Sequences and Limits

- A sequence of real numbers can be viewed as a set of numbers \( \{x_1, x_2, \ldots, x_k, \ldots\} \), which is often also denoted as \( \{x_k\} \) or \( \{x_k\}_{k=1}^{\infty} \).

- A sequence \( \{x_k\} \) is **increasing** if \( x_1 < x_2 < \cdots < x_k < \cdots \). If \( x_k \leq x_{k+1} \), then we say that the sequence is **nondecreasing**. Similarly, we can define **decreasing** and **nonincreasing** sequences. Nonincreasing and nondecreasing sequences are called **monotone sequences**.

- A number \( x^* \in R \) is called the **limit** of the sequence \( \{x_k\} \) if for any positive \( \epsilon \) there is a number \( K \) (which may depend on \( \epsilon \)) such that for all \( k > K \), \( |x_k - x^*| < \epsilon \). In this case, we write
  \[ x^* = \lim_{k \to \infty} x_k \quad \text{or} \quad x_k \to x^* \]

- A sequence that has a limit is called a **convergent sequence**.
Sequences and Limits

- A sequence in $R^n$ is a function whose domain is the set of natural numbers 1, 2, ..., $k$, ... and whose range is contained in $R^n$. We use the notation $\{x^{(1)}, x^{(2)}, \ldots\}$ or $\{x^{(k)}\}$ for sequences in $R^n$.

- For limits of sequences in $R^n$, we need to replace absolute values with vector norms. In other words, $x^*$ is the limit of $\{x^{(k)}\}$ if for any positive $\epsilon$ there is a number $K$ such that $k > K$, $\|x^{(k)} - x^*\| < \epsilon$.

- If a sequence $\{x^{(k)}\}$ is convergent, we write $x^* = \lim_{k \to \infty} x^{(k)}$ or $x^{(k)} \to x^*$. 

Sequences and Limits

- Theorem 5.1: A convergent sequence has only one limit.
- A sequence \( \{x^{(k)}\} \) in \( \mathbb{R}^n \) is **bounded** if there exists a number \( B \geq 0 \) such that \( \|x^{(k)}\| \leq B \) for all \( k = 1, 2, \ldots \).
- Theorem 5.2: Every convergent sequence is bounded.
- For a sequence \( \{x_k\} \) in \( \mathbb{R} \), a number \( B \) is called an **upper bound** if \( x_k \leq B \) for all \( k = 1, 2, \ldots \). In this case, we say \( \{x_k\} \) is **bounded above**.
- A number \( B \) is called an **lower bound** if \( x_k \geq B \) for all \( k = 1, 2, \ldots \). In this case, we say \( \{x_k\} \) is **bounded below**.
Sequences and Limits

- Any sequence \( \{x_k\} \) in \( \mathbb{R} \) that has an upper bound has a **least upper bound** (also called the **supremum**), which is the smallest number \( B \) that is an upper bound of \( \{x_k\} \). Similarly, it has a **greatest lower bound** (also called **infimum**).

- If \( B \) is the least upper bound of the sequence \( \{x_k\} \), then \( x_k \leq B \) for all \( k \), and for any \( \epsilon > 0 \), there exists a number \( K \) such that \( x_K > B - \epsilon \)

- If \( B \) is the greatest lower bound of \( \{x_k\} \), then \( x_k \geq B \) for all \( k \), and for any \( \epsilon > 0 \), there exists a number \( K \) such that \( x_K < B + \epsilon \)

- Theorem 5.3: Every monotone bounded sequence in \( \mathbb{R} \) is convergent.
Sequences and Limits

- Given a sequence \( \{x^{(k)}\} \) and an increasing sequence of natural numbers \( \{m_k\} \), the sequence
  \[
  \{x^{(m_k)}\} = \{x^{(m_1)}, x^{(m_2)}, \ldots\}
  \]
  is called a **subsequence** of the sequence \( \{x^{(k)}\} \).

- Theorem 5.4: Consider a convergent sequence \( \{x^{(k)}\} \) with limit \( x^* \). Then any subsequence of \( \{x^{(k)}\} \) also converges to \( x^* \).

- It turns out that any bounded sequence contains a convergent subsequence (**Bolzano-Weierstrass Theorem**).
Sequences and Limits

- Consider a function $f : \mathbb{R}^n \to \mathbb{R}^m$ and a point $x_0 \in \mathbb{R}^n$. Suppose that there exists $f^*$ such that for any convergent sequence $\{x^{(k)}\}$ with limit $x_0$, we have
  \[ \lim_{k \to \infty} f(x^{(k)}) = f^* \]
  Then, we use the notation $\lim_{x \to x_0} f(x)$ to represent $f^*$

- It turns out that $f$ is continuous at $x_0$ if and only if for any convergent sequence $\{x^{(k)}\}$ with limit $x_0$, we have
  \[ \lim_{k \to \infty} f(x^{(k)}) = f \left( \lim_{k \to \infty} x^{(k)} \right) = f(x_0) \]

- Therefore, using the notation introduced above, the function $f$ is continuous at $x_0$ if and only if
  \[ \lim_{x \to x_0} f(x) = f(x_0) \]
Sequences and Limits

- We say that a sequence \( \{A_k\} \) of \( m \times n \) matrices converges to the \( m \times n \) matrix \( A \) if \( \lim_{k \to \infty} \|A - A_k\| = 0 \).

- Lemma 5.1: Let \( A \in \mathbb{R}^{n \times n} \). Then, \( \lim_{k \to \infty} A^k = O \) if and only if the eigenvalues of \( A \) satisfy \( |\lambda_i(A)| < 1, i = 1, \ldots, n \).

- Lemma 5.2: The series of \( n \times n \) matrices
  \[
  I_n + A + A^2 + \cdots + A^k + \cdots
  \]
  converges if and only if \( \lim_{k \to \infty} A^k = O \). In this case the sum of the series equals \( (I_n - A)^{-1} \).
A matrix-valued function $A : \mathbb{R}^r \rightarrow \mathbb{R}^{n \times n}$ is continuous at a point $\xi_0 \in \mathbb{R}^r$ if
\[ \lim_{\|\xi - \xi_0\|} \|A(\xi) - A(\xi_0)\| = 0 \]

Lemma 5.3: Let $A : \mathbb{R}^r \rightarrow \mathbb{R}^{n \times n}$ be an $n \times n$ matrix-valued function that is continuous at $\xi_0$. If $A(\xi_0)^{-1}$ exists, then $A(\xi)^{-1}$ exists for $\xi$ sufficiently close to $\xi_0$ and $A(\cdot)^{-1}$ is continuous at $\xi_0$. 
Differential calculus is based on the idea of approximating an arbitrary function by an **affine function**.

A function $\mathcal{A} : R^n \rightarrow R^m$ is **affine** if there exists a **linear function** $\mathcal{L} : R^n \rightarrow R^m$ and a vector $y \in R^m$ such that

$$\mathcal{A}(x) = \mathcal{L}(x) + y$$

for every $x \in R^n$.

Consider a function $f : R^n \rightarrow R^m$ and a point $x_0 \in R^n$. We wish to find an affine function $\mathcal{A}$ that approximates $f$ near the point $x_0$.

First, it is natural to impose the condition

$$\mathcal{A}(x_0) = f(x_0)$$
\[ A(x_0) = f(x_0) \]

**Differentiability**

- Because \( A(x) = L(x) + y \), we obtain \( y = f(x_0) - L(x_0) \)
- By the linearity of \( L \),
  \[ L(x) + y = L(x) - L(x_0) + f(x_0) = L(x - x_0) + f(x_0) \]
- Hence, we may write
  \[ A(x) = L(x - x_0) + f(x_0) \]
- Next, we require that \( A(x) \) approaches \( f(x) \) faster than \( x \) approaches \( x_0 \); that is,
  \[ \lim_{x \to x_0, x \in \Omega} \frac{\| f(x) - A(x) \|}{\| x - x_0 \|} = 0 \]
- The conditions ensure that \( A \) approximates \( f \) near \( x_0 \) in the sense that the approximation error is “small” compared with the distance of the point from \( x_0 \).
In summary, a function $f : \Omega \to \mathbb{R}^m, \Omega \subset \mathbb{R}^n$, is said to be \textit{differentiable} at $x_0 \in \Omega$ if there is an affine function that approximates $f$ near $x_0$; that is, there exists a linear function $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$
\lim_{x \to x_0, x \in \Omega} \frac{\|f(x) - (\mathcal{L}(x - x_0) + f(x_0))\|}{\|x - x_0\|} = 0
$$

The linear function $\mathcal{L}$ is determined uniquely by $f$ and $x_0$ and is called the \textit{derivative} of $f$ at $x_0$.

The function is said to be \textit{differentiable} on $\Omega$ if $f$ is differentiable at every point of its domain $\Omega$. 

$$
A(x) = \mathcal{L}(x - x_0) + f(x_0)
$$
Differentiability

- In $\mathbb{R}$, an affine function has the form $ax + b$, with $a, b \in \mathbb{R}$. Hence, a real-valued function $f(x)$ of a real variable $x$ that is differentiable at $x_0$ can be approximated by a function $A(x) = ax + b$.

- Because $f(x_0) = A(x_0) = ax_0 + b$, we obtain $A(x) = ax + b = a(x - x_0) + f(x_0)$.

- The linear part of $A(x)$, denoted by $L(x)$ earlier, is just $ax$. The norm of a real number is its absolute value, so by the definition of differentiability

$$
\lim_{x \to x_0} \frac{|f(x) - (a(x - x_0) + f(x_0))|}{|x - x_0|} = 0
$$

$\implies$ $\lim_{x \to x_0} \frac{|f(x) - f(x_0)|}{|x - x_0|} = a$
The number \( a \) is commonly denoted \( f'(x_0) \) and is called the derivative of \( f \) at \( x_0 \).

The affine function \( \mathcal{A} \) is therefore given by
\[
\mathcal{A}(x) = f(x_0) + f'(x_0)(x - x_0)
\]

The affine function is tangent to \( f \) at \( x_0 \).
The Derivative Matrix

- Any linear transformation from $R^n$ to $R^m$, and in particular the derivative $L$ of $f : R^n \rightarrow R^m$, can be represented by an $m \times n$ matrix.

- To find the matrix $L$ of the derivative $L$, we use the natural basis $\{e_1, e_2, ..., e_n\}$ for $R^n$. Consider the vectors $x_j = x_0 + te_j, j = 1, 2, .., n$

By the definition of the derivative, we have

$$\lim_{t \to 0} \frac{f(x_j) - (tLe_j + f(x_0))}{t} = 0 \quad j = 1, 2, .., n$$

This means that

$$\lim_{t \to 0} \frac{f(x_j) - f(x_0)}{t} = Le_j$$

$$L(x_j - x_0) = L(x_0 + te_j - x_0)$$
$$= L(te_j) = L(te_j) = tLe_j$$
The Derivative Matrix

\[ \lim_{t \to 0} \frac{f(x_j) - f(x_0)}{t} = L e_j \]

- But \( L e_j \) is the \( j \)th column of the matrix \( L \). The vector \( x_j \) differs from \( x_0 \) only in the \( j \)th coordinate, and in that coordinate the difference is just the number \( t \). Therefore, the left side is the partial derivative \( \frac{\partial f}{\partial x_j}(x_0) \)

- Because vector limits are computed by taking the limit of each coordinate function, it follows that if

\[ f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix} \quad \left( f_i : \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, m \right) \]

then \( \frac{\partial f}{\partial x_j}(x_0) = \begin{bmatrix} \frac{\partial f_1}{\partial x_j}(x_0) \\ \vdots \\ \frac{\partial f_m}{\partial x_j}(x_0) \end{bmatrix} \) and the matrix \( L \) has the form

\[
\begin{bmatrix}
\frac{\partial f}{\partial x_1}(x_0) & \cdots & \frac{\partial f}{\partial x_n}(x_0)
\end{bmatrix}
= \begin{bmatrix}
\frac{\partial f_1}{\partial x_1}(x_0) & \cdots & \frac{\partial f_1}{\partial x_n}(x_0) \\
\vdots & & \vdots \\
\frac{\partial f_m}{\partial x_1}(x_0) & \cdots & \frac{\partial f_m}{\partial x_n}(x_0)
\end{bmatrix}
\]
The Derivative Matrix

- The matrix $L$ is called the **Jacobian matrix**, or **derivative matrix**, of $f$ at $x_0$, and is denoted $Df(x_0)$.
- For convenience, we often refer to $Df(x_0)$ simply as the derivative of $f$ at $x_0$.

$$
\begin{bmatrix}
\frac{\partial f_1(x_0)}{\partial x_1} & \cdots & \frac{\partial f_1(x_0)}{\partial x_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial f_m(x_0)}{\partial x_1} & \cdots & \frac{\partial f_m(x_0)}{\partial x_n}
\end{bmatrix}
$$
The Derivative Matrix

Theorem 5.5: If a function \( f : \mathbb{R}^n \to \mathbb{R}^m \) is differentiable at \( x_0 \), then the derivative of \( f \) at \( x_0 \) is determined uniquely and is represented by an \( m \times n \) derivative matrix \( Df(x_0) \). The best affine approximation of \( f \) near \( x_0 \) is then given by

\[
A(x) = f(x_0) + Df(x_0)(x - x_0),
\]

in the sense that

\[
f(x) = A(x) + r(x)
\]

and \( \lim_{x \to x_0} \frac{\|r(x)\|}{\|x - x_0\|} = 0 \).

The columns of the derivative matrix \( Df(x_0) \) are vector partial derivatives. The vector \( \frac{\partial f}{\partial x_j}(x_0) \) is a tangent vector at \( x_0 \) to the curve \( f \) obtained by varying only the \( j \)th coordinate of \( x \).
The Derivative Matrix

- If \( f : \mathbb{R}^n \to \mathbb{R} \) is differentiable, then the function \( \nabla f \) defined by
  \[
  \nabla f(x) = \begin{bmatrix}
  \frac{\partial f}{\partial x_1}(x) \\
  \vdots \\
  \frac{\partial f}{\partial x_n}(x)
  \end{bmatrix}
  = Df(x)^T
  \]
  is called the **gradient** of \( f \).

- Given \( f : \mathbb{R}^n \to \mathbb{R} \), if \( \nabla f \) is differentiable, we say that \( f \) is **twice differentiable**, and we write the derivative of \( \nabla f \) as
  \[
  D^2 f = \begin{bmatrix}
  \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\
  \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\
  \vdots & \vdots & \ddots & \vdots \\
  \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2}
  \end{bmatrix}
  \]
  \( \frac{\partial^2 f}{\partial x_i \partial x_j} \) represents taking the partial derivative with respect to \( x_j \) first, then with respect to \( x_i \).
The Derivative Matrix

- The matrix $D^2 f(x)$ is called the **Hessian matrix** of $f$ at $x$, and is often also denoted $F(x)$.

- A function $f : \Omega \to R^m, \Omega \subset R^n$, is said to be **continuously differentiable** on $\Omega$ if it is differentiable (on $\Omega$), and $Df : \Omega \to R^{m \times n}$ is continuous; that is, the components of $f$ have continuous partial derivatives. In this case, we write $f \in C^1$. If the components of $f$ have continuous partial derivatives of order $p$, we write $f \in C^p$.

- The Hessian matrix of a function $f : R^n \to R$ at $x$ is symmetric if $f$ is twice continuously differentiable at $x$. This is a well-known result from calculus called **Clairaut's theorem** or **Schwarz's theorem**.
Consider the function \( f(\mathbf{x}) = \begin{cases} \frac{x_1 x_2 (x_1^2 - x_2^2)}{(x_1^2 + x_2^2)} & \text{if } \mathbf{x} \neq \mathbf{0} \\ 0 & \text{if } \mathbf{x} = \mathbf{0} \end{cases} \)

Compute its Hessian at the point \( \mathbf{0} = [0, 0]^T \)

Start with 
\[
\frac{\partial^2 f}{\partial x_1^2} = \frac{\partial}{\partial x_1} \left( \frac{\partial f}{\partial x_1} \right) 
\]

\[
\frac{\partial f}{\partial x_1}(\mathbf{x}) = \begin{cases} \frac{x_2(x_1^4 - x_2^4 + 4x_1^2x_2^2)}{(x_1^2 + x_2^2)^2} & \text{if } \mathbf{x} \neq \mathbf{0} \\ 0 & \text{if } \mathbf{x} = \mathbf{0} \end{cases}
\]

Note that
\[
\frac{\partial f}{\partial x_1}([x_1, 0]^T) = 0 \quad \frac{\partial^2 f}{\partial x_1^2}(\mathbf{0}) = 0
\]
\[
\frac{\partial f}{\partial x_1}([0, x_2]^T) = -x_2 \quad \frac{\partial f}{\partial x_2x_1}(\mathbf{0}) = -1
\]
Example

- We next compute
  \[
  \frac{\partial^2 f}{\partial x_2^2} = \frac{\partial}{\partial x_2} \left( \frac{\partial f}{\partial x_2} \right)
  \]
  \[
  \frac{\partial f}{\partial x_2}(\mathbf{x}) = \begin{cases} 
  x_1(x_1^4 - x_2^4 - 4x_1^2x_2^2)/(x_1^2 + x_2^2)^2 & \text{if } \mathbf{x} \neq \mathbf{0} \\
  0 & \text{if } \mathbf{x} = \mathbf{0}
  \end{cases}
  \]

- \[
  \frac{\partial f}{\partial x_2}([0, x_2]^T) = 0 \\
  \frac{\partial f}{\partial x_2^2}(\mathbf{0}) = 0
  \]

- \[
  \frac{\partial f}{\partial x_2}([x_1, 0]^T) = x_1 \\
  \frac{\partial f}{\partial x_1 x_2}([x_1, 0]^T) = 1
  \]

- Therefore, the Hessian evaluated at the point \( \mathbf{0} = [0, 0]^T \) is
  \[
  \mathbf{F}(\mathbf{0}) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}
  \]
The **chain rule** for differentiating the composition \( g(f(t)) \), of a function \( f : R \to R^n \) and a function \( g : R^n \to R \).

**Theorem 5.6:** Let \( g : \mathcal{D} \to R \) be differentiable on an open set \( \mathcal{D} \subset R^n \), and let \( f : (a, b) \to \mathcal{D} \) be differentiable on \((a, b)\). Then, the composite function \( h : (a, b) \to R \) given by \( h(t) = g(f(t)) \) is differentiable on \((a, b)\), and

\[
h'(t) = Dg(f(t)) D f(t) = \bigtriangledown g(f(t))^T \begin{bmatrix} f'_1(t) \\ \vdots \\ f'_n(t) \end{bmatrix}
\]
Differentiation Rules

\[
\frac{d}{dx}(fg) = f'g + fg'
\]

- **Product rule**: Let \( f : \mathbb{R}^n \to \mathbb{R}^m \) and \( g : \mathbb{R}^n \to \mathbb{R}^m \) be two differentiable functions. Define the function \( h : \mathbb{R}^n \to \mathbb{R} \) by \( h(x) = f(x)^T g(x) \). Then \( h \) is also differentiable and

\[
Dh(x) = f(x)^T Dg(x) + g(x)^T Df(x)
\]

- Some useful formulas: Let \( A \in \mathbb{R}^{m \times n} \) and \( y \in \mathbb{R}^m \), the derivative with respect to \( x \)

\[
D(y^T A x) = y^T A
\]

\[
D(x^T A x) = x^T (A + A^T), \quad \text{if} \ m = n
\]

- If \( y \in \mathbb{R}^n \), then \( D(y^T x) = y^T \)

- If \( Q \) is a symmetric matrix, then \( D(x^T Q x) = 2x^T Q \). In particular,

\[
D(x^T x) = 2x^T
\]
Level Sets and Gradients

- The **level set** of a function \( f : \mathbb{R}^n \to \mathbb{R} \) at level \( c \) is the set of points \( S = \{ x : f(x) = c \} \). For \( f : \mathbb{R}^2 \to \mathbb{R} \), we are usually interested in \( S \) when it is a curve. For \( f : \mathbb{R}^3 \to \mathbb{R} \), the sets \( S \) most often considered are surfaces.

- Example: Consider \( f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 \), \( x = [x_1, x_2]^T \). It is called **Rosenbrock’s function**.
Level Sets and Gradients

- A point $x_0$ is on the level set $S$ at level $c$ means that $f(x_0) = c$. Suppose that there is a curve $\gamma$ lying in $S$ and parameterized by a continuously differentiable function $g : R \rightarrow R^n$. Suppose also that $g(t_0) = x_0$ and $Dg(t_0) = v \neq 0$, so that $v$ is a tangent vector to $\gamma$ at $x_0$.

- Applying the chain rule to the function $h(t) = f(g(t))$

$\begin{align*}
    h'(t_0) &= Df(g(t_0))Dg(t_0) = Df(x_0)v \\
    \text{Since } \gamma \text{ lies on } S, \text{ we have } h(t) &= f(g(t)) = c \\
    \text{That is, } h \text{ is constant. Thus, } h'(t_0) &= 0 = Df(x_0)v = \nabla f(x_0)^Tv
\end{align*}$
Level Sets and Gradients

- Theorem 5.7: The vector $\nabla f(x_0)$ is orthogonal to the tangent vector to an arbitrary smooth curve passing through $x_0$ on the level set determined by $f(x) = f(x_0)$.

- It is natural to say that $\nabla f(x_0)$ is \textit{orthogonal} or \textit{normal} to the level set $S$ corresponding to $x_0$, and to take as the tangent plane (or line) to $S$ at $x_0$ the set of all points $x$ satisfying $\nabla f(x_0)^T(x - x_0) = 0$, if $\nabla f(x_0) \neq 0$. 
Level Sets and Gradients

- $\nabla f(x_0)$ is the direction of \textit{maximum rate of increase} of $f$ at $x_0$.
- The direction of maximum rate of increase of a real-valued differentiable function at a point is orthogonal to the level set of the function through that point.
- An example about $f : R^2 \rightarrow R$
  - The curve on the surface running from bottom to top has the property that its projection onto the $(x_1, x_2)$-plane is always orthogonal to the level curves and is called a \textit{path of steepest ascent}.
Level Sets and Gradients

- The **graph** of \( f : \mathbb{R}^n \to \mathbb{R} \) is the set \( \{ [x^T, f(x)]^T : x \in \mathbb{R}^n \} \subset \mathbb{R}^{n+1} \). The notion of the gradient of a function has an alternative useful interpretation in terms of the tangent hyperplane to its graph.

- Let \( x_0 \in \mathbb{R}^n \) and \( z_0 = f(x_0) \). The point \( [x_0^T, z_0]^T \in \mathbb{R}^{n+1} \) is a point on the graph of \( f \). If \( f \) is differentiable at \( \xi \), then the graph admits a nonvertical tangent hyperplane at \( \xi = [x^T, z_0]^T \). The hyperplane through \( \xi \) is the set of all points \( [x_1, \ldots, x_n, z]^T \in \mathbb{R}^{n+1} \) satisfying the equation

\[
 u_1(x_1 - x_{01}) + \cdots + u_n(x_n - x_{0n}) + v(z - z_0) = 0
\]

where the vector \( [u_1, \ldots, u_n, v] \in \mathbb{R}^{n+1} \) is normal to the hyperplane.

\[
\langle (u_1, \ldots, u_n, v), (x_1 - x_{01}, \ldots, x_n - x_{0n}, z - z_0) \rangle
\]

is normal to \( (u_1, \ldots, u_n, v) \), which is a vector on the hyperplane. 

\[
x_0 = (x_{01}, \ldots, x_{0n})
\]
Level Sets and Gradients

\[ u_1(x_1 - x_{01}) + \cdots + u_n(x_n - x_{0n}) + v(z - z_0) = 0 \]

- Assuming that this hyperplane is nonvertical (that is, \( v \neq 0 \)), let
  
  \[ d_i = -\frac{u_i}{v} \]

  Thus, we can rewrite the hyperplane equation as

  \[ z = d_1(x_1 - x_{01}) + \cdots + d_n(x_n - x_{0n}) + z_0 \]

- We can think of the right side as a function \( z : \mathbb{R}^n \to \mathbb{R} \). Observe that for the hyperplane to be tangent to the graph of \( f \), the functions \( f \) and \( z \) must have the same partial derivatives at the point \( x_0 \). Hence, if \( f \) is differential at \( x_0 \), its tangent hyperplane can be written in terms of its gradient, as given by the equation

  \[ z - z_0 = Df(x_0)(x - x_0) = (x - x_0)^T \nabla f(x_0) \]

  \[ Df(x_0) = \frac{z - z_0}{x - x_0} \]
Theorem 5.8 **Taylor’s Theorem**: Assume that a function \( f : \mathbb{R} \to \mathbb{R} \) is \( m \) times continuously differentiable (i.e. \( f \in C^m \)) on an interval \([a, b]\). Denote \( h = b - a \). Then,

\[
f(b) = f(a) + \frac{h}{1!} f^{(1)}(a) + \frac{h^2}{2!} f^{(2)}(a) + \cdots + \frac{h^{m-1}}{(m-1)!} f^{(m-1)}(a) + R_m
\]

(called Taylor’s formula) where \( f^{(i)} \) is the \( i \)th derivative of \( f \), and

\[
R_m = \frac{h^m(1-\theta)^{m-1}}{(m-1)!} f^{(m)}(a + \theta h) = \frac{h^m}{m!} f^{(m)}(a + \theta' h) \quad \theta, \theta' \in (0, 1)
\]
Taylor Series

- An important property of Taylor’s theorem arises from the forms of the remainder $R_m$.
- We introduce the order symbols, $O$ and $o$.
- Let $g$ be a real-valued function defined in some neighborhood of $0 \in \mathbb{R}^n$, with $g(x) \neq 0$ if $x \neq 0$. Let $f : \Omega \to \mathbb{R}^m$ be defined in a domain $\Omega \subset \mathbb{R}^n$ that includes $0$. Then, we write
  1. $f(x) = O(g(x))$ to mean that the quotient $\|f(x)\|/|g(x)|$ is bounded near $0$; that is, these exist numbers $K > 0$ and $\delta > 0$ such that if $\|x\| < \delta$, $x \in \Omega$, then $\|f(x)\|/|g(x)| \leq K$ or $\|f(x)\| \leq Kg(x)$
  2. $f(x) = o(g(x))$ to mean that

$$
\lim_{x \to 0, x \in \Omega} \frac{\|f(x)\|}{|g(x)|} = 0
$$
Taylor Series

- The symbol $O(g(x))$ [read “big-oh” of $g(x)$] is used to represent a function that is bounded by a scaled version of $g$ in a neighborhood of 0.

- Examples:
  - $x = O(x)$
  - $\left[ \frac{x^3}{2x^2 + 3x^4} \right] = O(x^2)$
  - $\cos x = O(1)$
  - $\sin x = O(x)$
Taylor Series

- On the other hand, \( o(g(x)) \) [read “little-oh” of \( g(x) \) ] represents a function that goes to zero “faster” than \( g(x) \) in the sense that \( \lim_{x \to 0, x \in \Omega} \| o(g(x)) \| / |g(x)| = 0 \)

- Examples:
  - \( x^2 = o(x) \)
  - \[
    \begin{bmatrix}
    x^3 \\
    2x^2 + 3x^4
    \end{bmatrix}
    = o(x)
  \]
  - \( x^3 = o(x^2) \)
  - \( x = o(1) \)
Taylor Series

- Note that if \( f(x) = o(g(x)) \), then \( f(x) = O(g(x)) \) (but the converse is not necessarily true). Also, if \( f(x) = O(\|x\|^p) \), then \( f(x) = o(\|x\|^{p-\epsilon}) \) for any \( \epsilon > 0 \).

- Suppose that \( f \in C^m \). Recall that the remainder term in Taylor’s theorem has the form

\[
R_m = \frac{h^m}{m!} f^{(m)}(a + \theta h) \quad \theta \in (0, 1)
\]

Substituting this into Taylor’s formula, we get

\[
f(b) = f(a) + \frac{h}{1!} f^{(1)}(a) + \frac{h^2}{2!} f^{(2)}(a) + \cdots + \frac{h^{m-1}}{(m-1)!} f^{(m-1)}(a) + \frac{h^m}{m!} f^{(m)}(a + \theta h)
\]

By the continuity of \( f^{(m)} \), we have \( f^{(m)}(a + \theta h) \rightarrow f^{(m)}(a) \) as \( h \rightarrow 0 \) that is, \( f^{(m)}(a + \theta h) = f^{(m)}(a) + o(1) \). Therefore,

\[
R_m = \frac{h^m}{m!} f^{(m)}(a + \theta h) = \frac{h^m}{m!} f^{(m)}(a) + o(h^m) \quad h^m o(1) = o(h^m)
\]
Taylor Series

- We may then write Taylor’s formula as

$$f(b) = f(a) + \frac{h}{1!} f^{(1)}(a) + \frac{h^2}{2!} f^{(2)}(a) + \cdots + \frac{h^{m-1}}{(m-1)!} f^{(m-1)}(a) + \frac{h^m}{m!} f^{(m)}(a) + o(h^m)$$

- If, in addition, we assume that $f \in C^{m+1}$, we may replace the term $o(h^m)$ above by $O(h^{m+1})$

$$f(b) = f(a) + \frac{h}{1!} f^{(1)}(a) + \frac{h^2}{2!} f^{(2)}(a) + \cdots + \frac{h^{m-1}}{(m-1)!} f^{(m-1)}(a) + \frac{h^m}{m!} f^{(m)}(a) + O(h^{m+1})$$
Theorem 5.9 *Mean value theorem*: If a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable on an open set $\Omega \subset \mathbb{R}^n$, then for any pair of points $x, y \in \Omega$, there exists a matrix $M$ such that

$$f(x) - f(y) = M(x - y)$$

The mean value theorem follows from Taylor’s theorem (for the case where $m = 1$) applied to each component of $f$. $M$ is a matrix whose rows are the rows of $Df$ evaluated at points that lie one the line segment joining $x$ and $y$.

若函數$f(x)$在$[a,b]$區間上連續並可維分，則在該區間內必存在一點$c$，使

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