# Probabilistic Approach to Inverse Problems*

Klaus Mosegaard† & Albert Tarantola‡

October 14, 2000

**Abstract**

In 'inverse problems' data from indirect measurements are used to estimate unknown parameters of physical systems. Uncertain data, (possibly vague) prior information on model parameters, and a physical theory relating the model parameters to the observations are the fundamental elements of any inverse problem. Using concepts from probability theory, a consistent formulation of inverse problems can be made, and, while the most general solution of the inverse problem requires extensive use of Monte Carlo methods, special hypotheses (e.g., Gaussian uncertainties) allow, in some cases, an analytical solution to part of the problem (e.g., using the method of least squares).

*Chapter for the *International Handbook of Earthquake and Engineering Seismology*, to be published by *Academic Press* for the *International Association of Seismology and Physics of the Earth Interior* (IASPEI). Appendices to this text are available on CD-ROM. Publication date: Summer 2001. The authors retain the right of publishing this text and its CD-ROM supplement, or any development of it, elsewhere.

†Niels Bohr Institute; Juliane Maries Vej 30; 2100 Copenhagen OE; Denmark; mailto:klaus@gfy.ku.dk

‡Institut de Physique du Globe; 4, place Jussieu; 75005 Paris; France; mailto:tarantola@ipgp.jussieu.fr

# Contents

# 1 Introduction

## 1.1 General Comments

Given a physical system, the 'forward' or 'direct' problem consists, by definition, in using a physical theory to predict the outcome of possible experiments. In classical physics this problem has a unique solution. For instance, given a seismic model of the whole Earth (elastic constants, attenuation, etc. at every point inside the Earth) and given a model of a seismic source, we can use current seismological theories to predict which seismograms should be observed at given locations at the Earth's surface.

The 'inverse problem' arises when we do not have a good model of the Earth, or a good model of the seismic source, but we have a set of seismograms, and we wish to use these observations to infer the internal Earth structure or a model of the source (typically we try to infer both).

There are many reasons that make the inverse problem underdetermined (non-unique). In the seismic example, two different Earth models may predict the same seismograms[1], the finite bandwidth of our data will never allow us to resolve very small features of the Earth model, and there are always experimental uncertainties that allow different models to be 'acceptable'.

The name 'inverse problem' is widely used. The authors of this text only like this name moderately, as we see the problem more as a problem of 'conjunction of states of information' (theoretical, experimental and prior information). In fact, the equations used below have a range of applicability well beyond 'inverse problems': they can be used, for instance, to predict the values of observations in a realistic situation where the parameters describing the Earth model are not 'given', but only known approximately.

We take here a probabilistic point of view. The axioms of probability theory apply to different situations. One is the traditional statistical analysis of random phenomena, another one is the description of (more or less) subjective states of information on a system. For instance, estimation of the uncertainties attached to any measurement usually involves both uses of probability theory: some uncertainties contributing to the total uncertainty are estimated using statistics, while some other uncertainties are estimated using informed scientific judgement about the quality of an instrument, about effects not explicitly taken into account, etc. The International Organization for Standardization (ISO) in *Guide to the Expression of Uncertainty in Measurement* (1993), recommends that the uncertainties evaluated by statistical methods are named 'type A' uncertainties, and those evaluated by other means (for instance, using Bayesian arguments) are named 'type B' uncertainties. It also recommends that former classifications, for instance into 'random' and 'systematic uncertainties', should be avoided. In the present text, we accept ISO's basic point of view, and extend it by downplaying the role assigned by ISO to the particular Gaussian model for uncertainties (see section 4.3) and by not assuming that the uncertainties are 'small'.

In fact, we like to think of an 'inverse' problem as merely a 'measurement'. A measurement that can be quite complex, but the basic principles and the basic equations to be used are the same for a relatively complex 'inverse problem' as for a relatively simple 'measurement'.

We do not normally use, in this text, the term 'random variable', as we assume that we have probability distributions over 'physical quantities'. This a small shift in terminology that we hope will not disorient the reader.

An important theme of this paper is *invariant formulation* of inverse problems, in the sense that solutions obtained using different, equivalent, sets of parameters should be consistent, i.e., probability densities obtained as the solution of an inverse problem, using two different set of parameters, should be related through the well known rule of multiplication by the Jacobian of the transformation.

This paper is organized as follows. After a brief historical review of inverse problem theory, with special emphasis on seismology, we give a small introduction to probability theory. In addition to being a tutorial, this introduction also aims at fixing a serious problem of classical probability, namely the non-invariant definition of conditional probability. This problem, which materializes in the so-called Borel paradox, has profound consequences for inverse problem theory.

A probabilistic formulation of inverse theory for general inverse problems (usually called 'nonlinear inverse problems') is not complete without the use of Monte Carlo methods. Section 3 is an introduction to the most versatile of these methods, the Metropolis sampler. Apart from being versatile, it also turns out to be the most natural method for implementing our probabilistic approach.

In sections 4, 5 and 6 time has come for applying probability theory and Monte Carlo methods to inverse problems. All the steps of a careful probabilistic formulations are described, including parameterization, prior information over the parameters, and experimental uncertainties. The hitherto overlooked problem of uncertain

[1] For instance, we could fit our observations with a heterogeneous but isotropic Earth model or, alternatively, with an homogeneous but anisotropic Earth.

physical laws ('forward relations') is given special attention in this text, and it is shown how this problem is profoundly linked to the resolution of the Borel paradox.

Section 7 treats the special case of the mildly nonlinear inverse problems, where deterministic (non Monte Carlo) methods can be employed. In this section, invariant forms of classical inversion formulæ are given.

## 1.2 Brief Historical Review

For a long time scientists have estimated parameters using optimization techniques. Laplace explicitly stated the least absolute values criterion. This, and the least squares criterion were later popularized by Gauss (1809). While Laplace and Gauss were mainly interested in overdetermined problems, Hadamard (1902, 1932) introduced the notion of an "ill-posed problem", that can be viewed in many cases as an underdetermined problem.

The late sixties and early seventies was a golden age for the theory of inverse problems. In this period the first uses of Monte Carlo theory to obtain Earth models were made by Keilis-Borok and Yanovskaya (1967) and by Press (1968). At about the same time, Backus and Gilbert, and Backus alone, in the years 1967–1970, made original contributions to the theory of inverse problems, focusing on the problem of obtaining an unknown *function* from discrete data. Although the resulting mathematical theory is elegant, its initial predominance over the more 'brute force' (but more powerful) Monte Carlo theory was only possibly due to the quite limited capacities of the computers at that time. It is our feeling that Monte Carlo methods will play a more important role in the future (and this is the reason why we put emphasis on these methods in this article). An investigation of the connection between analogue models, discrete models and Monte Carlo models can be found in a paper by Kennett and Nolet (1978).

Important developments of inverse theory in the fertile period around 1970 were also made by Wiggins (1969), with his method of suppressing 'small eigenvalues', and by Franklin (1970), by introducing the right mathematical setting for the Gaussian, functional (i.e., infinite dimensional) inverse problem (see also Lehtinen et al., 1989). Other important papers from the period are Gilbert (1971) and Wiggins (1972).

A reference that may interest some readers is Parzen et al. (1998), where the probabilistic approach of Akaike is described.

To the 'regularizing techniques' of Tikhonov (1963), Levenberg (1944) and Marquardt (1970), we prefer, in this paper, the approach where the a priori information is used explicitly.

For seismologists, the first bona fide solution of an inverse problem was the estimation of the hypocenter coordinates of an earthquake using the 'Geiger method' (Geiger, 1910), that present-day computers have made practical. In fact, seismologists have been the originators of the theory of inverse problems (for data interpretation), and this is because the problem of understanding the structure of the Earth's interior using only surface data is a difficult problem.

3-D tomography of the Earth, using travel times of seismic waves, was developed by Keiiti Aki and his coworkers in a couple of well known papers (Aki and Lee, 1976; Aki, Christofferson and Husebye 1977). Minster and Jordan (1978) applied the theory of inverse problems to the reconstruction of the tectonic plate motions, introducing the concept of 'data importance'. Later, tomographic studies have provided spectacular images of the Earth's interior. Interesting papers on these inversions are van der Hilst et al. (1997) and Su et al. (1992).

One of the major current challenges in seismic inversion is the nonlinearity of wave field inversions. This is accentuated by the fact that major experiments in the future most likely will allow us to sample the whole seismic wave field. For the low frequencies wave field inversion is linear. Dahlen (1976) investigated the influence of lateral heterogeneity on the free oscillations. He showed that the the inverse problem of estimating lateral heterogeneity of even degree from multiplet variance and skewance is linear. At the time this was published, data accuracy and unknown ellipticity splitting parameters hindered its application to real data, but later developments, including the works of Woodhouse and Dahlen (1978) on discontinuous Earth models, led to present-days successful inversions of low frequency seismograms. In this connection the works of Woodhouse, Dziewonski and others spring to mind[2]. Later, the first attempts to go to higher frequencies and nonlinear inversion were made by Nolet et al. (1986), and Nolet (1990)

Purely probabilistic formulations of inverse theory saw the light around 1970 (see, for instance, Kimeldorf and Wahba, 1970). In an interesting paper, Rietsch (1977) made nontrivial use of the notion of a 'noninformative' prior distribution for positive parameters. Jackson (1979) explicitly introduced prior information in the context of linear inverse problems, an approach that was generalized by Tarantola and Valette (1982a, 1982b) to nonlinear problems.

[2]Preliminary Earth Reference Model (PREM), Dziewonski and Anderson, PEPI, 1981. Inversion for Centroid Moment Tensor (CMT), Dziewonski, Chou and Woodhouse, JGR, 1982. First global tomographic model, Dziewonski, JGR, 1984.

There are three monographs in the area of Inverse Problems (from the view point of data interpretation). In Tarantola (1987), the general, probabilistic formulation for nonlinear inverse problems is proposed. The small book by Menke (1984) covers several viewpoints on discrete, linear and nonlinear inverse problems, and is easy to read. Finally, Parker (1994) exposes his view of the general theory of linear problems.

Recently, the interest in Monte Carlo methods, for the solution of Inverse Problems, has been increasing. Mosegaard and Tarantola (1995) proposed a generalization of the Metropolis algorithm (Metropolis et al., 1953) for analysis of general inverse problems, introducing explicitly prior probability distributions, and they applied the theory to a synthetic numerical example. Monte Carlo analysis was recently applied to real data inverse problems by Mosegaard et al. (1997), Dahl-Jensen et al. (1998), Mosegaard and Rygaard-Hjalsted (1999), and Khan et al. (2000).

# 2 Elements of Probability

Probability theory is essential to our formulation of inverse theory. This chapter therefore contains a review of important elements of probability theory, with special emphasis on results that are important for the analysis of inverse problems. Of particular importance is our explicit introduction of *distance* and *volume* in data and model spaces. This has profound consequences for the notion of *conditional probability density* which plays an important role in probabilistic inverse theory.

Also, we replace the concept of conditional probability by the more general notion of 'conjunction' of probabilities, this allowing us to address the more general problem where not only the data, but also the physical laws, are uncertain.

## 2.1 Volume

Let us consider an abstract space $\mathcal{S}$, where a point $\mathbf{x}$ is represented by some coordinates $\{x^1, x^2, \ldots\}$, and let $\mathcal{A}$ be some region (subspace) of $\mathcal{S}$. The measure associating a volume $V(\mathcal{A})$ to any region $\mathcal{A}$ of $\mathcal{S}$ will be denoted the *volume measure*

$$V(\mathcal{A}) \;=\; \int_{\mathcal{A}} d\mathbf{x}\; v(\mathbf{x}) \;, \tag{1}$$

where the function $v(\mathbf{x})$ is the *volume density*, and where we write $d\mathbf{x} = dx^1 dx^2 \ldots$ The *volume element* is then[3]

$$dV(\mathbf{x}) \;=\; v(\mathbf{x})\; d\mathbf{x} \;, \tag{2}$$

and we may write $V(\mathcal{A}) = \int_{\mathcal{A}} dV(\mathbf{x})$. A manifold is called a *metric manifold* if there is a definition of distance between points, such that the distance $ds$ between the point of coordinates $\{x^i\}$ and the point of coordinates $\{x^i + dx^i\}$ can be expressed as[4]

$$ds^2 \;=\; g_{ij}(\mathbf{x})\; dx^i\; dx^j \quad , \tag{3}$$

i.e., if the notion of distance is 'of the $L_2$ type'[5]. The matrix whose entries are $g_{ij}$ is the *metric matrix*, and an important result of differential geometry and integration theory is that the volume density of the space, $v(\mathbf{x})$, equals the square root of the determinant of the metric:

$$v(\mathbf{x}) \;=\; \sqrt{\det \mathbf{g}(\mathbf{x})} \quad . \tag{4}$$

**Example 1** *In the Euclidean 3D space, using spherical coordinates, the distance element is* $ds^2 = dr^2 + r^2\, d\theta^2 + r^2 \sin^2\theta\, d\varphi^2$ *, from where it follows that the metric matrix is*

$$\begin{pmatrix} g_{rr} & g_{r\theta} & g_{r\varphi} \\ g_{\theta r} & g_{\theta\theta} & g_{\theta\varphi} \\ g_{\varphi r} & g_{\varphi\theta} & g_{\varphi\varphi} \end{pmatrix} \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2\theta \end{pmatrix} \quad . \tag{5}$$

[3] The capacity element associated to the vector elements $d\mathbf{r}_1, d\mathbf{r}_2 \ldots d\mathbf{r}_n$ is defined as $d\tau = \varepsilon_{ij\ldots k}\, dr_1^i\, dr_2^j \ldots dr_n^k$, where $\varepsilon_{ij\ldots k}$ is the Levi-Civita capacity (whose components take the values $\{0, \pm 1\}$). If the metric tensor of the space is $\mathbf{g}(\mathbf{x})$, then $\eta_{ij\ldots k} = \sqrt{\det \mathbf{g}}\; \varepsilon_{ij\ldots k}$ is a true tensor, as it is the product of a density $\sqrt{\det \mathbf{g}}$ by a capacity $\varepsilon_{ij\ldots k}$. Then, the volume element, defined as $dV = \eta_{ij\ldots k}\, dr_1^i\, dr_2^j \ldots dr_n^k = \sqrt{\det \mathbf{g}}\; d\tau$, is a (true) scalar.

[4] This is a property that is valid for any coordinate system that can be chosen over the space.

[5] As a counterexample, the distance defined as $ds = |dx| + |dy|$ is not of the $L_2$ type (it is $L_1$).

*The volume density equals the metric determinant* $v(r,\theta,\varphi) = \sqrt{\det \mathbf{g}(r,\theta,\varphi)} = r^2 \sin\theta$ *and therefore the volume element is* $dV(r,\vartheta,\varphi) = v(r,\vartheta,\varphi)\, dr\, d\vartheta\, d\varphi = r^2 \sin\theta\, dr\, d\vartheta\, d\varphi$ . [END OF EXAMPLE.]

Assume that we have defined over the space, not only the volume $V(\mathcal{A})$ of a region $\mathcal{A}$ of the space, but also its *probability* $P(\mathcal{A})$ , that is assumed to satisfy the Kolmogorov axioms (Kolmogorov, 1933). This probability is assumed to be descriptible in terms of a *probability density* $f(x)$ through the expression

$$P(\mathcal{A}) \;=\; \int_{\mathcal{A}} d\mathbf{x}\; f(\mathbf{x}). \tag{6}$$

It is well known that, in a change of coordinates over the space, a probability density changes its value: it is multiplied by the Jacobian of the transformation (this is the *Jacobian rule*). Normally, the probability of the whole space is normalized to one. If it is not normalizable, we do not say that we have a probability, but a 'measure'. We can state here the

**Postulate 1** *Given a space* $\mathcal{X}$ *over which a volume measure* $V(\,\cdot\,)$ *is defined. Any other measure (normalizable or not)* $M(\,\cdot\,)$ *considered over* $\mathcal{X}$ *is* absolutely continuous *with respect to* $V(\,\cdot\,)$ *, i.e., the measure* $M(\mathcal{A})$ *of any region* $\mathcal{A} \subset \mathcal{X}$ *with vanishing volume must be zero:* $V(\mathcal{A}) = 0 \;\;\Rightarrow\;\; M(\mathcal{A}) = 0$ .

## 2.2 Homogeneous Probability Distributions

In some parameter spaces, there is an obvious definition of distance between points, and therefore of volume. For instance, in the 3D Euclidean space the distance between two points is just the Euclidean distance (which is invariant under translations and rotations). Should we choose to parameterize the position of a point by its Cartesian coordinates $\{x,y,z\}$ , the volume element in the space would be $dV(x,y,z) \;=\; dx\, dy\, dz$ , while if we choose to use geographical coordinates, the volume element would be $dV(r,\theta,\varphi) \;=\; r^2 \sin\theta\, dr\, d\vartheta\, d\varphi$ .

**Definition.** *The* homogeneous probability distribution *is the probability distribution that assigns to each region of the space a probability proportional to the volume of the region.*

Then, which probability density represents such a homogeneous probability distribution? Let us give the answer in three steps.

- If we use Cartesian coordinates $\{x,y,z\}$ , as we have $dV(x,y,z) \;=\; dx\, dy\, dz$ , the probability density representing the homogeneous probability distribution is constant: $f(x,y,z) \;=\; k$ .
- If we use geographical coordinates $\{r,\theta,\varphi\}$ , as we have $dV(r,\theta,\varphi) \;=\; r^2 \sin\theta\, dr\, d\theta\, d\varphi$ , the probability density representing the homogeneous probability distribution is $g(r,\theta,\varphi) \;=\; k\, r^2 \sin\theta$ .
- Finally, if we use an arbitrary system of coordinates $\{u,v,w\}$ , in which the volume element of the space is $dV(u,v,w) = v(u,v,w)\, du\, dv\, dw$ , the homogeneous probability distribution is represented by the probability density $h(u,v,w) \;=\; k\, v(u,v,w)$ .

This is obviously true, since if we calculate the probability of a region $\mathcal{A}$ of the space, with volume $V(\mathcal{A})$ , we get a number proportional to $V(\mathcal{A})$ .

From these observations we can arrive at conclusions that are of general validity. First, the homogeneous probability distribution over some space is represented by a constant probability density **only** if the space is flat (in which case rectilinear systems of coordinates exist) and if we use Cartesian (or rectilinear) coordinates. The other conclusions can be stated as rules:

**Rule 1** *The probability density representing the homogeneous probability distribution is easily obtained if the expression of the volume element* $dV(u_1,u_2,\ldots) = v(u_1,u_2,\ldots)\, du_1\, du_2\, \ldots$ *of the space is known, as it is then given by* $h(u_1,u_2,\ldots) = k\, v(u_1,u_2,\ldots)$ *, where* $k$ *is a proportionality constant (that may have physical dimensions).*

**Rule 2** *If there is a metric* $g_{ij}(u_1,u_2,\ldots)$ *in the space, then the volume element is given by* $dV(u_1,u_2,\ldots) = \sqrt{\det \mathbf{g}(u_1,u_2,\ldots)}\, du_1\, du_2\, \ldots$ *, i.e., we have* $v(u_1,u_2,\ldots) = \sqrt{\det \mathbf{g}(u_1,u_2,\ldots)}$ *. The probability density representing the homogeneous probability distribution is, then,* $h(u_1,u_2,\ldots) = k\, \sqrt{\det \mathbf{g}(u_1,u_2,\ldots)}$ .

**Rule 3** *If the expression of the probability density representing the homogeneous probability distribution is known in one system of coordinates, then it is known in any other system of coordinates, through the Jacobian rule.*

Indeed, in the expression above, $g(r,\theta,\varphi) = k\, r^2 \sin\theta$ , we recognize the Jacobian between the geographical and the Cartesian coordinates (where the probability density is constant).

For short, when we say *the homogeneous probability density* we mean *the probability density representing the homogeneous probability distribution*. **One should remember that, in general, the homogeneous probability density is *not* constant**.

Let us now examine 'positive parameters', like a temperature, a period, or a seismic wave propagation velocity. One of the properties of the parameters we have in mind is that they occur in pairs of mutually reciprocal parameters:

| | | | | |
|---|---|---|---|---|
| Period | $T = 1/\nu$ | ; | Frequency | $\nu = 1/T$ |
| Resistivity | $\rho = 1/\sigma$ | ; | Conductivity | $\sigma = 1/\rho$ |
| Temperature | $T = 1/(k\beta)$ | ; | Thermodynamic parameter | $\beta = 1/(kT)$ |
| Mass density | $\rho = 1/\ell$ | ; | Lightness | $\ell = 1/\rho$ |
| Compressibility | $\gamma = 1/\kappa$ | ; | Bulk modulus (uncompressibility) | $\kappa = 1/\gamma$ |
| Wave velocity | $c = 1/n$ | ; | Wave slowness | $n = 1/c$ . |

When working with physical theories, one may freely choose one of these parameters or its reciprocal.

Sometimes these pairs of equivalent parameters come from a definition, like when we define frequency $\nu$ as a function of the period $T$ , by $\nu = 1/T$ . Sometimes these parameters arise when analyzing an idealized physical system. For instance, Hooke's law, relating stress $\sigma_{ij}$ to strain $\varepsilon_{ij}$ can be expressed as $\sigma_{ij} = {c_{ij}}^{k\ell}\, \varepsilon_{k\ell}$ , thus introducing the stiffness tensor $c_{ijk\ell}$ , or as $\varepsilon_{ij} = {d_{ij}}^{k\ell}\, \sigma_{k\ell}$ , thus introducing the compliance tensor $d_{ijk\ell}$ , the inverse of the stiffness tensor. Then the respective eigenvalues of these two tensors belong to the class of scalars analyzed here.

Let us take, as an example, the pair conductivity-resistivity (this may be thermal, electric, etc.). Assume we have two samples in the laboratory $S_1$ and $S_2$ whose resistivities are respectively $\rho_1$ and $\rho_2$ . Correspondingly, their conductivities are $\sigma_1 = 1/\rho_1$ and $\sigma_2 = 1/\rho_2$ . How should we define the 'distance' between the 'electrical properties' of the two samples? As we have $|\rho_2 - \rho_1| \neq |\sigma_2 - \sigma_1|$ , choosing one of the two expressions as the 'distance' would be arbitrary. Consider the following definition of 'distance' between the two samples

$$D(S_1, S_2) \;=\; \left| \log \frac{\rho_2}{\rho_1} \right| = \left| \log \frac{\sigma_2}{\sigma_1} \right| \quad . \tag{7}$$

This definition (i) treats symmetrically the two equivalent parameters $\rho$ and $\sigma$ and, more importantly, (ii) has an *invariance of scale* (what matters is how many 'octaves' we have between the two values, not the plain difference between the values). In fact, it is the only definition of distance between the two samples $S_1$ and $S_2$ that has an invariance of scale and is additive (i.e., $D(S_1, S_2) + D(S_2, S_3) = D(S_1, S_3)$ ).

Associated to the distance $D(x_1, x_2) = |\log\,(x_2/x_1)\,|$ is the distance element (differential form of the distance)

$$dL(x) \;=\; \frac{dx}{x} \quad . \tag{8}$$

This being a 'one-dimensional volume' we can apply now the rule 1 above to get the expression of the homogeneous probability density for such a positive parameter:

$$f(x) \;=\; \frac{k}{x} \quad . \tag{9}$$

Defining the reciprocal parameter $y = 1/x$ and using the Jacobian rule we arrive at the homogeneous probability density for $y$ :

$$g(y) \;=\; \frac{k}{y} \quad . \tag{10}$$

These two probability densities have the same form: the two reciprocal parameters are treated symmetrically. Introducing the logarithmic parameters

$$x^* \;=\; \log \frac{x}{x_0} \qquad ; \qquad y^* \;=\; \log \frac{y}{y_0} \quad , \tag{11}$$

where $x_0$ and $y_0$ are arbitrary positive constants, and using the Jacobian rule, we arrive at the homogeneous probability densities

$$f'(x^*) \;=\; k \qquad ; \qquad g'(y^*) \;=\; k \quad . \tag{12}$$

This shows that the logarithm of a positive parameter (of the type considered above) is a 'Cartesian' parameter. In fact, it is the consideration of equations 12, together with the Jacobian rule, that allows full understanding of the (homogeneous) probability densities 9–10.

The association of the probability density $f(u) = k/u$ to positive parameters was first made by Jeffreys (1939). To honor him, we propose to use the term *Jeffreys parameters* for all the parameters of the type considered above. The $1/u$ probability density was advocated by Jaynes (1968), and a nontrivial use of it was made by Rietsch (1977) in the context of inverse problems.

**Rule 4** *The homogeneous probability density for a Jeffreys quantity* $u$ *is* $f(u) = k/u$ .

**Rule 5** *The homogeneous probability density for a 'Cartesian parameter'* $u$ *(like the logarithm of a Jeffreys parameter, an actual Cartesian coordinate in an Euclidean space, or the Newtonian time coordinate) is* $f(u) = k$ . *The homogeneous probability density for an angle describing the position of a point in a circle is also constant.*

If a parameter $u$ is a Jeffreys parameter with the homogeneous probability density $f(u) = k/u$ , then its inverse, its square, and, in general, any power of the parameter is also a Jeffreys parameter, as it can easily be seen using the Jacobian rule.

**Rule 6** *Any power of a Jeffreys quantity (including its inverse) is a Jeffreys quantity.*

It is important to recognize when we do **not** face a Jeffreys parameter. Among the many parameters used in the literature to describe an isotropic linear elastic medium we find parameters like the Lamé's coefficients $\lambda$ and $\mu$ , the bulk modulus $\kappa$ , the Poisson ratio $\sigma$ , etc. A simple inspection of the theoretical range of variation of these parameters shows that the first Lamé parameter $\lambda$ and the Poisson ratio $\sigma$ may take negative values, so they are certainly not Jeffreys parameters. In contrast, Hooke's law $\sigma_{ij} = c_{ijk\ell}\,\varepsilon^{k\ell}$ , defining a linearity between stress $\sigma_{ij}$ and strain $\varepsilon_{ij}$ , defines the positive definite stiffness tensor $c_{ijk\ell}$ or, if we write $\varepsilon_{ij} = d_{ijk\ell}\,\sigma^{k\ell}$ , defines its inverse, the compliance tensor $d_{ijk\ell}$ . The two reciprocal tensors $c_{ijk\ell}$ and $d_{ijk\ell}$ are 'Jeffreys tensors'. This is a notion whose development is beyond the scope of this paper, but we can give the following rule:

**Rule 7** *The eigenvalues of a Jeffreys tensor are Jeffreys quantities*[6].

As the two (different) eigenvalues of the stiffness tensor $c_{ijk\ell}$ are $\lambda_\kappa = 3\kappa$ (with multiplicity 1) and $\lambda_\mu = 2\mu$ (with multiplicity 5) , we see that the incompressibility modulus $\kappa$ and the shear modulus $\mu$ are Jeffreys parameters[7] (as are any parameter proportional to them, or any power of them, including the inverses). If for some reason, instead of working with $\kappa$ and $\mu$ , we wish to work with other elastic parameters, like for instance the Young modulus $Y$ and the Poisson ratio $\sigma$ , or the two elastic wave velocities, then the homogeneous probability distribution must be found using the Jacobian of the transformation (see appendix H).

Some probability densities have conspicuous 'dispersion parameters', like the $\sigma$'s in the normal probability density $f(x) = k\,\exp\left(-\frac{(x-x_0)^2}{2\,\sigma^2}\right)$ , in the lognormal probability $g(X) = \frac{k}{X}\,\exp\left(-\frac{(\log\ X/X_0)^2}{2\,\sigma^2}\right)$ or in the Fisher probability density (Fischer, 1953) $h(\vartheta,\varphi) = k\,\sin\theta\,\exp\left(\cos\theta\ /\ \sigma^2\right)$ . A consistent probability model requires that when the dispersion parameter $\sigma$ tends to infinity, the probability density tends to the homogeneous probability distribution. For instance, in the three examples just given, $f(x) \to k$ , $g(X) \to k/X$ and $h(\theta,\varphi) \to k\,\sin\theta$ , which are the respective homogeneous probability densities for a Cartesian quantity, a Jeffreys quantity and the geographical coordinates on the surface of the sphere. We can state the

**Rule 8** *If a probability density has some 'dispersion parameters', then, in the limit where the dispersion parameters tend to infinity, the probability density must tend to the homogeneous one.*

[6]This solves the complete problem for isotropic tensors only. It is beyond the scope of this text to propose rules valid for general anisotropic tensors: the necessary mathematics have not yet been developed.

[7]The definition of the elastic constants was made before the tensorial structure of the theory was understood. Seismologists today should not use, at a theoretical level, parameters like the first Lamé coefficient $\lambda$ or the Poisson ratio. Instead they should use $\kappa$ and $\mu$ (and their inverses). In fact, our suggestion in this IASPEI volume is to use the true eigenvalues of the stiffness tensor, $\lambda_\kappa = 3\kappa$ , and $\lambda_\mu = 2\mu$ , which we propose to call the *eigen-bulk-modulus* and the *eigen-shear-modulus*, respectively.

As an example, using the normal probability density $f(x) = k \exp\left(-\frac{(x-x_0)^2}{2\,\sigma^2}\right)$ , for a Jeffreys parameter is not consistent. Note that it would assign a finite probability to negative values of a positive parameter that, by definition, is positive. More technically, this would violate our postulate 1. Using the log-normal probability density for a Jeffreys parameter is consistent.

There is a problem of terminology in the Bayesian literature. The homogeneous probability distribution is a very special distribution. When the problem of selecting a 'prior' probability distribution arises in the absence of any information, except the fundamental symmetries of the problem, one may select as prior probability distribution the homogeneous distribution. But enthusiastic Bayesians do not call it 'homogeneous', but 'noninformative'. We cannot recommend using this terminology. The homogeneous probability distribution is as informative as any other distribution, it is just the homogeneous one (see appendix D).

In general, each time we consider an abstract parameter space, each point being represented by some parameters $\mathbf{x} = \{x^1, x^2 \dots x^n\}$ , we will start by solving the (sometimes nontrivial) problem of defining a distance between points that respects the necessary symmetries of the problem. Only exceptionally this distance will be a quadratic expression of the parameters (coordinates) being used (i.e., only exceptionally our parameters will correspond to 'Cartesian coordinates' in the space). From this distance, a volume element $dV(\mathbf{x}) = v(\mathbf{x})\, d\mathbf{x}$ will be deduced, from where the expression $f(\mathbf{x}) = k\, v(\mathbf{x})$ of the homogeneous probability density will follow. Sometimes, we can directly define volume element, without the need of a distance. We emphasize the need of defining a distance —or a volume element— in the parameter space, from which the notion of homogeneity will follow. With this point of view we slightly depart from the original work by Jeffreys and Jaynes.

## 2.3 Conjunction of Probabilities

We shall here consider two probability distributions $P$ and $Q$ . We say that a probability $R$ is a product of the two given probabilities, and is denoted $(P \wedge Q)$ if

- $P \wedge Q = Q \wedge P$ ;
- for any subset $\mathcal{A}$ , $(P \wedge Q)(\mathcal{A}) \neq 0 \qquad \Longrightarrow \qquad P(\mathcal{A}) \neq 0 \quad \text{and} \quad Q(\mathcal{A}) \neq 0$ ;
- if $M$ denotes the homogeneous probability distribution, then $P \wedge M = P$ .

The realization of these conditions leading to the simplest results can easily be expressed using probability densities (see appendix G for details). If the two probabilities $P$ and $Q$ are represented by the two probability densities $p(\mathbf{x})$ and $q(\mathbf{x})$ , respectively, and if the homogeneous probability density is represented by $\mu(\mathbf{x})$ , then the probability $P \wedge Q$ is represented by a probability density, denoted $(p \wedge q)(\mathbf{x})$ , that is given by

$$(p \wedge q)(\mathbf{x}) \;=\; k\,\frac{p(\mathbf{x})\,q(\mathbf{x})}{\mu(\mathbf{x})} \quad , \tag{13}$$

where $k$ is a normalization constant[8].

The two left columns of figure 1 represent these probability densities.

**Example 2** *On the surface of the Earth, using geographical coordinates (latitude $\vartheta$ and longitude $\varphi$ ), the homogeneous probability distribution is represented by the probability density $\mu(\vartheta,\varphi) = \frac{1}{4\pi}\cos\vartheta$ . An estimation of the position of a floating object at the surface of the sea by an airplane navigator gives a probability distribution for the position of the object corresponding to the probability density $p(\vartheta,\varphi)$ , and an independent, simultaneous estimation of the position by another airplane navigator gives a probability distribution corresponding to the probability density $q(\vartheta,\varphi)$ . How do we 'combine' the two probability densities $p(\vartheta,\varphi)$ and $q(\vartheta,\varphi)$ to obtain a 'resulting' probability density? The answer is given by the conjunction of the two probability densities:*

$$(p \wedge q)(\vartheta,\varphi) \;=\; k\,\frac{p(\vartheta,\varphi)\,q(\vartheta,\varphi)}{\mu(\vartheta,\varphi)} \quad . \tag{14}$$

[END OF EXAMPLE.]

[8] Assume that $p(\mathbf{x})$ and $q(\mathbf{x})$ are normalized by $\int_{\mathcal{X}} d\mathbf{x}\; p(\mathbf{x}) = 1$ and $\int_{\mathcal{X}} d\mathbf{x}\; q(\mathbf{x}) = 1$ . Then, irrespective of the normalizability of $\mu(\mathbf{x})$ (as explained above, $p(\mathbf{x})$ and $q(\mathbf{x})$ are assumed to be absolutely continuous with respect to the homogeneous distribution), $(p \wedge q)(\mathbf{x})$ is normalizable, and its normalized expression is $(p \wedge q)(\mathbf{x}) = \frac{p(\mathbf{x})\,q(\mathbf{x})/\mu(\mathbf{x})}{\int_{\mathcal{X}} d\mathbf{x}\; p(\mathbf{x})\,q(\mathbf{x})/\mu(\mathbf{x})}$ .

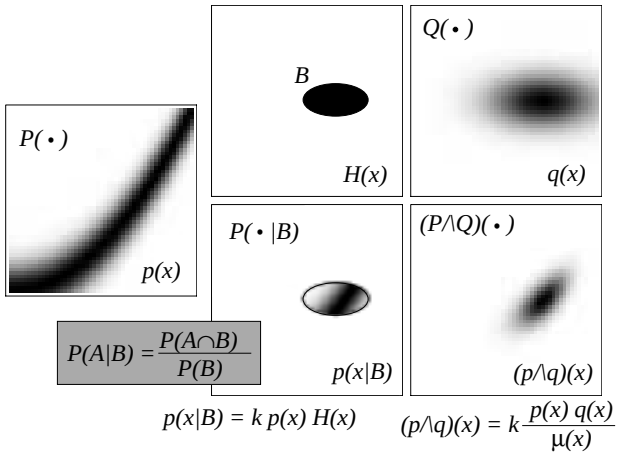



Figure 1: The two left columns of the figure illustrate the definition of conditional probability (see text for details). The right of the figure explains that the definition of the AND operation is a generalization of the notion of conditional probability. While a conditional probability combines a probability distribution $P(\,\cdot\,)$ with an 'event' $\mathcal{B}$, the AND operation combines two probability distributions $P(\,\cdot\,)$ and $Q(\,\cdot\,)$ defined over the same space. See text for a detailed explanation.

We emphasize here the following:

**Example 2 is at the basis of the paradigm that we use below to solve inverse problems.**

More generally, the conjunction of the probability densities $f_1(\mathbf{x})$, $f_2(\mathbf{x})$ ... is

$$h(\mathbf{x}) \;=\; (f_1 \wedge f_2 \wedge f_3 \ldots)(\mathbf{x}) \ldots \;=\; k\,\mu(\mathbf{x})\,\frac{f_1(\mathbf{x})}{\mu(\mathbf{x})}\,\frac{f_2(\mathbf{x})}{\mu(\mathbf{x})}\,\frac{f_3(\mathbf{x})}{\mu(\mathbf{x})}\;\ldots \quad . \tag{15}$$

For a formalization of the notion of conjunction of probabilities, the reader is invited to read appendix G.

## 2.4 Conditional Probability Density

Given a probability distribution over a space $\mathcal{X}$, represented by the probability density $f(\mathbf{x})$, and given a subspace $\mathcal{B}$ of $\mathcal{X}$ of lower dimension, can we, in a consistent way, infer a probability distribution over $\mathcal{B}$, represented by a probability density $f(\mathbf{x}|\mathcal{B})$ (to be named the conditional probability density 'given $\mathcal{B}$')?

The answer is: using only the elements given, NO, THIS IS NOT POSSIBLE.

The usual way to induce a probability distribution on a subspace of lower dimension is to assign a 'thickness' to the subspace $\mathcal{B}$, to apply the general definition of conditional probability (this time to a region of $\mathcal{X}$, not to a subspace of it) and to take the limit when the 'thickness' tends to zero. But, as suggested in figure 2, there are infinitely many ways to take this limit, each defining a different 'conditional probability density' on $\mathcal{B}$. Among the infinitely many ways to define a conditional probability density there is one that is based on the notion of distance between points in the space, and therefore corresponds to an intrinsic definition (see figure 2).

Assume that the space $\mathcal{U}$ has $p$ dimensions, the space $\mathcal{V}$ has $q$ dimensions and define in the $(p+q)$-dimensional space $\mathcal{X} = (\mathcal{U}, \mathcal{V})$ a $p$-dimensional subspace by the $p$ relations

$$\begin{array}{rcl} v_1 &=& v_1(u_1, u_2, \ldots, u_p) \\ v_2 &=& v_2(u_1, u_2, \ldots, u_p) \\ \ldots &=& \ldots \\ v_q &=& v_q(u_1, u_2, \ldots, u_p) \quad . \end{array} \tag{16}$$

The restriction of a probability distribution, represented by the probability density $f(\mathbf{x}) = f(\mathbf{u}, \mathbf{v})$ into the subspace defined by the constraint $\mathbf{v} = \mathbf{v}(\mathbf{u})$, can be defined with all generality when it is assumed that we have a metric defined over the $(p+q)$-dimensional space $\mathcal{X} = (\mathcal{U}, \mathcal{V})$. Let us limit here to the special circumstance (useful for a vast majority of inverse problems[9]) where there the $(p+q)$-dimensional space $\mathcal{X}$ is built as the Cartesian product of $\mathcal{U}$ and $\mathcal{V}$ (then we write, as usual, $\mathcal{X} = \mathcal{U} \times \mathcal{V}$). In this case, there is a metric $\mathbf{g}_u$ over $\mathcal{U}$, with associated volume element $dV_u(\mathbf{u}) = \sqrt{\det \mathbf{g}_u}\; d\mathbf{u}$, there is a metric $\mathbf{g}_v$ over $\mathcal{V}$, with associated volume element $dV_v(\mathbf{v}) = \sqrt{\det \mathbf{g}_v}\; d\mathbf{v}$, and the global volume element is simply $dV(\mathbf{u}, \mathbf{v}) = dV_u(\mathbf{u})\; dV_v(\mathbf{v})$.

The restriction of the probability distribution represented by the probability density $f(\mathbf{u}, \mathbf{v})$ on the subspace $\mathbf{v} = \mathbf{v}(\mathbf{u})$ (i.e., the conditional probability density given $\mathbf{v} = \mathbf{v}(\mathbf{u})$) is a probability distribution *on* the submanifold

[9] As a counter example, working at the surface of the sphere with geographical coordinates $(\mathbf{u}, \mathbf{v}) = (u, v) = (\vartheta, \varphi)$ this condition is **not** fulfilled, as $g_\varphi = \sin\theta$ is a function of $\vartheta$ : the surface of the sphere is not the Cartesian product of two 1D spaces.

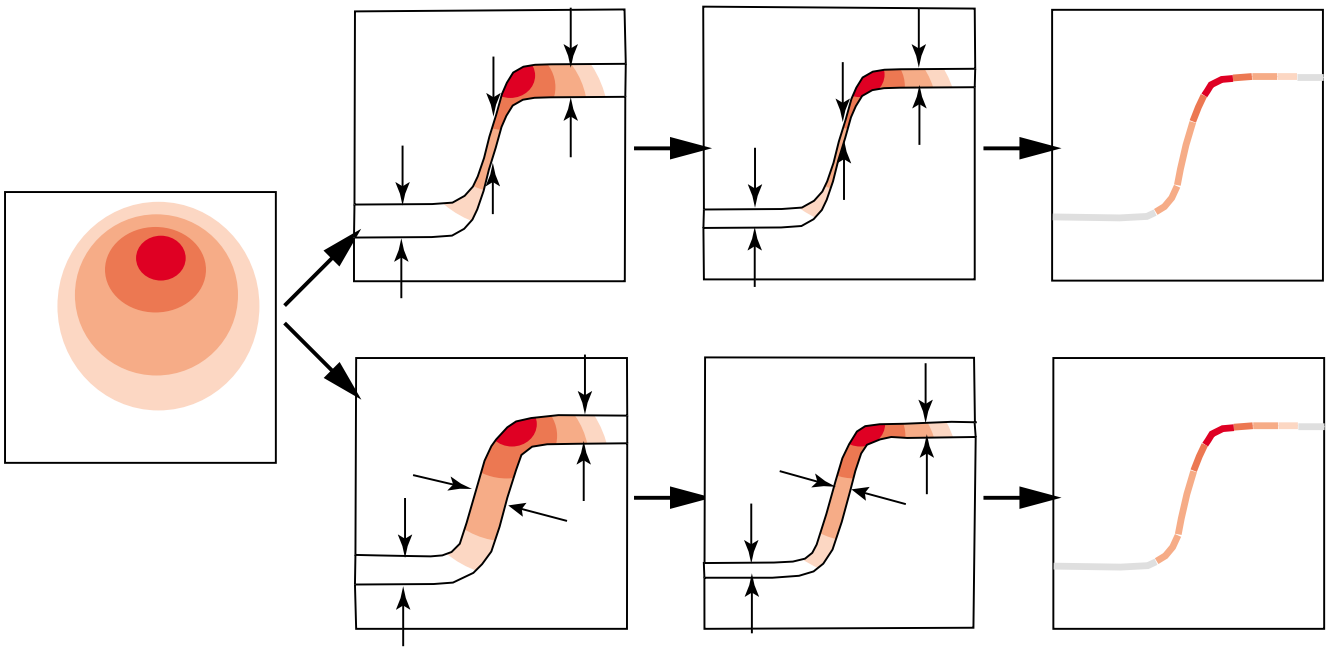


Figure 2: An original 2D probability density, and two possible ways (among many) of defining a region of the space whose limit is a given curve. At the top is the 'vertical' limit, while at the bottom is the normal (or orthogonal) limit. Each possible limit defines a different 'induced' or 'conditional' probability density. Only the orthogonal limit gives an intrinsic definition (i.e., a definition invariant under any change of variables). It is, therefore, the only one examined in this work.

$\mathbf{v} = \mathbf{v}(\mathbf{u})$. We could choose ad-hoc coordinates over this manifold, but as there is a one-to-one correspondence between the coordinates $\mathbf{u}$ and the points on the manifold, the conditional probability density can be expressed using the coordinates $\mathbf{u}$. The restriction of $f(\mathbf{u}, \mathbf{v})$ over the submanifold $\mathbf{v} = \mathbf{v}(\mathbf{u})$ defines the probability density (see appendix B for the more general case)

$$f_{u|v(u)}(\mathbf{u}|\mathbf{v} = \mathbf{v}(\mathbf{u})) \;=\; k\; f(\mathbf{u}, \mathbf{v}(\mathbf{u})) \;\; \frac{\sqrt{\det\left(\mathbf{g}_u + \mathbf{V}^T\, \mathbf{g}_v\, \mathbf{V}\right)}}{\sqrt{\det \mathbf{g}_u}\;\sqrt{\det \mathbf{g}_v}} \;\Bigg|_{\mathbf{v}=\mathbf{v}(\mathbf{u})} \quad , \tag{17}$$

where $k$ is a normalizing constant, and where $\mathbf{V} = \mathbf{V}(\mathbf{u})$ is the matrix of partial derivatives (see appendix K for a simple explicit calculation of such partial derivatives)

$$\begin{pmatrix} V_{11} & V_{12} & \cdots & V_{1p} \\ V_{21} & V_{22} & \cdots & V_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ V_{q1} & V_{q2} & \cdots & V_{qp} \end{pmatrix} \;=\; \begin{pmatrix} \frac{\partial v_1}{\partial u_1} & \frac{\partial v_1}{\partial u_2} & \cdots & \frac{\partial v_1}{\partial u_p} \\ \frac{\partial v_2}{\partial u_1} & \frac{\partial v_2}{\partial u_2} & \cdots & \frac{\partial v_2}{\partial u_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial v_q}{\partial u_1} & \frac{\partial v_q}{\partial u_2} & \cdots & \frac{\partial v_q}{\partial u_p} \end{pmatrix} \quad . \tag{18}$$

**Example 3** *If the hypersurface* $\mathbf{v} = \mathbf{v}(\mathbf{u})$ *is defined by a constant value of* $\mathbf{v}$*, say* $\mathbf{v} = \mathbf{v}_0$*, then equation 17 reduces to*

$$f_{u|v}(\mathbf{u}|\mathbf{v} = \mathbf{v}_0) \;=\; k\; f(\mathbf{u}, \mathbf{v}_0) \;=\; \frac{f(\mathbf{u}, \mathbf{v}_0)}{\int_{\mathcal{U}} d\mathbf{u}\; f(\mathbf{u}, \mathbf{v}_0)} \quad . \tag{19}$$

[END OF EXAMPLE.]

Elementary definitions of conditional probability density are not based on this notion of distance-based uniform convergence, but use other, ill-defined limits. This is a mistake that, unfortunately, pollutes many scientific works. See appendix P, in particular, for a discussion on the "Borel paradox".

Equation 17 defines the conditional $f_{u|v(u)}(\mathbf{u}|\mathbf{v} = \mathbf{v}(\mathbf{u}))$. Should the relation $\mathbf{v} = \mathbf{v}(\mathbf{u})$ be invertible, it would correspond to a change of variables. It is then possible to show that the alternative conditional $f_{v|u(v)}(\mathbf{v}|\mathbf{u} = \mathbf{u}(\mathbf{v}))$, is related to $f_{u|v(u)}(\mathbf{u}|\mathbf{v} = \mathbf{v}(\mathbf{u}))$ through the Jacobian rule. This is a property that elementary definitions of conditional probability do not share.

### 2.5 Marginal Probability Density

In the special circumstance described above, where we have a Cartesian product of two spaces, $\mathcal{X} = \mathcal{U} \times \mathcal{V}$ , given a 'joint' probability density $f(\mathbf{u}, \mathbf{v})$ , it is possible to give an intrinsic sense to the definitions

$$f_u(\mathbf{u}) \;=\; \int_{\mathcal{V}} d\mathbf{v}\; f(\mathbf{u}, \mathbf{v}) \qquad ; \qquad f_v(\mathbf{v}) \;=\; \int_{\mathcal{U}} d\mathbf{u}\; f(\mathbf{u}, \mathbf{v}) \quad . \tag{20}$$

These two densities are called *marginal probability densities*. Their intuitive interpretation is clear, as the 'projection' of the joint probability density respectively over $\mathcal{U}$ and over $\mathcal{V}$ .

### 2.6 Independence and Bayes Theorem

Dropping the index $_0$ in equation 19 and using the second of equations 20 gives

$$f_{u|v}(\mathbf{u}|\mathbf{v}) \;=\; \frac{f(\mathbf{u}, \mathbf{v})}{f_v(\mathbf{v})} \quad , \tag{21}$$

or, equivalently, $f(\mathbf{u}, \mathbf{v}) \;=\; f_{u|v}(\mathbf{u}|\mathbf{v})\; f_v(\mathbf{v})$ . As we can also define $f_{v|u}(\mathbf{v}|\mathbf{u})$ , we have the two equations

$$\begin{array}{rcl} f(\mathbf{u}, \mathbf{v}) & = & f_{u|v}(\mathbf{u}|\mathbf{v})\; f_v(\mathbf{v}) \\ f(\mathbf{u}, \mathbf{v}) & = & f_{v|u}(\mathbf{v}|\mathbf{u})\; f_u(\mathbf{u}) \quad , \end{array} \tag{22}$$

that can be read as follows: 'when we work in a space that is the Cartesian product $\mathcal{U} \times \mathcal{V}$ of two subspaces, a joint probability density can always be expressed as the product of a conditional times a marginal'.

From these last equations it follows the expression

$$f_{u|v}(\mathbf{u}|\mathbf{v}) \;=\; \frac{f_{v|u}(\mathbf{v}|\mathbf{u})\; f_u(\mathbf{u})}{f_v(\mathbf{v})} \quad , \tag{23}$$

known as the *Bayes theorem*, and generally used as the starting point to solve inverse problems. We do not think this is a useful setting, and we prefer in this paper *not* to use the Bayes theorem (or, more precisely, not to use the intuitive paradigm usually associated to it).

It also follows from equations 22 that the two conditions

$$f_{u|v}(\mathbf{u}|\mathbf{v}) \;=\; f_u(\mathbf{u}) \qquad ; \qquad f_{v|u}(\mathbf{v}|\mathbf{u}) \;=\; f_v(\mathbf{v}) \tag{24}$$

are equivalent. It is then said that $\mathbf{u}$ and $\mathbf{v}$ are *independent parameters* (with respect to the probability density $f(\mathbf{u}, \mathbf{v})$ ). The term 'independent' is easy to understand, as the conditional of any of the two (vector) variables, given the other variable equals the (unconditional) marginal of the variable. Then, one clearly has

$$f(\mathbf{u}, \mathbf{v}) \;=\; f_u(\mathbf{u})\; f_v(\mathbf{v}) \quad , \tag{25}$$

i.e., for independent variables, the joint probability density can be simply expressed as the product of the two marginals.

## 3 Monte Carlo Methods

When a probability distribution has been defined, we face the problem of how to 'use' it. The definition of 'central estimators' (like the mean or the median) and 'estimators of dispersion' (like the covariance matrix) lacks generality as it is quite easy to find examples (like multimodal distributions in highly-dimensional spaces) where these estimators fail to have any interesting meaning.

When a probability distribution has been defined over a space of low dimension (say, from one to four dimensions) we can directly represent the associated probability density. This is trivial in one or two dimensions. It is easy in three dimensions, and some tricks may allow us to represent a four-dimensional probability distribution, but clearly this approach cannot be generalized to the high dimensional case.

Let us explain the only approach that seems practical, with help of figure 3. At the left of the figure, there is an explicit representation of a 2D probability distribution (by means of the associated probability density or the associated (2D) volumetric probability). In the middle, some random points have been generated (using the

Monte Carlo method about to be described). It is clear that if we make a histogram with these points, in the limit of a sufficiently large number of points we recover the representation at the left. Disregarding the histogram possibility we can concentrate on the individual points. In the 2D example of the figure we have actual points in a plane. If the problem is multi-dimensional, each 'point' may correspond to some abstract notion. For instance, for a geophysicist a 'point' may be a given model of the Earth. This model may be represented in some way, for instance a by color plot. Then a collection of 'points' is a collection of such pictures. Our experience shows that, given a collection of randomly generated 'models', the human eye-brain system is extremely good at apprehending the basic characteristics of the underlying probability distribution, including possible multimodalities, correlations, etc.

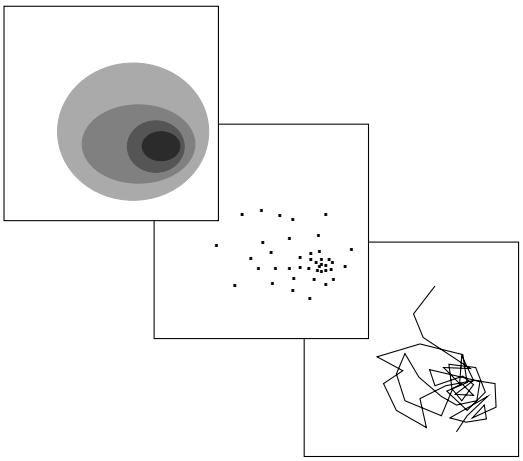


Figure 3: An explicit representation of a 2D probability distribution and the sampling of it, using Monte Carlo methods. While the representation at the top-left cannot be generalized to high dimensions, the examination of a collection of points can be done in arbitrary dimensions. Practically, Monte Carlo generation of points is done through a 'random walk' where a 'new point' is generated in the vicinity of the previous point.

When such a (hopefully large) collection of random models is available we can also answer quite interesting questions. For instance, a geologist may ask: *at which depth is that subsurface strucure?* To answer this, we can make an histogram of the depth of the given geological structure over the collection of random models, and the histogram *is* the answer to the question. *What is the probability of having a low velocity zone around a given depth?* The ratio of the number of models presenting such a low velocity zone over the total number of models in the collection gives the answer (if the collection of models is large enough).

This is essentially what we propose: looking at a large number of randomly generated models in order to intuitively apprehend the basic properties of the probability distribution, followed by calculation of the probabilities of all interesting 'events'.

Practically, as we shall see, the random sampling is not made by generating points independently of each other. Rather, as suggested in the last image of figure 3, through a 'random walk' where a 'new point' is generated in the vicinity of the previous point.

Monte Carlo methods have a random generator at their core. At present, Monte Carlo methods are typically implemented on digital computers, and are based on pseudorandom generation of numbers[10]. As we shall see, any conceivable operation on probability densities (e.g., computing marginals and conditionals, integration, conjunction (the AND operation), etc.) has its counterpart in an operation on/by their corresponding Monte Carlo algorithms.

Inverse problems are often formulated in high-dimensional spaces. In this case a certain class of Monte Carlo algorithms, the so-called *importance sampling algorithms*, come to the rescue, allowing us to sample the space with a sampling density proportional to the given probability density. In this case excessive (and useless) sampling of low-probability areas of the space is avoided. This is not only important, but in fact vital in high dimensional spaces.

Another advantage of the importance sampling Monte Carlo algorithms is that we need not have a closed form mathematical expression for the probability density we want to sample. Only an algorithm that allows us to evaluate it at a given point in the space is needed. This has considerable practical advantage in analysis of inverse problems where computer intensive evaluation of, e.g., misfit functions plays an important role in calculation of certain probability densities.

Given a probability density that we wish to sample, and a class of Monte Carlo algorithms that samples this density, which one of the algorithms should we choose? Practically, the problem is here to find the most efficient of these algorithms. This is an interesting and difficult problem that we will not go into detail with here. We

[10] I.e., series of numbers that appear random if tested with any reasonable statistical test.

will, later in this chapter, limit ourselves to only two general methods which are recommendable in many practical situations.

## 3.1 Random Walks

To escape the dimensionality problem, *any* sampling of a probability density for which point values are available only upon request has to be based on a *random walk*, i.e., in a generation of successive points with the constraint that point $\mathbf{x}_{i+1}$ sampled in iteration $(i+1)$ is in the vicinity of the point $\mathbf{x}_i$ sampled in iteration $i$. The simplest of the random walks are generated by the so-called Markov Chain Monte Carlo (MCMC) algorithms, where the point $\mathbf{x}_{i+1}$ depends on the point $\mathbf{x}_i$, but not on previous points. We will concentrate on these algorithms here.

If random rules have been defined to select points such that the probability of selecting a point in the infinitesimal "box" $dx_1 \dots dx_N$ is $p(\mathbf{x})\, dx_1 \dots dx_N$, then the points selected in this way are called *samples* of the probability density $p(\mathbf{x})$. Depending on the rules defined, successive samples $i, j, k, \dots$ may be dependent or independent.

## 3.2 The Metropolis Rule

The most common Monte Carlo sampling methods are the Metropolis sampler (described below) and the Gibbs sampler (Geman and Geman, 1984). As we believe that the Gibbs sampler is only superior to the Metropolis sampler in low-dimensional problems, we restrict ourselves here to the presentation of the latter.

Consider the following situation. Some random rules define a random walk that samples the probability density $f(\mathbf{x})$ . At a given step, the random walker is at point $\mathbf{x}_j$, and the application of the rules would lead to a transition to point $\mathbf{x}_i$. By construction, when all such 'proposed transitions' $\mathbf{x}_i \leftarrow \mathbf{x}_j$ are always accepted, the random walker will sample the probability density $f(\mathbf{x})$ . Instead of always accepting the proposed transition $\mathbf{x}_i \leftarrow \mathbf{x}_j$, we reject it sometimes by using the following rule to decide if it is allowed to move to $\mathbf{x}_i$ or if it must stay at $\mathbf{x}_j$:

- if $g(\mathbf{x}_i)/\mu(\mathbf{x}_i) \geq g(\mathbf{x}_j)/\mu(\mathbf{x}_j)$ , then accept the proposed transition to $\mathbf{x}_i$ ,
- if $g(\mathbf{x}_i)/\mu(\mathbf{x}_i) < g(\mathbf{x}_j)/\mu(\mathbf{x}_j)$ , then decide randomly to move to $\mathbf{x}_i$ , or to stay at $\mathbf{x}_j$ , with the following probability of accepting the move to $\mathbf{x}_i$ :

$$P \;=\; \frac{g(\mathbf{x}_i)/\mu(\mathbf{x}_i)}{g(\mathbf{x}_j)/\mu(\mathbf{x}_j)} \quad . \tag{26}$$

Then we have the following

**Theorem 1** *The random walker samples the conjunction $h(\mathbf{x})$ of the probability densities $f(\mathbf{x})$ and $g(\mathbf{x})$*

$$h(\mathbf{x}) \;=\; k\, f(\mathbf{x})\, \frac{g(\mathbf{x})}{\mu(\mathbf{x})} \;=\; k\, \frac{f(\mathbf{x})\, g(\mathbf{x})}{\mu(\mathbf{x})} \tag{27}$$

*(see appendix O for a demonstration).*

It should be noted here that this algorithm nowhere requires the probability densities to be normalized. This is of vital importance in practice, since it allows sampling of probability densities whose values are known only in points already sampled by the algorithm. Obviously, such probability densities cannot be normalized. Also, the fact that our theory allows unnormalizable probability densities will not cause any trouble in the application of the above algorithm.

The algorithm above is reminiscent (see appendix O) of the Metropolis algorithm (Metropolis et al., 1953), originally designed to sample the Gibbs-Boltzmann distribution [11]. Accordingly, we will refer to the above acceptance rule as the *Metropolis rule*.

[11] To see this, put $f(\mathbf{x}) = \mathbf{1}$, $\mu(\mathbf{x}) = \mathbf{1}$ , and $g(\mathbf{x}) = \frac{\exp(-E(\mathbf{x})/T)}{\int \exp(-E(\mathbf{x})/T) d\mathbf{x}}$ , where $E(\mathbf{x})$ is an "energy" associated to the point $\mathbf{x}$, and $T$ is a "temperature". The summation in the denominator is over the entire space. In this way, our acceptance rule becomes the classical Metropolis rule: point $\mathbf{x}_i$ is always accepted if $E(\mathbf{x}_i) \leq E(\mathbf{x}_j)$, but if $E(\mathbf{x}_i) > E(\mathbf{x}_j)$, it is only accepted with probability $p_{ij}^{\mathrm{acc}} = \exp\left(-\left(E(\mathbf{x}_i) - E(\mathbf{x}_j)\right)/T\right)$ .

### 3.3 The Cascaded Metropolis Rule

As above, assume that some random rules define a random walk that samples the probability density $f_1(\mathbf{x})$ . At a given step, the random walker is at point $\mathbf{x}_j$ ;

1 apply the rules that unthwarted would generate samples distributed according to $f_1(\mathbf{x})$ , to propose a new point $\mathbf{x}_i$ ,

2 if $f_2(\mathbf{x}_i)/\mu(\mathbf{x}_i) \geq f_2(\mathbf{x}_j)/\mu(\mathbf{x}_j)$ , go to point 3; if $f_2(\mathbf{x}_i)/\mu(\mathbf{x}_i) < f_2(\mathbf{x}_j)/\mu(\mathbf{x}_j)$ , then decide randomly to go to point 3 or to go back to point 1, with the following probability of going to point 3: $P = (f_2(\mathbf{x}_i)/\mu(\mathbf{x}_i))/(f_2(\mathbf{x}_j)/\mu(\mathbf{x}_j))$ ;

3 if $f_3(\mathbf{x}_i)/\mu(\mathbf{x}_i) \geq f_3(\mathbf{x}_j)/\mu(\mathbf{x}_j)$ , go to point 4; if $f_3(\mathbf{x}_i)/\mu(\mathbf{x}_i) < f_3(\mathbf{x}_j)/\mu(\mathbf{x}_j)$ , then decide randomly to go to point 4 or to go back to point 1, with the following probability of going to point 4: $P = (f_3(\mathbf{x}_i)/\mu(\mathbf{x}_i))/(f_3(\mathbf{x}_j)/\mu(\mathbf{x}_j))$ ;

... ...

$n$ if $f_n(\mathbf{x}_i)/\mu(\mathbf{x}_i) \geq f_n(\mathbf{x}_j)/\mu(\mathbf{x}_j)$ , then accept the proposed transition to $\mathbf{x}_i$ ; if $f_n(\mathbf{x}_i)/\mu(\mathbf{x}_i) < f_n(\mathbf{x}_j)/\mu(\mathbf{x}_j)$ , then decide randomly to move to $\mathbf{x}_i$ , or to stay at $\mathbf{x}_j$ , with the following probability of accepting the move to $\mathbf{x}_i$ : $P = (f_n(\mathbf{x}_i)/\mu(\mathbf{x}_i))/(f_n(\mathbf{x}_j)/\mu(\mathbf{x}_j))$ ;

Then we have the following

**Theorem 2** *The random walker samples the conjunction* $h(\mathbf{x})$ *of the probability densities* $f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x})$ *:*

$$h(\mathbf{x}) \;=\; k\, f_1(\mathbf{x})\, \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})} \ldots \frac{f_n(\mathbf{x})}{\mu(\mathbf{x})} \quad . \tag{28}$$

*(see the CD-ROM supplement for a demonstration).*

### 3.4 Initiating a Random Walk

Consider the problem of obtaining samples of a probability density $h(\mathbf{x})$ defined as the conjunction of some probability densitites $f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x}) \ldots$ ,

$$h(\mathbf{x}) \;=\; k\, f_1(\mathbf{x})\, \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})}\, \frac{f_3(\mathbf{x})}{\mu(\mathbf{x})} \ldots \quad , \tag{29}$$

and let us examine three common situations.

**We start with a random walk that samples $f_1(\mathbf{x})$ (optimal situation):** This corresponds to the basic algorithm where we know how to produce a random walk that samples $f_1(\mathbf{x})$ , and we only need to modify it, taking into account the values $f_2(\mathbf{x})/\mu(\mathbf{x})$ , $f_3(\mathbf{x})/\mu(\mathbf{x}) \ldots$ , using the cascaded Metropolis rule, to obtain a random walk that samples $h(\mathbf{x})$ .

**We start with a random walk that samples the homogeneous probability density $\mu(\mathbf{x})$ :** We can write equation 29 as

$$h(\mathbf{x}) \;=\; k \left( \left( \left( \mu(\mathbf{x})\, \frac{f_1(\mathbf{x})}{\mu(\mathbf{x})} \right) \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})} \right) \ldots \right) \quad . \tag{30}$$

The expression corresponds to the case where we are not able to start with a random walk that samples $f_1(\mathbf{x})$ , but we have a random walk that samples the homogeneous probability density $\mu(\mathbf{x})$ . Then, with respect to the example just mentioned, there is one extra step to be added, taking into account the values of $f_1(\mathbf{x})/\mu(\mathbf{x})$ .

**We start with an arbitrary random walk (worst situation):** In the situation where we are not able to directly define a random walk that samples the homogeneous probability distribution, but only one that samples some arbitrary (but known) probability distribution $\psi(\mathbf{x})$ , we can write equation 29 in the form

$$h(\mathbf{x}) \;=\; k \left( \left( \left( \left( \psi(\mathbf{x})\, \frac{\mu(\mathbf{x})}{\psi(\mathbf{x})} \right) \frac{f_1(\mathbf{x})}{\mu(\mathbf{x})} \right) \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})} \right) \ldots \right) \quad . \tag{31}$$

Then, with respect to the example just mentioned, there is one more extra step to be added, taking into account the values of $\mu(\mathbf{x}))/\psi(\mathbf{x})$ . Note that the closer $\psi(\mathbf{x})$ will be to $\mu(\mathbf{x})$ , the more efficient will be the first modification of the random walk.

### 3.5 Convergence Issues

When has a random walk visited enough points in the space so that a probability density has been sufficiently sampled? This is a complex issue, and it is easy to overlook its importance. There is no general rule: each problem has its own 'physics', and the experience of the 'implementer' is, here, crucial.

Many methods that work for low dimension completely fail when the number of dimensions is high. Typically, a random walk select a random direction and, then, a random step along that direction. The notion of 'direction' in a high-dimensional space is far from the intuitive one we get in the familar three-dimensional space. Any serious discussion on this issue must be problem-dependent, so we don't even attempt one here.

Obviously, a necessary condition for adequate sampling is that any 'output' from the algorithm must 'look stationary'.

## 4 Probabilistic Formulation of Inverse Problems

A so-called 'inverse problem' arises when a usually complex measurement is made, and information on unknown parameters of the physical system is sought. Any measurement is indirect (we may weigh a mass by observing the displacement of the cursor of a balance), and therefore a possibly nontrivial analysis of uncertainties must be done. Any guide describing good experimental practice (see, for instance ISO's *Guide to the expression of uncertainty in measurement* [ISO, 1993] or the shorter description by Taylor and Kuyatt, 1994) acknowledges that a measurement involves, at least, two different sources of uncertainties: those estimated using statistical methods, and those estimated using subjective, common-sense estimations. Both are described using the axioms of probability theory, and this article clearly takes the probabilistic point of view for developing inverse theory.

### 4.1 Model Parameters and Observable Parameters

Although the separation of all the variables of a problem in two groups, 'directly observable parameters' (or 'data') and 'model parameters', may sometimes be artificial, we take this point of view here, since it allows us to propose a simple setting for a wide class of problems.

We may have in mind a given physical system, like the whole Earth, or a small crystal under our microscope. The system (or a given state of the system) may be described by assigning values to a given set of parameters $\mathbf{m} = \{m^1, m^2, \ldots, m^{\mathrm{NM}}\}$ that we will name the *model parameters*.

Let us assume that we make observations on this system. Although we are interested in the parameters $\mathbf{m}$, they may not be directly observable, so we make indirect measurements like obtaining seismograms at the Earth's surface for analyzing the Earth's interior, or making spectroscopic measurements for analyzing the chemical properties of a crystal. The set of *(directly) observable parameters* (or, by language abuse, the set of *data parameters*) will be represented by $\mathbf{d} = \{d^1, d^2, \ldots, d^{\mathrm{ND}}\}$.

We assume that we have a physical theory that can be used to solve the *forward problem*, i.e., that given an arbitrary model $\mathbf{m}$, it allows us to predict the theoretical data values $\mathbf{d}$ that an ideal measurement should produce (if $\mathbf{m}$ was the actual system). The generally nonlinear function that associates to any model $\mathbf{m}$ the theoretical data values $\mathbf{d}$ may be represented by a notation like

$$d^i \;=\; f^i(m^1, m^2, \ldots, m^{\mathrm{NM}}) \qquad ; \qquad i = 1, 2, \ldots, \mathrm{ND} \;, \tag{32}$$

or, for short,

$$\mathbf{d} \;=\; \mathbf{f}(\mathbf{m}) \;. \tag{33}$$

It is in fact this expression that separates the whole set of our parameters into the subsets $\mathbf{d}$ and $\mathbf{m}$, although sometimes there is no difference in nature between the parameters in $\mathbf{d}$ and the parameters in $\mathbf{m}$. For instance, in the classical inverse problem of estimating the hypocenter coordinates of an earthquake, we may put in $\mathbf{d}$ the arrival times of the seismic waves at seismic observatories, and we need to put in $\mathbf{m}$, besides the hypocentral coordinates, the coordinates defining the location of the seismometers —as these are parameters that are needed to compute the travel times—, although we estimate arrival times of waves and coordinates of the seismic observatories using similar types of measurements.

## 4.2 Prior Information on Model Parameters

In a typical geophysical problem, the model parameters contain geometrical parameters (positions and sizes of geological bodies) and physical parameters (values of the mass density, of the elastic parameters, the temperature, the porosity, etc.).

The *prior information* on these parameters is all the information we possess independently of the particular measurements that will be considered as 'data' (to be described below). This prior probability distribution is generally quite complex, as the model space may be high-dimensional, and the parameters may have nonstandard probability densities.

To this generally complex probability distribution over the model space corresponds a probability density that we denote $\rho_{\mathcal{M}}(\mathbf{m})$ .

If an explicit expression for the probability density $\rho_{\mathcal{M}}(\mathbf{m})$ is known, it can be used in analytical developments. But such an explicit expression is, by no means, necessary. Using Monte Carlo methods, all that is needed is a set of probabilistic rules that allows us to generate samples distributed according to $\rho_{\mathcal{M}}(\mathbf{m})$ in the model space (Mosegaard and Tarantola, 1995).

**Example 4** *Appendix E presents an example of prior information for the case of an Earth model consisting of a stack of horizontal layers with variable thickness and uniform mass density.* [END OF EXAMPLE.]

## 4.3 Measurements and Experimental Uncertainties

Observation of geophysical phenomena is represented by a set of parameters $\mathbf{d}$ that we usually call data. These parameters result from prior measurement operations, and they are typically seismic vibrations on the instrument site, arrival times of seismic phases, gravity or electromagnetic fields. As in any measurement, the data is determined with an associated uncertainty, described by a probability density over the data parameter space, that we denote here $\rho_{\mathcal{D}}(\mathbf{d})$. This density describes, not only marginals on individual datum values, but also possible cross-relations in data uncertainties.

Although the instrumental errors are an important source of data uncertainties, in geophysical measurements there are other sources of uncertainty. The errors associated with the positioning of the instruments, the environmental noise, and the human factor (like for picking arrival times) are also relevant sources of uncertainty.

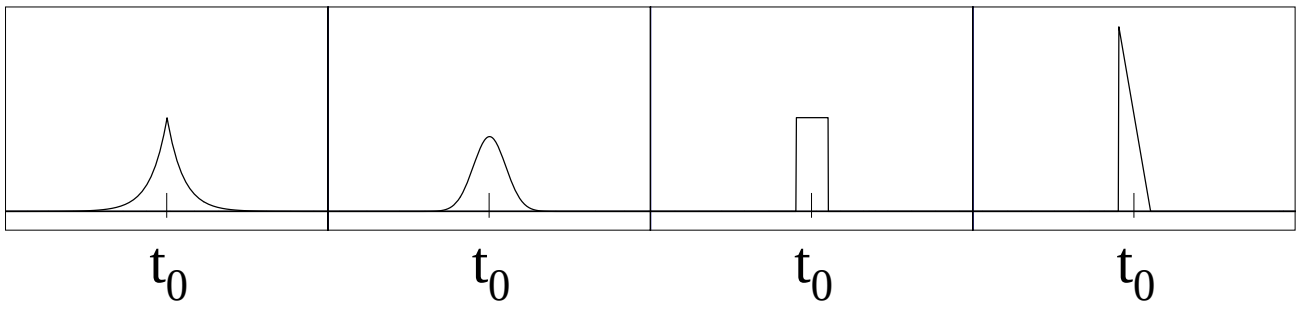



Figure 4: What has an experimenter in mind when she/he describes the result of a measurement by something like $t = t_0 \pm \sigma$ ?

**Example 5 Non-analytic Probability Density**

*Assume that we wish to measure the time* $t$ *of occurrence of some physical event. It is often assumed that the result of a measurement corresponds to something like*

$$t \;=\; t_0 \pm \sigma \quad . \tag{34}$$

*An obvious question is the exact meaning of the* $\pm\sigma$ *. Has the experimenter in mind that she/he is absolutely certain that the actual arrival time satisfies the strict conditions* $t_0 - \sigma \leq t \leq t_0 + \sigma$ *, or has she/he in mind something like a Gaussian probability, or some other probability distribution (see figure 4)? We accept, following ISO's recommendations (1993) that the result of any measurement has a probabilistic interpretation, with some sources of uncertainty being analyzed using statistical methods ('type A' uncertainties), and other sources of uncertainty being evaluated by other means (for instance, using Bayesian arguments) ('type B' uncertainties). But, contrary to ISO suggestions, we do not assume that the Gaussian model of uncertainties should play any central role. In an extreme example, we may well have measurements whose probabilistic description may correspond to a multimodal probability density. Figure 5 shows a typical example for a seismologist: the measurement on a seismogram of the arrival time of a certain seismic wave, in the case one hesitates in the phase identification or in the identification of noise and signal. In this case the probability density for the arrival of the seismic phase does not have an explicit expression like* $f(t) = k\exp(-(t-t_0)^2/(2\sigma^2))$ *, but is a numerically defined function.* [END OF EXAMPLE.]

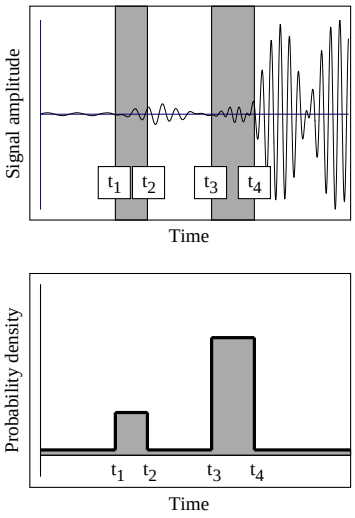



Figure 5: A seismologist tries to measure the arrival time of a seismic wave at a seismic station, by 'reading' the seismogram at the top of the figure. The seismologist may find quite likely that the arrival time of the wave is between times $t_3$ and $t_4$, and believe that what is before $t_3$ is just noise. But if there is a significant probability that the signal between $t_1$ and $t_2$ is not noise but the actual arrival of the wave, then the seismologist should define a bimodal probability density, as the one suggested at the bottom of the figure. Typically, the actual form of each peak of the probability density is not crucial (here, box-car functions are chosen), but the position of the peaks is important. Rather than assigning a zero probability density to the zones outside the two intervals, it is safer (more 'robust') to attribute some small 'background' value, as we may never exclude some unexpected source of error.

**Example 6** *The Gaussian model for uncertainties. The simplest probabilistic model that can be used to describe experimental uncertainties is the Gaussian model*

$$\rho_{\mathcal{D}}(\mathbf{d}) \;=\; k\,\exp\left(-\frac{1}{2}(\mathbf{d}-\mathbf{d}_{\rm obs})^T\,\mathbf{C}_D^{-1}\,(\mathbf{d}-\mathbf{d}_{\rm obs})\right) \quad . \tag{35}$$

*It is here assumed that we have some 'observed data values'* $\mathbf{d}_{\rm obs}$ *with uncertainties described by the covariance matrix* $\mathbf{C}_D$ *. If the uncertainties are uncorrelated,*

$$\rho_{\mathcal{D}}(\mathbf{d}) \;=\; k\exp\left(-\frac{1}{2}\sum_i\left(\frac{d^i-d^i_{\rm obs}}{\sigma^i}\right)^2\right) \quad , \tag{36}$$

*where the* $\sigma^i$ *are the 'standard deviations'.* [END OF EXAMPLE.]

**Example 7** *The Generalized Gaussian model for uncertainties. An alternative to the Gaussian model is to use the Laplacian (double exponential) model for uncertainties,*

$$\rho_{\mathcal{D}}(\mathbf{d}) \;=\; k\exp\left(-\sum_i\frac{|d^i-d^i_{\rm obs}|}{\sigma^i}\right) \quad . \tag{37}$$

*While the Gaussian model leads to least-squares related methods, this Laplacian model leads to absolute-values methods (see section4.5.2), well known for producing robust*[12] *results. More generally, there is the* $L_p$ *model of uncertainties*

$$\rho_p(\mathbf{d}) \;=\; k\,\exp\left(-\frac{1}{p}\sum_i \frac{|d^i - d^i_{\text{obs}}|^p}{(\sigma^i)^p}\right) \tag{38}$$

*(see figure 6).* [END OF EXAMPLE.]

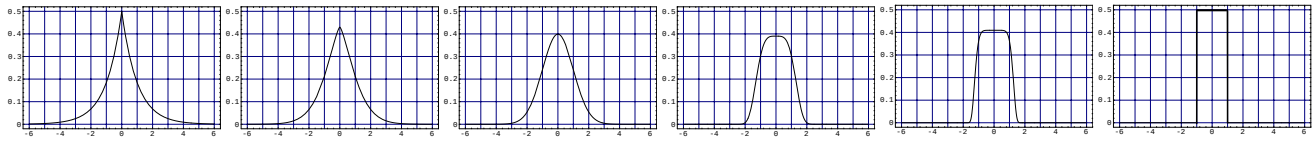


Figure 6: Generalized Gaussian for values of the parameter $p = 1, \sqrt{2}, 2, 4, 8$ and $\infty$ .

## 4.4 Joint 'Prior' Probability Distribution in the $(\mathcal{M}, \mathcal{D})$ Space

We have just seen that the prior information on model parameters can be described by a probability density in the model space, $\rho_m(\mathbf{m})$ , and that the result of measurements can be described by a probability density in the data space $\rho_d(\mathbf{d})$ . As by 'prior' information on model parameters we mean information obtained *independently* from the measurements (it often represents information we had before the measurements were made), we can use the notion of indepedency of variables of section 2.6 to define a joint probability density in the $\mathcal{X} = (\mathcal{M}, \mathcal{D})$ space as the product of the two 'marginals'

$$\rho(\mathbf{x}) \;=\; \rho(\mathbf{m}, \mathbf{d}) = \rho_m(\mathbf{m})\,\rho_d(\mathbf{d}) \quad . \tag{39}$$

Although we have introduced $\rho_m(\mathbf{m})$ and $\rho_d(\mathbf{d})$ separately, and we have suggested to build a probability distribution in the $(\mathcal{M}, \mathcal{D})$ space by the multiplication 39, we may have more general situation where the information we have on $\mathbf{m}$ and on $\mathbf{d}$ is not independent So, in what follows, let us assume that we have some information in the $\mathcal{X} = (\mathcal{M}, \mathcal{D})$ space, represented by the 'joint' probability density

$$\rho(\mathbf{x}) = \rho(\mathbf{m}, \mathbf{d}) \quad , \tag{40}$$

and let us contemplate equation 39 as just a special case.

Let us in the rest of this paper denote by $\mu(\mathbf{x})$ the probability density representing the homogeneous probability distribution, as introduced in section 2.2. We may remember here the rule 8, stating that the limit of a consistent probability density must the the homogeneous one, so we may formally write

$$\mu(\mathbf{x}) \quad = \quad \lim_{\text{infinite dispersions}} \quad \rho(\mathbf{x}) \;. \tag{41}$$

When the partition 39 holds, then, typically (see rule 8),

$$\mu(\mathbf{x}) \;=\; \mu(\mathbf{m}, \mathbf{d}) \;=\; \mu_m(\mathbf{m})\,\mu_d(\mathbf{d}) \quad . \tag{42}$$

## 4.5 Physical Laws as Mathematical Functions

### 4.5.1 Physical Laws

Physics analyzes the correlations existing between physical parameters. In standard mathematical physics, these correlations are represented by 'equalities' between physical parameters (like when we write $\mathbf{F} = m\,\mathbf{a}$ to relate the force $\mathbf{F}$ applied to a particle, the mass $m$ of the particle and the acceleration $\mathbf{a}$ ). In the context of inverse problems this corresponds to assuming that we have a function from the 'parameter space' to the 'data space' that we may represent as

$$\mathbf{d} \;=\; \mathbf{f}(\mathbf{m}) \quad . \tag{43}$$

[12] A numerical method is called robust if it is not sensitive to a small number of large errors.

We do not mean that the relation is necessarily explicit. Given $\mathbf{m}$ we may need to solve a complex system of equations in order to get $\mathbf{d}$, but this nevertheless defines a function $\mathbf{m} \rightarrow \mathbf{d} = \mathbf{f}(\mathbf{m})$.

At this point, given the probability density $\rho(\mathbf{m}, \mathbf{d})$ and given the relation $\mathbf{d} = \mathbf{f}(\mathbf{m})$, we can define the associated conditional probability density $\rho_{m|d(m)}(\mathbf{m}|\mathbf{d} = \mathbf{f}(\mathbf{m}))$. We could here use the more general definition of conditional probability density of appendix B, but let us simplify the text by using a simplification assumption: that the total parameter space $(\mathcal{M}, \mathcal{D})$ is just the cartesian product $\mathcal{M} \times \mathcal{D}$ of the model parameter space $\mathcal{M}$ times the space of directly observable parameters (or 'data space') $\mathcal{D}$. Then, rather than a general metric in the total space, we have a metric $\mathbf{g}_m$ over the model parameter space $\mathcal{M}$ and a metric $\mathbf{g}_d$ over the data space, and the total metric is just the Cartesian product of the two metrics. In particular, then, the total volume element in the space, $dV(\mathbf{m}, \mathbf{d})$ is just the product of the two volume elements in the model parameter space and the data space: $dV(\mathbf{m}, \mathbf{d}) = dV_m(\mathbf{m})\, dV_d(\mathbf{d})$. Most of inverse problems satisfy this assumption[13]. In this setting, the formulas of section 2.4 are valid.

### 4.5.2 Inverse Problems

In the $(\mathcal{M}, \mathcal{D}) = \mathcal{M} \times \mathcal{D}$ space, we have the probability density $\rho(\mathbf{m}, \mathbf{d})$ and we have the hypersurface defined by the relation $\mathbf{d} = \mathbf{f}(\mathbf{m})$. The natural way to 'compose' these two kinds of information is by defining the conditional probability density induced by $\rho(\mathbf{m}, \mathbf{d})$ on the hypersurface $\mathbf{d} = \mathbf{f}(\mathbf{m})$,

$$\sigma_m(\mathbf{m}) \;\equiv\; \rho_{m|d(m)}(\mathbf{m}|\mathbf{d} = \mathbf{f}(\mathbf{m})) \;, \tag{44}$$

this giving (see equation 17)

$$\sigma_m(\mathbf{m}) \;=\; k\; \rho(\mathbf{m}, \mathbf{f}(\mathbf{m}))\;\; \frac{\sqrt{\det\left(\mathbf{g}_m + \mathbf{F}^T\, \mathbf{g}_d\, \mathbf{F}\right)}}{\sqrt{\det \mathbf{g}_m}\; \sqrt{\det \mathbf{g}_d}} \;\Bigg|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} \;, \tag{45}$$

where $\mathbf{F} = \mathbf{F}(\mathbf{m})$ is the matrix of partial derivatives, with components $F_{i\alpha} = \partial f_i / \partial m_\alpha$, where $\mathbf{g}_m$ is the metric in the model parameter space $\mathcal{M}$ and where $\mathbf{g}_d$ is the metric in the data space $\mathcal{D}$.

**Example 8** *Quite often,* $\rho(\mathbf{m}, \mathbf{d}) = \rho_m(\mathbf{m})\; \rho_d(\mathbf{d})$. *Then, equation 45 can be written*

$$\sigma_m(\mathbf{m}) \;=\; k\; \rho_m(\mathbf{m})\;\; \left( \frac{\rho_d(\mathbf{d})}{\sqrt{\det \mathbf{g}_d}}\;\; \frac{\sqrt{\det\left(\mathbf{g}_m + \mathbf{F}^T\, \mathbf{g}_d\, \mathbf{F}\right)}}{\sqrt{\det \mathbf{g}_m}} \right) \Bigg|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} \;. \tag{46}$$

[END OF EXAMPLE.]

**Example 9** *If* $\mathbf{F}^T\, \mathbf{g}_d\, \mathbf{F}$ *is negligible compared to of* $\mathbf{g}_m$, *then equation 46 reduces to*

$$\sigma_m(\mathbf{m}) \;=\; k\; \rho_m(\mathbf{m})\;\; \frac{\rho_d(\mathbf{d})}{\mu_d(\mathbf{d})} \;\Bigg|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} \;, \tag{47}$$

*where we have used* $\mu_d(\mathbf{d}) = k \sqrt{\det \mathbf{g}_d(\mathbf{d})}$ *(see rule 2).* [END OF EXAMPLE.]

**Example 10** *We examine here the simplification that we arrive at when assuming that the 'input' probability densities are Gaussian:*

$$\rho_m(\mathbf{m}) \;=\; k\, \exp\left( -\frac{1}{2} (\mathbf{m} - \mathbf{m}_{\text{prior}})^t\, \mathbf{C}_M^{-1}\, (\mathbf{m} - \mathbf{m}_{\text{prior}}) \right) \tag{48}$$

$$\rho_d(\mathbf{d}) \;=\; k\, \exp\left( -\frac{1}{2} (\mathbf{d} - \mathbf{d}_{\text{obs}})^t\, \mathbf{C}_D^{-1}\, (\mathbf{d} - \mathbf{d}_{\text{obs}}) \right) \;. \tag{49}$$

*In this circumstance, quite often, it is the covariance operators* $\mathbf{C}_M$ *and* $\mathbf{C}_D$ *that are used to define the metrics over the spaces* $\mathcal{M}$ *and* $\mathcal{D}$. *Then,* $\mathbf{g}_m = \mathbf{C}_M^{-1}$ *and* $\mathbf{g}_d = \mathbf{C}_D^{-1}$. *Grouping some of the constant factors in the factor* $k$, *equation 45 becomes here*

$$\sigma_m(\mathbf{m}) \;=$$

[13] It would be violated, for instance, if we use the pair of elastic parameters longitudinal wave velocity – shear wave velocity, as the volume element in the space of elastic wave velocities does not factorize (see appendix H).

$$= \quad k \exp\left(-\frac{1}{2}\left((\mathbf{m}-\mathbf{m}_{\text{prior}})^t\,\mathbf{C}_M^{-1}\,(\mathbf{m}-\mathbf{m}_{\text{prior}})+(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})^t\,\mathbf{C}_D^{-1}\,(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})\right)\right)\;\times$$

$$\times \quad \frac{\sqrt{\det\left(\mathbf{C}_M^{-1}+\mathbf{F}^T(\mathbf{m})\,\mathbf{C}_D^{-1}\,\mathbf{F}(\mathbf{m})\right)}}{\sqrt{\det\mathbf{C}_M^{-1}}} \tag{50}$$

*(the constant factor* $\sqrt{\det\mathbf{C}_M^{-1}}$ *has been left for subsequent simplifications). Defining the* misfit

$$S(\mathbf{m}) \;=\; -2\,\log\frac{\sigma_m(\mathbf{m})}{\sigma_0} \quad , \tag{51}$$

*where* $\sigma_0$ *is an arbitrary value of* $\sigma_m(\mathbf{m})$ *, gives, up to an additive constant,*

$$S(\mathbf{m}) \;=\; S_1(\mathbf{m})-S_2(\mathbf{m}) \quad , \tag{52}$$

*where* $S_1(\mathbf{m})$ *is the usual least-squares misfit function*

$$S_1(\mathbf{m}) \;=\; (\mathbf{m}-\mathbf{m}_{\text{prior}})^t\,\mathbf{C}_M^{-1}\,(\mathbf{m}-\mathbf{m}_{\text{prior}})+(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})^t\,\mathbf{C}_D^{-1}\,(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}}) \tag{53}$$

*and where*[14]

$$S_2(\mathbf{m}) \;=\; \log\det\left(\mathbf{I}+\mathbf{C}_M\,\mathbf{F}^T(\mathbf{m})\,\mathbf{C}_D^{-1}\,\mathbf{F}(\mathbf{m})\right) . \tag{54}$$

[END OF EXAMPLE.]

**Example 11** *If, in the context of example 10, we have*[15] $\mathbf{C}_M\mathbf{F}^T\mathbf{C}_D^{-1}\mathbf{F} << \mathbf{I}$ *, we can use the low order approximation for* $S_2(\mathbf{m})$ *, that is*[16]

$$S_2(\mathbf{m}) \quad\approx\quad \text{trace}\;\; \mathbf{C}_M\,\mathbf{F}^T(\mathbf{m})\,\mathbf{C}_D^{-1}\,\mathbf{F}(\mathbf{m}) \; . \tag{55}$$

[END OF EXAMPLE.]

**Example 12** *If in the context of example 10 we assume that the nonlinearities are weak, then the matrix of partial derivatives* $\mathbf{F}$ *is approximately constant, and equation 50 simplifies to*

$$\sigma_m(\mathbf{m}) \;= \tag{56}$$

$$= \; k\,\exp\left(-\frac{1}{2}\left((\mathbf{m}-\mathbf{m}_{\text{prior}})^t\,\mathbf{C}_M^{-1}\,(\mathbf{m}-\mathbf{m}_{\text{prior}})+(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})^t\,\mathbf{C}_D^{-1}\,(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})\right)\right) \quad ,$$

*and the function* $S_2(\mathbf{m})$ *is just a constant.* [END OF EXAMPLE.]

**Example 13** *If the 'relation solving the forward problem'* $\mathbf{d}=\mathbf{f}(\mathbf{m})$ *happens to be a linear relation,* $\mathbf{d} \;=\; \mathbf{F}\,\mathbf{m}$ *, then one get the standard equations for linear problems (see appendix F).* [END OF EXAMPLE.]

**Example 14** *We examine here the simplifications at we arrive at when assuming that the 'input' probability densities are Laplacian:*

$$\rho_m(\mathbf{m}) \;=\; k\,\exp\left(-\sum_\alpha \frac{|m^\alpha - m^\alpha_{\text{prior}}|}{\sigma_\alpha}\right) \tag{57}$$

$$\rho_d(\mathbf{d}) \;=\; k\,\exp\left(-\sum_i \frac{|d^i - d^i_{\text{obs}}|}{\sigma_i}\right) \quad . \tag{58}$$

[14] We use here the properties $\log\sqrt{\mathbf{A}}=\frac{1}{2}\,\log\mathbf{A}$ , and $\det\mathbf{A}\,\mathbf{B}=\det\mathbf{B}\,\mathbf{A}$

[15] Typically, this may happen because the derivatives $\mathbf{F}$ are small or because the variances in $\mathbf{C}_M$ are large.

[16] We first use $\log\det\mathbf{A}=\text{trace}\,\log\mathbf{A}$ , and then the series expansion of the logarithm of an operator, $\log(\mathbf{I}+\mathbf{A}) \;=\; \mathbf{A}-\frac{1}{2}\mathbf{A}^2+\ldots$ .

*Equation 45 becomes, here*

$$\sigma_m(\mathbf{m}) \;=\; k\,\exp\left(-\left(\sum_\alpha \frac{|m^\alpha - m^\alpha_{\text{prior}}|}{\sigma_\alpha} + \sum_i \frac{|f^i(\mathbf{m}) - d^i_{\text{obs}}|}{\sigma_i}\right)\right)\;\Psi(\mathbf{m}) \quad , \tag{59}$$

*where* $\Psi(\mathbf{m})$ *is a complex term containing, in particular, the matrix of partial derivatives* $\mathbf{F}$ *. If this term is approximately constant (weak nonlinearities, constant metrics), then*

$$\sigma_m(\mathbf{m}) \;=\; k\,\exp\left(-\left(\sum_\alpha \frac{|m^\alpha - m^\alpha_{\text{prior}}|}{\sigma_\alpha} + \sum_i \frac{|f^i(\mathbf{m}) - d^i_{\text{obs}}|}{\sigma_i}\right)\right) \quad . \tag{60}$$

[END OF EXAMPLE.]

The formulas in the examples above give expressions that contain analytic parts (like the square roots containing the matrix of partial derivatives $\mathbf{F}$ ). What we write as $\mathbf{d} = \mathbf{f}(\mathbf{m})$ may sometimes correspond to an explicit expression; sometimes it may corresponds to the solution of an implicit equation[17]. Should $\mathbf{d} = \mathbf{f}(\mathbf{m})$ be an explicit expression, and should the 'prior probability densities' $\rho_m(\mathbf{m})$ and $\rho_d(\mathbf{d})$ (or the joint $\rho(\mathbf{m}, \mathbf{d})$ ) also be given by explicit expressions (like when we have Gaussian probability densities), then the formulas of this section would give explicit expressions for the posterior probability density $\sigma_m(\mathbf{m})$ .

If the relation $\mathbf{d} = \mathbf{f}(\mathbf{m})$ is a linear relation, then the expression giving $\sigma_m(\mathbf{m})$ can sometimes be simplified easily (as with the linear Gaussian case to be examined below). More often than not the relation $\mathbf{d} = \mathbf{f}(\mathbf{m})$ is a complex nonlinear relation, and the expression we are left with for $\sigma_m(\mathbf{m})$ is explicit, but complex.

Once the probability density $\sigma_m(\mathbf{m})$ has been defined, there are different ways of 'using' it.

If the 'model space' $\mathcal{M}$ has a small number of dimensions (say between one and four) the values of $\sigma_m(\mathbf{m})$ can be computed at every point of a grid and a graphical representation of $\sigma_m(\mathbf{m})$ can be attempted. A visual inspection of such a representation is usually worth a thousand 'estimators' (central estimators or estimators of dispersion). But, of course, if the values of $\sigma_m(\mathbf{m})$ are known at all points where $\sigma_m(\mathbf{m})$ has a significant value, these estimators can also be computed.

If the 'model space' $\mathcal{M}$ has a large number of dimensions (say from five to many millions or billions), then an exhaustive exploration of the space is not possible, and we must turn to Monte Carlo sampling methods to extract information from $\sigma_m(\mathbf{m})$ . We discuss the application of Monte Carlo methods to inverse problems, and optimization techniques in section 6 and 7, respectively.

## 4.6 Physical Laws as Probabilistic Correlations

### 4.6.1 Physical Laws

We return here to the general case where it is not assumed that the total space $(\mathcal{M}, \mathcal{D})$ is the Cartesian product of two spaces.

In section 4.5 we have examined the situation where the physical correlation between the parameters of the problem are expressed using an exact, analytic expression $\mathbf{d} = \mathbf{g}(\mathbf{m})$ . In this case, the notion of conditional probability density has been used to combine the 'physical theory' with the 'data' and the 'a priori information' on model parameters.

But we have seen that in order to properly define the notion of conditional probability density, it has been necessary to introduce a metric over the space, and to take a limit using the metric of the space. This is equivalent to put some 'thickness' around the theoretical relation $\mathbf{d} = \mathbf{g}(\mathbf{m})$ , and to take the limit when the thickness tends to zero.

But actual theories have some uncertainties, and, for more generality, it is better to explicitly introduce these uncertainties. Assume, then, that the physical correlations between the model parameters $\mathbf{m}$ and the data parameters $\mathbf{d}$ are not represented by an analytical expression like $\mathbf{d} = \mathbf{f}(\mathbf{m})$ , but by a probability density

$$\vartheta(\mathbf{m}, \mathbf{d}) \quad . \tag{61}$$

[17] Practically, it may correspond to the output of some 'black box' solving the 'forward problem'.

**Example: Realistic 'Uncertainty Bars' Around a Functional Relation**

In the approximation of a constant gravity field, with acceleration $\mathbf{g}$, the position at time $t$ of an apple in free fall is $\mathbf{r}(t) = \mathbf{r}_0 + \mathbf{v}_0\, t + \frac{1}{2}\,\mathbf{g}\, t^2$, where $\mathbf{r}_0$ and $\mathbf{v}_0$ are, respectively, the position and velocity of the object at time $t = 0$. More simply, if the movement is 1D,

$$x(t) \;=\; x_0 + v_0\, t + \frac{1}{2}\, g\, t^2 \quad . \tag{62}$$

Of course, or many reasons this equation can never be exact: air friction, wind effects, inhomogeneity of the gravity field, effects of the Earth rotation, forces from the Sun and the Moon (not to mention Pluto), relativity (special and general), etc.

It is not a trivial task, given very careful experimental conditions, to estimate the size of the leading uncertainty. Although one may think of an equation $x = x(t)$ as a line, infinitely thin, there will always be sources of uncertainty (at least due to the unknown limits of validity of general relativity): looking at the line with a magnifying glass should reveal a fuzzy object of finite thickness. As a simple example, let us examine here the mathematical object we arrive at when assuming that the leading sources of uncertainty in the relation $x = x(t)$ are the uncertainties in the initial position and velocity of the falling apple. Let us assume that:

- the initial position of the apple is random, with a Gaussian distribution centered at $x_0$, and with standard deviation $\sigma_x$;
- the initial velocity of the apple is random, with a Gaussian distribution centered at $v_0$, and with standard deviation $\sigma_v$;

Then, it can be shown that at a given time $t$, the possible positions of the apple are random, with probability density

$$\vartheta(x|t) \;=\; \frac{1}{\sqrt{2\,\pi}\,\sqrt{\sigma_x^2 + \sigma_v^2\, t^2}}\; \exp\left( -\frac{1}{2}\frac{\left(x - (x_0 + v_0\, t + \frac{1}{2}\, g\, t^2)\right)^2}{\sigma_x^2 + \sigma_v^2\, t^2} \right) \quad . \tag{63}$$

This is obviously a conditional probability density for $x$, given $t$. Should we have any reason to choose some marginal probability density $\vartheta_t(t)$, then, the 'law' for the fall of the apple would be

$$\vartheta(x,t) \;=\; \vartheta(x|t)\; \vartheta_t(t) \quad . \tag{64}$$

See appendix C for more details.

### 4.6.2 Inverse Problems

We have seen that the result of measurements can be represented by a probability density $\rho_d(\mathbf{d})$ in the data space. We have also seen that the a priori information on the model parameters can be represented by another probability density $\rho_m(\mathbf{m})$ in the model space. When we talk about 'measurements' and about 'a priori information on model parameters', we usually mean that we have a joint probability density in the $(\mathcal{M}, \mathcal{D})$ space, that is $\rho(\mathbf{m}, \mathbf{d}) = \rho_m(\mathbf{m})\, \rho_d(\mathbf{d})$. But let us consider the more general situation where for the whole set of parameters $(\mathcal{M}, \mathcal{D})$ we have some information that can be represented by a joint probability density $\rho(\mathbf{m}, \mathbf{d})$. Having well in mind the interpretation of this information, let us use the simple name of 'experimental information' for it

$$\rho(\mathbf{m}, \mathbf{d}) \qquad \text{(experimental information)} \; . \tag{65}$$

We have also seen that we have information coming from physical theories, that predict correlations between the parameters, and it has been argued that a probabilistic description of these correlations is well adapted to the resolution of inverse problems[18]. Let $\vartheta(\mathbf{m}, \mathbf{d})$ be the probability density representing this 'theoretical information':

$$\vartheta(\mathbf{m}, \mathbf{d}) \qquad \text{(theoretical information)} \; . \tag{66}$$

A quite fundamental assumption is that in all the spaces we consider, there is a notion of volume which allows to give sense to the notion of 'homogeneous probability distribution' over the space. The corresponding probability density is not constant, but is proportional to the volume element of the space (see section 2.2):

$$\mu(\mathbf{m}, \mathbf{d}) \qquad \text{(homogeneous probability distribution)} \; . \tag{67}$$

[18] Remember that, even if we wish to use a simple method based on the notion of conditional probability density, an analytic expression like $\mathbf{d} = \mathbf{f}(\mathbf{m})$ needs some 'thickness' before going to the limit defining the conditional probability density. This limit crucially depends on the 'thickness', i.e., on the type of uncertainties the theory contains.

Finally, we have seen examples suggesting that the conjunction of of the experimental information with the theoretical information corresponds exactly to the AND operation defined over the probability densities, to obtain the 'conjunction of information', as represented by the probability density

$$\sigma(\mathbf{m},\mathbf{d}) \;=\; k\,\frac{\rho(\mathbf{m},\mathbf{d})\;\vartheta(\mathbf{m},\mathbf{d})}{\mu(\mathbf{m},\mathbf{d})} \qquad\qquad \text{(conjunction of informations)} \quad , \tag{68}$$

with marginal probability densities

$$\sigma_m(\mathbf{m}) \;=\; \int_{\mathcal{D}} d\mathbf{d}\;\sigma(\mathbf{m},\mathbf{d}) \qquad ; \qquad \sigma_d(\mathbf{d}) \;=\; \int_{\mathcal{M}} d\mathbf{m}\;\sigma(\mathbf{m},\mathbf{d}) \quad . \tag{69}$$

**Example 15** *We may assume that the physical correlations between the parameters* $\mathbf{m}$ *and* $\mathbf{d}$ *are of the form*

$$\vartheta(\mathbf{m},\mathbf{d}) \;=\; \vartheta_{D|M}(\mathbf{d}|\mathbf{m})\;\vartheta_M(\mathbf{m}) \quad , \tag{70}$$

*this expressing that a 'physical theory' gives, one the one hand, the conditional probability for* $\mathbf{d}$ *, given* $\mathbf{m}$ *, and on the other hand, the marginal probability density for* $\mathbf{m}$ *. See appendix C for more details.* [END OF EXAMPLE.]

**Example 16** *Many applications concern the special situation where we have*

$$\mu(\mathbf{m},\mathbf{d}) \;=\; \mu_m(\mathbf{m})\;\mu_d(\mathbf{d}) \qquad ; \qquad \rho(\mathbf{m},\mathbf{d}) \;=\; \rho_m(\mathbf{m})\;\rho_d(\mathbf{d}) \quad . \tag{71}$$

*In this case, equations 68–69 give*

$$\sigma_m(\mathbf{m}) \;=\; k\;\frac{\rho_m(\mathbf{m})}{\mu_m(\mathbf{m})}\;\int_{\mathcal{D}} d\mathbf{d}\;\frac{\rho_d(\mathbf{d})\;\vartheta(\mathbf{m},\mathbf{d})}{\mu_d(\mathbf{d})} \quad . \tag{72}$$

*If equation 70 holds, then*

$$\sigma_m(\mathbf{m}) \;=\; k\;\rho_m(\mathbf{m})\;\frac{\vartheta_m(\mathbf{m})}{\mu_m(\mathbf{m})}\;\int_{\mathcal{D}} d\mathbf{d}\;\frac{\rho_d(\mathbf{d})\;\vartheta_{D|M}(\mathbf{d}\mid\mathbf{m})}{\mu_d(\mathbf{d})} \quad . \tag{73}$$

*Finally, if the simplification* $\vartheta_M(\mathbf{m}) = \mu_m(\mathbf{m})$ *arises (see appendix C for an illustration) then,*

$$\sigma_m(\mathbf{m}) \;=\; k\;\rho_m(\mathbf{m})\;\int_{\mathcal{D}} d\mathbf{d}\;\frac{\rho_d(\mathbf{d})\;\vartheta(\mathbf{d}|\mathbf{m})}{\mu_d(\mathbf{d})} \quad . \tag{74}$$

[END OF EXAMPLE.]

**Example 17** *In the context of the previous example, assume that observational uncertainties are Gaussian,*

$$\rho_d(\mathbf{d}) \;=\; k\,\exp\left(-\frac{1}{2}\,(\mathbf{d}-\mathbf{d}_{\text{obs}})^t\;\mathbf{C}_D^{-1}\;(\mathbf{d}-\mathbf{d}_{\text{obs}})\right) \quad . \tag{75}$$

*Note that the limit for infinite variances gives the homogeneous probability density* $\mu_d(\mathbf{d}) = k$ *. Furthermore, assume that uncertainties in the physical law are also Gaussian:*

$$\vartheta(\mathbf{d}|\mathbf{m}) \;=\; k\,\exp\left(-\frac{1}{2}\,(\mathbf{d}-\mathbf{f}(\mathbf{m}))^t\;\mathbf{C}_T^{-1}\;(\mathbf{d}-\mathbf{f}(\mathbf{m}))\right) \quad . \tag{76}$$

*Here 'the physical theory says' that the data values must be 'close' to the 'computed values'* $\mathbf{f}(\mathbf{m})$ *, with a notion of closeness defined by the 'theoretical covariance matrix'* $\mathbf{C}_T$ *. As demonstrated in Tarantola (1987, page 158), the integral in equation 74 can be analytically evaluated, and gives*

$$\int_{\mathcal{D}} d\mathbf{d}\;\frac{\rho_d(\mathbf{d})\;\vartheta(\mathbf{d}|\mathbf{m})}{\mu_d(\mathbf{d})} \;=\; k\,\exp\left(-\frac{1}{2}\,(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})^t\;(\mathbf{C}_D+\mathbf{C}_T)^{-1}\;(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})\right) \quad . \tag{77}$$

*This shows that when using the Gaussian probabilistic model, observational and theoretical uncertainties combine through addition of the respective covariance operators (a nontrivial result).* [END OF EXAMPLE.]

**Example 18** *In the 'Galilean law' example developed in section 4.6.1, we described the correlation between the position* $x$ *and the time* $t$ *of a free falling object through a probability density* $\vartheta(x,t)$ *. This law says than falling objects describe, approximately, a space-time parabola. Assume that in a particular experiment the falling object explodes at some point of its space-time trajectory A plain measurement of the coordinates* $(x,t)$ *of the event gives the probability density* $\rho(x,t)$ *. By 'plain measurement' we mean here that we have used a measurement technique that is not taking into account the particular parabolic character of the fall (i.e., the measurement is designed to work identically for any sort of trajectory). The conjunction of the physical law* $\vartheta(x,t)$ *and the experimental result* $\rho(x,t)$ *, using expression 68, gives*

$$\sigma(x,t) \;=\; k\,\frac{\rho(x,t)\,\vartheta(x,t)}{\mu(x,t)} \quad , \tag{78}$$

*where, as the coordinates* $(x,t)$ *are 'Cartesian',* $\mu(x,t) = k$ *. Taking the explicit expression given for* $\vartheta(x,t)$ *in equations 63–64, with* $\vartheta_t(t) = k$ *,*

$$\vartheta(x,t) \;=\; \frac{1}{\sqrt{2\,\pi}\,\sqrt{\sigma_x^2+\sigma_v^2\,t^2}}\;\exp\left(-\frac{1}{2}\frac{\left(x-(x_0+v_0\,t+\frac{1}{2}\,g\,t^2)\right)^2}{\sigma_x^2+\sigma_v^2\,t^2}\right) \;, \tag{79}$$

*and assuming the Gaussian form*[19] *for* $\rho(x,t)$ *,*

$$\rho(x,t) \;=\; \rho_x(x)\,\rho_t(t) \;=\; k\,\exp\left(-\frac{1}{2}\frac{(x-x_{\rm obs})^2}{\Sigma_x^2}\right)\;\exp\left(-\frac{1}{2}\frac{(t-t_{\rm obs})^2}{\Sigma_t^2}\right) \quad , \tag{80}$$

*we obtain the combined probability density*

$$\sigma(x,t) \;=\; \frac{k}{\sqrt{\sigma_x^2+\sigma_v^2\,t^2}}\;\exp\left(-\frac{1}{2}\left(\frac{(x-x_{\rm obs})^2}{\Sigma_x^2}+\frac{(t-t_{\rm obs})^2}{\Sigma_t^2}+\frac{\left(x-(x_0+v_0\,t+\frac{1}{2}\,g\,t^2)\right)^2}{\sigma_x^2+\sigma_v^2\,t^2}\right)\right) \;. \tag{81}$$

*Figure 7 illustrates the three probability densities* $\vartheta(x,t)$ *,* $\rho(x,t)$ *and* $\sigma(x,t)$ *. See appendix C for a more detailed examination of this problem.* [END OF EXAMPLE.]

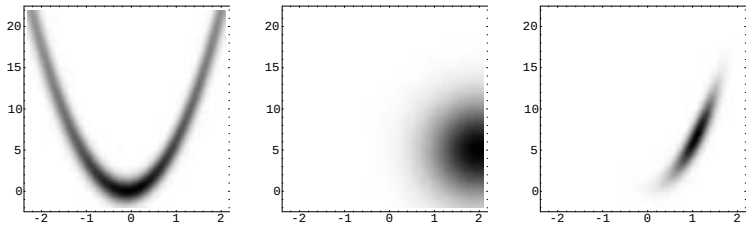



Figure 7: This figure has been was made with the numerical values mentioned in figure 17 (see appendix C) with, in addition, $x_{\rm obs} = 5.0\,{\rm m}$ , $\Sigma_x = 4.0\,{\rm m}$ , $t_{\rm obs} = 2.0\,{\rm s}$ and $\Sigma_t = 0.75\,{\rm s}$ .

# 5 Solving Inverse Problems (I): Examination of the Probability Density

The next two sections deal with Monte Carlo and optimization methods. The implementation of these methods takes some programming effort that is not required when we face problems with fewer degrees of freedom (say, between one to five).

When we have a small number of parameters we should directly 'plot' the probability density.

In appendix L the problem of estimation of a seismic hypocenter is treated, and it is shown there that the examination of the probability density for the location of the hypocenter offers a much better possibility for analysis than any other method.

[19] Note that taking the limit of $\vartheta(x,t)$ or of $\rho(x,t)$ for infinite variances we obtain $\mu(x,t)$ , as we should.

# 6 Solving Inverse Problems (II): Monte Carlo Methods

## 6.1 Basic Equations

The starting point could be the explicit expression (equation 46) for $\sigma_m(\mathbf{m})$ given in section 4.5.2:

$$\sigma_m(\mathbf{m}) \;=\; k\,\rho_m(\mathbf{m})\,L(\mathbf{m}) \quad . \tag{82}$$

where

$$L(\mathbf{m}) \;=\; \left( \frac{\rho_d(\mathbf{d})}{\sqrt{\det \mathbf{g}_d(\mathbf{d})}} \; \frac{\sqrt{\det \left(\mathbf{g}_m(\mathbf{m}) + \mathbf{F}^T(\mathbf{m})\,\mathbf{g}_d(\mathbf{d})\,\mathbf{F}(\mathbf{m})\right)}}{\sqrt{\det \mathbf{g}_m(\mathbf{m})}} \right) \Bigg|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} \quad . \tag{83}$$

In this expression the matrix of partial derivatives $\mathbf{F} = \mathbf{F}(\mathbf{m})$ , with components $D_{i\alpha} = \partial f_i / \partial m_\alpha$ , appears. The 'slope' $\mathbf{F}$ enters here because the steeper the slope for a given $\mathbf{m}$ , the greater the accumulation of points we will have with this particular $\mathbf{m}$ . This is because we use explicitly the analytic expression $\mathbf{d} = \mathbf{f}(\mathbf{m})$ . One should realize that using the more general approach based on equation 68 of section 4.6.2, the effect is automatically accounted for, and there is no need to explicitly consider the partial derivatives.

Equation 82 has the standard form of a conjunction of two probability densities, and is therefore ready to be integrated in a Metropolis algorithm. But one should note that, contrary to many 'nonlinear' formulations of inverse problems, the partial derivatives $\mathbf{F}$ are needed even if we use a Monte Carlo method.

In some weakly nonlinear problems, we have $\mathbf{F}^T(\mathbf{m})\,\mathbf{g}_d(\mathbf{d})\,\mathbf{F}(\mathbf{m}) << \mathbf{g}_m(\mathbf{m})$ and, then, equation 83 becomes

$$L(\mathbf{m}) \;=\; \frac{\rho_d(\mathbf{d})}{\mu_d(\mathbf{d})} \Bigg|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} \quad , \tag{84}$$

where we have used $\mu_d(\mathbf{d}) = k\,\sqrt{\det \mathbf{g}_d(\mathbf{d})}$ (see rule 2).

This expression is also ready for use in the Metropolis algorithm. In this way sampling of the prior $\rho_m(\mathbf{m})$ is modified into a sampling of the posterior $\sigma_m(\mathbf{m})$, and the Metropolis Rule uses the "Likelihood function" $L(\mathbf{m})$ to calculate acceptance probabilities.

## 6.2 Sampling the Homogeneous Probability Distribution

If we do not have an algorithm that samples the prior probability density directly, the first step in a Monte Carlo analysis of an inverse problem is to design a random walk that samples the model space according to the homogeneous probability distribution $\mu_m(\mathbf{m})$ . In some cases this is easy, but in other cases only an algorithm (a *primeval random walk*) that samples an arbitrary (possibly constant) probability density $\psi(\mathbf{m}) \neq \mu_m(\mathbf{m})$ is available. Then the Metropolis Rule can be used to modify $\psi(\mathbf{m})$ into $\mu_m(\mathbf{m})$ (see section 3.4). This way of generating samples from $\mu_m(\mathbf{m})$ is efficient if $\psi(\mathbf{m})$ is close to $\mu_m(\mathbf{m})$, otherwise it may be very inefficient.

Once $\mu(\mathbf{m})$ can be sampled, the Metropolis Rule allows us to modify this sampling into an algorithm that samples the prior.

## 6.3 Sampling the Prior Probability Distribution

The first step in the Monte Carlo analysis is to temporarily 'switch off' the comparison between computed and observed data, thereby generating samples of the prior probability density. This allows us to verify statistically that the algorithm is working correctly, and it allows us to understand the prior information we are using. We will refer to a large collection of models representing the prior probability distribution as the "prior movie" (in a computer screen, when the models are displayed one after the other, we have a 'movie'). The more models present in this movie, the more accurate representation of the prior probability density.

## 6.4 Sampling the Posterior Probability Distribution

If we now switch on the comparison between computed and observed data using, e.g., the Metropolis Rule for the actual equation 82, the random walk sampling the prior distribution is modified into a walk sampling the posterior distribution.

Since data rarely put strong constraints on the Earth, the "posterior movie" typically shows that many different models are possible. But even though the models in the posterior movie may be quite different, all of them predict

data that, within experimental uncertainties, are models with high likelihood. In other words, we must accept that data alone cannot have a preferred model.

The posterior movie allows us to perform a proper resolution analysis that helps us to choose between different interpretations of a given data set. Using the movie we can answer complicated questions about the correlations between several model parameters. To answer such questions, we can view the posterior movie and try to discover structure that is well resolved by data. Such structure will appear as "persistent" in the posterior movie.

The 'movie' can be used to answer quite complicated questions. For instance, to answer the question '*which is the probability that the Earth has this special characteristic, but not having this other special characteristic?*' we can just count the number $n$ of models (samples) satisfying the criterion, and the probability is $P = n/m$ , where $m$ is the total number of samples.

Once this 'movie' is generated, it is, of course, possible to represent the 1D or 2D marginal probability densities for all or for some selected parameters: it is enough to concentrate one's attention to those selected parameters in each of the samples generated. Those marginal probability densities may have some pathologies (like being multimodal, or having infinite dispersions), but those are the general characteristics of the joint probability density. Our numerical experience shows that these marginals are, quite often, 'stable' objects, in the sense that they can be accurately determined with only a small number of samples.

If the marginals are, essentially, beautiful bell-shaped distributions, then, one may proceed to just computing mean vales and standard deviations (or median values and mean deviations), using each of the samples and the elementary statistical formulas.

Another, more traditional, way of investigating resolution is to calculate covariances and higher order moments. For this we need to evaluate integrals of the form

$$R_f \;=\; \int_{\mathcal{A}} d\mathbf{m}\; f(\mathbf{m})\; \sigma_m(\mathbf{m}) \tag{85}$$

where $f(\mathbf{m})$ is a given function of the model parameters and $\mathcal{A}$ is an event in the model space $\mathcal{M}$ containing the models we are interested in. For instance,

$$\mathcal{A} \;=\; \{\mathbf{m}\,|\text{a given range of parameters in } \mathbf{m} \text{ is } \textit{cyclic}\}\,. \tag{86}$$

In the special case when $\mathcal{A} = \mathcal{M}$ is the entire model space, and $f(\mathbf{m}) = m_i$, the $R_f$ in equation (85) equals the mean $\langle m_i \rangle$ of the $i$'th model parameter $m_i$. If $f(\mathbf{m}) = (m_i - \langle m_i \rangle)\,(m_j - \langle m_j \rangle)$, $R_f$ becomes the covariance between the $i$'th and $j$'th model parameters. Typically, in the general inverse problem we cannot evaluate the integral in (85) analytically because we have no analytical expression for $\sigma(\mathbf{m})$. However, from the samples of the posterior movie $\mathbf{m}_1, \ldots, \mathbf{m}_n$ we can approximate $R_f$ by the simple average

$$R_f \approx \frac{1}{\text{total number of models}} \sum_{\{i|\mathbf{m}_i \in \mathcal{A}\}} f(\mathbf{m}_i) \quad . \tag{87}$$

# 7 Solving Inverse Problems (III): Deterministic Methods

As we have seen, the solution of an inverse problem essentially consists of a probability distribution over the space of all possible models of the physical system under study. In general, this 'model space' is high-dimensional, and the only general way to explore it is by using the Monte Carlo methods developed in section 3.

If the probability distributions are 'bell-shaped' (i.e., if they look like a Gaussian or like a generalized Gaussian), then one may simplify the problem by calculating only the point around which the probability is maximum, with an approximate estimation of the variances and covariances. This is the problem addressed in this section. Among the many methods available to obtain the point at which a scalar function reaches its maximum value (relaxation methods, linear programming techniques, etc.) we limit our scope here to the methods using the gradient of the function, which we assume can be computed analytically or, at least, numerically. For more general methods, the reader may have a look at Fletcher, (1980, 1981), Powell (1981), Scales (1985), Tarantola (1987) or Scales et al. (1992).

## 7.1 Maximum Likelihood Point

Let us consider a space $\mathcal{X}$ , with a volume element $dV$ defined. If the coordinates $\mathbf{x} \equiv \{x^1, x^2, \ldots, x^n\}$ are chosen over the space, the volume element has an expression $dV(\mathbf{x}) = v(\mathbf{x})\, d\mathbf{x}$ , and each probability distribution over $\mathcal{X}$ can be represented by a probability density $f(\mathbf{x})$ . For any fixed small volume $\Delta V$ we can search for the point $\mathbf{x}_{\text{ML}}$ such that the probability $dP$ of the small volume, when centered around $\mathbf{x}_{\text{ML}}$ , attains a maximum. In the limit $\Delta V \to 0$ this defines the *maximum likelihood point*. The maximum likelihood point may be unique (if the probability distribution is unimodal), may be degenerated (if the probability distribution is 'chevron-shaped') or may be multiple (as when we have the sum of a few bell-shaped functions).

The maximum likelihood point is **not** the point at which the probability density is maximum. Our definition implies that a maximum must be attained by the ratio between the probability density and the function $v(\mathbf{x})$ defining the volume element [20]:

$$\mathbf{x} = \mathbf{x}_{\text{ML}} \iff F(\mathbf{x}) = \frac{f(\mathbf{x})}{v(\mathbf{x})} \quad \text{maximum} \quad . \tag{88}$$

As the homogeneous probability density is $\mu(\mathbf{x}) = k\, v(\mathbf{x})$ (see rule 2), we can equivalently define the maximum likelihood point by the condition

$$\mathbf{x} = \mathbf{x}_{\text{ML}} \iff \frac{f(\mathbf{x})}{\mu(\mathbf{x})} \quad \text{maximum} \quad . \tag{89}$$

The point at which a probability density has its maximum is, in general, not $\mathbf{x}_{\text{ML}}$ . In fact, the maximum of a probability density does not correspond to an intrinsic definition of a point: a change of coordinates $\mathbf{x} \mapsto \mathbf{y} = \boldsymbol{\psi}(\mathbf{x})$ would change the probability density $f(\mathbf{x})$ into the probability density $g(\mathbf{y})$ (obtained using the Jacobian rule), but the point of the space at which $f(\mathbf{x})$ is maximum is not the same as the point of the space where $g(\mathbf{y})$ is maximum (unless the change of variables is linear). This contrasts with the maximum likelihood point, as defined by equation 89, that is an intrinsically defined point: no matter which coordinates we use in the computation we always obtain the same point of the space.

## 7.2 Misfit

One of the goals here is to develop gradient-based methods for obtaining the maximum of $F(\mathbf{x}) = f(\mathbf{x})/\mu(\mathbf{x})$ . As a quite general rule, gradient-based methods perform quite poorly for (bell-shaped) probability distributions, as when one is far from the maximum the probability densities tend to be quite flat, and it is difficult to get, reliably, the direction of steepest ascent. Taking a logarithm transforms a bell-shaped distribution into a paraboloid-shaped distribution on which gradient methods work well.

The logarithmic volumetric probability, or *misfit*, is defined as $S(\mathbf{x}) = -\log(F(\mathbf{x})/F_0)$ , where $p'$ and $F_0$ are two constants, and is given by

$$S(\mathbf{x}) = -\log \frac{f(\mathbf{x})}{\mu(\mathbf{x})} \quad . \tag{90}$$

The problem of maximization of the (typically) bell-shaped function $f(\mathbf{x})/\mu(\mathbf{x})$ has been transformed into the problem of minimization of the (typically) paraboloid-shaped function $S(\mathbf{x})$ :

$$\mathbf{x} = \mathbf{x}_{\text{ML}} \iff S(\mathbf{x}) \quad \text{minimum} \quad . \tag{91}$$

**Example 19** *The conjunction $\sigma(\mathbf{x})$ of two probability densities $\rho(\mathbf{x})$ and $\vartheta(\mathbf{x})$ was defined (equation 13) as*

$$\sigma(\mathbf{x}) = p\, \frac{\rho(\mathbf{x})\, \vartheta(\mathbf{x})}{\mu(\mathbf{x})} \quad . \tag{92}$$

*Then,*

$$S(\mathbf{x}) = S_\rho(\mathbf{x}) + S_\vartheta(\mathbf{x}) \quad , \tag{93}$$

*where*

$$S_\rho(\mathbf{x}) = -\log \frac{\rho(\mathbf{x})}{\mu(\mathbf{x})} \qquad ; \qquad S_\vartheta(\mathbf{x}) = -\log \frac{\vartheta(\mathbf{x})}{\mu(\mathbf{x})} \quad . \tag{94}$$

[END OF EXAMPLE.]

[20] The ratio $F(\mathbf{x}) = f(\mathbf{x})/v(\mathbf{x})$ is what we refer to as *the volumetric probability* associated to the probability density $f(\mathbf{x})$ . See appendix A.

**Example 20** *In the context of Gaussian distributions we have found the probability density (see example 12)*

$$\sigma_m(\mathbf{m}) \;=\; \tag{95}$$

$$=\; k\,\exp\left(-\frac{1}{2}\left((\mathbf{m}-\mathbf{m}_{\text{prior}})^t\,\mathbf{C}_M^{-1}\,(\mathbf{m}-\mathbf{m}_{\text{prior}})+(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})^t\,\mathbf{C}_D^{-1}\,(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})\right)\right) \quad .$$

*The limit of this distribution for infinite variances is a constant, so in this case* $\mu_m(\mathbf{m}) = k$ *. The misfit function* $S(\mathbf{m}) = -\log(\,\sigma_m(\mathbf{m})/\mu_m(\mathbf{m})\,)$ *is then given by*

$$2\,S(\mathbf{m}) \;=\; (\mathbf{m}-\mathbf{m}_{\text{prior}})^t\,\mathbf{C}_M^{-1}\,(\mathbf{m}-\mathbf{m}_{\text{prior}})+(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}})^t\,\mathbf{C}_D^{-1}\,(\mathbf{f}(\mathbf{m})-\mathbf{d}_{\text{obs}}) \quad . \tag{96}$$

*The reader should remember that this misfit function is valid only for weakly nonlinear problems (see examples 10 and 12). The maximum likelihood model here is the one that minimizes the sum of squares 96. This correpponds to the* least squares criterion. [END OF EXAMPLE.]

## 7.3 Gradient and Direction of Steepest Ascent

One must not consider as synonymous the notions of 'gradient' and 'direction of steepest ascent'. Consider, for instance, an *adimensional* misfit function[21] $S(P,T)$ over a pressure $P$ and a temperature $T$. Any sensible definition of the gradient of $S$ will lead to an expression like

$$\text{grad}\; S \;=\; \begin{pmatrix} \frac{\partial S}{\partial P} \\ \\ \frac{\partial S}{\partial T} \end{pmatrix} \tag{97}$$

and this by no means can be regarded as a 'direction' in the $(P,T)$ space (for instance, the components of this 'vector' does not have the dimensions of pressure and temperature, but of inverse pressure and inverse temperature).

Mathematically speaking, *the gradient of a function* $S(\mathbf{x})$ *at a point* $\mathbf{x}_0$ *is the linear function that is tangent to* $S(\mathbf{x})$ *at* $\mathbf{x}_0$. This definition of gradient is consistent with the more elementary one, based on the use of the first order expansion

$$S(\mathbf{x}_0+\delta\mathbf{x}) \;=\; S(\mathbf{x}_0)+\widehat{\boldsymbol{\gamma}}_0^T\,\delta\mathbf{x}+\ldots \tag{98}$$

Here $\widehat{\boldsymbol{\gamma}}_0$ is called the gradient of $S(\mathbf{x})$ at point $\mathbf{x}_0$. It is clear that $S(\mathbf{x}_0)+\widehat{\boldsymbol{\gamma}}_0^T\,\delta\mathbf{x}$ is a linear function, and that it is tangent to $S(\mathbf{x})$ at $\mathbf{x}_0$, so the two definitions are in fact equivalent. Explicitly, the components of the gradient at point $\mathbf{x}_0$ are

$$(\widehat{\boldsymbol{\gamma}}_0)_p \;=\; \frac{\partial S}{\partial x^p}(\mathbf{x}_0) \quad . \tag{99}$$

Everybody is well trained at computing the gradient of a function (event if the interpretation of the result as a direction in the original space is wrong). How can we pass from the gradient to the direction of steepest ascent (a bona fide direction in the original space)? In fact, the gradient (at a given point) of a function defined over a given space $\mathcal{E}$ ) is an element of the dual of the space. To obtain a direction in $\mathcal{E}$ we must pass from the dual to the primal space. As usual, it is the metric of the space that maps the dual of the space into the space itself. So if $\mathbf{g}$ is the metric of the space where $S(\mathbf{x})$ is defined, and if $\widehat{\boldsymbol{\gamma}}$ is the gradient of $S$ at a given point, the *direction of steepest ascent* is

$$\boldsymbol{\gamma} \;=\; \mathbf{g}^{-1}\,\widehat{\boldsymbol{\gamma}} \quad . \tag{100}$$

The direction of steepest ascent must be interpreted as follows: if we are at a point $\mathbf{x}$ of the space, we can consider a very small hypersphere around $\mathbf{x}_0$. The direction of steepest ascent points towards the point of the sphere at which $S(\mathbf{x})$ attains its maximum value.

**Example 21** *In the context of least squares, we consider a misfit function* $S(\mathbf{m})$ *and a covariance matrix* $\mathbf{C}_M$. *If* $\widehat{\boldsymbol{\gamma}}_0$ *is the gradient of* $S$, *at a point* $\mathbf{x}_0$, *and if we use* $\mathbf{C}_M$ *to define distances in the space, the direction of steepest ascent is*

$$\boldsymbol{\gamma}_0 \;=\; \mathbf{C}_M\,\widehat{\boldsymbol{\gamma}}_0 \quad . \tag{101}$$

[END OF EXAMPLE.]

[21] We take this example because typical misfit functions are adimensional (have no physical dimensions) but the argument has general validity.

## 7.4 The Steepest Descent Method

Consider that we have a probability distribution defined over an $n$-dimensional space $\mathcal{X}$. Having chosen the coordinates $\mathbf{x} \equiv \{x^1, x^2, \ldots, x^n\}$ over the space, the probability distribution is represented by the probability density $f(\mathbf{x})$ whose homogeneous limit (in the sense developed in section 2.2) is $\mu(\mathbf{x})$. We wish to calculate the coordinates $\mathbf{x}_{\mathrm{ML}}$ of the maximum likelihood point. By definition (equation 89),

$$\mathbf{x} = \mathbf{x}_{\mathrm{ML}} \quad \Longleftrightarrow \quad \frac{f(\mathbf{x})}{\mu(\mathbf{x})} \quad \text{maximum} \quad , \tag{102}$$

i.e.,

$$\mathbf{x} = \mathbf{x}_{\mathrm{ML}} \quad \Longleftrightarrow \quad S(\mathbf{x}) \quad \text{minimum} \quad , \tag{103}$$

where $S(\mathbf{x})$ is the misfit (equation90)

$$S(\mathbf{x}) = -k \log \frac{f(\mathbf{x})}{\mu(\mathbf{x})} \quad . \tag{104}$$

Let us denote by $\widehat{\boldsymbol{\gamma}}(\mathbf{x}_k)$ the gradient of $S(\mathbf{x})$ at point $\mathbf{x}_k$, i.e. (equation 99),

$$(\widehat{\boldsymbol{\gamma}}_0)_p = \frac{\partial S}{\partial x^p}(\mathbf{x}_0) \quad . \tag{105}$$

We have seen above that $\widehat{\boldsymbol{\gamma}}(\mathbf{x})$ should not be interpreted as a direction in the space $\mathcal{X}$ but as a direction in the dual space. The gradient can be converted into a direction using a metric $\mathbf{g}(\mathbf{x})$ over $\mathcal{X}$. In simple situations the metric $\mathbf{g}$ will be the one used to define the volume element of the space, i.e., we will have $\mu(\mathbf{x}) = k\, v(\mathbf{x}) = k\, \sqrt{\det \mathbf{g}(\mathbf{x})}$, but this is not a necessity, and iterative algorithms may be accelerated by astute introduction of ad-hoc metrics.

Given, then, the gradient $\widehat{\boldsymbol{\gamma}}(\mathbf{x}_k)$ (at some particular point $\mathbf{x}_k$) to any possible choice of metric $\mathbf{g}(\mathbf{x})$ we can define the direction of steepest ascent associated to the metric $\mathbf{g}$, by (equation 101)

$$\boldsymbol{\gamma}(\mathbf{x}_k) = \mathbf{g}^{-1}(\mathbf{x}_k)\, \widehat{\boldsymbol{\gamma}}(\mathbf{x}_k) \quad . \tag{106}$$

The algorithm of steepest descent is an iterative algorithm passing from point $\mathbf{x}_k$ to point $\mathbf{x}_{k+1}$ by making a 'small jump' along the local direction of steepest descent,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \varepsilon_k\, \mathbf{g}_k^{-1}\, \widehat{\boldsymbol{\gamma}}_k \quad , \tag{107}$$

where $\varepsilon_k$ is an ad-hoc (real, positive) value adjusted to force the algorithm to converge rapidly (if $\varepsilon_k$ is chosen too small the convergence may be too slow; it is it chosen too large, the algorithm may even diverge).

Many elementary presentations of the steepest descent algorithm just forget to include the metric $\mathbf{g}_k$ in expression 107. These algorithms are not consistent. Even the physical dimensionality of the equation is not assured. 'Numerical' problems in computer implementations of steepest descent algorithms can often be traced to the fact that the metric has been neglected.

**Example 22** *In the context of example 20, where the misfit function $S(\mathbf{m})$ is given by*

$$2\, S(\mathbf{m}) = (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\mathrm{obs}})^t\, \mathbf{C}_D^{-1}\, (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\mathrm{obs}}) + (\mathbf{m} - \mathbf{m}_{\mathrm{prior}})^t\, \mathbf{C}_M^{-1}\, (\mathbf{m} - \mathbf{m}_{\mathrm{prior}}) \quad , \tag{108}$$

*the gradient $\widehat{\boldsymbol{\gamma}}$, whose components are $\widehat{\gamma}_\alpha = \partial S / \partial m^\alpha$, is given by the expression*

$$\widehat{\boldsymbol{\gamma}}(\mathbf{m}) = \mathbf{F}^t(\mathbf{m})\, \mathbf{C}_D^{-1}\, (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\mathrm{obs}}) + \mathbf{C}_M^{-1}\, (\mathbf{m} - \mathbf{m}_{\mathrm{prior}}) \quad , \tag{109}$$

*where $\mathbf{F}$ is the matrix of partial derivatives*

$$F^{i\alpha} = \frac{\partial f^i}{\partial m^\alpha} \quad . \tag{110}$$

*An example of computation of partial derivatives is given in appendix K.* [END OF EXAMPLE.]

**Example 23** *In the context of example 22 the model space* $\mathcal{M}$ *has an obvious metric, namely that defined by the inverse of the 'a priori'covariance operator* $\mathbf{g} = \mathbf{C}_M^{-1}$ *. Using this metric and the gradient given by equation 109, the steepest descent algorithm 107 becomes*

$$\mathbf{m}_{k+1} \;=\; \mathbf{m}_k - \varepsilon_k \left( \mathbf{C}_M \, \mathbf{F}_k^t \, \mathbf{C}_D^{-1} \, (\mathbf{f}_k - \mathbf{d}_{\text{obs}}) + (\mathbf{m}_k - \mathbf{m}_{\text{prior}}) \right) \quad , \tag{111}$$

*where* $\mathbf{F}_k \equiv \mathbf{F}(\mathbf{m}_k)$ *and* $\mathbf{f}_k \equiv \mathbf{f}(\mathbf{m}_k)$ *. The real positive quantities* $\varepsilon_k$ *can be fixed after some trial and error by accurate linear search, or by using a linearized approximation*[22]*.* [END OF EXAMPLE.]

**Example 24** *In the context of example 22 the model space* $\mathcal{M}$ *has a less obvious metric, namely that defined by the inverse of the 'posterior' covariance operator,* $\mathbf{g} = \widetilde{\mathbf{C}}_M^{-1}$ [23]*. Using this metric and the gradient given by equation 109, the steepest descent algorithm 107 becomes*

$$\mathbf{m}_{k+1} \;=\; \mathbf{m}_k - \varepsilon_k \Big( \mathbf{F}_k^t \, \mathbf{C}_D^{-1} \, \mathbf{F}_k + \mathbf{C}_M^{-1} \Big)^{-1} \left( \mathbf{F}_k^t \, \mathbf{C}_D^{-1} \, (\mathbf{f}_k - \mathbf{d}_{\text{obs}}) + \mathbf{C}_M^{-1} \, (\mathbf{m}_k - \mathbf{m}_{\text{prior}}) \right) \quad , \tag{113}$$

*where* $\mathbf{F}_k \equiv \mathbf{F}(\mathbf{m}_k)$ *and* $\mathbf{f}_k \equiv \mathbf{f}(\mathbf{m}_k)$ *. The real positive quantities* $\varepsilon_k$ *can be fixed, after some trial and error, by accurate linear search, or by using a linearized approximation that simply gives*[24] $\varepsilon_k \approx 1$ *.* [END OF EXAMPLE.]

The algorithm 113 is usually called a 'quasi-Newton algorithm'. This name is not well chosen: a Newton method applied to minimization of a misfit function $S(\mathbf{m})$ would be a method using the second derivatives of $S(\mathbf{m})$ , and thus the derivatives $H^i_{\alpha\beta} = \frac{\partial^2 f^i}{\partial m^\alpha \partial m^\beta}$ , that are not computed (or not estimated) when using this algorithm. It is just a steepest descent algorithm with a nontrivial definition of the metric in the working space. In this sense it belongs to the wider class of 'variable metric methods', not discussed in this article.

## 7.5 Estimating Posterior Uncertainties

In the Gaussian context, the Gaussian probability density that is tangent to $\sigma_m(\mathbf{m})$ has its center at the point given by the iterative algorithm

$$\mathbf{m}_{k+1} \;=\; \mathbf{m}_k - \varepsilon_k \left( \mathbf{C}_M \, \mathbf{F}_k^t \, \mathbf{C}_D^{-1} \, (\mathbf{f}_k - \mathbf{d}_{\text{obs}}) + (\mathbf{m}_k - \mathbf{m}_{\text{prior}}) \right) \quad , \tag{114}$$

(equation 111) or, equivalently, by the iterative algorithm

$$\mathbf{m}_{k+1} \;=\; \mathbf{m}_k - \varepsilon_k \Big( \mathbf{F}_k^t \, \mathbf{C}_D^{-1} \, \mathbf{F}_k + \mathbf{C}_M^{-1} \Big)^{-1} \left( \mathbf{F}_k^t \, \mathbf{C}_D^{-1} \, (\mathbf{f}_k - \mathbf{d}_{\text{obs}}) + \mathbf{C}_M^{-1} \, (\mathbf{m}_k - \mathbf{m}_{\text{prior}}) \right) \tag{115}$$

(equation 113). The covariance of the tangent gaussian is

$$\widetilde{\mathbf{C}}_M \;\approx\; \Big( \mathbf{F}_\infty^t \, \mathbf{C}_D^{-1} \, \mathbf{F}_\infty + \mathbf{C}_M^{-1} \Big)^{-1} \quad , \tag{116}$$

where $\mathbf{F}_\infty$ refers to the value of the matrix of partial derivatives at the convergence point.

## 7.6 Some Comments on the Use of Deterministic Methods

### 7.6.1 Linear, Weakly Nonlinear and Nonlinear Problems

There are different degrees of nonlinearity. Figure 9 illustrates four domains of nonlinearity, calling for different optimization algorithms. In this figure the abscissa symbolically represents the model space, and the ordinate represents the data space. The gray oval represents the combination of prior information on the model parameters, and information from the observed data[25]. It is the probability density $\rho(\mathbf{d}, \mathbf{m}) = \rho_d(\mathbf{d}) \; \rho_m(\mathbf{m})$ . seen elsewhere.

To fix ideas, the oval suggests here a Gaussian probability, but our distinction between problems according to their nonlinearity will not depend fundamentally on this.

[22] As shown in Tarantola (1987), if $\boldsymbol{\gamma}_k$ is the direction of steepest ascent at point $\mathbf{m}_k$ , i.e., $\boldsymbol{\gamma}_k = \mathbf{C}_M \, \mathbf{F}_k^t \, \mathbf{C}_D^{-1} \, (\mathbf{f}_k - \mathbf{d}_{\text{obs}}) + (\mathbf{m}_k - \mathbf{m}_{\text{prior}})$ , then, a local linearized approximation for the optimal $\varepsilon_k$ gives $\varepsilon_k = \dfrac{\boldsymbol{\gamma}_k^t \, \mathbf{C}_M^{-1} \, \boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_k^t (\mathbf{F}_k^t \, \mathbf{C}_D^{-1} \, \mathbf{F}_k + \mathbf{C}_M^{-1}) \boldsymbol{\gamma}_k}$ .

[23] The 'best estimator' of $\widetilde{\mathbf{C}}_M$ is

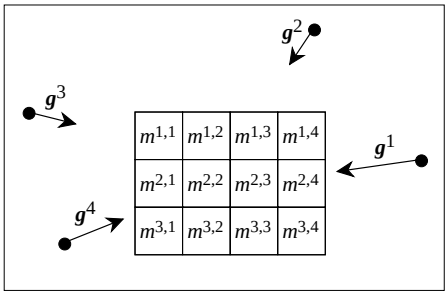
$$\widetilde{\mathbf{C}}_M \;\approx\; \Big( \mathbf{F}_k^t \, \mathbf{C}_D^{-1} \, \mathbf{F}_k + \mathbf{C}_M^{-1} \Big)^{-1} \quad . \tag{112}$$

See, e.g., Tarantola (1987)

[24] While a sensible estimation of the optimal values of the real positive quantities $\varepsilon_k$ is crucial for the algorithm 111, they can in many usual circumstances be dropped from the algorithm 113.

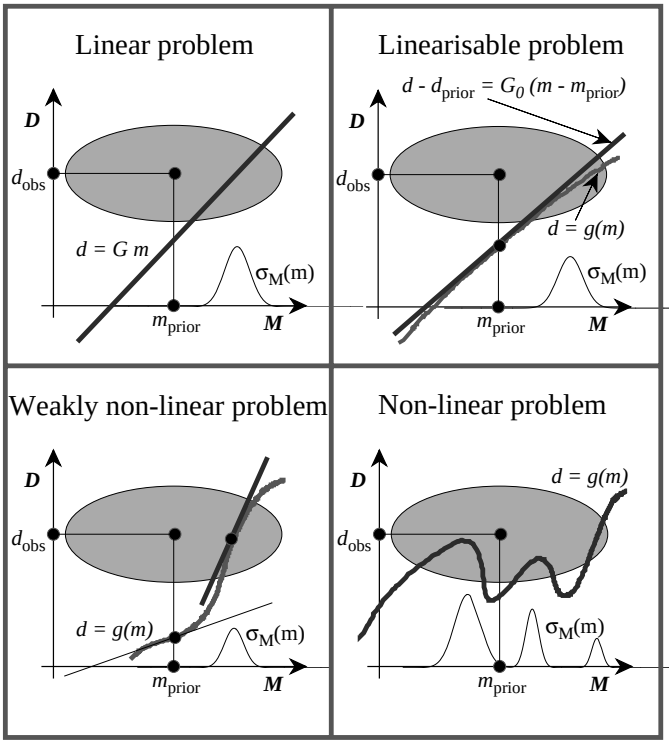
[25] The gray oval is the product of the probability density over the model space, representing the prior information, and the probability density over the data space representing the experimental results.

Figure 8: A simple example where we are interested in predicting the gravitational field $\mathbf{g}$ generated by a 2-D distribution of mass.




First, there are strictly linear problems. For instance, in the example illustrated by figure 8 the gravitational field $\mathbf{g}$ depends linearly on the masses inside the blocks[26]

Figure 9: Illustration of the four domains of nonlinearity, calling for different optimization algorithms. The model space is symbolically represented by the abscissa, and the data space is represented by the ordinate. The gray oval represents the combination of prior information on the model parameters, and information from the observed data. What is important is not an intrinsic nonlinearity of the function relating model parameters to data, but how linear the function is *inside the domain of significant probabilty*.




Strictly linear problems are illustrated at the top left of figure 9. The linear relationship between data and model parameters, $\mathbf{d} = \mathbf{G}\ \mathbf{m}$ , is represented by a straight line. The prior probability density $\rho(\mathbf{d},\mathbf{m})$ "induces" on this straight line the posterior probability density[27] $\sigma(\mathbf{d},\mathbf{m})$ whose "projection" over the model space gives gives the posterior probability density over the model parameter space, $\sigma_m(\mathbf{m})$ . Should the prior probability densities be Gaussian, then the posterior probability distribution would also be Gaussian: this is the simplest

[26]The gravitational field at point $\mathbf{x}_0$ generated by a distribution of volumetric mass $\rho(\mathbf{x})$ is given by

$$\mathbf{g}(\mathbf{x}_0) = \int dV(\mathbf{y})\ \frac{\mathbf{x}_0 - \mathbf{y}}{\|\mathbf{x}_0 - \mathbf{x}\|^3}\ \rho(\mathbf{x})\ .$$

When the volumetric mass is constant inside some predefined (2-D) volumes, as suggested in figure 8, this gives

$$\mathbf{g}(\mathbf{x}_0) = \sum_A \sum_B \mathbf{G}^{\mathrm{A,B}}(\mathbf{x}_0)\ m^{\mathrm{A,B}}\ .$$

This is a strictly linear equation between data (the gravitational field at a given observation point) and the model parameters (the masses inside the volumes). Note that if instead of choosing as model parameters the total masses inside some predefined volumes one chooses the geometrical parameters defining the sizes of the volumes, then the gravity field is not a linear function of the parameters. More details can be found in Tarantola and Valette (1982b, page 229).

[27]using the 'orthogonal-limit' method described in section 2.4.

situation.

Quasi-linear problems are illustrated at the bottom-left of figure 9. If the relationship linking the observable data $\mathbf{d}$ to the model parameters $\mathbf{m}$ ,

$$\mathbf{d} = \mathbf{g}(\mathbf{m}) \ , \tag{117}$$

is approximately linear *inside the domain of significant prior probability* (i.e., inside the gray oval of the figure), then the posterior distribution is just as simple as the prior distribution. For instance, if the prior is Gaussian the posterior is also Gaussian.

In this case also, the problem can be reduced to the computation of the mean and the covariance of the Gaussian. Typically, one begins at some "starting model" $\mathbf{m}_0$ (typically, one takes for $\mathbf{m}_0$ the "a priori model" $\mathbf{m}_{\text{prior}}$ ) [28], linearizing the function $\mathbf{d} = \mathbf{g}(\mathbf{m})$ around $\mathbf{m}_0$ and one looks for a model $\mathbf{m}_1$ "better than $\mathbf{m}_0$ ".

Iterating such an algorithm, one tends to the model $\mathbf{m}_\infty$ at which the "quasi-Gaussian" $\sigma_m(\mathbf{m})$ is maximum. The linearizations made in order to arrive to $\mathbf{m}_\infty$ are so far not an approximation: the point $\mathbf{m}_\infty$ is perfectly defined, independently of any linearization and any method used to find it. But once the convergence to this point has been obtained, a linearization of the function $\mathbf{d} = \mathbf{g}(\mathbf{m})$ around this point,

$$\mathbf{d} - \mathbf{g}(\mathbf{m}_\infty) = \mathbf{G}_\infty \ (\mathbf{m} - \mathbf{m}_\infty) \quad , \tag{118}$$

allows to obtain a good approximation to the posterior uncertainties. For instance, if the prior distribution is Gaussian this will give the covariance of the "tangent Gaussian".

Between linear and quasi-linear problems there are the "linearizable problems". The scheme at the top-right of figure 9 shows the case where the linearization of the function $\mathbf{d} = \mathbf{g}(\mathbf{m})$ around the prior model,

$$\mathbf{d} - \mathbf{g}(\mathbf{m}_{\text{prior}}) = \mathbf{G}_{\text{prior}} \ (\mathbf{m} - \mathbf{m}_{\text{prior}}) \quad , \tag{119}$$

gives a function that, inside the domain of significant probability, is very similar to the true (nonlinear) function.

In this case, there is no practical difference between this problem and the strictly linear problem, and the iterative procedure necessary for quasi-linear problems is here superfluous.

It remains to analyze the true nonlinear problems that, using a pleonasm, are sometimes called *strongly nonlinear problems*. They are illustrated at the bottom-right of figure 9.

In this case, even if the prior distribution is simple, the posterior distribution can be quite complicated. For instance, it can be multimodal. These problems are in general quite complex to solve, and only a Monte Carlo analysis, as described in the previous chapter, is feasible.

If full Monte Carlo methods cannot be used, because they are too expensive, then one can mix a random part (for instance, to choose the starting point) and a deterministic part. The optimization methods applicable to quasi-linear problems can, for instance, allow us to go from the randomly chosen starting point to the "nearest" optimal point. Repeating these computations for different starting points one can arrive at a good idea of the posterior distribution in the model space.

### 7.6.2 The Maximum Likelihood Model

The *most likely model* is, by definition, that at which the volumetric probability (see appendix A) $\sigma_\beta(\mathbf{m})$ attains its maximum. As $\sigma_\beta(\mathbf{m})$ is maximum when $S(\mathbf{m})$ is minimum, we see that the most likely model is also the the 'best model' obtained when using a 'least squares criterion'. Should we have used the double exponential model for all the uncertainties, then the most likely model would be defined by a 'least absolute values' criterion.

There are many circumstances where the most likely model is not an interesting model. One trivial example is when the volumetric probability has a 'narrow maximum', with small total probability (see figure 10). A much less trivial situation arises when the number of parameters is very large, as for instance when we deal with a random function (that, strictly speaking, corresponds to an infinite number of random variables). Figure 11 for instance, shows a few realizations of a Gaussian function with zero mean and an (approximately) exponential correlation. The most likely function is the center of the Gaussian, i.e., the null function shown at the left. But this is not a representative sample of the probability distribution, as any realization of the probability distribution will have, with a probability very close to one, the 'oscillating' characteristics of the three samples shown at the right.

[28] The term "a priori model" is an abuse of language. The correct term is "mean a priori model".

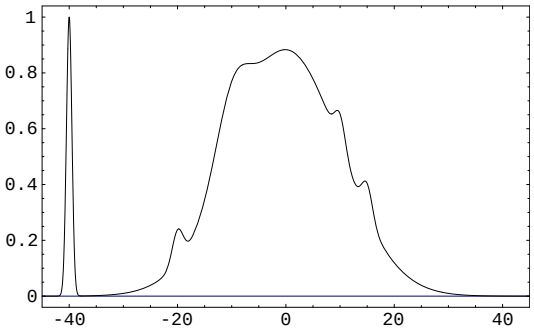



Figure 10: One of the circumstances where the 'maximum likelihood model' may not be very interesting is when it corresponds to a narrow maximum with small total probability, as the peak in the left part of this probability distribution.

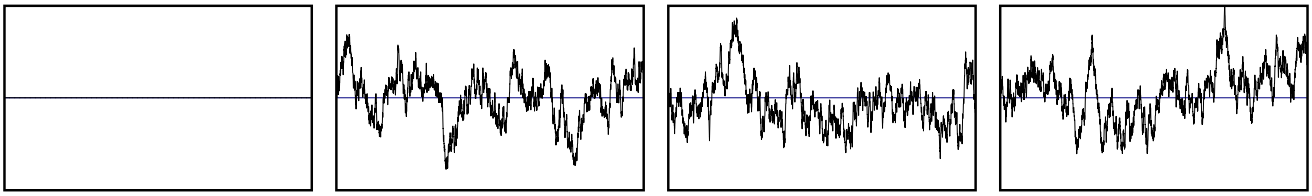


Figure 11: At the right, three random realizations of a Gaussian random function with zero mean and (approximately) exponential correlation function. The most likely function, i.e., the center of the Gaussian, is shown at the left. We see that the most likely function is not a representative of the probability distribution.

# 8 Conclusions

Probability theory is well adapted to the formulation of inverse problems, although its formulation must be rendered intrinsic (introducing explicitly the definition of distances in the working spaces, by redefining the notion of conditional probability density, and by introducing the notion of conjunction of states of information). The Metropolis algorithm is well adapted to the solution of inverse problems, as its inherent structure allows us to sequentially combine prior information, theoretical information, etc., and allows us to take advantage of the 'movie philosophy'. When a general Monte Carlo approach cannot be afforded, one can use simplified optimization techniques (like least squares). However, this usually requires strong simplifications that can only be made at the cost of realism.

# 9 Acknowledgements

We are very indebted to our colleagues (Bartolomé Coll, Miguel Bosch, Guillaume Évrard, John Scales, Christophe Barnes, Frédéric Parrenin and Bernard Valette) for illuminating discussions. We are also grateful to the students of the Geophysical Tomography Group, and the students at our respective institutes (in Paris and Copenhagen).

# 10 Bibliography

Aki, K. and Lee, W.H.K., 1976, Determination of three-dimensional velocity anomalies under a seismic array using first P arrival times from local earthquakes, J. Geophys. Res., **81**, 4381–4399.

Aki, K., Christofferson, A., and Husebye, E.S., 1977, Determination of the three-dimensional seismic structure of the lithosphere, J. Geophys. Res., **82**, 277-296.

Backus, G., 1970a. Inference from inadequate and inaccurate data: I, Proceedings of the National Academy of Sciences, 65, 1, 1-105.

Backus, G., 1970b. Inference from inadequate and inaccurate data: II, Proceedings of the National Academy of Sciences, 65, 2, 281-287.

Backus, G., 1970c. Inference from inadequate and inaccurate data: III, Proceedings of the National Academy of Sciences, 67, 1, 282-289.

Backus, G., 1971. Inference from inadequate and inaccurate data, Mathematical problems in the Geophysical Sciences: Lecture in applied mathematics, 14, American Mathematical Society, Providence, Rhode Island.

Backus, G., and Gilbert, F., 1967. Numerical applications of a formalism for geophysical inverse problems, Geophys. J. R. astron. Soc., 13, 247-276.

Backus, G., and Gilbert, F., 1968. The resolving power of gross Earth data, Geophys. J. R. astron. Soc., 16, 169-205.

Backus, G., and Gilbert, F., 1970. Uniqueness in the inversion of inaccurate gross Earth data, Philos. Trans. R. Soc. London, 266, 123-192.

Dahlen, F. A., Models of the lateral heterogeneity of the Earth consistent with eigenfrequency splitting data, *Geophys. J. R. Astron. Soc.*, **44**, 77–105, 1976.

Dahl-Jensen, D., Mosegaard, K., Gundestrup, N., Clow, G. D., Johnsen, S. J., Hansen, A. W., and Balling, N., 1998, Past temperatures directly from the Greenland Ice Sheet, *Science*, Oct. 9, 268–271.

Fisher, R.A., 1953, Dispersion on a sphere, Proc. R. Soc. London, A, **217**, 295–305.

Fletcher, R., 1980. Practical methods of optimization, Volume 1: Unconstrained optimization, Wiley.

Fletcher, R., 1981. Practical methods of optimization, Volume 2: Constrained optimization, Wiley.

Franklin, J.N., 1970. Well posed stochastic extensions of ill posed linear problems, J. Math. Anal. Applic., 31, 682-716.

Gauss, C.F., 1809, Theoria Motus Corporum Cœlestium.

Geiger, L., 1910, Herdbestimmung bei Erdbeben aus den Ankunftszeiten, Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, **4**, 331–349.

Geman, S., and Geman, D. 1984, IEEE Trans. Patt. Anal. Mach. Int., PAMI-6 6, 721.

Gilbert, F., 1971, Ranking and winnowing gross Earth data for inversion and resolution, *Geophys. J. R. Astron. Soc.*, **23**, 215–128.

Hadamard, J., 1902, Sur les problémes aux dérivées partielles et leur signification physique, Bull. Univ. Princeton, **13**.

Hadamard, J., 1932, Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques, Hermann, Paris.

ISO, 1993, Guide to the expression of uncertainty in measurement, International Organization for Standardization, Switzerland.

Jackson, D.D., The use of a priori data to resolve non-uniqueness in linear inversion, *Geophys. J. R. Astron. Soc.*, **57**, 137–157, 1979.

Jaynes, E.T., Prior probabilities, *IEEE Transactions on systems, science, and cybernetics*, Vol. SSC–4, No. 3, 227–241, 1968.

Jeffreys, H., 1939, Theory of probability, Clarendon Press, Oxford. Reprinted in 1961 by Oxford University Press.
Kandel, A., 1986, Fuzzy Mathematical Techniques with Applications, Addison-Wesley Pub. Co., Reading.

Keilis-Borok, V.J., and Yanovskaya, T.B., Inverse problems in seismology (structural review), *Geophys. J. R. astr. Soc.*, **13**, 223–234, 1967.

Kennett, B.L.N., and Nolet, G., Resolution analysis for discrete systems, Geophys, *Geophys. J. R. astr. Soc.*, **53**, 413–425, 1978.

Khan, A., Mosegaard, K., and Rasmussen, K. L., 2000, A New Seismic Velocity Model for the Moon from a Monte Carlo Inversion of the Apollo Lunar Seismic Data, *Geophys. Res. Lett.*, **37, 11**, 1,591–1,594, 2000.

Kimeldorf, G. and Wahba, G., 1970, A correspondence between Bayesian estimation of stochastic processes and smooting by splines, Ann. Math. Stat., **41**, 495–502.

Kullback, S., 1967, The two concepts of information, J. Amer. Statist. Assoc., **62**, 685–686.

Lehtinen, M.S., Päivärinta, L., and Somersalo, E., 1989, Linear inverse problems for generalized random variables, Inverse Problems, **5**,599–612.

Levenberg, K., 1944, A method for the solution of certain nonlinear problems in least-squares, Quart. Appl. Math., Vol. 2, 164–168.

Marquardt, D.W., 1963, An algorithm for least squares estimation of nonlinear parameters, SIAM J., **11**, 431–441.

Marquardt, D.W., 1970, Generalized inverses, ridge regression, biased linear estimation and non-linear estimation, Technometrics, **12**, 591–612.

Menke, W., 1984, Geophysical data analysis: discrete inverse theory, Academic Press.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E., Equation of State Calculations by Fast Computing Machines, *J. Chem. Phys.*, Vol. 1, No. 6, 1087–1092, 1953.

Minster, J.B. and Jordan, T.M., 1978, Present-day plate motions, J. Geophys. Res., **83**, 5331–5354.

Mosegaard, K., and Rygaard-Hjalsted, C., 1999, Bayesian analysis of implicit inverse problems, *Inverse Problems*, **15**, 573–583.

Mosegaard, K., Singh, S.C., Snyder, D., and Wagner, H., 1997, Monte Carlo Analysis of seismic reflections from Moho and the W-reflector, *J. Geophys. Res. B* , **102**, 2969–2981.

Mosegaard, K., and Tarantola, A., 1995, Monte Carlo sampling of solutions to inverse problems, *J. Geophys. Res.*, Vol. 100, No. B7, 12,431–12,447.

Nolet, G., Partitioned wave-form inversion and 2D structure under the NARS array, *J. Geophys. Res.*, **95**, 8499–8512, 1990.

Nolet, G., van Trier, J., and R. Huisman, A formalism for nonlinear inversion of seismic surface waves, *Geoph. Res. Lett.* **13**, 26–29, 1986.

Parker, R.L., 1994, Geophysical Inverse Theory, Princeton University Press.

Parzen, E., Tanabe, K. and Kitagawa G., eds., 1998, Selected papers of Hirotugu Akaike, Springer Series in Statistics, Springer-Verlag, New York.

Powell, M.J.D., 1981. Approximation theory and methods, Cambridge University Press.

Press, F., Earth models obtained by Monte Carlo inversion, *J. Geophys. Res.*, **73**, 5223–5234, 1968.

Rietsch, E., The maximum entropy approach to inverse problems, *J. Geophys.*, **42**, 489–506, 1977.

Scales, L. E., 1985. Introduction to non-linear optimization, Macmillan.

Scales, J.A., Smith, M.L., and Fischer, T.L., 1992, Global optimization methods for multimodal inverse problems, *Journal of Computational Physics*, **102**, 258-268.

Shannon, C.E., 1948, A mathematical theory of communication, Bell System Tech. J., 27, 379–423.

Su, W.-J., R.L. Woodward and A.M. Dziewonski, 1992, Deep origin of mid-oceanic ridge velocity anomalies, *Nature*, **360**, 149–152.

Tarantola, A., and Valette, B., 1982a, Inverse Problems = Quest for Information, J. Geophys., 50, 159-170.

Tarantola, A., and Valette, B., 1982b, Generalized nonlinear inverse problems solved using the least-squares criterion, Rev. Geophys. Space Phys., 20, No. 2, 219-232.

Tarantola, A., 1984, Inversion of seismic reflection data in the acoustic approximation, Geophysics, **49**, 1259–1266.

Tarantola, A., 1986, A strategy for nonlinear elastic inversion of seismic reflection data, Geophysics, **51**, 1893–1903.

Tarantola, A., 1987, Inverse problem theory; methods for data fitting and model parameter estimation, Elsevier.

Taylor, B.N., and C.E. Kuyatt, 1994, Guidelines for evaluating and expressing the uncertainty of NIST measurement results, NIST Technical note 1297.

Taylor A.E. and Lay D.C., 1980, Introduction to functional analysis, John Wiley and Sons, New York.

Tikhonov, A.N., 1963, Resolution of ill-posed problems and the regularization method (in russian), Dokl. Akad. Nauk SSSR, **151**, 501-504.

van der Hilst, R.D., S. Widiyantoro, and E.R. Engdahl, Evidence for deep mantle circulation from global tomography, *Nature*, **386**, 578–584, 1997.

Wiggins, R.A., Monte Carlo Inversion of Body-Wave Observations, *J. Geoph. Res.*, Vol. 74, No. 12, 3171–3181, 1969.

Wiggins, R.A., The General Linear Inverse Problem: Implication of Surface Waves and Free Oscillations for earth Structure, *Rev. Geoph. and Space Phys.*, Vol. 10, No. 1, 251–285, 1972.

Woodhouse, J.H., and F.A. Dahlen, The effect of general aspheric perturbation on the free oscillations of the Earth, *Geophys. J. R. astr. Soc.* **53**), 335–354 1978.

# Appendixes

## A Volumetric Probability and Probability Density

A probability distribution $\mathcal{A} \rightarrow P(\mathcal{A})$ over a manifold can be represented by a *volumetric probability* $F(\mathbf{x})$ , defined through

$$P(\mathcal{A}) \;=\; \int_{\mathcal{A}} dV(\mathbf{x})\; F(\mathbf{x}) \quad . \tag{120}$$

or by a *probability density* $f(\mathbf{x})$ , defined through

$$P(\mathcal{A}) \;=\; \int_{\mathcal{A}} d\mathbf{x}\; f(\mathbf{x}) \tag{121}$$

where $d\mathbf{x} = dx^1\, dx^2 \ldots$

While, under a change of variables, a probability density behaves as a density (i.e., its value at a point gets multiplied by the Jacobian of the transformation), a volumetric probability is a scalar (i.e., its value at a point remains invariant: it is defined indepenedently of any coordinate system).

Defining the *volume density* through

$$V(\mathcal{A}) \;=\; \int_{\mathcal{A}} d\mathbf{x}\; v(\mathbf{x}) \tag{122}$$

and considering the expression

$$V(\mathcal{A}) \;=\; \int_{\mathcal{A}} dV(\mathbf{x}) \quad , \tag{123}$$

we obtain

$$dV(\mathbf{x}) \;=\; v(\mathbf{x})\, d\mathbf{x} \quad . \tag{124}$$

It is then clear that the relation between volumetric probability and probability density is

$$f(\mathbf{x}) \;=\; v(\mathbf{x})\, F(\mathbf{x}) \quad . \tag{125}$$

While the homogeneous probability distribution (the one assigning equal probabilities to equal volumes of the space) is, in general, **not** represented by a constant probability density, it is always represented by a constant volumetric probability.

The authors of this paper favor, in their own work, the use of volumetric probabilites. For pedagogical reasons, we have chosen in this work to use probability densities.

## B Conditional and Marginal Probability Densities

### B.1 Conditional Probability Density

Let be $\mathcal{A}_p$ a set of the submanifold $\mathcal{M}_p$ and let be $\mathcal{A}_n(\delta)$ the set of points in $\mathcal{M}_n$ whose distance to the submanifold $\mathcal{M}_p$ is less or equal to $\delta$ , and whose normal projection on the submanifold $\mathcal{M}_p$ falls inside $\mathcal{A}_p$ . Given $\mathcal{A}_p$ and given $\delta$ , the set $\mathcal{A}_n(\delta)$ is uniquely defined (see figure 12). In particular, $\mathcal{M}_n(\delta)$ is the set of all points of $\mathcal{M}_n$ whose distance to the submanifold $\mathcal{M}_p$ is less or equal to $\delta$ (i.e., $\mathcal{M}_n(\delta)$ is the 'tube' or radius $\delta$ around $\mathcal{M}_p$ ). It is clear that for any value of $\delta$ ,

$$Q_\delta(\mathcal{A}_p) = \frac{P(\,\mathcal{A}_n(\delta)\,)}{P(\,\mathcal{M}_n(\delta)\,)} \tag{126}$$

defines a probability over the submanifold $\mathcal{M}_p$ . The *conditional probability* over the submanifold $\mathcal{M}_p$ is defined as the limit of $Q_\delta$ for $\delta \rightarrow 0$ :

$$Q(\mathcal{A}_p) = \lim_{\delta \rightarrow 0} \frac{P(\,\mathcal{A}_n(\delta)\,)}{P(\,\mathcal{M}_n(\delta)\,)} \quad . \tag{127}$$

The expression of the probability density (or the volumetric probability) associated to $Q$ may be complicated is the hypersurface is nor flat, if the metric is not Euclidean, or if the coordinates being used are not Cartesian.

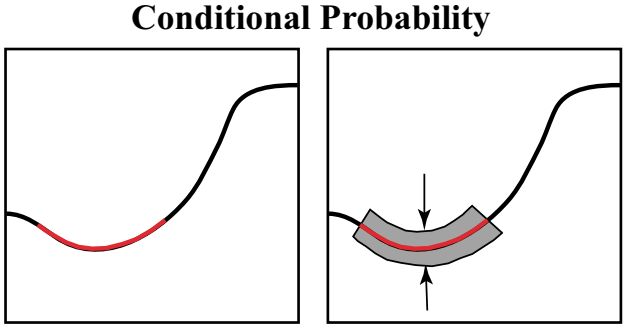



Figure 12: Definition of conditional volumetric probability (or conditional probability density).

Assume, for instance, the the hypersurface we consider is defined by the explicit equation

$$\mathbf{v} = \mathbf{v}(\mathbf{u}) \quad , \tag{128}$$

and let be

$$\mathbf{g} = \begin{pmatrix} \mathbf{g}_{uu} & \mathbf{g}_{uv} \\ \mathbf{g}_{vu} & \mathbf{g}_{vv} \end{pmatrix} \tag{129}$$

the metric of the space. Consider, then, a 'joint' volumetric probability $F(\mathbf{u}, \mathbf{v})$ , or, equivalently, a 'joint' probability density $f(\mathbf{u}, \mathbf{v})$ .

It can be shown, using the standard techniques of differential geometry, that the conditional volumetric probability over the submanifold $\mathbf{v} = \mathbf{v}(\mathbf{u})$ can be expressed as

$$F(\mathbf{u}|\mathbf{v} = \mathbf{v}(\mathbf{u})) = k\ F(\mathbf{u}, \mathbf{v}(\mathbf{u}))\ \left. \frac{\sqrt{\det(\mathbf{g}_{uu} + \mathbf{g}_{uv}\mathbf{V} + \mathbf{V}^T\mathbf{g}_{vu} + \mathbf{V}^T\mathbf{g}_{vv}\mathbf{V})}}{\sqrt{\det \mathbf{g}_{uu}}} \right|_{\mathbf{v}=\mathbf{v}(\mathbf{u})} \quad , \tag{130}$$

while the conditional probability density is

$$f(\mathbf{u}|\mathbf{v} = \mathbf{v}(\mathbf{u})) = k\ f(\mathbf{u}, \mathbf{v}(\mathbf{u}))\ \left. \frac{\sqrt{\det(\mathbf{g}_{uu} + \mathbf{g}_{uv}\mathbf{V} + \mathbf{V}^T\mathbf{g}_{vu} + \mathbf{V}^T\mathbf{g}_{vv}\mathbf{V})}}{\sqrt{\det \mathbf{g}}} \right|_{\mathbf{v}=\mathbf{v}(\mathbf{u})} \quad . \tag{131}$$

In these two equations, $\mathbf{V} = \mathbf{V}(\mathbf{u})$ is the matrix of partial derivatives

$$\begin{pmatrix} V_{11} & V_{12} & \cdots & V_{1p} \\ V_{21} & V_{22} & \cdots & V_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ V_{q1} & V_{q2} & \cdots & V_{qp} \end{pmatrix} = \begin{pmatrix} \frac{\partial v_1}{\partial u_1} & \frac{\partial v_1}{\partial u_2} & \cdots & \frac{\partial v_1}{\partial u_p} \\ \frac{\partial v_2}{\partial u_1} & \frac{\partial v_2}{\partial u_2} & \cdots & \frac{\partial v_2}{\partial u_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial v_q}{\partial u_1} & \frac{\partial v_q}{\partial u_2} & \cdots & \frac{\partial v_q}{\partial u_p} \end{pmatrix} \quad . \tag{132}$$

## B.2 Marginal Probability

The starting point here is the same as above, where the probability $Q_\delta$ is defined by equation 126, but now we take the limit when $\delta$ grows indefinitely (see figure 13).

Formally, we define the *marginal probability* over the submanifold $\mathcal{M}_p$ as the limit of $Q_\delta$ for $\delta \to \infty$ :

$$Q(\mathcal{A}_p) = \lim_{\delta \to \infty} \frac{P(\,\mathcal{A}_n(\delta)\,)}{P(\,\mathcal{M}_n(\delta)\,)} \quad . \tag{133}$$

It is clear that this limit will make sense only in special circumstances. Figure 14, for instance, suggests the case of a linear submanifold in an Euclidean space, and the case of a geodesic submanifold in an space with constant curvature. While in the Euclidean space, the limit $\delta \to \infty$ can actually be taken, in the spherical case, the limit is only taken up to the pole of the sphere (where all geodesics orthogonal to a geodesic submanifold meet).

If, in the Euclidean example, we use Cartesian coodinates, we have (e.g. in 2D)

$$f_x(x) = \int_{-\infty}^{+\infty} dy\ f(x, y) \quad , \tag{134}$$

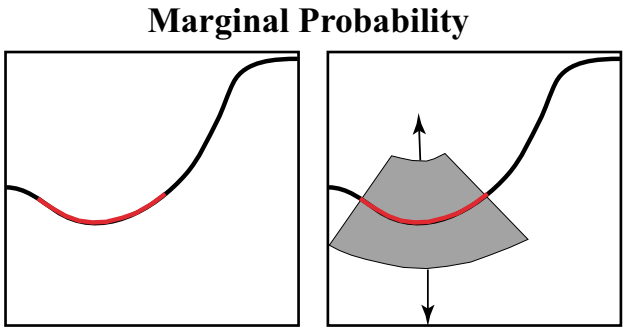



Figure 13: Definition of marginal volumetric probability (or marginal probability density.

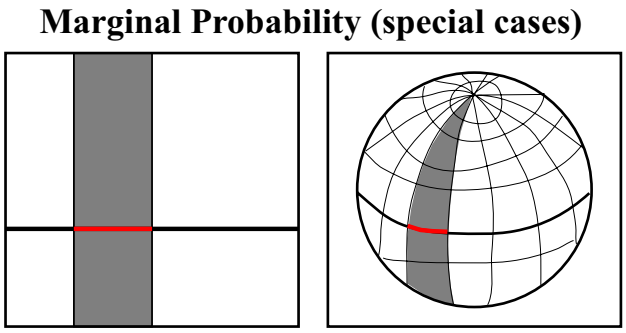



Figure 14: Two special cases where one can obtain an explicit expression for the marginal volumetric probability (or marginal probability density): in the Euclidean plane over a stright line or on the surface of the sphere, on a great circle.

equation valid for both, probability densitites and volumetric probabilities.

Over the 2D sphere, using colatitude $\lambda$ and longitude $\varphi$ , if we compute the marginal of a probability over the equator we obtain

$$f_\varphi(\varphi) = \int_{-\pi/2}^{+\pi/2} d\lambda \; f(\lambda,\varphi) \qquad \text{(probability densities)} \tag{135}$$

or, equivalently,

$$F_\varphi(\varphi) = \int_{-\pi/2}^{+\pi/2} dL(\lambda) \; F(\lambda,\varphi) \qquad \text{(volumetric probabilities)} \tag{136}$$

where

$$dL(\lambda) = \cos\lambda \; d\lambda \quad . \tag{137}$$

In the case where we build the total space using the Cartesian product of two spaces $\mathcal{U} \times \mathcal{V}$ , with

$$dV(\mathbf{u},\mathbf{v}) \;=\; dV_u(\mathbf{u})\, dV_v(\mathbf{v}) \quad , \tag{138}$$

then, from a joint volumetric probability $F(\mathbf{u},\mathbf{v})$ we can intrinsically define the two marginal volumetric probabilities

$$F_u(\mathbf{u}) = \int_v dV(\mathbf{v}) \; F(\mathbf{u},\mathbf{v}) \qquad ; \qquad F_v(\mathbf{v}) = \int_u dV(\mathbf{u}) \; F(\mathbf{u},\mathbf{v}) \tag{139}$$

or, equivalently, from a joint probability density $f(\mathbf{u},\mathbf{v})$ we can obtain the two marginal probability densities

$$f_u(\mathbf{u}) = \int_v d\mathbf{v} \; f(\mathbf{u},\mathbf{v}) \qquad ; \qquad f_v(\mathbf{v}) = \int_u d\mathbf{u} \; f(\mathbf{u},\mathbf{v}) \quad . \tag{140}$$

# C Combining Data and Theories: a Conceptual Example

In this section the same basic problem is solved in four different circumstances. As we shall see, each circumstance forces the use of a different approach.

The basic problem is the following. A particle follows a simple trajectory in space-time, the characteristics of the trajectory being not known a priori. An 'event' happens on the trajectory (like the particle emitting a light signal), and we use some experimental equipment to measure the space-time coordinates $(x,t)$ of the event. As there are always some experimental uncertainties, we may assume, with some generality, that the result of the

measurement can be described using a probability density $g(t,x)$ . In particular, we have some information on $t$ , as represented by the marginal probability density $g_t(t) = \int_{-\infty}^{+\infty} dx\ g(x,t)$ , and some information on $x$ , as represented by the marginal probability density $g_x(x) = \int_{-\infty}^{+\infty} dt\ g(x,t)$ . Now, the characteristics of the trajectory are given to us. This extra information can be used to ameliorate our knowledge on $x$ and $t$ , defining some new probability densities $h_t(t)$ and $h_x(x)$ that may have smaller uncertainties than the experimentally obtained $g_t(t)$ and $g_x(x)$ .

As an example, the trajectory of the particle is assumed to be a free-fall trajectory in a space with constant gravity field[29].

## C.1 Contemplative Approach (Conjunction of Probabilities)

Imagine that we have a "blinking mass in free-fall", i.e., a mass that is in free fall and which emits intermittent flashes. We do not know if the flashes are emitted at constant time intervals, or at constant position intervals, or with some other predetermined or random sequence.

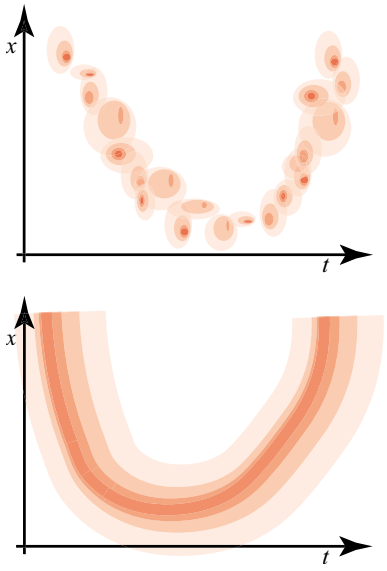



Figure 15: Measuring the position of blinking masses. Top: measurement (with uncertainties) of the blinks of a single mass. Bottom: the 'addition' of the measurements of a great number of similar blinking masses.

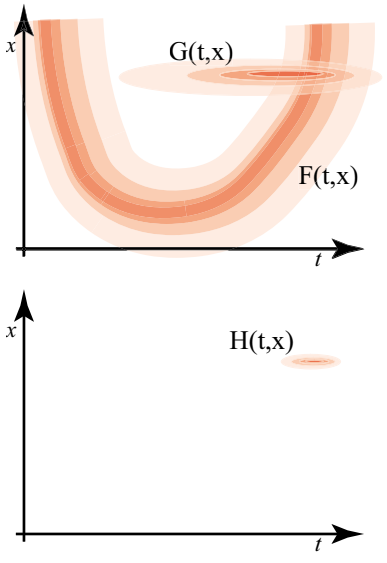



Figure 16: A 'measurement' $G(x,t)$ can be refined by combining it with the 'law' $F(x,t)$ . See equation 161.

[29] We shall assume Newtonian physics or special relativity, and shall not enter in the complications induced by the curved space-time of general relativity.

The top of figure 15 represents the measured 'trajectory' of a single particle. The bottom of the figure shows the 'free-fall law' of a blinking mass. It is built as follows: if each individual measurement of a blink produces the probability density $f_i(t,x)$ , and the OR operation produces, using all the measurements,

$$f(t,x) = k \sum_i f_i(t,x) \quad . \tag{141}$$

Once this 'law' is known, having used as many measurement as possible, we can consider a new blink of a new particle. A messurement of this particular blink gives the probability density $g(t,x)$ . As special cases, we may have, in fact, only have measured the position of the blink,

$$g(t,x) = g_x(x) \qquad \text{position measured only} \tag{142}$$

of the instant of the blink

$$g(t,x) = g_t(t) \qquad \text{instant measured only} \quad , \tag{143}$$

but let us continue with the general case $g(t,x)$ .

To 'combine' the 'law' $f(t,x)$ with the 'measurement' $g(t,x)$ , we use the AND operation (here, as the use Galilean coordinates, the homogeneous probability density is a constant)

$$h(t,x) = k\ f(t,x)\ g(t,x) \quad . \tag{144}$$

See figure 16 for a graphical illustration.

## C.2 "Ideal Theory" (Conditional Probability Density)

In order to have a natural metric in the space-time, assume that we are in the context of the (special) relativistic space-time, where the distance element is assumed to be

$$ds^2 = dt^2 - \frac{1}{c^2}\, dx^2 \quad , \tag{145}$$

this meaning that the metric tensor of the working space is

$$\mathbf{g} = \begin{pmatrix} g_{tt} & g_{tx} \\ g_{xt} & g_{xx} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1/c^2 \end{pmatrix} \quad . \tag{146}$$

In special relativity, a particle of mass $m$ submitted to a force $f$ satisfies the dynamic equation

$$f = \frac{d}{dt} \frac{m\, v}{\sqrt{1-(v/c)^2}} \quad , \tag{147}$$

where $v$ is the particle's velocity. If the force is constant, this integrates (assuming $v(0) = 0$ ) into

$$v(t) = \frac{a\, t}{\sqrt{1+(at/c)^2}} \quad , \tag{148}$$

where the constant $a$ , having the dimensions of an acceleration, is the ratio

$$a = \frac{f}{m} \quad . \tag{149}$$

In turn, the expression 148 integrates into the expression for the trajectory:

$$x(t) = \frac{c^2}{a} \left( \sqrt{1+(a\, t/c)^2} - 1 \right) \quad . \tag{150}$$

The developments for the velocity and the position for $t \to 0$ and for $t \to \infty$ are given as a footnote[30].

[30] For $t \to 0$ we obtain $v(t) = a\, t \left(1 - \frac{1}{2} \left(\frac{a\, t}{c}\right)^2 + \ldots\right)$ and $x(t) = \frac{1}{2}\, a\, t^2 \left(1 - \frac{1}{4} \left(\frac{a\, t}{c}\right)^2 + \ldots\right)$ , while for $t \to \infty$ we obtain $v(t) = c \left(1 - \frac{1}{2} \left(\frac{c}{a\, t}\right)^2 + \ldots\right)$ and $x(t) = c \left(t - \frac{c}{a} + \ldots\right)$ .

Assume that a measurement of the coordinates $(t, x)$ of an event (like a blinking of the particle), produces the probability density $h(t, x)$ . If we are told that the particle is necessarily on the trajectory given by equation 150, and we are not given any other information, we can modify $h(t, x)$ by the condition $x = x(t)$ , i.e., we can define the conditional probability density $h(t|x = x(t))$ . To do this consistently, we have to use some distance over our working space (here, the space-time), in order to define the conditional probability density, from the more general notion of conditional probability, using a notion of uniform convergence.

This is what we have done above. Adapting to this special case equation 17 gives

$$h_t(t) = h(t|x = x(t)) = k\ h(t, x(t))\ \left.\frac{\sqrt{g_{tt} + g_{xx}\,\dot{x}^2}}{\sqrt{g_{tt}}}\right|_{x=x(t)} \quad , \tag{151}$$

i.e. (redefining the constant),

$$h_t(t) = k\ h(t, x(t))\ \sqrt{g_{tt} + g_{xx} v(t)^2} \quad . \tag{152}$$

Using the results above this gives

$$h_t(t) = k\ h(t, x(t))\ \left(1 - \frac{a^2\,t^2}{a^2\,t^2 + c^2}\right) \quad , \tag{153}$$

and this solves our problem. While for small $t$ this gives

$$h_t(t) = k\ h(t, x(t)) \quad , \tag{154}$$

for large $t$ we obtain, instead,

$$h_t(t) = k\ \frac{1}{t^2}\ h(t, x(t)) \quad . \tag{155}$$

Reversing the use we have made of the variables $t$ and $x$ , we could have calculated $h_x(x)$ , rather than $h_t(t)$ . But the invariance property mentioned in section 2.4 warrants us that these two probability densities are related through the Jacobian rule, i.e., we shall have

$$h_t(t) = \frac{dx}{dt}\,h_x(x) = \dot{x}(t)\,h_x(x) \ . \tag{156}$$

## C.3 Uncertain Analytical Theory (Conjunction of Probabilities)

We prepare particles that have to follow a free fall. It is assumed that the trajectory of the particles is, approximately,

$$x(t) \approx x_0 + v_0\,t + \frac{1}{2}\,g\,t^2 \quad , \tag{157}$$

with some uncertainties, principally due to uncertainties in the values $x_0$ and $v_0$ . These uncertainties are assumed to dominate all other sources of uncertainty (air friction, variations in the gravity field, etc.). The value of $g$ is assumed to be known with high accuracy.

Of course, equation 157 can, equivalently, be written

$$t(x) \approx \pm\frac{1}{g}\sqrt{v_0^2 + 2g(x - x_0)} - \frac{v_0}{g} \quad , \tag{158}$$

The probability density for $x_0$ is $Q(x_0)$ , and that of $v_0$ is $R(v_0)$ .

**Case** $x = x(t)$

The particles are prepared so that they desintegrate (or that they blink) at some time instant $t$ chosen homogeneously at random inside some (large) time interval.

Using equation 157 it is possible to compute $f(x|t)$ , the conditional probability density for the particle to be at $x$ when it desintegrates at $t$ . For instance, if the probability densityfor $v(t_0)$ is a Gaussian centered at $v_0$ with standard deviation $\sigma_v$ and if the probability density for $x(t_0)$ is another Gaussian centered at $x_0$ with standard deviation $\sigma_x$ , we get

$$f(x|t) = k\ \frac{1}{\sqrt{\sigma_x^2 + \sigma_v^2\,t^2}}\ \exp\left(-\frac{1}{2}\,\frac{\left(x - (x_0 + v_0\,t + \frac{1}{2}\,g\,t^2)\right)^2}{\sigma_x^2 + \sigma_v^2\,t^2}\right) \tag{159}$$

As the particle is prepared to desintegrate at some instant $t$ chosen homogeneously at random, the joint probability density is

$$f(t,x) = k\ f(x|t) \quad . \tag{160}$$

This probability density is represented in figure 17, together with the two marginals, and the conditional probability density at three different times is represented in figure 18.

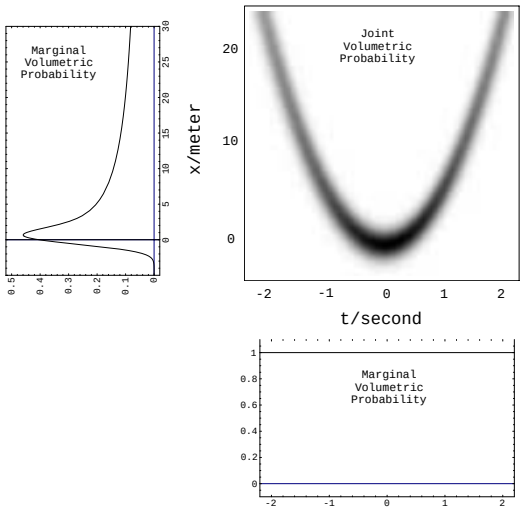



Figure 17: A typical parabola representing the free fall of an object (position $x$ as a function of time $t$ ). Here, rather than an infinitely thin line we have a fuzzy object (a probability distribution) because the initial position and initial velocity is uncertain. This figure represents the probability density defined by equations 159–160 , with $x_0 = 0$ , $v_0 = 1\,\mathrm{m/s}$ , $\sigma_x = 1\,\mathrm{m}$ , $\sigma_v = 1\,\mathrm{m/s}$ and $g = 9.91\,\mathrm{m/s^2}$ . While, by definition, the marginal of the probability density with respect to the time $t$ is homogeneous, the marginal for the position $x$ is not: there is a pronounced maximum for $x = 0$ (when the falling object is slower), and the distribution is very asymmetric (as the object is falling 'downwards').

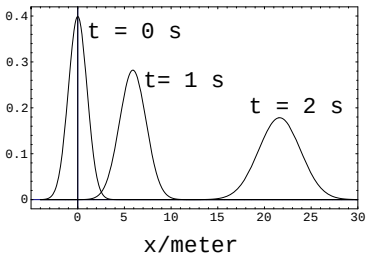



Figure 18: Three conditional probability densities from the joint distribution of the previous figure at times $t = 0$ , $t = 1\,\mathrm{s}$ and $t = 2\,\mathrm{s}$ . The width increases with time because of the uncertainty in the initial velocity.

Consider a new blink of a new particle. A measurement of this particular blink gives the probability density $g(t,x)$ . To 'combine' the 'law' $f(t,x)$ with the 'measurement' $g(t,x)$ , means to use the AND operation (note that, here, as we use Galilean coordinates, the homogeneous probability density is a constant):

$$h(t,x) = k\ f(t,x)\ g(t,x) \quad . \tag{161}$$

Let us evaluate the marginal

$$h_t(t) = \int dx\ h(t,x) = k\ \int dx\ f(t,x)\ g(t,x) \quad . \tag{162}$$

For sufficiently small $\sigma_x$ and $\sigma_v$ , we have the approximation

$$h_t(t) \approx k\ \frac{1}{\sqrt{\sigma_x^2 + \sigma_v^2\, t^2}}\ g(\ t\ ,\ x(t)\ ) \quad , \tag{163}$$

where $x(t)$ is defined as

$$x(t) = x_0 + v_0\, t + \frac{1}{2}\, g\, t^2 \quad . \tag{164}$$

We can take two limits here. If $\sigma_x \to 0$ , then, for whatever value of $\sigma_v$ (but that was assumed above to be small), we have, redefining the constant $k$ ,

$$h_t(t) = k\ \frac{1}{t}\ g(\ t\ ,\ x(t)\ ) \quad . \tag{165}$$

If, on the contrary, it is $\sigma_v$ which tends to zero, then,

$$h_t(t) = k\ g(\ t\ ,\ x(t)\ )\quad . \tag{166}$$

These equations are to be compared with equations 153–155. We see that the result we obtain crucially depends on the 'preparation' of the particle. To combine the result of a measurement with a theory is not a simple matter.

**Case** $t = t(x)$

We leave as an exercise to the reader the solution of the same problem, but where the particles are prepared so that they desintegrate (or that they blink) at some position $x$ chosen homogeneously at random inside some (large) space interval.

# D Information Content

Shannon's definition of information content (Shannon, 1948) of a discrete probability $I = \sum_i p_i \log p_i$ does not generalize into a definition of the information content of a probability density (the 'definition' $I = \int d\mathbf{x}\ f(\mathbf{x})\ \log f(\mathbf{x})$ is not invariant under a change of variables). Rather, one may define the 'Kullback distance' (Kullback, 1967) from the probability density $g(\mathbf{x})$ to the probability density $f(\mathbf{x})$ as

$$I(f|g) = \int d\mathbf{x}\ f(\mathbf{x}) \log \frac{f(\mathbf{x})}{g(\mathbf{x})}\ . \tag{167}$$

This means in particular that we never know if a single probability density is, by itself, informative or not. The equation above defines the information gain when we pass from $g(\mathbf{x})$ to $f(\mathbf{x})$ ( $I$ is always positive). But there is also an information gain when we pass from $f(\mathbf{x})$ to $g(\mathbf{x})$ : $I(g|f) = \int d\mathbf{x}\ g(\mathbf{x}) \log\ g(\mathbf{x})/f(\mathbf{x})$ . One should note that (i) the 'Kullback distance' is not a distance (the distance from $f(\mathbf{x})$ to $g(\mathbf{x})$ does not equal the distance from $g(\mathbf{x})$ to $f(\mathbf{x})$ ); (ii) for the 'Kullback distance' $I(f|g) = \int d\mathbf{x}\ f(\mathbf{x}) \log\ f(\mathbf{x})/g(\mathbf{x})$ to be defined, the probability density $f(\mathbf{x})$ has to be 'absolutely continuous' with respect to $g(\mathbf{x})$ , which amounts to say that $f(\mathbf{x})$ can only be zero where $g(\mathbf{x})$ is zero. We have postulated that any probability density $f(\mathbf{x})$ is absolutely continuous with respect to the homogeneous probability distribution $\mu(\mathbf{x})$ , since the homogeneous probability distribution 'fills the space'. Then one may use the convention that the information content of any probability density $f(\mathbf{x})$ is measured with respect to the homogeneous probability density:

$$I(f) \equiv f(f|\mu) = \int d\mathbf{x}\ f(\mathbf{x}) \log \frac{f(\mathbf{x})}{\mu(\mathbf{x})}\ . \tag{168}$$

The homogeneous probability density is then 'noninformative', $I(\mu) = I(\mu|\mu) = 0$ , but this is just by definition.

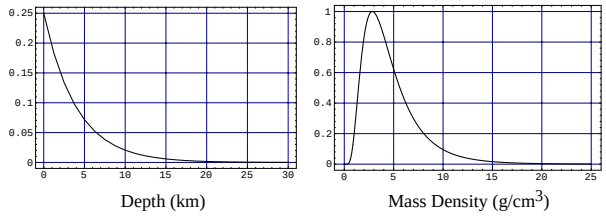
# E Example: Prior Information for a 1D Mass Density Model

Of course, the simplest example of a probability distribution is the Gaussian (or 'normal') distribution. Not many physical parameters accept the Gaussian as a probabilistic model (we have, in particular, seen that many positive parameters are Jeffreys parameters, for which the simplest consistent probability density is not the normal, but the *log-normal* probability density [see rule 8]), and we shall therefore consider the problem of describing a model consisting of a stack of horizontal layers with variable thickness and uniform mass density. The prior information is shown in figure 19, involving marginal distributions of the mass density and the layer thickness. Spatial statistical homogeneity is assumed, hence marginals are not dependent on depth in this example. Additionally, they are independent of neighbor layer parameters.

The model parameters consist of a sequence of thicknesses and a sequence of mass density parameters, $\mathbf{m} = \{\ell_1, \ell_2, \ldots, \ell_{NL}, \rho_1, \rho_2, \ldots, \rho_{NL}\}$ . The marginal prior probability densities for the layer thicknesses are all assumed to be identical and of the form (exponential probability density)

$$f(\ell)\ =\ \frac{1}{\ell_0}\ \exp\left(-\frac{\ell}{\ell_0}\right)\quad , \tag{169}$$

Figure 19: At left, the probability density for the layer thickness. At right, the probability density for the density of mass.




where the constant $\ell_0$ has the value $\ell_0 = 4$ km (see the left of figure 19), while all the marginal prior probability densities for the mass density are also assumed to be identical and of the form (log-normal probability density)

$$g(\rho) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\,\frac{1}{\rho}\,\exp\left(-\frac{1}{2\,\sigma^2}\left(\log\frac{\rho}{\rho_0}\right)^2\right) \quad , \tag{170}$$

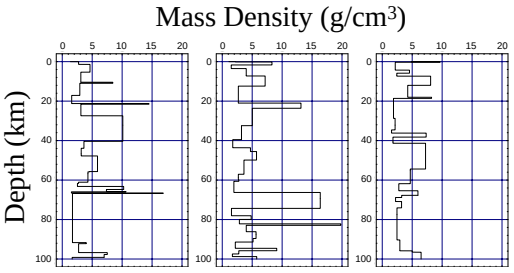
where $\rho_0 = 3.98$ g/cm$^3$ and $\sigma = 0.58$ (see the right of figure 19).

Assuming that the probability distribution of any layer thickness is independent of the thicknesses of the other layers, that the probability distribution of any mass density is independent of the mass densities of the other layers, and that layer thicknesses are independent of mass densities, the prior probability density in this problem is the product of prior probability densities (equations 169 and 170) for each parameter,

$$\rho_{\mathcal{M}}(\mathbf{m}) \;=\; \rho_{\mathcal{M}}(\ell_1, \ell_2, \ldots, \ell_{NL}, \rho_1, \rho_2, \ldots, \rho_{NL}) = k \prod_i^{NL} f(\rho_i)\, g(\rho_i) \quad . \tag{171}$$

Figure 20 shows (pseudo-) random models generated according to this probability distribution. Of course, the explicit expression 171 has not been used to generate these random models. Rather, consecutive layer thicknesses and consecutive mass densities have been generated using the univariate probability densities defined by equations 169 and 170.

Figure 20: Three random Earth models generated according to the prior probability density in the model space.




# F Gaussian Linear Problems

If the 'relation solving the forward problem' $\mathbf{d} = \mathbf{f}(\mathbf{m})$ happens to be a linear relation,

$$\mathbf{d} \;=\; \mathbf{F}\,\mathbf{m} \quad , \tag{172}$$

then the probability density $\sigma_m(\mathbf{m})$ in equation 50 becomes[31]

$$\sigma_m(\mathbf{m}) \;= \tag{173}$$

$$k\,\exp\left(-\frac{1}{2}\left((\mathbf{m}-\mathbf{m}_{\text{prior}})^t\,\mathbf{C}_M^{-1}\,(\mathbf{m}-\mathbf{m}_{\text{prior}}) + (\mathbf{F}\,\mathbf{m}-\mathbf{d}_{\text{obs}})^t\,\mathbf{C}_D^{-1}\,(\mathbf{F}\,\mathbf{m}-\mathbf{d}_{\text{obs}})\right)\right) \quad .$$

As the argument of the exponential is a quadratic function of $\mathbf{m}$ we can write it in standard form,

$$\sigma(\mathbf{m}) \;=\; k\;\exp\left(-\frac{1}{2}\left((\mathbf{m}-\widetilde{\mathbf{m}})^T\widetilde{\mathbf{C}}_M^{-1}(\mathbf{m}-\widetilde{\mathbf{m}})\right)\right) \quad , \tag{174}$$

[31] The last multiplicative factor in equation 50 is a constant that can be integrated into the constant $k$ .

This implies that $\sigma_m(\mathbf{m})$ is a Gaussian probability density. The values $\widetilde{\mathbf{m}}$ and $\widetilde{\mathbf{C}}_M$ of the center and covariance matrix, respectively, of the Gaussian representing the posterior information in the model space, can be computed using certain matrix identities (see, for instance, Tarantola, 1987, problem 1.19). This gives

$$\begin{aligned}
\widetilde{\mathbf{m}} &= \left(\mathbf{F}^T\mathbf{C}_D^{-1}\mathbf{F} + \mathbf{C}_M^{-1}\right)^{-1}\left(\mathbf{F}^T\mathbf{C}_D^{-1}\mathbf{d}_{\text{obs}} + \mathbf{C}_M^{-1}\mathbf{m}_{\text{prior}}\right) && (175)\\
&= \mathbf{m}_{\text{prior}} + \left(\mathbf{F}^T\mathbf{C}_D^{-1}\mathbf{F} + \mathbf{C}_M^{-1}\right)^{-1}\mathbf{F}^T\mathbf{C}_D^{-1}\left(\mathbf{d}_{\text{obs}} - \mathbf{F}\mathbf{m}_{\text{prior}}\right) && (176)\\
&= \mathbf{m}_{\text{prior}} + \mathbf{C}_M\mathbf{F}^T\left(\mathbf{F}\mathbf{C}_M\mathbf{F}^T + \mathbf{C}_D\right)^{-1}\left(\mathbf{d}_{\text{obs}} - \mathbf{F}\mathbf{m}_{\text{prior}}\right) && (177)
\end{aligned}$$

and

$$\begin{aligned}
\widetilde{\mathbf{C}}_M &= \left(\mathbf{F}^T\mathbf{C}_D^{-1}\mathbf{F} + \mathbf{C}_M^{-1}\right)^{-1} && (178)\\
&= \mathbf{C}_M - \mathbf{C}_M\mathbf{F}^T\left(\mathbf{F}\mathbf{C}_M\mathbf{F}^T + \mathbf{C}_D\right)^{-1}\mathbf{F}\mathbf{C}_M \quad . && (179)
\end{aligned}$$

If we do not have any prior information on the model parameters, then $\mathbf{C}_M \to \infty\,\mathbf{I}$ , i.e.,

$$\mathbf{C}_M^{-1} \to 0\,. \tag{180}$$

Formulas 175 and 178 then become

$$\widetilde{\mathbf{m}} = \left(\mathbf{F}^T\mathbf{C}_D^{-1}\mathbf{F}\right)^{-1}\left(\mathbf{F}^T\mathbf{C}_D^{-1}\mathbf{d}_{\text{obs}}\right) \tag{181}$$

and

$$\widetilde{\mathbf{C}}_M = \left(\mathbf{F}^T\mathbf{C}_D^{-1}\mathbf{F}\right)^{-1} \quad . \tag{182}$$

In the very special circumstance where we have the same number of 'data parameters' and 'model parameters', i.e., the case where the matrix $\mathbf{F}$ is a square matrix. Assume that the matrix is regular so its inverse exists. It is easy to see that equation 181 then becomes

$$\widetilde{\mathbf{m}} = \mathbf{F}^{-1}\,\mathbf{d}_{\text{obs}} \quad . \tag{183}$$

We see that in this special case $\widetilde{\mathbf{m}}$ is just the Cramer solution of the linear equation $\mathbf{d}_{\text{obs}} = \mathbf{F}\,\widetilde{\mathbf{m}}$ .

# G The Structure of an Inference Space

*Note: This appendix is a reproduction of a section of the e-paper* `arXiv:math-ph/0009029` *that can be found at http://arXiv.org/abs/math-ph/0009029.*

Before Kolmogorov, probability calculus was made using the intuitive notions of "chance" or "hazard". Kolmogorov's axioms clarified the underlying mathematical structure and brought probability calculus inside well defined mathematics. In this section we will recall these axioms. Our opinion is that the use in physical theories (where we have invariance requirements) of probability distributions, through the notions of conditional probability or the so-called Bayesian paradigm suffers today from the same defects as probability calculus suffered from before Kolmogorov. To remedy this, we introduce in this section, in the space of all probability distributions, two logical operations (OR and AND) that give the necessary mathematical structure to the space.

## G.1 Kolmogorov's Concept of Probability

A point $\mathbf{x}$ , that can materialize itself anywhere inside a domain $\mathcal{D}$ , may be realized, for instance, inside $\mathcal{A}$ , a subdomain of $\mathcal{D}$ . The probability of realization of the point is completely described if we have introduced a *probability distribution* (in Kolmogorov's sense) on $\mathcal{D}$ , i.e., if to every subdomain $\mathcal{A}$ of $\mathcal{D}$ we are able to associate a real number $P(\mathcal{A})$ , called *the probability of* $\mathcal{A}$ , having the three properties:

- For any subdomain $\mathcal{A}$ of $\mathcal{D}$ , $P(\mathcal{A}) \geq 0$ .
- If $\mathcal{A}_i$ and $\mathcal{A}_j$ are two disjoint subsets of $\mathcal{D}$ , then, $P(\mathcal{A}_i \cup \mathcal{A}_j) = P(\mathcal{A}_i) + P(\mathcal{A}_j)$ .
- For a sequence of events $\mathcal{A}_1 \supseteq \mathcal{A}_2 \supseteq \cdots$ tending to the empty set, we have $P(\mathcal{A}_i) \to 0$ .

We will not necessarily assume that a probability distribution is normed to unity ( $P(\mathcal{D}) = 1$ ). Although one refers to this as a *measure*, instead of a probability, we will not use this distinction. Sometimes, our probability

distributions will not be normalizable at all ( $P(\mathcal{D}) = \infty$ ). We can only then compute the *relative probabilities* of subdomains.

These axioms apply to probability distributions over discrete or continuous spaces. Below, we will consider probability distributions over spaces of physical parameters, that are continuous spaces. Then, a probability distribution is represented by a probability density.

In the next section, given a space $\mathcal{D}$ , we will consider different probability distributions $P, Q \ldots$ Each probability distribution will represent a particular *state of information* over $\mathcal{D}$ . In what follows, we will use as synonymous the terms "probability distribution" and "state of information".

## G.2 Inference Space

We will now give a structure to the space of all the probability distributions over a given space, by introducing two operations, the OR and the AND operation. This contrasts with the basic operations introduced in deductive logic, where the negation ("NOT"), nonexistent here, plays a central role. In what follows, the OR and the AND operation will be denoted, symbolically, by $\vee$ and $\wedge$ . They are assumed to satisfy the set of axioms here below.

The first axiom states that if an event $\mathcal{A}$ is possible for $(P \text{ OR } Q)$ , then the event is either possible for $P$ or possible for $Q$ (which his is consistent with the usual logical sense for the "or"): *For any subset* $\mathcal{A}$ *, and for any two probability distributions* $P$ *and* $Q$ *, the* OR *operation satisfies*

$$(P \vee Q)(\mathcal{A}) \neq 0 \qquad \Longrightarrow \qquad P(\mathcal{A}) \neq 0 \quad \text{or} \quad Q(\mathcal{A}) \neq 0 \ ,$$

*the word* "or" *having here its ordinary logical sense.*

The second axiom states that if an event $\mathcal{A}$ is possible for $(P \text{ AND } Q)$ , then the event is possible for both $P$ and $Q$ (which is consistent with the usual logical sense for the "and"): *For any subset* $\mathcal{A}$ *, and for any two probability distributions* $P$ *and* $Q$ *, the* AND *operation satisfies*

$$(P \wedge Q)(\mathcal{A}) \neq 0 \qquad \Longrightarrow \qquad P(\mathcal{A}) \neq 0 \quad \text{and} \quad Q(\mathcal{A}) \neq 0 \ ,$$

*the word* "and" *having here its ordinary logical sense.*

The third axiom ensures the existence of a neutral element, that will be interpreted below as the probability distribution carrying no information at all: *There is a* neutral element*,* $M$ *for the* AND *operation, i.e., it exists a* $M$ *such that for any probability distribution* $P$ *and for any subset* $\mathcal{A}$ *,*

$$(M \wedge P)(\mathcal{A}) = (P \wedge M)(\mathcal{A}) = P(\mathcal{A}) \ .$$

The fourth axiom imposes that *the* OR *and the* AND *operations are* commutative *and* associative, and, by analogy with the algebra of propositions of ordinary logic, have a distributivity property: *the* AND *operation is* distributive *with respect to the* OR *operation.*

The structure obtained when furnishing the space of all probability distributions (over a given space $\mathcal{D}$ ) with two operations OR and AND, satisfying the given axioms constitutes what we propose to call an *inference space*.

These axioms do not define uniquely the operations. Let $\mu(\mathbf{x})$ be the particular probability density representing $M$ , the neutral element for the AND operation, and let $p(\mathbf{x}), q(\mathbf{x}) \ldots$ be the probability densities representing the probability distributions $P, Q \ldots$ Using the notations $(p \vee q)(\mathbf{x})$ and $(p \wedge q)(\mathbf{x})$ for the probability densities representing the probability distributions $P \vee Q$ and $P \wedge Q$ respectively, one realization of the axioms (the one we will retain) is given by

$$(p \vee q)(\mathbf{x}) = p(\mathbf{x}) + q(\mathbf{x}) \qquad ; \qquad (p \wedge q)(\mathbf{x}) = \frac{p(\mathbf{x})\, q(\mathbf{x})}{\mu(\mathbf{x})} \ , \tag{184}$$

where one should remember that we do not impose to our probability distributions to be normalized.

The structure of an inference space, as defined, contains other useful solutions. For instance, the theory of fuzzy sets (Kandel, 1986) uses positive functions $p(\mathbf{x}), q(\mathbf{x}) \ldots$ quite similar to probability densities, but having a different interpretation: the are normed by the condition that their maximum value equals one, and are interpreted as the "grades of membership" of a point $\mathbf{x}$ to the "fuzzy sets" $P, Q \ldots$ . The operations OR and AND correspond then respectively to the *union* and *intersection* of fuzzy sets, and to the following realization of our axioms:

$$(p \vee q)(\mathbf{x}) = \max(p(\mathbf{x}), q(\mathbf{x})) \qquad ; \qquad (p \wedge q)(\mathbf{x}) = \min(p(\mathbf{x}), q(\mathbf{x})) \ , \tag{185}$$

where the neutral element for the AND operation (intersection of fuzzy sets) is simply the function $\mu(\mathbf{x}) = 1$ .

While fuzzy set theory is an alternative to classical probability (and is aimed at the solution of a different class of problems), our aim here is only to complete the classical probability theory. As explained below the solution given by equations 184 correspond to the natural generalisation of two fundamental operations in classical probability theory: that of "making histograms" and that of taking "conditional probabilities". To simplify our language, we will sometimes use this correspondence between our theory and the fuzzy set theory, and will say that the OR operation, when applied to two probability distributions, corresponds to the *union* of the two states of information, while the AND operation corresponds to their *intersection*.

It is easy to write some extra conditions that distinguish the two solutions given by equations 184 and 185. For instance, as probability densities are normed using a multiplicative constant (this is not the case with the grades of membership in fuzzy set theory), it makes sense to impose the simplest possible algebra for the multiplication of probability densities $p(\mathbf{x}), q(\mathbf{x}) \ldots$ by constants $\lambda, \mu \ldots$ :

$$[(\lambda + \mu)p]\,(\mathbf{x}) = (\lambda\, p \vee \mu p)\,(\mathbf{x}) \qquad ; \qquad [\lambda(p \wedge q)]\,(\mathbf{x}) = (\lambda p \wedge q)\,(\mathbf{x}) = (p \wedge \lambda q)\,(\mathbf{x})\ . \tag{186}$$

This is different from finding a (minimal) set of axioms characterizing (uniquely) the proposed solution, which is an open problem.

One important property of the two operations OR and AND just introduced is that of *invariance* with respect to a change of variables. As we consider probability distribution over a continuous space, and as our definitions are independent of any choice of coordinates over the space, it must happen that we obtain equivalent results in any coordinate system. Changing for instance from the coordinates $\mathbf{x}$ to some other coordinates $\mathbf{y}$ , will change a probability density $p(\mathbf{x})$ to $\widetilde{p}(\mathbf{y}) = p(\mathbf{x})\,|\partial\mathbf{x}/\partial\mathbf{y}|$ . It can easily be seen that performing the OR or the AND operation, then changing variables, gives the same result than first changing variables, then, performing the OR or the AND operation.

Let us mention that the equivalent of equations 184 for discrete probability distributions is:

$$(p \vee q)_i = p_i + q_i \qquad ; \qquad (p \wedge q)_i = \frac{p_i\, q_i}{\mu_i}\ . \tag{187}$$

Although the OR and AND notions just introduced are consistent with classical logic, they are here more general, as they can handle states of information that are more subtle than just the "possible" or "impossible" ones.

## G.3 The Interpretation of the OR and the AND Operation

If an experimenter faces realizations of a random process and wants to investigate the probability distribution governing the process, he may start making histograms of the realizations. For instance, for realizations of a probability distribution over a continuous space, he will obtain histograms that, in some sense, will approach the probability density corresponding to the probability distribution.

A histogram is typically made by dividing the working space into cells, and by counting how many realizations fall inside each cell. A more subtle approach is possible. First, we have to understand that, in the physical sciences, when we say "a random point has materialized in an abstract space", we may mean something like "this object, one among many that may exist, vibrates with some fixed period; let us measure as accurately as possible its period of oscillation". Any physical measure of a real quantity will have attached uncertainties. This means that when, mathematically speaking, we measure "the coordinates of a point in an abstract space" we will not obtain a point, but a state of information over the space, i.e., a probability distribution.

If we have measured the coordinates of many points, the results of each measurement will be described by a probability density $p_i(\mathbf{x})$ . The union of all these, i.e., the probability density

$$(p_1 \vee p_2 \vee \ldots)\,(\mathbf{x}) = \sum_i p_i(\mathbf{x}) \tag{188}$$

is a finer estimation of the background probability density than an ordinary histogram, as actual measurement uncertainties are used, irrespective of any division of the space into cells. If it happens that the measurement uncertainties can be described using box-car functions at fixed positions, then, the approach we propose reduces to the conventional making of histograms.

Figure 1 explains that our definition of the AND operation is a generalization of the notion of conditional probability. A probability distribution $P(\,\cdot\,)$ is represented, in the figure, by its probability density. To any region $\mathcal{A}$ of the plane, it associates the probability $P(\mathcal{A})$ . If a point has been realized following the probability distribution $P(\,\cdot\,)$ and we are given the information that, in fact, the point is "somewhere" inside the region $\mathcal{B}$ , then we can

update the prior probability $P(\,\cdot\,)$, replacing it by the conditional probability $P(\,\cdot\,|\mathcal{B}) = P(\,\cdot\,\cap\mathcal{B})/P(\mathcal{B})$. It equals $P(\,\cdot\,)$ inside $\mathcal{B}$ and is zero outside (center of the figure). If instead of the hard constraint $x \in \mathcal{B}$ we have a soft information about the location of $x$, represented by the probability distribution $Q(\,\cdot\,)$ (right of the figure), the intersection of the two states of information $P$ and $Q$ gives a new state of information (here, $\mu(x)$ is the probability density representing the state of null information, and, to simplify the figure, has been assumed to be constant). The comparison of the right with the center of the figure shows that the AND operation generalizes the notion of conditional probability. In the special case where the probability density representing the second state of information, $Q(\,\cdot\,)$, equals the null information probability density inside the domain $\mathcal{B}$ and is zero outside, then, the notion of intersection of states of information exactly reduces to the notion of conditional probability.

Now the interpretation of the neutral element for the AND operation can be made clear. We postulated that the neutral probability distribution $M$ is such that for any probability distribution $P$, $P \wedge M = P$. This means that if a point is realized according to a probability distribution $P$, and if a (finite accuracy) measure of the coordinates of the point produces the information represented by $M$, the posterior probability distribution, $P \wedge M$ is still $P$: the probability distribution $M$ is not carrying any information at all. Accordingly, we call $M$ the *null information probability distribution*. Sometimes, the probability density representing this state of null information is constant over all the space; sometimes, it is not, as explained in section 2.2. It is worth mentioning that this particular state of information enters in the Shannon's definition of Information Content (Shannon, 1948).

It is unfortunate that, when dealing with probability distributions over continous spaces, conditional probabilities are often misused. Section P describes the so-called Borel-Kolmogorov paradox: using conditional probability densities in a space with coordinates $(x, y)$ will give results that will not be consistent with those obtained by the use of conditional probability densities on the same space but where other coordinates $(u, v)$ are used (if the change of coordinates is nonlinear). Jaynes (1995) gives an excellent, explicit, account of the paradox. But his choice for resolving the paradox is different from our's: while Jaynes just insists on the technical details of how some limits have to be taken in order to ensure consistency, we radically decide to abandon the notion of conditional probability, and replace it by the intersection of states of information (the AND operation) which is naturally consistent under a change of variables.

# H Homogeneous Probability for Elastic Parameters

In this appendix, we start from the assumption that the uncompressibility modulus and the shear modulus are Jeffreys parameters (they are the eigenvalues of the stiffness tensor $c_{ijk\ell}$), and find the expression of the homogeneous probability density for other sets of elastic parameters, like the set { `Young's modulus - Poisson ratio` } or the set { `Longitudinal wave velocity - Tranverse wave velocity` }.

## H.1 Uncompressibility Modulus and Shear Modulus

The 'Cartesian parameters' of elastic theory are the logarithm of the uncompressibility modulus and the logarithm of the shear modulus

$$\kappa^* \;=\; \log\frac{\kappa}{\kappa_0} \qquad ; \qquad \mu^* \;=\; \log\frac{\mu}{\mu_0} \quad , \tag{189}$$

where $\kappa_0$ and $\mu_0$ are two arbitrary constants. The homogeneous probability density is just constant for these parameters (a constant that we set arbitrarily to one)

$$f_{\kappa^*\mu^*}(\kappa^*, \mu^*) \;=\; 1 \quad . \tag{190}$$

As is often the case for homogeneous 'probability' densities, $f_{\kappa^*\mu^*}(\kappa^*, \mu^*)$ is not normalizable. Using the jacobian rule, it is easy to transform this probability density into the equivalent one for the positive parameters themselves

$$f_{\kappa\mu}(\kappa, \mu) \;=\; \frac{1}{\kappa\,\mu} \quad . \tag{191}$$

This $1/x$ form of the probability density remains invariant if we take any power of $\kappa$ and of $\mu$. In particular, if instead of using the uncompressibility $\kappa$ we use the compressibility $\gamma = 1/\kappa$, the Jacobian rule simply gives $f_{\gamma\mu}(\gamma, \mu) = 1/(\gamma\,\mu)$.

Associated to the probability density 190 there is the Euclidean definition of distance

$$ds^2 \;=\; (d\kappa^*)^2 + (d\mu^*)^2 \quad , \tag{192}$$

that corresponds, in the variables $(\kappa, \mu)$ , to

$$ds^2 \;=\; \left(\frac{d\kappa}{\kappa}\right)^2 + \left(\frac{d\mu}{\mu}\right)^2 \quad , \tag{193}$$

i.e., to the metric

$$\begin{pmatrix} g_{\kappa\kappa} & g_{\kappa\mu} \\ g_{\mu\kappa} & g_{\mu\mu} \end{pmatrix} \;=\; \begin{pmatrix} 1/\kappa^2 & 0 \\ 0 & 1/\mu^2 \end{pmatrix} \quad . \tag{194}$$

## H.2 Young Modulus and Poisson Ratio

The Young modulus $Y$ and the Poisson ration $\sigma$ can be expressed as a function of the uncompressibility modulus and the shear modulus as

$$Y \;=\; \frac{9\,\kappa\,\mu}{3\kappa+\mu} \qquad ; \qquad \sigma \;=\; \frac{1}{2}\,\frac{3\kappa-2\mu}{3\kappa+\mu} \tag{195}$$

or, reciprocally,

$$\kappa \;=\; \frac{Y}{3(1-2\sigma)} \qquad ; \qquad \mu \;=\; \frac{Y}{2(1+\sigma)} \quad . \tag{196}$$

The absolute value of the Jacobian of the transformation is easily computed,

$$J \;=\; \frac{Y}{2(1+\sigma)^2(1-2\sigma)^2} \quad , \tag{197}$$

and the Jacobian rule transforms the probability density 191 into

$$f_{Y\sigma}(Y,\sigma) \;=\; \frac{1}{\kappa\,\mu}\;J \;=\; \frac{3}{Y\,(1+\sigma)(1-2\sigma)} \quad , \tag{198}$$

which is the probability density representing the homogeneous probability distribution for elastic parameters using the variables $(Y, \sigma)$ . This probability density is the product of the probability density $1/Y$ for the Young modulus and the probability density

$$g(\sigma) \;=\; \frac{3}{Y\,(1+\sigma)(1-2\sigma)} \tag{199}$$

for the Poisson ratio. This probability density is represented in figure 21. From the definition of $\sigma$ it can be demonstrated that its values must range in the interval $-1 < \sigma < 1/2$ , and we see that the homogeneous probability density is singular at these points. Although most rocks have positive values of the Poisson ratio, there are materials where $\sigma$ is negative (e.g., Yeganeh-Haeri et al., 1992).

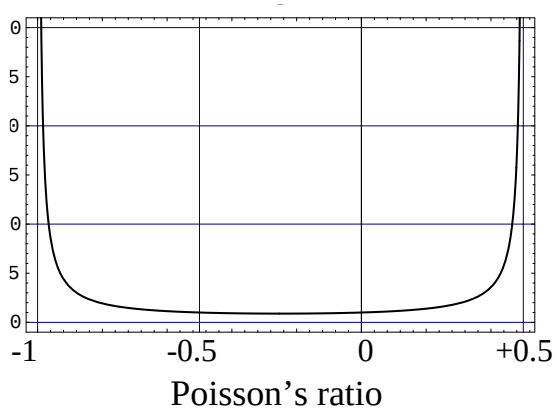



Figure 21: The homogeneous probability density for the Poisson ratio, as deduced from the condition that the uncompressibility and the shear modulus are Jeffreys parameters.

It may be surprising that the probability density in figure 21 corresponds to a homogeneous distribution. If we have many samples of elastic materials, and if their logarithmic uncompressibility modulus $\kappa^*$ and their logarithmic shear modulus $\mu^*$ have a constant probability density (what *is* the definition of homogeneous distribution of elastic materials), then, $\sigma$ will be distributed according to the $g(\sigma)$ of the figure.

To be complete, let us mention that in a change of variables $x^i \rightleftharpoons x^I$ , a metric $g_{ij}$ changes to

$$g_{IJ} \;=\; \Lambda_I{}^i \, \Lambda_J{}^j \, g_{ij} \;=\; \frac{\partial x^i}{\partial x^I} \, \frac{\partial x^j}{\partial x^J} \, g_{ij} \quad . \tag{200}$$

The metric 193 then transforms into

$$\begin{pmatrix} g_{YY} & g_{Y\sigma} \\ g_{\sigma Y} & g_{\sigma\sigma} \end{pmatrix} \;=\; \begin{pmatrix} \frac{2}{Y^2} & \frac{2}{(1-2\,\sigma)\,Y} - \frac{1}{(1+\sigma)\,Y} \\ \frac{2}{(1-2\,\sigma)\,Y} - \frac{1}{(1+\sigma)\,Y} & \frac{4}{(1-2\,\sigma)^2} + \frac{1}{(1+\sigma)^2} \end{pmatrix} \quad . \tag{201}$$

The surface element is

$$dS_{Y\sigma}(Y,\sigma) \;=\; \sqrt{\det g}\; dY\, d\sigma \;=\; \frac{3\, dY\, d\sigma}{Y\,(1+\sigma)(1-2\sigma)} \quad , \tag{202}$$

a result from which expression 198 can be inferred.

Although the Poisson ratio has a historical interest, it is not a simple parameter, as shown by its theoretical bounds $-1 < \sigma < 1/2$ , or the form of the homogeneous probability density (figure 21). In fact, the Poisson ratio $\sigma$ depends only on the ratio $\kappa/\mu$ (incompressibility modulus over shear modulus), as we have

$$\frac{1+\sigma}{1-2\sigma} = \frac{3}{2}\,\frac{\kappa}{\mu}\,. \tag{203}$$

The ratio $J = \kappa/\mu$ of two Jeffreys parameters being a Jeffreys parameter, a useful pair of Jeffreys parameters may be $\{\kappa, J\}$ . The ratio $J = \kappa/\mu$ has a physical interpretation easy to grasp (as the ratio between the uncompressibility and the shear modulus), and should be preferred, in theoretical developments, to the Poisson ratio, as it has simpler theoretical properties. As the name of the nearest metro station to the university of one of the authors (A.T.) is *Jussieu*, we accordingly call $J$ the *Jussieu's ratio*.

## H.3 Longitudinal and Transverse Wave Velocities

Equation 191 gives the probability density representing the homogeneous homogeneous probability distribution of elastic media, when parameterized by the uncompressibility modulus and the shear modulus:

$$f_{\kappa\mu}(\kappa,\mu) \;=\; \frac{1}{\kappa\,\mu} \quad . \tag{204}$$

Should we have been interested, in addition, to the mass density $\rho$ , then we would have arrived (as $\rho$ is another Jeffreys parameter), to the probability density

$$f_{\kappa\mu\rho}(\kappa,\mu,\rho) \;=\; \frac{1}{\kappa\,\mu\,\rho} \quad . \tag{205}$$

This is the starting point for this section.

What about the probability density representing the homogeneous probability distribution of elastic materials when we use as parameters the mass density and the two wave velocities? The longitudinal wave velocity $\alpha$ and the shear wave velocity $\beta$ are related to the uncompressibility modulus $\kappa$ and the shear modulus $\mu$ through

$$\alpha = \sqrt{\frac{\kappa + 4\mu/3}{\rho}} \qquad ; \qquad \beta = \sqrt{\frac{\mu}{\rho}}\,, \tag{206}$$

and a direct use of the Jacobian rule transforms the probability density 205 into

$$f_{\alpha\beta\rho}(\alpha,\beta,\rho) \;=\; \frac{1}{\rho\,\alpha\,\beta\left(\frac{3}{4} - \frac{\beta^2}{\alpha^2}\right)} \quad . \tag{207}$$

which is the answer to our question.

That this function becomes singular for $\alpha = \frac{2}{\sqrt{3}}\beta$ is just due to the fact that the "boundary" $\alpha = \frac{2}{\sqrt{3}}\beta$ can not be crossed: the fundamental inequalities $\kappa > 0$ ; $\mu > 0$ impose that the two velocities are linked by the inequality constraint

$$\alpha \;>\; \frac{2}{\sqrt{3}}\,\beta \quad . \tag{208}$$

Let us focus for a moment on the homogeneous probability density for the two wave velocities $(\alpha, \beta)$ existing in an elastic solid (disregard here the mass density $\rho$ ). We have

$$f_{\alpha\beta}(\alpha, \beta) \; = \; \frac{1}{\alpha \, \beta \, \left( \frac{3}{4} - \frac{\beta^2}{\alpha^2} \right)} \quad . \tag{209}$$

It is displayed in figure 22.

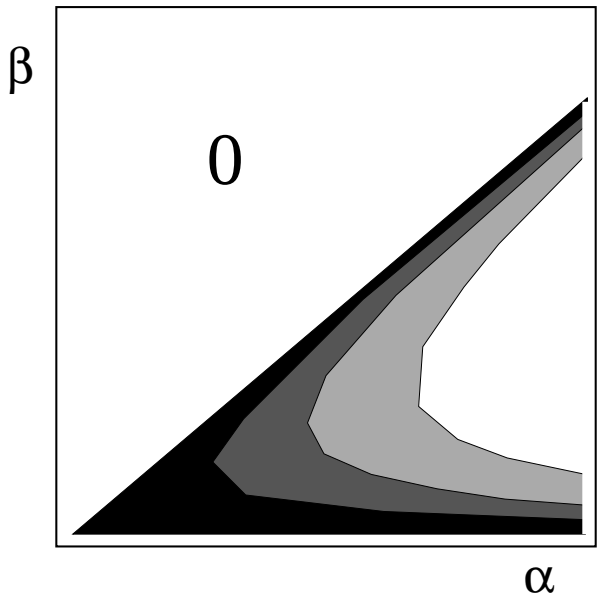



Figure 22: The joint homogeneous probability density for the velocities $(\alpha, \beta)$ of the longitudinal and transverse waves propagating in an elastic solid. Contrary to the incompressibility and the shear modulus, that are independent parameters, the longitudinal wave velocity and the transversal wave velocity are not independent (see text for an explanation). The scales for the velocities are unimportant: it is possible to multiply the two velocity scales by any factor without modifying the form of the probability (which is itself defined up to a multiplicative constant).

Let us demonstrate that the marginal probability density for both $\alpha$ and $\beta$ is of the form $1/x$ . For we have to compute

$$f_\alpha(\alpha) = \int_0^{\sqrt{3}\,\alpha/2} d\beta \; f(\alpha, \beta) \tag{210}$$

and

$$f_\beta(\beta) = \int_{2\,\beta/\sqrt{3}}^{+\infty} d\alpha \; f(\alpha, \beta) \tag{211}$$

(the bounds of integration can easily be understood by a look at figure 22). These integrals can be evaluated as

$$f_\alpha(\alpha) = \lim_{\varepsilon \to 0} \int_{\sqrt{\varepsilon}\,\sqrt{3}\,\alpha/2}^{\sqrt{1-\varepsilon}\,\sqrt{3}\,\alpha/2} d\beta \; f(\alpha, \beta) = \lim_{\varepsilon \to 0} \left( \frac{4}{3} \, \log \frac{1-\varepsilon}{\varepsilon} \right) \, \frac{1}{\alpha} \tag{212}$$

and

$$f_\beta(\beta) = \lim_{\varepsilon \to 0} \int_{\sqrt{1+\varepsilon}\,2\,\beta/\sqrt{3}}^{2\,\beta/(\sqrt{\varepsilon}\,\sqrt{3})} d\alpha \; f(\alpha, \beta) = \lim_{\varepsilon \to 0} \left( \frac{2}{3} \, \log \frac{1/\varepsilon - 1}{\varepsilon} \right) \, \frac{1}{\beta} \; . \tag{213}$$

The numerical factors tend to infinity, but this is only one more manifestation of the fact that the homogeneous probability densities are usually improper (not normalizable). Dropping these numerical factors gives

$$f_\alpha(\alpha) \; = \; \frac{1}{\alpha} \tag{214}$$

and

$$f_\beta(\beta) \; = \; \frac{1}{\beta} \; . \tag{215}$$

It is interesting to note that we have here an example where two parameters that look like Jeffreys parameters, but are not, because they are not independent (the homogeneous joint probability density is not the product of the homogeneous marginal probability densities.).

It is also worth to know that using slownesses instead of velocities ( $n = 1/\alpha, \eta = 1/\beta$ ) leads, as one would expect, to

$$f_{n\eta\rho}(n,\eta,\rho) = \frac{1}{\rho\, n\, \eta \left(\frac{3}{4} - \frac{n^2}{\eta^2}\right)} \,. \tag{216}$$

# I Homogeneous Distribution of Second Rank Tensors

The usual definition of the norm of a tensor provides the only natural definition of distance in the space of all possible tensors. This shows that, when using a Cartesian system of coordinates, the components of a tensor are the 'Cartesian coordinates' in the 6D space of symmetric tensors. The homogeneous distribution is then represented by a constant (nonnormalizable) probability density:

$$f(\sigma_{xx},\sigma_{yy},\sigma_{zz},\sigma_{xy},\sigma_{yz},\sigma_{zx}) \;=\; k \quad . \tag{217}$$

Instead of using the components, we may use the three eigenvalues $\{\lambda_1,\lambda_2,\lambda_3\}$ of the tensor and the three Euler angles $\{\psi,\theta,\varphi\}$ defining the orientation of the eigendirections in the space. As the Jacobian of the transformation

$$\{\sigma_{xx},\sigma_{yy},\sigma_{zz},\sigma_{xy},\sigma_{yz},\sigma_{zx}\} \rightleftharpoons \{\lambda_1,\lambda_2,\lambda_3,\psi,\theta,\varphi\} \tag{218}$$

is

$$\left|\frac{\partial(\sigma_{xx},\sigma_{yy},\sigma_{zz},\sigma_{xy},\sigma_{yz},\sigma_{zx})}{\partial(\lambda_1,\lambda_2,\lambda_3,\psi,\theta,\varphi)}\right| \;=\; (\lambda_1-\lambda_2)(\lambda_2-\lambda_3)(\lambda_3-\lambda_1)\;\sin\theta \quad , \tag{219}$$

the homogeneous probability density 217 transforms into

$$g(\lambda_1,\lambda_2,\lambda_3,\psi,\theta,\varphi) \;=\; k\,(\lambda_1-\lambda_2)(\lambda_2-\lambda_3)(\lambda_3-\lambda_1)\;\sin\theta \quad . \tag{220}$$

Although this is not obvious, this probability density is isotropic in spatial directions (i.e., the 3D referentials defined by the three Euler angles are isotropically distributed). In this sense, we recover 'isotropy' as a special case of 'homogeneity'.

The rule 8, imposing that any probability density on the variables $\{\lambda_1,\lambda_2,\lambda_3,\psi,\theta,\varphi\}$ has to tend to the homogeneous probability density 220 when the 'dispersion parameters' tend to infinity imposes a strong constraint on the form of acceptable probability densities, that is, generally, overlooked.

For instance, a Gaussian model for the variables $\{\sigma_{xx},\sigma_{yy},\sigma_{zz},\sigma_{xy},\sigma_{yz},\sigma_{zx}\}$ is consistent (as the limit of Gaussian is a constant). This induces, via the Jacobian rule, a probability density for the variables $\{\lambda_1,\lambda_2,\lambda_3,\psi,\theta,\varphi\}$, a probability density that is not simple, but consistent. A Gaussian model for the parameters $\{\lambda_1,\lambda_2,\lambda_3,\psi,\theta,\varphi\}$ would not be consistent.

# J Example of Ideal (Although Complex) Geophysical Inverse Problem

Assume we wish to explore a complex medium, like the Earth's crust, using elastic waves. Figure 23 suggests an Earth model and a set of seismograms produced by the waves generated by an earthquake (or an artificial source). The seismometers (not represented) may be at the Earth's surface or inside boreholes. Although only four seismograms are displayed, actual experiments may generate thousands or millions of them. The problem here is to use a set of observed seismograms to infer the structure of the Earth.

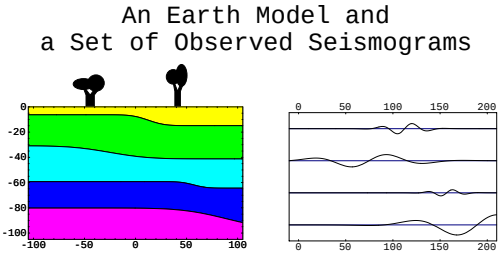



Figure 23: A set of observed seismograms (at the right) is to be used to infer the structure of the Earth (at the left). A couple of trees suggest an scale (the numbers could correspond to meters), although the same principle can be used for global Earth tomography.

The first step is to define the set of parameters to be used to represent an Earth model. These parameters have to qualitatively correspond to the ideas we have about the Earth's interior: Thicknes and curvature of the

geological layers, position and dip of the geological faults, etc. Inside of the bodies so defined, different types of rocks will correspond to different values of some geophysical quantities (volumetric mass, elastic rigidity, porosity, etc.). These quantities, that have a smooth space variation (inside a given body), may be discretized by considering a grid of points, by using a discrete basis of functions to represent them, etc. If the source of seismic waves is not perfectly known (this is alwaus the case it the source is an earthquake), then, the parameters describing the source also belong to the 'model parameter set'.

A given Earth model (including the source of the waves), then, will consist in a huge set of values: the 'numerical values' of all the parameters being used in the description. For instance, we may use the parameters $\mathbf{m} = \{m^1, m^2, \ldots, m^M\}$ to decribe an Earth model, where $M$ may be a small number (for simple 1D models) or a large number (by the millions or billions for comlex 3D models). Then, we may consider an 'Earth model number one', denoted $\mathbf{m}_1$, an 'Earth model number two', denoted $\mathbf{m}_2$, and so on.

Now, what is a seismogram? It is, in fact, one of the components of a vectorial function $\mathbf{s}(t)$ that depends on the vectorial displacement $\mathbf{r}(t)$ of the particles 'at the point' where the seismometer is located. Given the manufacturing parameters of the seismometers, then, it is possible to calculate the 'output' (seismogram) $\mathbf{s}(t)$ that corresponds to a given 'input' (soil displacement) $\mathbf{r}(t)$. In some loosy sense, the instrument acts as a 'nonlinear filter' (the nonlinearity coming from the possible saturation of the sensors for large values of the input, or from their insensivity to small values). While the displacement of the soil is measured, say, in micrometers, the output of the seismometer, typically an electric tension, is measured, say in millivolts. In our digital era, seismograms are not recorded as 'functions'. Rather, a discrete value of the output is recorded with a given frequency (for instance, one value every millisecond). A *seismogram set* consists, then, in a large number of (discrete) values, say, $s_{iam} = (s_i(t_a))_m$ representing the value at time $t_a$ of the $i$-th component of the $m$-th seismogram. Such a seismogram set is what is schematically represented at the right of figure 23. For our needs, the particular structure of a seismogram set is not interesting, and we will simply represent such a set using the notation $\mathbf{d} = \{d^1, d^2, \ldots, d^N\}$, where the number $N$ may range in the thousands (if we only have one seismogram), or in the trillions for global Earth data or data from seismic exploration for minerals.

An exact theory then defines a function $\mathbf{d} = \mathbf{f}(\mathbf{m})$ : given an arbitrary Earth model $\mathbf{m}$, the associated theoretical seismograms $\mathbf{d} = \mathbf{f}(\mathbf{m})$ can be computed.

A 'theory' able to predict seismograms has to encompass the whole way between the Earth model and the instrument output, the millivolts. An 'exact theory' would define a functional relationship $\mathbf{d} = \mathbf{f}(\mathbf{m})$ associating, to any Earth model $\mathbf{m}$ a precisely defined point in the data space. This theory would essentially consist in the theory of elastic waves in anisotropic and heterogeneous media, perhaps modified to include attenuation, nonlinear effects, the descrotion of the recording instrument, etc.

As mentioned in section 4.6 there are many reasons for which a 'theory' is not an exact functional relationship but, rather, a conditional volumetric probability $\vartheta(\mathbf{d}|\mathbf{m})$. Realistic estimations of this probability distribution may be extremely complex. Sometimes we may limit ourselves to 'putting uncertainty bars around a functional relation', as suggested in section 4.6.1. Then, for instance, using a Gaussian model, we may write

$$\vartheta(\mathbf{d}|\mathbf{m}) = k \, \exp\left( -\frac{1}{2} \left( \mathbf{d} - \mathbf{f}(\mathbf{m}) \right)^T \mathbf{C}_T^{-1} \left( \mathbf{d} - \mathbf{f}(\mathbf{m}) \right) \right) \,, \tag{221}$$

where the uncertaintity on the predicted data point, $\mathbf{d} = \mathbf{f}(\mathbf{m})$ is described by the 'theory covariance operator' $\mathbf{C}_T$. With a simple probability model, lime this one, or by any other means, it is assumed the the conditional probability volumetric probability $\vartheta(\mathbf{d}|\mathbf{m})$ is defined. Then, given any point $\mathbf{m}$ representing an Earth model, we should be able to sample the volumetric probability $\vartheta(\mathbf{d}|\mathbf{m})$, i.e., to obtain as many samples (specimens) of $\mathbf{d}$ as we may wish. Figure 24 gives a schematic illustration of this.

Assume that we do not have yet collected the seismograms. At this moment, the information we have on the Earth is called 'a priori' information. As explained elsewhere in this text, it may always be represented by a probability distribution over the model parameter space, corresponding to a volumetric probability $\rho_m(m)$. The expression of this volumetric probability is, in realistic problems, never explicitly known. Let us see this with some detail.

In some very simple situations, we may have an 'average a priori model' $\mathbf{m}_{\text{prior}}$ and a priori uncertainties that can be modeled by a Gaussian distribution with covariance operator $\mathbf{C}_m$. Then,

$$\rho_m(\mathbf{m}) = k \, \exp\left( -\frac{1}{2} \left( \mathbf{m} - \mathbf{m}_{\text{prior}} \right)^T \mathbf{C}_m^{-1} \left( \mathbf{m} - \mathbf{m}_{\text{prior}} \right) \right) \,. \tag{222}$$

Other probability models (Laplace, Pareto, etc.) may, or course, be used. In more realistic situations, the a

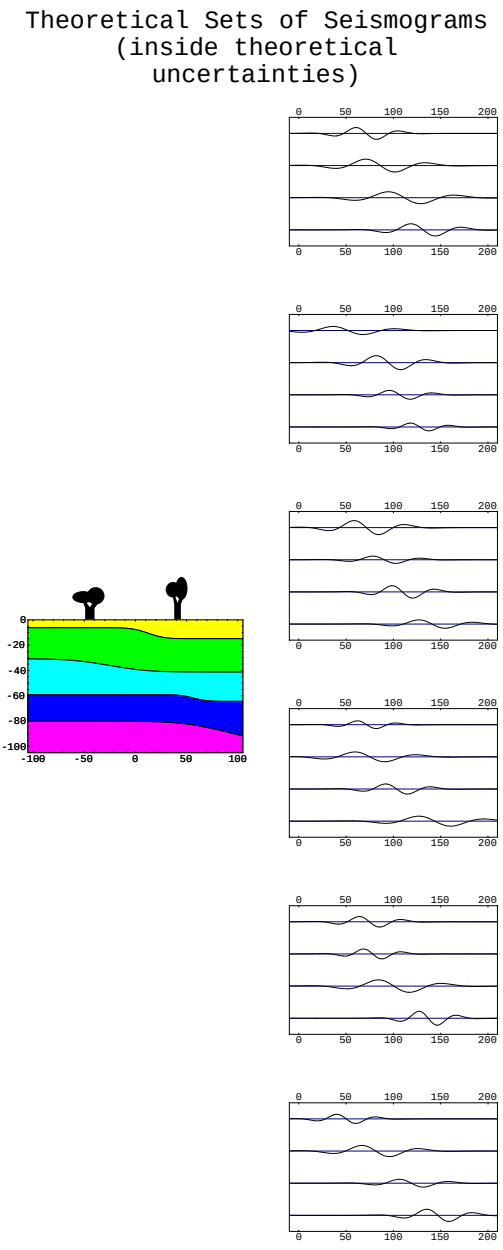



Figure 24: Given an arbitrary Earth model $\mathbf{m}$ , a (non exact) theory given a probability distribution for the data, $\vartheta(\mathbf{d}|\mathbf{m})$ , than can be smpled, producing the sets of seismograms shown here.

priori information we have over the model space is not easily expressible as an explicit expression of a volumetric probability. Rather, a large set of rules, some of them probabilistic, is expressed.

Already, the very definition of the parameters contains a fundamental topological information (the type of objects being considered: geological layers, faults, etc.). Then, we may have rules of the type 'a sedimentary layer may never be below a layer of igneous origin' or 'with probability 2/3, a layer with a thichness larger that $D$ is followed by a layer with a thickness smaller than $d$ ', etc. There are, also, explicit volumetric probabilities, like 'the joint volumetric probability for porosity $\pi$ and rigidity $\mu$ for a calcareous layer is $g(\pi, \mu) = \dots$ '. They may come from statistical studies made using large petrophysical data banks, or from qualitative 'Bayesian' estimations of the correlations existing between different parameters.

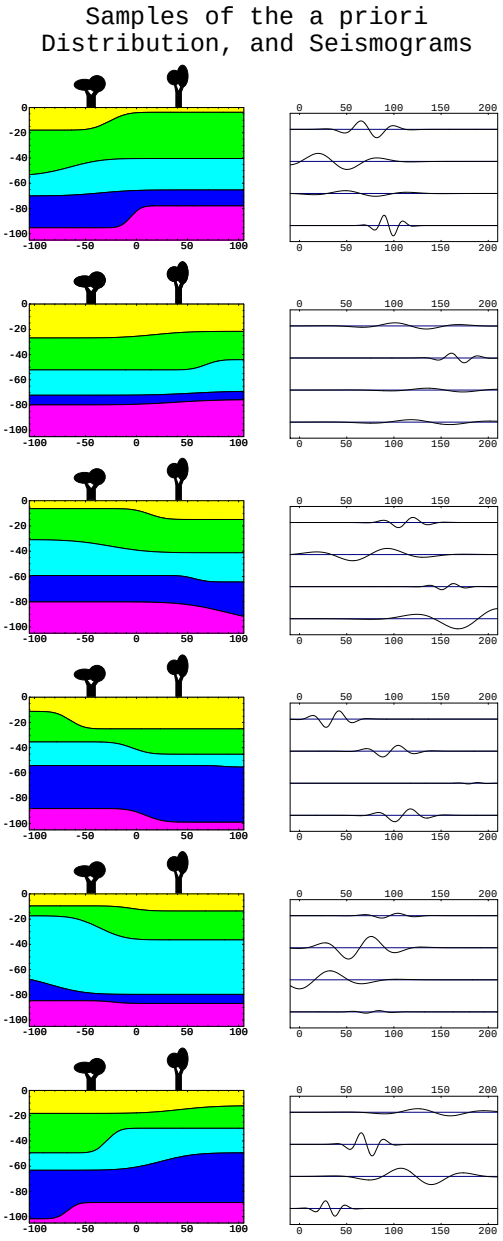



Figure 25: Samples of the a priori distribution of Earth models, each accompanied by the predicted set of seismograms. A set of rules, some determistic, some random, is used to randomly generate Earth models. These are assumed to be samples from a probability distribution over the model space corresponding to a volumetric probability $\rho_m(\mathbf{m})$ whose explicit expression may be difficult to obtain. But it is not this expression that is required for proceeding with the method, only the possibility of obtaining as many samples of it as we may wish. Although a large number of samples may be necessary to grasp all the details of a probability distribution, as few as the six samples shown here already provide some elementary information. For instance, there are always five geological layers, separated by smooth interfaces. In each model, all the four interfaces are dipping 'leftwards' or all the four are dipping 'rightwards'. These observations may be confirmed, and other properties become conspicuous as more and more samples are displayed. The theoretical set of the seismograms associated to each model, displayed at rigth, are as different as the models are different. Thse are only 'schematic' seismograms, bearing no relation with any actual situation.

The fundamental hypothesis of the approach that follows is that we are able to use this set of rules to randomly generate Earth models. And as many as may wish. Figure 25 suggests the results obtained using such a procedure. In a computer screen, when the models are displayed one after the other, we have a 'movie'. A geologist (knowing nothing about mathematics) should, when observing such a movie for long enough, agree with a sentence like the following. *All models displayed are possible models; the more likely models appear quite frequently; some unlikely models appear, but unfrequently; if we wait long enough we may well reach a model that may be arbitrarily close to*

*the actual Earth.*

This means that (i) we have described the a priori information, by defining a probability distribution over the model space, (ii) we are sampling this probability distribution, event if an expression for the associated volumetric probability $\rho_m(\mathbf{m})$ has not been developed explicitly.

Assume now that we collect a data set, i.e., in our example, the set of seismograms generated by a given set of earthquakes, or by a given set of artificial sources. In the notation introduced above, a given set of seismograms corresponds to a particular point $\mathbf{d}$ in the data space. As any measurement has attached uncertainties, rather than 'a point' in the data space, we have, as explained elsewhere in this text, a probability distribution in the data space, corresponding to a volumetric probability $\rho_d(\mathbf{d})$ .

The simplests examples of probability distribution in the data space are obtained when using simple probability models. For instance, the assumption of Gaussian uncertainties would give

$$\rho_d(\mathbf{d}) = k \, \exp\left( -\frac{1}{2} \, (\mathbf{d} - \mathbf{d}_{\rm obs})^T \, \mathbf{C}_d^{-1} \, (\mathbf{d} - \mathbf{d}_{\rm obs}) \right) \, , \tag{223}$$

where $\mathbf{d}_{\rm obs}$ represents the 'observed data values', with 'experimental uncertainties' described by the covariance operator $\mathbf{C}_d$ . As always, other probability models may, of course, be used.

Note that from

$$\begin{array}{rcl} \sigma(\mathbf{d},\mathbf{m}) & = & k \; \rho(\mathbf{d},\mathbf{m}) \; \vartheta(\mathbf{d},\mathbf{m}) \\ \rho(\mathbf{d},\mathbf{m}) & = & \rho_d(\mathbf{d}) \; \rho_m(\mathbf{m}) \\ \vartheta(\mathbf{d},\mathbf{m}) & = & \vartheta(\mathbf{d}|\mathbf{m}) \; \vartheta_m(\mathbf{m}) = k \; \vartheta(\mathbf{d}|\mathbf{m}) \end{array} \tag{224}$$

it follows

$$\sigma(\mathbf{d},\mathbf{m}) = k \; \rho_d(\mathbf{d}) \; \vartheta(\mathbf{d}|\mathbf{m}) \; \rho_m(\mathbf{m}) \; . \tag{225}$$

Assume that we are able to generate a random walk that samples the a priori probability distribution of Earth models, $\rho_m(\mathbf{m})$ (we have seen above how to do this; see also section XXX). Consider the following algorithm:

1. Initialize the algorithm at an arbitrary point $(\mathbf{m}_1, \mathbf{d}_1)$ , the first 'accepted' point.
2. Relabel the last accepted point $(\mathbf{m}_n, \mathbf{d}_n)$ . Given $\mathbf{m}_n$ , use the rules that sample the volumetric probability $\rho_m(\mathbf{m})$ to generate a candidate point $\mathbf{m}_c$ .
3. Given $\mathbf{m}_c$ , randomly generate a sample data point, according to the volumetric probability $\vartheta(\mathbf{d}|\mathbf{m}_c)$ , and name it $\mathbf{d}_c$ .
4. Compare the values $\rho_d(\mathbf{d}_n)$ and $\rho_d(\mathbf{d}_c)$ , and decide to accept or to reject the candidate point $\mathbf{d}_c$ according to the logistic or to the Metropolis rule (or any equivalent rule). If the candidate point is accepted, set $(\mathbf{m}_{n+1}, \mathbf{d}_{n+1}) = (\mathbf{m}_c, \mathbf{d}_c)$ and go to 2. If the candidate point is rejected, set $(\mathbf{m}_{n+1}, \mathbf{d}_{n+1}) = (\mathbf{m}_n, \mathbf{d}_n)$ and go to 2.

Figure 27 shows some samples of the a posteriori probability distribution.

Note that the marginal for $\mathbf{m}$ corresponds to the same 'movie', just looking to the models, and disregarding the date sets. Reciprocally, the marginal for $\mathbf{d}$ ...

'Things' can be considerably simplified if uncertainties in the theory can be neglected (i.e., if the 'theory' is assumed to be exact):

$$\vartheta(\mathbf{d}|\mathbf{m}) = \delta(\, \mathbf{d} - \mathbf{f}(\mathbf{m}) \,) \; . \tag{226}$$

Then, the marginal for $\mathbf{m}$ , $\sigma_m(\mathbf{m}) = \int dV_d(d) \; \sigma(\mathbf{d},\mathbf{m})$ , is using 225,

$$\sigma_m(\mathbf{m}) = k \; \rho_m(\mathbf{m}) \; \rho_d(\, \mathbf{f}(\mathbf{m}) \,) \; . \tag{227}$$

The algorithm proposed above, simplifies to:

1. Initialize the algorithm at an arbitrary point $\mathbf{m}_1$ , the first 'accepted' point.
2. Relabel the last accepted point $\mathbf{m}_n$ . Use the rules that sample the volumetric probability $\rho_m(\mathbf{m})$ to generate a candidate point $\mathbf{m}_c$ .
3. Compute $d_c = \mathbf{f}(\mathbf{m}_c)$ .
4. Compare the values $\rho_d(\mathbf{d}_n)$ and $\rho_d(\mathbf{d}_c)$ , and decide to accept or to reject the candidate point $\mathbf{d}_c$ according to the logistic or to the Metropolis rule (or any equivalent rule). If the candidate point is accepted, set $\mathbf{m}_{n+1} = \mathbf{m}_c$ and go to 2. If the candidate point is rejected, set $\mathbf{m}_{n+1} = \mathbf{m}_n$ and go to 2.

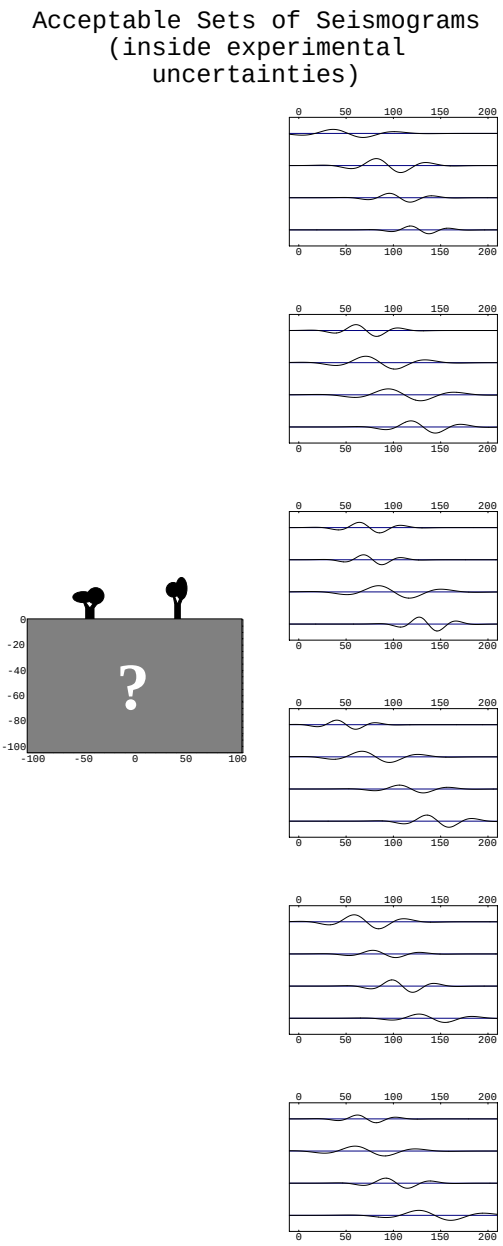



Figure 26: We have one 'observed set of seismograms', together with a description of the uncertainties in the data. The corresponding probability distribution may be complex (correlation of uncertainties, non Gaussianity of the noise, etc.). Rather than plotting the 'oberved set of seismograms', pseudorandom realizations of the probability distribution in the data space are displayed here.

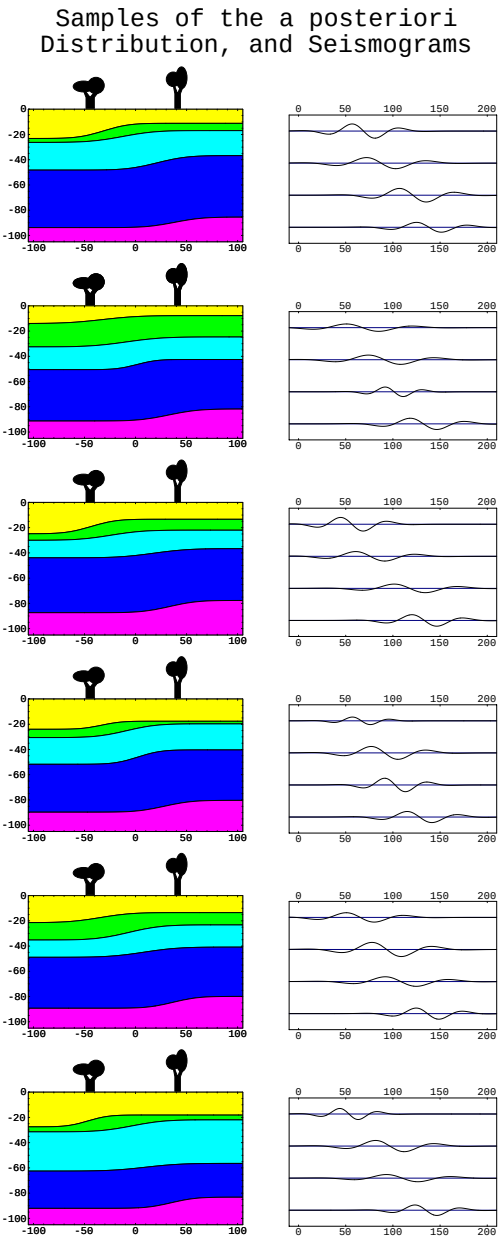



Figure 27: Samples of the a posteriori distribution of Earth models, each accompanied by the predicted set of seismograms. Note that, contrary to what happens with the a priori samples, all the models presented here have 'left dipping interfaces'. The second layer is quite thin. Etc.

# K An Example of Partial Derivatives

Let us consider the problem of locating a point in space using a system like the Global Positioning System (GPS), where some sources (satellites) send waves to a receiver, which measures the travel times. Let us use Cartesian coordinates, denote by $(x^i, y^i, z^i)$ the position of the $i$-th source, and by $(x^R, y^R, z^R)$ the position of the receiver. Simplify the problem here by assuming that the the medium where the waves propagate is homogeneous, so the velocity of the waves is constant (say $v$ ) and the rays are straight lines. Then, the travel time from the $i$-th source to the receiver is

$$t^i = g^i(x^R, y^R, z^R, v) = \frac{\sqrt{(x^R - x^i)^2 + (y^R - y^i)^2 + (z^R - z^i)^2}}{v} \ . \tag{228}$$

The dependence of $t^i$ on the variables describing the source positions $(x^i, y^i, z^i)$ is not explicitly considered, as the typical GPS problem consists in assuming the position of the sources exactly known, and to estimate the receiver position $(x^R, y^R, z^R)$ . At no extra cost we can also try to estimate the velocity of propagation of waves $v$ . The partial derivatives of the problem are then

$$\begin{pmatrix} \frac{\partial g^1}{\partial x^R} & \frac{\partial g^1}{\partial y^R} & \frac{\partial g^1}{\partial z^R} & \frac{\partial g^1}{\partial v} \\ \frac{\partial g^2}{\partial x^R} & \frac{\partial g^2}{\partial y^R} & \frac{\partial g^2}{\partial z^R} & \frac{\partial g^2}{\partial v} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial g^i}{\partial x^R} & \frac{\partial g^i}{\partial y^R} & \frac{\partial g^i}{\partial z^R} & \frac{\partial g^i}{\partial v} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} = \begin{pmatrix} \frac{x^R - x^1}{v\, D^1} & \frac{y^R - y^1}{v\, D^1} & \frac{z^R - z^1}{v\, D^1} & -\frac{D^i}{v^2} \\ \frac{x^R - x^2}{v\, D^2} & \frac{y^R - y^2}{v\, D^2} & \frac{z^R - z^2}{v\, D^2} & -\frac{D^i}{v^2} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{x^R - x^i}{v\, D^i} & \frac{y^R - y^i}{v\, D^i} & \frac{z^R - z^i}{v\, D^i} & -\frac{D^i}{v^2} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \ , \tag{229}$$

where $D^i$ is a short notation for the distance

$$D^i = \sqrt{(x^R - x^i)^2 + (y^R - y^i)^2 + (z^R - z^i)^2} \ . \tag{230}$$

In order to keep notations simple, it has not been explicitly indicated that these partial derivatives are functions of the variables of the problem, i.e., as functions of $(x^R, y^R, z^R, v)$ (remember that the locations of the satellites, $(x^i, y^i, z^i)$ are assumed exactly known, so they are not "variables"). Assigning particular values to the variables $(x^R, y^R, z^R, v)$ gives particular values for the travel times $t^i$ (through equation 228) and for the partial derivatives (through equation 229).

# L Probabilistic Estimation of Hypocenters

Earthquakes generate waves, and the arrival times of the waves at a network of seismic observatories carries information on the location of the hypocenter. This information is better understood by a direct examination of the probability density $f(X, Y, Z)$ defined by the arrival times, rather than just estimating a particular location $(X, Y, Z)$ and the associated uncertainties.

Provided that a 'black box' is available that rapidly computes the travel times to the seismic station from any possible location of the earthquake, this probabilistic approach can be relatively efficient. Tjhis appendix shows that it is quite trivial to write a computer code that uses this probabilistic approach (much easier than to write a code using the traditional Geiger method, that seeks to obtain the 'best' hypocentral coordinates).

## L.1 A Priori Information on Model Parameters

The 'unknowns' (morel parameters) of the problem are the hypocentral coordinates of an Earthquake[32] $\{X, Z\}$ , as well as the origin time $T$ . We assume to have some a priori information about the location of the earthquake, as well as about ots origin time. This a priori information is assumed to be represented using the probability density

$$\rho_m(X, Z, T) \quad . \tag{231}$$

Because we use Cartesian coordinates and Newtonian time, the homogeneous probability density is just a constant,

$$\mu_m(X, Y, T) \ = \ k \quad . \tag{232}$$

For consistency, we must assume (rule 8) that the limit of $\rho_m(X, Z, T)$ for infinite 'dispersions' is $\mu_m(X, Z, T)$ .

[32] To simplify, here, we consider a 2D flat model of the Earth, and use Cartesian coordinates.

**Example 25** *We assume that the a priori probability density for* $(X, Z)$ *is constant inside the region* $0 < X < 60\,\mathrm{km}$ , $0 < Z < 50\,\mathrm{km}$ *, and that the (unnormalizable) probability density for* $T$ *is constant.* [END OF EXAMPLE.]

## L.2 Data

The data of the problem are the arrival times $\{t^1, t^2, t^3, t^4\}$ of the seismic waves at a set of four seismic observatories whose coordinates are $\{x^i, z^i\}$ . The measurement of the arrival times will produce a probability density

$$\rho_d(t^1, t^2, t^3, t^4) \tag{233}$$

over the 'data space'. As these are Newtonian times, the associated homogeneous probability density is constant:

$$\mu_d(t^1, t^2, t^3, t^4) \;=\; k \quad . \tag{234}$$

For consistency, we must assume (rule 8) that the limit of $\rho_d(t^1, t^2, t^3, t^4)$ for infinite 'dispersions' is $\mu_d(t^1, t^2, t^3, t^4)$ .

**Example 26** *Assuming Gaussian, independent uncertainties, we have*

$$\begin{aligned} \rho_d(t^1, t^2, t^3, t^4) \;&=\; k\, \exp\left(-\frac{1}{2}\frac{(t^1 - t^1_{\mathrm{obs}})^2}{\sigma_1^2}\right)\, \exp\left(-\frac{1}{2}\frac{(t^2 - t^2_{\mathrm{obs}})^2}{\sigma_2^2}\right) \\ &\times \quad \exp\left(-\frac{1}{2}\frac{(t^3 - t^3_{\mathrm{obs}})^2}{\sigma_3^2}\right)\, \exp\left(-\frac{1}{2}\frac{(t^4 - t^4_{\mathrm{obs}})^2}{\sigma_4^2}\right) \quad . \end{aligned} \tag{235}$$

[END OF EXAMPLE.]

## L.3 Solution of the Forward Problem

The forward problem consists in calculating the arrival times $t^i$ as a function of the hypocentral coordinates $\{X, Z\}$ , and the origin time $T$ :

$$t^i \;=\; f^i(X, Z, T) \quad . \tag{236}$$

**Example 27** *Assuming that the velocity of the medium is constant, equal to* $v$ *,*

$$t^1_{\mathrm{cal}} \;=\; T + \frac{\sqrt{(X - x^i)^2 + (Z - z^i)^2}}{v} \quad . \tag{237}$$

## L.4 Solution of the Inverse Problem

Putting all this together gives

$$\sigma_m(X, Z, T) \;=\; k\, \rho_m(X, Z, T)\;\; \rho_d(t^1, t^2, t^3, t^4)\big|_{t^i = f^i(X,Z,T)} \quad . \tag{238}$$

## L.5 Numerical Implementation

To show how simple is to implement an estimation of the hypocentral coordinates using the solution given by equation 238, we give, in extenso, all the commands that are necessary to the implementation, using a commercial mathematical software (Mathematica). Unfortunately, while it is perfectly possible, using this software, to explicitly use quantities with their physical dimensions, the plotting routines require adimensional numbers. This is why the dimensions have been suppresed in whay follows. We use kilometers for the space positions and seconds for the time positions.

We start by defining the geometry of the seismic network (the vertical coordinate z is oriented with positive sign upwards):

```
x1 = 5;
z1 = 0;
x2 = 10;
z2 = 0;
x3 = 15;
z3 = 0;
x4 = 20;
z4 = 0;
```

The velocity model is simply defined, in this toy example, by giving its constant value ( 5 km/s ):

```
v = 5;
```

The 'data' of the problem are those of example 26. Explicitly:

```
t1OBS = 30.3;
s1 = 0.1;
t2OBS = 29.4;
s2 = 0.2;
t3OBS = 28.6;
s3 = 0.1;
t4OBS = 28.3;
s4 = 0.1;

rho1[t1_] := Exp[ - (1/2) (t1 - t1OBS)^2/s1^2 ]
rho2[t2_] := Exp[ - (1/2) (t2 - t2OBS)^2/s2^2 ]
rho3[t3_] := Exp[ - (1/2) (t3 - t3OBS)^2/s3^2 ]
rho4[t4_] := Exp[ - (1/2) (t4 - t4OBS)^2/s4^2 ]

rho[t1_,t2_,t3_,t4_]:=rho1[t1] rho2[t2] rho3[t3] rho4[t4]
```

Although an arbitrarily complex velocity velocity model could be considered here, let us take, for solving the forward problem, the simple model in example 27:

```
t1CAL[X_, Z_, T_] := T + (1/v) Sqrt[ (X - x1)^2 + (Z - z1)^2 ]
t2CAL[X_, Z_, T_] := T + (1/v) Sqrt[ (X - x2)^2 + (Z - z2)^2 ]
t3CAL[X_, Z_, T_] := T + (1/v) Sqrt[ (X - x3)^2 + (Z - z3)^2 ]
t4CAL[X_, Z_, T_] := T + (1/v) Sqrt[ (X - x4)^2 + (Z - z4)^2 ]
```

The posterior probability density is just that defined in equation 238:

```
sigma[X_,Z_,T_] := rho[t1CAL[X,Z,T],t2CAL[X,Z,T],t3CAL[X,Z,T],t4CAL[X,Z,T]]
```

We should have multiplied by the $\rho_m(X,Z,T)$ defined in example 25, but as this just corresponds to a 'trimming' of the values of the probability density outside the 'box' $0 < X < 60\,\text{km}$ , $0 < Z < 50\,\text{km}$ , we can do this afterwards.

The defined probability density is 3D, and we could try to represent it. Instead, let us just represent the marginal probabilty densities. First, we ask the software to evaluate analytically the space marginal:

```
sigmaXZ[X_,Z_] = Integrate[ sigma[X,Z,T], {T,-Infinity,Infinity} ];
```

This gives a complicated result, with hypergeometric functions[33]. Representing this probability density is easy, as we just need to type the command

```
ContourPlot[-sigmaXZ[X,Z],{X,15,35},{Z,0,-25},
            PlotRange->All,PlotPoints->51]
```

The result is represented in figure 28 (while the level lines are those directly produced by the software, there has been some additional editing to add the labels). When using `ContourPlot`, we change the sign of sigma, because we wish to reverse the software's convention of using light colors for positive values. We have chosen the right region of the space to be plotted (significant values of sigma) by a preliminary plotting of 'all' the space (not represented here).

Should we have some a priori probability density on the location of the earthquake, represented by the probability density f(X,Y,Z), then, the theory says that we should multiply the density just plotted by f(X,Y,Z). For instance, if we have the a priori information that the hypocenter is above the level $z = -10$ km, we just put to zero everyhing below this level in the figure just plotted.

Let us now evaluate the marginal probability density for the time, by typing the command

```
sigmaT[T_] := NIntegrate[ sigma[X,Z,T], {X,0,+60}, {Z,0,+50} ]
```

[33] Typing `sigmaXZ[X,Z]` presents the result.

Here, we ask Mathematica NOT to try to evaluate analytically the result, but to perform a numerical computation (as we have checked that no analytical result is found). We use the 'a priori information' that the hypocenter must be inside a region $0 < X < 60\,\text{km}$ , $0 < Z < 50\,\text{km}$ but limiting the integration domain to that area (see example 25). To represent the result, we enter the command

```
p = Table[0,{i,1,400}];
Do[ p[[i]] = sigmaT[i/10.] , {i,100,300}]
ListPlot[ p,PlotJoined->True, PlotRange->{{100,300},All}]
```

and the produced result is shown (after some editing) in figure 29. The software was not very stable in producing the results of the numerical integhration.

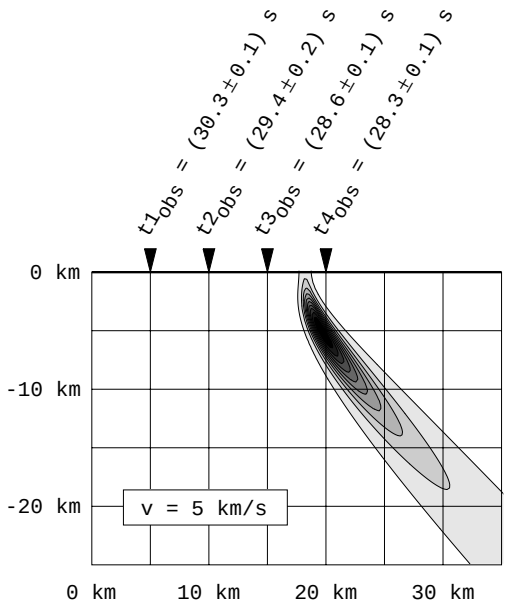



Figure 28: The probability density for the location of the hypocenter. Its asymmetric shape is quite typical, as seismic observatories tend to be asymmetrically placed.

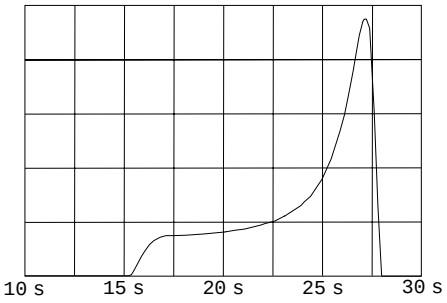



Figure 29: The marginal probability density for the origin time. The asymmetry seen in the probability density in figure 28, where the decay of probability is slow downwards, translates here in significant probabilities for early times. The sharp decay of the probability density for $t < 17s$ does not come from the values of the arrival times, but from the a priori information that the hypocenters must be above the depth $Z = -50\,\text{km}$ .

## L.6 An Example of Bimodal Probability Density for an Arrival Time.

As an exercise, the reader could reformulate the problem replacing the assumtion of Gaussian uncertainties in the arrival times by multimodal probability densities. For instance, figure 5 suggested the use of a bimodal probability density for the reading of the arrival time of a seismic wave. Using the Mathematica software, the command

```
rho[t_] := (If[8.0<t<8.8,5,1]  If[9.8<t<10.2,10,1])
```

defines a probability density that, when plotted using the command

```
Plot[ rho[t],{t,7,11} ]
```

produces the result displayed in figure 30.

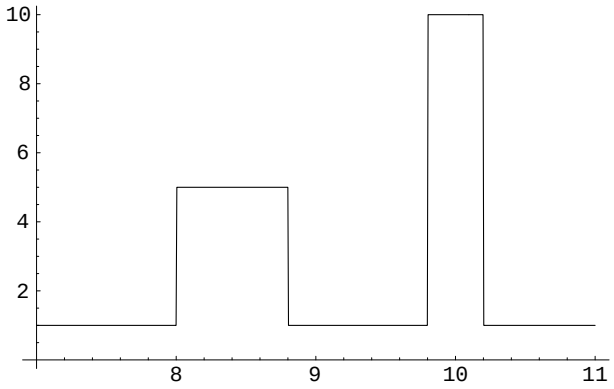



Figure 30: In figure 5 it was suggested that the probability density for the arrival time of a seismic phase may be multimodal. This is just an example to show that it is quite easy to define such multimodal probability densities in computer codes, even if they are not analytic.

# M Functional Inverse Problems

## M.1 Introduction

The main concern of this article is with discrete problems, i.e., problems where the number of data/parameters is finite. When functions are involved, it was assumed that a sampling of the function could be made that was fine enough for subsequent refinements of the sampling having no effect on the results. This, of course, means replacing any step (Heaviside) function by a sort of discretized Erf function[34]. The limit of a very steep Erf function being the step function, any functional operation involving the Erf will have as limit the same functional operation involving the step (unless very pathological problems are considered).

The major reason for this limitation is that probability theory is easily developed in finite-dimensional spaces, but not in infinite-dimensional spaces. In fact, the only practical infinite-dimensional probability theory, where 'measures' are replaced by 'cylinder measures', is nothing but the assumption that the probabilities calculated have a well behaved limit when the dimensions of the space tend to infinity. Then, the 'cylinder measure' or 'probability' of a region of the infinite-dimensional space is defined as the limit of the probability calculated in a finite-dimensional subspace, when the dimensions of this subspace tend to infinity.

There are, nevertheless, some parcels of the theory whose generalization to the infinite dimensional case is possible and well understood. For instance, infinite dimensional Gaussian probability distributions have been well studied. This is not well surprised, because the random realizations of an infinite dimensional Gaussian probability distribution are $L_2$ functions, la crème de la crème of the functions.

Most of what will be said here will concern $L_2$ functions[35], and formulas presented will be the functional equivalent to the least-squares formalism developed above for discrete problems. In fact, most results will be valid for $L_p$ functions. The difference, of course, between an $L_2$ space and an $L_p$ space is the existence of an scalar product in the $L_2$ spaces, scalar product intimately related, as we will see, with the covariance operator typical of Gaussian probability distributions.

We face here an unfortunate fact that plagues some mathematical literature: the abuse of the term 'adjoint operator' where the simple 'transpose operator' would suffice. As we will see below, the transposed of a linear operator is something as simple as the original operator (like the transpose of a matrix is as simple as the original matrix), but the adjoint of an operator is a different thing. It is defined only in spaces that have a scalar product (i.e., in $L_2$ spaces), and depends essentially of the particular scalar product of the space. As the scalar product is, usually, nontrivial (it will always involve covariance operators in our examples), the adjoint operator is generally an object more complex than the transpose operator. What we need, for using optimization methods in functional spaces, is to be able to define the norm of a function, and the transposed of an operator, so the ideal setting is that of $L_p$ spaces. Unfortunately, most mathematical results that, in fact, are valid for $L_p$ , are demonstrated only for $L_2$ .

The steps necessary for the solution of an inverse problem involving functions are: (i) definition of the functional norms; (ii) definition of the (generally nonlinear) application between parameters and data (forward problem); (iii) calculation of its tangent linear application (characterized by a linear operator); (iv) understanding of the transposed of this operator; (v) setting an iterative procedure that leads to the function minimizing the norm of the 'misfit'.

[34]The Erf function, or error function, is the primitive of a Gaussian. It is a simple example of a 'sigmoidal' function.

[35]Grossly speaking, a function $f(x)$ belongs to $L_2$ if $\| f \| = \left( \int dx f(x)^2 \right)^{1/2}$ is finite. A function $f(x)$ belongs to $L_p$ if $\| f \| = \left( \int dx |f(x)|^p \right)^{1/p}$ is finite. The limit for $p \to \infty$ corresponds to the $l_\infty$ space.

Let us see here the main mathematical points to be understood prior to any attempt of 'functional inversion'. There are not many good books on functional analysis, the best probably is the 'Introduction to Functional Analysis' by Taylor and Lay (1980).

## M.2 The Functional Spaces Under Investigation

A seismologist may consider a (three-component) seismogram

$$\mathbf{u} = \{ \quad u^i(t) \quad ; \quad i = 1,2,3 \quad ; \quad t_0 \leq t \leq t_1 \quad \} \qquad , \tag{239}$$

representing the displacement of a given material point of an elastic body, as a function of time. She/he may wish to define the norm of the function (in fact of 'the set of three functions') $\mathbf{u}$ , denoted $\| \mathbf{u} \|$ , as

$$\| \mathbf{u} \|^2 = \int_{t_0}^{t_1} dt \; u_i(t) \, u^i(t) \; , \tag{240}$$

where, as usual, $u_i \, u^i$ stands for the Euclidean scalar product. The space of all the elements $\mathbf{u}$ where this norm $\| \mathbf{u} \|$ is finite, is, by definition, an $L_2$ space.

This plain example is here to warn against wrong definitions of norm. For instance, we may measure a resistivity-versus-depth profile

$$\boldsymbol{\rho} = \{ \quad \rho(z) \quad ; \quad z_0 \leq z \leq z_1 \quad \} \qquad , \tag{241}$$

but it will generally not make sense to define

$$\| \boldsymbol{\rho} \|^2 = \int_{z_0}^{z_1} dz \; \rho(\mathbf{z})^2 \qquad \text{(bad definition)} \quad . \tag{242}$$

For the resistivity-versus-depth profile is equivalent to the conductivity-versus-depth profile

$$\boldsymbol{\sigma} = \{ \quad \sigma(z) \quad ; \quad z_0 \leq z \leq z_1 \quad \} \qquad , \tag{243}$$

where, for any $z$ , $\rho(z) \, \sigma(z) = 1$ , and the definition of the norm

$$\| \boldsymbol{\sigma} \|^2 = \int_{z_0}^{z_1} dz \; \sigma(\mathbf{z})^2 \qquad \text{(bad definition)} \quad , \tag{244}$$

would not be consistent with that of the norm $\| \boldsymbol{\rho} \|$ (we do not have, in general, any reason to assume that $\sigma(z)$ sould be 'more $L_2$ ' than $\rho(z)$ , or vice-versa). This is a typical example where the logarithmic variables $r = \log \rho / \rho_0$ and $s = \log \sigma / \sigma_0$ (where $\rho_0$ and $\sigma_0$ are arbitrary constants) allow the only sensible definition of norm

$$\| \mathbf{r} \|^2 = \| \mathbf{s} \|^2 = \int_{z_0}^{z_1} dz \; r(\mathbf{z})^2 = \int_{z_0}^{z_1} dz \; s(\mathbf{z})^2 \qquad \text{(good definition)} \quad , \tag{245}$$

or, in terms of $\rho$ and $\sigma$ ,

$$\| \boldsymbol{\rho} \|^2 = \| \boldsymbol{\sigma} \|^2 = \int_{z_0}^{z_1} dz \; \left( \log \frac{\rho(z)}{\rho_0} \right)^2 = \int_{z_0}^{z_1} dz \; \left( \log \frac{\sigma(z)}{\sigma_0} \right)^2 \qquad \text{(good definition)} \quad , \tag{246}$$

We see that the right functional spaces for the resistivity $\rho(z)$ or the conductivity $\sigma(z)$ is not $L_2$ , but, to speak grossly, *the exponential of* $L_2$ .

Although these examples concern the $L_2$ norm, the same comments apply to any $L_p$ norm. We will see below an example with the $L_1$ norm.

## M.3 Duality Product

Every time we define a functional space, and we start developing mathematical properties (for instance, analyzing the existence and unicity of solutions to partial differential equations), we face another function space, with the same degrees of freedom.

For instance, in elastic theory we may define the strain field $\boldsymbol{\varepsilon} = \{ \varepsilon^{ij}(\mathbf{x}, t) \}$ . It will automatically appear another field, with the same variables (degrees of freedom) that, in this case, is the stress $\boldsymbol{\sigma} = \{ \sigma_{ij}(\mathbf{x}, t) \}$ . The

'contacted multiplication' will consist in making the sum (over discrete indices) and the integral (over continuous variables) of the product of the two fields, as in

$$\langle\ \boldsymbol{\sigma}\ ,\ \boldsymbol{\varepsilon}\ \rangle = \int dt \int dV(\mathbf{x})\ \sigma_{ij}(\mathbf{x},t)\,\varepsilon^{ij}(\mathbf{x},t)\ , \tag{247}$$

where the sum over $i,j$ is implicitly notated.

The space of strains and the space of stresses is just one example of *dual spaces*. When one space is called 'the primal space', the other one is calles 'the dual space', but this is just a matter of convention.

The product 247 is one example of *duality product*, where one element of the primal space and one element of the dual space are 'mutiplied' to form a scalar (that may be a real number or that may have physical dimensions). This implies the sum or the integral over the variables of the functions. Mathematicians say that 'the dual of an space $\mathcal{X}$ is the space of all linear forms over $\mathcal{X}$ '. It is true that a given $\boldsymbol{\sigma}$ associates, to any $\boldsymbol{\varepsilon}$ , the number defined by equation 247; and that this association defines a linear application. But this rough definition of duality doesn't help readers to understand the actual mathematical structure.

## M.4 Scalar Product in $L_2$ Spaces

When we consider a functional space, its dual appears spontaneously, and we can say that any space is *always* accompanied by its dual space (as in the example strain-stress seen above). Then, the duality product is always defined.

Things are completely different with the scalar product, that it is only defined *sometimes*.

If, for instance, we consider functions $\mathbf{f} = \{f(x)\}$ belonging to a space $\mathcal{F}$ , the scalar product is a bilinear form that associates, to any pair of elements $\mathbf{f}_1$ and $\mathbf{f}_2$ of $\mathcal{F}$ , a number[36] denoted $(\ \mathbf{f}_1\ ,\ \mathbf{f}_2\ )$ .

Practically, to define a scalar product over a space $\mathcal{F}$ , we must first define a symmetric, positive definite operator $\mathbf{C}^{-1}$ mapping $\mathcal{F}$ into its dual, $\widehat{\mathcal{F}}$ . The dual of a function $\mathbf{f} = \{f(x)\}$ , that we may denote $\widehat{\mathbf{f}} = \{\widehat{f}(x)\}$ , is then

$$\widehat{\mathbf{f}} = \mathbf{C}^{-1}\,\mathbf{f}\ . \tag{248}$$

The scalar product of two elements $\mathbf{f}_1$ and $\mathbf{f}_2$ of $\mathcal{F}$ is then defined as

$$(\ \mathbf{f}_1\ ,\ \mathbf{f}_2\ ) = \langle\ \widehat{\mathbf{f}}_1\ ,\ \mathbf{f}_2\ \rangle = \langle\ \mathbf{C}^{-1}\,\mathbf{f}_1\ ,\ \mathbf{f}_2\ \rangle \tag{249}$$

In the context of an infinite-dimensional Gaussian process, some mean and some covariance are always defined. If, for instance, we consider functions $\mathbf{f} = \{f(x)\}$ , the mean function may be denoted $\mathbf{f}_0 = \{f_0(x)\}$ and the covariance function (the kernel of the covariance operator) may be denoted $\mathbf{C} = \{C(x,x')\}$ . The space of functions we work with, say $\mathcal{F}$ , is the set of all the possible random realization of such a Gaussian process with the given mean and the given covariance. The dual of $\mathcal{F}$ can be here identified with the image of $\mathcal{F}$ under $\mathbf{C}^{-1}$ , the inverse of the covariance operator (that is a symmetric, positive definite operator). So, denoting $\widehat{\mathcal{F}}$ the dual of $\mathcal{F}$ , we can formally write $\widehat{\mathcal{F}} = \mathbf{C}^{-1}\,\mathcal{F}$ o, equivalently, $\mathcal{F} = \mathbf{C}\,\widehat{\mathcal{F}}$ . The explicit expression of the equation

$$\mathbf{f} = \mathbf{C}\,\widehat{\mathbf{f}} \tag{250}$$

is

$$f(x) = \int dx'\ C(x,x')\,\widehat{f}(x)\ . \tag{251}$$

Let us denote $\mathbf{W}$ the inverse of the covariance operator,

$$\mathbf{W} = \mathbf{C}^{-1}\ , \tag{252}$$

that is usually named the *weight operator*. As $\mathbf{C}\,\mathbf{W} = \mathbf{W}\,\mathbf{C} = \mathbf{I}$ , its kernel, $W(x,x')$ , the *weight function*, satisfyes

$$\int dx' C(x,x')\,W(x',x'') = \int dx' W(x,x')\,C(x',x'') = \delta(x-x'')\ , \tag{253}$$

where $\delta(\,\cdot\,)$ is the Dirac's delta 'function'. Typically, the covariance function $C(x,x')$ is a smooth function; then, the weight function $W(x,x')$ is a distribution (sum of Dirac delta 'functions' and its derivatives).

[36] It is usually a real number, but it may have physical dimensions.

Equations 250–251 can equivalently be written

$$\widehat{\mathbf{f}} = \mathbf{W}\,\mathbf{f} \tag{254}$$

and

$$\widehat{f}(x) = \int dx'\; W(x,x')\, f(x)\;. \tag{255}$$

If the duality product between $\widehat{\mathbf{f}}_1$ and $\mathbf{f}_2$ is written

$$\langle\; \widehat{\mathbf{f}}_1\;,\; \mathbf{f}_2\;\rangle = \int dx \widehat{f}_1(x)\, f_2(x)\;, \tag{256}$$

the scalar product, as defined by equation 249, becomes

$$\begin{aligned}
(\; \mathbf{f}_1\;,\; \mathbf{f}_2\;) &= \langle\; \widehat{\mathbf{f}}_1\;,\; \mathbf{f}_2\;\rangle = \langle\; \mathbf{C}^{-1}\,\mathbf{f}_1\;,\; \mathbf{f}_2\;\rangle = \langle\; \mathbf{W}\,\mathbf{f}_1\;,\; \mathbf{f}_2\;\rangle \\
&= \int dx \left( \int dx'\; W(x,x') f_1(x') \right)\; f_2(x) \\
&= \int dx \int dx'\; f_1(x)\, W(x,x')\, f_2(x')\;.
\end{aligned} \tag{257}$$

The norm of $\mathbf{f}$, denoted $\|\;\mathbf{f}\;\|$ and defined as

$$\|\;\mathbf{f}\;\|^2 = (\;\mathbf{f}\;,\;\mathbf{f}\;)\;, \tag{258}$$

is expressed, in this example, as

$$\|\;\mathbf{f}\;\|^2 = \int dx \int dx'\; f(x)\, W(x,x')\, f(x')\;. \tag{259}$$

This is the $L_2$ norm of the function $f(x)$ (the case where $W(x,x') = \delta(x-x')$ being a very special case).

One final remark. If $\widehat{f}(x)$ is a random realization of a Gaussian white noise with zero mean, then, the function $f(x)$ defined by equation 251 is a random realization of a Gaussian process with zero mean and covariance function $C(x,x')$. This means that if the space $\mathcal{F}$ is the space of all the random realizations of a Gaussian process with covariance operator $\mathbf{C}$, then, its dual, $\widehat{\mathcal{F}}$, is the space of all the realizations of a Gaussian white noise.

**Example 28** *Consider the covariance operator* $\mathbf{C}$ *, with covariance function* $C(x,x')$ *,*

$$\mathbf{f} = \mathbf{C}\,\widehat{\mathbf{f}} \qquad \Longleftrightarrow \qquad f(x) = \int_{-\infty}^{+\infty} dx\; C(x,x')\,\widehat{f}(x') \quad , \tag{260}$$

*in the special case where the covariance function is the exponential function,*

$$C(x,x') = \sigma^2\;\exp\left(-\frac{|x-x'|}{X}\right)\;, \tag{261}$$

*where* $X$ *is a constant. The results of this example are a special case of those demonstrated in Tarantola (1987, page 572). The inverse covariance operator is*

$$\widehat{\mathbf{f}} = \mathbf{C}^{-1}\,\mathbf{f} \qquad \Longleftrightarrow \qquad \widehat{f}(t) = \frac{1}{2\,\sigma^2}\left(\frac{1}{X}\;f(x) - X\;\ddot{f}(x)\right) \quad , \tag{262}$$

*where the double dot means second derivative. As noted above, if* $f(x)$ *is a random realization of a Gaussian process having the exponential covariance function considered here, then, the* $\widehat{f}(x)$ *given by this equation is a random realization of a white noise. Formally, this means that the weighting function (kernel of* $\mathbf{C}^{-1}$ *) is* $W(x,x') = \frac{1}{2\,\sigma^2}\left(\frac{1}{X}\;\delta(x) - X\;\ddot{\delta}(x)\right)$ *. The squared norm of a function* $f(x)$ *is obtained integrating by parts:*

$$\|\;\mathbf{f}\;\|^2 = \langle\;\widehat{\mathbf{f}}\;,\;\mathbf{f}\;\rangle = \frac{1}{2\,\sigma^2}\left(\frac{1}{X}\;\int_{\infty}^{+\infty} dx\; f^2(x) + X\;\int_{-\infty}^{+\infty} dx\;\dot{f}^2(x)\right)\;. \tag{263}$$

*This is the usual norm in the so-called Sobolev space* $H^1$ *.* [END OF EXAMPLE.]

## M.5 The Transposed Operator

Let $\mathbf{G}$ a linear operator mapping an space $\mathcal{E}$ into an space $\mathcal{F}$ (we have in mind functional spaces, but the definition is general). We denote, as usual

$$\mathbf{G} \ : \ \mathcal{E} \to \mathcal{F} \ . \tag{264}$$

If $\mathbf{e} \in \mathcal{E}$ and $\mathbf{f} \in \mathcal{F}$, then we write

$$\mathbf{f} = \mathbf{G}\,\mathbf{e} \ . \tag{265}$$

Let $\widehat{\mathcal{E}}$ and $\widehat{\mathcal{F}}$ be the respective duals of $\mathcal{E}$ and $\mathcal{F}$, and denote $\langle \ \cdot \ , \ \cdot \ \rangle_{\mathcal{E}}$ and $\langle \ \cdot \ , \ \cdot \ \rangle_{\mathcal{F}}$ the respective duality products. A linear operator $\mathbf{H}$ mapping the dual of $\mathcal{F}$ into the dual of $\mathcal{E}$, is named the transpose of $\mathbf{G}$ if for any $\widehat{\mathbf{f}} \in \widehat{\mathcal{F}}$ and for any $\mathbf{e} \in \mathcal{E}$ we have $\langle \ \widehat{\mathbf{f}} \ , \ \mathbf{G}\,\mathbf{e} \ \rangle_{\mathcal{F}} = \langle \ \mathbf{H}\widehat{\mathbf{f}} \ , \ \mathbf{e} \ \rangle_{\mathcal{E}}$, and, in this case, we use the notation $\mathbf{H} = \mathbf{G}^T$. The whole definition then reads

$$\mathbf{G}^T \ : \ \widehat{\mathcal{F}} \to \widehat{\mathcal{E}} \tag{266}$$

$$\forall \mathbf{e} \in \mathcal{E} \quad ; \quad \forall \widehat{\mathbf{f}} \in \widehat{\mathcal{F}} \quad : \qquad \langle \ \widehat{\mathbf{f}} \ , \ \mathbf{G}\,\mathbf{e} \ \rangle_{\mathcal{F}} = \langle \ \mathbf{G}^T \ \widehat{\mathbf{f}} \ , \ \mathbf{e} \ \rangle_{\mathcal{E}} \ . \tag{267}$$

**Example 29 The Transposed of a Matrix.** *Let us consider a discrete situation where*

$$\mathbf{f} = \mathbf{G}\,\mathbf{e} \qquad \Longleftrightarrow \qquad f_i = \sum_\alpha G_{i\alpha}\, e_\alpha \ . \tag{268}$$

*In this circumstance, the duality products in each space will read*

$$\langle \ \widehat{\mathbf{f}} \ , \ \mathbf{f} \ \rangle_{\mathcal{F}} = \sum_i \widehat{f}_i \, f_i \qquad ; \qquad \langle \ \widehat{\mathbf{e}} \ , \ \mathbf{e} \ \rangle_{\mathcal{E}} = \sum_\alpha \widehat{e}_\alpha \, e_\alpha \quad . \tag{269}$$

*The linear operator* $\mathbf{H}$ *is the transposed of* $\mathbf{G}$ *if for any* $\widehat{\mathbf{f}}$ *and for any* $\mathbf{e}$ *(equation 267),*

$$\langle \ \widehat{\mathbf{f}} \ , \ \mathbf{G}\,\mathbf{e} \ \rangle_{\mathcal{F}} = \langle \ \mathbf{H}\ \widehat{\mathbf{f}} \ , \ \mathbf{e} \ \rangle_{\mathcal{E}} \ , \tag{270}$$

*i.e., if*

$$\sum_i \widehat{f}_i \, (\mathbf{G}\,\mathbf{e})_i = \sum_\alpha (\mathbf{H}\,\widehat{\mathbf{f}})_\alpha \, e_\alpha \tag{271}$$

*or, explicitly,*

$$\sum_i \widehat{f}_i \left( \sum_\alpha G_{i\alpha}\, e_\alpha \right) = \sum_\alpha \left( \sum_i H_{\alpha i}\, \widehat{f}_i \right) \, e_\alpha \ . \tag{272}$$

*The condition can be written*

$$\sum_i \sum_\alpha \widehat{f}_i \, G_{i\alpha}\, e_\alpha = \sum_i \sum_\alpha \widehat{f}_i \, H_{\alpha i}\, e_\alpha \ , \tag{273}$$

*and it is clear that this true for any* $\widehat{\mathbf{f}}$ *and for any* $\mathbf{e}$ *iff*

$$H_{\alpha i} = G_{i\alpha} \ , \tag{274}$$

*i.e., if the matrix representing* $\mathbf{H}$ *is the transposed (in the elementary matricial sense) of the matrix representing* $\mathbf{G}$ *:*

$$\mathbf{H} = \mathbf{G}^T \ . \tag{275}$$

*This demonstrates that the abstract definition given above of the transpose of a linear operator is consistent with the matricial notion of transpose.* [END OF EXAMPLE.]

**Example 30 The Transposed of the Derivative Operator.** *Let us consider a situation where*

$$\mathbf{v} = \mathbf{D}\,\mathbf{x} \qquad \Longleftrightarrow \qquad v(t) = \frac{dx}{dt}(t) \ , \tag{276}$$

*i.e., where the linear operator* $\mathbf{D}$ *is the* derivative operator*. In this circumstance, the duality products in each space will typically read*

$$\langle \ \widehat{\mathbf{v}} \ , \ \mathbf{v} \ \rangle_{\mathcal{V}} = \int_{t_1}^{t_2} dt \ \widehat{v}(t) \ v(t) \qquad ; \qquad \langle \ \widehat{\mathbf{x}} \ , \ \mathbf{x} \ \rangle_{\mathcal{X}} = \int_{t_1}^{t_2} dt \ \widehat{x}(t) \ x(t) \quad . \tag{277}$$

*If the linear operator* $\mathbf{D}^T$ *has to be the transposed of* $\mathbf{D}$ *, for any* $\widehat{\mathbf{v}}$ *and for any* $\mathbf{x}$ *we mst have (equation 267)*

$$\langle\ \widehat{\mathbf{v}}\ ,\ \mathbf{D}\,\mathbf{x}\ \rangle_{\mathcal{V}} = \langle\ \mathbf{D}^T\ \widehat{\mathbf{v}}\ ,\ \mathbf{x}\ \rangle_{\mathcal{X}}\ . \tag{278}$$

[END OF EXAMPLE.]

Let us demonstrate that *the derivative operator is an antisymmetric operator* i.e, that

$$\mathbf{D}^T = -\mathbf{D}\ . \tag{279}$$

To demonstrate this, we will need to make a restrictive condition, interesting to analyze.

Using 279, equation 278 writes

$$\int_{t_1}^{t_2} dt\ \widehat{v}(t)\ (\mathbf{D}\,\mathbf{x})(t) = -\int_{t_1}^{t_2} dt\ (\mathbf{D}\,\widehat{\mathbf{v}})(t)\ x(t) \tag{280}$$

i.e.,

$$\int_{t_1}^{t_2} dt\ \widehat{v}(t)\ \frac{dx}{dt}(t) + \int_{t_1}^{t_2} dt\ \frac{d\widehat{v}}{dt}(t)\ x(t) = 0\ . \tag{281}$$

We have to check if this equation holds for any $x(t)$ and any $v(t)$ .

The condition is equivalent to

$$\int_{t_1}^{t_2} dt\ \left(\widehat{v}(t)\ \frac{dx}{dt}(t) + \frac{d\widehat{v}}{dt}(t)\ x(t)\right) = 0\ , \tag{282}$$

i.e., to

$$\int_{t_1}^{t_2} dt\ \frac{d}{dt}\left(\widehat{v}(t)\ x(t)\right) = 0\ , \tag{283}$$

or, using the elementary properties of the integral, to

$$\widehat{v}(t_2)\ x(t_2) + \widehat{v}(t_1)\ x(t_1) = 0\ . \tag{284}$$

In general, there is no reason for this being true. So, in general, we can not say that $\mathbf{D}^T = -\mathbf{D}$ .

If the spaces of functions we work with (here, the space of functions $v(t)$ and the space of functions $x(t)$ ) satisfy the condition 284 it is said that the spaces satisfy *dual boundary conditions*. If the spaces satisfy dual boundary conditions, then it is true that $\mathbf{D}^T = -\mathbf{D}$ , i.e., that the derivative operator is antisymmetric.

A typical example of dual boundary conditions being satisfied is in the case where all the functions $x(t)$ vanish at the initial time, and all the functions $\widehat{v}(t)$ vanish at the final time:

$$x(t_1) = 0 \quad ; \qquad \widehat{v}(t_2) = 0 \quad . \tag{285}$$

The notation $\mathbf{D}^T = -\mathbf{D}$ is very suggestive. One has, nevertheless, to remember that (with the boundary conditions chose) while $\mathbf{D}$ acts on functions that vanish at the initial time, $\mathbf{D}^T$ acts on functions $\widehat{v}(t)$ that vanish at the final time.

Consider now the operator $\mathbf{D}^2$ (second derivative)

$$\gamma(t) = \frac{dx^2}{dt^2}(t)\ . \tag{286}$$

Following the same lines of reasoning as above, the reader may easily demonstrate that the second derivative operator is symmetrical, i.e., $(\mathbf{D}^2)^T = \mathbf{D}^2$ , provided that the functional spaces into consideration satisfy the dual doundary condition

$$\widehat{\gamma}(t_2)\,\frac{dx}{dt}(t_2) - \frac{d\widehat{\gamma}}{dt}(t_2)\,x(t_2) = \widehat{\gamma}(t_1)\,\frac{dx}{dt}(t_1) - \frac{d\widehat{\gamma}}{dt}(t_1)\,x(t_1)\ . \tag{287}$$

A typical example where this condition is satisfied is when we have

$$x(t_1) = 0 \qquad ; \qquad \frac{dx}{dt}(t_1) = 0 \qquad ; \qquad \widehat{\gamma}(t_2) = 0 \qquad ; \qquad \frac{d\widehat{\gamma}}{dt}(t_2) = 0 \quad , \tag{288}$$

i.e., when the functions $x(t)$ have zero value and zero derivative value at the initial time and the functions $\widehat{\gamma}(t)$ have zero value and zero derivative value at the final time.

This is the sort of boundary conditions found when working with the wave equation, as it contains second order time derivatives. Further details are given in section M.7 below.

As an exercise, the reader may try to understand why the quite obvious property

$$\left(\frac{\partial}{\partial x^i}\right)^T = -\left(\frac{\partial}{\partial x^i}\right) \tag{289}$$

corresponds, in fact, to the properties

$$\mathbf{grad}^T = -\mathrm{div} \qquad ; \qquad \mathrm{div}^T = -\mathbf{grad} \tag{290}$$

(hint: if an operator maps $\mathcal{E}$ into $\mathcal{F}$, its transpose maps $\widehat{\mathcal{F}}$ into $\widehat{\mathcal{E}}$; the dual of an space has the same 'variables' as the original space).

Let us formally demostrate that the operator representing the acoustic wave equation is symmetric. Starting from[37]

$$\mathbf{L} = \frac{1}{\kappa(\mathbf{x})}\frac{\partial^2}{\partial t^2} - \mathrm{div}\frac{1}{\rho(\mathbf{x})}\,\mathbf{grad}\ , \tag{291}$$

we have

$$\begin{aligned} \mathbf{L}^T &= \left(\frac{1}{\kappa(\mathbf{x})}\frac{\partial^2}{\partial t^2} - \mathrm{div}\frac{1}{\rho(\mathbf{x})}\,\mathbf{grad}\right)^T \\ &= \left(\frac{1}{\kappa(\mathbf{x})}\frac{\partial^2}{\partial t^2}\right)^T - \left(\mathrm{div}\frac{1}{\rho(\mathbf{x})}\,\mathbf{grad}\right)^T\ . \end{aligned} \tag{292}$$

Using the property $(\mathbf{A}\,\mathbf{B})^T = \mathbf{B}^T\,\mathbf{A}^T$, we arrive at

$$\mathbf{L}^T = \left(\frac{\partial^2}{\partial t^2}\right)^T \left(\frac{1}{\kappa(\mathbf{x})}\right)^T - (\mathbf{grad})^T \left(\frac{1}{\rho(\mathbf{x})}\right)^T (\mathrm{div})^T\ . \tag{293}$$

Now, (i) the transposed of a scalar is the scalar itself; (ii) the second derivative (as we have seen) is a symmetric operator; (iii) we have (as it has been mentioned above) $\mathbf{grad}^T = -\mathrm{div}$ and $\mathrm{div}^T = -\mathbf{grad}$. We then have

$$\mathbf{L}^T = \frac{\partial^2}{\partial t^2}\frac{1}{\kappa(\mathbf{x})} - \mathrm{div}\frac{1}{\rho(\mathbf{x})}\,\mathbf{grad}\ , \tag{294}$$

and, as the uncompressibility $\kappa$ is assumed to be independent on time,

$$\mathbf{L}^T = \frac{1}{\kappa(\mathbf{x})}\frac{\partial^2}{\partial t^2} - \mathrm{div}\frac{1}{\rho(\mathbf{x})}\,\mathbf{grad} = \mathbf{L}\ , \tag{295}$$

and we see that the acoustic wave operator is symmetric. As we have seen above, this conclusion has to be understood with the condition that the wavefields $p(\mathbf{x},t)$ on which acts $\mathbf{L}$ satisfy boundary conditions that are dual with those satisfied by the fields $\widehat{p}(\mathbf{x},t)$ on which acts $\mathbf{L}^T$. Typically the fields $p(\mathbf{x},t)$ satisfy initial conditions of rest, and the fields $\widehat{p}(\mathbf{x},t)$ satisfy final conditions of rest.

Tarantola (1988) demostrates that the transposed of the operator corresponding to the 'wave equation with attenuation' corresponds to the wave equation with 'anti-attenuation'. But it has to be understood that any physical or numerical implementation of the operator $\mathbf{L}^T$ is made 'backwards in time', so, in that sense of time, we face an ondinary attenuation: there is no difficulty in the implementation of $\mathbf{L}^T$.

**Example 31 The Kernel of the Transposed Operator** *If the explicit expression of the equation*

$$\mathbf{f} = \mathbf{G}\,\mathbf{e} \tag{296}$$

[37] Here, and below, an expression like $\mathbf{A}\,\mathbf{B}\,\mathbf{C}$, means, as usual, $\mathbf{A}(\mathbf{B}\,\mathbf{C})$. This means, for instance, that the div operator in this equation is to be understood as being applied not to $1/\rho(\mathbf{x})$ only, but to 'everything at its right'.

*is*

$$f(t) = \int dt\ G(t,x)\, e(t)\ , \tag{297}$$

*where* $G(t,x)$ *is an ordinary function*[38]*, then, it is said that* $\mathbf{G}$ *is an* integral operator*, and that the function* $G(t,x)$ *is its kernel.* [END OF EXAMPLE.]

The transpose of $\mathbf{G}$ will map an element $\widehat{\mathbf{f}}$ into an element $\widehat{\mathbf{e}}$, these two elements belonging to the respective duals of the spaces where the elements $\mathbf{e}$ and $\mathbf{f}$ mentioned in equation 296 belong. An equation like

$$\widehat{\mathbf{e}} = \mathbf{G}^T \widehat{\mathbf{f}} \tag{298}$$

will correspond, explicitly, to

$$\widehat{e}(t) = \int dx\ G^T(x,t)\, \widehat{f}(t)\ . \tag{299}$$

The reader may easily verify that the definition of transpose operator imposes that the kernel of $\mathbf{G}^T$ is related to the kernel of $\mathbf{G}$ by the simple expression

$$G^T(x,t) = G(t,x)\ . \tag{300}$$

We see that the kernels of $\mathbf{G}$ and of $\mathbf{G}^T$ are, in fact, identical, via a simple 'transposition' of the variables.

## M.6 The Adjoint Operator

Let $\mathbf{G}$ be a linear operator mapping an space $\mathcal{E}$ into an space $\mathcal{F}$:

$$\mathbf{G}\ :\ \mathcal{E} \to \mathcal{F}\ . \tag{301}$$

If $\mathbf{e} \in \mathcal{E}$ and $\mathbf{f} \in \mathcal{F}$, then we write

$$\mathbf{f} = \mathbf{G}\,\mathbf{e}\ . \tag{302}$$

Assume that both, $\mathcal{E}$ and $\mathcal{F}$ are furnished with an scalar product each (see section M.4), that we denote, respectively, as $(\ \mathbf{e}_1\ ,\ \mathbf{e}_2\ )_{\mathcal{E}}$ and $(\ \mathbf{f}_1\ ,\ \mathbf{f}_2\ )_{\mathcal{F}}$

A linear operator $\mathbf{H}$ mapping $\mathcal{F}$ into $\mathcal{E}$, is named the adjoint of $\mathbf{G}$ if for any $\mathbf{f} \in \mathcal{F}$ and for any $\mathbf{e} \in \mathcal{E}$ we have $(\ \mathbf{f}\ ,\ \mathbf{G}\,\mathbf{e}\ )_{\mathcal{F}} = (\ \mathbf{H}\,\mathbf{f}\ ,\ \mathbf{e}\ )_{\mathcal{E}}$, and, in this case, we use the notation $\mathbf{H} = \mathbf{G}^*$. The whole definition then reads

$$\mathbf{G}^*\ :\ \mathcal{F} \to \mathcal{E} \tag{303}$$

$$\forall \mathbf{e} \in \mathcal{E}\quad ;\quad \forall \mathbf{f} \in \mathcal{F}\quad :\qquad (\ \mathbf{f}\ ,\ \mathbf{G}\,\mathbf{e}\ )_{\mathcal{F}} = (\ \mathbf{G}^*\,\mathbf{f}\ ,\ \mathbf{e}\ )_{\mathcal{E}}\ . \tag{304}$$

Let $\widehat{\mathcal{E}}$ and $\widehat{\mathcal{F}}$ be the respective duals of $\mathcal{E}$ and $\mathcal{F}$, and denote $\langle\ \cdot\ ,\ \cdot\ \rangle_{\mathcal{E}}$ and $\langle\ \cdot\ ,\ \cdot\ \rangle_{\mathcal{F}}$ the respective duality products. We have seen above that a scalar product is defined through a symmetric, positive operator mapping a space into its dual. Then, as $\mathcal{E}$ and $\mathcal{F}$ are assumed to have a scalar product defined, there are two 'covariance' operators $\mathbf{C}_{\mathcal{E}}$ and $\mathbf{C}_{\mathcal{F}}$ such that the respective scalar products are given by

$$\begin{aligned}(\ \mathbf{e}_1\ ,\ \mathbf{e}_2\ )_{\mathcal{E}} &= \langle\ \mathbf{C}_{\mathcal{E}}^{-1}\,\mathbf{e}_2\ ,\ \mathbf{e}_1\ \rangle_{\mathcal{E}} \\ (\ \mathbf{f}_1\ ,\ \mathbf{f}_2\ )_{\mathcal{F}} &= \langle\ \mathbf{C}_{\mathcal{F}}^{-1}\,\mathbf{f}_2\ ,\ \mathbf{f}_1\ \rangle_{\mathcal{F}}\ .\end{aligned} \tag{305}$$

Then, equation 304 writes $\langle\ \mathbf{C}_{\mathcal{F}}^{-1}\,\mathbf{f}\ ,\ \mathbf{G}\,\mathbf{e}\ \rangle_{\mathcal{F}} = \langle\ \mathbf{C}_{\mathcal{E}}^{-1}\,\mathbf{G}^*\,\mathbf{f}\ ,\ \mathbf{e}\ \rangle_{\mathcal{E}}$, or, denoting $\widehat{\mathbf{f}} = \mathbf{C}_{\mathcal{F}}^{-1}\,\mathbf{f}$,

$$\langle\ \widehat{\mathbf{f}}\ ,\ \mathbf{G}\,\mathbf{e}\ \rangle_{\mathcal{F}} = \langle\ \mathbf{C}_{\mathcal{E}}^{-1}\,\mathbf{G}^*\ \mathbf{C}_{\mathcal{F}}\,\widehat{\mathbf{f}}\ ,\ \mathbf{e}\ \rangle_{\mathcal{E}}\ . \tag{306}$$

The comparison with equation 304 defining the transposed operator gives the relation between adjoint and transpose, $\mathbf{G}^T = \mathbf{C}_{\mathcal{E}}^{-1}\,\mathbf{G}^*\,\mathbf{C}_{\mathcal{F}}$, that can be written, equivalently, as

$$\mathbf{G}^* = \mathbf{C}_{\mathcal{E}}\,\mathbf{G}^T\,\mathbf{C}_{\mathcal{F}}^{-1}\ . \tag{307}$$

The transposed operator is an elementary operator. Its definition only requires the existence of the dual of the considered spaces, that is automatic. If, for instance, a linear operator $\mathbf{G}$ has the kernel $G(u,v)$, the transposed operator $\mathbf{G}^T$ will have the kernel $G^T(v,u) = G(u,v)$.

The adjoint operator is not an elementary operator. Its definition requires the existence of scalar products in the working spaces, that are necessarily defoned through symmetric, positive definite operators. This means that (excepted degenerated cases) the adjoint operator is a complex object, depending on three elementary objects: this is how equation 307 is to be interpreted.

[38] If $G(t,x)$ is a distribution (like the derivative of a Dirac's delta) then equation 296 may be a disguised expression for a differential operator.

## M.7 The Green Operator

The pressure field $p(\mathbf{x},t)$ propagating in an elastic medium with uncompressibility modulus $\kappa(\mathbf{x})$ and volumetric mass $\rho(\mathbf{x})$ satisfies the 'acoustic wave equation'

$$\frac{1}{\kappa(\mathbf{x})}\,\frac{\partial^2 p}{\partial t^2}(\mathbf{x},t) - \mathrm{div}\left(\frac{1}{\rho(\mathbf{x})}\,\mathbf{grad}\,p(\mathbf{x},t)\right) = S(\mathbf{x},t)\ . \tag{308}$$

Here, $\mathbf{x}$ denotes a point inside the medium (the coordinate system being still unspecified), $t$ is the Newtonian time, and $S(\mathbf{x},t)$ is a source function. To simplify the notations, the variables $\mathbf{x}$ and $t$ will be dropped when there is no risk of confusion. For instance, the equation above will be written

$$\frac{1}{\kappa}\,\frac{\partial^2 p}{\partial t^2} - \mathrm{div}\left(\frac{1}{\rho}\,\mathbf{grad}\,p\right) = S\ . \tag{309}$$

Also, I shall denote $\mathbf{p}$ the function $\{p(\mathbf{x},t)\}$ as a whole, and not its value at a given point of space and time. Similarly, $\mathbf{S}$ shall denote the source function $S(\mathbf{x},t)$ .

For fixed $\kappa(\mathbf{x})$ and $\rho(\mathbf{x})$ , the wave equation above can be written, for short,

$$\mathbf{L}\,\mathbf{p} = \mathbf{S}\ , \tag{310}$$

where $\mathbf{L}$ is the second order differential operator defined through equation 309. In order to define an unique wavefield $\mathbf{p}$ , we have to prescribe some boundary and initial conditions. An example of those are, if we work inside the time interval $(t_1,t_2)$ , and inside a volume $V$ bounded by the surface $S$ ,

$$\begin{array}{rcl} p(\mathbf{x},t_1) = 0 & ; & \mathbf{x}\in V \\ \dot{p}(\mathbf{x},t_1) = 0 & ; & \mathbf{x}\in V \\ p(\mathbf{x},t) = 0 & ; & \mathbf{x}\in S\ ;\ t\in(t_1,t_2) \quad . \end{array} \tag{311}$$

Here, a dot means time derivative. With prescribed initial and boundary conditions, then, there is an one to one correspondence between the source field $\mathbf{S}$ and the wavefield $\mathbf{p}$ . The inverse of the wave equation operator, $\mathbf{L}^{-1}$ , is called the *Green operator*, and is denoted $\mathbf{G}$ :

$$\mathbf{G} = \mathbf{L}^{-1}\ . \tag{312}$$

We can then write

$$\mathbf{L}\,\mathbf{p} = \mathbf{S} \qquad \Longleftrightarrow \qquad \mathbf{p} = \mathbf{G}\,\mathbf{S} \quad . \tag{313}$$

As $\mathbf{L}$ is a differential operator, its inverse $\mathbf{G}$ is an integral operator. The kernel of the Green operator is named the *Green function*, and is usually denoted $G(\mathbf{x},t;\mathbf{x}',t')$ . The explicit expression for $\mathbf{p} = \mathbf{G}\,\mathbf{S}$ is then

$$p(\mathbf{x},t) = \int_V dV(\mathbf{x}') \int_{t_1}^{t_2} dt'\ G(\mathbf{x},t;\mathbf{x}',t')\ S(\mathbf{x}',t')\ . \tag{314}$$

It is easy to demonstrate[39] that the wave equation operator $\mathbf{L}$ is a symmetric operator, so this is also true for the Green operator $\mathbf{G}$ . But we have seen that the transpose operators work in spaces with have *dual boundary conditions* (see section 30 above).

Using the method outlined in section 30, the boundary conditions dual to those in equations 311 are

$$\begin{array}{rcl} p(\mathbf{x},t_2) = 0 & ; & \mathbf{x}\in V \\ \dot{p}(\mathbf{x},t_2) = 0 & ; & \mathbf{x}\in V \\ p(\mathbf{x},t) = 0 & ; & \mathbf{x}\in S\ ;\ t\in(t_1,t_2) \quad , \end{array} \tag{315}$$

i.e., we have *final* conditions of rest instead of initial conditions of rest (and the same surface condition). We have to underdstand that while the equation $\mathbf{L}\,\mathbf{p} = \mathbf{S}$ is associated to the boundary conditions 311, equations like

$$\mathbf{L}^T\,\widehat{\mathbf{p}} = \widehat{\mathbf{S}} \qquad ; \qquad \widehat{\mathbf{p}} = \mathbf{G}^T\,\widehat{\mathbf{S}} \tag{316}$$

[39] This comes from the property that the derivative operator is antisymmetric, (so that the second derivative is a symmetric operator) and from the properties $\mathbf{grad}^T = -\mathrm{div}$ and $\mathrm{div}^T = -\mathbf{grad}$ , mentioned in section protect30.

are associated to the dual boundary conditions 315 (the hats here mean that the transpose operator operator operates in the dual spaces (see section M.3). This being understood, we can write $\mathbf{L}^T = \mathbf{L}$ and $\mathbf{G}^T = \mathbf{G}$ , and rewrite equations 316 as

$$\mathbf{L}\,\widehat{\mathbf{p}} = \widehat{\mathbf{S}} \qquad ; \qquad \widehat{\mathbf{p}} = \mathbf{G}\,\widehat{\mathbf{S}} \quad . \tag{317}$$

The hats have to be maintained, to remember that the fields with a hat must satisfy boundary conditions dual to those satisfied by the fields without a hat.

Using the transposed of the Green operator, we can write

$$\widehat{p}(\mathbf{x},t) = \int dV(\mathbf{x}') \int_{t_2}^{t_1} dt'\; G^T(\mathbf{x},t;\mathbf{x}',t')\,\widehat{S}(\mathbf{x}',t') \;, \tag{318}$$

equation identical to

$$\widehat{p}(\mathbf{x},t) = \int dV(\mathbf{x}') \int_{t_1}^{t_2} dt'\; G(\mathbf{x}',t';\mathbf{x},t)\,\widehat{S}(\mathbf{x}',t') \;. \tag{319}$$

## M.8 Born Approximation for the Acoustic Wave Equation

Let us start from equation 309, using the same notations:

$$\frac{1}{\kappa}\,\frac{\partial^2 p}{\partial t^2} - \mathrm{div}\left(\frac{1}{\rho}\,\mathbf{grad}\,p\right) = S \;. \tag{320}$$

I shall denote $\mathbf{p}$ the function $\{p(\mathbf{x},t)\}$ as a whole, and not its value at a given point of space and time. Similarly, $\boldsymbol{\kappa}$ and $\boldsymbol{\rho}$ will denote the functions $\{\kappa(\mathbf{x})\}$ and $\{\rho(\mathbf{x})\}$ .

Given appropriate boundary and initial conditions, and given a source function, the acoustic wave equation defines an application $\{\boldsymbol{\kappa},\boldsymbol{\rho}\} \to \mathbf{p} = \psi(\boldsymbol{\kappa},\boldsymbol{\rho})$ , i.e., an application that associates to each medium $\{\boldsymbol{\kappa},\boldsymbol{\rho}\}$ the (unique) pressure field $\mathbf{p}$ that satisfies the wave equation (with given boundary and initial conditions).

Let $\mathbf{p}_0$ be the pressure field propagating in the medium defined by $\boldsymbol{\kappa}_0$ and $\boldsymbol{\rho}_0$ , i.e., $\mathbf{p}_0 = \psi(\boldsymbol{\kappa}_0,\boldsymbol{\rho}_0)$ , and let $\mathbf{p}$ be the pressure field propagating in the medium defined by $\boldsymbol{\kappa}$ and $\boldsymbol{\rho}$ , i.e., $\mathbf{p} = \psi(\boldsymbol{\kappa},\boldsymbol{\rho})$ . Clearly, if $\boldsymbol{\kappa}$ and $\boldsymbol{\rho}$ are close (in a sense to be defined) to $\boldsymbol{\kappa}_0$ and $\boldsymbol{\rho}_0$ , then, the wavefield $\mathbf{p}$ will be close to $\mathbf{p}_0$ .

Let us obtain an explicit expression for the first order approximation to $\mathbf{p}$ . This is known as the (first) Born approximation of the wavefield. Both $\kappa$ and $\rho$ could be perturbed, but I simplify the discussion here by considering only perturbations in the uncompressibility $\kappa$ . The reader may easily obtain the general case.

The pressure $P$ is, in thermodynamics, a positive quantity. When considering small variations around some 'ambient pressure' $P_0$ , we can define

$$p = P_0\,\log\frac{P}{P_0} \;. \tag{321}$$

For small pressure perturbations, we have

$$p \;=\; P_0\,\log\left(1 + \frac{(P-P_0)}{P_0}\right) \;\approx\; P - P_0 \;. \tag{322}$$

So defined, the *pressure perturbation* $p$ may take positive or negative values, corresponding to an elastic medium that is compressed or stretched. In the terminology of section 2, this is a Cartesian quantity.

The uncompressibility and the volumetric mass are positive, Jeffreys quantities.

In most texts, the difference $\mathbf{p}-\mathbf{p}_0$ is calculated as a function of the difference $\boldsymbol{\kappa}-\boldsymbol{\kappa}_0$ , but we have seen that this is not the right way, as the resulting approximation will depend on the fact that we are using uncompressibilty $\kappa(\mathbf{x})$ instead of compressibility $\gamma(\mathbf{x}) = 1/\kappa(\mathbf{x})$ .

At this point we may introduce the logarithmic parameters, and proceed trivially. The logarithmic uncompressibilities for the reference medium and for the perturbed medium are

$$\kappa_0^* = \log\frac{\kappa_0}{K} \qquad ; \qquad \kappa^* = \log\frac{\kappa}{K} \quad , \tag{323}$$

where $K$ and $R$ are arbitrary constants (having the right physical dimension). Reciprocally,

$$\kappa_0 = K\,\exp\kappa_0^* \qquad ; \qquad \kappa = K\,\exp\kappa^* \quad . \tag{324}$$

In particular, we have

$$\kappa = \kappa_0 \, \exp(\delta\kappa^*) \; , \tag{325}$$

where

$$\delta\kappa^* = \kappa^* - \kappa_0^* = \log \frac{\kappa}{\kappa_0} \; . \tag{326}$$

Note that we have here a perturbation $\delta\kappa^*$ of a logarithmic (Cartesian) quantity, not of the positive (Jeffreys) one. We also write

$$p = p_0 + \delta p \; . \tag{327}$$

The reference solution satisfies

$$\frac{1}{\kappa_0} \frac{\partial^2 p_0}{\partial t^2} - \mathrm{div} \left( \frac{1}{\rho_0} \, \mathbf{grad} \, p_0 \right) = S \; , \tag{328}$$

while the perturbed solution satisfies

$$\frac{1}{\kappa} \frac{\partial^2 p}{\partial t^2} - \mathrm{div} \left( \frac{1}{\rho_0} \, \mathbf{grad} \, p \right) = S \; . \tag{329}$$

In this equation, $\kappa$ can be replaced by the expression 325, and $p$ by the expression 327. Using then the first order approximation $\exp(-\delta\kappa^*) = 1 - \delta\kappa^*$ leads to

$$\left( \frac{1}{\kappa_0} - \frac{\delta\kappa^*}{\kappa_0} \right) \left( \frac{\partial^2 p_0}{\partial t^2} + \frac{\partial^2 \delta p}{\partial t^2} \right) - \mathrm{div} \left( \frac{1}{\rho_0} \left( \mathbf{grad} \, p_0 + \mathbf{grad} \delta p \right) \right) = S \; . \tag{330}$$

Some of the terms in this equation correspond to the terms in the reference equation 328, and can be simplified. Keeping only first order terms then leads to

$$\frac{1}{\kappa_0} \frac{\partial^2 \delta p}{\partial t^2} - \mathrm{div} \left( \frac{1}{\rho_0} \, \mathbf{grad} \, \delta p \right) = \frac{\delta\kappa^*}{\kappa_0} \frac{\partial^2 p_0}{\partial t^2} \; . \tag{331}$$

Explicitly, replacing $\delta p = p - p_0$ and $\delta\kappa^* = \log \kappa/\kappa_0$ , gives

$$\frac{1}{\kappa_0} \frac{\partial^2 (p - p_0)}{\partial t^2} - \mathrm{div} \left( \frac{1}{\rho_0} \, \mathbf{grad} \, (p - p_0) \right) = \frac{1}{\kappa_0} \, \log \frac{\kappa}{\kappa_0} \frac{\partial^2 p_0}{\partial t^2} \; . \tag{332}$$

This is the equation we were looking for. It says that the field $p - p_0$ satisfies the wave equation *with the unperturbed value of the uncompressibility* $\kappa_0$ , and is generated by the 'Born secondary source'

$$S_{\mathrm{Born}} = \frac{1}{\kappa_0} \, \log \frac{\kappa}{\kappa_0} \frac{\partial^2 p_0}{\partial t^2} \; . \tag{333}$$

Should we have made the development using the compressibility $\gamma = 1/\kappa$ instead of the uncompressibility, we would have arrived at the secondary source

$$S_{\mathrm{Born}} = \gamma_0 \, \log \frac{\gamma_0}{\gamma} \frac{\partial^2 p_0}{\partial t^2} \tag{334}$$

that is identical to the previous one.

The expression here obtained for the secondary source is not the usual one, as it depends on the *distance* $\log \kappa/\kappa_0$ and not on the difference $\kappa - \kappa_0$ . For an additive perturbation $\kappa = \kappa_0 + \delta\kappa$ of the positive parameter $\kappa$ would have lead to the Born secondary source

$$S_\kappa = \frac{\delta\kappa}{\kappa^2} \frac{\partial^2 p_0}{\partial t^2} = \frac{\kappa - \kappa_0}{\kappa^2} \frac{\partial^2 p_0}{\partial t^2} \tag{335}$$

while an additive perturbation $\gamma = \gamma_0 + \delta\gamma$ of the positive parameter $\gamma = 1/\kappa$ would have lead to the Born secondary source

$$S_\gamma = -\delta\gamma \frac{\partial^2 p_0}{\partial t^2} = (\gamma - \gamma_0) \frac{\partial^2 p_0}{\partial t^2} \; , \tag{336}$$

and these two sources are not identical. I mean here that they finite expression is not identical. Of course, in the limit for an infinitesimal perturbation they tend to be identical.

The approach followed here has two advantages. First, mathematical consistence, in the sense that the secondary source is defined independently of the quantities used to make the computation (covariance of the results). Second advantage, in a numerical computation, the perturbations may be small, but they are finite. 'Large contrasts' in the parameters may give, when inserting the differences in expressions 335 or 336 quite bad approximations, while the logarithmic expressions in the right Born source (equation 333 or 334) may remain good.

## M.9 Tangent Application of Data With Respect to Parameters

In the context of an inverse problem, assume that we observe the pressure field $p(\mathbf{x},t)$ at some points $\mathbf{x}_i$ inside the volume. The solution of the forward problem is obtained by solving the wave equation, or by using the Green's function. We are here interested in the tangent linear application. Let us write the first order perturbation $\delta p(\mathbf{x}_i,t)$ of the pressure wavefield produced when the logarithmic uncompressibility is perturbed by the amount $\delta\kappa^*(\mathbf{x})$ as (linear tangent application)

$$\delta\mathbf{p} = \mathbf{F}\,\delta\boldsymbol{\kappa}^* \ , \tag{337}$$

or, introducing the kernel of the Fréchet derivative $\mathbf{F}$ ,

$$\delta p(\mathbf{x}_i,t) = \int_V dV(\mathbf{x}')\ F(\mathbf{x}_i,t;\mathbf{x}')\,\delta\kappa^*(\mathbf{x}') \ . \tag{338}$$

Let us express the kernel $F(\mathbf{x}_i,t;\mathbf{x}')$ .

We have seen that a perturbation $\delta\kappa^*$ is equivalent, up to the first order, to have the secondary Born source (equation 331)

$$S_{\rm Born}(\mathbf{x},t) = \frac{\delta\kappa^*(\mathbf{x})}{\kappa_0(\mathbf{x})}\,\ddot{p}_0(\mathbf{x},t) \ . \tag{339}$$

Then, using the Green function,

$$\begin{aligned}\delta p(\mathbf{x}_i,t) &= \int_V dV(\mathbf{x}') \int_{t_2}^{t_1} dt'\ G(\mathbf{x}_i,t;\mathbf{x}',t')\,S_{\rm Born}(\mathbf{x}',t') \\ &= \int_V dV(\mathbf{x}') \int_{t_2}^{t_1} dt'\ G(\mathbf{x}_i,t;\mathbf{x}',t')\,\frac{\delta\kappa^*(\mathbf{x}')}{\kappa_0(\mathbf{x}')}\,\ddot{p}_0(\mathbf{x}',t') \ .\end{aligned} \tag{340}$$

The last expression can be rearranged into the form used in equation 338, this showing that $F(\mathbf{x}_i,t;\mathbf{x}',t')$ is given by

$$F(\mathbf{x}_i,t;\mathbf{x}') = \frac{1}{\kappa_0(\mathbf{x}')} \int_{t_2}^{t_1} dt'\ G(\mathbf{x}_i,t;\mathbf{x}',t')\,\ddot{p}_0(\mathbf{x}',t') \tag{341}$$

This is the kernel of the Fréchet derivative of the data with respect to the parameter $\kappa^*(\mathbf{x})$ .

## M.10 The Transpose of the Fréchet Derivative Just Computed

Now that we are able to understand the expression $\delta\mathbf{p} = \mathbf{F}\,\delta\boldsymbol{\kappa}^*$ , let us face the dual problem. Which is the meaning of an expresion like

$$\delta\widehat{\boldsymbol{\kappa}}^* = \mathbf{F}^T\,\delta\widehat{\mathbf{p}} \ ? \tag{342}$$

Denoting by $F^T(\mathbf{x}';\mathbf{x}_i,t)$ the kernel of $\mathbf{F}^T$ , such an expression writes

$$\delta\widehat{\kappa}(\mathbf{x}') = \sum_i \int_{t_2}^{t_1} dt\ F^T(\mathbf{x}';\mathbf{x}_i,t)\,\delta\widehat{p}(\mathbf{x}_i,t) \ , \tag{343}$$

but we know that the kernel of the transpose operator equals the kernel of the original operator, with variables transposed, so that we can write this equation as

$$\delta\widehat{\kappa}(\mathbf{x}') = \sum_i \int_{t_2}^{t_1} dt\ F(\mathbf{x}_i,t;\mathbf{x}')\,\delta\widehat{p}(\mathbf{x}_i,t) \ , \tag{344}$$

where $F(\mathbf{x}_i,t;\mathbf{x}')$ is the kernel given in equation 341. Replacing the kernel by its expression gives

$$\delta\widehat{\kappa}^*(\mathbf{x}') = \sum_i \int_{t_2}^{t_1} dt\ \frac{1}{\kappa_0(\mathbf{x}')} \int_{t_2}^{t_1} dt'\ G(\mathbf{x}_i,t;\mathbf{x}',t')\,\ddot{p}_0(\mathbf{x}',t')\,\delta\widehat{p}(\mathbf{x}_i,t) \ , \tag{345}$$

and this can be rearranged into (note that primed and nonprimed variables have been exchanged)

$$\delta\widehat{\kappa}^*(\mathbf{x}) = \frac{1}{\kappa_0(\mathbf{x})} \int_{t_2}^{t_1} dt\ \psi(\mathbf{x},t)\,\ddot{p}_0(\mathbf{x},t) \ , \tag{346}$$

where

$$\psi(\mathbf{x},t) = \sum_i \int_{t_2}^{t_1} dt'\ G(\mathbf{x}_i,t';\mathbf{x},t)\,\delta\widehat{p}(\mathbf{x}_i,t')\ , \tag{347}$$

or, using the kernel of the tranposed Green's operator,

$$\psi(\mathbf{x},t) = \sum_i \int_{t_2}^{t_1} dt'\ G^T(\mathbf{x},t;\mathbf{x}_i,t')\,\delta\widehat{p}(\mathbf{x}_i,t')\ . \tag{348}$$

The field $\psi(\mathbf{x},t)$ can be interpreted as the solution of the transposed wave equation, with a point source at each point $\mathbf{x}_i$ where we have a receiver, radiating the value $\delta\widehat{p}(\mathbf{x}_i,t')$ . As we have the transposed of the Green's operator, the field $\psi(\mathbf{x},t)$ must satisfy dual boundary conditions, i.e., in our case, final conditions of rest.

### M.11 The Continuous Inverse Problem

Let be $\mathbf{p} = \mathbf{f}(\boldsymbol{\kappa}^*)$ the function calculating the theoretical data associated to the model $\boldsymbol{\kappa}$ (resolution of the forward problem). We seek the model minimizing the sum

$$S(\boldsymbol{\kappa}^*) \ = \ \frac{1}{2}\ \left(\ \parallel \mathbf{f}(\boldsymbol{\kappa}^*) - \mathbf{p}_{\rm obs}\ \parallel^2 + \parallel\ \boldsymbol{\kappa}^* - \boldsymbol{\kappa}^*_{\rm prior}\ \parallel^2\ \right) \tag{349}$$

$$= \ \frac{1}{2}\ \left(\ \langle\ \mathbf{C}_p^{-1}(\mathbf{f}(\boldsymbol{\kappa}^*) - \mathbf{p}_{\rm obs})\ ,\ \mathbf{f}(\boldsymbol{\kappa}^*) - \mathbf{p}_{\rm obs}\ \rangle + \langle\ \mathbf{C}_{\kappa^*}^{-1}(\boldsymbol{\kappa}^* - \boldsymbol{\kappa}^*_{\rm prior})\ ,\ \boldsymbol{\kappa}^* - \boldsymbol{\kappa}^*_{\rm prior}\ \rangle\ \right) \quad .$$

Using, in this functional context, the steepest descent algorithm proposed in section 7.4, we arrive at

$$\kappa^*_{n+1} = \kappa^*_n - \epsilon\left(\mathbf{C}_{\kappa^*}\,\mathbf{F}_n^T\,\mathbf{C}_p^{-1}\ (\mathbf{p}_n - \mathbf{p}_{\rm obs}) + (\boldsymbol{\kappa}^*_n - \boldsymbol{\kappa}^*_{\rm prior})\right)\ , \tag{350}$$

where $\mathbf{p}_n = \mathbf{f}(\boldsymbol{\kappa}^*_n)$ and where $\mathbf{F}_n^T$ is the transposed operatoir defined above, at point $\boldsymbol{\kappa}^*_n$ .

Covariances aside, we see that the fundamental object appearing in this inversion algorithm is the transposed operator $\mathbf{F}^T$ . As it has been interpreted above, we have all the elements to understand how this sort of inverse problems are solved. For more details, see Tarantola (1984, 1986, 1987).

## N Random Walk Design

The design of a random walk that equilibrates at a desired distribution $p(\mathbf{x})$ can be formulated as the design of an equilibrium flow having a throughput of $p(\mathbf{x}_i)\mathbf{dx}_i$ particles in the neighborhood of point $\mathbf{x}_i$. The simplest equilibrium flows are *symmetric*, that is, they satisfy

$$F(\mathbf{x}_i,\mathbf{x}_j) = F(\mathbf{x}_j,\mathbf{x}_i) \tag{351}$$

That is, the transition $\mathbf{x}_i \leftarrow \mathbf{x}_j$ is as likely as the transition $\mathbf{x}_i \rightarrow \mathbf{x}_j$. It is easy to define a symmetric flow, but it will in general not have the required throughput of $p(\mathbf{x}_j)\mathbf{dx}_j$ particles in the neighborhood of point $\mathbf{x}_j$. This requirement can be satisfied if the following adjustment of the flow density is made: first multiply $F(\mathbf{x}_i,\mathbf{x}_j)$ with a positive constant $c$. This constant must be small enough to assure that the throughput of the resulting flow density $cF(\mathbf{x}_i,\mathbf{x}_j)$ at every point $\mathbf{x}_j$ is smaller than the desired probability $p(\mathbf{x}_j)\mathbf{dx}_j$ of its neighborhood. Finally, at every point $\mathbf{x}_j$, add a flow density $F(\mathbf{x}_j,\mathbf{x}_j)$, going from the point to itself, such that the throughput at $\mathbf{x}_j$ gets the right size $p(\mathbf{x}_j)\mathbf{dx}_j$. Neither the flow scaling nor the addition of $F(\mathbf{x}_j,\mathbf{x}_j)$ will destroy the equilibrium property of the flow. In practice, it is unnecessary to add a flow density $F(\mathbf{x}_j,\mathbf{x}_j)$ explicitly, since it is implicit in our algorithms that if no move away from the current point takes place, the move goes from the current point to itself. This rule automatically adjusts the throughput at $\mathbf{x}_j$ to the right size $p(\mathbf{x}_j)\mathbf{dx}_j$

## O The Metropolis Algorithm

Characteristic of a random walk is that the probability of going to a point $\mathbf{x}_i$ in the space $\mathcal{X}$ in a given step (iteration) depends only on the point $\mathbf{x}_j$ it came from. We will define the conditional probability density $P(\mathbf{x}_i \mid \mathbf{x}_j)$ of the location of the next destination $\mathbf{x}_i$ of the random walker, given that it currently is at neighbouring point $\mathbf{x}_j$.

The $P(\mathbf{x}_i \mid \mathbf{x}_j)$ is called the *transition probability density*. As, at each step, the random walker must go somewhere (including the possibility of staying at the same point), then

$$\int_{\mathcal{X}} P(\mathbf{x}_i \mid \mathbf{x}_j) d\mathbf{x}_i = 1. \tag{352}$$

For convenience we shall assume that $P(\mathbf{x}_i \mid \mathbf{x}_j)$ is nonzero everywhere (but typically negligibly small everywhere, except in a certain neighborhood around $\mathbf{x}_j$). For this reason, staying in an infinitesimal neighborhood of the current point $\mathbf{x}_j$ has nonzero probability, and therefore is considered a "transition" (from the point $\mathbf{x}_j$ to itself). The current point, having been reselected, contributes then with one more sample.

Given a random walk defined by the transition probability density $P(\mathbf{x}_i \mid \mathbf{x}_j)$. Assume that the point, where the random walk is initiated, is only known probabilistically: there is a probability density $q(\mathbf{x})$ that the random walk is initiated at point $\mathbf{x}$. Then, when the number of steps tends to infinity, the probability density that the random walker is at point $\mathbf{x}$ will "equilibrate" at some other probability density $p(\mathbf{x})$. It is said that $p(\mathbf{x})$ is an *equilibrium probability density* of $P(\mathbf{x}_i \mid \mathbf{x}_j)$. Then, $p(\mathbf{x})$ is an eigenfunction with eigenvalue 1 of the linear integral operator with kernel $P(\mathbf{x}_i \mid \mathbf{x}_j)$:

$$\int_{\mathcal{X}} P(\mathbf{x}_i \mid \mathbf{x}_j) p(\mathbf{x}_j) d\mathbf{x}_j = p(\mathbf{x}_i). \tag{353}$$

If for any initial probability density $q(\mathbf{x})$ the random walk equilibrates to the same probability density $p(\mathbf{x})$, then $p(\mathbf{x})$ is called *the* equilibrium probability of $P(\mathbf{x}_i \mid \mathbf{x}_j)$. Then, $p(\mathbf{x})$ is the unique eigenfunction of with eigenvalue 1 of the integral operator.

If it is possible for the random walk to go from any point to any other point in $\mathcal{X}$ it is said that the random walk is *irreducible*. Then, there is only one equilibrium probability density.

Given a probability density $p(\mathbf{x})$, many random walks can be defined that have $p(\mathbf{x})$ as their equilibrium density. Some tend more rapidly to the final probability density than others. Samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \ldots$ obtained by a random walk where $P(\mathbf{x}_i \mid \mathbf{x}_j)$ is negligibly small everywhere, except in a certain neighborhood around $\mathbf{x}_j$ will, of course, not be independent unless we only consider points separated by a sufficient number of steps.

Instead of considering $p(\mathbf{x})$ to be the probability density of the position of a (single) random walker (in which case $\int_{\mathcal{X}} p(\mathbf{x})) d\mathbf{x} = 1$), we can consider a situation where we have a "density $p(\mathbf{x})$ of random walkers" in point $\mathbf{x}$. Then, $\int_{\mathcal{X}} p(\mathbf{x})) d\mathbf{x}$ represents the total number of random walkers. None of the results presented below will depend on the way $p(\mathbf{x})$ is normed.

If at some moment the density of random walkers at a point $\mathbf{x}_j$ is $p(\mathbf{x}_j)$, and the transitions probability density is $P(\mathbf{x}_i \mid \mathbf{x}_j)$, then

$$F(\mathbf{x}_i, \mathbf{x}_j) = P(\mathbf{x}_i \mid \mathbf{x}_j) p(\mathbf{x}_j) \tag{354}$$

represents the probability density of transitions from $\mathbf{x}_j$ to $\mathbf{x}_i$: while $P(\mathbf{x}_i \mid \mathbf{x}_j)$ is the *conditional* probability density of the next point $\mathbf{x}_i$ visited by the random walker, given that it currently is at $\mathbf{x}_j$, $F(\mathbf{x}_i, \mathbf{x}_j)$ is the *unconditional* probability density that the next step will be a transition from $\mathbf{x}_j$ to $\mathbf{x}_i$, given only the probability density $p(\mathbf{x}_j)$.

When $p(\mathbf{x}_j)$ is interpreted as the density of random walkers at a point $\mathbf{x}_j$, $F(\mathbf{x}_i, \mathbf{x}_j)$ is called the *flow density*, as $F(\mathbf{x}_i, \mathbf{x}_j) d\mathbf{x}_i d\mathbf{x}_j$ can be interpreted as the number of particles going to a neighborhood of volume $d\mathbf{x}_i$ around point $\mathbf{x}_i$ from a neighborhood of volume $d\mathbf{x}_j$ around point $\mathbf{x}_j$ in a given step. The flow corresponding to an equilibrated random walk has the property that the particle density $p(\mathbf{x}_i)$ at point $\mathbf{x}_i$ is constant in time. Thus, that a random walk has equilibrated at a distribution $p(\mathbf{x})$ means that, in each step, the total flow into an infinitesimal neighborhood of a given point is equal to the total flow out of this neighborhood

Since each of the particles in a neighborhood around point $\mathbf{x}_i$ must move in each step (possibly to the neighborhood itself), the flow has the property that the total flow out from the neighborhood, and hence the total flow into the neighborhood, must equal $p(\mathbf{x}_i)\mathbf{dx}_i$:

$$\int_{\mathcal{X}} F(\mathbf{x}_i, \mathbf{x}_j) d\mathbf{x}_j = \int_{\mathcal{X}} F(\mathbf{x}_k, \mathbf{x}_i) d\mathbf{x}_k = p(\mathbf{x}_i) \tag{355}$$

Consider a random walk with transition probability density $P(\mathbf{x}_i \mid \mathbf{x}_j)$ with equilibrium probability density $p(\mathbf{x})$ and equilibrium flow density $F(\mathbf{x}_i, \mathbf{x}_j)$. We can multiply $F(\mathbf{x}_i, \mathbf{x}_j)$ with any symmetric flow density $\psi(\mathbf{x}_i, \mathbf{x}_j)$, where $\psi(\mathbf{x}_i, \mathbf{x}_j) \leq q(\mathbf{x}_j)$, for all $\mathbf{x}_i$ and $\mathbf{x}_j$, and the resulting flow density

$$\varphi(\mathbf{x}_i, \mathbf{x}_j) = F(\mathbf{x}_i, \mathbf{x}_j) \psi(\mathbf{x}_i, \mathbf{x}_j) \tag{356}$$

will also be symmetric, and hence an equilibrium flow density. A "modified" algorithm with flow density $\psi(\mathbf{x}_i, \mathbf{x}_j)$ and equilibrium probability density $r(\mathbf{x}_j)$ is obtained by dividing $\varphi(\mathbf{x}_i, \mathbf{x}_j)$ with the product probability density $r(\mathbf{x}_j) = p(\mathbf{x}_j)q(\mathbf{x}_j)$. This gives the transition probability density

$$\begin{aligned} P(\mathbf{x}_i, \mathbf{x}_j)^{\text{modified}} &= F(\mathbf{x}_i, \mathbf{x}_j)\frac{\psi(\mathbf{x}_i, \mathbf{x}_j)}{p(\mathbf{x}_j)q(\mathbf{x}_j)} \\ &= P(\mathbf{x}_i \mid \mathbf{x}_j)\frac{\psi(\mathbf{x}_i, \mathbf{x}_j)}{q(\mathbf{x}_j)}, \end{aligned}$$

which is the product of the original transition probability density, and a new probability — the acceptance probability

$$P_{ij}^{\text{acc}} = \frac{\psi(\mathbf{x}_i, \mathbf{x}_j)}{q(\mathbf{x}_j)}. \tag{357}$$

If we choose to multiply $F(\mathbf{x}_i, \mathbf{x}_j)$ with the symmetric flow density

$$\psi_{ij} = \text{Min}(q(\mathbf{x}_i), q(\mathbf{x}_j)), \tag{358}$$

we obtain the Metropolis acceptance probability

$$P_{ij}^{\text{metrop}} = \text{Min}\left(1, \frac{q(\mathbf{x}_i)}{q(\mathbf{x}_j)}\right), \tag{359}$$

which is one for $q(\mathbf{x}_i) \geq q(\mathbf{x}_j)$, and equals $q(\mathbf{x}_i)/q(\mathbf{x}_j)$ when $q(\mathbf{x}_i) < q(\mathbf{x}_j)$.

The *efficiency* of an acceptance rule can be defined as the sum of acceptance probabilities for all possible transitions. The acceptance rule with maximum efficiency is obtained by simultaneously maximizing $\psi(\mathbf{x}_i, \mathbf{x}_j)$ for all pairs of points $\mathbf{x}_j$ and $\mathbf{x}_i$. Since the only constraint on $\psi(\mathbf{x}_i, \mathbf{x}_j)$ (except for positivity) is that $\psi(\mathbf{x}_i, \mathbf{x}_j)$ is symmetric and $\psi(\mathbf{x}_k, \mathbf{x}_l) \leq q(\mathbf{x}_l)$, for all $k$ and $l$, we have $\psi(\mathbf{x}_i, \mathbf{x}_j) \leq q(\mathbf{x}_j)$ and $\psi(\mathbf{x}_i, \mathbf{x}_j) \leq q(\mathbf{x}_i)$. This means that the acceptance rule with maximum efficiency is the Metropolis rule, where

$$\psi_{ij} = \text{Min}\left(q(\mathbf{x}_i), q(\mathbf{x}_j)\right). \tag{360}$$

# P The Borel 'Paradox'

A description of the paradox is given, for instance, by Kolmogorov (1933), in his *Foundations of the Theory of Probability* (see figure 31 here).

Figure 31: A reproduction of a section of Kolmogorov's book *Foundations of the theory of probability* (1950, pp. 50–51). He describes the so-called "Borel paradox". We do not agree with his conclusion as he ignores the fact that the surface of the the sphere is a metric space, so it allows an intrinsic definition of conditional probability density (see main text).

---

**§ 2. Explanation of a Borel Paradox**

Let us choose for our basic set $E$ the set of all points on a spherical surface. Our $\mathcal{F}$ will be the aggregate of all Borel sets of the spherical surface. And finally, our $P(A)$ is to be proportional to the measure set of $A$. Let us now choose two diametrically opposite points for our poles, so that each meridian circle will be uniquely defined by the longitude $\psi$, $0 \leq \psi < \pi$. Since $\psi$ varies from 0 only to $\pi$, — in other words, we are considering *complete* meridian circles (and not merely semicircles) — the latitude $\theta$ must vary

from $-\pi$ to $+\pi$ (and not from $-\frac{\pi}{2}$ to $+\frac{\pi}{2}$ ). Borel set the following problem: Required to determine "the conditional probability distribution" of latitude $\theta$ , $-\pi \leq \theta < +\pi$ , for a given longitude $\psi$ .

It is easy to calculate that

$$P_{\psi}(\theta_1 \leq \theta < \theta_2) = \frac{1}{4} \int_{\theta_1}^{\theta_2} |\cos\theta| \; d\theta \quad .$$

The probability distribution of $\theta$ for a given $\psi$ is not uniform.

If we assume the the conditional probability distribution of $\theta$ "with the hypothesis that $\xi$ lies on the given meridian circle" must be uniform, then we have arrived at a contradiction.

This shows that the concept of a conditional probability with regard to an isolated given hypothesis whose probability equals 0 is inadmissible. For we van obtain a probability distribution for $\theta$ on the meridian circle only if we regard this circle as an element of the decomposition of the entire spherical surface into meridian circles with the given poles.

---

A probability distribution is considered over the surface of the unit sphere, associating, as it should, to any region $\mathcal{A}$ of the surface of the sphere, a positive real number $P(\mathcal{A})$ . To any possible choice of coordinates $\{u, v\}$ on the surface of the sphere will correspond a probability density $f(u, v)$ representing the given probability distribution, through $P(\mathcal{A}) = \int du \int dv \; f(u, v)$ (integral over the region $\mathcal{A}$ ). At this point of the discussion, the coordinates $\{u, v\}$ may be the standard spherical coordinates or any other system of coordinates (as, for instance, the Cartesian coordinates in a representation of the surface of the sphere as a 'geographical map', using any 'geographical projection').

A great circle is given on the surface of the sphere, that, should we use spherical coordinates, is not necessarily the 'equator' or a 'meridian'. Points on this circle may be parameterized by a coordinate $\alpha$ , that, for simplicity, we may take to be the circular angle (as measured from the center of the sphere).

The probability distribution $P(\,\cdot\,)$ defined over the surface of the sphere will induce a probability distribution over the circle. Said otherwise, the probability density $f(u, v)$ defined over the surface of the sphere will induce a probability density $g(\alpha)$ over the circle. This is the situation one has in mind when defining the notion of conditional probability density, so we may say that $g(\alpha)$ is the conditional probability density induced on the circle by the probability density $f(u, v)$ , given the condition that points must lie on the great circle.

The Borel-Kolmogorov paradox is obtained when the probability distribution over the surface of the sphere is homogeneous. If it is homogeneous over the sphere, the conditional probability distribution over the great circle must be homogeneous too, and as we parameterize by the circular angle $\alpha$ , the conditional probability density over the circle must be

$$g(\alpha) = \frac{1}{2\pi} \; , \tag{361}$$

and this is *not* what one gets from the standard definition of conditional probability density, as we will see below.

From now on, assume that the spherical coordinates $\{\vartheta, \varphi\}$ are used, where $\vartheta$ is the latitude (rather than the colalitude $\theta$ ), so the domains of definition of the variables are

$$-\pi/2 < \vartheta \leq +\pi/2 \qquad ; \qquad -\pi < \varphi \leq +\pi \quad . \tag{362}$$

As the surface element is $dS(\vartheta, \varphi) = \cos\vartheta \, d\vartheta \, d\varphi$ , the homogeneous probability distribution over the surface of the sphere is represented, in spherical coordinates, by the probability density

$$f(\vartheta, \varphi) = \frac{1}{4\pi} \cos\vartheta \; , \tag{363}$$

and we satisfy the normalization condition

$$\int_{-\pi/2}^{+\pi/2} d\vartheta \int_{-\pi}^{+\pi} d\varphi \; f(\vartheta, \varphi) = 1 \; . \tag{364}$$

The probability of any region equals the relative surface of the region (i.e., the ratio of the surface of the region divided by the surface of the sphere, $4\pi$ ), so the probability density in equation 363 do represents the homogeneous probability distribution.

Two different computations follow. Both are aimed at computing the conditional probability density over a great circle.

The first one uses the nonconventional definition of conditional probability density introduced in section in appendix B of this article (and claimed to be 'consistent'). No paradox appears. No matter if we take as great circle a meridian or the equator.

The second computation is the conventional one. The traditional Borel-Kolmogorov paradox appears, when the great circle is taken to be a meridian. We interpret this as a sign of the inconsistency of the conventional theory. Let us develop the example.

We have the line element (taking a sphere of radius 1 ),

$$ds^2 = d\vartheta^2 + \cos^2\vartheta \, d\varphi^2 \ , \tag{365}$$

which gives the metric components

$$g_{\vartheta\vartheta}(\vartheta,\varphi) = 1 \qquad ; \qquad g_{\varphi\varphi}(\vartheta,\varphi) = \cos^2\vartheta \tag{366}$$

and the surface element

$$dS(\vartheta,\varphi) = \cos\vartheta \, d\vartheta \, d\varphi \ . \tag{367}$$

Letting $f(\vartheta,\varphi)$ be a probability density over the sphere, consider the restriction of this probability on the (half) meridian $\varphi = \varphi_0$ , i.e., the conditional probability density on this (half) meridian. It is, following equation 131 ,

$$f_\vartheta(\vartheta|\varphi=\varphi_0) \ = \ k \, \frac{f(\vartheta,\varphi_0)}{\sqrt{g_{\varphi\varphi}(\vartheta,\varphi_0)}} \quad . \tag{368}$$

In our case, using the second of equations 366

$$f_\vartheta(\vartheta|\varphi=\varphi_0) \ = \ k \, \frac{f(\vartheta,\varphi_0)}{\cos\vartheta} \quad , \tag{369}$$

or, in normalized version,

$$f_\vartheta(\vartheta|\varphi=\varphi_0) \ = \ \frac{f(\vartheta,\varphi_0)/\cos\vartheta}{\int_{-\pi/2}^{+\pi/2} d\vartheta \ f(\vartheta,\varphi_0)/\cos\vartheta} \quad . \tag{370}$$

If the original probability density $f(\vartheta,\varphi)$ represents an homogeneous probability, then it must be proportional to the surface element $dS$ (equation 367), so, in normalized form, the homogeneous probability density is

$$f(\vartheta,\varphi) = \frac{1}{4\pi} \ \cos\vartheta \ . \tag{371}$$

Then, equation 369 gives

$$f_\vartheta(\vartheta|\varphi=\varphi_0) = \frac{1}{\pi} \ . \tag{372}$$

We see that this conditional probability density is constant[40].

This is in contradiction with usual 'definitions' of conditional probability density, where the metric of the space is not considered, and where instead of the correct equation 368, the conditional probability density is 'defined' by

$$f_\vartheta(\vartheta|\varphi=\varphi_0) \ = \ k \, f(\vartheta,\varphi_0) \ = \ \frac{f(\vartheta,\varphi_0)}{\int_{-\pi/2}^{+\pi/2} d\vartheta \ f(\vartheta,\varphi_0)/\cos\vartheta} \qquad \text{wrong definition} \quad , \tag{373}$$

this leading, in the considered case, to the conditional probability density

$$f_\vartheta(\vartheta|\varphi=\varphi_0) \ = \ \frac{\cos\vartheta}{2} \qquad \text{wrong result} \quad . \tag{374}$$

[40]This constant value is $1/\pi$ if we consider half a meridian, or it is $1/2\pi$ if we consider a whole meridian.

This result is the celebrated 'Borel paradox'. As any other 'mathematical paradox', it is not a paradox, it is just the result of an inconsistent calculation, with an arbitrary definition of conditional probability density.

The interpretation of the paradox by Kolmogorov (1933) sounds quite strange to us (see figure 31). Jaynes (1995) says "*Whenever we have a probability density on one space and we wish to generate from it one on a subspace of measure zero, the only safe procedure is to pass to an explicitly defined limit [...]. In general, the final result will and must depend on which limiting operation was specified. This is extremely counter-intuitive at first hearing; yet it becomes obvious when the reason for it is understood.*"

We agree with Jaynes, and go one step further. We claim that usual parameter spaces, where we define probability densities, normally accept a natural definition of distance, and that the 'limiting operation' (in the words of Jaynes) must the the *uniform convergence associated to the metric*. This is what we have done to define the notion of conditional probability. Many examples of such distances are shown in this article.

End Of Paper
Klaus Mosegaard & Albert Tarantola, October 2000