^2}(x)
\end{pmatrix}$$

$f \in C^2 \Rightarrow \frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial^2 f}{\partial x_j \partial x_i}(x)$

$\Rightarrow D^2 f(x)$ is symmetric

$\Rightarrow D^2 f(x)$ has an orthonormal basis of eigenvectors

and thus can be diagonalized
Theorem 15 (Stronger Version of Thm. 4.4) Let \( X \subseteq \mathbb{R}^n \) be open, \( f : X \rightarrow \mathbb{R} \), \( f \in C^2(X) \), and \( x \in X \). Then

\[
f(x + h) = f(x) + Df(x)h + \frac{1}{2}h^\top (D^2 f(x))h + o \left( |h|^2 \right) \quad \text{as} \quad h \rightarrow 0
\]

If \( f \in C^3 \),

\[
f(x + h) = f(x) + Df(x)h + \frac{1}{2}h^\top (D^2 f(x))h + O \left( |h|^3 \right) \quad \text{as} \quad h \rightarrow 0
\]

Remark: de la Fuente assumes \( X \) is convex. \( X \) is said to be convex if, for every \( x, y \in X \) and \( \alpha \in [0, 1] \), \( \alpha x + (1 - \alpha)y \in X \). Notice we don’t need this. Since \( X \) is open,

\[x \in X \Rightarrow \exists \delta > 0 \text{ s.t. } B_\delta(x) \subseteq X\]

and \( B_\delta(x) \) is convex.

Definition 16 We say \( f \) has a saddle at \( x \) if \( Df(x) = 0 \) but \( f \) has neither a local maximum nor a local minimum at \( x \).

Corollary 17 Suppose \( X \subseteq \mathbb{R}^n \) is open and \( x \in X \). If \( f : X \rightarrow \mathbb{R} \) is \( C^2 \), then there is an orthonormal basis \( \{v_1, \ldots, v_n\} \) and corresponding eigenvalues \( \lambda_1, \ldots, \lambda_n \in \mathbb{R} \) of \( D^2 f(x) \) such that

\[
f(x + h) = f(x + \gamma_1 v_1 + \cdots + \gamma_n v_n)
\]

\[
= f(x) + \sum_{i=1}^{n} (Df(x) v_i) \gamma_i + \frac{1}{2} \sum_{i=1}^{n} \lambda_i \gamma_i^2 + o \left( |\gamma|^2 \right)
\]

where \( \gamma_i = h \cdot v_i \).

1. If \( f \in C^3 \), we may strengthen \( o(|\gamma|^2) \) to \( O(|\gamma|^3) \).

2. If \( f \) has a local maximum or local minimum at \( x \), then

\[Df(x) = 0\]

3. If \( Df(x) = 0 \), then

\[
\lambda_1, \ldots, \lambda_n > 0 \Rightarrow f \text{ has a local minimum at } x
\]

\[
\lambda_1, \ldots, \lambda_n < 0 \Rightarrow f \text{ has a local maximum at } x
\]

\[
\lambda_i < 0 \text{ for some } i, \lambda_j > 0 \text{ for some } j \Rightarrow f \text{ has a saddle at } x
\]

\[
\lambda_1, \ldots, \lambda_n \geq 0, \lambda_i > 0 \text{ for some } i \Rightarrow f \text{ has a local minimum or a saddle at } x
\]

\[
\lambda_1, \ldots, \lambda_n \leq 0, \lambda_i < 0 \text{ for some } i \Rightarrow f \text{ has a local maximum or a saddle at } x
\]

\[
\lambda_1 = \cdots = \lambda_n = 0 \quad \text{gives no information.}
\]
**Proof:** (Sketch) From our study of quadratic forms, we know the behavior of the quadratic terms is determined by the signs of the eigenvalues. If $\lambda_i = 0$ for some $i$, then we know that the quadratic form arising from the second partial derivatives is identically zero in the direction $v_i$, and the higher derivatives will determine the behavior of the function $f$ in the direction $v_i$. For example, if $f(x) = x^3$, then $f'(0) = 0$, $f''(0) = 0$, but we know that $f$ has a saddle at $x = 0$; however, if $f(x) = x^4$, then again $f'(0) = 0$ and $f''(0) = 0$ but $f$ has a local (and global) minimum at $x = 0$. ■
Figure 1: The Mean Value Theorem.