C477: Computing for Optimal Decisions
Background in Linear Algebra & Calculus

Ruth Misener & Panos Parpas

October 7, 2018

Note: These notes summarise the background material in linear algebra and calculus necessary for C447. If this material is unfamiliar, please revise independently. It is in your best interest to revise before the first week of classes ends; C477 will assume that you understand the following notes. To assist revision, we offer the following:

- Post questions to Piazza: `piazza.com/imperial.ac.uk/fall2018/477/home`;
- Quiz 0 posted on CATE will help you assess your level of preparation for the module. It does not affect your final mark. The answers to the quiz will be posted on the morning of 16 October.
- Revise using notes from C145 (Mathematical Methods) and C233 (Computational Techniques); we have posted these notes on Piazza: `https://piazza.com/imperial.ac.uk/fall2018/477/resources`.

We will also assume familiarity with the material in C496 (Mathematics for Inference and Machine Learning).

Contents

1 Linear Algebra

Alternate revision sources: C145 notes *Matrices* and *Vectors* by Dr Mahdi Cheraghchi & Dr Marc Deisenroth (provided on Piazza: `https://piazza.com/imperial.ac.uk/fall2018/477/resources`); Chapters 2-3, *An Introduction to Optimization*, Chong & Żak; C233 *General Lecture Notes* by Prof István Maros gives a complete development.

1.1 Vectors & Matrices

Vector Definitions (column vector, row vector, transpose)

- We define a **column vector** in $\mathbb{R}^n$ as an array of n numbers,

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

- A **row vector** in $\mathbb{R}^n$ is

$$\boldsymbol{x} = (x_1,\, x_2,\, \ldots,\, x_n)$$

- The **transpose** of a column vector is a row vector. If x is a column vector in $\mathbb{R}^n$ then,

$$\boldsymbol{x}^\top = (x_1,\, x_2,\, \ldots, x_n)$$

- In C477 all vectors are column vectors; we use lower-case bold font to represent vectors $\boldsymbol{x}$ rather than arrow notation $\vec{x}$. Note that this is opposite from C496. In C496, all vectors are row vectors. The courses are following two different books; this is the reason for the discrepancy.

Matrix Definitions (matrix, symmetric matrix)

- A **matrix** is a rectangular array of entries with n rows and m columns.
- A matrix B with n rows and m columns belongs to $\mathbb{R}^{n\times m}$
- A **symmetric matrix** is one that is equal to its transpose, i.e., $B = B^\top$.

Vector Operations (equality test, summation, scalar multiplication)

- Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are **equal**, $\boldsymbol{x} = \boldsymbol{y}$ if $x_i = y_i \; \forall \, i = 1, \ldots, n$
- The **sum of two vectors** $\boldsymbol{x} + \boldsymbol{y}$ is the vector

$$(x_1 + y_1, \, x_2 + y_2, \, \ldots, \, x_n + y_n)^\top$$

- **Multiplication** of a vector $\boldsymbol{x}$ by a scalar α is defined as,

$$\alpha \; \boldsymbol{x} = (\alpha \; x_1, \, \alpha \; x_2, \, \ldots, \, \alpha \; x_n)^\top$$

Matrix Multiplication

- **Multiplication:** If $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times k}$, then $C = A \; B$ is a matrix in $\mathbb{R}^{n \times k}$. The $(i, j)^{th}$ entry of C is,

$$C_{ij} = \sum_{l=1}^{m} A_{il} \cdot B_{lj}$$

- For two matrices A and B, the **transpose of their product** is $(AB)^\top = B^\top A^\top$.

1.2 Linear Systems of Equations

Definitions Suppose that we are given m equations in n unknowns of the form,

$$\begin{array}{rcccl} a_{11} \; x_1 & + a_{12} \; x_2 & + \ldots & + a_{1n} \; x_n & = b_1 \\ a_{21} \; x_1 & + a_{22} \; x_2 & + \ldots & + a_{2n} \; x_n & = b_2 \\ \vdots & & \ddots & & \vdots \\ a_{m1} \; x_1 & + a_{m2} \; x_2 & + \ldots & + a_{mn} \; x_n & = b_m. \end{array}$$

We can also represent the set of equations above in vector notation,

$$\boldsymbol{a_1} \; x_1 + \boldsymbol{a_2} \; x_2 + \cdots + \boldsymbol{a_n} \; x_n = \boldsymbol{b}$$

where

$$\boldsymbol{a_j} = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix} \quad \text{and} \quad \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

We associate the matrix,

$$A = [\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_n}]$$

with the system of equations, and represent the system as follows,

$$A\ \boldsymbol{x} = \boldsymbol{b}$$

where

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Linear Independence & Matrix Rank

- A vector $\boldsymbol{y}$ is said to be a **linear combination of vectors** $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_m}$ if there are scalars $\alpha_1, \alpha_2, \ldots, \alpha_m$ such that,

$$\boldsymbol{y} = \alpha_1\ \boldsymbol{x_1} + \alpha_2\ \boldsymbol{x_2} + \ldots + \alpha_m\ \boldsymbol{x_m} = \sum_{i=1}^{m} \alpha_i\ \boldsymbol{x_i}$$

- A set of vectors $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_m}$ is said to be **linearly independent** if the equality,

$$\alpha_1\ \boldsymbol{x_1} + \alpha_2\ \boldsymbol{x_2} + \ldots + \alpha_m\ \boldsymbol{x_m} = \sum_{i=1}^{m} \alpha_i\ \boldsymbol{x_i} = 0,$$

  implies that all coefficients α_i, $i = 1, \ldots, m$ are equal to zero.

- **Exercise:** A set of vectors contains the zero vector, is the set linearly independent?

Definition 1.1 (Rank of a Matrix)**.** *The maximum number of linearly independent columns of A is called the rank of the matrix, denoted by rank(A)*

Theorem 1.1 (Solution of linear systems of equations)**.** *The system of equations*

$$A\ \boldsymbol{x} = \boldsymbol{b},$$

where $A \in \mathbb{R}^{m \times n}$, $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{b} \in \mathbb{R}^m$ has a solution **if and only if***,*

$$rank(A) = rank([A, \boldsymbol{b}]) = n.$$

If $m = n$ and the condition above holds then the system has a unique solution

1.3 Eigenvalues & Eigenvectors

Definitions

- An **eigenvector** of square matrix $A \in \mathbb{R}^{n \times n}$ is a vector $\boldsymbol{v} \in \mathbb{R}^n$ such that the product $A\ \boldsymbol{v}$ is equal to a scalar multiple ($\lambda \in \mathbb{R}$) of $\boldsymbol{v}$:

$$A\ \boldsymbol{v} = \lambda\ \boldsymbol{v}$$

- The scalars λ are called **eigenvalues**; a matrix is positive definite if all eigenvalues are positive.

Positive definite & positive semidefinite matrices

- A matrix $A \in \mathbb{R}^{n \times n}$ is called **positive semidefinite** if for all $\boldsymbol{d} \in \mathbb{R}^n$,

$$\boldsymbol{d}^\top\ A\ \boldsymbol{d} \geq 0.$$

  We use the notation $A \succeq 0$. All eigenvalues are of $\left(A + A^\top\right)/2$ are non-negative.

- If the above inequality is satisfied strictly, i.e. if

$$\boldsymbol{d}^\top\ A\ \boldsymbol{d} > 0\ \forall \boldsymbol{d} \in \mathbb{R}^n \backslash 0,$$

  then A is called **positive definite**. We use the notation $A \succ 0$. All eigenvalues are of $\left(A + A^\top\right)/2$ are positive.

1.4 Inner product

Definition: $\boldsymbol{x}^\top\ \boldsymbol{y}$ is called the **inner product** between two vectors and it is sometimes written as $\langle \boldsymbol{x},\ \boldsymbol{y} \rangle$, where

$$\langle \boldsymbol{x},\ \boldsymbol{y} \rangle = \boldsymbol{x}^\top\ \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$$

Orthogonality

Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are **orthogonal** if $\langle \boldsymbol{x},\ \boldsymbol{y} \rangle = 0$. By this definition, the zero vector is orthogonal to every other vector. An example of two nonzero orthogonal vectors:

$$\langle \boldsymbol{x},\ \boldsymbol{y} \rangle = [1,\ 2] \begin{bmatrix} -3 \\ 3/2 \end{bmatrix}$$

1.5 Norms

Definition: A vector norm is a function $\|\cdot\| : \mathbb{R}^n \mapsto \mathbb{R}$ satisfying the following properties:

1. Positivity: $\|\boldsymbol{x}\| \geq 0$ and $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = 0$;
2. Homogeneity: $\|r\ \boldsymbol{x}\| = |r|\ \|\boldsymbol{x}\|$ for all $r \in \mathbb{R}$;
3. Triangle Inequality: $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$

Examples: 1-norm, 2-norm, ∞-norm

1-norm $\|\boldsymbol{x}\|_1 = |x_1| + |x_2| + \ldots + |x_n|$

2-norm $\|\boldsymbol{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$

∞-norm $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$

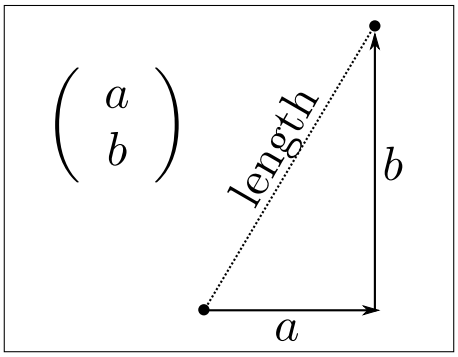
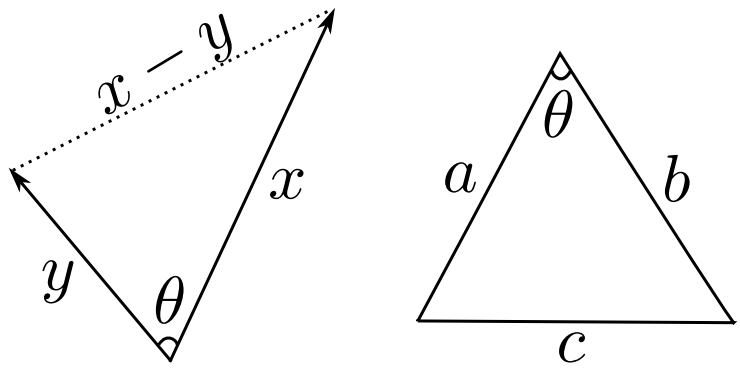


Figure 1: The length of vector $\boldsymbol{x} = [a,\, b]^\top$ is the Euclidean norm $\|\boldsymbol{x}\|_2 = \sqrt{a^2 + b^2}$. The 1-norm of vector $\boldsymbol{x}$ is $a + b$ and the ∞-norm is b.

- The **2-norm** is also called the **Euclidean norm**. When no index is specified on a norm, e.g., $\|\cdot\|$, this is considered to be the Euclidean norm.
- $\|\boldsymbol{x}\|_1 \geq \|\boldsymbol{x}\|_2 \geq \|\boldsymbol{x}\|_\infty$



(a) Angle between vectors (b) Cosine Law

Figure 2: Illustration of angles between vectors and the cosine law in Section 1.6

1.6 Angles Between Vectors

The **angle between two vectors** $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^n$ is the angle $\theta \in [0, \pi]$,

$$\cos(\theta) = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\|\boldsymbol{x}\|_2 \ \|\boldsymbol{y}\|_2}.$$

Cosine Law: For any triangle with sides of length a, b and c the angle opposite c satisfies (see Fig. 2),

$$c^2 = a^2 + b^2 - 2\ a\ b\ \cos(\theta)$$

1.7 Cauchy-Schwarz Inequality

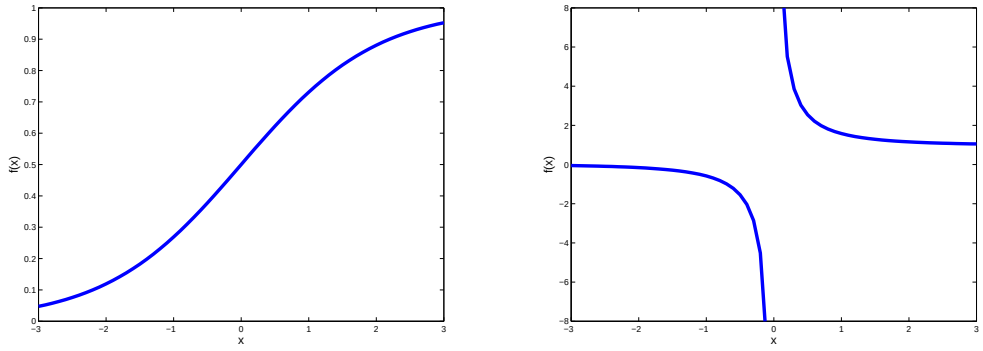
The **Cauchy-Schwarz Inequality** states: $|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\|_2 \ \|\boldsymbol{y}\|_2$

2 Analysis

2.1 Continuous Functions

Definitions

- We write $f : X \to Y$ to mean that a function takes points from the set X (domain) to the set Y (range).



(a) $f(x) = 1/(1 + \exp(-x))$ (b) $f(x) = 1/(1 - \exp(-x))$

Figure 3: Example of a continuous and a discontinuous function at $x = 0$

- Let $X \subset \mathbb{R}^n$, then the function $f : X \to \mathbb{R}^m$ is a **continuous function** at $\boldsymbol{x_0}$ if and only if for every $\epsilon > 0$ there exists a $\delta > 0$ such that when $\|\boldsymbol{x} - \boldsymbol{x_0}\|_2 < \delta$ then $\|f(\boldsymbol{x}) - f(\boldsymbol{x_0})\|_2 < \epsilon$.

Example: The function,

$$f(\boldsymbol{x}) = f(x_1, x_2) = \begin{bmatrix} x_1^2 + x_2 \\ \exp(x_1) + x_2 \\ x_2 \end{bmatrix},$$

evaluated at the point $(2, -1)$ is,

$$f(\boldsymbol{x}) = f(2, -1) = \begin{bmatrix} 3 \\ \exp(2) - 1 \\ -1 \end{bmatrix}$$

3 Multivariable Calculus

Alternate revision sources: C145 notes *Calculus* by Dr Mahdi Cheraghchi & Dr Marc Deisenroth (provided on Piazza: `https://piazza.com/imperial.ac.uk/fall2018/477/resources`); Chapter 5, *An Introduction to Optimization*, Chong & Żak; C233 *General Lecture Notes* by Prof István Maros gives a complete development.

Definitions

- The derivative of a function in one dimension is defined below.

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

  When the above holds, in the sense that the limit exists, then we say that the **function is differentiable at the point** x.

- When a function is differentiable and its derivative is also continuous we say that the function is **continuously differentiable**.

- If a function is defined over $\mathbb{R}^n$, then its **partial derivative** with respect to dimension i is defined as,

$$\frac{\partial f(\boldsymbol{x})}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_i + h, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{h}$$

3.1 Gradients & Jacobian

Definitions

- The vector of the partial derivatives is the **gradient:** $\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix}$;

- If $f : \mathbb{R}^n \to \mathbb{R}^m$ we evaluate the gradient of each function, and call the matrix of derivatives $\nabla f(\boldsymbol{x})$ the gradient of f:

$$= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix};$$

- The transpose of the gradient is called the **Jacobian** matrix.

Example

Find the gradient matrix of,

$$f(x_1, x_2) = \begin{bmatrix} x_1 x_2 \\ \sin(x_1 + x_2) \\ x_1^2 - x_2^2 \end{bmatrix}$$

Then,

$$\nabla f_1(\boldsymbol{x}) = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix}, \ \nabla f_2(\boldsymbol{x}) = \begin{bmatrix} \cos(x_1 + x_2) \\ \cos(x_1 + x_2) \end{bmatrix}, \ \nabla f_3(\boldsymbol{x}) = \begin{bmatrix} 2x_1 \\ -2x_2 \end{bmatrix}$$

We therefore have the gradient matrix.

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \nabla f_1(\boldsymbol{x}) & \nabla f_2(\boldsymbol{x}) & \nabla f_3(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} x_2 & \cos(x_1 + x_2) & 2x_1 \\ x_1 & \cos(x_1 + x_2) & -2x_2 \end{bmatrix}$$

3.2 Hessian

Definitions

- Given a function $f : \mathbb{R}^n \to \mathbb{R}$,
$$\frac{\partial^2 f}{\partial x_i \partial x_j},$$
to denote the i^{th} partial derivative of $\frac{\partial f}{\partial x_j}$.
- We define the matrix $\nabla^2 f(\boldsymbol{x})$ to denote the matrix whose $(i, j)^{th}$ entry is given by $\frac{\partial^2 f}{\partial x_i \partial x_j}$ as the **Hessian** matrix of f, i.e.

$$Hf(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_n} \end{bmatrix}$$

Example: Find the Hessian matrix of the function $f : \mathbb{R}^3 \to \mathbb{R}$, given by $f(x_1, x_2, x_3) = x_1^2 + x_1 x_3 + x_1 x_2 + x_3^2 + \exp(x_1 x_3)$

$$\nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} 2 + x_3^2 \exp(x_1 x_3) & 1 & 1 + (1 + x_1 x_3) \exp(x_1 x_3) \\ 1 & 0 & 0 \\ 1 + (1 + x_1 x_3) \exp(x_1 x_3) & 0 & 2 + x_1^2 \exp(x_1 x_3) \end{bmatrix}$$

3.3 The Chain Rule

The *chain rule* is used for differentiating the composition $g(f(t))$ of a function $g : \mathbb{R}^n \to \mathbb{R}$, and $f : \mathbb{R} \to \mathbb{R}^n$. In particular, suppose that $g : \mathbb{R}^n \to \mathbb{R}$ is differentiable,

and suppose that $f : \mathbb{R} \to \mathbb{R}^n$ is differentiable in $\mathbb{R}$. Then the derivative of the composite function $h(t) = g(f(t))$ is given by,

$$\frac{dh(t)}{dt} = h'(t) = \nabla g(f(t))^\top \begin{pmatrix} \frac{df_1(t)}{dt} \\ \frac{df_2(t)}{dt} \\ \vdots \\ \frac{df_n(t)}{dt} \end{pmatrix}$$

Let $x(t) = [e^t + t^3, t^2, t+1]^\top$, $t \in \mathbb{R}$, and $f(\boldsymbol{x}) = x_1^3 x_2 x_3^2 + x_1 x_2 + x_3$, $x = [x_1, x_2, x_3] \in \mathbb{R}^3$. Find $\frac{df(x(t))}{dt}$ in terms of t.

$$\begin{aligned}
\frac{df(x(t))}{dt} &= \nabla f(x(t))^\top \frac{dx(t)}{dt} \\
&= [3x_1(t)^2 x_2(t) x_3(t)^2 + x_2(t), x_1(t)^3 x_3(t)^2 + x_1(t), 2x_1(t)^3 x_2(t) x_3(t) + 1] \begin{bmatrix} e^t + 3t^2 \\ 2t \\ 1 \end{bmatrix} \\
&= 12t(e^t + 3t^2)^3 + 2te^t + 6t^2 + 2t + 1
\end{aligned}$$

3.4 Mean Value Theorem

1-D: For function defined over the closed interval $[a, b]$ there must exist a point $a < c < b$ such that:

$$f(b) = f(a) + \frac{df(x)}{dx}\Big|_{x=c} (b - a)$$

n-D If a function is differentiable in U and suppose that the interval $[a, b]$ is contained in U, then there exists a point $c \in [a, b]$ such that,

$$f(b) = f(a) + \nabla f(\boldsymbol{x})^T\Big|_{\boldsymbol{x}=c} (b - a).$$

Consider the following function, illustrated in Figure 4, $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = (x - 2)\ (x - 3) + (x - 1)^3.$$

On the interval $[-2, 3]$. Show that there exist two points c_1 and c_2 such that,

$$f(3) = f(-2) + 5\frac{df}{dx}\Big|_{x=c_i}.$$

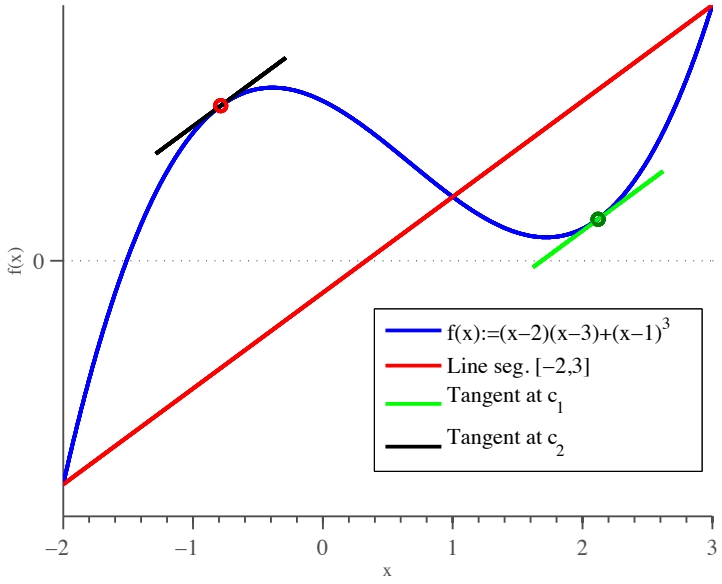


Figure 4: Illustration of the example from Section 3.4

Since $\frac{df}{dx} = 2x - 5 + 3(x-1)^2$ we obtain the following quadratic equation for c,

$$3c^2 - 4c - 5 = 0$$

It has two roots $c_1 = (4 - \sqrt{76})/6$ and $c_2 = (4 + \sqrt{76})/6$. It is easy to verify that these two points satisfy,

$$f(b) = f(a) + \nabla f(\boldsymbol{x})\Big|_{x=c}(b-a),$$

with $a = -2$ and $b = 3$.

3.5 Taylor Series

In many dimensions a function $f : \mathbb{R}^n \to \mathbb{R}$ has the first order expansion about $\boldsymbol{x_0}$

$$f(\boldsymbol{x}) = f(\boldsymbol{x_0}) + \nabla f(\boldsymbol{x_0})^T(\boldsymbol{x} - \boldsymbol{x_0}) + r(\boldsymbol{x} - \boldsymbol{x_0})$$

Where the residual $r(\boldsymbol{x} - \boldsymbol{x_0})$ satisfies $\lim_{\boldsymbol{x}-\boldsymbol{x_0}\to 0} r(\boldsymbol{x} - \boldsymbol{x_0}) = 0$.

The second order expansion is given by,

$$f(\boldsymbol{x}) = f(\boldsymbol{x_0}) + \nabla f(\boldsymbol{x_0})^T(\boldsymbol{x} - \boldsymbol{x_0}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x_0})^T\nabla^2 f(\boldsymbol{x_0})(\boldsymbol{x} - \boldsymbol{x_0}) + r(\boldsymbol{x} - \boldsymbol{x_0})$$