国外数学名著系列(续一)

(影印版)40

Albert Tarantola

Inverse Problem Theory and Methods for Model Parameter Estimation

模型参数估计的反问题理论与方法

科学出版社
www.sciencep.com

图字：01-2008-5111

图书在版编目(CIP)数据

模型参数估计的反问题理论与方法=Inverse Problem Theory and Methods for Model Parameter Estimation / (意)塔兰托拉(Tarantola, A.)著. —影印版. —北京: 科学出版社, 2009
(国外数学名著系列; 40)
ISBN 978-7-03-023484-1
I. 模… II. 塔… III. 参数估计-逆问题-研究-英文 IV. O211.67 O175
中国版本图书馆 CIP 数据核字(2008) 第 186197 号

责任编辑: 范庆奎/责任印刷: 钱玉芬/封面设计: 黄华斌

科学出版社 出版
北京东黄城根北街 16 号
邮政编码: 100717
http://www.sciencep.com
中国科学院印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2009 年 1 月第 一 版 开本: B5(720 × 1000)
2009 年 1 月第一次印刷 印张: 22 1/2
印数: 1—2 000 字数: 431 000
定价: 75.00 元
(如有印装质量问题, 我社负责调换〈科印〉)

《国外数学名著系列》(影印版)序

要使我国的数学事业更好地发展起来，需要数学家淡泊名利并付出更艰苦地努力。另一方面，我们也要从客观上为数学家创造更有利的发展数学事业的外部环境，这主要是加强对数学事业的支持与投资力度，使数学家有较好的工作与生活条件，其中也包括改善与加强数学的出版工作。

从出版方面来讲，除了较好较快地出版我们自己的成果外，引进国外的先进出版物无疑也是十分重要与必不可少的。从数学来说，施普林格（Springer）出版社至今仍然是世界上最具权威的出版社。科学出版社影印一批他们出版的好的新书，使我国广大数学家能以较低的价格购买，特别是在边远地区工作的数学家能普遍见到这些书，无疑是对推动我国数学的科研与教学十分有益的事。

这次科学出版社购买了版权，一次影印了 23 本施普林格出版社出版的数学书，就是一件好事，也是值得继续做下去的事情。大体上分一下，这 23 本书中，包括基础数学书 5 本，应用数学书 6 本与计算数学书 12 本，其中有些书也具有交叉性质。 这些书都是很新的，2000 年以后出版的占绝大部分，共计 16 本，其余的也是 1990 年以后出版的。这些书可以使读者较快地了解数学某方面的前沿，例如基础数学中的数论、代数与拓扑三本，都是由该领域大数学家编著的“数学百科全书”的分册。对从事这方面研究的数学家了解该领域的前沿与全貌很有帮助。按照学科的特点，基础数学类的书以“经典”为主，应用和计算数学类的书以“前沿”为主。这些书的作者多数是国际知名的大数学家，例如《拓扑学》一书的作者诺维科夫是俄罗斯科学院的院士，曾获“菲尔兹奖”和“沃尔夫数学奖”。这些大数学家的著作无疑将会对我国的科研人员起到非常好的指导作用。

当然，23 本书只能涵盖数学的一部分，所以，这项工作还应该继续做下去。更进一步，有些读者面较广的好书还应该翻译成中文出版，使之有更大的读者群。

总之，我对科学出版社影印施普林格出版社的部分数学著作这一举措表示热烈的支持，并盼望这一工作取得更大的成绩。

王　元

2005 年 12 月 3 日

To my parents,
Joan and Fina

Preface

Physical theories allow us to make predictions: given a complete description of a physical system, we can predict the outcome of some measurements. This problem of predicting the result of measurements is called the *modelization problem*, the *simulation problem*, or the *forward problem*. The *inverse problem* consists of using the actual result of some measurements to infer the values of the parameters that characterize the system.

While the forward problem has (in deterministic physics) a unique solution, the inverse problem does not. As an example, consider measurements of the gravity field around a planet: given the distribution of mass inside the planet, we can uniquely predict the values of the gravity field around the planet (forward problem), but there are different distributions of mass that give *exactly* the same gravity field in the space outside the planet. Therefore, the inverse problem — of inferring the mass distribution from observations of the gravity field — has multiple solutions (in fact, an infinite number).

Because of this, in the inverse problem, one needs to make explicit any available a priori information on the model parameters. One also needs to be careful in the representation of the data uncertainties.

The most general (and simple) theory is obtained when using a probabilistic point of view, where the a priori information on the model parameters is represented by a probability distribution over the 'model space.' The theory developed here explains how this a priori probability distribution is transformed into the a posteriori probability distribution, by incorporating a physical theory (relating the model parameters to some observable parameters) and the actual result of the observations (with their uncertainties).

To develop the theory, we shall need to examine the different types of parameters that appear in physics and to be able to understand what a total absence of a priori information on a given parameter may mean.

Although the notion of the inverse problem could be based on conditional probabilities and Bayes's theorem, I choose to introduce a more general notion, that of the 'combination of states of information,' that is, in principle, free from the special difficulties appearing in the use of conditional probability densities (like the well-known Borel paradox).

The general theory has a simple (probabilistic) formulation and applies to any kind of inverse problem, including linear as well as strongly nonlinear problems. Except for very simple examples, the probabilistic formulation of the inverse problem requires a resolution in terms of 'samples' of the a posteriori probability distribution in the model space. This, in particular, means that the solution of an inverse problem is not a model but a collection of models (that are consistent with both the data and the a priori information). This is

why Monte Carlo (i.e., random) techniques are examined in this text. With the increasing availability of computer power, Monte Carlo techniques are being increasingly used.

Some special problems, where nonlinearities are weak, can be solved using special, very efficient techniques that do not differ essentially from those used, for instance, by Laplace in 1799, who introduced the 'least-absolute-values' and the 'minimax' criteria for obtaining the best solution, or by Legendre in 1801 and Gauss in 1809, who introduced the 'least-squares' criterion.

The first part of this book deals exclusively with discrete inverse problems with a finite number of parameters. Some real problems are naturally discrete, while others contain functions of a continuous variable and can be discretized if the functions under consideration are smooth enough compared to the sampling length, or if the functions can conveniently be described by their development on a truncated basis. The advantage of a discretized point of view for problems involving functions is that the mathematics is easier. The disadvantage is that some simplifications arising in a general approach can be hidden when using a discrete formulation. (Discretizing the forward problem and setting a discrete inverse problem is not always equivalent to setting a general inverse problem and discretizing for the practical computations.)

The second part of the book deals with general inverse problems, which may contain such functions as data or unknowns. As this general approach contains the discrete case in particular, the separation into two parts corresponds only to a didactical purpose.

Although this book contains a lot of mathematics, it is not a mathematical book. It tries to explain how a method of acquisition of information can be applied to the actual world, and many of the arguments are heuristic.

This book is an entirely rewritten version of a book I published long ago (Tarantola, 1987). Developments in inverse theory in recent years suggest that a new text be proposed, but that it should be organized in essentially the same way as my previous book. In this new version, I have clarified some notions, have underplayed the role of optimization techniques, and have taken Monte Carlo methods much more seriously.

I am very indebted to my colleagues (Bartolomé Coll, Georges Jobert, Klaus Mosegaard, Miguel Bosch, Guillaume Évrard, John Scales, Christophe Barnes, Frédéric Parrenin, and Bernard Valette) for illuminating discussions. I am also grateful to my collaborators at what was the *Tomography Group* at the Institut de Physique du Globe de Paris.

Albert Tarantola
Paris, June 2004

Contents

Chapter 1

The General Discrete Inverse Problem

Far better an approximate answer to the right question,
which is often vague,
than an exact answer to the wrong question,
which can always be made precise.

John W. Tukey, 1962

Central to this chapter is the concept of the 'state of information' over a parameter set. It is postulated that the most general way to describe such a state of information is to define a probability density over the parameter space. It follows that the results of the measurements of the observable parameters (data), the a priori information on model parameters, and the information on the physical correlations between observable parameters and model parameters can all be described using probability densities. The general inverse problem can then be set as a problem of 'combining' all of this information. Using the point of view developed here, the solution of inverse problems, and the analysis of uncertainty (sometimes called 'error and resolution analysis'), can be performed in a fully nonlinear way (but perhaps with a large amount of computing time). In all usual cases, the results obtained with this method reduce to those obtained from more conventional approaches.

1.1 Model Space and Data Space

Let $\mathcal{S}$ be the *physical system* under study. For instance, $\mathcal{S}$ can be a galaxy for an astrophysicist, Earth for a geophysicist, or a quantum particle for a quantum physicist.

The scientific procedure for the study of a physical system can be (rather arbitrarily) divided into the following three steps.

i) *Parameterization of the system*: discovery of a minimal set of *model parameters* whose values completely characterize the system (from a given point of view).

ii) *Forward modeling*: discovery of the *physical laws* allowing us, for given values of the model parameters, to make predictions on the results of measurements on some *observable parameters*.

iii) *Inverse modeling*: use of the actual results of some measurements of the observable parameters to infer the actual values of the model parameters.

Strong feedback exists between these steps, and a dramatic advance in one of them is usually followed by advances in the other two. While the first two steps are mainly inductive, the third step is deductive. This means that the rules of thinking that we follow in the first two steps are difficult to make explicit. On the contrary, the mathematical theory of logic (completed with probability theory) seems to apply quite well to the third step, to which this book is devoted.

1.1.1 Model Space

The choice of the model parameters to be used to describe a system is generally not unique.

Example 1.1. *An anisotropic elastic sample* $\mathcal{S}$ *is analyzed in the laboratory. To describe its elastic properties, it is possible to use the tensor* $c^{ij}{}_{k\ell}(\mathbf{x})$ *of elastic* stiffnesses *relating stress,* $\sigma^{ij}(\mathbf{x})$, *to strain,* $\varepsilon^{ij}(\mathbf{x})$, *at each point* $\mathbf{x}$ *of the solid:*

$$\sigma^{ij}(\mathbf{x}) = c^{ij}{}_{k\ell}(\mathbf{x})\, \varepsilon^{k\ell}(\mathbf{x}) \quad . \tag{1.1}$$

Alternatively, it is possible to use the tensor $s^{ij}{}_{k\ell}(\mathbf{x})$ *of elastic* compliances *relating strain to stress,*

$$\varepsilon^{ij}(\mathbf{x}) = s^{ij}{}_{k\ell}(\mathbf{x})\, \sigma^{k\ell}(\mathbf{x}) \quad , \tag{1.2}$$

where the tensor $\mathbf{s}$ *is the inverse of* $\mathbf{c}$, $c^{ij}{}_{k\ell}\, s^{k\ell}{}_{mn} = \delta^i_m\, \delta^j_n$. *The use of stiffnesses or of compliances is completely equivalent, and there is no 'natural' choice.*

A particular choice of model parameters is a *parameterization* of the system. Two different parameterizations are *equivalent* if they are related by a bijection (one-to-one mapping).

Independently of any particular parameterization, it is possible to introduce an abstract space of points, a *manifold*,[1] each point of which represents a conceivable model of the system. This manifold is named the *model space* and is denoted $\mathfrak{M}$. Individual models are points of the model space manifold and could be denoted $\mathcal{M}_1$, $\mathcal{M}_2$, ... (but we shall use another, more common, notation).

For quantitative discussions on the system, a particular parameterization has to be chosen. To define a parameterization means to define a set of experimental procedures allowing, at least in principle, us to measure a set of physical quantities that characterize the system. Once a particular parameterization has been chosen, with each point $\mathcal{M}$ of the

[1] The reader interested in the theory of differentiable manifolds may refer, for instance, to Lang (1962), Narasimhan (1968), or Boothby (1975).

model space $\mathfrak{M}$ a set of numerical values $\{m^1, \ldots, m^n\}$ is associated. This corresponds to the definition of a system of *coordinates* over the model manifold $\mathfrak{M}$.

Example 1.2. *If the elastic sample mentioned in Example* 1.1 *is, in fact, isotropic and homogeneous, the model manifold* $\mathfrak{M}$ *is two-dimensional (as such a medium is characterized by two elastic constants). As parameters to characterize the sample, one may choose, for instance,* $\{m^1, m^2\} = \{$ Young modulus , Poisson ratio $\}$ *or* $\{m^1, m^2\} = \{$ bulk modulus , shear modulus $\}$. *These two possible choices define two different coordinate systems over the model manifold* $\mathfrak{M}$.

Each point $\mathcal{M}$ of $\mathfrak{M}$ is named a *model*, and, to conform to usual notation, we may represent it using the symbol $\mathbf{m}$. By no means is $\mathbf{m}$ to be understood as a vector, i.e., as an element of a linear space. For the manifold $\mathfrak{M}$ may be linear or not, and even when the model space $\mathfrak{M}$ is linear, the coordinates being used may not be a set of Cartesian coordinates.

Example 1.3. *Let us choose to characterize the elastic samples mentioned in Example* 1.2 *using the bulk modulus and the shear modulus,* $\{m^1, m^2\} = \{\kappa, \mu\}$. *A convenient*[2] *definition of the distance between two elastic media is*

$$d = \sqrt{\left(\log \frac{\kappa_2}{\kappa_1}\right)^2 + \left(\log \frac{\mu_2}{\mu_1}\right)^2} \quad . \tag{1.3}$$

This clearly shows that the two coordinates $\{m^1, m^2\} = \{\kappa, \mu\}$ *are* not *Cartesian. Introducing the logarithmic bulk modulus* $\kappa^* = \log(\kappa/\kappa_0)$ *and the logarithmic shear modulus* $\mu^* = \log(\mu/\mu_0)$ *(where* κ_0 *and* μ_0 *are arbitrary constants) gives*

$$d = \sqrt{(\kappa_2^* - \kappa_1^*)^2 + (\mu_2^* - \mu_1^*)^2} \quad . \tag{1.4}$$

The logarithmic bulk modulus and the logarithmic shear modulus are *Cartesian coordinates over the model manifold* $\mathfrak{M}$.

The number of model parameters needed to completely describe a system may be either finite or infinite. This number is infinite, for instance, when we are interested in a property $\{m(\mathbf{x}) ; \mathbf{x} \in \mathcal{V}\}$ that depends on the position $\mathbf{x}$ inside some volume $\mathcal{V}$.

The theory of infinite-dimensional manifolds needs a greater technical vocabulary than the theory of finite-dimensional manifolds. In what follows, and in all of the first part of this book, I assume that the model space is *finite dimensional*. This limitation to systems with a finite number of parameters may be severe from a mathematical point of view. For instance, passing from a continuous field $m(\mathbf{x})$ to a discrete set of quantities $m^\alpha = m(\mathbf{x}^\alpha)$ by discretizing the space will only make sense if the considered fields are smooth. If this is indeed the case, then there will be no practical difference between the numerical results given by functional approaches and those given by discrete approaches to

[2] This definition of distance is invariant of form when changing these positive elastic parameters by their inverses, or when multiplying the values of the elastic parameters by a constant. See Appendix 6.3 for details.

inverse problem theory (although the numerical algorithms may differ considerably, as can be seen by comparing the continuous formulation in sections 5.6 and 5.7 and the discrete formulation in Problem 7.3).

Once we agree, in the first part of this book, to deal only with a finite number of parameters, it remains to decide if the parameters may take continuous or discrete values (i.e., in fact, if the quantities are real numbers or integer numbers). For instance, if a parameter m^α represents the mass of the Sun, we can assume that it can take any value from zero to infinity; if m^α represents the spin of a quantum particle, we can assume a priori that it can only take discrete values. As the use of 'delta functions' allows us to consider parameters taking discrete values as a special case of parameters taking continuous values, we shall, to simplify the discussion, use the terminology corresponding to the assumption that all the parameters under consideration take their values in a continuous set. If this is not the case in a particular problem, the reader will easily make the corresponding modifications.

When a particular parameterization of the system has been chosen, each point of $\mathcal{M}$ (i.e., each model) can be represented by a particular set of values for the model parameters $\mathbf{m} = \{m^\alpha\}$, where the index α belongs to some discrete finite index set. As we have interpreted any particular parameterization of the physical system $\mathcal{S}$ as a choice of coordinates over the manifold $\mathcal{M}$, the variables m^α can be named the *coordinates* of $\mathbf{m}$, but not the 'components' of $\mathbf{m}$, unless a linear space can be introduced. But, more often than not, the model space is not linear. For instance, when trying to estimate the geographical coordinates $\{\theta, \varphi\}$ of the (center of the) meteoritic impact that killed the dinosaurs, the model space $\mathcal{M}$ is the surface of Earth, which is intrinsically curved.

When it can be demonstrated that the model manifold $\mathcal{M}$ has no curvature, to introduce a linear (vector) space still requires a proper definition of the 'components' of vectors. When such a structure of linear space has been introduced, then we can talk about the *linear model space*, denoted $\mathbb{M}$, and, by definition, the *sum of two models*, $\mathbf{m}_1$ and $\mathbf{m}_2$, corresponds to the sum of their *components*, and the *multiplication of a model by a real number* corresponds to the multiplication of all its components:[3]

$$(\mathbf{m}_1 + \mathbf{m}_2)^\alpha = {m_1}^\alpha + {m_2}^\alpha \quad , \quad (\lambda\, \mathbf{m})^\alpha = \lambda\, m^\alpha \quad . \tag{1.5}$$

Example 1.4. *For instance, in the elastic solid considered in Example* 1.3, *to have a structure of linear (vector) space, one must select an arbitrary point of the manifold* $\{\kappa_0, \mu_0\}$ *and define the vector* $\mathbf{m} = \{m^1, m^2\}$ *whose components are*

$$m^1 = \log(\kappa/\kappa_0) \quad , \quad m^2 = \log(\mu/\mu_0) \quad . \tag{1.6}$$

Then, the distance between two models, as defined in Example 1.3, *equals* $\| \mathbf{m}_2 - \mathbf{m}_1 \|$, *the norm here being understood in its ordinary sense (for vectors in a Euclidean space).*

One must keep in mind, however, that the basic definitions of the theory developed here will not depend in any way on the assumption of the linearity of the model space. We are about to see that the only mathematical objects to be defined in order to deal with the most general formulation of inverse problems are probability distributions over the model space

[3] The index α in equation (1.5) may just be a shorthand notation for a multidimensional index (see an example in Problem 7.3). For details of array algebra see Snay (1978) or Rauhala (2002).

manifold. A probability over $\mathfrak{M}$ is a mapping that, with any subset $\mathcal{A}$ of $\mathfrak{M}$, associates a nonnegative real number, $P(\mathcal{A})$, named the probability of $\mathcal{A}$, with $P(\mathfrak{M}) = 1$. Such probability distributions can be defined over any finite-dimensional manifold $\mathfrak{M}$ (curved or linear) and irrespective of any particular parameterization of $\mathfrak{M}$, i.e., independently of any particular choice of coordinates. But if a particular coordinate system $\{m^\alpha\}$ has been chosen, it is then possible to describe a probability distribution using a probability density (and we will make extensive use of this possibility).

1.1.2 Data Space

To obtain information on model parameters, we have to perform some observations during a physical experiment, i.e., we have to perform a measurement of some observable parameters.[4]

Example 1.5. *For a nuclear physicist interested in the structure of an atomic particle, observations may consist in a measurement of the flux of particles diffused at different angles for a given incident particle flux, while for a geophysicist interested in understanding Earth's deep structure, observations may consist in recording a set of seismograms at Earth's surface.*

We can thus arrive at the abstract idea of a *data space*, which can be defined as the space of all conceivable instrumental responses. This corresponds to another manifold, the *data manifold* (or data space), which we may represent by the symbol $\mathfrak{D}$. Any conceivable (exact) result of the measurements then corresponds to a particular point $\mathcal{D}$ on the manifold $\mathfrak{D}$.

As was the case with the model manifold, it shall sometimes be possible to endow the data space with the structure of a linear manifold. When this is the case, then we can talk about the *linear data space*, denoted by $\mathbb{D}$; the coordinates $\mathbf{d} = \{d^i\}$ (where i belongs to some discrete and finite index set) are then *components*,[5] and, as usual,

$$(\mathbf{d}_1 + \mathbf{d}_2)^i = d_1{}^i + d_2{}^i \quad , \quad (r\,\mathbf{d})^i = r\,d^i \quad . \tag{1.7}$$

Each possible realization of $\mathbf{d}$ is then named a *data vector*.

1.1.3 Joint Manifold

The separation suggested above between the model parameters $\{m^\alpha\}$ and the data parameters $\{d^i\}$ is sometimes clear-cut. In other circumstances, this may require some argumentation, or may not even be desirable. It is then possible to introduce one single manifold $\mathfrak{X}$ that represents all the parameters of the problem. A point of the manifold $\mathfrak{X}$ can be represented by the symbol $\mathcal{X}$ and a system of coordinates by $\{x^A\}$.

[4] The task of experimenters is difficult not only because they have to perform measurements as accurately as possible, but, more essentially, because they have to *imagine* new experimental procedures allowing them to measure observable parameters that carry a maximum of information on the model parameters.

[5] As mentioned above for the model space, the index i here may just be a shorthand notation for a multidimensional index (see an example in Problem 7.3).

As the quantities $\{d^i\}$ were termed observable parameters and the quantities $\{m^\alpha\}$ were termed model parameters, we can call $\{x^A\}$ the *physical parameters* or simply the *parameters*. The manifold $\mathcal{X}$ is then named the *parameter manifold*.

1.2 States of Information

The probability theory developed here is self-sufficient. For good textbooks with some points in common with the present text, see Jeffreys (1939) and Jaynes (2003).

1.2.1 Definition of Probability

We are going to work with a finite-dimensional manifold $\mathcal{X}$ (for instance, the model or the data space) and the field of all its subsets $\mathcal{A}$, $\mathcal{B}$, ... These subsets can be individual points, disjoint collections of points, or contiguous collections of points (whole regions of the manifold $\mathcal{X}$). As is traditional in probability theory, a subset $\mathcal{A} \subseteq \mathcal{X}$ is called an *event*. The *union* and the *intersection* of two events $\mathcal{A}$ and $\mathcal{B}$ are respectively denoted $\mathcal{A} \cup \mathcal{B}$ and $\mathcal{A} \cap \mathcal{B}$.

The field of events is called, in technical terms, a σ-field, meaning that the complement of an event is also an event. The notion of a σ-field could allow us to introduce probability theory with great generality, but we limit ourselves here to probabilities defined over a finite-dimensional manifold.

By definition, a *measure* over the manifold $\mathcal{X}$ is an application $P(\cdot)$ that with any event $\mathcal{A}$ of $\mathcal{X}$ associates a real positive number $P(\mathcal{A})$, named the *measure of* $\mathcal{A}$, that satisfies the following two properties (Kolmogorov axioms):

- If $\mathcal{A}$ and $\mathcal{B}$ are two disjoint events, then

$$P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) \quad . \tag{1.8}$$

- There is *continuity at zero*, i.e., if a sequence $\mathcal{A}_1 \supseteq \mathcal{A}_2 \supseteq \cdots$ tends to the empty set, then $P(\mathcal{A}_i) \to 0$.

This last condition implies that the probability of the empty event is zero,

$$P(\emptyset) = 0 \quad , \tag{1.9}$$

and it immediately follows from condition (1.8) that if the two events $\mathcal{A}$ and $\mathcal{B}$ are not necessarily disjoint, then

$$P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B}) \quad . \tag{1.10}$$

The probability of the whole manifold, $P(\mathcal{X})$, is not necessarily finite. If it is, then P is termed a *probability* over $\mathcal{X}$. In that case, P is usually normalized to unity: $P(\mathcal{X}) = 1$. In what follows, the term 'probability' will be reserved for a value, like $P(\mathcal{A})$ for the probability of $\mathcal{A}$. The function $P(\cdot)$ itself will rather be called a *probability distribution*.

An important notion is that of a sample of a distribution, so let us give its formal definition. A randomly generated point $\mathcal{P} \in \mathcal{X}$ is a *sample* of a probability distribution

$P(\cdot)$ if the probability that the point $\mathcal{P}$ is generated inside any $\mathcal{A} \subset \mathfrak{X}$ equals $P(\mathcal{A})$, the probability of $\mathcal{A}$. Two points $\mathcal{P}$ and $\mathcal{Q}$ are *independent samples* if (i) both are samples and (ii) the generation of the samples is independent (i.e., if the actual place where each point has materialized is, by construction, independent of the actual place where the other point has materialized).[6]

Let P be a probability distribution over a manifold $\mathfrak{X}$ and assume that a particular coordinate system $\mathbf{x} = \{x^1, x^2, \ldots\}$ has been chosen over $\mathfrak{X}$. For any probability distribution P, there exists (Radon–Nikodym theorem) a positive function $f(\mathbf{x})$ such that, for any $\mathcal{A} \subseteq \mathfrak{X}$, $P(\mathcal{A})$ can be obtained as the integral

$$\boxed{P(\mathcal{A}) = \int_{\mathcal{A}} d\mathbf{x}\, f(\mathbf{x}) \quad ,} \tag{1.11}$$

where

$$\int_{\mathcal{A}} d\mathbf{x} \equiv \underbrace{\int dx^1 \int dx^2 \cdots}_{\text{over } \mathcal{A}} \quad . \tag{1.12}$$

Then, $f(\mathbf{x})$ is termed the *probability density* representing P (with respect to the given coordinate system). The functions representing probability densities may, in fact, be distributions, i.e., generalized functions containing in particular Dirac's delta function.

Example 1.6. *Let $\mathfrak{X}$ be the 2D surface of the sphere endowed with a system of spherical coordinates $\{\theta, \varphi\}$. The probability density*

$$f(\theta, \varphi) = \frac{\sin\theta}{4\pi} \tag{1.13}$$

associates with every region $\mathcal{A}$ of $\mathfrak{X}$ a probability that is proportional to the surface of $\mathcal{A}$. Therefore, the probability density $f(\theta, \varphi)$ is 'homogeneous' (although the function does not take constant values).

Example 1.7. *Let $\mathfrak{X} = \mathbb{R}^+$ be the positive part of the real line, and let $f(x)$ be the function $1/x$. The integral $P(x_1 < x < x_2) = \int_{x_1}^{x_2} dx\, f(x)$ then defines a measure over $\mathfrak{X}$, but not a probability (because $P(0 < x < \infty) = \infty$). The function $f(x)$ is then a* measure density *but not a probability density.*

To develop our theory, we will effectively need to consider nonnormalizable measures (i.e., measures that are not a probability). These measures cannot describe the probability of a given event $\mathcal{A}$: they can only describe the *relative probability* of two events $\mathcal{A}_1$ and $\mathcal{A}_2$. We will see that this is sufficient for our needs. To simplify the discussion, we will sometimes use the linguistic abuse of calling probability a nonnormalizable measure.

It should be noticed that, as a probability is a real number, and as the parameters $x^1, x^2, \ldots$ in general have physical dimensions, the physical dimension of a probability

[6] Many of the algorithms used to generate samples in large-dimensional spaces (like the Gibbs sampler of the Metropolis algorithm) do *not* provide independent samples.