Chapter 2
Fundamental Concepts in Inverse Problems

The final answer to several problems can be reduced to evaluating a function—the solution function or the solution operator—and in the case of inverse problems it is not different. This is the point of view of a mathematician — insisting in the use of the notion of function. Not that one can always come about with the solution operator explicitly, but we can think abstractly on it and deduce its properties. This justifies the treatment that we present in this chapter of some of the aspects and complications that arise in the evaluation of functions. Next, we discuss some general aspects of mathematical models and inverse problems. A few classification schemes of inverse problems, illuminating different aspects, are presented. These classifications are used in subsequent chapters.

At times, the functions we are dealing with are quite complex, or are given in an extremely intricate way (for example, the function happens to be the solution of a partial differential equation [PDE]). In such a case it is almost always unavoidable to resort to a computer to produce a numerical value, i.e., a numerical solution. In this case the knowledge of the properties of the solution operator turns out to be very useful, even when we do not have the solution operator explicitly. In any case, the evaluation of the function (solution)—the final result—will be within a certain range of error, which varies mainly due to the following characteristics:

1. The problem is more complicated, or ill-behaved;
2. The function is more ill-behaved;
3. The way in which the function is evaluated in the computer (the algorithm) can be better or worse.

Case 1 is related to the need of regularizing the problem and we will present that concept in Chapter 3. Presently we deal with cases 2 and 3. Once a (small) error is introduced in a computation —through an experimental datum or due to round-off— it affects the final outcome. The error in the result can: (i) be reduced; (ii) remain small; (iii) be amplified.

There are two notions that can help us to understand *error dynamics* when evaluating a function: (a) the *condition* of the function being evaluated; and (b) the *stability*[1] of the algorithm used to evaluate it.

Condition will be dealt with in Sections 2.1 to 2.4 and algorithm stability in Section 2.5. Section 2.6 covers some questions related to existence and uniqueness. The notion of well-posed problem in the sense of Hadamard is considered in Section 2.7. In Section 2.8 a very simple classification of inverse problems is presented.

[1] The word stability can have several meanings. It is used even to name the notion of condition, in the sense that will be defined in this chapter, depending on authors. Caution is to be exercised as its meaning depends on the context.

2.1 Condition of Function Evaluation

Evaluation of a function at a given point can be *well* or *ill* conditioned. This is an intrinsic property of the function being evaluated and it does not depend on approximations.

Qualitatively it is said that the evaluation is *well-conditioned* if a small error in the point where the function is evaluated does not affect greatly the value of the function. If, however, a small error in the evaluation point leads to a large error in the value of the function, the evaluation is *ill-conditioned*.

It is possible to identify the notion of well-conditioned evaluation with continuity. Nevertheless, we desire a more restrictive notion, in the sense that well-conditioned implies continuity, but not the other way around. This shall be important when working with finite-precision arithmetic, and will allow us to distinguish different behaviours among continuous functions. Given a function, its qualitative behaviour can even depend on the region of the domain, having places where its evaluation is well-conditioned, and others where it is ill-conditioned. Let us see an example to illustrate this discussion.

Example 2.1. Evaluation of a rational function. Consider the evaluation of the function $f(x) = 1/(1-x)$. The computation of $f(x)$ is:

(a) *ill-conditioned*, if x lies near 1, (but, of course, x must be different from 1);

(b) *well-conditioned*, otherwise.

We shall treat these two cases next:

(a) When x is near 1.

Assume $x = 1.00049$ and that in the computation we use an approximate value, $x^* = 1.0005$. In this case, the absolute error of the evaluation is:

$$e_{\text{abs}} \quad = \quad f(x^*) - f(x) = -10^3/24.5\,.$$

We remark that an error of $10^{-5} = x^* - x$ in the data led to an evaluation error of $-10^3/24.5$. The error is magnified by the multiplication factor

$$m = \frac{\text{error in the result (of the evaluation)}}{\text{error in the point (of the domain)}} = -\frac{10^3/24.5}{10^{-5}}\,,$$

which, in absolute value satisfies $c = |m| > 10^6$.

(b) When x is far from 1.

When x is far from 1, the previous magnification phenomenon does not occur. Let $x = 1998$, and $x^* = 2000$ an approximation of x. Then the absolute error in the evaluation is

$$e_{abs} = \frac{1}{1-2000} - \frac{1}{1-1998} \;=\; \frac{2}{1999 \cdot 1997}\,.$$

Thus, the amplification factor of the error is $(1999 \cdot 1997)^{-1} < 10^{-6}$, *effectively* reducing the error. ■

2.2 Condition as a Derivative Bound

To study how a data error affects the evaluation of a function f, let x be the point of interest and let x^* be its approximation, and consider the quotient

$$m = \frac{\text{error in the result (evaluation)}}{\text{error in the datum (domain)}} = \frac{f(x^*) - f(x)}{x^* - x} . \tag{2.1}$$

Of course, Eq. (2.1) is *Newton's quotient* of f, a preliminary step in the definition of the derivative of a function. In the limit, $x^* \to x$, we have $m \to f'(x)$ and we define $f'(x)$ as the error *multiplication factor* of the evaluation error of f at x. This is a *local* quantity, coming from a local operator, $\frac{d}{dx}$.

Definition 2.1. Given[2] $f : \mathcal{D} \subset \mathbb{R} \to \mathbb{R}$ of class C^1, $c_f(x) = |f'(x)|$ is the *condition number* of the (*evaluation*) of f at x. We also say that the evaluation of f at x is *well-conditioned* if $c_f(x) \leq 1$ and *ill-conditioned* if $c_f(x) > 1$.

Thus, if the absolute value of the derivative of f is bounded by 1 at all the points of its domain of definition, i.e., if $|f'(x)| \leq 1$ for all $x \in \mathcal{D}$, then the evaluation of f is always well-conditioned. A simple example is the evaluation of $\sin x$, for any $x \in \mathbb{R}$.

2.3 Other Derivatives and Other Notions of Condition

It is not always convenient to access the sensitivity of the evaluation of a function by means of Eq. (2.1). As a matter of fact, Eq. (2.1) is written in terms of the quotient of two absolute errors: the evaluation absolute error and the data absolute error. At times, it is more interesting to consider relative errors. This leads to other possibilities to define the multiplication factor of the error in the evaluation. We present several alternatives in Table 2.1.

Table 2.1 Possible definitions for the multiplication factor of the error

numerator → / denominator ↓	absolute error in the evaluation	relative error in the evaluation
data absolute error	(a) $\frac{f(x^*)-f(x)}{x^*-x} \sim f'(x)$	(b) $\frac{(f(x^*)-f(x))/f(x)}{x^*-x} \sim \frac{f'(x)}{f(x)}$
data relative error	(c) $\frac{f(x^*)-f(x)}{(x^*-x)/x} \sim x f'(x)$	(d) $\frac{(f(x^*)-f(x))/f(x)}{(x^*-x)/x} \sim \frac{x f'(x)}{f(x)}$

[2] We recall that $\mathbb{R}^n \supset \Omega \ni \mathbf{x} \mapsto \mathbf{f}(\mathbf{x}) \in \mathbb{R}^m$ is called a function of *class* C^k if the derivatives of its component functions of order at least k exist and are continuous. A function of class C^0 is just a continuous function.

As done in the case of Eq. (2.1), we can propose the notion of multiplication factor by considering the limit $x^* \to x$, in the quotients of Table 2.1. Thus, we have the following *multiplication factor*[3], $m_f(x)$, at x:

(a) $f'(x)$: the (usual) derivative of f at x;

(b) $f'(x)/f(x)$: the *logarithmic derivative* of f at x (in fact, the derivative of $\ln f(x)$);

(c) $xf'(x)$: derivative (differential operator) without a special name;

(d) $xf'(x)/f(x)$: the *elasticity* of f at x (much used in economics).

Likewise, we define the *condition number* as the absolute value of the multiplication factor, $c_f(x) = |m_f(x)|$. In this case, we can talk about well or ill-conditioned evaluation for all of the condition numbers presented.

For all cases, from (a) to (d), the notion of well-conditioned evaluation is that the absolute value of the condition be less than or equal to one.

On the other hand, to be bounded by one can, sometimes, be considered too restrictive. We opt here for this criterion because in this case there is always a reduction of the error. That can be unnecessary, however. It must be pointed out that, in some applications, it is possible to work with values greater than one: a small amplification of the error in the data can be manageable. The transition value between well and ill-conditioned evaluations depends on the given problem and objectives.

2.4 Condition of Vector-Valued Functions of Several Variables

Consider now a vector-valued function of several variables,

$$\begin{aligned} \mathbf{f}\colon \mathbb{R}^n \supset \Omega &\to \mathbb{R}^m \\ \mathbf{x} = (x_1, \ldots, x_n)^T &\mapsto \mathbf{f}(\mathbf{x}) = (f_1(x_1, \ldots, x_n), \ldots, f_m(x_1, \ldots, x_n))^T \end{aligned}$$

where Ω is a subset of $\mathbb{R}^n$. As done previously for functions of a single variable, we can define different *condition numbers* (of the evaluation) of function $\mathbf{f}$ at $\mathbf{x}$. The important thing to keep in mind is that the condition is the norm of the multiplier of the error in the data determining the error in the evaluation.

To begin, let us recall that *Taylor's formula*[4] gives

$$\mathbf{f}(\mathbf{x}^*) - \mathbf{f}(\mathbf{x}) = \mathcal{J}\mathbf{f}_x \cdot (\mathbf{x}^* - \mathbf{x}) + \mathcal{O}\left(|\mathbf{x}^* - \mathbf{x}|^2\right), \text{ as } \mathbf{x}^* \to \mathbf{x}\,, \tag{2.2a}$$

[3] When one desires to be more specific, one may say, for example, that $f'(x)/f(x)$ is the multiplication factor (of the evaluation) of the *relative error* (in the result) with respect to the *absolute error* (in the datum).

[4] See precise statements of Taylor's formulae on page 204 of the Appendix A.

Here, $\mathcal{J}\mathbf{f}_x$ is the *Jacobian matrix* of $\mathbf{f}$, the $m \times n$ matrix of first-order derivatives of $\mathbf{f}$, evaluated at $\mathbf{x}$,

$$\mathcal{J}\mathbf{f}_x = \left(\begin{array}{ccc} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{array} \right)\Bigg|_x .$$

Also, O is the usual big-O order symbol, and the norm of a vector $\mathbf{v}$ is given by

$$|\mathbf{v}| = \left(v_1^2 + \ldots + v_n^2\right)^{\frac{1}{2}} .$$

Equation (2.2a) allows us to write

$$\frac{\mathbf{f}(\mathbf{x}^*) - \mathbf{f}(\mathbf{x})}{|\mathbf{f}(\mathbf{x})|} = \frac{1}{|\mathbf{f}(\mathbf{x})|}\mathcal{J}\mathbf{f}_x \cdot (\mathbf{x}^* - \mathbf{x}) + O\left(\frac{|\mathbf{x}^* - \mathbf{x}|^2}{|\mathbf{f}(\mathbf{x})|}\right) , \tag{2.2b}$$

$$\mathbf{f}(\mathbf{x}^*) - \mathbf{f}(\mathbf{x}) = |\mathbf{x}|\mathcal{J}\mathbf{f}_x \cdot \frac{(\mathbf{x}^* - \mathbf{x})}{|\mathbf{x}|} + O\left(|\mathbf{x}^* - \mathbf{x}|^2\right) , \tag{2.2c}$$

$$\frac{\mathbf{f}(\mathbf{x}^*) - \mathbf{f}(\mathbf{x})}{|\mathbf{f}(\mathbf{x})|} = \frac{|\mathbf{x}|}{|\mathbf{f}(\mathbf{x})|}\mathcal{J}\mathbf{f}_x \cdot \frac{(\mathbf{x}^* - \mathbf{x})}{|\mathbf{x}|} + O\left(\frac{|\mathbf{x}^* - \mathbf{x}|^2}{|\mathbf{f}(\mathbf{x})|}\right) , \tag{2.2d}$$

as $\mathbf{x}^* \to \mathbf{x}$, in all cases.

Due to this result, Eq. (2.2), we define the *multiplication matrices* (of the evaluation):

(a) absolute (error in the result) to absolute (error in the datum),

$$\mathcal{M}_{aa} = \mathcal{J}\mathbf{f}_x ; \tag{2.3a}$$

(b) relative (error in the result) to absolute (error in the datum),

$$\mathcal{M}_{ra} = \mathcal{J}\mathbf{f}_x / |\mathbf{f}(\mathbf{x})| ; \tag{2.3b}$$

(c) absolute (error in the result) to relative (error in the datum),

$$\mathcal{M}_{ar} = |\mathbf{x}|\mathcal{J}\mathbf{f}_x ; \tag{2.3c}$$

(d) relative (error in the result) to relative (error in the datum),

$$\mathcal{M}_{rr} = |\mathbf{x}|\mathcal{J}\mathbf{f}_x / |\mathbf{f}(\mathbf{x})| . \tag{2.3d}$$

It is worthwhile to compare these matrices with the one dimensional case, as shown in Table 2.1. These are generalizations, and therefore the notation is slightly more cumbersome, but the fundamental meaning remains.

Note that, from the definition of the norm of a matrix, Eq. (A5), and its properties, Eq. (A6), we have that

$$|\mathcal{J}_x\mathbf{f} \cdot (\mathbf{x}^* - \mathbf{x})| \leq |\mathcal{J}_x\mathbf{f}| \cdot |\mathbf{x}^* - \mathbf{x}| .$$

The condition number of the evaluation, in these cases, is the norm of the multiplication matrix,

$$c_f(\mathbf{x}) = |\mathcal{J}\mathbf{f}_x| \,, \text{ absolute to absolute} \tag{2.4a}$$

$$c_f(\mathbf{x}) = |\mathcal{J}\mathbf{f}_x|/|\mathbf{f}(\mathbf{x})| \,, \text{ relative to absolute} \tag{2.4b}$$

$$c_f(\mathbf{x}) = |\mathbf{x}| \cdot |\mathcal{J}\mathbf{f}_x| \,, \text{ absolute to relative} \tag{2.4c}$$

$$c_f(\mathbf{x}) = |\mathbf{x}| \cdot |\mathcal{J}\mathbf{f}_x|/|\mathbf{f}(\mathbf{x})| \,, \text{ relative to relative} \tag{2.4d}$$

Example 2.2. Condition number of a matrix. Given a matrix A, we will analyze the condition of the function $\mathbf{f}(\mathbf{x}) = A\mathbf{x}$.

For the linear function $\mathbf{f}$,

$$\mathcal{J}\mathbf{f}_x = A \,, \text{ for all } \mathbf{x}$$

The condition number, relative to relative, is given by (see Eq. (2.3d))

$$|\mathbf{x}|\,|A|/|A\mathbf{x}| \,.$$

Now, assume that A is invertible, and let $\mathbf{y} = A\mathbf{x}$. Therefore, $\mathbf{x} = A^{-1}\mathbf{y}$, and

$$\frac{|\mathbf{x}|\,|A|}{|A\mathbf{x}|} = \frac{|A^{-1}\mathbf{y}|}{|\mathbf{y}|}\,|A| \leq |A^{-1}|\,|A| \,,$$

due to Eq. (A6).

Thus, notwithstanding the point where the function $\mathbf{f}$ is being evaluated, the condition number is bounded by $|A^{-1}|\,|A|$. In this case, it is customary to say that

$$k(A) \quad = \quad |A^{-1}|\,|A|$$

is the *condition of matrix A*.

In general it is clear that the condition of the evaluation of $\mathbf{f}(\mathbf{x}) = A\mathbf{x}$ depends on the point $\mathbf{x}$ where such evaluation is to be done, and it can be less than $k(A)$. ■

Example 2.3. Condition in the *resolution* of a linear system of equations. Consider the linear problem

$$K\mathbf{x} \quad = \quad \mathbf{y} \,,$$

where $\mathbf{y} \in \mathbb{R}^n$ is a datum and $\mathbf{x} \in \mathbb{R}^n$ is the unknown. Assume that K is invertible so that $\mathbf{x} = K^{-1}\mathbf{y}$. We want to analyze the condition of the solution operator

$$\mathbf{y} \quad \mapsto \quad K^{-1}\mathbf{y} \,.$$

The condition number in this case, when considering the relation of the absolute error in the evaluation to the absolute error in the datum, Eq. (2.3a), is $|K^{-1}|$. ■

The previous example is used in Chapters 3 and 4.

Example 2.4. Difference between close numbers. It is known that the difference between two numbers that are close is *ill-behaved* when finite-precision arithmetic is used, as is usually the case with digital computers.

This fact is not directly related to finite-precision arithmetic, but to the intrinsic nature of the difference function, and to the fact that it is ill-conditioned for close numbers.

At the beginning, consider the function

$$\mathbb{R} \times \mathbb{R} \ni (x, y) \quad \mapsto \quad m(x, y) = x - y \in \mathbb{R} \,,$$

that calculates the difference between two numbers and let us compute its *elasticity* as defined in the following equation (compare with Eq. (2.3d) and verify their dissimilarity). We have

$$E\, m(x,y) = \left(\frac{x}{m} \frac{\partial m}{\partial x} \,,\; \frac{y}{m} \frac{\partial m}{\partial y} \right) = \left(\frac{x}{x-y} \,,\; -\frac{y}{x-y} \right) \,,$$

from whence

$$|E\, m(x, y)| = \sqrt{\frac{x^2 + y^2}{(x-y)^2}} \,. \tag{2.5}$$

Now, $|Em(x, y)|$ cannot be less than or equal to 1 (unless x and y have opposite signs, but then m would not be a difference, it would be a sum). Verify. Yet, it is possible to obtain regions of $\mathbb{R}^2$ in which the difference has a moderate elasticity, say, less than 2. Although there is an amplification of the error, it is a "small" amplification. On the other hand, if we choose y near x, $y = x + \epsilon$, with ϵ small, then the elasticity,

$$|E\, m(x, x + \epsilon)| = \sqrt{(x^2 + (x + \epsilon)^2)/\epsilon^2} \,,$$

can be arbitrarily large subject to ϵ being sufficiently small and $x \neq 0$.

Summing up, the problem of computing the difference between two numbers can be: (a) well-conditioned, if the numbers are far apart; (b) ill-conditioned if the numbers are close. Notice that this is not related to the use of finite-precision arithmetic, but will be observed in the presence of round-off errors when using such arithmetic.

Finally, it would be more correct to use the norm of the multiplication factor — the condition number of m — as obtained from Eq. (2.3d), instead of the norm of the elasticity, Eq. (2.4d). However, in this case, these two notions coincide. ■

2.5 Stability of Algorithms of Function Evaluation

As mentioned before, the notions of condition and stability are important in understanding how the errors alter the final outcome of a computation. The notion of condition of the evaluation of a function is intrinsic to the function whose outcome (image value) is to be computed, so it is unavoidable to deal with it as long as we work with that particular function. On the other hand, the notion of stability depends on the algorithm that is used to compute the value of the function.

Since there are several ways to evaluate a function, it is possible to select one that best fits our needs. An algorithm to compute the value of a function is *unstable* when the errors occurring through the intermediate steps (of the computation) are amplified. If this is not the case, the algorithm is *stable*.

An algorithm to compute the value of a function is constituted of several elementary steps. To study its stability every step must be analyzed separately, and the effect of the introduction of an error must be tracked.

Example 2.5. Evaluation of an algebraic function. The evaluation of $f(x) = \sqrt{x+1} - \sqrt{x}$ is well-conditioned, for all $x \geq 0$. We will show two algorithms for this evaluation: the more natural one is unstable, the other one is stable.

The evaluation of f is always well-conditioned, since the absolute value of its elasticity,

$$|Ef(x)| = \frac{1}{2}\left|\frac{x}{\sqrt{x+1}\,\sqrt{x}}\right| \leq \frac{1}{2}\,,$$

is always bounded by $\frac{1}{2}$.

We will consider two algorithms to evaluate f. The first one corresponds to the direct computation of

$$\sqrt{x+1} - \sqrt{x}\,,$$

and the second one is based on the algebraic expression,

$$\frac{1}{\sqrt{x+1} + \sqrt{x}}\,.$$

The two expressions are algebraically equivalent. In fact,

$$\sqrt{x+1} - \sqrt{x} \quad = \quad \frac{1}{\sqrt{x+1} + \sqrt{x}}\,,$$

for all $x > 0$.

To analyze the stability of the natural algorithms streaming from these two equivalent algebraic expressions, we depict in Fig. 2.1 the steps that make up each algorithm. The schemes in the figure help to prove that the second algorithm is preferable, as we shall see.

Each algorithm is assembled with the intermediate steps defined by the following functions,

$$\begin{aligned} p_1(x) &= \sqrt{x}\,, & p_2(x) &= x+1\,,\\ p_3(x) &= \sqrt{x}\,, & p_4(y,z) &= z-y\,,\\ p_5(y,z) &= y+z\,, & p_6(x) &= 1/x\,. \end{aligned}$$

The steps of each one of the two algorithms correspond to some of these functions and the algorithms can be interpreted as compositions of these functions.

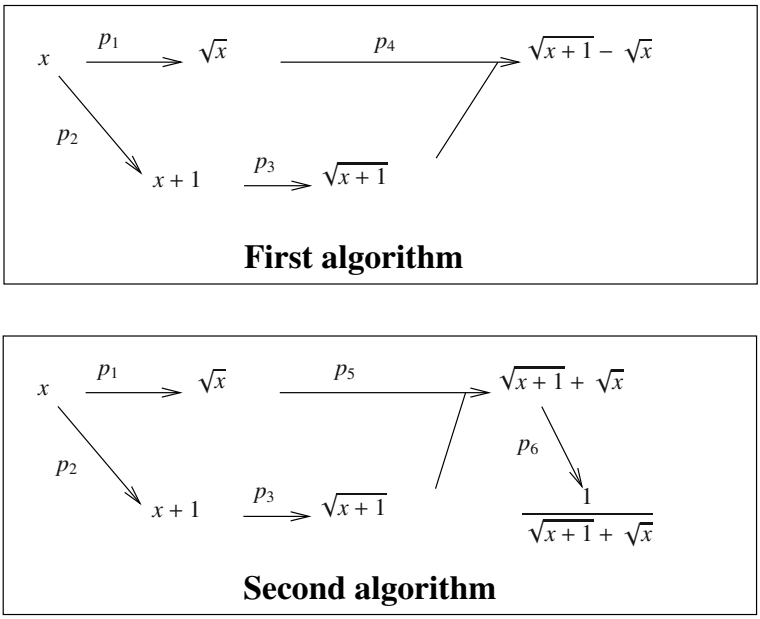

Fig. 2.1 Diagrams of the algorithms

Since the *elasticity of the composition* of functions is the product of the elasticity of the functions[5],

$$E(p_1 \circ p_2 \circ \ldots \circ p_n) = E(p_1) \times E(p_2) \times \ldots \times E(p_n) \ , \tag{2.6}$$

it suffices to analyze each step. We have that

$$|Ep_1(y)| = 1/2 \ , \ |Ep_2(y)| = |y/(y+1)| \leq 1 \ , \ Ep_3(y) = Ep_1(y) \ ,$$

and, finally, we recall that the elasticity of the difference function, p_4, has already been analyzed in example 2.4.

Now, we see that step p_4 is crucial to decide if the first algorithm is stable or not. It will be stable if

$$y = \sqrt{x} \text{ and } z = \sqrt{x+1}$$

are not too close. However, due to the elasticity of p_4, we see that, if they are close, the multiplication factor will be large. Therefore, this algorithm will only work well

[5] We recall that the *composition* of two functions, p_1 and p_2, denoted by $p_1 \circ p_2$, is the function,

$$(p_1 \circ p_2)(x) = p_1(p_2(x)) \ .$$

if x is very close to zero. For x far from zero, y and z will be close, and the algorithm will behave inadequately (it will be unstable).

Now let us check the second algorithm. Steps p_1, p_2 and p_3 are common to the first algorithm. Besides,

$$|Ep_5(y,z)| = \sqrt{(y^2+z^2)/(y+z)^2} \leq 1 ,$$

since we are dealing only with non-negative values of y and z. Also, $|Ep_6(y)| = 1$. Thus, we see that all the multiplication factors in the various steps of the second algorithm are less than or equal to 1, rendering it stable. ■

Example 2.6. Evaluation of a polynomial. Given the cubic polynomial

$$p(x) = (x - 99\pi)(x - 100\pi)(x - 101\pi) ,$$

where π is given by the value of a pocket calculator, we will present two algorithms to evaluate p in $x_o = 314.15$. The first one will use directly the previous expression, the second one will be based in the power form of $p(x)$ and will use Horner's method.

Let us start by analyzing the condition of the evaluation. Since the *elasticity of the product* is the sum of the elasticities,

$$E(f \cdot g) = Ef + Eg , \tag{2.7}$$

we see that the elasticity of the evaluation of $p(x)$ is given by

$$Ep(x) = \frac{x}{x - 99\pi} + \frac{x}{x - 100\pi} + \frac{x}{x - 101\pi} .$$

Thus, this evaluation is ill-conditioned at $x_o = 314.15$ due to the necessity of evaluating $(x - 100\pi)$ in this point. In fact,

$$\begin{aligned} Ep(314.15) &= \frac{314.15}{314.15 - 99\pi} + \frac{314.15}{314.15 - 100\pi} + \frac{314.15}{314.15 - 101\pi} \\ &\approx 100.2928 - 33905.86 - 99.7030 \approx -33905 . \end{aligned}$$

Notwithstanding, the first algorithm is superior to the second one. In fact, in the first case, we obtain

$$p(x_o) = (x_o - 99\pi)(x_o - 100\pi)(x_o - 101\pi) = 0.091446 ,$$

while using *Horner's method* (starting the algorithm by the innermost expression, $x_0 - 300\pi$ and proceeding outwards), we obtain:

$$p(x_o) = ((x_o - 300\pi)x_o - 29999\pi^2)x_o - 999900\pi^3 = -0.229 .$$

Evidently, this result contains numerical errors, since the answer should be positive. ■

The fact that a problem is ill-conditioned does not prevent us from computing its result. We must choose carefully the algorithm to be used, no matter if it is an ill or well-conditioned problem.

2.6 Questions on Existence and Uniqueness

Some of the difficulties in solving inverse problems are related to the available information: quantity (not sufficient data or seemingly overabundance thereof) and quality of information. We will illustrate these points by means of simple examples.

Let us suppose that the function that truly generates the phenomenon is

$$f(x) = 2x+1 \, .$$

In the inverse identification problem we assume that such function is unknown to us. We do assume, however, that we can determine to which class of functions the phenomenon, $f(x)=2x+1$, belongs, i.e., we characterize the *model*. The observation of the phenomenon allows us to characterize it as, say, belonging to the class of functions of the form

$$f_{a,b}(x) = ax + b \, ,$$

where a and b are arbitrary constants. From the available data we must then determine a and b, i.e., we must identify or select the model. We shall consider two situations, when one has exact data, or, otherwise, has real (noisy) data.

2.6.1 Exact Data

For the sake of the present analysis, we assume that the available data are exact. We then have three possibilities:

(a) **Not sufficient data.** The basic unit of information in this example corresponds to a point in the graph of the model. Assume known that the point $(1, 3)$ belongs to the graph of f. It is obvious that this datum is not sufficient for us to determine a and b. As a matter of fact we only know that

$$f(1)=3 \text{ or } a+b=3 \, ,$$

It is then impossible to determine the model (find the values of a and b uniquely). ■

(b) **Sufficient data.** We know data $(1, 3)$ and $(2, 5)$. Thus, $a+b = 3$ and $2a+b = 5$, from whence we can determine that $a = 2$ and $b = 1$, and select the model $f(x) = 2x + 1$. ■

(c) **Too much data.** Assume now that it is known that the points $(0, 1)$, $(1, 3)$, and also $(2, 5)$ belong to the graph of f. Then

$$a=2 \text{ and } b=1 \, .$$

It is plain to see that we have too much data, since we could determine a and b without knowing the three ordered pairs, being for that matter sufficient to know any two of such pairs. We point out that too much data does not cause problems in the determination of the model when exact data is used.

2.6.2 Real Data

In practice we do not have exact data, so it is best to have too much data, even if some repetition occurs. In spite of it, in many occasions, we only have a sufficient number of (non-repeated) data due, for example, to costs of acquiring data. Sometimes, real data will be called *noisy* data. We still have three possibilities, discussed below:

(a) **Insufficient data.** Datum $(1, 3.1)$ has an error—as we know, for our "phenomenon," $f(1) = 3$ and not 3.1. Moreover, this datum is insufficient, because we obtain only one relation between a and b,

$$a + b = 3.1 \, ,$$

but cannot determine them individually, not even approximately. Another restriction must be imposed so the inverse problem has a unique solution. A concrete and unexpected example of how this can be achieved in real problems can be seen in Section 4.6. ■

(b) **Sufficient data.** Consider that we have the following approximate data: (1, 3.1) and (2, 4.9). Then, an approximation for a and b is obtained by substituting the data in a class of models,

$$\begin{cases} a + b = 3.1 \\ 2a + b = 4.9 \end{cases} , \tag{2.8}$$

which gives $a = 1.8$ and $b = 1.3$. ■

However, even with sufficient (but with errors, i.e., noisy) data, it is not always possible to estimate the parameters by imposing that the model fits or interpolates the data. Later, we will see on example 2.7 that clarifies this remark.

Alternatively, in these cases, we try to minimize a certain measure of discrepancy between the model and the data. In the example, for every proposed model within the characterized class, that is, for every pair (a, b), we would compute the difference between the data and what is established by the model and combine these differences in some way and minimize it.

For example, if the pair (a, b) is known, the model is given by $f_{a,b}$, and the point

$$(1, f_{a,b}(1))$$

should belong to the graph. This would be what the model establishes, the so-called *prediction*[6] of the model. The data however indicates that the point should be $(1, 3.1)$, so

$$f_{a,b}(1) - 3.1$$

[6] Here the word prediction means not only "foretelling the future" as is usual; it is foretelling in regions where data is not available. In models involving time, however, it carries the former meaning. In any case, we search for scientific predictions, as discussed in Chapter 1.

is a measure of the error due to the difference between what the model foretells and what actually happens. This must be done with all the experimental data. The results may be combined to create a *discrepancy measure* between what the model foretells with (a, b) and what the real data shows. Finally, we must find the value of the pair (a, b) that minimizes that measure.

As an example, let us consider the discrepancy (or error) measure given by

$$\begin{aligned} E(a,b) &= \frac{1}{2}\left[(f_{a,b}(1) - 3.1)^2 + (f_{a,b}(2) - 4.9)^2\right] \\ &= \frac{1}{2}\left[(a + b - 3.1)^2 + (2a + b - 4.9)^2\right] . \end{aligned}$$

The minimum point of E is given by its critical point, i.e., the point where the gradient of E is null,

$$\begin{aligned} 0 &= \frac{\partial E}{\partial a} = (a + b - 3.1) + 2(2a + b - 4.9) \\ 0 &= \frac{\partial E}{\partial b} = (a + b - 3.1) + (2a + b - 4.9) . \end{aligned}$$

We thus conclude that

$$a + b = 3.1 \quad \text{and} \quad 2a + b = 4.9 .$$

It is a coincidence that, due to the form of the function $f_{a,b}$, the system in Eq. (2.8), and the one just obtained, are the same. At times the problem obtained by interpolation as in Eq. (2.8) does not have a solution, while the one obtained by least squares, like the one we just deduced, is solvable. This is a way to reach a solution in real inverse problems. This subject is addressed in example 2.7.

(c) **Too much data.** Assume that

$$(x_1, y_1),\ (x_2, y_2), \ldots, (x_n, y_n), \text{ with } n \geq 3 ,$$

are several experimental points associated with the "phenomenon" $f(x) = 2x + 1$.

It is unavoidable that these experimental data are contaminated by errors and imprecisions intrinsic to the measuring process. Thus, the data are usually *incompatible*, i.e., it is impossible to solve for a and b the system

$$\begin{aligned} y_1 - f_{a,b}(x_1) &= y_1 - (ax_1 + b) = 0 \\ y_2 - f_{a,b}(x_2) &= y_2 - (ax_2 + b) = 0 \\ &\vdots \\ y_n - f_{a,b}(x_n) &= y_n - (ax_n + b) = 0 . \end{aligned}$$

Usually we say that, since the system has n equations and only two unknown variables, it possibly has no solution.

In general we would say that there is no model in the characterized class of models that interpolates the data, i.e., the data cannot be fitted by the model.

We deal with this question from a geometrical point of view, in the context we are discussing. Note that the system can be rewritten as

$$a\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + b\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} .$$

We introduce the notation $\mathbf{x} = (x_1, \ldots, x_n)^T$, $\mathbb{1} = (1, \ldots, 1)^T$ and $\mathbf{y} = (y_1, \ldots, y_n)^T$. This notation allows the vector equation above to be written as

$$a\mathbf{x} + b\mathbb{1} = \mathbf{y} .$$

Thus, the solution of the system of equations can be rephrased as the problem of finding a linear combination of the vectors $\mathbf{x}$ and $\mathbb{1}$ to obtain $\mathbf{y}$. These vectors, $\mathbf{x}$ and $\mathbb{1}$, belong to $\mathbb{R}^n$, and they are only two. It is necessary n linearly independent vectors to represent an arbitrary vector of $\mathbb{R}^n$ as a linear combination of them. Therefore, it is very rare for one to be able to choose a and b in order that $\mathbf{y} = a\mathbf{x} + b\mathbb{1}$. That this is the case is easily visualized by means of a simple figure, see Fig. 2.2.

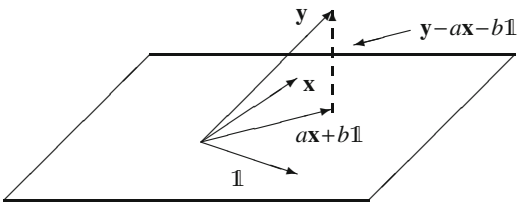

Fig. 2.2 Plane (2-dimensional subspace) in $\mathbb{R}^n$. Vector $\mathbf{y}$ does not belong to the plane spanned by $\mathbf{x}$ and $\mathbb{1}$, span $\{\mathbf{x}, \mathbb{1}\} = \{a\mathbf{x} + b\mathbb{1}$, for all $a, b \in \mathbb{R}\}$.

Let us consider the method of least squares. Define the *error* or residual vector,

$$\mathbf{r} = \mathbf{y} - (a\mathbf{x} + b\mathbb{1}) ,$$

given as the vector of differences between the experimental measurements, $\mathbf{y}$, and the predictions of the model with coefficients a and b, $a\mathbf{x} + b\mathbb{1}$.

Effectively, what can be done is to choose a linear combination between $\mathbf{x}$ and $\mathbb{1}$ (i.e., choose a and b), in such a way that the functional error,

$$E(a,b) = \frac{1}{2}|\mathbf{r}|^2 = \frac{1}{2}|\mathbf{y} - a\mathbf{x} - b\mathbb{1}|^2 = \frac{1}{2}\sum_{i=1}^{n}(y_i - (ax_i + b))^2 ,$$

is minimized. Since the sum stands for the error vector's squared norm, $|\mathbf{y}-a\mathbf{x}-b\mathbb{1}|^2$, a look at Fig. 2.2 suggests that it is equivalent to requiring that the error vector be

orthogonal to the plane spanned by $\mathbf{x}$ and $\mathbb{1}$. Or, else,[7]

$$\langle \mathbf{x}, \mathbf{y} - a\mathbf{x} - b\mathbb{1} \rangle = 0 \quad \text{and} \quad \langle \mathbb{1}, \mathbf{y} - a\mathbf{x} - b\mathbb{1} \rangle = 0 \,.$$

This can be written as

$$\mathbf{x}^T (\mathbf{y} - a\mathbf{x} - b\mathbb{1}) = 0 \,, \quad \text{and} \quad \mathbb{1}^T (\mathbf{y} - a\mathbf{x} - b\mathbb{1}) = 0 \,,$$

which leads to,

$$\begin{aligned} a\mathbf{x}^T\mathbf{x} + b\mathbf{x}^T\mathbb{1} &= \mathbf{x}^T\mathbf{y} \\ a\mathbb{1}^T\mathbf{x} + b\mathbb{1}^T\mathbb{1} &= \mathbb{1}^T\mathbf{y} \,, \end{aligned}$$

whence,

$$\begin{pmatrix} \mathbf{x}^T\mathbf{x} & \mathbf{x}^T\mathbb{1} \\ \mathbb{1}^T\mathbf{x} & \mathbb{1}^T\mathbb{1} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \mathbf{x}^T\mathbf{y} \\ \mathbb{1}^T\mathbf{y} \end{pmatrix} . \tag{2.9}$$

Defining $A = (\mathbf{x}, \mathbb{1})$, an $n \times 2$ matrix, Eq. (2.9) can be rewritten as

$$A^T A \begin{pmatrix} a \\ b \end{pmatrix} = A^T \mathbf{y} \tag{2.10}$$

which is usually called *normal equation*. Therefore, the determination of the model reduces to solving the linear system, Eq. (2.10). ■

We remark that, even though matrix $A^T A$ may not be invertible, Eq. (2.10) will always have a solution. We shall treat this question later on. Assume, however, that $A^T A$ is invertible. Then, the solution to the inverse problem can be represented by

$$\begin{pmatrix} a \\ b \end{pmatrix} = \left(A^T A\right)^{-1} A^T \mathbf{y} \,. \tag{2.11}$$

This is the solution of the inverse problem given by the *least squares method*, which corresponds to the evaluation of the function,

$$\mathbf{y} \quad \mapsto \quad \left(A^T A\right)^{-1} A^T \mathbf{y} \,.$$

It is pertinent here to recall the discussion of Section 2.5, regarding the stability of function evaluation algorithms. As a matter of fact, the algorithm suggested by the expression, $\left(A^T A\right)^{-1} A^T \mathbf{y}$ is not the best way to evaluate it; it can be *unstable* (depending on matrix A) and even inefficient from a computational point of view[8],

[7] The notation of inner product is recalled on page 189, Appendix A.

[8] The question of the inefficiency of the algorithms must be considered. Non-efficient methods can render impractical the use of a given algorithm. In the present case, depending on the size of $A^T A$, it may be very time consuming to find its inverse. Recall however that one wants to find $(a,b)^T$ satisfying Eq. (2.10).

because it presumes that the inverse of A^TA will be computed. The geometric interpretation of the problem at hand is the basis for the construction of alternative algorithms, stable and more efficient, see [35].

Assume now that A is invertible. Then A^T is also invertible, since

$$(A^T)^{-1} \quad = \quad (A^{-1})^T \, .$$

Therefore,

$$\begin{pmatrix} a \\ b \end{pmatrix} \quad = \quad A^{-1} \left(A^T\right)^{-1} A^T \mathbf{y} = A^{-1} \mathbf{y} \, . \tag{2.12}$$

This result, Eq. (2.12), is valid whether the data are exact or not. ■

Example 2.7. Sufficient data and the least squares method. We consider now an example in which, although there are sufficient data, it is impossible to identify the model due to the existence of experimental errors.

Assume a phenomenon is given precisely by

$$g(x) = \frac{1}{1 + x^2/99} \, .$$

Thus, in particular, $g(1) = 0.99$. Again, assume that g is unknown and that, after the model has been characterized, the following class of models is obtained,

$$C = \left\{ g_c \text{ for all } c \in \mathbb{R} \text{ where } g_c(x) = \frac{1}{1 + (x-c)^2/99} \right\} \, .$$

Finally, let $(1, 1.1)$ be the only experimental datum.

Well the given datum is, in principle, sufficient to select the model, since only one parameter, c, is to be determined. However, if we try to determine c by interpolation, i.e., by means of the equation

$$g_c(1) = \frac{1}{1 + (1-c)^2/99} = 1.1 \, ,$$

we see that it is impossible. In fact, for every value of c, $g_c(1)$ will be less than one. An adequate approach is to use the approximation of least squares previously presented.

Let

$$E(c) = \frac{1}{2}(1.1 - g_c(1))^2 = \frac{1}{2}\left(1.1 - \frac{1}{1 + \frac{(1-c)^2}{99}}\right)^2 \, .$$

The minimum of E is reached when $dE/dc = 0$, thus $c = 1$, (see Fig. 2.3). ■

It must be noted that, in the case of inexact data, it is certainly better to use "too much" data. In the example just discussed, if we have access to more data points, even if they contain errors, we may be able to estimate a model more symmetric with relation to the y-axis, like the phenomenon that is being modeled.

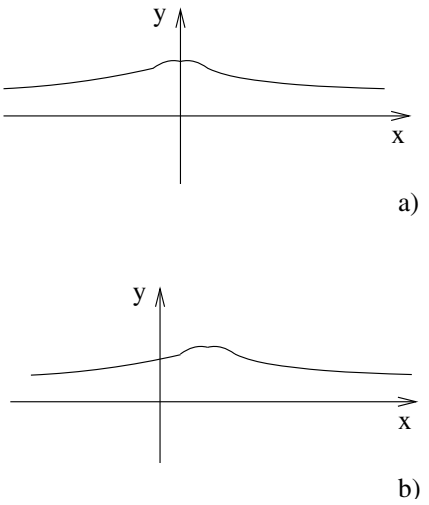

Fig. 2.3 a) The graph of a phenomenon: $g(x) = \frac{1}{1+x^2/99}$. b) Estimated graph: $g_1(x) = \frac{1}{1+(x-1)^2/99}$.

2.7 Well-Posed Problems

Hadamard defined the notion of a *well-posed problem* as being one that:

(i) has a solution (*existence*);

(ii) the solution is unique (*uniqueness*);

(iii) the solution depends "smoothly" on the data (*regularity*).

When any of these properties is not satisfied, we say that the problem is *ill-posed*. As we have already seen, inverse problems do not always satisfy properties (i)–(iii).

Sometimes, property (i) is not satisfied because it is impossible, once the class of models is characterized, to interpolate the data with any model within the class. This has been exemplified in the previous section. We may surpass this by relaxing the notion of solution — an approximation instead of an interpolation, for example in a least squares sense.

If property (ii) is not satisfied, additional restrictions must be found to force the solution to be unique. It is not possible to obtain a unique solution if information is lacking—there are no mathematical tricks to circumvent lack of knowledge. The difficulty here steams from the modelization of the phenomenon.

It is said implicitly that the problem involves data. In this case we can talk about the set of data and properties (i) and (ii) implies that the attribution

$$\text{data} \longrightarrow \text{solution}$$

is a function called *solution operator*, since for any data there is a solution and it is unique. Property (iii) asks for additional features of the solution operator.

Property (iii) is more complex from a technical point of view. The notion of smoothness has to be specified for each problem and, sometimes, can be translated as continuity or differentiability. It is common for inverse problems in infinite dimension spaces to be discontinuous. These problems must be rendered discrete to be solved by means of a computer. It is almost certain that a discrete problem, coming from the discretization of a differential equation model, is continuous. Even in this case it may be difficult to obtain the solution, since it can, still, be *ill conditioned*.

Thus, in practice, if "well-posed" is to mean a reasonable behaviour in the numerical solution to problems, property (iii) may be substituted by *well-conditioned* when we deal with finite dimension spaces. The goal of regularization techniques is to move around the difficulties associated with the lack of smooth dependence between the input data and the solution. In some texts this is called a stability problem.

2.8 Classification of Inverse Problems

We give a brief overview of general classes of mathematical models and a discussion of several ways to classify inverse problems.

2.8.1 Classes of Mathematical Models

When investigating a phenomenon or a physical system, one of the first things to do is to characterize a mathematical model, i.e., to select a class of models. This question, which is of the utmost importance, was considered very superficially in Chapter 1.

For several purposes, the characterization of the system leads to choosing a set of functions or equations (algebraic, ordinary and/or partial differential equations, integral equations, integro-differential, algebraic-differential equations, etc), containing certain unknown constant parameters and/or functions. Of course, other classes of models can be considered, expressing, nonetheless, basic relations satisfied by the system or phenomenon under investigation, but we shall only consider the previous ones.

Taking a somewhat more general point of view, we remark that models can be either discrete or continuous, either deterministic or stochastic, and either given by a function (kinematic) or by an equation (dynamic). Each pair of these concepts, although not exaustive, are exclusive.

Even though each kind of model set forth previously distinguishes itself by its own technical tools of the trade, we just focus in their conceptual differences, as far as modeling is concerned. We may also split most of the problems between linear and nonlinear types. So, when characterizing a model, one possible first thing to do is deciding that it will be, say, a linear, discrete, stochastic, dynamic model. See Table 2.2.

From a standpoint of knowledge level, the 'dynamic' or equation model is more fundamental than the 'kinematic' or function model. The distinction between

Table 2.2 Mathematical models

categories	values	
information content	full: deterministic	lacking: stochastic
nature of variables	discrete	continuous
nature of modelization	descriptive (kinematic)	explanatory (dynamic)
mathematical structure	function	equation
mathematical property	linear	non-linear

models, that are given by a function or by an equation, is best understood through examples. One such example is given by *summing integers*.

Example 2.8. Sum of integers. Consider the 'phenomenon' resulting from progressively adding integers,

$$1,\ 1+2=3,\ 1+2+3=6,\ 10,\ 15,\ 21,\ldots$$

This phenomenon can be modeled by the function $F : \mathbb{N} \to \mathbb{N}$ such that

$$F(n) = F_n = \frac{n(n+1)}{2}, \quad \text{for all } n \in \mathbb{N}\,.$$

This is a 'kinematic' or descriptive model. The corresponding *recurrence relation*

$$F_{n+1} = F_n + (n+1), \ \text{ for } n \geq 1\,,$$

together with initial condition $F_1 = 1$, is a 'dynamic' or explanatory model of the same 'phenomenon', the sum of the first n integers. Both are discrete, deterministic models. The kinematic model is *nonlinear* and the dynamic is (nonhomogeneous) linear. ■

In the same line a simple classical example from mechanics is, perhaps, the best. We shall consider it next.

Example 2.9. Uniform motion. For uniform motion of a pointwise particle in a straight line, the kinematic model (a function) is

$$[0,\infty[\ni t \mapsto x(t) = x_0 + v_0 t \in \mathbb{R}\,,$$

where $x(t)$ represents the position at a given time t, v_0 is the constant velocity, and x_0 is the initial position (position at time $t = 0$). Clearly, this is a continuous, deterministic, linear (better, affine function), kinematic model.

The corresponding dynamic model is given by Newton's second law (differential equation),

$$m\frac{d^2x}{dt^2} = F, \quad \text{for } t > 0, \quad \text{with } F = 0\,,$$

subjected to the following initial conditions,

$$x(0) = x_0, \quad \text{and} \quad \left.\frac{dx}{dt}\right|_{t=0} = v(0) = v_0\,.$$

Here m represents the mass of the particle, and F the resultant of forces acting on it. This is a continuous, deterministic, linear, dynamic model. ■

We shall not consider here the characterization of models any further. See, however, Afterword, page 177.

2.8.2 Classes of Inverse Problems

A classification scheme of inverse problems arises from the process point of view, represented by the *black box* (see Fig. 1.1 on page 7). There, the black box set-up could represent the interaction between an *external observer* (researcher) and a *physical system*, where the observer could interact with the system by stimulating it and collecting information about its behaviour or reaction[9].

In line with what was said in Chapter 1, we can consider three general classes of problems:

P_1: ***Direct problem*** — Given an arbitrary stimulus, tell what the corresponding reaction will be;

P_2: ***Inverse reconstruction problem*** — Given the reaction, tell what stimulus produced it;

P_3: ***Inverse identification problem*** — From a data set, determine the parameters and/or functions that specify the system.

This book emphasizes the solution of inverse problems, either reconstruction or identification of models, and we shall only consider problems related to classes of models represented by:

- linear or non-linear equations in spaces of finite dimension,

  $$A(\mathbf{x}) = \mathbf{y}\,,$$

  where A is a linear or non-linear function,

  $$\mathbb{R}^n \supset \Omega \ni \mathbf{x} \mapsto A(\mathbf{x}) \in \mathbb{R}^m$$

[9] Of course, it is not always possible to have such clear cut separation between observer and physical systems. However, we shall only consider these.

and $\mathbf{y} \in \mathbb{R}^m$;

- initial and/or boundary value problems for differential or integro-differential equations.

Once a particular class of models is selected, we *solve the inverse problem* by approximately determining the value of the unknown model's constants or functions, using experimental data.

When dealing with differential equation models, we classify the inverse problem with respect to the role, in the differential equation, of the object to be estimated, which can be:

(i) initial conditions;

(ii) boundary conditions;

(iii) forcing terms;

(iv) coefficients of the equation, ie. properties of the system[10].

Here, (i) to (iii) are *reconstruction* problems and (iv) is an *identification* problem.

Moreover, we have a natural splitting of inverse problems to be considered, either the models are of finite dimension (such as a system of n equations and m unknowns) or of infinite dimension (such as an initial value problem for a partial differential equation), and if the object being estimated is of finite dimension (some parameters or constants of the model) or of infinite dimension (such as a function, or an infinite number of parameters). Problems are then classified as belonging to one of the following types:

Type I: Estimation of a finite number of parameters in a model of finite dimension;

Type II: Estimation of an infinite number of parameters or a function in a model of finite dimension;

Type III: Estimation of a finite number of parameters in a model of infinite dimension;

Type IV: Estimation of an infinite number of parameters or a function in a model of infinite dimension.

[10] Just to mention a few, the coefficients could represent material or geometrical properties as the viscosity of a fluid, the thermal conductivity of a solid material, the permeability of a porous medium, and so on. If the coefficient varies throughout the space (it would be a function, not just a constant number) one says that the medium is *heterogeneous*, otherwise, if it is constant everywhere in space, one says that the medium is *homogeneous*. If the coefficients vary with respect to the direction in space, then the medium is said to be *anisotropic*, otherwise it is said *isotropic*.

Table 2.3 Classification of inverse problems according to the dimension of the model and of the object being estimated

Estimation of quantity → Dimension of the model ↓	Finite	Infinite
Finite	Type I	Type II
Infinite	Type III	Type IV

This is summed up in Table 2.3. Inverse problems of Type I are considered in Chapters 1–4; Chapters 5–8 deal mainly with inverse problems of Types III and IV. In this book we do not consider Type II inverse problems. Section 8.1 further elaborates on this classification.

Beck [11, 10] proposed the classification of inverse problems with respect to the type of the unknown of the inverse problem, either parameters or functions. Our classification is just an extension of his, in the sense that we split his classification taking into account the dimension of the model, essentially, either finite when, typically, the model is given by an algebraic function/equation, or infinite when, most of the times, the model is a differential/integral equation. Therefore, Beck's estimation of parameters corresponds to Type I or III problems, and Beck's estimation of functions corresponds to Type II or IV inverse problems. This further splitting is justified by the increased level of mathematical complexity of going from a model of finite dimension to one of infinite dimension.

Exercises

2.1. Let f be a real function of one variable, $f : \mathbb{R} \to \mathbb{R}$. It is worthwhile to see what is the consequence of the fact that the different multiplication factors are constant. Determine the functions such that:

(a) $f'(x) = c$. What can you say about its graph?

(b) $f'(x)/f(x) = c$;

(c) $xf'(x) = c$;

(d) $xf'(x)/f(x) = c$.

Verify that the graph of a function satisfying (b), when plotted in a *log-normal* scale, is a straight line. Analogously, verify that the graphs of a function satisfying (c), in a *normal-log* scale, and of a function satisfying (d), in a *log-log* scale, are straight lines.

2.2. For functions $f(x) = \ln x$, $g(x) = e^{\alpha x}$, $h(x) = x^{\beta}$, $l(x) = \frac{1}{x-a}$ discuss the regions of well and ill-conditioning, for the four types of condition numbers defined, in general, in Eq. (2.4), or on page 30, for a scalar function of one real variable.

2.3. (a) Since $\mathcal{I} = AA^{-1}$ use Eq. (A6) to show that

$$k(A) \geq 1 \,,$$

that is, the condition of any matrix is greater than or equal to 1.

(b) Show that

$$k(A) \quad = \quad k(A^{-1}) \,,$$

that is, the condition of a matrix and of its inverse are equal.

2.4. Let $p_j : \mathbb{R} \to \mathbb{R}$, $j = 1, \ldots, N$, be differentiable functions that do not vanish at any point, i.e.,

$$p_j(x) \quad \neq \quad 0 \text{ for all } x \in \mathbb{R} \; j = 1, \ldots, N \,.$$

(a) Show that the *elasticity of the composition is the product of the elasticities*, that is,

$$E(p_1 \circ p_2) = Ep_1 \cdot Ep_2 \,.$$

(b) Show, by induction, that

$$E(p_1 \circ p_2 \circ \ldots \circ p_N) = Ep_1 \cdot Ep_2 \cdot \ldots Ep_N \,.$$

(c) Show that the *elasticity of the product is the sum of the elasticities*, that is,

$$E(p_1 \cdot p_2) = Ep_1 + Ep_2 \,.$$

(d) Show, therefore, that

$$E(p_1 \cdot p_2 \cdot \ldots \cdot p_N) = Ep_1 + Ep_2 + \ldots + Ep_N \,.$$

2.5. The *difference* function between non-negative numbers is given by

$$[0, +\infty[\times [0, +\infty[\ni (x,y) \mapsto m(x,y) \; = \; x - y \in \mathbb{R} \,.$$

(a) Show that

$$| \, Em(x,y) \, | \geq 1 \,.$$

(b) Determine and sketch the region in $[0, +\infty[\times [0, +\infty[$ satisfying

$$| \, Em(x,y) \, | \leq 2 \,.$$

Hint. Example A.1, on page 192, can be useful here.

(c) Do the same for

$$| \, Em(x,y) \, | \geq 4 \,.$$

2.6. (a) Check the assertion on the last paragraph of Section 2.4 on page 33.

(b) Let $f : \mathbb{R}^n \to \mathbb{R}$ and define the elasticity of f,

$$Ef(\mathbf{x}) = \left(\frac{x_1}{f}\frac{\partial f}{\partial x_1}, \frac{x_2}{f}\frac{\partial f}{\partial x_2}, \ldots, \frac{x_n}{f}\frac{\partial f}{\partial x_n} \right) .$$

Let $c_f(\mathbf{x})$ denote the relative to relative condition number of f. Show that $|Ef(\mathbf{x})| \leq c_f(\mathbf{x})$.

(c) Conclude that if $|Ef(\mathbf{x})| \leq 1$ then f is well-conditioned with respect to the relative to relative condition number.

2.7. Let m be the difference function as in Exercise 2.5. Compute the condition number of m, for each notion of condition number set forth in Eq. (2.4).

2.8. We should have not used the elasticity in Example 2.5. Compute the relative to relative condition number of function p_5 in that example.

2.9. Let $c_f(\mathbf{x})$ denote either one of the condition numbers of $\mathbf{f}$ as defined by Eq. (2.4). Let $\mathbf{h} = \mathbf{f} \circ \mathbf{g}$. Show that

$$c_h(\mathbf{x}) \leq c_f(\mathbf{g}(\mathbf{x}))\, c_g(\mathbf{x}) .$$

Hint. Recall chain's rule, $\mathcal{J}\mathbf{h}_x = \mathcal{J}\mathbf{f}_{g(x)}\mathcal{J}\mathbf{g}_x$.

2.10. Write down the algorithms discussed in Example 2.5 as composition of functions.

2.11. Relate normal equation (2.10) and its solution, Eq. (2.11), with the results of Exercise 1.7. (Pay attention: the roles of constants a and b are interchanged.)

2.12. QR method for the solution of least squares. From Eq. (2.11) and recalling that $A = (\mathbf{x},\mathbb{1})$ the solution of the least squares problem,

$$(\hat{a},\hat{b})^T = \operatorname{argmin}_{(a,b)} E(a,b) = \operatorname{argmin}_{(a,b)} \frac{1}{2}|\mathbf{y} - a\mathbf{x} - \mathbf{b}|^2 ,$$

is given by

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \left(A^T A\right)^{-1} A^T \mathbf{y} = \begin{pmatrix} \mathbf{x}^T\mathbf{x} & \mathbf{x}^T\mathbb{1} \\ \mathbb{1}^T\mathbf{x} & \mathbb{1}^T\mathbb{1} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}^T \\ \mathbb{1} \end{pmatrix} \mathbf{y} . \tag{2.13a}$$

Consider the function

$$\mathbb{R}^n \times \mathbb{R}^n \ni (\mathbf{x},\mathbf{y}) \overset{G}{\mapsto} \begin{pmatrix} \hat{a}(\mathbf{x},\mathbf{y}) \\ \hat{b}(\mathbf{x},\mathbf{y}) \end{pmatrix} . \tag{2.13b}$$

The condition of the algorithm to compute $(\hat{a},\hat{b})^T$ suggested by Eq. (2.13) is very high since it depends on the computation of the triple product $\left(A^TA\right)^{-1}A^T$. This undermines the stability of the algorithm.

This exercise proposes an alternative algorithm based on the method QR, [35]. Consider the vector

$$\mathbf{v} = \hat{a}\mathbf{x} + \hat{b}\mathbb{1} = (\mathbf{x}\ \mathbb{1})\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = A\left(A^TA\right)^{-1}A^T\mathbf{y}$$

which represents the vector, in the plane generated by $\mathbf{x}$ and $\mathbb{1}$, closer to $\mathbf{y}$, representing its orthogonal projection[11].

(a) Consider the function

$$\mathbf{x} \mapsto \Lambda(\mathbf{x}) = A\left(A^TA\right)^{-1}A^T$$

which represents the projection matrix from $\mathbb{R}^n$ to the space generated by the vectors $\mathbf{x}$ and $\mathbb{1}$. Show that $\Lambda(\mathbf{x} - \lambda\mathbb{1}) = \Lambda(\mathbf{x})$, for all $\lambda \in \mathbb{R}$. (From a geometrical point of view this result is expected since the space generated by $\mathbf{x}$ and $\mathbb{1}$ is the same as the space generated by $\mathbf{x} - \lambda\mathbb{1}$ and $\mathbb{1}$. Of course, we are assuming that $\mathbf{x}$ and $\mathbb{1}$ are independent, i.e., that $\mathbf{x}$ is not a multiple of $\mathbb{1}$.)

(b) In particular, choose $\lambda_\star$ such that $\mathbf{x} - \lambda_\star\mathbb{1} \perp \mathbb{1}$. Check how the quadruple product present in $\Lambda(\mathbf{x} - \lambda_\star\mathbb{1})$ simplifies.
Hint. A^TA is diagonal, therefore easily invertible.

(c) In this case, obtain a simpler expression for $\mathbf{v}$.

(d) Determine α, β such that

$$\alpha(\mathbf{x} - \lambda_\star\mathbb{1}) + \beta\mathbb{1} = \mathbf{v}\,.$$

Hint. Use item (b) and Fourier-Pythagoras trick, page 209.

(e) From the result in item (d), determine an expression for $(\hat{a},\hat{b})$, where

$$\mathbf{v} = \hat{a}\mathbf{x} + \hat{b}\mathbb{1}\,,$$

obtaining a more stable algorithm for least squares.

2.13. Using the concepts introduced in Sections 2.1 and 2.5 study the evaluation and algorithms to compute

$$f(x) = \sin x - x\,.$$

Hint. You may consider using a truncated Taylor's series of $\sin x$.

[11] Further discussion on orthogonal projections can be seen in Section A.4.1.

2.14. Heat conduction problem. Consider a one-dimensional bar, isolated in its lateral surface, being heated in one extremity, and in contact with the ambient at the temperature T_{amb} in the other extremity. Assume also that heat is being transferred to its interior. Let $T = T(x,t)$ denote its temperature on position x, at time t . Then T satisfies the following initial and boundary value problem for a partial differential equation,

$$\rho c_p \frac{\partial T}{\partial t}(x,t) = \frac{\partial}{\partial x}\left(k(T,x)\frac{\partial T}{\partial x}\right) + g(x,t)\,,$$

for $x \in [0,L]$, $t > 0$,

$$-k\frac{\partial T}{\partial x}\bigg|_{x=0} = q''(t)\,, \quad \text{for } t > 0\,,$$

(left boundary: prescribed heat flux),

$$-k\frac{\partial T}{\partial x}\bigg|_{x=L} = h\,[T(L,t) - T_{amb}]\,, \quad \text{for } t > 0\,,$$

(right boundary: contact with ambient),

$$T(x,0) = T_0(x), \quad \text{for } t > 0\,,$$

(initial condition). Here, $\rho = \rho(\mathbf{x})$ is the specific mass of the material, $c_p = c_p(x)$ is the *specific heat* of the material, $k = k(T,x)$ is the *thermal conductivity*, $g = g(x,t)$ is an internal heat source/sink, q'' is the interfacial heat flux, h is the convection coefficient, and $T_0 = T_0(\mathbf{x})$ is the initial temperature of the bar.

Classify the following problems with respect to being direct, inverse identification, or inverse reconstruction problems:

(a) Given k, g, ρ, c_p, q'', h, and T_0, determine $T = T(x,t)$;

(b) Given k, g, ρ, c_p, h, T_0, and measurements of temperature at certain times, on some locations of the bar, determine q'';

(c) Given k, ρ, c_p, q'', h, T_0, and measurements of temperature at certain times, on some locations of the bar, determine g;

(d) Given k, ρ, g, c_p, q'', h, and measurements of temperature at certain times, on some locations of the bar, determine T_0;

(e) Given ρ, g, c_p, q'', h, T_0, and measurements of temperature at certain times, on some locations of the bar, determine k;

(f) Given ρ, g, c_p, q'', T_0, and measurements of temperature at certain times, on some locations of the bar, determine k and h;

(g) Given ρ, g, c_p, h, T_0, and measurements of temperature at certain times, on some locations of the bar, determine k and q''.