Chapter 2

Fundamentals of Multivariable Optimization

2.1 Introduction

We start by introducing the notation to be used throughout the course in the area of multivariable optimization, which will be the subject of the balance of the course. We then continue 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 (DVV), or *design vector* for brevity, $\mathbf{x}$. The main result here is the *normality conditions* (NC) of the problem at hand. We derive the *first-order* NC, which are *necessary* for a stationary point (SP); then, we derive the *second-order* NC, which are *sufficient* for a *minimum*, a *maximum* or a *saddle point*. These three kinds of SP are duly characterized.

In the next stage, 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 *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 NC of the problem at hand, which readily leads to the *left Moore-Penrose generalized inverse* (LMPGI) of the coefficient matrix, namely, rectangular matrix, and for which an *inverse* proper cannot be defined. It is shown that computing the best fit 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 computation of such a generalized inverse is to be avoided.

2.2 Notation

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

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

$$\mathbf{A}^I \equiv (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T \tag{2.1}$$

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

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

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

$$\mathbf{C}^\dagger \equiv \mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1} \tag{2.2}$$

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

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. $\mathbf{x}$

$\mathbf{H}_i$: ith 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 $i = 1, \ldots, l$, expressed in vector form as $\mathbf{h}(\mathbf{x}) = \mathbf{0}$

m: number of equations $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{0}$

n: number of design variables $\mathbf{x}$

$\mathbf{O}$: The zero matrix

p: number of constraint equations $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ or $\mathbf{C}\mathbf{x} = \mathbf{d}$

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

$\mathbf{U}$: (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

$\mathbf{x}_0$: minimum-norm solution of an underdetermined linear system

$\mathbf{x}_1$: least-square solution of an overdetermined linear system

$\mathbf{1}$: The identity matrix

∇: 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 *Hesdian operator*; its operand being a scalar, it produces a square, symmetric matrix

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

2.3 The Numerical Solution of Linear Systems of Equations

We consider the system

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{2.3}$$

where

A: $n \times n$ matrix of *known* coefficients

b: n-dimensional right-hand side *known* vector

x: n-dimensional vector of *unknowns*

Definition 2.3.1 If

$$\det(\mathbf{A}) = 0 \tag{2.4}$$

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

Fact: *If* $\mathbf{A}$ *is nonsingular, then eq.(2.3) has a* unique solution*, which is given by*

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \tag{2.5}$$

Caveat: Never compute $\mathbf{A}^{-1}$ explicitly. It 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 certain roundoff error.

Avoid roundoff-error amplification!

Various methods for computing a good approximation to *the* solution (2.5):

Gaussian elimination, a.k.a. *LU-decomposition*: 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.

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.)

If $\mathbf{A}$ is nonsingular, but otherwise arbitrary, of $n \times n$, decompose $\mathbf{A}$ into the form

$$\mathbf{A} = \mathbf{LU} \tag{2.6}$$

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

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & 1 \end{bmatrix} \tag{2.7}$$

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix} \tag{2.8}$$

Now eq.(2.3) is rewritten as

$$\mathbf{LUx} = \mathbf{b} \quad \Rightarrow \quad \begin{cases} \mathbf{Ly} = \mathbf{b} \\ \mathbf{Ux} = \mathbf{y} \end{cases} \tag{2.9}$$

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

$$\det(\mathbf{A}) = \det(\mathbf{L})\det(\mathbf{U}) \tag{2.10a}$$

But, apparently,

$$\det(\mathbf{L}) = 1, \quad \det(\mathbf{U}) = \Pi_1^n u_{ii} \tag{2.10b}$$

$$\Rightarrow \quad \det(\mathbf{A}) = \det(\mathbf{U}) = \Pi_1^n u_{ii} \tag{2.10c}$$

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

2.3.1 Cholesky Decomposition

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

$$\mathbf{A} = \mathbf{U}^T\mathbf{U} \tag{2.11}$$

$$\mathbf{U} = \begin{bmatrix} u_{11} & 0 & \cdots & 0 \\ u_{21} & u_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & \cdots & u_{nn} \end{bmatrix} \tag{2.12}$$

where $\mathbf{U}$ is a *real*, upper-triangular matrix.

The solution of system (2.3) proceeds as in the general case, in two steps:

$$\mathbf{U}^T\mathbf{y} = \mathbf{b} \tag{2.13}$$

$$\mathbf{U}\mathbf{x} = \mathbf{y} \tag{2.14}$$

2.3.2 Condition Numbers

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 an n-dimensional $\mathbf{a}$ with components a_i, for $i = 1, \ldots, n$:

$$\|\mathbf{a}\|_E \equiv \sqrt{a_1^2 + \cdots + a_n^2} \tag{2.15}$$

Computing this norm thus requires n multiplications, n additions, and one square root. Not very "cheap."

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

$$\|\mathbf{a}\|_\infty \equiv \max_i\{|a_i|\}_1^n \tag{2.16}$$

Notice that this norm requires no floating-point operations (flops). Quite economical.

The *p-norm*:

$$\|a\|_p \equiv (\sum_{j=1}^{n} |a_i|^p)^{1/p} \tag{2.17}$$

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

Likewise, matrix norms can be defined in various ways:

- The *Euclidean norm*: the square root of the largest (nonnegative) eigenvalue of the positive-semidefinite product of the matrix by its transpose. For example, for the $n \times n$ matrix $\mathbf{A}$,

$$\|\mathbf{A}\|_E \equiv \max_i\{\sqrt{\lambda_i}\} \tag{2.18}$$

  where $\{\, \lambda_i \,\}_1^n$ is the set of eigenvalues of $\mathbf{A}\mathbf{A}^T$.

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

$$\|\mathbf{A}\|_F \equiv \sqrt{\sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}^2} \equiv \sqrt{\mathrm{tr}(\mathbf{A}\mathbf{A}^T)} \tag{2.19}$$

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

$$\|\mathbf{A}\|_\infty \equiv \max_{i,j} |\{a_{ij}\} \tag{2.20}$$

- The *p-norm*:

$$\|A\|_p \equiv (\sum_{j=1}^{n}\sum_{i=1}^{n} |a_{ij}|^p)^{1/p} \tag{2.21}$$

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

Remarks:

The *trace* of $\mathbf{A}$, $\mathrm{tr}(\mathbf{A})$, is defined as the sum of its diagonal entries: $\mathrm{tr}(\mathbf{A}) \equiv \sum_{i=1}^{n} a_{ij}$.

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*

Now, regarding the roundoff-error amplification when solving the system (2.3), let

$\delta\mathbf{A}$: the matrix roundoff error in $\mathbf{A}$

$\delta\mathbf{b}$: the vector roundoff-error in $\mathbf{b}$

$\delta\mathbf{x}$: the vector roundoff-error incurred when solving eq.(2.3), 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

$$\epsilon_{\mathbf{A}} \equiv \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}, \quad \epsilon_{\mathbf{b}} \equiv \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}, \quad \epsilon_{\mathbf{x}} \equiv \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \tag{2.22}$$

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)

$$\epsilon_{\mathbf{x}} \leq \kappa(\mathbf{A})(\epsilon_{\mathbf{A}} + \epsilon_{\mathbf{b}}) \tag{2.23}$$

where $\kappa(\mathbf{A})$ is the *condition number* of matrix $\mathbf{A}$, which is defined, *for nonsingular square matrices*, as

$$\kappa(\mathbf{A}) \equiv \|\mathbf{A}\|\|\mathbf{A}^{-1}\| \tag{2.24}$$

If the Euclidean norm is adopted, then we have

$$\kappa \equiv \kappa_E = \sqrt{\frac{\lambda_l}{\lambda_s}} \equiv \sqrt{\frac{\lambda_l}{\lambda_s}} \tag{2.25}$$

where

- λ_s: smallest eigenvalue of $\mathbf{A}\mathbf{A}^T$
- λ_l: largest eigenvalue of $\mathbf{A}\mathbf{A}^T$

It is now apparent that κ_E is bounded from below but unbounded from above:

$$\kappa_E \geq 1 \tag{2.26}$$

In fact, the above result holds for κ defined based on any norm.

Remarks:

- The condition number of a singular matrix tends to ∞
- 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 $\kappa = 1$. They are optimally conditioned.

Definition 2.3.2 An $n\times n$ matrix $\mathbf{A}$ is *symmetric* if it equals its transpose: $\mathbf{A} = \mathbf{A}^T$

Definition 2.3.3 An $n \times n$ matrix $\mathbf{A}$ is *skew-symmetric* if it equals *the negative* of its transpose: $\mathbf{A} = -\mathbf{A}^T$

Fact 2.3.1 (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:

$$\mathbf{A} = \mathbf{A}_s + \mathbf{A}_{ss} \tag{2.27a}$$

$$\mathbf{A}_s = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T) \tag{2.27b}$$

$$\mathbf{A}_{ss} = \frac{1}{2}(\mathbf{A} - \mathbf{A}^T) \tag{2.27c}$$

Equation (2.27a) is termed the *Cartesian decomposition* of $\mathbf{A}$.

Definition 2.3.4 A quadratic form q of an n-dimensional vector $\mathbf{x}$ is associated with an $n \times n$ matrix $\mathbf{A}$:

$$q \equiv \mathbf{x}^T\mathbf{A}\mathbf{x} \tag{2.28}$$

Fact 2.3.2 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 $\mathbf{x}$,

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = 0 \tag{2.29}$$

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

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

Expand 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 2.3.5 An $n \times n$ (symmetric) matrix $\mathbf{A}$ is *positive-definite* (*positive-semidefinite*) if, for every n-dimensional vector $\mathbf{x}$, the quadratic form $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is greater than (greater than or equal to) zero.

Characerization of positive-definiteness (semidefiniteness): An $n \times n$ (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.*

2.4 The Least-Square Solution of Overdetermined Linear Systems

Definition 2.4.1 A system of linear equations of the form

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{2.30}$$

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, no $\mathbf{x}$ that verifies *all* the equations is available.

Definition 2.4.2 $\mathbf{A}$ is of full rank if its n $(< q)$ q-dimensional columns are linearly independent.

Remark: If $\mathbf{A}$ is of full rank, then

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

$$\det(\mathbf{A}^T\mathbf{A}) > 0 \tag{2.31}$$

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

$$\mathbf{e} \equiv \mathbf{b} - \mathbf{A}\mathbf{x} \tag{2.32}$$

Problem: 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

$$f \equiv \frac{1}{2}\|\mathbf{e}\|^2 \quad \rightarrow \quad \min_{\mathbf{x}} \tag{2.33}$$

The *normality conditions* of Problem (2.33) are obtained upon zeroing the gradient of f with respect to $\mathbf{x}$.

$$\nabla f \equiv \frac{\partial f}{\partial \mathbf{x}} = 0 \tag{2.34}$$

Moreover, ∇f is obtained from the "chain rule":

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

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

$$\nabla f \equiv \left(\frac{\partial \mathbf{e}}{\partial \mathbf{x}}\right)^T \frac{\partial f}{\partial \mathbf{e}} \tag{2.35}$$

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

$$\frac{\partial \mathbf{e}}{\partial \mathbf{x}} = -\mathbf{A}, \quad \frac{\partial f}{\partial \mathbf{e}} = \mathbf{e} \equiv \mathbf{b} - \mathbf{A}\mathbf{x} \tag{2.36}$$

Upon plugging expressions (2.36) into eq.(2.34):

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b} \tag{2.37}$$

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

If $\mathbf{A}$ is of full-rank, then eq.(2.37) admits one *unique* solution—determined case—which is the *least-square solution* of the given system:

$$\mathbf{x}_L = \mathbf{A}^I\mathbf{b} \tag{2.38a}$$

where

$$\mathbf{A}^I \equiv (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T \tag{2.38b}$$

Here, $\mathbf{A}^I$ is termed the *left Moore-Penrose generalized inverse* of the rectangular matrix $\mathbf{A}$.

Remarks:

- The condition number κ_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 κ_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, κ_E is given by the ratio of the largest to the smallest eigenvalues of $(\mathbf{A}^T\mathbf{A})$, i.e.,
$$\kappa_E(\mathbf{A}^T\mathbf{A}) = \kappa_E^2(\mathbf{A}) \tag{2.39}$$

- Thus, the roundoff-error amplification factor incurred in solving the normal equations (2.37) is the square of that incurred when "solving" eq.(2.3) in the determined case.

- Not only this. Formula (2.38a) 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 $n^2 \times q$ products and $n^2(q-1)$ additions;
  - the computation of the right-hand side of eq.(2.37), 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 normal equations should be avoided!**

The good news is that there are alternative to normal-equation solving. One of these relies on *Householder reflections* (to be described presently): Premultiply both sides of eq. (2.30) by n Householder reflections—$q \times q$ *improper* orthogonal matrices— $\mathbf{H}_i$, for $i = 1, \ldots, n$, i.e.,

$$\mathbf{HAx} = \mathbf{Hb} \tag{2.40}$$

where

$$\mathbf{H} = \mathbf{H}_n\mathbf{H}_{n-1}\ldots\mathbf{H}_1$$

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

$$\mathbf{HA} = \begin{bmatrix} \mathbf{U} \\ \mathbf{O} \end{bmatrix}, \quad \mathbf{Hb} = \begin{bmatrix} \mathbf{b}_U \\ \mathbf{b}_L \end{bmatrix} \tag{2.41}$$

in which

- $\mathbf{U}$: an $n \times n$ *upper-triangular matrix*
- $\mathbf{O}$: the $(q-n) \times n$ zero matrix
- $\mathbf{b}_U$: an n-dimensional vector containing the first n components of $\mathbf{A}^T\mathbf{b}$
- $\mathbf{b}_L$: an $(q-n)$-dimensional vector containing the last $q-n$ components of $\mathbf{A}^T\mathbf{b}$

Thus, eq.(2.40) leads to two subsystems of equations:

$$\mathbf{Ux} = \mathbf{b}_U \tag{2.42a}$$

$$\mathbf{Ox} = \mathbf{b}_L \neq \mathbf{0} \tag{2.42b}$$

The least-square solution can be readily calculated by backward substitution from eq.(2.42a), and symbolically expressed as

$$\mathbf{x}_L = \mathbf{U}^{-1}\mathbf{b}_U. \tag{2.43}$$

Remark: Equation (2.42b) expresses a contradiction: The left-hand side is the product of the $(q-n) \times n$ zero matrix times the unknown vector; the right-hand side is not necessarily zero

Thus, eq.(2.42b) yields the least-square error associated with the solution $\mathbf{x}_L$: $\|\mathbf{b}_L\|$. Now we have an important result:

Theorem 2.4.1 (The Projection Theorem) Let $\mathbf{e}_0$ denote the error vector of minimum Euclidean norm, i.e.,

$$\mathbf{e}_0 \equiv \mathbf{b} - \mathbf{Ax}_L \tag{2.44}$$

Then, $\mathbf{e}_0$ is *orthogonal* to the image of $\mathbf{x}_L$ under $\mathbf{A}$.

Proof: We have

$$\mathbf{e}_0^T\mathbf{Ax}_L = (\mathbf{b} - \mathbf{Ax}_L)^T\mathbf{Ax}_L$$

Upon expansion,

$$\mathbf{e}_0^T\mathbf{A}\mathbf{x}_L = \mathbf{b}^T\mathbf{A}\mathbf{x}_L - \mathbf{x}_L^T\mathbf{A}^T\mathbf{A}\mathbf{x}_L$$

Plugging expressions (2.38a & b) into the above equation,

$$\begin{aligned}\mathbf{e}_0^T\mathbf{A}\mathbf{x}_L &= \mathbf{b}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} - \mathbf{b}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} \\ &= \mathbf{b}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} - \mathbf{b}^T\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} = 0 \end{aligned} \tag{2.45}$$

thereby completing the proof. The Projection Theorem is illustrated in Fig. 2.1.

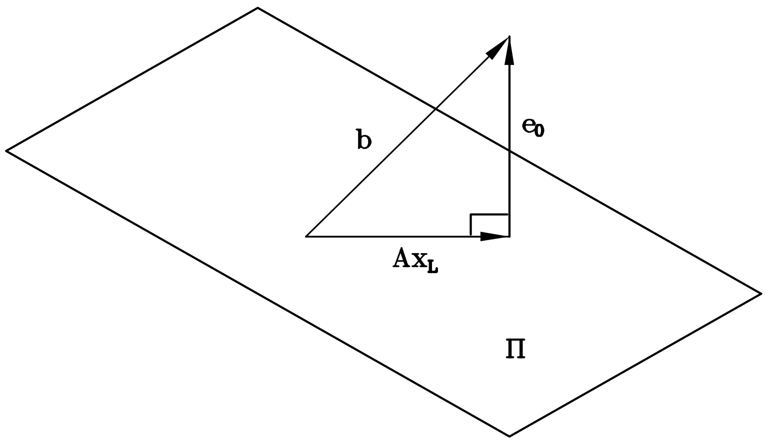

Figure 2.1: The Projection Theorem

2.4.1 Householder Reflections

Definition 2.4.3 An $n \times n$ matrix $\mathbf{R}$ is **improper orthogonal** if (i) $\mathbf{R}\mathbf{R}^T = \mathbf{R}^T\mathbf{R} = \mathbf{1}$ and (ii) $\det(\mathbf{R}) = -1$

Remark: An improper orthogonal matrix of $n \times n$ represents *reflections*, i.e., linear transformations of an n-dimensional vector space that preserve both the *magnitude* of vectors—their Euclidean norm—and the inner product of any two vectors.
Problem: Find a linear transformation of the columns of the $q \times n$ matrix $\mathbf{A}$ that will render this matrix in upper-triangular form *without changing the geometric relations among the columns*, i.e., while preserving the inner products of any two of these columns, including the product of a column by itself

Solution: Assume that we have applied reflections $\mathbf{H}_1$, $\mathbf{H}_2$, ..., $\mathbf{H}_{k-1}$, in this order, to $\mathbf{A}$ that have rendered it in *upper-trapezoidal form*, i.e.,

$$
\mathbf{A}_{i-1} \equiv \mathbf{H}_{i-1} \ldots \mathbf{H}_2 \mathbf{H}_1 \mathbf{A}
$$
$$
= \begin{bmatrix}
a_{11}^* & a_{12}^* & \cdots & a_{1,i-1}^* & a_{1i}^* & \cdots & a_{1n}^* \\
0 & a_{22}^* & \cdots & a_{2,i-1}^* & a_{2i}^* & \cdots & a_{2n}^* \\
0 & 0 & \cdots & a_{3,i-1}^* & a_{3i}^* & \cdots & a_{3n}^* \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & a_{i-1,i-1}^* & a_{i-1,i}^* & \cdots & a_{i-1,n}^* \\
0 & 0 & \cdots & 0 & a_{i,i}^* & \cdots & a_{i,n}^* \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & a_{q,i}^* & \cdots & a_{qn}^*
\end{bmatrix} \tag{2.46}
$$

The next Householder reflection, $\mathbf{H}_i$, is determined so as to render the last $q-i$ components of the ith 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

$$
\alpha_i = \operatorname{sgn}(a_{ii}^*)\sqrt{(a_{ii}^*)^2 + (a_{i+1,i}^*)^2 + \cdots + (a_{qi}^*)^2} \tag{2.47}
$$
$$
\mathbf{u}_i = [\,0 \quad 0 \quad \cdots \quad 0 \quad a_{ii}^* + \alpha_i \quad a_{i+1,i}^* \quad \cdots \quad a_{qi}^*\,]^T \tag{2.48}
$$
$$
\mathbf{H}_i = \mathbf{1} - 2\frac{\mathbf{u}_i\mathbf{u}_i^T}{\|\mathbf{u}_i\|^2} \tag{2.49}
$$

where $\operatorname{sgn}(x)$ is defined as $+1$ if $x > 0$, as -1 if $x < 0$, and is left undefined when $x = 0$.

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 a single multiplication.

Exercise: 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. 2.2.

It is noteworthy that

(a) α_i is defined with the sign of a_{ii}^* because β_i is a multiple of the ith component of $\mathbf{u}_i$, which is, in turn, the sum of a_{ii}^* and α_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 β_i a

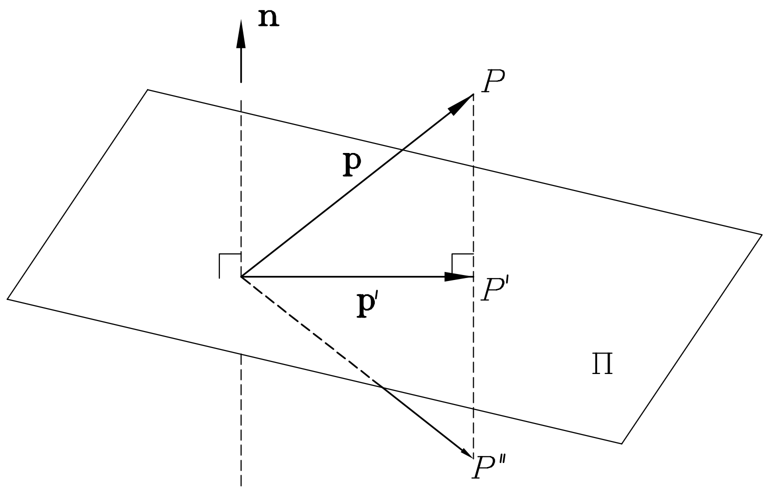

Figure 2.2: The geometric interpretation of the ith Householder reflection

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 β_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 defined, the system at hand becomes

$$\mathbf{HAx} = \mathbf{Hb} \tag{2.50}$$

with $\mathbf{H}$ defined as

$$\mathbf{H} \equiv \mathbf{H}_n \ldots \mathbf{H}_2\mathbf{H}_1 \tag{2.51}$$

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

$$\mathbf{HA} = \begin{bmatrix} \mathbf{U} \\ \mathbf{O}_{q'n} \end{bmatrix}, \quad \mathbf{Hb} = \begin{bmatrix} \mathbf{b}_U \\ \mathbf{b}_L \end{bmatrix} \tag{2.52}$$

where: $q' \equiv q - n$; $\mathbf{O}_{q'n}$ is the $(q-n)\times n$ zero matrix; $\mathbf{b}_U$ is an n-dimensional vector; and $\mathbf{b}_L$ is a q'-dimensional vector, normally different from zero.

The unknown $\mathbf{x}$ can thus be calculated from eq.(2.50) by back-substitution.

Remarks:

- The last m' components of the left-hand side of eq.(2.50) are zero

- However, the corresponding components of the RHS of the same equation are not necessarily zero. WWW?

- Nothing! Recall that the overdetermined system (2.30) in general has no solution. The lower part of $\mathbf{b}$, $\mathbf{b}_L$, is then nothing but a q'-dimensional array containing the nonzero components of the approximation error in the new coordinates. That is, the least-square error $\mathbf{e}_0$ in these coordinates, takes the form

$$\mathbf{e}_0 = \begin{bmatrix} \mathbf{0}_n \\ \mathbf{b}_L \end{bmatrix} \tag{2.53a}$$

  Therefore,

$$\|\mathbf{e}_0\| = \|\mathbf{b}_L\| \tag{2.53b}$$

2.5 Unconstrained Optimization

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

$$f(\mathbf{x}) \quad \rightarrow \quad \min_{\mathbf{x}} \tag{2.54}$$

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* ∇f of the objective function with respect to the design vector must vanish:

$$\nabla f \equiv \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\mathbf{x}_o} = \mathbf{0} \tag{2.55a}$$

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 ∇f

vector, namely,

$$\nabla f \equiv \frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \partial f/\partial x_1 \\ \partial f/\partial x_2 \\ \vdots \\ \partial f/\partial x_n \end{bmatrix} \tag{2.55b}$$

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 $\mathbf{x} = \mathbf{x}_o$:

$$f(\mathbf{x}) = f(\mathbf{x}_o) + \nabla f\,|_{\mathbf{x}_o} (\mathbf{x} - \mathbf{x}_o) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_o)^T \nabla\nabla f\,|_{\mathbf{x}_o}(\mathbf{x} - \mathbf{x}_o) + \mathrm{HOT} \tag{2.56a}$$

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,

$$\nabla\nabla f \equiv \frac{\partial^2 f}{\partial \mathbf{x}^2} = \begin{bmatrix} \partial^2 f/\partial x_1^2 & \partial^2 f/\partial x_1 \partial x_2 & \cdots & \partial^2 f/\partial x_1 \partial x_n \\ \partial^2 f/\partial x_2 \partial x_1 & \partial^2 f/\partial x_2^2 & \cdots & \partial^2 f/\partial x_2 \partial x_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2 f/\partial x_n \partial x_1 & \partial^2 f/\partial x_n \partial x_2 & \cdots & \partial^2 f/\partial x_n^2 \end{bmatrix} \tag{2.56b}$$

Notice that, by virtue of the smoothness assumption,

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_i \partial x_j}, \quad \text{for} \quad i, j = 1, 2, \ldots, n \tag{2.57}$$

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.(2.57), 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.(2.56a) leads to

$$\Delta f \equiv f(\mathbf{x}) - f(\mathbf{x}_o) \approx \frac{1}{2}(\mathbf{x} - \mathbf{x}_o)^T \nabla\nabla f\,|_{\mathbf{x}_o}(\mathbf{x} - \mathbf{x}_o) \tag{2.58}$$

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 Δf for every possible $\Delta \mathbf{x}$. However, it is possible to characterize the nature of the stationary point $\mathbf{x}_o$ by means of a test based on 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.

2.6 Nonlinear-Equation Solving: Determined Case

Definition 2.6.1 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*.

Example: Find the intersection of the circle and the hyperbola depicted in Fig. 2.3.

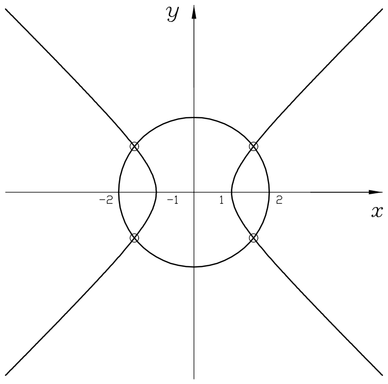

Figure 2.3: Intersection of a circle and a hyperbola

Solution: The equations of the circle and the hyperbola are

$$\phi_1(x,y) \equiv x^2 + y^2 - 4 = 0$$
$$\phi_2(x,y) \equiv x^2 - y^2 - 1 = 0$$

The solution to a nonlinear system of equations, when one exists at all, is usually *multiple*: The circle and the hyperbola of Fig. 2.3 intersect at four points:

$$\left(\sqrt{\frac{5}{2}}, \sqrt{\frac{3}{2}}\right), \quad \left(\sqrt{\frac{5}{2}}, -\sqrt{\frac{3}{2}}\right),$$
$$\left(-\sqrt{\frac{5}{2}}, \sqrt{\frac{3}{2}}\right), \quad \left(-\sqrt{\frac{5}{2}}, -\sqrt{\frac{3}{2}}\right)$$

The problem may have **no real solution**, e.g., the circle and the hyperbola of Fig. 2.4 do not intersect. The system of equations from which the coordinates of the intersection points are to be computed is given below:

$$\phi_1(x,y) \equiv x^2 + y^2 - 1 = 0$$
$$\phi_2(x,y) \equiv x^2 - y^2 - 16 = 0$$

This system of equations admits no real solution!

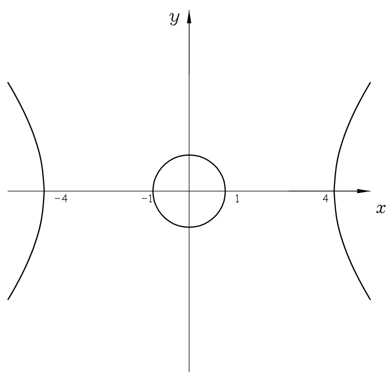

Figure 2.4: A circle and a hyperbola that do not intersect

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

$$\boldsymbol{\phi}(\mathbf{x}) = \mathbf{0} \tag{2.59}$$

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

$$\mathbf{x} \equiv \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \boldsymbol{\phi} \equiv \begin{bmatrix} \phi_1(x_1,\, x_2,\, \ldots,\, x_n) \\ \phi_2(x_1,\, x_2,\, \ldots,\, x_n) \\ \vdots \\ \phi_n(x_1,\, x_2,\, \ldots,\, x_n) \end{bmatrix} \tag{2.60}$$

2.6.1 The Newton-Raphson Method

A value $\mathbf{x}_0$ of $\mathbf{x}$ is given as an *initial guess*:

$$\mathbf{x}_0 \equiv [\, p_1 \quad p_2 \quad \ldots \quad p_n \,]^T$$

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):

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

Now, $\boldsymbol{\phi}(\mathbf{x}_0 + \Delta\mathbf{x})$ is evaluated up to its linear approximation (all quadratic and higher-order terms are dropped from its series expansion):

$$\boldsymbol{\phi}(\mathbf{x}_0 + \Delta\mathbf{x}) \approx \boldsymbol{\phi}(\mathbf{x}_0) + \left.\frac{\partial \boldsymbol{\phi}}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}_0} \Delta\mathbf{x} \tag{2.61}$$

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

$$\mathbf{\Phi} \equiv \frac{\partial \boldsymbol{\phi}}{\partial \mathbf{x}} = \begin{bmatrix} \partial\phi_1/\partial x_1 & \partial\phi_1/\partial x_2 & \cdots & \partial\phi_1/\partial x_n \\ \partial\phi_2/\partial x_1 & \partial\phi_2/\partial x_2 & \cdots & \partial\phi_2/\partial x_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial\phi_n/\partial x_1 & \partial\phi_n/\partial x_2 & \cdots & \partial\phi_n/\partial x_n \end{bmatrix} \tag{2.62}$$

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

$$\boldsymbol{\phi}_0 + \mathbf{\Phi}(\mathbf{x}_0)\Delta\mathbf{x} = \mathbf{0}$$

or

$$\mathbf{\Phi}(\mathbf{x}_0)\Delta\mathbf{x} = -\boldsymbol{\phi}_0 \tag{2.63}$$

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

$$\Delta\mathbf{x} = -\mathbf{\Phi}_0^{-1}\boldsymbol{\phi}_0, \quad \mathbf{\Phi}_0 \equiv \mathbf{\Phi}(\mathbf{x}_0) \tag{2.64}$$

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

$$\mathbf{x}_0 \quad \leftarrow \quad \mathbf{x}_0 + \Delta\mathbf{x} \tag{2.65}$$

the procedure stopping when

$$\|\Delta\mathbf{x}\| \leq \epsilon_x \tag{2.66}$$

for a prescribed tolerance ϵ_x.

Remarks:

- Use the maximum norm to test convergence in eq.(2.66);
- no guarantee that the Newton-Raphson method will converge at all;
- whether the Newton-Raphson method converges is dependent upon the initial guess, $\mathbf{x}_0$;
- the boundary between regions of convergence and divergence is a *fractal*;
- when the Newton-Raphson method converges, it does so *quadratically*: At every iteration, *two* decimal places of accuracy are gained.

2.7 Overdetermined Nonlinear Systems of Equations

A system of nonlinear equations of the form

$$\boldsymbol{\phi}(\mathbf{x}) = \mathbf{0} \tag{2.67}$$

where $\mathbf{x}$ is an n-dimensional vector and $\boldsymbol{\phi}$ 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 below:

Problem: Find an *approximate* solution to system (2.67) that verifies those equations with the *least-square error*:

$$f(\mathbf{x}) = \frac{1}{2}\boldsymbol{\phi}^T\mathbf{W}\boldsymbol{\phi} \quad \rightarrow \quad \min_{\mathbf{x}} \tag{2.68}$$

where $\mathbf{W}$ is a $q \times q$ positive-definite *weighting matrix.*
Solution: We follow a procedure similar to Newton-Raphson's: First, an initial guess $\mathbf{x}^0$ of $\mathbf{x}$ is given. Then, we produce the sequence

$$\mathbf{x}^1, \mathbf{x}^2, \ldots, \tag{2.69}$$

such that

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \Delta\mathbf{x}^k \tag{2.70}$$

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

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

$$\mathbf{W} = \mathbf{V}^T\mathbf{V} \tag{2.71}$$

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

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

$$\mathbf{V}\mathbf{\Phi}(\mathbf{x}^k)\Delta\mathbf{x}^k = -\mathbf{V}\boldsymbol{\phi}(\Delta\mathbf{x}^k) \tag{2.72}$$

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

$$\mathbf{\Phi}(\mathbf{x}) = \frac{\partial\boldsymbol{\phi}(\mathbf{x})}{\partial\mathbf{x}} \tag{2.73}$$

  Drop superscripts and recall eqs.(2.38a & b):

$$\Delta\mathbf{x} = -(\mathbf{\Phi}^T\mathbf{W}\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{W}\boldsymbol{\phi} \tag{2.74}$$

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

2.7.1 Convergence Criterion

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

$$\nabla f(\mathbf{x}) \equiv \frac{\partial f}{\partial\mathbf{x}} = \left(\frac{\partial\boldsymbol{\phi}}{\partial\mathbf{x}}\right)^T \frac{\partial f}{\partial\boldsymbol{\phi}} \tag{2.75}$$

$$\frac{\partial\boldsymbol{\phi}}{\partial\mathbf{x}} \equiv \mathbf{\Phi}, \quad \frac{\partial f}{\partial\boldsymbol{\phi}} = \mathbf{W}\boldsymbol{\phi} \tag{2.76}$$

Hence, the condition for a stationary point is

$$\mathbf{\Phi}^T\mathbf{W}\boldsymbol{\phi} = \mathbf{0} \tag{2.77}$$

which is the *normality condition* of eq.(2.68).

It is thus apparent that, at a stationary point of f, $\boldsymbol{\phi}(\mathbf{x})$ **need not vanish**; however, $\boldsymbol{\phi}(\mathbf{x})$ **must lie in the nullspace of $\boldsymbol{\Phi}^T\mathbf{W}$**. Moreover, from eq.(2.74) follows that, at a stationary point, $\Delta\mathbf{x}$ vanishes. Hence, the stopping criterion is

$$\|\Delta\mathbf{x}\| < \epsilon \tag{2.78}$$

where ϵ is a prescribed tolerance.

Remarks:

- The normality condition (2.77) alone does not guarantee a minimum, but only a *stationary point*.
- However, it turns out that, if the procedure converges, then it does so, to a second-order approximation, to a minimum, and neither to a maximum nor a to saddle point, as we prove below.

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

$$\Delta f = \left(\frac{\partial f}{\partial \mathbf{x}}\right)^T \Delta\mathbf{x} \tag{2.79}$$

i.e.,

$$\Delta f = \boldsymbol{\phi}^T\mathbf{W}\boldsymbol{\Phi}\Delta\mathbf{x} \tag{2.80}$$

Upon plugging expression (2.74) of $\Delta\mathbf{x}$ into eq. (2.80), we obtain

$$\begin{aligned}\Delta f &= -\boldsymbol{\phi}^T\mathbf{W}\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\mathbf{W}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{W}\boldsymbol{\phi} \\ &= -\boldsymbol{\phi}^T\mathbf{M}\boldsymbol{\phi}\end{aligned} \tag{2.81}$$

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, Δf is negative definite. Thus, the second-order approximation of $f(\mathbf{x})$ is negative-definite, and hence, the sequence of f values *decreases monotonically*. That is, in the neighbourhood 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**.