

# **MECH 577 Optimum Design**

## **Lecture Notes**

**Jorge Angeles**

Department of Mechanical Engineering &  
Centre for Intelligent Machines  
McGill University, Montreal (Quebec), Canada

© January 14, 2013

These lecture notes are not as yet in final form.

Please report corrections & suggestions to

Prof. J. Angeles

Department of Mechanical Engineering &

McGill Centre for Intelligent Machines

McGill University

817 Sherbrooke St. W.

Montreal, Quebec

CANADA H3A 2K6

FAX: (514) 398-7348

[angeles@cim.mcgill.ca](mailto:angeles@cim.mcgill.ca)

# Contents

<b>Notation</b>	<b>7</b>
<b>1 Preliminaries</b>	<b>11</b>
1.1 The Role of Optimization Within the Design Process . . . . .	11
1.2 The Structure of Optimum Design Problems . . . . .	16
1.2.1 Case Study: The Optimum Design of the Links of a Robotic Wrist . . . . .	17
<b>2 Single-Variable Optimization</b>	<b>19</b>
2.1 Methods of Single-Variable Optimization . . . . .	19
2.2 Dichotomous Search . . . . .	21
2.3 Interval-Halving . . . . .	29
2.4 Fibonacci Numbers . . . . .	30
2.5 Golden-Section Search . . . . .	40
<b>3 Numerical Equation Solving</b>	<b>47</b>
3.1 Introduction . . . . .	47
3.2 Background Facts and Definitions . . . . .	48
3.3 Background on Linear Transformations . . . . .	52
3.3.1 Rotations . . . . .	53
3.3.2 Reflections . . . . .	56
3.3.3 Projections . . . . .	57
3.4 The Numerical Solution of Determined Linear Systems of Equations .	59
3.4.1 Roundoff Error of the Solution and Condition Numbers . . . .	60
3.4.2 LU-Decomposition, a.k.a. Gaussian Elimination . . . . .	61
3.4.3 Cholesky Decomposition . . . . .	62
3.5 The Least-square Solution of Overdetermined Linear Systems . . . .	62
3.5.1 The Normal Equations for Plain Least Squares . . . . .	63

3.5.2	The Normal Equations for Weighted Least Squares . . . . .	66
3.5.3	The Numerical Solution of Least-square Problems . . . . .	67
3.6	Nonlinear-equation Solving: the Determined Case . . . . .	73
3.6.1	The Newton-Raphson Method . . . . .	75
3.7	Overdetermined Systems of Nonlinear Equations . . . . .	79
3.7.1	The Newton-Gauss Method . . . . .	79
3.7.2	Convergence Criterion . . . . .	80
3.8	Computer Implementation Using ODA— C-Library of Routines for Optimum Design . . . . .	85
<b>4</b>	<b>Unconstrained Optimization</b>	<b>87</b>
4.1	Introduction . . . . .	87
4.2	The Normality Conditions . . . . .	87
4.3	Methods of Solution . . . . .	89
4.4	Direct Methods . . . . .	90
4.4.1	The Hooke and Jeeves Method . . . . .	90
4.4.2	The Powell Method (Conjugate Directions) . . . . .	91
4.4.3	The Nelder-Mead Simplex Method . . . . .	96
4.5	Gradient Methods . . . . .	100
4.5.1	The Method of Steepest Descent(Cauchy) . . . . .	100
4.5.2	The Conjugate-Gradient Method (Fletcher-Reeves) . . . . .	101
4.5.3	Quasi-Newton Methods . . . . .	109
4.6	Newton Methods . . . . .	111
4.6.1	The Newton-Raphson Method . . . . .	111
4.6.2	The Levenberg-Marquardt Method . . . . .	112
<b>5</b>	<b>Equality-Constrained Optimization: Normality Conditions</b>	<b>115</b>
5.1	Introduction . . . . .	115
5.2	The First-Order Normality Conditions . . . . .	116
5.2.1	The Primal Form . . . . .	116
5.2.2	The Dual Form . . . . .	119
5.3	The Second-Order Normality Conditions . . . . .	123
5.3.1	A Mechanical Interpretation of the Lagrange Multipliers . . .	141
5.4	Linear-Quadratic Problems . . . . .	143
5.4.1	The Minimum-Norm Solution of Underdetermined Systems . .	143
5.4.2	Least-Square Problems Subject to Linear Constraints . . . . .	149

5.5	Equality-Constrained Nonlinear Least Squares . . . . .	151
5.6	Linear Least-Square Problems Under Quadratic Constraints . . . . .	153
<b>6</b>	<b>Equality-Constrained Optimization:</b>	
	<b>The Orthogonal-Decomposition Algorithm</b>	<b>161</b>
6.1	Introduction . . . . .	161
6.2	Linear Least-square Problems Subject to Linear Equality Constraints: the ODA . . . . .	162
6.3	Equality-Constrained Nonlinear Least-Square Problems . . . . .	166
6.3.1	A Geometric Interpretation of the ODA . . . . .	182
6.4	Equality-Constrained Optimization with Arbitrary Objective Function	190
6.4.1	A Sequential-Quadratic Programming Approach . . . . .	192
6.4.2	A Gradient-based Approach . . . . .	222
<b>7</b>	<b>Inequality-Constrained Optimization</b>	<b>223</b>
7.1	Introduction . . . . .	223
7.2	The Karush-Kuhn-Tucker Conditions . . . . .	224
7.3	Second-Order Normality Conditions . . . . .	231
7.3.1	Overconstrained Problems . . . . .	233
7.4	Methods of Solution . . . . .	234
7.5	Indirect Methods . . . . .	236
7.5.1	Slack Variables . . . . .	236
7.5.2	Penalty Functions . . . . .	239
7.6	Direct Methods . . . . .	244
7.6.1	The Generalized Reduced-Gradient Method . . . . .	244
7.6.2	The Complex Method . . . . .	255
	<b>Bibliography</b>	<b>259</b>
	<b>Index</b>	<b>263</b>



# Notation

$\mathbf{0}_n$ : The  $n$ -dimensional zero vector

$\mathbf{1}$ : The  $n \times n$  identity matrix, when  $n$  is obvious

$\mathbf{1}_k$ : the  $k \times k$  identity matrix, when  $k$  should be specified

$\mathbf{A}$ :  $q \times n$  coefficient matrix of the linear system  $\mathbf{Ax} = \mathbf{b}$

$\mathbf{A}^I$ : the *left* Moore-Penrose generalized inverse (LMPGI) of the *full-rank*  $q \times n$  matrix  $\mathbf{A}$ , with  $q > n$ :

$$\mathbf{A}^I \equiv (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \quad (1)$$

$\mathbf{b}$ :  $q$ -dimensional vector of the linear system  $\mathbf{Ax} = \mathbf{b}$

$\mathbf{C}$ :  $p \times n$ , with  $p < n$ , coefficient matrix of the *underdetermined* linear system  $\mathbf{Cx} = \mathbf{d}$

$\mathbf{C}^\dagger$ : the *right* Moore-Penrose generalized inverse (RMPGI) of the *full-rank*  $p \times n$  matrix  $\mathbf{C}$ , with  $p < n$ :

$$\mathbf{C}^\dagger \equiv \mathbf{C}^T (\mathbf{C} \mathbf{C}^T)^{-1} \quad (2)$$

$\mathbf{d}$ :  $p$ -dimensional vector of the linear system  $\mathbf{Cx} = \mathbf{d}$

$f$ : scalar objective function  $f(\mathbf{x})$  to be minimized

$\mathbf{g}(\mathbf{x})$ :  $p$ -dimensional nonlinear vector function of the set of inequalities  $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$

$\mathbf{G}$ :  $p \times n$  Jacobian matrix of vector function  $\mathbf{g}(\mathbf{x})$  w.r.t.  $\mathbf{x}$

$\mathbf{H}$ : the  $n \times n$  Hessian matrix of the objective function  $f(\mathbf{x})$

$\mathbf{H}_i$ :  $i$ th Householder reflection used to render a rectangular matrix into upper-triangular form; a square matrix

$\mathbf{h}(\mathbf{x})$ :  $l$ -dimensional nonlinear vector function of  $\mathbf{x}$ , occurring in the equality constraints  $\mathbf{h}(\mathbf{x}) = \mathbf{0}$

$\mathbf{J}(\mathbf{x})$ :  $l \times n$  gradient of  $\mathbf{h}$  w.r.t.  $\mathbf{x}$

$\mathbf{L}$ : lower-triangular matrix of the LU-decomposition of a square matrix  $\mathbf{A}$ . Also used to denote the *orthogonal complement* of  $\mathbf{C}$  or  $\mathbf{G}$ ; confusion is avoided because of the two different contexts in which these matrices occur

$l$ : number of equality constraints  $h_i(\mathbf{x}) = 0$ , for  $i = 1, \dots, l$ , expressed in vector form as  $\mathbf{h}(\mathbf{x}) = \mathbf{0}$

$m$ : number of equations  $\phi_i(\mathbf{x}) = 0$ , for  $i = 1, \dots, m$ , expressed in vector form as  $\phi(\mathbf{x}) = \mathbf{0}$

$n$ : number of design variables  $x_i$ , for  $i = 1, \dots, n$ , expressed in vector form as  $\mathbf{x}$

$\mathbf{O}_{mn}$ : the  $m \times n$  zero matrix

$p$ : number of constraint equations  $g_i(x) = 0$ , for  $i = 1, \dots, p$ , expressed in vector form as  $\mathbf{g}(\mathbf{x}) = \mathbf{0}$  or  $\mathbf{C}\mathbf{x} = \mathbf{d}$

$q$ : number of equations in  $\mathbf{A}\mathbf{x} = \mathbf{b}$

$\mathbb{R}^n$ : the  $n$ -dimensional vector space defined over the real field  $\mathbb{R}$

$\mathbf{U}$ : (square) upper-triangular matrix

$\mathbf{V}$ :  $m \times m$  lower-triangular matrix, a factor of  $\mathbf{W}$ , i.e.,  $\mathbf{W} = \mathbf{V}^T \mathbf{V}$

$\mathbf{W}$ :  $m \times m$  symmetric and positive-semidefinite weighting matrix

$\mathbf{x}$ :  $n$ -dimensional vector of design variables

$\mathbf{x}_o$ : minimum-norm solution of an underdetermined linear system

$\mathbf{x}_L$ : least-square solution of an overdetermined linear system

$x_i$ : the  $i$ th component of vector  $\mathbf{x}$

$\mathbf{x}^k$ : the  $k$ th entry of a sequence  $\mathbf{x}^0, \mathbf{x}^1, \dots$

$\nabla$ : the *gradient operator*, pronounced “nabla”; when its operand is a scalar, it yields a vector; when a vector, it yields a matrix



$\nabla\nabla$ : the *Hessian operator*, its operand being a scalar, it produces a square, symmetric matrix

$\| \cdot \|$ : a norm of either vector or matrix ( $\cdot$ )



# Chapter 1

## Preliminaries

### 1.1 The Role of Optimization Within the Design Process

The English word *design* derives from the Latin word *designare*, which means “to mark out”—as found, for example, in the *Random College Dictionary*. The word thus implies a goal, an objective. As such, the meaning of the word is extremely broad, encompassing the general activity of producing concepts aimed at a given goal, be this pure intellectual pleasure, in the realm of art, or pragmatic, in the realm of engineering.

The product of the design activity is a good, whether tangible, e.g., a fountain pen in the realm of *industrial design*, or intangible, e.g., a business plan, in the realm of *management*. We focus here on *engineering design*, but this does not mean that we exclude intangible goods. An important branch of engineering is *production systems*, whereby the design good is many a time intangible, such as the *organization* of a healthcare system.

Design is an extremely complex process. Various models have been proposed in the literature, e.g., the one proposed by French (1992) and shown in Fig. 1.1, which divides the process into two parts, (i) stages/descriptions and (ii) activities. The process is represented as a flow diagram, in which stages or descriptions are included in circles, while activities in rectangles. In this model, the process starts with a need and ends with “working drawings.” In-between, we have a sequence, starting by the *analysis of the problem*, an activity, followed by the *statement of the problem*, a stage. Once the problem has been formulated, in design engineering terms, we

suppose, as opposed to “client-needs” terms, the activity leading to the *conceptual design* follows, out of which comes (come) the “selected scheme(s),” apparently a description. Then comes the “embodiment of scheme(s),” i.e., of the scheme(s) selected in the previous part. The embodiment is then followed by the “detailing” of the designed object, also an activity.

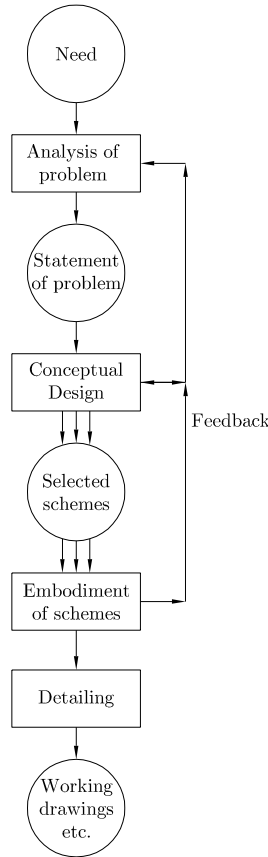


Figure 1.1: French’s model of the design process

However, the design process is recognized as being anything but a one-way street. Two feedback loops are thus included in French’s model: one at the conceptual design level, in which a revision may indicate that the problem needs further analysis, and hence, the designer or the design team must return to the first activity. A second loop arises at the embodiment level, at which the designer may realize that either a revision of the conceptual design is needed, or even a revision of the problem analysis is warranted.

The foregoing model contemplates the design process as a consultant’s activity, in which a consultant—an individual or a company—participates in a project within

an organization, to either develop a new product or improve an existing one. The model does not contemplate the prototyping aspects—usually outside of the scope of the consultant’s activities—that lead to the realization of the design motivated by the client’s need. Prototyping is needed when either an innovative product is under design or when an improvement on an existing design is planned that will affect thousands or millions of produced objects. Prototype tests may bring about various feedback loops in turn. We propose in Fig. 1.2 an alternative model that should help better understand the role of optimization within the design process and its place therein.

In the model of Fig. 1.2, the design process is initiated by a client’s need. The designer then translates the need into a form that allows the designer to analyze the client’s needs within a design context, which is the block indicated as “Problem Definition.” Once the problem is well-defined, free of the fuzziness of the client’s description, a search for alternative solutions begins, leading to a design candidate. These activities take place within the “Preliminary Design<sup>1</sup>” box in the proposed model. Once a design candidate has been selected, a detailed design follows, as included in the dashed box of the same model. In this phase, the input is a preliminary design solution, devoid of details. That is, the preliminary design is nothing but a rough layout needing an *embodiment*, i.e., a detailed definition in terms of materials, dimensions, and so on. The first step in this stage is the *synthesis* of the embodiment, i.e., a *topological* layout of the design—number of moving parts; types of moving parts; individual functions of the parts; etc.—involving only a qualitative description of the design structure. Once a topological layout has been produced, its design features are identified and labelled, probably using mathematical symbols, such as  $\ell$  for length;  $m$  for mass;  $R$  for resistor;  $P$  for compressive load;  $T$  for tensile load; etc.

A key step in the dashed box consists in assigning numerical values to the foregoing features, which is a task calling for discipline-specific knowledge—fluid mechanics; structural engineering; machine design; multibody dynamics; etc.—either theoretical or empirical. Out of this knowledge comes a *mathematical model* relating all the foregoing features. The importance of a mathematical model cannot be overstressed, as it is the foundation on which optimization methods rely. A detailed discussion, with a few examples, on mathematical modelling, and modelling at large, in fact, is given by Papalambros and Wilde (2000).

Once a mathematical model of the object under design is available, numerical

---

<sup>1</sup>Also known as *Conceptual Design*.

values can be assigned to its features using most often good *engineering judgment*, which comes only from experience and common sense. As a means of verifying decisions on dimensioning, the designer can resort to well-developed design methods leading systematically to the best possible values of those features, while satisfying the client’s needs and budgetary constraints. Under “budgetary” we understand not only financial resources, but also time, for deadlines must be respected. How to assign values systematically to the foregoing features is the role of *optimization*.

Optimization is thus a process by which the decision-maker, in our case the designer, arrives at *optimum* values of the features defining the design solution proposed. A set of optimum values has been achieved when a cost has been minimized or a profit has been maximized, while respecting the mathematical model, i.e., the *functional relations* among all quantities at stake and the budgetary constraints expressed in the form of equality or inequality relations. Once all design features have been determined, and validated by means of simulation using discipline-specific tools—computational fluid dynamics code; finite element code; electromagnetic design code; code implementing Monte Carlo methods; etc.—an embodiment of the design solution can be produced, probably as a virtual prototype. Such embodiment is the output of the dashed box in Fig. 1.2.

The embodiment is then further developed to the last detail, in order to allow for third parties, e.g., a machine-tool shop, to produce all the parts leading to the physical prototype upon assembly. Finally, the physical prototype is subjected to validation tests before it is certified and ready to go either into mass production or to the client as an end user, thereby completing the design process. Current trends dictate that the design contemplate not only delivery to the end user, but also disposability of the designed object upon completion of its life cycle.

Engineering design problems have increasingly become model-based, in that their complexity calls for a mathematical model describing not only the physical laws governing the operation of the product under design, but also corporative policies, government regulations (codes and standards) and legislation. The work of the design engineer starts by producing a mathematical model that best reflects the relations among all quantities involved, some which are constant values either provided by the operating conditions—e.g., gravity acceleration  $g$ ; number  $\pi$ ; freezing temperature at sea level, i.e., 273.2° K—or to be determined by the designer; some are variable during the operation of the designed object—engine temperature, current in a low-pass filter, speed of the electrode tip in arc-welding, etc. In model-based design, variables are usually grouped into *state variables*, *output variables* and *control*

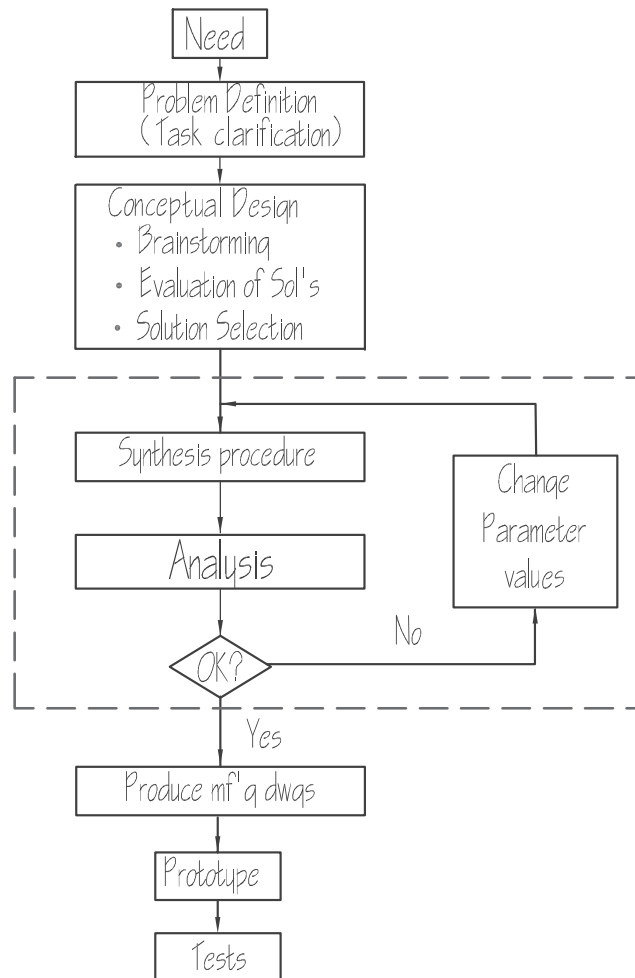


Figure 1.2: The role of optimization within the design process

*variables*. State variables are those governed by the dynamics of the system under modelling, e.g., altitude and angle of attack in aircraft-piloting; output variables are those that are available to the operator via instruments, e.g., altitude, as this is available to the pilot as an altimeter display; control variables are those that the operator can set at will either manually or automatically by means of a feedback control system, e.g., rate of climbing. Constant variables whose values are to be determined optimally by the designer are usually called *design variables*, as these attain different values within the optimization process, which more often than not is *iterative*.

## 1.2 The Structure of Optimum Design Problems

In optimum design problems, the conditions under which the designed object will operate are given by the client either explicitly or implicitly. The designer then assumes that these bear *ideal values* that are *representative*, and do not change. The fact of the matter is that the *operation conditions* entail values that are *random* and hence, are known, if at all, only through their statistics, i.e., their probability distributions. In this case, then, the designer can choose either the mean values of the variables representing the operation conditions, or their extreme values, if safety is the main design concern, as the *nominal values* of these variables.

The foregoing approach is classical, and will be followed here. An alternative approach, due to Taguchi (1988), consists in admitting that the operation conditions are not fixed, but varying in a random manner, beyond the control of the designer, the purpose of the design task then being to select the design variables in such a way that, under arbitrary variations of the operation conditions within a certain range, the performance of the designed object exhibits “small” variations. This approach is known as *robust design*. For an introductory course, we will not dwell on this approach.

In an optimum design problem the designer formulates an *objective function*, to be either minimized, when this function represents a cost, or maximized, when the same represents a profit. As a matter of fact, profit-maximization can be readily turned into cost-minimization if the profit is redefined as a cost by, for example, reversing its sign or taking its reciprocal. Moreover, a large class of optimum design problems lends itself to a *least-square* formulation, which inherently aims at *minimizing* a sum of squares. For these reasons, and unless otherwise stated, we will aim in this course at the *minimization* of an objective function  $f(\mathbf{x})$ . In defining objective functions, it will prove convenient to use dimensionless quantities. As a matter of fact, mathematical models based on dimensionless quantities have the merit that their results can be more generally applicable, as every time the operation conditions or design specifications change numerically, a simple scaling will yield the optimum values applicable to the specific conditions or specifications.

The foregoing discussion contemplates one single objective function. However, some applications problems lead to several objectives, which, moreover, are conflicting. Such is the realm of aircraft design, which calls for lightweight for fuel economy. However, in aircraft design reliability plays a major role, which calls, in turn, for stronger, and hence, heavier structural elements. Such situations are best



formulated as *multi-objective optimization problems*. We will not dwell on problems of this kind in this course, but will provide an example, if time permits.

We illustrate the structure of an optimum design problem with the aid of a Case Study in the subsection below:

### **1.2.1 Case Study: The Optimum Design of the Links of a Robotic Wrist**

See:

Bidaud, F., Teng, C.-P. and Angeles, J., 2001, “Structural Optimization of a Spherical Parallel Manipulator Using a Two-Level Approach,” *Proc. ASME 2001 Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Pittsburgh, Pennsylvania, September 9–12.



# Chapter 2

## Single-Variable Optimization

### 2.1 Methods of Single-Variable Optimization

While real-life design problems involve multiple variables, some techniques developed to find the optimum of these problems rely on a search along each of the variables at a time. Moreover, the designer in many instances is interested in the role played by one single variable, in which case the search for the optimum value can be conducted with techniques specific to this case. For this reason, it is convenient to study techniques applicable to the solution of single-variable optimization problems, which is the subject of this chapter. We start by introducing a **definition**:

*A function  $f(x)$  is unimodal in the interval  $[0, 1]$  if it attains one single extremum—a minimum or a maximum—within this interval.*

**Remarks:**

- We will deal only with function minimization in explaining the methods of interest—function maximization can be handled by paraphrasing the corresponding method accordingly.
- A unimodal function need neither be continuous nor smooth.
- Defining the interval of interest as  $[0, 1]$  is not restrictive. If this interval is  $[a, b]$ , where  $a$  and  $b$  are any real numbers, then a simple linear transformation of the variable in question can lead to the above interval.

We introduce, moreover, the **basic assumption**: Function  $f(x)$ , to be minimized, is *unimodal* in the interval  $[0, 1]$ , which means that  $f(x)$  attains exactly one minimum (or one maximum) in the given interval.

As a consequence of the above definition, we have

**Lemma 2.1.1** *Let  $f(x)$  be unimodal in  $[0, 1]$  and attain a minimum within this interval. Then, its maximum lies necessarily at the extremes of the interval, i.e., either at  $x = 0$  or at  $x = 1$ .*

The proof of this lemma is left to the reader as an exercise. Moreover, note that:

- The objective function can be evaluated *only at a discrete, finite set of sample values* of its argument  $x$ ,  $\{x_i\}_1^n$ . Each function evaluation,  $f(x_i) \equiv f_i$ , is termed an *experiment*. The name is quite appropriate because in some instances it may happen that the evaluation of function  $f(x)$  can be done only by physical experiments, e.g., when this function is the steady-state temperature of an engine, that is known to change as the proportion of a mixture of fuel and air varies;
- We assume that the interval in which the minimum lies is *known*, and termed the *interval of uncertainty* (i.o.u) of the problem at hand. Upon a suitable transformation of the design variable, this interval is mapped into the *normal interval*  $[0, 1]$ , which is of unit length. The length of the interval of uncertainty when the series of experiments is initiated is thus 1, the purpose of the minimization exercise being to bring down the interval of uncertainty to an acceptable low, which is dictated mostly by the cost of each experiment;
- If the cost of each experiment is not an issue, then the function can be evaluated in a rich sample of argument values within the interval  $[0, 1]$  and plot the corresponding values; the optimum can then be located by inspection, possibly at the click of a mouse. This is termed an *exhaustive search*;
- If the foregoing cost is high, then proceed *iteratively*: At each iteration, the interval of uncertainty is cut by a certain factor using a suitable *strategy*, i.e., a search method;
- Any strategy exploits the unimodality assumption. We can cite four strategies that are the most commonly employed:
  - Dichotomous search
  - Interval-halving
  - Fibonacci numbers
  - Golden search.

## 2.2 Dichotomous Search

The qualifier “dichotomous” derives from Greek, meaning *to cut into two parts*. The strategy to follow thus consists in splitting the interval into two subintervals, not necessarily of the same length, with one not containing the minimum, and is hence, rejected; the other subinterval then is bound to contain the minimum sought.

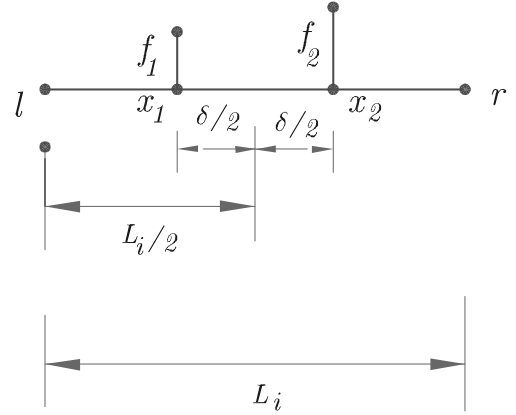


Figure 2.1: Dichotomous search

The search strategy of this method is described below:

- Assume that, at iteration  $i$ , the current search interval is  $\mathcal{I}_i = [l, r]$ , of length  $L_i = r - l < 1$  (Fig.2.1);
- locate **two** points of abscissae  $x_1$  and  $x_2$  around the centre of the interval: For a “small”  $\delta > 0$ , prescribed by the user,

$$x_1 \equiv \frac{r + l - \delta}{2}, \quad x_2 \equiv \frac{r + l + \delta}{2}; \quad f_i \equiv f(x_i); \quad f_1 \neq f_2$$

Note: If  $f_1 = f_2$ , then we have two cases: (i)  $f(x)$  is symmetric about  $x = (r + l)/2$ , in which case the minimum lies at  $x = 1/2$ , and we are done; and (ii)  $f(x)$  is not symmetric about  $x = 1/2$ , in which case we just change  $\delta$ , and eliminate this special case.

- if  $f_2 > f_1$ , then eliminate the interval segment to the right of  $x_2$ , the new search interval being  $[l, x_2]$ . If, on the contrary,  $f_1 > f_2$ , then eliminate the interval segment to the left of  $x_1$ , the new search interval being  $[x_1, r]$ .
- After the last pair of experiments, one is left with two values,  $x_1$ , and  $x_2$ . The obvious choice as the “optimum” is the mean value of these,  $x_{opt} = (x_1 + x_2)/2$ .

Notice that the new search interval  $\mathcal{I}_{i+1}$  is of length  $L_{i+2} = (L_i + \delta)/2$ , i.e., slightly over one half the length of the previous one.

Now we determine the length  $L_{2k}$  of interval  $\mathcal{I}_{2k}$  after  $2k$  experiments—this number is always even! To this end, we notice how the length of the i.o.u. evolves

as the search progresses:

$$\begin{aligned}
L_2 &= \frac{1}{2} + \frac{\delta}{2} \\
L_4 &= \frac{L_2}{2} + \frac{\delta}{2} = \frac{1}{4} + \frac{\delta}{4} + \frac{\delta}{2} \\
&= \frac{1}{4} + \frac{3\delta}{4} \\
L_6 &= \frac{L_4}{2} + \frac{\delta}{2} = \frac{1}{8} + \frac{3\delta}{8} + \frac{\delta}{2} \\
&= \frac{1}{8} + \frac{7\delta}{8} \\
&\vdots
\end{aligned}$$

The length of the interval after  $2n$  experiments is thus

$$L_{2n} = \frac{1}{2^n} + \left(1 - \frac{1}{2^n}\right) \delta \quad (2.1)$$

Usually,  $L_{2n}$  is prescribed, but  $n$  is not. Computing  $n$  from  $L_{2n}$  is, nevertheless, straightforward, as described below: Solving eq.(2.1) for  $2^n$  yields

$$2^n = \frac{1 - \delta}{L_{2n} - \delta} \quad (2.2)$$

and hence, upon taking  $\ln(-)$  of both sides and solving for  $n$

$$n = \left\lceil \frac{\ln[(1 - \delta)/(L_{2n} - \delta)]}{\ln(2)} \right\rceil \quad (2.3)$$

where  $\lceil(\cdot)\rceil$  is the *ceiling function*, defined as the smallest integer that is greater than the real argument  $(\cdot)$ . Note:  $1 > \delta$ ,  $2^n > 0 \Rightarrow L_{2n} > \delta$ .

**Example 2.2.1 (Finding the Maximum Dexterity Posture of a Two-Phalanx Robotic Finger)** *We study here the optimum dimensioning of the two-phalanx robotic finger, as depicted in Fig. 2.2. The geometry of the finger is thus completely specified by the angles that the phalanx axes make with given lines. In the figure, only the angle  $\theta$  made by the distal phalanx with the first one is indicated because only this angle is relevant to the problem under study.*

*In optimizing the performance of robotic hands, one is interested in maximizing their dexterity, a performance index that comes into play as explained below.*

*In robot control, a velocity  $\mathbf{v}$  of the operation point of the end link is to be produced by a suitable set of joint rates, grouped in vector  $\dot{\mathbf{q}}$ , the relation between the*

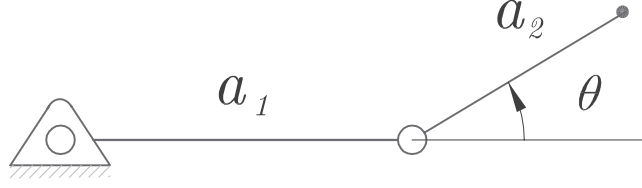


Figure 2.2: Two-phalanx robotic finger

two vectors being linear:  $\mathbf{J}\dot{\mathbf{q}} = \mathbf{v}$ . Hence, the Jacobian matrix  $\mathbf{J}(\mathbf{q})$  must be inverted in order to compute the joint-rate vector, for a given posture of the manipulator, as specified by vector  $\mathbf{q}$ , and a given desired velocity  $\mathbf{v}$ . Dexterity measures, essentially, how invertible the Jacobian matrix is. If we assume that  $a_1 = l$  and  $a_2 = l\sqrt{2}/2$ , which bear the optimum proportion found by Salisbury and Craig (1982), then dexterity can be quantified by means of the product  $\mathbf{J}^T\mathbf{J} = \ell^2\mathbf{K}$ , where  $\mathbf{K}$  is given by

$$\mathbf{K} \equiv \begin{bmatrix} 3 + 2\sqrt{2}\cos\theta & 1 + \sqrt{2}\cos\theta \\ 1 + \sqrt{2}\cos\theta & 1 \end{bmatrix}$$

It should be apparent that, when  $\theta = 0$  or  $\pi$ , matrix  $\mathbf{K}$ , that we shall term here the dexterity matrix, is singular, and hence, not invertible, as is  $\mathbf{J}$ . Between these two values, 0 and  $\pi$ , there is one specific value  $\theta_o$  optimum, at which the dexterity matrix is maximally invertible. To find  $\theta_o$ , we start by defining the dexterity as the ratio of the smallest ( $\lambda_m$ ) to the largest ( $\lambda_M$ ) eigenvalues of  $\mathbf{K}$ . In this regard, note that  $\mathbf{K}$  is symmetric, and hence, its eigenvalues are real. Moreover, one can readily verify that  $\mathbf{K}$  is positive-definite, and becomes singular only for the two values of  $\theta$  given above. We thus have the dexterity function  $D(\theta)$  defined below:

$$D(\theta) = \frac{\lambda_m}{\lambda_M} \geq 0, \quad 0 \leq D(\theta) \leq 1$$

Now, maximizing  $D(\theta)$  is equivalent to minimizing  $f(\theta) \equiv 1/D(\theta)$ , which will be defined as the loss of dexterity, and becomes, then, the objective function of the problem at hand. Given the form of the objective function, then, each experiment involves four steps:

1. For a given value of  $\theta$ , compute the two eigenvalues of  $\mathbf{K}$ , a task that can be readily implemented using an eigenvalue routine, a quadratic-equation solver, or even the Mohr circle Angeles (2011).
2. Order the two eigenvalues in ascending order:  $\lambda_m, \lambda_M$ .

3. Compute  $D(\theta)$  as

$$D(\theta) = \frac{\lambda_m}{\lambda_M} \geq 0$$

4. Compute  $f(\theta)$  as

$$f(\theta) = \frac{1}{D(\theta)} \quad 0 \leq f(\theta) < \infty$$

*An expert roboticist claims that the dexterity is maximum—the finger is at the peak of its positioning accuracy—when  $\theta$  lies “somewhere between  $90^\circ$  and  $150^\circ$ .” Find an estimate of  $\theta_{opt}$  within an interval of uncertainty of 5% of the given interval length of  $60^\circ$ .*

*Solution:* We implemented the dichotomous search in the Maple worksheet described below, which is posted in the course Web page.

```
> restart: with(linalg):
```

We start by producing a procedure  $K$  that will allow us to evaluate matrix  $\mathbf{K}$  for a given value  $\theta$ :

```
> K:=proc(theta)
matrix([[3+2*sqrt(2)*cos(th),1+sqrt(2)*cos(th)],
[1+sqrt(2)*cos(th), 1]])
end;
```

```
K := proc(theta)
matrix([[3 + 2 * sqrt(2) * cos(th), 1 + sqrt(2) * cos(th)], [1 + sqrt(2) * cos(th), 1]])
end proc
```

```
> argu:= 3*Pi/4; K(argu);
# Testing procedure, which should yield the 2 by 2 identity matrix for
this value of argument theta:
```

$$argu := \frac{3}{4} \pi$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Apparently, procedure is OK.



We introduce now a transformation that maps  $\theta$ , given in degrees, into the normal interval  $[0, 1]$ . Let the associated “normal” variable be  $x$ , to be produced by a second procedure  $x$ . By the same token, we need a third procedure  $\theta$  to return the angle in radians, for a given value of  $x$ . Thus,

```

> x:=proc(th)(th-Pi/2)/(5*Pi/6-Pi/2)
end;

      x := proc(th) 3 * (th - 1/2 *  $\pi$ ) /  $\pi$  end proc
> thet:=proc(x)(Pi/3)*x+Pi/2
end;

      thet := proc(x) 1/3 *  $\pi$  * x + 1/2 *  $\pi$  end proc

```

Now we determine the number  $2k$  of experiments needed to attain the prescribed length of the i.o.u. We recall that the length  $L_{2k}$  of this interval is given by

$$L_{2k} = \frac{1}{2^k} + \delta \left( 1 - \frac{1}{2^k} \right) \quad (2.4)$$

```

> L[2*k] := (1/2^k) + delta * (1 - (1/2^k));

       $L_{2k} := \frac{1}{2^k} + \delta \left( 1 - \frac{1}{2^k} \right)$ 

```

Let  $2^k = N$ . Then,

```

> N:=solve(L_N= (1/N) + delta*(1 - 1/N), N);

       $N := \frac{-1 + \delta}{-L_N + \delta}$ 

```

We want the length of the final i.o.u. to be 5% of original length, for a value of  $\delta$  of 0.01:

```

> delta:=0.01; L_N:= 0.05;

       $\delta := 0.01$ 
       $L_N := 0.05$ 
> N:=subs((delta=0.01, L_N=0.05), N);

       $N := 24.75000000$ 

```

Hence,

```
> k:=ceil(solve(2^k=N,k));
k := 5
```

where the Maple `ceil(·)` command has been used.

We thus need  $2k = 10$  experiments. Hence, the two points  $x_1$  and  $x_2$  within  $I_0$  are defined as

```
> x[1]:= (1-delta)/2; x[2]:= (1+delta)/2;
x1 := .4950000000
x2 := .5050000000
> theta[1]:=evalf(thet(x[1]));
theta[2]:=evalf(thet(x[2]));
K1:=evalf(K(theta[1]));
K2:=evalf(K(theta[2]));
theta1 := 2.089159115
theta2 := 2.099631090
K1 :=  $\begin{bmatrix} 1.598631263 & .2993156313 \\ .2993156313 & 1. \end{bmatrix}$ 
K2 :=  $\begin{bmatrix} 1.572980385 & .2864901922 \\ .2864901922 & 1. \end{bmatrix}$ 
> lambda:=eigenvals(K1); mu:=eigenvals(K2);
lambda := .8760194061, 1.722611857
mu := .8813318770, 1.691648508
> f[1]:=lambda[2]/lambda[1]; f[2]:=mu[2]/mu[1];
f1 := 1.966408330
f2 := 1.919422810
```

$f_1 > f_2 \Rightarrow$  delete subinterval  $[0, x_1]$ . Let  $l$  and  $r$  denote, respectively, the abscissae of the left and right ends of the new subinterval:

```
> l:=x[1]; r:=1; L:=r-l;
l := .4950000000
r := 1
L := .5050000000
```

$L$  is length of i.o.u. at the end of the first two experiments. Carry on:

```

> x[1]:=(1+r-delta)/2; x[2]:=(1+r+delta)/2;
       $x_1 := 0.7425000000$ 
       $x_2 := 0.7525000000$ 
> theta[1]:=evalf(thet(x1));
theta[2]:=evalf(thet(x2));
K1:=evalf(K(theta[1])); K2:=evalf(K(theta[2]));
       $\theta_1 := 2.348340509$ 
       $\theta_2 := 2.358812484$ 
       $K1 := \begin{bmatrix} 1.015769486 & .0078847430 \\ .0078847430 & 1. \end{bmatrix}$ 
       $K2 := \begin{bmatrix} .994770873 & -.002614564 \\ -.002614564 & 1. \end{bmatrix}$ 
> lambda:=eigenvals(K1); mu:=eigenvals(K2);
       $\lambda := .9967340325, 1.019035453$ 
       $\mu := .9936878850, 1.001082988$ 
> f[1]:=lambda[2]/lambda[1]; f[2]:=mu[2]/mu[1];
       $f_1 := 1.022374495$ 
       $f_2 := 1.007442078$ 

```

$f_1 > f_2 \Rightarrow$  delete subinterval  $[l, x_1]$ . Redefine  $l$  and  $r$ :

```

> l:=x[1]; L:=r-l; #r remains unchanged
       $l := .7425000000$ 
       $L := .2575000000$ 

```

$L$  is length of i.o.u. at the end of 3rd & 4th experiments. Carry on:

```

> x[1]:=(1+r-delta)/2; x[2]:=(1+r+delta)/2;
       $x_1 := .8662500000$ 
       $x_2 := .8762500000$ 
> theta[1]:=evalf(thet(x[1]));
theta[2]:=evalf(thet(x[2]));
K1:=evalf(K(theta[1])); K2:=evalf(K(theta[2]));

```

```

 $\theta_1 := 2.477931206$ 
 $\theta_2 := 2.488403181$ 
 $K1 := \begin{bmatrix} .771929030 & -.114035485 \\ -.114035485 & 1. \end{bmatrix}$ 
 $K2 := \begin{bmatrix} .753805937 & -.123097032 \\ -.123097032 & 1. \end{bmatrix}$ 
> lambda:=eigenvals(K1); mu:=eigenvals(K2);
 $\lambda := .7246939855, 1.047235044$ 
 $\mu := .7028174767, 1.050988460$ 
> f[1]:=lambda[2]/lambda[1]; f[2]:=mu[2]/mu[1];
 $f_1 := 1.445072079$ 
 $f_2 := 1.495393178$ 

```

The ensuing computations follow the same pattern. In the interest of brevity, we record here only the last two experiments:

At the end of 7th and 8th experiments, we have

```

 $r := .8143750000$ 
 $L := .0718750000$ 
> x[1]:=(1+r-delta)/2; x[2]:=(1+r+delta)/2;
 $x_1 := 0.7734375000$ 
 $x_2 := 0.7834375000$ 
> theta[1]:=evalf(thet(x[1]));
theta[2]:=evalf(thet(x[2]));
K1:=evalf(K(theta[1])); K2:=evalf(K(theta[2]));
 $\theta_1 := 2.380738183$ 
 $\theta_2 := 2.391210159$ 
 $K1 := \begin{bmatrix} .951519906 & -.024240047 \\ -.024240047 & 1. \end{bmatrix}$ 
 $K2 := \begin{bmatrix} .931208945 & -.034395528 \\ -.034395528 & 1. \end{bmatrix}$ 
> lambda:=eigenvals(K1); mu:=eigenvals(K2);
 $\lambda := .9414793498, 1.010040556$ 
 $\mu := .9169618507, 1.014247094$ 
> f[1]:=lambda[2]/lambda[1]; f[2]:=mu[2]/mu[1];

```

$$f_1 := 1.072822847$$

$$f_2 := 1.106095192$$

$f_2 > f_1 \Rightarrow$  delete subinterval  $[x_2, r]$ . Redefine  $l$  and  $r$ :

```
> r:=x[2]; L:=r-l; #l remains unchanged
```

$$r := .7834375000$$

$$L := .0409375000$$

$L$  is length of i.o.u. at the end of 9th & 10th experiments. Since  $L$  is smaller than 0.05, we're done. The best estimate of  $\theta_{opt}$  is obviously the mid point of current i.o.u., i.e.,

```
> x[opt]:=(l+r)/2; theta:=evalf(thet(x[opt]));
```

$$x_{opt} := .7629687500$$

$$\theta := 2.369775334$$

```
> K[opt]:=evalf(K(theta));
```

$$K_{opt} := \begin{bmatrix} .973023585 & -.013488208 \\ -.013488208 & 1. \end{bmatrix}$$

```
> lambda:=eigenvals(K[opt]);
```

$$\lambda := .9674365862, 1.005586999$$

```
> f[o]:=lambda[2]/lambda[1];
```

$$f_o := 1.039434536$$

```
> theta[opt]:=evalf(th*180/Pi);
```

$$\theta_{opt} := 135.7781250$$

Note that optimum value of theta is  $135^\circ$ , i.e.,  $3\pi/4$  rad, which yields a value of  $f=1.0$ .

## 2.3 Interval-Halving

To be included.

## 2.4 Fibonacci Numbers

Fibonacci numbers are named after the Italian mathematician Leonardo Pisano (1175), son of Guglielmo Bonaccio, and hence, referred to as *Filius Bonacci* in Latin, or *Fibonacci* for brevity (Livio, 2002). These numbers form a *sequence*, defined *recursively* as

$$F_0 = F_1 = 1 \quad (2.5a)$$

$$F_k = F_{k-2} + F_{k-1} \quad (2.5b)$$

**Remark:** The sequence is *monotonically increasing*, for all numbers are positive integers and the current one equals the sum of the two previous ones. From eq.(2.5b),

$$F_k - F_{k-1} = F_{k-2} \quad (2.6)$$

Moreover, by virtue of the above remark,

$$F_{k-1} > F_{k-2} \quad (2.7)$$

Upon addition of eq. (2.6) to inequality (2.7) sidewise, we obtain

$$F_k - F_{k-1} + F_{k-1} > 2F_{k-2}$$

i.e.,

$$\frac{F_{k-2}}{F_k} < \frac{1}{2} \quad (2.8)$$

Furthermore, from eq.(2.5b),

$$\frac{F_{k-1}}{F_k} = 1 - \frac{F_{k-2}}{F_k} \quad (2.9)$$

Now we outline the strategy to follow in this method:

- Let  $\mathcal{I}_0 \equiv [0, 1]$  be the initial interval of uncertainty, of length 1, where the minimum is *known* to lie.
- **Prescribe** the number  $n$  of experiments to be conducted
- **Define** a length  $L_2^*$  as<sup>1</sup>

$$L_2^* \equiv \frac{F_{n-2}}{F_n} L_0 \equiv \frac{F_{n-2}}{F_n} \left( < \frac{1}{2} \right) \quad (2.10)$$

where  $L_0$  is the length of the original interval, which has been defined as unity.

---

<sup>1</sup>Length  $L_1^*$  is skipped because we want to make the subscript of  $L^*$  match that of the corresponding Fibonacci number; since  $F_1 = F_0$ , the first two Fibonacci numbers are undistinguishable, and we can arbitrary set  $L_1^* = 1$ .  $L_1 = L_0$  as well.

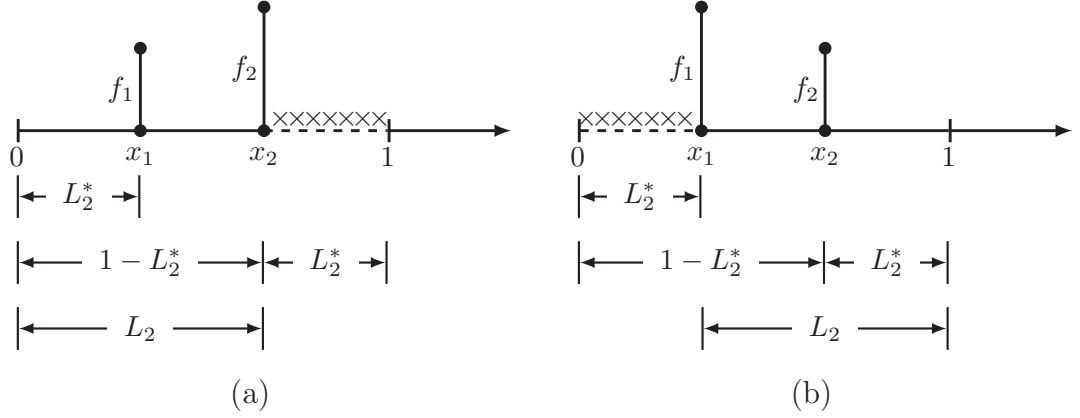


Figure 2.3: Shrinking of the i.o.u. when: (a) right subinterval is eliminated; (b) left subinterval is eliminated

- **Let**  $P_1(x_1)$  and  $P_2(x_2)$  be two points *equidistant* from the left and the right ends of  $\mathcal{I}_0$ , respectively, by a distance  $L_2^*$ , i.e., with abscissae

$$x_1 = 0 + L_2^* = \frac{F_{n-2}}{F_n} < \frac{1}{2} \quad (2.11a)$$

$$x_2 = 1 - L_2^* = 1 - \frac{F_{n-2}}{F_n} > \frac{1}{2} > x_1 \quad (2.11b)$$

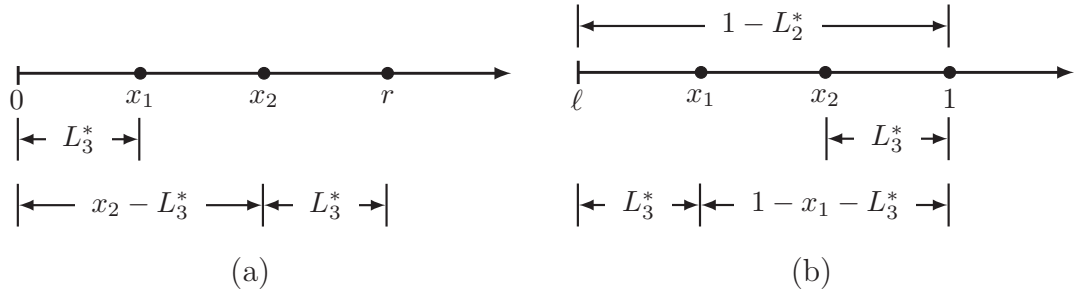


Figure 2.4: Subdivision of the new interval into three subintervals when: (a) right subinterval is eliminated; (b) left subinterval is eliminated

Note that

$$x_2 = 1 - \frac{F_{n-2}}{F_n} = \frac{F_n - F_{n-2}}{F_n} = \frac{F_{n-1}}{F_n} \quad (2.11c)$$

Use the unimodality assumption to eliminate the subinterval, left or right, where the minimum **cannot** lie. Whether the subinterval eliminated is the right or the left, the length  $L_2$  of the new, shorter interval  $\mathcal{I}_2$ —again, such as we do not define  $L_1^*$ , we neither define  $\mathcal{I}_1$ —is given by

$$L_2 = 1 - L_2^* = 1 - \frac{F_{n-2}}{F_n} = \frac{F_n - F_{n-2}}{F_n} = \frac{F_{n-1}}{F_n} < 1 \quad (2.12)$$

as depicted in Fig. 2.3.

Once the first iteration is completed, and as illustrated in Fig. 2.3,

- Let  $l$  and  $r > l$  denote, respectively, the abscissae of the left and the right ends of the current, smaller interval of uncertainty  $\mathcal{I}_2$ . The abscissa of one of the ends of  $\mathcal{I}_2$ , left or right, is either  $x_1$  or  $x_2$ ;
- if  $x_1$  is the abscissa of one of the ends of  $\mathcal{I}_2$ , case of Fig. 2.3(b), then  $x_2 \in \mathcal{I}_2$ ; else, as in Fig. 2.3(a),  $x_1 \in \mathcal{I}_2$ . Let  $x_I$  be the abscissa of point  $P_I$ , the interior point of  $\mathcal{I}_2$ , either  $P_1$  or  $P_2$ ;
- note that  $P_I$  lies a distance  $L_3^*$  from one of the ends of  $\mathcal{I}_2$ . Now, define  $x_3 \in \mathcal{I}_2$ , so that its associated point  $P_3$  also lies a distance  $L_3^*$  from the other end, as depicted in Fig. 2.4;
- the process is continued until the interval  $\mathcal{I}_n$ , of length  $L_n$ , is obtained.  $L_n$  is the length of the final interval of uncertainty.

The abscissa  $x_j$  computed at the  $j$ th experiment is determined by length  $L_j^*$ , so that its associated point  $P_j$ , as well as the interior point  $P_I$  of interval  $\mathcal{I}_{j-1}$  are a distance  $L_j^*$  from the ends of  $\mathcal{I}_j$ , with  $L_j^*$  given by<sup>2</sup>

$$L_j^* = \frac{F_{n-j}}{F_{n-(j-2)}} L_{j-1} \quad (2.13a)$$

while the length  $L_j$  of the  $j$ th interval of uncertainty is<sup>3</sup>

$$L_j = \frac{F_{n-(j-1)}}{F_n} \quad (2.13b)$$

Hence, for  $j = n$ ,

$$L_n = \frac{F_1}{F_n} \quad (2.14)$$

which allows us to find  $n$  for a prescribed length  $L_n$ . Notice that, by virtue of relations (2.8) and (2.10),

$$\text{For } n \geq 2, \quad L_n^* < \frac{1}{2} \quad (2.15)$$

---

<sup>2</sup>see eq.(2.10), which is identical to eq.(2.13a) evaluated at  $j = 2$ ; see also footnote 1.

<sup>3</sup>see eq.(2.12), which is identical to eq.(2.13b) evaluated at  $j = 2$ .



## The Location of the Final Experiment

Let  $\mathcal{I}_{n-1} = [l, r]$ , of length  $L_{n-1}$ , be the one-before-the-last interval of uncertainty. According with eqs.(2.11a & b), the abscissae of the last two experiments,  $x_{n-1}$  and  $x_n$ , are given as

$$x_{n-1} = l + L_n^* \quad (2.16a)$$

$$x_n = r - L_n^* = l + L_{n-1} - L_n^* \quad (2.16b)$$

where  $L_n^*$  is given by eq.(2.13a), with  $j = n$ :

$$L_n^* = \frac{F_0}{F_2} L_{n-1} = \frac{1}{2} L_{n-1} \quad (2.17)$$

Upon substitution of  $L_n^*$ , as given by eq.(2.17), into eq.(2.16b), it is apparent that

$$x_n = l + \frac{1}{2} L_{n-1} = x_{n-1}$$

$x_{n-1}$  and  $x_n$  thus coinciding, and hence, the last experiment fails to produce two distinct points in  $\mathcal{I}_{n-1}$ . To cope with this outcome, we have to define points  $P_{n-1}$  and  $P_n$ , of abscissae  $x_1$  and  $x_2$ , in an alternative manner. For example, we can define them as in the strategy employed by the dichotomous search, with a  $\delta$  small enough with respect to  $L_n^*$ .

Fibonacci numbers are tabulated in many manuals, with short tables available in textbooks (Rao, 1996). Also note that scientific software is provided with Fibonacci numbers. For example, Maple includes the command

```
with(combinat, Fibonacci):
```

that allows the user to invoke the Fibonacci number  $F(i)$  by typing

```
fibonacci(i)
```

However, note that **not all Fibonacci sequences are identical**. For example, the first two Fibonacci numbers in Maple are defined as

$$f(0) = 0 \text{ and } f(1) = 1$$

**Example 2.4.1 (Finding the Maximum Dexterity Posture of a Two-Phalanx Robotic Finger)** *For the problem of Example 2.2.1, apply the Fibonacci search strategy using exactly 10 experiments.*

Solution: We implement the Fibonacci search in a Maple worksheet:

```
> restart:
> with(linalg): with(combinat, fibonacci):
```

However, because of the way Maple defines the Fibonacci sequence, we must use  $F(n+1)$  when we would normally use  $F(n)$ . Moreover, we shall also use  $K$ ,  $\theta$  and  $x$  exactly as described in Subsection 2.2.

We want to have

$$L_2^* = \frac{F_{10-2}}{F_{10}}, \quad (2.18)$$

but must shift the subscript by 1:

```
> Lstar[2]:=evalf(fibonacci(9)/fibonacci(11));
                                Lstar2 := .3820224719
> l:=0; r:=1; L[0]:=r - l;
> # Abscissae of extremes of left- & right-hand
> sides of the initial (normal) interval, and length of this interval
                                l := 0
                                r := 1
                                L0 := 1
> x[1]:= l + L^star[2]; x[2]:= r - Lstar[2];
                                x1 := .3820224719
                                x2 := .6179775281
> theta[1]:=evalf(thet(x1));
> theta[2]:=evalf(thet(x2));
> K1:=K(theta[1]); K2:=K(theta[2]);
                                θ1 := 1.970849324
                                θ2 := 2.217940881
                                K1 :=  $\begin{bmatrix} 3 - .7789343108 \sqrt{2} & 1 - .3894671554 \sqrt{2} \\ 1 - .3894671554 \sqrt{2} & 1 \end{bmatrix}$ 
                                K2 :=  $\begin{bmatrix} 3 - 1.205821535 \sqrt{2} & 1 - .6029107675 \sqrt{2} \\ 1 - .6029107675 \sqrt{2} & 1 \end{bmatrix}$ 
> lambda:=<eigenvals(K1)>;
> mu:=<eigenvals(K2)>;
```

$$\lambda := \begin{bmatrix} .8139310151 \\ 2.084489519 \end{bmatrix}$$

$$\mu := \begin{bmatrix} .9389633883 \\ 1.355747444 \end{bmatrix}$$

```

> f[1]:=lambda[2]/lambda[1];
> f[2]:=mu[2]/mu[1];

```

$$f_1 := 2.561014976$$

$$f_2 := 1.443876791$$

$f_1 > f_2 \Rightarrow$  drop left end:

```

> l:=x[1]; L[2]:=L[0] - Lstar[2];
> Lstar[3]:=(fibonacci(8)/fibonacci(10))*L[2];

```

$$l := .3820224719$$

$$L_2 := .6179775281$$

$$Lstar_3 := .2359550562$$

```

> x[1]:= l + Lstar[3]; x[2]:= r - Lstar[3];

```

$$x_1 := .6179775281$$

$$x_2 := .7640449438$$

```

> theta[1]:=evalf(thet(x[1]));
> K1:=K(theta[1]);
> theta[2]:=evalf(thet(x[2])); K2:=K(theta[2]);

```

$$\theta_1 := 2.217940881$$

$$K1 := \begin{bmatrix} 3 - 1.205821535 \sqrt{2} & 1 - .6029107675 \sqrt{2} \\ 1 - .6029107675 \sqrt{2} & 1 \end{bmatrix}$$

$$\theta_2 := 2.370902321$$

$$K2 := \begin{bmatrix} 3 - 1.434859867 \sqrt{2} & 1 - .7174299337 \sqrt{2} \\ 1 - .7174299337 \sqrt{2} & 1 \end{bmatrix}$$

```

> lambda:=<eigenvals(K1)>;
> mu:=<eigenvals(K2)>;

```

$$\lambda := \begin{bmatrix} .9389633883 \\ 1.355747444 \end{bmatrix}$$

$$\mu := \begin{bmatrix} .9647545542 \\ 1.006047163 \end{bmatrix}$$

```

> f[1]:=lambda[2]/lambda[1];
> f[2]:=mu[2]/mu[1];

```

$$f_1 := 1.443876791$$

$$f_2 := 1.042801155$$

$f_1 > f_2 \Rightarrow$  drop left end:

```

> l:= x[1]; L[3]:= L[2] - Lstar[3];
> Lstar[4]:= (fibonacci(7)/fibonacci(9))*L[3];
      l := .6179775281
      L3 := .3820224719
      Lstar4 := .1460674157
> x[1]:= l + Lstar[4]; x[2]:= r - Lstar[4];
      x1 := .7640449438
      x2 := .8539325843
> theta[1]:=evalf(thet(x[1])); K1:=K(theta[1]);
> theta[2]:=evalf(thet(x[2])); K2:=K(theta[2]);

      θ[1] := 2.370902321
      K1 :=  $\begin{bmatrix} 3 - 1.434859867\sqrt{2} & 1 - .7174299337\sqrt{2} \\ 1 - .7174299337\sqrt{2} & 1 \end{bmatrix}$ 
      θ2 := 2.465032438
      K2 :=  $\begin{bmatrix} 3 - 1.559462054\sqrt{2} & 1 - .7797310268\sqrt{2} \\ 1 - .7797310268\sqrt{2} & 1 \end{bmatrix}$ 
> lambda:=<eigenvals(K1)>;
> mu:=<eigenvals(K2)>;

      λ :=  $\begin{bmatrix} .9647545542 \\ 1.006047163 \end{bmatrix}$ 
      μ :=  $\begin{bmatrix} .7520453159 \\ 1.042542298 \end{bmatrix}$ 
> f[1]:=lambda[2]/lambda[1];
> f[2]:=mu[2]/mu[1];

      f1 := 1.042801155
      f2 := 1.386275901

```

In the interest of brevity, we skip the intermediate results, and display only the last two experiments:

```

> x[1]:= 1 + Lstar[8]; x[2]:= r - Lstar[8];
       $x_1 := .7303370787$ 
       $x_2 := .7415730337$ 
> theta[1]:=evalf(thet(x1)); K1:=K(theta[1]);
> theta[2]:=evalf(thet(x2)); K2:=K(theta[2]);
       $\theta_1 := 2.335603527$ 
       $K1 := \begin{bmatrix} 3 - 1.384795807\sqrt{2} & 1 - .6923979033\sqrt{2} \\ 1 - .6923979033\sqrt{2} & 1 \end{bmatrix}$ 
       $\theta_2 := 2.347369792$ 
       $K2 := \begin{bmatrix} 3 - 1.401678651\sqrt{2} & 1 - .7008393253\sqrt{2} \\ 1 - .7008393253\sqrt{2} & 1 \end{bmatrix}$ 
> lambda:=<eigenvals(K1)>;
> mu:=<eigenvals(K2)>;
       $\lambda := \begin{bmatrix} .9913837386 \\ 1.050219250 \end{bmatrix}$ 
       $\mu := \begin{bmatrix} .9963286091 \\ 1.021398433 \end{bmatrix}$ 
> f[1]:=lambda[2]/lambda[1]; f[2]:=mu[2]/mu[1];
       $f_1 := 1.059346859$ 
       $f_2 := 1.025162204$ 
> l:= x[1]; L[8]:= L[7] - Lstar[8];
> Lstar[9]:= (fibonacci(2)/fibonacci(4))*L[8];
       $l := .7303370787$ 
       $L_8 := .03370786516$ 
       $Lstar_9 := .01123595505$ 

```

Note that the length of the i.o.u. at the end of the 8th experiment is 3.4% of original length, i.e., smaller than at the end of 10 experiments with the dichotomous search!

```

> x[1]:= 1 + Lstar[9]; x[2]:= r - Lstar[9];
       $x_1 := .7415730338$ 
       $x_2 := .7528089888$ 
> theta[1]:=evalf(thet(x[1])); K1:=K(theta[1]);
> theta[2]:=evalf(thet(x[2]));
> K2:=K(theta[2]);

```

$$\theta_1 := 2.347369792$$

$$K1 := \begin{bmatrix} 3 - 1.401678651 \sqrt{2} & 1 - .7008393253 \sqrt{2} \\ 1 - .7008393253 \sqrt{2} & 1 \end{bmatrix}$$

$$\theta_2 := 2.359136057$$

$$K2 := \begin{bmatrix} 3 - 1.418367442 \sqrt{2} & 1 - .7091837208 \sqrt{2} \\ 1 - .7091837208 \sqrt{2} & 1 \end{bmatrix}$$

```

> lambda:=<eigenvals(K1)>;
> mu:=<eigenvals(K2)>;


$$\lambda := \begin{bmatrix} .9963286091 \\ 1.021398433 \end{bmatrix}$$


$$\mu := \begin{bmatrix} .9929088850 \\ 1.001216643 \end{bmatrix}$$

> f[1]:=lambda[2]/lambda[1];
> f[2]:=mu[2]/mu[1];


$$f_1 := 1.025162204$$


$$f_2 := 1.008367090$$


```

$f_1 > f_2 \Rightarrow$  drop left end:

```

> l:= x[1]; L[9]:= L[8] - Lstar[9];
> Lstar[10]:= (fibonacci(1)/fibonacci(3))*L[9];


$$l := .7415730338$$


$$L_9 := .02247191011$$


$$Lstar_{10} := .01123595506$$

> x[1]:= l + Lstar[10]; x[2]:= r - Lstar[10];


$$x_1 := .7528089889$$


$$x_2 := .7528089887$$


```

As expected,  $x_1 = x_2$ . Let us estimate the optimum by dichotomous search over the last i.o.u.: Let  $\delta = Lstar_{10}/10$

```

> delta:= Lstar[10]/10;


$$\delta := .001123595506$$

> x[1]:= (1 + r - delta)/2;
> x[2]:= (1 + r + delta)/2;

```

$$x_1 := .7522471910$$

$$x_2 := .7533707866$$

```
> theta[1]:=evalf(thet(x[1])); K1:=K(theta[1]);
> theta[2]:=evalf(thet(x[2])); K2:=K(theta[2]);
```

$$\theta_1 := 2.358547744$$

$$K1 := \begin{bmatrix} 3 - 1.417537647 \sqrt{2} & 1 - .7087688235 \sqrt{2} \\ 1 - .7087688235 \sqrt{2} & 1 \end{bmatrix}$$

$$\theta_2 := 2.359724370$$

$$K2 := \begin{bmatrix} 3 - 1.419196745 \sqrt{2} & 1 - .7095983727 \sqrt{2} \\ 1 - .7095983727 \sqrt{2} & 1 \end{bmatrix}$$

```
> lambda:=<eigenvals(K1)>;
> mu:=<eigenvals(K2)>;
```

$$\lambda := \begin{bmatrix} .9943254329 \\ 1.000973602 \end{bmatrix}$$

$$\mu := \begin{bmatrix} .9914931757 \\ 1.001459540 \end{bmatrix}$$

```
> f[1]:=lambda[2]/lambda[1];
> f[2]:=mu[2]/mu[1];
```

$$f_1 := 1.006686110$$

$$f_2 := 1.010051874$$

$f_2 > f_1 \Rightarrow$  delete  $[x_2, r]$  and take as most likely estimate of the optimum the midpoint of remaining interval  $[l, x_1]$ :

```
> x[o]:=(1+x[1])/2; th_opt:=evalf(th(x[o]));
> K_opt:= K(th_opt);
```

$$x_o := .7469101124$$

$$\theta_{opt} := 2.352958768 \longrightarrow 134.8146^\circ$$

$$Kopt := \begin{bmatrix} 3 - 1.409630165 \sqrt{2} & 1 - .7048150824 \sqrt{2} \\ 1 - .7048150824 \sqrt{2} & 1 \end{bmatrix}$$

```
> lambda:=<eigenvals(Kopt)>;
```

$$\lambda := \begin{bmatrix} .9986575537 \\ 1.007824349 \end{bmatrix}$$

> f[opt]:=lambda[2]/lambda[1];

$$f_{opt} := 1.009179118$$

## 2.5 Golden-Section Search

This method is similar to the method based on Fibonacci numbers, but its implementation is much simpler. The outcome is that its convergence is a bit slower than that of the former. *A major difference with the Fibonacci search is that the number of elimination stages is not prescribed.*

The basis of the golden-search method is the Fibonacci sequence. Indeed, the golden-search strategy is derived from the Fibonacci search under the assumption that  $n$  in the Fibonacci search is “large”; we denote a large  $n$  appropriately by  $N$ . The length of the interval of uncertainty is shrunken at every iteration by the same proportion, as opposed to the Fibonacci search.

In order to find the length  $L_k$  of the interval  $\mathcal{I}_k$  at the  $k$ th iteration of the golden search, we compute the corresponding lengths of the Fibonacci search for  $n = N \rightarrow \infty$ , namely,

$$L_2 = \lim_{N \rightarrow \infty} \frac{F_{N-1}}{F_N} \quad (2.19a)$$

$$L_3 = \lim_{N \rightarrow \infty} \frac{F_{N-2}}{F_N} = \lim_{N \rightarrow \infty} \frac{F_{N-2}}{F_{N-1}} \frac{F_{N-1}}{F_N} = \lim_{N \rightarrow \infty} \left( \frac{F_{N-1}}{F_N} \right)^2 \quad (2.19b)$$

In general,

$$L_k = \lim_{N \rightarrow \infty} \left( \frac{F_{N-1}}{F_N} \right)^{k-1} \quad (2.19c)$$

Hence, all we need to implement this method is the above limit, which is computed below:

Recall eq.(2.5b), for  $k = N$ :

$$F_N = F_{N-1} + F_{N-2} \quad (2.20a)$$

Therefore,

$$\frac{F_N}{F_{N-1}} = 1 + \frac{F_{N-2}}{F_{N-1}} \quad (2.20b)$$

Now we define the *golden section* or *golden ratio*  $\phi$  as

$$\phi \equiv \lim_{N \rightarrow \infty} \frac{F_N}{F_{N-1}} \quad (2.20c)$$



Upon taking limits, eq.(2.20b) can be rewritten as

$$\phi = 1 + \frac{1}{\phi}$$

or

$$\phi^2 - \phi - 1 = 0, \quad \phi > 0 \quad (2.20d)$$

whence, with three decimals,

$$\phi = 1.618 \quad \text{or} \quad \phi = -0.618 \quad (2.20e)$$

Obviously, we need only the positive root, and hence,  $L_k$  in eq.(2.19c) becomes

$$L_k = \left(\frac{1}{\phi}\right)^{k-1} = (0.618)^{k-1} \quad (2.21)$$

Greeks in the classical period, around the fifth century B.C.E., coined the expression *golden section* to refer to a rectangle of *divine proportions*, whose base  $b$  and height  $h$  observe the relation

$$\frac{b+h}{b} = \frac{b}{h} \quad (2.22)$$

The foregoing equation readily leads to one on the ratio  $b/h$  identical to eq.(2.20d), namely,

$$\left(\frac{b}{h}\right)^2 - \frac{b}{h} - 1 = 0$$

thereby showing that the positive solution to eq.(2.22) is, indeed,  $\phi$ , the golden section.

This relation is claimed to appear in the facade of the Parthenon, although Livio (2002) disputed brilliantly this claim and others along the same lines. Nevertheless, the golden section is irrefutably present in nature and in many artifacts<sup>4</sup>. Shown in Fig. 2.5 is a rectangle with sides obeying the divine proportion.

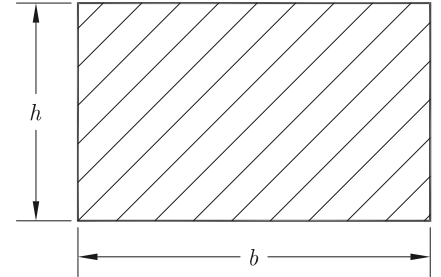


Figure 2.5: A rectangle with sides obeying the divine proportion

To implement the search, we need the quantity  $L_2^*$ , which is defined below:

$$L_2^* = \frac{F_{N-2}}{F_N} = \frac{F_{N-2}}{F_{N-1}} \frac{F_{N-1}}{F_N} = \frac{1}{\phi^2} = 0.382 \quad (2.23)$$

In general,

$$L_3^* = x_2 - x_1, \quad L_4^* = x_3 - x_2, \quad \dots, \quad L_k^* = x_{k-1} - x_{k-2}$$

---

<sup>4</sup>For example, Microsoft Word uses the golden-section ratio to proportion its margins.

**Example 2.5.1 (Finding the Maximum Dexterity Posture of a Two-Phalanx Robotic Finger)** Repeat Example 2.2.1 using the golden-section search strategy to reduce the i.o.u. to 5% of its original length.

Solution: We implement below the golden-section search strategy by means of a Maple worksheet. We shall resort to the  $K$ ,  $\theta$  and  $x$  procedures introduced earlier.

Let us calculate  $\phi$ :

```
> eq:=(x^2 - x - 1);
                                eq := x^2 - x - 1
> racines:=<solve(eq, x)>;
                                racines :=  $\begin{bmatrix} \frac{1}{2} + \frac{1}{2}\sqrt{5} \\ \frac{1}{2} - \frac{1}{2}\sqrt{5} \end{bmatrix}$ 
> phi:=evalf(racines[1]);
                                 $\phi := 1.618033989$ 
> ihp :=1.0/phi; #we'll also need the reciprocal of phi
                                ihp := .6180339887
> Lstar[2]:=ihp ^2;
                                Lstar_2 := .3819660112
```

Now let us find  $n$  from the problem specification:  $L_n = 0.05$ , which, in light of eq.(2.21), leads to

$$\frac{1}{\phi^{n-1}} = 0.05 \quad (2.24)$$

```
> eq:= (n-1)*ln(ihp ) - ln(0.05)=0;
                                eq := -.4812118251 n + 3.476944099 = 0
> n:=ceil(solve(eq, n));
                                n := 8
> l:=0; r:=1; #extremes of initial normal
> interval
                                l := 0
                                r := 1
> L[0]:= r - l; #length of initial interval
                                L_0 := 1
```

```

> x[1]:=l+Lstar[2]; x[2]:=r-Lstar[2];
       $x_1 := .3819660112$ 
       $x_2 := .6180339888$ 
> theta[1]:=evalf(thet(x[1])); K1:=K(theta[1]);
> theta[2]:=evalf(thet(x[2])); K2:=K(theta[2]);
> lambda:<eigenvals(K1)>; mu:<eigenvals(K2)>;
       $\theta_1 := 1.970790199$ 
       $K1 := \begin{bmatrix} 3 - .7788253964 \sqrt{2} & 1 - .3894126982 \sqrt{2} \\ 1 - .3894126982 \sqrt{2} & 1 \end{bmatrix}$ 
       $\theta_2 := 2.218000007$ 
       $K2 := \begin{bmatrix} 3 - 1.205915875 \sqrt{2} & 1 - .6029579377 \sqrt{2} \\ 1 - .6029579377 \sqrt{2} & 1 \end{bmatrix}$ 
       $\lambda := \begin{bmatrix} .8138991148 \\ 2.084675447 \end{bmatrix}$ 
       $\mu := \begin{bmatrix} .9389910200 \\ 1.355586395 \end{bmatrix}$ 
> f[1]:=lambda[2]/lambda[1];
> f[2]:=mu[2]/mu[1];
       $f_1 := 2.561343794$ 
       $f_2 := 1.443662789$ 

```

$f_1 > f_2 \Rightarrow$  drop the left end:

```

> L[2]:=L[0] - Lstar[2]; Lstar[3]:= x[2] - x[1];
> # You should be able to prove that ihp^3 = x_2 - x_1
       $L_2 := .6180339888$ 
       $Lstar_3 := .2360679776$ 
> x[3]:= r - Lstar[3];
> # x[2] is now a distance Lstar_3 from new left end, x[1],
> to the left of x[3]
       $x_3 := .7639320224$ 
> theta[3]:=evalf(th(x[3])); K3:=K(theta[3]);
> lambda:<eigenvals(K3)>;
       $\theta_3 := 2.370784070$ 

```

$$K_3 := \begin{bmatrix} 3 - 1.434695103 \sqrt{2} & 1 - .7173475515 \sqrt{2} \\ 1 - .7173475515 \sqrt{2} & 1 \end{bmatrix}$$

$$\lambda := \begin{bmatrix} .9650358237 \\ 1.005998904 \end{bmatrix}$$

> f[3]:=lambda[2]/lambda[1];f[2];

> #evaluate f[3] & recall f[2]

$$f_3 := 1.042447212$$

$$f_2 := 1.443662789$$

$f_2 > f_3 \Rightarrow$  drop left end:

> L[3]:=L[2] - Lstar[3];

> Lstar[4]:= x3 - x2;

$$L_3 := .3819660112$$

$$Lstar_4 := .1458980336$$

> x[4]:= r - Lstar[4];

$$x_4 := .8541019664$$

> theta[4]:=evalf(thet(x[4])); K4:=K(th4);

> lambda:=<eigenvals(K4)>;

$$\theta_4 := 2.465209815$$

$$K_4 := \begin{bmatrix} 3 - 1.559684146 \sqrt{2} & 1 - .7798420728 \sqrt{2} \\ 1 - .7798420728 \sqrt{2} & 1 \end{bmatrix}$$

$$\lambda := \begin{bmatrix} .7516661806 \\ 1.042607347 \end{bmatrix}$$

> f[4]:=lambda[2]/lambda[1];f[3];

> #evaluate f[4] & recall f[3]

$$f_4 := 1.387061669$$

$$f_3 := 1.042447212$$

Again, for brevity we introduce only the last two experiments. We have the interval  $[x_5, x_6]$  and hence,

> L[6]:=L[5] - Lstar[6]; Lstar[7]:= x6 - x3;

$$L_6 := .0901699440$$

$$Lstar_7 := .0344418544$$

```

> x[7]:= x[5] + Lstar[7];
                                 $x_7 := .7426457872$ 
> theta[7]:=evalf(thet(x[7]));
> K7:=K(theta[7]);lambda:=<eigenvals(K7)>;
                                 $\theta_7 := 2.348493177$ 
                                
$$K_7 := \begin{bmatrix} 3 - 1.403280430 \sqrt{2} & 1 - .7016402150 \sqrt{2} \\ 1 - .7016402150 \sqrt{2} & 1 \end{bmatrix}$$

                                
$$\lambda := \begin{bmatrix} .9967977596 \\ 1.018664025 \end{bmatrix}$$

> f[7]:=lambda[2]/lambda[1];f[3];
> #evaluate f[7] & recall f[3]:
                                 $f_7 := 1.021936511$ 
                                 $f_3 := 1.042447212$ 

```

$f_7 < f_3 \Rightarrow$  drop right end and perform last experiment:

```

> L[7]:=L[6] - Lstar[7]; Lstar[8]:=x3-x7;
                                 $L_7 := .0557280896$ 
                                 $Lstar_8 := .0212862352$ 
> x[8]:= x[5] + Lstar[8];
                                 $x_8 := .7294901680$ 
> theta[8]:=evalf(thet(x8)); K8:=K(theta[8]);
> lambda:=<eigenvals(K8)>;
                                 $\theta_8 := 2.334716645$ 
                                
$$K_8 := \begin{bmatrix} 3 - 1.383515463 \sqrt{2} & 1 - .6917577317 \sqrt{2} \\ 1 - .6917577317 \sqrt{2} & 1 \end{bmatrix}$$

                                
$$\lambda := \begin{bmatrix} .9910087350 \\ 1.052404934 \end{bmatrix}$$

> f[8]:=lambda[2]/lambda[1];f[7];
> #evaluate f[8] & recall f[7]:
                                 $f_8 := 1.061953237$ 
                                 $f_7 := 1.021936511$ 

```

$f_8 > f_7 \Rightarrow$  delete left end and accept midpoint, of abscissa  $x_o$ , as best estimate of optimum:

```
> x[o] := (x[8] + x[3])/2;
```

$$x_o := .7467110952$$

Evaluate  $f_o = f(x_o)$ :

```
> theta[o] := evalf(thet(x[o]));
```

```
> Ko := K(th[o]); lambda := <eigenvals(Ko)>;
```

$$\theta_o := 2.352750357$$

$$Ko := \begin{bmatrix} 3 - 1.409334444 \sqrt{2} & 1 - .7046672222 \sqrt{2} \\ 1 - .7046672222 \sqrt{2} & 1 \end{bmatrix}$$

$$\lambda := \begin{bmatrix} .9985709394 \\ 1.008329177 \end{bmatrix}$$

```
> L[8] := x[3] - x[8];
```

```
> #Length of final i.o.u.
```

$$L_8 := .0344418544$$

Notice that length of final i.o.u. is 3.4% the length of original i.o.u.

```
> f[opt] := lambda[2]/lambda[1];
```

$$f_{opt} := 1.009772203$$

```
> theta[opt] := evalf(thet[o]*180/Pi);
```

```
> # theta_optimum in degrees
```

$$\theta_{opt} := 134.8026657$$

# Chapter 3

## Numerical Equation Solving

### 3.1 Introduction

Multivariable optimization frequently calls for the solution of systems of equations that can be linear, nonlinear, or a combination thereof. If linear, then a solution can be found numerically by means of a *direct method*<sup>1</sup>, as opposed to *iterative methods*. This is a major difference, because direct methods involve a fixed number of operations; on the contrary, iterative methods involve a fixed number of operations *per iteration*, but the number of iterations the method will take until convergence is reached cannot be predicted. Furthermore, if the nonlinear equations of a system are *algebraic*, i.e., *multivariate polynomials*, then the system can be reduced, at least in principle, to a single univariate polynomial, if of a degree higher than that of any of the individual equations.

When the objective function and the constraints are multivariate polynomials in the design variables, the optimization problem leads to a system of multivariate polynomials, if with extra variables, namely, the Lagrange multipliers, to be introduced in Ch. 5. Under these conditions, it is possible to use elimination methods, as implemented in computer-algebra code, to eliminate all but one of the design variables, thereby ending up with a single univariate polynomial. Having reduced the optimization problem to polynomial-root finding is advantageous, because the roots of the polynomial provide all *stationary points*—as defined in Chs. 2 and 4, these are points where the objective function ceases to change locally—and, hence,

---

<sup>1</sup>Some linear systems of a large number of unknowns and weakly coupled equations, frequently arising in some contexts, like structural mechanics, can be solved to a great advantage using an iterative method like Gauss-Seidel's. Such methods are left out of this course.

all *local minima*. The *global minimum* can then be found by inspection.

As an alternative to univariate-polynomial reduction, the optimization problem at hand can be reduced to two (or more) *bivariate equations*, polynomial or trigonometric, whose plots appear as contours in the plane of those two variables. All solutions can then be obtained visually, by contour-intersection.

Prior to the discussion of equation-solving, we revisit the fundamental concepts of linear algebra that will be needed in the sequel. Then, we recall the basic problem of solving a system of  $n$  linear equations in  $n$  unknowns, what is called a *determined system*. The issue of roundoff-error amplification is given due attention, which takes us to the concept of *condition number*.

As a natural extension of the above problem, we undertake the problem of *linear least squares*. That is, we now study the solution of a system of  $q$  linear equations in  $n$  unknowns, when  $q > n$ , what is called an *overdetermined* system of linear equations. In this case, in general, it is not possible to find a single vector  $\mathbf{x}$  that verifies the *redundant* and, most likely *inconsistent*, set of equations. Hence, we aim at finding the *best fit* in the least-square sense, i.e., the vector  $\mathbf{x}$  that approximates the whole set of  $q$  equations with the minimum *Euclidean norm*. We derive a closed-form expression, i.e., a *formula*, for the best fit  $\mathbf{x}$  directly from the *normality conditions* (NC) of the problem at hand. This derivation readily leads to the *left Moore-Penrose generalized inverse* (LMPGI) of the coefficient matrix, which is rectangular, and for which an *inverse* proper cannot be defined. It is shown that computing the best fit *directly* from the NC is prone to *ill-conditioning*, a phenomenon characterized by a “large” roundoff-error amplification. Hence, the reader is strongly advised against computing the best fit with the said formula. Instead, *orthogonalization algorithms* are to be used. The difference between a formula, like that giving the best fit in terms of the LMPGI, and an *algorithm* is stressed here: The LMPGI is seldom needed as such, in the same way that the inverse of a nonsingular (square) matrix is seldom needed. Therefore, the *verbatim* computation of such a generalized inverse is to be avoided.

## 3.2 Background Facts and Definitions

We begin by recalling the concept of vector and matrix norms:

A *norm* is to an array of numbers, be it a column vector, a row vector, or a matrix, what the absolute value is to real numbers and the module is to complex numbers.



Vector norms can be defined in various ways:

The *Euclidean norm*: The best known. For a  $n$ -dimensional vector  $\mathbf{a}$  with components  $a_i$ , for  $i = 1, \dots, n$ :

$$\|\mathbf{a}\|_E \equiv \sqrt{a_1^2 + \dots + a_n^2} \quad (3.1)$$

Computing this norm thus requires  $n$  multiplications,  $n$  additions, and one square root. Not very “cheap” to compute!

The *Chebyshev norm*, a.k.a. the *maximum norm*, or the *infinity norm*:

$$\|\mathbf{a}\|_\infty \equiv \max_i \{|a_i|\}_1^n \quad (3.2)$$

Notice that this norm requires no floating-point operations (flops), which makes it quite economical in iterative processes that require a monitoring of the norm of a vector or a matrix quantity as a convergence criterion.

The  $p$ -norm:

$$\|\mathbf{a}\|_p \equiv \left( \sum_{j=1}^n |a_j|^p \right)^{1/p} \quad (3.3)$$

This is the most general case. For  $p = 2$ , the  $p$ -norm becomes the Euclidean norm; for  $p \rightarrow \infty$ , the  $p$ -norm becomes the Chebyshev norm.

Likewise, matrix norms can be defined in various ways:

- The *Euclidean norm*, a.k.a. the 2-norm: the square root of the largest (non-negative) eigenvalue of the positive-semidefinite product of the matrix by its transpose, regardless of the ordering of the factors. For example, for the  $n \times n$  matrix  $\mathbf{A}$ ,

$$\|\mathbf{A}\|_E \equiv \max_i \{\sqrt{\lambda_i}\} \quad (3.4)$$

where  $\{\lambda_i\}_1^n$  is the set of non-negative eigenvalues of  $\mathbf{A}\mathbf{A}^T$ , or of  $\mathbf{A}^T\mathbf{A}$  for that matter. This norm is also called the *spectral norm*.

- The *Frobenius norm*: the square root of the sum of the squares of the entries of the matrix. For the same matrix  $\mathbf{A}$ ,

$$\|\mathbf{A}\|_F \equiv \sqrt{\sum_{j=1}^n \sum_{i=1}^n a_{ij}^2} \equiv \sqrt{\text{tr}(\mathbf{A}\mathbf{A}^T)} \quad (3.5)$$

- The *Chebyshev* norm or *infinity norm*: the maximum absolute value of the entries of the matrix. For the above matrix  $\mathbf{A}$ ,

$$\|\mathbf{A}\|_{\infty} \equiv \max_{i,j} \{|a_{ij}|\} \quad (3.6)$$

- The *p-norm*:

$$\|\mathbf{A}\|_p \equiv \left( \sum_{j=1}^n \sum_{i=1}^n |a_{ij}|^p \right)^{1/p} \quad (3.7)$$

For  $p = 2$ , the  $p$ -norm becomes the Frobenius norm; for  $p \rightarrow \infty$ , the  $p$ -norm becomes, such as in the vector case, the Chebyshev norm.

**Remarks:**

- The *trace* of  $\mathbf{A}$ ,  $\text{tr}(\mathbf{A})$ , is defined as the sum of its diagonal entries:  $\text{tr}(\mathbf{A}) \equiv \sum_{i=1}^n a_{ii}$ .
- The counterpart of the vector Euclidean norm is **not** the Euclidean matrix norm, but rather the *Frobenius norm*.
- The counterpart of the vector Chebyshev norm is the *matrix Chebyshev norm*.

Furthermore,

**Definition 3.2.1** A  $n \times n$  matrix  $\mathbf{A}$  is *symmetric* if it equals its transpose:  $\mathbf{A} = \mathbf{A}^T$

**Definition 3.2.2** A  $n \times n$  matrix  $\mathbf{A}$  is *skew-symmetric* if it equals the negative of its transpose:  $\mathbf{A} = -\mathbf{A}^T$

**Exercise 3.2.1** Prove that the eigenvalues of  $\mathbf{A}^T \mathbf{A}$ , or of  $\mathbf{A} \mathbf{A}^T$  for that matter, with  $\mathbf{A}$  defined as a  $n \times n$  matrix, are the modules of the, in general complex, eigenvalues of  $\mathbf{A}$ .

**Fact 3.2.1 (The Matrix Cartesian Decomposition)** Every  $n \times n$  matrix  $\mathbf{A}$  can be decomposed into the sum of a symmetric and a skew-symmetric components:

$$\mathbf{A} = \mathbf{A}_s + \mathbf{A}_{ss} \quad (3.8a)$$

$$\mathbf{A}_s = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T) \quad (3.8b)$$

$$\mathbf{A}_{ss} = \frac{1}{2}(\mathbf{A} - \mathbf{A}^T) \quad (3.8c)$$

Equation (3.8a) is termed the *Cartesian decomposition* of  $\mathbf{A}$ , because of its resemblance with the Cartesian representation of a complex number  $z$  as  $x + jy$ , with  $x, y \in \mathbb{R}$  and  $j = \sqrt{-1}$ . Notice that the Cartesian decomposition is *unique*.

**Exercise 3.2.2** *Prove that the eigenvalues of a skew-symmetric matrix are either 0 or imaginary.* Hint: Show first that, given any  $n \times n$  matrix  $\mathbf{A}$ , the eigenvalues of  $\mathbf{A}^2$  are the squares of the eigenvalues of  $\mathbf{A}$ , and that the two matrices,  $\mathbf{A}$  and  $\mathbf{A}^2$ , share the same eigenvectors.

The eigenvalues of the symmetric component are real and its eigenvectors are mutually orthogonal. As a consequence of Exercise 3.2.2, moreover, the eigenvalues of  $\mathbf{A}_{ss}$  of the same equation are either 0 or imaginary.

**Definition 3.2.3** A quadratic form  $q$  of a  $n$ -dimensional vector  $\mathbf{x}$  is associated with a  $n \times n$  matrix  $\mathbf{A}$ :

$$q \equiv \mathbf{x}^T \mathbf{A} \mathbf{x} \quad (3.9)$$

**Fact 3.2.2** The quadratic form associated with a skew-symmetric matrix vanishes identically. That is, if  $\mathbf{A} = -\mathbf{A}^T$ , then, for any  $n$ -dimensional vector  $\mathbf{x}$ ,

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = 0 \quad (3.10)$$

*Proof:* Note that, since  $q \equiv \mathbf{x}^T \mathbf{A} \mathbf{x}$  is a scalar,  $q = q^T$ , and hence,

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

Expanding the left-hand side,

$$\mathbf{x}^T \mathbf{A}^T \mathbf{x} = \mathbf{x}^T \mathbf{A} \mathbf{x}$$

However, by assumption,  $\mathbf{A}^T = -\mathbf{A}$ , and hence,

$$-\mathbf{x}^T \mathbf{A} \mathbf{x} = \mathbf{x}^T \mathbf{A} \mathbf{x}$$

whence the proof follows immediately.

**Definition 3.2.4** A  $n \times n$  matrix  $\mathbf{A}$  is *positive-definite* (*positive-semidefinite*) if it is symmetric and, for every  $n$ -dimensional vector  $\mathbf{x}$ , the quadratic form  $\mathbf{x}^T \mathbf{A} \mathbf{x}$  is greater than (or equal to) zero.

**Characterization of positive-definiteness (semidefiniteness):** A  $n \times n$  (symmetric) matrix  $\mathbf{A}$  is *positive-definite* (*positive-semidefinite*) if and only if its eigenvalues are all positive (nonnegative).

**Remarks:**

- Negative-definiteness and negative-semidefiniteness are defined and characterized likewise;
- If a matrix is neither positive- nor negative-definite, or semidefinite, then it is said to be *sign-indefinite*.

### 3.3 Background on Linear Transformations

The general form of a *linear transformation* mapping a vector space  $\mathcal{U}$  of dimension  $n$  into a  $m$ -dimensional vector space  $\mathcal{V}$  is

$$\mathbf{v} = \mathbf{L}\mathbf{u} \quad (3.11)$$

where  $\mathbf{u}$  and  $\mathbf{v}$  are  $n$ - and  $m$ -dimensional vectors, respectively, with  $\mathbf{u} \in \mathcal{U}$  and  $\mathbf{v} \in \mathcal{V}$ . Apparently,  $\mathbf{L}$  is a  $m \times n$  matrix.

We distinguish two vector subspaces associated with  $\mathbf{L}$ , namely,

The *range* of  $\mathbf{L}$ , denoted by  $\mathcal{R}(\mathbf{L})$ : the set of vectors  $\mathbf{v}$  that are *images* of  $\mathbf{u}$  under transformation (3.11). Notice that, if the  $n$  columns of  $\mathbf{L}$  are not linearly independent, then  $\mathcal{R}(\mathbf{L})$  is not all of  $\mathcal{V}$ , but only a *proper* subspace of it, of dimension  $m' < n$ , i.e.,  $\mathcal{R}(\mathbf{L}) \subset \mathcal{V}$ . The dimension of  $\mathcal{R}(\mathbf{L})$ , known as the rank of  $\mathbf{L}$ , is denoted by  $\rho(\mathbf{L})$ .

The *null space* or *kernel* of  $\mathbf{L}$ , denoted by  $\mathcal{N}(\mathbf{L})$ : the set of all vectors  $\mathbf{u}$  of  $\mathcal{U}$  that are mapped by  $\mathbf{L}$  into  $\mathbf{0}_m$ , the zero of  $\mathcal{V}$ . The dimension of  $\mathcal{N}$  is termed the *nullity* of  $\mathbf{L}$ , and is denoted by  $\nu(\mathbf{L})$ . Obviously,  $\nu < n$ , with  $\nu = n$  occurring only when  $\mathbf{L} = \mathbf{O}_{mn}$ ,  $\mathbf{O}_{mn}$  denoting the  $m \times n$  zero matrix.

A fundamental result of linear algebra follows:

$$\rho(\mathbf{L}) + \nu(\mathbf{L}) = n \quad (3.12)$$

The most frequent linear transformations used in optimum design are studied in the balance of this section. They all pertain to square matrices.

**Example 3.3.1** Consider the linear transformation

$$\mathbf{L} = \mathbf{1} - \mathbf{e}\mathbf{e}^T$$

with  $\mathbf{e}$  defined as a unit vector in  $\mathbb{R}^3$ , the three-dimensional vector space defined over the real field, a.k.a. the three-dimensional Euclidean space. That is,  $\mathbf{L}$  maps  $\mathcal{U} = \mathbb{R}^3$  into itself, i.e.,  $\mathcal{V} = \mathcal{U}$  in the above notation. More specifically,  $\mathbf{L}$  maps vectors in  $\mathbb{R}^3$  into vectors lying in a plane normal to  $\mathbf{e}$ . By the same token, vectors parallel to  $\mathbf{e}$  are mapped into  $\mathbf{0}_3$ , the zero vector of  $\mathbb{R}^3$ . Therefore,  $\mathcal{R}(\mathbf{L})$  is a plane normal to  $\mathbf{e}$  and passing through the origin, while  $\mathcal{N}(\mathbf{L})$  is a line parallel to  $\mathbf{e}$  and passing through the origin. Therefore,  $\rho = 2$  and  $\nu = 1$ , thereby verifying  $n = \rho + \nu$ .

### 3.3.1 Rotations

A rotation  $\mathbf{Q}$  is an *orthogonal* transformation of  $\mathcal{U}$  into itself, with a constraint on its determinant, as we shall outline presently. Orthogonality requires that the inverse of  $\mathbf{Q}$  be its transpose, i.e.,

$$\mathbf{Q}\mathbf{Q}^T = \mathbf{Q}^T\mathbf{Q} = \mathbf{1} \quad (3.13)$$

where  $\mathbf{1}$  denotes the  $n \times n$  identity matrix. Hence, taking the determinant of both sides of the above equation,

$$\det(\mathbf{Q}\mathbf{Q}^T) = \det(\mathbf{Q}^T\mathbf{Q}) = \det(\mathbf{Q})^T \det(\mathbf{Q}) = [\det(\mathbf{Q})]^2 = 1$$

whence

$$\det(\mathbf{Q}) = \pm 1$$

A *proper orthogonal matrix*  $\mathbf{Q}$  is one whose determinant is positive, and hence,

$$\det(\mathbf{Q}) = +1 \quad (3.14)$$

Proper orthogonal transformations of  $\mathcal{U}$  into itself represent rotations about the origin of  $\mathcal{U}$ .

The best-known rotations are those in two and three dimensions. Thus, for two dimensions, the  $2 \times 2$  matrix  $\mathbf{Q}$  rotating vectors through an angle  $\phi$  ccw takes the form

$$\mathbf{Q} = \begin{bmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{bmatrix} \quad (3.15a)$$

which can be expressed alternatively as

$$\mathbf{Q} = (\cos \phi)\mathbf{1}_2 + (\sin \phi)\mathbf{E}_2 \quad (3.15b)$$

with  $\mathbf{1}_2$  defined as the  $2 \times 2$  identity matrix and  $\mathbf{E}_2$  as a skew-symmetric matrix, namely,

$$\mathbf{E}_2 \equiv \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \quad (3.15c)$$

The definition of the rotation matrix in three dimensions requires the concept of cross-product matrix (CPM), which is introduced with the aid of the partial derivative of a vector with respect to another vector.

### Partial differentiation with respect to vectors

Let  $\mathbf{u}$  and  $\mathbf{v}$  be vectors of spaces  $\mathcal{U}$  and  $\mathcal{V}$ , of dimensions  $n$  and  $m$ , respectively. The partial derivative of a scalar function  $f(\mathbf{u}, \mathbf{v})$  with respect to  $\mathbf{u}$  is a  $m$ -dimensional vector whose  $i$ th component is the partial derivative of  $f$  with respect to  $u_i$ , with a corresponding definition for the partial derivative of  $f$  with respect to  $\mathbf{v}$ . The foregoing derivatives, as all other vectors, will be assumed, henceforth, to be *column* arrays. Thus,

$$\frac{\partial f}{\partial \mathbf{u}} \equiv \begin{bmatrix} \partial f / \partial u_1 \\ \partial f / \partial u_2 \\ \vdots \\ \partial f / \partial u_n \end{bmatrix}, \quad \frac{\partial f}{\partial \mathbf{v}} \equiv \begin{bmatrix} \partial f / \partial v_1 \\ \partial f / \partial v_2 \\ \vdots \\ \partial f / \partial v_m \end{bmatrix} \quad (3.16)$$

Furthermore, let  $\mathbf{v} = \mathbf{v}(\mathbf{u})$ . In order to derive  $\partial \mathbf{u} / \partial \mathbf{v}$ , first the differential  $d\mathbf{v}$  upon a differential  $d\mathbf{u}$  is computed:

$$d\mathbf{v} = \frac{\partial \mathbf{v}}{\partial u_1} du_1 + \frac{\partial \mathbf{v}}{\partial u_2} du_2 + \dots + \frac{\partial \mathbf{v}}{\partial u_n} du_n \quad (3.17a)$$

or, in array form,

$$d\mathbf{v} = \underbrace{\begin{bmatrix} \frac{\partial \mathbf{v}}{\partial u_1} & \frac{\partial \mathbf{v}}{\partial u_2} & \dots & \frac{\partial \mathbf{v}}{\partial u_n} \end{bmatrix}}_{\frac{\partial \mathbf{v}}{\partial \mathbf{u}}} \underbrace{\begin{bmatrix} du_1 \\ du_2 \\ \vdots \\ du_n \end{bmatrix}}_{d\mathbf{u}} \quad (3.17b)$$

That is, the partial derivative of  $\mathbf{v}$  with respect to  $\mathbf{u}$  is a  $m \times n$  array whose  $(i, j)$  entry is defined as  $\partial v_i / \partial u_j$ , i.e.,

$$\frac{\partial \mathbf{v}}{\partial \mathbf{u}} \equiv \begin{bmatrix} \partial v_1 / \partial u_1 & \partial v_1 / \partial u_2 & \dots & \partial v_1 / \partial u_n \\ \partial v_2 / \partial u_1 & \partial v_2 / \partial u_2 & \dots & \partial v_2 / \partial u_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial v_m / \partial u_1 & \partial v_m / \partial u_2 & \dots & \partial v_m / \partial u_n \end{bmatrix} \quad (3.18)$$

**Example 3.3.2 (The cross-product matrix (CPM) of a vector)** *Let the components of  $\mathbf{v}$  and  $\mathbf{u}$  in a certain reference frame  $\mathcal{F}$  be given as*

$$[\mathbf{v}]_{\mathcal{F}} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}, \quad [\mathbf{u}]_{\mathcal{F}} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \quad (3.19a)$$

*Then,*

$$[\mathbf{v} \times \mathbf{u}]_{\mathcal{F}} = \begin{bmatrix} v_2 u_3 - v_3 u_2 \\ v_3 u_1 - v_1 u_3 \\ v_1 u_2 - v_2 u_1 \end{bmatrix} \quad (3.19b)$$

*Hence,*

$$\left[ \frac{\partial(\mathbf{v} \times \mathbf{u})}{\partial \mathbf{u}} \right]_{\mathcal{F}} = \begin{bmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{bmatrix} \quad (3.19c)$$

Henceforth, the partial derivative of the cross product of any three-dimensional vector  $\mathbf{v}$  will be denoted by the  $3 \times 3$  matrix  $\mathbf{V}$ , which is termed the *cross-product matrix* of vector  $\mathbf{v}$ . Thus, the foregoing cross product admits the alternative representations

$$\mathbf{v} \times \mathbf{u} = \mathbf{V}\mathbf{u} \quad (3.20)$$

Note that given *any* 3-dimensional vector  $\mathbf{a}$ , its cross-product matrix  $\mathbf{A}$  is *uniquely* defined. Moreover, this matrix is skew-symmetric. The converse also holds, i.e., given any  $3 \times 3$  skew-symmetric matrix  $\mathbf{A}$ , its associated *vector* is uniquely defined as well. This result is made apparent from Example 3.3.2.

The cross-product matrix is defined only for three-dimensional vectors. Let  $\mathbf{a}$  and  $\mathbf{x}$  be two arbitrary three-dimensional vectors. We define

$$\text{CPM}(\mathbf{a}) \equiv \frac{\partial(\mathbf{a} \times \mathbf{x})}{\partial \mathbf{x}} \equiv \mathbf{A} \quad \forall \quad \mathbf{x} \in \mathbb{R}^3 \quad (3.21)$$

Because of the relation  $\mathbf{a} \times \mathbf{x} = -\mathbf{x} \times \mathbf{a}$ , the CPM is skew-symmetric.

### The rotation matrix in three dimensions

In three dimensions, the rotation matrix takes the form

$$\mathbf{Q} = \mathbf{e}\mathbf{e}^T + \cos \phi (\mathbf{1}_3 - \mathbf{e}\mathbf{e}^T) + \sin \phi \mathbf{E}_3 \quad (3.22a)$$

where  $\mathbf{e}$  is the unit vector indicating the direction of the axis about which the rotation takes place,  $\mathbf{1}_3$  is the  $3 \times 3$  identity matrix, and  $\mathbf{E}_3$  is defined as

$$\mathbf{E}_3 \equiv \text{CPM}(\mathbf{e}) \quad (3.22b)$$

Rotations in higher dimensions can be defined as well, but then, the axis and the angle of rotation are not unique.

### 3.3.2 Reflections

Reflections are improper orthogonal matrices, preserving the distance between any two points of the  $n$ -dimensional space. For any  $n \times n$  reflection  $\mathbf{R}$ , we have

$$\det(\mathbf{R}) = -1 \quad (3.23)$$

In two dimensions, a reflection  $\mathbf{R}$  about a line passing through the origin normal to the unit vector  $\mathbf{e}$  maps a vector  $\mathbf{p}$  into  $\mathbf{p}'$  in the form

$$\mathbf{p}' = \mathbf{p} - 2(\mathbf{p}^T \mathbf{e})\mathbf{e} = (\mathbf{1} - 2\mathbf{e}\mathbf{e}^T)\mathbf{p} \equiv \mathbf{R}\mathbf{p}$$

and hence, the reflection  $\mathbf{R}$  sought is given by

$$\mathbf{R} = \mathbf{1} - 2\mathbf{e}\mathbf{e}^T \quad (3.24)$$

In three dimensions, the reflection about a plane passing through the origin, of unit normal  $\mathbf{e}$ , takes exactly the same form as  $\mathbf{R}$  in the two-dimensional case, eq.(3.24). However, in this case,  $\mathbf{e}$  is three-dimensional, while  $\mathbf{R}$  is of  $3 \times 3$ , and  $\mathbf{1}$  is the  $3 \times 3$  identity matrix.

In all foregoing instances, the reflections are represented by symmetric matrices, and are hence termed *pure reflections*. However, this need not always be the case, for reflections can combine with rotations, thereby yielding a new reflection—notice that the product of a rotation by a pure reflection is a reflection!—but this time, the matrix representing the reflection is no longer symmetric. A rotation can be distinguished from a reflection by the sign of its determinant.

**Remark:** A *reflection*, like a rotation, defined on a  $n$ -dimensional vector space, preserves the inner product of any two vectors. That is, rotations and reflections preserve both the *magnitude* of vectors—their Euclidean norm—and the angle between any two vectors. For this reason, reflections and rotations—orthogonal transformations—are said *to preserve the geometry of  $\mathbb{R}^n$* .



### 3.3.3 Projections

Henceforth, a *projection*  $\mathbf{P}$  means an *orthogonal projection* onto a plane in  $n$  dimensions, which we call the *projection plane*. When  $n = 2$ , the “projection plane” becomes a line in the plane.

Let us consider a plane  $\Pi$  in a  $n$ -dimensional space, of unit normal  $\mathbf{n}$ . Any point  $P$  in this space is given by its  $n$ -dimensional position vector  $\mathbf{p}$ . Let the projection of  $P$  onto  $\Pi$  be  $P'$ , which is given by its position vector  $\mathbf{p}'$ , namely,

$$\mathbf{p}' = \mathbf{p} - (\mathbf{n}^T \mathbf{p})\mathbf{n} = (\mathbf{1} - \mathbf{n}\mathbf{n}^T)\mathbf{p} \equiv \mathbf{P}\mathbf{p} \quad (3.25)$$

where  $\mathbf{P}$  is obviously defined as

$$\mathbf{P} \equiv \mathbf{1} - \mathbf{n}\mathbf{n}^T \quad (3.26)$$

Matrix  $\mathbf{P}$  is also called a *projector*. A projector  $\mathbf{P}$  is represented by a symmetric, singular, *idempotent* matrix. Symmetry is obvious; singularity is less so, but rather straightforward. To prove that  $\mathbf{P}$  is singular, all we have to do is prove that its null space is non-empty. However, this is so because all vectors  $\mathbf{r}$  of the form  $\alpha\mathbf{n}$ , for a scalar  $\alpha \neq 0$ , are mapped by  $\mathbf{P}$  onto the zero vector. Indeed,

$$\mathbf{P}\mathbf{r} = \alpha\mathbf{P}\mathbf{n} = \alpha(\mathbf{1} - \mathbf{n}\mathbf{n}^T)\mathbf{n} = \alpha(\mathbf{n} - \mathbf{n}) = \mathbf{0}$$

A matrix is idempotent of degree  $k$  when it equals its  $k$ th power, but is different from any lower power. When  $k = 2$ , the degree is self-understood and need not be spelled out. To prove idempotency, let us calculate

$$\mathbf{P}^2 = (\mathbf{1} - \mathbf{n}\mathbf{n}^T)(\mathbf{1} - \mathbf{n}\mathbf{n}^T) = \mathbf{1} - 2\mathbf{n}\mathbf{n}^T + \underbrace{\mathbf{n}\mathbf{n}^T\mathbf{n}\mathbf{n}^T}_{=1} = \mathbf{1} - \mathbf{n}\mathbf{n}^T \equiv \mathbf{P}$$

thereby completing the proof.

The foregoing projection has a nullity of 1, its nullspace being spanned by vector  $\mathbf{n}$ . In three-dimensional space, we can have projections onto a subspace of dimension 1, namely, a line  $\mathcal{L}$  passing through the origin and parallel to the unit vector  $\mathbf{e}$ . In this case, the projection  $P'$  of  $P$  onto  $\mathcal{L}$  is given by

$$\mathbf{p}' = (\mathbf{p}^T \mathbf{e})\mathbf{e} \equiv \mathbf{e}(\mathbf{e}^T \mathbf{p}) = (\mathbf{e}\mathbf{e}^T)\mathbf{p}$$

whence the projection  $\mathbf{P}$  sought takes the form:

$$\mathbf{P} = \mathbf{e}\mathbf{e}^T \quad (3.27)$$

Notice that this projection is symmetric, singular and idempotent as well, its null space being of dimension two. Indeed, we can find two mutually-orthogonal unit vectors  $\mathbf{f}$  and  $\mathbf{g}$ , lying in a plane normal to  $\mathbf{e}$ , which are mapped by  $\mathbf{P}$  onto the zero vector. These two linearly-independent vectors lie in the null space of  $\mathbf{P}$ . For  $n$  dimensions, the projection “plane” can in fact be a subspace of dimension  $\nu \leq n - 1$ .

Also notice that the projection of eq.(3.25) maps vectors in three-dimensional space onto the null space of the *rank-one matrix*  $\mathbf{nn}^T$ , while that of eq.(3.27) does so onto the range of the rank-one matrix  $\mathbf{ee}^T$ . Now, the range of this matrix is the null space of a matrix  $\mathbf{A}$  defined as

$$\mathbf{A} \equiv \begin{bmatrix} \mathbf{f}^T \\ \mathbf{g}^T \end{bmatrix} \quad (3.28)$$

where  $\mathbf{f}$  and  $\mathbf{g}$  are mutually orthogonal unit vectors normal to  $\mathbf{e}$ . Then, we can define a projector  $\mathbf{P}$  in the form

$$\mathbf{P} = \mathbf{1} - \mathbf{A}^T \mathbf{A} = \mathbf{1} - (\mathbf{ff}^T + \mathbf{gg}^T) \quad (3.29)$$

This projector maps three-dimensional vectors onto the null space of  $\mathbf{A}$ , which is vector  $\mathbf{e}$ , as the reader can readily verify.

In general, if we have a full-rank  $m \times n$  matrix  $\mathbf{A}$ , with  $m < n$ , then,  $\text{rank}(\mathbf{A}) = \min\{m, n\} = m$ . This means that the  $m$   $n$ -dimensional rows of  $\mathbf{A}$  are linearly independent. By virtue of the basic relation (3.12), then,  $\nu = n - m$ . A projector that maps  $n$ -dimensional vectors onto the null space of  $\mathbf{A}$  is defined below:

$$\mathbf{P} = \mathbf{1} - \mathbf{A}^T (\mathbf{A} \mathbf{A}^T)^{-1} \mathbf{A} \quad (3.30)$$

Note that, by virtue of the definition of  $\mathbf{f}$  and  $\mathbf{g}$ , matrix  $\mathbf{A}$  of eq.(3.28) produces  $\mathbf{A} \mathbf{A}^T = \mathbf{1}_2$ , the  $2 \times 2$  identity matrix.

**Exercise 3.3.1** *Prove that  $\mathbf{P}$ , as given by eq.(3.30), is a projector; then prove that its projection maps  $n$ -dimensional vectors onto the null space of  $\mathbf{A}$ .*

**Example 3.3.3** *Let*

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix} \equiv \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \end{bmatrix}$$

*The null space of  $\mathbf{A}$  is spanned by a unit vector  $\mathbf{u}$  that can be found as*

$$\mathbf{u} \equiv \frac{\mathbf{b}}{\|\mathbf{b}\|}, \quad \mathbf{b} \equiv \mathbf{a}_1 \times \mathbf{a}_2$$

The projector  $\mathbf{P}$  mapping vectors in three-dimensional space onto the null space of  $\mathbf{A}$ , spanned by  $\mathbf{u}$ , is given by

$$\mathbf{P} = \mathbf{1} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A} = \frac{1}{3} \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 1 \end{bmatrix}$$

In this case,

$$\mathbf{b} = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{u} = \frac{\sqrt{3}}{3} \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}$$

Notice that the image of any vector  $\mathbf{p} = [x, y, z]^T$  under  $\mathbf{P}$  can be expressed as the product of a scalar times  $\mathbf{u}$ :

$$\mathbf{P}\mathbf{p} = \frac{1}{3} \begin{bmatrix} x - y - z \\ -x + y + z \\ -x + y + z \end{bmatrix} = \frac{1}{3}(-x + y + z) \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} = \frac{\sqrt{3}}{3}(-x + y + z)\mathbf{u}$$

which apparently lies in the null space of  $\mathbf{A}$ .

### 3.4 The Numerical Solution of Determined Linear Systems of Equations

We consider the system

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{3.31}$$

where

$\mathbf{A}$ :  $n \times n$  matrix of *known* coefficients,

$\mathbf{b}$ :  $n$ -dimensional right-hand side *known* vector,

$\mathbf{x}$ :  $n$ -dimensional vector of *unknowns*.

**Definition 3.4.1** If

$$\det(\mathbf{A}) = 0 \tag{3.32}$$

then  $\mathbf{A}$  is said to be **singular**. Otherwise,  $\mathbf{A}$  is **nonsingular**.

**Fact 3.4.1** If  $\mathbf{A}$  is nonsingular, then eq.(3.31) has a *unique solution*, which is given by

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \tag{3.33}$$

**Caveat:** Never compute  $\mathbf{A}^{-1}$  explicitly, ... unless explicitly instructed to do so. The inverse is seldom needed as such, and incurs a waste of precious CPU time! Instead, find a *good* numerical approximation to the solution, while taking into account that  $\mathbf{A}$  and  $\mathbf{b}$  are usually known only up to a random roundoff error.

**Beware of roundoff-error amplification!**

### 3.4.1 Roundoff Error of the Solution and Condition Numbers

Regarding the roundoff-error amplification when solving the system (3.31), let  $\delta\mathbf{A}$  be the matrix roundoff error in  $\mathbf{A}$ ,  $\delta\mathbf{b}$  be the vector roundoff-error in  $\mathbf{b}$ , and  $\delta\mathbf{x}$  be the vector roundoff-error incurred when solving eq.(3.31), by virtue of  $\delta\mathbf{A}$  and  $\delta\mathbf{b}$ .

The *relative roundoff errors* in the data,  $\epsilon_{\mathbf{A}}$  and  $\epsilon_{\mathbf{b}}$ , and in the computed solution,  $\epsilon_{\mathbf{x}}$ , are defined as

$$\epsilon_{\mathbf{A}} \equiv \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}, \quad \epsilon_{\mathbf{b}} \equiv \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}, \quad \epsilon_{\mathbf{x}} \equiv \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \quad (3.34)$$

where  $\|\cdot\|$  denotes *any* vector or matrix norm.

The relative roundoff error in the computed solution is known to be related to the relative roundoff error in the data via the relation (Golub and Van Loan, 1983)

$$\epsilon_{\mathbf{x}} \leq \kappa(\mathbf{A})(\epsilon_{\mathbf{A}} + \epsilon_{\mathbf{b}}) \quad (3.35)$$

with  $\kappa(\mathbf{A})$  defined as the *condition number* of matrix  $\mathbf{A}$ , which is defined, for *non-singular square matrices*, as

$$\kappa(\mathbf{A}) \equiv \|\mathbf{A}\| \|\mathbf{A}^{-1}\| \quad (3.36)$$

where  $\|\mathbf{A}\|$  is a *norm* of a matrix  $\mathbf{A}$ .

Now, if the Euclidean norm is adopted for the condition number, then we have

$$\kappa \equiv \kappa_E = \sqrt{\frac{\lambda_l}{\lambda_s}} \equiv \sqrt{\frac{\lambda_l}{\lambda_s}} \quad (3.37)$$

in which  $\lambda_s$  is the smallest and  $\lambda_l$  is the largest eigenvalue of  $\mathbf{A}\mathbf{A}^T$ . It is now apparent that  $\kappa_E$  is bounded from below but unbounded from above:

$$\kappa_E \geq 1 \quad (3.38)$$

In fact, the above result holds for  $\kappa$  defined based on any norm. Moreover, if  $\kappa(\mathbf{A})$  is based on the Frobenius norm  $\|\mathbf{A}\|_F$ , then

$$\begin{aligned}\kappa = \kappa_F(\mathbf{A}) &= \sqrt{\frac{1}{n}\text{tr}(\mathbf{A}\mathbf{A}^T)}\sqrt{\frac{1}{n}\text{tr}(\mathbf{A}^{-1}\mathbf{A}^{-T})} \\ &= \frac{1}{n}\sqrt{\text{tr}(\mathbf{A}\mathbf{A}^T)\text{tr}[(\mathbf{A}\mathbf{A}^T)^{-1}]} \equiv \frac{1}{n}\sqrt{\text{tr}(\mathbf{A}^T\mathbf{A})\text{tr}[(\mathbf{A}^T\mathbf{A})^{-1}]} \quad (3.39)\end{aligned}$$

**Remarks:**

- The condition number of a singular matrix is unbounded (tends to  $\infty$ )
- If a matrix  $\mathbf{A}\mathbf{A}^T$  has all its eigenvalues identical, then  $\mathbf{A}$  is said to be *isotropic*. Isotropic matrices have a  $\kappa = 1$  for  $\kappa$  defined in any matrix norm. Isotropic matrices are optimally conditioned.

### 3.4.2 LU-Decomposition, a.k.a. Gaussian Elimination

Various methods for computing a good approximation to *the* solution (3.33):

*Iteratively:* Various types of methods, by the names Gauss-Jordan, Gauss-Seidel, successive-overrelaxation (SOR), etc. Used mainly for “large” systems (thousands of unknowns) that are *weakly coupled*; we will not handle such systems.

*Symbolically:* Only possible for certain classes of  $\mathbf{A}$  matrices, like tridiagonal, and for arbitrary matrices of modest size ( $n$  is below 5 or so.)

*LU-decomposition, a.k.a. Gaussian elimination:* This is based on the observation that a *triangular system* is readily solved by either *backward* or *forward substitution*.  $\mathbf{A}$  is decomposed into a *lower-triangular* and an *upper-triangular* factor,  $\mathbf{L}$  and  $\mathbf{U}$ , respectively.

If  $\mathbf{A}$  is nonsingular, but otherwise arbitrary, of  $n \times n$ , then, using **Gaussian elimination** we decompose  $\mathbf{A}$  into

$$\mathbf{A} = \mathbf{L}\mathbf{U} \quad (3.40a)$$

where  $\mathbf{L}$  is *lower-triangular* and  $\mathbf{U}$  is *upper-triangular*, namely,

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & 1 \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix} \quad (3.40b)$$

Now, eq.(3.31) is rewritten as

$$\mathbf{LU}\mathbf{x} = \mathbf{b} \quad \Rightarrow \quad \begin{cases} \mathbf{L}\mathbf{y} = \mathbf{b} \\ \mathbf{U}\mathbf{x} = \mathbf{y} \end{cases} \quad (3.41)$$

and hence,  $\mathbf{x}$  is computed in two stages: First  $\mathbf{y}$  is computed from a lower-triangular system; then,  $\mathbf{x}$  is computed from an upper-triangular system. The lower-triangular system is solved for  $\mathbf{y}$  by *forward substitution*; the upper-triangular system is solved for  $\mathbf{x}$  by *backward substitution*. Note that

$$\det(\mathbf{A}) = \det(\mathbf{L})\det(\mathbf{U}) \quad (3.42a)$$

But, apparently,

$$\det(\mathbf{L}) = 1, \quad \det(\mathbf{U}) = \prod_{i=1}^n u_{ii} \quad \Rightarrow \quad \det(\mathbf{A}) = \det(\mathbf{U}) = \prod_{i=1}^n u_{ii} \quad (3.42b)$$

Hence,  $\mathbf{A}$  is singular iff any of the diagonal entries of  $\mathbf{U}$  vanishes.

### 3.4.3 Cholesky Decomposition

If  $\mathbf{A}$  is *symmetric and positive-definite*, then it admits the **Cholesky decomposition**:

$$\mathbf{A} = \mathbf{L}^T \mathbf{L} \quad (3.43)$$

$$\mathbf{L} = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix} \quad (3.44)$$

where  $\mathbf{L}$  is a *real*, lower-triangular matrix. Notice that, contrary to the lower-triangular matrix  $\mathbf{L}$  of the LU-decomposition of eqs.(3.40a), the diagonal entries of  $\mathbf{L}$  of the Cholesky-decomposition need not all be equal to unity!

The solution of system (3.31) proceeds as in the general case, in two steps:

$$\mathbf{L}^T \mathbf{y} = \mathbf{b} \quad \rightarrow \quad \mathbf{L}\mathbf{x} = \mathbf{y} \quad (3.45)$$

## 3.5 The Least-square Solution of Overdetermined Linear Systems

We start with

**Definition 3.5.1** A system of linear equations of the form

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad (3.46)$$

is *overdetermined* if  $\mathbf{A}$  is rectangular, of  $q \times n$ , with  $q > n$ .

This means that the system has more equations than unknowns. In general, it is not possible to find a  $\mathbf{x} \in \mathbb{R}^n$  that verifies *all* the  $q > n$  equations.

**Definition 3.5.2**  $\mathbf{A}$  is of full rank if its  $n$  ( $< q$ )  $q$ -dimensional columns are linearly independent.

**Remark:** If  $\mathbf{A}$  is of full rank, i.e., if  $\text{rank}(\mathbf{A}) = n$ , then

- The product  $\mathbf{A}^T \mathbf{A}$  is nonsingular, and hence, positive-definite;
- as a consequence,

$$\det(\mathbf{A}^T \mathbf{A}) > 0 \quad (3.47)$$

For an arbitrary  $\mathbf{x}$ , there will be an error  $\mathbf{e}$ :

$$\mathbf{e} \equiv \mathbf{b} - \mathbf{A}\mathbf{x} \neq \mathbf{0}_q \quad (3.48)$$

### 3.5.1 The Normal Equations for Plain Least Squares

**Problem (Plain Least Square):** find a particular  $\mathbf{x}$ ,  $\mathbf{x}_L$ , that minimizes the Euclidean norm of the error, or its square, for that matter:  $\|\mathbf{e}\|^2 = \mathbf{e}^T \mathbf{e}$ .

*Solution:* Define the *objective function*  $f$  to be minimized as

$$f \equiv \frac{1}{2} \|\mathbf{e}\|^2 \quad \rightarrow \quad \min_{\mathbf{x}} \quad (3.49)$$

This problem is termed “plain” to distinguish it from its “weighted” counterpart, to be defined presently.

The *normality conditions* (NC) of Problem (3.49) are obtained upon zeroing the gradient of  $f$  with respect to  $\mathbf{x}$ :

$$\nabla f \equiv \frac{\partial f}{\partial \mathbf{x}} = \mathbf{0} \quad (3.50)$$

Moreover,  $\nabla f$  is obtained from the “chain rule”:

$$\frac{\partial f}{\partial x_i} = \frac{\partial e_j}{\partial x_i} \frac{\partial f}{\partial e_j}, \quad i = 1, \dots, n$$

where the repeated index  $j$  indicates summation, for  $j = 1, \dots, q$ . The foregoing relation can be written in compact form as

$$\nabla f \equiv \left( \frac{\partial \mathbf{e}}{\partial \mathbf{x}} \right)^T \frac{\partial f}{\partial \mathbf{e}} \quad (3.51)$$

Apparently, from the definitions of  $f$  and  $\mathbf{e}$ ,

$$\frac{\partial \mathbf{e}}{\partial \mathbf{x}} = -\mathbf{A}, \quad \frac{\partial f}{\partial \mathbf{e}} = \mathbf{e} \equiv \mathbf{b} - \mathbf{A}\mathbf{x} \quad (3.52)$$

Upon plugging expressions (3.52) into eq.(3.50),

$$\mathbf{A}^T \mathbf{A} \mathbf{x} = \mathbf{A}^T \mathbf{b} \quad (3.53)$$

which is a system of  $n$  linear equations in  $n$  unknowns. This set of equations yields the NC of the problem at hand; the set is known as the *normal equations* of the given problem.

If  $\mathbf{A}$  is of full-rank, then eq.(3.53) admits one *unique* solution—determined case—which is the *least-square solution* of the given system:

$$\mathbf{x}_L = \mathbf{A}^I \mathbf{b} \quad (3.54a)$$

with  $\mathbf{A}^I$ , termed the *left Moore-Penrose generalized inverse* (LMPGI) of the rectangular matrix  $\mathbf{A}$ , defined as

$$\mathbf{A}^I \equiv (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \quad (3.54b)$$

#### Remarks:

- The condition number  $\kappa_E$  of the rectangular matrix  $\mathbf{A}$  of  $q \times n$ , with  $q > n$ , based on the Euclidean norm, is defined in a similar way to that of a square matrix, with the difference that, in the case at hand, this is done in terms of the eigenvalues of  $\mathbf{A}^T \mathbf{A}$ ;
- The condition number  $\kappa_E$  of  $\mathbf{A}^T \mathbf{A}$  is the square root of the ratio of the largest to the smallest eigenvalues of  $(\mathbf{A}^T \mathbf{A})(\mathbf{A}^T \mathbf{A})^T = (\mathbf{A}^T \mathbf{A})^2$ ;
- Hence,  $\kappa_E$  is given by the ratio of the largest to the smallest eigenvalues of  $(\mathbf{A}^T \mathbf{A})$ , i.e.,

$$\kappa_E(\mathbf{A}^T \mathbf{A}) = \kappa_E^2(\mathbf{A}) \quad (3.55)$$



- Thus, the roundoff-error amplification factor incurred in solving the normal equations (3.53) is the square of that incurred when “solving” eq.(3.31) in the determined case.
- Not only this. Formula (3.54a) is computationally expensive, for it involves:
  - the multiplication of  $\mathbf{A}$  by its transpose from the left, which consumes  $n^2$  scalar products of two  $q$ -dimensional vectors. Hence,  $\mathbf{A}^T \mathbf{A}$  requires  $n^2 q$  products and  $n^2(q - 1)$  additions;
  - the computation of the right-hand side of eq.(3.53), which entails, in turn,  $n$  scalar products of two  $q$ -dimensional vectors, i.e.,  $q \times n$  multiplications and  $(q - 1)n$  additions.
- In consequence,
 

**solving *numerically* normal equations should be avoided!**
- In some cases, the normal equations allow for handling them with computer algebra, in which case roundoff-error amplification is not an issue. In these cases it is safe to work with these equations.

Now we have an important result:

**Theorem 3.5.1 (The Projection Theorem)** *Let  $\mathbf{e}_o$  denote the error vector of minimum Euclidean norm, i.e.,*

$$\mathbf{e}_o \equiv \mathbf{b} - \mathbf{A}\mathbf{x}_L \quad (3.56)$$

*Then,  $\mathbf{e}_o$  is orthogonal to the image of  $\mathbf{x}_L$  under  $\mathbf{A}$ .*

*Proof:* We have

$$\mathbf{e}_o^T \mathbf{A}\mathbf{x}_L = (\mathbf{b} - \mathbf{A}\mathbf{x}_L)^T \mathbf{A}\mathbf{x}_L$$

Upon expansion,

$$\mathbf{e}_o^T \mathbf{A}\mathbf{x}_L = \mathbf{b}^T \mathbf{A}\mathbf{x}_L - \mathbf{x}_L^T \mathbf{A}^T \mathbf{A}\mathbf{x}_L$$

Plugging expressions (3.54a & b) into the above equation,

$$\begin{aligned} \mathbf{e}_o^T \mathbf{A}\mathbf{x}_L &= \mathbf{b}^T \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} - \mathbf{b}^T \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} \\ &= \mathbf{b}^T \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} - \mathbf{b}^T \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} = 0 \end{aligned} \quad (3.57)$$

thereby completing the proof. The Projection Theorem is illustrated in Fig. 3.1.

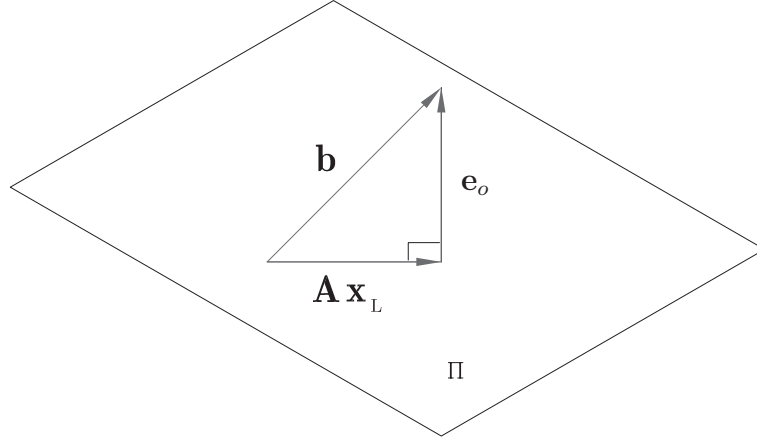


Figure 3.1: The Projection Theorem

### 3.5.2 The Normal Equations for Weighted Least Squares

If in Problem (3.54) the *Weighted Euclidean norm* is to be minimized instead, with a weighting matrix  $\mathbf{W}$  that is symmetric and positive definite, then the least-square problem thus arising is termed, correspondingly, “weighted least square”. The weighted norm is given by

$$\|\mathbf{e}\|_W \equiv \mathbf{e}^T \mathbf{W} \mathbf{e} \quad (3.58)$$

and hence,

**Problem** (Weighted Least Squares): Find  $\mathbf{x}_{WL}$  that minimizes the square of the weighted Euclidean norm of the error, i.e.,

$$f \equiv \frac{1}{2} \|\mathbf{e}\|_W^2 \rightarrow \min_{\mathbf{x}} \quad (3.59)$$

The reader should be able to verify that the normal equations are now

$$\mathbf{A}^T \mathbf{W} \mathbf{A} \mathbf{x} = \mathbf{A}^T \mathbf{W} \mathbf{b} \quad (3.60)$$

whose solution  $\mathbf{x}_{WL}$  is, in *symbolic form*:

$$\mathbf{x}_{WL} = \mathbf{A}_W^I \mathbf{b} \quad (3.61a)$$

with  $\mathbf{A}_W^I$ , termed the *weighted left Moore-Penrose generalized inverse* (LMPGI) of  $\mathbf{A}$ , defined as

$$\mathbf{A}_W^I \equiv (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W} \quad (3.61b)$$

### 3.5.3 The Numerical Solution of Least-square Problems

In the same way as the explicit numerical inversion of a square matrix coefficient is advised against, the explicit numerical computation of the LMPGI of eq.(3.61b), whether weighted or not—if  $\mathbf{W} = \mathbf{1}$ , the LMPGI is said to be unweighted whenever the qualifier is needed to avoid confusion—is also advised against. Then again, similar to the LU-decomposition of square matrices, that obviates the numerical inversion of a square matrix when solving a determined system of equations, the *QR-decomposition* obviates the numerical computation of the LMPGI. Indeed, as will be made apparent below, every  $q \times n$  rectangular matrix  $\mathbf{A}$ , with  $q > n$ , admits the factoring

$$\mathbf{A} = \mathbf{QR} \quad (3.62a)$$

where  $\mathbf{Q}$  is an orthogonal  $q \times q$  matrix, whether proper or improper, and  $\mathbf{R}$  is an upper-triangular  $q \times n$  rectangular matrix. More specifically, these matrices are represented below in block form:

$$\mathbf{QR} \equiv \underbrace{\begin{bmatrix} \mathbf{Q}_L & \mathbf{Q}_R \end{bmatrix}}_{\mathbf{Q}} \underbrace{\begin{bmatrix} \mathbf{U} \\ \mathbf{O}_{nq'} \end{bmatrix}}_{\mathbf{R}} \quad (3.62b)$$

In the above *matrix partitionings*,  $\mathbf{Q}_L$  and  $\mathbf{Q}_R$  are  $q \times n$  and  $q \times q'$  blocks, with  $q' \equiv q - n$ , while  $\mathbf{U}$  is a  $n \times n$  upper-triangular matrix and, according to the notation adopted at the outset,  $\mathbf{O}_{nq'}$  is the  $n \times (q - n)$  zero matrix. Therefore, eq.(3.46) can now be cast in the form

$$\mathbf{QRx} = \mathbf{b} \quad \Rightarrow \quad \mathbf{Rx} = \mathbf{Q}^T \mathbf{b} \quad (3.63)$$

the second equation following by virtue of the orthogonality of  $\mathbf{Q}$ . Upon taking into account the block structures of  $\mathbf{Q}$  and  $\mathbf{R}$  the same equation leads to two systems of equations, namely,

$$\mathbf{Ux} = \mathbf{Q}_L^T \mathbf{b} \quad (3.64a)$$

$$\mathbf{O}_{nq'} \mathbf{x} = \mathbf{Q}_R^T \mathbf{b} \neq \mathbf{0}_{q'} \quad (3.64b)$$

The computation of  $\mathbf{x}$  from eq.(3.64a) is now straightforward: since  $\mathbf{U}$  is upper-triangular,  $\mathbf{x}$  can be computed from this equation by backward substitution, in the same way as the second of equations (3.41) is solved for  $\mathbf{x}$ . However, eq.(3.64b) poses an apparent problem:  $\mathbf{x}$ , being pre-multiplied by the  $(q - n) \times n$  zero matrix, is bound to yield  $\mathbf{0}_{q'}$  as a product; however, the right-hand side of the same equation is, by definition, non-zero. WWW?<sup>2</sup> Nothing! This inconsistency is to be expected,

---

<sup>2</sup>What went wrong?

for, in general, no vector  $\mathbf{x} \in \mathbb{R}^n$  is capable of satisfying all  $q > n$  given linear equations. The non-zero right-hand side of eq.(3.64b) is a result of the non-zero nature of the least-square error.

All we need now is an algorithm for the computation of the QR-decomposition of a rectangular matrix like  $\mathbf{A}$ , i.e., with more rows than columns. Notice that the assumed orthogonality of  $\mathbf{Q}$  of eq.(3.63a) guarantees that the columns of  $\mathbf{A}$  are either rotated or reflected versions of the columns of  $\mathbf{R}$ , depending on the sign of  $\det(\mathbf{Q})$ . In either case, the magnitudes of all columns and the angles made by any pair thereof are preserved. The outcome is that matrices  $\mathbf{A}$  and  $\mathbf{Q}$  bear the same condition number if this is based on either the Euclidean or the Frobenius norm. Two algorithms are described below: (i) Householder reflections and (ii) Gram-Schmidt orthogonalization.

### Householder Reflections

This method is based on a sequence of  $n$  orthogonal transformations, represented by  $q \times q$  *improper orthogonal* matrices<sup>3</sup>, termed *Householder reflections* (HHR),  $\mathbf{H}_i$ , for  $i = 1, \dots, n$ . Upon successively multiplying both sides of eq. (3.46) by the  $n$  HHR, the effect is equivalent to multiplying both sides by the product  $\mathbf{H}$ :

$$\mathbf{H} \equiv \mathbf{H}_n \mathbf{H}_{n-1} \cdots \mathbf{H}_1 \quad (3.65)$$

equation (3.46) then becoming

$$\mathbf{H}\mathbf{A}\mathbf{x} = \mathbf{H}\mathbf{b} \quad (3.66)$$

The set  $\{\mathbf{H}_i\}_1^n$  is chosen so that

$$\mathbf{H}\mathbf{A} = \begin{bmatrix} \mathbf{U} \\ \mathbf{O} \end{bmatrix} \quad (3.67)$$

in which  $\mathbf{U}$  is a  $n \times n$  *upper-triangular matrix*, while  $\mathbf{O}$  the  $(q - n) \times n$  zero matrix.

The procedure to find each  $\mathbf{H}_i$  matrix is now described: Assume that we have applied reflections  $\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_{i-1}$ , in this order, to  $\mathbf{A}$ , thus obtaining a new matrix *'nigrAi* - 1:

$$\mathbf{A}_{i-1} \equiv \mathbf{H}_{i-1} \dots \mathbf{H}_2 \mathbf{H}_1 \mathbf{A} \quad (3.68)$$

---

<sup>3</sup>Proper orthogonal matrices, characterized by a determinant equal to +1, represent rotations, their improper counterparts, characterized by a determinant equal to -1, represent reflections.

which now has its first  $i - 1$  columns in *upper-triangular form*, i.e., in the form<sup>4</sup>

$$\mathbf{A}_{i-1} = \begin{bmatrix} a_{11}^* & a_{12}^* & \cdots & a_{1,i-1}^* & a_{1i}^* & \cdots & a_{1n}^* \\ 0 & a_{22}^* & \cdots & a_{2,i-1}^* & a_{2i}^* & \cdots & a_{2n}^* \\ 0 & 0 & \cdots & a_{3,i-1}^* & a_{3i}^* & \cdots & a_{3n}^* \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{i-1,i-1}^* & a_{i-1,i}^* & \cdots & a_{i-1,n}^* \\ 0 & 0 & \cdots & 0 & a_{i,i}^* & \cdots & a_{i,n}^* \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & a_{q,i}^* & \cdots & a_{qn}^* \end{bmatrix} \quad (3.69)$$

The next Householder reflection,  $\mathbf{H}_i$ , is determined so as to render the last  $q - i$  components of the  $i$ th column of  $\mathbf{H}_i \mathbf{A}_{i-1}$  equal to zero, while leaving its first  $i - 1$  columns unchanged. We do this by setting

$$\alpha_i = \text{sgn}(a_{ii}^*) \sqrt{(a_{ii}^*)^2 + (a_{i+1,i}^*)^2 + \cdots + (a_{qi}^*)^2} \quad (3.70a)$$

$$\mathbf{u}_i = [0 \quad 0 \quad \cdots \quad 0 \quad a_{ii}^* + \alpha_i \quad a_{i+1,i}^* \quad \cdots \quad a_{qi}^*]^T \quad (3.70b)$$

$$\mathbf{H}_i = \mathbf{1} - 2 \frac{\mathbf{u}_i \mathbf{u}_i^T}{\|\mathbf{u}_i\|^2} \quad (3.70c)$$

where the *signum* of  $x$ ,  $\text{sgn}(x)$ , is defined as  $+1$  if  $x > 0$ , as  $-1$  if  $x < 0$ , and is left undefined when  $x = 0$ . The reason for this provision lies in ensuring that  $|a_{ii}^* + \alpha_i|$  in eq.(3.70b) is never smaller than the absolute value of any of its two terms; else, the denominator in eq.(3.70c) could become negligibly small and lead to roundoff error amplification.

Notice that

$$\frac{1}{2} \|\mathbf{u}_i\|^2 = \alpha_i (\mathbf{u}_i)_i = \alpha_i (a_{ii}^* + \alpha_i) \equiv \beta_i$$

and hence, the denominator appearing in the expression for  $\mathbf{H}_i$  is calculated with one single addition and one single multiplication.

**Exercise 3.5.1** Show that  $\mathbf{H}_i \mathbf{H}_i^T = \mathbf{H}_i^T \mathbf{H}_i = \mathbf{1}$  and  $\det(\mathbf{H}_i) = -1$ .

**Remark:**  $\mathbf{H}_i$  reflects vectors in  $q$ -dimensional space onto a hyperplane of unit normal  $\mathbf{n} \equiv \mathbf{u}_i / \|\mathbf{u}_i\|$ , as depicted in Fig. 3.2. It is noteworthy that

---

<sup>4</sup>The entries of  $\mathbf{A}_{i-1}$  are superscripted with an asterisk to distinguish them from the entries of the original  $\mathbf{A}$ .

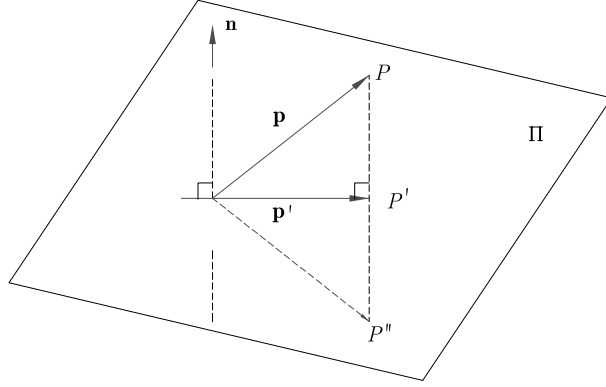


Figure 3.2: The geometric interpretation of the  $i$ th Householder reflection

- (a)  $\alpha_i$  is defined with the sign of  $a_{ii}^*$  because  $\beta_i$  is a multiple of the  $i$ th component of  $\mathbf{u}_i$ , which is, in turn, the sum of  $a_{ii}^*$  and  $\alpha_i$ , thereby guaranteeing that the absolute value of this sum will always be greater than the absolute value of each of its terms. If this provision were not made, then the resulting sum could be of a negligibly small absolute value, which would thus render  $\beta_i$  a very small positive number, thereby introducing unnecessarily an inadmissibly large roundoff-error amplification upon dividing the product  $\mathbf{u}_i \mathbf{u}_i^T$  by  $\beta_i$ ;
- (b) an arbitrary  $q$ -dimensional vector  $\mathbf{v}$  is transformed by  $\mathbf{H}_i$  with unusually few flops, namely,

$$\mathbf{H}_i \mathbf{v} = \mathbf{v} - \frac{1}{\beta_i} (\mathbf{v}^T \mathbf{u}_i) \mathbf{u}_i$$

Upon application of the  $n$  Householder reflections thus, matrix  $\mathbf{A}$  becomes

$$\mathbf{H}\mathbf{A} = \begin{bmatrix} \mathbf{U} \\ \mathbf{O}_{q'n} \end{bmatrix} \quad (3.71)$$

with  $\mathbf{H}$  defined as

$$\mathbf{H} \equiv \mathbf{H}_n \dots \mathbf{H}_2 \mathbf{H}_1 \quad (3.72)$$

Notice that  $\mathbf{H}\mathbf{A}$  is in *upper-triangular* form. That is,

$$\mathbf{H}\mathbf{A} = \begin{bmatrix} \mathbf{U} \\ \mathbf{O}_{q'n} \end{bmatrix} \quad (3.73)$$

where  $\mathbf{O}_{q'n}$  is the  $(q - n) \times n$  zero matrix.

Now, if  $\mathbf{H}^T \equiv \mathbf{Q}$  and the right-hand side of eq.(3.71) is defined as  $\mathbf{R}$ , i.e.,

$$\mathbf{R} = \begin{bmatrix} \mathbf{U} \\ \mathbf{O}_{q'n} \end{bmatrix} \quad (3.74)$$

then  $\mathbf{A}$  can be expressed as the product

$$\mathbf{A} = \mathbf{Q}\mathbf{R} \quad (3.75)$$

which is the QR-decomposition of  $\mathbf{A}$ .

**Remarks:**

- The solution of the weighted least-square problem proceeds likewise, if with eq.(3.46) replaced with

$$\mathbf{V}\mathbf{A}\mathbf{x} = \mathbf{V}\mathbf{b} \quad (3.76)$$

and  $\mathbf{V}$  found from the Cholesky-decomposition of  $\mathbf{W}$ :

$$\mathbf{W} = \mathbf{V}^T\mathbf{V} \quad (3.77)$$

- Any  $q \times n$  matrix admits the QR-decomposition eq.(3.75), provided that  $q > n$ , which is the case of  $\mathbf{V}\mathbf{A}$ . Hence, for weighted least squares, rather than the QR-decomposition of  $\mathbf{A}$ , that of  $\mathbf{V}\mathbf{A}$  is implemented.
- If  $\mathbf{Q}$  is partitioned into a  $q \times n$  left block  $\mathbf{Q}_L$  and a  $q' \times n$  right block  $\mathbf{Q}_R$ ,

$$\mathbf{Q} = [\mathbf{Q}_L \quad \mathbf{Q}_R] \quad (3.78)$$

then

$$\mathbf{V}\mathbf{A} = \mathbf{Q}_L\mathbf{U} \quad (3.79)$$

- The least-square approximation  $\mathbf{x}_L$  eq.(3.46), displayed in eq.(3.54a), is now readily found by backward substitution from the upper-triangular system

$$\mathbf{U}\mathbf{x}_L = \mathbf{Q}_L^T\mathbf{b} \quad (3.80)$$

with a similar expression for the solution  $\mathbf{x}_{WL}$  of the weighted least squares, displayed in eqs.(3.61a & b),

### The Gram-Schmidt Orthogonalization Procedure

This procedure consists in regarding the columns of  $\mathbf{A}$  as a set of  $n$   $m$ -dimensional vectors  $\{\mathbf{a}_k\}_1^n$ . From this set, a new set  $\{\mathbf{e}_k\}_1^n$  is obtained that is *orthonormal*. The first step consists in defining a *normal* vector—i.e, of *unit norm*— $\mathbf{e}_1$  as<sup>5</sup>

$$\mathbf{e}_1 = \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|} \quad (3.81)$$

---

<sup>5</sup>In the balance of this section only the Euclidian norm is used; for the sake of brevity, this norm is simply referred to as “the norm”.

Further, we define  $\mathbf{e}_2$  as the *normal component* of  $\mathbf{a}_2$  onto  $\mathbf{e}_1$ , namely,

$$\mathbf{b}_2 \equiv (\mathbf{1} - \mathbf{e}_1 \mathbf{e}_1^T) \mathbf{a}_2 \quad (3.82a)$$

$$\mathbf{e}_2 \equiv \frac{\mathbf{b}_2}{\|\mathbf{b}_2\|} \quad (3.82b)$$

In the next step, we define  $\mathbf{e}_3$  as the unit vector normal to the plane defined by  $\mathbf{e}_1$  and  $\mathbf{e}_2$  and in the direction in which the inner product  $\mathbf{e}_3^T \mathbf{a}_3$  is positive, which is possible because all vectors of the set  $\{\mathbf{a}_k\}_1^m$  have been assumed to be linearly independent—remember that  $\mathbf{A}$  has been assumed to be of full rank. To this end, we subtract from  $\mathbf{a}_3$  its projection onto the plane mentioned above, i.e.,

$$\mathbf{b}_3 = (\mathbf{1} - \mathbf{e}_1 \mathbf{e}_1^T - \mathbf{e}_2 \mathbf{e}_2^T) \mathbf{a}_3 \quad (3.83a)$$

$$\mathbf{e}_3 \equiv \frac{\mathbf{b}_3}{\|\mathbf{b}_3\|} \quad (3.83b)$$

and so on, until we obtain  $\mathbf{e}_{n-1}$ . The last unit vector of the orthogonal set,  $\mathbf{e}_n$ , is obtained as

$$\mathbf{b}_n = (\mathbf{1} - \mathbf{e}_1 \mathbf{e}_1^T - \mathbf{e}_2 \mathbf{e}_2^T - \cdots - \mathbf{e}_{n-1} \mathbf{e}_{n-1}^T) \mathbf{a}_n \quad (3.84a)$$

whence,

$$\mathbf{e}_n \equiv \frac{\mathbf{b}_n}{\|\mathbf{b}_n\|} \quad (3.84b)$$

In the next stage, we represent all vectors of the set  $\{\mathbf{a}_k\}_1^n$  in *orthogonal coordinates*, i.e., in the base  $\mathcal{O} = \{\mathbf{e}_k\}_1^n$ , which are then arranged in a  $m \times n$  array  $\mathbf{A}_o$ . By virtue of the form in which the set  $\{\mathbf{e}_k\}_1^n$  was defined, the last  $m - k$  components of vector  $\mathbf{a}_k$  vanish. We thus have, in the said orthonormal basis,

$$[\mathbf{a}_k]_{\mathcal{O}} = \begin{bmatrix} \alpha_{1k} \\ \alpha_{2k} \\ \vdots \\ \alpha_{kk} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad k = 1, 2, \dots, n \quad [\mathbf{b}]_{\mathcal{O}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$$



Therefore, eq.(3.46), when expressed in  $\mathcal{O}$ , becomes

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ 0 & \alpha_{22} & \cdots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_{nn} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \\ \beta_{n+1} \\ \vdots \\ \beta_m \end{bmatrix} \quad (3.85)$$

whence  $\mathbf{x}$  can be computed by backward substitution. It is apparent, then, that the last  $m - n$  equations of the foregoing system are incompatible: their left-hand sides are zero, while their right-hand sides are not necessarily so. What the right-hand sides of these equations represent, then, is the approximation error in orthogonal coordinates. Its Euclidean norm is, then,

$$\|\mathbf{e}_0\| \equiv \sqrt{\beta_{n+1}^2 + \cdots + \beta_m^2} \quad (3.86)$$

Apparently, then, the QR-decomposition of  $\mathbf{A}$  was implemented, with the upper-triangular block of the matrix coefficient of eq.(3.85) playing the role of  $\mathbf{U}$ , the block of zeros being  $\mathbf{O}_{q'n}$  in eq.(3.62b). The two foregoing blocks make up matrix  $\mathbf{R}$  of the QR-decomposition of  $\mathbf{A}$ . While the factor  $\mathbf{Q}$  is not apparent in eq.(3.85), it is implicit, as it is responsible for the rendering of  $\mathbf{A}$  in upper-triangular form.

The solution of linear least-square problems using commercial software is a standard feature. For example, in Maple, the `LeastSquares(A, B)` function returns a vector that best fits eq.(3.46) in the least-square sense. The QR-decomposition of a rectangular matrix with more rows than columns is standard in both computer-algebra and numerical software.

**Remark:** A square matrix also admits the QR-decomposition, which can then be used to solve determined systems of linear equations. However, the LU-decomposition in these cases is by far more economical.

## 3.6 Nonlinear-equation Solving: the Determined Case

**Definition 3.6.1** A system of algebraic equations containing some that are not linear is termed *nonlinear*. If the number of equations is identical to the number of unknowns, the system is *determined*.

Unlike their linear counterparts, determined systems of nonlinear equations admit multiple solutions. Consider, for example, the intersection of the circle  $\mathcal{C}$  and the hyperbola  $\mathcal{H}$  depicted in Fig. 3.3 and represented by the equations below:

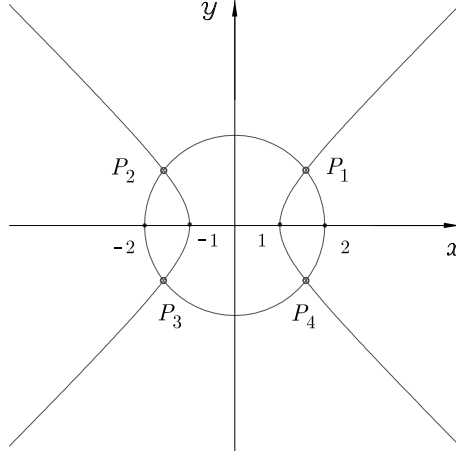


Figure 3.3: Intersection of a circle and a hyperbola

$$\mathcal{C} : \phi_1(x, y) \equiv x^2 + y^2 - 4 = 0$$

$$\mathcal{H} : \phi_2(x, y) \equiv x^2 - y^2 - 1 = 0$$

The circle and the hyperbola intersect at four points  $\{P_i\}_1^4$ , of coordinates  $(x_i, y_i)$ , as displayed in Table 3.1. The problem may have **no real solution**, e.g., the circle

$P_i$	$x_i$	$y_i$
1	$\sqrt{5/2}$	$\sqrt{3/2}$
2	$\sqrt{5/2}$	$-\sqrt{3/2}$
3	$-\sqrt{5/2}$	$\sqrt{3/2}$
4	$-\sqrt{5/2}$	$-\sqrt{3/2}$

Table 3.1: The four intersection points of the circle and the hyperbola of Fig. 3.3

and the hyperbola of Fig. 3.4 do not intersect. The system of equations from which the coordinates of the intersection points are to be computed is given below:

$$\mathcal{C} : \phi_1(x, y) \equiv x^2 + y^2 - 1 = 0$$

$$\mathcal{H} : \phi_2(x, y) \equiv x^2 - y^2 - 16 = 0$$

This system of equations admits no real solution!

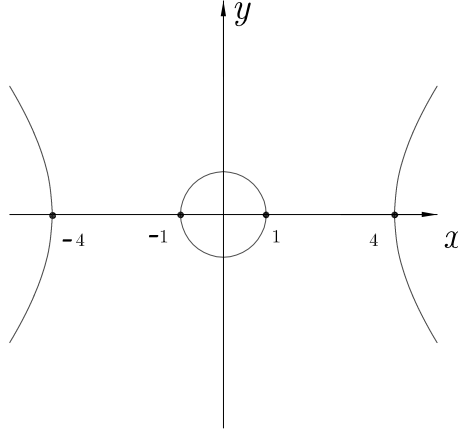


Figure 3.4: A circle and a hyperbola that do not intersect

In general, a determined nonlinear system of equations takes the form

$$\boldsymbol{\phi}(\mathbf{x}) = \mathbf{0} \quad (3.87)$$

where  $\mathbf{x}$  and  $\boldsymbol{\phi}$  are both  $n$ -dimensional vectors:

$$\mathbf{x} \equiv \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \boldsymbol{\phi} \equiv \begin{bmatrix} \phi_1(x_1, x_2, \dots, x_n) \\ \phi_2(x_1, x_2, \dots, x_n) \\ \vdots \\ \phi_n(x_1, x_2, \dots, x_n) \end{bmatrix} \quad (3.88)$$

### 3.6.1 The Newton-Raphson Method

We outline below the method of solution of determined nonlinear systems using the *Newton-Raphson method*. This is an *iterative method*, whereby a sequence of approximations is obtained that, if converging, it approaches the solution in a finite number of iterations within a prescribed *tolerance*.

A value  $\mathbf{x}^0$  of  $\mathbf{x}$  is given as an *initial guess*:

$$\mathbf{x}^0 \equiv [p_1 \quad p_2 \quad \dots \quad p_n]^T$$

and  $\boldsymbol{\phi}$  is evaluated at  $\mathbf{x}^0$ :

$$\boldsymbol{\phi}^0 \equiv \boldsymbol{\phi}(\mathbf{x}^0)$$

If the value  $\mathbf{x}^0$  was chosen randomly, most likely it will not verify the given system of equations, i.e.,

$$\boldsymbol{\phi}^0 \neq \mathbf{0}$$

Next, we look for a “small” increment  $\Delta \mathbf{x}$  of  $\mathbf{x}$  (the increment is small if its norm—any norm—is small):

$$\Delta \mathbf{x} \equiv [\Delta x_1 \quad \Delta x_2 \quad \dots \quad \Delta x_n]^T$$

Now,  $\phi(\mathbf{x}^0 + \Delta \mathbf{x})$  is evaluated up to its linear approximation (all quadratic and higher-order terms are dropped from its series expansion):

$$\phi(\mathbf{x}^0 + \Delta \mathbf{x}) \approx \phi(\mathbf{x}^0) + \left. \frac{\partial \phi}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}^0} \Delta \mathbf{x} \quad (3.89)$$

The *Jacobian matrix* of  $\phi$  with respect to  $\mathbf{x}$  is defined as the matrix of partial derivatives of the components of  $\phi$  with respect to all the components of  $\mathbf{x}$ :

$$\Phi \equiv \frac{\partial \phi}{\partial \mathbf{x}} \equiv \nabla \phi = \begin{bmatrix} \partial \phi_1 / \partial x_1 & \partial \phi_1 / \partial x_2 & \cdots & \partial \phi_1 / \partial x_n \\ \partial \phi_2 / \partial x_1 & \partial \phi_2 / \partial x_2 & \cdots & \partial \phi_2 / \partial x_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial \phi_n / \partial x_1 & \partial \phi_n / \partial x_2 & \cdots & \partial \phi_n / \partial x_n \end{bmatrix} \quad (3.90)$$

In the next step, we find  $\Delta \mathbf{x}$  that renders zero the linear approximation of  $\phi(\mathbf{x}_0 + \Delta \mathbf{x})$ :

$$\phi_0 + \Phi(\mathbf{x}^0) \Delta \mathbf{x} = \mathbf{0}$$

or

$$\Phi(\mathbf{x}^0) \Delta \mathbf{x} = -\phi^0 \quad (3.91)$$

whence  $\Delta \mathbf{x}$  can be found using, for example, Gaussian elimination:

$$\Delta \mathbf{x} = -\Phi_0^{-1} \phi^0, \quad \Phi_0 \equiv \Phi(\mathbf{x}^0) \quad (3.92)$$

Next,  $\mathbf{x}$  is updated:

$$\mathbf{x} \leftarrow \mathbf{x}^0 + \Delta \mathbf{x} \quad (3.93)$$

the procedure stopping when

$$\|\Delta \mathbf{x}\| \leq \epsilon_x \quad (3.94)$$

for a prescribed tolerance  $\epsilon_x$ .

### Remarks:

- Use the maximum norm to test convergence in eq.(3.94), for it costs virtually nothing;
- no guarantee that the Newton-Raphson method will converge at all;

- whether the Newton-Raphson method converges or not is dependent upon the initial guess,  $\mathbf{x}^0$ ;
- the boundary between regions of convergence and divergence is a *fractal* (Mandelbrot, 1983; Gleick, 1988);
- when the Newton-Raphson method converges, it does so *quadratically*: At every iteration, *two* decimal places of accuracy are gained (Dahlquist and Björck, 1974).

**Example 3.6.1 (Intersection of a circle and a hyperbola)** *The Newton-Raphson method is applied here to find one of the intersections of the circle and the hyperbola displayed in Fig. 3.3. The functions  $\phi_1$  and  $\phi_2$  describing the two curves were given above. The corresponding Jacobian matrix is displayed below, with variables  $x$  and  $y$  labelled  $x_1$  and  $x_2$ , respectively.*

$$\Phi = \begin{bmatrix} \partial\phi_1/\partial x_1 & \partial\phi_1/\partial x_2 \\ \partial\phi_2/\partial x_1 & \partial\phi_2/\partial x_2 \end{bmatrix} = \begin{bmatrix} 2x_1 & 2x_2 \\ 2x_1 & -2x_2 \end{bmatrix}$$

*Given the initial guess  $\mathbf{x}^0 = [1.0, 1.0]^T$ , function  $\phi$  and its Jacobian  $\Phi$  are evaluated at this value of  $\mathbf{x}$  as*

$$\phi^0 = \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix}, \quad \Phi_0 = \begin{bmatrix} 2 & 2 \\ 2 & -2 \end{bmatrix}$$

*which yield the increment  $\Delta\mathbf{x}^0 = [0.75, 0.25]^T$ . Successive iterations, recorded in Circle-Hype.mw, are listed below:*

$$\mathbf{x}^1 = \begin{bmatrix} 1.75 \\ 1.25 \end{bmatrix}, \quad \phi^1 = \begin{bmatrix} 0.625 \\ 0.500 \end{bmatrix}, \quad \Phi_1 = \begin{bmatrix} 3.5 & 2.5 \\ 3.5 & -2.5 \end{bmatrix}$$

$$\mathbf{x}^2 = \begin{bmatrix} 1.5893 \\ 1.2250 \end{bmatrix}, \quad \phi^2 = \begin{bmatrix} 0.0265 \\ 0.0252 \end{bmatrix}, \quad \Phi_2 = \begin{bmatrix} 3.1786 & 2.4500 \\ 3.1786 & -2.4500 \end{bmatrix}$$

$$\mathbf{x}^3 = \begin{bmatrix} 1.5812 \\ 1.2247 \end{bmatrix}, \quad \phi^3 = \begin{bmatrix} 6.610 \\ 6.600 \end{bmatrix} \times 10^{-5}, \quad \Phi_3 = \begin{bmatrix} 3.162 & 2.449 \\ 3.162 & -2.449 \end{bmatrix}$$

$$\mathbf{x}^4 = \begin{bmatrix} 1.581 \\ 1.225 \end{bmatrix}, \quad \Delta\mathbf{x}^4 = \begin{bmatrix} -2.10 \times 10^{-5} \\ -2.66 \times 10^{-8} \end{bmatrix}$$

In light of the small-enough increment, the procedure is stopped here. Figure 3.5a illustrates the iteration sequence, with a zoom-in included in Fig. 3.5b.

The reader should notice that, given the quadratic nature of the Newton-Raphson method, the norm of the error  $\phi^k$  at the  $k$ th iteration is roughly two orders of magnitude smaller than at the previous iteration in the proximity of a solution. The error upon convergence can be computed from the exact results listed in Table 3.1.

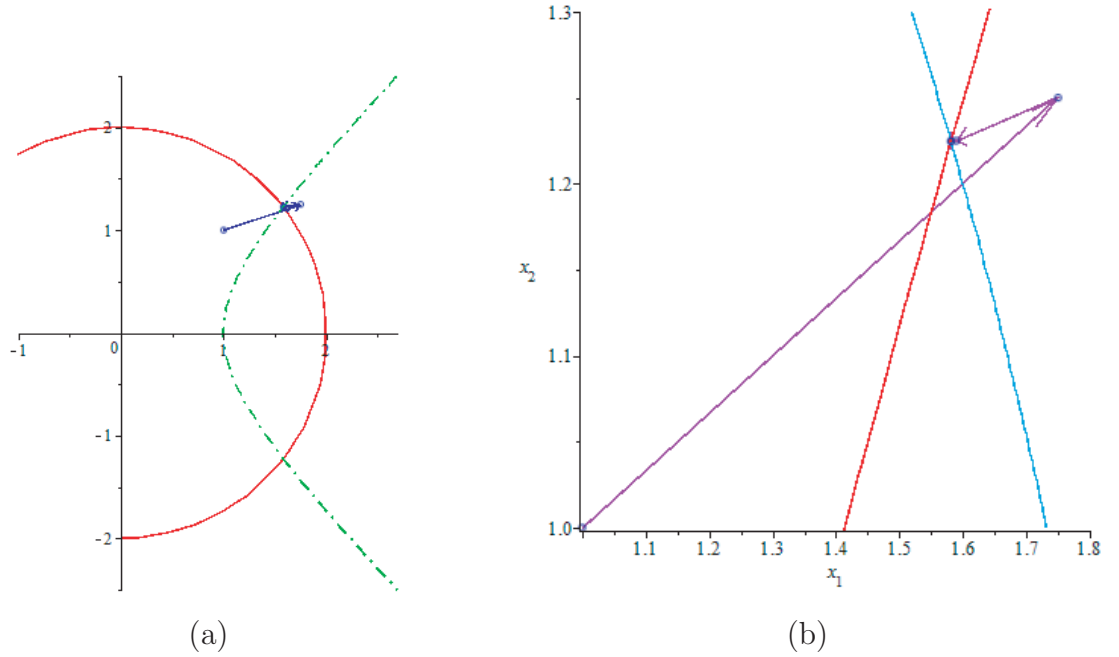


Figure 3.5: Iteration sequence in the Newton-Raphson procedure to find the intersection of a circle and a hyperbola: (a) the big picture; (b) a zoom-in close to the solution

### The Damping Factor

When implementing the Newton-Raphson method, the norm of the error  $\phi^k$  may increase upon correcting  $\mathbf{x}^k$ , i.e.,

$$f(\mathbf{x}^{k+1}) > f(\mathbf{x}^k) \quad (3.95)$$

This increase gives rise to oscillations and sometimes even leads to divergence. One way to cope with this situation is by introducing *damping*. Instead of using the whole increment  $\Delta\mathbf{x}^k$ , we use a fraction of it, i.e.

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha\Delta\mathbf{x}^k, \quad 0 < \alpha < 1 \quad (3.96)$$

where  $\alpha$  is known as the *damping factor*.

## 3.7 Overdetermined Systems of Nonlinear Equations

A system of nonlinear equations of the form

$$\phi(\mathbf{x}) = \mathbf{0} \quad (3.97)$$

where  $\mathbf{x}$  is a  $n$ -dimensional vector and  $\phi$  is a  $q$ -dimensional vector, is *overdetermined* if  $q > n$ . Just as in the linear case, in general, no vector  $\mathbf{x}$  can be found that verifies *all* the  $q$  scalar equations of the system. However, approximations can be found that minimize the least-square error of the approximation, as described in the balance of this Section. The method of solution adopted here is the overdetermined counterpart of the Newton-Raphson method.

### 3.7.1 The Newton-Gauss Method

**Problem:** Find an *approximate solution* to system (3.97) that verifies those equations with the *least-square error*:

$$f(\mathbf{x}) = \frac{1}{2} \phi^T \mathbf{W} \phi \rightarrow \min_{\mathbf{x}} \quad (3.98)$$

where  $\mathbf{W}$  is a  $q \times q$  positive-definite *weighting matrix*.

**Solution:** We follow a procedure similar to Newton-Raphson's, which is known as the *Newton-Gauss method*, as described below:

First, an initial guess  $\mathbf{x}^0$  of  $\mathbf{x}$  is given; then, we produce the sequence

$$\mathbf{x}^1, \mathbf{x}^2, \dots, \quad (3.99)$$

such that

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \Delta \mathbf{x}^k \quad (3.100)$$

**Calculation of  $\Delta \mathbf{x}^k$ :**

- Factor  $\mathbf{W}$  into its two Cholesky factors:

$$\mathbf{W} = \mathbf{V}^T \mathbf{V} \quad (3.101)$$

which is possible because  $\mathbf{W}$  is assumed positive-definite.

- Compute  $\Delta \mathbf{x}^k$  as the *weighted least-square solution* of the unconstrained overdetermined linear system

$$\mathbf{V}\Phi(\mathbf{x}^k)\Delta \mathbf{x}^k = -\mathbf{V}\phi(\mathbf{x}^k) \quad (3.102)$$

with  $\Phi(\mathbf{x})$  defined as the  $q \times n$  Jacobian matrix of the vector function  $\phi(\mathbf{x})$ , i.e.,

$$\Phi(\mathbf{x}) = \frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}} \quad (3.103)$$

Drop superscripts for the sake of notation-simplicity and recall eq.(3.61a):

$$\Delta \mathbf{x} = -(\Phi^T \mathbf{W} \Phi)^{-1} \Phi^T \mathbf{W} \phi \quad (3.104)$$

This procedure is iterative, stopping when a *convergence criterion is met*.

### 3.7.2 Convergence Criterion

Calculate first  $\nabla f(\mathbf{x})$ :

$$\nabla f(\mathbf{x}) \equiv \frac{\partial f}{\partial \mathbf{x}} = \left( \frac{\partial \phi}{\partial \mathbf{x}} \right)^T \frac{\partial f}{\partial \phi} \quad (3.105)$$

$$\frac{\partial \phi}{\partial \mathbf{x}} \equiv \Phi, \quad \frac{\partial f}{\partial \phi} = \mathbf{W} \phi \quad (3.106)$$

The condition for a stationary point is that  $\nabla f(\mathbf{x})$  vanish, and hence,

$$\Phi^T \mathbf{W} \phi = \mathbf{0} \quad (3.107)$$

which is the *normality condition* of eq.(3.98).

It is thus apparent that, at a stationary point of  $f$ ,  $\phi(\mathbf{x})$  **need not vanish**, as is the case of unconstrained optimization, to be discussed in Chapter 4; however,  $\phi(\mathbf{x})$  **must lie in the null space of  $\Phi^T \mathbf{W}$** . Moreover, from eqs.(3.104) and (3.107) follows that, at a stationary point,  $\Delta \mathbf{x}$  vanishes. Hence, the convergence criterion is

$$\|\Delta \mathbf{x}\| < \epsilon \quad (3.108)$$

where  $\epsilon$  is a prescribed tolerance.

#### Remarks:

- The normality condition (3.107) alone does not guarantee a minimum, but only a *stationary point*.



- However, as it turns out, if the procedure converges, then it does so, to a first-order approximation, to a minimum, and neither to a maximum nor a saddle point, as we prove below.

The sequence  $f(\mathbf{x}^0), f(\mathbf{x}^1), \dots, f(\mathbf{x}^k), f(\mathbf{x}^{k+1}), \dots$ , obtained from the sequence of  $\mathbf{x}$  values, evolves, to a first order, as  $\Delta f(\mathbf{x})$ , given by

$$\Delta f = \left( \frac{\partial f}{\partial \mathbf{x}} \right)^T \Delta \mathbf{x} \quad (3.109)$$

i.e.,

$$\Delta f = \boldsymbol{\phi}^T \mathbf{W} \boldsymbol{\Phi} \Delta \mathbf{x} \quad (3.110)$$

Upon plugging expression (3.104) of  $\Delta \mathbf{x}$  into eq. (3.110), we obtain

$$\Delta f = -\boldsymbol{\phi}^T \underbrace{\mathbf{W} \boldsymbol{\Phi} (\boldsymbol{\Phi}^T \mathbf{W} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{W}}_{\mathbf{M}} \boldsymbol{\phi} = -\boldsymbol{\phi}^T \mathbf{M} \boldsymbol{\phi} \quad (3.111)$$

where, apparently,  $\mathbf{M}$  is a  $q \times q$  positive-definite matrix. As a consequence,  $\boldsymbol{\phi}^T \mathbf{M} \boldsymbol{\phi}$  becomes a positive-definite quadratic expression of  $\boldsymbol{\phi}$ ; hence,  $\Delta f$  is negative definite. Thus, the first-order approximation of  $\Delta f(\mathbf{x})$  is negative-definite, and hence, the sequence of  $f$  values *decreases monotonically*. That is, in the neighborhood of a stationary point, the first-order approximation of  $\boldsymbol{\phi}(\mathbf{x})$  is good enough, and hence, if the procedure **converges**, it does so to a **minimum**.

**Example 3.7.1 (“Intersection” of a circle, a hyperbola and a parabola)** *Given the circle  $\mathcal{C}$ , the hiperbola  $\mathcal{H}$  and the parabola  $\mathcal{P}$  displayed in Fig. 3.6, find one point closest to the three curves in the least-square sense.*

*Solution: the three curves have the representations displayed below:*

$$\begin{aligned} \mathcal{C} : \quad & x_1^2 + x_2^2 - 4 = 0 \\ \mathcal{H} : \quad & x_1^2 - x_2^2 - 1 = 0 \\ \mathcal{P} : \quad & x_1^2 - 2.4x_2 = 0 \end{aligned}$$

*The vector function  $\boldsymbol{\phi}$  and its gradient  $\boldsymbol{\Phi}$  are readily derived as*

$$\boldsymbol{\phi} = \begin{bmatrix} x_1^2 + x_2^2 - 4 \\ x_1^2 - x_2^2 - 1 \\ x_1^2 - 2.4x_2 \end{bmatrix}, \quad \boldsymbol{\Phi} = \begin{bmatrix} 2x_1 & 2x_2 \\ 2x_1 & -2x_2 \\ 2x_1 & -2.4 \end{bmatrix}$$

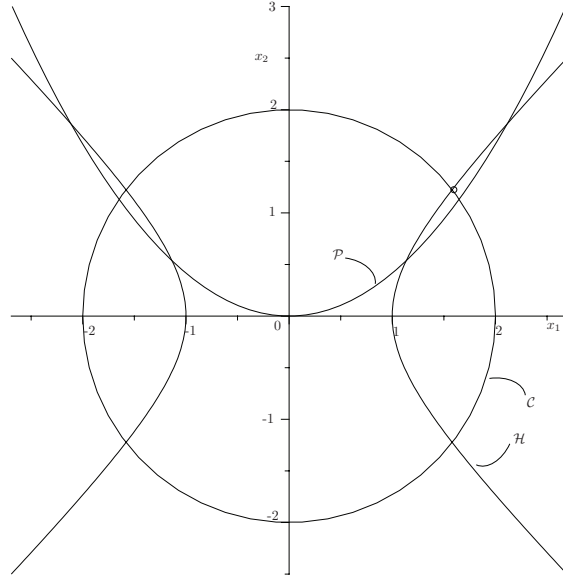


Figure 3.6: Finding the point closest to a circle  $\mathcal{C}$ , a hyperbola  $\mathcal{H}$  and a parabola  $\mathcal{P}$

Given the initial guess  $\mathbf{x}^0 = [1, 1]^T$ , function  $\phi$  and its gradient  $\Phi$  are evaluated as

$$\phi = \begin{bmatrix} -2.0 \\ -1.0 \\ -1.4 \end{bmatrix}, \quad \Phi = \begin{bmatrix} 2 & 2 \\ 2 & -2 \\ -0.4 & 0 \end{bmatrix}$$

which yield the increment  $\Delta \mathbf{x}^0 = [0.6667, 0.2500]^T$ . Successive iterations, recorded in `Circle-Hype-Para-plot.mw`, are listed below:

$$\mathbf{x}^1 = \begin{bmatrix} 1.667 \\ 1.250 \end{bmatrix}, \quad \phi^1 = \begin{bmatrix} 0.3403 \\ 0.2153 \\ -0.2222 \end{bmatrix}, \quad \Phi_1 = \begin{bmatrix} 3.333 & 2.500 \\ 3.333 & -2.500 \\ 0.933 & 0 \end{bmatrix}$$

$$\mathbf{x}^2 = \begin{bmatrix} 3.333 & 2.500 \\ 3.333 & -2.500 \\ 0.9333 \end{bmatrix}, \quad \phi^2 = \begin{bmatrix} 0.0461 \\ 0.0449 \\ -0.394 \end{bmatrix}, \quad \Phi_2 = \begin{bmatrix} 3.191 & 2.450 \\ 3.1901 & -2.450 \\ 0.791 & 0 \end{bmatrix}$$

$$\mathbf{x}^3 = \begin{bmatrix} 1.596 \\ 1.225 \end{bmatrix}, \quad \phi^3 = \begin{bmatrix} 0.0488 \\ 0.04878 \\ -0.3906 \end{bmatrix}, \quad \Phi_3 = \begin{bmatrix} 3.193 & 2.449 \\ 3.193 & -2.449 \\ 0.793 & 0 \end{bmatrix}$$

$$\mathbf{x}^4 = \begin{bmatrix} 1.596 \\ 1.225 \end{bmatrix}, \quad \phi^4 = \begin{bmatrix} 0.0485 \\ 0.0485 \\ -0.391 \end{bmatrix}, \quad \Phi_4 = \begin{bmatrix} 3.193 & 2.450 \\ 3.193 & -2.450 \\ 0.793 & 0 \end{bmatrix}$$

$$\mathbf{x}^5 = \begin{bmatrix} 1.596 \\ 1.225 \end{bmatrix}, \quad \phi^5 = \begin{bmatrix} 0.0485 \\ 0.0485 \\ -0.391 \end{bmatrix}, \quad \Phi_5 = \begin{bmatrix} 3.193 & 2.450 \\ 3.193 & -2.450 \\ 0.793 & 0 \end{bmatrix}$$

At this stage, the normality condition (3.107) is tested:

$$\Phi_5^T \phi^5 = \begin{bmatrix} 7.243 \times 10^{-6} \\ -5.412 \times 10^{-16} \end{bmatrix}$$

which shows that the second component is acceptably small, but the first could still be improved. Therefore, one more iteration is conducted:

$$\mathbf{x}^6 = \begin{bmatrix} 1.596 \\ 1.225 \end{bmatrix}, \quad \phi^6 = \begin{bmatrix} 0.0485 \\ 0.0485 \\ -0.391 \end{bmatrix}, \quad \Phi_6 = \begin{bmatrix} 3.193 & 2.450 \\ 3.193 & -2.450 \\ 0.793 & 0 \end{bmatrix}$$

the corresponding normality condition now being

$$\Phi_6^T \phi^6 = \begin{bmatrix} -4.532 \times 10^{-7} \\ -5.551 \times 10^{-16} \end{bmatrix}$$

which is acceptably small, and hence,  $\mathbf{x}^6$  is taken as the least-square approximation sought.

Figure 3.7a illustrates the iteration sequence, with a zoom-in included in Fig. 3.7b.

The reader may wonder whether the Newton-Raphson method can be used to solve nonlinear least-square problems. Although the answer is *yes*, the Newton-Raphson method is not advisable in this case, as made apparent below.

Recall  $\nabla f$  from eqs.(3.105) and (3.106):

$$\nabla f(\mathbf{x}) = \frac{\partial f}{\partial \mathbf{x}} = \underbrace{\Phi^T(\mathbf{x})}_{n \times q} \underbrace{\mathbf{W}}_{q \times q} \underbrace{\phi(\mathbf{x})}_{q-\text{dim}}$$

$$\nabla f(\mathbf{x}) = \mathbf{0} \Rightarrow \underbrace{\Phi^T(\mathbf{x})\mathbf{W}\phi(\mathbf{x})}_{\equiv \psi(\mathbf{x}) \in \mathbb{R}^n} = \mathbf{0} \quad (\text{NC})$$

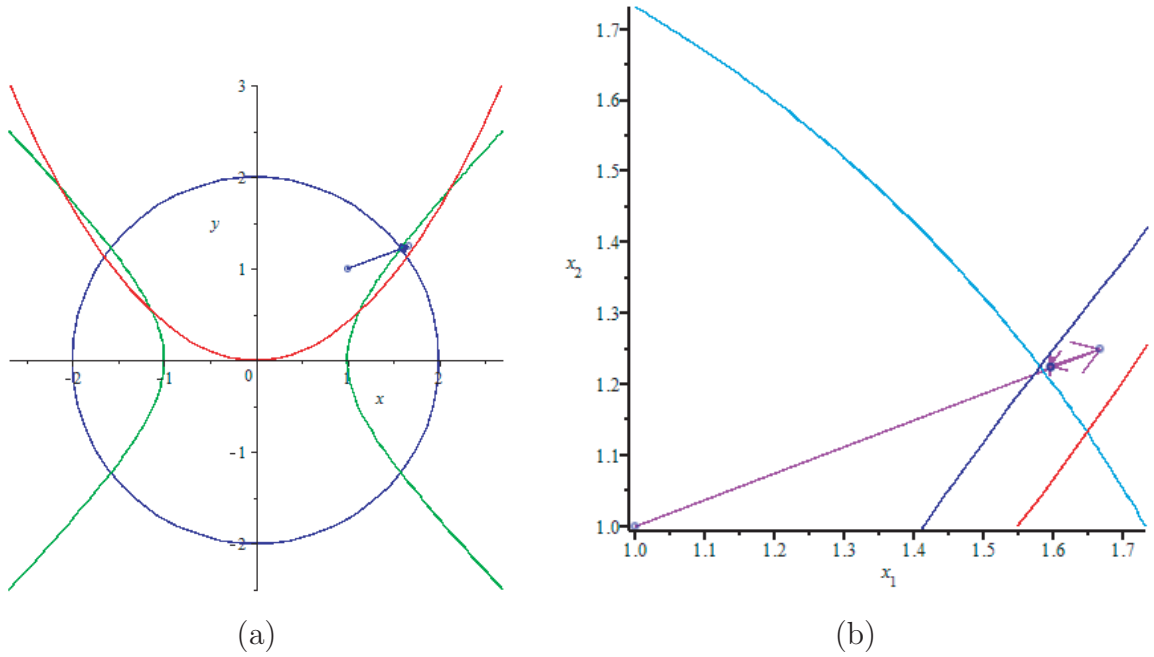


Figure 3.7: Iteration sequence in the Newton-Raphson procedure to find the intersection of a circle and a hyperbola: (a) the big picture; (b) a zoom-in close to the solution

thereby obtaining a determined system of  $n$  equations in  $n$  unknowns. This system can be solved using Newton-Raphson method which requires  $\nabla\psi(\mathbf{x})$ :

$$\nabla\psi(\mathbf{x}) = \frac{\partial\psi}{\partial\mathbf{x}} = \frac{\partial}{\partial\mathbf{x}} \left[ \underbrace{\Phi^T(\mathbf{x})}_{(\partial\phi/\partial\mathbf{x})^T} \mathbf{W}\phi(\mathbf{x}) \right]$$

That is,  $\nabla\psi(\mathbf{x})$  involves second-order derivatives of  $\phi$  with respect to  $\mathbf{x}$ :

$$\frac{\partial^2\phi_i}{\partial x_j\partial x_i}, \quad i = 1, \dots, n$$

In summary, the Newton-Raphson method is too cumbersome, besides being prone to ill-conditioning, for it is based on the normality conditions of the problem at hand.

The solution of nonlinear least-square problems is standard in commercial software. For example, Maple features the `LSSolve` command to find one—of several— $\mathbf{x}$  that best fits the overdetermined system of nonlinear equations (3.97) in the least-square sense.

### 3.8 Computer Implementation Using ODA— C-Library of Routines for Optimum Design

ODA is a C library of subroutines for optimization problems. The source file of this package, implemented in C, consists of a number of subroutines designed and classified based on their application. At the beginning of each subroutine a detailed description of the purpose and usage of the subroutine is included. Moreover, data validation has been considered in the software. In order to solve a problem, the user simply calls one corresponding C subroutine.

Since the solutions for linear problems are *direct*—as opposed to *iterative*—the use of ODA to solve linear problems requires only information on the problem parameters, such as matrices  $\mathbf{A}$ ,  $\mathbf{C}$ , and  $\mathbf{W}$ , as well as vectors  $\mathbf{b}$  and  $\mathbf{d}$ , as applicable. For nonlinear problems, the solution is iterative, and hence, the user is required to provide functions describing  $\phi(\mathbf{x})$ ,  $\mathbf{h}(\mathbf{x})$ ,  $\Phi(\mathbf{x})$  and  $\mathbf{J}(\mathbf{x})$ , as needed. These functions are provided via subroutines in forms that can be called by the package. In addition to this information, the user is also required to provide an initial guess  $\mathbf{x}_0$  of  $\mathbf{x}$ , so that the iterative procedure can be started.

1. **Unconstrained linear problems:** Subroutine MNSLS is used to find the minimum-norm solution of an underdetermined linear system, while subroutine LSSLS is used to find the least-square approximation of an overdetermined linear system. LSSLS can also handle determined systems, i.e., systems of as many equations as unknowns.
2. **Unconstrained nonlinear problems:** Subroutine LSSNLS is used to solve this type of problems. Since the nonlinear functions and their associated gradient matrices are problem-dependent, the user is required to provide two subroutines that are used to evaluate the foregoing items, namely,
  - FUNPHI: This subroutine is used to evaluate the  $q$ -dimensional vector function  $\phi(\mathbf{x})$  in terms of the given  $n$ -dimensional vector  $\mathbf{x}$ .
  - DPHIDX: This subroutine is used to evaluate the  $q \times n$  gradient matrix  $\Phi$  of the vector-function  $\phi(\mathbf{x})$  with respect to  $\mathbf{x}$ , at the current value of  $\mathbf{x}$ . Moreover, an initial guess of  $\mathbf{x}$  is required when calling this subroutine.
3. **Constrained linear problems:** Subroutine LSSCLS is used to solve this type of problems.

4. **Constrained nonlinear problems:** Subroutine `LSSCNL` is used for solving this type of problems. Before calling `LSSCNL`, the user is required to provide four problem-dependent subroutines: Two of these are `FUNPHI` and `DPHIDX`, already described in item 2 above. The other two are used to evaluate the left-hand sides of the constraint equations and their gradient matrix, as listed below:

- **FUNH:** This subroutine is used to evaluate the  $l$ -dimensional constraint function  $\mathbf{h}$  in terms of the given  $n$ -dimensional vector  $\mathbf{x}$ .
- **DHDX:** This subroutine is used to evaluate the  $l \times n$  gradient matrix  $\mathbf{J}$  of the vector-function  $\mathbf{h}(\mathbf{x})$  in terms of the given  $n$ -dimensional vector  $\mathbf{x}$ . Moreover, an initial guess of  $\mathbf{x}$  is required when calling `LSSCNL`.

5. **Constrained problems with arbitrary objective function:** Subroutine `ARBITRARY` is used for solving this type of problems. Before calling `ARBITRARY`, the user is required to provide four problem-dependent subroutines: Two of these are `FUNPHI` and `DPHIDX`, as described in item 2 above. The other two subroutines are used to evaluate the left-hand sides of the constraint equations and their gradient matrix, as listed below:

- **phi:** Subroutine used to evaluate the objective function  $\phi(\mathbf{x})$  in terms of the given  $n$ -dimensional vector  $\mathbf{x}$ .
- **h:** Subroutine used to evaluate the  $l$ -dimensional constraint function  $\mathbf{h}$  in terms of the given  $n$ -dimensional vector  $\mathbf{x}$ .
- **J:** Subroutine used to evaluate the  $l \times n$  gradient matrix  $\mathbf{J}$  of the vector-function  $\mathbf{h}(\mathbf{x})$  at the current value of  $\mathbf{x}$ .
- **gradient:** Subroutine used to evaluate the  $n$ -dimensional gradient  $\nabla f$  of the objective function  $f(\mathbf{x})$  at the current value of vector  $\mathbf{x}$ .
- **Hessian:** Subroutine used to evaluate the  $n \times n$  Hessian matrix  $\nabla \nabla f$  of the objective function  $f(\mathbf{x})$  at the current value of vector  $\mathbf{x}$ . Moreover, an initial guess of  $\mathbf{x}$  is required when calling `ARBITRARY`.

# Chapter 4

## Unconstrained Optimization

### 4.1 Introduction

We start by studying the simplest problem in multivariable optimization, namely, the *unconstrained minimization* of a *smooth* scalar objective function  $f(\mathbf{x})$  of the  $n$ -dimensional *design-variable vector*, or *design vector* (DV) for brevity, that we denote by  $\mathbf{x}$ . The main result here is the *normality conditions* (NC) of the problem at hand. We derive the *first-order normality conditions* (FONC), which are *necessary* for a stationary point (SP); then, we derive the *second-order normality conditions* (SONC), which are *sufficient* for a *minimum*, a *maximum* or a *saddle point*. These three kinds of SP are duly characterized.

### 4.2 The Normality Conditions

Under the smoothness assumption, the objective function is *continuous* and has *continuous first- and second-order derivatives*. The problem at hand is, moreover,

$$f(\mathbf{x}) \rightarrow \min_{\mathbf{x}} \quad (4.1)$$

Since the problem under study is unconstrained, the search of the minimum is conducted over the whole design space  $\mathbb{R}^n$ , which eases the search tremendously. Notice that every point of the design space is characterized by a position vector  $\mathbf{x}$ , which defines a *design*, and hence, every such point represents one design. For conciseness, we will refer to a point and the design that the point represents by its position vector.

Now, for  $f(\mathbf{x})$  to attain a minimum at a certain point  $\mathbf{x}_o$  of the design space, the point must be, first and foremost, *stationary*, i.e., the *gradient*  $\nabla f$  of the objective function with respect to the design vector must vanish:

$$\nabla f \equiv \left. \frac{\partial f}{\partial \mathbf{x}} \right|_{\mathbf{x}_o} = \mathbf{0} \quad (4.2a)$$

which is known as the *first-order normality condition*. As a matter of fact, the above relation is short-hand for  $n$  normality conditions, one for each component of the  $\nabla f$  vector, namely,

$$\nabla f \equiv \frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \partial f / \partial x_1 \\ \partial f / \partial x_2 \\ \vdots \\ \partial f / \partial x_n \end{bmatrix} \quad (4.2b)$$

However, a stationary point can be a *minimum*, a *maximum* or a *saddle point*, to a second-order approximation. To characterize each case, we expand, to this order of approximation,  $f(\mathbf{x})$  around  $\mathbf{x} = \mathbf{x}_o$ :

$$f(\mathbf{x}) = f(\mathbf{x}_o) + (\nabla f|_{\mathbf{x}_o})^T (\mathbf{x} - \mathbf{x}_o) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_o)^T \nabla \nabla f|_{\mathbf{x}_o} (\mathbf{x} - \mathbf{x}_o) + \text{HOT} \quad (4.3a)$$

where HOT stands for “higher-order-terms”, while  $\nabla \nabla f$ , the *Hessian* of  $f$  with respect to  $\mathbf{x}$ , is a matrix of second derivatives, namely,

$$\nabla \nabla f \equiv \frac{\partial^2 f}{\partial \mathbf{x}^2} = \begin{bmatrix} \partial^2 f / \partial x_1^2 & \partial^2 f / \partial x_1 \partial x_2 & \cdots & \partial^2 f / \partial x_1 \partial x_n \\ \partial^2 f / \partial x_2 \partial x_1 & \partial^2 f / \partial x_2^2 & \cdots & \partial^2 f / \partial x_2 \partial x_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2 f / \partial x_n \partial x_1 & \partial^2 f / \partial x_n \partial x_2 & \cdots & \partial^2 f / \partial x_n^2 \end{bmatrix} \quad (4.3b)$$

Notice that, by virtue of the smoothness assumption,

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}, \quad \text{for } i, j = 1, 2, \dots, n \quad (4.4)$$

which follows after *Schwartz's Theorem*: *Given a continuous function  $f(\mathbf{x})$  with first- and second-order continuous derivatives, the order of differentiation in computing the second derivatives is immaterial.*

As a consequence of eq.(4.4), then,

- The Hessian of  $f$  with respect to  $\mathbf{x}$  is a symmetric  $n \times n$  matrix, and
- the eigenvalues of the Hessian matrix are all real and its eigenvectors are mutually orthogonal.



At a stationary point  $\mathbf{x}_o$ , then, and up to a second-order approximation, eq.(4.3a) leads to

$$\Delta f \equiv f(\mathbf{x}) - f(\mathbf{x}_o) \approx \frac{1}{2}(\mathbf{x} - \mathbf{x}_o)^T \nabla \nabla f|_{\mathbf{x}_o} (\mathbf{x} - \mathbf{x}_o) \quad (4.5)$$

Now we have that

- If, for any  $\Delta \mathbf{x} \equiv \mathbf{x} - \mathbf{x}_o$ ,  $\Delta f(\mathbf{x}) > 0$ , then the stationary point (SP)  $\mathbf{x}_o$  is a *local minimum* of  $f(\mathbf{x})$ ;
- if, for any  $\Delta \mathbf{x} \equiv \mathbf{x} - \mathbf{x}_o$ ,  $\Delta f(\mathbf{x}) < 0$ , then the SP  $\mathbf{x}_o$  is a *local maximum* of  $f(\mathbf{x})$ ; and
- otherwise, the SP  $\mathbf{x}_o$  is a *saddle point*.

It is not practical to test a stationary point for the sign of  $\Delta f$  for every possible  $\Delta \mathbf{x}$ . However, it is possible to characterize the nature of the stationary point  $\mathbf{x}_o$  by means of the signs of the eigenvalues of the Hessian matrix. To this end, we recall the characterization of positive-definite, positive-semidefinite and sign-indefinite matrices given above. In this light, then,

- the stationary point  $\mathbf{x}_o$  is a *local minimum* if the Hessian evaluated at this point is positive-definite;
- the SP is a *local maximum* if the Hessian evaluated at this point is negative-definite;
- the SP is a *saddle point* if the Hessian evaluated at this point is sign-indefinite.

## 4.3 Methods of Solution

The variety of methods available is immensely rich. In a nutshell, the various methods can be classified according to one criterion: the requirement of partial derivatives of the objective function. We thus have:

- (i) **Direct methods:** No derivatives are required.
- (ii) **Gradient Methods:** Only first-order derivatives of the objective function with respect to all design variables are required.
- (iii) **Newton methods:** First- and second-order derivatives of the objective function with respect to all all design variables are required.

Needless to say, direct methods are the most general—the least demanding—and simplest to implement, the price to be paid for the lack of information on derivatives being the speed of convergence. These methods are the slowest to converge. Gradient methods are faster, with a *linear convergence rate*, which means that the error between the current *iterate*  $\mathbf{x}^k$  and *the closest local minimum* decreases by one order of magnitude at each iteration. The Newton-Raphson method, and variations thereof, generically termed *Newton methods*, resort to first- and second-order derivatives. Newton methods converge *quadratically*, which means that the aforementioned error decreases by two orders of magnitude per iteration.

## 4.4 Direct Methods

Direct methods are based on function evaluations and nothing else. While these methods are slow to converge, they can handle discontinuous functions. There are various of these: random (random jumps, random walks); Hooke and Jeeves; Powell; and the simplex (Nelder-Mead) method. We outline the last three of these methods below.

The main concept behind direct methods is the *pattern directions*, namely, the directions of search, along which the minimum is approached. These three methods differ on the way of defining the pattern directions.

### 4.4.1 The Hooke and Jeeves Method

In this method, the search directions are fixed. The method thus starts by defining a set of unit vectors  $\{\mathbf{u}_i\}_1^n$  in the directions of the  $n$  design variables, along which the search is conducted.

#### Hooke & Jeeves Algorithm

1. Define starting base point  $\mathbf{x}^1$  and prescribed length  $\Delta x$  of search step
2.  $f_k \leftarrow f(\mathbf{x}^k)$ ;  $1 \leftarrow i$ ;  $\mathbf{y}^{k,0} \leftarrow \mathbf{x}^k$ , where  $\mathbf{y}^{k,j}$  denotes a temporary base point obtained from  $\mathbf{x}^k$  upon perturbing  $(\mathbf{x}^k)_j$
3. For  $i = 1$  to  $n$  do
 

$$f^+ \leftarrow f(\mathbf{y}^{k,i-1} + \mathbf{u}_i \Delta x_i)$$

$$f^- \leftarrow f(\mathbf{y}^{k,i-1} - \mathbf{u}_i \Delta x_i)$$

```

    if  $f^+ < f(\mathbf{y}^{k,i-1})$  then  $\mathbf{y}^{k,i} \leftarrow \mathbf{y}^{k,i-1} + \mathbf{u}_i \Delta x_i$ 
    if  $f^- < f(\mathbf{y}^{k,i-1})$  then  $\mathbf{y}^{k,i} \leftarrow \mathbf{y}^{k,i-1} - \mathbf{u}_i \Delta x_i$ 
    else  $\mathbf{y}^{k,i} \leftarrow \mathbf{y}^{k,i-1}$ 
enddo

4.   if  $\mathbf{y}^{k,n} = \mathbf{x}^k$  then  $\Delta \mathbf{x} \leftarrow \frac{1}{2} \Delta \mathbf{x}$  go to 3
    else  $\mathbf{x}^{k+1} \leftarrow \mathbf{y}^{k,n}$ 

5.    $\mathbf{s} \leftarrow \mathbf{x}^{k+1} - \mathbf{x}^k$  % pattern direction
    Find  $\lambda$  that minimizes  $f(\mathbf{x}^{k+1} + \lambda \mathbf{s})$ 
     $\mathbf{y}^{k+1,0} \leftarrow \mathbf{x}^{k+1} + \lambda \mathbf{s}$ 

6.    $k \leftarrow k + 1$ ,  $f_k \leftarrow f(\mathbf{y}^{k,0})$ ,  $1 \leftarrow i$ 
    go to 3
    if  $f(\mathbf{y}^{k,n}) < f(\mathbf{x}^k)$  then  $\mathbf{x}^{k+1} \leftarrow \mathbf{y}^{k,n}$ 
    else  $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k$ ;  $\Delta \mathbf{x} \leftarrow \frac{1}{2} \Delta \mathbf{x}$  go to 2

7.   if  $\|\Delta \mathbf{x}\| < \epsilon$  then stop
    else go to 2

```

#### 4.4.2 The Powell Method (Conjugate Directions)

Let  $\mathbf{A} = \mathbf{A}^T \in \mathbb{R}^{n \times n}$  be positive-definite. Vectors  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  are said to be  $\mathbf{A}$ -conjugate if

$$\mathbf{x}^T \mathbf{A} \mathbf{y} = \mathbf{y}^T \mathbf{A} \mathbf{x} = 0 \quad (4.6)$$

In the area of mechanical systems under linear vibrations, for  $n$  degrees of freedom, the  $i$ th and  $j$ th modal vectors  $\mathbf{u}^i$  and  $\mathbf{u}^j$  are  $\mathbf{K}$ - and  $\mathbf{M}$ -conjugate, i.e.,

$$(\mathbf{u}^i)^T \mathbf{K} \mathbf{u}^j = 0, \quad \text{and} \quad (\mathbf{u}^i)^T \mathbf{M} \mathbf{u}^j = 0$$

where  $\mathbf{K}$  and  $\mathbf{M}$  are the  $n \times n$  positive-definite stiffness and mass matrices, respectively.

**Theorem 4.4.1** (cf. Theorem 6.1 of (Rao, 1996)) *Let  $\mathbf{x} \in \mathbb{R}^n$  and*

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c, \quad \mathbf{A} = \mathbf{A}^T > \mathbf{O} \quad (4.7)$$

Further, let  $\Pi_1$  and  $\Pi_2$  be two hyperplanes in  $\mathbb{R}^n$  parallel to each other. If  $\mathbf{x}^i$  is the minimum of  $f(\mathbf{x})$ , with  $\mathbf{x}^i \in \Pi_i$ ,  $i = 1, 2$ , then  $\mathbf{x}^2 - \mathbf{x}^1$  is  $\mathbf{A}$ -conjugate with any vector  $\mathbf{y} \in \mathbb{R}^n$  parallel to  $\Pi_1$  and  $\Pi_2$ , i.e.,

$$(\mathbf{x}^2 - \mathbf{x}^1)^T \mathbf{A} \mathbf{y} = 0$$

*Proof:*  $\Pi_1$  and  $\Pi_2$  are defined by

$$\Pi_i : \mathbf{C} \mathbf{x} = \mathbf{d}^i, \quad i = 1, 2; \quad \mathbf{C} \in \mathbb{R}^{p \times n}, \quad p < n \quad (4.8)$$

The algebraic interpretation of  $\mathbf{y}$  parallel to  $\Pi_1$  and  $\Pi_2$  is that  $\mathbf{y} \in \mathcal{N}(\mathbf{C})$  and hence, if  $\mathbf{L}$  is an *orthogonal complement* of  $\mathbf{C}$ , i.e., if

$$\underbrace{\mathbf{C}}_{p \times n} \underbrace{\mathbf{L}}_{n \times n'} = \underbrace{\mathbf{O}}_{p \times n'}, \quad n' \equiv n - p$$

then we can write

$$\mathbf{y} = \mathbf{L} \mathbf{u}, \quad \mathbf{u} \in \mathbb{R}^{n-p}$$

Now,  $\mathbf{x}^i$  is found as the solution to

$$\min_{\mathbf{x}} f(\mathbf{x})$$

subject to eq.(4.8). We solve the foregoing problem as an unconstrained problem by means of the Lagrangian<sup>1</sup>

$$F_i \equiv f(\mathbf{x}) + (\boldsymbol{\lambda}^i)^T (\mathbf{C} \mathbf{x} - \mathbf{d}^i) \rightarrow \min_{\mathbf{x}, \boldsymbol{\lambda}^i}, \quad i = 1, 2$$

subject to no constraints, and denote the solution  $\mathbf{x}^i$ . The normality conditions of the foregoing problem are, for  $i = 1, 2$ ,

$$\nabla F_i = \mathbf{A} \mathbf{x} + \mathbf{b} + \mathbf{C}^T \boldsymbol{\lambda}^i = \mathbf{0}_n, \quad (4.9a)$$

$$\mathbf{C} \mathbf{x} - \mathbf{d}^i = \mathbf{0}_p \quad (4.9b)$$

which yield a system of  $p + n$  equations for the  $p + n$  unknowns  $\mathbf{x}$  and  $\boldsymbol{\lambda}^i$ . Upon solving for  $\mathbf{x} = \mathbf{x}^i$  from eq.(4.9a), for  $i = 1, 2$ , we obtain

$$\mathbf{x}^i = -\mathbf{A}^{-1}(\mathbf{C}^T \boldsymbol{\lambda}^i + \mathbf{b}) \quad (4.10)$$

---

<sup>1</sup>Constrained optimization is the subject of Chapter 5.

Substitution of the foregoing expression into eq.(4.8) yields, always for  $i = 1, 2$ ,

$$-\mathbf{CA}^{-1}(\mathbf{C}^T \boldsymbol{\lambda}^i + \mathbf{b}) = \mathbf{d}^i$$

or

$$\mathbf{CA}^{-1}\mathbf{C}^T \boldsymbol{\lambda}^i = -\mathbf{CA}^{-1}\mathbf{b} - \mathbf{d}^i$$

Hence,

$$\boldsymbol{\lambda}^i = -(\mathbf{CA}^{-1}\mathbf{C}^T)^{-1}(\mathbf{CA}^{-1}\mathbf{b} + \mathbf{d}^i) \quad (4.11)$$

Further, substitution of eq.(4.11) into eq.(4.10) leads to

$$\begin{aligned} \mathbf{x}^i &= \mathbf{A}^{-1}[\mathbf{C}^T(\mathbf{CA}^{-1}\mathbf{C}^T)^{-1}(\mathbf{CA}^{-1}\mathbf{b} + \mathbf{d}^i) + \mathbf{b}] \\ &= \mathbf{A}^{-1}[\mathbf{C}^T(\mathbf{CA}^{-1}\mathbf{C}^T)^{-1}\mathbf{CA}^{-1} + \mathbf{1}]\mathbf{b} + \mathbf{A}^{-1}[\mathbf{C}^T(\mathbf{CA}^{-1}\mathbf{C}^T)^{-1}\mathbf{d}^i] \end{aligned}$$

Therefore,

$$\mathbf{x}^2 - \mathbf{x}^1 = \mathbf{A}^{-1}\mathbf{C}^T(\mathbf{CA}^{-1}\mathbf{C}^T)^{-1}(\mathbf{d}^2 - \mathbf{d}^1)$$

Hence,

$$(\mathbf{x}^2 - \mathbf{x}^1)^T \mathbf{A} \mathbf{y} \equiv (\mathbf{x}^2 - \mathbf{x}^1)^T \mathbf{A} \mathbf{L} \mathbf{u} = (\mathbf{d}^2 - \mathbf{d}^1)^T (\mathbf{CA}^{-1}\mathbf{C}^T)^{-1} \underbrace{\mathbf{C} \mathbf{A}^{-1} \mathbf{A} \mathbf{L}}_{\substack{1 \\ \mathbf{CL}=\mathbf{O}_{pn'}}} \mathbf{u}$$

That is,

$$(\mathbf{x}^2 - \mathbf{x}^1)^T \mathbf{A} \mathbf{y} = 0$$

thereby completing the proof.

### The Powell Algorithm

One iteration of the basic procedure of the Powell algorithm (Powell, 1964) is summarized below:

1. For  $i = 1, 2, \dots, n$ , calculate  $\lambda_i$  that minimizes  $f(\mathbf{x}_{i-1} + \lambda_i \boldsymbol{\xi}_i)$ ; then  $\mathbf{x}_i \leftarrow \mathbf{x}_{i-1} + \lambda_i \boldsymbol{\xi}_i$ .
2. For  $i = 1, 2, \dots, n - 1$ , replace  $\boldsymbol{\xi}_i$  by  $\boldsymbol{\xi}_{i+1}$ .
3. Replace  $\boldsymbol{\xi}_n$  by  $\mathbf{x}_n - \mathbf{x}_0$ .
4. Find  $\lambda$  that minimizes  $f(\mathbf{x}_n + \lambda(\mathbf{x}_n - \mathbf{x}_0))$ , and replace  $\mathbf{x}_0$  by  $\mathbf{x}_0 + \lambda(\mathbf{x}_n - \mathbf{x}_0)$ .

Powell includes, in the same paper, an improvement of this method to accelerate its convergence. The improved algorithm follows:

Data: A set of linearly independent directions  $\mathcal{D} = \{\xi^i\}_1^n$ , which, in the absence of any background information, can be taken as the coordinate axes, and an initial guess  $\mathbf{x}^0$

1. For  $k = 1$  to  $n$  do  
    find minimizer  $\lambda_k$  of  $f(\mathbf{x}^{k-1} + \lambda_k \xi^k)$   
     $\mathbf{x}^k \leftarrow \mathbf{x}^{k-1} + \lambda_k \xi^k$   
  enddo
2. Find  $m \in \{1, \dots, n\}$  such that  $f(\mathbf{x}^{m-1}) - f(\mathbf{x}^m)$  is a maximum; then  
 $\Delta \leftarrow f(\mathbf{x}^{m-1}) - f(\mathbf{x}^m)$
3.  $f_1 \leftarrow f(\mathbf{x}^0)$ ,  $f_2 \leftarrow f(\mathbf{x}^n)$ ,  $f_3 \leftarrow f(2\mathbf{x}^n - \mathbf{x}^0)$
4. if  $\begin{cases} f_3 \geq f_1, \text{ or} \\ (f_1 - 2f_2 + f_3)(f_1 - f_2)^2 \geq \frac{1}{2}(f_1 - f_3)^2 \Delta \end{cases}$   
    then keep  $\mathcal{D}$ ; else
5.  $\xi \leftarrow \mathbf{x}^n - \mathbf{x}^0$   
    find  $\lambda$  that minimizes  $f(\mathbf{x}^n + \lambda \xi)$   
 $\mathcal{D} \leftarrow \{\xi^1, \dots, \xi^{m-1}, \xi^{m+1}, \dots, \xi^n, \xi\}$   
 $\mathbf{x}^0 \leftarrow \mathbf{x}^n + \lambda \xi$   
    if  $\mathbf{x}^0$  is a minimum, stop; else, go to 1.

We shall prove presently that, if the function to be minimized is *quadratic* in its  $n$  arguments, then the Powell Algorithm converges to the minimum in *at most*  $n$  steps. In the sequel, we shall need a previous result:

**Lemma 4.4.1** *Let  $\mathbf{y} = \mathbf{\Lambda} \mathbf{x}$  denote a linear transformation, with  $\mathbf{\Lambda} \in \mathbb{R}^{n \times n}$  nonsingular, and  $Q(\mathbf{x})$  a quadratic function of  $\mathbf{x} \in \mathbb{R}^n$ , of the form of eq.(4.7), with  $\mathbf{A}$  positive-definite. Under the foregoing transformation,  $Q = Q(\mathbf{y})$ , its minimum being found at  $\mathbf{y}_{\text{opt}} = \mathbf{\Lambda} \mathbf{x}_{\text{opt}}$ , with  $\mathbf{x}_{\text{opt}}$  denoting the minimizer of  $Q(\mathbf{x})$ .*

*Proof:* The minimizer of  $Q(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c$  is found upon zeroing the gradient of  $Q$  with respect to  $\mathbf{x}$ :

$$\nabla Q = \mathbf{A} \mathbf{x} + \mathbf{b} \Rightarrow \mathbf{A} \mathbf{x} + \mathbf{b} = \mathbf{0} \quad (4.12)$$

whence,

$$\mathbf{x}_{\text{opt}} = -\mathbf{A}^{-1} \mathbf{b} \quad (4.13)$$

On the other hand,

$$Q(\mathbf{y}) = \frac{1}{2} \mathbf{y}^T \mathbf{\Lambda}^{-T} \mathbf{A} \mathbf{\Lambda}^{-1} \mathbf{y} + \mathbf{b}^T \mathbf{\Lambda}^{-1} \mathbf{y} + c$$

whence,

$$\nabla Q(\mathbf{y}) = \Lambda^{-T} \mathbf{A} \Lambda^{-1} \mathbf{y} + \Lambda^{-T} \mathbf{b}$$

which, upon zeroing, yields

$$\mathbf{y}_{\text{opt}} = -\Lambda \mathbf{A}^{-1} \mathbf{b} = \Lambda \mathbf{x}_{\text{opt}}$$

thereby completing the proof. Now we have

**Theorem 4.4.2** *If the quadratic function  $Q(\mathbf{x})$ , of the form (4.7), with  $\mathbf{A}$  positive-definite is minimized sequentially, along each direction of a set  $\mathcal{D} = \{\boldsymbol{\xi}^i\}_1^n$  of  $\mathbf{A}$ -conjugate directions, then the minimum of  $Q(\mathbf{x})$  will be found in at most  $n$  steps, irrespective of the initial guess.*

*Proof:* Since  $\mathbf{A}$  is positive-definite, it admits a real Cholesky decomposition, namely,

$$\mathbf{A} = \mathbf{B}^T \mathbf{B} \quad (4.14)$$

Alternatively,  $\mathbf{A}$  can be factored into its two real square roots,  $\mathbf{A} = \sqrt{\mathbf{A}} \sqrt{\mathbf{A}}$ —an arbitrary  $n \times n$  matrix has  $2^n$  square roots, one of which is positive-definite if the matrix is so. Now let us introduce the linear transformation

$$\mathbf{y} = \mathbf{B} \mathbf{x} \quad (4.15)$$

under which  $Q$  becomes

$$Q(\mathbf{y}) = \frac{1}{2} \|\mathbf{y}\|^2 + (\mathbf{B}^{-T} \mathbf{b})^T \mathbf{y} + c \quad (4.16)$$

which is a quadratic form associated with the  $n \times n$  identity matrix. Hence, a set of  $\mathbf{1}$ -conjugate directions is obviously the set  $\mathcal{E} = \{\mathbf{e}_i\}_1^n$ , with  $\mathbf{e}_i$  denoting the unit vector in the  $y_i$ -direction, i.e., a  $n$ -dimensional array with zeros everywhere, except for the  $i$ th entry, which is unity.  $\mathbf{1}$ -conjugacy can thus be readily verified, as

$$\mathbf{e}_i^T \mathbf{e}_j = 0, \quad i \neq j$$

We shall prove the theorem for  $n = 3$ . For  $n > 3$ , the proof follows the same pattern, although the pattern cannot be visualized. The set of vectors  $\mathbf{y}$  verifying

$$Q(\mathbf{y}) = Q_0 = \text{const}$$

defines a sphere with centre  $C$  of position vector  $\mathbf{c}$ , not to be confused with the scalar  $c$  of eq.(4.16) and radius  $r$ , to be determined presently. Vector  $\mathbf{c}$  and scalar  $r$  thus verify

$$\|\mathbf{y} - \mathbf{c}\|^2 = r^2$$

Upon comparing the above expression with its counterpart of eq.(4.16), it is apparent that

$$\mathbf{c} = -\mathbf{B}^{-T}\mathbf{b} \quad \& \quad r^2 = 2(Q_0 - c) + \|\mathbf{c}\|^2$$

Under a search along the set  $\mathcal{E}$  of conjugate directions, let  $\mathbf{y}^0$  be the initial guess, the position vector of point  $P_0$ . The first search direction  $\mathbf{e}_1$  thus takes place along the  $y_1$ -axis. Along this direction,  $Q(\mathbf{y})$  reaches a minimum at a point  $P_1$ , of position vector  $\mathbf{y}^1$ , such that  $(\mathbf{y}^1 - \mathbf{c})^T \mathbf{e}_1 = 0$ . That is,  $P_1$  lies on a line  $\mathcal{L}_1$  passing through  $P_0$  and parallel to the  $y_1$ -axis.  $P_1$  is found as the point of  $\mathcal{L}_1$  tangent to a sphere of centre  $C$ .

The second search takes places along a line  $\mathcal{L}_2$  passing through  $P_1$  and parallel to the  $y_2$ -axis. Along this line, a point  $P_2$  is determined, at which  $\mathcal{L}_2$  is tangent to a sphere centred at  $C$ . Notice that  $\mathcal{L}_1$  and  $\mathcal{L}_2$  define a plane  $\Pi_{12}$  parallel to the  $y_1$ - $y_2$  plane.

The third search thus takes place along a line  $\mathcal{L}_3$  passing through  $P_2$  and parallel to the  $y_3$ -axis. A sketch should help the reader visualize that the orthogonal projection of point  $C$  onto  $\Pi_{12}$ , labelled  $C'$ , is exactly point  $P_2$ . Consequently,  $P_3$ , the point at which  $\mathcal{L}_3$  is tangent to a sphere centred at  $C$  is exactly  $C$ , the minimum of  $Q(\mathbf{y})$  finding itself at  $C$ , thereby completing the search in exactly  $n = 3$  steps. Apparently, the optimum  $\mathbf{x}_{\text{opt}}$  is given by

$$\mathbf{x}_{\text{opt}} = \mathbf{B}^{-1}\mathbf{c} \tag{4.17}$$

**Exercise 4.4.1 (cf. Example 6.6 of (Rao, 1996))** *Using the Powell Algorithm, find the minimum of*

$$Q(\mathbf{x}) = 6x_1^2 + 2x_2^2 - 6x_1x_2 - x_1 - 2x_2$$

*To this end, transform the above quadratic form into one associated with the  $2 \times 2$  identity matrix. Hint: For  $2 \times 2$  positive-definite matrices, their square root can be found graphically using the Mohr circle, as explained in Angeles (2011).*

### 4.4.3 The Nelder-Mead Simplex Method

With a few changes in the notation, this subsection is taken from (Rao, 1996). This method, first proposed by Spendley et al. (1962), and later improved by Nelder and Mead (1965), is based on the concept of *simplex*  $\mathcal{S}_n$ . A simplex is a  $(n + 1)$ -vertex hyperpolyhedron in  $\mathbb{R}^n$ . The search for the minimum of the objective function  $f(\mathbf{x})$



is conducted by means of function evaluations at all  $n + 1$  vertices of the simplex. The strategy followed is outlined below.

We start by defining an *initial simplex*, which is done by means of a *base point*  $P_0 \in \mathbb{R}^n$ , of position vector  $\mathbf{x}_0$ . The remaining  $n$  vertices of  $\mathcal{S}_n$  are generated so as to yield a *regular hyperpolyhedron* of unit-length edges. To this end, let

$$p = \frac{1}{\sqrt{2}n}(\sqrt{n+1} + n - 1), \quad q = \frac{1}{\sqrt{2}n}(\sqrt{n+1} - 1) \quad (4.18a)$$

Then, if  $\mathbf{e}_i$  denotes the unit vector in the direction of the  $i$ th coordinate axis, corresponding to  $x_i$ , let

$$\mathbf{x}_i = \mathbf{x}_0 + p\mathbf{e}_i + \sum_{j=1, j \neq i}^n q\mathbf{e}_j, \quad i = 1, 2, \dots, n \quad (4.18b)$$

Shown in Fig. 4.1 are displays of the simplexes in  $\mathbb{R}^2$  and  $\mathbb{R}^3$ , respectively, defined as described in eqs.(4.18a & 4.18b), with the base point *at the origin*.

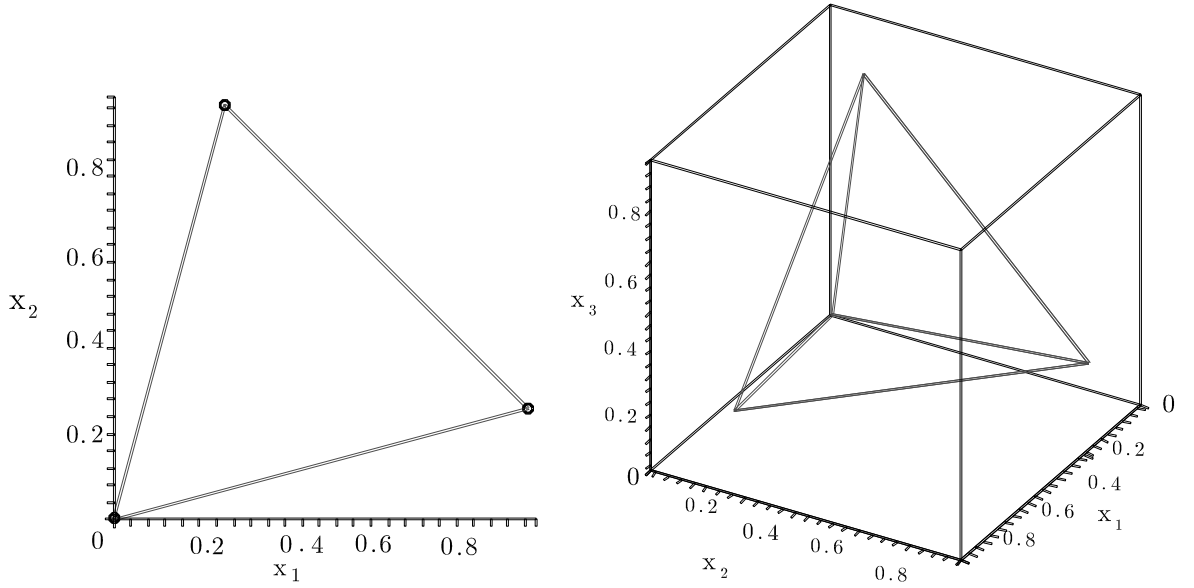


Figure 4.1: The initial simplex in  $\mathbb{R}^2$

The initial simplex in  $\mathbb{R}^3$

The search strategy is based on three operations: a) *reflection*; (b) *contraction*; and (c) *expansion*.

## Reflection

Let  $f_i = f(\mathbf{x}_i)$ , for  $i = 0, 1, \dots, n$ , and

$$f_M = \max_i \{ f_i \}_0^n, \quad f_m = \min_i \{ f_i \}_0^n \quad (4.19)$$

the corresponding vertices being  $P_M$  and  $P_m$ , of position vectors  $\mathbf{x}_M$  and  $\mathbf{x}_m$ , respectively. With the foregoing information, we now seek a new simplex, by replacing *the worst* vertex  $P_M$  of the current simplex by a new one,  $P_{n+1}$ , of position vector  $\mathbf{x}_{n+1}$ . The new vertex is found by means of a reflection of  $P_M$  about the centroid  $\overline{P}$ , of position vector  $\overline{\mathbf{x}}$ , of all the simplex vertices, except for  $P_M$ , namely,

$$\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=0, i \neq M}^n \mathbf{x}_i \quad (4.20)$$

Let, moreover,  $\alpha > 0$  be the user-prescribed *reflection coefficient*, which is used to define the new vertex a distance  $\alpha \|\overline{\mathbf{x}} - \mathbf{x}_M\|$  from the centroid  $\overline{P}$ , the position vector  $\mathbf{x}_{n+1}$  of the new vertex thus being

$$\mathbf{x}_{n+1} = \overline{\mathbf{x}} + \alpha(\overline{\mathbf{x}} - \mathbf{x}_M) \quad (4.21)$$

The intended effect is that  $P_{n+1}$  lie *the farthest* from the worst vertex  $P_M$ , so that this new vertex,  $P_{n+1}$ , will *very likely* be *the best* of all the vertices of the new simplex,  $\{P_i\}_{i=0, i \neq M}^{n+1}$ . An unlikely, although quite possible scenario, is that  $f_{n+1} = f_M$ , and hence, no gain will be made by defining the new simplex. In this case, we can define the new vertex in one of two possible ways:

1.  $P_M$  is preserved, the rejected vertex being  $P_{M'}$ , which is the next worst vertex, i.e., with the subscript  $M'$  defined such that

$$f_{M'} = \max_i \{f_i\}_{i=0, i \neq M}^n \quad (4.22)$$

2. Alternatively, redefine  $\alpha$ —make it either larger or smaller—while rejecting always the same worst vertex  $P_M$ .

## Expansion

If the outcome of the reflection stage yields  $f_{n+1} \equiv f(\mathbf{x}_{n+1}) < f_m$ , then the direction defined by the vector difference  $\mathbf{x}_{n+1} - \overline{\mathbf{x}}$  is very likely to point towards the minimum, and hence, it may be advisable to place a new vertex  $P_e$  away from  $\overline{P}$  in the said direction, i.e., by defining the position vector  $\mathbf{x}_e$  of the new vertex  $P_e$  in the form

$$\mathbf{x}_e = \overline{\mathbf{x}} + \gamma(\mathbf{x}_{n+1} - \overline{\mathbf{x}}) \quad (4.23)$$

with  $\gamma > 1$ , for expansion.

Now, there are two possible outcomes:

- If  $f_e \equiv f(\mathbf{x}_e) < f(\mathbf{x}_m)$ , then replace  $P_M$  by  $P_e$  and start a new reflection;
- if  $f_e > f(\mathbf{x}_m)$ , then the expansion failed, and the new simplex obtained by the reflection, with  $P_M$  replaced by  $P_{n+1}$ , is kept, and a new reflection is started.

An alternative outcome of the reflection is described below.

### Contraction

We have two possible outcomes of the reflection: (i)  $f_{n+1} > f_i \equiv f(\mathbf{x}_i)$ , for  $i = 0, 1, \dots, M-1, M+1, \dots, n$ , but  $f_{n+1} < f_M$ , and (ii)  $f_{n+1} > f_M$ . In case (i), the reflection failed to give a direction towards the minimum, but a slightly improved simplex is obtained upon replacing  $P_M$  with  $P_{n+1}$ . In this case, we *contract* the simplex by finding a new vertex  $P_c$ , of position vector  $\mathbf{x}_c$ , a distance  $\beta\|\mathbf{x}_M - \bar{\mathbf{x}}\|$  from  $\bar{P}$ , with  $0 \leq \beta \leq 1$ :

$$\mathbf{x}_c = \bar{\mathbf{x}} + \beta(\mathbf{x}_M - \bar{\mathbf{x}}) \quad (4.24)$$

which can be readily proven to yield a  $P_c$  lying between  $P_M$  and  $\bar{P}$ , for  $\mathbf{x}_c$  has been defined as a *convex combination*<sup>2</sup> of  $\mathbf{x}_c$  and  $\bar{\mathbf{x}}$ . In case (ii),  $P_M$  is kept and proceed according with one of two outcomes, as described below:

- if  $f_c \equiv f(\mathbf{x}_c) < \min\{f_M, f_{n+1}\}$ , then  $P_M$  is replaced by  $P_c$ , a new simplex thus being obtained, and a new reflection operation is started;
- if  $f_c > \min\{f_M, f_{n+1}\}$ , then the contraction failed, in which case  $P_i$  is replaced by  $\bar{P}_i$ , of position vector  $\bar{\mathbf{x}}_i$ , halfway between  $P_i$  and  $P_m$ , i.e.,

$$\bar{\mathbf{x}}_i = \frac{1}{2}(\mathbf{x}_i + \mathbf{x}_m), \quad i = 0, 1, \dots, n, \quad i \neq m \quad (4.25)$$

thereby defining a new simplex, and a new reflection is started.

### Convergence criterion

The method converges when the rms value  $f_{rms}$  of the objective function is smaller than a prescribed tolerance  $\epsilon$ , i.e., when

$$f_{rms} \equiv \sqrt{\frac{1}{n+1} \sum_1^{n+1} (f_i - \bar{f})^2} < \epsilon \quad (4.26)$$

---

<sup>2</sup>See Section 5.4 for a definition of this term.

where (i) a relabelling of the vertices has been assumed, with the order  $f_1 \leq f_2 \leq \dots \leq f_{n+1}$  and (ii)  $\bar{f}$  is the mean value of the objective function evaluated at all the vertices of the current complex, i.e.,

$$\bar{f} \equiv \frac{1}{n+1} \sum_{i=1}^{n+1} f_i \quad (4.27)$$

## 4.5 Gradient Methods

### 4.5.1 The Method of Steepest Descent(Cauchy)

**Algorithm:**

1. Pick up an *initial guess*  $\mathbf{x}^0$  to start the iterations. Set the iteration counter  $i$  at  $i = 0$
2. Define the *i*th *search direction*  $\mathbf{s}_i$  as

$$\mathbf{s}^i = -\nabla f|_{\mathbf{x}=\mathbf{x}^i} \quad (4.28)$$

3. Define the next test point,  $\mathbf{x}^{i+1}$ , as

$$\mathbf{x}^{i+1} = \mathbf{x}^i + \lambda \mathbf{s}^i = \mathbf{x}^i - \lambda \nabla f|_{\mathbf{x}=\mathbf{x}^i} \quad (4.29)$$

To find  $\lambda$ , conduct a *one-dimensional search* along the direction  $\mathbf{s}^i$  so that  $\lambda_{\text{opt}}$  is the value of  $\lambda$  that minimizes  $F(\lambda) = f(\mathbf{x}^i - \lambda \nabla f|_{\mathbf{x}=\mathbf{x}^i})$

4. If  $\mathbf{x}^{i+1}$  satisfies the *convergence criteria* adopted at the outset, stop; else, go to step 5.
5. Update the iteration counter:  $i+1 \leftarrow i$ . Go to step 2.

**Convergence criteria:** Use one or more of those applicable, namely,

$$|f(\mathbf{x}^{i+1}) - f(\mathbf{x}^i)| \leq \epsilon_1 |f(\mathbf{x}^i)| \quad (4.30a)$$

$$\|\nabla f|_{\mathbf{x}=\mathbf{x}^i}\| \leq \epsilon_2 \quad (4.30b)$$

$$\|\mathbf{x}^{i+1} - \mathbf{x}^i\| \leq \epsilon_3 \quad (4.30c)$$

**Remark:** In criteria (4.30b & c), *any* norm can be used  $\Rightarrow$  Use the most economic one, i.e., the Chebyshev or maximum norm.

### 4.5.2 The Conjugate-Gradient Method (Fletcher-Reeves)

We use here the concept of  $\mathbf{A}$ -conjugacy introduced in Subsection 4.4.2, for a symmetric, positive-definite  $\mathbf{A}$ . **Preliminary Remarks:**

- The conjugate-gradient method of Fletcher and Reeves (1964), the FR method, is aimed at minimizing a  $C^2$ -continuous function  $f(\mathbf{x})$  under no constraints.
- The FR method works on the concept of *sequential quadratic programming* (SQP).
- The FR method is based on the *quadratic approximation* of  $f(\mathbf{x})$ : It is assumed that

$$f(\mathbf{x}) \approx f_0 + (\nabla f)^T \mathbf{x} + \text{HOT} = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + f_0 + \text{HOT} \quad (4.31)$$

where HOT stands for *higher-order terms*

- It is assumed that  $\mathbf{A}$  in eq.(4.31) is *positive-definite*

**Algorithm Overview:** The FR algorithm works on the base of two items:

- A search direction  $\mathbf{s}^i$  at each iteration, and
- a step of optimum length  $\lambda^*$  in direction  $\mathbf{s}^i$  that minimizes  $f$  in that direction

Begin with an initial guess  $\mathbf{x}^0$

Next, a *search direction*  $\mathbf{s}^0$  is defined in the direction of *steepest descent* of  $f(\mathbf{x})$ , i.e., along  $-\nabla f$  at  $\mathbf{x}^0$ , namely,

$$\mathbf{s}^0 \equiv -\nabla f|_{\mathbf{x}=\mathbf{x}^0} \quad (4.32)$$

Further, a new iterate  $\mathbf{x}^1$  is sought along the above direction, from the current iterate  $\mathbf{x}^0$ :

$$\mathbf{x}^1 \equiv \mathbf{x}^0 + \lambda \mathbf{s}^0 \quad (4.33)$$

where  $\lambda$  is a real number, as yet to be determined. This is done by imposing that  $\nabla f$ , when evaluated at  $\mathbf{x} = \mathbf{x}^1$ , be *normal* to  $\mathbf{s}^0$ , i.e.,

$$(\nabla f|_{\mathbf{x}=\mathbf{x}^1})^T \mathbf{s}^0 = 0 \quad (4.34)$$

Now substitute  $\mathbf{x}^1$  as given by eq.(4.33) and the quadratic approximation of  $f(\mathbf{x})$  given in eq.(4.31) into eq.(4.34), to obtain, with  $\nabla f \approx \mathbf{A}\mathbf{x} + \mathbf{b}$ ,

$$[\mathbf{A}(\underbrace{\mathbf{x}^0 + \lambda \mathbf{s}^0}_{\mathbf{x}^1}) + \mathbf{b}]^T \mathbf{s}^0 = 0 \quad (4.35)$$

whence the optimum value of  $\lambda$ ,  $\lambda^*$ , is readily derived<sup>3</sup>:

$$\lambda^* = \frac{-(\mathbf{A}\mathbf{x}^0 + \mathbf{b})^T \mathbf{s}^0}{(\mathbf{s}^0)^T \mathbf{A}\mathbf{s}^0} \equiv -\frac{(\mathbf{s}^0)^T \nabla f|_{\mathbf{x}=\mathbf{x}^0}}{(\mathbf{s}^0)^T \mathbf{A}\mathbf{s}^0} \quad (4.36)$$

where, from the assumed positive-definiteness of  $\mathbf{A}$ ,

$$(\mathbf{s}^0)^T \mathbf{A}\mathbf{s}^0 > 0$$

and hence,  $\mathbf{s}^0$  can be expressed, upon recalling eq.(4.33), as

$$\mathbf{s}^0 = \frac{1}{\lambda^*}(\mathbf{x}^1 - \mathbf{x}^0) \quad (4.37)$$

The new search direction,  $\mathbf{s}^1$ , is defined as a *linear combination* of  $\mathbf{s}^0$  and  $-\nabla f|_{\mathbf{x}=\mathbf{x}^1}$ , i.e.,

$$\mathbf{s}^1 \equiv -\nabla f|_{\mathbf{x}=\mathbf{x}^1} + \beta_1 \mathbf{s}^0 \quad (4.38)$$

where  $\beta_1$  is chosen so as to make  $\mathbf{s}^1$  conjugate to  $\mathbf{s}^0$  with respect to  $\mathbf{A}$ :

$$(\mathbf{s}^0)^T \mathbf{A}\mathbf{s}^1 = 0 \quad \Rightarrow \quad (\mathbf{s}^0)^T \mathbf{A}(-\nabla f|_{\mathbf{x}=\mathbf{x}^1} + \beta_1 \mathbf{s}^0) = 0$$

or

$$(-\nabla f|_{\mathbf{x}=\mathbf{x}^1} + \beta_1 \mathbf{s}^0)^T \mathbf{A}\mathbf{s}^0 = 0 \quad (4.39)$$

Recall eq.(4.37) and substitute that expression into eq.(4.39), to obtain

$$(-\nabla f|_{\mathbf{x}=\mathbf{x}^1} + \beta_1 \mathbf{s}^0)^T \mathbf{A} \left[ \frac{1}{\lambda^*}(\mathbf{x}^1 - \mathbf{x}^0) \right] = 0 \quad (4.40)$$

Next, we find an expression for the difference  $\mathbf{x}^1 - \mathbf{x}^0$  in terms of the gradients at  $\mathbf{x}^0$  and  $\mathbf{x}^1$ . Indeed, recalling the quadratic approximation, eq.(4.31),

$$\nabla f|_{\mathbf{x}=\mathbf{x}^1} - \nabla f|_{\mathbf{x}=\mathbf{x}^0} \approx \mathbf{A}\mathbf{x}^1 - \mathbf{b} - (\mathbf{A}\mathbf{x}^0 - \mathbf{b}) = \mathbf{A}(\mathbf{x}^1 - \mathbf{x}^0) \quad (4.41)$$

Substitute expression (4.41) into eq.(4.40), after clearing the denominator:

$$(-\nabla f|_{\mathbf{x}=\mathbf{x}^1} + \beta_1 \mathbf{s}^0)^T (\nabla f|_{\mathbf{x}=\mathbf{x}^1} - \nabla f|_{\mathbf{x}=\mathbf{x}^0}) = 0 \quad (4.42)$$

Upon expansion,

$$-(\nabla f|_{\mathbf{x}=\mathbf{x}^1})^T \nabla f|_{\mathbf{x}=\mathbf{x}^1} + \underbrace{(\nabla f|_{\mathbf{x}=\mathbf{x}^0})^T \nabla f|_{\mathbf{x}=\mathbf{x}^1}}_{-\mathbf{s}^0} + \beta_1 \underbrace{(\nabla f|_{\mathbf{x}=\mathbf{x}^1})^T \mathbf{s}^0}_0 - \beta_1 (\nabla f|_{\mathbf{x}=\mathbf{x}^0})^T \mathbf{s}^0 = 0$$

---

<sup>3</sup>Although  $\mathbf{A}$  is apparently needed to compute  $\lambda^*$ , in reality the latter can be computed as the root of  $(\mathbf{s}^0)^T \nabla f|_{\mathbf{x}=\mathbf{x}^1} = 0$ , which is linear in  $\lambda$

where we have recalled eqs.(4.32) and (4.34). Hence, the above equation simplifies to

$$(\nabla f|_{\mathbf{x}=\mathbf{x}^1})^T \nabla f|_{\mathbf{x}=\mathbf{x}^1} + \beta_1 (\nabla f|_{\mathbf{x}=\mathbf{x}^0})^T \underbrace{\mathbf{s}^0}_{-\nabla f|_{\mathbf{x}=\mathbf{x}^0}} = 0$$

whence we can solve for  $\beta_1$  as

$$\beta_1 = \frac{(\nabla f|_{\mathbf{x}=\mathbf{x}^1})^T \nabla f|_{\mathbf{x}=\mathbf{x}^1}}{(\nabla f|_{\mathbf{x}=\mathbf{x}^0})^T \mathbf{s}^0} \equiv \frac{(\nabla f|_{\mathbf{x}=\mathbf{x}^1})^T \nabla f|_{\mathbf{x}=\mathbf{x}^1}}{(\nabla f|_{\mathbf{x}=\mathbf{x}^0})^T \nabla f|_{\mathbf{x}=\mathbf{x}^0}}$$

or

$$\beta_1 = \frac{\|\nabla f|_{\mathbf{x}=\mathbf{x}^1}\|^2}{\|\nabla f|_{\mathbf{x}=\mathbf{x}^0}\|^2} \quad (4.43)$$

whence  $\mathbf{x}^2 = \mathbf{x}^1 + \lambda \mathbf{s}^1$ ; find  $\lambda = \lambda^*$  that makes  $(\mathbf{s}^1)^T \nabla f|_{\mathbf{x}^2} = 0$ .

A third search direction  $\mathbf{s}^2$  is now defined as a linear combination of  $\mathbf{s}^1$  and  $-\nabla f|_{\mathbf{x}=\mathbf{x}^2}$ :

$$\mathbf{s}^2 = -\nabla f|_{\mathbf{x}=\mathbf{x}^2} + \beta_2 \mathbf{s}^1 \quad (4.44)$$

Now impose the *conjugacy condition* (4.6):

$$(\mathbf{s}^1)^T \mathbf{A} \mathbf{s}^2 = 0 \quad \Rightarrow \quad \beta_2 = \frac{(\nabla f|_{\mathbf{x}=\mathbf{x}^2})^T \nabla f|_{\mathbf{x}=\mathbf{x}^2}}{(\nabla f|_{\mathbf{x}=\mathbf{x}^1})^T \nabla f|_{\mathbf{x}=\mathbf{x}^1}} = \frac{\|\nabla f|_{\mathbf{x}=\mathbf{x}^2}\|^2}{\|\nabla f|_{\mathbf{x}=\mathbf{x}^1}\|^2} \quad (4.45)$$

In general, we have

$$\mathbf{s}^i = -\nabla f|_{\mathbf{x}=\mathbf{x}^i} + \beta_i \mathbf{s}^{i-1}, \quad \beta_i = \frac{\|\nabla f|_{\mathbf{x}=\mathbf{x}^i}\|^2}{\|\nabla f|_{\mathbf{x}=\mathbf{x}^{i-1}}\|^2}, \quad i = 1, 2, \dots \quad (4.46)$$

The procedure stops when  $\|\nabla f|_{\mathbf{x}=\mathbf{x}^i}\| < \epsilon$ , for a user-prescribed tolerance  $\epsilon$ , which indicates that the normality condition (4.2a) has been satisfied.

### Summary of the Fletcher-Reeves Algorithm

1. Choose an initial guess  $\mathbf{x}^0$
2. Let  $\mathbf{s}^0 = -\nabla f|_{\mathbf{x}=\mathbf{x}^0}$
3. Let

$$\mathbf{x}^1 = \mathbf{x}^0 + \lambda^* \mathbf{s}^0 \quad (4.47)$$

where  $\lambda^*$  is the value of  $\lambda$  that makes  $(\mathbf{s}^0)^T \nabla f|_{\mathbf{x}=\mathbf{x}^1} = 0$

4. Let  $i = 1$

5. Let

$$\mathbf{s}^i = -\nabla f|_{\mathbf{x}=\mathbf{x}^i} + \frac{\|\nabla f|_{\mathbf{x}=\mathbf{x}^i}\|^2}{\|\nabla f|_{\mathbf{x}=\mathbf{x}^{i-1}}\|^2} \mathbf{s}^{i-1}$$

6. Find the value of  $\lambda, \lambda^*$ , that makes  $(\mathbf{s}^i)^T \nabla f|_{\mathbf{x}=\mathbf{x}^{i+1}} = 0$ . Then,

$$\mathbf{x}^{i+1} = \mathbf{x}^i + \lambda^* \mathbf{s}^i$$

7. If  $\|\nabla f|_{\mathbf{x}=\mathbf{x}^{i+1}}\| < \epsilon$ , stop; else  $i \leftarrow i + 1$  and go to step 5

**Example 4.5.1** Shown in Fig. 4.2 is a three-joint robot with every pair of neighbouring joint axes at right angles and a distance  $a$  apart (Angeles, 2007). In designing this robot for a given maximum reach  $r_M$ , it is necessary to find  $r_M$  as a function of  $a$ , namely,  $r_M = Ra$ . The problem thus reduces to finding the coefficient  $R$ . **N.B.:** it is known that the position vector  $\mathbf{c}$  of the end-point  $C$ , for arbitrary values of the joint angles  $\{\theta_i\}_1^3$ , where  $\theta_i$  is the angle made by  $X_i$  and  $X_{i+1}$ , measured positive in the direction of  $Z_i$ , is given by

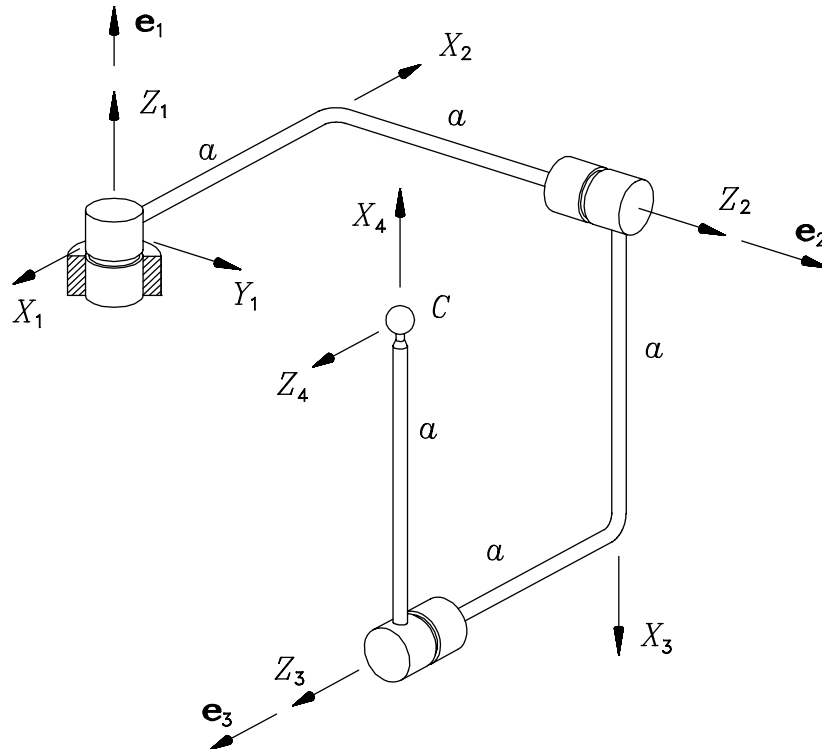


Figure 4.2: Manipulator configuration for  $C(0, a, 0)$ .



$$\mathbf{c} = a \begin{bmatrix} \cos(\theta_1) + \cos(\theta_1)(\cos(\theta_2) + \cos(\theta_2)\cos(\theta_3) + \sin(\theta_2)) + \sin(\theta_1)(1 + \sin(\theta_3)) \\ \sin(\theta_1) + \sin(\theta_1)(\cos(\theta_2) + \cos(\theta_2)\cos(\theta_3) + \sin(\theta_2)) - \cos(\theta_1)(1 + \sin(\theta_3)) \\ \sin(\theta_2) + \sin(\theta_2)\cos(\theta_3) - \cos(\theta_2) \end{bmatrix}$$

As  $r_M$  is reached when  $C$  lies farthest from the  $Z_1$  axis, the (nondimensional) distance to maximize is

$$r = \frac{\sqrt{c_1^2 + c_2^2}}{a}$$

The objective function is thus defined as

$$f(\mathbf{x}) = \frac{1}{2}r^2$$

which, upon expansion, shows that it is independent of  $\theta_1$ . This is not surprising, as this joint variable does not play any role in the distance from the  $Z_1$  axis to  $C$ . That is, if the second and third joints are locked, and the first joint goes through a full revolution, at a given robot posture, the distance of  $C$  from  $Z_1$  remains constant.

The Fletcher-Reeves conjugate-gradient method is chosen to solve the problem. However, this method was introduced above as applicable to function-minimization. For this reason, the objective function has to be redefined as the negative of that given above:

$$f(\mathbf{x}) = -\frac{1}{2}r^2$$

which is displayed below:

$$f = -2 - \cos(\theta_2)\cos(\theta_3)\sin(\theta_2) - (\cos(\theta_2))^2\cos(\theta_3) - (1/2)(\cos(\theta_2))^2(\cos(\theta_3))^2 \\ + (1/2)(\cos(\theta_3))^2 - \cos(\theta_2) - \sin(\theta_2) - \sin(\theta_3) - \cos(\theta_2)\sin(\theta_2) - \cos(\theta_2)\cos(\theta_3)$$

The above function is thus a function of only two variables,  $\theta_2$  and  $\theta_3$ , which then become the two components of the vector  $\mathbf{x}$  of design variables.

Now let us compute  $\nabla f = \partial f / \partial \mathbf{x}$ , with  $\mathbf{x} = [\theta_2, \theta_3]^T$ :

$$\frac{\partial f}{\partial \theta_2} = (\sin(\theta_2))^2\cos(\theta_3) - (\cos(\theta_2))^2\cos(\theta_3) + 2\cos(\theta_2)\cos(\theta_3)\sin(\theta_2) \\ + \cos(\theta_2)(\cos(\theta_3))^2\sin(\theta_2) + \sin(\theta_2) - \cos(\theta_2) + (\sin(\theta_2))^2 - (\cos(\theta_2))^2 \\ + \sin(\theta_2)\cos(\theta_3)$$

$$\begin{aligned}\frac{\partial f}{\partial \theta_3} = & \cos(\theta_2) \sin(\theta_3) \sin(\theta_2) + (\cos(\theta_2))^2 \sin(\theta_3) + (\cos(\theta_2))^2 \cos(\theta_3) \sin(\theta_3) \\ & - \cos(\theta_3) \sin(\theta_3) - \cos(\theta_3) + \cos(\theta_2) \sin(\theta_3)\end{aligned}$$

Since we have a symbolic expression for  $\nabla f$ , and this is given by two scalar functions of two variables, we can plot the contour defined by each function in the  $\theta_2$ - $\theta_3$  plane, all real stationary points, maxima, minima and saddle points, being found at the intersections of the two contours, which are plotted in Fig. 4.3.

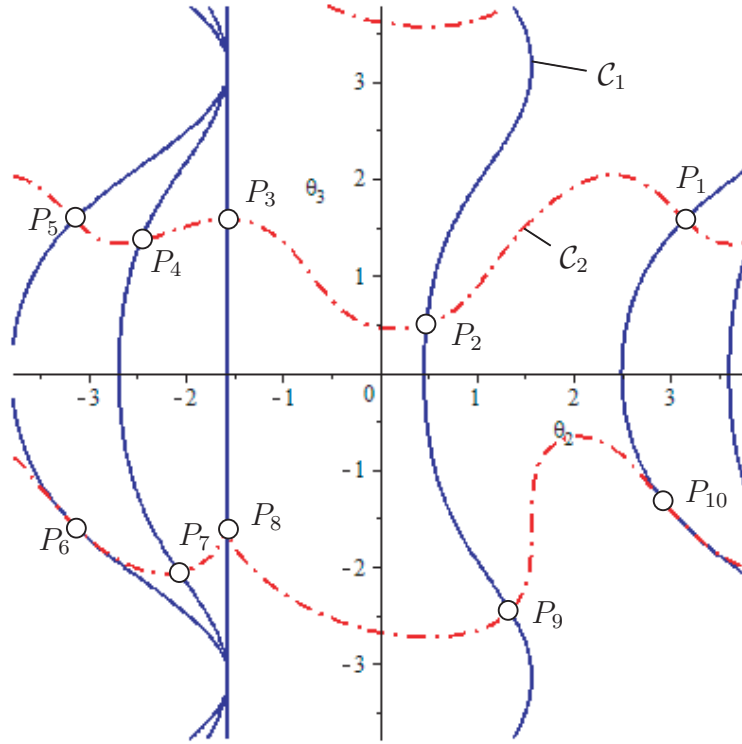


Figure 4.3: Contours  $\mathcal{C}_1$  and  $\mathcal{C}_2$  defined by the plots of  $\partial f/\partial \theta_2 = 0$  and  $\partial f/\partial \theta_3 = 0$ , respectively, which yield all the real SPs of  $f(\mathbf{x})$

Apparently, the two contours intersect at 10 points, although the two at the leftmost end,  $P_5$  and  $P_6$ , bear coordinates with values that differ from those of the points at the rightmost end,  $P_1$  and  $P_{10}$ , by exactly  $2\pi$ . Therefore, only eight intersections lead to distinct SPs. Some of these are local maxima, some are local minima, and some are saddle points. In order to gain insight into the nature of the SPs, the surface  $z = r^2(\theta_2, \theta_3)$  is plotted in Fig. 4.4a, its projection onto the  $\theta_2$ - $\theta_3$  plane being displayed in Fig. 4.4b. From these plots it is apparent that  $P_2$  is the maximum maximorum, i.e., the global maximum in the case at hand.

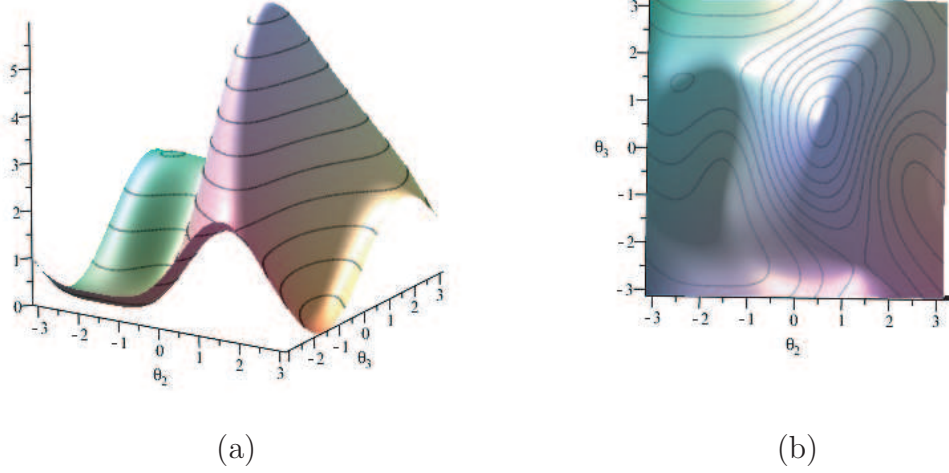


Figure 4.4: Plots of the isocontours—equal values of  $z$ —for the distance of  $C$  from  $Z_1$  of Fig. 4.2

Within the Fletcher-Reeves (FR) method, point  $P_2$  is targeted by defining the initial guess as  $\theta_2 = \theta_3 = 0$ , which yielded:

$$f_0 = -5.0, \quad (\nabla f)^0 = \begin{bmatrix} -3.0 \\ -1.0 \end{bmatrix}, \quad \mathbf{s}^0 = \begin{bmatrix} 3.0 \\ 1.0 \end{bmatrix} \Rightarrow \lambda \mathbf{s}^0 = \lambda \begin{bmatrix} 3.0 \\ 1.0 \end{bmatrix}$$

and hence,  $\mathbf{x}^1 = [3.0\lambda, \lambda]^T$ . Next, a value of  $\lambda$  is sought that satisfies

$$F \equiv (\mathbf{s}^0)^T (\nabla f)^1 = 0$$

The equation thus resulting is transcendental, in powers of the harmonic functions of products of the form  $\lambda \theta_i$ , for  $i = 2, 3$ . For this reason, a closed-form solution is not possible. Using the plotting capabilities of computer-algebra software, a plot of  $F(\lambda)$  vs.  $\lambda$  is obtained, as displayed in Fig. 4.5, in the interval  $[-1, +1]$ . The smallest positive root, estimated by inspection as  $\lambda^* = 0.18$ , is adopted here. With this value, an updated  $\mathbf{x}^1$  is calculated as

$$\mathbf{x}^1 = \mathbf{x}^0 + \lambda \mathbf{s}^0 = \begin{bmatrix} 0.54 \\ 0.18 \end{bmatrix}$$

whence updated values  $(\nabla f)^1$ ,  $\mathbf{s}^1$  and  $\lambda \mathbf{s}^1$  follow, with a new equation  $F(\lambda) = 0$ . Its smallest positive value is found, again, by inspection of the  $F(\lambda)$ -vs.- $\lambda$  plot. The procedure was conducted for seven iterations, with the results

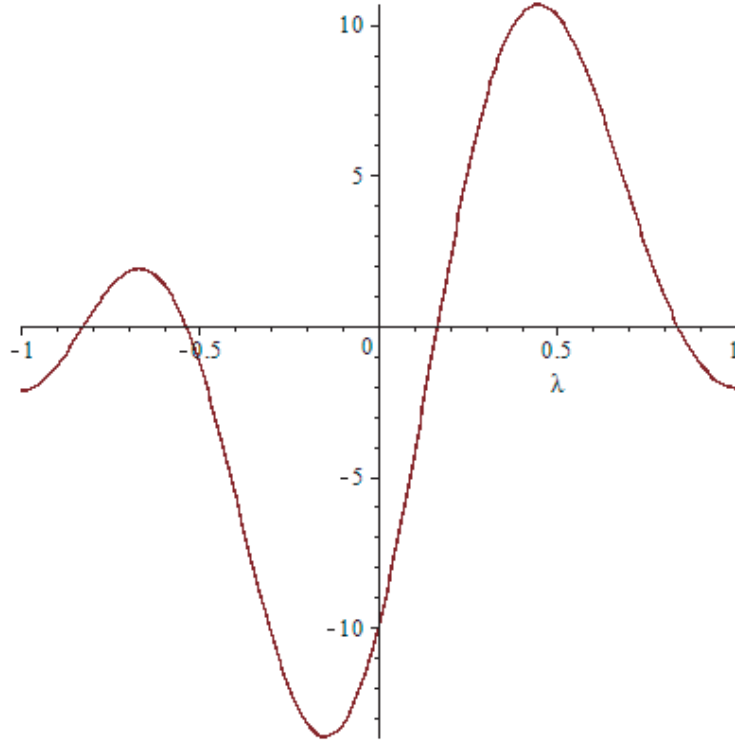


Figure 4.5: Plot of  $F = (\mathbf{s}^0)^T (\nabla f)^1$  vs.  $\lambda$ , showing its intersections with the  $\lambda$ -axis

$$\mathbf{x}^7 = \begin{bmatrix} 0.4851 \\ 0.4594 \end{bmatrix}, \quad (\nabla f)^7 = \begin{bmatrix} -0.001162 \\ -0.0605 \end{bmatrix}$$

which was accepted as the optimum value, given that the Chebyshev norm of the gradient dropped from 3.0 to 0.06, i.e., to 2% of its original value. The objective function and the maximum value of  $r$  are

$$f_7 = -5.9836 \quad \Rightarrow \quad r_M = \sqrt{-2f_7a} = 3.4594a$$

A more precise value of  $r_M$  will be calculated in Example 5.3.2 using a constrained-minimization approach. A plot of the isocontours of the objective function and the iterative FR-sequence is included in Fig. 4.6

For the record, an alternative sequence is displayed in Fig. 4.7, generated with a faulty code: the norms in numerator and denominator of the expression for  $\beta_i$  in eq.(4.46) were left unsquared! The sequence appears, after the 7th iteration, to be converging to the minimum sought. However, the convergence is rather slow, for, at the same iteration, the procedure yields

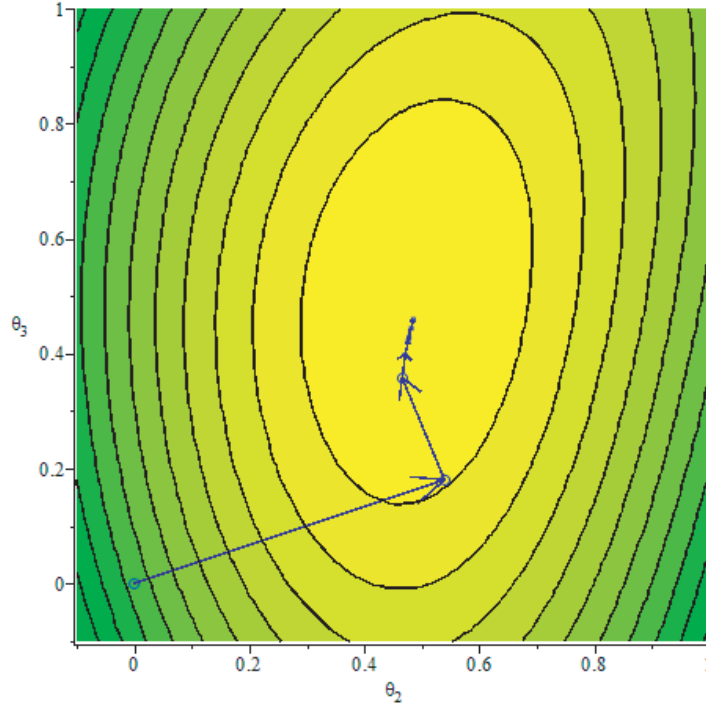


Figure 4.6: Sequence of iterations of the FR method for Example 4.5.1

$$\mathbf{x}^7 = \begin{bmatrix} 0.4700 \\ 0.5116 \end{bmatrix}, \quad (\nabla f)^7 = \begin{bmatrix} -0.1647 \\ 0.15222 \end{bmatrix}$$

where, apparently, the Chebyshev norm of the gradient has gone down only to slightly over 5% of its original value. The details of the procedure are included in `RRRorthoConGrad.mw`, those of the faulty code in `RRRorthoConGradGlitch.mw`.

### 4.5.3 Quasi-Newton Methods

As we will see in Section 4.6, Newton methods rely on the normality conditions, which lead to a determined system of  $n$  nonlinear equations in  $n$  unknowns. In applying those methods it is assumed that the Hessian of the objective function, which is the Jacobian  $\Phi$  of the Newton-Raphson method, is available, and hence, the Hessian can be used to update the iterations. Quasi-Newton methods replace the update  $\Delta \mathbf{x} = -\Phi_0^{-1} \phi^0$  of eq.(3.92) by an expression that a) does not rely on the Hessian, but only on the gradient of the objective function and b) does not require any matrix inversion. These features make quasi-Newton methods quite

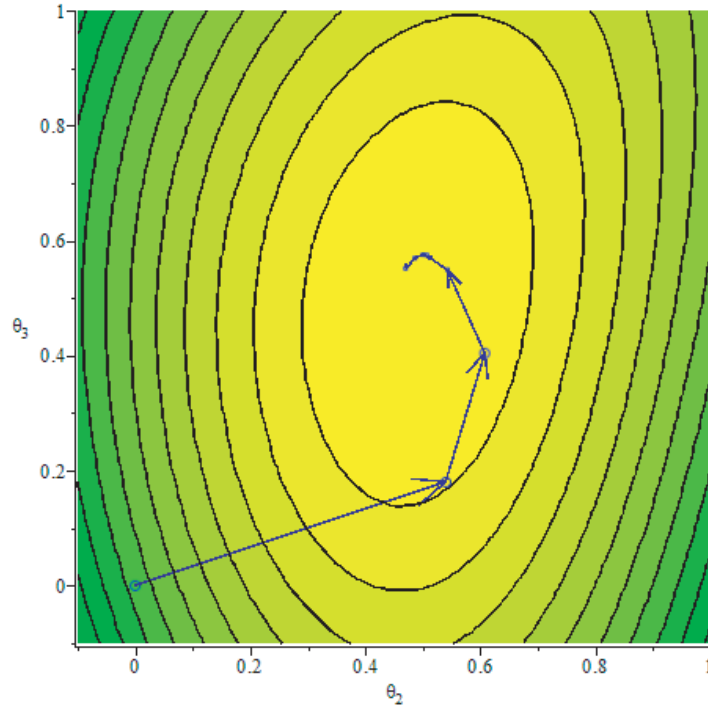


Figure 4.7: Sequence of faulty iterations of the FR method for Example 4.5.1

attractive, and many times, preferable over Newton methods. The two quasi-Newton methods outlined below differ only in the form in which the update of the solution is computed. These two methods aim to find an approximation to  $(\nabla \nabla f)^{-1}$  using only information on  $\nabla f$ .

Moreover, while the Newton-Raphson method is known to have a *quadratic convergence rate*, quasi-Newton methods show a convergence rate that lies between gradient methods and Newton methods. That is, quasi-Newton methods have a *superlinear convergence rate*.

### The Davidon-Fletcher-Powell Method

The method is summarized below:

#### Algorithm

1. Give an initial guess:  $\mathbf{x}^0$ ;  $0 \leftarrow i$
2. Define an initial search direction:  $\mathbf{s}^0 = -\nabla f|_{\mathbf{x}=\mathbf{x}^0}$
3. Define an initial Hessian-inverse:  $\mathbf{B}_0 = \mathbf{1}$ , the  $n \times n$  identity matrix

4. Let

$$\mathbf{x}^{i+1} = \mathbf{x}^i + \lambda_i \mathbf{s}^i$$

Then find  $\lambda_i$  that minimizes<sup>4</sup>  $f(\mathbf{x}^{i+1})$ .

5. 
$$\mathbf{g}^i = \nabla f|_{\mathbf{x}=\mathbf{x}^{i+1}} - \nabla f|_{\mathbf{x}=\mathbf{x}^i}$$

6.

$$\mathbf{M}_i = \lambda_i \frac{\mathbf{s}^i (\mathbf{s}^i)^T}{(\mathbf{s}^i)^T \mathbf{g}^i}, \quad \mathbf{N}_i = -\frac{\mathbf{B}_i \mathbf{g}^i (\mathbf{B}_i \mathbf{g}^i)^T}{(\mathbf{g}^i)^T \mathbf{B}_i \mathbf{g}^i}, \quad \mathbf{B}_{i+1} = \mathbf{B}_i + \mathbf{M}_i + \mathbf{N}_i$$

7. 
$$\mathbf{s}^{i+1} = -\mathbf{B}_{i+1} \nabla f|_{\mathbf{x}=\mathbf{x}^{i+1}}$$

8. if convergence criterion reached, stop; else, go to 4

### The Broyden-Fletcher-Goldfarb-Shanno Method

This is an improved DFP method, but still with superlinear convergence. Only difference with the DFP Algorithm lies in step 6, which is replaced by:

$$\begin{aligned} \mathbf{M}'_i &= \alpha_i \frac{\mathbf{s}^i (\mathbf{s}^i)^T}{(\mathbf{s}^i)^T \mathbf{g}^i}, \quad \alpha_i = 1 + \frac{(\mathbf{g}^i)^T \mathbf{B}_i \mathbf{g}^i}{(\mathbf{s}^i)^T \mathbf{g}^i} \\ \mathbf{N}'_i &= -\frac{\mathbf{s}^i (\mathbf{g}^i)^T \mathbf{B}_i}{(\mathbf{s}^i)^T \mathbf{g}^i} \\ \mathbf{B}_{i+1} &= \mathbf{B}_i + \mathbf{M}'_i + \mathbf{N}'_i + (\mathbf{N}'_i)^T \end{aligned}$$

## 4.6 Newton Methods

### 4.6.1 The Newton-Raphson Method

Here, we resort to the normality condition (4.2a), and let

$$\phi(\mathbf{x}) \equiv \nabla f$$

the normality condition thus leading to a system of  $n$  nonlinear equations in  $n$  unknowns of the form of eq.(3.87), repeated below for quick reference:

$$\phi(\mathbf{x}) = \mathbf{0}$$

---

<sup>4</sup>In this step, any of the methods studied in Chapter 2 can be applied. A thorough discussion of univariable minimization is available in (Brent, 1972).

which can be solved using the Newton-Raphson method because, by assumption, second-order derivatives of the objective function are available, and hence, the Jacobian  $\Phi$  of  $\phi(\mathbf{x})$  with respect to  $\mathbf{x}$  is nothing but the Hessian matrix of  $f(\mathbf{x})$ , i.e.,

$$\Phi = \nabla \nabla f$$

Let us introduce the notation

$$\Phi_k \equiv \Phi(\mathbf{x}^k), \quad \phi^k \equiv \phi(\mathbf{x}^k) \quad (4.48)$$

at the  $k$ th iteration

This method, while offering a quadratic convergence, is not as favoured as methods of the gradient type. One reason argued in the past is the cost of solving a system of linear equations, namely, eq.(3.91) at each iteration, although with fast processors this argument loses weight. Another reason why Newton methods are not popular is the inherent requirement of a Hessian. In many practical problems, e.g., structural optimization, the objective function is not an *analytic function* of the design parameters. This is the case when the objective function is the maximum von Mises stress in a structure. As this function is not analytic, its gradient, not to speak of its Hessian, is not available. That is, not even with finite differences is it possible to approximate the gradient of a non-analytic function of the design variables.

## 4.6.2 The Levenberg-Marquardt Method

The Levenberg-Marquardt method aims at enhancing the robustness of the Newton-Raphson method, when the Hessian becomes ill-conditioned, by adding to the Hessian, which is assumed positive-definite, a symmetric, isotropic matrix  $\alpha \mathbf{1}$ , where  $\alpha > 0$  and  $\mathbf{1}$  is the  $n \times n$  identity matrix:

$$\overline{\nabla \nabla f} \leftarrow \nabla \nabla f + \alpha \mathbf{1} \quad (4.49)$$

Notice that the eigenvalues<sup>5</sup> of  $\nabla \nabla f$ , denoted by  $\{\lambda_i\}_1^n$ , and those of  $\overline{\nabla \nabla f}$ , denoted by  $\{\mu_i\}_1^n$ , are related by

$$\mu_i = \lambda_i + \alpha, \quad i = 1, 2, \dots, n$$

---

<sup>5</sup>By virtue of the assumed positive-definiteness of the Hessian, its eigenvalues are identical to their singular values.



If we denote by  $\kappa$  the 2-norm condition number of  $\nabla\nabla f$  and by  $\overline{\kappa}$  that of  $\overline{\nabla\nabla f}$ , we have

$$\kappa = \frac{\lambda_M}{\lambda_m}, \quad \overline{\kappa} = \frac{\lambda_M + \alpha}{\lambda_m + \alpha} \quad (4.50)$$

the result being that  $\overline{\kappa} < \kappa$ , and hence, the numerical behaviour of the Hessian is stabilized.



## Chapter 5

# Equality-Constrained Optimization: Normality Conditions

### 5.1 Introduction

In this chapter we introduce the simplest class of constrained-optimization problems, namely, those subject to equality constraints. The problem statement at hand is

$$f(\mathbf{x}) \rightarrow \min_{\mathbf{x}} \quad (5.1a)$$

subject to

$$\mathbf{h}(\mathbf{x}) = \mathbf{0}_l \quad (5.1b)$$

where  $\mathbf{h}$  is a *smooth*<sup>1</sup>  $l$ -dimensional vector function of the  $n$ -dimensional vector argument  $\mathbf{x}$ ,  $\mathbf{0}_l$  denoting the  $l$ -dimensional zero vector.

Moreover,  $l < n$ , for a  $n$ -dimensional design vector  $\mathbf{x}$ , as otherwise  $\mathbf{x}$  would be either fully constrained or overconstrained, thereby leaving no room for optimization!

The main outcome is the derivation of the two *normality conditions* of the problems at hand. We derive the first of these in two forms: (i) the *primal form*, in terms of the gradients of the objective function  $f(\mathbf{x})$  to be minimized and of the constraint functions of (5.1b),  $\mathbf{h}$ ; and (ii) the *dual form*, in terms of an *orthogonal complement* of the gradient of  $\mathbf{h}$  with respect to  $\mathbf{x}$ , and  $\nabla f(\mathbf{x})$ .

---

<sup>1</sup>Smoothness implies that each component of  $\mathbf{h}(\mathbf{x})$ , i.e., each constraint, is continuous and has a continuous gradient and a continuous Hessian with respect to  $\mathbf{x}$ .

As a special case, that lends itself to a closed-form solution, we study *minimum-norm* problems, whereby a *weighted* Euclidean norm of the design vector is to be minimized subject to  $l$  linear equality constraints. In this vein, we introduce the *right Moore-Penrose generalized inverse*.

## 5.2 The First-Order Normality Conditions

### 5.2.1 The Primal Form

We derive here the first-order normality conditions of problem (5.1a) in *primal form*. To this end, we resort to *Lagrange multipliers*  $\lambda_1, \lambda_2, \dots, \lambda_l$ , one for each scalar constraint  $h_i(\mathbf{x}) = 0$ , and group them in the  $l$ -dimensional array  $\boldsymbol{\lambda}$ . Upon adjoining the  $l$  constraints to the objective function  $f(\mathbf{x})$ , we obtain the *Lagrangian*  $F(\mathbf{x}; \boldsymbol{\lambda})$  that we aim to minimize under no constraints, while choosing  $\boldsymbol{\lambda}$  in such a way that the  $l$  equality constraints are satisfied. That is,

$$F(\mathbf{x}; \boldsymbol{\lambda}) \equiv f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{h} \quad \rightarrow \quad \min_{\mathbf{x}, \boldsymbol{\lambda}} \quad (5.2)$$

subject to no constraints. We have thus transformed the equality-constrained minimization problem into an unconstrained one. We derive now the first-order normality conditions (FONC) of the problem at hand by recalling those of Ch. 4, requiring that the gradient of the objective function with respect to the whole set of design variables— $\mathbf{x}$  and  $\boldsymbol{\lambda}$  in the case at hand—vanish. However, note that we now have  $l$  additional variables besides the original  $n$  design variables. Hence, the design-variable vector must be augmented correspondingly, which we do by defining an augmented  $(n + l)$ -dimensional design vector  $\mathbf{y}$ :

$$\mathbf{y} \equiv \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{bmatrix} \quad (5.3)$$

Therefore, the unconstrained minimization problem (5.2) can be formulated in a more compact form, namely,

$$F(\mathbf{y}) \quad \rightarrow \quad \min_{\mathbf{y}} \quad (5.4)$$

subject to no constraints. The FONC of the above problem are, thus,

$$\frac{\partial F}{\partial \mathbf{y}} = \begin{bmatrix} \partial F / \partial \mathbf{x} \\ \partial F / \partial \boldsymbol{\lambda} \end{bmatrix} = \mathbf{0}_{n+l} \quad (5.5a)$$

where  $\mathbf{0}_{n+l}$  denotes the  $(n+l)$ -dimensional zero vector. The above equation can thus be broken down into two, namely,

$$\frac{\partial F}{\partial \mathbf{x}} = \mathbf{0}_n \quad (5.5b)$$

$$\frac{\partial F}{\partial \boldsymbol{\lambda}} = \mathbf{0}_l \quad (5.5c)$$

To gain insight into the geometric significance of the foregoing normality conditions, we expand the left-hand side of eq.(5.5b) componentwise:

$$\begin{aligned} \frac{\partial F}{\partial x_1} &\equiv \frac{\partial f}{\partial x_1} + \lambda_1 \frac{\partial h_1}{\partial x_1} + \lambda_2 \frac{\partial h_2}{\partial x_1} + \cdots + \lambda_l \frac{\partial h_l}{\partial x_1} = 0 \\ \frac{\partial F}{\partial x_2} &\equiv \frac{\partial f}{\partial x_2} + \lambda_1 \frac{\partial h_1}{\partial x_2} + \lambda_2 \frac{\partial h_2}{\partial x_2} + \cdots + \lambda_l \frac{\partial h_l}{\partial x_2} = 0 \\ &\vdots \\ \frac{\partial F}{\partial x_n} &\equiv \underbrace{\frac{\partial f}{\partial x_n}}_{\nabla f} + \lambda_1 \frac{\partial h_1}{\partial x_n} + \lambda_2 \frac{\partial h_2}{\partial x_n} + \cdots + \lambda_l \frac{\partial h_l}{\partial x_n} = 0 \end{aligned} \quad (5.6)$$

where the first term of the  $i$ th equation can be readily identified as the  $i$ th component of  $\nabla f = \partial f / \partial \mathbf{x}$ . The sum of the remaining terms of the same equation can be identified as the  $i$ th component of an inner product  $p_i$  defined as

$$p_i \equiv \begin{bmatrix} \partial h_1 / \partial x_i & \partial h_2 / \partial x_i & \cdots & \partial h_l / \partial x_i \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_l \end{bmatrix} \equiv \left( \frac{\partial \mathbf{h}}{\partial x_i} \right)^T \boldsymbol{\lambda}$$

Therefore, if we let  $\mathbf{p} = \begin{bmatrix} p_1 & p_2 & \cdots & p_n \end{bmatrix}^T$ , then

$$\mathbf{p} = \underbrace{\begin{bmatrix} \partial h_1 / \partial x_1 & \partial h_2 / \partial x_1 & \cdots & \partial h_l / \partial x_1 \\ \partial h_1 / \partial x_2 & \partial h_2 / \partial x_2 & \cdots & \partial h_l / \partial x_2 \\ \vdots & \vdots & \ddots & \vdots \\ \partial h_1 / \partial x_n & \partial h_2 / \partial x_n & \cdots & \partial h_l / \partial x_n \end{bmatrix}}_{(\nabla \mathbf{h})^T: n \times l} \underbrace{\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_l \end{bmatrix}}_{\boldsymbol{\lambda}} \quad (5.7)$$

whose first factor can be readily identified as  $(\nabla \mathbf{h})^T$ . Indeed, compared with expression (3.90) for the partial derivative  $\nabla \boldsymbol{\phi}$  of a vector function  $\boldsymbol{\phi}(\mathbf{x})$  with respect to its vector argument, this factor is the transpose of  $\nabla \mathbf{h}$ , henceforth represented as

the *Jacobian*  $\mathbf{J}$ , or  $\mathbf{J}(\mathbf{x})$ , whenever the argument should be explicitly stated. The first  $n$  normality conditions, displayed in eq.(5.5b), thus amount to

$$\nabla f + \mathbf{J}^T \boldsymbol{\lambda} = \mathbf{0}_n \quad (5.8a)$$

Since variables  $\mathbf{x}$  and  $\boldsymbol{\lambda}$  are independent at the outset, the second term of the left-hand side of eq.(5.8a) is the gradient of  $\mathbf{h}^T \boldsymbol{\lambda}$ , i.e.,

$$\mathbf{J}^T \boldsymbol{\lambda} = \nabla(\mathbf{h}^T \boldsymbol{\lambda}) \quad (5.8b)$$

The remaining  $l$  normality conditions, displayed in eq.(5.5c), yield nothing but the constraints themselves, namely

$$\mathbf{h}(\mathbf{x}) = \mathbf{0}_l \quad (5.8c)$$

The FONC can thus be rewritten in a more illustrative form, namely,

$$\nabla f + \nabla(\mathbf{h}^T \boldsymbol{\lambda}) = \mathbf{0}_n \quad (5.8d)$$

which states that, at a SP, the two above gradients *must cancel each other*. This representation lends itself to a mechanical interpretation: as the gravity force is the gradient of a potential field, eq.(5.8d) can be interpreted as a force-equilibrium condition.

Equation (5.8a) is the vector representation of the *first-order normality conditions* (FONC) sought. What eq.(5.8a) represents has a geometric significance that will be made apparent upon rewriting it in the alternative form

$$\mathbf{J}^T \boldsymbol{\lambda} = -\nabla f \quad (5.9)$$

The foregoing equation states that, at a stationary point  $\mathbf{x}_o$ ,  $-\nabla f$ , or  $\nabla f$  for that matter, lies in the range of the transpose of the gradient of the constraints. Notice that the range  $\mathcal{J}'$  of  $\mathbf{J}^T$  is a subspace of the  $n$ -dimensional space of design variables. In fact,  $\dim(\mathcal{J}') = l < n$ , for this subspace is spanned by  $l$  linearly independent vectors, the columns of  $\mathbf{J}^T$ , or the  $n$ -dimensional rows of  $\mathbf{J}$ .

Algebraically, eq.(5.9) represents an overdetermined system of  $n$  linear equations in the  $l < n$  unknowns  $\{\lambda_i\}_1^l$ . The normality condition then states that the least-square approximation of this overdetermined system yields a zero error. That is, at a stationary point, the  $n$  ( $> l$ ) equations (5.9) become all consistent. Note that the least-square approximation  $\boldsymbol{\lambda}_o$  of the foregoing equations can be expressed in terms of the left Moore-Penrose generalized inverse of  $\mathbf{J}^T$ , namely,

$$\boldsymbol{\lambda}_o = -(\mathbf{J}\mathbf{J}^T)^{-1} \mathbf{J} \nabla f \quad (5.10)$$

The least-square error  $\mathbf{e}_o$  of this approximation is thus

$$\mathbf{e}_o = \mathbf{J}^T \boldsymbol{\lambda}_o - (-\nabla f) = -\mathbf{J}^T (\mathbf{J}\mathbf{J}^T)^{-1} \mathbf{J} \nabla f + \nabla f = [\mathbf{1} - \mathbf{J}^T (\mathbf{J}\mathbf{J}^T)^{-1} \mathbf{J}] \nabla f \quad (5.11)$$

with  $\mathbf{1}$  denoting the  $n \times n$  identity matrix.

We can now express the first-order normality condition (5.9) in an alternative form:

$$[\mathbf{1} - \mathbf{J}^T (\mathbf{J}\mathbf{J}^T)^{-1} \mathbf{J}] \nabla f = \mathbf{0}_n \quad (5.12)$$

The matrix inside the brackets in the foregoing equation can be readily identified as a projector, of the form of  $\mathbf{P}$  introduced in eq.(3.30). This projector maps vectors in  $\mathbb{R}^n$  onto the null space of  $\mathbf{J}$ , as the reader is invited to verify. In other words, *at a stationary point*  $P_o$  the gradient of the objective function need not vanish; only its projection onto the null space of the gradient of the constraints must vanish, which is an alternative form of stating the first-order normality condition. Sometimes the product  $\bar{\nabla} f$ , defined as

$$\bar{\nabla} f \equiv [\mathbf{1} - \mathbf{J}^T (\mathbf{J}\mathbf{J}^T)^{-1} \mathbf{J}] \nabla f \in \mathbb{R}^n \quad (5.13)$$

is referred to as the *constrained gradient*. The FONC (5.12) can then be simply stated as:

*At a stationary point of the equality-constrained problem (5.1a & b), the constrained gradient vanishes.*

### Exercise 5.2.1

Prove that

$$\mathbf{P} \equiv \mathbf{1} - \mathbf{J}^T (\mathbf{J}\mathbf{J}^T)^{-1} \mathbf{J} \quad (5.14)$$

(i) is a projector, and

(ii) indeed maps vectors in  $\mathbb{R}^n$  onto the null space of  $\mathbf{J}$ .

## 5.2.2 The Dual Form

One *alternative representation* of the FONC of the problem under discussion is now derived in what can be termed the *dual form*. This can be obtained rather simply if the concept of *orthogonal complement* of the Jacobian  $\mathbf{J}$  of the constraint function  $\mathbf{h}(\mathbf{x})$  of eq.(5.1b) is invoked. An orthogonal complement  $\mathbf{L}$  of a full-rank<sup>2</sup>  $\mathbf{J}$ —or of

---

<sup>2</sup>An orthogonal complement can also be defined for rank-deficient matrices, but such matrices will be left aside, for the sake of brevity.

any rectangular  $l \times n$  full-rank matrix, with  $l < n$ , for that matter—is a  $n \times n'$  matrix, for  $n' \equiv n - l$ , whose columns span the null space of  $\mathbf{J}$ . That is,

$$\mathbf{J}\mathbf{L} = \mathbf{O}_{ln'} \quad (5.15)$$

If now both sides of eq.(5.8a) are multiplied from the left by  $\mathbf{L}^T$ , the second term of the left-hand side disappears, and hence, the alternative form of the FONC is obtained:

$$\mathbf{L}^T \nabla f = \mathbf{0}_{n'} \quad (5.16)$$

That is, at a stationary point, the gradient of  $f$  need not vanish; however, it must lie in the null space of  $\mathbf{L}^T$ , i.e., in the range of  $\mathbf{J}^T$ —both spaces are of dimension  $n$ . The latter is, in fact, a restatement of the primal form of the FONC, as per eq.(5.9). We can thus call  $\mathbf{L}^T \nabla f$  the *feasible gradient*, and represent it by  $\nabla_u f$ , i.e.,

$$\nabla_u f = \mathbf{L}^T \nabla f \quad (5.17)$$

which is a  $(n - l)$ -dimensional vector. Sometimes the feasible gradient is also called the *reduced gradient*.

A geometric interpretation of the range and the null space of each of each of  $\mathbf{J}^T$  and  $\mathbf{L}^T$ , and their relations with the two gradients  $\nabla f$  and  $\nabla h = \mathbf{J}^T$ , for a simple case in which  $n = 2$  and  $l = 1$ , is shown in Fig. 5.1.

A *differential interpretation* of  $\mathbf{L}$  can be obtained if we realize that the solution sought  $\mathbf{x}_o$  must lie in a subset  $\mathcal{F}$  of the space  $\mathbb{R}^n$  of design variables, of reduced dimension  $n - l$ . This subset contains all design vectors satisfying the constraints. Such vectors will be labelled  $\mathbf{x}_F$ , to distinguish them from all other vectors of the design space. Now, this set need not be a vector space, and in general it is not. Indeed, if the constraints are nonlinear, then a linear combination of two distinct *feasible vectors*  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , i.e., which satisfy the constraint eqs.(5.1b), will most likely not satisfy the constraints, even if these two designs do so independently. Neither need the zero vector satisfy the constraints, which thus disqualifies  $\mathcal{F}$  from being a subspace of  $\mathbb{R}^n$ . What we have as a *feasible subset* of the design space is a *manifold*, i.e., a smooth surface embedded in  $\mathbb{R}^n$ . We shall term the subset  $\mathcal{F}$  the *feasible manifold*.

Finding  $\mathcal{F}$  may be a tremendous task when the constraints are nonlinear and algebraically complicated. The good news is that we do not actually need the feasible manifold to obtain a feasible solution. What we really need is a *feasible subspace*  $\mathcal{T}_F$  tangent to the said manifold at a *feasible point*  $P_F$ , of position vector  $\mathbf{x}_F$ . This



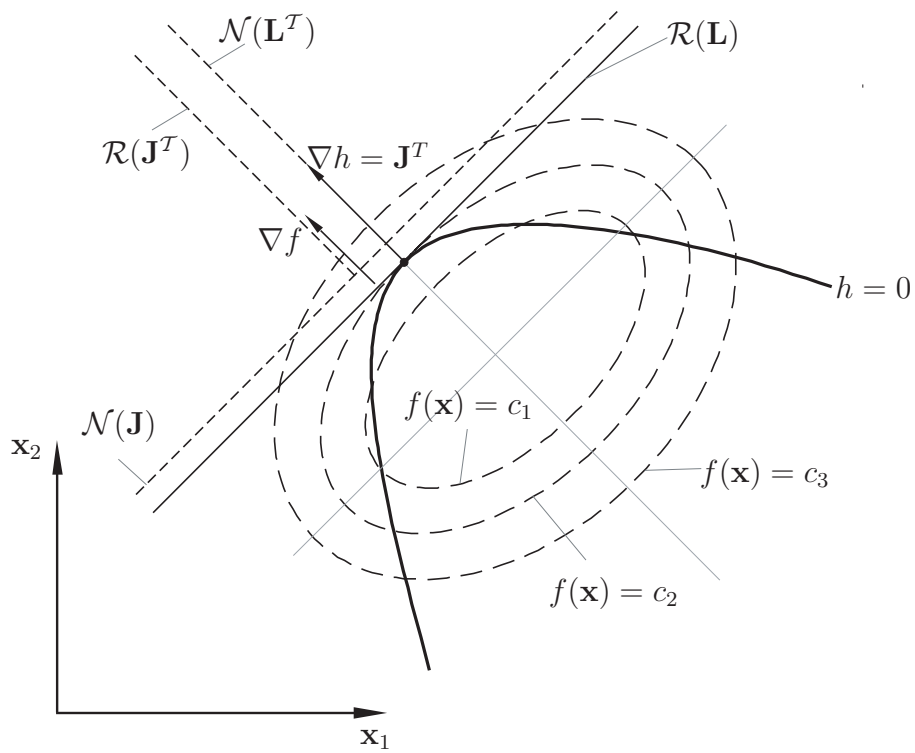


Figure 5.1: Geometric interpretation of the subspaces associated with matrices  $\mathbf{J}$  and  $\mathbf{L}$  and the gradients  $\nabla f$  and  $\nabla h$ , for  $n = 1$  and  $l = 1$ .

subspace is nothing but the null space of  $\mathbf{J}^T$ . Indeed, since  $\mathbf{x}_F$  is a feasible design vector, it verifies the constraints, i.e.,

$$\mathbf{h}(\mathbf{x}_F) = \mathbf{0}_l \quad (5.18)$$

An arbitrary “move”  $\Delta\mathbf{x}$  from  $\mathbf{x}_F$  will most likely take  $P_F$  away from the constraints  $\mathbf{h}(\mathbf{x}) = \mathbf{0}_l$ . What we would like to find is a *feasible move*, i.e., a vector  $\Delta\mathbf{x}_F$  such that  $\mathbf{x}_F + \Delta\mathbf{x}_F$  still verifies the constraints, i.e.,

$$\mathbf{h}(\mathbf{x}_F + \Delta\mathbf{x}_F) = \mathbf{h}(\mathbf{x}_F) + \mathbf{J}(\mathbf{x}_F)\Delta\mathbf{x}_F + \text{HOT} = \mathbf{0}_l \quad (5.19)$$

Since we assumed at the outset that  $\mathbf{x}_F$  is feasible, as per eq.(5.18), we have, from the foregoing equation and to a *first-order approximation*, i.e., neglecting HOT,

$$\mathbf{J}(\mathbf{x}_F)\Delta\mathbf{x}_F = \mathbf{0}_l \quad (5.20)$$

i.e.,  $\Delta\mathbf{x}_F$  lies in  $\mathcal{N}[\mathbf{J}(\mathbf{x}_F)]$ , the null space of  $\mathbf{J}(\mathbf{x}_F)$ . In light of eq.(5.15), then, an arbitrary  $\Delta\mathbf{u} \in \mathbb{R}^{n'}$  is mapped by  $\mathbf{L}$  into  $\mathcal{N}[\mathbf{J}(\mathbf{x}_F)]$ . That is, a feasible move  $\Delta\mathbf{x}_F$  can be produced as

$$\Delta\mathbf{x}_F = \mathbf{L}\Delta\mathbf{u} \quad (5.21a)$$

for

$$\mathbf{J}\Delta\mathbf{x}_F = \mathbf{J}\mathbf{L}\Delta\mathbf{u} = \mathbf{0}_{n'} \quad (5.21b)$$

A geometric interpretation of the subspaces associated with  $\mathbf{J}$  and  $\mathbf{L}$ , and their transposes, is illustrated in Fig. 5.1.

Notice that, from eq.(5.21a),  $\mathbf{L}$  has the differential interpretation

$$\mathbf{L} = \frac{\partial\mathbf{x}}{\partial\mathbf{u}} \quad (5.22)$$

and hence, the FONC (5.16) can be restated as

$$\left(\frac{\partial\mathbf{x}}{\partial\mathbf{u}}\right)^T \left(\frac{\partial f}{\partial\mathbf{x}}\right) \equiv \frac{\partial f}{\partial\mathbf{u}} = \mathbf{0}_{n'} \quad \text{or} \quad \nabla_u f = \mathbf{0}_{n'} \quad (5.23)$$

thereby justifying the subscript in  $\nabla_u$ . That is, the FONC (5.23) states that, at a stationary point of problem (5.1a & b), the gradient of  $f(\mathbf{x})$  with respect to the vector of independent design variables  $\mathbf{u}$  vanishes.

**Remark:** When comparing the two forms of the FONC, eqs.(5.12) and (5.16), the simplicity of the latter with respect to the former is apparent. This simplicity, however, is more than formal, for eq.(5.12) involves  $n$  scalar equations, while eq.(5.16) involves only  $n - l$  scalar equations.

### 5.3 The Second-Order Normality Conditions

The *second-order normality conditions* (SONC) of the problem at hand are now derived. To this end, the *second variation*  $\Delta\Delta F$  of the Lagrangian at a stationary, feasible point  $\mathbf{y}_o = [\mathbf{x}_o^T, \boldsymbol{\lambda}_o^T]^T$  is first derived. This is defined up to second-order terms, i.e.,

$$\Delta\Delta F \equiv \left. \frac{\partial F}{\partial \mathbf{y}} \right|_{\mathbf{y}=\mathbf{y}_o} \Delta\mathbf{y} + \frac{1}{2} \Delta\mathbf{y}^T \left. \frac{\partial^2 F}{\partial \mathbf{y}^2} \right|_{\mathbf{y}=\mathbf{y}_o} \Delta\mathbf{y} \quad (5.24)$$

where

$$\Delta\mathbf{y} = [\Delta\mathbf{x}_F^T \quad \Delta\boldsymbol{\lambda}_F^T]^T \quad (5.25)$$

As we have assumed that the above variation is computed at a stationary point  $\mathbf{y}_o$ , the first variation, i.e., the first term in the right-hand side of eq.(5.24), vanishes. Moreover, let us assume that the variation  $\Delta\mathbf{x}_F$  of the design-variable vector is *feasible*, and hence, verifies the constraints (5.1b) *to a first order*, i.e.,

$$\mathbf{h}(\mathbf{x}_o + \Delta\mathbf{x}_F) = \mathbf{h}(\mathbf{x}_o) + \mathbf{J}\Delta\mathbf{x}_F = \mathbf{0}_l \quad (5.26a)$$

Now, since  $\mathbf{x}_o$  is feasible, the first term of the above expression vanishes, and hence, the *feasible move*  $\Delta\mathbf{x}_F$  verifies

$$\mathbf{J}(\mathbf{x}_o)\Delta\mathbf{x}_F = \mathbf{0}_l \quad (5.26b)$$

Moreover, upon expansion of  $\Delta\Delta F$ , as given by eq.(5.24),

$$\begin{aligned} \Delta\Delta F &\equiv \frac{1}{2} \Delta\mathbf{y}^T \frac{\partial^2 F}{\partial \mathbf{y}^2} \Delta\mathbf{y} \\ &= \frac{1}{2} [\Delta\mathbf{x}_F^T \quad \Delta\boldsymbol{\lambda}^T] \begin{bmatrix} \nabla\nabla f + \partial(\mathbf{J}^T \boldsymbol{\lambda})/\partial \mathbf{x} & \mathbf{J}^T \\ \mathbf{J} & \mathbf{O}_l \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x}_F \\ \Delta\boldsymbol{\lambda} \end{bmatrix} \\ &= \frac{1}{2} \Delta\mathbf{x}_F^T \left[ \nabla\nabla f + \frac{\partial(\mathbf{J}^T \boldsymbol{\lambda})}{\partial \mathbf{x}} \right] \Delta\mathbf{x}_F + \Delta\mathbf{x}_F^T \mathbf{J}^T(\mathbf{x}_o) \Delta\boldsymbol{\lambda} \end{aligned}$$

where, as usual,  $\nabla\nabla f$  represents the Hessian of  $f(\mathbf{x})$  with respect to  $\mathbf{x}$ , and  $\mathbf{O}_l$  denotes the  $l \times l$  zero matrix.

Further, by virtue of eq.(5.26b), the second term of the above expansion vanishes, the expansion thus reducing to

$$\Delta\Delta F = \frac{1}{2} \Delta\mathbf{x}_F^T \left[ \nabla\nabla f + \frac{\partial(\mathbf{J}^T \boldsymbol{\lambda})}{\partial \mathbf{x}} \right] \Delta\mathbf{x}_F \quad (5.27)$$

which is a quadratic form associated with the  $n \times n$  matrix  $\mathbf{H}_c$ , defined as

$$\mathbf{H}_c \equiv \nabla \nabla f + \frac{\partial(\mathbf{J}^T \boldsymbol{\lambda})}{\partial \mathbf{x}} \quad (5.28)$$

which thus plays the role of the Hessian of the objective function under constrained minimization. For this reason, we term this matrix the *constrained Hessian*. Notice that  $\mathbf{H}_c$  can be expressed in a more illustrative form, namely,

$$\mathbf{H}_c = \nabla \nabla f + \nabla \nabla (\mathbf{h}^T \boldsymbol{\lambda}) \quad (5.29)$$

Moreover, in light of the smoothness assumptions on both  $f$  and  $\mathbf{h}$ , the two Hessians appearing in the right-hand side of eq.(5.29) are symmetric. Therefore, the constrained Hessian is necessarily symmetric as well, its eigenvalues thus being real and its eigenvectors mutually orthogonal. Hence, the SONC can be stated in terms of the sign-definition of the constrained Hessian, to read:

A stationary point  $\mathbf{x}_o$  of the constrained problem (5.1) is

- (i) a *local minimum* if the constrained Hessian is *positive-definite*;
- (ii) a *local maximum* if the constrained Hessian is *negative-definite*;

#### Remarks:

1. The above conditions are *sufficient*, but not necessary, as we have to ensure that  $\mathbf{x}_F$  is *feasible*<sup>3</sup>;
2. the verification of the SONC requires the computation of: the unconstrained Hessian  $\nabla \nabla f$  of the objective function; the Hessian of each scalar constraint; and the Lagrange multipliers;
3. the verification of the SONC requires, additionally, the solution of an eigenvalue problem associated with a  $n \times n$  symmetric matrix, although the problem, in fact, involves only  $n - l$  independent variables.
4. if the constrained Hessian is sign-indefinite, no conclusion can be drawn about the nature of the SP at hand, as the Hessian eigenvalues are distributed between the feasible and the unfeasible regions.

---

<sup>3</sup>This point is made apparent in Example 5.3.1

We can ensure that  $\Delta \mathbf{x}_F$  is feasible if we define it as the image of a  $(n - l)$ -dimensional vector  $\Delta \mathbf{u}$  under an orthogonal complement  $\mathbf{L}$  of  $\mathbf{J}$ :

$$\Delta \mathbf{x}_F = \mathbf{L} \Delta \mathbf{u} \quad (5.30)$$

If now this expression is substituted into eq.(5.27), an alternative expression for the second variation is derived, namely,

$$\Delta \Delta F = \frac{1}{2} \Delta \mathbf{u}^T \mathbf{L}^T [\nabla \nabla f + \nabla \nabla (\mathbf{h}^T \boldsymbol{\lambda})] \mathbf{L} \Delta \mathbf{u} \quad (5.31)$$

where we have recalled that the second term inside the brackets in eq.(5.27) is the Hessian of  $\mathbf{h}^T \boldsymbol{\lambda}$ , thereby deriving the  $(n - l) \times (n - l)$  matrix  $\mathbf{H}_u$ , that we shall call the *reduced Hessian*:

$$\mathbf{H}_u \equiv \mathbf{L}^T [\nabla \nabla f + \nabla \nabla (\mathbf{h}^T \boldsymbol{\lambda})] \mathbf{L} \quad (5.32)$$

Therefore,  $\mathbf{H}_u$  is a  $(n - l) \times (n - l)$  symmetric matrix. The reduced Hessian can be regarded, in fact, as a *projection* of the constrained Hessian onto the space of *increments*  $\Delta \mathbf{u}$  of the feasible variables. We can thus state the SONC as:

At a feasible, stationary point  $\mathbf{x}_o$ , the objective function  $f(\mathbf{x})$  attains

- (i) a local minimum iff  $\mathbf{H}_u$  is positive-definite;
- (i) a local maximum iff  $\mathbf{H}_u$  is negative-definite;
- (i) a saddle point iff  $\mathbf{H}_u$  is sign-indefinite.

Notice that the above conditions are necessary and sufficient.

**Example 5.3.1 (Taken from Luenberger (1984))** Find all the stationary points (SPs) of

$$f = x_1 x_2 + x_2 x_3 + x_1 x_3$$

subject to

$$h(\mathbf{x}) = x_1 + x_2 + x_3 - 3 = 0$$

and identify the nature of each SP, i.e., maximum, minimum or saddle point.

*Solution:* We start by calculating  $\nabla f$ ,  $\mathbf{J}^T = \nabla h$  and  $\nabla \nabla f$ :

$$\nabla f = \begin{bmatrix} x_2 + x_3 \\ x_1 + x_3 \\ x_1 + x_2 \end{bmatrix}, \quad \nabla \nabla f = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \quad \nabla h = \mathbf{J}^T = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

whence  $\mathbf{J}$  is *constant*. The four FONC of the Lagrangian are

$$x_2 + x_3 + \lambda = 0$$

$$x_1 + x_3 + \lambda = 0$$

$$x_1 + x_2 + \lambda = 0$$

$$x_1 + x_2 + x_3 = 3$$

with only one Lagrange multiplier, given that there is one single constraint. Moreover, the FONC form a determined system of linear equations, and hence, they admit, if the coefficient matrix is not singular, one *unique* solution.

Summation of the first three equations leads to

$$2(x_1 + x_2 + x_3) = -3\lambda$$

which, by virtue of the fourth equation, leads to  $\lambda = -2$ . Hence,  $x_1 = x_2 = x_3 = 1$  is the solution of the foregoing equations. Now we verify the FONC in dual form. To this end, we first find an orthogonal complement of  $\mathbf{J}$ . Given that  $\mathbf{J}$  is a  $1 \times 3$  matrix in this case, i.e., a row vector, an orthogonal complement  $\mathbf{L}$  of  $\mathbf{J}$  is a  $3 \times 2$  matrix, whose two three-dimensional columns are orthogonal to the single row of  $\mathbf{J}$ . A suitable candidate is given below, normalized so that its two columns be unit vectors, although this normalization is not essential. It is convenient from a numerical viewpoint, though, in order to avoid multiplication by too large or too small numbers. Thus,

$$\mathbf{L} = \frac{\sqrt{2}}{2} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}$$

Further,

$$\mathbf{J}^T \boldsymbol{\lambda} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \lambda = \begin{bmatrix} \lambda \\ \lambda \\ \lambda \end{bmatrix} \Rightarrow \frac{\partial(\mathbf{J}^T \boldsymbol{\lambda})}{\partial \mathbf{x}} = \mathbf{O}_{33}$$

It is noteworthy that the second term of the constrained Hessian vanishes because  $\mathbf{J}$  is constant, a consequence of the linearity of the constraint  $h$  in the design vector  $\mathbf{x}$ . We can thus state that

*In the case of linear equality constraints  $\mathbf{h}(\mathbf{x})$ , the constrained Hessian is identical to its unconstrained counterpart.*

Therefore,

$$\nabla\nabla f + \frac{\partial(\mathbf{J}^T\boldsymbol{\lambda})}{\partial\mathbf{x}} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

whose eigenvalues are  $\{-1, -1, 2\}$ , and hence, the constrained Hessian is sign-indefinite. We cannot thus decide on the nature of the SP, for which reason we resort to the reduced Hessian, which is

$$\begin{aligned} \mathbf{H}_u &= \mathbf{L}^T \left( \nabla\nabla f + \frac{\partial(\mathbf{J}^T\boldsymbol{\lambda})}{\partial\mathbf{x}} \right) \mathbf{L} \\ &= \frac{1}{2} \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \underbrace{\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}}_{\begin{bmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{bmatrix}} = \frac{1}{2} \begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix} \end{aligned}$$

and is hence constant throughout the whole feasible manifold. Moreover, notice that: a)  $\mathbf{H}_u$  is a  $2 \times 2$  symmetric matrix, and hence, has two real eigenvalues; b)  $\text{tr}(\mathbf{H}_u) = -2 < 0$ ; and c)  $\det(\mathbf{H}_u) = 3/4 > 0$ . From b) it is apparent that the larger eigenvalue is negative; from c), it is apparent that the two eigenvalues bear the same sign, and hence, the two eigenvalues of  $\mathbf{H}_u$  are negative, the reduced Hessian thus being negative-definite, and the unique SP is, consequently, a maximum. Moreover, this is the *global maximum* of the given  $f(\mathbf{x})$ .

### Example 5.3.2 (The Design of a Positioning Robot for a Given Reach)

We revisit here the design of the three-joint robot of Fig. 4.2, reproduced as Fig. 5.2 for quick reference. In Example 4.5.1, an expression was found for the maximum reach of the robot as a linear function of  $a$ . The same reach is found here using a constrained-optimization approach, with the purpose of finding the value of the length  $a$  that will produce the maximum reach of a Puma 560 robot, namely, 0.8772 m. Thus, the global maximum of the distance  $d$  of the robot operation point  $C$  from the  $Z_1$ -axis has to be found. This can be done by: (i) finding all stationary points of  $d$ ; (ii) identifying each SP as a local maximum, minimum or saddle point; and (iii) upon comparing all local maxima, select the maximum maximorum.

*Solution:* It is apparent that the maximum reach is independent of  $\theta_1$ , the angle of rotation of the first joint, for motions about the first joint do not affect the reach. So,

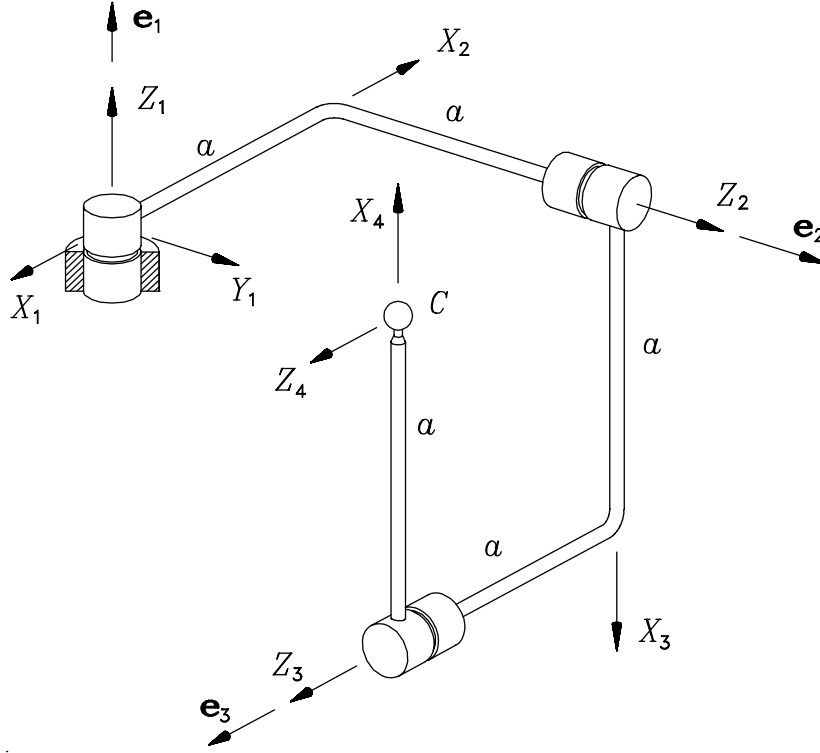


Figure 5.2: Manipulator configuration for  $C(0, a, 0)$ .

we lock the first joint and, in the posture of Fig. 5.2, rotate the third joint through one full turn, point  $C$  thus describing a circle  $\mathcal{C}$  of radius  $a$  lying in the  $Y_1$ - $Z_1$  plane, with centre at point  $O'_3$  of coordinates  $(0, a, -a)$ . Next, upon going through a full rotation of the second joint, the circle describes a toroid—not a torus, for this is generated by a circle turning about an axis contained in the plane of the circle—of axis  $Z_2$ , the problem now reducing to one of finding the point of the surface of the toroid lying the farthest from the  $Z_1$  axis. Figure 5.3 includes side views of circle  $\mathcal{C}$ .

Let the trace of the toroid with the  $X_2$ - $Z_2$  plane be the contour  $\mathcal{T}$  of Fig. 5.4. It is most convenient to represent this contour with the aid of the non-dimensional variables  $u$  and  $v$ , which are defined as

$$u \equiv \frac{x_2}{a}, \quad v \equiv \frac{z_2}{a} \quad (5.33)$$

In terms of these new variables, the equation of  $\mathcal{T}$  becomes

$$\mathcal{T}: \quad h(u, v) \equiv (u^2 + v^2)(u^2 + v^2 - 4v) - 4(u^2 - v^2 - 1) = 0 \quad (5.34)$$

The contour  $\mathcal{T}$  defined by the implicit function  $h(u, v) = 0$  is displayed in Fig. 5.4. The distance of point  $C$  from the  $Z_1$ -axis can be shown to be, in non-dimensional



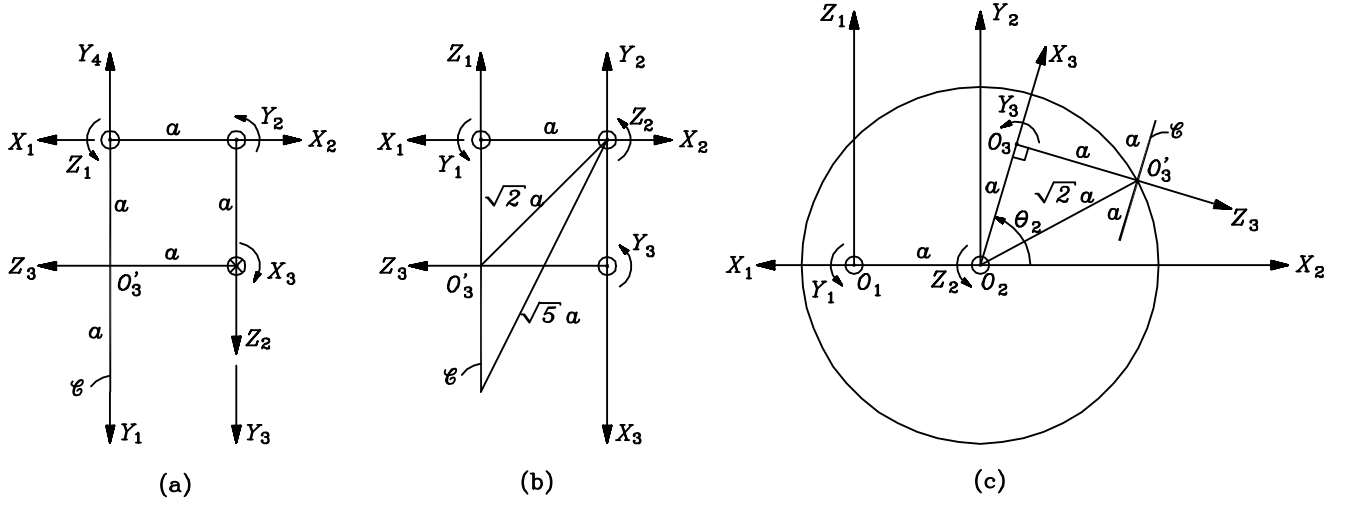


Figure 5.3: Side views of circle  $\mathcal{C}$ : (a) and (b) at the position of Fig. 5.2; and (c) at an arbitrary position for a given value of  $\theta_2$

coordinates  $u$  and  $v$ ,  $d = (u + 1)^2 + v^2$ . Now, the maximum distance  $r_M$  of  $O_1$  to  $\mathcal{T}$  can be found as the solution of the optimization problem defined below:

$$f(u, v) \equiv \frac{1}{2}[(u + 1)^2 + v^2] \rightarrow \max_{u, v} \quad (5.35)$$

subject to eq.(5.34). Notice that, rather than maximizing  $d$ , we aim to maximize  $d/2$ , which amounts exactly to the same goal, the difference being that, upon differentiation, the factor  $1/2$  will help eliminate an inconvenient factor of 2 in the gradient of  $f(\mathbf{x})$ .

We thus have an equality-constrained maximization problem. In order to find the normality conditions of this problem, we resort to Lagrange multipliers, thus defining a new, unconstrained, maximization problem:

$$F(u, v, \lambda) \equiv f + \lambda h \rightarrow \max_{u, v, \lambda} \quad (5.36)$$

The normality conditions of the foregoing problem are, thus,

$$\frac{\partial F}{\partial u} \equiv u + 1 + 4\lambda u(u^2 + v^2 - 2v - 2) = 0 \quad (5.37a)$$

$$\frac{\partial F}{\partial v} \equiv v + 4\lambda(v - 1)(u^2 + v^2 - 2v) = 0 \quad (5.37b)$$

$$\frac{\partial F}{\partial \lambda} \equiv (u^2 + v^2)(u^2 + v^2 - 4v) - 4(u^2 - v^2 - 1) = 0 \quad (5.37c)$$

the last equation being just a restatement of the constraint, eq.(5.34). Now we eliminate  $\lambda$ , the Lagrange multiplier, *dialytically* (Salmon, 1885) from eqs.(5.37a & b). We do this by rewriting these two equations in linear homogeneous form in the “variables”  $\lambda$  and 1, namely,

$$\begin{bmatrix} 4u(u^2 + v^2 - 2v - 2) & u + 1 \\ 4(v - 1)(u^2 + v^2 - 2v) & v \end{bmatrix} \begin{bmatrix} \lambda \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (5.38a)$$

Obviously, the foregoing equation requires a nontrivial solution—note that one component of the vector of “unknowns” is unity!—which in turn requires that the coefficient matrix be singular, i.e.,

$$\det \begin{bmatrix} 4u(u^2 + v^2 - 2v - 2) & u + 1 \\ 4(v - 1)(u^2 + v^2 - 2v) & v \end{bmatrix} = 0 \quad (5.38b)$$

Upon expansion,

$$4u(u^2 + v^2 - 2v - 2)v - 4(v - 1)(u^2 + v^2 - 2v)(u + 1) = 0$$

or

$$\mathcal{S}: \quad u^3 - v^3 - u^2v + uv^2 + u^2 + 3v^2 - 4uv - 2v = 0 \quad (5.38c)$$

Now, the maximum reach is found via the solution of the system of *polynomial equations* (5.34) and (5.38c). The former is a quartic equation, the latter cubic. The *Bezout number*<sup>4</sup> of the foregoing system of equations is defined as the product of the degrees of those equations, i.e.,  $4 \times 3 = 12$ , which gives an upper bound of 12 for the number of solutions, both real and imaginary, of the problem at hand. One graphical means of obtaining estimates of the real solutions of this system consists in plotting the two corresponding contours in the  $u$ - $v$  plane, as shown in Fig. 5.5. The global maximum reach occurs apparently, at point  $A$ , of *approximate* coordinates (2.2, 1.4), estimated by inspection, which leads to a visual estimate of  $r_M$ , namely,

$$r_M \approx 3.5a \quad (5.39)$$

The four intersections of these two curves correspond to four stationary values of the distance from a point in the trace  $\mathcal{T}$  to the point  $O_1$  in the  $u$ - $v$  plane. Of these four intersections, two are local maxima and two local minima. The **normality** of  $\nabla f$ —which in this case is identical to the vector from  $O_1$  to  $\mathcal{T}$ —at the intersection points, is to be highlighted.

---

<sup>4</sup>To define the Bezout number of a system of  $p$  polynomial equations in  $p$  variables  $x_1, x_2, \dots, x_p$ , we look first at the  $i$ th equation: A typical term of this equation involves the product  $x_1^{d_{1i}} x_2^{d_{2i}} \dots x_p^{d_{pi}}$ . The degree  $d_i$  of this equation is the maximum of  $d_{1i} + d_{2i} + \dots + d_{pi}$ , for  $i = 1, \dots, N_i$ , where  $N_i$  denotes the number of terms of the  $i$ th equation. The Bezout number  $N_B$  of this system is defined as  $N_B = d_1 d_2 \dots d_p$ .

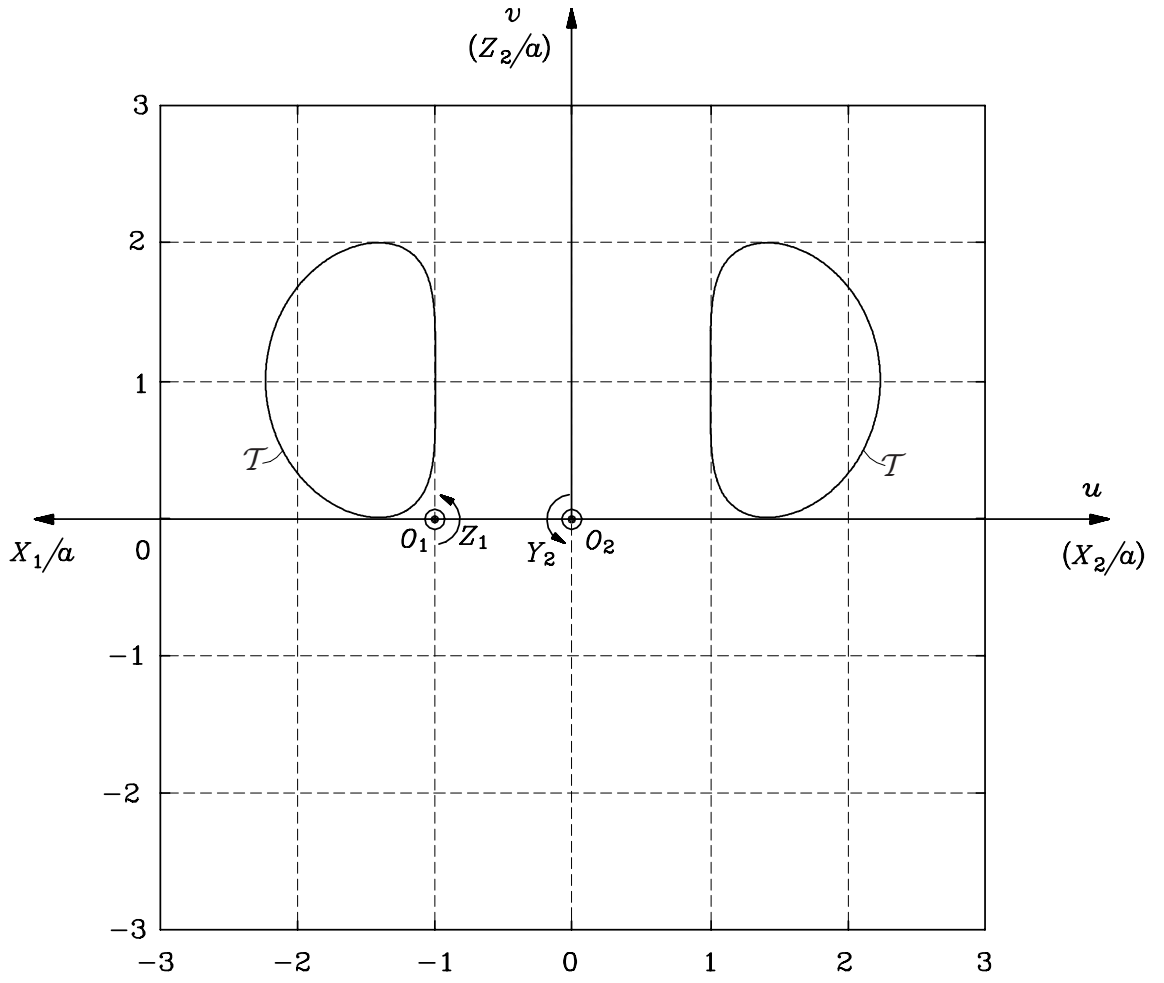


Figure 5.4: Contour of the trace  $\mathcal{T}$  of the toroid on the  $u$ - $v$  plane

The foregoing system is solved more precisely using computer algebra, thus obtaining the four real solutions given below:

$$\begin{aligned}
 (u)_A &= 2.132242, & (v)_A &= 1.4689944 \\
 (u)_B &= -1.578095, & (v)_B &= 1.975316 \\
 (u)_C &= -1.132242, & (v)_C &= 0.116796 \\
 (u)_D &= 1.025308, & (v)_D &= 0.366325
 \end{aligned}$$

which lead to *stationary* reach values of

$$r_A = 3.459606a, \quad r_B = 2.058171a, \quad r_C = 0.176435a, \quad r_D = 2.058171a$$

for a global maximum reach of

$$r_M = 3.459606a$$

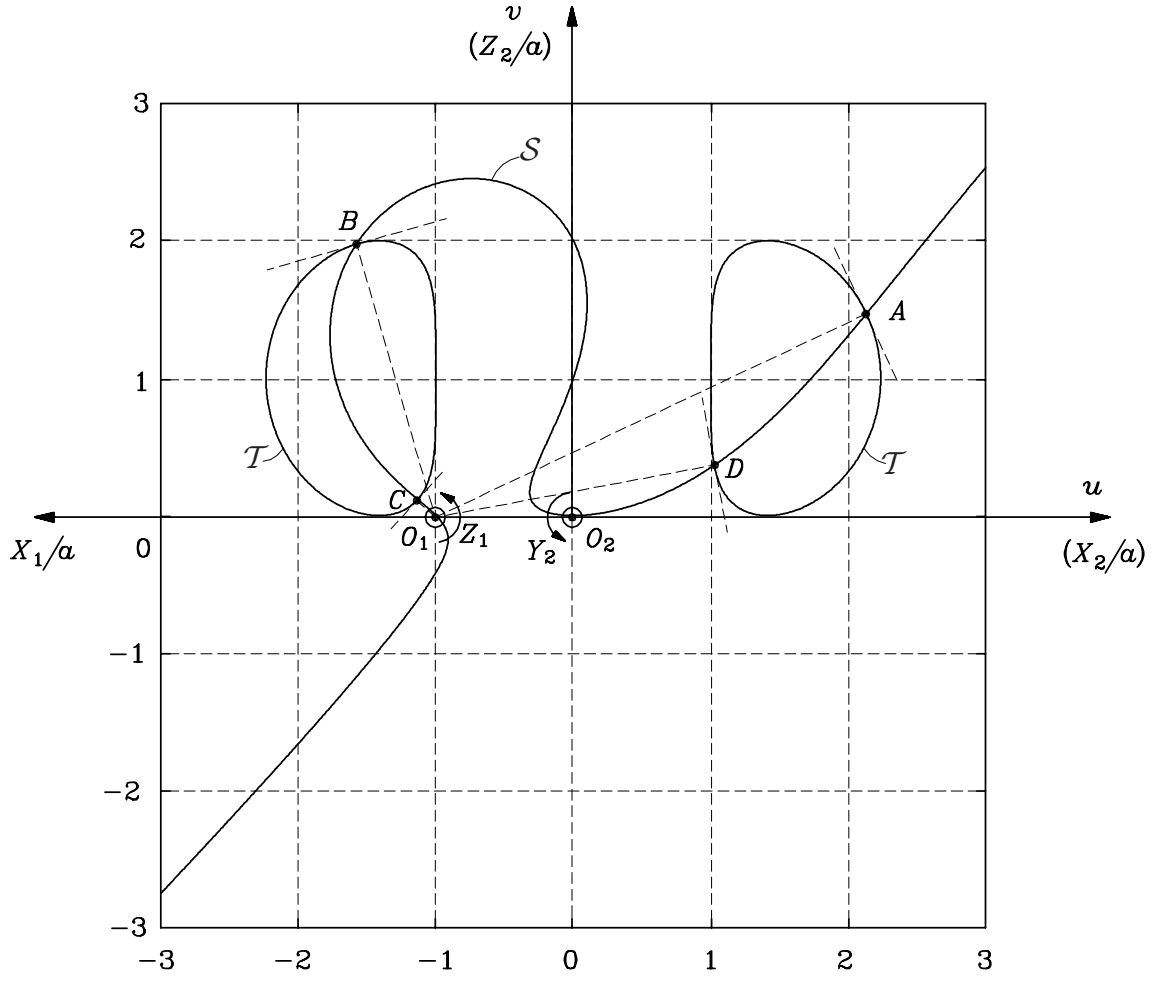


Figure 5.5: Plots of the two contours,  $\mathcal{S}$  and  $\mathcal{T}$

The value of  $a$  that will yield the foregoing maximum reach is thus found as

$$3.460a = 0.8772 \quad \Rightarrow \quad a = 0.2535 \text{ m}$$

thereby completing the solution.

The first- and second-order normality conditions are next verified. It is apparent that points  $A$  and  $B$  yield local maxima,  $C$  and  $D$  local minima. We verify the normality conditions for each point. First, we list below the general expressions for the various items needed in the sequel:

$$\nabla f = \begin{bmatrix} u+1 \\ v \end{bmatrix}, \quad \mathbf{J} = (\nabla h)^T = [4u(u^2 + v^2 - 2v - 2) \quad 4(v-1)(u^2 + v^2 - 1)],$$

$$\nabla \nabla f = \mathbf{1}, \quad \frac{\partial(\mathbf{J}^T \boldsymbol{\lambda})}{\partial \mathbf{x}} \equiv \frac{\partial \lambda \nabla h}{\partial \mathbf{x}} = \lambda \nabla \nabla h$$

i.e.,

$$\frac{\partial(\mathbf{J}^T \boldsymbol{\lambda})}{\partial \mathbf{x}} = \lambda \begin{bmatrix} 12u^2 + 4v^2 - 8v - 8 & 2u(2v - 4) + 4uv \\ 2u(2v - 4) + 4uv & 4u^2 + 4v^2 - 8v + 4v(2v - 4) + 8 \end{bmatrix}$$

the expression for  $\partial(\mathbf{J}^T \boldsymbol{\lambda})/\partial \mathbf{x}$  in terms of the Hessian  $\nabla \nabla h$  following in this particular case because  $\lambda$  is a scalar. Moreover, numerical values for  $\lambda$ , at each of the four stationary points, will be needed. These are computed from the FONC in primal form, eq.(5.9):

$$\begin{bmatrix} 4u(u^2 + v^2 - 2v - 2) \\ 4(v - 1)(u^2 + v^2 - 1) \end{bmatrix} \lambda = - \begin{bmatrix} u + 1 \\ v \end{bmatrix}$$

which is, as expected, an overdetermined system of two linear equations in one single unknown,  $\lambda$ . The latter is computed by means of the left Moore-Penrose generalized inverse of  $\mathbf{J}^T$ , as in eq.(5.10)<sup>5</sup>, using *computer algebra*,

$$\lambda = -\frac{1}{4} \frac{N}{D}$$

where

$$\begin{aligned} N &= u^4 + u^3 + 2vu^2v^2 + uv^2 - 3vu^2 - 2uv - 2u^2 - 2u + v^4 - 3v^3 + 2v^2 \\ D &= u^6 + 3u^4v^2 - 6u^4v - 3u^4 + 3u^2v^4 - 12u^2v^3 + 4vu^2 + 4u^2 + 13v^4 \\ &\quad + 4v^2 + v^6 - 6v^5 - 12v^3 + 10u^2v^2 \end{aligned}$$

At this stage the only outstanding item is an orthogonal complement of  $\mathbf{J}$ . As  $\mathbf{J}$  is of  $1 \times 2$ , it can be regarded as a two-dimensional vector, and hence, an orthogonal complement thereof is readily obtained as a  $2 \times 1$  matrix, i.e., a two-dimensional column vector premultiplied by the  $2 \times 2$  matrix  $\mathbf{E}_2$  introduced in eq.(3.15c). Indeed, this matrix rotates two-dimensional vectors through  $90^\circ$  counterclockwise. We thus define

$$\mathbf{L} = \mathbf{E}_2 \mathbf{J}^T = \begin{bmatrix} -4(v - 1)(u^2 + v^2 - 1) \\ 4u(u^2 + v^2 - 2v - 2) \end{bmatrix} \quad (5.40)$$

which, upon normalization (division by  $\|\mathbf{L}\|$ ), becomes

$$\mathbf{L} \leftarrow \frac{1}{D} \begin{bmatrix} -4(v - 1)(u^2 + v^2 - 1) \\ 4u(u^2 + v^2 - 2v - 2) \end{bmatrix} \quad (5.41)$$

with  $D$  given by

$$D \equiv 4\sqrt{((v - 1)(u^2 + v^2 - 1))^2 + (u(u^2 + v^2 - 2v - 2))^2}$$

---

<sup>5</sup>While the verbatim use of the formula of this equation is not recommended for numerical computations, because of the inherent round-off error amplification, the use of the formula here is safe, as all computations are done *symbolically*, and hence, without roundoff error.

Table 5.1: Summary of FONC & SONC for Example 5.3.2

SP	$A$	$B$
$f$	5.984442	2.118034
$h$	0.000033	-0.000002
$\nabla f$	$[3.132242, 1.468994]^T$	$[-0.578095, 1.975316]^T$
$\mathbf{J}$	$[15.065669, 7.065704]$	$[-2.787706, 9.525424]$
$\mathbf{L}^T$ (norm'zed)	$[-0.424615, 0.905374]$	$[-0.959744, -0.280878]$
$\nabla\nabla f$	$\mathbf{1}_{22}$	$\mathbf{1}_{22}$
$\mathbf{H}_c$	$\begin{bmatrix} -8.030863 & -1.663262 \\ -1.663262 & -2.498075 \end{bmatrix}$	$\begin{bmatrix} -3.497832 & 2.553411 \\ 2.553411 & -2.603403 \end{bmatrix}$
$\mathbf{H}_u$	-2.216791	-2.050618

SP	$C$	$D$
$f$	0.015565	2.118033
$h$	0.000001	0.000002
$\nabla f$	$[-0.132242, 0.116796]^T$	$[2.025308, 0.366325]^T$
$\mathbf{J}$	$[4.248076, -3.751926]$	$[-6.345424, -1.147713]$
$\mathbf{L}^T$ (norm'zed)	$[0.661981, 0.749521]$	$[0.177985, -0.9840333]$
$\nabla\nabla f$	$\mathbf{1}_{22}$	$\mathbf{1}_{22}$
$\mathbf{H}_c$	$\begin{bmatrix} 1.202463 & 0.249038 \\ 0.249038 & 1.326504 \end{bmatrix}$	$\begin{bmatrix} 1.708971 & -1.658981 \\ -1.658981 & 2.603399 \end{bmatrix}$
$\mathbf{H}_u$	1.519277	3.156182

Using the above expressions, the various items appearing in the FONC and SONC are now calculated at each SP. These are displayed in Table 5.1.

From the numerical results recorded in Table 5.1, it is apparent that:

The constraint is respected to the precision of the data, as the values of  $h$  at the four stationary points is smaller than  $10^{-4}$ ;

the maximum maximorum, i.e., the global maximum of the distance  $d$  is found at the SP  $A$ ;

$\nabla f$  does not vanish at any of the SPs, but it need not vanish anyway; however, this vector lies in the range of  $\mathbf{J}$  at the four SP, which is reflected by the proportionality between the two components of the two arrays;

$\nabla\nabla f$  equals the  $2 \times 2$  identity matrix *everywhere* in the  $u$ - $v$  plane, but this value alone doesn't give information on the nature of each SP. What gives this information is the reduced Hessian  $\mathbf{H}_c$ , which turns out to be a  $1 \times 1$  matrix, i.e., a scalar; this is negative at SPs  $A$  and  $B$ , thereby indicating local maxima, and positive at the two other SPs, which indicates local minima. The global minimum is attained at  $C$ . Interestingly, for this example, the smaller local maximum equals the greater local minimum, up to the ninth digit. This means that both SPs are, in fact, identical.

A 3D visualization of the toroid defining the curve  $\mathcal{T}$  is included in Fig. 5.6, while the intersection of the paraboloid  $f(u, v) = (1/2)[(u + 1)^2 + v^2]$  with the cylindrical surface generated by  $h = (u^2 + v^2)(u^2 + v^2 - 4v) - 4(u^2 - v^2 - 1)$  is illustrated in Fig. 5.7.

A Maple worksheet, `x5-3-1-fonc+sonc.mw`, is available on the course website with all calculations leading to the results of the above table.

### Example 5.3.3 (The Equilibrium Configuration of a Four-Link Chain)

*We consider here the problem of determining the equilibrium configuration of a chain composed of four identical links of length  $\ell$  each, suspended at two points located at the same level, a distance  $d$  apart, as illustrated in Fig. 5.8. This problem was proposed by Luenberger (1984) to illustrate methods of nonlinear programming. Here, we use a simplified version of this problem with the purpose of obtaining a solution by simple equation-solving. The principle determining the equilibrium configuration is one of minimum energy. This leads to the need of finding the stationary points of its potential energy, and then finding the nature of these points.*

At the outset, we exploit the symmetry of the problem, which enables us to reduce the number of design variables to only two, namely, the inclination of the two links on the left half of the chain. Let  $\theta_i$ , for  $i = 1, 2$ , denote the angle made by the axis of the  $i$ th link from the vertical and  $\mu$  denote the mass distribution per unit length, while  $g$  represents the gravity acceleration. The potential energy  $V$  of the whole chain is, thus, for an arbitrary configuration of the chain,

$$V(\theta_1, \theta_2) = -2\mu g \ell \left( \frac{1}{2} \cos \theta_1 + \cos \theta_1 + \frac{1}{2} \cos \theta_2 \right)$$

which attains a minimum at an equilibrium configuration. However, notice that the two design variables are not independent, for their horizontal span must be exactly

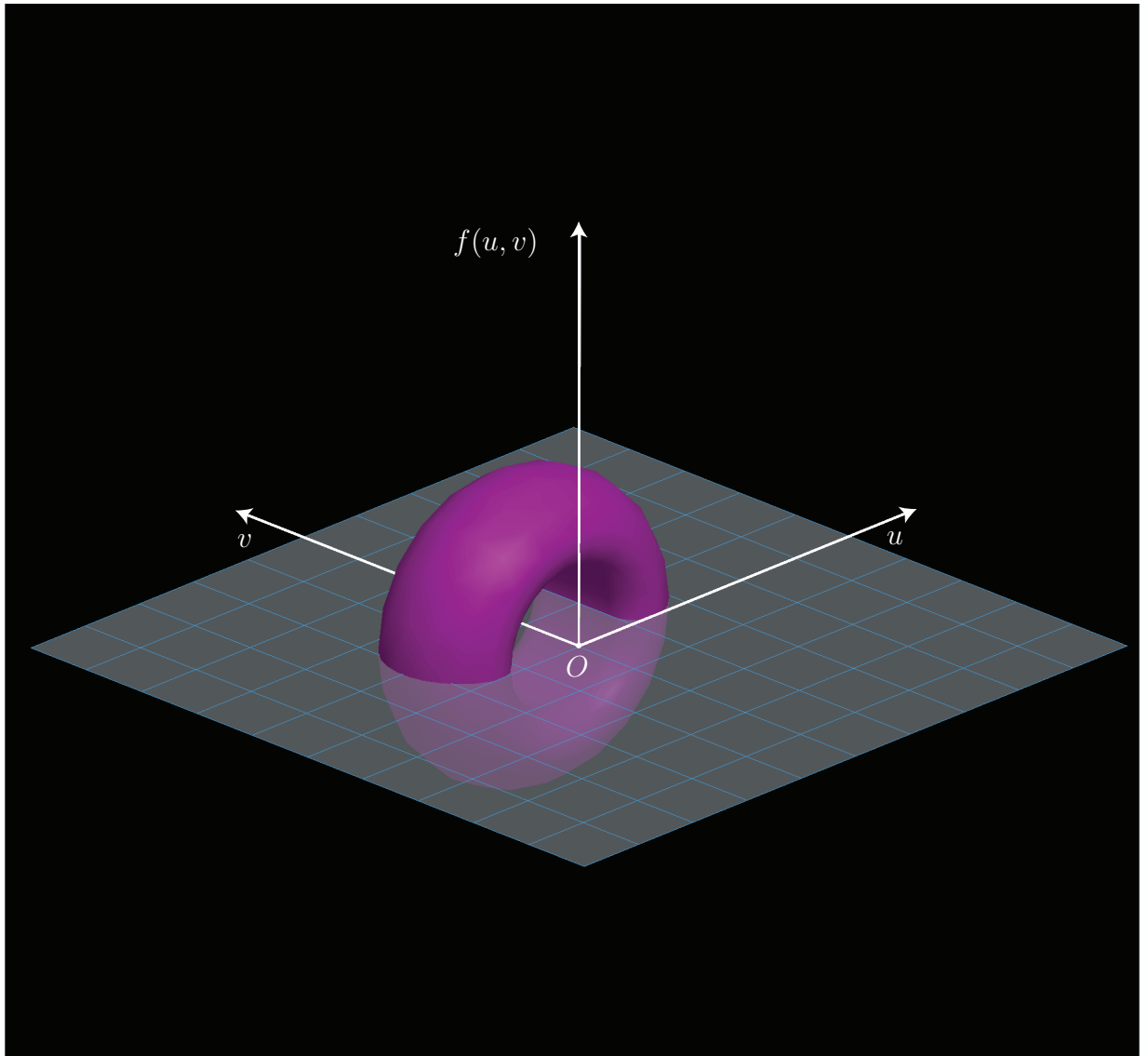


Figure 5.6: The torus generated by the last two joints of the given robot



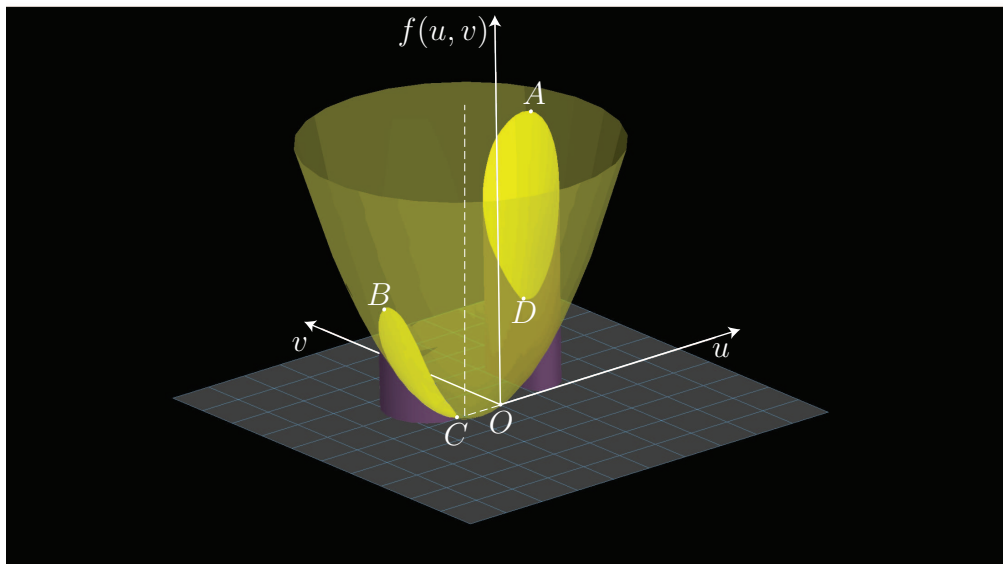


Figure 5.7: The distance paraboloid and its intersection with the cylindrical surface generated by  $h(\mathbf{x}) = 0$

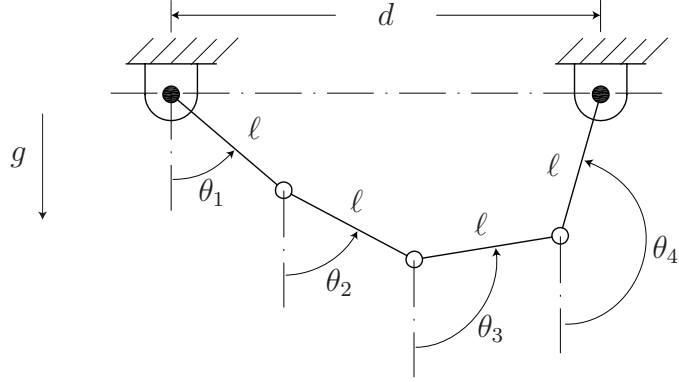


Figure 5.8: A four-link chain

$d/2$ , i.e.,

$$\ell(\sin \theta_1 + \sin \theta_2) - \frac{d}{2} = 0$$

The optimum design problem at hand now has the form

$$f(\theta_1, \theta_2) \equiv \frac{V}{\mu g \ell} = -3 \cos \theta_1 - \cos \theta_2 \quad \rightarrow \quad \min_{\theta_1, \theta_2}$$

subject to

$$h(\theta_1, \theta_2) = \sin \theta_1 + \sin \theta_2 - p = 0, \quad p \equiv \frac{d}{2\ell}$$

The Lagrangian of the problem is to be minimized, i.e.,

$$F(\theta_1, \theta_2) = -3 \cos \theta_1 - \cos \theta_2 + \lambda(\sin \theta_1 + \sin \theta_2 - p) \quad \rightarrow \quad \min_{\theta_1, \theta_2, \lambda}$$

subject to no constraints. The normality conditions of the unconstrained problem are, thus

$$\begin{aligned} \frac{\partial F}{\partial \theta_1} &= 3 \sin \theta_1 + \lambda \cos \theta_1 = 0 \\ \frac{\partial F}{\partial \theta_2} &= \sin \theta_2 + \lambda \cos \theta_2 = 0 \\ \frac{\partial F}{\partial \lambda} &= \sin \theta_1 + \sin \theta_2 - p = 0 \end{aligned}$$

The problem has thus been reduced to solving the foregoing system of three nonlinear equations in three unknowns,  $\theta_1$ ,  $\theta_2$  and  $\lambda$ . While this nonlinear system can be solved using the Newton-Raphson method, the simplicity of the equations lends itself to a more comprehensive approach. Indeed, the Newton-Raphson method yields one single solution at a time, the user never knowing whether any other solutions exist.

Moreover, there is no guarantee that the solution found is a minimum and not a maximum or a saddle point.

For starters, we can eliminate  $\lambda$  from the above equations, for it appears linearly in the first two of those. We thus rewrite those two equations in the form

$$\mathbf{A}\boldsymbol{\lambda} = \mathbf{0}_2$$

with  $\mathbf{0}_2$  denoting the two-dimensional zero vector, while  $\mathbf{A}$  and  $\boldsymbol{\lambda}$  are defined as

$$\mathbf{A} \equiv \begin{bmatrix} \cos \theta_1 & 3 \sin \theta_1 \\ \cos \theta_2 & \sin \theta_2 \end{bmatrix}, \quad \boldsymbol{\lambda} \equiv \begin{bmatrix} \lambda \\ 1 \end{bmatrix} \neq \mathbf{0}_2$$

Since the solution  $\boldsymbol{\lambda}$  sought cannot be zero, the above homogeneous system must admit a nontrivial solution, which calls for  $\mathbf{A}$  to be singular, i.e.,

$$\Delta(\theta_1, \theta_2) \equiv \det(\mathbf{A}) = 0$$

Upon expansion,

$$\Delta(\theta_1, \theta_2) = \cos \theta_1 \sin \theta_2 - 3 \sin \theta_1 \cos \theta_2 = 0$$

which we shall call the *reduced normality condition*. We thus have eliminated  $\lambda$  *dialytically* (Salmon, 1885), the problem thus reducing to a system of two equations in two unknowns,  $h(\theta_1, \theta_2) = 0$  and  $\Delta(\theta_1, \theta_2) = 0$ . We can further reduce the same system to one single equation in one single unknown, which can be done by dialytic elimination as well. However, notice that dialytic elimination is applicable to systems of polynomial equations, while the two equations at hand are not polynomial; they are trigonometric. Nevertheless, by application of the well-known trigonometric “tan-half” identities:

$$\cos x \equiv \frac{1 - T^2}{1 + T^2}, \quad \sin x \equiv \frac{2T}{1 + T^2}, \quad T \equiv \tan\left(\frac{x}{2}\right)$$

the two equations can be transformed into polynomial equations. We will not pursue here this elimination procedure. Instead, we plot the two foregoing functions in the  $\theta_1$ - $\theta_2$  plane, the solutions sought being found visually at the intersection of the corresponding contours. In order to plot the contours, however, we must assign a numerical value to parameter  $p$ . By assuming  $d = 1.25$  m and  $\ell = 0.5$  m, we obtain  $p = 1.25$ . These contours are plotted in Fig. 5.9.

The contours apparently intersect at two points, of coordinates estimated visually at

$$\theta_1 = 0.45, \quad \theta_2 = 1.00 \quad \text{and} \quad \theta_1 = 2.70, \quad \theta_2 = 2.20$$

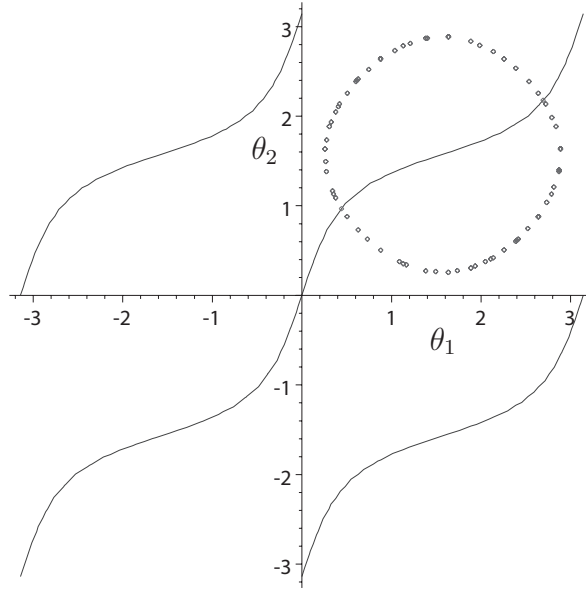


Figure 5.9: The reduced normality condition and the equality constraint (dashed contour)

with all values in radians. These values are quite rough. More precise values can be obtained by means of Newton-Raphson's method applied to the two nonlinear equations, using the foregoing estimates as initial guesses. Alternatively, the two equations can be solved dialytically by means of computer algebra. For example, upon invoking Maple's `solve` procedure, the real roots below were reported:

$$\theta_1 = 0.4449420670, \theta_2 = 0.9607027573 \quad \text{and} \quad \theta_1 = 2.696650587, \theta_2 = 2.180889896$$

Translated into degrees, the foregoing angles read:

$$\theta_1 = 25.49330256^\circ, \theta_2 = 55.04421335^\circ \quad \text{and} \quad \theta_1 = 154.5066974^\circ, \theta_2 = 124.9557866^\circ$$

The first solution corresponds, apparently, to a minimum, the second to a maximum. If this is the case, then the sum of the corresponding roots for the two solutions should be  $\pi$ , which is the case. Moreover, upon evaluation of the objective function at the two solutions, we obtain

$$\begin{aligned} f(0.4449420670, 0.9607027573) &= -1.640425478 \\ f(2.696650587, 2.180889896) &= 1.640425479 \end{aligned}$$

which clearly shows that the first solution is a minimum, the second a maximum. Notice the symmetry of the objective function at the two foregoing extrema.

The reader is invited to verify the first- and second-order normality conditions. The chain at its *lower* equilibrium configuration is displayed in Fig. 5.10. A similar display of the second stationary point should show that the two stationary points are symmetric with respect to the line joining the two supports. The difference between the two equilibrium—stationary—points lies in the sign-definition of the Hessian: that associated with the lower equilibrium of Fig. 5.10 is positive-definite, which means a *stable equilibrium*, or a *minimum*. The Hessian associated with the second SP is negative-definite, which means an *unstable equilibrium*, or a *maximum*.

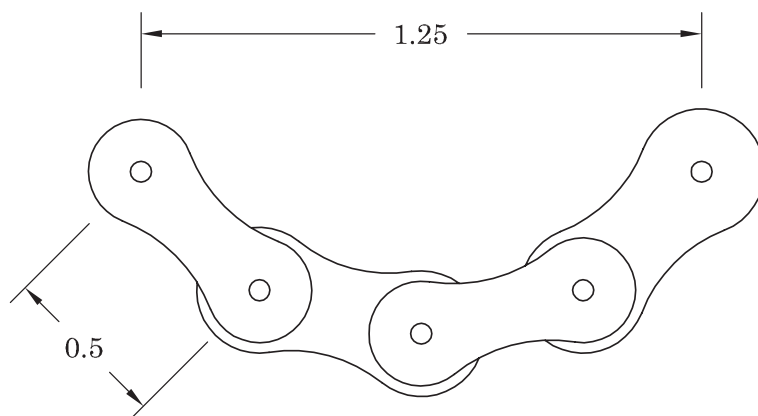


Figure 5.10: The equilibrium configuration of a four-link chain

### 5.3.1 A Mechanical Interpretation of the Lagrange Multipliers

The Lagrange multipliers bear a mechanical interpretation, as illustrated below. This is done with the aid of Example 5.3.3.

The force  $\mathbf{f}$  of *constraint* acting at the bottom joint is first calculated. To this end, the chain is cut into two halves, the left half being illustrated in Fig. 5.11, which depicts the force  $\mathbf{f}$  required to balance the weight of the links. This force is provided by the right half. By symmetry, this force is bound to be horizontal.

The constraint force is found by application of the *Principle of Virtual Work*:  $\sum \pi_i = 0$ , where  $\pi_i$  is the power developed by the  $i$ th external force. Further, let  $\pi_1$  and  $\pi_2$  denote the power developed by the weight of the first and the second links, respectively, while  $\pi_3$  denotes the power developed by the constraint force, i.e.,

$$\pi_1 = \mu \ell \mathbf{g} \cdot \dot{\mathbf{c}}_1, \quad \pi_2 = \mu \ell \mathbf{g} \cdot \dot{\mathbf{c}}_2, \quad \pi_3 = \mathbf{f} \cdot \dot{\mathbf{o}}_2$$

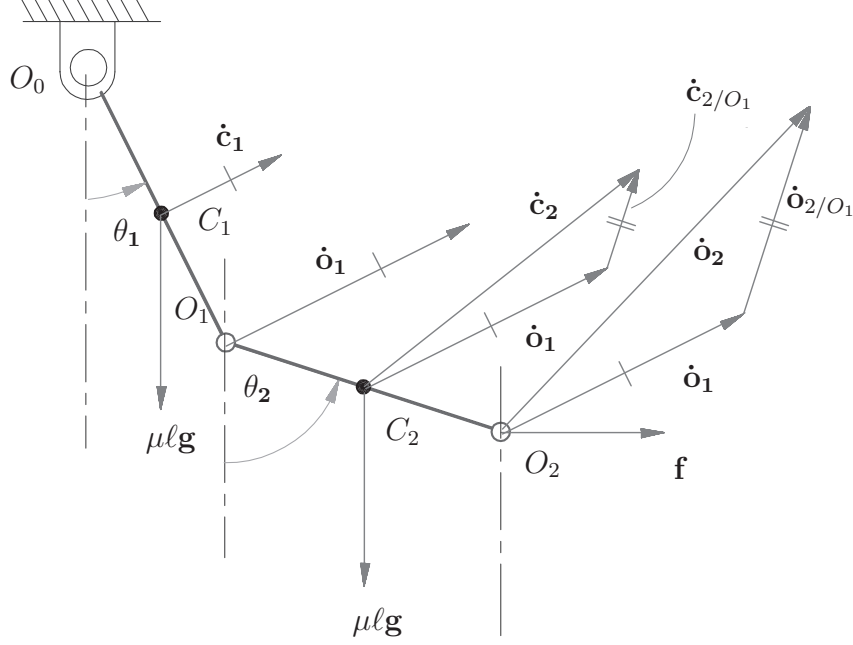


Figure 5.11: The left-hand half of a four-link chain

where, as illustrated in Fig. 5.11,  $\dot{\mathbf{c}}_i$  denotes the velocity of the centre of mass  $C_i$  of link  $i$ , while  $\dot{\mathbf{o}}_i$  denotes the velocity of the  $i$ th joint  $O_i$ , for  $i = 1, 2$ . In the same figure, the velocity triangles of  $C_2$  and  $O_2$  are included.

The power developed by each of the external forces is calculated as the scalar product of the corresponding force by its velocity, and hence,

$$\begin{aligned}\mathbf{g} \cdot \dot{\mathbf{c}}_1 &= -g \frac{\ell}{2} \dot{\theta}_1 \sin \theta_1 \\ \mathbf{g} \cdot \dot{\mathbf{c}}_2 &= \mathbf{g} \cdot (\dot{\mathbf{o}}_1 + \dot{\mathbf{c}}_{2/o_1}) = 2\mathbf{g} \cdot \dot{\mathbf{c}}_1 + \mathbf{g} \cdot \dot{\mathbf{c}}_{2/o_1} = -g\ell \left( \dot{\theta}_1 \sin \theta_1 + \frac{1}{2} \dot{\theta}_2 \sin \theta_2 \right) \\ \mathbf{f} \cdot \dot{\mathbf{o}}_2 &= \mathbf{f} \cdot (\dot{\mathbf{o}}_1 + \dot{\mathbf{o}}_{2/o_1}) = F\ell(\dot{\theta}_1 \cos \theta_1 + \dot{\theta}_2 \cos \theta_2)\end{aligned}$$

where  $F \equiv \|\mathbf{f}\|$ . The Principle of Virtual Work thus yields

$$-\mu g \frac{\ell^2}{2} \dot{\theta}_1 \sin \theta_1 - \mu g \ell^2 \left( \dot{\theta}_1 \sin \theta_1 + \frac{1}{2} \dot{\theta}_2 \sin \theta_2 \right) + F\ell(\dot{\theta}_1 \cos \theta_1 + \dot{\theta}_2 \cos \theta_2) = 0$$

or

$$\left[ -\mu g \ell \left( \frac{1}{2} \sin \theta_1 + \sin \theta_1 \right) + F \cos \theta_1 \right] \dot{\theta}_1 + \left( -\frac{1}{2} \mu g \ell \sin \theta_2 + F \cos \theta_2 \right) \dot{\theta}_2 = 0$$

Since  $\dot{\theta}_1$  and  $\dot{\theta}_2$  are independent, the foregoing relation holds if and only if their respective coefficient vanishes, namely,

$$-\mu g \ell \frac{3}{2} \sin \theta_1 + F \cos \theta_1 = 0 \quad \text{and} \quad -\mu g \ell \frac{1}{2} \sin \theta_2 + F \cos \theta_2 = 0$$

or, with  $\lambda$  defined as

$$\lambda \equiv \frac{F}{\mu g \ell / 2}$$

a simpler form of the foregoing equations is obtained, namely,

$$3 \sin \theta_1 - \lambda \cos \theta_1 = 0 \quad \text{and} \quad \sin \theta_2 - \lambda \cos \theta_2 = 0$$

which are identical to the first two of the FONC of Example 5.3.3. Hence the mechanical interpretation:

*The Lagrange multiplier of the optimization problem under study is proportional to the force required to keep the two halves of the chain together.*

## 5.4 Linear-Quadratic Problems

### 5.4.1 The Minimum-Norm Solution of Underdetermined Systems

We start by recalling a concept of paramount importance in optimization:

**Definition 5.4.1 (Convex set)** *A set of points  $\mathcal{C}$  is convex if, given any two distinct points  $P_1$  and  $P_2$  of the set, then any point  $P$  of  $\mathcal{C}$  comprised between  $P_1$  and  $P_2$  also belongs to the set. Otherwise, the set is nonconvex.*

More formally, if  $\mathbf{x}_i$  denotes the position vector of  $P_i$ , for  $i = 1, 2$ , and  $\mathbf{x}$  that of  $P$ , then, for any scalar  $\alpha$  comprised in the interval  $[0, 1]$ , we can express the position vector of  $P$  as a *convex combination* of those of  $P_1$  and  $P_2$ , namely,

$$\mathbf{x} = \alpha \mathbf{x}_1 + (1 - \alpha) \mathbf{x}_2, \quad 0 \leq \alpha \leq 1 \quad (5.42)$$

We can thus rephrase the definition of convex set as

**Definition 5.4.2 (Convex set—an alternative definition)** *A set of points  $\mathcal{C}$  is convex if, given any two distinct points of position vectors  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , then the point whose position vector is a convex combination of  $\mathbf{x}_1$  and  $\mathbf{x}_2$  also belongs to  $\mathcal{C}$ .*

Germane to the concept of convex set is that defined below:

**Definition 5.4.3 (convex function)** *A function  $f(\mathbf{x})$  is convex if, for any  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , and a  $\mathbf{x}$  defined as a convex combination of  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , and given, e.g., as in eq.(5.42),*

$$f(\mathbf{x}) \leq \alpha f(\mathbf{x}_1) + (1 - \alpha)f(\mathbf{x}_2) \quad (5.43)$$

Now we study the *underdetermined system* of linear equations

$$\mathbf{C}\mathbf{x} = \mathbf{d} \quad (5.44)$$

where  $\mathbf{C}$  is a  $p \times n$  matrix with  $p < n$ , all equations being assumed linearly independent. As a consequence,  $\mathbf{C}$  is of full rank, i.e.,  $\text{rank}(\mathbf{C}) = p$ . Apparently, the system *admits infinitely-many solutions*. Notice that the set of solutions of this equation does not form a vector space. Indeed, since  $\mathbf{0}$  is not a solution, the solution set does not include the origin, which disqualifies the set from being a vector space. However, the same set has a quite interesting property:

**Fact 5.4.1** *The set of solutions of the system (5.44) is convex.*

*Proof:* Assume that  $\mathbf{x}_1$  and  $\mathbf{x}_2$  are two distinct solutions of eq.(5.44), i.e.,

$$\mathbf{C}\mathbf{x}_1 = \mathbf{d} \quad (5.45a)$$

$$\mathbf{C}\mathbf{x}_2 = \mathbf{d} \quad (5.45b)$$

Now, for a real  $\alpha$  such that  $0 \leq \alpha \leq 1$ , we have

$$\mathbf{C}(\alpha\mathbf{x}_1) = \alpha\mathbf{d} \quad (5.46a)$$

$$\mathbf{C}[(1 - \alpha)\mathbf{x}_2] = (1 - \alpha)\mathbf{d} \quad (5.46b)$$

Upon adding sidewise eqs.(5.46a & b), we obtain

$$\mathbf{C}[\alpha\mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2] = \mathbf{d} \quad (5.47)$$

thereby completing the proof.

Geometrically, eq.(5.44) represents a plane embedded in  $n$ -dimensional space, offset from the origin. Each point of the plane thus has a position vector that is a solution of eq.(5.44). Out of the infinity of solutions satisfying the equation, then,



there is one that lies closest to the origin. This is the *minimum-norm solution* of eq.(5.44). We derive below this solution upon solving the problem:

$$f(\mathbf{x}) \equiv \frac{1}{2}\|\mathbf{x}\|^2 \quad \rightarrow \quad \min_{\mathbf{x}} \quad (5.48)$$

subject to eq.(5.44). As before, we transform the above constrained problem into an unconstrained one. We do this by means of Lagrange multipliers:

$$F(\mathbf{x}) \equiv f(\mathbf{x}) + \boldsymbol{\lambda}^T(\mathbf{C}\mathbf{x} - \mathbf{d}) \quad \rightarrow \quad \min_{\mathbf{x}, \boldsymbol{\lambda}} \quad (5.49)$$

subject to no constraints. The normality conditions of this problem are, thus,

$$\frac{\partial F}{\partial \mathbf{x}} \equiv \nabla f + \mathbf{C}^T \boldsymbol{\lambda} = \mathbf{x} + \mathbf{C}^T \boldsymbol{\lambda} = \mathbf{0}_n \quad (5.50a)$$

$$\frac{\partial F}{\partial \boldsymbol{\lambda}} \equiv \mathbf{C}\mathbf{x} - \mathbf{d} = \mathbf{0}_p \quad (5.50b)$$

the second set of the above equations thus being just a restatement of the system of underdetermined equations (5.44). Solving for  $\mathbf{x}$  from eq.(5.50a) yields

$$\mathbf{x} = -\mathbf{C}^T \boldsymbol{\lambda} \quad (5.51a)$$

which, when substituted into eq.(5.50b), leads to

$$-\mathbf{C}\mathbf{C}^T \boldsymbol{\lambda} - \mathbf{d} = \mathbf{0}_p \quad (5.51b)$$

Since we assumed at the outset that the given eqs.(5.44) are linearly-independent,  $\mathbf{C}$  is of full rank, and hence, the  $p \times p$  symmetric matrix  $\mathbf{C}\mathbf{C}^T$  is nonsingular. As a result, this matrix is, in fact, positive-definite, the outcome being that eq.(5.51b) can be solved for  $\boldsymbol{\lambda}$  by means of the Cholesky decomposition. The result is, symbolically, the minimum-norm solution  $\mathbf{x}_o$  sought:

$$\mathbf{x}_o = \mathbf{C}^\dagger \mathbf{d} \quad (5.52a)$$

where

$$\mathbf{C}^\dagger = \mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1} \quad (5.52b)$$

which is the *right Moore-Penrose generalized inverse* (RMPGI) of the rectangular matrix  $\mathbf{C}$ . One can see that the straightforward evaluation of  $\mathbf{C}^\dagger$  by its definition, eq. (5.52a), involves the inversion of a matrix product, which is computationally costly and prone to ill conditioning, similar to the case of the left Moore-Penrose generalized inverse of eq.(3.54b). Moreover, the solution of eq. (5.52a) does not hold when  $\mathbf{C}$  is rank-deficient.

**Exercise 5.4.1** Let the QR-decomposition of  $\mathbf{C}^T$  be

$$\mathbf{C}^T = \mathbf{Q}\mathbf{R} = [\mathbf{Q}_L \quad \mathbf{Q}_R] \begin{bmatrix} \mathbf{U} \\ \mathbf{O}_{n'p} \end{bmatrix}, \quad n' = n - p \quad (5.53)$$

whence  $\mathbf{C}^T = \mathbf{Q}_L \mathbf{U}$  with  $\mathbf{Q}_L$  of  $n \times p$ ,  $\mathbf{Q}_R$  of  $n \times (n - p)$ ,  $\mathbf{U}$  of  $p \times p$  and  $\mathbf{O}_{n'p}$  representing the  $(n - p) \times p$  zero matrix. Devise a two-step algorithm to compute the minimum-norm solution of eq.(5.44) à la eq.(3.41).

**Exercise 5.4.2 (An orthonormal basis for the null space of  $\mathbf{C}$ )** Show that the columns of  $\mathbf{Q}_R$  span  $\mathcal{N}(\mathbf{C})$ , the null space of  $\mathbf{C}$ , and hence, form a basis for  $\mathcal{N}(\mathbf{C})$ . Moreover, show that this basis is orthonormal. Hint: show that  $\mathbf{Q}_L^T \mathbf{Q}_R = \mathbf{O}_{pn'}$ , the  $p \times (n - p)$  zero matrix.

An efficient and robust method to computing explicitly the RMPGI relies on the QR-decomposition of  $\mathbf{C}^T$ , as explained below:

First, notice that by virtue of the orthogonality of  $\mathbf{Q}$ ,

$$\mathbf{C}\mathbf{C}^T = \mathbf{R}^T \mathbf{Q}^T \mathbf{Q} \mathbf{R} = \mathbf{R}^T \mathbf{R} = \mathbf{U}^T \mathbf{U}$$

whence,

$$(\mathbf{C}\mathbf{C}^T)^{-1} = \mathbf{U}^{-1} \mathbf{U}^{-T} \quad (5.54)$$

Upon substitution of eqs.(5.53) and (5.54) into eq.(5.52b),  $\mathbf{C}^\dagger$  is obtained as

$$\mathbf{C}^\dagger = \mathbf{Q} \begin{bmatrix} \mathbf{U}\mathbf{U}^{-1}\mathbf{U}^{-T} \\ \mathbf{O}_{n'p} \end{bmatrix} = [\mathbf{Q}_L \quad \mathbf{Q}_R] \begin{bmatrix} \mathbf{U}^{-T} \\ \mathbf{O}_{n'p} \end{bmatrix}$$

Therefore,

$$\mathbf{C}^\dagger = \mathbf{Q}_L \mathbf{U}^{-T} \quad (5.55a)$$

which means that  $\mathbf{C}^\dagger$  is to be computed numerically, once the QR-decomposition of  $\mathbf{C}^T$  is available, by forward substitution from the lower-triangular system

$$\mathbf{U}^T \mathbf{C}^\dagger = \mathbf{Q}_L \quad (5.55b)$$

**Example 5.4.1 (The Solution of  $\mathbf{a} \times \mathbf{x} = \mathbf{b}$ )**

Let  $\mathbf{a}$ ,  $\mathbf{b}$ , and  $\mathbf{x}$  be three 3-dimensional Cartesian vectors. We would like to solve the equation

$$\mathbf{a} \times \mathbf{x} = \mathbf{b}$$

for  $\mathbf{x}$ .

Solution: It is well known, however, that the foregoing equation contains only two linearly independent scalar equations, which prevents us from finding “the  $\mathbf{x}$ ” that verifies that equation. Thus, we can proceed by finding a specific  $\mathbf{x}$ ,  $\mathbf{x}_o$ , that verifies any two of these three equations and that is of minimum norm, for example. To this end, we expand that equation into its three components:

$$a_2x_3 - a_3x_2 = b_1$$

$$a_3x_1 - a_1x_3 = b_2$$

$$a_1x_2 - a_2x_1 = b_3$$

Note that the foregoing equation can be cast in the form of eq.(3.31) if we define matrix  $\mathbf{A}$  as  $\text{CPM}(\mathbf{a})$ —see eq.(3.19c)—namely,

$$\mathbf{A} \equiv \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix}$$

which is known to be singular, the given equations then becoming

$$\mathbf{Ax} = \mathbf{b}$$

Picking up, for example, the first two the three foregoing scalar equations, we obtain an underdetermined system of the form (5.44), with

$$\mathbf{C} \equiv \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

and hence, the corresponding minimum-norm solution  $\mathbf{x}_o$  is given by eqs.(5.52a & b), with

$$\mathbf{CC}^T = \begin{bmatrix} a_2^2 + a_3^2 & -a_1a_2 \\ -a_1a_2 & a_1^2 + a_3^2 \end{bmatrix}$$

Hence,

$$(\mathbf{CC}^T)^{-1} = \frac{1}{\Delta} \begin{bmatrix} a_1^2 + a_3^2 & a_1a_2 \\ a_1a_2 & a_2^2 + a_3^2 \end{bmatrix}$$

where

$$\Delta \equiv \det(\mathbf{CC}^T) = (a_2^2 + a_3^2)(a_1^2 + a_3^2) - a_1^2a_2^2 = a_3^2\|\mathbf{a}\|^2 > 0$$

a relation that shows that  $\mathbf{CC}^T$  is singular if  $a_3$  happens to vanish. If this is not the case, then

$$\mathbf{C}^\dagger = \frac{a_3}{\Delta} \begin{bmatrix} a_1a_2 & a_2^2 + a_3^2 \\ -(a_1^2 + a_3^2) & -a_1a_2 \\ a_2a_3 & -a_1a_3 \end{bmatrix}$$

whence,

$$\mathbf{x}_o = \frac{a_3}{\Delta} \begin{bmatrix} a_1 a_2 b_1 + (a_2^2 + a_3^2) b_2 \\ -(a_1^2 + a_3^2) b_1 - a_1 a_2 b_2 \\ a_2 a_3 b_1 - a_1 a_3 b_2 \end{bmatrix} \quad (5.57)$$

Notice that the foregoing solution depends on the condition  $a_3 \neq 0$ . If  $a_3 = 0$ , or very close to 0, then  $\mathbf{C}$  becomes either rank-deficient or, correspondingly, ill-conditioned, which is bad news. Apparently, the foregoing solution has an element of arbitrariness that may lead either to rank-deficiency or to ill-conditioning. There is no guarantee that the two equations singled out are the best choice from the condition-number viewpoint. Besides, that approach leaves aside useful information, that of the deleted equation. The alternative approach uses all three equations, to which one fourth equation is adjoined, namely, the minimum-norm condition, as described below.

First we observe that, if  $\mathbf{x}$  has been found that verifies the given cross-product equation, then any other vector  $\mathbf{x} + \alpha \mathbf{a}$ , for  $\alpha \in \mathbb{R}$ , verifies that equation. Apparently, then, the minimum-norm  $\mathbf{x}$  is that whose component along  $\mathbf{a}$  vanishes, i.e.,

$$\mathbf{a}^T \mathbf{x} = 0$$

Upon adjoining the foregoing equation to the original three, we end up with an apparently overdetermined system of four equations with three unknowns, of the form

$$\mathbf{M} \mathbf{x} = \mathbf{n}$$

where  $\mathbf{M}$  and  $\mathbf{n}$  are given by

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} \\ \mathbf{a}^T \end{bmatrix}, \quad \mathbf{n} = \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix}$$

Hence,  $\mathbf{M}$  is a  $4 \times 3$  matrix, while  $\mathbf{n}$  is a 4-dimensional vector. The least-square approximation of the new system is, then, the minimum-norm solution of the original system, provided that the latter is verified exactly, which it is, as will become apparent. Indeed, the least-square approximation of the new system takes the form

$$\mathbf{x}_L = (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T \mathbf{n} \quad (5.58)$$

While we have strongly advised against the explicit computation of generalized inverses, our advice is valid only as pertaining to numerical computations. In the case at hand, we will pursue not a numerical, but rather a symbolic computation of the solution sought.

The first issue now is whether  $\mathbf{M}^T \mathbf{M}$  is invertible, but it is so and, moreover, its inverse is extremely simple to find:

$$\mathbf{M}^T \mathbf{M} = \begin{bmatrix} \mathbf{A}^T & \mathbf{a} \end{bmatrix} \begin{bmatrix} \mathbf{A} \\ \mathbf{a}^T \end{bmatrix} = \mathbf{A}^T \mathbf{A} + \mathbf{a} \mathbf{a}^T$$

But, since  $\mathbf{A}$  is skew-symmetric,

$$\mathbf{M}^T \mathbf{M} = -\mathbf{A}^2 + \mathbf{a} \mathbf{a}^T$$

as the reader can readily verify; moreover,

$$\mathbf{A}^2 = -\|\mathbf{a}\|^2 \mathbf{1} + \mathbf{a} \mathbf{a}^T$$

Hence,

$$\mathbf{M}^T \mathbf{M} = \|\mathbf{a}\|^2 \mathbf{1}$$

which means that  $\mathbf{M}$  is isotropic, i.e.,  $\kappa(\mathbf{M}) = 1$ , and hence,  $\mathbf{M}$  is optimally-conditioned. Therefore,

$$(\mathbf{M}^T \mathbf{M})^{-1} = \frac{1}{\|\mathbf{a}\|^2} \mathbf{1}$$

That is ,

$$\mathbf{x}_L = \frac{1}{\|\mathbf{a}\|^2} \mathbf{1} \begin{bmatrix} \mathbf{A}^T & \mathbf{a} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix} = \frac{1}{\|\mathbf{a}\|^2} \mathbf{A}^T \mathbf{b}$$

which can be further expressed as

$$\mathbf{x}_L = -\frac{\mathbf{a} \times \mathbf{b}}{\|\mathbf{a}\|^2} \quad (5.59)$$

thereby obtaining a much simpler, and robust, expression than that displayed above as  $\mathbf{x}_o$ . Robustness, in this case, means that the above expression does not break down under any special value of any specific component of  $\mathbf{a}$ . This is a benefit of an invariant expression, as opposed to one depending on components.

**Exercise 5.4.3** Find the number of floating-point operations (flops) required to compute  $\mathbf{x}_o$  as given by eq.(5.57) and compare them with those required to compute  $\mathbf{x}_L$  as given by eq.(5.59).

## 5.4.2 Least-Square Problems Subject to Linear Constraints

Given the system of linear equations

$$\mathbf{A} \mathbf{x} = \mathbf{b} \quad (5.60)$$

where  $\mathbf{A}$  is a full-rank  $q \times n$  matrix, with  $q > n$ , and  $\mathbf{b}$  is a  $q$ -dimensional vector, find a  $n$ -dimensional vector  $\mathbf{x}$  that verifies the above system with the least-square error, subject to the linear equality constraints

$$\mathbf{C}\mathbf{x} = \mathbf{d} \quad (5.61)$$

with  $\mathbf{C}$  a full-rank  $p \times n$  matrix and  $\mathbf{d}$  a  $p$ -dimensional vector. Moreover,  $\mathbf{W}$  is a  $q \times q$  positive-definite weighting matrix, with  $q$ ,  $p$  and  $n$  subject to

$$q + p > n \quad \text{and} \quad n > p \quad (5.62)$$

The error-squared of eqs.(5.60) is defined as

$$f \equiv \frac{1}{2}(\mathbf{A}\mathbf{x} - \mathbf{b})^T \mathbf{W}(\mathbf{A}\mathbf{x} - \mathbf{b}) \quad (5.63)$$

As usual, we solve this problem by introducing Lagrange multipliers:

$$F(\mathbf{x}; \boldsymbol{\lambda}) \equiv f(\mathbf{x}) + \boldsymbol{\lambda}^T (\mathbf{C}\mathbf{x} - \mathbf{d}) \rightarrow \min_{\mathbf{x}, \boldsymbol{\lambda}} \quad (5.64)$$

subject to no constraints.

The first-order normality conditions of the foregoing problem are

$$\frac{\partial F}{\partial \mathbf{x}} \equiv \mathbf{A}^T \mathbf{W}(\mathbf{A}\mathbf{x} - \mathbf{b}) + \mathbf{C}^T \boldsymbol{\lambda} = \mathbf{0}_n \quad (5.65a)$$

$$\frac{\partial F}{\partial \boldsymbol{\lambda}} \equiv \mathbf{C}\mathbf{x} - \mathbf{d} = \mathbf{0}_p \quad (5.65b)$$

Since  $\mathbf{A}$  is assumed of full rank and  $\mathbf{W}$  is positive-definite, we can solve eq.(5.65a) for  $\mathbf{x}$  in terms of  $\boldsymbol{\lambda}$ , namely,

$$\mathbf{x} = (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} (\mathbf{A}^T \mathbf{W} \mathbf{b} - \mathbf{C}^T \boldsymbol{\lambda}) \quad (5.66)$$

Upon substituting the above expression into eq.(5.65b), we obtain

$$\mathbf{C}(\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{C}^T \boldsymbol{\lambda} = \mathbf{C}(\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W} \mathbf{b} - \mathbf{d}$$

whence,

$$\boldsymbol{\lambda} = [\mathbf{C}(\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{C}^T]^{-1} [\mathbf{C}(\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W} \mathbf{b} - \mathbf{d}] \quad (5.67)$$

ow, the foregoing expression for  $\boldsymbol{\lambda}$  is substituted, in turn, into eq.(5.66), thereby obtaining the optimum value of  $\mathbf{x}$ ,  $\mathbf{x}_o$ , namely,

$$\mathbf{x}_o = \mathbf{P}\mathbf{B}\mathbf{b} + \mathbf{D}\mathbf{d} \quad (5.68a)$$

where  $\mathbf{P}$ ,  $\mathbf{B}$  and  $\mathbf{D}$  are the  $n \times n$ -,  $n \times m$ - and  $n \times p$  matrices given below:

$$\mathbf{P} = \mathbf{1}_n - \mathbf{D}\mathbf{C} \quad (5.68b)$$

$$\mathbf{B} = (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W} \quad (5.68c)$$

$$\mathbf{D} = (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{C}^T [\mathbf{C}(\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{C}^T]^{-1} \quad (5.68d)$$

with  $\mathbf{1}_n$  standing for the  $n \times n$  identity matrix. The solution derived above, while being exact, for it is symbolic, is unsuitable for numerical implementation. Indeed, this solution contains inversions of products of several matrices times their transposes, which brings about ill-conditioning. Various approaches to the numerical solution of this problem will be studied in Ch. 6.

**Exercise 5.4.4** *Using nothing but: a) the Cholesky-decomposition of positive-definite matrices; b) the QR-decomposition of rectangular matrices with more rows than columns; and c) forward and backward substitutions in triangular systems of linear equations, devise an algorithm to compute numerically the solution  $\mathbf{x}_o$  of eq.(5.68a).*

## 5.5 Equality-Constrained Nonlinear Least Squares

We consider here the problem of finding the least-square error  $f$  of an *overdetermined* system of  $q$  nonlinear equations in the  $n$ -dimensional vector of unknowns  $\mathbf{x}$ , namely,

$$\boldsymbol{\phi}(\mathbf{x}) = \mathbf{0}_q \quad (5.69a)$$

subject to the  $l$  nonlinear constraints

$$\mathbf{h}(\mathbf{x}) = \mathbf{0}_l \quad (5.69b)$$

whereby

$$q > n, \quad n > l \quad (5.70)$$

The problem thus consists in finding a  $\mathbf{x}$  that makes  $\boldsymbol{\phi}$  as close as possible to zero, while observing *strictly* eqs.(5.69b).

In general, moreover, the various scalar equations of eq.(5.69a) have different relevance and are, hence, assigned different weights, which then leads to a problem of *weighted least squares*, namely,

$$f(\mathbf{x}) = \frac{1}{2} \boldsymbol{\phi}^T \mathbf{W} \boldsymbol{\phi} \rightarrow \min_{\mathbf{x}} \quad (5.71)$$

subject to eq.(5.69b).

The normality conditions of the problem at hand are derived directly from those of the general equality-constrained problem, namely, eq.(5.12) or its dual counterpart, eq.(5.16). In our case,

$$\nabla f = \left( \frac{\partial \phi}{\partial \mathbf{x}} \right)^T \frac{\partial f}{\partial \phi} \quad (5.72a)$$

where

$$\frac{\partial \phi}{\partial \mathbf{x}} \equiv \mathbf{\Phi}(\mathbf{x}), \quad \frac{\partial f}{\partial \phi} = \mathbf{W}\phi(\mathbf{x}) \quad (5.72b)$$

i.e.,  $\mathbf{\Phi}(\mathbf{x})$  denotes the  $q \times n$  gradient of  $\phi(\mathbf{x})$  with respect to  $\mathbf{x}$ . Hence,

$$\nabla f = \mathbf{\Phi}^T \mathbf{W} \phi \quad (5.72c)$$

where we have dispensed with the argument  $\mathbf{x}$  for the sake of simplicity.

The normality condition (5.12) thus reduces to

$$[\mathbf{1} - \mathbf{J}^T(\mathbf{J}\mathbf{J}^T)^{-1}\mathbf{J}]\mathbf{\Phi}^T \mathbf{W} \phi = \mathbf{0}_n \quad (5.73)$$

with  $\mathbf{J}$  defined, as usual, as  $\mathbf{J} = \partial \mathbf{h} / \partial \mathbf{x} = \nabla \mathbf{h}$ . What the foregoing condition states is that, at a stationary point,  $\nabla f = \mathbf{\Phi}^T \mathbf{W} \phi$  need not vanish<sup>6</sup>; only the projection of  $\nabla f$  onto the null space of the gradient of the constraints must vanish.

The dual form of the same normality conditions, in turn, reduces to

$$\mathbf{L}^T \mathbf{\Phi}^T \mathbf{W} \phi = \mathbf{0}_{n'} \quad (5.74)$$

with  $\mathbf{L}$  indicating a  $n \times (n - l)$  orthogonal complement of  $\mathbf{J}$ , as defined in eq.(5.15), and  $\mathbf{0}_{n'}$  denoting the  $(n - l)$ -dimensional zero vector.

The second-order normality conditions of  $f(\mathbf{x})$  are now derived by assuming that we have found a stationary value of the design-variable vector,  $\mathbf{x}_o$ . This means that the FONC are satisfied at  $\mathbf{x} = \mathbf{x}_o$ . More specifically, the FONC in dual form are verified at the given SP, which means that

$$\mathbf{L}^T(\mathbf{x}_o) \mathbf{\Phi}^T(\mathbf{x}_o) \mathbf{W} \phi(\mathbf{x}_o) = \mathbf{0}_{n'} \quad (5.75)$$

Now, in order to set up the SONC, we need the constrained Hessian  $\mathbf{H}_c$  of the problem at hand, namely,

$$\mathbf{H}_c = \nabla \nabla f + \frac{\partial(\mathbf{J}^T \boldsymbol{\lambda})}{\partial \mathbf{x}} \quad (5.76)$$

---

<sup>6</sup>Certainly,  $\phi$  need neither vanish, as the system of nonlinear equations (5.69a) is overdetermined.



Given the form (5.72c) of  $\nabla f$ , its gradient, i.e., the *unconstrained Hessian*  $\nabla\nabla f$ , requires partial derivatives of matrix  $\Phi$ , leading to a three-dimensional array. To avoid the encumbrance of such arrays, we calculate the unconstrained Hessian in two steps: first, we differentiate the rightmost-hand factor  $\phi$  of  $\nabla f$  in eq. (5.72c) with respect to  $\mathbf{x}$ ; then, upon fixing  $\mathbf{x}$  at  $\mathbf{x}_o$  in  $\phi$ , a constant vector  $\phi_o = \phi(\mathbf{x}_o)$  is obtained; finally, differentiation of the whole product  $\Phi^T \mathbf{W} \phi_o$  with respect to  $\mathbf{x}$  yields a matrix, which is just added to the first term:

$$\nabla\nabla f = \Phi^T \mathbf{W} \Phi + \frac{\partial(\Phi^T \mathbf{W} \phi_o)}{\partial \mathbf{x}} \quad (5.77)$$

We recall that the constrained Hessian  $\mathbf{H}_c$  as well as its unconstrained counterpart is symmetric, and takes the form

$$\mathbf{H}_c = \Phi^T \mathbf{W} \Phi + \frac{\partial(\Phi^T \mathbf{W} \phi_o)}{\partial \mathbf{x}} + \frac{\partial(\mathbf{J}^T \boldsymbol{\lambda})}{\partial \mathbf{x}} \quad (5.78)$$

Furthermore, the *reduced Hessian*  $\mathbf{H}_u$  of  $f$  is now

$$\mathbf{H}_u = \mathbf{L}^T \left[ \Phi^T \mathbf{W} \Phi + \frac{\partial(\Phi^T \mathbf{W} \phi_o)}{\partial \mathbf{x}} + \frac{\partial(\mathbf{J}^T \boldsymbol{\lambda})}{\partial \mathbf{x}} \right] \mathbf{L} \quad (5.79)$$

In summary, then, the optimization problem at hand can have local minima, local maxima or saddle points, depending on the sign-definition of  $\mathbf{H}_u$ .

## 5.6 Linear Least-Square Problems Under Quadratic Constraints

An important family of design problems lends itself to a formulation whereby the objective function is quadratic in a linear function of the design vector  $\mathbf{x}$ , while the constraints are quadratic in  $\mathbf{x}$ . Contrary to the case of linear least-squares subject to linear constraints, this family of problems does not allow, in general, for closed-form solutions, the reason being that their normal equations are nonlinear. Let us consider

$$f(\mathbf{x}) \equiv \frac{1}{2}(\mathbf{b} - \mathbf{A}\mathbf{x})^T \mathbf{W}(\mathbf{b} - \mathbf{A}\mathbf{x}) \rightarrow \min_{\mathbf{x}} \quad (5.80a)$$

subject to

$$\mathbf{h}(\mathbf{x}) = \mathbf{0}_l \quad (5.80b)$$

where  $\mathbf{A}$  is a  $q \times n$  full-rank matrix, with  $q \geq n$ ,  $\mathbf{W}$  is a  $q \times q$  positive-definite weighting matrix, while  $\mathbf{h}$ ,  $\mathbf{x}$  and  $\mathbf{b}$  are  $l$ -,  $n$ - and  $q$ -dimensional vectors, respectively, with

$$q + l > n, \quad n > l \quad (5.80c)$$

Moreover, in this particular case, the  $i$ th component of vector  $\mathbf{h}$  is quadratic, namely,

$$h_i(\mathbf{x}) \equiv \frac{1}{2}\mathbf{x}^T \mathbf{P}_i \mathbf{x} + \mathbf{q}_i^T \mathbf{x} + r_i, \quad i = 1, 2, \dots, l \quad (5.80d)$$

in which  $\mathbf{P}_i$  is a known  $n \times n$  symmetric matrix, while  $\mathbf{q}_i$  is a  $n$ -dimensional given vector and  $r_i$  is a given scalar. Apparently, then, the  $i$ th row of  $\mathbf{J}$ , the Jacobian of  $\mathbf{h}$  with respect to  $\mathbf{x}$ , takes the form

$$\left( \frac{\partial h_i}{\partial \mathbf{x}} \right)^T = \mathbf{P}_i \mathbf{x} + \mathbf{q}_i, \quad i = 1, 2, \dots, l \quad (5.81)$$

Therefore,

$$\mathbf{J}^T = [\mathbf{P}_1 \mathbf{x} + \mathbf{q}_1 \quad \mathbf{P}_2 \mathbf{x} + \mathbf{q}_2 \quad \cdots \quad \mathbf{P}_l \mathbf{x} + \mathbf{q}_l] \quad (5.82)$$

whence  $\mathbf{J}$  is *linear* in  $\mathbf{x}$ . To derive the FONC, we need  $\nabla f$ , which is readily derived as

$$\nabla f = -\mathbf{A}^T \mathbf{W}(\mathbf{b} - \mathbf{A}\mathbf{x}) \quad (5.83)$$

the first-order normality conditions (5.12) thus taking the primal form

$$[\mathbf{1}_n - \mathbf{J}^T(\mathbf{J}\mathbf{J}^T)^{-1}\mathbf{J}]\mathbf{A}^T \mathbf{W}(\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{0}_n \quad (5.84)$$

It is thus apparent that, although  $\mathbf{J}$  is linear in  $\mathbf{x}$ , the normality conditions are rational<sup>7</sup>

The FONC in dual form are, then,

$$\mathbf{L}^T \mathbf{A}^T \mathbf{W}(\mathbf{b} - \mathbf{A}\mathbf{x}) = \mathbf{0}_{n'} \quad (5.85)$$

The SONC are now derived. Differentiation of  $\nabla f$  with respect to  $\mathbf{x}$  leads to the unconstrained Hessian, namely,

$$\nabla \nabla f = \mathbf{A}^T \mathbf{W} \mathbf{A} \quad (5.86)$$

which is, apparently, positive-definite, but this Hessian does not tell the nature of a SP. We need the constrained Hessian, as defined in eq.(5.76), which calls for

---

<sup>7</sup>A rational function of a real variable  $x$  bears the form  $P(x)/Q(x)$ , with  $P$  and  $Q$  denoting polynomials in  $x$  of degrees  $p$  and  $q$ , respectively. A similar definition follows for a  $n$ -dimensional vector  $\mathbf{x}$ , instead of  $x$ ,  $P$  and  $Q$  then changing to *multi-variable polynomials*., stemming from the inverse of  $\mathbf{J}\mathbf{J}^T$ , thereby leading to a problem lacking a closed-form solution, except for special cases, like the example included in this section.

the computation of an additional term, namely,  $\partial(\mathbf{J}^T \boldsymbol{\lambda})/\partial \mathbf{x}$ ; this term is computed below. First, we have the product

$$\begin{aligned} \mathbf{J}^T \boldsymbol{\lambda} &= [\mathbf{P}_1 \mathbf{x} + \mathbf{q}_1 \quad \mathbf{P}_2 \mathbf{x} + \mathbf{q}_2 \quad \cdots \quad \mathbf{P}_l \mathbf{x} + \mathbf{q}_l] \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_l \end{bmatrix} \\ &= \sum_1^l \lambda_i (\mathbf{P}_i \mathbf{x} + \mathbf{q}_i) \end{aligned} \quad (5.87)$$

whence,

$$\frac{\partial(\mathbf{J}^T \boldsymbol{\lambda})}{\partial \mathbf{x}} = \lambda_1 \mathbf{P}_1 + \lambda_2 \mathbf{P}_2 + \cdots + \lambda_l \mathbf{P}_l = \sum_1^l \lambda_i \mathbf{P}_i = \sum_1^l \lambda_i \mathbf{P}_i \quad (5.88)$$

Therefore,

$$\mathbf{H}_c = \mathbf{A}^T \mathbf{W} \mathbf{A} + \sum_1^l \lambda_i \mathbf{P}_i \quad (5.89)$$

which is the sum of a positive-definite term plus one that is sign-indefinite. However, notice that the two terms are symmetric, and hence, the constrained Hessian is symmetric as well, similar to the general case. Now, the reduced  $(n-l) \times (n-l)$  Hessian  $\mathbf{H}_u$  is readily obtained as

$$\mathbf{H}_u = \mathbf{L}^T (\mathbf{A}^T \mathbf{W} \mathbf{A} + \sum_1^l \lambda_i \mathbf{P}_i) \mathbf{L} \quad (5.90)$$

In conclusion, then, a SP is a local minimum if, at this SP,  $\mathbf{H}_u$  is positive-definite; it is a local maximum if negative-definite; it is a saddle point if sign-indefinite.

### Example 5.6.1 (A Quadratic Objective Function with a Quadratic Constraint)

*Find all the SPs of the objective function*

$$f(\mathbf{x}) = \frac{1}{2}(9x_1^2 - 8x_1x_2 + 3x_2^2) \quad \rightarrow \quad \min_{x_1, x_2}$$

*subject to the quadratic constraint*

$$h(\mathbf{x}) = x_1^2 + x_2^2 - 1 = 0$$

*and determine the nature of each SP.*

*Solution:* The objective function can be shown to be associated to a positive-definite matrix  $\mathbf{Q}$ :

$$\mathbf{Q} = \nabla \nabla f = \begin{bmatrix} 9 & -4 \\ -4 & 3 \end{bmatrix}$$

Moreover,  $f(\mathbf{x})$  can be factored as

$$f(\mathbf{x}) = \frac{1}{2}(\mathbf{b} - \mathbf{A}\mathbf{x})^T \mathbf{W}(\mathbf{b} - \mathbf{A}\mathbf{x})$$

with<sup>8</sup>

$$\mathbf{A} = \mathbf{1}_2, \quad \mathbf{b} = \mathbf{0}_2, \quad \mathbf{W} = \begin{bmatrix} 9 & -4 \\ -4 & 3 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

where  $\mathbf{W}$  is, apparently, identical to  $\nabla \nabla f$  in this case, and hence, *positive-definite*, which means that the problem belongs to the *least-square class*.

Upon adjoining the constraint to the objective function, we obtain the Lagrangian  $F(\mathbf{x}; \lambda)$ , namely,

$$F(\mathbf{x}; \lambda) = \frac{1}{2}(9x_1^2 - 8x_1x_2 + 3x_2^2) + \lambda(x_1^2 + x_2^2 - 1)$$

which we want to minimize subject to no constraints. The normality conditions of the unconstrained problem are, thus

$$\begin{aligned} \frac{\partial F}{\partial x_1} &= 9x_1 - 4x_2 + 2\lambda x_1 = 0 \\ \frac{\partial F}{\partial x_2} &= -4x_1 + 3x_2 + 2\lambda x_2 = 0 \\ \frac{\partial F}{\partial \lambda} &= x_1^2 + x_2^2 - 1 = 0 \end{aligned}$$

Let us now eliminate  $\lambda$  from the first and the second of the above equations. We do this dialytically, i.e., we write these two equations in linear homogeneous form in  $\lambda$  and 1, i.e.,

$$\mathbf{M}\mathbf{y} = \mathbf{0}_2$$

where

$$\mathbf{M} = \begin{bmatrix} 2x_1 & 9x_1 - 4x_2 \\ 2x_2 & -4x_1 + 3x_2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \lambda \\ 1 \end{bmatrix} \neq \mathbf{0}_2$$

---

<sup>8</sup>Actually, the choice of  $\mathbf{A}$  and  $\mathbf{W}$  in the factoring of the quadratic form is not unique. An alternative definition of these matrices could be  $\mathbf{W} = \mathbf{1}_2$  and  $\mathbf{A}$  one of the Cholesky factros of  $\mathbf{Q}$ , as in this case,  $q = n$ . For  $q > n$ , the factoring of  $\nabla \nabla f$  into the product  $\mathbf{A}^T \mathbf{A}$  becomes indeterminate.

with  $\mathbf{0}_2$  denoting the 2-dimensional zero vector. Now, the above linear homogeneous equation in  $\mathbf{y}$  cannot be zero, for  $\mathbf{y} \neq \mathbf{0}_2$ , and hence, matrix  $\mathbf{M}$  must be singular, which is stated as

$$\Delta \equiv \det(\mathbf{M}) = 0$$

Upon expansion, the foregoing equation leads to

$$\Delta = 2x_1(-4x_1 + 3x_2) - 2x_2(9x_1 - 4x_2) = 0$$

or, after simplification,

$$x_1^2 + \frac{3}{2}x_1x_2 - x_2^2 = 0$$

thereby reducing the problem to the solution of two quadratic equations in two unknowns, the above equation and the third normality condition, i.e., the constraint. While it is not too difficult to manipulate two quadratic bivariate equations to derive one single monovariate equation, one must avoid cumbersome algebraic manipulations. Notice that

$$\Delta - h = \frac{3}{2}x_1x_2 - 2x_2^2 + 1 = 0$$

which is linear in  $x_1$ , and hence, readily yields an expression for  $x_1$  in terms of  $x_2$ , namely,

$$x_1 = \frac{2}{3} \frac{2x_2^2 - 1}{x_2}$$

Substitution of the above expression into  $h = 0$  yields, after simplifications,

$$25x_2^4 - 25x_2^2 + 4 = 0$$

which is a quadratic equation in  $x_2^2$ , its four roots being

$$(x_2)_1 = \frac{\sqrt{5}}{5}, \quad (x_2)_2 = -\frac{\sqrt{5}}{5}, \quad (x_2)_3 = \frac{2\sqrt{5}}{5}, \quad (x_2)_4 = -\frac{2\sqrt{5}}{5}$$

which lead to the corresponding values of  $x_1$ , namely,

$$(x_1)_1 = -\frac{2\sqrt{5}}{5}, \quad (x_1)_2 = \frac{2\sqrt{5}}{5}, \quad (x_1)_3 = \frac{\sqrt{5}}{5}, \quad (x_1)_4 = -\frac{\sqrt{5}}{5}$$

In fact,  $\Delta = 0$  is a degenerate conic, that breaks down into the product of two linear equations. The four foregoing solutions of the problem at hand are thus the four intersections of two lines passing through the origin with the unit circle centred at the origin, as shown in Fig. 5.12.

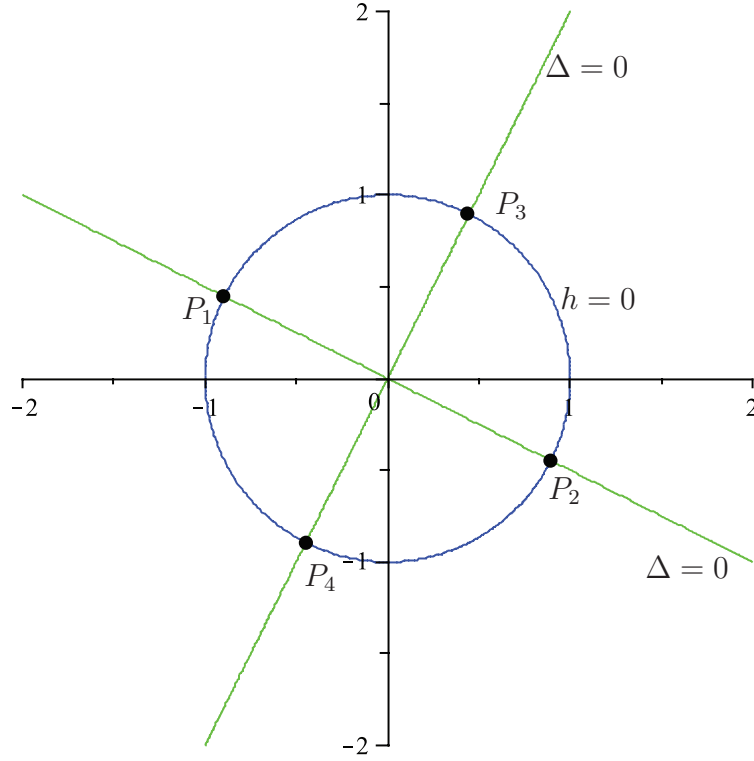


Figure 5.12: The intersection of  $\Delta = 0$ , a degenerate conic, with  $h = 0$

Now we determine the nature of the SPs. To this end, we need an orthogonal complement  $\mathbf{L}$  of  $\mathbf{J}$ . This is readily derived, by regarding  $\mathbf{J}$  as a two-dimensional vector row array, namely,

$$\mathbf{J} = [2x_1 \quad 2x_2]$$

Upon rotating this vector through  $90^\circ$  ccw, normalizing the array thus resulting so that it will be of unit norm, and casting it in column form, we obtain

$$\mathbf{L} = \frac{1}{\sqrt{x_1^2 + x_2^2}} \begin{bmatrix} -x_2 \\ x_1 \end{bmatrix}$$

Moreover, upon adding the first two normality conditions for the Lagrangian, one expression for  $\lambda$  is obtained in terms of the design variables, namely,

$$\lambda = \frac{-5x_1 + x_2}{2(x_1 + x_2)}$$

Further, a general expression for  $\mathbf{H}_c$  is

$$\mathbf{H}_c = \begin{bmatrix} 9 + \lambda & -4 \\ -4 & 3 + \lambda \end{bmatrix}$$

SP1: The first SP is  $\mathbf{x} = (\sqrt{5}/5)[2, -1]^T$ , whence

$$\lambda = -\frac{11}{2}, \quad \mathbf{L} = \frac{\sqrt{5}}{5} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Therefore,

$$\mathbf{H}_u = \frac{1}{5} \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 7/2 & -4 \\ -4 & -5/2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = -\frac{9}{2} < 0$$

which indicates a *local maximum*. A quick calculation shows that  $f_{\max} = 37/10$ .

SP2: The second SP is  $\mathbf{x} = (\sqrt{5}/5)[-2, 1]^T$ , for which

$$\lambda = -\frac{11}{2}, \quad \mathbf{L} = \frac{\sqrt{5}}{5} \begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

Therefore,

$$\mathbf{H}_u = \frac{1}{5} \begin{bmatrix} -1 & -2 \end{bmatrix} \begin{bmatrix} 7/2 & -4 \\ -4 & -5/2 \end{bmatrix} \begin{bmatrix} -1 \\ -2 \end{bmatrix} = -\frac{9}{2} < 0$$

which indicates a *local maximum* as well.

SP3: The third SP is  $\mathbf{x} = (\sqrt{5}/5)[1, 2]^T$ , for which

$$\lambda = -\frac{1}{2}, \quad \mathbf{L} = \frac{\sqrt{5}}{5} \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

Therefore,

$$\mathbf{H}_u = \frac{1}{5} \begin{bmatrix} -2 & 1 \end{bmatrix} \begin{bmatrix} 17/2 & -4 \\ -4 & 5/2 \end{bmatrix} \begin{bmatrix} -2 \\ 1 \end{bmatrix} = \frac{21}{2} > 0$$

thereby identifying a *local minimum*. Again, a quick calculation leads to  $f_{\min} = 1/2$ .

SP4: The fourth SP is  $\mathbf{x} = (\sqrt{5}/5)[-1, -2]^T$ , for which

$$\lambda = -\frac{1}{2}, \quad \mathbf{L} = \frac{\sqrt{5}}{5} \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

Therefore,

$$\mathbf{H}_u = \frac{1}{5} \begin{bmatrix} 2 & -1 \end{bmatrix} \begin{bmatrix} 17/2 & -4 \\ -4 & 5/2 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \frac{21}{2} > 0$$

which indicates a *local minimum* as well.

The four stationary points and the ellipses they define, tangent to the unit circle, are displayed in Fig. 5.13.

More general problems of this family can be solved using the methods discussed in Ch. 6 for arbitrary objective functions subject to nonlinear equality constraints.

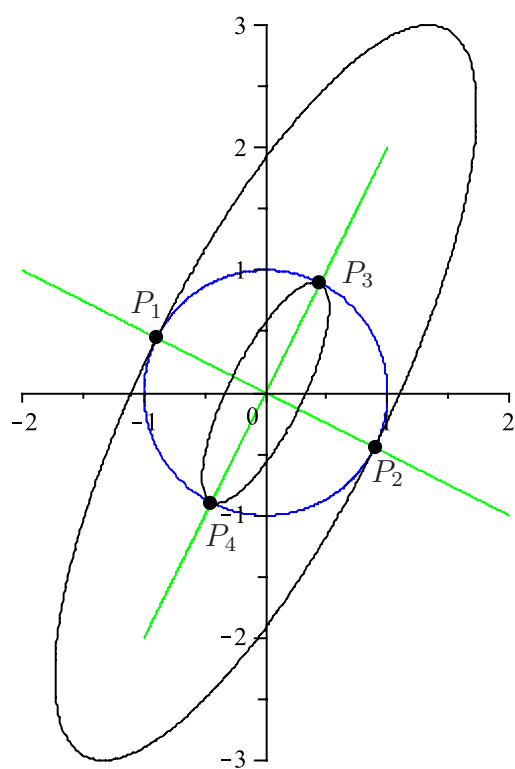


Figure 5.13: The four stationary points of Example 5.6.1



## Chapter 6

# Equality-Constrained Optimization: The Orthogonal-Decomposition Algorithm

### 6.1 Introduction

The main objective of this chapter is the development of the *orthogonal-decomposition algorithm* (ODA), which is derived first in the context of equality-constrained linear least-square problems; then, it is applied to equality-constrained nonlinear least-square problems. Several numerical techniques, such as the QR-decomposition, the Cholesky-decomposition, the Newton-Gauss method, etc., are applied in order to obtain numerical solutions by means of procedures that are both *efficient* and *robust*. What we mean by the former is procedures that use as few floating-point operations (flops) as possible; by the latter we mean procedures that keep the roundoff error in the solution as low as possible with respect to that of the data, an item that falls in the realm of *numerical conditioning*, a subject that was introduced in Subsection 3.4.1.

The orthogonal-decomposition algorithm is implemented in a C library of routines called ODA and its Matlab version; both are available at

`www.cim.mcgill.ca/~rmsl/index/index.htm`

In Section 6.4, two methods applicable to arbitrary objective functions, not stem-

ming from a least-square problem, are developed. In the first method, the assumptions are made that *i*) the objective function is  $C^2$ -continuous and *ii*) its second partial derivatives with respect to the design variables are available. Assumption *i*) depends both on the nature of the problem *and* on the objective function to minimize (or maximize). For example, in structural optimization, as long as the stress levels are below rupture, stress values are smooth functions of the loads, and hence,  $C^2$ -continuous. If the objective function, under these conditions, is the von Mises stress, which is a smooth scalar function of the stress tensor, then  $C^2$ -continuity is guaranteed. However, if the Tresca stress is the objective function, then the nature of the latter changes dramatically, as this function is not even  $C^1$ -continuous. Moreover, even in the presence of  $C^2$ -continuity, assumption *ii*) may not hold, as second-order derivatives may be prohibitively expensive to compute. In FEA<sup>1</sup>, even first-order derivatives are usually out of the question. Thus, in practice both  $C^2$ -continuity and second derivatives are seldom found. We propose therefore a second method that relies only on  $C^1$ -continuity and first-order derivatives of the objective and the constraint functions.

## 6.2 Linear Least-square Problems Subject to Linear Equality Constraints: the ODA

We recall below the linear least-square problem subject to linear equality constraints: Given the overdetermined system of linear equations

$$\mathbf{Ax} = \mathbf{b} \tag{6.1}$$

find a vector  $\mathbf{x}$  that verifies the above system with the least-square error. For the sake of generality, the error in the approximation is defined as in the case of weighted least-square problems—see Subsection 3.5.2—i.e., as

$$f \equiv \frac{1}{2}(\mathbf{Ax} - \mathbf{b})^T \mathbf{W}(\mathbf{Ax} - \mathbf{b}) \rightarrow \min_{\mathbf{x}} \tag{6.2}$$

subject to the linear constraints

$$\mathbf{Cx} = \mathbf{d} \tag{6.3}$$

Here,  $\mathbf{x}$  is the  $n$ -dimensional vector of design variables, while  $\mathbf{A}$  and  $\mathbf{C}$  are  $q \times n$  and  $p \times n$  matrices, while  $\mathbf{b}$  and  $\mathbf{d}$  are  $q$ - and  $p$ -dimensional vectors. Moreover,  $\mathbf{W}$  is a  $q \times q$  positive-definite<sup>2</sup> weighting matrix, with  $q$ ,  $p$  and  $n$  subject to

---

<sup>1</sup>Finite element analysis.

<sup>2</sup>Should  $\mathbf{W}$  fail to be positive-definite, we wouldn't have a least-square problem!

$$q > n \quad \text{and} \quad p < n \quad (6.4a)$$

Note that the first of the foregoing inequalities excludes the possibility of a unique solution upon solving for  $\mathbf{x}$  from eq.(6.1), the second preventing a unique solution from eq.(6.3). Moreover, for the foregoing constrained problem to be overdetermined, the number  $q$  of eqs.(6.1) must be greater than the *degree of freedom*<sup>3</sup> allowed by the constraints (6.3), and hence,  $p$ ,  $q$  and  $n$  must obey one more constraint:

$$q > n - p \quad (6.4b)$$

If  $\mathbf{A}$  and  $\mathbf{C}$  are *full-rank matrices*, then the foregoing problem was shown to have a *unique* solution, given by eqs.(5.68a-d), which is reproduced below for quick reference:

$$\mathbf{x} = \mathbf{P}\mathbf{B}\mathbf{b} + \mathbf{D}\mathbf{d} \quad (6.5a)$$

where,

$$\mathbf{P} = \mathbf{1}_n - \mathbf{D}\mathbf{C} \quad (6.5b)$$

$$\mathbf{B} = (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W} \quad (6.5c)$$

$$\mathbf{D} = (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{C}^T [\mathbf{C}(\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{C}^T]^{-1} \quad (6.5d)$$

and  $\mathbf{1}_n$  is, as usual, the  $n \times n$  identity matrix.

As pointed out in Subsection 5.4.2, the above expression is unsuitable for numerical implementation. A popular approach to obtaining the solution under study consists in partitioning  $\mathbf{C}$  into a  $p \times p$  and a  $p \times (n - p)$  submatrices, where care should be taken so as to choose a well-conditioned  $p \times p$  matrix, for safe inversion. Correspondingly, vector  $\mathbf{x}$  should be partitioned into a *master* part  $\mathbf{x}_M$ , of  $n - p$  components, and a *slave* part  $\mathbf{x}_S$  of  $p$  components. Thus, the constraint equations would be solved for the slave part in terms of the master part and the problem would reduce to an unconstrained least-square problem of dimension  $n - p$ . However, an arbitrary partitioning of  $\mathbf{C}$  may lead to an ill-conditioned  $p \times p$  block, even if  $\mathbf{C}$  itself is well-conditioned. This situation can be prevented if, out of all  $N$  possible partitionings of  $\mathbf{C}$ , the one with the lowest condition number is chosen. Note that the number of partitionings is given by

$$N = \frac{n!}{p!(n - p)!}$$

---

<sup>3</sup>The degree of freedom in question is obviously the difference between the number  $n$  of variables, unknowns in the case at hand, and that of constraints,  $p$ .

and hence,  $N$  can become quite large, even for modest values of  $n$  and  $p$ . Since calculating the condition number of a matrix is a computationally costly procedure, this approach is to be avoided.

Alternatively, by introduction of the singular values of  $\mathbf{C}$  (Strang, 1988), a subsystem of  $p$  equations in  $p$  unknowns, which are linear combinations of the components of  $\mathbf{x}$ , can be found that is optimally conditioned. The computation of singular values, however, similar to that of eigenvalues, is a problem even more difficult to solve than the one at hand, for it is nonlinear and must be solved iteratively. Therefore, the singular-value approach is strongly recommended against.

One more approach is introduced here, which stems from the *geometric interpretation* of the solution (6.5a). Indeed, vector  $\mathbf{Bb}$  of that solution represents the unconstrained *weighted* least-square approximation of eq.(6.2). The second term of the right-hand side of eq.(6.5a) is the minimum-norm solution of the underdetermined system (6.3), based on the weighted norm of  $\mathbf{x}$ , i.e.,

$$\|\mathbf{x}\|_W^2 = \mathbf{x}^T \mathbf{A}^T \mathbf{W} \mathbf{A} \mathbf{x} \quad (6.6)$$

in which the weighting matrix is the positive-definite product  $\mathbf{A}^T \mathbf{W} \mathbf{A}$ . Moreover,  $\mathbf{P}$  is a *projector*<sup>4</sup> onto the null space of  $\mathbf{C}$ . Indeed, one can readily prove that every  $n$ -dimensional vector  $\mathbf{x}$  is mapped by  $\mathbf{P}$  onto the null space of  $\mathbf{C}$ . Moreover,  $\mathbf{P}^2$  can be proven to equal  $\mathbf{P}$ , thereby making apparent that  $\mathbf{P}$  is, in fact, a projector.

**Exercise 6.2.1** *Prove that, under the assumptions spelled out above, (a)  $\mathbf{A}^T \mathbf{W} \mathbf{A}$  is positive-definite, and (b)  $\mathbf{P}$ , as given by eq.(6.5b), is a projector.*

Furthermore, for any  $p \times n$  matrix  $\mathbf{C}$ ,  $\mathcal{R}(\mathbf{C})$  and  $\mathcal{N}(\mathbf{C})$  are orthogonal subspaces<sup>5</sup> of  $\mathbb{R}^n$ , their *direct sum* producing all of  $\mathbb{R}^n$ ; i.e., every  $n$ -dimensional vector  $\mathbf{x}$  can be *uniquely* decomposed into a vector lying in the range of  $\mathbf{C}^T$  and a second one lying in the null space of  $\mathbf{C}$ . Now let  $\mathbf{L}$  be a  $n \times (n - p)$  matrix spanning the null space of  $\mathbf{C}$ , i.e., its  $n - p$   $n$ -dimensional columns lie in  $\mathcal{N}(\mathbf{C})$ , which means that

$$\mathbf{C} \mathbf{L} = \mathbf{O}_{pn'} \quad (6.7)$$

with  $\mathbf{O}_{pn'}$  representing the  $p \times (n - p)$  zero matrix. Matrix  $\mathbf{L}$  is thus an *orthogonal complement* of matrix  $\mathbf{C}$ . Now, the solution to the above problem can be decomposed into two parts, namely,

---

<sup>4</sup>Note that  $\mathbf{P}$  is apparently not symmetric!

<sup>5</sup>Note that, for two subspaces to be orthogonal, they must be embedded in the same space. In our case,  $\mathcal{R}(\mathbf{C}) \subseteq \mathbb{R}^p \subset \mathbb{R}^n$  and  $\mathcal{N}(\mathbf{C}) \subset \mathbb{R}^n$ .

$$\mathbf{x} = \mathbf{x}_o + \mathbf{x}_u \quad (6.8)$$

in which  $\mathbf{x}_o$  represents the minimum-norm solution to the constraint equation (6.3), i.e.,  $\mathbf{x}_o$  lies in the range of  $\mathbf{C}^T$ , while  $\mathbf{x}_u$  lies in the null space of  $\mathbf{C}$ . Vector  $\mathbf{x}_o$  is computed by means of the QR-decomposition of  $\mathbf{C}^T$ , as discussed in Subsection 5.4.1, while vector  $\mathbf{x}_u$  is computed by means of a linear least-square problem. We outline below the computation of  $\mathbf{x}_u$ .

Let us define a  $q \times q$  matrix  $\mathbf{V}$  as the right Cholesky factor of the given weighting matrix  $\mathbf{W}$ , i.e.,

$$\mathbf{W} = \mathbf{V}^T \mathbf{V} \quad (6.9)$$

Moreover, with  $\mathbf{x}_o$  known,  $\mathbf{x}_u$  is found as the least-square approximation of

$$\mathbf{V} \mathbf{A} \mathbf{x}_u = \mathbf{V}(\mathbf{b} - \mathbf{A} \mathbf{x}_o) \quad (6.10)$$

subject to the constraints

$$\mathbf{C} \mathbf{x}_u = \mathbf{0}_p \quad (6.11)$$

Further, let us represent  $\mathbf{x}_u$  as the image of a  $(n - p)$ -dimensional vector under a transformation given by a  $n \times (n - p)$  matrix  $\mathbf{L}$ , namely,

$$\mathbf{x}_u = \mathbf{L} \mathbf{u} \quad (6.12)$$

with  $\mathbf{L}$  defined, in turn, as introduced in eq.(6.7). Equation (6.10) thus becomes

$$\mathbf{V} \mathbf{A} \mathbf{L} \mathbf{u} = \mathbf{V}(\mathbf{b} - \mathbf{A} \mathbf{x}_o) \quad (6.13)$$

which is an overdetermined system of  $q$  linear equations in  $n - p$  unknowns. It is thus apparent that  $\mathbf{u}$  can be computed as the *unconstrained* least-square solution of eq.(6.13).

However, matrix  $\mathbf{L}$ , an orthogonal complement of  $\mathbf{C}$ , *is not unique*. We have thus reached a crucial point in the solution of the constrained linear least-square problem at hand: how to define  $\mathbf{L}$ . While  $\mathbf{L}$  can be defined in infinitely many forms—notice that, once *any*  $\mathbf{L}$  has been found, a multiple of it also satisfies eq.(6.7)—we invoke here Exercise 5.4.2, and hence, define  $\mathbf{L}$  as the  $n \times (n - p)$   $\mathbf{Q}_R$  block of matrix  $\mathbf{Q}$  of the QR-decomposition of  $\mathbf{C}^T$ , namely,

$$\mathbf{L} = \mathbf{Q}_R \quad (6.14)$$

In this way,  $\mathbf{L}$  has an extremely simple LMPGI, namely,

$$\mathbf{L}^I = \mathbf{L}^T = \mathbf{Q}_R^T \quad (6.15)$$

which is a consequence of the orthogonality of  $\mathbf{Q}$ : its columns, or its rows for that matter, form an orthonormal set. In particular, the columns of  $\mathbf{Q}_L$  and  $\mathbf{Q}_R$  are orthonormal subsets of  $p$  and  $n - p$  vectors in  $\mathbb{R}^n$ , respectively. In this light, both  $\mathbf{Q}_L$  and  $\mathbf{Q}_R$  are isotropic matrices, i.e., of  $\kappa = 1$ , whose *weighted* Frobenius norms, with weighting matrices  $\mathbf{W}_L = (1/p)\mathbf{1}_p$  and  $\mathbf{W}_R = (1/n')\mathbf{1}_{n'}$ , respectively— $n'$  is defined as  $n - p$ —are, moreover, unity.

Once  $\mathbf{L}$  is known, eq.(6.13) can be solved for  $\mathbf{u}$  as the least-square approximation of that system. Then,  $\mathbf{x}_u$  is calculated from eq.(6.12). Notice that, in calculating  $\mathbf{u}$ , the QR-decomposition of the product  $\mathbf{V}\mathbf{A}\mathbf{L}$  will be needed, rather than the one of  $\mathbf{A}$  alone.

As the reader can readily prove, the two components of  $\mathbf{x}$ ,  $\mathbf{x}_o$  and  $\mathbf{x}_u$ , are orthogonal. For this reason, the foregoing procedure is known as the *Orthogonal-Decomposition Algorithm* (ODA).

### 6.3 Equality-Constrained Nonlinear Least-Square Problems

The solution of nonlinear least-square problems by means of the ODA is now straightforward: The problem consists in finding the least-square error  $f$  of an overdetermined system of nonlinear equations,  $\phi(\mathbf{x}) = \mathbf{0}_q$ , i.e.,

$$f(\mathbf{x}) = \frac{1}{2}\phi^T \mathbf{W} \phi \rightarrow \min_{\mathbf{x}} \quad (6.16a)$$

subject to the nonlinear constraints

$$\mathbf{h}(\mathbf{x}) = \mathbf{0}_p \quad (6.16b)$$

where  $\phi$  and  $\mathbf{x}$  are  $q$ - and  $n$ -dimensional vectors, respectively, with  $q > n$ , and  $\mathbf{W}$  is a  $q \times q$  positive-definite weighting matrix. Moreover,  $\mathbf{h}$  is a  $p$ -dimensional vector of nonlinear constraints.

The normality condition of the foregoing constrained problem was derived in Ch. 5 in its dual form, eq.(5.74), and recalled below for quick reference:

$$\mathbf{L}^T \Phi^T \mathbf{W} \phi = \mathbf{0}_{n'} \quad (6.17)$$

with  $\mathbf{0}_{n'}$  denoting the  $(n - p)$ -dimensional zero vector,  $\Phi$  the  $q \times n$  gradient of  $\phi$

with respect to  $\mathbf{x}$ , and  $\mathbf{L}$  the  $n \times (n - p)$  *isotropic* orthogonal complement<sup>6</sup> of  $\mathbf{J}$ , the  $p \times n$  gradient of  $\mathbf{h}$  with respect to  $\mathbf{x}$ , or  $\nabla \mathbf{h}$ .

As the problem is nonlinear, its solution is obtained *iteratively*, within the spirit of the Newton-Gauss method introduced in Subsection 3.7.1: From an initial guess  $\mathbf{x}^0$ , *not necessarily feasible*, i.e., with  $\mathbf{h}(\mathbf{x}^0) \neq \mathbf{0}_l$ , the sequence  $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^k, \mathbf{x}^{k+1}$  is generated as

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \Delta \mathbf{x}^k \quad (6.18)$$

The increment  $\Delta \mathbf{x}^k$  is computed as the solution of an equality-constrained linear least-square problem, namely,

$$\min_{\Delta \mathbf{x}^k} \frac{1}{2} [\phi(\mathbf{x}^{k+1})]^T \mathbf{W} \phi(\mathbf{x}^{k+1}) \quad (6.19a)$$

subject to

$$\mathbf{J}(\mathbf{x}^k) \Delta \mathbf{x}^k = -\mathbf{h}(\mathbf{x}^k) \quad (6.19b)$$

with  $\phi(\mathbf{x}^{k+1})$  given, to a first-order approximation, by

$$\phi(\mathbf{x}^{k+1}) \equiv \phi(\mathbf{x}^k) + \Phi(\mathbf{x}^k) \Delta \mathbf{x}^k \quad (6.19c)$$

Now, for compactness, we introduce a few definitions:

$$\mathbf{h}^k \equiv \mathbf{h}(\mathbf{x}^k), \quad \phi^k \equiv \phi(\mathbf{x}^k), \quad \Phi_k \equiv \Phi(\mathbf{x}^k), \quad \mathbf{J}_k \equiv \mathbf{J}(\mathbf{x}^k) \quad (6.20)$$

while  $\mathbf{L}_k$  is defined as the *isotropic* orthogonal complement of  $\mathbf{J}_k$  and of *unit weighted Frobenius norm*, namely, the  $\mathbf{Q}_R$  block of matrix  $\mathbf{Q}$  of the QR-decomposition of  $\mathbf{J}_k^T$ . Indeed,  $\mathbf{J}_k$  now plays the role of matrix  $\mathbf{C}$  of eq.(6.3) in the linear case. Moreover,  $\Phi_k$  and  $\mathbf{J}_k$  will be assumed to be of full rank throughout, the solution  $\Delta \mathbf{x}^k$  of problem (6.19a–c) thus being expressed as in the linear case, namely,

$$\Delta \mathbf{x}^k = \Delta \mathbf{v}^k + \mathbf{L}_k \Delta \mathbf{u}^k \quad (6.21)$$

A procedure to compute  $\Delta \mathbf{v}^k$  and  $\Delta \mathbf{u}^k$  is introduced below. To this end, eq.(6.19b) is first rewritten in terms of the orthogonal decomposition of  $\Delta \mathbf{x}^k$ :

$$\mathbf{J}_k \Delta \mathbf{v}^k + \mathbf{J}_k \mathbf{L}_k \Delta \mathbf{u}^k = -\mathbf{h}^k \quad (6.22)$$

By virtue of the orthogonal-complementarity relation between  $\mathbf{J}_k$  and  $\mathbf{L}_k$ , however, eq.(6.22) reduces to

$$\mathbf{J}_k \Delta \mathbf{v}^k = -\mathbf{h}^k \quad (6.23)$$

---

<sup>6</sup>As a matter of fact, matrix  $\mathbf{L}$  appearing in eq.(6.17), need not be isotropic for the normality condition to hold. *Any* orthogonal complement of  $\mathbf{J}$  will do.

Next, an expression is derived for  $f_{k+1} \equiv f(\mathbf{x}^{k+1})$ :

$$f_{k+1} \equiv \frac{1}{2}(\phi^k + \Phi_k \Delta \mathbf{x}^k)^T \mathbf{V}^T \mathbf{V}(\phi^k + \Phi_k \Delta \mathbf{x}^k) \equiv \frac{1}{2} \|\mathbf{V}(\phi^k + \Phi_k \Delta \mathbf{x}^k)\|^2 \quad (6.24)$$

where the Cholesky-decomposition of  $\mathbf{W}$ , introduced in eq.(6.9), has been recalled. Hence, in computing  $\Delta \mathbf{x}^k$  the aim will be *to render vector  $\mathbf{V}(\phi^k + \Phi_k \Delta \mathbf{x}^k)$  as close to zero as possible*, i.e.,

$$\mathbf{V}\phi^k + \mathbf{V}\Phi_k \Delta \mathbf{x}^k \rightarrow \mathbf{0}_p$$

or, in terms of the orthogonal decomposition of  $\Delta \mathbf{x}^k$  given in eq.(6.21),

$$\mathbf{V}\Phi_k(\Delta \mathbf{v}^k + \mathbf{L}_k \Delta \mathbf{u}^k) + \mathbf{V}\phi^k = \mathbf{0}_p$$

whence<sup>7</sup>,

$$\underbrace{\mathbf{V}\Phi_k}_{\mathbf{B}_k} \mathbf{L}_k \Delta \mathbf{u}^k = -\mathbf{V}(\phi^k + \Phi_k \Delta \mathbf{v}^k) \quad (6.25)$$

which is an overdetermined system of  $q$  linear equations in  $n - p < q$  unknowns, the components of  $\Delta \mathbf{u}^k$ .

The computation of  $\Delta \mathbf{v}^k$  and  $\Delta \mathbf{u}^k$  now proceeds *sequentially*:

1. Compute  $\Delta \mathbf{v}^k$  as the *minimum-norm solution* of the underdetermined system of eq.(6.23):

$$\Delta \mathbf{v}^k = -\mathbf{J}_k^\dagger \mathbf{h}^k \quad (6.26)$$

where  $\mathbf{J}_k^\dagger$  is computed as  $\mathbf{Q}_L \mathbf{U}^{-T}$ ,  $\mathbf{Q}_L$  and  $\mathbf{U}$  being blocks of the QR-decomposition of  $\mathbf{J}_k^T$ , similar to the blocks of the same decomposition of  $\mathbf{C}^T$  in eq.(5.53).

2. with  $\Delta \mathbf{v}^k$  known, compute  $\Delta \mathbf{u}^k$  as the least-square approximation of eq.(6.25):

$$\Delta \mathbf{u}^k = -(\mathbf{V}\Phi_k \mathbf{L}_k)^I \mathbf{V}(\phi^k - \Phi_k \mathbf{J}_k^\dagger \mathbf{h}^k) \quad (6.27)$$

the LMPGI of  $\mathbf{V}\Phi_k \mathbf{L}_k$  being computed with the QR-decomposition of the foregoing product.

The stopping criteria of the procedure are, then,

$$\|\Delta \mathbf{x}^k\| \leq \epsilon_1 \quad \text{and} \quad \|\mathbf{h}(\mathbf{x}^k)\| \leq \epsilon_2 \quad (6.28)$$

for prescribed tolerances  $\epsilon_1$  and  $\epsilon_2$ . The second inequality of eq.(6.28) apparently implies that the constraints, eq.(6.16b), are satisfied within the prescribed tolerance

---

<sup>7</sup>Although  $\mathbf{B}_k$  is not needed in the discussion below, it is defined here because it is an auxiliary variable in `oda-quad.mw`.



$\epsilon_2$ . It is shown in the sequel that the first inequality of eq.(6.28) guarantees that the normality condition (6.17) is verified within the prescribed tolerance  $\epsilon_1$ .

With  $\Delta \mathbf{v}^k$  known,  $\Delta \mathbf{u}^k$  is computed as the least-square approximation of eq.(6.25)<sup>8</sup>:

$$\Delta \mathbf{u} = -(\mathbf{V}\Phi\mathbf{L})^T \mathbf{V}(\phi - \Phi\mathbf{J}^\dagger \mathbf{h})$$

The LMPGI of the product  $\mathbf{V}\Phi\mathbf{Q}_R$  is displayed below for purposes of analysis only, its computation to be implemented with the aid of the QR-decomposition of the product in question.

$$(\mathbf{V}\Phi\mathbf{L})^T = (\mathbf{L}^T \Phi^T \mathbf{V}^T \mathbf{V} \Phi \mathbf{L})^{-1} \mathbf{L}^T \Phi^T \mathbf{V} = (\mathbf{L}^T \Phi^T \mathbf{W} \Phi \mathbf{L})^{-1} \mathbf{L}^T \Phi^T \mathbf{V}$$

Hence,

$$\Delta \mathbf{u} = -(\mathbf{L}^T \Phi^T \mathbf{W} \Phi \mathbf{L})^{-1} \mathbf{L}^T \Phi^T \mathbf{W}(\phi - \Phi\mathbf{J}^\dagger \mathbf{h}) \quad (6.29)$$

Now, from eq.(6.21)  $\Delta \mathbf{x}$  can be expressed as

$$\Delta \mathbf{x} = (\mathbf{M}\Phi^T \mathbf{W} \Phi - \mathbf{1}_n) \mathbf{J}^\dagger \mathbf{h} - \mathbf{M}\Phi^T \mathbf{W} \phi \quad (6.30)$$

where  $\mathbf{M}$  is the  $n \times n$  matrix defined as

$$\mathbf{M} = \mathbf{L}(\mathbf{L}^T \Phi^T \mathbf{W} \Phi \mathbf{L})^{-1} \mathbf{L}^T \quad (6.31)$$

Upon convergence, *i*)  $\Delta \mathbf{x} \rightarrow \mathbf{0}_n$ , as its norm does within  $\epsilon_1$ , and *ii*) the constraint equations hold within  $\epsilon_2$ , and hence,  $\mathbf{h} \rightarrow \mathbf{0}_l$ . Now, setting  $\Delta \mathbf{x} = \mathbf{0}_n$ ,  $\mathbf{h} = \mathbf{0}_p$  and substituting  $\mathbf{M}$  by its expression (6.31) into eq.(6.30), one obtains

$$\mathbf{L}(\mathbf{L}^T \Phi^T \mathbf{W} \Phi \mathbf{L})^{-1} \underbrace{\mathbf{L}^T \Phi^T \mathbf{W} \phi}_{\mathbf{n}_c} = \mathbf{0}_n \quad (6.32)$$

which implies that  $\mathbf{n}_c = \mathbf{0}_{n-p}$ . Upon comparing  $\mathbf{n}_c$  with the left-hand side of eq.(6.17), it becomes apparent that eq.(6.32) implies the verification of the normality condition.

In summary, then if the ODA converged as per eq.(6.28), then both the normality conditions and the constraints are verified. The sequence  $\{\Delta \mathbf{x}^k\}$  produces a sequence  $\{f_k\}$ ; the increment  $\Delta f$  between two consecutive values of the former is given, to first order, by

$$\Delta f = (\nabla f)^T \Delta \mathbf{x} \quad (6.33)$$

---

<sup>8</sup>For simplicity, upper- and subscript  $k$  is dropped henceforth.

where  $\nabla f = \Phi^T \mathbf{W} \phi$ , and hence,

$$\begin{aligned} \Delta f &= (\Phi^T \mathbf{W} \phi)^T \Delta \mathbf{x} \\ &= -\phi^T \mathbf{W} \Phi \mathbf{M} \Phi^T \mathbf{W} \phi - \phi^T \mathbf{W} \Phi (\mathbf{1}_n - \mathbf{M} \Phi^T \mathbf{W} \Phi) \mathbf{J}^\dagger \mathbf{h} \end{aligned} \quad (6.34)$$

From eq.(6.34), if the current value of  $\mathbf{x}$  is feasible, i.e., if  $\mathbf{h} = \mathbf{0}$ , then  $\Delta f$  is negative-definite, and the procedure yields an improved value of  $f$ . Therefore, the ODA iterations lead to a feasible local minimum in the proximity of a feasible SP.

Furthermore,

$$\Delta(\mathbf{h}^T \mathbf{h}) = (\Delta \mathbf{h})^T \mathbf{h} + \mathbf{h}^T \Delta \mathbf{h} = 2\mathbf{h}^T \Delta \mathbf{h}$$

However, since  $\mathbf{J} = \nabla \mathbf{h}$ ,

$$\Delta \mathbf{h} = \mathbf{J} \Delta \mathbf{x}$$

But, from eq.(6.19b),  $\mathbf{J} \Delta \mathbf{x} = -\mathbf{h}$  at each iteration, and hence,

$$\Delta \mathbf{h} = -\mathbf{h} \quad (6.35)$$

Therefore,

$$\Delta(\mathbf{h}^T \mathbf{h}) = -2\mathbf{h}^T \mathbf{h} \quad (6.36)$$

the result being that the ODA converges towards a *feasible solution*, and hence, the initial guess need not be feasible. This is important, as many a method requires that the initial guess be feasible. Finding a feasible guess may be as hard a task as that of finding a feasible stationary point. From eq.(6.34), if the current value of  $\mathbf{x}$  is feasible, i.e., if  $\mathbf{h} = \mathbf{0}$ , then  $\Delta f$  is negative-definite, and the procedure yields an improved value of  $f$ . Therefore, the ODA iterations lead always to a local minimum, although the least-square problem may contain local maxima, as found in Example 6.3.1, or even saddle points.

**Example 6.3.1 (A Quadratic Objective Function with a Quadratic Constraint)** We recall Example 5.6.1, which is reproduced below for quick reference:

$$f(\mathbf{x}) = \frac{1}{2}(9x_1^2 - 8x_1x_2 + 3x_2^2) \rightarrow \min_{x_1, x_2}$$

subject to

$$h(\mathbf{x}) = x_1^2 + x_2^2 - 1 = 0$$

The above objective function is quadratic in the design-variable vector and the associated matrix is positive-definite, as found in Example 5.6.1, which means that the

problem is of the least-square class. However, the constraint is quadratic, which disqualifies this problem from a direct solution—see Section 6.2 for linear least-squares subject to linear constraints.

*Solution:* This problem, due to its simplicity, could be solved exactly in Section 5.6. Here, we solve this problem numerically, using the ODA. First, we recall that the objective function  $f(\mathbf{x})$  can be factored as

$$f(\mathbf{x}) = \frac{1}{2}\boldsymbol{\phi}^T \mathbf{W} \boldsymbol{\phi}$$

with

$$\mathbf{W} = \begin{bmatrix} 9 & -4 \\ -4 & 3 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\phi} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{\Phi} = \mathbf{1}_2$$

i.e,  $f(\mathbf{x})$  is a special case of the  $f(\mathbf{x})$  defined in eq.(6.2), with  $\mathbf{A} = \mathbf{1}$  and  $\mathbf{b} = \mathbf{0}$ . We include below the Maple worksheet `oda-quad.mw` describing the step-by-step implementation of the ODA in solving the foregoing problem *iteratively*. In this worksheet, the *QR-decomposition* is implemented with *Householder reflections (HHR)*, as introduced in Subsection 3.5.3

```
> restart:with(LinearAlgebra):
```

```
Warning, the protected names norm and trace have been redefined and
unprotected
```

```
> with(plots): with(plottools):
```

```
Warning, the name changecoords has been redefined
```

```
Warning, the name arrow has been redefined
```

### Linear-least square problem subject to a quadratic constraint

$$f(x) = (1/2)(9x_1^2 - 8x_1x_2 + 3x_2^2) \quad \rightarrow \quad \min_{x_1, x_2}$$

subject to

$$h(x_1, x_2) = x_1^2 + x_2^2 - 1 = 0$$

```
> obj:= proc(x) (1/2)*(9*x[1]^2 - 8*x[1]*x[2] +
> 3*x[2]^2)
> end; #procedure to compute the objective function
```

```

      obj := proc(x) 9/2 * x1^2 - 4 * x1 * x2 + 3/2 * x2^2 end proc
> constr:= proc(x) x[1]^2+x[2]^2 - 1 end;
> #procedure computing the constraint
      constr := proc(x) x1^2 + x2^2 - 1 end proc
> dhdx:= proc(x) Matrix([[2*x[1], 2*x[2]]])
> end; #procedure computing the gradient of the constraint
      dhdx := proc(x) Matrix([[2 * x1, 2 * x2]]) end proc
> alfa:= proc(J)
> evalf(signum(J[1,1])*sqrt(J[1,1]^2 + J[1,2]^2))
> end; #procedure computing "alpha" of Householder reflections in
> least-square solution at each iteration
      alfa := proc(J) evalf(signum(J1,1) * sqrt(J1,1^2 + J1,2^2)) end proc
> W:=Matrix([[9, -4], [-4, 3]]); #weighting matrix

```

$$W := \begin{bmatrix} 9 & -4 \\ -4 & 3 \end{bmatrix}$$

```

> V:=Transpose(cholesky(W)); #Maple returns a
> lower-triangular matrix with procedure "cholesky"!

```

$$V := \begin{bmatrix} 3 & \frac{-4}{3} \\ 0 & \frac{1}{3}\sqrt{11} \end{bmatrix}$$

```

> V:= map(evalf, V);

```

$$V := \begin{bmatrix} 3. & -1.333333333 \\ 0. & 1.105541597 \end{bmatrix}$$

```

> ID:=Matrix(2,2,shape=identity); E:=
> Matrix([[0], [1]]);#right-hand side of eq.(6.13)
> Phi:= ID;
> B:=MatrixMatrixMultiply(V, Phi); #Defining various auxiliary matrices;
for B see eq.(6.25)

```

$$ID := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$E := \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\Phi := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

```


$$B := \begin{bmatrix} 3. & -1.333333333 \\ 0. & 1.105541597 \end{bmatrix}$$

> x:=Vector([2, 2]); x0:= map(evalf,x); #initial guess, x^0, stored
as x0 for plotting

$$x := [2, 2]$$


$$x0 := [2, 2]$$

> f:= evalf(obj(x)); #f_0

$$f := 8.$$

> phi:= map(evalf,x); #phi^0

$$\phi := [2, 2]$$

> h:= constr(x); #h^0

$$h := 7$$

> J:= dhdx(x); #J_0

$$J := \begin{bmatrix} 4 & 4 \end{bmatrix}$$

> alpha:= alfa(J); #local variable

$$\alpha := 5.656854248$$

> t:=Vector([J[1,1] + alpha, J[1,2]]);
> #u in HHR algorithm, a local variable

$$t := [9.656854248, 4]$$

> normt2:=evalf(DotProduct(t,t)/2); #half of
> Euclidean norm-squared of t, a local variable

$$normt2 := 54.62741700$$

> H:=evalf(ID-MatrixMatrixMultiply(t, Transpose(t)/normt2));
> #evaluating Householder reflection

$$H := \begin{bmatrix} -.707106781 & -.7071067812 \\ -.7071067812 & .7071067811 \end{bmatrix}$$

> P:=MatrixMatrixMultiply(H, Transpose(H)); #checking whether H
> is a reflection

$$P := \begin{bmatrix} .9999999997 & 0. \\ 0. & .9999999999 \end{bmatrix}$$

> detH:=Determinant(H);

$$detH := -.9999999998$$


```

**H** is indeed a reflection!

```

> HJT:=MatrixMatrixMultiply(H, Transpose(J));

$$HJT := \begin{bmatrix} -5.656854249 \\ -.110^{-8} \end{bmatrix}$$

> HJT[2,1]:=0; print(HJT); #setting last entry
> of HJ^T equal to zero

$$HJT_{2,1} := 0$$


$$\begin{bmatrix} -5.656854249 \\ 0 \end{bmatrix}$$

> w:= Vector([-h/HJT[1,1], 0]); #w = Hv, eq.(??) with Deltav & Deltaw
replaced with v & w here

$$w := [1.237436867, 0]$$

> v:=MatrixMatrixMultiply(H, w); #v^0

$$v := [-.8749999997, -.8750000000]$$

> L := MatrixMatrixMultiply(H, E);
> BL := MatrixMatrixMultiply(B, L); #L_0 & (BL)_0

$$L := \begin{bmatrix} -.7071067812 \\ .7071067811 \end{bmatrix}$$


$$BL := \begin{bmatrix} -3.064129385 \\ .7817359600 \end{bmatrix}$$

> p:=Add(phi, MatrixMatrixMultiply(Phi, v)); #auxiliary variable

$$p := [1.125000000, 1.125000000]$$

> r:= MatrixMatrixMultiply(-V, p);
> #RHS of overdetermined system to compute u in ODA

$$r := [-1.875000000, -1.243734297]$$

> u:= LeastSquares(BL, r); #u^0

$$u := [.4772970772]$$

> Deltax:= Add(v, MatrixMatrixMultiply(L, u)); #Deltax^0

$$Deltax := [-1.212500000, -.5375000001]$$


```

First iteration is complete. Update **x**:

```

> x:= map(evalf,x + Deltax); x1:= map(evalf,x); #x^1

$$x := [.787500000, 1.462500000]$$


```

```

                                 $x1 := [.787500000, 1.462500000]$ 
> f:= evalf(obj(x)); #f_1
                                 $f := 1.392187500$ 
> phi:= map(evalf,x); #phi^1
                                 $\phi := [.787500000, 1.462500000]$ 
> h:= constr(x); #h^1
                                 $h := 1.759062500$ 
> J:= dhdx(x); #J_1
                                 $J := \begin{bmatrix} 1.575000000 & 2.925000000 \end{bmatrix}$ 
> alpha:= alfa(J);
                                 $\alpha := 3.322085189$ 
> t:=Vector([J[1,1] + alpha, J[1,2]]);
> #u in HHR algorithm
                                 $t := [4.897085189, 2.925000000]$ 
> normt2:=evalf(DotProduct(t,t)/2);
                                 $normt2 := 16.26853418$ 
> H:=evalf(ID-MatrixMatrixMultiply(t, Transpose(t)/normt2));
> #evaluating Householder reflection
                                 $H := \begin{bmatrix} -.474099823 & -.8804710997 \\ -.8804710997 & .4740998233 \end{bmatrix}$ 
> HJT:=MatrixMatrixMultiply(H, Transpose(J));
                                 $HJT := \begin{bmatrix} -3.322085188 \\ .1 \cdot 10^{-8} \end{bmatrix}$ 
> HJT[2,1]:=0; print(HJT);
                                 $HJT_{2,1} := 0$ 
                                 $\begin{bmatrix} -3.322085188 \\ 0 \end{bmatrix}$ 
> w:= Vector([-h/HJT[1,1], 0]); #w = Hv
                                 $w := [.5295055366, 0]$ 
> v:=MatrixMatrixMultiply(H, w); # v^1
                                 $v := [-.2510384812, -.4662143221]$ 
> L:= MatrixMatrixMultiply(H, E);
> BL:=MatrixMatrixMultiply(MatrixMatrixMultiply(V, Phi), L);
> #L_1 & (BL)_1

```

$$L := \begin{bmatrix} -.8804710997 \\ .4740998233 \end{bmatrix}$$

$$BL := \begin{bmatrix} -3.273546397 \\ .5241370758 \end{bmatrix}$$

```

> p:=Add(phi, MatrixMatrixMultiply(Phi, v)); #auxiliary variable
      p := [.5364615188, .9962856779]
> r:= MatrixMatrixMultiply(-V, p);
> #RHS of overdetermined system to compute u in ODA
      r := [-.281003652, -1.101435259]
> u:= LeastSquares(BL, r); #u^1
      u := [.03116921755]
> Deltax:= Add(v, MatrixMatrixMultiply(L, u)); #delta x^1
      Deltax := [-.2784820764, -.4514370016]

```

Second iteration is complete. Update x:

```

> x:= Add(x, Deltax); x2:= map(evalf,x); #x^2
      x := [.5090179236, 1.011062998]
      x2 := [.5090179236, 1.011062998]
> f:= obj(x); #f_2
      f := .640722436
> phi:= map(evalf,x); #phi^2
      phi := [.5090179236, 1.011062998]
> h:= constr(x); #h^2
      h := .281347632
> J:= dhdx(x); #J_2
      J := [ 1.018035847  2.022125996 ]
> alpha:= alfa(J);
      alpha := 2.263932537
> t:=Vector([J[1,1] + alpha, J[1,2]]);
> #u in HHR algorithm
      t := [3.281968384, 2.022125996]
> normt2:=evalf(DotProduct(t,t)/2);

```



```

                                normt2 := 7.430155005
> H:=evalf(ID-MatrixMatrixMultiply(t, Transpose(t)/normt2));
> #evaluating Householder
> reflection


$$H := \begin{bmatrix} -.449675877 & -.8931918088 \\ -.8931918088 & .4496758759 \end{bmatrix}$$

> HJT:=MatrixMatrixMultiply(H, Transpose(J));


$$HJT := \begin{bmatrix} -2.263932538 \\ -.12 \cdot 10^{-8} \end{bmatrix}$$

> HJT[2,1]:=0; print(HJT);


$$HJT_{2,1} := 0$$


$$\begin{bmatrix} -2.263932538 \\ 0 \end{bmatrix}$$

> w:= Vector([-h/HJT[1,1], 0]); #w = Hv

$$w := [.1242738586, 0]$$

> v:=MatrixMatrixMultiply(H, w); #v^2

$$v := [-.05588295635, -.1110003925]$$

> L:= MatrixMatrixMultiply(H, E);
> BL:=MatrixMatrixMultiply(MatrixMatrixMultiply(V, Phi), L);
> #L_2 & (BL)_2


$$L := \begin{bmatrix} -.8931918088 \\ .4496758759 \end{bmatrix}$$


$$BL := \begin{bmatrix} -3.279143260 \\ .4971353860 \end{bmatrix}$$

> p:=Add(phi, MatrixMatrixMultiply(Phi, v)); #auxiliary variable

$$p := [.4531349672, .9000626055]$$

> r:= MatrixMatrixMultiply(-V, p);
> #RHS of overdetermined system to compute u in ODA

$$r := [-.159321428, -.9950566503]$$

> u:= LeastSquares(BL, r); #u^2

$$u := [.002523646038]$$

> Deltax:= Add(v, MatrixMatrixMultiply(L, u)); #Deltax^2

$$Deltax := [-.05813705632, -.1098655698]$$


```

Third iteration is complete. Update  $\mathbf{x}$ :

```

> x:= Add(x, Deltax); x3:=map(evalf,x); #x^3


$$x := [.4508808673, .9011974282]$$


$$x^3 := [.4508808673, .9011974282]$$

> f:= obj(x); #f_3


$$f := .5077254992$$

> phi:= map(evalf,x); #phi^3


$$\phi := [.4508808673, .9011974282]$$

> h:= constr(x); #h^3


$$h := .015450361$$

> J:= dhdx(x); #J_3


$$J := \begin{bmatrix} .9017617346 & 1.802394856 \end{bmatrix}$$

> alpha:= alfa(J);


$$\alpha := 2.015391139$$

> t:=Vector([J[1,1] + alpha, J[1,2]]);
> #u in HHR algorithm


$$t := [2.917152874, 1.802394856]$$

> normt2:=evalf(DotProduct(t,t)/2);


$$normt2 := 5.879204055$$

> H:=evalf(ID-MatrixMatrixMultiply(t, Transpose(t)/normt2));
> #evaluating Householder reflection


$$H := \begin{bmatrix} -.447437580 & -.8943151635 \\ -.8943151635 & .4474375804 \end{bmatrix}$$

> HJT:=MatrixMatrixMultiply(H, Transpose(J));


$$HJT := \begin{bmatrix} -2.015391138 \\ .2 \cdot 10^{-9} \end{bmatrix}$$

> HJT[2,1]:=0; print(HJT);


$$HJT_{2,1} := 0$$


$$\begin{bmatrix} -2.015391138 \\ 0 \end{bmatrix}$$

> w:= Vector([-h/HJT[1,1], 0]); #w = Hv


$$w := [.007666184846, 0]$$

> v:=MatrixMatrixMultiply(H, w); #v^3


$$v := [-.003430139195, -.006855985354]$$


```

```

> L:= MatrixMatrixMultiply(H, E);
> BL:=MatrixMatrixMultiply(MatrixMatrixMultiply(V, Phi), L);
> #L_3 & (BL)_3

$$L := \begin{bmatrix} -.8943151635 \\ .4474375804 \end{bmatrix}$$


$$BL := \begin{bmatrix} -3.279528930 \\ .4946608572 \end{bmatrix}$$

> p:=Add(phi, MatrixMatrixMultiply(Phi, v)); #auxiliary variable

$$p := [.4474507281, .8943414428]$$

> r:= MatrixMatrixMultiply(-V, p);
> #RHS of overdetermined system to compute u in ODA

$$r := [-.149896927, -.9887316669]$$

> u:= LeastSquares(BL, r); #u^3

$$u := [.0002276777133]$$

> Deltax:= Add(v, MatrixMatrixMultiply(L, u)); #Deltax^3

$$Deltax := [-.003633754826, -.006754113789]$$


```

Fourth iteration is complete. Update  $\mathbf{x}$ :

```

> x:= Add(x, Deltax); x4:=map(evalf,x); # x^4

$$x := [.4472471125, .8944433144]$$


$$x4 := [.4472471125, .8944433144]$$

> f:= obj(x); #f_4

$$f := .5000294132$$

> phi:= map(evalf,x); #phi^4

$$\phi := [.4472471125, .8944433144]$$

> h:= constr(x); #h^4

$$h := .000058822$$

> J:= dhdx(x); #J_4

$$J := \begin{bmatrix} .8944942250 & 1.788886629 \end{bmatrix}$$

> alpha:= alfa(J);

$$\alpha := 2.000058822$$

> t:= Vector([J[1,1] + alpha, J[1,2]]);
> #u in HHR algorithm

```

```

t := [2.894553047, 1.788886629]
> normt2:=evalf(DotProduct(t,t)/2);
normt2 := 5.789276355
> H:=evalf(ID-MatrixMatrixMultiply(t, Transpose(t)/normt2));
> #evaluating Householder
> reflection

$$H := \begin{bmatrix} -.447233960 & -.8944170093 \\ -.8944170093 & .4472339590 \end{bmatrix}$$

> HJT:=MatrixMatrixMultiply(H, Transpose(J));

$$HJT := \begin{bmatrix} -2.000058823 \\ -.3 \cdot 10^{-9} \end{bmatrix}$$

> HJT[2,1]:=0; print(HJT);

$$HJT_{2,1} := 0$$


$$\begin{bmatrix} -2.000058823 \\ 0 \end{bmatrix}$$

> w:= Vector([-h/HJT[1,1], 0]); #w = Hv
w := [.00002941013500, 0]
> v:=MatrixMatrixMultiply(H, w); #v^4
v := [-.00001315321114, -.00002630492499]
> L:= MatrixMatrixMultiply(H, E);
> BL:=MatrixMatrixMultiply(MatrixMatrixMultiply(V, Phi), L);
> #L_4 & (BL)_4

$$L := \begin{bmatrix} -.8944170093 \\ .4472339590 \end{bmatrix}$$


$$BL := \begin{bmatrix} -3.279562973 \\ .4944357453 \end{bmatrix}$$

> p:=Add(phi, MatrixMatrixMultiply(Phi, v)); #auxiliary variable
p := [.4472339593, .8944170095]
> r:= MatrixMatrixMultiply(-V, p);
> #RHS of overdetermined system to compute u in ODA
r := [-.149145866, -.9888152091]
> u:= LeastSquares(BL, r); #u^4
u := [.00002069770909]
> Deltax:= Add(v, MatrixMatrixMultiply(L, u)); #Deltax^4
Deltax := [-.00003166559420, -.00001704820661]

```

Fourth iteration is complete. Update  $\mathbf{x}$ :

```
> x:= Add(x, Deltax); x4:= map(evalf,x); #x^4
      x := [.4472154469, .8944262662]
      x4 := [.4472154469, .8944262662]
> f:= obj(x); #f_4
      f := .50000000006
```

Given the norm of Deltax, we declare convergence here, and plot the iteration history in  $x[1]$ - $x[2]$  plane:

```
> o0:= map(evalf,x0); o1:= map(evalf,x1);
> o2:= map(evalf,x2); o3:= map(evalf,x3); o4:= map(evalf,x4);
      o0 := [2, 2]
      o1 := [.787500000, 1.462500000]
      o2 := [.5090179236, 1.011062998]
      o3 := [.4508808673, .9011974282]
      o4 := [.4472154469, .8944262662]

> p0:=point(convert(o0,list), symbol=circle,
> color=blue);
> p1:=point(convert(o1,list), symbol=circle, color=blue);
> p2:=point(convert(o2,list), symbol=circle, color=blue);
> p3:=point(convert(o3,list), symbol=circle, color=blue);
> p4:=point(convert(o4,list), symbol=circle, color=blue);
p0 := POINTS([2., 2.], COLOUR(RGB, 0., 0., 1.00000000), SYMBOL(CIRCLE))

p1 := POINTS([.787500000, 1.462500000], COLOUR(RGB, 0., 0., 1.00000000),
SYMBOL(CIRCLE))
p2 := POINTS([.5090179236, 1.011062998], COLOUR(RGB, 0., 0., 1.00000000),
SYMBOL(CIRCLE))
p3 := POINTS([.4508808673, .9011974282], COLOUR(RGB, 0., 0., 1.00000000),
SYMBOL(CIRCLE))
p4 := POINTS([.4472154469, .8944262662], COLOUR(RGB, 0., 0., 1.00000000),
SYMBOL(CIRCLE))
```

```

> l1 :=
> arrow(convert(o0,list),convert(o1,list), 10.0, 0.1, .1, arrow,
> color=red, thickness=2): l2 :=
> arrow(convert(o1,list),convert(o2,list), 6.0, 0.1, .2, arrow,
> color=green, thickness=2): l3 :=
> arrow(convert(o2,list),convert(o3,list), 6.0, 0.1, .7, arrow,
> color=red, thickness=2): l4 :=
> arrow(convert(o3,list),convert(o4,list), 6.0, 0.1, 6.0, arrow,
> color=green, thickness=2): c1 := arc([0,0],
> 1,-Pi/6..4*Pi/6,color=black, thickness=2): obj_plot:= proc(ax,ay)
> (1/2)*(9*ax^2 - 8*ax*ay + 3*ay^2)
> end:f1:=implicitplot(obj_plot-3,-0.5..2,
> -0.5..2.5,numpoints=3000,
> linestyle=4,color=blue):
> f2:=implicitplot(obj_plot-2,-0.5..2,-0.5..2,
> numpoints=3000,linestyle=4,color=blue):
> f3:=implicitplot(obj_plot-1,-0.5..2,-0.5..2,numpoints=6000,
> linestyle=4,color=blue):f4:=implicitplot(obj_plot-0.5,-0.5..2,-0.5..2,
> numpoints=6000, linestyle=4,color=blue):

> display({c1,p0,p1, p2, p3, p4, p1, l1, l2,
> l3, l4,f1,f2,f3,f4},
> insequence = false, color=red, scaling=constrained);

```

The plots produced by the plotting commands in the Maple worksheet are reproduced in Fig. 6.1.

A plot of the contours of the objective function and the constraint, showing all four stationary points, is displayed in Fig. 6.2.

### 6.3.1 A Geometric Interpretation of the ODA

The ODA admits a geometric interpretation: the objective function can be regarded as a surface  $\mathcal{F}$  embedded in a  $(n+1)$ -dimensional space  $\mathbb{R}^{n+1}$ , namely, the union of  $\mathbb{R}^n$ , the space of the design-variable vector  $\mathbf{x}$ , and  $\mathbb{R}$ , the space of  $f(\mathbf{x})$ . Likewise, the set of constraint functions  $h_i(\mathbf{x})$ , for  $i = 1, 2, \dots, l$ , can be regarded as a set of  $l$  surfaces  $\mathcal{H}_i$  embedded in the same  $(n+1)$ -dimensional space.

The gradient  $\nabla f$  thus becomes the projection onto  $\mathbb{R}^n$  of a vector normal to  $\mathcal{F}$ , while the  $i$ th row of the Jacobian  $\mathbf{J}$  represents the projection onto  $\mathbb{R}^n$  of a normal  $\nabla h_i$  to  $\mathcal{H}_i$  at a given point  $Q_k$ , projected onto  $\mathbb{R}^n$  as  $P_k$ , of position vector  $\mathbf{x}^k$ . Furthermore, all  $n-l$  unit vectors  $\mathbf{l}_i$  denoting the  $n$ -dimensional columns of the

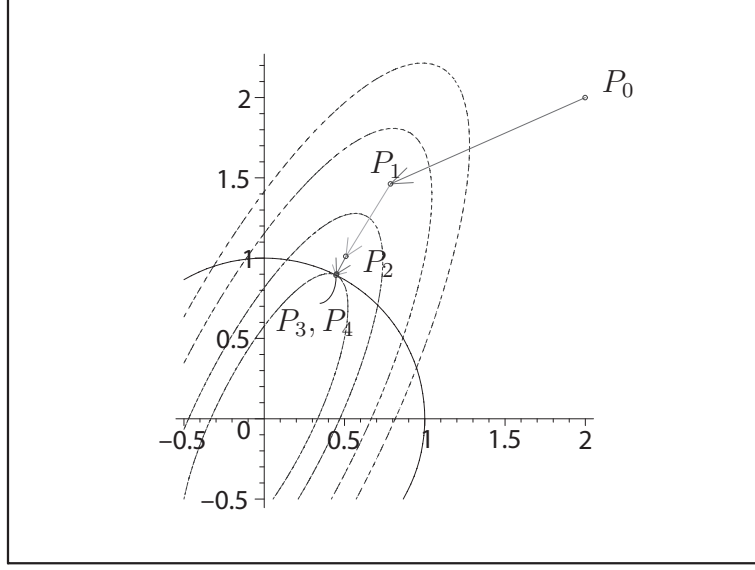


Figure 6.1: The four iterations leading to the solution of the linear least-square problem subject to one quadratic constraint

orthogonal complement  $\mathbf{L}_k$  of  $\mathbf{J}_k$  are normal to all  $l$  vectors  $\nabla h_i$ . Vectors  $\mathbf{l}_i$  thus span a hyperplane, of dimension  $n - l + 1$ , of  $\mathbb{R}^{n+1}$  that is tangent to all  $\mathcal{H}_i$  surfaces at  $Q_k$ . The intersection of this hyperplane with  $\mathbb{R}^n$  is thus a  $(n - l)$ -dimensional space  $\mathcal{L}$  on which vector  $\Delta \mathbf{u}^k$ , of the orthogonal decomposition of  $\mathbf{x}^k$ , lies. Matrix  $\mathbf{L}_k$  then maps this vector onto  $\mathbb{R}^n$ .

Illustrated in Fig. 6.3 is the paraboloid  $\mathcal{H}$  defined by  $h(\mathbf{x}) = x_1^2 + x_2^2 - 1$  of Example 6.3.1, with point  $Q_0$  indicating the location on the paraboloid of the initial guess  $P_0$  in  $\mathbb{R}^2$ , of position vector  $\mathbf{x}^0$  in this space, which is, the  $x_1$ - $x_2$  plane. This is a paraboloid of revolution, of axis of symmetry passing through the origin of the  $x_1$ - $x_2$  plane and normal to this plane; the intersection of  $\mathcal{H}$  with the  $x_1$ - $x_2$  plane is the circle  $x_1^2 + x_2^2 - 1 = 0$ . Moreover, the intersection of the plane  $\mathcal{T}$  tangent to  $\mathcal{H}$  at  $Q_0$  with the  $x_1$ - $x_2$  plane is the line  $\mathcal{L}$  containing vector  $\Delta \mathbf{u}^0$ . In the same figure, vector  $\Delta \mathbf{v}^0$  is the minimum-norm solution to the corresponding underdetermined system (6.23), line  $\mathcal{L}$  thus representing the linearized version of  $h(\mathbf{x}) = 0$ . In this figure,  $\Delta \mathbf{v}^0$  is normal to  $\mathcal{L}$ , point  $P_0$  becoming the origin of the vector space  $\mathbb{R}^2$  in which  $\Delta \mathbf{x}^0$  lies.

Furthermore, the paraboloid  $\mathcal{F}$  given by  $f(\mathbf{x}) = (1/2)(9x_1^2 - 8x_1x_2 + 3x_2^2)$  is shown in Fig. 6.4; contrary to  $\mathcal{H}$ ,  $\mathcal{F}$  is not of revolution, its cross sections, parallel to the  $x_1$ - $x_2$  plane, being ellipses of equal eccentricities, i.e., of equal axis-length ratios. Point  $P_1$ , of position vector  $\mathbf{x}^1 = \mathbf{x}^0 + \Delta \mathbf{x}^0$ , is found as the minimum of

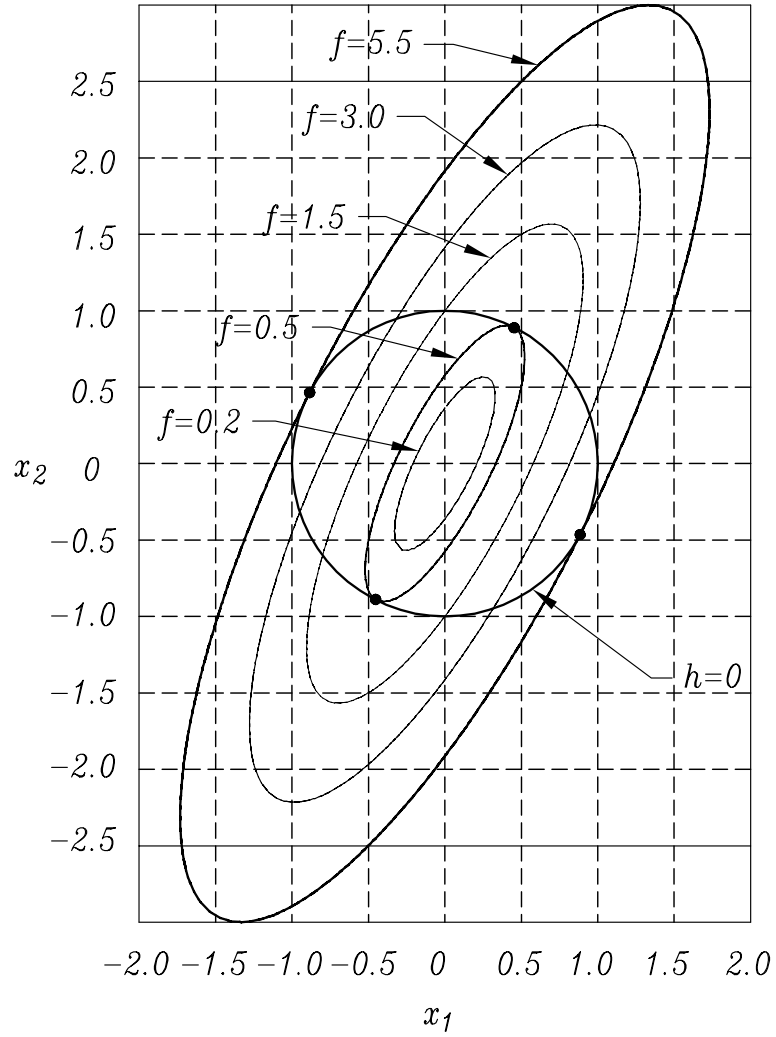


Figure 6.2: The contours of constant  $f$  and the constraint  $h = 0$

$f(\mathbf{x})$  along  $\mathcal{L}$ . This minimum is located at the projection  $P_1$  of point  $Q_1$  of the parabola defined by the intersection of a plane normal to  $x_1$ - $x_2$  passing through  $\mathcal{L}$  with  $\mathcal{F}$ . The one-dimensional vector  $\Delta \mathbf{u}^0$ , directed from  $P'_0$  to  $P_1$ , is parallel to  $\mathcal{L}$ . This vector is mapped into a two-dimensional vector of the  $x_1$ - $x_2$  plane by matrix  $\mathbf{L}_0$ .

Details of the layout of the foregoing points and vectors in the  $x_1$ - $x_2$  plane are displayed in Figs. 6.5 and 6.6.

**Example 6.3.2 (Finding the Eigenvalues and Eigenvectors of a Symmetric Matrix)** *The problem of finding the eigenvalues and corresponding eigenvectors of a  $n \times n$  symmetric positive-definite matrix  $\mathbf{M}$  is solved as a linear least-square*



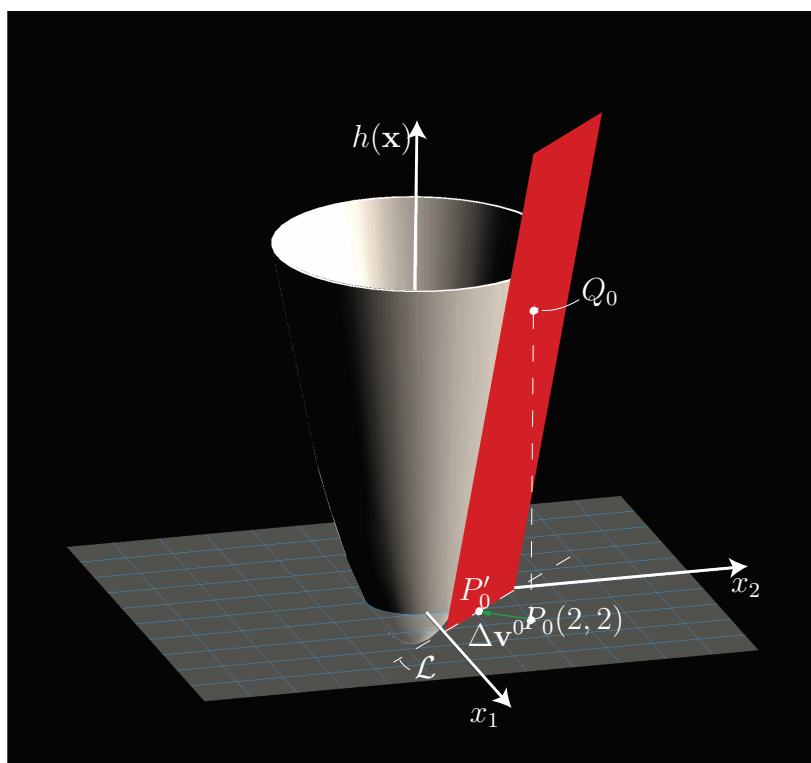


Figure 6.3: The paraboloid  $\mathcal{H}$ , its tangent plane  $\mathcal{T}$  at  $Q_0$ , whose projection onto the  $x_1$ - $x_2$  plane is  $P_0$ , the initial guess, and the intersection  $\mathcal{L}$  of  $\mathcal{T}$  with the  $x_1$ - $x_2$  plane

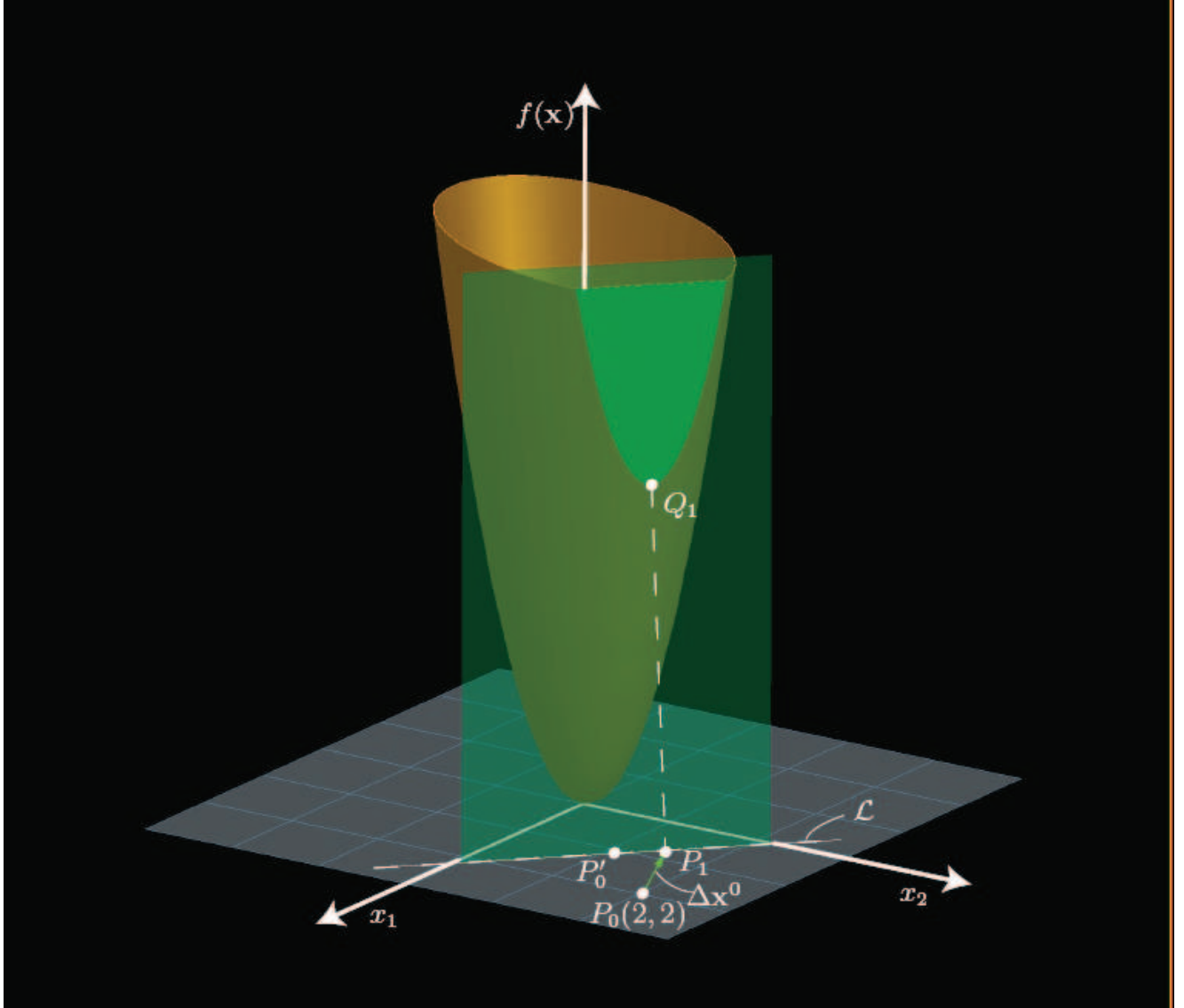
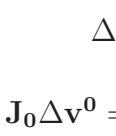
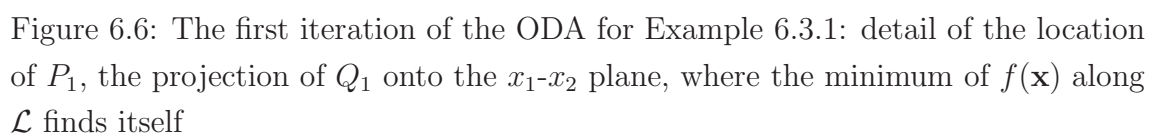


Figure 6.4: The paraboloid  $\mathcal{F}$ , its contours, and its intersection with the plane  $\Pi$  normal to the  $x_1$ - $x_2$  plane, the minimum of the intersection curve denoted by  $Q_1$ , its projection onto the  $x_1$ - $x_2$  plane being  $P_1$



187



problem subject to quadratic constraints: for  $i = 1, 2, \dots, n$ , and  $k = 1, 2, \dots, i$ , find  $\lambda_i$  and  $\mathbf{x}_i$  such that

$$\lambda_i = \min_{\mathbf{x}_i} \frac{1}{2} \mathbf{x}_i^T \mathbf{M} \mathbf{x}_i$$

subject to

$$\mathbf{x}_k^T \mathbf{x}_i = \begin{cases} 0, & \text{if } k = 1, 2, \dots, i-1; \\ 1, & \text{if } k = i \end{cases}$$

where  $\lambda_i$  is the  $i$ th eigenvalue of matrix  $\mathbf{M}$  and  $\mathbf{x}_i$  is the corresponding eigenvector. In order to use the ODA package to solve the problem, we define, for  $i = 1, 2, \dots, n$ :

$$q = n \quad \text{and} \quad l = i, \quad (6.37)$$

$$\mathbf{x} = \mathbf{x}_i,$$

$$\phi(\mathbf{x}) = \mathbf{x},$$

$$\mathbf{h}(\mathbf{x}) = [\mathbf{x}_1^T \mathbf{x} \quad \dots \quad \mathbf{x}_{i-1}^T \mathbf{x} \quad \mathbf{x}^T \mathbf{x} - 1]^T,$$

$$\mathbf{W} = \mathbf{M}$$

where  $\mathbf{x}_k$ , for  $k = 1, 2, \dots, i-1$ , are the previously calculated eigenvectors of  $\mathbf{M}$ , and hence, are known. With the above definitions, for  $i = 1, 2, \dots, n$ , subroutine LSSCNL of the ODA package is called  $n$  times. After each call, one eigenvalue and its corresponding eigenvector are obtained. Notice that, in the last call, the number of constraints is equal to the number of variables, namely,  $l = n$ . Matrix  $\mathbf{M}$  is given as

$$\mathbf{M} = \begin{bmatrix} 4 & 2 & 1 & 1 & 1 \\ 2 & 4 & 2 & 1 & 1 \\ 1 & 2 & 4 & 2 & 1 \\ 1 & 1 & 2 & 4 & 2 \\ 1 & 1 & 1 & 2 & 4 \end{bmatrix}.$$

We use the initial guess

$$\mathbf{x}^0 = [0.1 \quad 0.1 \quad 0.1 \quad 0.1 \quad 0.1]^T.$$

The eigenvalues and the eigenvectors computed with the ODA package are listed in Table 6.1. In that table, the number of iterations that the package took till convergence was reached with  $\epsilon_1 = 0.0001$  and  $\epsilon_2 = 0.0001$ , is indicated.

**Example 6.3.3 (The Constrained Minimization of the Rosenbrock Function)** In this example, we find its equality-constrained minimum of the Rosenbrock function using SQP via the ODA. We thus have

$$f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 \quad \rightarrow \quad \min_{x_1, x_2}$$

$i$	1	2	3	4	5
$\lambda_i$	1.27738	3.08749	9.63513	2.0	4.0
1st comp. of $\mathbf{x}_i$	0.26565	-0.51369	0.40689	0.5	0.5
2nd comp. of $\mathbf{x}_i$	-0.51853	0.10368	0.46944	-0.5	0.5
3rd comp. of $\mathbf{x}_i$	0.56667	0.67138	0.47764	0.0	0.0
4th comp. of $\mathbf{x}_i$	-0.51853	0.10368	0.46944	0.5	-0.5
5th comp. of $\mathbf{x}_i$	0.26565	-0.51369	0.40689	-0.5	-0.5
# of iterations	12	11	6	38	54

Table 6.1: Eigenvalues and eigenvectors of  $\mathbf{M}$

subject to

$$h((x_1, x_2)) = 0.7525x_1^2 - 1.1202x_1 - 0.8574x_1x_2 \\ + 0.6168x_2 + 0.2575x_2^2 + 0.4053$$

The function is notorious for its ill-conditioning, which is apparent from its contours, as shown in Fig. 6.7, showing elongated valleys. The outcome is that the quadratic approximation of this function within those valleys is a family of ellipses that have one semiaxis much greater than the other one, thereby leading to ill-conditioning. Notice that the constraint is a rather elongated ellipse that contributes to the ill-conditioning of the problem.

Starting from the initial guess  $\mathbf{x} = [1.5 \ 1.5]^T$  with a damping ratio of 0.025, the optimum solution is found in 312 ODA iterations, the result being

$$\mathbf{x}_{opt} = \begin{bmatrix} 0.9176 \\ 0.5873 \end{bmatrix}$$

which yields

$$f_{\min} = 6.6963$$

## 6.4 Equality-Constrained Optimization with Arbitrary Objective Function

The problem to be solved is defined as:

$$f = f(\mathbf{x}) \quad \rightarrow \quad \min_{\mathbf{x}} \quad (6.38)$$

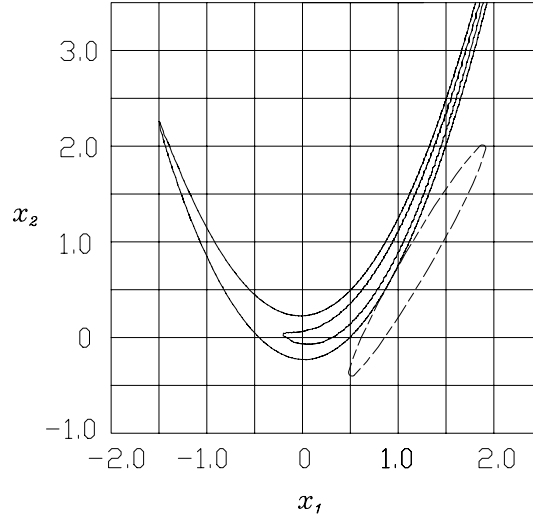


Figure 6.7: The contours of the Rosenbrock (banana) function and its quadratic constraint (dashed)

subject to the nonlinear equality constraints

$$\mathbf{h}(\mathbf{x}) = \mathbf{0} \quad (6.39)$$

where  $\mathbf{x}$  is the  $n$ -dimensional design-variable vector, the objective function  $f(\mathbf{x})$  being a nonlinear function of  $\mathbf{x}$ , not necessarily quadratic and positive-definite, which was the case studied in Section 6.3. Moreover,  $\mathbf{h}(\mathbf{x})$  is a  $l$ -dimensional vector of nonlinear equality constraints.

In the problem defined in eq.(6.38), if the constraints in eq.(6.39) are *analytic*, then there exists a *feasible manifold*  $\mathcal{F} \subset \mathbb{R}^n$ , of dimension  $l$ , such that, if  $\mathbf{u} \in \mathcal{F}$ , then

$$\mathbf{h}(\mathbf{x}(\mathbf{u})) = \mathbf{0} \quad (6.40)$$

In the particular case in which  $\mathbf{h}(\mathbf{x}(\mathbf{u}))$  is linear,  $\mathcal{F}$  is a vector space, i.e., the *feasible space* of the problem at hand.

An example of a problem that falls in the category studied here is the objective function

$$f(\mathbf{x}) = \frac{1}{2}(-9x_1^2 - 8x_1x_2 + 3x_2^2) \quad (6.41a)$$

to be minimized subject to

$$h(\mathbf{x}) = x_1^2 + x_2^2 - 1 = 0 \quad (6.41b)$$

It can be readily shown that this problem does not belong to the least-square class, and hence, cannot be handled with the ODA directly. Indeed, as the objective function is quadratic in the design variables, for the problem to be of the least-square class, it should be possible to cast the objective function in the form (6.2). This means that the Hessian of this function should admit the form  $\mathbf{A}^T \mathbf{W} \mathbf{A}$ , which means that the unconstrained Hessian should be positive-definite, as  $\mathbf{W}$  is so by hypothesis. A quick computation shows that the Hessian in the case at hand is

$$\mathbf{H} = \begin{bmatrix} -9 & -4 \\ -4 & 3 \end{bmatrix}$$

Further,  $\text{tr}(\mathbf{H}) = -6$ , which means that its larger eigenvalue is negative, while  $\det(\mathbf{H}) = -43$ , and hence, its two eigenvalues bear opposite signs, the outcome being that the Hessian is sign-indefinite. Therefore, no matrices  $\mathbf{A}$  and  $\mathbf{W}$  can be found that would allow the casting of the given objective function in the form of eq.(6.2). The numerical solution of this problem thus cannot be found by application of the method of Section 6.2, or that of Section 6.3 for that matter. Nevertheless, ODA can still be applied, if with some additional work, as described below.

Two cases are studied:

1. The objective and the constraint functions are assumed to be continuous and to have *continuous derivatives up to the second order*; and
2. the objective and the constraint functions are assumed to be continuous and to have *continuous derivatives up to the first order*.

In both cases it is assumed that the derivatives in question are available, which is a big assumption. In real-life problems, second-order derivatives are seldom available; first-order derivatives are more likely to be available. If the latter are not, then these can be approximated, provided that they are continuous, by finite differences, for which reliable algorithms exist.

The first case is included here not only for pedagogical reasons but also because many problems in design can be formulated with  $C^2$ -continuous functions. These arise from conservation laws in mechanics and electromagnetism.

### 6.4.1 A Sequential-Quadratic Programming Approach

In this subsection the objective and the constraint functions are assumed to be continuous and to have continuous derivatives up to the second order, while the



Hessian  $\mathbf{H} \equiv \nabla \nabla f$  of  $f(\mathbf{x})$  with respect to  $\mathbf{x}$  is not assumed positive definite. Furthermore, we assume that, at  $\mathbf{x} = \mathbf{x}^k$ , in general,  $\mathbf{h}(\mathbf{x}^k) = \mathbf{h}^k \neq \mathbf{0}_l$ , i.e., the current  $\mathbf{x}$  is not feasible, and note that  $f(\mathbf{x}^k + \Delta \mathbf{x}^k)$  can be expanded, to a second order, as

$$f(\mathbf{x}^k + \Delta \mathbf{x}^k) \approx f(\mathbf{x}^k) + (\nabla f)_k^T \Delta \mathbf{x}^k + \frac{1}{2} (\Delta \mathbf{x}^k)^T (\nabla \nabla f)_k \Delta \mathbf{x}^k \rightarrow \min_{\Delta \mathbf{x}^k} \quad (6.42a)$$

To find the increment  $\Delta \mathbf{x}^k$ , we can still resort to the ODA, as introduced in Section 6.2. To this end, we decompose the foregoing vector into its two orthogonal components, as per eq.(6.21)

$$\Delta \mathbf{x}^k = \Delta \mathbf{v}^k + \mathbf{L}_k \Delta \mathbf{u}^k$$

where  $\mathbf{L}_k$  is the isotropic orthogonal complement of  $\mathbf{J}_k$ , found to be the  $\mathbf{Q}_R$  block of the QR-decomposition of  $\mathbf{J}_k^T$ , while  $\mathbf{J}_k$  itself is defined as the gradient of  $\mathbf{h}$  with respect to  $\mathbf{x}$ , evaluated at  $\mathbf{x} = \mathbf{x}^k$ . Moreover,  $\mathbf{L}_k$  and  $\Delta \mathbf{u}^k$  are found with the procedure described in Section 6.2 for linearly constrained linear least-square problems. Under these conditions, the  $\Delta \mathbf{v}^k$  component of  $\Delta \mathbf{x}^k$  in eq.(6.42a) is subject to

$$\mathbf{J}_k \Delta \mathbf{v}^k = -\mathbf{h}^k \quad (6.42b)$$

We have thus formulated a quadratic optimization problem in  $\Delta \mathbf{x}^k$ , eq.(6.42a), subject to the linear constraints (6.42b). This problem, however, need not belong to the least-square class. It does if and only if the Hessian matrix  $(\nabla \nabla f)_k$  is positive-definite. This, however, need not be the case, as stated at the outset. Moreover,  $\mathbf{J}_k^\dagger$  and  $\mathbf{L}_k$  are now substituted by  $\mathbf{Q}_L \mathbf{U}^{-T}$  and  $\mathbf{Q}_R$ , respectively, of the QR-decomposition of  $\mathbf{J}_k^T$ , as introduced in eq.(5.53) for  $\mathbf{C}^T$ . Then, the minimum-norm solution of eq.(6.42b) is given by

$$\Delta \mathbf{v}^k = -\mathbf{J}_k^\dagger \mathbf{h}^k \quad (6.43)$$

where  $\mathbf{J}_k^\dagger$  is computed using the QR-decomposition of  $\mathbf{J}_k^T$ .

Furthermore, with  $\Delta \mathbf{v}^k$  given by eq.(6.43),  $f(\mathbf{x}^k + \Delta \mathbf{x}^k)$  becomes a function solely of  $\Delta \mathbf{u}^k$ , i.e.,

$$\begin{aligned} f(\Delta \mathbf{u}^k) \approx \tilde{f}(\Delta \mathbf{u}^k) \equiv & f(\mathbf{x}^k) + (\nabla f)_k^T (\Delta \mathbf{v}^k + \mathbf{L}_k \Delta \mathbf{u}^k) \\ & + \frac{1}{2} (\Delta \mathbf{v}^k + \mathbf{L}_k \Delta \mathbf{u}^k)^T (\nabla \nabla f)_k (\Delta \mathbf{v}^k + \mathbf{L}_k \Delta \mathbf{u}^k) \rightarrow \min_{\Delta \mathbf{u}^k} \end{aligned}$$

which can be cast in the form

$$\begin{aligned}\tilde{f}(\Delta \mathbf{u}^k) = & \frac{1}{2}(\Delta \mathbf{u}^k)^T \mathbf{L}_k^T (\nabla \nabla f)_k \mathbf{L}_k \Delta \mathbf{u}^k + [\mathbf{L}_k^T (\nabla \nabla f)_k \Delta \mathbf{v}^k + \mathbf{L}_k^T (\nabla f)_k]^T \Delta \mathbf{u}^k \\ & + \frac{1}{2}(\Delta \mathbf{v}^k)^T (\nabla \nabla f)_k \Delta \mathbf{v}^k + (\nabla f)_k^T \Delta \mathbf{v}^k + f(\mathbf{x}^k) \rightarrow \min_{\Delta \mathbf{u}^k}\end{aligned}\quad (6.44)$$

subject to *no constraints*,  $\tilde{f}(\Delta \mathbf{u}^k)$  being *quadratic* in  $\Delta \mathbf{u}^k$ . Function  $\tilde{f}(\Delta \mathbf{u}^k)$  has a minimum if its Hessian with respect to  $\Delta \mathbf{u}^k$ , what is termed the *reduced Hessian*  $\mathbf{H}_{uk} = \mathbf{L}_k^T (\nabla \nabla f)_k \mathbf{L}_k$ , is positive-definite. Under the assumption that  $\mathbf{H}$  is not necessarily positive definite,  $\mathbf{H}_{uk}$  can still be positive-definite. In this case, a minimum  $\Delta \mathbf{u}^k$  of  $\tilde{f}(\Delta \mathbf{u}^k)$  can be readily computed upon zeroing its gradient with respect to  $\Delta \mathbf{u}^k$ , which yields

$$\mathbf{H}_{uk} \Delta \mathbf{u}^k = -\mathbf{L}_k^T [(\nabla \nabla f)_k \Delta \mathbf{v}^k + (\nabla f)_k] \quad (6.45)$$

Under the foregoing assumption,  $\mathbf{H}_{uk}$  is invertible, and hence,

$$\Delta \mathbf{u}^k = -\mathbf{H}_{uk}^{-1} \mathbf{L}_k^T [(\nabla \nabla f)_k \Delta \mathbf{v}^k + (\nabla f)_k] \quad (6.46)$$

We have thus reduced the original problem to a *sequence of linear-quadratic programs*, within an iterative procedure. At each iteration, moreover, we find the correction to the current approximation  $\Delta \mathbf{x}^k$  by means of a combination of two linear problems, one being a minimum-norm problem, the other involving a determined linear system of equations. As the objective function at each iteration is quadratic in  $\Delta \mathbf{u}$ , and its minimization is unconstrained, the above procedure is called *sequential quadratic programming*.

The foregoing procedure relies on the rather daring assumption that the reduced Hessian  $\mathbf{H}_{uk}$  at each iteration is positive-definite. Below we study the more realistic case of a non-positive-definite Hessian.

### Sequential quadratic programming with Hessian stabilization

In the presence of a non-positive-definite Hessian  $\mathbf{H}$ , we aim at a *perturbation*  $\Delta \mathbf{H}$  of the Hessian that will render the *perturbed Hessian*  $\tilde{\mathbf{H}}$  positive-definite, thus producing

$$\tilde{\mathbf{H}} \equiv \mathbf{H} + \Delta \mathbf{H} \quad (6.47)$$

The rationale behind forcing the Hessian to be positive-definite lies in that, if a  $C^2$ -continuous function has a minimum, then in a region around a local minimum—or a local maximum for that matter—then the function behaves as a quadratic function

associated with a positive-definite—correspondingly, as a negative-definite—matrix, which is the Hessian of the function. In the case of an objective function defined over  $\mathbb{R}^2$ , the minimum (maximum) appears in a valley (summit) whose level curves are quite close to ellipses, as illustrated in Fig. 4.4. This figure shows the contours of equal distances from  $Z_1$  to point  $C$  of the manipulator of Fig. 5.2, with its first joint locked and the other two sweeping angles  $\theta_2$  and  $\theta_3$  that vary from 0 to  $2\pi$ .

How to obtain  $\Delta\mathbf{H}$  that is guaranteed to produce a positive-definite-Hessian is the key issue here. We describe in the subsection below a method for the determination of  $\Delta\mathbf{H}$ . Note that, once the perturbed Hessian, which is most frequently referred to as the *stabilized Hessian*, is available, the reduced Hessian is necessarily positive-definite, and given by

$$\tilde{\mathbf{H}}_u = \mathbf{L}^T(\mathbf{H} + \Delta\mathbf{H})\mathbf{L} \quad (6.48)$$

Hence,  $\Delta\mathbf{u}^k$  is found from eq.(6.45), which is rewritten below with  $\mathbf{H}_{uk}$  replaced by  $\tilde{\mathbf{H}}_{uk}$ , its stabilized counterpart:

$$\tilde{\mathbf{H}}_{uk}\Delta\mathbf{u}^k = -\mathbf{L}_k^T[(\nabla\nabla f)_k\Delta\mathbf{v}^k + (\nabla f)_k] \quad (6.49)$$

The process of finding a positive-definite  $\tilde{\mathbf{H}}$  is termed *Hessian stabilization*. The idea behind the process lies in the property that, if the eigenvalues of a  $n \times n$  matrix  $\mathbf{M}$  are  $\{\mu_k\}_1^n$ , then the eigenvalues of matrix  $\mathbf{M} + \alpha\mathbf{1}$ , where  $\alpha$  is a real number and  $\mathbf{1}$  is the  $n \times n$  identity matrix, are  $\{\mu_k + \alpha\}_1^n$ . Thus, the effect of adding the isotropic<sup>9</sup> matrix  $\alpha\mathbf{1}$  to  $\mathbf{M}$  is to shift the eigenvalues of the latter to the right of the complex plane by an amount  $\alpha$  if  $\alpha > 0$ ; if  $\alpha < 0$ , then the shift occurs to the left of the complex plane by an amount  $|\alpha|$ . If the Hessian of interest is not positive-definite, this means that it has some negative eigenvalues, in which case Hessian stabilization consists in finding the right value of  $\alpha$  in the foregoing scheme, that will shift the Hessian eigenvalues to the right of the real axis—since the Hessian is necessarily symmetric, its eigenvalues are all real—so that none of the shifted eigenvalues will lie on the left half of the real axis. Notice that, if  $\alpha > 0$  is underestimated, then the associated isotropic matrix will fail to shift some of the negative Hessian eigenvalues to the right; if overestimated, then all shifted eigenvalues will lie on the right half of the real axis, but the Hessian will be overly perturbed. In this case, either convergence will slow down or divergence will ensue.

Obviously, if we know the eigenvalues of the Hessian  $\mathbf{H}$  at the  $k$ th iteration, then we can find the right  $\mu_k$  that will shift all its eigenvalues to the right. However,

---

<sup>9</sup>See the paragraph below eq.(6.15) for a definition of this concept.

computing eigenvalues is an iterative process, except for special cases of simple matrices, and hence, we cannot rely on knowledge of those eigenvalues. We discuss below how to estimate the right amount of shift  $\alpha$  in the absence of knowledge of the Hessian eigenvalues. The basis of the procedure is a result on positive-definite matrices that we recall below.

### Diagonal-dominance in Positive-definite Matrices

The *Gerschgorin Theorem* (Varga, 2000) establishes a region in the complex plane containing all the eigenvalues of a  $n \times n$  matrix  $\mathbf{A}$ , defined over the complex field  $\mathbb{C}$ , namely,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

According to the Gerschgorin Theorem, all the eigenvalues of  $\mathbf{A}$  lie within a region  $\mathcal{S}$  of the complex plane defined as the union of  $n$  disks  $\mathcal{D}_i$  centered at  $a_{ii}$ , with radius  $r_i$ , in the complex plane, for  $i = 1, \dots, n$ ,  $r_i$  being given by

$$r_i = \sum_{j=1, j \neq i}^n |a_{ij}|$$

in which  $|\cdot|$  denotes the module of  $(\cdot)$ . The Gerschgorin Theorem is illustrated in Fig. 6.8, the region  $\mathcal{S}$  thus being

$$\mathcal{S} = \bigcup_{i=1}^n \mathcal{D}_i$$

If  $\mathbf{A}$  is symmetric and real, which is so for Hessian matrices, then its eigenvalues lie in the union of the real intervals

$$I_i = [a_{ii} - r_i, a_{ii} + r_i], \quad i = 1, 2, \dots, n$$

A lower bound  $l$  of the set  $\{\lambda_i\}_1^n$  of eigenvalues of  $\mathbf{A}$  is, thus,

$$l \equiv \min_i \{a_{ii} - r_i\}_1^n \quad (6.50a)$$

the corresponding upper bound being

$$u \equiv \max_i \{a_{ii} + r_i\}_1^n \quad (6.50b)$$

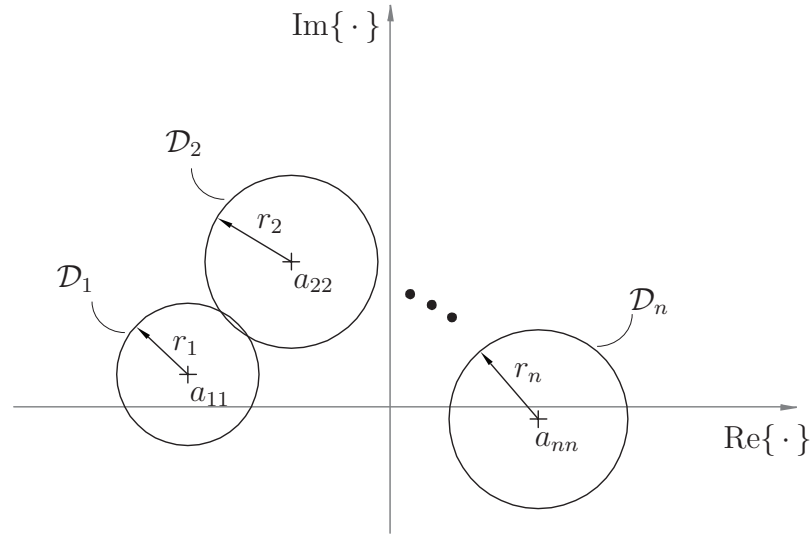


Figure 6.8: The Gerschgorin disks of an arbitrary  $n \times n$  matrix

If  $\mathbf{A}$  is positive-definite, all eigenvalues of  $\mathbf{A}$  must be positive, which means that the lower bound  $l$  should be positive as well. If, on the other hand,  $\mathbf{A}$  is either sign-indefinite or positive-definite, but close to singular, then  $l$  can be negative.

Now we have, with the foregoing notation,

**Definition 6.4.1 (Diagonal Dominance)** A  $n \times n$  matrix  $\mathbf{A}$  is said to be *diagonally dominant* if

$$a_{ii} > r_i$$

Further, we have a result allowing us to characterize positive-definite matrices without the burden of computing their eigenvalues.

**Theorem 6.4.1** *If a symmetric matrix  $\mathbf{A}$  is diagonally dominant, then it is positive-definite.*

Note, however, that the converse is not true, i.e., a positive-definite matrix need not be diagonally dominant. Consider the matrix

$$\mathbf{P} = \begin{bmatrix} 10 & 2.6 & -7.6 \\ 2.6 & 3.3 & -2.7 \\ -7.6 & -2.7 & 6.1 \end{bmatrix}$$

which it is not diagonally dominant, yet its eigenvalues are all positive: 0.080, 2.411, 16.91, with four digits.

## Hessian stabilization with the aid of diagonal-dominance

The reduced Hessian matrix  $\mathbf{L}_k^T(\nabla\nabla f)|_k\mathbf{L}_k$  of the objective function can fail to be positive-definite when the Hessian  $(\nabla\nabla f)|_k$  fails to be so. However, the latter may well fail to be positive-definite, while the former is positive-definite. In this light, apparently we need not stabilize the Hessian itself, but only its feasible projection. The question of the feasibility of this approach is cleared in the sequel.

We need first a criterion to tell when  $(\nabla\nabla f)_k$  is *suspected* of being sign-indefinite. The criterion is simple:

*If  $(\nabla\nabla f)_k$  fails to be diagonally dominant, then sign-indefiniteness is likely to occur and hence, Hessian stabilization is warranted.*

Notice that, although lack of diagonally dominance (DD) is not a guarantee of sign-indefiniteness, or of negative-definiteness for that matter, DD is a guarantee of positive-definiteness. For this reason, means to enforce DD are pursued below.

The stabilizing procedure is applied by introducing a scalar  $\mu_k > 0$ , at the  $k$ th iteration, that we will term the *Gerschgorin shift*, such that a new diagonally-dominant matrix  $\tilde{\mathbf{H}}_k$  is used to replace  $(\nabla\nabla f)_k$ , with  $\tilde{\mathbf{H}}_k$  defined as

$$\tilde{\mathbf{H}}_k = (\nabla\nabla f)_k + \mu_k \mathbf{1} \quad (6.51)$$

$\tilde{\mathbf{H}}_k$  thus being guaranteed to be positive-definite, based on the diagonal-dominance theorem above.

The stabilized Hessian thus yields the *stabilized reduced Hessian*

$$\tilde{\mathbf{H}}_{uk} = \mathbf{L}_k^T \tilde{\mathbf{H}}_k \mathbf{L}_k = \mathbf{L}_k^T (\nabla\nabla f)_k \mathbf{L}_k + \mu_k \mathbf{L}_k^T \mathbf{L}_k \quad (6.52)$$

Moreover, since  $\mathbf{L}_k$  is isotropic, for it has been chosen as the  $\mathbf{Q}_L$  block of the QR-decomposition of  $\mathbf{J}_k^T$ , it turns out that

$$\mathbf{L}_k^T \mathbf{L}_k = \mathbf{1}_{n'} \quad (6.53)$$

with  $\mathbf{1}_{n'}$  denoting the  $(n-l) \times (n-l)$  identity matrix. Therefore, the stabilized reduced Hessian  $\tilde{\mathbf{H}}_{uk}$  becomes

$$\tilde{\mathbf{H}}_{uk} = \mathbf{L}_k^T \tilde{\mathbf{H}}_k \mathbf{L}_k = \mathbf{L}_k^T (\nabla\nabla f)_k \mathbf{L}_k + \mu_k \mathbf{1}_{n'} \quad (6.54)$$

which is now *a fortiori* positive-definite, problem (6.44) thus admitting one minimum  $\Delta \mathbf{u}^k$ , which is computed from eq.(6.49). Notice the economy of computations brought about by an isotropic orthogonal complement  $\mathbf{L}_k$  in eq.(6.54).

### Choice of the Gerschgorin Shift

In this subsection we stress the importance of the selection of  $\mu_k$ , where subscript  $k$  denotes the iteration number. With a proper selection, the number of iterations can be effectively reduced. First, we assume that  $(\nabla\nabla f)_k$  was found to fail the diagonal-dominance test, and hence, the lower bound  $l_k$  of its eigenvalues is negative.

The selection of  $\mu_k$  is suggested to be *slightly* greater than the absolute value of the lower bound  $l_k$  of the eigenvalues, as obtained by the diagonal-dominance criterion (or test), i.e.,

$$\mu_k = -(1 + \Delta_k)l_k, \quad l_k < 0 \quad (6.55)$$

where  $\Delta_k$  is a positive number, that is to be chosen as small as possible.

The value of  $\Delta_k$  is related to the *bandwidth*  $b_k$  of the Hessian eigenvalues, with  $b_k$  defined as

$$b_k = u_k - l_k$$

The value of  $\Delta_k$  can then be chosen as a small fraction  $10^{-p}$ , with  $p = 1$  or  $2$ , probably even smaller.

**Example 6.4.1 (Powell's Function)** *A problem proposed by Powell (1969) is solved here:*

$$f(\mathbf{x}) = e^{x_1 x_2 x_3 x_4 x_5} \quad \rightarrow \quad \min_{\mathbf{x}}, \quad \mathbf{x} \equiv [x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5]^T$$

*subject to the nonlinear equality constraints*

$$h_1 = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 - 10 = 0$$

$$h_2 = x_2 x_3 + x_4 x_5 = 0$$

$$h_3 = x_1^3 + x_2^3 + 1 = 0$$

*Solution:* A word of caution is in order here: while the univariate exponential function  $e^x$  is convex—its second derivative with respect to  $x$  is positive everywhere—the bivariate exponential function  $e^{x_1 x_2}$  is not convex everywhere, and neither is so the above objective function. In fact, the Hessian of the bivariate exponential becomes sign-indefinite in a region of the  $x_1$ - $x_2$  plane. This statement is illustrated with the plot of this function, displayed in Fig. 6.9. The conclusion of the foregoing remark is, then, that the multivariable exponential function, like Powell's function, has a Hessian that is sign-indefinite in a region of  $\mathbb{R}^5$ . Optimum solutions were obtained with two different algorithms, the corresponding results being listed in Table 6.2.

Results were obtained under the same environment, a Silicon Graphics 64-bit Octane SE workstation, with a 250 MHz R10000 processor, running the IRIX 6.5 operating system<sup>10</sup>. An initial guess is taken as

$$\mathbf{x}_0 = [-1 \quad 2 \quad -0.5 \quad 1 \quad 2]^T$$

with tolerance of  $10^{-6}$ . The ODA package requires only 55 iterations, as compared with 186 required by the Matlab Optimization Toolbox. Moreover, the CPU time required by the ODA is only 8.9 % of the CPU time consumed by Matlab.

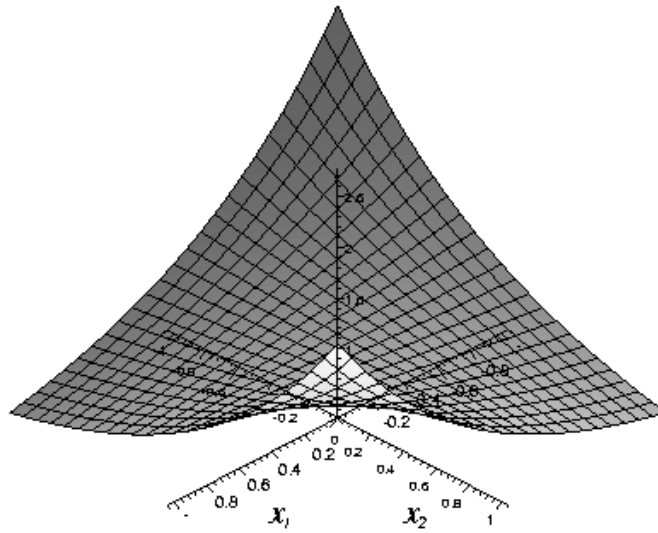


Figure 6.9: The bivariate exponential  $e^{x_1 x_2}$

**Example 6.4.2 (The Equilibrium Configuration of a  $N$ -link Chain)** *Shown in Fig. 6.10a is a chain with  $N$  links in its equilibrium configuration, which spans a distance  $d$ , with each link of length  $\ell$ . Knowing that the chain reaches its equilibrium configuration when its potential energy attains a minimum value, find the equilibrium configuration. This problem, originally proposed by Luenberger (1984), was solved for the case of two design variables, exactly, in Example 5.3.3.*

*Solution:* Angles  $\theta_i$ , used to define the configuration of the chain, are measured from the vertical ccw, with  $\theta_i$  corresponding to the angle that the axis of the  $i$ th link makes with the vertical, as shown in Fig. 6.11.

<sup>10</sup>Although the Octane SE was discontinued many years ago, the absolute CPU-time values of Table 6.4.1 are not relevant. What matters is that the ODA took less than 10% of the time of the Matlab `fmincon` function



Table 6.2: A performance comparison based on Powell's function

	Matlab	ODA
$f$	0.05395	0.05395
$x_1$	-1.7172	-1.7171
$x_2$	1.5957	1.5957
$x_3$	1.8272	-1.8272
$x_4$	0.7636	-0.7636
$x_5$	0.7636	0.7636
# of iterations	186	55
CPU time (s)	0.2903	0.0259

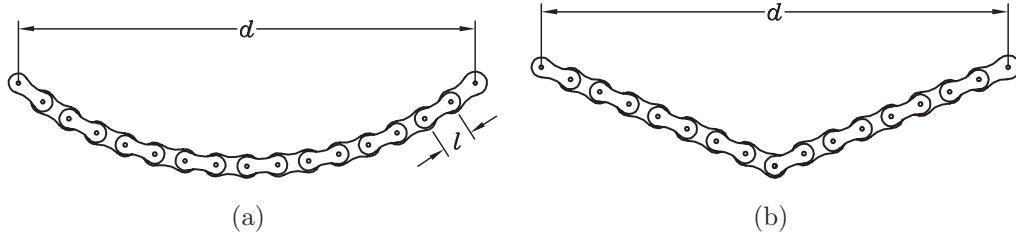


Figure 6.10: An  $N$ -link chain in: (a) its unknown equilibrium configuration; and (b) a configuration to be used as an initial guess

If  $V \equiv \mu \ell f(\theta_1, \theta_2, \dots, \theta_N)$  denotes the potential energy of the chain, and  $\mu$  is the mass density of the links per unit length, then minimizing  $V$  is equivalent to minimizing  $f$ , which is given by

$$f(\theta_1, \theta_2, \dots, \theta_N) = - \left[ \frac{1}{2} \cos \theta_1 + \left( \cos \theta_1 + \frac{1}{2} \cos \theta_2 \right) + \dots \right. \\ \left. + \left( \cos \theta_1 + \dots + \cos \theta_{N-1} + \frac{1}{2} \cos \theta_{N-1} \right) \right] \rightarrow \min_{\{\theta_i\}_1^N}$$

or, in compact form,

$$f(\theta_1, \theta_2, \dots, \theta_N) = -\frac{1}{2} \sum_{i=1}^N [2(N-i) + 1] \cos \theta_i \rightarrow \min_{\{\theta_i\}_1^N}$$

subject to two constraints: the two ends (1) must lie at the same level, and (2) are

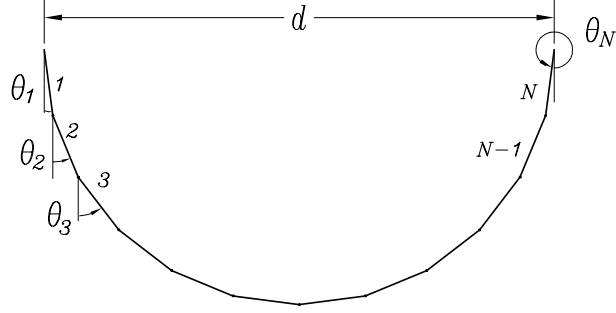


Figure 6.11: Definition of  $\theta_i$  for the  $N$ -link chain

separated by a distance  $d$ , as shown in Fig. 6.11. The constraints are

$$h_1 = \sum_{i=1}^N \cos \theta_i = 0$$

$$h_2 = \sum_{i=1}^N \sin \theta_i - \frac{d}{\ell} = 0$$

Under the assumption that the configuration is symmetric, and that  $N$  is even, then  $M = N/2$  is an integer. Thus, only one half of the chain need be considered. The problem is, thus, simplified as

$$f(\theta_1, \theta_2, \dots, \theta_M) = - \left[ \frac{1}{2} \cos \theta_1 + (\cos \theta_1 + \frac{1}{2} \cos \theta_2) + \dots + (\cos \theta_1 + \dots + \cos \theta_{M-1} + \frac{1}{2} \cos \theta_M) \right] = -\frac{1}{2} \sum_{i=1}^M [2(M-i) + 1] \cos \theta_i \rightarrow \min_{\{\theta_i\}_1^M}$$

The two constraints then reduce to only one:

$$h = \sum_{i=1}^M \sin \theta_i - \frac{d}{2\ell} = 0$$

This problem, with  $M = 5$ , i.e., with  $N = 10$ , is solved now using the configuration of Fig. 6.10b as an initial guess. The equilibrium configuration of the chain is given in Table 6.3 with a comparison between ODA and Matlab.

With the same tolerance set at 0.0001, the ODA takes less than half the number of iterations than Matlab; additionally, the CPU time consumed by ODA is about 10% of that consumed by Matlab.

It is noteworthy that the convergence of ODA is dependent on the choice of  $\mu_k$  in eq.(6.51), for a given value of  $d/(2\ell)$ .

Table 6.3: Luenberger's chain with  $M = 5$ 

	Matlab	The ODA
$\theta_1$	0.0893	0.0893
$\theta_2$	0.1147	0.1147
$\theta_3$	0.1599	0.1599
$\theta_4$	0.2625	0.2625
$\theta_5$	0.67856	0.6785
$f_{min}$	-12.2650	-12.2650
Iterations	16	7
CPU time (s)	0.2825	0.003261

### Hessian stabilization vs. reduced-Hessian stabilization

The reader may wonder why bother stabilizing the  $n \times n$  Hessian  $\mathbf{H} \equiv \nabla \nabla f$  when all that is needed in the procedure outlined in the first part of this subsection is the reduced Hessian. In other words, would it suffice with stabilizing the latter? The question is quite pertinent, and worth a detailed answer, as given below.

Stabilizing the reduced Hessian is plausible if the whole search is conducted within the feasible region, i.e., if the iterative process visits only points where the equality constraints are satisfied. As this is not the case—it would be very cumbersome to impose this condition—and, in fact, not being the case makes the ODA specially attractive, the full  $n \times n$  Hessian  $\mathbf{H}$  must be stabilized. This is best illustrated with a numerical example.

**Example 6.4.3 (Minimization of a quadratic, sign-indefinite objective function under one quadratic constraint)** *Using the ODA, find a local minimum of the problem below:*

$$f(\mathbf{x}) = \frac{1}{2}(-9x_1^2 - 8x_1x_2 + 3x_2^2) \quad \rightarrow \quad \min_{x_1, x_2}$$

*subject to*

$$h(\mathbf{x}) = x_1^2 + x_2^2 - 1 = 0$$

*Solution:* The above problem was introduced in Section 6.4, where it became apparent that its constant Hessian is sign-indefinite, and hence, calls for stabilization.

Before proceeding with the solution using the ODA, let us solve it symbolically: the gradients of  $f$  and  $h$  as well as the Hessian  $\mathbf{H} \equiv \nabla \nabla f$  are calculated below:

$$\nabla f = \begin{bmatrix} -9x_1 - 4x_2 \\ -4x_1 + 3x_2 \end{bmatrix}, \quad \mathbf{J} = [2x_1 \quad 2x_2], \quad \mathbf{H} = \begin{bmatrix} -9 & -4 \\ -4 & 3 \end{bmatrix}$$

The FONC are those given by eq.(5.9), i.e.,

$$\begin{bmatrix} -9x_1 - 4x_2 \\ -4x_1 + 3x_2 \end{bmatrix} + 2 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \lambda = \mathbf{0}_2$$

Given the simplicity of the Jacobian  $\mathbf{J}$ , an orthogonal complement  $\mathbf{L}$  thereof can be readily obtained upon rotating the vector array  $\mathbf{J}^T$  through  $90^\circ$  ccw. This is readily achieved upon first transposing  $\mathbf{J}$ , then swapping the two components of this array, and, finally reversing the sign of the first component, which yields  $\mathbf{L} = [-x_2, x_1]^T$ . Upon multiplying the above vector normality condition by  $\mathbf{L}^T$ , the term in  $\lambda$  is eliminated and a scalar equation, free of  $\lambda$ , is obtained:

$$12x_1x_2 + 4x_2^2 - 4x_1^2 = 0 \quad \text{or} \quad D(\mathbf{x}) \equiv x_2^2 + 3x_1x_2 - x_1^2 = 0$$

which is, in fact, the FONC in dual form. Upon subtracting the constraint equation  $h(x_1, x_2) = 0$  from the foregoing equation, an equation linear in  $x_2$  is obtained, namely,

$$2x_1^2 - 1 - 3x_1x_2 = 0 \quad \Rightarrow \quad x_2 = \frac{2x_1^2 - 1}{3x_1}$$

which, when substituted into  $h = 0$ , yields a quartic equation in  $x_1$ :

$$13x_1^4 - 13x_1^2 + 1 = 0$$

whose roots are

$$(x_1)_1 = 0.2897841489, \quad (x_1)_2 = -0.2897841489, \quad (x_1)_3 = 0.9570920265, \\ (x_1)_4 = -0.9570920265$$

These roots yield, correspondingly, four values of  $x_2$ :

$$(x_2)_1 = -0.9570920257, \quad (x_2)_2 = 0.9570920257, \quad (x_2)_3 = 0.2897841486, \\ (x_2)_4 = -0.2897841486$$

A display of  $D(\mathbf{x}) = 0$  is shown in Fig. 6.12a, while a superposition of the contours of  $h(\mathbf{x}) = 0$  and  $D(\mathbf{x}) = 0$  is included in Fig. 6.12b. Interestingly,  $z = D(\mathbf{x})$

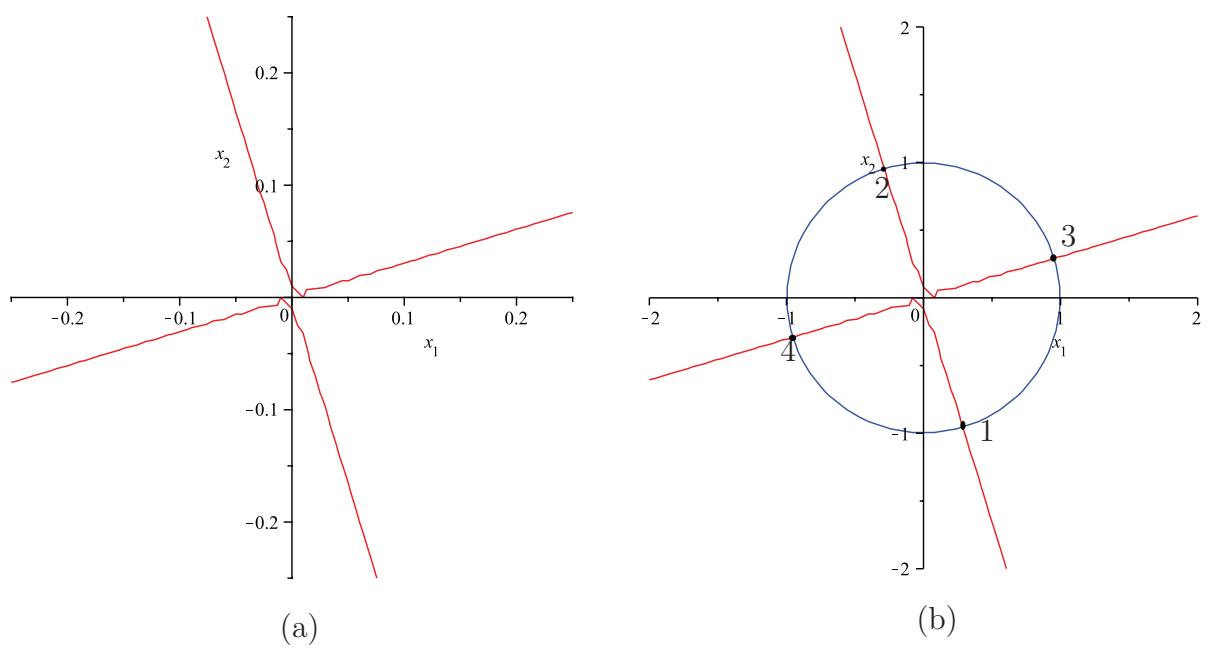


Figure 6.12: Plots of (a) the contour of  $D(\mathbf{x}) = 0$  and (b) the two contours  $h(\mathbf{x}) = 0$  and  $D(\mathbf{x}) = 0$  in the  $x_1$ - $x_2$  plane, their intersections being the four stationary points (SP)

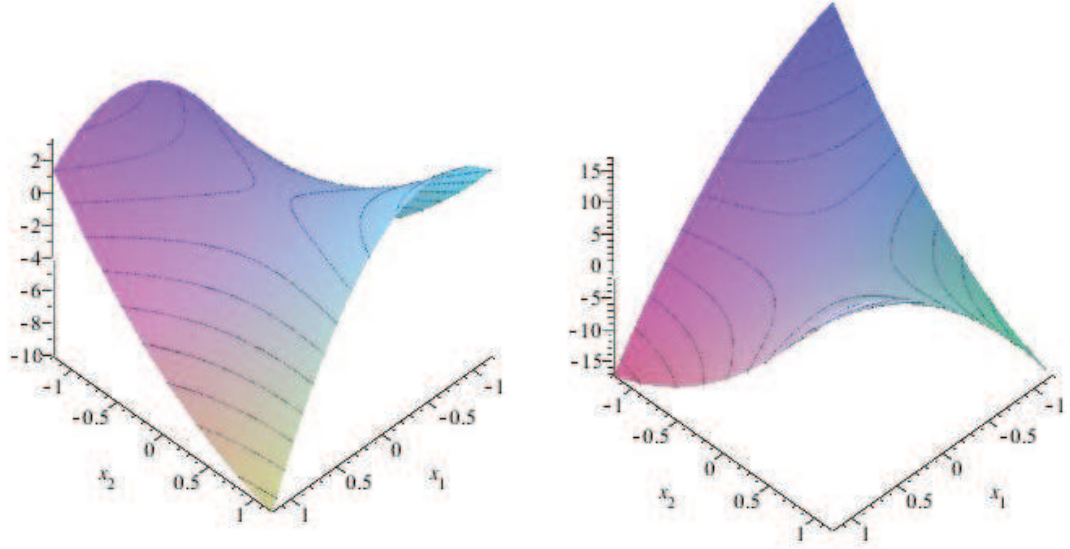


Figure 6.13: Display of the surface  $D(\mathbf{x})$  representing the dual form of the FONC of Example 6.4.3

represents a hyperbolic paraboloid, whose level contours are *equilateral* hyperbolas, that lie on the  $x_1$ - $x_2$  plane, degenerating into two lines at right angles passing through the origin, which are the asymptotes of all projected equilateral hyperbolas.

For the record, a display of the surface  $D(\mathbf{x})$  in 3D is shown in Fig. 6.13.

In order to determine the nature of each stationary point, the reduced Hessian  $H_u \equiv \mathbf{L}^T(\nabla\nabla f)\mathbf{L}$  is evaluated at each point, which, in this particular case, turns out to be a scalar. The four values are

$$(H_u)_1 = -10.21110254, (H_u)_2 = -10.21110254, (H_u)_3 = 4.211102551, \\ (H_u)_4 = 4.211102551$$

which thus show that SPs 1 and 2 are maxima, SPs 3 and 4 minima.

Now, in computing the SPs using the ODA, we use two approaches here:

- I. Stabilizing the reduced Hessian. The ODA was implemented in Maple, as per the code listed below.

Application of the ODA to a quadratic objective function that is not positive-definite, subject to a quadratic constraint. Precomputations are done inside

arb-obj-fn-bkgd.mw

Hessian stabilization done at reduced-Hessian level. See arb-obj-fnODA1.mw

$$f := -9/2 x_1^2 - 4 x_1 x_2 + 3/2 x_2^2$$

$$h := x_1^2 + x_2^2 - 1$$

```
> obj:=proc(x) 1/2*(-9*x[1]^2 -
> 8*x[1]*x[2] + 3*x[2]^2) end; #procedure to evaluate the
> objective function
obj := proc(x)    - 9/2 * x[1]^2 - 4 * x[1] * x[2] + 3/2 * x[2]^2 end proc;
> graf:=proc(x)
> Vector([-9*x[1]-4*x[2], -4*x[1]+3*x[2]]) end;
> #procedure to evaluate
> the gradient of the objective function
graf := proc(x)    Vector([-9 * x[1] - 4 * x[2], -4 * x[1] + 3 * x[2]]) end proc;
> Hf:= Matrix([[-9,-4], [-4,3]]); #procedure
> not needed to evaluate the Hessian of the objective function, as
> Hessian is constant!
```

$$Hf := \begin{bmatrix} -9 & -4 \\ -4 & 3 \end{bmatrix}$$

```
constr := proc(x)    x[1]^2 + x[2]^2 - 1 end proc;
```

```
grah := proc(x)    Vector([2 * x[1], 2 * x[2]]) end proc;
```

$$Hh := \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$E := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

$$x0 := \begin{bmatrix} 2 \\ -2 \end{bmatrix}$$

$$f0 := 4$$

$$gf0 := \begin{bmatrix} -10 \\ -14 \end{bmatrix}$$

$$h0 := 7$$

$$J0 := \begin{bmatrix} 4 \\ -4 \end{bmatrix}$$

$$nJ0 := 4\sqrt{2}$$

$$JMP := \begin{bmatrix} 1/8 \\ -1/8 \end{bmatrix}$$

$$L0 := \begin{bmatrix} 4 \\ 4 \end{bmatrix}$$

$$L0 := \begin{bmatrix} 1/2\sqrt{2} \\ 1/2\sqrt{2} \end{bmatrix}$$

$$Delv := \begin{bmatrix} -\frac{7}{8} \\ \frac{7}{8} \end{bmatrix}$$

$$pro := \begin{bmatrix} -13/2\sqrt{2} \\ -1/2\sqrt{2} \end{bmatrix}$$

$$H_u := -7$$

$$\mu := 8$$

$$H_{us} := 1$$

$$brack := \begin{bmatrix} \frac{35}{8} \\ \frac{49}{8} \end{bmatrix}$$

$$brack := \begin{bmatrix} -\frac{45}{8} \\ -\frac{63}{8} \end{bmatrix}$$



$$r := \frac{27}{4} \sqrt{2}$$

$$Deluk := 9.545941544$$

$$Ldelu := \begin{bmatrix} 4.772970772 \sqrt{2} \\ 4.772970772 \sqrt{2} \end{bmatrix}$$

$$Delx := \begin{bmatrix} -\frac{7}{8} + 4.772970772 \sqrt{2} \\ \frac{7}{8} + 4.772970772 \sqrt{2} \end{bmatrix}$$

$$Delx := \begin{bmatrix} 5.874999997 \\ 7.624999997 \end{bmatrix}$$

$$x1 := \begin{bmatrix} 7.87499999699999974 \\ 5.62499999699999975 \end{bmatrix}$$

$$f1 := -408.7968746$$

$$gf1 := \begin{bmatrix} -93.37499996 \\ -14.62500000 \end{bmatrix}$$

$$h1 := 92.65624992$$

$$J1 := \begin{bmatrix} 15.74999999 \\ 11.24999999 \end{bmatrix}$$

$$nJ1 := 19.3552318368961913$$

$$JMP := \begin{bmatrix} 0.0420420420678066370 \\ 0.0300300300408066385 \end{bmatrix}$$

$$L1 := \begin{bmatrix} -11.2499999899999992 \\ 15.7499999899999992 \end{bmatrix}$$

$$L1 := \begin{bmatrix} -0.581238193545843718 \\ 0.813733471170843714 \end{bmatrix}$$

$$Delv := \begin{bmatrix} -3.89545795698184526 \\ -2.78246996856608764 \end{bmatrix}$$

$$pro := \begin{bmatrix} 1.97620985722921860 \\ 4.76615318769590602 \end{bmatrix}$$

$$H_u := 2.72972973007237085$$

$$brack := \begin{bmatrix} 46.1890014871009598 \\ 7.23442192222911818 \end{bmatrix}$$

$$brack := \begin{bmatrix} -47.1859984728990369 \\ -7.39057807777088182 \end{bmatrix}$$

$$r := -21.4123437598611340$$

$$Deluk := -7.844125931$$

$$Ldelu := \begin{bmatrix} 4.55930558608054958 \\ -6.38302782213385632 \end{bmatrix}$$

$$Delx := \begin{bmatrix} 0.663847629098704317 \\ -9.16549779069994486 \end{bmatrix}$$

$$Delx := \begin{bmatrix} 0.663847629099999948 \\ -9.16549779099999994 \end{bmatrix}$$

$$x2 := \begin{bmatrix} 8.53884762609999903 \\ -3.54049779400000020 \end{bmatrix}$$

$$f2 := -188.3738629$$

$$gf2 := \begin{bmatrix} -62.68763745 \\ -44.77688388 \end{bmatrix}$$

$$h2 := 84.44704341$$

$$J2 := \begin{bmatrix} 17.07769525 \\ -7.080995588 \end{bmatrix}$$

$$nJ2 := 18.4875139910467148$$

$$JMP := \begin{bmatrix} 0.0499657289799323446 \\ -0.0207174973717899476 \end{bmatrix}$$

$$L2 := \begin{bmatrix} 7.08099558800000040 \\ 17.0776952500000014 \end{bmatrix}$$

$$L2 := \begin{bmatrix} 0.383015022589675624 \\ 0.923742113756327088 \end{bmatrix}$$

$$Delv := \begin{bmatrix} -4.21945808418064239 \\ 1.74953139990210672 \end{bmatrix}$$

$$pro := \begin{bmatrix} -7.14210365833238914 \\ 1.23916625091027877 \end{bmatrix}$$

$$H_u := -1.59086294212262080$$

$$\mu := 1.68$$

$$H_{us} := 0.089137058$$

$$brack := \begin{bmatrix} 30.9769971580173546 \\ 22.1264265364288910 \end{bmatrix}$$

$$brack := \begin{bmatrix} -31.7106402919826422 \\ -22.6504573435711088 \end{bmatrix}$$

$$r := 33.0688329518647066$$

$$Deluk := 370.9886067$$

$$Ldelu := \begin{bmatrix} 142.094209575712796 \\ 342.697799732572719 \end{bmatrix}$$

$$Delx := \begin{bmatrix} 137.874751491532152 \\ 344.447331132474801 \end{bmatrix}$$

$$Delx := \begin{bmatrix} 137.8747515000000002 \\ 344.4473310999999985 \end{bmatrix}$$

$$x\mathcal{J} := \begin{bmatrix} 146.413599126099996 \\ 340.906833306000010 \end{bmatrix}$$

$$f\mathcal{J} := -121793.6212$$

$$gf\mathcal{J} := \begin{bmatrix} -2681.349725 \\ 437.0661036 \end{bmatrix}$$

$$h\mathcal{J} := 137653.4110$$

$$J\mathcal{J} := \begin{bmatrix} 292.8271982 \\ 681.8136666 \end{bmatrix}$$

$$nJ\mathcal{J} := 742.036147346069697$$

$$JMP := \begin{bmatrix} 0.000531815864169775821 \\ 0.00123827064744849399 \end{bmatrix}$$

$$L\mathcal{J} := \begin{bmatrix} -681.8136666000000033 \\ 292.8271981999999998 \end{bmatrix}$$

$$L\mathcal{J} := \begin{bmatrix} -0.918841580778544830 \\ 0.394626595020851790 \end{bmatrix}$$

$$Delv := \begin{bmatrix} -73.2062677268823166 \\ -170.452178362463627 \end{bmatrix}$$

$$pro := \begin{bmatrix} 6.69106784692349608 \\ 4.85924610817673486 \end{bmatrix}$$

$$H_u := -4.23044361152556903$$

$$\mu := 4.4$$

$$H_{us} := 0.169556388$$

$$brack := \begin{bmatrix} 1340.66512299179544 \\ -218.531464179861614 \end{bmatrix}$$

$$brack := \begin{bmatrix} -1340.68460200820460 \\ 218.534639420138404 \end{bmatrix}$$

$$r := -1318.11633968315187$$

$$Deluk := -7773.911414$$

$$Ldelu := \begin{bmatrix} 7142.99305247213488 \\ -3067.79219130055618 \end{bmatrix}$$

$$Delx := \begin{bmatrix} 7069.78678474525260 \\ -3238.24436966301983 \end{bmatrix}$$

$$Delx := \begin{bmatrix} 7069.78678500000024 \\ -3238.24436999999989 \end{bmatrix}$$

$$x_4' := \begin{bmatrix} 7216.20038412610029 \\ -2897.33753669399994 \end{bmatrix}$$

$$f_4' := -138108045.7$$

$$gf_4' := \begin{bmatrix} -53356.45331 \\ -37556.81415 \end{bmatrix}$$

**Procedure is diverging. Stop!**

- II. Stabilizing the full  $2 \times 2$  Hessian. The ODA was implemented in Maple, as per the code listed below.

Application of the ODA to a quadratic objective function that is not positive-definite, subject to a quadratic constraint. Precomputations are done inside `arb-obj-fn-bkgd.mw`

**Hessian stabilization is done at the  $n \times n$  Hessian level.**

Look at accompanying file `ArbObjFnODA1-semigraph.mw` for further details

$$f := -9/2 x_1^2 - 4 x_1 x_2 + 3/2 x_2^2$$

$$h := x_1^2 + x_2^2 - 1$$

```
> obj:=proc(x) 1/2*(-9*x[1]^2 -
> 8*x[1]*x[2] + 3*x[2]^2) end; #procedure to e
> valuate the
> objective function
obj := proc(x)    - 9/2 * x[1]^2 - 4 * x[1] * x[2] + 3/2 * x[2]^2 end proc;
> graf:=proc(x)
> Vector([-9*x[1]-4*x[2],-4*x[1]+3*x[2]]) end; #procedure to
> evaluate
> the gradient of the objective function
graf := proc(x)    Vector([-9 * x[1] - 4 * x[2], -4 * x[1] + 3 * x[2]]) end proc;
> Hf:= Matrix([[ -9,-4],[ -4,3]]); #procedure
> not needed to evaluate the Hessian
> of the objective function, as
> Hessian is constant!
```

$$Hf := \begin{bmatrix} -9 & -4 \\ -4 & 3 \end{bmatrix}$$

$$ID := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Delta Hf := \begin{bmatrix} 13.2000000000000 & 0.0 \\ 0.0 & 13.2000000000000 \end{bmatrix}$$

$$Hfstab := \begin{bmatrix} 4.200000000000000 & -4.0 \\ -4.0 & 16.200000000000000 \end{bmatrix}$$

$$constr := \mathbf{proc}(x) \quad x[1]^2 + x[2]^2 + -1 \mathbf{end} \mathbf{proc};$$

$$grah := \mathbf{proc}(x) \quad \mathbf{Vector}([2 * x[1], 2 * x[2]]) \mathbf{end} \mathbf{proc};$$

$$E := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

$$x0 := \begin{bmatrix} 2 \\ -2 \end{bmatrix}$$

$$f0 := 4$$

$$gf0 := \begin{bmatrix} -10 \\ -14 \end{bmatrix}$$

$$h0 := 7$$

$$J0 := \begin{bmatrix} 4 \\ -4 \end{bmatrix}$$

$$nJ0 := 4\sqrt{2}$$

$$JMP := \begin{bmatrix} 1/8 \\ -1/8 \end{bmatrix}$$

$$L0 := \begin{bmatrix} 4 \\ 4 \end{bmatrix}$$

$$L0 := \begin{bmatrix} 1/2\sqrt{2} \\ 1/2\sqrt{2} \end{bmatrix}$$

$$Delv := \begin{bmatrix} -\frac{7}{8} \\ \frac{7}{8} \end{bmatrix}$$

$$pro := \begin{bmatrix} 0.0999999999999996 \sqrt{2} \\ 6.100000000000000 \sqrt{2} \end{bmatrix}$$

$$H_u := 6.200000000000000$$

$$brack := \begin{bmatrix} -7.175000000000000 \\ 17.675000000000000 \end{bmatrix}$$

$$brack := \begin{bmatrix} -17.175000000000000 \\ 3.675000000000000 \end{bmatrix}$$

$$r := 6.750000000000000 \sqrt{2}$$

$$Deluk := 1.53966799088710$$

$$Ldelu := \begin{bmatrix} 0.769833995443548 \sqrt{2} \\ 0.769833995443548 \sqrt{2} \end{bmatrix}$$

$$Delx := \begin{bmatrix} -\frac{7}{8} + 0.769833995443548 \sqrt{2} \\ \frac{7}{8} + 0.769833995443548 \sqrt{2} \end{bmatrix}$$

$$Delx := \begin{bmatrix} 0.213709676844912 \\ 1.96370967684491 \end{bmatrix}$$

$$x1 := \begin{bmatrix} 2.21370967684491 \\ -0.0362903231550875 \end{bmatrix}$$

$$f1 := -21.7289769605966$$

$$gf1 := \begin{bmatrix} -19.7782257989839 \\ -8.96370967684491 \end{bmatrix}$$

$$h1 := 3.90182752091151$$

$$J1 := \begin{bmatrix} 4.42741935368982 \\ -0.0725806463101750 \end{bmatrix}$$



$$nJ1 := 4.42801423706451302$$

$$JMP := \begin{bmatrix} 0.225804525699389 \\ -0.00370171359561084 \end{bmatrix}$$

$$L1 := \begin{bmatrix} 0.0725806463101750 \\ 4.42741935368982 \end{bmatrix}$$

$$L1 := \begin{bmatrix} 0.0163912405046195 \\ 0.999865654695380 \end{bmatrix}$$

$$Delv := \begin{bmatrix} -0.881050312720246 \\ 0.0144434479818867 \end{bmatrix}$$

$$pro := \begin{bmatrix} -3.93061940866212 \\ 16.1322586440467 \end{bmatrix}$$

$$H_u := 16.0656636227854399$$

$$brack := \begin{bmatrix} 7.87167902255467 \\ 3.56753159482664 \end{bmatrix}$$

$$brack := \begin{bmatrix} -11.9065467764292 \\ -5.39617808201827 \end{bmatrix}$$

$$r := 5.59061620262201052$$

$$Deluk := 0.3479853889$$

$$Ldelu := \begin{bmatrix} 0.00570391220155345 \\ 0.347938638696925 \end{bmatrix}$$

$$Delx := \begin{bmatrix} -0.875346400518692 \\ 0.362382086678812 \end{bmatrix}$$

$$Delx := \begin{bmatrix} -0.875346400518692 \\ 0.362382086678812 \end{bmatrix}$$

$$x2 := \begin{bmatrix} 1.33836327632622 \\ 0.326091763523724 \end{bmatrix}$$

$$f2 := -9.64668637407735$$

$$gf2 := \begin{bmatrix} -13.3496365410309 \\ -4.37517781473371 \end{bmatrix}$$

$$h2 := 0.897552097656666$$

$$J2 := \begin{bmatrix} 2.67672655265244 \\ 0.652183527047449 \end{bmatrix}$$

$$nJ2 := 2.75503328303428408$$

$$JMP := \begin{bmatrix} 0.352655212517412 \\ 0.0859243242845873 \end{bmatrix}$$

$$L2 := \begin{bmatrix} -0.652183527047449 \\ 2.67672655265244 \end{bmatrix}$$

$$L2 := \begin{bmatrix} -0.236724373201412 \\ 0.971576847818982 \end{bmatrix}$$

$$Delv := \begin{bmatrix} -0.316526425744561 \\ -0.0771215575013630 \end{bmatrix}$$

$$pro := \begin{bmatrix} -4.88054975872186 \\ 16.6864424274732 \end{bmatrix}$$

$$H_u := 17.3675062175090354$$

$$brack := \begin{bmatrix} 3.15722406170650 \\ 1.03474103047416 \end{bmatrix}$$

$$brack := \begin{bmatrix} -10.1924124793244 \\ -3.34043678425955 \end{bmatrix}$$

$$r := 0.832698585611156617$$

$$Deluk := 0.04794577731$$

$$Ldelu := \begin{bmatrix} -0.0113499340813642 \\ 0.0465830071850807 \end{bmatrix}$$

$$Delx := \begin{bmatrix} -0.327876359825925 \\ -0.0305385503162823 \end{bmatrix}$$

$$Delx := \begin{bmatrix} -0.327876359825925 \\ -0.0305385503162823 \end{bmatrix}$$

$$x\mathcal{J} := \begin{bmatrix} 1.01048691650029 \\ 0.295553213207442 \end{bmatrix}$$

$$f\mathcal{J} := -5.65846020542933$$

$$gf\mathcal{J} := \begin{bmatrix} -10.2765951013324 \\ -3.15528802637885 \end{bmatrix}$$

$$h\mathcal{J} := 0.108435510255517$$

$$J\mathcal{J} := \begin{bmatrix} 2.02097383300059 \\ 0.591106426414884 \end{bmatrix}$$

$$nJ\mathcal{J} := 2.10564527901117948$$

$$JMP := \begin{bmatrix} 0.455816737670390 \\ 0.133319985892344 \end{bmatrix}$$

$$L\mathcal{J} := \begin{bmatrix} -0.591106426414884 \\ 2.02097383300059 \end{bmatrix}$$

$$L\mathcal{J} := \begin{bmatrix} -0.280724598936670 \\ 0.959788361922486 \end{bmatrix}$$

$$Delv := \begin{bmatrix} -0.0494267205322938 \\ -0.0144566206974946 \end{bmatrix}$$

$$pro := \begin{bmatrix} -5.01819676322396 \\ 16.6714698588910 \end{bmatrix}$$

$$H_u := 17.4098140204463832$$

$$brack := \begin{bmatrix} 0.502666967580623 \\ 0.154337020036691 \end{bmatrix}$$

$$brack := \begin{bmatrix} -9.77392813375180 \\ -3.00095100634216 \end{bmatrix}$$

$$r := 0.136495795203464886$$

$$Deluk := 0.007840163889$$

$$Ldelu := \begin{bmatrix} -0.00220092686333729 \\ 0.00752489805622713 \end{bmatrix}$$

$$Delx := \begin{bmatrix} -0.0516276473956311 \\ -0.00693172264126750 \end{bmatrix}$$

$$Delx := \begin{bmatrix} -0.0516276473956311 \\ -0.00693172264126750 \end{bmatrix}$$

$$x4 := \begin{bmatrix} 0.958859269104663 \\ 0.288621490566175 \end{bmatrix}$$

$$f4 := -5.11938595950945$$

$$gf4 := \begin{bmatrix} -9.78421938420667 \\ -2.96957260472013 \end{bmatrix}$$

$$h4 := 0.00271346276456974$$

$$J_4 := \begin{bmatrix} 1.91771853820933 \\ 0.577242981132349 \end{bmatrix}$$

$$nJ_4 := 2.00271162453766127$$

$$JMP := \begin{bmatrix} 0.478132240237905 \\ 0.143920223031338 \end{bmatrix}$$

$$L_4 := \begin{bmatrix} -0.577242981132349 \\ 1.91771853820933 \end{bmatrix}$$

$$L_4 := \begin{bmatrix} -0.288230703763967 \\ 0.957560995899831 \end{bmatrix}$$

$$Delv := \begin{bmatrix} -0.00129739403042587 \\ -0.000390522166264107 \end{bmatrix}$$

$$pro := \begin{bmatrix} -5.04081293940798 \\ 16.6654109486331 \end{bmatrix}$$

$$H_u := 17.4110645661211550$$

$$brack := \begin{bmatrix} 0.0132386349388893 \\ 0.00401800962291116 \end{bmatrix}$$

$$brack := \begin{bmatrix} -9.77098074926778 \\ -2.96555459509722 \end{bmatrix}$$

$$r := 0.0234027536509913148$$

$$Deluk := 0.001344131116$$

$$Ldelu := \begin{bmatrix} -0.000387419857515726 \\ 0.00128708753005691 \end{bmatrix}$$

$$\begin{aligned}
Delx &:= \begin{bmatrix} -0.00168481388794160 \\ 0.000896565363792804 \end{bmatrix} \\
Delx &:= \begin{bmatrix} -0.00168481388794160 \\ 0.000896565363792804 \end{bmatrix} \\
x5 &:= \begin{bmatrix} 0.957174455216722 \\ 0.289518055929967 \end{bmatrix} \\
f5 &:= -5.10556931251391 \\
gf5 &:= \begin{bmatrix} -9.77264232067037 \\ -2.96014365307699 \end{bmatrix}
\end{aligned}$$

The foregoing value of  $x5$  is accepted as the solution of the problem, which corresponds to SP#3 of Fig. 6.12.

### 6.4.2 A Gradient-based Approach

Now we relax the assumption of Subsection 6.4.1, and assume that both the objective and the constraint functions are continuous and have continuous derivatives up to the first order only. Moreover, we assume that these derivatives are available. In this light, the objective function can be expanded at the  $k$ th iteration in the form

$$f(\mathbf{x}^k + \Delta\mathbf{x}^k) = f(\mathbf{x}^k) + (\nabla f_k)^T \Delta\mathbf{x}^k + \text{HOT}(\Delta\mathbf{x}^k) \quad (6.56)$$

with the usual notation. Furthermore, the constraint function is approximated to a first order as appearing in eq.(6.42b), which then allows us to express the increment  $\Delta\mathbf{x}^k$  as

$$\Delta\mathbf{x}^k = \underbrace{-\mathbf{J}_k^\dagger \mathbf{h}^k}_{\Delta\mathbf{v}^k} + \mathbf{L}_k \Delta\mathbf{u}^k \quad (6.57)$$

The expansion (6.56) can thus be expressed as

$$f(\mathbf{x}^k + \Delta\mathbf{x}^k) = f(\mathbf{x}^k) + \tilde{f}(-\mathbf{J}_k^\dagger \mathbf{h}^k + \mathbf{L}_k \Delta\mathbf{u}^k) \quad (6.58)$$

Apparently, the second term of the foregoing expansion is a function of the  $(n - p)$ -dimensional variable  $\Delta\mathbf{u}^k$ , the problem at hand thus reducing to

$$\tilde{f}(\Delta\mathbf{u}^k) \rightarrow \min_{\Delta\mathbf{u}^k} \quad (6.59)$$

subject to no constraints. The dimension of the problem has thus been reduced to  $n - l$ , lower than  $n$ , that of the original problem. Not only this, the problem has also been reduced to one of the unconstrained type. So far, we have then the same simplifications as in Subsection 6.4.1. However, rather than relying on the availability of second derivatives, we resort, for the numerical solution of the problem at hand, only to gradient evaluations.

Under the assumption that the objective and the constraint functions have continuous gradients, and these are available, we can use any of the methods introduced in Section 4.5 to solve the problem. Of these methods, we recommend two of the most effective ones, the Fletcher-Reeves and the BFGS methods.





# Chapter 7

## Inequality-Constrained Optimization

### 7.1 Introduction

The conditions under which a design problem is formulated involve, more often than not, inequality constraints, in addition to equalities. In fact, inequality constraints arise naturally in design because the resources available to accomplish a design job are *finite*. For example, a designed object must: fit into a designated region; be realizable within a given budget; and be delivered by a certain date.

In this chapter we address two issues around inequality-constrained problems: the *normality conditions* and the *methods of solution*. As to the former, we will not dwell into their rigorous derivation, which are elusive to a simple analysis with the tools of linear algebra; rather, we will introduce the first-order normality conditions without derivation, and illustrate their validity with examples. The second-order normality conditions will be derived using an intuitive approach, more so than a mathematical formulation.

The reason why linear algebra is no longer sufficient to derive the normality conditions of inequality-constrained problems lies in the nature of inequalities, which define a *region* of  $\mathbb{R}^n$  that is neither a vector subspace nor a manifold, as we encountered when studying equality-constrained problems. Now we will speak, more generally, of the *feasible region*  $\mathcal{F}$ , which can have sharp edges and vertices, not occurring in manifolds. For this reason, a simple transformation of the form  $\mathbf{x}(\mathbf{u})$  is not sufficient, in general, to guarantee the fulfillment of the inequality constraints.

## 7.2 The Karush-Kuhn-Tucker Conditions

The first-order normality conditions of equality-constrained problems are classical results, first proposed by Joseph Louis de Lagrange, brilliant mathematician born in Turin in 1736 and dead in Paris in 1813. Lagrange founded in Turin a society that would become the Academy of Sciences, then went to Berlin, to the Academy of Friedrich II, to succeed Euler. Rather late in his life, in 1787, did Lagrange move to Paris, invited by Louis XVI to teach at *École normale*. Appointed senator and made count by Napoleon, Lagrange became one of the first professors at *École polytechnique*.

The first-order normality conditions for inequality-constrained problems had to wait until well into the XX century. These conditions were disclosed first by W. Karush in his M.S. thesis in the Department of Mathematics at the University of Chicago Karush (1939). Apparently, these results were never published in the archival literature, for which reason they remained unknown. Twelve years later, they were published in the *Proc. Second Berkeley Symposium* by Kuhn and Tucker (1951). The credit of these normality conditions has gone mostly to Kuhn and Tucker, but given their history, these conditions are sometimes referred to as the *Karush-Kuhn-Tucker conditions*, abbreviated *KKT conditions*.

An extensive discussion of the KKT conditions is found in Hillier and Liebermann (1995)

The problem under discussion is formulated as

$$f(\mathbf{x}) \rightarrow \min_{\mathbf{x}} \quad (7.1a)$$

subject to

$$\mathbf{g}(\mathbf{x}) \leq \mathbf{0}_p \quad (7.1b)$$

$$\mathbf{h}(\mathbf{x}) = \mathbf{0}_l \quad (7.1c)$$

where inequality (7.1b) is to be taken with a grain of salt: Arrays *not forming ordered sets*, this relation has no verbatim meaning. It is to be interpreted as shorthand for a set of  $p$  inequalities, namely,

$$g_1(\mathbf{x}) \equiv g_1(x_1, x_2, \dots, x_n) \leq 0$$

$$g_2(\mathbf{x}) \equiv g_2(x_1, x_2, \dots, x_n) \leq 0$$

$$\vdots$$

$$g_p(\mathbf{x}) \equiv g_p(x_1, x_2, \dots, x_n) \leq 0$$

To formulate the normality conditions, we proceed as before, namely, by defining a *Lagrangian* upon adjoining the equality and the inequality constraints to the objective function, namely,

$$F(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \equiv f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x}) + \boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}) \rightarrow \min_{\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}} \quad (7.2)$$

where  $\boldsymbol{\lambda}$  is found so that  $\mathbf{h}(\mathbf{x})$  will vanish at a stationary point, the role played by  $\boldsymbol{\mu}$  being described below. We start by introducing a

**Definition 7.2.1 (Active constraint)** *When, at a stationary point, the  $i$ th constraint of eq.(7.1b) holds with the equality sign, this constraint is said to be active.*

By opposition, we will define as *passive* any non-active constraint.

The components of  $\boldsymbol{\mu}$  are chosen so that the contribution of all passive constraints to the Lagrangian vanish at a stationary point, that of all active constraints necessarily vanishing.

While the normality conditions cannot be derived by simply making the gradient of the foregoing Lagrangian equal to zero, these conditions look very much like those associated with equality-constrained problems. Indeed,  $\mathbf{x}_o$  is a *feasible* stationary point if

$$\mathbf{h}(\mathbf{x}_o) = \mathbf{0}_l, \quad \mathbf{g}(\mathbf{x}_o) \leq \mathbf{0}_p \quad (7.3a)$$

$$\nabla f|_{\mathbf{x}=\mathbf{x}_o} + \mathbf{J}_o^T \boldsymbol{\lambda} + \mathbf{G}_o^T \boldsymbol{\mu} = \mathbf{0}_n \quad (7.3b)$$

$$\boldsymbol{\mu} \geq \mathbf{0}_m, \quad \boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}) = 0 \quad (7.3c)$$

where

$$\mathbf{G}_o \equiv \mathbf{G}(\mathbf{x}_o) \equiv \left. \frac{\partial \mathbf{g}}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}_o}, \quad \mathbf{J}_o \equiv \mathbf{J}(\mathbf{x}_o) \equiv \left. \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}_o} \quad (7.3d)$$

i.e.,  $\mathbf{G}$  and  $\mathbf{J}$  are, respectively, the  $p \times n$  and the  $l \times n$  gradients of the inequality- and equality-constraint functions  $\mathbf{g}(\mathbf{x})$  and  $\mathbf{h}(\mathbf{x})$ , namely,

$$\mathbf{G} \equiv \begin{bmatrix} (\nabla g_1)^T \\ (\nabla g_2)^T \\ \vdots \\ (\nabla g_p)^T \end{bmatrix}, \quad \mathbf{J} \equiv \begin{bmatrix} (\nabla h_1)^T \\ (\nabla h_2)^T \\ \vdots \\ (\nabla h_l)^T \end{bmatrix} \quad (7.4)$$

Hence, the KKT condition (7.3b) can be expressed alternatively as

$$\begin{aligned} \nabla f|_{\mathbf{x}=\mathbf{x}_o} + \lambda_1 \nabla h_1|_{\mathbf{x}=\mathbf{x}_o} + \lambda_2 \nabla h_2|_{\mathbf{x}=\mathbf{x}_o} + \cdots + \lambda_l \nabla h_l|_{\mathbf{x}=\mathbf{x}_o} \\ + \mu_1 \nabla g_1|_{\mathbf{x}=\mathbf{x}_o} + \mu_2 \nabla g_2|_{\mathbf{x}=\mathbf{x}_o} + \cdots + \mu_p \nabla g_p|_{\mathbf{x}=\mathbf{x}_o} = \mathbf{0}_n \end{aligned} \quad (7.5)$$

It is noteworthy that, the components of  $\boldsymbol{\mu}$  being non-negative and those of  $\mathbf{g}(\mathbf{x})$  non-positive, each  $\mu_i g_i(\mathbf{x})$  is non-positive. Hence,  $\boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}_o) = 0$  of eq.(7.3c) implies

$$\mu_i g_i(\mathbf{x}_o) = 0, \quad i = 1, \dots, p$$

at a feasible stationary point.

Relations (7.3a–c) are the *Karush-Kuhn-Tucker* (KKT) *conditions*. These are the first-order normality conditions of the inequality-constrained problem at hand, and hence, guarantee a feasible stationary point of  $\mathcal{F}$ , but not a minimum. The latter is guaranteed by the second-order normality conditions, to be studied in Section 7.3.

**Remark:** In the absence of inequality constraints, eq.(7.3b) reduces to the FONC of equality-constrained problems, eq.(5.8a).

Relations (7.3c) of the KKT conditions are referred to as the *complementary slackness*, sometimes as the *transversality condition*<sup>1</sup>. Complementary slackness thus guarantees that, at a SP, the Lagrangian equals the objective function, under the assumption that the equality constraints are verified.

Further, if  $a$  of the  $p$  inequality constraints are active, we can partition vector  $\mathbf{g}(\mathbf{x}_o)$ , very likely after a reshuffling of its components, in the form

$$\mathbf{g}(\mathbf{x}_o) = \begin{bmatrix} \mathbf{g}_a \\ \mathbf{g}_{p'} \end{bmatrix}, \quad p' = p - a \quad (7.6)$$

where  $\mathbf{g}_a$  and  $\mathbf{g}_{p'}$  are  $a$ - and  $(p - a)$ -dimensional vectors, respectively. Now, the Karush-Kuhn-Tucker conditions of eq.(7.5) can be restated as

$$\begin{aligned} \nabla f|_{\mathbf{x}=\mathbf{x}_o} + \lambda_1 \nabla h_1|_{\mathbf{x}=\mathbf{x}_o} + \lambda_2 \nabla h_2|_{\mathbf{x}=\mathbf{x}_o} + \dots + \lambda_l \nabla h_l|_{\mathbf{x}=\mathbf{x}_o} \\ + \mu_1 \nabla g_1|_{\mathbf{x}=\mathbf{x}_o} + \mu_2 \nabla g_2|_{\mathbf{x}=\mathbf{x}_o} + \dots + \mu_a \nabla g_a|_{\mathbf{x}=\mathbf{x}_o} = \mathbf{0}_n, \quad \mathbf{g}_a = \mathbf{0}_a \end{aligned} \quad (7.7)$$

where the first equation resembles the FONC of equality-constrained problems, this time with  $l + a$  equality constraints, as derived in eq.(5.8a).

Moreover, the KKT conditions can be cast in the *canonical form*

$$\nabla_{\mathbf{x}} F \equiv \frac{\partial F}{\partial \mathbf{x}} = \mathbf{0}_n, \quad \nabla_{\boldsymbol{\mu}} F \leq \mathbf{0}_p, \quad \boldsymbol{\mu}^T \nabla_{\boldsymbol{\mu}} F = 0 \quad (7.8)$$

A proof of the KKT can be found in (Culioli, 1994; Boyd and Vandenberghe, 2004).

---

<sup>1</sup>Not to be confused with the transversality condition of *calculus of variations*.

To illustrate the validity of the KKT conditions, we give in Fig. 7.1 a mechanical interpretation: A heavy ball of weight  $\mathbf{w}$  is constrained to lie in a box, under the action of the gravity field  $\mathbf{g}$ , as depicted in Fig. 7.1a; the ball is shown in its equilibrium position in Fig. 7.1b; in Fig. 7.1c, the weight of the ball, equal to  $-\nabla V$ , where  $V$  is the potential energy of the ball, is decomposed into the two forces normal to the box walls. Notice that these two components push the walls, but cannot pull them, which is the reason why  $\mu_i > 0$ , for  $i = 1, 2$ .

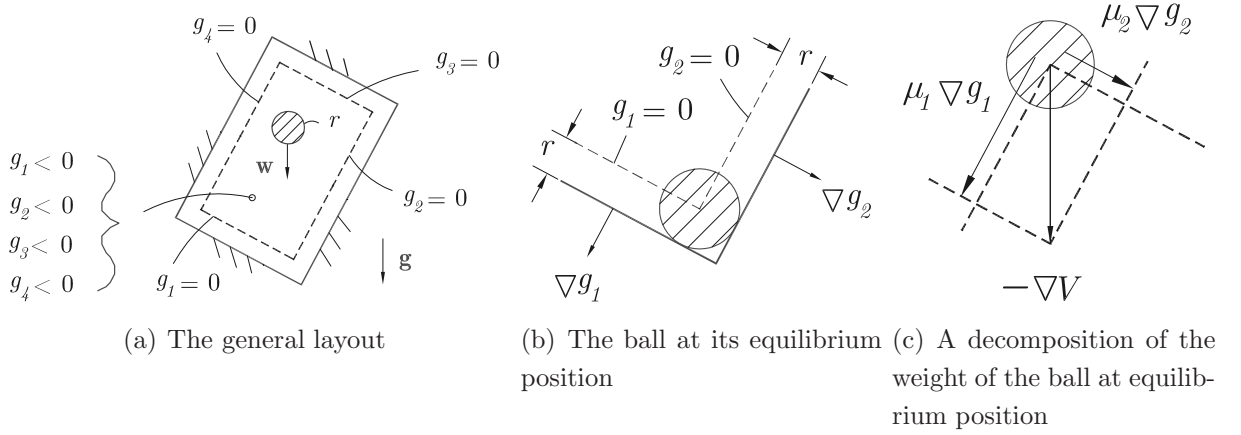


Figure 7.1: A heavy ball inside a box inclined with respect to the vertical

**Example 7.2.1** Consider the problem

$$f = \frac{1}{2}(x_1^2 + x_2^2) \rightarrow \min_{x_1, x_2}$$

subject to

$$x_1 + x_2 \geq 10$$

$$x_1 \geq 0$$

$$x_2 \geq 0$$

The objective function and the constraints are illustrated in Fig. 7.2.

Solution: For starters, we must express the inequality constraints in the standard form adopted at the outset, i.e., as  $g_i(\mathbf{x}) \leq 0$ , whence,

$$g_1 \equiv -x_1 - x_2 + 10 \leq 0, \quad g_2 \equiv -x_1 \leq 0, \quad g_3 \equiv -x_2 \leq 0$$

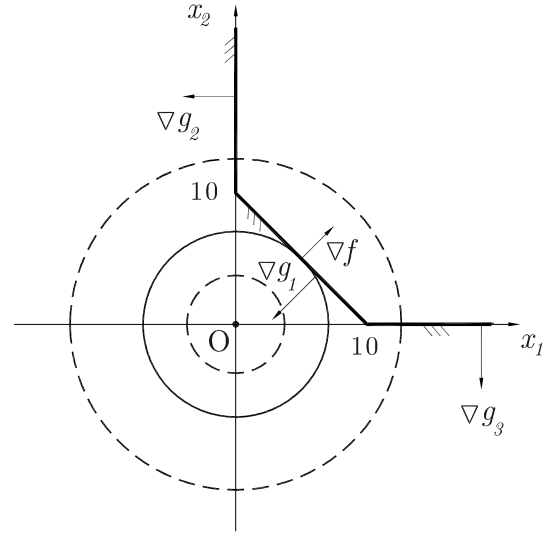


Figure 7.2: A quadratic objective function subject to linear inequality constraints

Apparently, the minimum is found at  $\mathbf{x}_o = [5, 5]^T$ . We evaluate then the items entering in the KKT conditions at  $\mathbf{x}_o$ :

$$\begin{aligned}\nabla g_1 &= \begin{bmatrix} -1 \\ -1 \end{bmatrix}, & \nabla g_2 &= \begin{bmatrix} -1 \\ 0 \end{bmatrix}, & \nabla g_3 &= \begin{bmatrix} 0 \\ -1 \end{bmatrix} \\ \nabla f &= \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, & \Rightarrow & \nabla f|_{\mathbf{x}=\mathbf{x}_o} = \begin{bmatrix} 5 \\ 5 \end{bmatrix}\end{aligned}$$

By inspection, only the first constraint is active, and hence,

$$g_1(\mathbf{x}_o) = 0, \quad \mu_1 > 0, \quad \mu_2 = \mu_3 = 0$$

The KKT condition (7.7) thus reduces to

$$\nabla f|_{\mathbf{x}=\mathbf{x}_o} + \mu_1 \nabla g_1|_{\mathbf{x}=\mathbf{x}_o} = \mathbf{0}_2, \quad \mu_1 > 0$$

or

$$\mu_1 \nabla g_1|_{\mathbf{x}=\mathbf{x}_o} = -\nabla f|_{\mathbf{x}=\mathbf{x}_o}$$

which states that, at the SP  $\mathbf{x}_o$  given above, the two gradients,  $\nabla f|_{\mathbf{x}=\mathbf{x}_o}$  and  $\nabla g_1|_{\mathbf{x}=\mathbf{x}_o}$ , are linearly-dependent. As a consequence, the above overdetermined system of two equations in one single unknown,  $\mu_1$ , admits one solution that verifies the two equations. Upon solving this system we obtain  $\mu_1 = 5 > 0$ , thereby verifying the second relation of conditions (7.3c).

### Example 7.2.2

$$f \equiv 8x_1^2 - 8x_1x_2 + 3x_2^2 \rightarrow \min_{\mathbf{x}}$$

subject to

$$x_1 \geq 3, \quad x_2 \leq \frac{3}{2}$$

The objective function and the constraints of this example are depicted in Fig. 7.3

*Solution:* Again, we start by restating the inequalities in our standard form:

$$g_1 \equiv 3 - x_1 \leq 0, \quad g_2 \equiv x_2 - \frac{3}{2} \leq 0$$

Therefore,

$$\nabla g_1|_{\mathbf{x}=\mathbf{x}_o} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad \nabla g_2|_{\mathbf{x}=\mathbf{x}_o} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

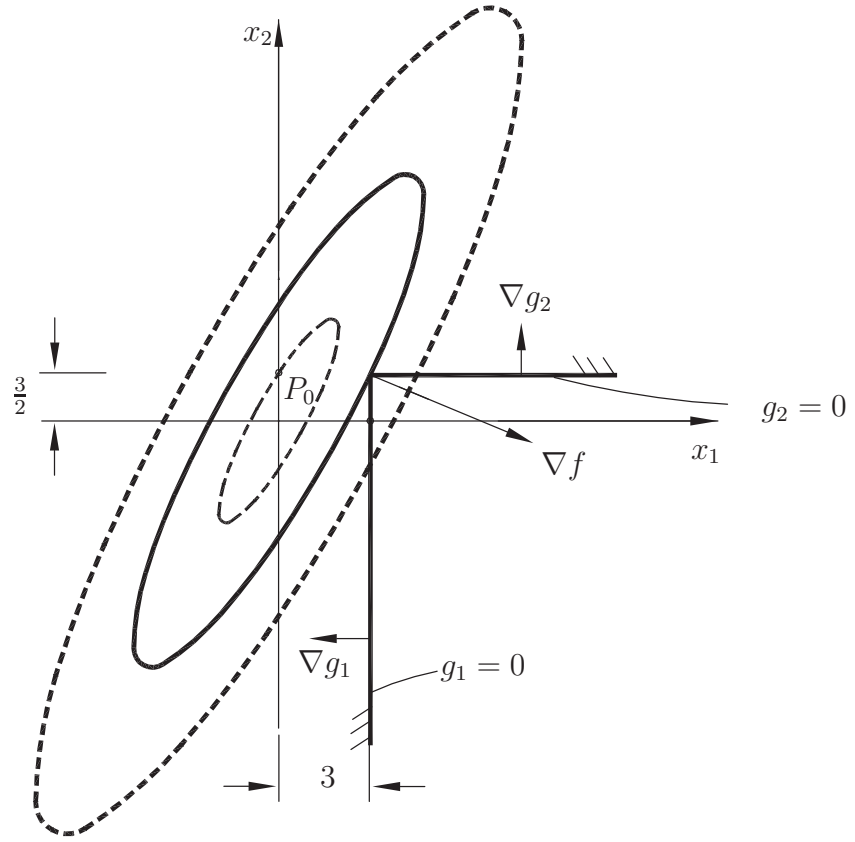


Figure 7.3: One more quadratic objective function subject to linear inequality constraints

condition (7.3b) thus leading to

$$\mu_1 \nabla g_1|_{\mathbf{x}=\mathbf{x}_o} + \mu_2 \nabla g_2|_{\mathbf{x}=\mathbf{x}_o} = -\nabla f|_{\mathbf{x}=\mathbf{x}_o}$$

where, apparently,

$$\mathbf{x}_o = \begin{bmatrix} 3 \\ 3/2 \end{bmatrix}, \quad \nabla f = \begin{bmatrix} 16x_1 - 8x_2 \\ -8x_1 + 6x_2 \end{bmatrix}$$

Hence,

$$\nabla f|_{\mathbf{x}=\mathbf{x}_o} = \begin{bmatrix} 36 \\ -15 \end{bmatrix}$$

The above normality condition thus leading to

$$\mu_1 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + \mu_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -36 \\ 15 \end{bmatrix}$$

which, in this case, turns out to be a determined system of two equations in two

unknowns, its solution being

$$\mu_1 = 36 > 0, \quad \mu_2 = 15 > 0$$

thereby verifying all KKT conditions.

**Example 7.2.3 (A Linear Program)** *A class of optimization problems finding a number of applications involves a linear objective function subject to linear equality and inequality constraints. This class is studied within the realm of linear programming. These problems cannot be solved with the tools described so far, for we have focused on least-square problems, with an extension to more general objective functions and equality constraints. By the same token, linear programs arise seldom in mechanical design. To be true, a family of design problems in structural engineering, known as limit design, pertain to the design of structural elements, beams, columns and plates, for minimum weight, in such a way that all modes of plastic failure are avoided. Problems in limit design lead to linear programs.*

*Linear programming is a first instance of application of the KKT conditions. We illustrate the concept with the problem below.*

$$f \equiv 2x_1 - x_2 \quad \rightarrow \quad \min_{x_1, x_2}$$

*subject to*

$$g_1(\mathbf{x}) \equiv -x_1 \leq 0$$

$$g_2(\mathbf{x}) \equiv -x_2 \leq 0$$

$$g_3(\mathbf{x}) \equiv x_1 + x_2 - 1 \leq 0$$

*Solution:* The gradient of the objective function and the constraints of this problem are illustrated in Fig. 7.4.

In this case,

$$\mathbf{G} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 1 \end{bmatrix}, \quad \nabla f = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \quad \mathbf{x}_o = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

The KKT conditions lead to

$$\begin{bmatrix} 2 \\ -1 \end{bmatrix} + \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$



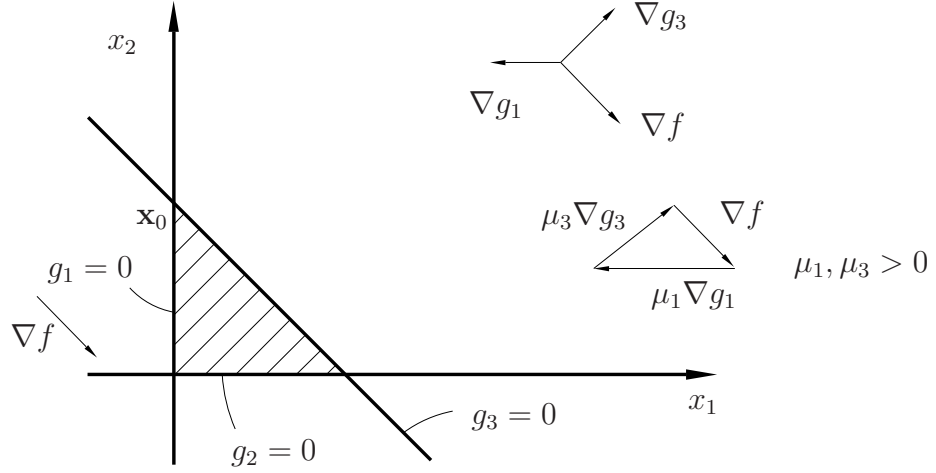


Figure 7.4: A linear program(the feasible region is shown dashed)

Apparently,  $g_1$  and  $g_3$  are active, and hence,

$$\mu_1, \mu_3 > 0, \quad \mu_2 = 0$$

thereby ending up with a system of two equations in two unknowns,  $\mu_1$  and  $\mu_3$ . Upon solving this system, we obtain, successively,

$$\mu_3 = 1 > 0, \quad \mu_1 = 3 > 0$$

thereby verifying the KKT conditions.

### 7.3 Second-Order Normality Conditions

The simplest way of stating the *sufficient* conditions for a minimum, i.e., the second-order normality conditions, is by imposing the condition that, at a stationary point  $\mathbf{x}_o$  within  $\mathcal{F}$ , any *feasible move*  $\Delta\mathbf{x}_F$  will produce a *growth*, and hence a *worsening*, of the objective function, while respecting the constraints. In other words, at a feasible minimum, we cannot decrease the objective function without violating the constraints. That is,

$$\Delta f \equiv f(\mathbf{x}_o + \Delta\mathbf{x}_F) - f(\mathbf{x}_o) > 0 \quad (7.9)$$

$$\Delta \mathbf{h} \equiv \mathbf{h}(\mathbf{x}_o + \Delta\mathbf{x}_F) - \mathbf{h}(\mathbf{x}_o) = \mathbf{0}_l \quad (7.10)$$

$$\mathbf{g}(\mathbf{x}_o + \Delta\mathbf{x}_F) \leq \mathbf{0}_p \quad (7.11)$$

A feasible move, moreover, is to be understood here at the first-order approximation of the objective function and the constraint functions  $\mathbf{g}(\mathbf{x})$  and  $\mathbf{h}(\mathbf{x})$ . In this light, then, eq.(7.10) constrains  $\Delta\mathbf{x}_F$  to lie in the null space of  $\mathbf{J}$ , i.e.,

$$\mathbf{J}\Delta\mathbf{x}_F = \mathbf{0}_l \quad (7.12)$$

If we recall now the partitioning of  $\mathbf{g}(\mathbf{x})$  introduced in eq.(7.6), relation (7.11) can be correspondingly partitioned as

$$\Delta\mathbf{g}_a(\mathbf{x}) \equiv \mathbf{g}_a(\mathbf{x}_0 + \Delta\mathbf{x}_F) - \mathbf{g}_a(\mathbf{x}_0) = \mathbf{0}_a \quad (7.13a)$$

$$\mathbf{g}_{p'}(\mathbf{x}) \equiv \mathbf{g}_{p'}(\mathbf{x}_0 + \Delta\mathbf{x}_F) < \mathbf{0}_{p'} \quad (7.13b)$$

the feasible move thus requiring that the passive constraints remain passive and that their active counterparts remain active. If we now let  $\mathbf{G}_a$  denote  $\nabla\mathbf{g}_a$ , the first-order approximation of eq.(7.13a) leads to

$$\mathbf{G}_a(\mathbf{x})\Delta\mathbf{x}_F = \mathbf{0}_a \quad (7.14)$$

and we need not worry about the passive constraints (7.13b), which will be respected as long as  $\|\Delta\mathbf{x}_F\|$  is “small enough.” Now we can adjoin eq.(7.14) to eq.(7.12) in the form

$$\bar{\mathbf{J}}\Delta\mathbf{x}_F = \mathbf{0}_{l+a}, \quad \bar{\mathbf{J}} \equiv \begin{bmatrix} \mathbf{J} \\ \mathbf{G}_a \end{bmatrix} \quad (7.15)$$

Further, we introduce, correspondingly, a  $n \times (n - l - a)$  orthogonal complement  $\bar{\mathbf{L}}$  of  $\bar{\mathbf{J}}$ , i.e.,

$$\bar{\mathbf{J}} \bar{\mathbf{L}} = \mathbf{0}_{(l+a) \times (n-l-a)} \quad (7.16)$$

The FONC can now be restated as

$$\nabla f + \bar{\mathbf{J}}^T \bar{\boldsymbol{\lambda}} = \mathbf{0}_n, \quad \bar{\boldsymbol{\lambda}} \in \mathbb{R}^{l+a}, \quad \bar{\boldsymbol{\mu}}_{p'} = \mathbf{0}_{p'} \quad (7.17)$$

where  $\bar{\boldsymbol{\mu}}_{p'}$  is the passive-constraint counterpart of  $\bar{\boldsymbol{\lambda}}$ .

The *sufficient* second-order normality conditions can now be stated in exactly the same form as for equality-constrained problems. That is, we define now the  $(n - l - a) \times (n - l - a)$  *reduced Hessian* as

$$\bar{\mathbf{H}}_r \equiv \bar{\mathbf{L}}^T \left[ \nabla \nabla f + \frac{\partial(\bar{\mathbf{J}}^T \bar{\boldsymbol{\lambda}})}{\partial \mathbf{x}} \right] \bar{\mathbf{L}} \quad (7.18)$$

and hence, the sufficient SONC can be stated as:

*A stationary point of an inequality-constrained problem is a minimum if the reduced Hessian, defined in eq.(7.18), is positive-definite.*

### 7.3.1 Overconstrained Problems

In the special case in which  $l + a > n$ , the orthogonal complement  $\bar{\mathbf{L}}$  does not exist for a full-rank  $\bar{\mathbf{J}}$ . In this case, the problem is overconstrained, and hence, ill-defined. However, if  $\bar{\mathbf{J}}$  is rank-deficient, of rank smaller than  $n$ , then some of the constraints are redundant, at least to a first order, and the problem may admit a minimum. If  $l + a = n$ , then the equality constraints yield a determined system of nonlinear equations, which can be solved using the Newton-Raphson method. Any of the solutions thus obtained is a solution candidate for the original optimization problem.

**Example 7.3.1 The KKT and Second-Order Conditions** *Consider the problem*

$$f(\mathbf{x}) = (x_1 - 1)^2 + (x_2 + 2)^2 \rightarrow \min_{x_1, x_2}$$

*subject to*

$$g(\mathbf{x}) = x_1 - x_2 + 0.5 \leq 0$$

*with  $f(\mathbf{x})$  and  $g(\mathbf{x})$  illustrated in Fig. 7.5*

*Solution:* We start by finding a feasible stationary point  $\mathbf{x}_{l.c.o}$  via the KKT conditions:

$$\nabla f + \mu \nabla g = 0$$

where

$$\nabla f = 2 \begin{bmatrix} x_1 - 1 \\ x_2 + 2 \end{bmatrix}, \quad \nabla g = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Hence, the KKT conditions lead to

$$2 \begin{bmatrix} x_1 - 1 \\ x_2 + 2 \end{bmatrix} + \mu \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Next, we investigate whether  $g(\mathbf{x}) \leq 0$  is active. As the unconstrained problem has a minimum at  $[1, -2]^T$ , the centre of the family of circles generated by the objective function, which violates the inequality constraint, this must be active. Hence, the above equations yield, with  $\mu > 0$ ,

$$2(x_1 - 1) + \mu = 0 \quad \text{and} \quad 2(x_2 + 2) - \mu = 0$$

which, upon summation, lead to

$$x_1 + x_2 = -1$$

while the inequality constraint, written as an active constraint, leads in turn to

$$x_1 - x_2 = -0.5$$

The solution of the two foregoing equations is  $x_1 = -0.25$ ,  $x_2 = -0.75$ , whence  $\mu = 2.5 > 0$  and hence, the KKT conditions are verified. Now we turn to the second-order sufficient conditions: In our case,

$$\bar{\mathbf{J}} \equiv [(\nabla g_a)^T] = [1 \quad -1] \Rightarrow \bar{\mathbf{L}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

and

$$\nabla \nabla f = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Moreover, as  $\bar{\mathbf{J}}$  is constant because the constraint is linear,

$$\bar{\mathbf{H}}_u = \frac{\partial(\bar{\mathbf{J}}^T \bar{\boldsymbol{\lambda}})}{\partial \mathbf{x}} = \mathbf{O}_{22}$$

Therefore,

$$\bar{\nabla} \bar{\nabla} f = [1 \quad 1] \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 4 > 0$$

thereby verifying the second-order sufficient condition, and hence showing that, indeed,  $\mathbf{x}_{l.c.o} = [-0.25 \quad -0.75]^T$  is a minimum.

Notice that, in this case,  $\nabla \nabla f$  happens to be positive-definite, and hence, the reduced Hessian is bound to be positive-definite as well. The computation of  $\bar{\mathbf{H}}_u$  in this case could thus have been dispensed with.

## 7.4 Methods of Solution

Two classes of methods are available to solve inequality-constrained problems: a) *direct methods*, which handle the inequalities as such, and b) *indirect methods*, which transform the problem into one of two types, either unconstrained or equality-constrained. Indirect methods being simpler to implement, we will focus on these, which we will study first. Direct methods will be outlined at the end of the chapter.

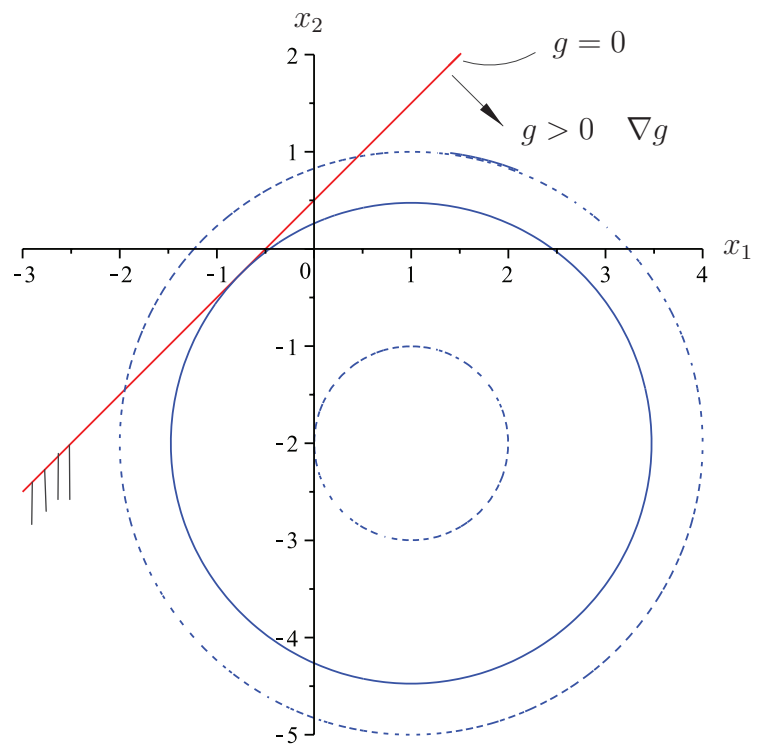


Figure 7.5: A quadratic objective function to be minimized under one inequality constraint

Inequality-constrained problems can be solved using the approach introduced for either unconstrained or equality-constrained problems, upon converting the problem at hand into an unconstrained or, correspondingly, an equality-constrained problem. This can be done by various methods; we focus on two, namely, slack variables and penalty functions.

## 7.5 Indirect Methods

In this section, the methods of *slack variables* and of *penalty functions* are discussed.

### 7.5.1 Slack Variables

Upon introducing the *slack variables*  $s_1, s_2, \dots, s_p$  into inequalities (7.1b), we convert these inequalities into equality constraints, namely,

$$\gamma(\mathbf{x}, \mathbf{s}) \equiv \begin{bmatrix} g_1 + s_1^2 \\ g_2 + s_2^2 \\ \vdots \\ g_p + s_p^2 \end{bmatrix} = \mathbf{0}, \quad \mathbf{x} \equiv \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{s} \equiv \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_p \end{bmatrix} \quad (7.19)$$

Notice that the slack variables being unknown, they have to be treated as additional design variables, the dimension of the design space being correspondingly increased. In consequence, the design vector is now of dimension  $n + p$ , i.e.,

$$\boldsymbol{\xi} \equiv \begin{bmatrix} \mathbf{x} \\ \mathbf{s} \end{bmatrix} \quad (7.20)$$

Further, the gradient of the objective function with respect to the new design-variable vector takes the form

$$\nabla_{\boldsymbol{\xi}} f \equiv \begin{bmatrix} \nabla f \\ \nabla_{\mathbf{s}} f \end{bmatrix} \quad (7.21a)$$

where

$$\nabla f \equiv \frac{\partial f}{\partial \mathbf{x}}, \quad \nabla_{\mathbf{s}} f \equiv \frac{\partial f}{\partial \mathbf{s}} = \mathbf{0}_p \quad (7.21b)$$

the second relation following because the slack variables do not appear explicitly in the objective function.

Likewise, the Hessian with respect to the new design-variable vector  $\boldsymbol{\xi}$  takes the form

$$\nabla_{\boldsymbol{\xi}} \nabla_{\boldsymbol{\xi}} f = \begin{bmatrix} \nabla \nabla f & \nabla(\nabla_{\mathbf{s}} f) \\ \nabla_{\mathbf{s}}(\nabla f) & \nabla_{\mathbf{s}} \nabla_{\mathbf{s}} f \end{bmatrix} \quad (7.22a)$$

with the notation

$$\nabla(\nabla_{\mathbf{s}}f) \equiv \frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{s}} = \left( \frac{\partial^2 f}{\partial \mathbf{s} \partial \mathbf{x}} \right)^T \equiv [\nabla_{\mathbf{s}}(\nabla f)]^T, \quad \nabla_{\mathbf{s}}\nabla_{\mathbf{s}}f \equiv \frac{\partial^2 f}{\partial \mathbf{s}^2} \quad (7.22b)$$

However, since  $\nabla_{\mathbf{s}}f = \mathbf{0}_p$ , the above Hessian expression reduces to

$$\nabla_{\boldsymbol{\xi}}\nabla_{\boldsymbol{\xi}}f = \begin{bmatrix} \nabla\nabla f & \mathbf{O}_{np} \\ \mathbf{O}_{np}^T & \mathbf{O}_p \end{bmatrix} \quad (7.23)$$

That is, the Hessian of the objective function with respect to the new design-variable vector  $\boldsymbol{\xi}$  is singular. In case  $\nabla\nabla f$  is positive-definite,  $\nabla_{\boldsymbol{\xi}}\nabla_{\boldsymbol{\xi}}f$  is positive-semidefinite. Hence, in applying the method of slack variables to solve inequality-constrained problems, Hessian stabilization—see Section 6.4.1—will always be needed.

Now, the problem can be formulated as equality-constrained, if we adjoin the  $p$  equality constraints (7.19) to the original  $l$ , thereby obtaining a new set of equality constraints:

$$\mathbf{h}(\boldsymbol{\xi}) = \mathbf{0}_{l+p} \quad (7.24)$$

Therefore, the problem at hand can be solved using the ODA.

**Example 7.5.1 (Minimization of the design error of a four-bar linkage with an input crank)** Determine the link-lengths of the four-bar linkage shown in Fig. 7.6, that will produce the set of input-output pairs  $\{\psi_i, \phi_i\}_1^{10}$  of Table 7.1 with the least-square error, and such that its input link have a full rotatability. In Table 7.1,  $\psi$  and  $\phi$  denote the input and output angles, respectively.

Table 7.1: The input-output pairs of  $\{\psi_i, \phi_i\}_1^{10}$

$i$	1	2	3	4	5
$\psi_i$	123.8668°	130.5335°	137.2001°	143.8668°	150.5335°
$\phi_i$	91.7157°	91.9935°	92.8268°	94.2157°	96.1601°
$i$	6	7	8	9	10
$\psi_i$	157.2001°	163.8668°	170.5335°	177.2001°	183.8668°
$\phi_i$	98.6601°	101.7157°	105.3268°	109.4935°	114.2157°

*Solution:* The link-lengths are obtained via the *Freudenstein parameters*  $k_1$ ,  $k_2$  and  $k_3$ , defined as

$$k_1 = \frac{a_1^2 + a_2^2 - a_3^2 + a_4^2}{2a_2a_4}, \quad k_2 = \frac{a_1}{a_2}, \quad k_3 = \frac{a_1}{a_4} \quad (7.25a)$$

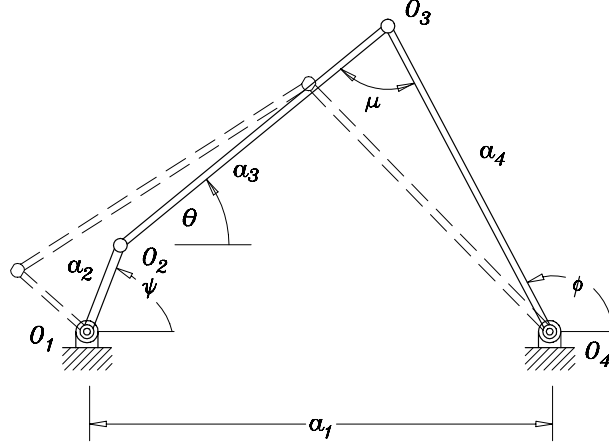


Figure 7.6: A four-bar linkage

with the inverse relations

$$a_2 = \frac{a_1}{k_1}, \quad a_3 = \frac{\sqrt{k_2^2 + k_3^2 + k_2^2 k_3^2 - 2k_1 k_2 k_3}}{|k_2 k_3|}, \quad a_4 = \frac{a_1}{k_3} \quad (7.25b)$$

for a given value of  $a_1$ . The synthesis equations for the planar four-bar linkage can be written in the form (Liu and Angeles, 1992):

$$\mathbf{S}\mathbf{k} = \mathbf{b} \quad (7.26)$$

where  $\mathbf{S}$  is the *synthesis matrix*, and  $\mathbf{k}$  is the vector of linkage parameters. Moreover,  $\mathbf{S}$ ,  $\mathbf{k}$  and  $\mathbf{b}$  are defined as

$$\mathbf{S} = \begin{bmatrix} 1 & \cos \phi_1 & -\cos \psi_1 \\ 1 & \cos \phi_2 & -\cos \psi_2 \\ \vdots & \vdots & \vdots \\ 1 & \cos \phi_{10} & -\cos \psi_{10} \end{bmatrix} \quad \mathbf{k} = \begin{bmatrix} k_1 \\ k_2 \\ k_3 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} \cos(\psi_1 - \phi_1) \\ \cos(\psi_2 - \phi_2) \\ \vdots \\ \cos(\psi_{10} - \phi_{10}) \end{bmatrix} \quad (7.27)$$

The *design error* is defined in turn as

$$\mathbf{d} \equiv \mathbf{b} - \mathbf{S}\mathbf{k} \quad (7.28)$$

the purpose of the optimization exercise being to minimize the Euclidean norm of the design error, while ensuring that its input link is a crank. The *conditions for full mobility* of the input link were reported by Liu and Angeles (1992)

$$\begin{aligned} g_1(\mathbf{x}) &= (k_1 + k_3)^2 - (1 + k_2)^2 < 0 \\ g_2(\mathbf{x}) &= (k_1 - k_3)^2 - (1 - k_2)^2 < 0 \end{aligned}$$



By introducing two slack-variables  $s_1$  and  $s_2$ , the inequality constraints are converted into equality constraints, i.e.,

$$h_1(\mathbf{x}) = (k_1 + k_3)^2 - (1 + k_2)^2 + s_1^2 = 0 \quad (7.29a)$$

$$h_2(\mathbf{x}) = (k_1 - k_3)^2 - (1 - k_2)^2 + s_2^2 = 0 \quad (7.29b)$$

The design vector  $\boldsymbol{\xi}$  thus becomes  $\boldsymbol{\xi} = [k_1 \ k_2 \ k_3 \ s_1 \ s_2]^T$ . From the initial guess  $\boldsymbol{\xi}_0 = [0.28 \ 0.74 \ 0.12 \ 1.69 \ 1.2]^T$ , the solution was found to be  $\boldsymbol{\xi}_{\text{opt}} = [0.3248 \ 0.5875 \ -0.009725 \ 1.556 \ 0.2415]^T$ , and the corresponding link lengths are  $a_1 \equiv 1$ ,  $a_2 = 1.702$ ,  $a_3 = 103.4$  and  $a_4 = 102.8$ . The Euclidean norm of the minimum design error is  $5 \times 10^{-2}$ .

The problem with this design is that it leads to a quite disproportionate linkage: two of its links have lengths two orders of magnitude bigger than those of the other two!

In the foregoing reference, a technique is introduced to eliminate this *dimensional unbalance* by means of a *penalty function*.

## 7.5.2 Penalty Functions

Penalty-function methods are based on a transformation of the problem under study into an unconstrained one, which is done by suitably modifying the objective function. The modification consists in adding one *penalty term* that penalizes any violation of the constraints—both equality and inequality constraints can be handled in this way, but we limit the discussion here to inequality constraints. The iterative procedure approaches the optimum solution *asymptotically*, by extrapolation of a sequence of optimum solutions to unconstrained problems. There are two possibilities: the solution is approached either *within* the feasible region or from *without*, the penalty function being correspondingly referred to as *interior* or *exterior*. It is noteworthy that exterior penalty-function methods are applicable *only* to problems whereby the optimum finds itself at the boundary of the feasible region, but misses interior optima. Hence, we focus here on interior penalty functions.

### Interior Penalty Functions

Given an objective function  $f(\mathbf{x})$  subject to inequality constraints, as defined in eq.(7.1b), a sequence of interior penalty functions  $\{\phi_k\}_1^\nu$  is constructed as

$$\phi_k(\mathbf{x}; r_k) \equiv f(\mathbf{x}) - r_k \sum_{i=1}^p \frac{1}{g_i(\mathbf{x})} \quad k = 1, 2, \dots, \nu \quad (7.30)$$

where the term  $-r_k \sum_{i=1}^p [1/g_i(\mathbf{x})]$  is called the *penalty term*, and all the  $r_k$  factors are positive and observe a decreasing order, i.e.,

$$r_1 > r_2 > r_3 \cdots > r_\nu > 0 \quad (7.31)$$

The idea here is that the search for the minimum is conducted within the feasible region. Under these conditions, the summation in the penalty term remains negative, and hence, a *positive* penalization is always *added* to the objective function. As the design-variable vector approaches the constraint  $g_i(\mathbf{x}) = 0$ , it does so *from the left*, i.e.,  $g_i(\mathbf{x}) \rightarrow 0^-$ , and  $1/g_i(\mathbf{x}) \rightarrow -\infty$ , the penalty term thus becoming a “large” positive quantity, whose value is kept finite thanks to the presence of the “small” factor  $r_k$ .

Now, a sequence of unconstrained minimization problems is defined:

$$\phi_k(\mathbf{x}; r_k) \equiv f(\mathbf{x}) - r_k \sum_{i=1}^p \frac{1}{g_i(\mathbf{x})} \rightarrow \min_{\mathbf{x}}, \quad k = 1, 2, \dots, \nu \quad (7.32)$$

Let  $\mathbf{x}_o^1, \mathbf{x}_o^2, \dots, \mathbf{x}_o^\nu$  be the sequence of corresponding unconstrained minima. Next, these minima are interpolated to a vector function  $\mathbf{x}_o(r)$ :

$$\mathbf{x}_o(r) \equiv \mathbf{c}_0 + \sum_{k=1}^{\nu-1} \mathbf{c}_k r^{k/2} \quad (7.33)$$

thereby obtaining a system of  $\nu n$  equations in  $\nu n$  unknowns, the  $n$  components of the  $\nu$  unknown vector coefficients  $\{\mathbf{c}_k\}_0^{\nu-1}$ . Note that the foregoing equations are all linear in the unknowns, and hence, they can be solved for the unknowns using Gaussian elimination, as described below. First, eq.(7.33) is written for  $r = r_i$ , with  $i = 1, 2, \dots, \nu$ :

$$\mathbf{x}_o(r_i) \equiv \mathbf{c}_0 + \mathbf{c}_1 r_i^{1/2} + \mathbf{c}_2 r_i^{2/2} + \cdots + \mathbf{c}_{\nu-1} r_i^{(\nu-1)/2} \quad (7.34)$$

or

$$\mathbf{x}_o(r_i) \equiv [\mathbf{c}_0 \quad \mathbf{c}_1 \quad \cdots \quad \mathbf{c}_{\nu-1}] \begin{bmatrix} 1 \\ r_i^{1/2} \\ \vdots \\ r_i^{(\nu-1)/2} \end{bmatrix}, \quad i = 1, 2, \dots, \nu \quad (7.35)$$

In the next step, we regroup all  $\nu$  vector equations above to produce a matrix equation. To this end, we define the matrices

$$\mathbf{R} \equiv \begin{bmatrix} 1 & 1 & \cdots & 1 \\ r_1^{1/2} & r_2^{1/2} & \cdots & r_\nu^{1/2} \\ \vdots & \vdots & \ddots & \vdots \\ r_1^{(\nu-1)/2} & r_2^{(\nu-1)/2} & \cdots & r_\nu^{(\nu-1)/2} \end{bmatrix} \quad (7.36a)$$

$$\mathbf{X}_o \equiv [\mathbf{x}_o(r_1) \quad \mathbf{x}_o(r_2) \quad \cdots \quad \mathbf{x}_o(r_\nu)] \quad (7.36b)$$

$$\mathbf{C} \equiv [\mathbf{c}_0 \quad \mathbf{c}_1 \quad \cdots \quad \mathbf{c}_{\nu-1}] \quad (7.36c)$$

It is noteworthy that square matrices with the gestalt of  $\mathbf{R}$  of eq.(7.36a) occur quite frequently in system theory, where they are termed *Vandermonde matrices*. For this reason, scientific code includes commands that ease the construction of such matrices. For example, Maple includes the command

$$\text{VandermondeMatrix}(\mathbf{r}, \text{output})$$

in which  $\mathbf{r}$  is the array  $[r_1, r_2, \dots, r_\nu]^T$ , and **output** is the name assigned to the matrix thus constructed. The command includes various options.

Thus, the  $\nu$  vector equations (7.34) become, in matrix form,

$$\mathbf{C}\mathbf{R} = \mathbf{X}_o \quad (7.37a)$$

whence,

$$\mathbf{C} = \mathbf{X}_o \mathbf{R}^{-1} \quad (7.37b)$$

or, if eq.(7.37a) is written in the usual form, with the unknown matrix  $\mathbf{C}$  to the right of its matrix coefficient, the foregoing equation should first be transposed, the result then being

$$\mathbf{C}^T = \mathbf{R}^{-T} \mathbf{X}_o^T \quad (7.37c)$$

with exponent  $-T$  indicating the inverse of the transpose or, equivalently, the transpose of the inverse. Once the  $\nu$  vector coefficients sought are available, the optimum of the inequality-constrained problem,  $\mathbf{x}_{\text{opt}}$ , is calculated as

$$\mathbf{x}_{\text{opt}} = \lim_{r \rightarrow 0} \mathbf{x}(r)$$

i.e.,

$$\mathbf{x}_{\text{opt}} = \mathbf{c}_0 \quad (7.38)$$

In computing the above value, note that  $\mathbf{c}_0$  is the first column of the unknown matrix  $\mathbf{C}$  or, equivalently, the first row of its transpose. In either case, it is not possible to obtain  $\mathbf{c}_0$  as the solution of one single vector equation. A matrix equation must be solved in order to obtain  $\mathbf{c}_0$ . Such an equation is to be solved as a sequence of linear systems using LU-decomposition, one column of the matrix at a time.

**Example 7.5.2 (A Two-dimensional Optimization Problem Subject to Inequality Constraints)**

Consider an optimization problem with an objective function defined as

$$f = x^2 + 2y^2 \rightarrow \min_{x,y} \quad (7.39)$$

subject to the inequality constraints

$$g_1 \equiv -x \leq 0 \quad (7.40a)$$

$$g_2 \equiv -y \leq 0 \quad (7.40b)$$

$$g_3 \equiv 1 - x - y \leq 0 \quad (7.40c)$$

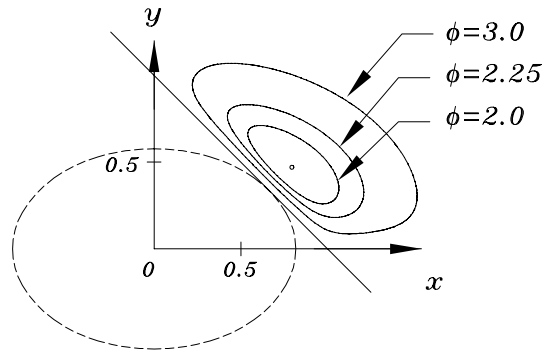


Figure 7.7: Isocontours of the penalty function with  $r_1 = 0.1$

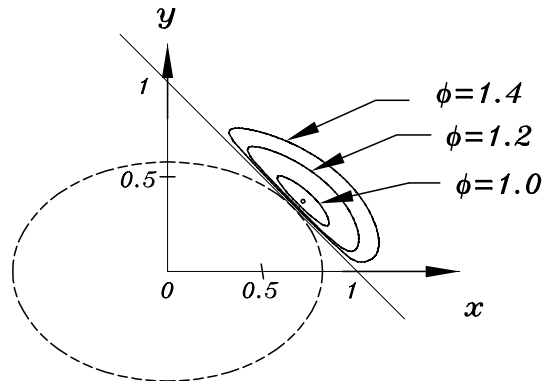


Figure 7.8: Isocontours of the penalty function with  $r_2 = 0.01$

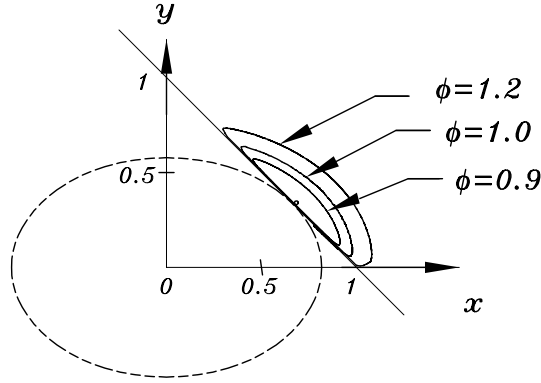


Figure 7.9: Isocontours of the penalty function with  $r_3 = 0.001$

*Solution:* From a sketch of the constraints and the contours of the objective function in the  $x$ - $y$  plane, it should be apparent that the minimum of  $f$  is attained at a point where the gradient  $\nabla f$  is parallel to the normal to the line  $g_3(x, y) = 0$ . The optimum values of  $x$  and  $y$  are, then

$$x_{\text{opt}} = \frac{2}{3}, \quad y_{\text{opt}} = \frac{1}{3}$$

We demonstrate below the application of penalty functions to obtain the foregoing optimum. We have

$$\phi_k \equiv x^2 + 2y^2 - r_k \left( -\frac{1}{x} - \frac{1}{y} - \frac{1}{x+y-1} \right), \quad k = 1, \dots, 3 \quad \rightarrow \quad \min_{x,y}$$

subject to no constraints, for

$$r_1 = 0.1, \quad r_2 = 0.01, \quad r_3 = 0.001$$

The penalty-function isocontours for different  $r_k$  values are shown in Figs. 7.7–7.9. In those figures, the isocontour of the objective function  $f$  that includes the constrained minimum is indicated with a dashed curve.

The optima  $\mathbf{x}_o(r_k) \equiv [x_o(r_k), y_o(r_k)]^T$  for the three given values of  $r_k$  were found by the ODA, using subroutine **ARBITRARY**, as

$$\mathbf{x}_o(r_1) = \begin{bmatrix} 0.7941 \\ 0.4704 \end{bmatrix}, \quad \mathbf{x}_o(r_2) = \begin{bmatrix} 0.7140 \\ 0.3703 \end{bmatrix}, \quad \mathbf{x}_o(r_3) = \begin{bmatrix} 0.6836 \\ 0.3434 \end{bmatrix} \quad (7.41)$$

We now fit the values of  $\{\mathbf{x}_o(r_k)\}_1^3$  to the function

$$\mathbf{x}_o(r) = \mathbf{c}_0 + \mathbf{c}_1 r^{1/2} + \mathbf{c}_2 r$$

We thus have

$$\mathbf{R} = \begin{bmatrix} 1 & 1 & 1 \\ 0.3163 & 0.1000 & 0.03163 \\ 0.1000 & 0.0100 & 0.00100 \end{bmatrix}, \quad \mathbf{X}_o = \begin{bmatrix} 0.7941 & 0.7140 & 0.6836 \\ 0.4704 & 0.3703 & 0.3434 \end{bmatrix} \quad (7.42)$$

The coefficient matrix  $\mathbf{C}$  is thus found to be

$$\mathbf{C} = \mathbf{X}_o \mathbf{R}^{-1} = \begin{bmatrix} 0.6687 & 0.4790 & -0.2605 \\ 0.3317 & 0.3612 & 0.2443 \end{bmatrix} \quad (7.43)$$

Therefore,

$$\mathbf{x}_{\text{opt}} = \mathbf{c}_0 = \begin{bmatrix} 0.6687 \\ 0.3317 \end{bmatrix} \quad (7.44)$$

which yields the optimum with two significant digits of accuracy.

## 7.6 Direct Methods

Of the various direct methods for the solution of inequality-constrained problems, we shall discuss here three:

- (i) The generalized reduced-gradient method;
- (ii) the complex method; and
- (iii) the method of the feasible directions.

Of these, the first two are described below, the third method, due to Zoutendijk (1960), to be included in a future edition.

### 7.6.1 The Generalized Reduced-Gradient Method

This method, abbreviated as the GRG method, is an evolution of the *gradient-projection method* proposed by Rosen (1960). Further developments led to the *reduced-gradient method*, as applicable to arbitrary objective functions with linear equality constraints and inequalities of the form  $\mathbf{x} \geq \mathbf{0}$ . The generalization of the reduced-gradient method lies in its applicability to nonlinear equality and inequality constraints.

The method is best described if the given problem is formulated in a format slightly different from the one we have used so far: Given the objective function  $f(\mathbf{x})$  and the  $C^1$ -continuous functions  $h_j(\mathbf{x})$ , for  $j = 1, \dots, l + p$ ,

$$f(\mathbf{x}) \rightarrow \min_{\mathbf{x} \in \mathbb{R}^n} \quad (7.45a)$$

subject to

$$h_j(\mathbf{x}) \leq 0, \quad j = 1, \dots, p \quad (7.45b)$$

$$h_{p+j}(\mathbf{x}) = 0, \quad j = 1, \dots, l \quad (7.45c)$$

$$l_i \leq x_i \leq u_i, \quad i = 1, \dots, n \quad (7.45d)$$

The problem is first reformulated upon elimination of the first  $p$  inequalities by means of *nonnegative slack variables*  $x_{n+1}, \dots, x_{n+p}$ . The constraints become now:

$$h_j(\mathbf{x}) + x_{n+j} = 0, \quad j = 1, \dots, p \quad (7.46a)$$

$$h_{p+j}(\mathbf{x}) = 0, \quad j = 1, \dots, l < n \quad (7.46b)$$

$$l_i \leq x_i \leq u_i, \quad i = 1, \dots, n \quad (7.46c)$$

$$x_{n+j} \geq 0, \quad j = 1, \dots, p \quad (7.46d)$$

Notice that the slack variables introduced in eq.(7.46a) are *non-negative*, while those of Subsection 7.5.1 are *quadratic*. The reason for the difference is that the latter were introduced in the framework of *least squares*; the former have a historic origin, greatly influenced by the *simplex method* of *linear programming*, whereby all decision, or design, variables are regarded as non-negative.

We thus end up with a new problem:

$$f(\mathbf{x}) \rightarrow \min_{\mathbf{x} \in \mathbb{R}^p} \quad (7.47a)$$

subject to

$$h_j(\mathbf{x}) = 0, \quad j = 1, \dots, l + p \quad (7.47b)$$

$$l_i \leq x_i \leq u_i, \quad i = 1, \dots, p + n \quad (7.47c)$$

Inequalities (7.47c) will be termed *bilateral*.

**Remarks:**

- $l_i = 0$ ,  $u_i < \infty$ ,  $i = n + 1, \dots, p$ , the latter meaning that the last  $p$  components of  $\mathbf{x}$  are unbounded from above
- In light of eqs.(7.45c), the problem at hand has  $n' \equiv (n - l)$  degrees of freedom, and hence,  $(n - l)$  design variables—or a combination thereof—can be freely prescribed.

**Strategy:** *Partition* the set of design variables into two sets:  $(n - l)$  independent design variables and  $(l + p)$  dependent design variables, a.k.a. state variables. Relabel the design variables, if necessary.

**Definitions:**

$$\mathbf{x} \equiv \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix}, \quad \mathbf{y} \equiv \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-l} \end{bmatrix}, \quad \mathbf{z} \equiv \begin{bmatrix} x_{n-l+1} \\ x_{n-l+2} \\ \vdots \\ x_{p+n} \end{bmatrix}, \quad \mathbf{h} \equiv \begin{bmatrix} h_1(\mathbf{x}) \\ h_2(\mathbf{x}) \\ \vdots \\ h_{l+p} \end{bmatrix} \quad (7.48)$$

- $\mathbf{y}$ :  $(n - l)$ -dimensional vector of independent variables
- $\mathbf{z}$ :  $(l + p)$ -dimensional vector of dependent variables or state variables
- The partial derivatives  $\partial f / \partial \mathbf{y}$ ,  $\partial f / \partial \mathbf{z}$ ,  $\partial \mathbf{h} / \partial \mathbf{y}$ , and  $\partial \mathbf{h} / \partial \mathbf{z}$  denote derivatives that **do not take into account** the dependence of  $\mathbf{z}$  on  $\mathbf{y}$
- $df/d\mathbf{y}$  denotes the *total derivative* of  $f$  with respect to  $\mathbf{y}$ , which takes into account the dependent variables, and is, hence, a  $(n - l)$ -dimensional vector
- $d\mathbf{h}/d\mathbf{y}$  denotes the *total derivative* of  $\mathbf{h}$  with respect to  $\mathbf{y}$ , which takes into account the dependent variables, and is, hence, a  $(l + p) \times (n - l)$  matrix

That is,

$$\frac{df}{d\mathbf{y}} \equiv \begin{bmatrix} df/dx_1 \\ df/dx_2 \\ \vdots \\ df/dx_{n-l} \end{bmatrix}, \quad \frac{d\mathbf{h}}{d\mathbf{y}} \equiv \begin{bmatrix} \frac{d\mathbf{h}}{dx_1} & \frac{d\mathbf{h}}{dx_2} & \cdots & \frac{d\mathbf{h}}{dx_{n-l}} \end{bmatrix} \quad (7.49)$$

**Remark:** Because of the constraints (7.47b), the partial derivatives with respect to  $\mathbf{z}$  are dependent upon those with respect to  $\mathbf{y}$ . Indeed, since the equalities (7.47b) must hold, we must have

$$\mathbf{h}(\mathbf{y}, \mathbf{z}(\mathbf{y})) = \mathbf{0} \quad (7.50a)$$

which means that the total derivative of  $\mathbf{h}$  with respect to  $\mathbf{y}$  must vanish. From the *chain rule*,

$$\frac{d\mathbf{h}}{d\mathbf{y}} \equiv \frac{\partial \mathbf{h}}{\partial \mathbf{y}} + \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{y}} = \mathbf{0} \quad (7.50b)$$

Let

$$\mathbf{C} \equiv \frac{\partial \mathbf{h}}{\partial \mathbf{y}} = \begin{bmatrix} \partial h_1 / \partial x_1 & \partial h_1 / \partial x_2 & \cdots & \partial h_1 / \partial x_{n-l} \\ \partial h_2 / \partial x_1 & \partial h_2 / \partial x_2 & \cdots & \partial h_2 / \partial x_{n-l} \\ \vdots & \vdots & \ddots & \vdots \\ \partial h_{l+p} / \partial x_1 & \partial h_{l+p} / \partial x_2 & \cdots & \partial h_{l+p} / \partial x_{n-l} \end{bmatrix} \quad (7.50c)$$



$$\mathbf{D} \equiv \frac{\partial \mathbf{h}}{\partial \mathbf{z}} = \begin{bmatrix} \partial h_1 / \partial z_1 & \partial h_1 / \partial z_2 & \cdots & \partial h_1 / \partial z_{l+p} \\ \partial h_2 / \partial z_1 & \partial h_2 / \partial z_2 & \cdots & \partial h_2 / \partial z_{l+p} \\ \vdots & \vdots & \ddots & \vdots \\ \partial h_{l+p} / \partial z_1 & \partial h_{l+p} / \partial z_2 & \cdots & \partial h_{l+p} / \partial z_{l+p} \end{bmatrix} \quad (7.50d)$$

whence  $\mathbf{C}$  is a  $(l+p) \times (n-l)$  matrix, while  $\mathbf{D}$  is a  $(l+p) \times (l+p)$  matrix. We thus have

$$\mathbf{C} + \mathbf{D} \frac{\partial \mathbf{z}}{\partial \mathbf{y}} = \mathbf{0}$$

Solving for  $\partial \mathbf{z} / \partial \mathbf{y}$  from the above equation yields, for  $\mathbf{D}$  invertible,

$$\frac{\partial \mathbf{z}}{\partial \mathbf{y}} = -\mathbf{D}^{-1} \mathbf{C} \quad (7.51)$$

Also notice that

$$\frac{\partial f}{\partial \mathbf{x}} \equiv \begin{bmatrix} \partial f / \partial \mathbf{y} \\ \partial f / \partial \mathbf{z} \end{bmatrix} \quad (7.52)$$

By application of the *chain rule*, the *total derivative* of  $f$  with respect to  $\mathbf{y}$  is given by

$$\frac{df}{d\mathbf{y}} = \frac{\partial f}{\partial \mathbf{y}} + \left( \frac{\partial \mathbf{z}}{\partial \mathbf{y}} \right)^T \frac{\partial f}{\partial \mathbf{z}}$$

Substitution of eq.(7.51) into the above equation leads to

$$\begin{aligned} \frac{df}{d\mathbf{y}} &= \frac{\partial f}{\partial \mathbf{y}} - (\mathbf{D}^{-1} \mathbf{C})^T \frac{\partial f}{\partial \mathbf{z}} \\ &= [\mathbf{1}_{n-l} \quad -(\mathbf{D}^{-1} \mathbf{C})^T] \begin{bmatrix} \partial f / \partial \mathbf{y} \\ \partial f / \partial \mathbf{z} \end{bmatrix} \equiv \mathbf{M} \frac{\partial f}{\partial \mathbf{x}} \end{aligned} \quad (7.53)$$

where

- $\mathbf{1}_{n-l}$ :  $(n-l) \times (n-l)$  identity matrix
- $\mathbf{M}$ : a  $(n-l) \times (p+n)$  matrix, namely,

$$\mathbf{M} \equiv [\mathbf{1}_{n-l} \quad -(\mathbf{D}^{-1} \mathbf{C})^T] \quad (7.54)$$

**Definition:** The *reduced gradient* of  $f$  is defined as the  $(n-l)$ -dimensional vector  $df/d\mathbf{y}$  of eq.(7.53).

**Remarks:**

- The reduced gradient of  $f$  is a linear transformation of the gradient of  $f$  with respect to  $\mathbf{x}$ , the transformation being given by matrix  $\mathbf{M}$
- In the absence of inequalities (7.47c), the normality condition of Problem (7.47a & b) are

$$\frac{df}{d\mathbf{y}} \equiv \mathbf{M} \frac{\partial f}{\partial \mathbf{x}} = \mathbf{0} \quad (7.55)$$

i.e., at a stationary point of Problem (7.47a & b), the gradient of  $f$  with respect to  $\mathbf{x}$  need not vanish; only its *projection* onto matrix  $\mathbf{M}$  must vanish.

Apart from inequalities (7.47c), which are relatively simple to handle, the problem can be treated as an *unconstrained* one.

It is noteworthy that  $\mathbf{M}^T$  plays the role of the isotropic orthogonal complement  $\mathbf{L}$  of  $\mathbf{J} \equiv \partial \mathbf{h} / \partial \mathbf{x}$ . Indeed, from eq.(7.51),

$$\Delta \mathbf{z} = \frac{\partial \mathbf{z}}{\partial \mathbf{y}} \Delta \mathbf{y} = -\mathbf{D}^{-1} \mathbf{C} \Delta \mathbf{y}$$

or

$$\mathbf{D}^{-1} \mathbf{C} \Delta \mathbf{y} + \Delta \mathbf{z} = \mathbf{0}_{l+p}$$

which can be cast in the form

$$\begin{bmatrix} \mathbf{D}^{-1} \mathbf{C} & \mathbf{1}_{l'} \end{bmatrix} \begin{bmatrix} \Delta \mathbf{y} \\ \Delta \mathbf{z} \end{bmatrix} = \mathbf{0}_{l+p}$$

where  $l' \equiv l + p$ . Further, recalling the constraints  $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ , we have

$$\frac{\partial \mathbf{h}}{\partial \mathbf{x}} \Delta \mathbf{x} \equiv \mathbf{J} \Delta \mathbf{x} = \mathbf{0}_{l+p}$$

whence,

$$\mathbf{J} = \begin{bmatrix} \mathbf{D}^{-1} \mathbf{C} & \mathbf{1}_{l'} \end{bmatrix}$$

Now,

$$\mathbf{J} \mathbf{M}^T = \begin{bmatrix} \mathbf{D}^{-1} \mathbf{C} & \mathbf{1}_{l'} \end{bmatrix} \begin{bmatrix} \mathbf{1}_{n-l} \\ -\mathbf{D}^{-1} \mathbf{C} \end{bmatrix} = \mathbf{D}^{-1} \mathbf{C} - \mathbf{D}^{-1} \mathbf{C} = \mathbf{0}_{l' \times n'}$$

with  $n' \equiv n - l$ . Therefore,  $\mathbf{M}^T$  is an orthogonal complement of  $\mathbf{J}$ . However, contrary to  $\mathbf{L}$ ,  $\mathbf{M}^T$ , or  $\mathbf{M}$  for that matter, **is not isotropic**.

We need two search directions at the  $i$ th iteration: one for  $\mathbf{y}$  and one for  $\mathbf{z}$ . To find these directions, we start by recalling eq.(7.50a), whence

$$\Delta \mathbf{h} = \underbrace{\frac{\partial \mathbf{h}}{\partial \mathbf{y}}}_{\equiv \mathbf{C}} \Delta \mathbf{y} + \underbrace{\frac{\partial \mathbf{h}}{\partial \mathbf{z}}}_{\equiv \mathbf{D}} \Delta \mathbf{z} = \mathbf{0}_{l+p} \quad (7.56)$$

Let, at the  $i$ th iteration,

$$\Delta \mathbf{y} = \lambda \mathbf{s}^i, \quad \Delta \mathbf{z} = \lambda \mathbf{t}^i, \quad \lambda > 0 \quad (7.57)$$

Substitution of eq.(7.57) into eq.(7.56) leads to

$$\lambda \mathbf{C} \mathbf{s}^i + \lambda \mathbf{D} \mathbf{t}^i = \mathbf{0}$$

with  $\lambda$ ,  $\mathbf{s}^i$  and  $\mathbf{t}^i$  as yet to be determined. Since  $\lambda > 0$ , the above equation leads, in turn, to

$$\mathbf{C} \mathbf{s}^i + \mathbf{D} \mathbf{t}^i = \mathbf{0}$$

whence,

$$\mathbf{t}^i = -\mathbf{D}^{-1} \mathbf{C} \mathbf{s}^i \quad (7.58)$$

Since we want to minimize  $f(\mathbf{x}, \mathbf{y})$ , a plausible choice of  $\mathbf{s}^i$  is

$$\mathbf{s}^i = - \left. \frac{df}{d\mathbf{y}} \right|_{(\mathbf{y}^i, \mathbf{z}^i)} \quad (7.59)$$

Further, substitution of eq.(7.59) into eq.(7.58) leads to the desired expression for  $\mathbf{t}^i$ , namely,

$$\mathbf{t}^i = \mathbf{D}^{-1} \mathbf{C} \left. \frac{df}{d\mathbf{y}} \right|_{(\mathbf{y}^i, \mathbf{z}^i)} \quad (7.60)$$

thereby obtaining the two desired search directions.

We thus have an update of  $\mathbf{x}$ :

$$\mathbf{x}^{i+1} \equiv \begin{bmatrix} \mathbf{y}^i + \lambda \mathbf{s}^i \\ \mathbf{z}^i + \lambda \mathbf{t}^i \end{bmatrix} \quad (7.61)$$

The optimum value  $\lambda^*$  of  $\lambda$  is found upon solving an unconstrained problem of one-dimensional minimization:

$$f(\mathbf{y}^i + \lambda \mathbf{s}^i, \mathbf{z}^i + \lambda \mathbf{t}^i) \rightarrow \min_{\lambda} \quad (7.62)$$

**Remark:** The foregoing optimization is implemented without consideration of inequalities (7.47c). Hence, a test must be conducted to verify whether those constraints are obeyed. If some of the foregoing inequalities are violated, an *adjustment* is in order, as explained below.

**Adjustment:** Let  $\lambda_i$ , for  $i = 1, \dots, p + n$ , be the *positive* value of  $\lambda$  that renders one of the two inequalities of each of relations (7.47c) active. Then, let  $\lambda_{\text{opt}}$  be the adjusted value of  $\lambda$  that *does not violate* the above inequalities, i.e.,

$$\lambda_{\text{opt}} = \min\{\lambda^*, \{\lambda_i\}_1^{p+n}\} \quad (7.63)$$

Hence,

$$\mathbf{x}^{i+1} \equiv \begin{bmatrix} \mathbf{y}^i + \lambda_{\text{opt}} \mathbf{s}^i \\ \mathbf{z}^i + \lambda_{\text{opt}} \mathbf{t}^i \end{bmatrix} \quad (7.64)$$

**Remark:** After the foregoing adjustment has taken place, nothing guarantees that the constraints (7.47b) are verified. Hence, one further adjustment is needed:

**Complying with the equality constraints:** With  $\mathbf{y}$  fixed to its current value,  $\mathbf{y}^{\text{curr}}$ , we correct the current value  $\mathbf{z}^{\text{curr}}$  of  $\mathbf{z}$  by means of the Newton-Raphson method, i.e.,

$$1. \mathbf{h}^{\text{curr}} \longleftarrow \mathbf{h}(\mathbf{z}^{\text{curr}}; \mathbf{y}^{\text{curr}})$$

2.

$$\mathbf{h}(\mathbf{z}^{\text{curr}} + \Delta \mathbf{z}; \mathbf{y}^{\text{curr}}) \approx \mathbf{h}^{\text{curr}} + \left. \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \right|_{\mathbf{y}=\mathbf{y}^{\text{curr}}, \mathbf{z}=\mathbf{z}^{\text{curr}}} \Delta \mathbf{z} \rightarrow \mathbf{0}$$

3.

$$\Delta \mathbf{z} = - \left[ \left. \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \right|_{\mathbf{y}=\mathbf{y}^{\text{curr}}, \mathbf{z}=\mathbf{z}^{\text{curr}}} \right]^{-1} \mathbf{h}^{\text{curr}} \equiv -\mathbf{D}^{-1} \mathbf{h}^{\text{curr}}$$

$$4. \mathbf{z}^{\text{curr}} \longleftarrow \mathbf{z}^{\text{curr}} + \Delta \mathbf{z}, \quad \mathbf{h}^{\text{curr}} \longleftarrow \mathbf{h}(\mathbf{z}^{\text{curr}}; \mathbf{y}^{\text{curr}})$$

5. If  $\|\mathbf{h}^{\text{curr}}\| \leq \epsilon$  stop; else, go to 1

where  $\epsilon$  is a prescribed tolerance. Once the Newton-Raphson adjustment is completed, the optimization algorithm proceeds to the next iteration. The overall iterative procedure is finished when a convergence criterion has been met.

**Drawbacks of the GRG method:**

- Success is heavily dependent upon the user's choice of independent and dependent variables: Its rate of convergence decreases as  $\kappa(\mathbf{D})$  grows (Luenberger, 1984)
- The speed of convergence is slowed down by the Newton-Raphson iterations: we have an iteration loop within an exterior iteration loop!

In spite of the foregoing drawbacks, however, the GRG method is rather popular, for it is even available in Excel, an office-automation software package that has been used in the teaching of optimum design (Tai, 1998).

**Example 7.6.1 The equilibrium configuration of a  $N$ -link chain (Luenberger, 1984)**

We revisit here the problem of Example 6.4.2, for a  $N$ -link chain, as shown in Fig. 6.10, which we reproduce in Fig. 7.10 for quick reference.

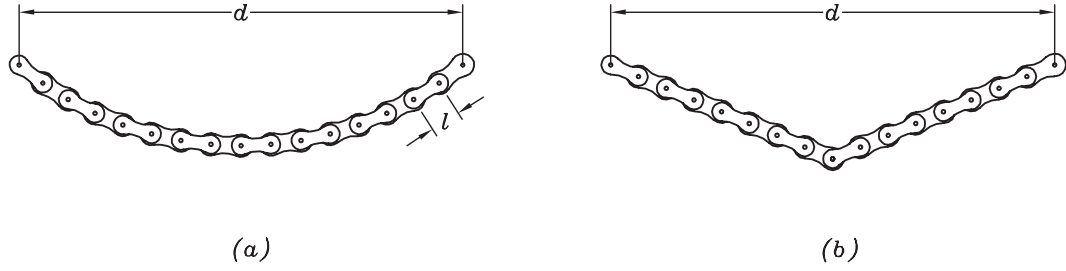


Figure 7.10: A  $N$ -link chain at: (a) its unknown equilibrium configuration; and (b) a known configuration, to be used as an initial guess

*We recall that the chain attains its equilibrium configuration when its potential energy attains its minimum value. As in the above-mentioned example, we use the configuration of Fig. 7.10b as initial guess and find the equilibrium configuration for the values  $N = 4$ ,  $d = 1.5$  m, and  $\ell = 0.5$  m*

*Solution:* In following Luenberger's formulation, we let the  $i$ th link span an  $x$  distance  $x_i$  and a  $y$  distance  $y_i$ . If  $V \equiv \mu f(x_1, y_1, x_2, y_2, \dots, x_N, y_N)$  denotes the potential energy of the chain, and  $\mu$  is the mass density of the links per unit length, then minimizing  $V$  is equivalent to minimizing  $f$ , which is given by

$$\begin{aligned}
 f(x_1, y_1, x_2, y_2, \dots, x_N, y_N) &= \frac{1}{2}y_1 + (y_1 + \frac{1}{2}y_2) + \dots \\
 &\quad + (y_1 + y_2 + \dots + y_{N-1} + \frac{1}{2}y_N) \\
 &= \frac{1}{2} \sum_{i=1}^N [2(N-i) + 1] \rightarrow \min_{\{x_i, y_i\}_1^N}
 \end{aligned}$$

subject to

$$\begin{aligned}
 \sum_{i=1}^N y_i &= 0 \\
 x_i^2 + y_i^2 - 0.5^2 &= 0 \\
 -0.5 \leq x_i, y_i &\leq +0.5, \quad i = 1, \dots, N
 \end{aligned}$$

For  $N = 4$ , we have

$$f(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4) = \frac{1}{2}(7y_1 + 5y_2 + 3y_3 + y_4) \rightarrow \min_{\{x_i, y_i\}_1^4}$$

subject to

$$\begin{aligned} h_1 &\equiv x_1 + x_2 + x_3 + x_4 - 1.5 = 0 \\ h_2 &\equiv y_1 + y_2 + y_3 + y_4 = 0 \\ h_3 &\equiv x_1^2 + y_1^2 - 0.5^2 = 0 \\ h_4 &\equiv x_2^2 + y_2^2 - 0.5^2 = 0 \\ h_5 &\equiv x_3^2 + y_3^2 - 0.5^2 = 0 \\ h_6 &\equiv x_4^2 + y_4^2 - 0.5^2 = 0 \\ -0.5 &\leq x_i, \quad y_i \leq +0.5, \quad i = 1, \dots, 4 \end{aligned}$$

Use symmetry to simplify the problem:

$$\begin{aligned} x_3 &= x_2, \quad x_4 = x_1, \quad y_3 = -y_2, \quad y_4 = -y_1 \\ \Rightarrow f(x_1, y_1, x_2, y_2) &\equiv 3y_1 + y_2 \rightarrow \min_{\{x_i, y_i\}_1^2} \end{aligned}$$

**Remark:** Now constraint  $h_2 = 0$  is identically verified and hence, is deleted. We are thus left with the constraints (note that  $h_2$  is redefined below)

$$\begin{aligned} h_1 &\equiv x_1 + x_2 - 0.75 = 0 \\ h_2 &\equiv x_1^2 + y_1^2 - 0.5^2 = 0 \\ h_3 &\equiv x_2^2 + y_2^2 - 0.5^2 = 0 \\ -0.5 &\leq x_i, \quad y_i \leq +0.5, \quad i = 1, \dots, 2 \end{aligned}$$

We thus have:

- Design variables:  $x_1, y_1, x_2, y_2 \Rightarrow n = 4$
- Equality constraints:  $h_i = 0$ , for  $i = 1, 2, 3 \Rightarrow l = 3$
- Degree of freedom:  $n - l = 1 \Rightarrow$  one single independent variable. Choose  $y_1$ .  
Hence,

$$\mathbf{y} \equiv [y_1], \quad \mathbf{z} \equiv \begin{bmatrix} x_1 \\ x_2 \\ y_2 \end{bmatrix}$$

Preliminary calculations:

$$\frac{\partial f}{\partial \mathbf{y}} = \left[ \frac{\partial f}{\partial y_1} \right] = [3], \quad \frac{\partial f}{\partial \mathbf{z}} = \begin{bmatrix} \partial f / \partial x_1 \\ \partial f / \partial x_2 \\ \partial f / \partial y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$\mathbf{h} \equiv \begin{bmatrix} x_1 + x_2 - 0.75 \\ x_1^2 + y_1^2 - 0.25 \\ x_2^2 + y_2^2 - 0.25 \end{bmatrix}$$

Hence,

$$\frac{\partial \mathbf{h}}{\partial \mathbf{y}} \equiv \begin{bmatrix} \partial \mathbf{h} / \partial y_1 & \partial \mathbf{h} / \partial y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 2y_1 \\ 0 \end{bmatrix} \equiv \mathbf{C},$$

$$\frac{\partial \mathbf{h}}{\partial \mathbf{z}} \equiv \begin{bmatrix} \partial \mathbf{h} / \partial x_1 & \partial \mathbf{h} / \partial x_2 & \partial \mathbf{h} / \partial y_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 2x_1 & 0 & 0 \\ 0 & 2x_2 & 2y_2 \end{bmatrix} \equiv \mathbf{D}$$

**Feasible initial guess:** From Fig. 7.10b, in m,

$$x_1 = x_2 = 0.37500, \quad y_1 = y_2 = \frac{1}{2} \sqrt{1^2 - 0.75^2} = 0.33072$$

$$\Rightarrow \quad \mathbf{x}^0 = \begin{bmatrix} 0.33072 \\ 0.37500 \\ 0.37500 \\ 0.33072 \end{bmatrix}$$

Hence,

$$f_{\text{curr}} \equiv f(\mathbf{x}^0) = 3 \times 0.33072 + 0.33072 = 1.32288$$

$$\mathbf{C} = \begin{bmatrix} 0 \\ 0.66144 \\ 0 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} 1 & 1 & 0 \\ 0.75000 & 0 & 0 \\ 0 & 0.7500 & 0.66144 \end{bmatrix}$$

and

$$\mathbf{D}^{-1} = \begin{bmatrix} 0 & 1.3333 & 0 \\ 1 & -1.3333 & 0 \\ -1.1339 & 1.5119 & 1.5119 \end{bmatrix}$$

Further,

$$\frac{\partial \mathbf{z}}{\partial \mathbf{y}} = -\mathbf{D}^{-1} \mathbf{C}$$

$$= - \begin{bmatrix} 0 & 1.3333 & 0 \\ 1 & -1.3333 & 0 \\ -1.1339 & 1.5119 & 1.5119 \end{bmatrix} \begin{bmatrix} 0 \\ 0.66144 \\ 0 \end{bmatrix} = \begin{bmatrix} -0.88192 \\ 0.88192 \\ -1.00000 \end{bmatrix}$$

$$\begin{aligned}
\Rightarrow \quad \frac{df}{d\mathbf{y}} &= \frac{\partial f}{\partial \mathbf{y}} + \left( \frac{\partial \mathbf{z}}{\partial \mathbf{y}} \right)^T \frac{\partial f}{\partial \mathbf{z}} = \frac{\partial f}{\partial \mathbf{y}} - (\mathbf{D}^{-1}\mathbf{C})^T \frac{\partial f}{\partial \mathbf{z}} \\
&= 3 + \begin{bmatrix} -0.88192 & 0.88192 & -1.0000 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = [3] - [1] = [2]
\end{aligned}$$

Search directions:

$$\mathbf{s}^1 = - \left. \frac{df}{d\mathbf{y}} \right|_{\mathbf{x}=\mathbf{x}^1} = -[2]$$

Hence,

$$\mathbf{t}^1 = -\mathbf{D}^{-1}\mathbf{C}\mathbf{s}^1 = \begin{bmatrix} -0.88192 \\ 0.88192 \\ -1.00000 \end{bmatrix} [-2] = \begin{bmatrix} 1.7638 \\ -1.7638 \\ 2.0000 \end{bmatrix}$$

$$\Rightarrow \quad \Delta \mathbf{x} = \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} = \begin{bmatrix} \lambda \mathbf{s}^1 \\ \lambda \mathbf{t}^1 \end{bmatrix} = \lambda \begin{bmatrix} -2.0000 \\ 1.7638 \\ -1.7638 \\ 2.0000 \end{bmatrix}$$

and

$$\mathbf{x}^{\text{new}} = \mathbf{x}^{\text{curr}} + \Delta \mathbf{x} = \begin{bmatrix} 0.33072 - 2.0000\lambda \\ 0.37500 + 1.7638\lambda \\ 0.37500 - 1.7638\lambda \\ 0.33072 + 2.0000\lambda \end{bmatrix}$$

Thus,

$$\begin{aligned}
f(\mathbf{x} + \Delta \mathbf{x}) &= f(\lambda) = 3(y_1 + \Delta y_1) + (y_2 + \Delta y_2) \\
&= 3(0.33072 - 2\lambda) + 0.33072 + 2\lambda = 1.3229 - 4\lambda
\end{aligned}$$

Hence, in order to minimize  $f(\lambda)$  we must make  $\lambda (> 0)$  as large as the constraints allow us. In the next step we find the set  $\{\lambda_i\}_1^4$  allowing us to attain the equality on one side of the set of inequalities (7.47c):

$$\begin{aligned}
y_1 : \quad 0.33072 - 2.0000\lambda_1 &= -0.5 \quad \Rightarrow \quad \lambda_1 = \frac{0.83072}{2.0000} = 0.41536 \\
x_1 : \quad 0.37500 + 1.7638\lambda_2 &= 0.5 \quad \Rightarrow \quad \lambda_2 = \frac{0.12500}{1.7638} = 0.070868 \\
x_2 : \quad 0.37500 - 1.7638\lambda_2 &= -0.5 \quad \Rightarrow \quad \lambda_3 = \frac{0.87500}{1.7638} = 0.49608 \\
y_2 : \quad 0.33072 + 2.0000\lambda_1 &= 0.5 \quad \Rightarrow \quad \lambda_4 = \frac{0.16928}{2.0000} = 0.084640
\end{aligned}$$



whence,  $\lambda_{\text{opt}} = \min\{0.41536, 0.070868, 0.49608, 0.084640\} = 0.070868$

Therefore,

$$\mathbf{x}_{\text{new}} = \begin{bmatrix} 0.33072 - 2.0000 \times 0.070868 \\ 0.37500 + 1.7638 \times 0.070868 \\ 0.37500 - 1.7638 \times 0.070868 \\ 0.33072 + 2.0000 \times 0.070868 \end{bmatrix} = \begin{bmatrix} 0.18898 \\ 0.5000 \\ 0.2500 \\ 0.47246 \end{bmatrix}$$

$$\Rightarrow f_{\text{new}} = 3 \times 0.18898 + 0.47246 = 1.0394$$

which means that we brought down the objective function by 21% of its original value in one single iteration. However, nothing guarantees that the equality constraints are satisfied. Let us verify:

$$\mathbf{h}^{\text{new}} = \mathbf{h}(\mathbf{x}^{\text{new}}) = \begin{bmatrix} 0.5000 + 0.25000 - 0.75000 \\ 0.5000^2 + 0.18898^2 - 0.5000^2 \\ 0.25000^2 + 0.47246^2 - 0.5000^2 \end{bmatrix} = \begin{bmatrix} 0.0000 \\ 0.035713 \\ 0.035718 \end{bmatrix} \neq \mathbf{0}$$

Hence, a correction to  $\mathbf{z}$ , with  $\mathbf{y}$  kept at its current value  $\mathbf{y}^{\text{curr}}$ , is warranted:

$$\mathbf{D}_{\text{new}} \Delta \mathbf{z} = -\mathbf{h}^{\text{new}}$$

where

$$\mathbf{D}_{\text{new}} \equiv \mathbf{D}(\mathbf{x}^{\text{new}}) = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0.5000 & 0.94492 \end{bmatrix}$$

Hence,

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0.5000 & 0.94492 \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta y_2 \end{bmatrix} = \begin{bmatrix} 0.0000 \\ 0.035713 \\ 0.035718 \end{bmatrix} \Rightarrow \Delta \mathbf{z} = \begin{bmatrix} -0.035713 \\ 0.035713 \\ -0.056698 \end{bmatrix}$$

Thus,

$$\mathbf{h}(\mathbf{z}^{\text{new}} + \Delta \mathbf{z}; \mathbf{y}^{\text{new}}) = \begin{bmatrix} 0.0000 \\ 0.0012786 \\ 0.0044866 \end{bmatrix}$$

Hence, the norm of  $\mathbf{h}$  has been brought down by one order of magnitude. Besides,  $f(\mathbf{z}^{\text{new}} + \Delta \mathbf{z}; \mathbf{y}^{\text{new}}) = 0.9827$ , and hence, the objective function was brought down by an additional 5.5%, thereby completing one full iteration. Further iterations are left to the reader as an exercise.

## 7.6.2 The Complex Method

The concept of *simplex* was introduced in Subsection 4.4.3 as the polyhedron in  $\mathbb{R}^n$  with the smallest number of vertices, namely  $n+1$ . A *complex* in  $\mathbb{R}^n$  is a polyhedron with  $m > n+1$  vertices; e.g., in 2D, a complex is a quadrilateral; in 3D a cube is an example of a complex. We describe below a method due to Box (1965) that is based on the concept of complex. In Box's method a complex  $\mathcal{C}$  with  $m = 2n$  vertices, for  $n \geq 2$ , is defined. Implementations are reported in (Kuester and Mize, 1973) and (Xu et al., 1994).

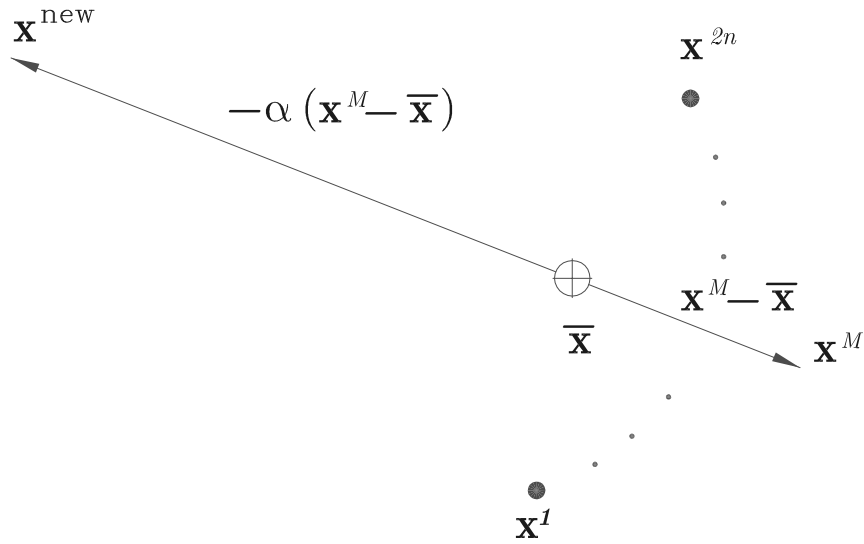


Figure 7.11: Replacement of the worst vertex of the complex by a reflection

### Box's Algorithm

1. Given one feasible vertex  $\mathbf{x}^1$  of the initial complex, generate the remaining  $2n-1$  vertices so that complex is feasible:  $\mathcal{C} = \{\mathbf{x}^i\}_1^{2n} \in \mathcal{F}$ , the feasible region
2. Let  $f_i \equiv f(\mathbf{x}^i)$  and  $f_M = \max\{f_i\}_1^{2n}$  %  $\mathbf{x}^M$  is the worst vertex
3. Let  $\mathcal{C}' = \{\mathbf{x}^i\}_{i=1, i \neq M}^{2n}$  and let  $\bar{\mathbf{x}}$  be the position vector of the centroid of  $\mathcal{C}'$ , i.e.,

$$\bar{\mathbf{x}} = \frac{1}{2n-1} \left( \sum_{i=1}^{2n} \mathbf{x}^i - \mathbf{x}^M \right)$$

4. Recover lost vertex by reflecting<sup>2</sup>  $\mathbf{x}^M$  about  $\bar{\mathbf{x}}$  by means of

$$\mathbf{x}^{\text{new}} \leftarrow \mathbf{x}^M \leftarrow \bar{\mathbf{x}} - \alpha(\mathbf{x}^M - \bar{\mathbf{x}}) \equiv (1 + \alpha)\bar{\mathbf{x}} - \alpha\mathbf{x}^M, \alpha > 0 \quad (\alpha_{\text{Box}} = 1.3)$$

5. if  $\mathbf{x}^{\text{new}} \in \mathcal{F}$ , continue; else

$$5.1 \quad \mathbf{x}^{\text{new}} \leftarrow \frac{1}{2}(\bar{\mathbf{x}} + \mathbf{x}^{\text{new}})$$

5.2 if  $\mathbf{x}^{\text{new}} \in \mathcal{F}$ , continue; else go to 5.1

abort if too many iterations

6. go to 2; if  $\mathbf{x}^{\text{new}}$  is not new worst vertex continue; else

$$\mathbf{x}^{\text{new}} \leftarrow \beta(\mathbf{x}^M - \bar{\mathbf{x}}), \quad 0 < \beta < 1$$

7. stop when convergence criterion has been met.

A possible convergence criterion is to stop when the difference between the maximum value  $f_M$  of the objective function and its minimum,  $f_m$ , is smaller than a prescribed ratio  $\epsilon_1$  times the same difference at the original complex. An alternative criterion involves the *size* of the current complex, given by the rms value of the distances of the vertices from the centroid: whenever this value is smaller than a prescribed ratio  $\epsilon_2$  times the corresponding value for the original complex, the procedure stops. A combination of the two criteria is advisable.

---

<sup>2</sup>A reflection in this context, depicted in Fig. 7.11, need not preserve the geometry of  $\mathbb{R}^n$ , as do improper orthogonal transformations—see Subsection 3.3.2



# Bibliography

- Angeles, J., 2007, *Fundamentals of Robotic Mechanical Systems. Theory, Methods, Algorithms*, Third Edition, Springer, New York.
- Angeles, J., 2011, *Dynamic Response of Linear Mechanical Systems. Modeling, Analysis and Simulation*, Springer, New York.
- Bertsekas, D. P., 1995. *Nonlinear Programming*, Athena Scientific Belmont, MA.
- Box, M.J., 1965, "A new method of constrained optimization and a comparison with other methods," *Computer Journal*, Vol. 8, pp. 42–52.
- Boyd, S. and Vandenberghe, L., 2004, *Convex Optimization*. Cambridge: Cambridge University Press.
- Brent, R.P., 1972, *Algorithms for Minimization Without Derivatives*, Prentice-Hall, Inc., Englewood Cliffs, NJ.
- Broyden, C.G., 1970, "The convergence of a class of double-rank minimization algorithm," *J. Inst. Math. Applications*, Vol. 6, pp. 76–90.
- Culioli, J.-C., 1994. *Introduction à l'optimisation*, Ellipses Publisher, Paris.
- Dahlquist, G. and Å. Björck. 1974, *Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ.
- Fletcher, R., 1970, "A new approach to variable metric algorithms," *Computer Journal*, Vol. 13, pp. 317–322.
- Fletcher, R. and Reeves, C.M., 1964, "Function minimization by conjugate gradients," *Computer Journal*, Vol. 7, No. 2, pp. 149–154.
- French, M.E., 1992. *Form, Structure and Mechanism*, Macmillan, London.

- Gleick, J., 1988. *Chaos. Making a New Science*, Penguin Books, New York.
- Goldfarb, D.. 1970, "A family of variable metric updates derived by variational means," *Mathematics of Computing*, Vol. 24, pp. 23–26.
- Hillier, S. and Liebermann, G.J., 1995, *Introduction to Mathematical Programming*, McGraw-Hill, Inc., New York.
- Karush, W., 1939, *Minima of Functions of Several Variables with Inequalities as Side Conditions*, M.S. Thesis, University of Chicago, Chicago, IL.
- Kuester, J.L. and Mize., J.H., 1973, *Optimization Techniques with Fortran*, McGraw-Hill Book Co., New York.
- Kuhn, H.W. and Tucker, A.W., 1951, "Nonlinear programming," in J. Neyman (ed.), *Second Berkeley Symposium*, pp. 481–492, University of California Press, Berkeley, CA.
- Lalee, M., Nocedal, J. and Plantenga, T., 1998, "On the implementation of an algorithm for large-scale equality constrained optimization," *SIAM J. Optim.*, Vol. 8, No. 3, pp. 682–706.
- Liu, Z. and Angeles, J., 1992, "Least-square optimization of planar and spherical four-bar function generator under mobility constraint," *ASME, J. Mech. Des.*, Vol. 114, pp. 569–573.
- Livio, M., 2002, *The Golden Ratio. The Story of Phi, the World's Most Astonishing Number*, Broadway Books, New York.
- Luenberger, D.G., 1984, *Linear and Nonlinear Programming*, Addison-Wesley Publishing Company, 2nd ed., Reading, MA.
- Mandelbrot, B.B., 1983, *The Fractal Geometry of Nature*, W.H. Freeman and Company, 3rd ed., New York.
- Murray, W., 1997, "Sequential quadratic programming methods for large problems," *Comput. Optim. Appl.*, Vol. 7, pp. 127–142.
- Nelder, J.A. and Mead, R., 1965, "A simplex method for function minimization," *Computer Journal*, Vol. 7, pp. 308–313.

- Papalambros, P. and Wilde, D.J., 2000, *Principles of Optimal Design*, Cambridge University Press, Cambridge, UK.
- Powell, M.J.D., 1964, "An efficient method for finding the minimum of a function of several variables without calculating derivatives," *Computer Journal*, Vol. 7, pp. 155–162.
- Powell, M.J.D., 1969, "A method for nonlinear constraints in minimizing problems," in R. Fletcher (ed.), *Optimization*, Academic Press, London, pp. 283–297.
- Rao, S.S., 1996, *Engineering Optimization*, John Wiley and Sons, Inc., 3rd ed., New York
- Rosen, J.B., 1960, "The gradient projection method for nonlinear programming. Part I. Linear constraints," *SIAM Journal*, Vol. 8, pp. 181–217.
- Rosen, J.B., 1961, "The gradient projection method for nonlinear programming. Part II. Nonlinear constraints," *SIAM Journal*, Vol. 9, pp. 414–452.
- Salmon, G., *Higher Algebra*, Fifth Edition (1885), reprinted 1964, Chelsea Publishing Co., New York.
- Salisbury, J.K. and Craig, J.J., 1982, "Articulated hands. Force and kinematic issues," *The Int. J. Robotics Res.*, Vol. 1, No. 1, pp. 4–17.
- Shanno, D.F., 1970, "Conditioning of quasi-Newton methods for function minimization," *Mathematics of Computing*, Vol. 24, pp. 647–656.
- Shigley, J. and Mischke, C., 1989, *Mechanical Engineering Design*, McGraw-Hill Book, Inc., New York.
- Spendley, W., Hext, G.R. and Himsworth, F.R., 1962, "Sequential application of simplex designs in optimization and evolutionary operation," *Technometrics*, Vol. 4, No. 4, pp. 441–461.
- Taguchi, G. 1988. The development of quality engineering. *The ASI Journal*, Vol. 1, No. 1, pp. ???–???
- Tai, K., 1998, "Integrated design optimization and analysis using a spreadsheet application," *Int. Journal of Mechanical Engineering Education*, Vol. 27, No. 1, pp. 29–40.

- Teng, C.P. and Angeles, J., 2001, "A sequential-quadratic-programming algorithm using orthogonal decomposition with Gerschgorin stabilization," *ASME J. Mechanical Design*, Vol. 123, pp. 501–509.
- Varga, R.S., 2000, *Matrix Iterative Analysis*, Springer, New York.
- Wilde, D. 1992. "Monotonicity analysis of Taguchi's robust circuit design problem," *ASME J. Mechanical Design*, Vol. 114, pp. 616–619.
- Xu, P., Ramden, T., Burton, R.T., Krus, P. and Sargent, C., 1994, "Feasibility of training a neural network based hydraulic component simulator using the complex method," *Proc. Seventh Bath International Fluid Power Workshop*.
- Zoutendijk, G., 1960, *Methods of Feasible Directions*, Elsevier Publishing Company, Amsterdam.



# Index

- École polytechnique, 226
- backward substitution, 62
- best fit, 48
- Bezout number, 130
- Cartesian decomposition, 51
- ceiling function, 22
- complementary slackness, 228
- complex, 258
- condition number, 48, 60
- constrained gradient, 119
- constrained Hessian, 124
- constraint
  - force of, 141
- control variables, 14
- convex, 143
  - combination, 143
- damping, 78
- damping factor, 78
- design, 11, 87
  - engineering, 11
  - industrial , 11
- design error, 240
- design variables, 15
- design vector (DV), 87
- design-variable vector, 87
- determined system, 73
- dexterity, 22
- dexterity matrix, 23
- diallytically, 130, 139
- dichotomous, 21
- differentiation with respect to
  - vectors, 54
- direct methods, 236
- divine proportions, 41
- dual form, 119
- equidistant points, 31
- exhaustive search, 20
- experiment, 20
- feasible
  - gradient, 120
  - manifold, 120, 191
  - move, 122, 233
  - region, 225
  - space, 191
  - subspace, 120
- feasible move, 123
- feasible subset, 120
- Fibonacci Numbers, 30
- Filius Bonacci, 30
- first-order normality condition, 88
- first-order normality conditions (FONC),
  - 118
- floating-point operations, 49
- flops, 49, 70
- forward substitution, 61, 62
- fractal, 77
- Freudenstein parameters, 239

- full mobility, 240
- full-rank, 7
- Gaussian elimination, 61
- Gerschgorin Theorem, 196
- golden ratio, 40
- golden section, 40, 41
- gradient, 8, 88
- gradient operator, 8
- Gram-Schmidt orthogonalization, 68
- Hessian, 88
  - perturbed , 194
  - stabilization, 195
  - stabilized, 195
- Hessian operator, 9
- Householder reflections, 68
- idempotent matrix, 57
- ill-conditioning, 48
- images, 52
- improper orthogonal matrix, 68
- inconsistent, 48
- Index, 265
- indirect methods, 236
- interval of uncertainty (i.o.u), 20
- isotropic matrix, 61, 149
- Jacobian matrix, 23, 76
- Karush-Kuhn-Tucker (KKT), 228
- Karush-Kuhn-Tucker conditions, 226
- kernel, 52
- Lagrange multipliers, 116
- Lagrangian, 116, 227
- least-square error, 79
- least-square formulation, 16
- least-square solution, 64, 80
- limit design, 232
- linear least squares, 48
- linear programming, 232
- linear system
  - underdetermined, 7
- linear transformation, 52
- linear-quadratic programs, 194
- local maximum, 89
- local minimum, 89
- lower-triangular matrix, 61
- LU-decomposition, 61
- manifold, 120
- master, 163
- mathematical model, 13, 14
- matrix
  - positive-definite, 51
  - positive-semidefinite, 51
  - sign-indefinite, 52
- maximally invertible, 23
- minimum-norm problem, 116
- minimum-norm solution, 145
- Moore-Penrose generalized inverse
  - left, 64
  - left (LMPGI), 7, 48
  - right, 145
  - right , 116
  - right (RMPGI), 7
- Newton-Gauss method, 79
- Newton-Raphson method, 75
- nominal values, 16
- nonconvex, 143
- nonlinear system, 73
- nonsingular square matrices, 60
- norm, 48
  - Chebyshev, 49, 50

- Euclidean, 48, 49
- Euclidean , 49
- Euclidean norm
  - weighted, 116
- Frobenius, 49, 50
- infinity, 49, 50
- maximum, 49
- normal equations, 64
- normal interval, 20
- normality condition, 80
- normality conditions, 115, 225
  - dual form, 115
  - in primal form, 115
- normality conditions (NC), 48, 63
  - first-order, 87
  - second-order, 87
- null space, 52
- nullity, 52
- numerical conditioning, 161
- objective function, 23
- objective function, 16
- of  $\phi$ 
  - gradient, 152
- operation conditions, 16
- operation point, 22
- ordered sets, 226
- orthogonal complement, 8, 119, 164
- orthogonal transformation, 53
- Orthogonal-Decomposition Algorithm (ODA),
  - 166
- orthogonal-decomposition algorithm (ODA),
  - 161
- orthogonalization algorithms, 48
- output variables, 14
- overdetermined system, 63, 79
- penalty term, 241
- perturbation, 194
- plastic failure, 232
- polynomial equations, 130
- production systems, 11
- projection, 57
  - orthogonal , 57
  - plane, 57
- projector, 57, 164
- proper orthogonal matrix, 53
- range, 52
- rank-one matrix, 58
- recursively, 30
- reduced
  - gradient, 120
  - normality condition, 139
- reduced Hessian, 125, 194
- redundant set of equations, 48
- reflection, 56
- reflections
  - pure , 56
- relative roundoff errors, 60
- robust design, 16
- saddle point, 89
- Schwartz's Theorem, 88
- second variation, 123
- second-order normality condition (SONC),
  - 123
- sequence, 30
  - monotonically increasing, 30
- sequential quadratic programming, 194
- slack variables, 238
- slave, 163
- stabilized reduced Hessian, 198
- state variables, 14

- stationary point, 80, 119
- strategy, 20
- synthesis matrix, 240
- system
  - determined, 48
  - overdetermined, 48
- trace, 50
- triangular system, 61
- unconstrained, 165
- unconstrained minimization, 87
- underdetermined system, 144
- unimodal function, 19
- upper-triangular form, 69
- upper-triangular matrix, 61, 68
- weakly coupled system, 61
- weighted least squares, 151
- weighting matrix, 79