

Lecture Notes
to be used in conjunction with

381 Computational Finance

Part I

Istvan Maros

Department of Computing
Imperial College

V1.4
February 2003

Contents

1	Introduction	1
2	Basic Notions and Definitions	1
2.1	Vectors and matrices	1
2.2	Functions	3
2.3	Functions of several variables	5
2.3.1	Partial differentiation, the gradient, the Hessian	6
2.3.2	Taylor expansion	7
2.3.3	Newton's method for solving $f(x) = 0$	8
2.3.4	Newton's method for $\min f(\mathbf{x})$	8
3	Partial Differential Equations	9
3.1	Solution of PDEs	10
3.2	Finite difference method	11
4	Random variables, probability	13
4.1	Discrete random variables	13
4.2	Continuous random variables	14
4.3	Skewness and kurtosis	16
4.4	Covariance, correlation	17
5	Optimization	18
5.1	Quadratic programming	18
5.2	Lagrangian function for QP	19
5.3	Optimality conditions	19

1 Introduction

Computational finance is a rapidly developing discipline. Its aim is to model and analyze and thus better understand events and processes in the financial world. This course gives an introduction to the most important notions and models of computational finance as well as to some tools for analyzing and solving the arising problems.

As most of the financial models lead to computational problems the knowledge of some computational techniques is essential. To facilitate the discussion, we briefly summarize the required notions and techniques.

Computational finance requires some level of familiarity with vectors, matrices, functions and the idea of probability. Elements of mathematical calculus together with basics of partial differential equations are also needed.

The first part of the lecture notes introduces a unified notation and the terminology used to describe the most basic models that arise in the practice of finance.

2 Basic Notions and Definitions

2.1 Vectors and matrices

A vector is a linearly ordered set of real numbers and is denoted by a boldface lower case letter, like \mathbf{v} . Unless stated otherwise, vectors are represented in columnar form:

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix}$$

The number of components determines the *dimension* of the vector. We say that \mathbf{v} is an m dimensional vector, or an element of the m dimensional space of real numbers: $\mathbf{v} \in \mathbb{R}^m$.

A matrix is a rectangular array of numbers. It is characterized by the number of rows and columns. A matrix is denoted by a boldface capital letter, its elements by the corresponding lower case letter. The rectangular structure is enclosed in square brackets. If \mathbf{A} has m rows and n columns it looks as follows:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

We say that matrix \mathbf{A} is an element of the $m \times n$ dimensional space of real numbers: $\mathbf{A} \in \mathbb{R}^{m \times n}$. The row dimension of \mathbf{A} is m , the column dimension is n .

A matrix \mathbf{A} can be conceived as a set of column vectors. Column j of \mathbf{A} is denoted by \mathbf{a}_j and \mathbf{A} can be written as

$$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n].$$

We can say that \mathbf{A} consists of n m -dimensional vectors.

Similarly, \mathbf{A} can be considered as a set of row vectors. Row i of \mathbf{A} is denoted by \mathbf{a}^i , that is, the index is in the superscript to indicate that this is a row vector. In this way

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}^1 \\ \mathbf{a}^2 \\ \vdots \\ \mathbf{a}^m \end{bmatrix}$$

and we say that \mathbf{A} consists of m n -dimensional row vectors.

Transpose of matrix \mathbf{A} is obtained by interchanging its rows by its columns. It is denoted by \mathbf{A}^T and is defined

$$a_{ij}^T = a_{ji}, \quad 1 \leq i \leq n, \quad 1 \leq j \leq m,$$

which means that the first row of \mathbf{A}^T is equal to the first column of \mathbf{A} , and so on. The transpose of the transpose is the original matrix: $(\mathbf{A}^T)^T = \mathbf{A}$.

The transpose of a column vector is a row vector with the same components.

Linear combination of vectors. Given a set of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ of the same dimension and a set of scalars $\lambda_1, \lambda_2, \dots, \lambda_k$, the *linear combination*, *LC*, of the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ is defined as:

$$\lambda_1 \mathbf{v}_1 + \lambda_2 \mathbf{v}_2 + \dots + \lambda_k \mathbf{v}_k = \sum_{j=1}^k \lambda_j \mathbf{v}_j. \quad (1)$$

Convex linear combination of vectors. If $\lambda_i \geq 0$ for $i = 1, \dots, k$ and $\sum_{i=1}^k \lambda_i = 1$ hold for the scalars in (1) then we say it is a *convex linear combination* of the vectors.

Dot product of two vectors \mathbf{u} and \mathbf{v} of the same dimension is denoted by $\mathbf{u}^T \mathbf{v}$ (sometimes by $\langle \mathbf{u}, \mathbf{v} \rangle$) and is defined in the following way:

$$\mathbf{u}^T \mathbf{v} = u_1 v_1 + u_2 v_2 + \dots + u_m v_m = \sum_{i=1}^m u_i v_i,$$

Operations involving vectors and matrices. The *sum of two matrices* \mathbf{A} and \mathbf{B} is defined if both have the same dimension (say $m \times n$). In this case:

$$\mathbf{C} = \mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{bmatrix}.$$

The $\mathbf{y} = \mathbf{A}\mathbf{x}$ product can be considered as the linear combination of the columns of \mathbf{A} with the set of scalars x_1, x_2, \dots, x_n :

$$\mathbf{y} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = \sum_{j=1}^n x_j\mathbf{a}_j.$$

The product of two matrices \mathbf{A} and \mathbf{B} is defined for the case when the column dimension of \mathbf{A} is equal to the row dimension of \mathbf{B} (the two matrices are *conformable* for multiplication). If $\mathbf{A} \in \mathbb{R}^{m \times p}$ and $\mathbf{B} \in \mathbb{R}^{p \times n}$ the resulting $\mathbf{C} = \mathbf{AB}$ will be an $m \times n$ matrix with c_{ij} defined by:

$$c_{ij} = \sum_{k=1}^p a_{ik}b_{kj}. \quad (2)$$

It is easy to notice that (2) is nothing but the dot product of row i of \mathbf{A} and column j of \mathbf{B} : $c_{ij} = \mathbf{a}^i \mathbf{b}_j$ which is well defined since both are p dimensional vectors. In the case of \mathbf{AB} we say that \mathbf{A} *premultiplies* \mathbf{B} or \mathbf{B} *postmultiplies* \mathbf{A} .

Example 1 *Matrix-matrix products:*

$$\begin{aligned} \text{If } \mathbf{A} &= \begin{bmatrix} 1 & 2 & 3 \\ 1 & -1 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} -1 & 2 \\ 2 & -1 \\ -3 & 0 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix} \\ \text{then } \mathbf{AB} &= \begin{bmatrix} -6 & 0 \\ -6 & 3 \end{bmatrix}, \quad \mathbf{BC} = \begin{bmatrix} 0 & 3 \\ 3 & -3 \\ -6 & 3 \end{bmatrix}, \quad \mathbf{CA} = \begin{bmatrix} 1 & 5 & 5 \\ 2 & 1 & 4 \end{bmatrix} \end{aligned}$$

Inverse matrix. Assume \mathbf{A} is square. If there exists an other square matrix, say \mathbf{X} , of the same dimension such that $\mathbf{AX} = \mathbf{XA} = \mathbf{I}$ then \mathbf{X} is called the *inverse* of \mathbf{A} and is denoted by \mathbf{A}^{-1} . If a square matrix has an inverse it is unique and such a matrix is called *nonsingular*. If it does not have an inverse it is called *singular*.

Two useful identities. The transpose and inverse of the product of two matrices \mathbf{A} and \mathbf{B} (if defined) are:

$$(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T \quad \text{and} \quad (\mathbf{AB})^{-1} = \mathbf{B}^{-1} \mathbf{A}^{-1}.$$

2.2 Functions

A function assigns a value that depends on its independent variable. A function and the variable are usually denoted by a letter, like f and x , respectively. The corresponding function value is denoted by $f(x)$. If the assigned value is a real number we say f is a *real-valued function*. The functional relationship is often denoted by $y = f(x)$.

The function may be defined over the entire \mathbb{R}^1 or a subset of it.

A real-valued function f is said to be *continuous* at x if $x_k \rightarrow x$ implies $f(x_k) \rightarrow f(x)$. The sum and product of continuous functions are also continuous.

Differentiation. A function $y = f(x)$ is said to be differentiable at x if the

$$\frac{f(x_k) - f(x)}{x_k - x}$$

expression has a well defined limit if $x_k \rightarrow x$. This value is called the derivative of f at x

$$f'(x) = \frac{df(x)}{dx} = \lim_{x_k \rightarrow x} \frac{f(x_k) - f(x)}{x_k - x} = \lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta x}. \quad (3)$$

$f(x)$ is differentiable in an interval if it is differentiable at every point of the interval. Differentiable functions are continuous but not all continuous functions are differentiable (counterexample: $y = |x|$ is continuous for all x but not differentiable at $x = 0$).

Some common derivatives are:

$$\begin{aligned} \text{If } f(x) = x^n & \quad \text{then } f'(x) = nx^{n-1}. \\ \text{If } f(x) = e^{cx} & \quad \text{then } f'(x) = ce^{cx}. \\ \text{If } f(x) = \ln(x) & \quad \text{then } f'(x) = 1/x. \end{aligned}$$

As the derivative is also a function of x it can be differentiated if it possesses the feature under (3). This leads to the notion of higher order derivatives.

The sum of differentiable functions is also differentiable.

Chain rule. Consider a composite function $y = f(g(x))$. If both f and g are differentiable (thus continuous) the derivative of f can be written as

$$\frac{dy}{dx} = \lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \rightarrow 0} \left(\frac{\Delta y}{\Delta g} \frac{\Delta g}{\Delta x} \right) = \left(\lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta g} \right) \left(\lim_{\Delta x \rightarrow 0} \frac{\Delta g}{\Delta x} \right) = \frac{dy}{dg} \frac{dg}{dx} = f'(g(x))g'(x).$$

This is the chain rule.

Example 2 Let $f(x) = (x^2 - 1)^3$.

Now, $g(x) = x^2 - 1$ and $f(x) = (g(x))^3$.

Therefore, $f'(x) = 3(g(x))^2 g'(x) = 3(x^2 - 1)^2 2x$.

Example 3 If $f(x) = e^{g(x)}$ and $g(x)$ is differentiable then $f'(x) = g'(x)e^{g(x)}$.

Derivative of a product. If g and h are differentiable functions then $f(x) = g(x)h(x)$ is also differentiable and $f'(x) = g'(x)h(x) + g(x)h'(x)$.

Concavity, convexity. A function $y = f(x)$ is said to be *concave* in an interval R if for any $x_1, x_2 \in R$ and any λ with $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \geq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

Convexity is defined similarly but with the opposite relationship:

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

The interpretation of concavity/convexity is that in any point $x(\lambda) = \lambda x_1 + (1 - \lambda)x_2$ between x_1 and x_2 the function lies above/under the line determined by points $P_1 = (x_1, f(x_1))$ and $P_2 = (x_2, f(x_2))$.

If in an interval, the second derivative of a differentiable function is always negative then the function is concave in this interval. Example: $R = \{x : -\infty \leq x \leq 0\}$, $f(x) = x^3$, $f''(x) = 6x$. On the other hand, if $f''(x)$ is positive in the interval then f is convex. Example: $R = \mathbb{R}^1$, $f(x) = x^2$, $f''(x) = 2$.

Maximum, minimum. Let $R \subseteq \mathbb{R}^1$. $f(x)$ reaches its *maximum* in R at $x^* \in R$ if relation $f(x^*) \geq f(x)$ holds for any other point $x \in R$. In case $f(x^*) > f(x)$ the maximum is unique. Example: $f(x) = -x^4 + 2x^2 - 1$ has a maximum at $x^* = 1$ with $f(1) = 0$. However, $x = 1$ is not a unique maximizer as $f(-1)$ is also equal to 0.

In a similar vein, $f(x)$ reaches its *minimum* at $x^* \in R$ if relation $f(x^*) \leq f(x)$ holds for any other point $x \in R$. In case $f(x^*) < f(x)$ the minimum is unique. Example: $f(x) = x^2 + 2x + 2$ has a unique minimum at $x^* = -1$ with $f(-1) = 1$.

If f is differentiable its derivatives can characterize the extreme (max/min) points of f . Let f be a function that attains its extreme inside the interval where it is defined. Then, a necessary condition for x^* to be an extreme point is that $f'(x^*) = 0$. If, additionally, the second derivative is negative at this point, i.e., $f''(x^*) < 0$, then x^* is a maximizer, otherwise, if $f''(x^*) > 0$ then x^* is a minimizer. The previous examples can be checked for this property.

2.3 Functions of several variables

There are functional relationships where several independent variables determine the value of the function. The n independent variables form a vector $\mathbf{x}^T = [x_1, \dots, x_n]$. The function can be written as $f(\mathbf{x}) = f(x_1, \dots, x_n)$. f may be defined only on a subset of \mathbb{R}^n . We assume that the value of the function is a real number and say that f is a *real-valued function*.

A real-valued function f defined on a subset of \mathbb{R}^n is said to be *continuous* at \mathbf{x} if $\mathbf{x}^k \rightarrow \mathbf{x}$ implies $f(\mathbf{x}^k) \rightarrow f(\mathbf{x})$. Equivalently, f is continuous at \mathbf{x} if for any given $\epsilon > 0$ there is a $\delta > 0$ such that $\|\mathbf{y} - \mathbf{x}\| < \delta$ implies $|f(\mathbf{y}) - f(\mathbf{x})| < \epsilon$, i.e., for points \mathbf{y} close enough to \mathbf{x} the function values $f(\mathbf{y})$ and $f(\mathbf{x})$ are also close (within a predetermined arbitrary accuracy ϵ).

2.3.1 Partial differentiation, the gradient, the Hessian

Let $f(\mathbf{x}) = f(x_1, \dots, x_n)$. If $n - 1$ variables are kept constant and just one is allowed to vary then we have a function of one variable. If this function is differentiable we can determine its derivative in the usual way. Let x_i be the selected variable. The partial differentiation of f with respect to x_i is denoted by $\frac{\partial f(\mathbf{x})}{\partial x_i}$. The C^1 function class is defined to be the set of functions that have continuous partial derivatives with respect to all variables.

Example 4 Let $f(x, y, z) = x^2 - 2xyz + y^2 + z^2/2$. The partial derivatives are:

$$\begin{aligned}\frac{\partial f(x, y, z)}{\partial x} &= 2x - 2yz \quad (\text{everything kept constant except } x), \\ \frac{\partial f(x, y, z)}{\partial y} &= -2xz + 2y, \\ \frac{\partial f(x, y, z)}{\partial z} &= -2xy + z.\end{aligned}$$

If $f \in C^1$ is a real-valued function on \mathbb{R}^n , $f(\mathbf{x}) = f(x_1, \dots, x_n)$, the *gradient* of f is defined to be the row vector

$$\nabla f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \dots, \frac{\partial f(\mathbf{x})}{\partial x_n} \right].$$

Though the entries of the gradient are functions, for any given \mathbf{x} they determine a row vector of numerical values.

Example 5 Let $f(\mathbf{x}) = (x_1 - 2x_2)^3 - 3x_3^3$. Then

$$\nabla f(\mathbf{x}) = [3(x_1 - 2x_2)^2, -6(x_1 - 2x_2)^2, -9x_3^2].$$

At point $\mathbf{x} = [1, 1, -1/3]^T$ we have $\nabla f(\mathbf{x}) = [3, -6, -1]$.

If $f \in C^2$ (partial second derivatives exist) then we define the *Hessian* of f at \mathbf{x} to be the $n \times n$ matrix, denoted by $\nabla^2 f(\mathbf{x})$ or $\mathbf{H}(\mathbf{x})$ as

$$\mathbf{H}(\mathbf{x}) = \left[\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \right]_{i,j=1}^n.$$

Since

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

it is easy to see that the Hessian is symmetric. The $\frac{\partial^2 f}{\partial x_i \partial x_i}$ term is denoted by $\frac{\partial^2 f}{\partial x_i^2}$.

Though the entries of the Hessian are functions, for any given \mathbf{x} they determine a matrix of numerical values.

Example 6 Let again $f(\mathbf{x}) = (x_1 - 2x_2)^3 - 3x_3^3$. Then

$$\mathbf{H}(\mathbf{x}) = \begin{bmatrix} 6(x_1 - 2x_2) & -12(x_1 - 2x_2) & 0 \\ -12(x_1 - 2x_2) & 24(x_1 - 2x_2) & 0 \\ 0 & 0 & -18x_3 \end{bmatrix}.$$

At point $\mathbf{x} = [1, 1, -1/3]^T$

$$\mathbf{H}(\mathbf{x}) = \begin{bmatrix} -6 & 12 & 0 \\ 12 & -24 & 0 \\ 0 & 0 & 6 \end{bmatrix}.$$

The partial derivatives of an n -variable function $f(\mathbf{x})$ can characterize the extreme points of the function. It can be proved that for a point \mathbf{x}^* to be an extreme point, a necessary condition is that the gradient in this point vanishes, i.e., $\nabla f(\mathbf{x}^*) = \mathbf{0}$. If, in this point, the Hessian exists and is positive definite this condition is sufficient and f has a minimum in \mathbf{x}^* .

2.3.2 Taylor expansion

A group of important results of multivariate analysis are often referred to under the general heading of Taylor's Theorem. It enables us to approximate a function with an m -degree polynomial in a domain if it satisfies certain conditions.

First, we demonstrate the expansion for a function $f(x)$ of a single variable with $m = 1$ (linear approximation). If f is twice differentiable in an $[x_0, x]$ interval then

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{1}{2}(x - x_0)^2 f''(\theta).$$

where f' and f'' denote the first and second derivatives, respectively, and θ lies in the $[x_0, x]$ interval. The $\frac{1}{2}(x - x_0)^2 f''(\theta)$ part is the error term which is small if x is close to x_0 and the second derivative is bounded in the interval.

For a quadratic approximation we have to assume that f is three times differentiable. Details are not discussed here.

For a function $f(\mathbf{x})$, the first order version of the theorem (linear approximation) assumes that $f \in C^1$ in a region containing the line segment $[\mathbf{x}_0, \mathbf{x}]$. It states that in this case there exists a θ , $0 \leq \theta \leq 1$ such that

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\theta\mathbf{x}_0 + (1 - \theta)\mathbf{x})(\mathbf{x} - \mathbf{x}_0).$$

Verbally, the function at \mathbf{x} can be expressed as its value at \mathbf{x}_0 plus the gradient evaluated at some point in the line segment of $[\mathbf{x}_0, \mathbf{x}]$ (which is given as a convex linear combination of the two end points) times the difference of \mathbf{x} and \mathbf{x}_0 . This formula can be used to estimate the behaviour of the function in the neighbourhood of a given point (\mathbf{x}_0 , in this case) when we have some information about the gradient.

If $f \in C^2$ then the second order version (quadratic approximation) says that there exists a θ , $0 \leq \theta \leq 1$ such that

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \mathbf{H}(\theta \mathbf{x}_0 + (1 - \theta)\mathbf{x})(\mathbf{x} - \mathbf{x}_0),$$

where \mathbf{H} denotes the Hessian of f .

2.3.3 Newton's method for solving $f(x) = 0$

An immediate application of the Taylor expansion is Newton's root finding method for a twice differentiable function.

In order to find a root of $f(x)$, i.e., find an x such that $f(x) = 0$, take the first order approximation of $f(x)$ without the error term and set it equal to zero

$$f(x) = f(x_0) + (x - x_0)f'(x_0) = 0$$

From this,

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}, \quad \text{if } f'(x_0) \neq 0. \quad (4)$$

Because of the dropped term, it cannot be expected that x of (4) provides a solution. However, (4) can be used to generate a sequence

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad \text{as long as } f'(x_k) \neq 0. \quad (5)$$

In the neighbourhood of a root this sequence converges to this root. (5) is a powerful and widely used procedure. There are well discussed criteria for the convergence of this method.

2.3.4 Newton's method for $\min f(\mathbf{x})$

The idea behind Newton's method is that the function $f(\mathbf{x})$ to be minimized is approximated locally by a quadratic function and this approximate function is minimized exactly. Thus, near a point \mathbf{x}_k we can approximate $f(\mathbf{x})$ by the truncated Taylor series (omitting error term)

$$f(\mathbf{x}) = f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k)^T \mathbf{H}(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k).$$

The necessary condition for a local minimum is that the gradient of $f(\mathbf{x})$ vanishes

$$\nabla f(\mathbf{x}) = \nabla f(\mathbf{x}_k) + (\mathbf{x} - \mathbf{x}_k)^T \mathbf{H}(\mathbf{x}_k) = \mathbf{0}^T,$$

which gives

$$\mathbf{x} = \mathbf{x}_k - \mathbf{H}^{-1}(\mathbf{x}_k) (\nabla f(\mathbf{x}_k))^T.$$

This leads to the following iterative scheme:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}^{-1}(\mathbf{x}_k) (\nabla f(\mathbf{x}_k))^T,$$

which is the pure form of Newton's method. If $\{\mathbf{x}_k\} \rightarrow \mathbf{x}^*$ and at \mathbf{x}^* the Hessian $\mathbf{H}(\mathbf{x}^*)$ is positive definite then function $f(\mathbf{x})$ has a local minimum at \mathbf{x}^* . This method has an excellent (quadratic) convergence rate in the neighbourhood of a local minimum. To make it convergent in a wider radius, some more involved study is necessary.

3 Partial Differential Equations

While Ordinary Differential Equations (ODEs) are equations that relate a function of a single variable and its derivatives, a Partial Differential Equation (PDE) expresses a relationship between an unknown function of several variables and its partial derivatives. Let u be a function of two variables, x (spatial variable) and t (time variable), $u = u(x, t)$. We are concerned with PDEs of the following form

$$a \frac{\partial^2 u}{\partial x^2} + b \frac{\partial^2 u}{\partial x \partial t} + c \frac{\partial^2 u}{\partial t^2} + d \frac{\partial u}{\partial x} + e \frac{\partial u}{\partial t} + fu = g, \quad (6)$$

where a, \dots, f may be functions of x, t and even u .

The *order* of a PDE is given by the order of the highest derivative. Thus, if one of the functions a, b, c in (6) is nonzero then the PDE is *second order*. If $a = b = c = 0$ but at least one of d and e is nonzero then the PDE is *first order*.

If functions a, \dots, f do not depend on the dependent variable u then (6) is *linear*, otherwise it is *nonlinear*.

If $g = 0$ in (6) then the PDE is *homogeneous*, otherwise it is *inhomogeneous*.

Example 7 *First order PDEs.*

1. A linear homogeneous PDE:

$$\frac{\partial u}{\partial x} - \frac{\partial u}{\partial t} = 0$$

2. A nonlinear homogeneous PDE:

$$u \frac{\partial u}{\partial x} - \frac{\partial u}{\partial t} = 0$$

3. A linear inhomogeneous PDE:

$$e^x \frac{\partial u}{\partial x} + 2 \frac{\partial u}{\partial t} = t$$

Example 8 *Second order PDEs.*

1. A linear homogeneous PDE:

$$\frac{\partial^2 u}{\partial x^2} + 4 \frac{\partial u}{\partial t} = 0$$

2. A nonlinear homogeneous PDE:

$$\frac{\partial^2 u}{\partial x^2} - e^{2x} \frac{\partial^2 u}{\partial t^2} - u^2 = 0$$

There are several important types of PDEs. For computational finance the most important one is the *linear parabolic PDE* in two variables. Such a PDE has a second (and may have also a first) derivative with respect to one variable and a first derivative with respect to the other. These PDEs are often referred to as *heat* or *diffusion equations*.

3.1 Solution of PDEs

A *solution* to the (6) PDE is a function $u(x, t)$ that satisfies the equation.

Example 9 *The first order wave equation is*

$$c \frac{\partial u}{\partial x} + \frac{\partial u}{\partial t} = 0, \tag{7}$$

where c is a constant (wave speed). It is easy to see that $u(x, t) = \sin(x - ct)$ solves (7) because

$$c \frac{\partial u}{\partial x} = c \cos(x - ct), \quad \frac{\partial u}{\partial t} = -c \cos(x - ct)$$

and their sum is equal to zero.

Interestingly, it is equally easy to see that $u(x, t) = F(x - ct)$ is also a solution to (7) with an arbitrary differentiable function F . This observation highlights an important fact. Namely, ODEs have arbitrary constants but PDEs have arbitrary functions. The function F is determined by an initial condition.

Assume in this example that the initial condition at $t = 0$ is $u(x, 0) = \varphi(x)$, where φ is a given function. Now, if $u(x, t) = F(x - ct)$ then $u(x, 0) = F(x)$. Therefore, the unknown function F is the given function φ but the argument is replaced by $x - ct$. Hence, the solution is

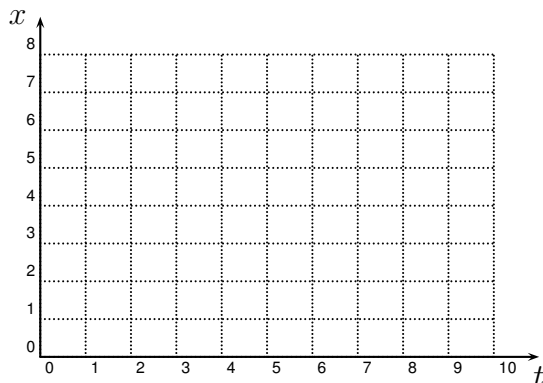
$$u(x, t) = \varphi(x - ct).$$

For instance, if $\varphi(x) = e^{-x^2}$ then $u(x, t) = e^{-(x-ct)^2}$ which satisfies (7) and the initial condition.

3.2 Finite difference method

Very often there is no analytical (closed-form) solution to a PDE or the form of the solution is awkward and unsuitable for obtaining numerical values easily. In such cases an approximate solution can be obtained by some numerical methods. The basic idea behind these methods is that the derivatives are approximated by the defining ratios, see (3) and the differential equations by finite difference equations. This leads to a discretization of the problem. In this way we can obtain the value of the unknown function in predefined discrete points.

For simplicity of the presentation of the ideas, we consider an unknown function in two variables, $u = u(x, t)$. A finite difference grid is created by equidistant points along both axes like this



Note, the stepsize can be different for the two variables. In this example the points along the t axis are defined by $t_j = jk$, $j = 0, \dots, n$, where k is the distance between adjacent grid points (increment). It gives $n + 1$ points. Similarly, grid points along the x axis are defined by $x_i = ih$, $i = 0, \dots, m$, where h is the x -increment resulting in $m + 1$ points. (There can be other shapes, for example, triangle, where m depends on the actual value of t_j .)

The approximate value of u in the grid points is denoted by $u_{ij} \approx u(x_i, t_j)$. The initial condition $u(x, 0) = \varphi(x)$ determines the exact values $u_{i0} = \varphi(x_i)$. The first derivative in the grid points with respect to t can be approximated in several different ways using the (assumed known) accurate or approximate values of $u(x, t)$. First, we consider the partial derivatives with respect to t . In practice, the *central difference*

$$\frac{\partial}{\partial t} u(x_i, t_j) \approx \frac{u(ih, (j+1)k) - u(ih, (j-1)k)}{2k} \quad (8)$$

$$\approx \frac{u_{i,j+1} - u_{i,j-1}}{2k}, \quad i = 0, \dots, m, \quad j = 1, \dots, n-1, \quad (9)$$

is more accurate than the *forward*

$$\frac{\partial}{\partial t} u(x_i, t_j) \approx \frac{u(ih, (j+1)k) - u(ih, jk)}{k} \quad (10)$$

$$\approx \frac{u_{i,j+1} - u_{i,j}}{k}, \quad i = 0, \dots, m, \quad j = 0, \dots, n-1, \quad (11)$$

or backward difference

$$\frac{\partial}{\partial t} u(x_i, t_j) \approx \frac{u(ih, (j-1)k) - u(ih, jk)}{k} \quad (12)$$

$$\approx \frac{u_{i,j-1} - u_{i,j}}{k}, \quad i = 0, \dots, m, \quad j = 1, \dots, n. \quad (13)$$

However, different situations may necessitate the use of any one of them (even if it is not the most accurate one).

The accuracy of the above approximations can be characterized by the error they carry. It can be obtained from the Taylor expansion. It turns out that the (8) central difference has an error of $O(k^2)$ while the (10) forward and (12) backward differences have error terms of $O(k)$ each. As $k < 1$ (typically $k \ll 1$) the central difference is substantially more accurate.

The second partial derivative with respect to x can be approximated in the following way (using the exact or approximate values of $u(x, t)$):

$$\frac{\partial^2}{\partial x^2} u(x_i, t_j) \approx \frac{u((i+1)h, jk) - 2u(ih, jk) + u((i-1)h, jk)}{h^2} \quad (14)$$

$$\approx \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}, \quad i = 1, \dots, m-1, \quad j = 0, \dots, n. \quad (15)$$

The error of the (14) approximation is $O(h^2)$. In practice, we usually have $h \ll 1$.

We can conclude that the error can be reduced if the stepsize is made smaller. However, in this case more grid points are generated for a given area and the method becomes computationally more demanding. The proper choice of h and k is an important decision. These values primarily depend on the required accuracy of the solution.

Example 10 Find the $u(x, t)$ function in the $t \geq 0$ half plain which, for $t > 0$, satisfies the

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

partial differential equation and also the

$$u(x, 0) = \varphi(x) \quad (16)$$

initial condition, where $\varphi(x)$ is a bounded continuous function.

Using (11) and (15) we obtain

$$\frac{\partial}{\partial t} u(ih, jk) \approx \frac{u_{i,j+1} - u_{i,j}}{k} \quad \text{and} \quad \frac{\partial^2}{\partial x^2} u(ih, jk) \approx \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}.$$

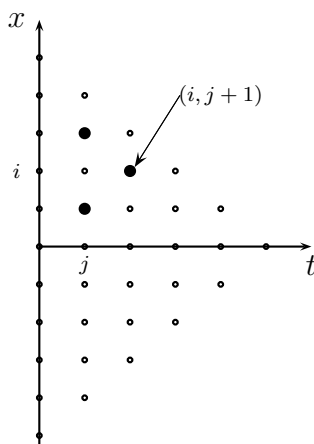
Equating and rearranging them we obtain

$$u_{i,j+1} - u_{i,j} = \frac{k}{h^2} (u_{i+1,j} - 2u_{i,j} + u_{i-1,j}).$$

If the stepsizes are chosen to satisfy the $k = \frac{1}{2}h^2$ relation $u_{i,j+1}$ can be expressed as

$$u_{i,j+1} = \frac{1}{2}(u_{i+1,j} + u_{i-1,j}). \quad (17)$$

The (16) initial condition enables us to compute u_{i0} directly from $u_{i0} = \varphi(ih)$ for $i = 0, \pm 1, \pm 2, \dots$. Therefore, the approximate function values of u can be obtained from (17) for $i = 0, \pm 1, \pm 2, \dots$ and $j = 0, 1, 2, \dots$ on a triangular domain as shown below. The value in position $(i, j + 1)$ is determined by positions $(i - 1, j)$ and $(i + 1, j)$.



For a rectangular domain additional boundary conditions are needed to uniquely determine the approximations of the derivatives at (or on) the boundary.

4 Random variables, probability

Computational finance deals with uncertain events. The most suitable way of modelling such events is using *random variables* (RVs). Below is a very brief survey of the most important concepts of probability theory that will be used later during this course.

There are two types of RVs: discrete and continuous.

4.1 Discrete random variables

They are usually denoted by ξ and can take one of many possible values with a given probability. As an example, consider throwing a die. The possible values on the top face are $\{1, 2, 3, 4, 5, 6\}$. In case of a fair die each outcome has an equal probability of $1/6$. In general, the possible outcomes of ξ are $\{x_1, \dots, x_n\}$ with probabilities $\{p_1, \dots, p_n\}$. The probability that ξ takes the value x_i is p_i . Formally, we write $P(\xi = x_i) = p_i$. Sometimes notation $\text{Prob}(\xi = x_i) = p_i$ is also used. Each $p_i \geq 0$ and $\sum_i p_i = 1$.

The *expected value* or *mean* of ξ is defined as

$$E(\xi) = \sum_{i=1}^n p_i x_i.$$

Sometimes it is denoted by μ_ξ . If there are infinitely many possible outcomes of the discrete event then summation goes from 1 to $+\infty$. The expected value gives the centre of the distribution of the random variable.

The *variance* of ξ , denoted by $\text{VAR}(\xi)$ or σ_ξ^2 , is defined to be the expected value of the squared deviation of ξ from its mean μ_ξ . When it is obvious what random variable is being considered it is possible to drop ξ from the subscript. Therefore,

$$\sigma^2 = E[(\xi - \mu)^2] = \sum_{i=1}^n p_i (x_i - \mu)^2.$$

It can easily be shown that

$$\sigma^2 = E[\xi^2] - (E[\xi])^2.$$

The variance is a measure of the spread of the distribution about the mean. It is the second central moment of ξ . The square root of the variance, σ , is called *standard deviation*.

Some frequently used discrete probability distributions are

1. *Binomial distribution.* There is an experiment with two possible outcomes, A and B , with respective probabilities p and $1 - p$. We perform the experiment n times. ξ denotes the number of occurrences of outcome A . The probability that A occurs exactly k times ($k \leq n$) is $P(\xi = k) = \binom{n}{k} p^k (1 - p)^{n-k}$. It can be shown that in this case $\mu = np$ and $\sigma^2 = np(1 - p)$.
2. *Poisson distribution.* The Poisson distribution occurs as the limiting form of the binomial distribution when the probability $p \rightarrow 0$ and the number of trials $n \rightarrow +\infty$, such that the mean $\mu = np$, remains finite. The probability of observing k events in this limit then reduces to $P(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda}$, where $\lambda = \lim_{n \rightarrow +\infty} np$. For the Poisson distribution $\mu = \lambda$ and $\sigma^2 = \lambda$.

4.2 Continuous random variables

Continuous random variables, also denoted also by ξ , can take any real value within a specified interval. In this case it is meaningless to ask whether ξ takes any specific value because its probability is zero. However, the probability of ξ falling into an interval with length greater than zero does make sense. The interval can be finite or infinite. The

probability distribution function of ξ is defined as $F(x) = \text{Prob}(\xi \leq x)$. Its derivative is called *density function*, i.e., $f(x) = F'(x) = dF(x)/dx$. $F(x)$ can be expressed as

$$F(x) = \int_{-\infty}^x f(t)dt.$$

Consequence: $\text{Prob}(a \leq \xi \leq b) = F(b) - F(a) = \int_a^b f(t)dt$.

The mean of a continuous ξ is defined as

$$\mu = E(\xi) = \int xf(x)dx.$$

The variance of ξ is

$$\sigma^2 = E[\xi^2] - (E[\xi])^2 = \int (x - \mu)^2 f(x)dx = \int x^2 f(x)dx - \mu^2.$$

Some frequently used continuous probability distributions are

1. *Uniform distribution* is defined by two parameters, a and b with $a \leq b$ and is referred to as $U(a, b)$. Its distribution and density functions, respectively, are

$$F(x) = \begin{cases} \frac{x-a}{b-a}, & \text{if } a \leq x \leq b, \\ 0, & \text{if } x < a, \\ 1, & \text{if } x > b, \end{cases} \quad f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

For $U(a, b)$: $\mu = (a + b)/2$ and $\sigma^2 = (b - a)^2/12$.

2. *Normal distribution* (also called *Gauss distribution*) is also defined by two parameters μ and σ and is referred to as $N(\mu, \sigma)$. The density function is

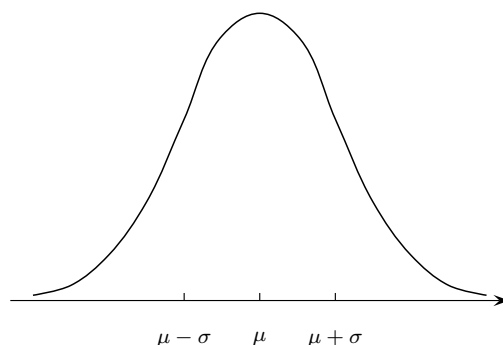
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty \leq x \leq +\infty.$$

The probability distribution function $F(x)$ has no closed form. It is given by the integral of the density function $f(x)$:

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt.$$

There are tables available that contain the value of $F(x)$ in discrete points for different combinations of the (μ, σ) pair.

The density function of $N(\mu, \sigma)$ is symmetric around μ and is known as the *bell shaped function*.



For $N(\mu, \sigma)$ the mean is μ and the variance is σ^2 .

There is an important special case, $N(0, 1)$, which is called the *standard normal distribution*.

The normal distribution is the most important probability distribution. In many real world models there is a strong empirical evidence for normality. Additionally, the sum of arbitrary random variables also behaves near normally if there are enough random variables involved. This is stated more precisely and proved by the *central limit theorem*.

3. A random variable has a *lognormal distribution* if its logarithm is normally distributed.

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma'} e^{-\frac{1}{2}\left(\frac{y-\mu'}{\sigma'}\right)^2}, \quad -\infty \leq y \leq +\infty,$$

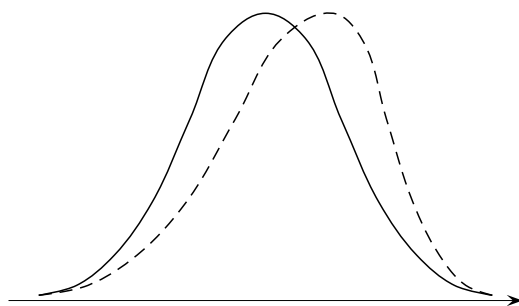
where $y = \ln x$, $\mu' = \ln \mu$ and $\sigma' = \ln \sigma$. Obviously, $x > 0$ is assumed.

4.3 Skewness and kurtosis

The variance of ξ is the second moment of ξ about the mean and it measures the spread of the distribution of ξ about the mean. The third and fourth moments of ξ about the mean also measure interesting features of the distribution. The third moment measures *skewness*, the lack of symmetry, while the fourth moment measures *kurtosis*, the degree to which the distribution is peaked. The actual numerical measures of these characteristics are standardized to eliminate the physical units, by dividing by an appropriate power of the standard deviation.

The skewness of ξ is defined to be

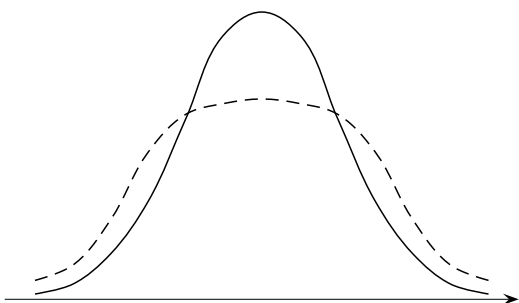
$$skew(\xi) = \frac{E[(\xi - \mu)^3]}{\sigma^3}.$$



Dashed line represents a negatively skewed distribution.

The kurtosis of ξ is defined to be

$$\text{kurt}(\xi) = \frac{E[(\xi - \mu)^4]}{\sigma^4} - 3.$$



Dashed line represents a distribution with negative kurtosis.

4.4 Covariance, correlation

If two or more random variables are considered, their mutual dependence can be characterized by a measure called *covariance*. Let ξ_1 and ξ_2 be two random variables with expected values μ_1 and μ_2 and variances σ_1 and σ_2 . The covariance of these variables (denoted by $\text{cov}(\xi_1, \xi_2)$ or σ_{12}) is defined as

$$\begin{aligned} \text{cov}(\xi_1, \xi_2) &= \sigma_{12} \\ &= E[(\xi_1 - \mu_1)(\xi_2 - \mu_2)] \\ &= E(\xi_1 \xi_2) - \mu_1 \mu_2 \end{aligned}$$

Notice that $\sigma_{12} = \sigma_{21}$.

The *correlation* of ξ_1 and ξ_2 is defined if their variances are positive. It is nothing but a scaled version of the covariance.

$$\text{cor}(\xi_1, \xi_2) = \frac{E[(\xi_1 - \mu_1)(\xi_2 - \mu_2)]}{\sigma_1 \sigma_2}.$$

If two random variables are independent they are called *uncorrelated* and $\sigma_{12} = 0$. If $\sigma_{12} > 0$, then two variables are said to be *positively correlated*. In this case, if one variable is above its mean, the other is likely to be above its mean as well. If $\sigma_{12} < 0$, the two variables are said to be *negatively correlated*.

5 Optimization

Optimization is a search for an extremal value (max or min) of a function $f(\mathbf{x})$ over a set $S \subseteq \mathbb{R}^n$. More formally, stating optimization as a minimization problem,

$$\min f(\mathbf{x}) \tag{18}$$

$$\text{subject to } \mathbf{x} \in S. \tag{19}$$

Function f is called the *objective function* and S is the *feasible set*. If there exists an $\mathbf{x}^* \in S$ such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for any $\mathbf{x} \in S$ then \mathbf{x}^* is called an *optimal solution*.

Any $\mathbf{x} \in S$ (see (19)) is called a *feasible solution* to the problem. S is usually defined as all points that satisfy a set of functional equalities and/or inequalities. In the general form

$$S = \{\mathbf{x} : g_i(\mathbf{x}) = 0, i = 1, \dots, m, h_j(\mathbf{x}) \geq 0, j = 1, \dots, p\},$$

or more concisely,

$$S = \{\mathbf{x} : \mathbf{g}(\mathbf{x}) = \mathbf{0}, \mathbf{h}(\mathbf{x}) \geq \mathbf{0}\}.$$

Quite often (many of) the variables are restricted to be nonnegative. This requirement can be included in the second set of constraints as $\mathbf{x} \geq \mathbf{0}$.

If all functions, $f(\mathbf{x})$, $\mathbf{g}(\mathbf{x})$, $\mathbf{h}(\mathbf{x})$ are linear then we have a *linear programming* (LP) problem. If any of these functions is nonlinear (18)–(19) is called a *nonlinear programming* (NLP) problem. Convexity of the feasible set S is a desirable feature.

If the functions and the constraints are completely (with certainty) known we talk about *deterministic optimization*. If uncertainty (by means of random variables) is included in the model in any way it is called *stochastic optimization*.

5.1 Quadratic programming

Of particular interest in NLP is *quadratic programming* (QP). In QP the objective function is a quadratic form, all constraints are linear and the variables are restricted to take nonnegative values. We consider the QP problem in the following form:

$$\min f(\mathbf{x}) = \mathbf{c}^T \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} \tag{20}$$

$$\text{subject to } \mathbf{A} \mathbf{x} \leq \mathbf{b} \tag{21}$$

$$\mathbf{x} \geq \mathbf{0}, \tag{22}$$

where \mathbf{c} , $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is symmetric, i.e., $\mathbf{Q} = \mathbf{Q}^T$. Obviously, any set of linear constraints can be brought into the form of (21). Now, the feasible points are the ones that satisfy (21) and (22). If $\mathbf{Q} = \mathbf{O}$ (null matrix) then QP reduces to LP. The (21) – (22) feasible set is convex.

If $f(\mathbf{x})$ is a strictly convex function for all feasible points the QP problem has a unique minimum. A sufficient condition for $f(\mathbf{x})$ to be strictly convex is that \mathbf{Q} is positive definite.

Reminder: \mathbf{Q} is positive definite if $\mathbf{x}^T \mathbf{Q} \mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$.

5.2 Lagrangian function for QP

Rearranging constraints (21) and (22) to $\mathbf{Ax} - \mathbf{b} \leq \mathbf{0}$ and $-\mathbf{x} \leq \mathbf{0}$ the *Lagrangian function* of the QP problem in (20) – (22) is defined as

$$L(\mathbf{x}, \boldsymbol{\mu}, \bar{\boldsymbol{\mu}}) = \mathbf{c}^T \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} + \boldsymbol{\mu}^T (\mathbf{Ax} - \mathbf{b}) - \bar{\boldsymbol{\mu}}^T \mathbf{x}. \quad (23)$$

where $\boldsymbol{\mu} = [\mu_1, \dots, \mu_m]^T$ and $\bar{\boldsymbol{\mu}} = [\bar{\mu}_1, \dots, \bar{\mu}_n]^T$ are the vectors of *Lagrange multipliers* of the constraints and the variables, respectively.

5.3 Optimality conditions

The *first order necessary conditions* for a solution to be optimal can be expressed in terms of the Lagrangian function. They are called the *Karush-Kuhn-Tucker (KKT) conditions* of optimality. We present them below without proof.

Component form	Matrix/vector form	
$\frac{\partial L}{\partial x_j} = 0, j = 1, \dots, n$	$\mathbf{c} + \mathbf{Qx} + \mathbf{A}^T \boldsymbol{\mu} - \bar{\boldsymbol{\mu}} = \mathbf{0}$	(24)

$\frac{\partial L}{\partial \mu_i} \leq 0, i = 1, \dots, m$	$\mathbf{Ax} - \mathbf{b} \leq \mathbf{0}$	(25)
---	--	------

$\frac{\partial L}{\partial \bar{\mu}_j} \leq 0, j = 1, \dots, n$	$-\mathbf{x} \leq \mathbf{0}$	(26)
---	-------------------------------	------

$\frac{\partial L}{\partial \mu_i} \mu_i = 0, i = 1, \dots, m$	$\boldsymbol{\mu}^T (\mathbf{Ax} - \mathbf{b}) = 0$	(27)
--	---	------

$\frac{\partial L}{\partial \bar{\mu}_j} \bar{\mu}_j = 0, j = 1, \dots, n$	$\bar{\boldsymbol{\mu}}^T \mathbf{x} = 0$	(28)
--	---	------

$\mu_i \geq 0, i = 1, \dots, m$	$\boldsymbol{\mu} \geq \mathbf{0}$	(29)
---------------------------------	------------------------------------	------

$\bar{\mu}_j \geq 0, j = 1, \dots, n$	$\bar{\boldsymbol{\mu}} \geq \mathbf{0}$	(30)
---------------------------------------	--	------

Constraint (25) can be converted into equalities by the introduction of a nonnegative vector of slack variables, $\mathbf{v} \in \mathbb{R}^m$. After moving the constants to the right-hand-side (RHS) we obtain

$$\mathbf{Qx} + \mathbf{A}^T \boldsymbol{\mu} - \bar{\boldsymbol{\mu}} = -\mathbf{c} \quad (31)$$

$$\mathbf{Ax} + \mathbf{v} = \mathbf{b} \quad (32)$$

$$\bar{\boldsymbol{\mu}}^T \mathbf{x} = 0 \quad (33)$$

$$\boldsymbol{\mu}^T \mathbf{v} = 0 \quad (34)$$

$$\mathbf{x} \geq \mathbf{0}, \quad \boldsymbol{\mu} \geq \mathbf{0}, \quad \bar{\boldsymbol{\mu}} \geq \mathbf{0}, \quad \mathbf{v} \geq \mathbf{0} \quad (35)$$

Constraints (33) and (34) are referred to as *complementarity* constraints, while (35) expresses the nonnegativity of the variables. In fact, $\boldsymbol{\mu} \geq \mathbf{0}$ says the Lagrange multipliers of inequality constraints are nonnegative. It is to be noted if there is an equality constraint $\mathbf{a}^i \mathbf{x} = b_i$ then the corresponding Lagrange multiplier μ_i is unrestricted in sign. An equality can also be written as two inequalities, like $\mathbf{a}^i \mathbf{x} \leq b_i$ and $-\mathbf{a}^i \mathbf{x} \leq -b_i$ and the two multipliers of these constraints will be nonnegative again.

Complementarity says that at an optimal solution \mathbf{x} for every i either $\mu_i = 0$ or the inequality is tight, i.e., satisfied as an equality, $\mathbf{a}^i \mathbf{x} = b_i$ (can be both).

It can be shown that the Hessian of the (23) Lagrangian function is \mathbf{Q} . Therefore, if \mathbf{Q} is positive definite the KKT conditions are also sufficient for optimality. Having found an \mathbf{x} that satisfies KKT we can be sure it solves (minimizes) the QP problem.

(31) and (32), together with (35) define a linear programming feasibility problem with $m + n$ constraints and $2(m + n)$ variables. Unfortunately, the (33) and (34) complementarity constraints are nonlinear. Still this problem can be solved by a slightly modified version of the simplex method (which is a powerful solution algorithm for LP) by applying a *restricted basis entry rule*. However, interior point methods seem to be more efficient in finding a solution to (31) – (35).

Example 11 Find the KKT conditions of the following QP problem

$$\begin{aligned} \min f(x_1, x_2) = & 3x_1 - 2x_2 + 4x_1^2 + 2x_1x_2 + 5x_2^2 \\ \text{s.t.} & 2x_1 + x_2 \geq 2 \\ & -x_1 + 2x_2 \leq 6 \\ & x_1, x_2 \geq 0 \end{aligned}$$

Solution:

After implied rewriting:

$$\begin{aligned} \min f(x_1, x_2) = & 3x_1 - 2x_2 + \frac{1}{2}(8x_1^2 + 2x_1x_2 + 2x_2x_1 + 10x_2^2) \\ \text{s.t.} & -2x_1 - x_2 \leq -2 \\ & -x_1 + 2x_2 \leq 6 \\ & x_1, x_2 \geq 0 \end{aligned}$$

The relevant vectors and matrices are

$$\mathbf{c} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 8 & 2 \\ 2 & 10 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} -2 & -1 \\ -1 & 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} -2 \\ 6 \end{bmatrix}.$$

The (31) – (35) form of the KKT conditions, without the complementarity equations,

$$\begin{aligned} 8x_1 + 2x_2 - 2\mu_1 - \mu_2 - \bar{\mu}_1 & = -3 \\ 2x_1 + 10x_2 - \mu_1 + 2\mu_2 - \bar{\mu}_2 & = 2 \\ -2x_1 - x_2 + v_1 & = -2 \\ -x_1 + 2x_2 + v_2 & = 6 \\ x_1, x_2, \mu_1, \mu_2, \bar{\mu}_1, \bar{\mu}_2, v_1, v_2 & \geq 0. \end{aligned}$$

Finally, the (33) – (34) complementarity constraints

$$\bar{\mu}_1 x_1 + \bar{\mu}_2 x_2 = 0 \quad \text{and} \quad \mu_1 v_1 + \mu_2 v_2 = 0.$$