

probability and measurement

ALBERT TARANTOLA

to be published by ...

Probability and Measurements ¹

Albert Tarantola

Université de Paris, Institut de Physique du Globe

4, place Jussieu; 75005 Paris; France

E-mail: `Albert.Tarantola@ipgp.jussieu.fr`

December 3, 2001

To the memory of my father.

To my mother and my wife.

Preface

In this book, I attempt to reach two goals. The first is purely mathematical: to clarify some of the basic concepts of probability theory. The second goal is physical: to clarify the methods to be used when handling the information brought by measurements, in order to understand how accurate are the predictions we may wish to make.

Probability theory is solidly based on Kolmogorov axioms, and there is no problem when treating discrete probabilities. But I am very unhappy with the usual way of extending the theory to continuous probability distributions. In this text, I introduce the notion of ‘volumetric probability’ different from the more usual notion of ‘probability density’. I claim that some of the more basic problems of the theory of continuous probability distributions can only be solved within this framework, and that many of the well known ‘paradoxes’ of the theory are fundamental misunderstandings, that I try to clarify.

I start the book with an introduction to tensor calculus, because I choose to develop the probability theory considering metric manifolds.

The second chapter deals with the probability theory per se. I try to use intrinsic notions everywhere, i.e., I only introduce definitions that make sense irrespectively of the particular coordinates being used in the manifold under investigation. The reader shall see that this leads to many developments that are at odds with those found in usual texts.

In physical applications one not only needs to define probability distributions over (typically) large-dimensional manifolds. One also needs to make use of them, and this is achieved by sampling the probability distributions using the ‘Monte Carlo’ methods described in chapter 3. There is no major discovery exposed in this chapter, but I make the effort to set Monte Carlo methods using the intrinsic point of view mentioned above.

The metric foundation used here allows to introduce the important notion of ‘homogeneous’ probability distributions. Contrary to the ‘noninformative’ probability distributions common in the Bayesian literature, the homogeneity notion is not controversial (provided one has agreed on a given metric over the space of interest).

After a brief chapter that explain what an ideal measuring instrument should be, the book enters in the four chapter developing what I see as the four more basic inference problems in physics: (i) problems that are solved using the notion of ‘sum of probabilities’ (just an elaborate way of ‘making histograms), (ii) problems that are solved using the ‘product of probabilities’ (and approach that seems to be original), (iii) problems that are solved using ‘conditional probabilities’ (these including the so-called ‘inverse problems’), and (iv) problems that are solved using the ‘transport of probabilities’ (like the typical [indirect] measurement problem, but solved here transporting probability distributions, rather than just transporting ‘uncertainties’).

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

Paris, December 3, 2001
Albert Tarantola

Contents

1	Introduction to Tensors	1
2	Elements of Probability	69
3	Monte Carlo Sampling Methods	153
4	Homogeneous Probability Distributions	169
5	Basic Measurements	185
6	Inference Problems of the First Kind (Sum of Probabilities)	207
7	Inference Problems of the Second Kind (Product of Probabilities)	211
8	Inference Problems of the Third Kind (Conditional Probabilities)	219
9	Inference Problems of the Fourth Kind (Transport of Probabilities)	287

Contents

1	Introduction to Tensors	1
1.1	Chapter's overview	3
1.2	Change of Coordinates (Notations)	4
1.3	Metric, Volume Density, Metric Bijections	7
1.4	The Levi-Civita Tensor	9
1.5	The Kronecker Tensor	11
1.6	Totally Antisymmetric Tensors	14
1.7	Integration, Volumes	19
1.8	Appendixes	23
2	Elements of Probability	69
2.1	Volume	70
2.2	Probability	78
2.3	Sum and Product of Probabilities	84
2.4	Conditional Probability	88
2.5	Marginal Probability	100
2.6	Transport of Probabilities	106
2.7	Central Estimators and Dispersion Estimators	116
2.8	Appendixes	120
3	Monte Carlo Sampling Methods	153
3.1	Introduction	154
3.2	Random Walks	155
3.3	Modification of Random Walks	157
3.4	The Metropolis Rule	158
3.5	The Cascaded Metropolis Rule	158
3.6	Initiating a Random Walk	159
3.7	Designing Primeval Walks	160
3.8	Multistep Iterations	161
3.9	Choosing Random Directions and Step Lengths	162
3.10	Appendixes	164
4	Homogeneous Probability Distributions	169
4.1	Parameters	169
4.2	Homogeneous Probability Distributions	171
4.3	Appendixes	176

5	Basic Measurements	185
5.1	Terminology	186
5.2	Old text: Measuring physical parameters	187
5.3	From ISO	189
5.4	The Ideal Output of a Measuring Instrument	194
5.5	Output as Conditional Probability Density	195
5.6	A Little Bit of Theory	195
5.7	Example: Instrument Specification	195
5.8	Measurements and Experimental Uncertainties	197
5.9	Appendixes	200
6	Inference Problems of the First Kind (Sum of Probabilities)	207
6.1	Experimental Histograms	208
6.2	Sampling a Sum	209
6.3	Further Work to be Done	209
7	Inference Problems of the Second Kind (Product of Probabilities)	211
7.1	The ‘Shipwrecked Person’ Problem	212
7.2	Physical Laws as Probabilistic Correlations	213
8	Inference Problems of the Third Kind (Conditional Probabilities)	219
8.1	Adjusting Measurements to a Physical Theory	220
8.2	Inverse Problems	222
8.3	Appendixes	231
9	Inference Problems of the Fourth Kind (Transport of Probabilities)	287
9.1	Measure of Physical Quantities	288
9.2	Prediction of Observations	299
9.3	Appendixes	300

Contents

1	Introduction to Tensors	1
1.1	Chapter's overview	3
1.2	Change of Coordinates (Notations)	4
1.2.1	Jacobian Matrices	4
1.2.2	Tensors, Capacities and Densities	5
1.3	Metric, Volume Density, Metric Bijections	7
1.3.1	Metric	7
1.3.2	Volume Density	8
1.3.3	Bijection Between Densities Tensors and Capacities	8
1.4	The Levi-Civita Tensor	9
1.4.1	Orientation of a Coordinate System	9
1.4.2	The Fundamental (Levi-Civita) Capacity	9
1.4.3	The Fundamental Density	9
1.4.4	The Levi-Civita Tensor	10
1.4.5	Determinants	10
1.5	The Kronecker Tensor	11
1.5.1	Kronecker Tensor	11
1.5.2	Kronecker Determinants	11
1.6	Totally Antisymmetric Tensors	14
1.6.1	Totally Antisymmetric Tensors	14
1.6.2	Dual Tensors	14
1.6.3	Exterior Product of Tensors	16
1.6.4	Exterior Derivative of Tensors	18
1.7	Integration, Volumes	19
1.7.1	The Volume Element	19
1.7.2	The Stokes' Theorem	20
1.8	Appendixes	23
1.8.1	Appendix: Tensors For Beginners	23
1.8.2	Appendix: Dimension of Components	41
1.8.3	Appendix: The Jacobian in Geographical Coordinates	42
1.8.4	Appendix: Kronecker Determinants in 2 3 and 4 D	44
1.8.5	Appendix: Definition of Vectors	45
1.8.6	Appendix: Change of Components	46
1.8.7	Appendix: Covariant Derivatives	47
1.8.8	Appendix: Formulas of Vector Analysis	48
1.8.9	Appendix: Metric, Connection, etc. in Usual Coordinate Systems	50

1.8.10	Appendix: Gradient, Divergence and Curl in Usual Coordinate Systems	56
1.8.11	Appendix: Connection and Derivative in Different Coordinate Systems	61
1.8.12	Appendix: Computing in Polar Coordinates	63
1.8.13	Appendix: Dual Tensors in 2 3 and 4D	65
1.8.14	Appendix: Integration in 3D	67
2	Elements of Probability	69
2.1	Volume	70
2.1.1	Notion of Volume	70
2.1.2	Volume Element	70
2.1.3	Volume Density and Capacity Element	71
2.1.4	Change of Variables	73
2.1.5	Conditional Volume	75
2.2	Probability	78
2.2.1	Notion of Probability	78
2.2.2	Volumetric Probability	79
2.2.3	Probability Density	79
2.2.4	Volumetric Histograms and Density Histograms	81
2.2.5	Change of Variables	82
2.3	Sum and Product of Probabilities	84
2.3.1	Sum of Probabilities	84
2.3.2	Product of Probabilities	85
2.4	Conditional Probability	88
2.4.1	Notion of Conditional Probability	88
2.4.2	Conditional Volumetric Probability	89
2.5	Marginal Probability	100
2.5.1	Marginal Probability Density	100
2.5.2	Marginal Volumetric Probability	102
2.5.3	Interpretation of Marginal Volumetric Probability	103
2.5.4	Bayes Theorem	103
2.5.5	Independent Probability Distributions	104
2.6	Transport of Probabilities	106
2.7	Central Estimators and Dispersion Estimators	116
2.7.1	Introduction	116
2.7.2	Center and Radius of a Probability Distribution	116
2.8	Appendixes	120
2.8.1	Appendix: Conditional Probability Density	120
2.8.2	Appendix: Marginal Probability Density	122
2.8.3	Appendix: Replacement Gymnastics	123
2.8.4	Appendix: The Gaussian Probability Distribution	125
2.8.5	Appendix: The Laplacian Probability Distribution	130
2.8.6	Appendix: Exponential Distribution	131
2.8.7	Appendix: Spherical Gaussian Distribution	137
2.8.8	Appendix: Probability Distributions for Tensors	140
2.8.9	Appendix: Determinant of a Partitioned Matrix	143
2.8.10	Appendix: The Borel ‘Paradox’	144

2.8.11	Appendix: Axioms for the Sum and the Product	148
2.8.12	Appendix: Random Points on the Surface of the Sphere	149
2.8.13	Appendix: Histograms for the Volumetric Mass of Rocks	151
3	Monte Carlo Sampling Methods	153
3.1	Introduction	154
3.2	Random Walks	155
3.3	Modification of Random Walks	157
3.4	The Metropolis Rule	158
3.5	The Cascaded Metropolis Rule	158
3.6	Initiating a Random Walk	159
3.7	Designing Primeval Walks	160
3.8	Multistep Iterations	161
3.9	Choosing Random Directions and Step Lengths	162
3.9.1	Choosing Random Directions	162
3.9.2	Choosing Step Lengths	163
3.10	Appendixes	164
3.10.1	Random Walk Design	164
3.10.2	The Metropolis Algorithm	165
3.10.3	Appendix: Sampling Explicitly Given Probability Densities	168
4	Homogeneous Probability Distributions	169
4.1	Parameters	169
4.2	Homogeneous Probability Distributions	171
4.3	Appendixes	176
4.3.1	Appendix: First Digit of the Fundamental Physical Constants	176
4.3.2	Appendix: Homogeneous Probability for Elastic Parameters	178
4.3.3	Appendix: Homogeneous Distribution of Second Rank Tensors	183
5	Basic Measurements	185
5.1	Terminology	186
5.2	Old text: Measuring physical parameters	187
5.3	From ISO	189
5.3.1	Proposed vocabulary to be used in metrology	189
5.3.2	Some basic concepts	191
5.4	The Ideal Output of a Measuring Instrument	194
5.5	Output as Conditional Probability Density	195
5.6	A Little Bit of Theory	195
5.7	Example: Instrument Specification	195
5.8	Measurements and Experimental Uncertainties	197
5.9	Appendixes	200
5.9.1	Appendix: Operational Definitions can not be Infinitely Accurate	200
5.9.2	Appendix: The International System of Units (SI)	201

6	Inference Problems of the First Kind (Sum of Probabilities)	207
6.1	Experimental Histograms	208
6.2	Sampling a Sum	209
6.3	Further Work to be Done	209
7	Inference Problems of the Second Kind (Product of Probabilities)	211
7.1	The ‘Shipwrecked Person’ Problem	212
7.2	Physical Laws as Probabilistic Correlations	213
7.2.1	Physical Laws	213
7.2.2	Example: Realistic ‘Uncertainty Bars’ Around a Functional Relation	213
7.2.3	Inverse Problems	214
8	Inference Problems of the Third Kind (Conditional Probabilities)	219
8.1	Adjusting Measurements to a Physical Theory	220
8.2	Inverse Problems	222
8.2.1	Model Parameters and Observable Parameters	223
8.2.2	A Priori Information on Model Parameters	223
8.2.3	Measurements and Experimental Uncertainties	225
8.2.4	Joint ‘Prior’ Probability Distribution in the (\mathbf{M}, \mathbf{D}) Space	225
8.2.5	Physical Laws	226
8.2.6	Inverse Problems	226
8.3	Appendixes	231
8.3.1	Appendix: Short Bibliographical Review	231
8.3.2	Appendix: Example of Ideal (Although Complex) Geophysical Inverse Problem	233
8.3.3	Appendix: Probabilistic Estimation of Earthquake Locations	241
8.3.4	Appendix: Functional Inverse Problems	246
8.3.5	Appendix: Nonlinear Inversion of Waveforms (by Charara & Barnes)	263
8.3.6	Appendix: Using Monte Carlo Methods	272
8.3.7	Appendix: Using Optimization Methods	275
9	Inference Problems of the Fourth Kind (Transport of Probabilities)	287
9.1	Measure of Physical Quantities	288
9.1.1	Example: Measure of Poisson’s Ratio	288
9.2	Prediction of Observations	299
9.3	Appendixes	300
9.3.1	Appendix: Mass Calibration	300
	Bibliography	501
	Index	601

Chapter 1

Introduction to Tensors

[Note: This is an old introduction, to be updated!]

The first part of this book recalls some of the mathematical tools developed to describe the geometric properties of a space. By “geometric properties” one understands those properties that Pythagoras (6th century B.C.) or Euclid (3rd century B.C.) were interested on. The only major conceptual progress since those times has been the recognition that the *physical* space may not be Euclidean, but may have curvature and torsion, and that the behaviour of clocks depends on their space displacements.

Still these representations of the space accept the notion of continuity (or, equivalently, of differentiability). New theories are being developed dropping that condition (e.g. Nottale, 1993). They will not be examined here.

A mathematical structure can describe very different physical phenomena. For instance, the structure “3-D vector space” may describe the combination of forces being applied to a particle, as well as the combination of colors. The same holds for the mathematical structure “differential manifold”. It may describe the 3-D physical space, any 2-D surface, or, more importantly, the 4-dimensional space-time space brought into physics by Minkowski and Einstein. The same theorem, when applied to the physical 3-D space, will have a geometrical interpretation (*stricto sensu*), while when applied to the 4-D space-time will have a dynamical interpretation.

The aim of this first chapter is to introduce the fundamental concepts necessary to describe geometrical properties: those of tensor calculus. Many books on tensor calculus exist. Then, why this chapter here? Essentially because no uniform system of notations exist (indices at different places, different signs . . .). It is then not possible to start any serious work without fixing the notations first. This chapter does not aim to give a complete discussion on tensor calculus. Among the many books that do that, the best are (of course) in French, and Brillouin (1960) is the best among them. Many other books contain introductory discussions on tensor calculus. Weinberg (1972) is particularly lucid. I do not pretend to give a complete set of demonstrations, but to give a complete description of interesting properties, some of which are not easily found elsewhere.

Perhaps original is a notation proposed to distinguish between densities and capacities.

While the trick of using indices in upper or lower position to distinguish between tensors or forms (or, in metric spaces, to distinguish between “contravariant” or “covariant” components) makes formulas intuitive, I propose to use a bar (in upper or lower position) to distinguish between densities (like a probability density) or capacities (like a volume element), this also leading to intuitive results. In particular the bijection existing between these objects in metric spaces becomes as “natural” as the one just mentioned between contravariant and covariant components.

1.1 Chapter's overview

[Note: This is an old introduction, to be updated!]

A vector at a point of an space can intuitively be imagined as an “arrow”. As soon as we can introduce vectors, we can introduce other objects, the *forms*. A form at a point of an space can intuitively be imagined as a series of parallel planes ... At any point of a space we may have tensors, of which the vectors of elementary texts are a particular case. Those tensors may describe the properties of the space itself (metric, curvature, torsion ...) or the properties of something that the space “contains”, like the stress at a point of a continuous medium.

If the space into consideration has a metric (i.e., if the notion of distance between two points has a sense), only tensors have to be considered. If there is not a metric, then, we have to simultaneously consider tensors and forms.

It is well known that in a transformation of coordinates, the value of a probability density \bar{f} at any point of the space is multiplied by ‘the Jacobian’ of the transformation. In fact, a probability density is a scalar field that has well defined tensor properties. This suggests to introduce two different notions where sometimes only one is found: for instance, in addition to the notion of mass density, $\bar{\rho}$, we will also consider the notion of volumetric mass ρ , identical to the former only in Cartesian coordinates. If $\bar{\rho}(\mathbf{x})$ is a mass density, and $v^i(\mathbf{x})$ a true vector, like a velocity. Their product $\bar{p}^i(\mathbf{x}) = \bar{\rho}(\mathbf{x})v^i(\mathbf{x})$ will not transform like a true vector: there will be an extra multiplication by the Jacobian. $\bar{p}^i(\mathbf{x})$ is a density too (of linear momentum).

In addition to tensors and to densities, the concept of “capacity” will be introduced. Under a transformation of coordinates, a capacity is *divided* by the Jacobian of the transformation. An example is the capacity element $d\underline{V} = dx^0 dx^1 \dots$, not to be assimilated to the volume element dV . The product of a capacity by a density gives a true scalar, like in $dM = \bar{\rho} d\underline{V}$.

It is well known that if there is a metric, we can define a bijection between forms and vectors (we can “raise and lower indices”) through $V_i = g_{ij}V^j$. The square root of the determinant of $\{g_{ij}\}$ will be denoted \bar{g} and we will see that it defines a natural bijection between capacities, tensors, and densities, like in $\bar{p}^i = \bar{g}p^i$, so, in addition to the rules concerning the indices, we will have rules concerning the “bars”.

Without a clear understanding of the concept of densities and capacities, some properties remain obscure. We can, for instance, easily introduce a Levi-Civita capacity $\underline{\varepsilon}_{ijk\dots}$, or a Levi-Civita density (the components of both take only the values -1, +1 or 0). A Levi-Civita pure tensor can be defined, but it does not have that simple property. The lack of clear understanding of the need to work simultaneously with densities, pure tensors, and capacities, forces some authors to juggle with “pseudo-things” like the pseudo-vector corresponding to the vector product of two vectors, or to the curl of a vector field.

Many of the properties of tensor spaces are not dependent on the fact that the space may have a metric (i.e., a notion of distance). We will only assume that we have a metric when the property to be demonstrated will require it. In particular, the definition of “covariant” derivative, in the next chapter, will not depend on that assumption.

Also, the dimension of the differentiable manifold (i.e., space) into consideration, is arbitrary (but finite). We will use Latin indices $\{i, j, k, \dots\}$ to denote the components of tensors.

In the second part of the book, as we will specifically deal with the physical space and space-time, the Latin indices $\{i, j, k, \dots\}$ will be reserved for the 3-D physical space, while the Greek indices $\{\alpha, \beta, \gamma, \dots\}$ will be reserved for the 4-D space-time.

1.2 Change of Coordinates (Notations)

1.2.1 Jacobian Matrices

Consider a change of coordinates, passing from the coordinate system $\mathbf{x} = \{x^i\} = \{x^1, \dots, x^n\}$ to another coordinate system $\mathbf{y} = \{y^i\} = \{y^1, \dots, y^n\}$. One may write the coordinate transformation using any of the two equivalent functions

$$\mathbf{y} = \mathbf{y}(\mathbf{x}) \quad ; \quad \mathbf{x} = \mathbf{x}(\mathbf{y}) \quad , \quad (1.1)$$

this being, of course, a short-hand notation for $y^i = y^i(x^1, \dots, x^n)$; $(i = 1, \dots, n)$ and $x^i = x^i(y^1, \dots, y^n)$; $(i = 1, \dots, n)$. We shall need the two sets of partial derivatives

$$Y^i_j = \frac{\partial y^i}{\partial x^j} \quad ; \quad X^i_j = \frac{\partial x^i}{\partial y^j} \quad . \quad (1.2)$$

One has

$$Y^i_k X^k_j = X^i_k Y^k_j = \delta^i_j \quad . \quad (1.3)$$

To simplify language and notations, it is useful to introduce a *matrices of partial derivatives*, ranging the elements X^i_j and Y^i_j as follows,

$$\mathbf{X} = \begin{pmatrix} X^1_1 & X^1_2 & X^1_3 & \cdots \\ X^2_1 & X^2_2 & X^2_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad ; \quad \mathbf{Y} = \begin{pmatrix} Y^1_1 & Y^1_2 & Y^1_3 & \cdots \\ Y^2_1 & Y^2_2 & Y^2_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad . \quad (1.4)$$

Then, equations 1.3 just tell that the matrices \mathbf{X} and \mathbf{Y} are mutually inverses:

$$\mathbf{YX} = \mathbf{XY} = \mathbf{I} \quad . \quad (1.5)$$

The two matrices \mathbf{X} and \mathbf{Y} are called *Jacobian matrices*. As the matrix \mathbf{Y} is obtained by taking derivatives of the variables y^i with respect to the variables x^i , one obtains the matrix $\{Y^i_j\}$ as a function of the variables $\{x^i\}$, so we can write $\mathbf{Y}(\mathbf{x})$ rather than just writing \mathbf{Y} . The reciprocal argument tells that we can write $\mathbf{X}(\mathbf{y})$ rather than just \mathbf{X} . We shall later use this to make some notations more explicit.

Finally, the *Jacobian determinants* of the transformation are the determinants¹ of the two Jacobian matrices:

$$Y = \det \mathbf{Y} \quad ; \quad X = \det \mathbf{X} \quad . \quad (1.6)$$

¹Explicitly, $Y = \det \mathbf{Y} = \frac{1}{n!} \varepsilon_{ijk\dots} Y^i_p Y^j_q Y^k_r \dots \bar{\varepsilon}^{pqr\dots}$, and $X = \det \mathbf{X} = \frac{1}{n!} \varepsilon_{ijk\dots} X^i_p X^j_q X^k_r \dots \bar{\varepsilon}^{pqr\dots}$, and where the Levi-Civita's "symbols" $\varepsilon_{ijk\dots}$ take the value $+1$ if $\{i, j, k, \dots\}$ is an *even* permutation of $\{1, 2, 3, \dots\}$, the value -1 if $\{i, j, k, \dots\}$ is an *odd* permutation of $\{1, 2, 3, \dots\}$, and the value 0 if some indices are identical. The Levi-Civita's tensors will be introduced with mre detail in section 1.4).

1.2.2 Tensors, Capacities and Densities

Consider an n -dimensional manifold, and let \mathcal{P} be a point of it. Also consider a tensor \mathbf{T} at point \mathcal{P} , and let $T_{\mathbf{x}}^{ij\dots kl\dots}$ be the components of \mathbf{T} on the local natural basis associated to some coordinates $\mathbf{x} = \{x^1, \dots, x^n\}$.

On a change of coordinates from \mathbf{x} into $\mathbf{y} = \{y^1, \dots, y^n\}$ (and the corresponding change of local natural basis) the components of \mathbf{T} shall become $T_{\mathbf{y}}^{ij\dots kl\dots}$. It is well known that the components are related through

$$T_{\mathbf{y}}^{pq\dots rs\dots} = \frac{\partial y^p}{\partial x^i} \frac{\partial y^q}{\partial x^j} \dots \frac{\partial x^k}{\partial y^r} \frac{\partial x^\ell}{\partial y^s} \dots T_{\mathbf{x}}^{ij\dots kl\dots} \quad , \quad (1.7)$$

or, using the notations introduced above,

$$\boxed{T_{\mathbf{y}}^{pq\dots rs\dots} = Y^p{}_i Y^q{}_j \dots X^k{}_r X^\ell{}_s \dots T_{\mathbf{x}}^{ij\dots kl\dots} \quad .} \quad (1.8)$$

In particular, for totally contravariant and totally covariant tensors,

$$T_{\mathbf{y}}^{kl\dots} = Y^k{}_i Y^\ell{}_j \dots T_{\mathbf{x}}^{ij\dots} \quad ; \quad T_{\mathbf{y}kl\dots} = X^i{}_k X^j{}_l \dots T_{\mathbf{x}ij\dots} \quad . \quad (1.9)$$

In addition to actual tensors, we shall encounter other objects, that ‘have indices’ also, and that transform in a slightly different way: *densities* and *capacities* (see for instance Weinberg [1972] and Winogradzki [1979]). Rather than a general exposition of the properties of densities and capacities, let us anticipate that we shall only find totally contravariant densities and totally covariant capacities (the most notable example being the Levi-Civita capacity, to be introduced below). From now on, in all this text,

- a density is denoted with an overline, like in $\bar{\mathbf{a}}$;
- a capacity is denoted with an underline, like in $\underline{\mathbf{b}}$.

It is time now to give what we can take as defining properties: Under the considered change of coordinates, a totally contravariant density $\bar{\mathbf{a}}$ changes components following the law

$$\boxed{\bar{\mathbf{a}}_{\mathbf{y}}^{kl\dots} = \frac{1}{Y} Y^k{}_i Y^\ell{}_j \dots \bar{\mathbf{a}}_{\mathbf{x}}^{ij\dots} \quad ,} \quad (1.10)$$

or, equivalently, $\bar{\mathbf{a}}_{\mathbf{y}}^{kl\dots} = X Y^k{}_i Y^\ell{}_j \dots \bar{\mathbf{a}}_{\mathbf{x}}^{ij\dots}$. Here $X = \det \mathbf{X}$ and $Y = \det \mathbf{Y}$ are the Jacobian determinants introduced in equation 1.6. This rule for the change of components for a totally contravariant density is the same as that for a totally contravariant tensor (equation at left in 1.9), excepted that there is an extra factor, the Jacobian determinant $X = 1/Y$.

Similarly, a totally covariant capacity $\underline{\mathbf{b}}$ changes components following the law

$$\boxed{\underline{\mathbf{b}}_{\mathbf{y}kl\dots} = \frac{1}{X} X^i{}_k X^j{}_l \dots \underline{\mathbf{b}}_{\mathbf{x}ij\dots} \quad ,} \quad (1.11)$$

or, equivalently, $\underline{\mathbf{b}}_{\mathbf{y}kl\dots} = Y X^i{}_k X^j{}_l \dots \underline{\mathbf{b}}_{\mathbf{x}ij\dots}$. Again, this rule for the change of components for a totally covariant capacity is the same as that for a totally covariant tensor (equation at right in 1.9), excepted that there is an extra factor, the Jacobian determinant $Y = 1/X$.

The number of terms in equations 1.10 and 1.11 depends on the ‘variance’ of the objects considered (i.e., in the number of indices they have). We shall find, in particular, scalar densities and scalar capacities, that do not have any index. The natural extension of equations 1.10 and 1.11 is, obviously,

$$\boxed{\bar{a}_{\mathbf{y}} = X \bar{a}_{\mathbf{x}} = \frac{1}{Y} \bar{a}_{\mathbf{x}}} \quad (1.12)$$

for a scalar density, and

$$\boxed{\underline{b}_{\mathbf{y}} = Y \underline{b}_{\mathbf{x}} = \frac{1}{X} \underline{b}_{\mathbf{x}}} \quad (1.13)$$

for a scalar capacity. Explicitly, these equations can be written, using \mathbf{y} as variable,

$$\bar{a}_{\mathbf{y}}(\mathbf{y}) = X(\mathbf{y}) \bar{a}_{\mathbf{x}}(\mathbf{x}(\mathbf{y})) \quad ; \quad \underline{b}_{\mathbf{y}}(\mathbf{y}) = \frac{1}{X(\mathbf{y})} \underline{b}_{\mathbf{x}}(\mathbf{x}(\mathbf{y})) \quad , \quad (1.14)$$

or, equivalently, using \mathbf{x} as variable,

$$\bar{a}_{\mathbf{y}}(\mathbf{y}(\mathbf{x})) = \frac{1}{Y(\mathbf{x})} \bar{a}_{\mathbf{x}}(\mathbf{x}) \quad ; \quad \underline{b}_{\mathbf{y}}(\mathbf{y}(\mathbf{x})) = Y(\mathbf{x}) \underline{b}_{\mathbf{x}}(\mathbf{x}) \quad . \quad (1.15)$$

1.3 Metric, Volume Density, Metric Bijections

1.3.1 Metric

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 $\mathbf{x} = \{x^i\}$ and the point of coordinates $\mathbf{x} + d\mathbf{x} = \{x^i + dx^i\}$ can be expressed as²

$$ds^2 = (d\mathbf{x})^2 = g_{ij}(\mathbf{x}) dx^i dx^j \quad , \quad (1.16)$$

i.e., if the notion of distance is ‘of the L_2 type’³. 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, $\bar{g}(\mathbf{x})$, equals the square root of the determinant of the metric:

$$\bar{g}(\mathbf{x}) = \sqrt{\det \mathbf{g}(\mathbf{x})} \quad . \quad (1.17)$$

Example 1.1 *In the Euclidean 3D space, using geographical coordinates (see example ??) the distance element is $ds^2 = dr^2 + r^2 \cos^2 \vartheta d\varphi^2 + r^2 d\vartheta^2$, from where it follows that the metric matrix is*

$$\begin{pmatrix} g_{rr} & g_{r\varphi} & g_{r\vartheta} \\ g_{\varphi r} & g_{\varphi\varphi} & g_{\varphi\vartheta} \\ g_{\vartheta r} & g_{\vartheta\varphi} & g_{\vartheta\vartheta} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 \cos^2 \vartheta & 0 \\ 0 & 0 & r^2 \end{pmatrix} \quad . \quad (1.18)$$

The volume density equals the metric determinant, $\bar{g}(r, \varphi, \vartheta) = \sqrt{\det \mathbf{g}(r, \varphi, \vartheta)} = r^2 \cos \vartheta$.
[End of example.]

Note: define here the contravariant components of the metric through

$$g^{ij} g_{jk} = \delta^i_k \quad . \quad (1.19)$$

Using equations 1.9, we see that the covariant and contravariant components of the metric change according to

$$g_{\mathbf{y}k\ell} = X^i_k X^j_\ell g_{\mathbf{x}ij} \quad \text{and} \quad g^{\mathbf{y}k\ell} = Y^k_i Y^\ell_j g^{\mathbf{x}ij} \quad . \quad (1.20)$$

In section 1.2, we introduced the matrices of partial derivatives. It is useful to also introduce two metric matrices, with respectively the covariant and contravariant components of the metric:

$$\mathbf{g} = \begin{pmatrix} g_{11} & g_{12} & g_{13} & \cdots \\ g_{21} & g_{22} & g_{23} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad ; \quad \mathbf{g}^{-1} = \begin{pmatrix} g^{11} & g^{12} & g^{13} & \cdots \\ g^{21} & g^{22} & g^{23} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad , \quad (1.21)$$

the notation \mathbf{g}^{-1} for the second matrix being justified by the definition 1.19, that now reads

$$\mathbf{g}^{-1} \mathbf{g} = \mathbf{I} \quad . \quad (1.22)$$

In matrix notation, the change of the metric matrix under a change of variables, as given by the two equations 1.20, is written

$$\mathbf{g}_{\mathbf{y}} = \mathbf{X}^t \mathbf{g}_{\mathbf{x}} \mathbf{X} \quad ; \quad \mathbf{g}_{\mathbf{y}}^{-1} = \mathbf{Y} \mathbf{g}_{\mathbf{x}}^{-1} \mathbf{Y}^t \quad . \quad (1.23)$$

²This is a property that is valid for any coordinate system that can be chosen over the space.

³As a counterexample, the distance defined as $ds = |dx| + |dy|$ is not of the L_2 type (it is L_1).

1.3.2 Volume Density

[Note: The text that follows has to be simplified.]

We have seen that the metric can be used to define a natural bijection between forms and vectors. Let us now see that it can also be used to define a natural bijection between tensors, densities, and capacities.

Let us denote by \bar{g} the square root of the determinant of the metric,

$$\bar{g} = \sqrt{\det \mathbf{g}} = \sqrt{\frac{1}{n!} \bar{\varepsilon}^{ijk\dots} \bar{\varepsilon}^{pqr\dots} g_{ip} g_{jq} g_{kr} \dots} \quad (1.24)$$

[Note: Explain here that this is a density (in fact, the **fundamental density**)].

In (Comment: where?) we demonstrate that we have

$$\partial_i \bar{g} = \bar{g} \Gamma_{is}^s. \quad (1.25)$$

Using expression (Comment: which one?) for the (covariant) derivative of a scalar density, this simply gives

$$\nabla_i \bar{g} = \partial_i \bar{g} - \bar{g} \Gamma_{is}^s = 0, \quad (1.26)$$

which is consistent with the fact that

$$\nabla_i g_{jk} = 0. \quad (1.27)$$

Note: define here the fundamental capacity

$$\underline{g} = \frac{1}{\bar{g}}, \quad (1.28)$$

and say that it is a capacity (obvious).

1.3.3 Bijection Between Densities Tensors and Capacities

Using the scalar density \bar{g} we can associate tensor densities, pure tensors, and tensor capacities. Using the same letter to designate the objects related through this natural bijection, we will write expressions like

$$\bar{\rho} = \bar{g} \rho \quad ; \quad \bar{V}^i = \bar{g} V^i \quad \text{or} \quad \underline{g} \bar{T}_{ij\dots}^{kl\dots} = T_{ij\dots}^{kl\dots} \quad (1.29)$$

So, if g_{ij} and g^{ij} can be used to “lower and raise indices”, \bar{g} and \underline{g} can be used to “put and remove bars”.

Comment: say somewhere that \bar{g} is the *density of volumetric content*, as the volume element of a metric space is given by

$$dV = \bar{g} d\mathcal{T}, \quad (1.30)$$

where $d\mathcal{T}$ is the *capacity element* defined in (Comment: where?), and which, when we take an element along the coordinate lines, equals $dx^1 \wedge dx^2 \wedge dx^3 \dots$.

Comment: Give somewhere the formula $\partial_i \bar{g} = \bar{g} \Gamma_i$. It can be justified by the fact that, for any density, \bar{s} , $\nabla_k \bar{s} = \partial_k \bar{s} - \Gamma_k \bar{s}$, and the result follows by using $\bar{s} = \bar{g}$ and remembering that $\nabla_k \bar{g} = 0$.

1.4 The Levi-Civita Tensor

1.4.1 Orientation of a Coordinate System

The Jacobian determinants associated to a change of variables $\mathbf{x} \Rightarrow \mathbf{y}$ have been defined in section 1.2. As their product must equal $+1$, they must be both positive or both negative. Two different coordinate systems $\mathbf{x} = \{x^1, x^2, \dots, x^n\}$ and $\mathbf{y} = \{y^1, y^2, \dots, y^n\}$ are said to have the ‘same orientation’ (at a given point) if the Jacobian determinants of the transformation, are positive. If they are negative, it is said that the two coordinate systems have ‘opposite orientation’. Precisely, the *orientation* of a coordinate system is the quantity η that may take the value $+1$ or the value -1 . The orientation η of *any* coordinate system is then unambiguously defined when a definite sign of η is assigned to a particular coordinate system.

Example 1.2 *In the Euclidean 3D space, a positive orientation is assigned to a Cartesian coordinate system $\{x, y, z\}$ when the positive sense of the z is obtained from the positive senses of the x axis and the y axis following the screwdriver rule. Another Cartesian coordinate system $\{u, v, w\}$ defined as $u = y, v = x, w = z$, then would have a negative orientation. A system of these spherical coordinates, if taken in their usual order $\{r, \theta, \varphi\}$, then also has a positive orientation, but when changing the order of two coordinates, like in $\{r, \varphi, \theta\}$, the orientation of the coordinate system is negative. For a system of geographical coordinates, the reverse is true, while $\{r, \varphi, \vartheta\}$ is a positively oriented system, $\{r, \vartheta, \varphi\}$ is negatively oriented. [End of example.]*

1.4.2 The Fundamental (Levi-Civita) Capacity

The *Levi-Civita capacity* can be defined by the condition

$$\epsilon_{ijk\dots} = \begin{cases} +\eta & \text{if } ijk\dots \text{ is an even permutation of } 12\dots n \\ 0 & \text{if some indices are identical} \\ -\eta & \text{if } ijk\dots \text{ is an odd permutation of } 12\dots n \end{cases}, \quad (1.31)$$

where η is the orientation of the coordinate system, as defined in section 1.4.1.

It can be shown [note: give here a reference or the demonstration] that the object so defined actually is a capacity, i.e., that in a change of coordinates, when it is imposed that the components of this ‘object’ change according to equation 1.11, the defining property 1.31 is preserved.

1.4.3 The Fundamental Density

Let \mathbf{g} the metric tensor of the manifold. For any positively oriented system of coordinates, we define the quantity \bar{g} , called the *volume density* (in the given coordinates) as

$$\bar{g} = \eta \sqrt{\det \mathbf{g}} \quad . \quad (1.32)$$

where η is the orientation of the coordinate system, as defined in section 1.4.1.

It can be shown [note: give here a reference or the demonstration] that the object so defined actually is a scalar density, i.e., that in a change of coordinates, this quantity changes according to equation 1.12 respectively, the property 1.32 is preserved.

1.4.4 The Levi-Civita Tensor

Then, the *Levi-Civita tensor* can be defined as ⁴

$$\epsilon_{ij\dots k} = \bar{g} \underline{\epsilon}_{ij\dots k} \quad , \quad (1.33)$$

i.e., explicitly,

$$\epsilon_{ijk\dots} = \begin{cases} +\sqrt{\det \mathbf{g}} & \text{if } ijk\dots \text{ is an } \textit{even} \text{ permutation of } 12\dots n \\ 0 & \text{if some indices are identical} \\ -\sqrt{\det \mathbf{g}} & \text{if } ijk\dots \text{ is an } \textit{odd} \text{ permutation of } 12\dots n \end{cases} . \quad (1.34)$$

It can be shown [*note: give here a reference or the demonstration*] that the object so defined actually is a tensor, i.e., that in a change of coordinates, when it is imposed that the components of this ‘object’ change according to equation 1.9, the property 1.34 is preserved.

1.4.5 Determinants

The Levi-Civita’s tensors can be used to define determinants. For instance, the determinants of the tensors Q_{ij} , R_i^j , S^i_j , and T^{ij} are defined by

$$Q = \frac{1}{n!} \epsilon^{ijk\dots} \epsilon^{mnr\dots} Q_{im} Q_{jn} Q_{kr} \dots \quad , \quad (1.35)$$

$$\begin{aligned} R &= \frac{1}{n!} \epsilon^{ijk\dots} \epsilon_{mnr\dots} R_i^m R_j^n R_k^r \dots \quad , \\ &= \frac{1}{n!} \bar{\epsilon}^{ijk\dots} \underline{\epsilon}_{mnr\dots} R_i^m R_j^n R_k^r \dots \quad , \end{aligned} \quad (1.36)$$

$$\begin{aligned} S &= \frac{1}{n!} \epsilon_{ijk\dots} \epsilon^{mnr\dots} S^i_m S^j_n S^k_r \dots \quad , \\ &= \frac{1}{n!} \underline{\epsilon}_{ijk\dots} \bar{\epsilon}^{mnr\dots} S^i_m S^j_n S^k_r \dots \quad , \end{aligned} \quad (1.37)$$

and

$$T = \frac{1}{n!} \epsilon_{ijk\dots} \epsilon_{mnr\dots} T^{im} T^{jn} T^{kr} \dots \quad , \quad (1.38)$$

where the Levi-Civita’s tensors $\underline{\epsilon}_{ijk\dots}$, $\epsilon_{ijk\dots}$, $\bar{\epsilon}^{ijk\dots}$ and $\epsilon^{ijk\dots}$ have as many indices as the space under consideration has dimensions.

⁴It can be shown that this, indeed, a tensor, i.e., in a change of coordinates, it transforms like a tensor should.

1.5 The Kronecker Tensor

1.5.1 Kronecker Tensor

There are two Kronecker's "symbols", g_i^j and g^i_j . They are defined similarly:

$$g_i^j = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are the same index} \\ 0 & \text{if } i \text{ and } j \text{ are different indices,} \end{cases} \quad (1.39)$$

and

$$g^i_j = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are the same index} \\ 0 & \text{if } i \text{ and } j \text{ are different indices.} \end{cases} \quad (1.40)$$

Comment: I should be avoid this last notation.

It can easily be seen

(Comment: how?)

that g^i_j are more than 'symbols': they are *tensors*, in the sense that, if when changing the coordinates, we *compute* the new components of the Kronecker's tensors using the rules applying to all tensors, the property (Comment: which equation?) remains satisfied.

The Kronecker's tensors are defined even if the space has not a metric defined on it. Note that, sometimes, instead of using the symbols g_i^j and g^i_j to represent the Kronecker's tensors, the symbols δ_i^j and δ^i_j are used. But then, using the metric g_{ij} to "lower an index" of δ_i^j gives

$$\delta_{ij} = g_{jk}\delta_i^k = g_{ij}, \quad (1.41)$$

which means that, if the space has a metric, the Kronecker's tensor and the metric tensor are the same object. Why, then, use a different symbol? The use of the symbol δ_i^j may lead, by inadvertence, after lowering an index, to assing to δ_{ij} the value 1 when i and j are the same index. This is obviously wrong: if there is not a metric, δ_{ij} is not defined, and if there is a metric, δ_{ij} equals g_{ij} , which is only 1 in Euclidean spaces using Cartesian coordinates.

There is only one Kronecker's tensor, and g_i^j and g^i_j can be deduced one from the other raising and lowering indices. But, even in that case, we dislike the notation g^i_j , where the place of each index is not indicated, and we will not use it sistematically.

Warning: a common error in beginners is to give the value 1 to the symbol g_i^i (or to g^i_i). In fact, the right value is n , the dimension of the space, as there is an implicit sum assumed: $g_i^i = g_0^0 + g_1^1 + \dots = 1 + 1 + \dots = n$.

1.5.2 Kronecker Determinants

Let us denote by n the dimension of the space into consideration. The Levi-Civita's tensor has then n indices. For any (non-negative) integer p satisfying $p \leq n$, consider the integer q such that $p + q = n$. The following property holds:

$$\underline{\varepsilon}_{i_1 \dots i_p s_1 \dots s_q} \bar{\varepsilon}^{j_1 \dots j_p s_1 \dots s_q} = q! \det \begin{pmatrix} \delta_{i_1}^{j_1} & \delta_{i_1}^{j_2} & \dots & \delta_{i_1}^{j_p} \\ \delta_{i_2}^{j_1} & \delta_{i_2}^{j_2} & \dots & \delta_{i_2}^{j_p} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{i_p}^{j_1} & \delta_{i_p}^{j_2} & \dots & \delta_{i_p}^{j_p} \end{pmatrix}, \quad (1.42)$$

where δ_i^j stands for the Kronecker's tensor. The determinant at the right-hand side is called the *Kronecker's determinant*, and is denoted $\delta_{i_1 i_2 \dots i_p}^{j_1 j_2 \dots j_p}$:

$$\delta_{i_1 i_2 \dots i_p}^{j_1 j_2 \dots j_p} = \det \begin{pmatrix} \delta_{i_1}^{j_1} & \delta_{i_1}^{j_2} & \dots & \delta_{i_1}^{j_p} \\ \delta_{i_2}^{j_1} & \delta_{i_2}^{j_2} & \dots & \delta_{i_2}^{j_p} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{i_p}^{j_1} & \delta_{i_p}^{j_2} & \dots & \delta_{i_p}^{j_p} \end{pmatrix}. \quad (1.43)$$

As the Kronecker's determinant is defined as a product of Levi-Civita's tensors, it is itself a tensor. It generalizes the definition of the Kronecker's tensor δ_i^j , as it has the properties

$$\delta_{i_1 i_2 \dots i_m}^{j_1 j_2 \dots j_m} = \begin{cases} +1 & \text{if } (j_1, j_2, \dots, j_m) \text{ is an even permutation of } (i_1, i_2, \dots, i_m) \\ -1 & \text{if } (j_1, j_2, \dots, j_m) \text{ is an odd permutation of } (i_1, i_2, \dots, i_m) \\ 0 & \text{if two of the } i\text{'s or two of the } j\text{'s are the same index} \\ 0 & \text{if } (i_1, i_2, \dots, i_m) \text{ and } (j_1, j_2, \dots, j_m) \text{ are different sets of indices.} \end{cases} \quad (1.44)$$

As applying the same permutation to the indices of the two Levi-Civita's tensors of equation 1.42 will not change the total sign of the expression, we have

$$\begin{aligned} \underline{\varepsilon}_{i_1 \dots i_p s_1 \dots s_q} \overline{\varepsilon}^{j_1 \dots j_p s_1 \dots s_q} &= \\ \underline{\varepsilon}_{s_1 \dots s_q i_1 \dots i_p} \overline{\varepsilon}^{s_1 \dots s_q j_1 \dots j_p} &= q! \delta_{i_1 i_2 \dots i_p}^{j_1 j_2 \dots j_p}, \end{aligned} \quad (1.45)$$

but we only perform a permutation in one of the Levi-Civita's tensors, then we must care about the sign of the permutation, and we obtain

$$\begin{aligned} \underline{\varepsilon}_{i_1 \dots i_p s_1 \dots s_q} \overline{\varepsilon}^{s_1 \dots s_q j_1 \dots j_p} &= \\ \underline{\varepsilon}_{s_1 \dots s_q i_1 \dots i_p} \overline{\varepsilon}^{j_1 \dots j_p s_1 \dots s_q} &= (-1)^{pq} q! \delta_{i_1 i_2 \dots i_p}^{j_1 j_2 \dots j_p}. \end{aligned} \quad (1.46)$$

This possible change of sign has only effect in spaces with even dimension ($n = 2, 4, \dots$), as in spaces with odd dimension ($n = 3, 5, \dots$) the condition $p + q = n$ implies that pq is an even number, and $(-1)^{pq} = +1$.

Remark that a multiplication and a division by \overline{g} will not change the value of an expression, so that, instead of using Levi-Civita's density and capacity we can use Levi-Civita's true tensors. For instance,

$$\underline{\varepsilon}_{i_1 \dots i_p s_1 \dots s_q} \overline{\varepsilon}^{j_1 \dots j_p s_1 \dots s_q} = \varepsilon_{i_1 \dots i_p s_1 \dots s_q} \varepsilon^{j_1 \dots j_p s_1 \dots s_q}. \quad (1.47)$$

Comment: explain better.

Appendix 1.8.4 gives special formulas to spaces with dimension 2, 3, and 4. As shown in appendix 1.8.8, these formulas replace more elementary identities between grad, div, rot, ...

As an example, a well known identity like

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) \quad (1.48)$$

is obvious, as the three formulas correspond to the expression $\varepsilon_{ijk} a^i b^j c^k$. The identity

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c}) \mathbf{b} - (\mathbf{a} \cdot \mathbf{b}) \mathbf{c} \quad (1.49)$$

is easily demonstrated, as

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \varepsilon_{ijk} a^j (\mathbf{b} \times \mathbf{c})^k = \varepsilon_{ijk} a^j \varepsilon^{klm} b_l c_m, \quad (1.50)$$

which, using XXX, gives

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (a^m c_m) \mathbf{b} - (a^m b_m) \mathbf{c} = (\mathbf{a} \cdot \mathbf{c}) \mathbf{b} - (\mathbf{a} \cdot \mathbf{b}) \mathbf{c}. \quad (1.51)$$

Comment: I should clearly say here that we have the identity

$$\underline{\varepsilon}_{ijk\dots} \overline{\varepsilon}^{\ell mn\dots} = \varepsilon_{ijk\dots} \varepsilon^{\ell mn\dots}. \quad (1.52)$$

Comment: say somewhere that if $B_{i_1\dots i_p}$ is a totally antisymmetric tensor, then

$$\frac{1}{p!} \delta_{i_1\dots i_p}^{\ell_1\dots \ell_p} B_{\ell_1\dots \ell_p} = B_{i_1\dots i_p} \quad (1.53)$$

Comment: give somewhere the property

$$\frac{1}{q!} \delta_{i_1\dots i_p j_1\dots j_q}^{k_1\dots k_p \ell_1\dots \ell_q} \delta_{m_1\dots m_q}^{j_1\dots j_q} = \delta_{i_1\dots i_p m_1\dots m_q}^{k_1\dots k_p \ell_1\dots \ell_q}. \quad (1.54)$$

Comment: give somewhere the property

$$\frac{1}{q!} \underline{\varepsilon}_{i_1\dots i_p j_1\dots j_q} \delta_{k_1\dots k_q}^{j_1\dots j_q} = \underline{\varepsilon}_{i_1\dots i_p k_1\dots k_q}. \quad (1.55)$$

Note: Check if there are not factors $(-1)^{pq}$ missing.

1.6 Totally Antisymmetric Tensors

1.6.1 Totally Antisymmetric Tensors

A tensor is completely antisymmetric if any *even* permutation of indices does not change the value of the components, and if any *odd* permutation of indices changes the sign of the value of the components:

$$t_{pqr\dots} = \begin{cases} +t_{ijk\dots} & \text{if } ijk\dots \text{ is an } \textit{even} \text{ permutation of } pqr\dots \\ -t_{ijk\dots} & \text{if } ijk\dots \text{ is an } \textit{odd} \text{ permutation of } pqr\dots \end{cases} \quad (1.56)$$

For instance, a fourth rank tensor t_{ijkl} is totally antisymmetric if

$$\begin{aligned} t_{ijkl} &= t_{iklj} = t_{iljk} = t_{jilk} = t_{jkil} = t_{jlki} \\ &= t_{kijl} = t_{kjli} = t_{klji} = t_{likj} = t_{ljik} = t_{lkij} \\ &= -t_{ijlk} = -t_{ikjl} = -t_{iljk} = -t_{jikl} = -t_{jklj} = -t_{jlik} \\ &= -t_{kilj} = -t_{kjil} = -t_{klji} = -t_{lijk} = -t_{ljki} = -t_{lkij} \end{aligned} \quad (1.57)$$

a third rank tensor t_{ijk} is totally antisymmetric if

$$t_{ijk} = t_{jki} = t_{kji} = -t_{ikj} = -t_{jik} = -t_{kji}, \quad (1.58)$$

a second rank tensor t_{ij} is totally antisymmetric if

$$t_{ij} = -t_{ji}, \quad (1.59)$$

and a first rank tensor t_i can always be considered totally antisymmetric.

Well known examples of totally antisymmetric tensors are the Levi-Civita's tensors of any rank, the rank-two electromagnetic tensors, the “vector product” of two vectors:

$$c_{ij} = a_i b_j - a_j b_i, \quad (1.60)$$

etc.

Comment: say somewhere that the Kronecker's tensors and determinants are totally antisymmetric.

1.6.2 Dual Tensors

In a space with n dimensions, let p and q be two (nonnegative) integers such that $p + q = n$. To any totally antisymmetric tensor of rank p , $B^{i_1 \dots i_p}$, we can associate a totally antisymmetric tensor of rank q , $b_{i_1 \dots i_q}$, defined by

$$b_{i_1 \dots i_q} = \frac{1}{p!} \varepsilon_{i_1 \dots i_q j_1 \dots j_p} B^{j_1 \dots j_p}. \quad (1.61)$$

The tensor \mathbf{b} is called the *dual* of \mathbf{B} , and we write

$$\mathbf{b} = \text{Dual}[\mathbf{B}] \quad (1.62)$$

or

$$\mathbf{b} = {}^* \mathbf{B} \quad (1.63)$$

From the properties of the product of Levi-Civita's tensors it follows that the dual of the dual gives the original tensor, excepted for a sign:

$${}^*({}^* \mathbf{B}) = \text{Dual}[\text{Dual}[\mathbf{B}]] = (-1)^{p(n-p)} \mathbf{B}. \quad (1.64)$$

For spaces with odd dimension ($n = 1, 3, 5, \dots$), the product $p(n-p)$ is even, and

$${}^*({}^* \mathbf{B}) = \mathbf{B} \quad (\text{spaces with odd dimension}). \quad (1.65)$$

For spaces with even dimension ($n = 2, 4, 6, \dots$), we have

$${}^*({}^* \mathbf{B}) = (-1)^p \mathbf{B} \quad (\text{spaces with even dimension}). \quad (1.66)$$

Although definition 1.61 has been written for pure tensors, it can obviously be written for densities and capacities,

$$\begin{aligned} b_{i_1 \dots i_q} &= \frac{1}{p!} \varepsilon_{i_1 \dots i_q j_1 \dots j_p} \overline{B}^{j_1 \dots j_p} \\ \underline{b}_{i_1 \dots i_q} &= \frac{1}{p!} \underline{\varepsilon}_{i_1 \dots i_q j_1 \dots j_p} B^{j_1 \dots j_p}, \end{aligned} \quad (1.67)$$

or for tensor where covariant and contravariant indices have replaced each other:

$$\begin{aligned} d^{i_1 \dots i_q} &= \frac{1}{p!} \varepsilon^{i_1 \dots i_q j_1 \dots j_p} D_{j_1 \dots j_p} \\ d^{i_1 \dots i_q} &= \frac{1}{p!} \overline{\varepsilon}^{i_1 \dots i_q j_1 \dots j_p} \underline{D}_{j_1 \dots j_p} \\ \overline{d}^{i_1 \dots i_q} &= \frac{1}{p!} \overline{\varepsilon}^{i_1 \dots i_q j_1 \dots j_p} D_{j_1 \dots j_p}, \end{aligned} \quad (1.68)$$

Appendix 1.8.13 gives explicitly the dual tensor relations in spaces with 2, 3, and 4 dimensions.

Example 1.3 Consider an antisymmetric tensor E_{ij} in three dimensions. It has components

$$\begin{pmatrix} E_{11} & E_{12} & E_{13} \\ E_{21} & E_{12} & E_{23} \\ E_{31} & E_{32} & E_{33} \end{pmatrix} = \begin{pmatrix} 0 & E_{12} & E_{13} \\ E_{21} & 0 & E_{23} \\ E_{31} & E_{32} & 0 \end{pmatrix}, \quad (1.69)$$

with $E_{ij} = -E_{ji}$. The definition

$$\overline{e}^i = \frac{1}{2!} \overline{\varepsilon}^{ijk} E_{jk} \quad (1.70)$$

gives

$$\begin{pmatrix} 0 & E_{12} & E_{13} \\ E_{21} & 0 & E_{23} \\ E_{31} & E_{32} & 0 \end{pmatrix} = \begin{pmatrix} 0 & \overline{e}^3 & -\overline{e}^2 \\ -\overline{e}^3 & 0 & \overline{e}^1 \\ \overline{e}^2 & -\overline{e}^1 & 0 \end{pmatrix}, \quad (1.71)$$

which is the classical relation between the three independent components of a 3-D antisymmetric tensor and the components of a vector density. **[End of example.]**

Example 1.4 The vector product of two vectors U_i and V_i can be either defined as the antisymmetric tensor

$$W_{ij} = U_i V_j - V_j U_i, \quad (1.72)$$

or as the vector density

$$\bar{w}^i = \frac{1}{2!} \bar{\varepsilon}^{ijk} U_j V_k. \quad (1.73)$$

The two definitions are equivalent, as W_{ij} and \bar{w}^i are mutually duals. **[End of example.]**

Definition 1.73 shows that the vector product of two vectors is not a pure vector, but a vector density. Changing the sense of one axis gives a Jacobian equal to -1 , thus *changing the sign of the vector product* \bar{w}^i .

1.6.3 Exterior Product of Tensors

In a space of dimension n , let $A_{i_1 i_2 \dots i_p}$ and $B_{i_1 i_2 \dots i_q}$, be two totally antisymmetric tensors with ranks p and q such that $p + q \leq n$. Note: check that total antisymmetry has been defined. The *exterior product* of the two tensors is denoted

$$\mathbf{C} = \mathbf{A} \wedge \mathbf{B} \quad (1.74)$$

and is the totally antisymmetric tensor of rank $p + q$ defined by

$$C_{i_1 \dots i_p j_1 \dots j_q} = \frac{1}{(p+q)!} \delta_{i_1 \dots i_p j_1 \dots j_q}^{k_1 \dots k_p \ell_1 \dots \ell_q} A_{k_1 i_2 \dots k_p} B_{\ell_1 i_2 \dots \ell_q}. \quad (1.75)$$

Permuting the set of indices $\{k_1 \dots k_p\}$ by the set $\{\ell_1 \dots \ell_q\}$ in the above definition gives the property

$$(\mathbf{A} \wedge \mathbf{B}) = (-1)^{pq} (\mathbf{B} \wedge \mathbf{A}). \quad (1.76)$$

It is also easy to see that the associativity property holds:

$$\mathbf{A} \wedge (\mathbf{B} \wedge \mathbf{C}) = (\mathbf{A} \wedge \mathbf{B}) \wedge \mathbf{C}. \quad (1.77)$$

Comment: say that $\delta_{i_1 i_2 \dots}^{j_1 j_2 \dots}$ are the components of the Kronecker's determinant defined in Section 1.5.2.

Say that it equation 1.54 gives the property

$$(\mathbf{A1} \wedge \mathbf{A2} \wedge \dots \wedge \mathbf{AP})_{i_1 i_2 \dots i_p} = \frac{1}{p!} \delta_{i_1 i_2 \dots i_p}^{j_1 j_2 \dots j_p} A1_{j_1} A2_{j_2} \dots AP_{j_p}. \quad (1.78)$$

1.6.3.1 Particular cases:

It follows from equation 1.53 that the exterior product of a tensor of rank zero (a scalar) by a totally antisymmetric tensor of any order is the simple product of the scalar by the tensor:

$$(A \quad , \quad B_{i_1 \dots i_q}) \quad \rightarrow \quad (\mathbf{A} \wedge \mathbf{B})_{i_1 \dots i_q} = A B_{i_1 \dots i_q}. \quad (1.79)$$

For the exterior product of two vectors we easily obtain (independently of the dimension of the space into consideration)

$$(A_i \quad , \quad B_j) \quad \rightarrow \quad (\mathbf{A} \wedge \mathbf{B})_{ij} = \frac{1}{2} (A_i B_j - A_j B_i). \quad (1.80)$$

The exterior product of a vector by a second rank (antisymmetric) tensor gives

$$(A_i \quad , \quad B_{ij}) \quad \rightarrow \quad (\mathbf{A} \wedge \mathbf{B})_{ijk} = \frac{1}{3} (A_i B_{jk} + A_j B_{ki} + A_k B_{ij}). \quad (1.81)$$

Finally, it can be seen that the exterior product of three vectors gives

$$\begin{aligned} (A_i \quad , \quad B_j \quad , \quad C_k) &\quad \rightarrow \quad (1.82) \\ (\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C})_{ijk} &= \frac{1}{6} (A_i (B_j C_k - B_k C_j) + A_j (B_k C_i - B_i C_k) + A_k (B_i C_j - B_j C_i)) \\ &= \frac{1}{6} (B_i (C_j A_k - C_k A_j) + B_j (C_k A_i - C_i A_k) + B_k (C_i A_j - C_j A_i)) \\ &= \frac{1}{6} (C_i (A_j B_k - A_k B_j) + C_j (A_k B_i - A_i B_k) + C_k (A_i B_j - A_j B_i)). \end{aligned}$$

Let us examine with more detail the formulas above in the special case of a 3-D space.

The dual of the exterior product of two vectors (equation 1.80) gives

$$*\overline{(\mathbf{a} \wedge \mathbf{b})}^i = \frac{1}{2} \bar{\varepsilon}^{ijk} a_j b_k, \quad (1.83)$$

i.e., one half the usual vector product of the two vectors:

$$*\overline{(\mathbf{a} \wedge \mathbf{b})} = \frac{1}{2} \overline{(\mathbf{a} \times \mathbf{b})}. \quad (1.84)$$

The dual of the exterior product of a vector by a second rank (antisymmetric) tensor (equation 1.81) is

$$*\overline{(\mathbf{a} \wedge \mathbf{b})} = \frac{1}{3} a_i \left(\frac{1}{2!} \bar{\varepsilon}^{ijk} b_{jk} \right), \quad (1.85)$$

or, introducing the vector $*\bar{b}^i$, dual of the tensor b_{ij} ,

$$*\overline{(\mathbf{a} \wedge \mathbf{b})} = \frac{1}{3} a_i * \bar{b}^i. \quad (1.86)$$

This shows that the exterior product contains, via the duals, the contraction of a form and a vector.

Finally, the dual of the exterior product of three vectors (equation 1.82) is

$$*\overline{(\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c})} = \frac{1}{3!} \bar{\varepsilon}^{ijk} a_i b_j c_k, \quad (1.87)$$

i.e., one sixth of the triple product of the three vectors.

Comment: explain that the triple product of three vectors is $\mathbf{a} \cdot \overline{(\mathbf{b} \times \mathbf{c})} = \mathbf{b} \cdot \overline{(\mathbf{c} \times \mathbf{a})} = \mathbf{c} \cdot \overline{(\mathbf{a} \times \mathbf{b})}$.

1.6.4 Exterior Derivative of Tensors

Let \mathbf{T} be a totally antisymmetric tensor with components $T_{i_1 i_2 \dots i_p}$. The exterior product of “nabla” with \mathbf{T} is called the *exterior derivative* of \mathbf{T} , and is denoted $\nabla \wedge \mathbf{T}$:

$$(\nabla \wedge \mathbf{T})_{i_1 j_2 \dots j_p} = \delta_{i_1 j_2 \dots j_p}^{k \ell_1 \ell_2 \dots \ell_p} \nabla_k T_{\ell_1 \ell_2 \dots \ell_p}. \quad (1.88)$$

Here, $\nabla_i T_{j k \dots}$ denotes the covariant derivative defined in section XXX.

The “nabla” notation allows to use directly the formulas developed for the exterior product of a vector by a tensor to obtain formulas for exterior derivatives. For instance, from equation 1.80 it follows the definition of the exterior derivative of a vector

$$(\nabla \wedge \mathbf{b})_{ij} = \frac{1}{2} (\nabla_i b_j - \nabla_j b_i), \quad (1.89)$$

or, if we use the dual (equations 1.83–1.84),

$$*(\nabla \wedge \mathbf{b})^i = \frac{1}{2} \bar{\varepsilon}^{ijk} \nabla_j b_k, \quad (1.90)$$

i.e.,

$$*(\nabla \wedge \mathbf{b}) = \frac{1}{2} \overline{(\nabla \times \mathbf{b})}. \quad (1.91)$$

The exterior derivative of a vector equals one-half the rotational (curl) of the vector.

The exterior derivative of a second rank (antisymmetric) tensor is directly obtained from equation 1.81:

$$(\nabla \wedge \mathbf{b})_{ijk} = \frac{1}{3} (\nabla_i b_{jk} + \nabla_j b_{ki} + \nabla_k b_{ij}). \quad (1.92)$$

Taking the dual of the expression and introducing the vector $*\bar{b}^i$, dual of the tensor b_{ij} , gives (see equation 1.86)

$$*(\nabla \wedge \mathbf{b}) = \frac{1}{3} \nabla_i *\bar{b}^i, \quad (1.93)$$

which shows that the dual of the exterior derivative of a second rank (antisymmetric) tensor equals one-third of the divergence of the dual of the tensor. The exterior derivative contains, via the duals, the divergence of a vector.

1.7 Integration, Volumes

1.7.1 The Volume Element

Consider, in a space with n dimensions, p linearly independent vectors $\{\mathbf{dr}_1, \mathbf{dr}_2, \dots, \mathbf{dr}_p\}$. As they are linear independent, $p \leq n$.

We define the “differential element”

$$d^{(p)}\boldsymbol{\sigma} = p! (\mathbf{dr}_1 \wedge \mathbf{dr}_2 \wedge \dots \wedge \mathbf{dr}_p) \quad . \quad (1.94)$$

Using equation 1.78 (Note: in fact this equation with indices changed of place) gives the components

$$d^{(p)}\sigma^{i_1 \dots i_p} = \delta_{j_1 \dots j_p}^{i_1 \dots i_p} dr_1^{j_1} dr_2^{j_2} \dots dr_p^{j_p} \quad . \quad (1.95)$$

In a space with n dimensions, the dual of the differential element of dimension p will have q indices, with $p + q = n$. The general definition of dual (equation 1.67) gives

$${}^*d^{(p)}\underline{\sigma}_{i_1 \dots i_q} = \frac{1}{p!} \underline{\varepsilon}_{i_1 \dots i_q j_1 \dots j_p} d^{(p)}\sigma^{j_1 \dots j_p} \quad (1.96)$$

The definition 1.95 and the property 1.55 give

$${}^*d^{(p)}\underline{\sigma}_{i_1 \dots i_q} = \underline{\varepsilon}_{i_1 \dots i_q j_1 \dots j_p} dr_1^{j_1} dr_2^{j_2} \dots dr_p^{j_p} \quad . \quad (1.97)$$

In order to simplify subsequent notations, it is better not to keep the $*$ notation. Instead, we will write

$${}^*d^{(p)}\underline{\sigma}_{i_1 \dots i_q} = d^{(p)}\underline{\Sigma}_{i_1 \dots i_q} \quad (1.98)$$

For reasons to be developed below, $d^{(p)}\underline{\Sigma}_{i_1 \dots i_q}$ will be called the *capacity element*.

We can easily see, for instance, that the differential elements of dimensions 0, 1, 2 and 3 have components

$$d^0\sigma = 1 \quad (1.99)$$

$$d^1\sigma^i = dr_1^i \quad (1.100)$$

$$d^2\sigma^{ij} = dr_1^i dr_2^j - dr_1^j dr_2^i \quad (1.101)$$

$$\begin{aligned} d^3\sigma^{ijk} &= dr_1^i (dr_2^j dr_3^k - dr_2^k dr_3^j) + dr_1^j (dr_2^k dr_3^i - dr_2^i dr_3^k) + dr_1^k (dr_2^i dr_3^j - dr_2^j dr_3^i) \\ &= dr_2^i (dr_3^j dr_1^k - dr_3^k dr_1^j) + dr_2^j (dr_3^k dr_1^i - dr_3^i dr_1^k) + dr_2^k (dr_3^i dr_1^j - dr_3^j dr_1^i) \\ &= dr_3^i (dr_1^j dr_2^k - dr_1^k dr_2^j) + dr_3^j (dr_1^k dr_2^i - dr_1^i dr_2^k) + dr_3^k (dr_1^i dr_2^j - dr_1^j dr_2^i) \quad . \end{aligned} \quad (1.102)$$

For a given dimension of the differential element, the number of indices of the capacity elements depends on the dimension of the space. **In a three-dimensional space**, for instance, we have

$$d^0\underline{\Sigma}_{ijk} = \underline{\varepsilon}_{ijk} \quad (1.103)$$

$$d^1\underline{\Sigma}_{ij} = \underline{\varepsilon}_{ijk} dr_1^k \quad (1.104)$$

$$d^2\underline{\Sigma}_i = \underline{\varepsilon}_{ijk} dr_1^j dr_2^k \quad (1.105)$$

$$d^3\underline{\Sigma} = \underline{\varepsilon}_{ijk} dr_1^i dr_2^j dr_3^k \quad . \quad (1.106)$$

Note: explain that I use the notation $d^{(p)}$ but d^1, d^2, \dots in order not to suggest that p is a tensor index and, at the same time, for not using too heavy notations..

Note: refer here to figure 1.1, and explain that we have, in fact, vector products of vectors and triple products of vectors.

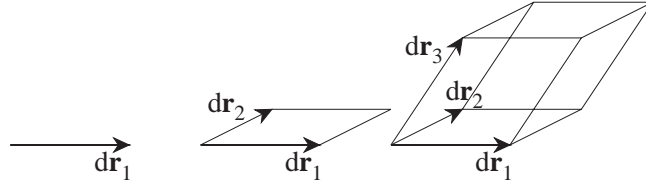


Figure 1.1: From vectors in a three-dimensional space we define the one-dimensional capacity element $d^1\underline{\Sigma}_{ij} = \varepsilon_{ijk} dr_1^k$, the two-dimensional capacity element $d^2\underline{\Sigma}_i = \varepsilon_{ijk} dr_1^j dr_2^k$ and the three-dimensional capacity element $d^3\underline{\Sigma} = \varepsilon_{ijk} dr_1^i dr_2^j dr_3^k$. In a metric space, the rank-two form $d^1\underline{\Sigma}_{ij}$ defines a surface perpendicular to dr_1 and with a surface magnitude equal to the length of dr_1 . The rank-one form $d^2\underline{\Sigma}_i$ defines a vector perpendicular to the surface defined by dr_1 and dr_2 and with length representing the surface magnitude (the vector product of the two vectors). The rank-zero form $d^3\underline{\Sigma}$ is a scalar representing the volume defined by the three vectors dr_1 , dr_2 and dr_3 (the triple product of the vectors). Note: clarify all this.

1.7.2 The Stokes' Theorem

Comment: I must explain here first what integration means.

Let, in a space with n dimensions, (\mathbf{T}) be a totally antisymmetric tensor of rank p , with $(p < n)$. The Stokes' theorem

$$\boxed{\int_{(p+1)D} d^{(p+1)}\sigma^{i_1 \dots i_{p+1}} (\nabla \wedge \mathbf{T})_{i_1 \dots i_{p+1}} = \int_{pD} d^{(p)}\sigma^{i_1 \dots i_p} T_{i_1 \dots i_p}} \quad (1.107)$$

holds. Here, the symbol $\int_{(p+1)D} d^{(p+1)}$ stands for an integral over a $(p+1)$ -dimensional “volume”, (embedded in an space of dimension n), and $\int_{pD} d^{(p)}$ for the integral over the p -dimensional boundary of the “volume”.

This fundamental theorem contains, as special cases, the divergence theorem of Gauss-Ostrogradsky, and the rotational theorem of Stokes (stricto sensu). Rather than deriving it here, we will explore its consequences. For a demonstration, see, for instance, Von Westenholz (1981).

In a three-dimensional space ($n = 3$), we may have p respectively equal to 2, 1 and 0. This gives the three theorems

$$\int_{3D} d^3\sigma^{ijk} (\nabla \wedge \mathbf{T})_{ijk} = \int_{2D} d^2\sigma^{ij} T_{ij} \quad (1.108)$$

$$\int_{2D} d^2\sigma^{ij} (\nabla \wedge \mathbf{T})_{ij} = \int_{1D} d^1\sigma^i T_i \quad (1.109)$$

$$\int_{1D} d^1\sigma^i (\nabla \wedge \mathbf{T})_i = \int_{0D} d^0\sigma T. \quad (1.110)$$

It is easy to see (appendix 1.8.14) that these equation can be written

$$\frac{1}{0!} \int_{3D} d^3\underline{\Sigma} \left(\frac{1}{2!} \bar{\varepsilon}^{ijk} \nabla_i T_{jk} \right) = \frac{1}{1!} \int_{2D} d^2\underline{\Sigma}_i \left(\frac{1}{2!} \bar{\varepsilon}^{ijk} T_{jk} \right) \quad (1.111)$$

$$\frac{1}{1!} \int_{2D} d^2\underline{\Sigma}_i \left(\frac{1}{1!} \bar{\varepsilon}^{ijk} \nabla_j T_k \right) = \frac{1}{2!} \int_{1D} d^1\underline{\Sigma}_{ij} \left(\frac{1}{1!} \bar{\varepsilon}^{ijk} T_k \right) \quad (1.112)$$

$$\frac{1}{2!} \int_{1D} d^1\underline{\Sigma}_{ij} \left(\frac{1}{0!} \bar{\varepsilon}^{ijk} \partial_k T \right) = \frac{1}{3!} \int_{0D} d^0\underline{\Sigma}_{ijk} \left(\frac{1}{0!} \bar{\varepsilon}^{ijk} T \right). \quad (1.113)$$

Simplifying equation 1.111 and introducing the vector density \bar{t}^i , dual to the tensor T_{ij} , (i.e., $\bar{t}^i = \frac{1}{2!} \bar{\varepsilon}^{ijk} T_{jk}$), gives

$$\int_{3D} d^3\underline{\Sigma} \nabla_i \bar{t}^i = \int_{2D} d^2\underline{\Sigma}_i \bar{t}^i. \quad (1.114)$$

This corresponds to the divergence theorem of Gauss-Ostrogradsky: The integral over a (3-D) volume of the divergence of a vector equals the flux of the vector across the surface bounding the volume.

It is worth to mention here that expression 1.114 has been derived without any mention to a metric in the space. We have sen elsewhere that densities and capacities can be defined even if there is no notion of distance. If there is a metric, then from the capacity element $d^3\underline{\Sigma}$ we can introduce the *volume element* $d^3\Sigma$ using the standard rule for putting on and taking off bars

$$d^3\Sigma = \bar{g} d^3\underline{\Sigma}, \quad (1.115)$$

as well as the *surface element*

$$d^2\Sigma_i = \bar{g} d^2\underline{\Sigma}_i. \quad (1.116)$$

$d^3\Sigma$ is now the familiar volume inside a prism, and $d^2\Sigma_i$ the vector (if we raise the index with the metric) representing the surface inside a lozenge.

Equation 1.114 then gives

$$\int_{3D} d^3\Sigma \nabla_i t^i = \int_{2D} d^2\Sigma_i t^i, \quad (1.117)$$

which is the familiar form for the divergence theorem.

Keeping the compact expression for the capacity element in the lefthand side of equation 1.112, but introducing its explicit expression in the right hand side gives, after simplification,

$$\int_{2D} d^2\underline{\Sigma}_i (\bar{\varepsilon}^{ijk} \nabla_j T_k) = \int_{1D} dr_1^i T_i, \quad (1.118)$$

which corresponds to the rotational theorem (theorem of Stokes stricto sensu): the integral of the rotational (curl) of a vector on a surface equals the circulation of the vector along the line bounding the surface.

Finally, introducing explicit expressions for the capacity elements at both sides of equation 1.113 gives

$$\int_{1D} dr^i \partial_i T = \int_{0D} T. \quad (1.119)$$

Writing this in the more familiar form gives

$$\int_a^b dr^i \partial_i T = T(b) - T(a), \quad (1.120)$$

which corresponds the fundamental theorem of integral calculus: the integral over a line of the gradient of a scalar equals the difference of the values of the scalar at the two end-points.

Note: say that more details can be found in appendix 1.8.14

Comment: explain here what the “capacity element” is. Explain that, in polar coordinates, it is given by $drd\varphi$, to be compared with the “surface element” $rdrd\varphi$. Comment figure 1.2.

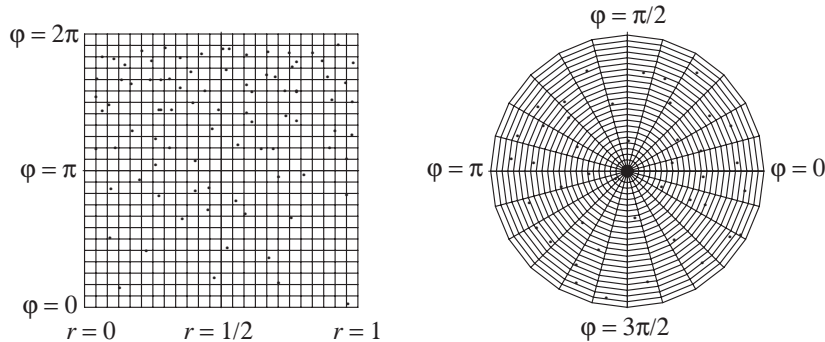


Figure 1.2: We consider, in an Euclidean space, a cylinder with a circular basis of radius 1, and cylindrical coordinates (r, φ, z) . Only a section of the cylinder is represented in the figure, with all its thickness, dz , projected on the drawing plane. At left, we have represented a “map” of the corresponding circle, and, at right, the coordinate lines on the circle itself. All the “cells” at left have the same capacity $d\underline{V} = drd\varphi dz$, while the cells at right have the volume $dV(r, \varphi, z) = r drd\varphi dz$. The points represent particles with given masses. If, at left, at point with coordinates (r, φ, z) the sum of all the masses inside the local cell is denoted, dM , then, the mass *density* at this point is estimated by $\bar{\rho}(r, \varphi, z) = dM/d\underline{V}$, i.e., $\bar{\rho}(r, \varphi) = dM/(drd\varphi dz)$. If, at right, at point (r, φ, z) the total mass inside the local cell is dM , the *volumetric* mass at this point is estimated by $\rho(r, \varphi, z) = dM/dV(r, \varphi, z)$, i.e., $\rho(r, \varphi, z) = dM/(r drd\varphi dz)$. By definition, then, the total mass inside a volume V will be found by $M = \int_V d\underline{V} \bar{\rho}(r, \varphi, z) = \int_V drd\varphi dz \bar{\rho}(r, \varphi, z)$ or by $M = \int_V dV(r, \varphi, z) \rho(r, \varphi, z) = \int_V r drd\varphi dz \rho(r, \varphi, z)$.

1.8 Appendixes

1.8.1 Appendix: Tensors For Beginners

1.8.1.1 Tensor Notations

The velocity of the wind at the top of Eiffel's tower, at a given moment, can be represented by a *vector* \mathbf{v} with components, in some local, given, basis, $\{v^i\}$ ($i = 1, 2, 3$). The velocity of the wind is defined at any point \mathbf{x} of the atmosphere at any time t : we have a *vector field* $v^i(\mathbf{x}, t)$.

The water's temperature at some point in the ocean, at a given moment, can be represented by a *scalar* T . The field $T(\mathbf{x}, t)$ is a *scalar field*.

The state of stress at a given point of the Earth's crust, at a given moment, is represented by a *second order tensor* $\boldsymbol{\sigma}$ with components $\{\sigma^{ij}\}$ ($i = 1, 2, 3; j = 1, 2, 3$). In a general model of continuous media, where it is not assumed that the stress tensor is symmetric, this means that we need 9 scalar quantities to characterize the state of stress. In more particular models, the stress tensor is symmetric, $\sigma^{ij} = \sigma^{ji}$, and only six scalar quantities are needed. The stress field $\sigma^{ij}(\mathbf{x}, t)$ is a *second order tensor field*.

Tensor fields can be combined, to give other fields. For instance, if n_i is a unit vector considered at a point inside a medium, the vector

$$\tau^i(\mathbf{x}, t) = \sum_{j=1}^3 \sigma^{ij}(\mathbf{x}, t) n_j(\mathbf{x}) = \sigma^{ij}(\mathbf{x}, t) n_j(\mathbf{x}) \quad ; \quad (i = 1, 2, 3) \quad (1.121)$$

represents the traction that the medium at one side of the surface defined by the normal n_i exerts the medium at the other side, at the considered point.

As a further example, if the deformations of an elastic solid are small enough, the stress tensor is related linearly to the strain tensor (Hooke's law). A linear relation between two second order tensors means that each component of one tensor can be computed as a linear combination of all the components of the other tensor:

$$\sigma^{ij}(\mathbf{x}, t) = \sum_{k=1}^3 \sum_{\ell=1}^3 c^{ijkl}(\mathbf{x}) \varepsilon_{k\ell}(\mathbf{x}, t) = c^{ijkl}(\mathbf{x}) \varepsilon_{k\ell}(\mathbf{x}, t) \quad ; \quad (i = 1, 2, 3; j = 1, 2, 3). \quad (1.122)$$

The *fourth order tensor* c^{ijkl} represents a property of an elastic medium: its elastic stiffness. As each index takes 3 values, there are $3 \times 3 \times 3 \times 3 = 81$ scalars to define the elastic stiffness of a solid at a point (assuming some symmetries we may reduce this number to 21, and assuming isotropy of the medium, to 2).

We are yet interested in the physical meaning of the equations above, but in their structure. First, tensor notations are such that they are independent on the coordinates being used. This is not obvious, as changing the coordinates implies changing the local basis where the components of vectors and tensors are expressed. That the two equalities above hold for any coordinate system, means that all the components of all tensors will change if we change the coordinate system being used (for instance, from Cartesian to spherical coordinates), but still the two sides of the expression will take equal values.

The mechanics of the notation, once understood, are such that it is only possible to write expressions that make sense (see a list of rules at the end of this section).

For reasons about to be discussed, indices may come in upper or lower positions, like in v^i , f_i or T_i^j . The definitions will be such that in all tensor expression (i.e., in all expressions that will be valid for all coordinate systems), the sums over indices will always concern an index in lower position an one index on upper position. For instance, we may encounter expressions like

$$\varphi = \sum_{i=1}^3 A_i B^i = A_i B^i \quad \text{or} \quad A_i = \sum_{j=1}^3 \sum_{k=1}^3 D_{ijk} E^{jk} = D_{ijk} E^{jk} \quad . \quad (1.123)$$

These two equations (as equations 1.121 and 1.122) have been written in two version, one with the sums over the indices explicitly indicated, and another where this sum is implicitly assumed. This implicit notation is useful as one easily forgets that one is dealing with sums, and that it happens that, with respect to the usual tensor operations (sum with another tensor field, multiplication with another tensor field, and derivation), a sum of such terms is handled as one single term of the sum could be handled.

In an expression like $A_i = D_{ijk} E^{jk}$ it is said that the indices j and k have been *contracted* (or are “dummy indices”), while the index i is a *free index*. A tensor equation is assumed to hold for all possible values of the free indices.

In some spaces, like our physical 3-D space, it is posible to define the distance between two points, and in such a way that, in a local system of coordinates, approximately Cartesian, the distance has approximately the Euclidean form (square root of a sum of squares). These spaces are called *metric spaces*. A mathematically convenient manner to introduce a metric is by defining the length of an arc Γ by $S = \int_{\Gamma} ds$, where, for instance, in Cartesian coordinates, $ds^2 = dx^2 + dy^2 + dz^2$ or, in spherical coordinates, $ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\varphi^2$. In general, we write $ds^2 = g_{ij} dx^i dx^j$, and we call $g_{ij}(\mathbf{x})$ the *metric field* or, simply, the *metric*.

The components of a vector \mathbf{v} are associated to a given basis (the vector will have different components on different basis). If a basis \mathbf{e}_i is given, then, the components v^i are defined through $\mathbf{v} = v^i \mathbf{e}_i$ (implicit sum). The *dual basis* of the basis $\{\mathbf{e}_i\}$ is denoted $\{\mathbf{e}^i\}$ and is defined by the equation $\mathbf{e}_i \mathbf{e}^j = \delta_i^j$ (equal to 1 if i are the same index and to 0 if not). When there is a metric, this equation can be interpreted as a scalar vector product, and the dual basis is just another basis (identical to the first one when working with Cartesian coordinates in Eucliden spaces, but different in general). The properties of the dual basis will be analyzed later in the chapter. Here we just need to recall that if v^i are the components of the vector \mathbf{v} on the basis $\{\mathbf{e}_i\}$ (remember the expression $\mathbf{v} = v^i \mathbf{e}_i$), we will denote by v_i are the components of the vector \mathbf{v} on the basis $\{\mathbf{e}^i\}$: $\mathbf{v} = v_i \mathbf{e}^i$. In that case (metric spaces) the components on the two basis are related by $v_i = g_{ij} v^j$: It is said that “the metric tensor ascends (or descends) the indices”.

Here is a list with some rules helping to recognize tensor equations:

- A tensor expression must have the same *free* indices, at the top and at the bottom, of the two sides of an equality. For instance, the expressions

$$\begin{aligned} \varphi &= A_i B^i \\ \varphi &= g_{ij} B^i C^j \\ A_i &= D_{ijk} E^{jk} \\ D_{ijk} &= \nabla_i F_{jk} \end{aligned} \quad (1.124)$$

are valid, but the expressions

$$\begin{aligned} A_i &= F_{ij} B^j \\ B^i &= A_j C^j \\ A_i &= B^i \end{aligned} \tag{1.125}$$

are not.

- Sum and multiplication of tensors (with eventual “contraction” of indices) gives tensors. For instance, if D_{ijk} , G_{ijk} and H_i^j are tensors, then

$$\begin{aligned} J_{ijk} &= D_{ijk} + G_{ijk} \\ K_{ijk\ell}^m &= D_{ijk} H_\ell^m \\ L_{ik\ell} &= D_{ijk} H_\ell^j \end{aligned} \tag{1.126}$$

also are tensors.

- True (or “covariant”) derivatives of tensor fields give tensor fields. For instance, if E^{ij} is a tensor field, then

$$\begin{aligned} M_i^{jk} &= \nabla_i E^{jk} \\ B^j &= \nabla_i E^{ij} \end{aligned} \tag{1.127}$$

also are tensor fields. But partial derivatives of tensors do not define, in general, tensors. For instance, if E^{ij} is a tensor field, then

$$\begin{aligned} M_i^{jk} &= \partial_i V^{jk} \\ B^j &= \partial_i V^{ij} \end{aligned} \tag{1.128}$$

are not tensors, in general.

- All “objects with indices” that are normally introduced are tensors, with four notable exceptions. The first exception are the coordinates $\{x^i\}$ (to see that it makes no sense to add coordinates, think, for instance, in adding the spherical coordinates of two points). But the differentials dx^i appearing in an expression like $ds^2 = g_{ij} dx^i dx^j$ do correspond to the components on a vector $d\mathbf{r} = dx^i \mathbf{e}_i$. Another notable exception is the “symbol” ∂_i mentioned above. The third exception is the “connection” Γ_{ij}^k to be introduced later in the chapter. In fact, it is because both of the symbols ∂_i and Γ_{ij}^k are not tensors than an expression like

$$\nabla_i V^j = \partial_i V^j + \Gamma_{ik}^j V^k \tag{1.129}$$

can have a tensorial sense: if one of the terms at right was a tensor and not the other, their sum could never give a tensor. The objects ∂_i and Γ_{ij}^k are both non tensors, and “what one term misses, the other term has”. The fourth and last case of “objects with indices” which are not tensors are the Jacobian matrices arising in coordinate changes $\mathbf{x} \rightleftharpoons \mathbf{y}$,

$$J^i_I = \frac{\partial x^i}{\partial y^J}. \tag{1.130}$$

That this is not a tensor is obvious when considering that, contrarily to a tensor, the Jacobian matrix is not defined per se, but it is only defined when two different coordinate systems have been chosen. A tensor exists even if no coordinate system at all has been defined.

1.8.1.2 Differentiable Manifolds

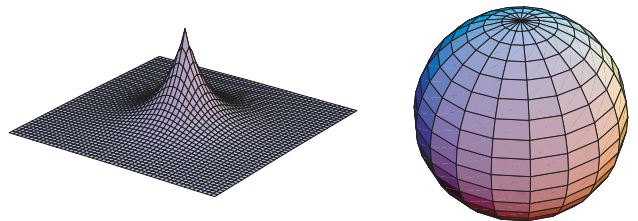
A manifold is a continuous space of points. In an n -dimensional manifold it is always possible to “draw” *coordinate lines* in such a way that to any point \mathcal{P} of the manifold correspond coordinates $\{x^1, x^2, \dots, x^n\}$ and vice versa.

Saying that the manifold is a continuous space of points is equivalent to say that the coordinates themselves are “continuous”, i.e., if they are, in fact, a part of \mathcal{R}^n . On such manifolds we define physical fields, and the continuity of the manifold will allow to define the derivatives of the considered fields. When derivatives of fields on a manifold can be defined, the manifold is then called a *differentiable manifold*.

Obvious examples of differentiable manifolds are the lines and surfaces of ordinary geometry. Our 3-D physical space (with, possibly, curvature and torsion) is also represented by a differentiable manifold. The space-time of general relativity is a four dimensional differentiable manifold.

A coordinate system may not “cover” all the manifold. For instance, the poles of a sphere are as ordinary as any other point in the sphere, but the coordinates are singular there (the coordinate φ is not defined). Changing the coordinate system around the poles will make any problem related to the coordinate choice to vanish there. A more serious difficulty appears when at some point, not the coordinates, but the manifold itself is singular (the linear tangent space is not defined at this point), as for instance, in the example shown in figure 1.3. Those are named “essential singularities”. No effort will be made on this book to classify them.

Figure 1.3: The surface at left has an essential singularity that will cause trouble for whatever system of coordinates we may choose (the tangent linear space is not defined at the singular point). The sphere at right has no essential singularity, but the coordinate system chosen is singular at the two poles. Other coordinate systems will be singular at different points.



1.8.1.3 Tangent Linear Space, Tensors.

Consider, for instance, in classical dynamics, a trajectory $x^i(t)$ on a space which may not be flat, as the surface of a sphere. The trajectory is “on” the sphere. If we define now the velocity at some point,

$$v^i = \frac{dx^i}{dt}, \quad (1.131)$$

we get a vector which is not “on” the sphere, but *tangent* to it. It belongs to what is called the *tangent linear space* to the considered point. At that point, we will have a basis for vectors. At another point, we will have another tangent linear space, and another vector basis.

More generally, at every point of a differential manifold, we can consider different vector or tensor quantities, like the *forces*, *velocities*, or *stresses* of mechanics of continuous media. As suggested by figure 1.4, those tensorial objects do not belong to the nonlinear manifold, but to the *tangent linear space* to the manifold at the considered point (that will only be introduced intuitively here).

At every point of an space, tensors can be added, multiplied by scalars, contracted, etc. This means that at every point of the manifold we have to consider a different vector space (in general, a tensor space). It is important to understand that two tensors at two different points of the space belong to two different tangent spaces, and can not be added as such (see figure 1.4). This is why we will later need to introduce the concept of “parallel transport of tensors”.

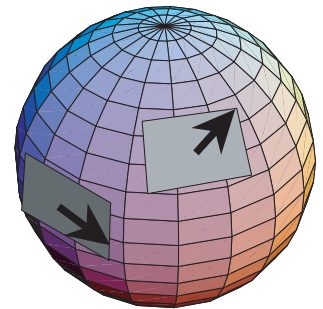
All through this book, the two names *linear space* and *vector space* will be used as completely equivalent.

The structure of vector space is too narrow to be of any use in physics. What is needed is the structure where equations like

$$\begin{aligned}\lambda &= R_i S^i \\ T^j &= U_i V^{ij} + \mu W^j \\ X^{ij} &= Y^i Z^j\end{aligned}\tag{1.132}$$

make sense. This structure is that of a *tensor space*. In short, a tensor space is a collection of vector spaces and rules of multiplication and differentiation that use elements of the vector spaces considered to get other elements of other vector spaces.

Figure 1.4: Surface with two planes tangent at two points, and a vector drawn at each point. As the vectors belong to two different vector spaces, their sum is not defined. Should we need to add them, for instance, to define true (or “covariant”) derivatives of the vector field, then, we would need to transport them (by “parallel transportation”) to a common point.



1.8.1.4 Vectors and Forms

When we introduce some vector space, with elements denoted, for instance, $\mathbf{V}, \mathbf{v}' \dots$, it often happens that a new, different, vector space is needed, with elements denoted, for instance $\mathbf{F}, \mathbf{F}' \dots$, and such that when taking an element of each space, we can “multiply” them and get a scalar,

$$\lambda = \langle \mathbf{F}, \mathbf{V} \rangle.\tag{1.133}$$

In terms of components, this will be written

$$\lambda = F_i V^i.\tag{1.134}$$

The product in 1.133–1.134, is called a *duality product*, and it has to be clearly distinguished from an inner (or scalar) product: in an inner product, we multiply two elements of a vector space; in a duality product, we multiply an element of a vector space by an element of a “dual space”.

This operation can always be defined, including the case where they do not have a metric (and, therefore, a scalar product). As an example, imagine that we work with pieces of metal and we need to consider the two parameters “electric conductivity” σ and “temperature” T . We may need to consider some (possibly nonlinear) function of σ and T , say $S(\sigma, T)$. For instance, $S(\sigma, T)$ may represent a “misfit function” on the (σ, T) space of those encountered when solving inverse problems in physics if we are measuring the parameters σ and T using indirect means. In this case, S is adimensional⁵. We may wish to know by which amount will S change when passing from point (σ_0, T_0) to a neighbouring point $(\sigma_0 + \Delta\sigma, T_0 + \Delta T)$. Writing only the first order term, and using matrix notations,

$$S(\sigma_0 + \Delta\sigma, T_0 + \Delta T) = S(\sigma_0, T_0) + \begin{pmatrix} \frac{\partial S}{\partial \sigma} \\ \frac{\partial S}{\partial T} \end{pmatrix}^T \begin{pmatrix} \Delta\sigma \\ \Delta T \end{pmatrix} + \dots, \quad (1.135)$$

where the partial derivatives are taken at point (σ_0, T_0) . Using tensor notations, setting $\mathbf{x} = (x^1, x^2) = (\sigma, T)$, we can write

$$\begin{aligned} S(\mathbf{x} + \Delta\mathbf{x}) &= S(\mathbf{x}) + \sum_i \frac{\partial S}{\partial x^i} \Delta x^i \\ &= S(\mathbf{x}) + \gamma_i \Delta x^i \\ &= S(\mathbf{x}) + \langle \boldsymbol{\gamma}, \Delta\mathbf{x} \rangle, \end{aligned} \quad (1.136)$$

where the notation introduced in equations 1.133–1.134 is used. As above, the partial derivatives are taken at point $\mathbf{x}_0 = (x_0^1, x_0^2) = (\sigma_0, T_0)$.

Note: say that figure 1.5 illustrates the definition of gradient as a tangent linear application. Say that the “mille-feuilles” are the “level-lines” of that tangent linear application.

Note: I have to explain somewhere the reason for putting an index in lower position to represent $\partial/\partial x^i$, i.e., to use the notation

$$\partial_i = \frac{\partial}{\partial x^i}.$$

Note: I have also to explain in spite of the fact that we have here partial derivatives, we have defined a tensorial object: the partial derivative of a scalar equals its true (covariant) derivative.

It is important that we realize that there is no “scalar product” involved in equations 1.136. Here are the arguments:

- The components of γ_i are **not** the components of a vector in the (σ, T) space. This can directly be seen by an inspection of their physical dimensions. As the function S is adimensional (see footnote 5), the components of $\boldsymbol{\gamma}$ have as dimensions the **inverse** of the physical dimensions of the components of the vector $\Delta\mathbf{x} = (\Delta x^1, \Delta x^2) = (\Delta\sigma, \Delta T)$. This clearly means that $\Delta\mathbf{x}$ and $\boldsymbol{\gamma}$ are “objects” that do not belong to the same space.

⁵ For instance, one could have the simple expression $S(\sigma, T) = \frac{|\sigma - \sigma_0|}{s_P} + \frac{|T - T_0|}{s_T}$, where s_P and s_T are standard deviations (or mean deviations) of some probability distribution.

- If equations 1.136 involved a scalar product we could define the norm of \mathbf{x} , the norm of γ and the angle between \mathbf{x} and γ . But these norms and angle are not defined. For instance, what could be the norm of $\mathbf{x} = (\Delta\sigma, \Delta T)$? Should we choose an L_2 norm? Or, as suggested by footnote 5, an L_1 norm? And, in any case, how could we make consistent such a definition of a norm with a change of variables where, instead of electric conductivity we use electric resistivity? (Note: make an appendix where the solution to this problem is given).

The product in equations 1.136 is not a scalar product (i.e., it is not the “product” of two elements belonging to the same space): it is a “duality product”, multiplying an element of a vector space and one element of a “dual space”.

Why this discussion is needed? Because of the tendency of imagining the gradient of a function $S(\sigma, T)$ as a vector (an “arrow”) in the $S(\sigma, T)$ space. If the gradient is not an arrow, then, what it is? Note: say here that figures 1.6 and 1.7 answer this by showing that an element of a dual space can be represented as a “mille-feuilles”.

Up to here we have only considered a vector space and its dual. But the notion generalizes to more general tensor spaces, i.e., to the case where “we have more than one index”. For instance, instead of equation 1.134 we could use an equation like

$$\lambda = F_{ij}^{.k} V^{ij}_k \quad (1.137)$$

to define scalars, consider that we are doing a duality product, and also use the notation of equation 1.133 to denote it. But this is not very useful, as, from a given “tensor” $F_{ij}^{.k}$ we can obtain scalar by operations like

$$\lambda = F_{ij}^{.k} V^i W^j_k . \quad (1.138)$$

It is better, in general, to just write explicitly the indices to indicate which sort of “product” we consider.

Sometimes (like in quantum mechanics), a “bra-ket” notation is used, where the name stands for the *bra* “ $\langle |$ ” and the *ket* “ $| \rangle$ ”. Then, instead of $\lambda = \langle \mathbf{F} , \mathbf{V} \rangle$ one writes

$$\lambda = \langle \mathbf{F} | \mathbf{V} \rangle = F_i V^i . \quad (1.139)$$

Then, the bra-ket notation is also used for the expression

$$\lambda = \langle \mathbf{V} | \mathbf{H} | \mathbf{W} \rangle = H_{ij} V^i W^j . \quad (1.140)$$

Note: say that the general rules for the change of component values in a change of coordinates, allow us to talk about “tensors” for “generalized vectors” as well as for “generalized forms”.

The “number of indices” that have to be used to represent the components of a tensor is called the *rank*, or the *order* of the tensor. Thus the tensors \mathbf{F} and \mathbf{V} just introduced are second rank, or second order. A tensor object with components R_{ijk}^ℓ could be called, in all rigor, a “(third-rank-form)-(first-rank-vector)” will we will not try to use this heavy terminology, the simple writing of the indices being explicit.

Note: say that if there is a metric, there is a trivial identification between a vector space and its dual, through equations like $F_i = g_{ij} V^j$, or $S^{ijk}_\ell = g^{ip} g^{jq} g^{kr} g_{ls} R_{pqr}^s$, and in that case, the same letter is used to designate one vector and *its* dual element, as in $V_i = g_{ij} V^j$, and $R^{ijk}_\ell = g^{ip} g^{jq} g^{kr} g_{ls} R_{pqr}^s$. But in *non metric* spaces (i.e., spaces without metric), there is usually a big difference between an space and its dual.

1.8.1.4.1 Gradient and Hessian Explain somewhere that if $\phi(\mathbf{x})$ is a scalar function, the Taylor development

$$\phi(\mathbf{x} + \Delta\mathbf{x}) = \phi(\mathbf{x}) + \langle \mathbf{g} | \Delta\mathbf{x} \rangle + \frac{1}{2!} \langle \Delta\mathbf{x} | \mathbf{H} | \Delta\mathbf{x} \rangle \quad (1.141)$$

defines the gradient \mathbf{g} and the Hessian \mathbf{H} .

1.8.1.4.2 Old text We may want the gradient to be “perpendicular” at the level lines of φ at \mathcal{O} , but there is no *natural* way to define a scalar product in the $\{P, T\}$ space, so we can not naturally define what “perpendicularity” is. That there is no natural way to define a scalar product does not mean that we can not define one: we can define many. For any symmetric, positive-definite matrix with the right physical dimensions (i.e., for any covariance matrix), the expression

$$\left(\left[\begin{array}{c} \delta P_1 \\ \delta T_1 \end{array} \right], \left[\begin{array}{c} \delta P_2 \\ \delta T_2 \end{array} \right] \right) = \left[\begin{array}{c} \delta P_1 \\ \delta T_1 \end{array} \right]^T \left[\begin{array}{cc} C_{PP} & C_{PT} \\ C_{TP} & C_{TT} \end{array} \right]^{-1} \left[\begin{array}{c} \delta P_2 \\ \delta T_2 \end{array} \right]$$

defines a scalar product. By an appropriate choice of the covariance matrix, we can make any of the two lines in figure 1.6 (or any other line) to be perpendicular to the level lines at the considered point: the gradient at a given point is something univocally defined, even in the absence of any scalar product; the “direction of steepest descent” is not, and there are as many as we may choose different scalar products. The gradient is not an arrow, i.e, it is not a *vector*. So, then, how to draw the gradient? Roughly speaking, the gradient is the *linear tangent application* at the considered point. It is represented in figure 1.7. As, by definition, it is a linear application, the level lines are straight lines, and the spacing of the level lines in the tangent linear application corresponds to the spacing of the level lines in the original function around the point where the gradient is computed. Speaking more technically, it is the development

$$\begin{aligned} \varphi(\mathbf{x} + \delta\mathbf{x}) &= \varphi(\mathbf{x}) + \langle \mathbf{g}, \delta\mathbf{x} \rangle + \dots \\ &= \varphi(\mathbf{x}) + g_i \delta x^i + \dots, \end{aligned}$$

when limited to its first order, that defines the tangent linear application. The *gradient* of φ is then \mathbf{g} . The gradient $\mathbf{g} = \{g_i\}$ at \mathcal{O} allows to associate a scalar to any vector $\mathbf{V} = \{V^i\}$ (also at \mathcal{O}): $\lambda = g_i V^i = \langle \mathbf{g}, \mathbf{V} \rangle$. This scalar is the difference of the values at the top and the bottom of the arrow representing the vector \mathbf{V} on the local tangent linear application to φ at \mathcal{O} . The index on the gradient can be a lower index, as the gradient is not a vector.

Note: say that figure 1.8 illustrates the fact that an element of the dual space can be represented as a “mille-feuilles” in the “primal” space or as an “arrow” in the dual space. And reciprocally.

Note: say that figure 1.9 illustrates the sum of arrows and the sum of “mille-feuilles”.

Note: say that figure 1.10 illustrates the sum of “mille-feuilles” in 3-D.

1.8.1.5 Natural Basis

A coordinate system associates to any point of the space, its coordinates. Each individual coordinate can be seen as a function associating, to any point of the space, the particular coordinate. We can define the gradient of this scalar function. We will have as many gradients

Figure 1.5: The gradient of a function (i.e., of an application) at a point x_0 is the tangent linear application at the given point. Let $x \mapsto f(x)$ represent the original (possibly nonlinear) application. The tangent linear application could be considered as mapping x into the values given by the linearized approximation for $f(x) : x \mapsto F(x) = \alpha + \beta x$. (Note: explain better). Rather, it is mathematically simpler to consider that the gradient maps *increments* of the independent variable x , $\Delta x = x - x_0$ into increments of the linearized dependent variable: $\Delta y = y - f(x_0) : \Delta x \mapsto \Delta y = \beta \Delta x$. (Note: explain this MUCH better).

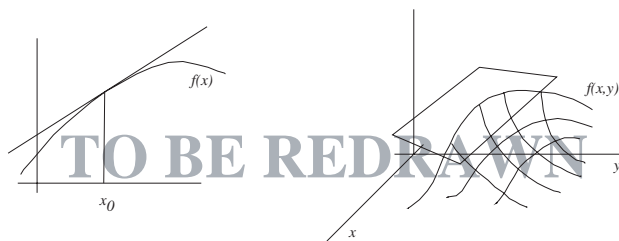


Figure 1.6: A scalar function $\varphi(P, T)$ depends on pressure and temperature. From a given point, two directions in the $\{P, T\}$ space are drawn. Which one corresponds to the gradient of $\varphi(P, T)$? In the figure at left, the pressure is indicated in International Units (m, kg, s), while in the figure at right, the c.g.s. units (cm, g, s) are used (remember that $1 \text{ Pa} = 10 \text{ dyne/cm}^{-2}$). From the left figure, we may think that the gradient is direction A, while from the figure at right we may think it is B. It is none: the right definition of gradient (see text) only allows, as graphic representation, the result shown in figure 1.7.

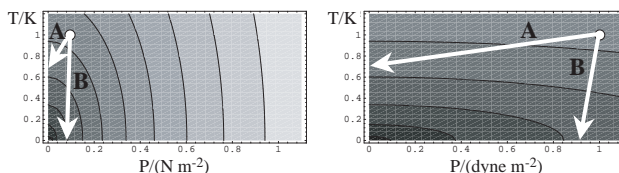


Figure 1.7: Gradient of the function displayed in figure 1.6, at the considered point. As the gradient is the linear tangent application at the given point, it is a linear application, and its level lines are straight lines. The value of the gradient at the considered point equals the value of the original function at that point. The spacing of the level lines in the gradient corresponds to the spacing of the level lines in the original function around the point where the gradient is computed. The two figures shown here are perfectly equivalent, as it should.

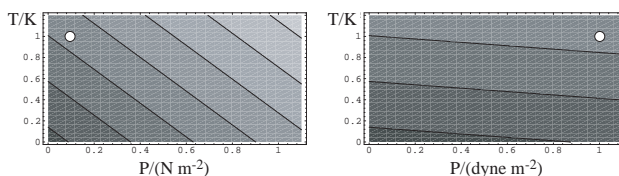


Figure 1.8: A point, at the left of the figure, may serve as the origin point for any vector we may want to represent. As usual, we may represent a vector \mathbf{V} by an arrow. Then, a form \mathbf{F} is represented by an oriented pattern of lines (or by an oriented pattern of surfaces in 3-D) with the line of zero value passing through the origin point. Each line has a value, that is the number that the form associates to any vector whose end point is on the line. Here, \mathbf{V} and \mathbf{F} are such that $\langle \mathbf{F}, \mathbf{V} \rangle = 2$. But a form is an element of the dual space, which is also a linear space. In the dual space, then, the form \mathbf{F} can be represented by an arrow (figure at right). In turn, \mathbf{V} is represented, in the dual space, by a pattern of lines.

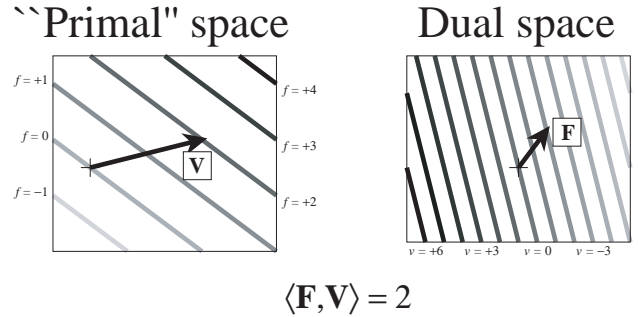


Figure 1.9: When representing vectors by arrows, the sum of two vectors is given by the main diagonal of the “parallelogram” drawn by two arrows. Then, a form is represented by a pattern of lines. The sum of two forms can be geometrically obtained using the “parallelogram” defined by the principal lozenge (containing the origin and with positive sense for both forms): the secondary diagonal of the lozenge is a line of the sum of the two forms. Note: explain this better.

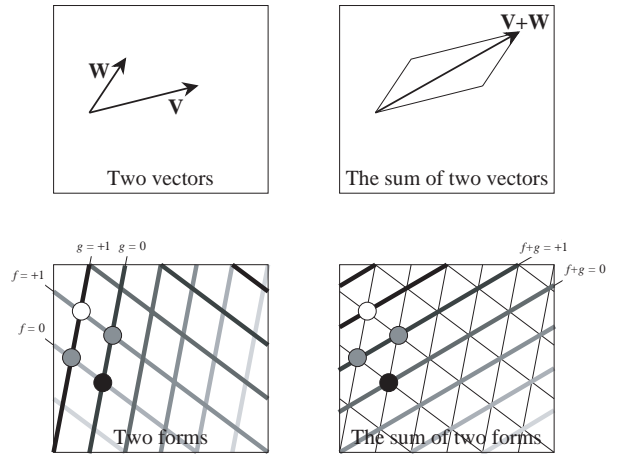


Figure 1.10: Sum of two forms, like in the previous figure, but here in 3-D. Note: explain that this figure can be “sheared” as one wants (we do not need to have a metric). Note: explain this better.

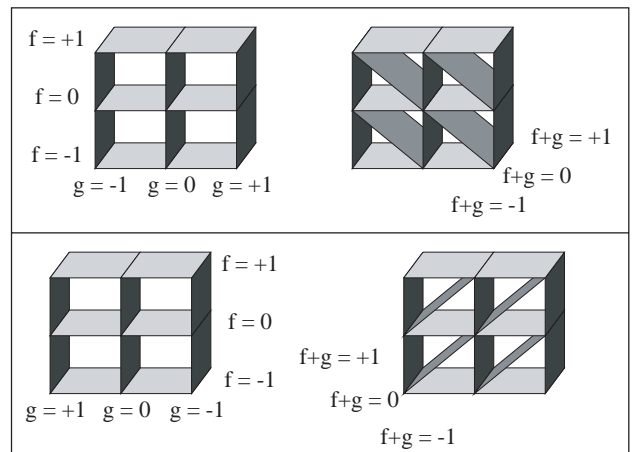
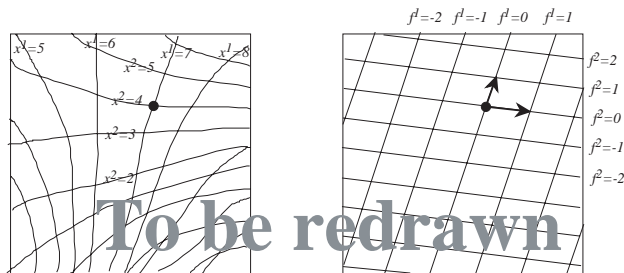


Figure 1.11: A system of coordinates, at left, and their gradients, at right. These gradient are forms. When in an n -dimensional space we have n forms, we can define n associate vortors by $\langle \mathbf{f}^i, \mathbf{e}_j \rangle = \delta^j_i$.



f^i as coordinates x^i . As a gradient, we have seen, is a form, we will have as many forms as coordinates. The usual requirements that coordinate systems have to fulfill (different points of the space have different coordinates, and vice versa) gives n linearly independent forms (we can not obtain one of them by linear combination of the others), i.e., a *basis* for the forms.

If we have a basis \mathbf{f}^i of forms, then we can introduce a basis \mathbf{e}_i of vectors, through

$$\langle \mathbf{f}^i, \mathbf{e}_j \rangle = \delta^j_i . \tag{1.142}$$

If we define the components V^i of a vector \mathbf{V} by

$$\mathbf{V} = V^i \mathbf{e}_i , \tag{1.143}$$

then, we can compute the components V^i by the formula

$$V^i = \langle \mathbf{f}^i, \mathbf{V} \rangle , \tag{1.144}$$

as we have

$$\langle \mathbf{f}^i, \mathbf{V} \rangle = \langle \mathbf{f}^i, V^j \mathbf{e}_j \rangle = \langle \mathbf{f}^i, \sum_j V^j \mathbf{e}_j \rangle = \sum_j V^j \langle \mathbf{f}^i, \mathbf{e}_j \rangle = \sum_j V^j \delta^i_j = V^j \delta^i_j = V^i . \tag{1.145}$$

Note that the computation of the components of a vector does not involve a scalar product, but a duality product.

To find the equivalent of equations 1.143 and 1.144 for forms, one defines the components F_i of a form \mathbf{F} by

$$\mathbf{F} = F_i \mathbf{f}^i , \tag{1.146}$$

and one easily gets

$$F_i = \langle \mathbf{F}, \mathbf{e}_i \rangle . \tag{1.147}$$

The notation \mathbf{e}_i for the basis of vectors is quite universal. Although the notation \mathbf{f}^i seems well adapted for a basis of forms, it is quite common to use the same letter for the basis of forms and for the basis of vectors. In what follows, we will use the notation

$$\mathbf{e}^i \equiv \mathbf{f}^i . \tag{1.148}$$

whose dangerousness vanishes only if we have a metric, i.e., when we can give sense to an expression like $\mathbf{e}_i = g_{ij} \mathbf{e}^j$. Using this notation the expressions

$$\mathbf{V} = V^i \mathbf{e}_i \iff V^i = \langle \mathbf{f}^i, \mathbf{V} \rangle \quad ; \quad \mathbf{F} = F_i \mathbf{f}^i , \iff F_i = \langle \mathbf{F}, \mathbf{e}_i \rangle \tag{1.149}$$

become

$$\mathbf{V} = V^i \mathbf{e}_i \iff V^i = \langle \mathbf{e}^i, \mathbf{V} \rangle \quad ; \quad \mathbf{F} = F_i \mathbf{e}^i, \iff F_i = \langle \mathbf{F}, \mathbf{e}_i \rangle. \quad (1.150)$$

We have now basis for vectors and forms, so we can write expressions like $\mathbf{V} = V^i \mathbf{e}_i$ and $\mathbf{F} = F_i \mathbf{e}^i$. We need basis for objects “with more than one index”, so we can write expressions like

$$\mathbf{B} = B^{ij} \mathbf{e}_{ij} \quad ; \quad \mathbf{C} = C_{ij} \mathbf{e}^{ij} \quad ; \quad \mathbf{D} = C_i^j \mathbf{e}^i_j \quad ; \quad \mathbf{E} = E_{ijk\dots}^{\ell mn\dots} \mathbf{e}^{ijk\dots}_{\ell mn\dots} \quad (1.151)$$

The introduction of these basis raises a difficulty. While we have an immediate intuitive representation for vectors (as “arrows”) and for forms (as “millefeuilles”), tensor objects of higher rank are more difficult to represent. If a symmetric 2-tensor, like the stress tensor σ^{ij} of mechanics, can be viewed as an ellipsoid, how could we view a tensor $T_{ijk}^{\ell m}$? It is the power of mathematics to suggest analogies, so we can work even without geometric interpretations. But this absence of intuitive interpretation of high-rank tensors tells us that we will have to introduce the basis for these objects in a non-intuitive way. Essentially, what we want is that the basis for high rank tensors is not independent for the basis of vectors and forms. We want, in fact, more than this. Given two vectors U^i and V^j , we understand what we mean when we define a 2-tensor \mathbf{W} by $W^{ij} = U^i V^j$. The basis for 2-tensors is perfectly defined by the condition that we wish that the components of \mathbf{W} are precisely $U^i V^j$ and not, for instance, the values obtained after some rotation or change of coordinates.

This is enough, and we could directly use the notations introduced by equations 1.151. Instead, common mathematical developments introduce the notion of “tensor product”, and, instead of notations like \mathbf{e}_{ij} , \mathbf{e}^{ij} , \mathbf{e}_i^j , or $\mathbf{e}^{ijk\dots}_{\ell mn\dots}$, introduce the notations $\mathbf{e}_i \otimes \mathbf{e}_j$, $\mathbf{e}^i \otimes \mathbf{e}^j$, or $\mathbf{e}^i \otimes \mathbf{e}^j \otimes \mathbf{e}^k \otimes \dots \otimes \mathbf{e}_\ell \otimes \mathbf{e}_m \otimes \mathbf{e}_n \otimes \dots$. Then, equations 1.151 are written

$$\begin{aligned} \mathbf{B} &= B^{ij} \mathbf{e}_i \otimes \mathbf{e}_j \quad ; \quad \mathbf{C} = C_{ij} \mathbf{e}^i \otimes \mathbf{e}^j \quad ; \quad \mathbf{D} = C_i^j \mathbf{e}^i \otimes \mathbf{e}^j \\ \mathbf{E} &= E_{ijk\dots}^{\ell mn\dots} \mathbf{e}^i \otimes \mathbf{e}^j \otimes \mathbf{e}^k \otimes \dots \otimes \mathbf{e}_\ell \otimes \mathbf{e}_m \otimes \mathbf{e}_n \otimes \dots \end{aligned} \quad (1.152)$$

What follows is an old text, to be updated.

The metric tensor has been introduced in section 1.3. Let us show here that if the space into consideration has a scalar product, then, the metric can be computed. Here, the scalar product of two vectors \mathbf{V} and \mathbf{W} is denoted $\mathbf{V} \cdot \mathbf{W}$. Then, defining

$$d\mathbf{r} = dx^i \mathbf{e}_i \quad (1.153)$$

and

$$ds^2 = d\mathbf{r} \cdot d\mathbf{r} \quad (1.154)$$

gives

$$ds^2 = d\mathbf{r} \cdot d\mathbf{r} = (dx^i \mathbf{e}_i) \cdot (dx^j \mathbf{e}_j) = (\mathbf{e}_i \cdot \mathbf{e}_j) dx^i dx^j. \quad (1.155)$$

Defining the metric tensor

$$g_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j \quad (1.156)$$

gives then

$$ds^2 = g_{ij} dx^i dx^j. \quad (1.157)$$

To emphasize that at every point of the manifold we have a different tensor space, and a different basis, we can always write explicitly the dependence of the basis vectors on the coordinates, as in $\mathbf{e}_i(\mathbf{x})$. Equation 1.143 is then just a short notation for

$$\mathbf{V}(\mathbf{x}) = V^i(\mathbf{x}) \mathbf{e}_i(\mathbf{x}), \quad (1.158)$$

while equation 1.146 is a short notation for

$$\mathbf{F}(\mathbf{x}) = F_i(\mathbf{x}) \mathbf{e}^i(\mathbf{x}). \quad (1.159)$$

Here and in most places of the book, the notation \mathbf{x} is a short-cut notation for $\{x^1, x^2, \dots\}$. The reader should just remember that \mathbf{x} represents a point in the space, but it is not a vector.

It is important to realize that, when dealing with tensor mathematics, a single basis is a basis for all the vector spaces at the considered point. For instance, the vector \mathbf{V} may be a velocity, and the vector \mathbf{E} may be an electric field. The two vectors belong to different vector spaces, but they are obtained as “linear combinations” of the same basis vectors:

$$\begin{aligned} \mathbf{V} &= V^i \mathbf{e}_i \\ \mathbf{E} &= E^i \mathbf{e}_i, \end{aligned} \quad (1.160)$$

but, of course, the components are not pure real numbers: they have dimensions. Box ?? recalls what the dimensions of components are.

Let us examine the components of the basis vectors (on the basis they define). Obviously,

$$(\mathbf{e}_i)^j = \delta_i^j \quad (\mathbf{e}^j)_i = \delta_i^j, \quad (1.161)$$

or, explicitly,

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix} \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix} \quad \dots \quad (1.162)$$

Equivalently, for the basis of 2-tensors we have

$$(\mathbf{e}_i \otimes \mathbf{e}_j)^{kl} = \delta_i^k \delta_j^l \quad (1.163)$$

$$\mathbf{e}_1 \otimes \mathbf{e}_1 = \begin{pmatrix} 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ \dots & \dots & \dots & \ddots \end{pmatrix} \quad \mathbf{e}_1 \otimes \mathbf{e}_2 = \begin{pmatrix} 0 & 1 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ \dots & \dots & \dots & \ddots \end{pmatrix} \quad \dots$$

$$\mathbf{e}_2 \otimes \mathbf{e}_1 = \begin{pmatrix} 0 & 0 & 0 & \dots \\ 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ \dots & \dots & \dots & \ddots \end{pmatrix} \quad \mathbf{e}_2 \otimes \mathbf{e}_2 = \begin{pmatrix} 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ \dots & \dots & \dots & \ddots \end{pmatrix} \quad \dots \quad (1.164)$$

and similar formular for other basis.

Note: say somewhere that the definition of basis vectors given above imposes that the vectors of the natural basis are, at any point, tangent to the coordinate lines at that point. The notion of tangency is independent of the existence, or not, of a metric, i.e., of the possibility of measuring distances in the space. This is not so for the notion of perpendicularity, that makes sense only if we can measure distances (and, therefore, angles). In, general, then, the vectors of the natural basis are tangent to the coordinate lines. When a metric has been introduced, the vectors in the natural basis at a given point will be mutually perpendicular only if the coordinate lines themselves are mutually perpendicular at that point. Ordinary coordinates in the Euclidean 3-D space (Cartesian, cylindrical, spherical, . . .) define coordinate lines that are orthogonal at every point. Then, the vectors of the natural basis will also be mutually orthogonal at all points. *But the vectors of the natural basis are not, in general, normed to 1.* For instance, figure XXX illustrates the fact that the norm of the vectors of the natural basis in polar coordinates are, at point (r, φ) , $\|\mathbf{e}_r\| = 1$ and $\|\mathbf{e}_\varphi\| = r$.

1.8.1.6 Tensor Components

Consider, over an n -dimensional manifold \mathcal{X} . At any point \mathcal{P} of the manifold, one can consider the linear space \mathcal{L} that is tangent to the manifold at that point, and its dual $\widehat{\mathcal{L}}$. One can also consider, at point \mathcal{P} , the ‘tensor product’ of spaces $\mathcal{L}(p, q) = \underbrace{\mathcal{L} \otimes \mathcal{L} \otimes \cdots \otimes \mathcal{L}}_{p \text{ times}} \otimes \underbrace{\widehat{\mathcal{L}} \otimes \widehat{\mathcal{L}} \otimes \cdots \otimes \widehat{\mathcal{L}}}_{q \text{ times}}$. A ‘ p -times contravariant, q -times covariant tensor’ at point \mathcal{P} of the manifold is an element of $\mathcal{L}(p, q)$.

When a coordinate system $\mathbf{x} = \{x^1, \dots, x^n\}$ is chosen over \mathcal{X} , one has, at point \mathcal{P} , the ‘natural basis’ for the linear tangent space \mathcal{L} , say $\{\mathbf{e}_i\}$, and, by virtue of the tensor product, also a basis for the space $\mathcal{L}(p, q)$, say $\{\mathbf{e}_{i_1} \otimes \mathbf{e}_{i_2} \otimes \cdots \otimes \mathbf{e}_{i_p} \otimes \mathbf{e}^{j_1} \otimes \mathbf{e}^{j_2} \otimes \cdots \otimes \mathbf{e}^{j_q}\}$. Any tensor \mathbf{T} at point \mathcal{P} of \mathcal{X} can then be developed on this basis,

$$\mathbf{T} = T_{\mathbf{x}}{}^{i_1 i_2 \dots i_p}{}_{j_1 j_2 \dots j_q} \mathbf{e}_{i_1} \otimes \mathbf{e}_{i_2} \otimes \cdots \otimes \mathbf{e}_{i_p} \otimes \mathbf{e}^{j_1} \otimes \mathbf{e}^{j_2} \otimes \cdots \otimes \mathbf{e}^{j_q} \quad , \quad (1.165)$$

to define the natural components of the tensor, $T_{\mathbf{x}}{}^{i_1 i_2 \dots i_p}{}_{j_1 j_2 \dots j_q}$. They are intimately linked to the coordinate system chosen over \mathcal{X} , as this coordinate system has induced the natural basis $\{\mathbf{e}_i\}$ at the considered point \mathcal{P} . The index \mathbf{x} in the components is there to recall this fact. It is essential when different coordinates are going to be simultaneously considered, but it can be dropped when there is no possible confusion about the coordinate system being used. Its lower or upper position may be chosen for typographical clarity, and, of course, has no special variance meaning.

1.8.1.7 Tensors in Metric Spaces

Comment: explain here that it is possible to give a lot of structure to a manifold (tangent linear space, (covariant) derivation, etc.) without the need of a metric. It is introduced here to simplify the text, as, if not, we would have need to come bak to most of the results to add the particular properties arising when there is a metric. But, in all rigor, it would be preferable to introduce the metric after, for instance, the definition of covariant differentiaition, that does not need it.

Having a metric in a differential manifold means being able to define the length of a line. This will then imply that we can define a scalar product at every local tangent linear space (and, thus, the angle between two crossing lines).

The metric will also allow to define a natural bijection between vectors and forms, and between tensors densities and capacities.

A metric is defined when a second rank symmetric form \mathbf{g} with components g_{ij} is given. The length L of a line $x^i(\lambda)$ is then defined by the line integral

$$L = \int_{\lambda} ds, \quad (1.166)$$

where

$$ds^2 = g_{ij} dx^i dx^j. \quad (1.167)$$

Once we have a metric, it is possible to define a bijection between forms and vectors. For, to the vector \mathbf{V} with components V^i we can associate the form \mathbf{F} with components

$$F_i = g_{ij} V^j. \quad (1.168)$$

Then, it is customary to use the same letter to designate a vector and a form that are linked by this natural bijection, as in

$$V_i = g_{ij} V^j. \quad (1.169)$$

The inverse of the previous equation is written

$$V^i = g^{ij} V_j, \quad (1.170)$$

where

$$g_{ij} g^{jk} = \delta_i^k. \quad (1.171)$$

The reader will easily give sense to the expression

$$\mathbf{e}_i = g_{ij} \mathbf{e}^j. \quad (1.172)$$

The equations above, and equations like

$$T_{ij\dots}{}^{kl\dots} = g_{ip} g_{jq} \dots g^{kr} g^{ls} \dots T^{pq\dots}{}_{rs\dots}, \quad (1.173)$$

are summarized by saying that “the metric tensor allows to raise and lower indices”.

The value of the metric at a particular point of the manifold allows to define a scalar product for the vectors in the local tangent linear space. Denoting the scalar product of two vectors \mathbf{V} and \mathbf{W} by $\mathbf{V} \cdot \mathbf{W}$, we can use any of the definitions

$$\mathbf{V} \cdot \mathbf{W} = g_{ij} V^i W^j = V_i W^j = V^i W_j. \quad (1.174)$$

To define parallel transportation of tensors, we have introduced a connection $\Gamma_{ij}{}^k$. Now that we have a metric we may wonder if when parallel-transporting a vector, it conserves constant

length. It is easy to show (see demonstration in [Comment: where?]) that this is true if we have the *compatibility condition*

$$\nabla_i g_{jk} = 0, \quad (1.175)$$

i.e.,

$$\partial_i g_{jk} = g_{sk} \Gamma_{ij}^s + g_{js} \Gamma_{ik}^s. \quad (1.176)$$

The compatibility condition 1.175 implies that the metric tensor and the nabla symbol commute:

$$\nabla_i (g_{jk} T^{pq\dots rs\dots}) = g_{jk} (\nabla_i T^{pq\dots rs\dots}), \quad (1.177)$$

which, in fact, means that it is equivalent to take a covariant derivative, then raise or lower an index, or first raise or lower an index, then take the covariant derivative.

Note: introduce somewhere the notation

$$\Gamma_{ijk} = g_{ks} \Gamma_{ij}^s, \quad (1.178)$$

warn the reader that this is just a *notation*: the connection coefficients are not the components of a tensor. and say that if the condition 1.175 holds, then, it is possible to compute the connection coefficients from the metric and the torsion:

$$\Gamma_{ijk} = \frac{1}{2} (\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}) + \frac{1}{2} (S_{ijk} + S_{kij} + S_{kji}). \quad (1.179)$$

As the basis vectors have components

$$(\mathbf{e}_i)^j = \delta_i^j, \quad (1.180)$$

we have

$$\mathbf{e}_i \cdot \mathbf{e}_j = g_{ij}. \quad (1.181)$$

Defining

$$d\mathbf{r} = dx^i \mathbf{e}_i \quad (1.182)$$

gives then

$$d\mathbf{r} \cdot d\mathbf{r} = ds^2. \quad (1.183)$$

We have seen that the metric can be used to define a natural bijection between forms and vectors. Let us now see that it can also be used to define a natural bijection between tensors, densities, and capacities.

We denote by \bar{g} the determinant of g_{ij} :

$$\bar{g} = \det(\{g_{ij}\}) = \frac{1}{n!} \bar{\varepsilon}^{ijk\dots} \bar{\varepsilon}^{pqr\dots} g_{ip} g_{jq} g_{kr} \dots \quad (1.184)$$

The two upper bars recall that $\overline{\overline{g}}$ is a second order density, as there is the product of two densities at the right-hand side.

For a reason that will become obvious soon, the square root of $\overline{\overline{g}}$ is denoted \overline{g} :

$$\overline{\overline{g}} = \overline{g} \overline{g}. \quad (1.185)$$

In (Comment: where?) we demonstrate that we have

$$\partial_i \overline{g} = \overline{g} \Gamma_{is}^s. \quad (1.186)$$

Using expression (Comment: which one?) for the (covariant) derivative of a scalar density, this simply gives

$$\nabla_i \overline{g} = \partial_i \overline{g} - \overline{g} \Gamma_{is}^s = 0, \quad (1.187)$$

which is consistent with the fact that

$$\nabla_i g_{jk} = 0. \quad (1.188)$$

We can also define the determinant of g^{ij} :

$$\underline{\underline{g}} = \det(\{g^{ij}\}) = \frac{1}{n!} \varepsilon_{ijk\dots} \varepsilon_{pqr\dots} g^{ip} g^{jq} g^{kr} \dots, \quad (1.189)$$

and its square root \underline{g} :

$$\underline{\underline{g}} = \underline{g} \underline{g}. \quad (1.190)$$

As the matrices g_{ij} and g^{ij} are mutually inverses, we have

$$\overline{g} \underline{g} = 1. \quad (1.191)$$

Using the scalar density \overline{g} and the scalar capacity \underline{g} we can associate tensor densities, pure tensors, and tensor capacities. Using the same letter to designate the objects related through this natural bijection, we will write expressions like

$$\overline{\rho} = \overline{g} \rho, \quad (1.192)$$

$$\overline{V}^i = \overline{g} V^i, \quad (1.193)$$

or

$$T_{ij\dots}{}^{kl\dots} = \underline{g} \overline{T}_{ij\dots}{}^{kl\dots}. \quad (1.194)$$

So, if g_{ij} and g^{ij} can be used to “lower and raise indices”, \overline{g} and \underline{g} can be used to “put and remove bars”.

Comment: say somewhere that \overline{g} is the *density of volumetric content*, as the volume element of a metric space is given by

$$dV = \overline{g} d\mathcal{T}, \quad (1.195)$$

where $d\tau$ is the *capacity element* defined in (Comment: where?), and which, when we take an element along the coordinate lines, equals $dx^1 dx^2 dx^3 \dots$.

Comment: Say that we can demonstrate that, in an Euclidean space, the matrix representing the metric equals the product of the Jacobian matrix times the transposed matrix:

$$\{g_{ij}\} = \begin{pmatrix} g_{11} & g_{12} & \cdots \\ g_{21} & g_{22} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} \frac{\partial X^1}{\partial x^1} & \frac{\partial X^1}{\partial x^2} & \cdots \\ \frac{\partial X^2}{\partial x^1} & \frac{\partial X^2}{\partial x^2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \times \begin{pmatrix} \frac{\partial X^1}{\partial x^1} & \frac{\partial X^2}{\partial x^1} & \cdots \\ \frac{\partial X^1}{\partial x^2} & \frac{\partial X^2}{\partial x^2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}. \quad (1.196)$$

In short,

$$g_{ij} = \sum_K \frac{\partial X^K}{\partial x^i} \frac{\partial X^K}{\partial x^j}. \quad (1.197)$$

This follows directly from the general equation

$$g_{ij} = \frac{\partial X^I}{\partial x^i} \frac{\partial X^J}{\partial x^j} g_{IJ} \quad (1.198)$$

using the fact that, if the $\{X^I\}$ are Cartesian coordinates,

$$\{g_{IJ}\} = \begin{pmatrix} g_{11} & g_{12} & \cdots \\ g_{21} & g_{22} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots \\ 0 & 1 & \cdots \\ 0 & 0 & \cdots \end{pmatrix}. \quad (1.199)$$

Comment: explain here that the metric introduces a bijection between forms and vectors:

$$V_i = g_{ij} V^j. \quad (1.200)$$

Comment: introduce here the notation

$$(\mathbf{V}, \mathbf{W}) = g_{ij} V^i W^j = V_i W^i = W_i V^i. \quad (1.201)$$

1.8.2 Appendix: Dimension of Components

Which dimensions have the components of a vector? Contrarily to the basis of elementary calculus, the vectors defining the natural basis are not normed to one. Rather, it follows from $g_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j$ that the length (i.e., the norm) of the basis vector \mathbf{e}_i is

$$\|\mathbf{e}_i\| = \sqrt{g_{ii}}.$$

For instance, if in the Euclidean 3-D space with Cartesian coordinates

$$\|\mathbf{e}_x\| = \|\mathbf{e}_y\| = \|\mathbf{e}_z\| = 1,$$

the use of spherical coordinates gives

$$\|\mathbf{e}_r\| = 1 \quad \|\mathbf{e}_\theta\| = r \quad \|\mathbf{e}_\varphi\| = r \sin \theta.$$

Denoting by $[\|\mathbf{V}\|]$ the physical dimension of (the norm of) a vector, this gives

$$[\|\mathbf{e}_i\|] = [\sqrt{g_{ii}}].$$

For instance, in Cartesian coordinates,

$$[\|\mathbf{e}_x\|] = [\|\mathbf{e}_y\|] = [\|\mathbf{e}_z\|] = 1,$$

and in spherical coordinates,

$$[\|\mathbf{e}_r\|] = 1 \quad [\|\mathbf{e}_\theta\|] = L \quad [\|\mathbf{e}_\varphi\|] = L,$$

where L represents the dimension of a *length*. A vector $\mathbf{V} = V^i \mathbf{e}_i$ has components with dimensions

$$[V^i] = \frac{[\|\mathbf{V}\|]}{[\|\mathbf{e}_i\|]} = \frac{[\|\mathbf{V}\|]}{[\sqrt{g_{ii}}]}.$$

For instance, in Cartesian coordinates,

$$[V^x] = [V^y] = [V^z] = [\|\mathbf{V}\|]$$

and in spherical coordinates,

$$[V^r] = [\|\mathbf{V}\|] \quad [V^\theta] = \frac{[\|\mathbf{V}\|]}{L} \quad [V^\varphi] = \frac{[\|\mathbf{V}\|]}{L}.$$

In general, the physical dimension of the component $T_{ij\dots}{}^{k\ell\dots}$ of a tensor \mathbf{T} is

$$\begin{aligned} [T_{ij\dots}{}^{k\ell\dots}] &= [\|\mathbf{T}\|] [\|\mathbf{e}_i\|] [\|\mathbf{e}_j\|] \dots \frac{1}{[\|\mathbf{e}_k\|]} \frac{1}{[\|\mathbf{e}_\ell\|]} \dots \\ &= [\|\mathbf{T}\|] [\sqrt{g_{ii}}] [\sqrt{g_{jj}}] \dots \frac{1}{[\sqrt{g_{kk}}]} \frac{1}{[\sqrt{g_{\ell\ell}}]} \dots \end{aligned}$$

1.8.3 Appendix: The Jacobian in Geographical Coordinates

Example 1.5 *Let*

$$\mathbf{x} = \{x, y, z\} \quad ; \quad \mathbf{y} = \{r, \varphi, \vartheta\} \quad (1.202)$$

respectively represent a Cartesian and a geographical system of coordinates over the Euclidean 3D space,

$$\begin{aligned} x &= r \cos \vartheta \cos \varphi \\ y &= r \cos \vartheta \sin \varphi \\ z &= r \sin \vartheta \quad . \end{aligned} \quad (1.203)$$

The matrix of partial derivatives defined at the right of equation 1.2 is

$$\mathbf{X} = \begin{pmatrix} \cos \vartheta \cos \varphi & r \sin \vartheta \cos \varphi & -r \cos \vartheta \sin \varphi \\ \cos \vartheta \sin \varphi & r \sin \vartheta \sin \varphi & r \cos \vartheta \cos \varphi \\ \sin \vartheta & -r \cos \vartheta & 0 \end{pmatrix} , \quad (1.204)$$

The matrix \mathbf{Y} defined at the left of equation 1.2 could be computed by, first, solving in equations 1.203 for the geographical coordinates as a function of the Cartesian ones, and, then, by computing the partial derivatives. This would give the matrix \mathbf{Y} as a function of $\{x, y, z\}$. More simply, we can just evaluate \mathbf{Y} as \mathbf{X}^{-1} (equation 1.5), but this, of course, gives \mathbf{Y} as a function of $\{r, \vartheta, \varphi\}$:

$$\mathbf{Y} = \begin{pmatrix} \cos \vartheta \cos \varphi & \cos \vartheta \sin \varphi & \sin \vartheta \\ \sin \vartheta \cos \varphi / r & \sin \vartheta \sin \varphi / r & -\cos \vartheta / r \\ -\sin \varphi / (r \cos \vartheta) & \cos \varphi / (r \cos \vartheta) & 0 \end{pmatrix} . \quad (1.205)$$

The two Jacobian determinants are

$$X = \frac{1}{Y} = r^2 \cos \vartheta \quad . \quad (1.206)$$

For the metric, as

$$ds^2 = dx^2 + dy^2 + dz^2 = dr^2 + r^2 \cos^2 \vartheta d\varphi^2 + r^2 d\vartheta^2 \quad , \quad (1.207)$$

one has the volume densities (remember that $\bar{g} = \sqrt{\det \mathbf{g}}$)

$$\bar{g}_{\mathbf{x}} = 1 \quad ; \quad \bar{g}_{\mathbf{y}} = r^2 \cos \vartheta \quad . \quad (1.208)$$

The comparison of these two last equations with equations 1.206 shows that one has

$$\bar{g}_{\mathbf{y}} = X \bar{g}_{\mathbf{x}} \quad , \quad (1.209)$$

in accordance with the general rule for the change of values of a scalar density under a change of variables (equation 1.12).

Here, the fundamental capacity elements are

$$d\underline{v}_x = dx \wedge dy \wedge dz \quad ; \quad d\underline{v}_y = dr \wedge d\varphi \wedge d\vartheta \quad . \quad (1.210)$$

Using the change of variables in equation 1.203 one obtains⁶

$$dx \wedge dy \wedge dz = r^2 \cos \vartheta \, dr \wedge d\varphi \wedge d\vartheta \quad , \quad (1.211)$$

and inserting this into equation 1.210 gives

$$d\underline{v}_y = \frac{1}{r^2 \cos \vartheta} d\underline{v}_x = \frac{1}{X} d\underline{v}_x \quad , \quad (1.212)$$

in accordance with the general rule for the change of values of a scalar capacity under a change of variables (equation 1.13). **[End of example.]**

⁶This results from the explicit computation of the exterior product $dx \wedge dy \wedge dz$, where $dx = \cos \varphi \cos \vartheta \, dr - r \sin \varphi \cos \vartheta \, d\varphi - r \cos \varphi \sin \vartheta \, d\vartheta$, $dy = \sin \varphi \cos \vartheta \, dr + r \cos \varphi \cos \vartheta \, d\varphi - r \sin \varphi \sin \vartheta \, d\vartheta$ and $dz = \sin \vartheta \, dr + r \cos \vartheta \, d\vartheta$.

1.8.4 Appendix: Kronecker Determinants in 2 3 and 4 D

1.8.4.1 The Kronecker's determinants in 2-D

$$\begin{aligned}
\delta_{ij}^{k\ell} &= (1/0!) \varepsilon_{ij} \varepsilon^{k\ell} = \delta_i^k \delta_j^\ell - \delta_i^\ell \delta_j^k \\
\delta_j^k &= (1/1!) \varepsilon_{ij} \varepsilon^{ik} = \delta_j^k \\
\delta &= (1/2!) \varepsilon_{ij} \varepsilon^{ij} = 1
\end{aligned} \tag{1.213}$$

1.8.4.2 The Kronecker's determinants in 3-D

$$\begin{aligned}
\delta_{ijk}^{\ell mn} &= (1/0!) \varepsilon_{ijk} \varepsilon^{\ell mn} = \delta_i^\ell \delta_j^m \delta_k^n + \delta_i^m \delta_j^n \delta_k^\ell + \delta_i^n \delta_j^\ell \delta_k^m - \delta_i^\ell \delta_j^n \delta_k^m - \delta_i^m \delta_j^\ell \delta_k^n - \delta_i^n \delta_j^m \delta_k^\ell \\
\delta_{jk}^{\ell m} &= (1/1!) \varepsilon_{ijk} \varepsilon^{ilm} = \delta_j^\ell \delta_k^m - \delta_j^m \delta_k^\ell \\
\delta_k^\ell &= (1/2!) \varepsilon_{ijk} \varepsilon^{ij\ell} = \delta_k^\ell \\
\delta &= (1/3!) \varepsilon_{ijk} \varepsilon^{ijk} = 1
\end{aligned} \tag{1.214}$$

1.8.4.3 The Kronecker's determinants in 4-D

$$\begin{aligned}
\delta_{ijkl}^{mnpq} &= (1/0!) \varepsilon_{ijkl} \varepsilon^{mnpq} \\
&= +\delta_i^m \delta_j^n \delta_k^p \delta_\ell^q + \delta_i^m \delta_j^p \delta_k^q \delta_\ell^n + \delta_i^m \delta_j^q \delta_k^n \delta_\ell^p + \delta_i^n \delta_j^q \delta_k^p \delta_\ell^m + \delta_i^n \delta_j^p \delta_k^m \delta_\ell^q + \delta_i^n \delta_j^m \delta_k^q \delta_\ell^p \\
&+ \delta_i^p \delta_j^q \delta_k^m \delta_\ell^n + \delta_i^p \delta_j^m \delta_k^n \delta_\ell^q + \delta_i^p \delta_j^n \delta_k^q \delta_\ell^m + \delta_i^q \delta_j^m \delta_k^p \delta_\ell^n + \delta_i^q \delta_j^n \delta_k^m \delta_\ell^p + \delta_i^q \delta_j^p \delta_k^n \delta_\ell^m \\
&- \delta_i^m \delta_j^n \delta_k^q \delta_\ell^p - \delta_i^m \delta_j^p \delta_k^n \delta_\ell^q - \delta_i^m \delta_j^q \delta_k^p \delta_\ell^n - \delta_i^n \delta_j^p \delta_k^q \delta_\ell^m - \delta_i^n \delta_j^q \delta_k^m \delta_\ell^p - \delta_i^n \delta_j^m \delta_k^p \delta_\ell^q \\
&- \delta_i^p \delta_j^q \delta_k^n \delta_\ell^m - \delta_i^p \delta_j^m \delta_k^q \delta_\ell^n - \delta_i^p \delta_j^n \delta_k^m \delta_\ell^q - \delta_i^q \delta_j^m \delta_k^n \delta_\ell^p - \delta_i^q \delta_j^n \delta_k^p \delta_\ell^m - \delta_i^q \delta_j^p \delta_k^m \delta_\ell^n
\end{aligned}$$

$$\begin{aligned}
\delta_{jkl}^{mnp} &= (1/1!) \varepsilon_{ijkl} \varepsilon^{imnp} \\
&= \delta_j^m \delta_k^n \delta_\ell^p + \delta_j^n \delta_k^p \delta_\ell^m + \delta_j^p \delta_k^m \delta_\ell^n - \delta_j^m \delta_k^p \delta_\ell^n - \delta_j^n \delta_k^m \delta_\ell^p - \delta_j^p \delta_k^n \delta_\ell^m
\end{aligned} \tag{1.215}$$

$$\delta_{kl}^{mn} = (1/2!) \varepsilon_{ijkl} \varepsilon^{ijmn} = (\delta_k^m \delta_\ell^n - \delta_k^n \delta_\ell^m)$$

$$\delta_\ell^m = (1/3!) \varepsilon_{ijkl} \varepsilon^{ijkm} = \delta_\ell^m$$

$$\delta = (1/4!) \varepsilon_{ijkl} \varepsilon^{ijkl} = 1$$

1.8.5 Appendix: Definition of Vectors

Consider the 3-D physical space, with coordinates $\{x^i\} = \{x^1, x^2, x^3\}$. In classical mechanics, the trajectory of a particle is described by the three functions of time $x^i(t)$. Obviously the three values $\{x^1, x^2, x^3\}$ are not the components of a vector, as an expression like $x^i(t) = x_I^i(t) + x_{II}^i(t)$ has, in general, no sense (think, for instance, in the case where we use spherical coordinates).

Define now the velocity of the particle at time t_0 :

$$v^i(t_0) = \left(\frac{dx^i}{dt} \right)_{t=t_0} .$$

If two particles coincide at some point of the space $\{x_0^1, x_0^2, x_0^3\}$, it makes sense to define, for instance, their relative velocity by $v^i(x_0^1, x_0^2, x_0^3, t_0) = v_I^i(x_0^1, x_0^2, x_0^3, t_0) - v_{II}^i(x_0^1, x_0^2, x_0^3, t_0)$. The v^i are the components of a vector.

If we change coordinates, $x'^I = x'^I(x^j)$, then the velocity is defined, in the new coordinate system, $v'^I = dx'^I/dt$, and we have $v'^I = dx'^I/dt = \partial x'^I / \partial x^i dx^i/dt$, i.e.,

$$v'^I = \frac{\partial x'^I}{\partial x^i} v^i ,$$

which is the standard rule for transformation of the components of a vector when the coordinates (and, so, the natural basis) change.

Objects with upper or lower indices not always are *tensors*. The four classical objects which do not have necessarily tensorial character are:

- the coordinates $\{x^i\}$,
- the partial differential operator ∂_i ,
- the Connection Coefficients Γ_{ij}^k ,
- the elements of the Jacobian matrix $J_i^I = \partial x'^I / \partial x^i$.

1.8.6 Appendix: Change of Components

	capacity	tensor	density
0-rank	$\underline{s}' = \mathcal{J} \underline{s}$	$s' = s$	$\bar{s}' = \mathcal{J}' \bar{s}$
1-form	$\underline{F}'_I = \mathcal{J} J_I^i \underline{F}_i$	$F'_I = J_I^i F_i$	$\bar{F}'_I = \mathcal{J}' J_I^i \bar{F}_i$
1-vector	$\underline{V}'^I = \mathcal{J} \underline{V}^i J_i^I$	$V'^I = V^i J_i^I$	$\bar{V}'^I = \mathcal{J}' \bar{V}^i J_i^I$
2-form	$\underline{Q}'_{IJ} = \mathcal{J} J_I^i J_J^j \underline{Q}_{ij}$	$Q'_{IJ} = J_I^i J_J^j Q_{ij}$	$\bar{Q}'_{IJ} = \mathcal{J}' J_I^i J_J^j \bar{Q}_{ij}$
(1-form)-(1-vector)	$\underline{R}'_I{}^J = \mathcal{J} J_I^i \underline{R}_i{}^j J_j^J$	$R'_I{}^J = J_I^i R_i{}^j J_j^J$	$\bar{R}'_I{}^J = \mathcal{J}' J_I^i \bar{R}_i{}^j J_j^J$
(1-vector)-(1-form)	$\underline{S}'^I{}_J = \mathcal{J} J_i^I \underline{S}^i{}_j J_j^J$	$S'^I{}_J = J_i^I S^i{}_j J_j^J$	$\bar{S}'^I{}_J = \mathcal{J}' J_i^I \bar{S}^i{}_j J_j^J$
2-vector	$\underline{T}'^{IJ} = \mathcal{J} \underline{T}^{ij} J_i^I J_j^J$	$T'^{IJ} = T^{ij} J_i^I J_j^J$	$\bar{T}'^{IJ} = \mathcal{J}' \bar{T}^{ij} J_i^I J_j^J$
\vdots	\vdots	\vdots	\vdots

Table 1.1: Changes of the components of the capacities, tensors and densities under a change of variables.

1.8.7 Appendix: Covariant Derivatives

Capacity	Tensor	Density
$\nabla_k \underline{s} = \partial_k \underline{s} + \Gamma_{k\underline{s}}$	$\nabla_k s = \partial_k s$	$\nabla_k \bar{s} = \partial_k \bar{s} - \Gamma_{k\bar{s}}$
$\nabla_k \underline{F}_i = \partial_k \underline{F}_i + \Gamma_{k\underline{F}_i} - \Gamma_{ki}{}^s \underline{F}_s$	$\nabla_k F_i = \partial_k F_i - \Gamma_{ki}{}^s F_s$	$\nabla_k \bar{F}_i = \partial_k \bar{F}_i - \Gamma_{k\bar{F}_i} - \Gamma_{ki}{}^s \bar{F}_s$
$\nabla_k \underline{V}^i = \partial_k \underline{V}^i + \Gamma_{k\underline{V}^i} + \Gamma_{ks}{}^i \underline{V}^s$	$\nabla_k V^i = \partial_k V^i + \Gamma_{ks}{}^i V^s$	$\nabla_k \bar{V}^i = \partial_k \bar{V}^i - \Gamma_{k\bar{V}^i} + \Gamma_{ks}{}^i \bar{V}^s$
$\nabla_k \underline{Q}_{ij} = \partial_k \underline{Q}_{ij} + \Gamma_{k\underline{Q}_{ij}} - \Gamma_{ki}{}^s \underline{Q}_{sj} - \Gamma_{kj}{}^s \underline{Q}_{is}$	$\nabla_k Q_{ij} = \partial_k Q_{ij} - \Gamma_{ki}{}^s Q_{sj} - \Gamma_{kj}{}^s Q_{is}$	$\nabla_k \bar{Q}_{ij} = \partial_k \bar{Q}_{ij} - \Gamma_{k\bar{Q}_{ij}} - \Gamma_{ki}{}^s \bar{Q}_{sj} - \Gamma_{kj}{}^s \bar{Q}_{is}$
$\nabla_k \underline{R}_i{}^j = \partial_k \underline{R}_i{}^j + \Gamma_{k\underline{R}_i{}^j} - \Gamma_{ki}{}^s \underline{R}_s{}^j + \Gamma_{ks}{}^j \underline{R}_i{}^s$	$\nabla_k R_i{}^j = \partial_k R_i{}^j - \Gamma_{ki}{}^s R_s{}^j + \Gamma_{ks}{}^j R_i{}^s$	$\nabla_k \bar{R}_i{}^j = \partial_k \bar{R}_i{}^j - \Gamma_{k\bar{R}_i{}^j} - \Gamma_{ki}{}^s \bar{R}_s{}^j + \Gamma_{ks}{}^j \bar{R}_i{}^s$
$\nabla_k \underline{S}^i{}_j = \partial_k \underline{S}^i{}_j + \Gamma_{k\underline{S}^i{}_j} + \Gamma_{ks}{}^i \underline{S}^s{}_j - \Gamma_{kj}{}^s \underline{S}^i{}_s$	$\nabla_k S^i{}_j = \partial_k S^i{}_j + \Gamma_{ks}{}^i S^s{}_j - \Gamma_{kj}{}^s S^i{}_s$	$\nabla_k \bar{S}^i{}_j = \partial_k \bar{S}^i{}_j - \Gamma_{k\bar{S}^i{}_j} + \Gamma_{ks}{}^i \bar{S}^s{}_j - \Gamma_{kj}{}^s \bar{S}^i{}_s$
$\nabla_k \underline{T}^{ij} = \partial_k \underline{T}^{ij} + \Gamma_{k\underline{T}^{ij}} + \Gamma_{ks}{}^i \underline{T}^{sj} + \Gamma_{ks}{}^j \underline{T}^{is}$	$\nabla_k T^{ij} = \partial_k T^{ij} + \Gamma_{ks}{}^i T^{sj} + \Gamma_{ks}{}^j T^{is}$	$\nabla_k \bar{T}^{ij} = \partial_k \bar{T}^{ij} - \Gamma_{k\bar{T}^{ij}} + \Gamma_{ks}{}^i \bar{T}^{sj} + \Gamma_{ks}{}^j \bar{T}^{is}$
⋮	⋮	⋮

Table 1.2: Covariant derivatives for capacities, tensors and densities.

1.8.8 Appendix: Formulas of Vector Analysis

Let be \mathbf{a} , \mathbf{b} , and \mathbf{c} vector fields, φ a scalar field, and $\Delta\mathbf{a}$ the vector Laplacian (the Laplacian applied to each component of the vector). The following list of identities holds:

$$\operatorname{div} \operatorname{rot} \mathbf{a} = 0 \quad (1.216)$$

$$\operatorname{rot} \operatorname{grad} \varphi = 0 \quad (1.217)$$

$$\operatorname{div}(\varphi\mathbf{a}) = (\operatorname{grad} \varphi) \cdot \mathbf{a} + \varphi(\operatorname{div} \mathbf{a}) \quad (1.218)$$

$$\operatorname{rot}(\varphi\mathbf{a}) = (\operatorname{grad} \varphi) \times \mathbf{a} + \varphi(\operatorname{rot} \mathbf{a}) \quad (1.219)$$

$$\operatorname{grad}(\mathbf{a} \cdot \mathbf{b}) = (\mathbf{a} \cdot \nabla)\mathbf{b} + (\mathbf{b} \cdot \nabla)\mathbf{a} + \mathbf{a} \times (\operatorname{rot} \mathbf{b}) + \mathbf{b} \times (\operatorname{rot} \mathbf{a}) \quad (1.220)$$

$$\operatorname{div}(\mathbf{a} \times \mathbf{b}) = \mathbf{b} \cdot (\operatorname{rot} \mathbf{a}) - \mathbf{a} \cdot (\operatorname{rot} \mathbf{b}) \quad (1.221)$$

$$\operatorname{rot}(\mathbf{a} \times \mathbf{b}) = \mathbf{a}(\operatorname{div} \mathbf{b}) - \mathbf{b}(\operatorname{div} \mathbf{a}) + (\mathbf{b} \cdot \nabla)\mathbf{a} - (\mathbf{a} \cdot \nabla)\mathbf{b} \quad (1.222)$$

$$\operatorname{rot} \operatorname{rot} \mathbf{a} = \operatorname{grad}(\operatorname{div} \mathbf{a}) - \Delta\mathbf{a} . \quad (1.223)$$

Using the nabla symbol everywhere, these equations become:

$$\nabla \cdot (\nabla \times \mathbf{a}) = 0 \quad (1.224)$$

$$\nabla \times (\nabla \cdot \mathbf{a}) = 0 \quad (1.225)$$

$$\nabla \cdot (\varphi\mathbf{a}) = (\nabla\varphi) \cdot \mathbf{a} + \varphi(\nabla \cdot \mathbf{a}) \quad (1.226)$$

$$\nabla \times (\varphi\mathbf{a}) = (\nabla\varphi) \times \mathbf{a} + \varphi(\nabla \times \mathbf{a}) \quad (1.227)$$

$$\nabla(\mathbf{a} \cdot \mathbf{b}) = (\mathbf{a} \cdot \nabla)\mathbf{b} + (\mathbf{b} \cdot \nabla)\mathbf{a} + \mathbf{a} \times (\nabla \times \mathbf{b}) + \mathbf{b} \times (\nabla \times \mathbf{a}) \quad (1.228)$$

$$\nabla \cdot (\mathbf{a} \times \mathbf{b}) = \mathbf{b} \cdot (\nabla \times \mathbf{a}) - \mathbf{a} \cdot (\nabla \times \mathbf{b}) \quad (1.229)$$

$$\nabla \times (\mathbf{a} \times \mathbf{b}) = \mathbf{a}(\nabla \cdot \mathbf{b}) - \mathbf{b}(\nabla \cdot \mathbf{a}) + (\mathbf{b} \cdot \nabla)\mathbf{a} - (\mathbf{a} \cdot \nabla)\mathbf{b} \quad (1.230)$$

$$\nabla \times (\nabla \times \mathbf{a}) = \nabla(\nabla \cdot \mathbf{a}) - \Delta\mathbf{a} . \quad (1.231)$$

The following three vector equations are also often useful:

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) \quad (1.232)$$

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c}) \cdot \mathbf{b} - (\mathbf{a} \cdot \mathbf{b}) \cdot \mathbf{c} \quad (1.233)$$

$$(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d}) = \mathbf{a} \cdot [\mathbf{b} \times (\mathbf{c} \times \mathbf{d})] = (\mathbf{a} \cdot \mathbf{c})(\mathbf{b} \cdot \mathbf{d}) - (\mathbf{a} \cdot \mathbf{d})(\mathbf{b} \cdot \mathbf{c}) \quad (1.234)$$

As, in tensor notations, the scalar product of two vectors is $\mathbf{a} \cdot \mathbf{b} = a_i b^i$, and the vector product has components $(\mathbf{a} \times \mathbf{b})_i = \varepsilon_{ijk} a^j b^k$ (see section XXX), the identities 1.232–1.234 correspond respectively to:

$$\nabla_i \varepsilon^{ijk} \nabla_j a_k = 0 \quad (1.235)$$

$$\varepsilon^{ijk} \nabla_j \nabla_k \varphi = 0 \quad (1.236)$$

$$a_i \varepsilon^{ijk} b_j c_k = b_i \varepsilon^{ijk} c_j a_k = c_i \varepsilon^{ijk} a_j b_k \quad (1.237)$$

$$\varepsilon^{ijk} a_j (\varepsilon_{klm} b^\ell c^m) = (a_j c^j) b^i - (a_j b^j) c^i \quad (1.238)$$

$$(\varepsilon^{ijk} a_j b_k)(\varepsilon_{ilm} c^\ell d^m) = a_i (\varepsilon^{ijk} b_j (\varepsilon_{klm} c^\ell d^m)) , \quad (1.239)$$

while the identities 1.226–1.231 correspond respectively to

$$\nabla_i (\varphi a^i) = (\nabla_i \varphi) a^i + \varphi (\nabla_i a^i) \quad (1.240)$$

$$\varepsilon^{ijk} \nabla_j (\varphi a_k) = \varepsilon^{ijk} (\nabla_j \varphi) a_k + \varphi \varepsilon^{ijk} \nabla_j a_k \quad (1.241)$$

$$\nabla_i (a_j b^j) = (a^j \nabla_j) b_i + (b^j \nabla_j) a_i + \varepsilon_{ijk} a^j (\varepsilon^{klm} \nabla_\ell b_m) + \varepsilon_{ijk} b^j (\varepsilon^{klm} \nabla_\ell a_m) \quad (1.242)$$

$$\nabla_i (\varepsilon^{ijk} a_j b_k) = b_k \varepsilon^{kij} \nabla_i a_j - a_j \varepsilon^{jik} \nabla_i b_k \quad (1.243)$$

$$\varepsilon^{ijk} \nabla_j (\varepsilon_{klm} a^\ell b^m) a^i \nabla_j b^j - b^i \nabla_j a^j + b^j \nabla_j a^i - a^j \nabla_j b^i \quad (1.244)$$

$$\varepsilon^{ijk} \nabla_j (\varepsilon_{klm} \nabla^\ell a^m) = \nabla^i (\nabla_j a^j) - \nabla^j \nabla_j a^i , \quad (1.245)$$

where the (inelegant) notation ∇^i represents $g^{ij} \nabla_j$.

The truth of the set of equations 1.235–1.245, when not obvious, is easily demonstrated by the simple use of the property (see section XXX)

$$\varepsilon_{ijk} \varepsilon^{klm} = \delta_i^\ell \delta_j^m - \delta_i^m \delta_j^\ell \quad (1.246)$$

1.8.9 Appendix: Metric, Connection, etc. in Usual Coordinate Systems

[Note: This appendix shall probably be suppressed.]

1.8.9.1 Cartesian Coordinates

1.8.9.1.1 Line element

$$ds^2 = dx^2 + dy^2 + dz^2 \quad (1.247)$$

1.8.9.1.2 Metric

$$\begin{pmatrix} g_{xx} & g_{xy} & g_{xz} \\ g_{yx} & g_{yy} & g_{yz} \\ g_{zx} & g_{zy} & g_{zz} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (1.248)$$

1.8.9.1.3 Fundamental density

$$\bar{g} = 1 \quad (1.249)$$

1.8.9.1.4 Connection

$$\begin{pmatrix} \Gamma_{xx}^x & \Gamma_{xy}^x & \Gamma_{xz}^x \\ \Gamma_{yx}^x & \Gamma_{yy}^x & \Gamma_{yz}^x \\ \Gamma_{zx}^x & \Gamma_{zy}^x & \Gamma_{zz}^x \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} \Gamma_{xx}^y & \Gamma_{xy}^y & \Gamma_{xz}^y \\ \Gamma_{yx}^y & \Gamma_{yy}^y & \Gamma_{yz}^y \\ \Gamma_{zx}^y & \Gamma_{zy}^y & \Gamma_{zz}^y \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} \Gamma_{xx}^z & \Gamma_{xy}^z & \Gamma_{xz}^z \\ \Gamma_{yx}^z & \Gamma_{yy}^z & \Gamma_{yz}^z \\ \Gamma_{zx}^z & \Gamma_{zy}^z & \Gamma_{zz}^z \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (1.250)$$

1.8.9.1.5 Contracted connection

$$\begin{pmatrix} \Gamma_x \\ \Gamma_y \\ \Gamma_z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad (1.251)$$

1.8.9.1.6 Relationship between covariant and contravariant components for first order tensors

$$\begin{pmatrix} V_x \\ V_y \\ V_z \end{pmatrix} = \begin{pmatrix} V^x \\ V^y \\ V^z \end{pmatrix} \quad (1.252)$$

1.8.9.1.7 Relationship between covariant and contravariant components for second order tensors

$$\begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{pmatrix} = \begin{pmatrix} T_x^x & T_x^y & T_x^z \\ T_y^x & T_y^y & T_y^z \\ T_z^x & T_z^y & T_z^z \end{pmatrix} = \begin{pmatrix} T^{xx} & T^{xy} & T^{xz} \\ T^{yx} & T^{yy} & T^{yz} \\ T^{zx} & T^{zy} & T^{zz} \end{pmatrix} \quad (1.253)$$

1.8.9.1.8 Norm of the vectors of the natural basis

$$\|\mathbf{e}_x\| = \|\mathbf{e}_y\| = \|\mathbf{e}_z\| = 1 \quad (1.254)$$

1.8.9.1.9 Norm of the vectors of the normed basis

$$\|\hat{\mathbf{e}}_x\| = \|\hat{\mathbf{e}}_y\| = \|\hat{\mathbf{e}}_z\| = 1 \quad (1.255)$$

1.8.9.1.10 Missing Comment: give also the norms of the vectors of the dual basis.

1.8.9.1.11 Relations between components on the natural and the normed basis for first order tensors

$$\begin{pmatrix} V_x \\ V_y \\ V_z \end{pmatrix} = \begin{pmatrix} \hat{V}_x \\ \hat{V}_y \\ \hat{V}_z \end{pmatrix} \quad ; \quad \begin{pmatrix} V^x \\ V^y \\ V^z \end{pmatrix} = \begin{pmatrix} \hat{V}^x \\ \hat{V}^y \\ \hat{V}^z \end{pmatrix} \quad (1.256)$$

1.8.9.1.12 Relations between components on the natural and the normed basis for second order tensors

$$\begin{aligned} \begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{pmatrix} &= \begin{pmatrix} \hat{T}_{xx} & \hat{T}_{xy} & \hat{T}_{xz} \\ \hat{T}_{yx} & \hat{T}_{yy} & \hat{T}_{yz} \\ \hat{T}_{zx} & \hat{T}_{zy} & \hat{T}_{zz} \end{pmatrix} \\ \begin{pmatrix} T_x^x & T_x^y & T_x^z \\ T_y^x & T_y^y & T_y^z \\ T_z^x & T_z^y & T_z^z \end{pmatrix} &= \begin{pmatrix} \hat{T}_x^x & \hat{T}_x^y & \hat{T}_x^z \\ \hat{T}_y^x & \hat{T}_y^y & \hat{T}_y^z \\ \hat{T}_z^x & \hat{T}_z^y & \hat{T}_z^z \end{pmatrix} \\ \begin{pmatrix} T^{xx} & T^{xy} & T^{xz} \\ T^{yx} & T^{yy} & T^{yz} \\ T^{zx} & T^{zy} & T^{zz} \end{pmatrix} &= \begin{pmatrix} \hat{T}^{xx} & \hat{T}^{xy} & \hat{T}^{xz} \\ \hat{T}^{yx} & \hat{T}^{yy} & \hat{T}^{yz} \\ \hat{T}^{zx} & \hat{T}^{zy} & \hat{T}^{zz} \end{pmatrix} \end{aligned} \quad (1.257)$$

1.8.9.2 Spherical Coordinates

1.8.9.2.1 Line element

$$ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\varphi^2 \quad (1.258)$$

1.8.9.2.2 Metric

$$\begin{pmatrix} g_{rr} & g_{r\theta} & g_{r\varphi} \\ g_{\theta r} & g_{\theta\theta} & g_{\theta\varphi} \\ g_{\varphi r} & g_{\varphi\theta} & g_{\varphi\varphi} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2 \theta \end{pmatrix} \quad (1.259)$$

1.8.9.2.3 Fundamental density

$$\bar{g} = r^2 \sin \theta \quad (1.260)$$

1.8.9.2.4 Connection

$$\begin{aligned} \begin{pmatrix} \Gamma_{rr}^r & \Gamma_{r\theta}^r & \Gamma_{r\varphi}^r \\ \Gamma_{\theta r}^r & \Gamma_{\theta\theta}^r & \Gamma_{\theta\varphi}^r \\ \Gamma_{\varphi r}^r & \Gamma_{\varphi\theta}^r & \Gamma_{\varphi\varphi}^r \end{pmatrix} &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & -r & 0 \\ 0 & 0 & -r \sin^2 \theta \end{pmatrix} \\ \begin{pmatrix} \Gamma_{rr}^\theta & \Gamma_{r\theta}^\theta & \Gamma_{r\varphi}^\theta \\ \Gamma_{\theta r}^\theta & \Gamma_{\theta\theta}^\theta & \Gamma_{\theta\varphi}^\theta \\ \Gamma_{\varphi r}^\theta & \Gamma_{\varphi\theta}^\theta & \Gamma_{\varphi\varphi}^\theta \end{pmatrix} &= \begin{pmatrix} 0 & 1/r & 0 \\ 1/r & 0 & 0 \\ 0 & 0 & -\sin \theta \cos \theta \end{pmatrix} \\ \begin{pmatrix} \Gamma_{rr}^\varphi & \Gamma_{r\theta}^\varphi & \Gamma_{r\varphi}^\varphi \\ \Gamma_{\theta r}^\varphi & \Gamma_{\theta\theta}^\varphi & \Gamma_{\theta\varphi}^\varphi \\ \Gamma_{\varphi r}^\varphi & \Gamma_{\varphi\theta}^\varphi & \Gamma_{\varphi\varphi}^\varphi \end{pmatrix} &= \begin{pmatrix} 0 & 0 & 1/r \\ 0 & 0 & \cotg \theta \\ 1/r & \cotg \theta & 0 \end{pmatrix} \end{aligned} \quad (1.261)$$

1.8.9.2.5 Contracted connection

$$\begin{pmatrix} \Gamma_r \\ \Gamma_\theta \\ \Gamma_\varphi \end{pmatrix} = \begin{pmatrix} 2/r \\ \cotg \theta \\ 0 \end{pmatrix} \quad (1.262)$$

1.8.9.2.6 Relationship between covariant and contravariant components for first order tensors

$$\begin{pmatrix} V_r \\ V_\theta \\ V_\varphi \end{pmatrix} = \begin{pmatrix} V^r \\ r^2 V^\theta \\ r^2 \sin^2 \theta V^\varphi \end{pmatrix} \quad (1.263)$$

1.8.9.2.7 Relationship between covariant and contravariant components for second order tensors

$$\begin{pmatrix} T_{rr} & \frac{1}{r^2} T_{r\theta} & \frac{1}{r^2 \sin^2 \theta} T_{r\varphi} \\ T_{\theta r} & \frac{1}{r^2} T_{\theta\theta} & \frac{1}{r^2 \sin^2 \theta} T_{\theta\varphi} \\ T_{\varphi r} & \frac{1}{r^2} T_{\varphi\theta} & \frac{1}{r^2 \sin^2 \theta} T_{\varphi\varphi} \end{pmatrix} = \begin{pmatrix} T_r^r & T_r^\theta & T_r^\varphi \\ T_\theta^r & T_\theta^\theta & T_\theta^\varphi \\ T_\varphi^r & T_\varphi^\theta & T_\varphi^\varphi \end{pmatrix} = \begin{pmatrix} T^{rr} & T^{r\theta} & T^{r\varphi} \\ r^2 T^{\theta r} & r^2 T^{\theta\theta} & r^2 T^{\theta\varphi} \\ r^2 \sin^2 \theta T^{\varphi r} & r^2 \sin^2 \theta T^{\varphi\theta} & r^2 \sin^2 \theta T^{\varphi\varphi} \end{pmatrix} \quad (1.264)$$

1.8.9.2.8 Norm of the vectors of the natural basis

$$\|\mathbf{e}_r\| = 1 \quad ; \quad \|\mathbf{e}_\theta\| = r \quad ; \quad \|\mathbf{e}_\varphi\| = r \sin \theta \quad (1.265)$$

1.8.9.2.9 Norm of the vectors of the normed basis

$$\|\hat{\mathbf{e}}_r\| = \|\hat{\mathbf{e}}_\theta\| = \|\hat{\mathbf{e}}_\varphi\| = 1 \quad (1.266)$$

1.8.9.2.10 **Missing** Comment: give also the norms of the vectors of the dual basis.

1.8.9.2.11 Relations between components on the natural and the normed basis for first order tensors

$$\begin{pmatrix} V_r \\ V_\theta \\ V_\varphi \end{pmatrix} = \begin{pmatrix} \hat{V}_r \\ r \hat{V}_\theta \\ r \sin \theta \hat{V}_\varphi \end{pmatrix} \quad ; \quad \begin{pmatrix} V^r \\ V^\theta \\ V^\varphi \end{pmatrix} = \begin{pmatrix} \hat{V}^r \\ \frac{1}{r} \hat{V}^\theta \\ \frac{1}{r \sin \theta} \hat{V}^\varphi \end{pmatrix} \quad (1.267)$$

1.8.9.2.12 Relations between components on the natural and the normed basis for second order tensors

$$\begin{aligned} \begin{pmatrix} T_{rr} & T_{r\theta} & T_{r\varphi} \\ T_{\theta r} & T_{\theta\theta} & T_{\theta\varphi} \\ T_{\varphi r} & T_{\varphi\theta} & T_{\varphi\varphi} \end{pmatrix} &= \begin{pmatrix} \hat{T}_{rr} & r \hat{T}_{r\theta} & r \sin \theta \hat{T}_{r\varphi} \\ r \hat{T}_{\theta r} & r^2 \hat{T}_{\theta\theta} & r^2 \sin \theta \hat{T}_{\theta\varphi} \\ r \sin \theta \hat{T}_{\varphi r} & r^2 \sin \theta \hat{T}_{\varphi\theta} & r^2 \sin^2 \theta \hat{T}_{\varphi\varphi} \end{pmatrix} \\ \begin{pmatrix} T_r^r & T_r^\theta & T_r^\varphi \\ T_\theta^r & T_\theta^\theta & T_\theta^\varphi \\ T_\varphi^r & T_\varphi^\theta & T_\varphi^\varphi \end{pmatrix} &= \begin{pmatrix} \hat{T}_r^r & \frac{1}{r} \hat{T}_r^\theta & \frac{1}{r \sin \theta} \hat{T}_r^\varphi \\ r \hat{T}_\theta^r & \hat{T}_\theta^\theta & \frac{1}{\sin \theta} \hat{T}_\theta^\varphi \\ r \sin \theta \hat{T}_\varphi^r & \sin \theta \hat{T}_\varphi^\theta & \hat{T}_\varphi^\varphi \end{pmatrix} \\ \begin{pmatrix} T^{rr} & T^{r\theta} & T^{r\varphi} \\ T^{\theta r} & T^{\theta\theta} & T^{\theta\varphi} \\ T^{\varphi r} & T^{\varphi\theta} & T^{\varphi\varphi} \end{pmatrix} &= \begin{pmatrix} \hat{T}^{rr} & \frac{1}{r} \hat{T}^{r\theta} & \frac{1}{r \sin \theta} \hat{T}^{r\varphi} \\ \frac{1}{r} \hat{T}^{\theta r} & \frac{1}{r^2} \hat{T}^{\theta\theta} & \frac{1}{r^2 \sin \theta} \hat{T}^{\theta\varphi} \\ \frac{1}{r \sin \theta} \hat{T}^{\varphi r} & \frac{1}{r^2 \sin \theta} \hat{T}^{\varphi\theta} & \frac{1}{r^2 \sin^2 \theta} \hat{T}^{\varphi\varphi} \end{pmatrix} \end{aligned} \quad (1.268)$$

Note: say somewhere in this appendix that the two following formulas are quite useful in deriving the formulas above.

$$\frac{1}{r^n} \frac{\partial}{\partial r} (r^n \psi) = \frac{\partial \psi}{\partial r} + \frac{n}{r} \psi \quad (1.269)$$

$$\frac{1}{\sin^n \vartheta} \frac{\partial}{\partial \vartheta} (\sin^n \vartheta \psi) = \frac{\partial \psi}{\partial \vartheta} + n \cot \vartheta \psi. \quad (1.270)$$

1.8.9.3 Cylindrical Coordinates: Metric, Connection ...

1.8.9.3.1 Line element

$$ds^2 = dr^2 + r^2 d\varphi^2 + dz^2 \quad (1.271)$$

1.8.9.3.2 Metric

$$\begin{pmatrix} g_{rr} & g_{r\varphi} & g_{rz} \\ g_{\varphi r} & g_{\varphi\varphi} & g_{\varphi z} \\ g_{zr} & g_{z\varphi} & g_{zz} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (1.272)$$

1.8.9.3.3 Fundamental density

$$\bar{g} = r \quad (1.273)$$

1.8.9.3.4 Connection

$$\begin{aligned} \begin{pmatrix} \Gamma_{rr}^r & \Gamma_{r\varphi}^r & \Gamma_{rz}^r \\ \Gamma_{\varphi r}^r & \Gamma_{\varphi\varphi}^r & \Gamma_{\varphi z}^r \\ \Gamma_{zr}^r & \Gamma_{z\varphi}^r & \Gamma_{zz}^r \end{pmatrix} &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & -r & 0 \\ 0 & 0 & 0 \end{pmatrix} \\ \begin{pmatrix} \Gamma_{rr}^\varphi & \Gamma_{r\varphi}^\varphi & \Gamma_{rz}^\varphi \\ \Gamma_{\varphi r}^\varphi & \Gamma_{\varphi\varphi}^\varphi & \Gamma_{\varphi z}^\varphi \\ \Gamma_{zr}^\varphi & \Gamma_{z\varphi}^\varphi & \Gamma_{zz}^\varphi \end{pmatrix} &= \begin{pmatrix} 0 & 1/r & 0 \\ 1/r & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\ \begin{pmatrix} \Gamma_{rr}^z & \Gamma_{r\varphi}^z & \Gamma_{rz}^z \\ \Gamma_{\varphi r}^z & \Gamma_{\varphi\varphi}^z & \Gamma_{\varphi z}^z \\ \Gamma_{zr}^z & \Gamma_{z\varphi}^z & \Gamma_{zz}^z \end{pmatrix} &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{aligned} \quad (1.274)$$

1.8.9.3.5 Contracted connection

$$\begin{pmatrix} \Gamma_r \\ \Gamma_\varphi \\ \Gamma_z \end{pmatrix} = \begin{pmatrix} 1/r \\ 0 \\ 0 \end{pmatrix} \quad (1.275)$$

1.8.9.3.6 Relationship between covariant and contravariant components for first order tensors

$$\begin{pmatrix} V_r \\ V_\varphi \\ V_z \end{pmatrix} = \begin{pmatrix} V^r \\ r^2 V^\varphi \\ V^z \end{pmatrix} \quad (1.276)$$

1.8.9.3.7 Relationship between covariant and contravariant components for second order tensors

$$\begin{pmatrix} T_{rr} & \frac{1}{r^2} T_{r\varphi} & T_{rz} \\ T_{\varphi r} & \frac{1}{r^2} T_{\varphi\varphi} & T_{\varphi z} \\ T_{zr} & \frac{1}{r^2} T_{z\varphi} & T_{zz} \end{pmatrix} = \begin{pmatrix} T_r^r & T_r^\varphi & T_r^z \\ T_\varphi^r & T_\varphi^\varphi & T_\varphi^z \\ T_z^r & T_z^\varphi & T_z^z \end{pmatrix} = \begin{pmatrix} T^{rr} & T^{r\varphi} & T^{rz} \\ r^2 T^{\varphi r} & r^2 T^{\varphi\varphi} & r^2 T^{\varphi z} \\ T^{zr} & T^{z\theta} & T^{zz} \end{pmatrix} \quad (1.277)$$

1.8.9.3.8 Norm of the vectors of the natural basis

$$\|\mathbf{e}_r\| = 1 \quad ; \quad \|\mathbf{e}_\varphi\| = r \quad ; \quad \|\mathbf{e}_z\| = 1 \quad (1.278)$$

1.8.9.3.9 Norm of the vectors of the normed basis

$$\|\widehat{\mathbf{e}}_r\| = \|\widehat{\mathbf{e}}_\varphi\| = \|\widehat{\mathbf{e}}_z\| = 1 \quad (1.279)$$

1.8.9.3.10 **Missing** Comment: give also the norms of the vectors of the dual basis.

1.8.9.3.11 Relations between components on the natural and the normed basis for first order tensors

$$\begin{pmatrix} V_r \\ V_\varphi \\ V_z \end{pmatrix} = \begin{pmatrix} \widehat{V}_r \\ r \widehat{V}_\varphi \\ \widehat{V}_z \end{pmatrix} \quad ; \quad \begin{pmatrix} V^r \\ V^\varphi \\ V^z \end{pmatrix} = \begin{pmatrix} \widehat{V}^r \\ \frac{1}{r} \widehat{V}^\varphi \\ \widehat{V}^z \end{pmatrix} \quad (1.280)$$

1.8.9.3.12 Relations between components on the natural and the normed basis for second order tensors

$$\begin{aligned} \begin{pmatrix} T_{rr} & T_{r\varphi} & T_{rz} \\ T_{\varphi r} & T_{\varphi\varphi} & T_{\varphi z} \\ T_{zr} & T_{z\varphi} & T_{zz} \end{pmatrix} &= \begin{pmatrix} \widehat{T}_{rr} & r\widehat{T}_{r\varphi} & \widehat{T}_{rz} \\ r\widehat{T}_{\varphi r} & r^2\widehat{T}_{\varphi\varphi} & r\widehat{T}_{\varphi z} \\ \widehat{T}_{zr} & r\widehat{T}_{z\varphi} & \widehat{T}_{zz} \end{pmatrix} \\ \begin{pmatrix} T_r^r & T_r^\varphi & T_r^z \\ T_\varphi^r & T_\varphi^\varphi & T_\varphi^z \\ T_z^r & T_z^\varphi & T_z^z \end{pmatrix} &= \begin{pmatrix} \widehat{T}_r^r & \frac{1}{r}\widehat{T}_r^\varphi & \widehat{T}_r^z \\ r\widehat{T}_\varphi^r & \widehat{T}_\varphi^\varphi & r\widehat{T}_\varphi^z \\ \widehat{T}_z^r & \frac{1}{r}\widehat{T}_z^\varphi & \widehat{T}_z^z \end{pmatrix} \\ \begin{pmatrix} T^{rr} & T^{r\varphi} & T^{rz} \\ T^{\varphi r} & T^{\varphi\varphi} & T^{\varphi z} \\ T^{zr} & T^{z\varphi} & T^{zz} \end{pmatrix} &= \begin{pmatrix} \widehat{T}^{rr} & \frac{1}{r}\widehat{T}^{r\varphi} & \widehat{T}^{rz} \\ \frac{1}{r}\widehat{T}^{\varphi r} & \frac{1}{r^2}\widehat{T}^{\varphi\varphi} & \frac{1}{r}\widehat{T}^{\varphi z} \\ \widehat{T}^{zr} & \frac{1}{r}\widehat{T}^{z\varphi} & \widehat{T}^{zz} \end{pmatrix} \end{aligned} \quad (1.281)$$

1.8.10 Appendix: Gradient, Divergence and Curl in Usual Coordinate Systems

Here we analyze the 3-D Euclidean space, using Cartesian, spherical or cylindrical coordinates. The words scalar, vector, and tensor mean “true” scalars, vectors and tensors, respectively. The scalar densities, vector densities and tensor densities (see section XXX) are named explicitly.

1.8.10.1 Definitions

If $\mathbf{x} \rightarrow \phi(\mathbf{x})$ is a scalar field, its *gradient* is the form defined by

$$G_i = \nabla_i \phi. \quad (1.282)$$

If $\mathbf{x} \rightarrow \bar{V}^i(\mathbf{x})$ is a vector density field, its *divergence* is the scalar density defined by

$$\bar{D} = \nabla_i \bar{V}^i. \quad (1.283)$$

If $\mathbf{x} \rightarrow F_i(\mathbf{x})$ is a form field, its *curl* (or *rotational*) is the vector density defined by

$$\bar{R}^i = \bar{\varepsilon}^{ijk} \nabla_j F_k. \quad (1.284)$$

1.8.10.2 Properties

These definitions are such that we can replace everywhere true (“covariant”) derivatives by partial derivatives (see exercise XXX). This gives, for the gradient of a density,

$$G_i = \nabla_i \phi = \partial_i \phi, \quad (1.285)$$

for the divergence of a vector density,

$$\bar{D} = \nabla_i \bar{V}^i = \partial_i \bar{V}^i, \quad (1.286)$$

and for the curl of a form,

$$\bar{R}^i = \bar{\varepsilon}^{ijk} \nabla_j F_k = \bar{\varepsilon}^{ijk} \partial_j F_k \quad (1.287)$$

[this equation is only valid for spaces without torsion; the general formula is $\bar{R}^i = \bar{\varepsilon}^{ijk} \nabla_j F_k = \bar{\varepsilon}^{ijk} (\partial_j F_k - \frac{1}{2} S_{jk}{}^\ell V_\ell)$].

These equations lead to particularly simple expressions. For instance, the following table shows that the explicit expressions have the same form for Cartesian, spherical and cylindrical coordinates (or for whatever coordinate system).

	Cartesian	Spherical	Cylindrical
Gradient	$G_x = \partial_x \phi$ $G_y = \partial_y \phi$ $G_z = \partial_z \phi$	$G_r = \partial_r \phi$ $G_\theta = \partial_\theta \phi$ $G_\varphi = \partial_\varphi \phi$	$G_r = \partial_r \phi$ $G_\varphi = \partial_\varphi \phi$ $G_z = \partial_z \phi$
Divergence	\bar{D} = $\partial_x \bar{V}^x + \partial_y \bar{V}^y + \partial_z \bar{V}^z$	\bar{D} = $\partial_r \bar{V}^r + \partial_\theta \bar{V}^\theta + \partial_\varphi \bar{V}^\varphi$	\bar{D} = $\partial_r \bar{V}^r + \partial_\varphi \bar{V}^\varphi + \partial_z \bar{V}^z$
Curl	$\bar{R}^x = \partial_y F_z - \partial_z F_y$ $\bar{R}^y = \partial_z F_x - \partial_x F_z$ $\bar{R}^z = \partial_x F_y - \partial_y F_x$	$\bar{R}^r = \partial_\theta F_\varphi - \partial_\varphi F_\theta$ $\bar{R}^\theta = \partial_\varphi F_r - \partial_r F_\varphi$ $\bar{R}^\varphi = \partial_r F_\theta - \partial_\theta F_r$	$\bar{R}^r = \partial_\varphi F_z - \partial_z F_\varphi$ $\bar{R}^\varphi = \partial_z F_r - \partial_r F_z$ $\bar{R}^z = \partial_r F_\varphi - \partial_\varphi F_r$

1.8.10.3 Remarks

Although we have only defined the gradient of a true scalar, the divergence of a vector density, and the curl of a form, the definitions can be immediately be extended by “putting bars on” and “taking bars off” (see section XXX).

As an example, from equation 1.282, we can immediately write the definition of the gradient of a scalar density,

$$\bar{G}_i = \nabla_i \bar{\phi}, \quad (1.288)$$

from equation 1.283 we can write the definition of the divergence of a (true) vector field,

$$D = \nabla_i V^i, \quad (1.289)$$

and from equation 1.284 we can write the definition of the curl of a form as a true vector,

$$R^i = \varepsilon^{ijk} \nabla_j F_k, \quad (1.290)$$

or a true form,

$$R_\ell = g_{\ell i} \varepsilon^{ijk} \nabla_j F_k. \quad (1.291)$$

Although equation 1.289 seems well adapted to the practical computation of the divergence of a true vector, it is better to use 1.286 instead. For we have successively

$$\bar{D} = \partial_i \bar{V}^i \quad \iff \quad \bar{g} D = \partial_i (\bar{g} V^i) \quad \iff \quad D = \frac{1}{\bar{g}} \partial_i (\bar{g} V^i). \quad (1.292)$$

This last expression provides directly compact expressions for the divergence of a vector. For instance, as the fundamental density \bar{g} takes, in Cartesian, spherical and cylindrical coordinates, respectively the values 1 , $r^2 \sin \theta$ and r , this leads to the results of the following table.

$$\text{Divergence, Cartesian coordinates} : D = \frac{\partial V^x}{\partial x} + \frac{\partial V^y}{\partial y} + \frac{\partial V^z}{\partial z} \quad (1.293)$$

$$\text{Divergence, Spherical coordinates} : D = \frac{1}{r^2} \frac{\partial(r^2 V^r)}{\partial r} + \frac{1}{\sin \theta} \frac{\partial(\sin \theta V^\theta)}{\partial \theta} + \frac{\partial V^\varphi}{\partial \varphi} \quad (1.294)$$

$$\text{Divergence, Cylindrical coordinates} : D = \frac{1}{r} \frac{\partial(r V^r)}{\partial r} + \frac{\partial V^\varphi}{\partial \varphi} + \frac{\partial V^z}{\partial z} \quad (1.295)$$

Replacing the components on the natural basis by the components on the normed basis (see section XXX) gives

$$\text{Divergence, Cartesian coordinates} : D = \frac{\partial \hat{V}^x}{\partial x} + \frac{\partial \hat{V}^y}{\partial y} + \frac{\partial \hat{V}^z}{\partial z} \quad (1.296)$$

$$\text{Divergence, Spherical coordinates} : D = \frac{1}{r^2} \frac{\partial(r^2 \hat{V}^r)}{\partial r} + \frac{1}{r \sin \theta} \frac{\partial(\sin \theta \hat{V}^\theta)}{\partial \theta} + \frac{1}{r \sin \theta} \frac{\partial \hat{V}^\varphi}{\partial \varphi} \quad (1.297)$$

$$\text{Divergence, Cylindrical coordinates} : D = \frac{1}{r} \frac{\partial(r \hat{V}^r)}{\partial r} + \frac{1}{r} \frac{\partial \hat{V}^\varphi}{\partial \varphi} + \frac{\partial \hat{V}^z}{\partial z} \quad (1.298)$$

These are the formulas given in elementary texts (not using tensor concepts).

Similarly, although 1.291 seems well adapted to a practical computation of the curl, it is better to go back to equation 1.287. We have, successively,

$$\overline{R}^i = \overline{\varepsilon}^{ijk} \partial_j F_k \iff \overline{g} R^i = \overline{\varepsilon}^{ijk} \partial_j F_k \iff R^i = \frac{1}{\overline{g}} \overline{\varepsilon}^{ijk} \partial_j F_k \iff R_\ell = \frac{1}{\overline{g}} g_{\ell i} \overline{\varepsilon}^{ijk} \partial_j F_k. \quad (1.299)$$

This last expression provides directly compact expressions for the curl. For instance, as the fundamental density \overline{g} takes, in Cartesian, spherical and cylindrical coordinates, respectively the values 1, $r^2 \sin \theta$ and r , this leads to the results of the following table.

$$\begin{aligned} & R_x = \partial_y F_z - \partial_z F_y \\ \text{Curl, Cartesian coordinates} : & R_y = \partial_z F_x - \partial_x F_z \\ & R_z = \partial_x F_y - \partial_y F_x \end{aligned} \quad (1.300)$$

$$\begin{aligned} & R_r = \frac{1}{r^2 \sin \theta} (\partial_\theta F_\varphi - \partial_\varphi F_\theta) \\ \text{Curl, Spherical coordinates} : & R_\theta = \frac{1}{\sin \theta} (\partial_\varphi F_r - \partial_r F_\varphi) \\ & R_\varphi = \sin \theta (\partial_r F_\theta - \partial_\theta F_r) \end{aligned} \quad (1.301)$$

$$\begin{aligned} & R_r = \frac{1}{r} (\partial_\varphi F_z - \partial_z F_\varphi) \\ \text{Curl, Cylindrical coordinates} : & R_\varphi = r (\partial_z F_r - \partial_r F_z) \\ & R_z = \frac{1}{r} (\partial_r F_\varphi - \partial_\varphi F_r) \end{aligned} \quad (1.302)$$

Replacing the components on the natural basis by the components on the normed basis (see section XXX) gives

$$\begin{aligned} \widehat{R}_x &= \partial_y \widehat{F}_z - \partial_z \widehat{F}_y \\ \text{Curl, Cartesian coordinates} &: \widehat{R}_y = \partial_z \widehat{F}_x - \partial_x \widehat{F}_z \\ &\widehat{R}_z = \partial_x \widehat{F}_y - \partial_y \widehat{F}_x \end{aligned} \quad (1.303)$$

$$\begin{aligned} \widehat{R}_r &= \frac{1}{r \sin \theta} \left(\frac{\partial(\sin \theta \widehat{F}_\varphi)}{\partial \theta} - \frac{\partial \widehat{F}_\theta}{\partial \varphi} \right) \\ \text{Curl, Spherical coordinates} &: \widehat{R}_\theta = \frac{1}{r} \left(\frac{1}{\sin \theta} \frac{\partial \widehat{F}_r}{\partial \varphi} - \frac{\partial(r \widehat{F}_\varphi)}{\partial r} \right) \\ &\widehat{R}_\varphi = \frac{1}{r} \left(\frac{\partial(r \widehat{F}_\theta)}{\partial r} - \frac{\partial \widehat{F}_r}{\partial \theta} \right) \end{aligned} \quad (1.304)$$

$$\begin{aligned} \widehat{R}_r &= \frac{1}{r} \left(\frac{\partial \widehat{F}_z}{\partial \varphi} - \frac{\partial(r \widehat{F}_\varphi)}{\partial z} \right) \\ \text{Curl, Cylindrical coordinates} &: \widehat{R}_\varphi = \frac{\partial \widehat{F}_r}{\partial z} - \frac{\partial \widehat{F}_z}{\partial r} \\ &\widehat{R}_z = \frac{1}{r} \left(\frac{\partial(r \widehat{F}_\varphi)}{\partial r} - \frac{\partial \widehat{F}_r}{\partial \varphi} \right) \end{aligned} \quad (1.305)$$

These are the formulas given in elementary texts (not using tensor concepts).

Comment: I should remember not to put this back in a table, as it is not very readable:

	Curl
Cartesian	$\begin{aligned} \widehat{R}_x &= \partial_y \widehat{F}_z - \partial_z \widehat{F}_y \\ \widehat{R}_y &= \partial_z \widehat{F}_x - \partial_x \widehat{F}_z \\ \widehat{R}_z &= \partial_x \widehat{F}_y - \partial_y \widehat{F}_x \end{aligned}$
Spherical	$\begin{aligned} \widehat{R}_r &= \frac{1}{r \sin \theta} \left(\frac{\partial \sin \theta \widehat{F}_\varphi}{\partial \theta} - \frac{\partial \widehat{F}_\theta}{\partial \varphi} \right) \\ \widehat{R}_\theta &= \frac{1}{r} \left(\frac{1}{\sin \theta} \frac{\partial \widehat{F}_r}{\partial \varphi} - \frac{\partial r \widehat{F}_\varphi}{\partial r} \right) \\ \widehat{R}_\varphi &= \frac{1}{r} \left(\frac{\partial r \widehat{F}_\theta}{\partial r} - \frac{\partial \widehat{F}_r}{\partial \theta} \right) \end{aligned}$
Cylindrical	$\begin{aligned} \widehat{R}_r &= \frac{\partial \widehat{F}_z}{\partial \varphi} - \frac{\partial r \widehat{F}_\varphi}{\partial z} \\ \widehat{R}_\varphi &= \frac{\partial \widehat{F}_r}{\partial z} - \frac{\partial \widehat{F}_z}{\partial r} \\ \widehat{R}_z &= \frac{1}{r} \left(\frac{\partial r \widehat{F}_\varphi}{\partial r} - \frac{\partial \widehat{F}_r}{\partial \varphi} \right) \end{aligned}$

1.8.10.3.1 Comment: What follows is not very interesting and should be suppressed.

From 1.288 we can write

$$\widehat{g} G_i = \nabla_i(\widehat{g} \phi), \quad (1.306)$$

which leads to the formula

$$G_i = \frac{1}{\widehat{g}} \nabla_i(\widehat{g}\phi). \quad (1.307)$$

For instance, as the fundamental density \widehat{g} takes, in Cartesian, spherical and cylindrical coordinates, respectively the values 1 , $r^2 \sin \theta$ and r , this leads to the results of the following table.

	Cartesian	Spherical	Cylindrical
Gradient	$\widehat{G}_x = \frac{\partial \overline{\phi}}{\partial x}$ $\widehat{G}_y = \frac{\partial \overline{\phi}}{\partial y}$ $\widehat{G}_z = \frac{\partial \overline{\phi}}{\partial z}$	$\widehat{G}_r = r^2 \frac{\partial}{\partial r} \left(\frac{1}{r^2} \overline{\phi} \right)$ $\widehat{G}_\theta = \sin \theta \frac{\partial}{\partial \theta} \left(\frac{1}{\sin \theta} \overline{\phi} \right)$ $\widehat{G}_\varphi = \frac{\partial \overline{\phi}}{\partial \varphi}$	$\widehat{G}_r = r \frac{\partial}{\partial r} \left(\frac{1}{r} \overline{\phi} \right)$ $\widehat{G}_\varphi = \frac{\partial \overline{\phi}}{\partial \varphi}$ $\widehat{G}_z = \frac{\partial \overline{\phi}}{\partial z}$

1.8.11 Appendix: Connection and Derivative in Different Coordinate Systems

(Comment: mention here the boxes with different coordinate systems).

1.8.11.1 Polar coordinates

(Two-dimensional Euclidean space with non-Cartesian coordinates).

$$ds^2 = r^2 + r^2 d\varphi^2 \quad (1.308)$$

$$\Gamma_{r\varphi}^\varphi = 1/r ; \quad \Gamma_{\varphi r}^\varphi = 1/r ; \quad \Gamma_{\varphi\varphi}^r = -r ; \quad (\text{the others vanish}) \quad (1.309)$$

$$R_{ij} = 0 \quad (1.310)$$

$$\nabla_i V^i = \frac{1}{r} \frac{\partial}{\partial r} (rV^r) + \frac{\partial V^\varphi}{\partial \varphi} \quad (1.311)$$

1.8.11.2 Cylindrical coordinates

(Three-dimensional Euclidean space with non-Cartesian coordinates).

$$ds^2 = r^2 + r^2 d\varphi^2 + dz^2 \quad (1.312)$$

$$\Gamma_{r\varphi}^\varphi = 1/r ; \quad \Gamma_{\varphi r}^\varphi = 1/r ; \quad \Gamma_{\varphi\varphi}^r = -r ; \quad (\text{the others vanish}) \quad (1.313)$$

$$R_{ij} = 0 \quad (1.314)$$

$$\nabla_i V^i = \frac{1}{r} \frac{\partial}{\partial r} (rV^r) + \frac{\partial V^\varphi}{\partial \varphi} + \frac{\partial V^z}{\partial z} \quad (1.315)$$

1.8.11.3 Geographical coordinates

Geographical coordinates

(Two-dimensional non-Euclidean space).

$$ds^2 = R^2(d\theta^2 + \sin^2\theta d\varphi^2) \quad (1.316)$$

$$\Gamma_{\theta\varphi}^\varphi = \cotg \theta ; \quad \Gamma_{\varphi\theta}^\varphi = \cotg \theta ; \quad \Gamma_{\varphi\varphi}^\theta = -\sin \theta \cos \theta ; \quad (\text{the others vanish}) \quad (1.317)$$

$$R_{\theta\theta} = 1/R^2 ; \quad R_{\varphi\varphi} = 1/R^2 ; \quad (\text{the others vanish}) ; \quad R = 2/R^2 \quad (1.318)$$

$$\nabla_i V^i = \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} (\sin \theta V^\theta) + \frac{\partial V^\varphi}{\partial \varphi} \quad (1.319)$$

1.8.11.4 Spherical coordinates

(Three-dimensional Euclidean space).

$$ds^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\varphi^2 \quad (1.320)$$

$$\begin{aligned} \Gamma_{r\theta}^\theta &= 1/r ; & \Gamma_{r\varphi}^\varphi &= 1/r ; & \Gamma_{\theta r}^\theta &= 1/r ; \\ \Gamma_{\theta\theta}^r &= -r ; & \Gamma_{\theta\varphi}^\varphi &= \cotg \theta ; & \Gamma_{\varphi r}^\varphi &= 1/r ; \\ \Gamma_{\varphi\theta}^\varphi &= \cotg \theta ; & \Gamma_{\varphi\varphi}^r &= -r \sin^2 \theta ; & \Gamma_{\varphi\varphi}^\theta &= -\sin \theta \cos \theta ; \\ & & & & & \text{(the others vanish)} \end{aligned} \quad (1.321)$$

$$R_{ij} = 0 \quad (1.322)$$

$$\nabla_i V^i = \frac{1}{r^2} \frac{\partial}{\partial r} (r^2 V^r) + \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} (\sin \theta V^\theta) + \frac{\partial V^\varphi}{\partial \varphi} \quad (1.323)$$

1.8.12 Appendix: Computing in Polar Coordinates

[Note: This appendix is probably to be suppressed.]

1.8.12.1 General formula

1.8.12.1.1 Simple-minded computation

From

$$\operatorname{div} \mathbf{V} = \frac{1}{r} \frac{\partial}{\partial r} (rV^r) + \frac{\partial V^\varphi}{\partial \varphi}, \quad (1.324)$$

we obtain, using a simple-minded discretisation, at

$$\begin{aligned} (\operatorname{div} \mathbf{V})(r, \varphi) &= \frac{1}{r} \frac{(r + \delta r)V^r(r + \delta r, \varphi) - (r - \delta r)V^r(r - \delta r, \varphi)}{2 \delta r} \\ &\quad + \frac{V^\varphi(r, \varphi + \delta \varphi) - V^\varphi(r, \varphi - \delta \varphi)}{2 \delta \varphi}. \end{aligned} \quad (1.325)$$

1.8.12.1.2 Computation through parallel transport

The notion of parallel transport

leads to

$$\begin{aligned} (\operatorname{div} \mathbf{V})(r, \varphi) &= \frac{V^r(r, \varphi \parallel r + \delta r, \varphi) - V^r(r, \varphi \parallel r - \delta r, \varphi)}{2 \delta r} \\ &\quad + \frac{V^\varphi(r, \varphi \parallel r, \varphi + \delta \varphi) - V^\varphi(r, \varphi \parallel r, \varphi - \delta \varphi)}{2 \delta \varphi}, \end{aligned} \quad (1.326)$$

which gives

$$\begin{aligned} (\operatorname{div} \mathbf{V})(r, \varphi) &= \frac{V^r(r + \delta r, \varphi) - V^r(r - \delta r, \varphi)}{2 \delta r} \\ &\quad + \cos(\delta \varphi) \frac{V^\varphi(r, \varphi + \delta \varphi) - V^\varphi(r, \varphi - \delta \varphi)}{2 \delta \varphi} \\ &\quad + \frac{\sin(\delta \varphi)}{\delta \varphi} \frac{1}{r} \frac{V^r(r + \delta r, \varphi) + V^r(r - \delta r, \varphi)}{2}. \end{aligned} \quad (1.327)$$

1.8.12.1.3 Note: Natural basis and “normed” basis

The components on the natural

basis V^r et V^φ are related with the components on the normed basis \widehat{V}^r and \widehat{V}^φ through

$$\widehat{V}^r = V^r \quad (1.328)$$

and

$$\widehat{V}^\varphi = r V^\varphi. \quad (1.329)$$

1.8.12.2 Divergence of a constant field

A constant vector field (oriented “as the x axis”) has components

$$V^r(r, \varphi) = k \cos \varphi \quad (1.330)$$

and

$$V^\varphi(r, \varphi) = -\frac{k}{r} \sin \varphi. \quad (1.331)$$

1.8.12.2.1 Simple-minded computation An exact evaluation of approximation 1.325 gives

$$(\operatorname{div} \mathbf{V})(r, \varphi) = \frac{k}{r} \cos \varphi \left(1 - \frac{\sin(\delta\varphi)}{\delta\varphi} \right), \quad (1.332)$$

expression with an error of order $(\delta\varphi)^2$.

1.8.12.2.2 Computation through parallel transport An exact evaluation of approximation 1.327 gives

$$(\operatorname{div} \mathbf{V})(r, \varphi) = 0, \quad (1.333)$$

as it should.

1.8.13 Appendix: Dual Tensors in 2 3 and 4D

1.8.13.1 Dual tensors in 2-D

In 2-D, we may need to take the following duals of contravariant (antisymmetric) tensors:

$${}^*B_{ij} = \frac{1}{0!} \varepsilon_{ij} B \quad {}^*B_{ij} = \frac{1}{0!} \underline{\varepsilon}_{ij} \overline{B} \quad {}^*\underline{B}_{ij} = \frac{1}{0!} \underline{\varepsilon}_{ij} B \quad (1.334)$$

$${}^*B_i = \frac{1}{1!} \varepsilon_{ij} B^j \quad {}^*B_i = \frac{1}{1!} \underline{\varepsilon}_{ij} \overline{B}^j \quad {}^*\underline{B}_i = \frac{1}{1!} \underline{\varepsilon}_{ij} B^j \quad (1.335)$$

$${}^*B = \frac{1}{2!} \varepsilon_{ij} B^{ij} \quad {}^*B = \frac{1}{2!} \underline{\varepsilon}_{ij} \overline{B}^{ij} \quad {}^*\underline{B} = \frac{1}{2!} \underline{\varepsilon}_{ij} B^{ij} \quad (1.336)$$

We may also need to take duals of covariant tensors:

$${}^*B^{ij} = \frac{1}{0!} \varepsilon^{ij} B \quad {}^*B^{ij} = \frac{1}{0!} \overline{\varepsilon}^{ij} \underline{B} \quad {}^*\overline{B}^{ij} = \frac{1}{0!} \overline{\varepsilon}^{ij} B \quad (1.337)$$

$${}^*B^i = \frac{1}{1!} \varepsilon^{ij} B_j \quad {}^*B^i = \frac{1}{1!} \overline{\varepsilon}^{ij} \underline{B}_j \quad {}^*\overline{B}^i = \frac{1}{1!} \overline{\varepsilon}^{ij} B_j \quad (1.338)$$

$${}^*B = \frac{1}{2!} \varepsilon^{ij} B_{ij} \quad {}^*B = \frac{1}{2!} \overline{\varepsilon}^{ij} \underline{B}_{ij} \quad {}^*\overline{B} = \frac{1}{2!} \overline{\varepsilon}^{ij} B_{ij} \quad (1.339)$$

As in a space with an even number of dimensions the dual of the dual of a tensor of rank p equals $(-1)^p$ the original tensor (see text), we have, in 2-D, that for a tensor with 0 or 2 indices, ${}^*({}^*\mathbf{B}) = \mathbf{B}$, while for a tensor with 1 index, ${}^*({}^*\mathbf{B}) = -\mathbf{B}$.

1.8.13.2 Dual tensors in 3-D

In 3-D, we may need to take the following duals of contravariant (totally antisymmetric) tensors:

$${}^*B_{ijk} = \frac{1}{0!} \varepsilon_{ijk} B \quad {}^*B_{ijk} = \frac{1}{0!} \underline{\varepsilon}_{ijk} \overline{B} \quad {}^*\underline{B}_{ijk} = \frac{1}{0!} \underline{\varepsilon}_{ijk} B \quad (1.340)$$

$${}^*B_{ij} = \frac{1}{1!} \varepsilon_{ijk} B^k \quad {}^*B_{ij} = \frac{1}{1!} \underline{\varepsilon}_{ijk} \overline{B}^k \quad {}^*\underline{B}_{ij} = \frac{1}{1!} \underline{\varepsilon}_{ijk} B^k \quad (1.341)$$

$${}^*B_i = \frac{1}{2!} \varepsilon_{ijk} B^{jk} \quad {}^*B_i = \frac{1}{2!} \underline{\varepsilon}_{ijk} \overline{B}^{jk} \quad {}^*\underline{B}_i = \frac{1}{2!} \underline{\varepsilon}_{ijk} B^{jk} \quad (1.342)$$

$${}^*B = \frac{1}{3!} \varepsilon_{ijk} B^{ijk} \quad {}^*B = \frac{1}{3!} \underline{\varepsilon}_{ijk} \overline{B}^{ijk} \quad {}^*\underline{B} = \frac{1}{3!} \underline{\varepsilon}_{ijk} B^{ijk} \quad (1.343)$$

We may also need to take duals of covariant tensors:

$$*B^{ijk} = \frac{1}{0!} \varepsilon^{ijk} B \quad *B^{ijk} = \frac{1}{0!} \bar{\varepsilon}^{ijk} \underline{B} \quad *\bar{B}^{ijk} = \frac{1}{0!} \bar{\varepsilon}^{ijk} B \quad (1.344)$$

$$*B^{ij} = \frac{1}{1!} \varepsilon^{ijk} B_k \quad *B^{ij} = \frac{1}{1!} \bar{\varepsilon}^{ijk} \underline{B}_k \quad *\bar{B}^{ij} = \frac{1}{1!} \bar{\varepsilon}^{ijk} B_k \quad (1.345)$$

$$*B^i = \frac{1}{2!} \varepsilon^{ijk} B_{jk} \quad *B^i = \frac{1}{2!} \bar{\varepsilon}^{ijk} \underline{B}_{jk} \quad *\bar{B}^i = \frac{1}{2!} \bar{\varepsilon}^{ijk} B_{jk} \quad (1.346)$$

$$*B = \frac{1}{3!} \varepsilon^{ijk} B_{ijk} \quad *B = \frac{1}{3!} \bar{\varepsilon}^{ijk} \underline{B}_{ijk} \quad *\bar{B} = \frac{1}{3!} \bar{\varepsilon}^{ijk} B_{ijk} \quad (1.347)$$

As in a space with an odd number of dimensions the dual of the dual of a tensor always equals the original tensor (see text), we have, in 3-D, that for all tensors above, $*(\mathbf{*B}) = \mathbf{B}$.

1.8.13.3 Dual tensors in 4-D

In 4-D, we may need to take the following duals of contravariant (totally antisymmetric) tensors:

$$*B_{ijkl} = \frac{1}{0!} \varepsilon_{ijkl} B \quad *B_{ijkl} = \frac{1}{0!} \underline{\varepsilon}_{ijkl} \bar{B} \quad *\underline{B}_{ijkl} = \frac{1}{0!} \underline{\varepsilon}_{ijkl} B \quad (1.348)$$

$$*B_{ijk} = \frac{1}{1!} \varepsilon_{ijkl} B^\ell \quad *B_{ijk} = \frac{1}{1!} \underline{\varepsilon}_{ijkl} \bar{B}^\ell \quad *\underline{B}_{ijk} = \frac{1}{1!} \underline{\varepsilon}_{ijkl} B^\ell \quad (1.349)$$

$$*B_{ij} = \frac{1}{2!} \varepsilon_{ijkl} B^{kl} \quad *B_{ij} = \frac{1}{2!} \underline{\varepsilon}_{ijkl} \bar{B}^{kl} \quad *\underline{B}_{ij} = \frac{1}{2!} \underline{\varepsilon}_{ijkl} B^{kl} \quad (1.350)$$

$$*B_i = \frac{1}{3!} \varepsilon_{ijkl} B^{jkl} \quad *B_i = \frac{1}{3!} \underline{\varepsilon}_{ijkl} \bar{B}^{jkl} \quad *\underline{B}_i = \frac{1}{3!} \underline{\varepsilon}_{ijkl} B^{jkl} \quad (1.351)$$

$$*B = \frac{1}{4!} \varepsilon_{ijkl} B^{ijkl} \quad *B = \frac{1}{4!} \underline{\varepsilon}_{ijkl} \bar{B}^{ijkl} \quad *\underline{B} = \frac{1}{4!} \underline{\varepsilon}_{ijkl} B^{ijkl} \quad (1.352)$$

We may also need to take duals of covariant tensors:

$$*B^{ijkl} = \frac{1}{0!} \varepsilon^{ijkl} B \quad *B^{ijkl} = \frac{1}{0!} \bar{\varepsilon}^{ijkl} \underline{B} \quad *\bar{B}^{ijkl} = \frac{1}{0!} \bar{\varepsilon}^{ijkl} B \quad (1.353)$$

$$*B^{ijk} = \frac{1}{1!} \varepsilon^{ijkl} B_\ell \quad *B^{ijk} = \frac{1}{1!} \bar{\varepsilon}^{ijkl} \underline{B}_\ell \quad *\bar{B}^{ijk} = \frac{1}{1!} \bar{\varepsilon}^{ijkl} B_\ell \quad (1.354)$$

$$*B^{ij} = \frac{1}{2!} \varepsilon^{ijkl} B_{kl} \quad *B^{ij} = \frac{1}{2!} \bar{\varepsilon}^{ijkl} \underline{B}_{kl} \quad *\bar{B}^{ij} = \frac{1}{2!} \bar{\varepsilon}^{ijkl} B_{kl} \quad (1.355)$$

$$*B^i = \frac{1}{3!} \varepsilon^{ijkl} B_{jkl} \quad *B^i = \frac{1}{3!} \bar{\varepsilon}^{ijkl} \underline{B}_{jkl} \quad *\bar{B}^i = \frac{1}{3!} \bar{\varepsilon}^{ijkl} B_{jkl} \quad (1.356)$$

$$*B = \frac{1}{4!} \varepsilon^{ijkl} B_{ijkl} \quad *B = \frac{1}{4!} \bar{\varepsilon}^{ijkl} \underline{B}_{ijkl} \quad *\bar{B} = \frac{1}{4!} \bar{\varepsilon}^{ijkl} B_{ijkl} \quad (1.357)$$

As in a space with an even number of dimensions the dual of the dual of a tensor of rank p equals $(-1)^p$ the original tensor (see text), we have, in 4-D, that for a tensor with 0, 2 or 4 indices, $*(\mathbf{*B}) = \mathbf{B}$, while for a tensor with 1 or 3 indices, $*(\mathbf{*B}) = -\mathbf{B}$.

1.8.14 Appendix: Integration in 3D

In a three-dimensional space ($n = 3$), we may have p respectively equal to 2, 1 and 0. This gives the three theorems

$$\int_{3D} d^3\sigma^{ijk} (\nabla \wedge \mathbf{T})_{ijk} = \int_{2D} d^2\sigma^{ij} T_{ij} \quad (1.358)$$

$$\int_{2D} d^2\sigma^{ij} (\nabla \wedge \mathbf{T})_{ij} = \int_{1D} d^1\sigma^i T_i \quad (1.359)$$

$$\int_{1D} d^1\sigma^i (\nabla \wedge \mathbf{T})_i = \int_{0D} d^0\sigma T. \quad (1.360)$$

Explicitly, using the results of sections 1.6.3 and 1.6.4, this gives

$$\int_{3D} d^3\sigma^{ijk} \frac{1}{3} (\nabla_i T_{jk} + \nabla_j T_{ki} + \nabla_k T_{ij}) = \int_{2D} d^2\sigma^{ij} T_{ij} \quad (1.361)$$

$$\int_{2D} d^2\sigma^{ij} \frac{1}{2} (\nabla_i T_j - \nabla_j T_i) = \int_{1D} d^1\sigma^i T_i \quad (1.362)$$

$$\int_{1D} d^1\sigma^i \nabla_i T = \int_{0D} d^0\sigma T, \quad (1.363)$$

or, we use the antisymmetry of the tensors,

$$\int_{3D} d^3\sigma^{ijk} \nabla_i T_{jk} = \int_{2D} d^2\sigma^{ij} T_{ij} \quad (1.364)$$

$$\int_{2D} d^2\sigma^{ij} \nabla_i T_j = \int_{1D} d^1\sigma^i T_i \quad (1.365)$$

$$\int_{1D} d^1\sigma^i \partial_i T = \int_{0D} d^0\sigma T. \quad (1.366)$$

We can now introduce the capacity elements instead of the differential elements:

$$\frac{1}{0!} \int_{3D} d^3\underline{\Sigma} \left(\frac{1}{2!} \bar{\varepsilon}^{ijk} \nabla_i T_{jk} \right) = \frac{1}{1!} \int_{2D} d^2\underline{\Sigma}_i \left(\frac{1}{2!} \bar{\varepsilon}^{ijk} T_{jk} \right) \quad (1.367)$$

$$\frac{1}{1!} \int_{2D} d^2\underline{\Sigma}_i \left(\frac{1}{1!} \bar{\varepsilon}^{ijk} \nabla_j T_k \right) = \frac{1}{2!} \int_{1D} d^1\underline{\Sigma}_{ij} \left(\frac{1}{1!} \bar{\varepsilon}^{ijk} T_k \right) \quad (1.368)$$

$$\frac{1}{2!} \int_{1D} d^1\underline{\Sigma}_{ij} \left(\frac{1}{0!} \bar{\varepsilon}^{ijk} \partial_k T \right) = \frac{1}{3!} \int_{0D} d^0\underline{\Sigma}_{ijk} \left(\frac{1}{0!} \bar{\varepsilon}^{ijk} T \right). \quad (1.369)$$

Introducing explicit expressions for the capacity elements gives

$$\int_{3D} (\underline{\varepsilon}_{jkl} dr_1^j dr_2^k dr_3^\ell) \nabla_i \bar{t}^i = \int_{2D} (\underline{\varepsilon}_{ijk} dr_1^j dr_2^k) \bar{t}^i \quad (1.370)$$

$$\int_{2D} (\underline{\varepsilon}_{ilm} dr_1^\ell dr_2^m) (\bar{\varepsilon}^{ijk} \nabla_j T_k) = \int_{1D} dr_1^i T_i \quad (1.371)$$

$$\int_{1D} dr_1^i \partial_i T = \int_{0D} T, \quad (1.372)$$

where, in equation 1.370, \bar{t}^i stands for the vector dual to the tensor T_{ij} , i.e., $\bar{t}^i = \frac{1}{2!} \bar{\varepsilon}^{ijk} T_{jk}$.

Equations 1.367 and 1.370 correspond to the divergence theorem of Gauss-Ostrogradsky, equations 1.368 and 1.369 correspond to the rotational theorem of Stokes (stricto sensu), and equation 1.372, when written in its more familiar form

$$\int_a^b dr^i \partial_i T = T(b) - T(a) \tag{1.373}$$

corresponds the fundamental theorem of integral calculus.

Chapter 2

Elements of Probability

As probability theory is essential to the formulation of the rules of physical inference —to be analyzed in subsequent chapters— we have to start by an introduction of the concept of probability. This chapter is, however, more than a simple review. I assume that the spaces we shall work with, have a natural definition of distance between points, and, therefore, a definition of volume. This allows the introduction of the notion of ‘volumetric probability’, as opposed to the more conventional ‘probability density’. The notion of conditional volumetric probability is carefully introduced (I disagree with usual definitions of conditional probability density), and finally, the whole concept of conditional probability is generalized into a more general notion: the product of probability distributions.

2.1 Volume

2.1.1 Notion of Volume

The axiomatic introduction of a ‘volume’ over an n -dimensional manifold is very similar to the introduction of a ‘probability’, and both can be reduced to the axiomatic introduction of a ‘measure’. For pedagogical reasons, I choose to separate the two notions, presenting the notion of volume as more fundamental than that of a probability, as the definition of a probability shall require the previous definition of the volume.

Of course, given an n -dimensional manifold \mathbf{X} , one may wish to associate to it different ‘measures’ of the volume of any region of it. But, in this text, we shall rather assume than, within a given context, there is one ‘natural’ definition of volume.

So it is assumed the to any region $\mathcal{A} \subset \mathbf{X}$ it is associated a real or imaginary¹ quantity $V(\mathcal{A})$, called the *volume* of \mathcal{A} , that satisfies

Postulate 2.1 for any region \mathcal{A} of the space, $V(\mathcal{A}) \geq 0$;

Postulate 2.2 if \mathcal{A}_1 and \mathcal{A}_2 are two disjoint regions of the space, then $V(\mathcal{A}_1 \cup \mathcal{A}_2) = V(\mathcal{A}_1) + V(\mathcal{A}_2)$.

We shall say that a *volume distribution* (or, for short, a ‘*volume*’) has been definer over \mathbf{X} . The volume of the whole space \mathbf{X} may be positive real, positive imaginary, it may be zero or it may be infinite.

2.1.2 Volume Element

Consider a region \mathcal{A} of an n -dimensional manifold \mathbf{X} , and an approximate subdivision of it into regions with individual volume ΔV_i (see illustration 2.1). Successively refining the subdivision, allows easily to relate the volume of the whole region to the volumes of the individual regions,

$$V(\mathcal{A}) = \lim_{\Delta V_i \rightarrow 0} \sum_i \Delta V_i \quad , \quad (2.1)$$

expression that we may take as an elementary definition for the integral

$$V(\mathcal{A}) = \int_{\mathcal{P} \in \mathcal{A}} dV(\mathcal{P}) \quad . \quad (2.2)$$

When some coordinates $\mathbf{x} = \{x^1, \dots, x^n\}$ are chosen over \mathbf{X} , we may rewrite this equation as

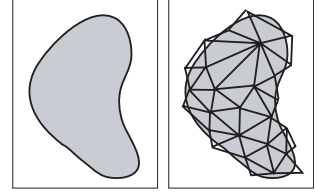
$$V(\mathcal{A}) = \int_{\mathbf{x} \in \mathcal{A}} dv(\mathbf{x}) \quad . \quad (2.3)$$

While $dV(\mathcal{P})$ stands for a function depending on the abstract notion of a ‘point’, $dv(\mathbf{x})$ stands for an ordinary function depending on some coordinates. A part this subtle difference, the two objects coincide: if by $\mathbf{x}(\mathcal{P})$ we designate the coordinates of the point \mathcal{P} , then,

$$dV(\mathcal{P}) = dv(\mathbf{x}(\mathcal{P})) \quad . \quad (2.4)$$

¹Some spaces having an ‘hyperbolic metric’, like the Minkowskian space-time of special relativity, have an imaginary volume. By convention, this volume is taken as imaginary positive.

Figure 2.1: The volume of an arbitrarily shaped, smooth, region of a space \mathbf{X} , can be defined as the limit of a sum, using elementary regions whose individual volume is known (for instance, triangles in this 2D illustration). This way of defining the volume of a region does not require the definition of a coordinate system over the space.



2.1.3 Volume Density and Capacity Element

Consider, at a given point \mathcal{P} of an n -dimensional manifold, n vectors (of the tangent linear space) $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$. These vectors may not have the same physical dimensions (for instance, \mathbf{v}_1 may represent a displacement, \mathbf{v}_2 a velocity, etc.). The *exterior product* of the n vectors, denoted $\mathbf{v}_1 \wedge \mathbf{v}_2 \wedge \dots \wedge \mathbf{v}_n$, is the scalar capacity

$$\mathbf{v}_1 \wedge \mathbf{v}_2 \wedge \dots \wedge \mathbf{v}_n = \epsilon_{i_1 i_2 \dots i_n} v_1^{i_1} v_2^{i_2} \dots v_n^{i_n} \quad , \quad (2.5)$$

where $\epsilon_{ij\dots}$ is the Levi-Civita capacity, defined in section 1.4.2. This is, of course, a totally antisymmetric expression. If some coordinates $\mathbf{x} = \{x^1, x^2, \dots, x^n\}$ have been defined over the manifold, then, at any given point we may consider the n infinitesimal vectors

$$d\mathbf{r}_1 = \begin{pmatrix} dx^1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} ; \quad d\mathbf{r}_2 = \begin{pmatrix} 0 \\ dx^2 \\ \vdots \\ 0 \end{pmatrix} ; \quad \dots ; \quad d\mathbf{r}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ dx^n \end{pmatrix} \quad (2.6)$$

corresponding to the respective perturbation of the n coordinates. The exterior product, at point \mathbf{x} , of these n vectors is called the *capacity element*, and is denoted $d\underline{v}(\mathbf{x})$:

$$\boxed{d\underline{v}(\mathbf{x}) = d\mathbf{r}_1 \wedge d\mathbf{r}_2 \wedge \dots \wedge d\mathbf{r}_n \quad .} \quad (2.7)$$

In view of expressions 2.6, and using a notational abuse, the capacity element so defined is usually written as

$$\boxed{d\underline{v}(\mathbf{x}) = dx^1 \wedge dx^2 \wedge \dots \wedge dx^n \quad .} \quad (2.8)$$

One of the major theorems of integration theory is that the volume element introduced in equation 2.3 is related to the capacity element $d\underline{v}$ through

$$\boxed{dv(\mathbf{x}) = \bar{g}(\mathbf{x}) d\underline{v}(\mathbf{x}) \quad ,} \quad (2.9)$$

where $\bar{g}(\mathbf{x})$ is the volume density in the coordinates \mathbf{x} , as defined in equation 1.32:

$$\bar{g}(\mathbf{x}) = \eta \sqrt{\det \mathbf{g}(\mathbf{x})} \quad . \quad (2.10)$$

Here η is the orientation of the coordinate system, as defined in section 1.4.1.

If the system of coordinates in use is positively oriented, the quantities $\bar{g}(\mathbf{x})$ and $d\underline{v}(\mathbf{x})$ are both positive. Alternatively, if the system of coordinates is negatively oriented, these two quantities are negative. The volume element $dv(\mathbf{x})$ is always a positive quantity.

The overbar in \bar{g} is to remember that the determinant of the metric tensor is a density, in the tensorial sense of section 1.2.2, while the underbar in \underline{dv} is to remember that the ‘capacity element’ is a capacity in the tensorial sense of the term. In equation 2.9, the product of a density times a capacity gives the volume element dv , that is an invariant scalar. In view of this equation, we can call $\bar{g}(\mathbf{x})$ the *volume density* in the coordinates $\mathbf{x} = \{x^1, \dots, x^n\}$. It is important to realize that $\bar{g}(\mathbf{x})$ does not represent any intrinsic property of the space, but, rather, a property of the coordinates being used.

Example 2.1 *In the Euclidean 3D space, using geographical coordinates² $\mathbf{x} = \{r, \varphi, \lambda\}$, it is well known that the volume element is*

$$dv(r, \varphi, \lambda) = r^2 \cos \lambda \, dr \wedge d\varphi \wedge d\lambda \quad , \quad (2.11)$$

so the volume density in to the geographical coordinates is

$$\bar{g}(r, \varphi, \lambda) = r^2 \cos \lambda \quad . \quad (2.12)$$

The metric in geographical coordinates is

$$ds^2 = dr^2 + r^2 \cos^2 \lambda \, d\varphi^2 + r^2 \, d\lambda^2 \quad , \quad (2.13)$$

so

$$\sqrt{\det \mathbf{g}} = r^2 \cos \lambda \quad . \quad (2.14)$$

Comparing this equation with equation 2.12 shows that one has

$$\bar{g} = \sqrt{\det \mathbf{g}} \quad . \quad (2.15)$$

as it should. **[End of example.]**

Figure 2.2: The geographical coordinates generalize better to n -dimensional spaces than the usual spherical coordinates. Note that the order of the angles, $\{\varphi, \lambda\}$, has to be the reverse of that of the angles $\{\theta, \varphi\}$, so as to define in both cases local referentials $dr \wedge d\theta \wedge d\varphi$ and $dr \wedge d\varphi \wedge d\lambda$ that have the same orientation as $dx \wedge dy \wedge dz$.

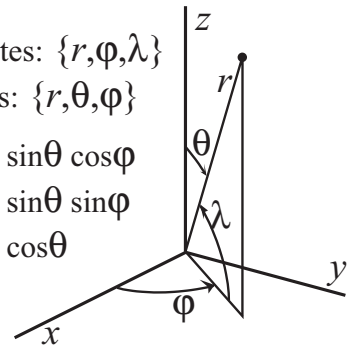
Geographical coordinates: $\{r, \varphi, \lambda\}$

Spherical coordinates: $\{r, \theta, \varphi\}$

$$x = r \cos \lambda \cos \varphi = r \sin \theta \cos \varphi$$

$$y = r \cos \lambda \sin \varphi = r \sin \theta \sin \varphi$$

$$z = r \sin \lambda = r \cos \theta$$



²The usual spherical coordinates are $\{r, \theta, \varphi\}$, and the domain of variation of θ is $0 \leq \theta \leq \pi$. These 3D coordinates do not generalize properly into ‘spherical’ coordinates in spaces of dimension larger than three. To these spherical coordinates one should prefer the ‘geographical coordinates’ $\{r, \varphi, \lambda\}$, where the domain of variation of λ is $-\pi/2 \leq \lambda \leq +\pi/2$. These are not ‘geographical coordinates’ in the normal sense used by geodesists, as r is here a radius (not the ‘height’ above some reference). See figure 2.2 for more details.

Example 2.2 In the 4D space-time of special relativity, with the Minkowskian coordinates $\{\tau_0, \tau_1, \tau_2, \tau_3\} = \{t, x/c, y/c, z/c\}$, the distance element ds satisfies

$$ds^2 = d\tau_0^2 - d\tau_1^2 - d\tau_2^2 - d\tau_3^2 \quad . \quad (2.16)$$

Then, the metric \mathbf{g} is diagonal, with the elements $\{+1, -1, -1, -1\}$ in the diagonal, and

$$\bar{g} = \sqrt{\det \mathbf{g}} = \sqrt{\det(-1)} = i \quad . \quad (2.17)$$

[End of example.]

Replacing 2.9, into equation 2.3 gives

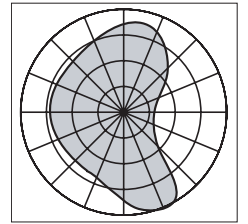
$$V(\mathcal{A}) = \int_{\mathbf{x} \in \mathcal{A}} d\underline{v}(\mathbf{x}) \bar{g}(\mathbf{x}) \quad . \quad (2.18)$$

Using expressions 2.8 and 1.32 we can write this in the more explicit (but not manifestly covariant) form

$$V(\mathcal{A}) = \eta \int_{\mathbf{x} \in \mathcal{A}} dx^1 \wedge \dots \wedge dx^n \sqrt{\det \mathbf{g}(\mathbf{x})} \quad . \quad (2.19)$$

These two (equivalent) expressions allow the usual interpretation of an integral as a limit involving the domains defined by constant increments of the coordinate values (see figure 2.3). Although such an expression is useful for analytic developments it is usually not well adapted to numerical evaluations (unless the coordinates are very specially chosen).

Figure 2.3: For the same shape of figure 2.1, the volume can be evaluated using, for instance, a polar coordinate system. In a numerical integration, regions near the origin may be oversampled, while regions far from the origin may be undersampled. In some situation, this problem may become crucial, so this sort of ‘coordinate integration’ is to be reserved to analytical developments only.



2.1.4 Change of Variables

2.1.4.1 Volume Element and Change of Variables

Consider an n -dimensional metric manifold with some coordinates \mathbf{x} . The defining property of the volume element, say $dv_{\mathbf{x}}(\mathbf{x})$, was (equation 2.3)

$$V(\mathcal{A}) = \int_{\mathbf{x} \in \mathcal{A}} dv_{\mathbf{x}}(\mathbf{x}) \quad . \quad (2.20)$$

Under a change of variables $\mathbf{x} \rightleftharpoons \mathbf{y}$, this expression shall become

$$V(\mathcal{A}) = \int_{\mathbf{y} \in \mathcal{A}} dv_{\mathbf{y}}(\mathbf{y}) \quad . \quad (2.21)$$

These two equations just correspond to a different labeling, respectively using the coordinates \mathbf{x} and the coordinates \mathbf{y} , of the fundamental equation 2.2 defining the volume element dV , so they are completely equivalent. In other words, the volume element is an invariant scalar, and one may write

$$\boxed{dv_{\mathbf{y}} = dv_{\mathbf{x}} \quad ,} \quad (2.22)$$

or, more explicitly,

$$dv_{\mathbf{y}}(\mathbf{y}) = dv_{\mathbf{x}}(\mathbf{x}(\mathbf{y})) \quad . \quad (2.23)$$

2.1.4.2 Volume Density, Capacity Element, and Change of Variables

In a change of variables $\mathbf{x} \rightleftharpoons \mathbf{y}$, the two capacity elements $dv_{\mathbf{x}}(\mathbf{x})$ and $dv_{\mathbf{y}}(\mathbf{y})$ are related via

$$\boxed{dv_{\mathbf{y}}(\mathbf{y}) = \frac{1}{X(\mathbf{y})} dv_{\mathbf{x}}(\mathbf{x}(\mathbf{y})) \quad ,} \quad (2.24)$$

where $X(\mathbf{y})$ is the Jacobian determinant $\det\{\partial x^i / \partial y^j\}$, as they are tensorial capacities, in the sense of section 1.2.2. Also, because a ‘volume density’ is a tensorial density, we have

$$\boxed{\bar{g}_{\mathbf{y}}(\mathbf{y}) = X(\mathbf{y}) \bar{g}_{\mathbf{x}}(\mathbf{x}(\mathbf{y})) \quad .} \quad (2.25)$$

Equation 2.18, that can be written, in the coordinates \mathbf{x} ,

$$V(\mathcal{A}) = \int_{\mathbf{x} \in \mathcal{A}} dv_{\mathbf{x}}(\mathbf{x}) \bar{g}_{\mathbf{x}}(\mathbf{x}) \quad , \quad (2.26)$$

$\bar{g}_{\mathbf{x}}(\mathbf{x})$ being the determinant of the metric matrix in the coordinates \mathbf{x} , becomes

$$V(\mathcal{A}) = \int_{\mathbf{y} \in \mathcal{A}} dv_{\mathbf{y}}(\mathbf{y}) \bar{g}_{\mathbf{y}}(\mathbf{y}) \quad , \quad (2.27)$$

$\bar{g}_{\mathbf{y}}(\mathbf{y})$ being the determinant of the metric matrix in the coordinates \mathbf{y} . Of course, the two capacity elements can be expressed as (equation 2.8)

$$dv_{\mathbf{x}}(\mathbf{x}) = dx^1 \wedge dx^2 \wedge \cdots \wedge dx^n \quad (2.28)$$

and

$$dv_{\mathbf{y}}(\mathbf{y}) = dy^1 \wedge dy^2 \wedge \cdots \wedge dy^n \quad . \quad (2.29)$$

If the two coordinate systems $\{x^1, \dots, x^n\}$ and $\{y^1, \dots, y^n\}$ have the same orientation, the two capacity elements $dv_{\mathbf{x}}(\mathbf{x})$ and $dv_{\mathbf{y}}(\mathbf{y})$ have the same sign. Otherwise, they have opposite sign.

2.1.5 Conditional Volume

Consider an n -dimensional manifold \mathbf{X}_n , with some coordinates $\mathbf{x} = \{x^1, \dots, x^n\}$, and a metric tensor $\mathbf{g}(\mathbf{x}) = \{g_{ij}(\mathbf{x})\}$. Consider also a p -dimensional submanifold \mathbf{X}_p of the n -dimensional manifold \mathbf{X}_n (with $p \leq n$). The n -dimensional volume over \mathbf{X}_n as characterized by the metric determinant $\sqrt{\det \mathbf{g}}$, induces a p -dimensional volume over the submanifold \mathbf{X}_p . Let us try to characterize it.

The simplest way to represent a p -dimensional submanifold \mathbf{X}_p of the n -dimensional manifold \mathbf{X}_n is by separating the n coordinates $\mathbf{x} = \{x^1, \dots, x^n\}$ of \mathbf{X}_n into one group of p coordinates $\mathbf{r} = \{r^1, \dots, r^p\}$ and one group of q coordinates $\mathbf{s} = \{s^1, \dots, s^q\}$, with

$$p + q = n \quad . \quad (2.30)$$

Using the notations

$$\mathbf{x} = \{x^1, \dots, x^n\} = \{r^1, \dots, r^p, s^1, \dots, s^q\} = \{\mathbf{r}, \mathbf{s}\} \quad , \quad (2.31)$$

the set of q relations

$$\begin{aligned} s_1 &= s_1(r_1, r_2, \dots, r_p) \\ s_2 &= s_2(r_1, r_2, \dots, r_p) \\ \dots &= \dots \\ s_q &= s_q(r_1, r_2, \dots, r_p) \quad , \end{aligned} \quad (2.32)$$

that, for short, may be written

$$\mathbf{s} = \mathbf{s}(\mathbf{r}) \quad , \quad (2.33)$$

define a p -dimensional submanifold \mathbf{X}_p in the $(p + q)$ -dimensional space \mathbf{X}_n . For later use, we can now introduce the matrix of partial derivatives

$$\mathbf{S} = \begin{pmatrix} S^1_1 & S^1_2 & \dots & S^1_p \\ S^2_1 & S^2_2 & \dots & S^2_p \\ \vdots & \vdots & \ddots & \vdots \\ S^q_1 & S^q_2 & \dots & S^q_p \end{pmatrix} = \begin{pmatrix} \frac{\partial s^1}{\partial r^1} & \frac{\partial s^1}{\partial r^2} & \dots & \frac{\partial s^1}{\partial r^p} \\ \frac{\partial s^2}{\partial r^1} & \frac{\partial s^2}{\partial r^2} & \dots & \frac{\partial s^2}{\partial r^p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial s^q}{\partial r^1} & \frac{\partial s^q}{\partial r^2} & \dots & \frac{\partial s^q}{\partial r^p} \end{pmatrix} \quad . \quad (2.34)$$

We can write $\mathbf{S}(\mathbf{r})$ for this matrix, as it is defined at a point $\{\mathbf{x}\} = \{\mathbf{r}, \mathbf{s}(\mathbf{r})\}$. Note also that the metric over \mathbf{X} can always be partitioned as

$$\mathbf{g}(\mathbf{x}) = \mathbf{g}(\mathbf{r}, \mathbf{s}) = \begin{pmatrix} \mathbf{g}_{rr}(\mathbf{r}, \mathbf{s}) & \mathbf{g}_{rs}(\mathbf{r}, \mathbf{s}) \\ \mathbf{g}_{sr}(\mathbf{r}, \mathbf{s}) & \mathbf{g}_{ss}(\mathbf{r}, \mathbf{s}) \end{pmatrix} \quad , \quad (2.35)$$

with $\mathbf{g}_{rs} = (\mathbf{g}_{sr})^T$.

In what follows, let us use the Greek indexes for the variables $\{r^1, \dots, r^p\}$, like in r^α ; $\alpha \in \{1, \dots, p\}$, and Latin indexes for the variables $\{s^1, \dots, s^q\}$, like in s^i ; $i \in \{1, \dots, q\}$. Consider an arbitrary point $\{\mathbf{r}, \mathbf{s}\}$ of the space \mathbf{X} . If the coordinates r^α are perturbed to $r^\alpha + dr^\alpha$, with the coordinates s^i kept unperturbed, one defines a p -dimensional subvolume of the n -dimensional manifold \mathbf{X}_n that can be written³ (middle panel in figure 2.4)

$$dv_p(\mathbf{r}, \mathbf{s}) = \sqrt{\det \mathbf{g}_{rr}(\mathbf{r}, \mathbf{s})} dr^1 \wedge \dots \wedge dr^p \quad . \quad (2.36)$$

³In all generality, we should write $dv_p(\mathbf{r}, \mathbf{s}) = \eta \sqrt{\det \mathbf{g}_{rr}(\mathbf{r}, \mathbf{s})} dr^1 \wedge \dots \wedge dr^p$, where η is ± 1 depending on the order of the coordinates $\{r^1, \dots, r^p\}$. Let us simplify the equations here but assuming that we have chosen the order of the coordinates so as to have a positively oriented capacity element $dr^1 \wedge \dots \wedge dr^p$.

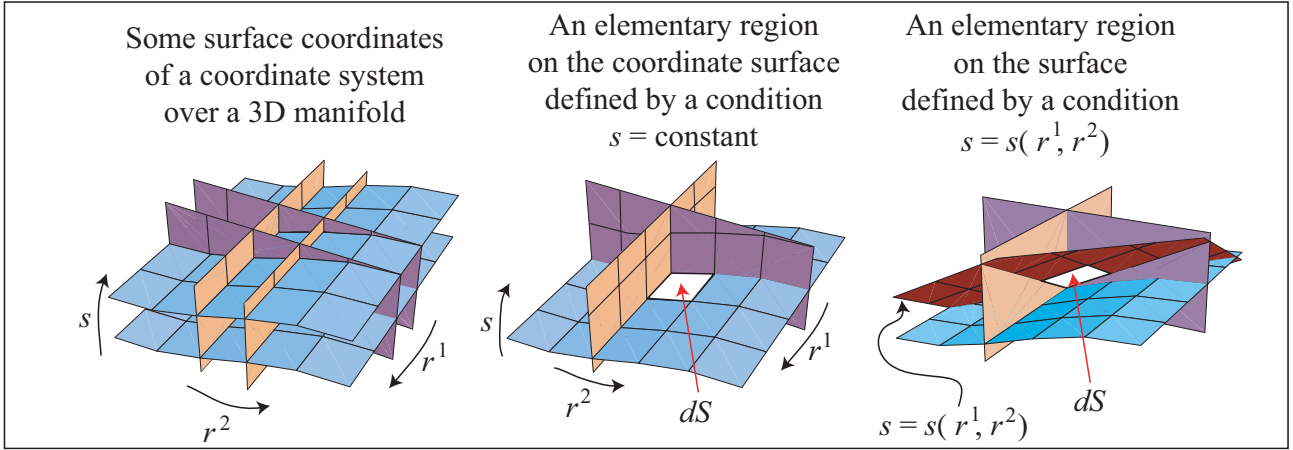


Figure 2.4: On a 3D space (3D manifold), a coordinate system $\{x^1, x^2, x^3\} = \{r^1, r^2, s\}$ is defined. Some characteristic surface coordinates are represented (left). In the middle, a surface element (2D volume element) on a coordinate surface $s = \text{const.}$ is represented, that corresponds to the expression in equation 2.36. In the right, a submanifold (surface) is defined by an equation $s = s(r^1, r^2)$. A surface element (2D volume element) is represented on the submanifold, that corresponds to the expression in equation 2.37.

Alternatively, consider a point (\mathbf{r}, \mathbf{s}) of \mathbf{X}_n that, in fact, is on the submanifold \mathbf{X}_p , i.e., a point that has coordinates of the form $(\mathbf{r}, \mathbf{s}(\mathbf{r}))$. It is clear that the variables $\{r^1 \dots r^p\}$ define a coordinate system over the submanifold, as it is enough to precise \mathbf{r} to define a point in \mathbf{X}_p . If the coordinates r^α are perturbed to $r^\alpha + dr^\alpha$, and the coordinates s^i are also perturbed to $s^i + ds^i$ in a way that one remains on the submanifold, (i.e., with $ds^i = S^i_\alpha dr^\alpha$), then, with the metric over \mathbf{X}_n partitioned as in equation 2.35, the general distance element $ds^2 = g_{ij} dx^i dx^j$ can be written $ds^2 = (\mathbf{g}_{rr})_{\alpha\beta} dr^\alpha dr^\beta + (\mathbf{g}_{rs})_{\alpha j} dr^\alpha ds^j + (\mathbf{g}_{sr})_{i\beta} ds^i dr^\beta + (\mathbf{g}_{ss})_{ij} ds^i ds^j$, and replacing ds^i by $ds^i = S^i_\alpha dr^\alpha$, we obtain $ds^2 = G_{\alpha\beta} dr^\alpha dr^\beta$, with $\mathbf{G} = \mathbf{g}_{rr} + \mathbf{g}_{rs} \mathbf{S} + \mathbf{S}^T \mathbf{g}_{sr} + \mathbf{S}^T \mathbf{g}_{ss} \mathbf{S}$. The ds^2 just expressed gives the distance between two any points of \mathbf{X}_p , i.e., \mathbf{G} is the metric matrix of the submanifold associated to the coordinates \mathbf{r} . The p -dimensional volume element on the manifold is, then, $dv_r = \sqrt{\det \mathbf{G}} dr^1 \wedge \dots \wedge dr^p$, i.e.,

$$dv_p(\mathbf{r}) = \sqrt{\det (\mathbf{g}_{rr} + \mathbf{g}_{rs} \mathbf{S} + \mathbf{S}^T \mathbf{g}_{sr} + \mathbf{S}^T \mathbf{g}_{ss} \mathbf{S})} dr^1 \wedge \dots \wedge dr^p \quad , \quad (2.37)$$

where $\mathbf{S} = \mathbf{S}(\mathbf{r})$, $\mathbf{g}_{rr} = \mathbf{g}_{rr}(\mathbf{r}, \mathbf{s}(\mathbf{r}))$, $\mathbf{g}_{rs} = \mathbf{g}_{rs}(\mathbf{r}, \mathbf{s}(\mathbf{r}))$, $\mathbf{g}_{sr} = \mathbf{g}_{sr}(\mathbf{r}, \mathbf{s}(\mathbf{r}))$ and $\mathbf{g}_{ss} = \mathbf{g}_{ss}(\mathbf{r}, \mathbf{s}(\mathbf{r}))$. Figure 2.4 illustrates this result. The expression 2.37 says that the p -dimensional volume density induced over the submanifold \mathbf{X}_p is

$$\bar{g}(\mathbf{x}) = \eta \sqrt{\det (\mathbf{g}_{rr} + \mathbf{g}_{rs} \mathbf{S} + \mathbf{S}^T \mathbf{g}_{sr} + \mathbf{S}^T \mathbf{g}_{ss} \mathbf{S})} \quad . \quad (2.38)$$

Note: say here that in the case the space \mathbf{X}_n is formed as the cartesian product of two spaces, $\mathbf{R}_p \times \mathbf{S}_q$, with the metric over \mathbf{X}_n induced from the metric \mathbf{g}_r over \mathbf{R}_p and the metric \mathbf{g}_s over \mathbf{S}_q by

$$ds_{\mathbf{x}}^2 = ds_{\mathbf{r}}^2 + ds_{\mathbf{s}}^2 \quad , \quad (2.39)$$

then, the expression of the metric 2.35 simplifies into

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} \mathbf{g}_r(\mathbf{r}) & \mathbf{0} \\ \mathbf{0} & \mathbf{g}_s(\mathbf{s}) \end{pmatrix}, \quad (2.40)$$

and equations 2.37–2.38 simplify into

$$dv_p(\mathbf{r}) = \sqrt{\det(\mathbf{g}_r + \mathbf{S}^T \mathbf{g}_s \mathbf{S})} dr^1 \wedge \dots \wedge dr^p \quad (2.41)$$

and

$$\bar{g}(\mathbf{x}) = \eta \sqrt{\det(\mathbf{g}_r + \mathbf{S}^T \mathbf{g}_s \mathbf{S})}. \quad (2.42)$$

2.2 Probability

2.2.1 Notion of Probability

Consider an n -dimensional metric manifold, over which a ‘volume distribution’ has been defined (satisfying the axioms in section 2.1.1), associating to any region (i.e., subset) \mathcal{A} of \mathbb{X} its volume

$$\mathcal{A} \mapsto V(\mathcal{A}) \quad . \quad (2.43)$$

A particular volume distribution having been introduced over \mathbb{X} , once for all, different ‘probability distributions’ may be considered, that we are about to characterize axiomatically.

We shall say that a *probability distribution* (or, for short, a *probability*) has been defined over \mathbb{X} if to any region $\mathcal{A} \subset \mathbb{X}$ we can associate an adimensional real number,

$$\mathcal{A} \mapsto P(\mathcal{A}) \quad (2.44)$$

called the *probability* of \mathcal{A} , that satisfies

Postulate 2.3 *for any region \mathcal{A} of the space,*

$$P(\mathcal{A}) \geq 0 \quad ; \quad (2.45)$$

Postulate 2.4 *for disjoint regions of the space, the probabilities are additive:*

$$\mathcal{A}_1 \cap \mathcal{A}_2 = \emptyset \quad \Rightarrow \quad P(\mathcal{A}_1 \cup \mathcal{A}_2) = P(\mathcal{A}_1) + P(\mathcal{A}_2) \quad ; \quad (2.46)$$

Postulate 2.5 *the probability distribution must be absolutely continuous with respect to the volume distribution, i.e., the probability $P(\mathcal{A})$ of any region $\mathcal{A} \subset \mathbb{X}$ with vanishing volume must be zero:*

$$V(\mathcal{A}) = 0 \quad \Rightarrow \quad P(\mathcal{A}) = 0 \quad . \quad (2.47)$$

The probability of the whole space \mathbb{X} may be zero, it may be finite, or it may be infinite. The first two axioms are due to Kolmogorov (1933). In common texts, there is usually an axiom concerning the behaviour of a probability when we consider an infinite collection⁴ of sets, $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3 \dots$, but this is a technical issue that I choose to ignore. Our third axiom here is not usually introduced, as the distinction between the ‘volume distribution’ and a ‘probability distribution’ is generally not made: both are just considered as examples of ‘measure distributions’. This distinction shall, in fact, play a major role in the theory that follows.

When the probability of the whole space is finite, a probability distribution can be renormalized, so as to have $P(\mathbb{X}) = 1$. We shall then say that we face an ‘absolute probability’. If a probability distribution is not normalizable, we shall say that we have a ‘relative probability’: in that case, what usually matters is not the probability $P(\mathcal{A})$ of a region $\mathcal{A} \in \mathbb{X}$, but the relative between probability two regions \mathcal{A} and \mathcal{B} , denoted $P(\mathcal{A}; \mathcal{B})$, and defined as

$$P(\mathcal{A}; \mathcal{B}) = \frac{P(\mathcal{A})}{P(\mathcal{B})} \quad . \quad (2.48)$$

⁴Presentations of measure theory that pretend to mathematical rigor, assume ‘finite additivity’ or, alternatively, ‘countable additivity’. See, for instance, the interesting discussion in Jaynes (1995).

2.2.2 Volumetric Probability

We have just defined a probability distribution over an n -dimensional manifold, that is absolutely continuous with respect to the volume distribution over the manifold. Then, by virtue of the Radon-Nikodym theorem (e.g., Taylor, 1966), one can define over \mathbf{X} a *volumetric probability* $f(\mathcal{P})$ such that the probability of any region \mathcal{A} of the space can be obtained as

$$P(\mathcal{A}) = \int_{\mathcal{P} \in \mathcal{A}} dV(\mathcal{P}) f(\mathcal{P}) \quad . \quad (2.49)$$

Note that this equation makes sense even if no particular coordinate system is defined over the manifold \mathbf{X} , as the integral here can be understood in the sense suggested in figure 2.1. If a coordinate system $\mathbf{x} = \{x^1, \dots, x^n\}$ is defined over \mathbf{X} , we may well wish to write equation 2.49 as

$$P(\mathcal{A}) = \int_{\mathbf{x} \in \mathcal{A}} dv_{\mathbf{x}}(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) \quad , \quad (2.50)$$

where, now, $dv_{\mathbf{x}}(\mathbf{x})$ is to be understood as the special expression of the volume element in the coordinates \mathbf{x} . One may be interested in using the volume element $dv_{\mathbf{x}}(\mathbf{x})$ directly for the integration (as suggested in figure 2.1). Alternatively, one may wish to use the coordinate lines for the integration (as suggested in figure 2.3). In this case, one writes (equation 2.9)

$$dv_{\mathbf{x}}(\mathbf{x}) = \bar{g}_{\mathbf{x}}(\mathbf{x}) d\underline{v}_{\mathbf{x}}(\mathbf{x}) \quad , \quad (2.51)$$

to get

$$P(\mathcal{A}) = \int_{\mathbf{x} \in \mathcal{A}} d\underline{v}_{\mathbf{x}}(\mathbf{x}) \bar{g}_{\mathbf{x}}(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) \quad . \quad (2.52)$$

Using $d\underline{v}_{\mathbf{x}}(\mathbf{x}) = dx^1 \wedge \dots \wedge dx^n$ (equation 2.8) and $\bar{g}_{\mathbf{x}}(\mathbf{x}) = \sqrt{\det \mathbf{g}(\mathbf{x})}$ (equation 1.32), this expression can be written in the more explicit (but not manifestly covariant) form

$$P(\mathcal{A}) = \eta \int_{\mathbf{x} \in \mathcal{A}} dx^1 \wedge \dots \wedge dx^n \sqrt{\det \mathbf{g}(\mathbf{x})} f_{\mathbf{x}}(\mathbf{x}) \quad , \quad (2.53)$$

where η is +1 if the system of coordinates is positively oriented and -1 if it is negatively oriented. These two (equivalent) expressions may be useful for analytical developments, but not for numerical evaluations, where one should choose a direct handling of expression 2.50.

2.2.3 Probability Density

In equation 2.52 we can introduce the definition

$$\bar{f}_{\mathbf{x}}(\mathbf{x}) = \bar{g}_{\mathbf{x}}(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) \quad , \quad (2.54)$$

to obtain

$$P(\mathcal{A}) = \int_{\mathbf{x} \in \mathcal{A}} d\underline{v}_{\mathbf{x}}(\mathbf{x}) \bar{f}_{\mathbf{x}}(\mathbf{x}) \quad , \quad (2.55)$$

where

$$d\underline{v}_{\mathbf{x}}(\mathbf{x}) = dx^1 \wedge \cdots \wedge dx^n \quad . \quad (2.56)$$

The function $\bar{f}_{\mathbf{x}}(\mathbf{x})$ is called the *probability density* (associated to the probability distribution P). It is a density, in the tensorial sense of the term, i.e., under a change of variables $\mathbf{x} \rightleftharpoons \mathbf{y}$ it change according to the Jacobian rule (see section 2.2.5.2).

Having defined a volumetric probability $f_{\mathbf{x}}(\mathbf{x})$ in section 2.2.2, why should one care at all about the probability density $\bar{f}_{\mathbf{x}}(\mathbf{x})$?

One possible advantage of a probability density over a volumetric probability appears when comparing equation 2.50 to equation 2.55. To integrate a volumetric probability one must have defined a volume element over the space, while to integrate a probability density, one only needs to have defined coordinates, irrespectively of any metric meaning they may have. This, is, of course, why usual expositions of the theory use probability densities.

In fact, I see this as a handicap. When probability theory is developed without the notion of volume and of distance, one is forced to include definitions that do not have the necessary invariances, the most striking example being the usual definition of ‘conditional probability density’. One does not obtain a correct definition unless a metric in the space is introduced (see section 2.4). The well-known ‘Borel paradox’ (see appendix 2.8.10) is the simplest example of this annoying situation. If I mention at all the notion of probability density is to allow the reader to make the connection between the formulas to be developed in this book and the formulas she/he may find elsewhere.

As we have chosen in this text to give signs to densities and capacities that are associated to the orientation of the coordinate system, it is clear from definition 2.54 that, contrary to a volumetric probability, a probability density is not necessarily positive: it has the sign of the capacity element, i.e., a positive sign in positively oriented coordinate systems, and a negative sign in negatively oriented coordinate systems.

Example 2.3 Consider a homogeneous probability distribution at the surface of a sphere of radius r . When parameterizing a point by its geographical coordinates (φ, λ) , the associated (2D) volumetric probability is

$$f(\varphi, \lambda) = \frac{1}{4\pi r^2} \quad . \quad (2.57)$$

The probability of a region \mathcal{A} of the surface is computed as

$$P(\mathcal{A}) = \int \int_{\{\varphi, \lambda\} \in \mathcal{A}} dS(\varphi, \lambda) f(\varphi, \lambda) \quad , \quad (2.58)$$

where $dS(\varphi, \lambda) = r^2 \cos \lambda d\varphi d\lambda$, and the total probability equals one. Alternatively, the probability density associated to the homogeneous probability distribution over the sphere is

$$\bar{f}(\varphi, \lambda) = \frac{1}{4\pi} \cos \lambda \quad . \quad (2.59)$$

The probability of a region \mathcal{A} of the surface is computed as

$$P(\mathcal{A}) = \int \int_{\{\varphi, \lambda\} \in \mathcal{A}} d\varphi d\lambda f(\varphi, \lambda) \quad , \quad (2.60)$$

and the probability of the whole surface also equals one. **[End of example.]**

2.2.4 Volumetric Histograms and Density Histograms

Note: explain here what is a volumetric histogram and a density histogram. Say that while the limit of a volumetric histogram is a volumetric probability, the limit of a density histogram is a probability density.

Introduce the notion of ‘naïve histogram’.

Consider a problem where we have two physical properties to analyze. The first is the property of electric resistance-conductance of a metallic wire, as it can be characterized, for instance, by its resistance R or by its conductance $C = 1/R$. The second is the ‘cold-warm’ property of the wire, as it can be characterized by its temperature T or its thermodynamic parameter $\beta = 1/kT$ (k being the Boltzmann constant). The ‘parameter space’ is, here, two-dimensional. In the ‘resistance-conductance’ space, the distance between two points, characterized by the resistances R_1 and R_2 , or by the conductances C_1 and C_2 is, as explained in section XXX,

$$D = \left| \log \frac{R_2}{R_1} \right| = \left| \log \frac{C_2}{C_1} \right| . \quad (2.61)$$

Similarly, in the ‘cold-warm’ space, the distance between two points, characterized by the temperatures T_1 and T_2 , or by the thermodynamic parameters β_1 and β_2 is

$$D = \left| \log \frac{T_2}{T_1} \right| = \left| \log \frac{\beta_2}{\beta_1} \right| . \quad (2.62)$$

An homogeneous probability distribution can be defined as ...

Bla, bla, bla ...

In figure 2.5, the two histograms that can be made from the two first diagrams give the volumetric probability. The naïve histogram that could be made from the diagram at the right would give a probability density.

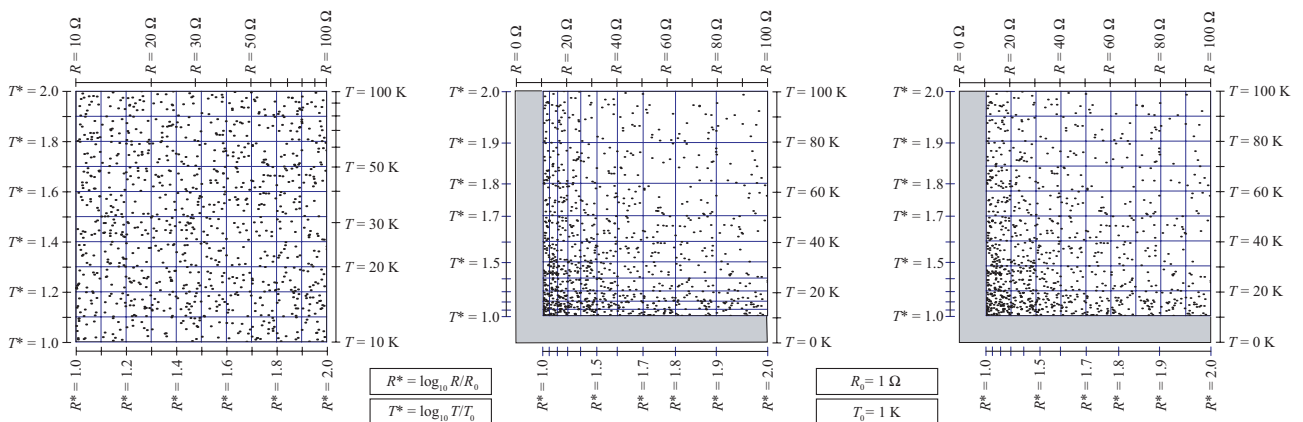


Figure 2.5: Note: explain here how to make a volumetric histogram. Explain that when the electric resistance or the temperature span orders of magnitude, the diagram at the right becomes totally impractical.

2.2.5 Change of Variables

2.2.5.1 Volumetric Probability and Change of Variables

In a change of coordinates $\mathbf{x} \mapsto \mathbf{y}(\mathbf{x})$, the expression 2.50

$$P(\mathcal{A}) = \int_{\mathbf{x} \in \mathcal{A}} dv_{\mathbf{x}}(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) \quad (2.63)$$

becomes

$$P(\mathcal{A}) = \int_{\mathbf{y} \in \mathcal{A}} dv_{\mathbf{y}}(\mathbf{y}) f_{\mathbf{y}}(\mathbf{y}) \quad (2.64)$$

where $dv_{\mathbf{y}}(\mathbf{y})$ and $f_{\mathbf{y}}(\mathbf{y})$ are respectively the expressions of the volume element and of the volumetric probability in the coordinates \mathbf{y} . These are actual invariants (in the tensorial sense), so, when comparing this equation (written in the coordinates \mathbf{y}) to equation 2.50 (written in the coordinates \mathbf{x}), one simply has, at every point,

$$\boxed{\begin{array}{l} f_{\mathbf{y}} = f_{\mathbf{x}} \\ dv_{\mathbf{y}} = dv_{\mathbf{x}} \end{array}}, \quad (2.65)$$

or, to be more explicit,

$$f_{\mathbf{y}}(\mathbf{y}) = f_{\mathbf{x}}(\mathbf{x}(\mathbf{y})) \quad ; \quad dv_{\mathbf{y}}(\mathbf{y}) = dv_{\mathbf{x}}(\mathbf{x}(\mathbf{y})) \quad . \quad (2.66)$$

That under a change of variables $\mathbf{x} \rightleftharpoons \mathbf{y}$ one has $f_{\mathbf{y}} = f_{\mathbf{x}}$ for volumetric probabilities, is an important property. It contrasts with the property found in usual texts (where the Jacobian of the transformation appears): remember that we are considering here volumetric probabilities, not the usual probability densities.

A volumetric probability can also be integrated using the expression 2.53

$$P(\mathcal{A}) = \eta_{\mathbf{x}} \int_{\mathbf{x} \in \mathcal{A}} dx^1 \wedge \cdots \wedge dx^n \sqrt{\det \mathbf{g}_{\mathbf{x}}(\mathbf{x})} f_{\mathbf{x}}(\mathbf{x}) \quad , \quad (2.67)$$

that, under the change of variables becomes

$$P(\mathcal{A}) = \eta_{\mathbf{y}} \int_{\mathbf{y} \in \mathcal{A}} dy^1 \wedge \cdots \wedge dy^n \sqrt{\det \mathbf{g}_{\mathbf{y}}(\mathbf{y})} f_{\mathbf{y}}(\mathbf{y}) \quad . \quad (2.68)$$

These equations contain each a capacity element and a volume density, that change, under the change of variables, following the rules given in section 1.2.2, but we do not need to be concerned with this here, as the meaning of $dy^1 \wedge \cdots \wedge dy^n$ is clear, and one usually obtains $\eta_{\mathbf{y}} \sqrt{\det \mathbf{g}_{\mathbf{y}}(\mathbf{y})}$ by an explicit computation of the determinant in the coordinates \mathbf{y} , rather than by multiplying the volume density $\eta_{\mathbf{x}} \sqrt{\det \mathbf{g}_{\mathbf{x}}(\mathbf{x}(\mathbf{y}))}$ by the Jacobian determinant $X(\mathbf{y})$ (see section 2.25).

[Note: Important: I have to erect as a basic principle to use, in a change of variables, the representation exemplified by figures 9.5, 9.8 and 9.9.]

2.2.5.2 Probability Density and Change of Variables

A probability density being defined as the product of an invariant times a density (equation 2.54) it is a density in the tensorial sense of the term. Under a change of variables $\mathbf{x} \rightleftharpoons \mathbf{y}$, expression 2.55

$$P(\mathcal{A}) = \int_{\mathbf{x} \in \mathcal{A}} d\underline{v}_{\mathbf{x}}(\mathbf{x}) \bar{f}_{\mathbf{x}}(\mathbf{x}) \quad , \quad (2.69)$$

where $d\underline{v}_{\mathbf{x}}(\mathbf{x}) = dx^1 \wedge \cdots \wedge dx^n$, becomes

$$P(\mathcal{A}) = \int_{\mathbf{y} \in \mathcal{A}} d\underline{v}_{\mathbf{y}}(\mathbf{y}) \bar{f}_{\mathbf{y}}(\mathbf{y}) \quad , \quad (2.70)$$

where $d\underline{v}_{\mathbf{y}}(\mathbf{y}) = dy^1 \wedge \cdots \wedge dy^n$. The two capacity elements $d\underline{v}_{\mathbf{x}}(\mathbf{x})$ and $d\underline{v}_{\mathbf{y}}(\mathbf{y})$ are related through the relation 2.24, and, more importantly, the two probability densities are related as tensorial densities should (see section 1.2.2),

$$\boxed{\bar{f}_{\mathbf{y}}(\mathbf{y}) = X(\mathbf{y}) \bar{f}_{\mathbf{x}}(\mathbf{x}(\mathbf{y})) \quad .} \quad (2.71)$$

This is called the *Jacobian rule* for the change of a probability density under a change of ‘variables’ (i.e., under a change of coordinates over the considered manifold). Note that the X appearing in this equation is the determinant of the matrix $\{X^i_j\} = \{\partial x^i / \partial y^j\}$, not that of the matrix $\{Y^i_j\} = \{\partial y^i / \partial x^j\}$.

Many authors take the absolute value of the Jacobian in this equation, which is not quite correct: it is the actual Jacobian that appears. The absolute value of the Jacobian is taken by these authors to force probability densities to always be positive, but this denies to probability densities the right to be densities, in the full tensorial sense of the term (see section 1.2.2).

In this text, I try to avoid the use of probability densities, and only mention them in the appendixes.

2.3 Sum and Product of Probabilities

Let \mathbf{X} be an n -dimensional metric manifold, with a volume distribution V , and let P and Q be two normalized probability distributions over \mathbf{X} . In what follows we shall deduce, from P and Q , two new probability distributions over \mathbf{X} , their sum, denoted $P \cup Q$ and their product, denoted $P \cap Q$.

2.3.1 Sum of Probabilities

P and Q being two probability distributions over \mathbf{X} , their *sum* (or *union*), denoted $P \cup Q$ is defined by the conditions

Postulate 2.6 for any $\mathcal{A} \subset \mathbf{X}$,

$$(P \cup Q)(\mathcal{A}) = (Q \cup P)(\mathcal{A}) \quad ; \quad (2.72)$$

Postulate 2.7 for any $\mathcal{A} \subset \mathbf{X}$,

$$P(\mathcal{A}) = 0 \quad \text{AND} \quad Q(\mathcal{A}) = 0 \quad \implies \quad (P \cup Q)(\mathcal{A}) = 0 \quad ; \quad (2.73)$$

Postulate 2.8 if there is some $\mathcal{A} \subset \mathbf{X}$ for which $P(\mathcal{A}) = 0$, then, necessarily, for any probability Q ,

$$(P \cup Q)(\mathcal{A}) = Q(\mathcal{A}) \quad . \quad (2.74)$$

Note: I have to explain here that these postulates do not characterize uniquely the sum operation. The solution I choose is the following one.

Property 2.1 If the probability distribution P is characterized by the volumetric probability $f(\mathcal{P})$, and the probability distribution Q is characterized by the volumetric probability $g(\mathcal{P})$, then, the probability distribution $P \cup Q$ is characterized by the volumetric probability, denoted $(f + g)(\mathcal{P})$ given by

$$(f + g)(\mathcal{P}) = \frac{\alpha f(\mathcal{P}) + \beta g(\mathcal{P})}{\alpha + \beta} \quad , \quad (2.75)$$

where α and β are two arbitrary constants.

Note: An alternative solution would be what is used in fuzzy set theory to define the union of fuzzy sets. Translated to the language of volumetric probabilities, and slightly generalized, this would correspond to

$$(f + g)(\mathcal{P}) = k \max(\alpha f(\mathcal{P}), \beta g(\mathcal{P})) \quad , \quad (2.76)$$

where α and β are two arbitrary constants, and k a normalizing one.

Let me try to give an interpretation of this sum of probabilities. If an experimenter faces realizations of a random process and wants to investigate the probability distribution governing the process, she/he may start making histograms of the realizations. As an example, for realizations of a probability distribution over a continuous space, the experimenter will obtain

histograms that, in some sense, will approach the volumetric probability corresponding to the probability distribution.

A histogram is typically made by dividing the working space into cells, by counting how many realizations fall inside each cell and by dividing the count by the cell volume. A more subtle approach is possible. First, we have to understand that, in the physical sciences, when “a random point materializes in an abstract space” we have to measure its coordinates. As any physical measure of a real quantity will have attached uncertainties, mathematically speaking, the measurement will not produce a ‘point’, but a state of information over the space, i.e., a volumetric probability. If we have measured the coordinates of many points, the results of each measurement will be described by a volumetric probability $f_i(\mathbf{x})$. The ‘sum’ of all these, i.e., the volumetric probability

$$(f_1 + f_2 + \dots)(\mathbf{x}) = \sum_i f_i(\mathbf{x}) \quad (2.77)$$

is a finer estimation of the background volumetric probability than an ordinary histogram, as actual measurement uncertainties are used, irrespective of any division of the space into cells.

2.3.2 Product of Probabilities

P and Q being two probability distributions over \mathbf{X} , their *product* (or *intersection*), denoted $P \cap Q$ is defined by the conditions

Postulate 2.9 for any $\mathcal{A} \subset \mathbf{X}$,

$$(P \cap Q)(\mathcal{A}) = (Q \cap P)(\mathcal{A}) \quad ; \quad (2.78)$$

Postulate 2.10 for any $\mathcal{A} \subset \mathbf{X}$,

$$P(\mathcal{A}) = 0 \quad \text{OR} \quad Q(\mathcal{A}) = 0 \quad \implies \quad (P \cap Q)(\mathcal{A}) = 0 \quad . \quad (2.79)$$

Postulate 2.11 if for whatever $\mathcal{B} \subset \mathbf{X}$, one has $P(\mathcal{B}) = k V(\mathcal{B})$, then, necessarily, for any $\mathcal{A} \subset \mathbf{X}$ and for any probability Q ,

$$(P \cap Q)(\mathcal{A}) = (Q \cap P)(\mathcal{A}) = Q(\mathcal{A}) \quad . \quad (2.80)$$

(The homogeneous probability distribution is the neutral element of the product operation).

Note: I have to explain here that these postulates do not characterize uniquely the product operation. The solution I choose is the following one.

Property 2.2 If the probability distribution P is characterized by the volumetric probability $f(\mathcal{P})$, and the probability distribution Q is characterized by the volumetric probability $g(\mathcal{P})$, then, the probability distribution $P \cup Q$ is characterized by the volumetric probability, denoted $(f \cdot g)(\mathcal{P})$ given by

$$(f \cdot g)(\mathcal{P}) = \frac{f(\mathcal{P}) g(\mathcal{P})}{\int_{\mathcal{P} \in \mathbf{X}} dV(\mathcal{P}) f(\mathcal{P}) g(\mathcal{P})} \quad . \quad (2.81)$$

More generally, the ‘product’ of the volumetric probabilities $f_1(\mathcal{P}), f_2(\mathcal{P}) \dots$ is

$$(f_1 \cdot f_2 \cdot f_3 \dots)(\mathcal{P}) = \frac{f_1(\mathcal{P}) f_2(\mathcal{P}) f_3(\mathcal{P}) \dots}{\int_{\mathcal{P} \in \mathbf{X}} dV(\mathcal{P}) f_1(\mathcal{P}) f_2(\mathcal{P}) f_3(\mathcal{P}) \dots} \quad (2.82)$$

Note: An alternative solution would be what is used in fuzzy set theory to define the intersection of fuzzy sets. Translated to the language of volumetric probabilities, and slightly generalized, this would correspond to

$$(f \cdot g)(\mathcal{P}) = k \min(\alpha f(\mathcal{P}), \beta g(\mathcal{P})) \quad , \quad (2.83)$$

where α and β are two arbitrary constants, and k a normalizing one.

It is easy to write some extra conditions that distinguish the solution to the axioms given by equation 2.75 end equation 2.81 and that given by equations 2.76 and 2.83. For instance, as volumetric probabilities 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 volumetric probabilities $f(\mathcal{P}), g(\mathcal{P}) \dots$ by constants $\lambda, \mu \dots$:

$$[(\lambda + \mu)f](\mathcal{P}) = (\lambda f + \mu f)(\mathcal{P}) \quad ; \quad [\lambda(f \cdot g)](\mathcal{P}) = (\lambda f \cdot g)(\mathcal{P}) = (f \cdot \lambda g)(\mathcal{P}) \quad . \quad (2.84)$$

One important property of the two operations ‘sum’ and ‘product’ 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, we obtain equivalent results in any coordinate system.

[Note: Say somewhere that the set of 11 postulates 2.1–2.11, defining the volume and a set of probability distributions furnished with two operations, define an inference space.]

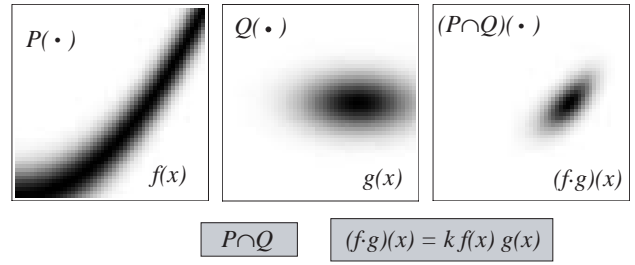
The interpretation of this product of volumetric probabilities, can be obtained by comparing figures 2.7 and 2.6. In figure 2.7, a probability distribution $P(\cdot)$ is represented by the volumetric probability associated to it. To any region \mathcal{A} of the plane, it associates the probability $P(\mathcal{A})$. 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 $P(\cdot | \mathcal{B}) = P(\cdot \cap \mathcal{B}) / P(\mathcal{B})$. This (classical) definition means that $P(\cdot | \mathcal{B})$ equals $P(\cdot)$ inside \mathcal{B} and is zero outside, as suggested in the center of the figure (the division by $P(\mathcal{B})$ just corresponds to a renormalization). If the probability $\mathcal{A} \rightarrow P(\mathcal{A})$ is represented by a volumetric probability $f(\mathcal{P})$, the probability $\mathcal{A} \rightarrow P(\mathcal{A} | \mathcal{B})$ is represented by the volumetric probability $f(\mathcal{P} | \mathcal{B})$ given by

$$f(\mathcal{P} | \mathcal{B}) = k f(\mathcal{P}) H(\mathcal{P}) = \frac{f(\mathcal{P}) H(\mathcal{P})}{\int_{\mathbf{X}} dV(\mathcal{P}) f(\mathcal{P}) H(\mathcal{P})} \quad , \quad (2.85)$$

where $H(\mathcal{P})$ takes a constant value inside \mathcal{B} , and vanishes outside. We see that $f(\mathcal{P} | \mathcal{B})$ is proportional to $f(\mathcal{P})$ inside \mathcal{B} and is zero outside \mathcal{B} .

While the elements entering the definition of a conditional probability are a probability distribution P and a subset $\mathcal{B} \subset \mathbf{X}$, we here consider two probability distributions P and Q , with volumetric probabilities $f(\mathcal{P})$ and $g(\mathcal{P})$. It is clear that equation 2.81 is a generalization of equation 2.85, as the set \mathcal{B} is now replaced a a probability distribution Q (see figure 2.6). In the special case where the probability Q is zero everywhere excepted inside a domain \mathcal{B} , where it is homogeneous, then, we recover the standard notion of conditional probability.

Figure 2.6: Illustration of the definition of the product of two probability distribution, interpreted here as a generalization of the notion of conditional probability (see figure 2.7). While a conditional probability combines a probability distribution $P(\cdot)$ with an ‘event’ \mathcal{B} , the *product* operation combines two probability distributions $P(\cdot)$ and $Q(\cdot)$ defined over the same space.



Example 2.4 Let \mathbf{S} represent the surface of the Earth, using geographical coordinates (longitude φ and latitude λ). 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 (2D) volumetric probability $f(\varphi, \lambda)$, and an independent, simultaneous estimation of the position by another airplane navigator gives a probability distribution corresponding to the volumetric probability $g(\varphi, \lambda)$. How the two volumetric probabilities $f(\varphi, \lambda)$ and $g(\varphi, \lambda)$ should be ‘combined’ to obtain a ‘resulting’ volumetric probability? The answer is given by the ‘product’ of the two volumetric probabilities densities:

$$(f \cdot g)(\varphi, \lambda) = \frac{f(\varphi, \lambda) g(\varphi, \lambda)}{\int_{\mathbf{S}} dS(\varphi, \lambda) f(\varphi, \lambda) g(\varphi, \lambda)} . \quad (2.86)$$

[End of example.]

2.4 Conditional Probability

2.4.1 Notion of Conditional Probability

Let $P(\cdot)$ represent a probability distribution over an n -dimensional manifold \mathbf{X}_n , i.e., a function $\mathcal{A} \rightarrow P(\mathcal{A})$ satisfying the Kolmogorov axioms. Letting now \mathcal{B} be a ‘fixed’ region of \mathbf{X}_n , we can define another probability distribution, say $P_B(\cdot)$, that to any region \mathcal{A} associates the probability $P_B(\mathcal{A})$ defined by $P_B(\mathcal{A}) = P(\mathcal{A} \cap \mathcal{B})/P(\mathcal{B})$. It can be shown that this, indeed, is a probability (i.e., satisfies the Kolmogorov axioms). Instead of the notation $P_B(\mathcal{A})$, it is customary to use the notation $P(\mathcal{A}|\mathcal{B}) = P(\mathcal{A} \cap \mathcal{B})/P(\mathcal{B})$ and the definition then reads

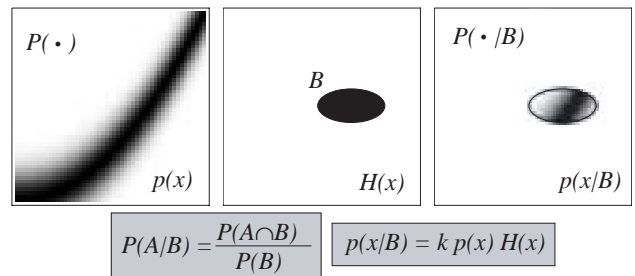
$$P(\mathcal{A}|\mathcal{B}) = \frac{P(\mathcal{A} \cap \mathcal{B})}{P(\mathcal{B})} . \quad (2.87)$$

It is important to intuitively understand this definition. The left of figure 2.7 (to be examined later with more detail) suggests a 2D probability distribution $P(\cdot)$, that to any region \mathcal{A} of the space associates the probability $P(\mathcal{A})$. Given now a fixed region \mathcal{B} , suggested in the figure by an ovoid, we can define another probability distribution, denoted $P(\cdot|\mathcal{B})$ that to any region \mathcal{A} of the space associates the probability $P(\mathcal{A}|\mathcal{B})$ defined by equation 2.87. The probability $P(\cdot|\mathcal{B})$ is to be understood as

- being identical to $P(\cdot)$ inside \mathcal{B} (except for a renormalization factor guaranteeing that $P(\mathcal{B}|\mathcal{B}) = 1$),
- vanishing outside \mathcal{B} .

This standard definition of conditional probability is mathematically consistent, and not prone to misinterpretations

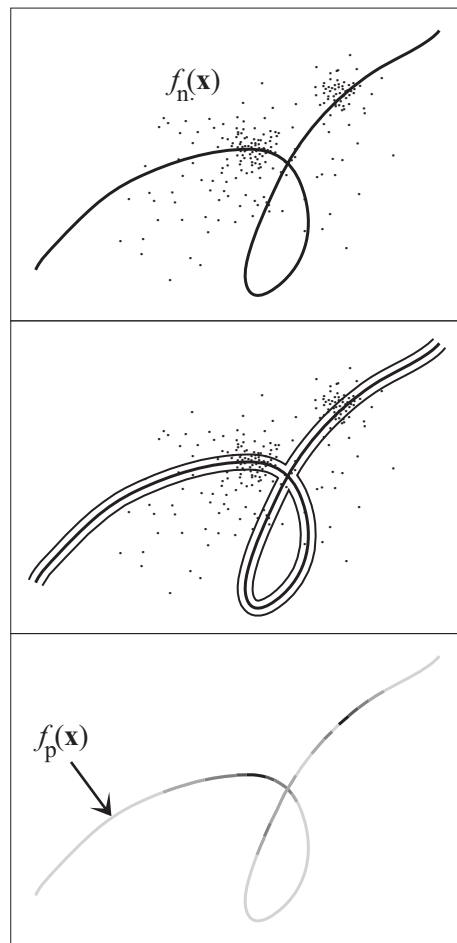
Figure 2.7: Illustration of the definition of conditional probability. Given an initial probability distribution $P(\cdot)$ (left of the figure) and a set \mathcal{B} (middle of the figure), $P(\cdot|\mathcal{B})$ is identical to $P(\cdot)$ inside \mathcal{B} (except for a renormalization factor guaranteeing that $P(\mathcal{B}|\mathcal{B}) = 1$) and vanishes outside \mathcal{B} (right of the figure)



2.4.2 Conditional Volumetric Probability

A volumetric probability over n -dimensional manifold induces a volumetric probability over any p -dimensional submanifold (see figure 2.8). We examine here the details of this important issue.

Figure 2.8: Top: A probability distribution in an n -dimensional metric manifold \mathbf{X}_n is suggested by some sample points. The probability distribution can be represented by a volumetric probability $f_n(\mathbf{x})$, proportional everywhere to the number of points per unit of n -dimensional volume. A p -dimensional submanifold \mathbf{X}_p is also suggested (by a line). Middle: To define the conditional volumetric probability on the submanifold \mathbf{X}_p , one considers a ‘tube’ of constant thickness around the submanifold, and counts the number of points the unit of n -dimensional volume. Bottom: In the limit where the thickness of the tube tends to zero, this defines a p -dimensional volumetric probability over the submanifold \mathbf{X}_p . The metric over \mathbf{X}_p is that induced by the metric over \mathbf{X}_n , as is the element of volume. When the n coordinates $\mathbf{x} = \{x^1, \dots, x^n\}$ can be separated into p coordinates $\mathbf{r} = \{r^1, \dots, r^p\}$ and q coordinates $\mathbf{s} = \{s^1, \dots, s^q\}$ (with $n = p + q$), so that the p -dimensional submanifold \mathbf{X}_p can be defined by the conditions $\mathbf{s} = \mathbf{s}(\mathbf{r})$, then, the coordinates \mathbf{r} can be used as coordinates over the submanifold \mathbf{X}_p , and the (p -dimensional) conditional volumetric probability, as given by equation 2.95, is simply $f_p(\mathbf{r}) = k f_n(\mathbf{r}, \mathbf{s}(\mathbf{r}))$, where k is a normalization constant. The probability of a region $\mathcal{A}_p \subset \mathbf{X}_p$ is to be evaluated as $P(\mathcal{A}_p) = \int_{\mathbf{r} \in \mathcal{A}_p} dv_p(\mathbf{x}) f_p(\mathbf{x})$, where the p -dimensional volume element $dv_p(\mathbf{x})$ is given in equations 2.97–2.99.



2.4.2.1 General Situation

As in section 2.1.5, consider an n -dimensional manifold \mathbf{X}_n , with some coordinates $\mathbf{x} = \{x^1, \dots, x^n\}$, and a metric tensor $\mathbf{g}(\mathbf{x}) = \{g_{ij}(\mathbf{x})\}$. The n -dimensional volume element is, then, $dV(\mathbf{x}) = \bar{g}(\mathbf{x}) d\underline{v}(\mathbf{x}) = \sqrt{\det \mathbf{g}(\mathbf{x})} dx^1 \wedge \dots \wedge dx^n$. In section 2.1.5, the n coordinates $\mathbf{x} = \{x^1, \dots, x^n\}$ of \mathbf{X} have been separated into one group of p coordinates $\mathbf{r} = \{r^1, \dots, r^p\}$ and one group of q coordinates $\mathbf{s} = \{s^1, \dots, s^q\}$, with $p + q = n$, and a p -dimensional submanifold \mathbf{X}_p of the n -dimensional manifold \mathbf{X} (with $p \leq n$) has been introduced via the constraint

$$\mathbf{s} = \mathbf{s}(\mathbf{r}) . \quad (2.88)$$

Consider a probability distribution P over \mathbf{X}_n , represented by the volumetric probability

$f(\mathbf{x}) = f(\mathbf{r}, \mathbf{s})$. We wish to define (and to characterize) the ‘conditional volumetric probability’ induced over the submanifold by the volumetric probability $f(\mathbf{x}) = f(\mathbf{r}, \mathbf{s})$.

Given the p -dimensional submanifold \mathbf{X}_p of the n -dimensional manifold \mathbf{X}_n , one can define a set $\mathcal{B}(\Delta s)$ as being the set of all points whose distance to the submanifold \mathbf{X}_p is less or equal than Δs . For any finite value of Δs , Kolmogorov’s definition of conditional probability applies, and the conditional probability so defined associates, to any $\mathcal{A} \subset \mathbf{X}_n$, the probability 2.87. Excepted for a normalization factor, this conditional probability equals the original one, excepted in that all the region whose points are at a distance larger than Δs have been ‘trimmed away’. This is still a probability distribution over \mathbf{X}_n . In the limit when $\Delta s \rightarrow 0$ this shall define a probability distribution over the submanifold \mathbf{X}_p that we are about to characterize.

Consider a volume element dv_p over the submanifold \mathbf{X}_n , and all the points of \mathbf{X}_n that are at a distance smaller or equal that Δs of the points inside the volume element. For small enough Δs the n -dimensional volume Δv_n so defined is

$$\Delta v_n \approx dv_p \Delta \omega_q \quad , \quad (2.89)$$

where $\Delta \omega_q$ is the volume of the q -dimensional sphere of radius Δs that is orthogonal to the submanifold at the considered point. This volume is proportional to $(\Delta s)^q$, so we have

$$\Delta v_n \approx k dv_p (\Delta s)^q \quad , \quad (2.90)$$

where k is a numerical factor. The conditional probability associated of this n -dimensional region by formula 2.87 is, by definition of volumetric probability,

$$dP_{(p+q)} \approx k' f \Delta v_n \approx k'' f dv_p (\Delta s)^q \quad , \quad (2.91)$$

where k' and k'' are constants. The conditional probability of the p -dimensional volume element dv_p of the submanifold \mathbf{X}_p is then defined as the limit

$$dP_p = \lim_{\Delta s \rightarrow 0} \frac{dP_{(p+q)}}{(\Delta s)^q} \quad , \quad (2.92)$$

this giving $dP_n = k'' f dv_p$, or, to put the variables explicitly,

$$dP_n(\mathbf{r}) = k'' f(\mathbf{r}, \mathbf{s}(\mathbf{r})) dv_p(\mathbf{r}) \quad . \quad (2.93)$$

We have thus arrived at a p -dimensional volumetric probability over the submanifold \mathbf{X}_p that is given by

$$f_p(\mathbf{r}) = k'' f(\mathbf{r}, \mathbf{s}(\mathbf{r})) \quad , \quad (2.94)$$

where k'' is a constant. If the probability is normalizable, and we choose to normalize it to one, then,

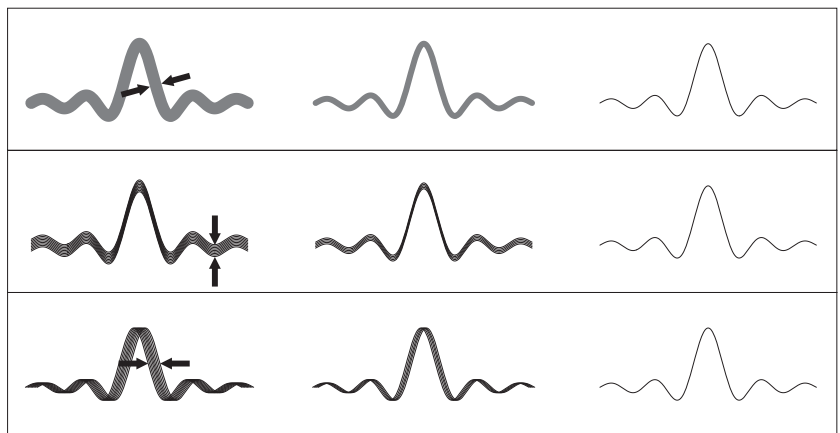
$$\boxed{f_p(\mathbf{r}) = \frac{f(\mathbf{r}, \mathbf{s}(\mathbf{r}))}{\int_{\mathbf{r} \in \mathbf{X}_p} dv_p(\mathbf{r}) f(\mathbf{r}, \mathbf{s}(\mathbf{r}))} \quad .} \quad (2.95)$$

With this volumetric probability, the probability of a region \mathcal{A}_p of the submanifold is computed as

$$P(\mathcal{A}_p) = \int_{\mathbf{r} \in \mathcal{A}_p} dv_p(\mathbf{x}) f_p(\mathbf{r}) \quad . \quad (2.96)$$

I must emphasize here the the limit we have used to define the conditional volumetric probability is an ‘orthogonal limit’ (see figure 2.9). This contrasts with usual texts, where, instead, a ‘vertical limit’ is used. The formal similarity of the result 2.95 with that proposed in the books that use the ‘vertical limit’ deserves explanation: we are handling here volumetric probabilities, not probability densities. The results for the ‘orthogonal limit’ used here, when translated to the language of probability densities, give results that are not the familiar results of common texts (see appendix 2.8.1).

Figure 2.9: The three limits that could be used to define a conditional volumetric probability over a submanifold. In the top, the ‘orthogonal’ or ‘natural’ limit. In the middle, the usual ‘vertical’ limit, and in the bottom a ‘horizontal’ limit. The last two although mentioned below (section 2.4.2.2), are not used in this book.



As already mentioned, the coordinates \mathbf{r} define a coordinate system over the submanifold \mathbf{X}_p . The volume element of the submanifold can, then, be written

$$dv_p(\mathbf{r}) = \bar{g}_p(\mathbf{r}) d\underline{v}_p(\mathbf{r}) \quad , \quad (2.97)$$

with $d\underline{v}_p(\mathbf{r}) = dr^1 \wedge \dots \wedge dr^p$. The volume density in the coordinates \mathbf{r} on the submanifold \mathbf{X}_p has been characterized in section 2.1.5 (equation 2.37):

$$\bar{g}_p(\mathbf{r}) = \sqrt{\det \mathbf{g}_p(\mathbf{r})} \quad , \quad (2.98)$$

with

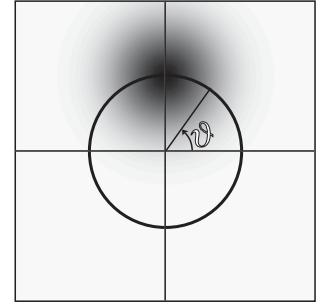
$$\mathbf{g}_p(\mathbf{r}) = \mathbf{g}_{rr} + \mathbf{g}_{rs} \mathbf{S} + \mathbf{S}^T \mathbf{g}_{sr} + \mathbf{S}^T \mathbf{g}_{ss} \mathbf{S} \quad . \quad (2.99)$$

It is understood that all the ‘matrices’ appearing at the right are taken at the point $(\mathbf{r}, \mathbf{s}(\mathbf{r}))$. The probability of a region \mathcal{A}_p of the submanifold can then either be computed using equation 2.96 or as

$$P(\mathcal{A}_p) = \int_{\mathbf{r} \in \mathcal{A}_p} d\underline{v}(\mathbf{r}) \bar{g}_p(\mathbf{r}) f_p(\mathbf{r}) \quad , \quad (2.100)$$

with the $\bar{g}_p(\mathbf{r})$ given in equation 2.98 and with $d\underline{v}(\mathbf{r}) = dr^1 \wedge \dots \wedge dr^p$.

Figure 2.10: The spherical Fisher distribution corresponds to the conditional probability distribution induced over a sphere by a Gaussian probability distribution in an Euclidean 3D space (see example 2.5). To have a full 3D representation of the property, this figure should be ‘rotated around the vertical axis’.



Example 2.5 In the Euclidean 3D space, consider an isotropic Gaussian probability distribution with standard deviation σ . Which is the conditional (2D) volumetric probability it induces on the surface of a sphere of unit radius whose center is at unit distance from the center of the Gaussian? Using geographical coordinates (see figure 2.10), the answer is given by the (2D) volumetric probability

$$f(\varphi, \lambda) = k \exp\left(\frac{\sin \lambda}{\sigma^2}\right) , \quad (2.101)$$

where k is a norming constant (see the demonstration in appendix XXX). This is the celebrated Fisher probability distribution, widely used as a model probability on the sphere’s surface. The surface element over the surface of the sphere could be obtained using the equations 2.98–2.99, but it is well known to be $dS(\varphi, \lambda) = \cos \lambda d\varphi d\lambda$. [End of example.]

Example 2.6 In the case where we work in a two-dimensional space \mathbf{X}_2 , with $p = q = 1$, we can use the notation r and s instead of \mathbf{r} and \mathbf{s} , so that the constraint 2.88 is written

$$s = s(r) , \quad (2.102)$$

and the ‘matrix’ of partial derivatives is now a simple real quantity

$$S = \frac{\partial s}{\partial r} . \quad (2.103)$$

The conditional volumetric probability on the line $s = s(r)$ induced by a volumetric probability $f(r, s)$ is (equation 2.95),

$$f_1(r) = \frac{f(r, s(r))}{\int dl(r') f(r', s(r'))} , \quad (2.104)$$

where, if the metric of the space \mathbf{X}_2 is written

$$\mathbf{g}(r, s) = \begin{pmatrix} g_{rr}(r, s) & g_{rs}(r, s) \\ g_{sr}(r, s) & g_{ss}(r, s) \end{pmatrix} , \quad (2.105)$$

the (1D) volume element is (equations 2.97–2.99)

$$dl(r) = \sqrt{g_{rr}(r, s(r)) + 2S(r)g_{rs}(r, s(r)) + S(r)^2g_{ss}(r, s(r))} dr . \quad (2.106)$$

The probability of an interval $(r_1 < r < r_2)$ along the line $s = s(r)$ is then

$$P = \int_{r_1}^{r_2} dl(r) f_1(r) . \quad (2.107)$$

If the constraint 2.102 is, in fact, $s = s_0$, then, equation 2.104 simplifies into

$$f_1(r) = \frac{f(r, s_0)}{\int d\ell(r') f(r', s_0)} \quad , \quad (2.108)$$

and, as the partial derive vanishes, $S = 0$, the length element 2.106 becomes

$$d\ell(r) = \sqrt{g_{rr}(r, s_0)} dr \quad . \quad (2.109)$$

[End of example.]

Example 2.7 Consider two Cartesian coordinates $\{x, y\}$ on the Euclidean plane, associated to the usual metric $ds^2 = dx^2 + dy^2$. It is easy to see (using, for instance, equation 1.23) that the metric matrix associated to the new coordinates (see figure 2.11)

$$r = x \quad ; \quad s = xy \quad (2.110)$$

is

$$\mathbf{g}(r, s) = \begin{pmatrix} 1 + s^2/r^4 & -s/r^3 \\ -s/r^3 & 1/r^2 \end{pmatrix} \quad , \quad (2.111)$$

with metric determinant $\sqrt{\det \mathbf{g}(r, s)} = 1/r$. Assume that all what we know about the position of a given point is described by the volumetric probability $f(r, s)$. Then, we are told that, in fact, the point is on the line defined by the equation $s = s_0$. What can we now say about the coordinate r of the point? This is clearly a problem of conditional volumetric probability, and the information we have now on the position of the point is represented by the volumetric probability (on the line $s = s_0$) given by equation 2.108:

$$f_1(r) = \frac{f(r, s_0)}{\int d\ell(r') f(r', s_0)} \quad . \quad (2.112)$$

Here, considering the special form of the metric in equation 2.111, the length element given by equation 2.109 is

$$d\ell(r) = \sqrt{1 + s_0^2/r^4} dr \quad . \quad (2.113)$$

The special case $s = s_0 = 0$ gives

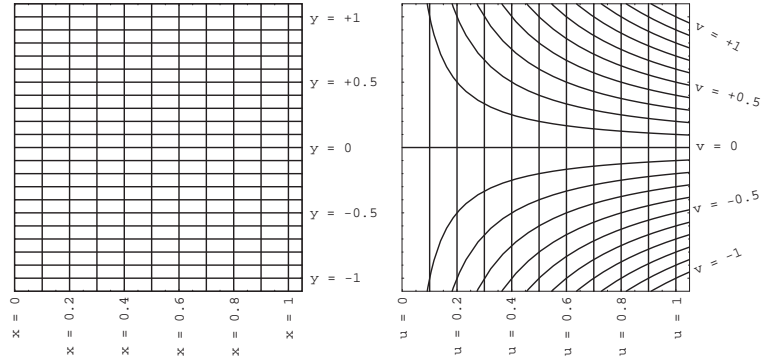
$$f_1(r) = \frac{f(r, 0)}{\int d\ell(r') f(r', 0)} \quad ; \quad d\ell(r) = dr \quad . \quad (2.114)$$

[End of example.]

Example 2.8 To address a paradox mentioned by E.T. Jaynes, let us solve the same problem as in the previous example, but using the Cartesian coordinates $\{x, y\}$. The information that was represented by the volumetric probability $f(r, s)$ is now represented by the volumetric probability $h(x, y)$ given by (as volumetric probabilities are invariant objects)

$$h(x, y) = f(r, s)|_{r=x ; s=xy} \quad . \quad (2.115)$$

Figure 2.11: The Euclidian plane, with, at the left, two Cartesian coordinates $\{x, y\}$, and, at the right the two coordinates $u = x ; v = x y$.



As the condition $s = 0$ is equivalent to the condition $y = 0$, and as the metric matrix is the identity, it is clear that we shall arrive, for the (1D) volumetric probability representing the information we have on the coordinate x to

$$h_1(x) = \frac{h(x, 0)}{\int d\ell(x') h(x', 0)} \quad ; \quad d\ell(x) = dx \quad . \quad (2.116)$$

Not only this equation is similar in form to equation 2.114; replacing here h by f (using equation 2.115) we obtain an identity that can be expressed using any of the two equivalent forms

$$h_1(x) = f_1(r)|_{r=x} \quad ; \quad f_1(r) = h_1(x)|_{x=r} \quad . \quad (2.117)$$

Along the line $s = y = 0$, the two coordinates r and s coincide, so we obtain the same volumetric probability (with the same length elements $d\ell(x) = dx$ and $d\ell(r) = dr$). Trivial as it may seem, this result is not that found in the traditional definition of conditional probability density. Jaynes, in the 15-th chapter of his unfinished *Probability Theory* book lists this as one of the paradoxes of probability theory. It is not a paradox, it is a mistake one makes when falling into the illusion that a conditional probability density (or a conditional volumetric probability) can be defined without invoking the existence of a metric (i.e., of a notion of distance) in the working space. This ‘paradox’ is related to the ‘Borel-Kolmogorov paradox’, that I address in appendix 2.8.10. **[End of example.]**

2.4.2.2 Case $\mathbf{X} = \mathbf{R} \times \mathbf{S}$

I shall show here that a ‘joint’ volumetric probability $f(\mathbf{r}, \mathbf{s})$ over a space $\mathbf{X}_{p+q} = \mathbf{R}_p \times \mathbf{S}_q$ can induce, via a relation $\mathbf{s} = \mathbf{s}(\mathbf{r})$, three different conditional volumetric probabilities: (i) a volumetric probability $f(\mathbf{r})$ over the submanifold $\mathbf{s} = \mathbf{s}(\mathbf{r})$ itself; (ii) a volumetric probability $f_{\mathbf{r}}(\mathbf{r})$ over \mathbf{R} ; and (iii) a volumetric probability $f_{\mathbf{s}}(\mathbf{r})$ (case $p \leq q$) or $f_{\mathbf{s}}(\mathbf{s})$ (case $p \geq q$) over \mathbf{S}_q . Figure 2.12 shows a schematical view of the properties we are about to analyze.

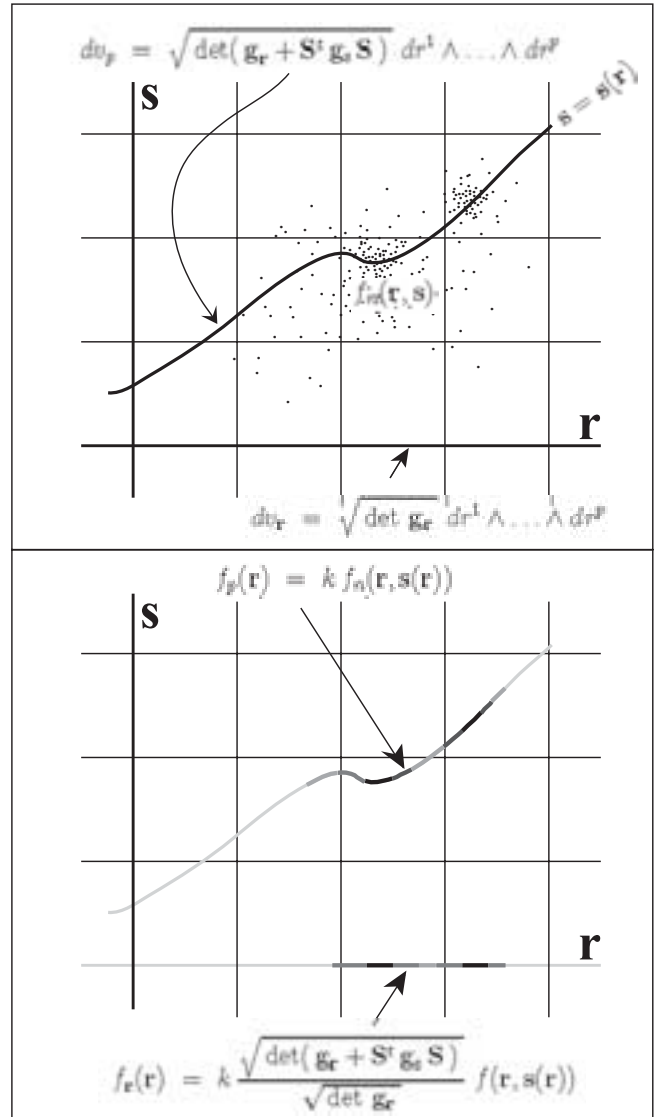


Figure 2.12: In an n -dimensional space \mathbf{X}_n that is the Cartesian product of two spaces \mathbf{R}_p and \mathbf{S}_q , with coordinates $\mathbf{r} = \{r^1, \dots, r^p\}$ and $\mathbf{s} = \{s^1, \dots, s^q\}$ and metric tensors \mathbf{g}_r and \mathbf{g}_s , there is a volume element on each of \mathbf{R}_p and \mathbf{S}_q , and an induced volume element in $\mathbf{X}_n = \mathbf{R}_p \times \mathbf{S}_q$. Given a p -dimensional submanifold manifold $\mathbf{s} = \mathbf{s}(\mathbf{r})$ of \mathbf{X}_n , there also is an induced volume element on it. A volumetric probability $f(\mathbf{r}, \mathbf{s})$ over \mathbf{X}_n , induces a (conditional) volumetric probability $f_x(\mathbf{r})$ over the submanifold $\mathbf{s} = \mathbf{s}(\mathbf{r})$ (equation 2.125), and, as the submanifold shares the same coordinates as \mathbf{R}_p , a volumetric probability $f_r(\mathbf{r})$ is also induced over \mathbf{R}_p (equation 2.127). This volumetric probability can, in turn, be transported into \mathbf{S}_q , using the concepts developed in section 2.6.

Consider a p -dimensional manifold \mathbf{R} with a coordinate system $\mathbf{r} = \{r^\alpha\}$ and metric tensor $\mathbf{g}_r(\mathbf{r})$, and a q -dimensional manifold \mathbf{S} with a coordinate system $\mathbf{s} = \{s^i\}$ and metric tensor $\mathbf{g}_s(\mathbf{s})$. Each space has, then, a distance element

$$ds_{\mathbf{r}}^2 = (\mathbf{g}_r)_{\alpha\beta} dr^\alpha dr^\beta \quad ; \quad ds_{\mathbf{s}}^2 = (\mathbf{g}_s)_{ij} ds^i ds^j \quad , \quad (2.118)$$

and a volume element

$$dv_{\mathbf{r}}(\mathbf{r}) = \bar{g}_{\mathbf{r}}(\mathbf{r}) d\underline{v}_{\mathbf{r}}(\mathbf{r}) \quad ; \quad dv_{\mathbf{s}}(\mathbf{s}) = \bar{g}_{\mathbf{s}}(\mathbf{s}) d\underline{v}_{\mathbf{s}}(\mathbf{s}) \quad , \quad (2.119)$$

that are related to the capacity elements

$$d\underline{v}_{\mathbf{r}}(\mathbf{r}) = dr^1 \wedge \cdots \wedge dr^p \quad ; \quad d\underline{v}_{\mathbf{s}}(\mathbf{r}) = ds^1 \wedge \cdots \wedge ds^q \quad (2.120)$$

via the volume densities

$$\bar{g}_{\mathbf{r}}(\mathbf{r}) = \eta_{\mathbf{r}} \sqrt{\det \mathbf{g}_{\mathbf{r}}(\mathbf{r})} \quad ; \quad \bar{g}_{\mathbf{s}}(\mathbf{s}) = \eta_{\mathbf{s}} \sqrt{\det \mathbf{g}_{\mathbf{s}}(\mathbf{s})} \quad . \quad (2.121)$$

We can build the *Cartesian product* $\mathbf{X} = \mathbf{R} \times \mathbf{S}$ of the two spaces, by defining the points of \mathbf{X} as being made by a point of \mathbf{R} and a point of \mathbf{S} (so we can write $\mathbf{x} = \{\mathbf{r}, \mathbf{s}\}$), and by introducing a metric tensor $\mathbf{g}(\mathbf{x})$ over \mathbf{X} through the definition⁵

$$ds^2 = ds_{\mathbf{r}}^2 + ds_{\mathbf{s}}^2 \quad . \quad (2.122)$$

This implies that the metric $\mathbf{g}(\mathbf{x}) = \mathbf{g}(\mathbf{r}, \mathbf{s})$ has the partitioned form

$$\mathbf{g}(\mathbf{r}, \mathbf{s}) = \begin{pmatrix} \mathbf{g}_{\mathbf{r}}(\mathbf{r}) & 0 \\ 0 & \mathbf{g}_{\mathbf{s}}(\mathbf{s}) \end{pmatrix} \quad . \quad (2.123)$$

Note: explain that what follows is on the submanifold.

With this partitioned metric, the metric tensor in equation 2.99 simplifies to

$$\mathbf{g}_p = \mathbf{g}_{\mathbf{r}} + \mathbf{S}^T \mathbf{g}_{\mathbf{s}} \mathbf{S} \quad , \quad (2.124)$$

or, more explicitly, $\mathbf{g}_p(\mathbf{r}) = \mathbf{g}_{\mathbf{r}}(\mathbf{r}) + \mathbf{S}^T(\mathbf{r}) \mathbf{g}_{\mathbf{s}}(\mathbf{s}(\mathbf{r})) \mathbf{S}(\mathbf{r})$. Collecting here equations 2.95, 2.98 and 2.100, we can write the conditional probability of a region \mathcal{A}_p of the submanifold $\mathbf{s} = \mathbf{s}(\mathbf{r})$ as

$$f_{\mathbf{x}}(\mathbf{r}) = k f(\mathbf{r}, \mathbf{s}(\mathbf{r})) \quad , \quad (2.125)$$

where k is a normalization constant. Using the volume element over the submanifold, the probability of a region \mathcal{A} of the submanifold $\mathbf{s} = \mathbf{s}(\mathbf{r})$ is computed via

$$P(\mathcal{A}) = \int_{\mathcal{C}} dr^1 \wedge \cdots \wedge dr^p \sqrt{\det(\mathbf{g}_{\mathbf{r}} + \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S})} f_{\mathbf{x}}(\mathbf{r}) \quad . \quad (2.126)$$

As the conditional volumetric probability $f_{\mathbf{x}}(\mathbf{r})$ is on the submanifold $\mathbf{s} = \mathbf{s}(\mathbf{r})$, it is integrated with the volume density of the submanifold (equation 2.126). Remember that the coordinates \mathbf{r} are not only the coordinates of the subspace \mathbf{R} , they also define a coordinate system over the submanifold $\mathbf{s} = \mathbf{s}(\mathbf{r})$.

Note: explain that what follows is on the space \mathbf{R}_p :

Equations 2.125–2.126 define a volumetric probability over the submanifold \mathbf{X}_p . As the coordinates \mathbf{r} are both, coordinates over \mathbf{R}_p and over the submanifold \mathbf{X}_p , if we define

$$f_{\mathbf{r}}(\mathbf{r}) = k \frac{\sqrt{\det(\mathbf{g}_{\mathbf{r}} + \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S})}}{\sqrt{\det \mathbf{g}_{\mathbf{r}}}} f(\mathbf{r}, \mathbf{s}(\mathbf{r})) \quad , \quad (2.127)$$

⁵Expression 2.122 is just a special situation. More generally, one should take $ds^2 = \alpha^2 ds_{\mathbf{r}}^2 + \beta^2 ds_{\mathbf{s}}^2$.

where the normalization factor k is given by

$$\frac{1}{k} = \int_{\mathbb{R}_p} dv_{\mathbf{r}}(\mathbf{r}) \frac{\sqrt{\det(\mathbf{g}_{\mathbf{r}} + \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S})}}{\sqrt{\det \mathbf{g}_{\mathbf{r}}}} f(\mathbf{r}, \mathbf{s}(\mathbf{r})) \quad , \quad (2.128)$$

a probability is then expressed as

$$P(\mathcal{A}) = \int_{\mathcal{A}} dv_{\mathbf{r}}(\mathbf{r}) f_{\mathbf{r}}(\mathbf{r}) \quad , \quad (2.129)$$

the volume element being

$$dv_{\mathbf{r}}(\mathbf{r}) = \sqrt{\det \mathbf{g}_{\mathbf{r}}} dr^1 \wedge \cdots \wedge dr^p \quad . \quad (2.130)$$

As this is the volume element of \mathbb{R}_p , we see that we have defined a volumetric probability over \mathbb{R}_p . [Note: This is very important, it has to be better explained.]

We see thus that, via $\mathbf{s} = \mathbf{s}(\mathbf{r})$, the volumetric probability $F(\mathbf{r}, \mathbf{s})$, has not only induced a conditional volumetric probability $f_{\mathbf{x}}(\mathbf{r})$ over the submanifold $\mathbf{s} = \mathbf{s}(\mathbf{r})$, but also a volumetric probability $f_{\mathbf{r}}(\mathbf{r})$ over \mathbb{R} . These two volumetric probabilities are completely equivalent, and one may focus in one or the other depending on the applications in view. We shall talk about the conditional volumetric probability $f_{\mathbf{x}}(\mathbf{r})$ on the submanifold $\mathbf{s} = \mathbf{s}(\mathbf{r})$ and about the conditional volumetric probability $f_{\mathbf{r}}(\mathbf{r})$ on the subspace \mathbb{R} .

If instead of the volumetric probabilities $f_{\mathbf{r}}(\mathbf{r})$ and $f(\mathbf{r}, \mathbf{s})$ we introduce the probability densities

$$\begin{aligned} \bar{f}(\mathbf{r}, \mathbf{s}) &= \bar{g}_{\mathbf{r}}(\mathbf{r}) f(\mathbf{r}, \mathbf{s}) = \sqrt{\det \mathbf{g}_{\mathbf{r}}(\mathbf{r})} f(\mathbf{r}, \mathbf{s}) \\ \bar{f}(\mathbf{r}, \mathbf{s}) &= \bar{g}(\mathbf{r}, \mathbf{s}) f(\mathbf{r}, \mathbf{s}) = \sqrt{\det \mathbf{g}_{\mathbf{r}}(\mathbf{r})} \sqrt{\det \mathbf{g}_{\mathbf{s}}(\mathbf{s})} f(\mathbf{r}, \mathbf{s}) \quad , \end{aligned} \quad (2.131)$$

then, equation 2.127 becomes

$$\bar{f}_{\mathbf{r}}(\mathbf{r}) = k \frac{\sqrt{\det(\mathbf{g}_{\mathbf{r}} + \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S})}}{\sqrt{\det \mathbf{g}_{\mathbf{r}}} \sqrt{\det \mathbf{g}_{\mathbf{s}}}} \bar{f}(\mathbf{r}, \mathbf{s}(\mathbf{r})) \quad , \quad (2.132)$$

where the normalization factor k is given by

$$\frac{1}{k} = \int_{\mathbb{R}_p} dv_{\mathbf{r}}(\mathbf{r}) \frac{\sqrt{\det(\mathbf{g}_{\mathbf{r}} + \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S})}}{\sqrt{\det \mathbf{g}_{\mathbf{r}}} \sqrt{\det \mathbf{g}_{\mathbf{s}}}} \bar{f}(\mathbf{r}, \mathbf{s}(\mathbf{r})) \quad , \quad (2.133)$$

the capacity element being

$$dv_{\mathbf{r}}(\mathbf{r}) = dr^1 \wedge \cdots \wedge dr^p \quad . \quad (2.134)$$

A probability is expressed as

$$P(\mathcal{A}) = \int_{\mathcal{A}} dv_{\mathbf{r}}(\mathbf{r}) \bar{f}_{\mathbf{r}}(\mathbf{r}) \quad . \quad (2.135)$$

Note: analyze here the case where the application $\mathbf{s} = \mathbf{s}(\mathbf{r})$ degenerates into

$$\mathbf{s} = \mathbf{s}_0 \quad , \quad (2.136)$$

in which case the matrix \mathbf{S} of partial derivatives vanishes. Then, using for the conditional volumetric probability the usual notation $f(\mathbf{r}|\mathbf{s}_0)$, equations 2.127–2.128 simply give

$$f(\mathbf{r}|\mathbf{s}_0) = \frac{f(\mathbf{r}, \mathbf{s}_0)}{\int dv_{\mathbf{r}}(\mathbf{r}) f(\mathbf{r}, \mathbf{s}_0)} . \quad (2.137)$$

Equivalently, in terms of probability densities, equations 2.132–2.133 become, in the case $\mathbf{s} = \mathbf{s}_0$,

$$\bar{f}(\mathbf{r}|\mathbf{s}_0) = \frac{\bar{f}(\mathbf{r}, \mathbf{s}_0) / \sqrt{\det \mathbf{g}_s(\mathbf{s}_0)}}{\int d\underline{v}_{\mathbf{r}}(\mathbf{r}) \bar{f}(\mathbf{r}, \mathbf{s}_0) / \sqrt{\det \mathbf{g}_s(\mathbf{s}_0)}} . \quad (2.138)$$

Note: I have to check if I can drop the constant term $\sqrt{\det \mathbf{g}_s(\mathbf{s}_0)}$ from this equation.

The assumption that the joint metric diagonalizes ‘in the variables’ $\{\mathbf{r}, \mathbf{s}\}$ is essential here. If from the variables $\{\mathbf{r}, \mathbf{s}\}$ we pass to some other variables $\{\mathbf{u}, \mathbf{v}\}$ through a general change of variables, the metric of the space \mathbf{X} shall no longer be diagonal in the new variables, and a definition of, say, $f(\mathbf{u}|\mathbf{v}_0)$ shall not be possible.

This difficulty is often disregarded in usual texts working with probability densities, this causing some confusions in applications of probability theory using the notion of conditional probability density, and the associated expression of the Bayes theorem (see section 2.5.4).

Example 2.9 *With the notations of this section, consider that the metric \mathbf{g}_r of the space \mathbb{R}_p and the metric \mathbf{g}_s of the space \mathbf{S}_q are constant (i.e., that both, the coordinates r^α and s^i are rectilinear coordinates in Euclidean spaces), and that the application $\mathbf{s} = \mathbf{s}(\mathbf{r})$ is a linear application, that we can write*

$$\mathbf{s} = \mathbf{S} \mathbf{r} , \quad (2.139)$$

as this is consistent with the definition of \mathbf{S} as the matrix of partial derivatives, $S^i_\alpha = \partial s^i / \partial r^\alpha$. Consider that we have a Gaussian probability distribution over the space \mathbb{R}_p , represented by the volumetric probability

$$f_p(\mathbf{r}) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} (\mathbf{r} - \mathbf{r}_0)^t \mathbf{g}_r (\mathbf{r} - \mathbf{r}_0)\right) , \quad (2.140)$$

that is normalized via $\int dr^1 \wedge \cdots \wedge dr^p \sqrt{\det \mathbf{g}_r} f_p(\mathbf{r}) = \sqrt{\det \mathbf{g}_r} \int dr^1 \wedge \cdots \wedge dr^p f_p(\mathbf{r}) = 1$. Similarly, consider that we also have a Gaussian probability distribution over the space \mathbf{S}_q , represented by the volumetric probability

$$f_q(\mathbf{s}) = \frac{1}{(2\pi)^{q/2}} \exp\left(-\frac{1}{2} (\mathbf{s} - \mathbf{s}_0)^t \mathbf{g}_s (\mathbf{s} - \mathbf{s}_0)\right) , \quad (2.141)$$

that is normalized via $\int ds^1 \wedge \cdots \wedge ds^q \sqrt{\det \mathbf{g}_s} f_q(\mathbf{s}) = \sqrt{\det \mathbf{g}_s} \int ds^1 \wedge \cdots \wedge ds^q f_q(\mathbf{s}) = 1$. Finally, consider the $p+q$ -dimensional probability distribution over the space \mathbf{X}_{p+q} defined as the product of these two volumetric probabilities,

$$f(\mathbf{r}, \mathbf{s}) = f_p(\mathbf{r}) f_q(\mathbf{s}) . \quad (2.142)$$

Given this $p + q$ -dimensional volumetric probability $f(\mathbf{r}, \mathbf{s})$ and given the p -dimensional hyperplane $\mathbf{s} = \mathbf{S} \mathbf{r}$, we obtain the conditional volumetric probability $f_{\mathbf{r}}(\mathbf{r})$ over \mathbb{R}_p as given by equation 2.127. All simplifications done⁶ one obtains the Gaussian volumetric probability⁷

$$f_{\mathbf{r}}(\mathbf{r}) = \frac{1}{(2\pi)^{p/2}} \frac{\sqrt{\det \mathbf{g}'_{\mathbf{r}}}}{\sqrt{\det \mathbf{g}_{\mathbf{r}}}} \exp\left(-\frac{1}{2} (\mathbf{r} - \mathbf{r}'_0)^t \mathbf{g}'_{\mathbf{r}} (\mathbf{r} - \mathbf{r}'_0)\right) , \quad (2.143)$$

where the metric $\mathbf{g}'_{\mathbf{r}}$ (inverse of the covariance matrix) is

$$\mathbf{g}'_{\mathbf{r}} = \mathbf{g}_{\mathbf{r}} + \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S} \quad (2.144)$$

and where the mean \mathbf{r}'_0 can be obtained solving the expression⁸

$$\mathbf{g}'_{\mathbf{r}} (\mathbf{r}'_0 - \mathbf{r}_0) = \mathbf{S}^t \mathbf{g}_{\mathbf{s}} (\mathbf{s}_0 - \mathbf{S}_0 \mathbf{r}_0) . \quad (2.145)$$

Note: I should now show here that $f_{\mathbf{s}}(\mathbf{s})$, the volumetric probability in the space \mathbf{S}_q is given, in all cases ($p \leq q$ or $p \geq q$) by

$$f_{\mathbf{s}}(\mathbf{s}) = \frac{1}{(2\pi)^{q/2}} \frac{\sqrt{\det \mathbf{g}'_{\mathbf{s}}}}{\sqrt{\det \mathbf{g}_{\mathbf{s}}}} \exp\left(-\frac{1}{2} (\mathbf{s} - \mathbf{s}'_0)^t \mathbf{g}'_{\mathbf{s}} (\mathbf{s} - \mathbf{s}'_0)\right) , \quad (2.146)$$

where the metric $\mathbf{g}'_{\mathbf{s}}$ (inverse of the covariance matrix) is

$$(\mathbf{g}'_{\mathbf{s}})^{-1} = \mathbf{S} (\mathbf{g}'_{\mathbf{r}})^{-1} \mathbf{S}^t \quad (2.147)$$

and where the mean \mathbf{s}'_0 is

$$\mathbf{s}'_0 = \mathbf{S} \mathbf{r}'_0 . \quad (2.148)$$

Note: say that this is illustrated in figure 2.13. **[End of example.]**

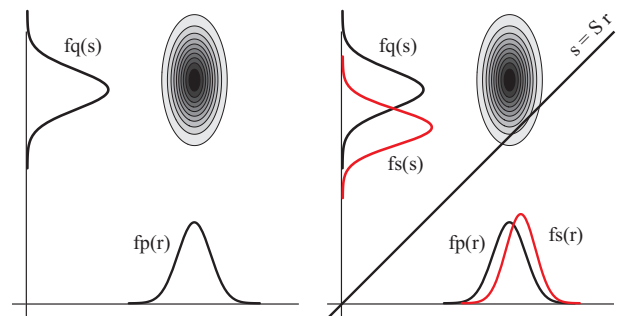


Figure 2.13: Provisional figure to illustrate example 2.9.

⁶Note: explain this.

⁷This volumetric probability is normalized by $\int dr^1 \wedge \dots \wedge dr^p \sqrt{\det \mathbf{g}_{\mathbf{r}}} f_{\mathbf{r}}(\mathbf{r}) = 1$.

⁸Explicitly, one can write $\mathbf{r}'_0 = \mathbf{r}_0 + (\mathbf{g}'_{\mathbf{r}})^{-1} \mathbf{S}^t \mathbf{g}_{\mathbf{s}} (\mathbf{s}_0 - \mathbf{S}_0 \mathbf{r}_0)$, but in numerical applications, the direct resolution of the linear system 2.145 is preferable.

2.5 Marginal Probability

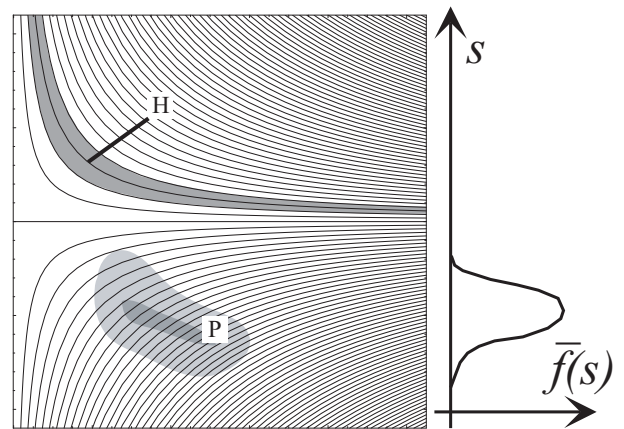
2.5.1 Marginal Probability Density

In a $p+q$ -dimensional space \mathbf{X}_{p+q} , consider a continuous, non intersecting set of p -dimensional hypersurfaces, parameterized by some parameters $\mathbf{s} = \{s^1, s^2, \dots, s^q\}$, as suggested in figure 2.14. Each given value of \mathbf{s} , say $\mathbf{s} = \mathbf{s}_0$, defines one such hypersurface.

Consider also a probability distribution over \mathbf{X}_{p+q} (suggested by the ovoidal shape marked 'P' in the figure). We have seen above that given a particular hypersurface $\mathbf{s} = \mathbf{s}_0$, we can define a conditional probability distribution, that associates a different value of a volumetric probability to each point of the hypersurface. We are not interested now in the 'variability' inside each hypersurface, but in defining a global 'probability' for each hypersurface, to analyze the variation of the probability from one hypersurface to another one.

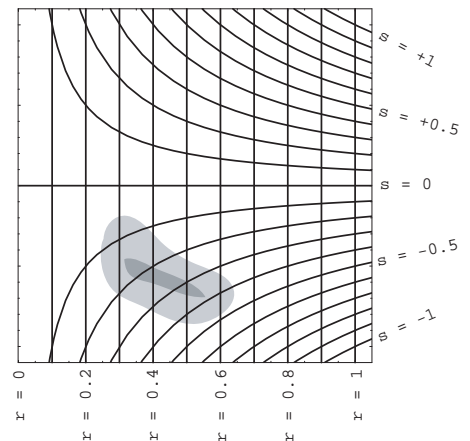
Crudely speaking, to the hypersurface marked 'H' in the figure, we are going to associate the probability of the small 'crescent' defined by two infinitely close hypersurfaces.

Figure 2.14: Figure por the definition of marginal probability. Caption to be written.



The easiest way to develop the idea (and to find explicit expressions) is to characterize the points inside each of the hypersurfaces by some coordinates $\mathbf{r} = \{r^1, r^2, \dots, r^p\}$. Still better, we can assume that the set $\{\mathbf{r}, \mathbf{s}\}$ individualizes one particular point of \mathbf{X}_{p+q} , i.e., the set $\mathbf{x} = \{\mathbf{r}, \mathbf{s}\}$ is a coordinate system over \mathbf{X}_{p+q} (see figure 2.15).

Figure 2.15: Figure por the definition of marginal probability. Caption to be written.



We shall verify at the end that the definition we are going to made of a probability distribution over \mathbf{s} is independent of the particular choice of coordinates \mathbf{r} .

Let, then, $f(\mathbf{r}, \mathbf{s})$ be a volumetric probability over \mathbf{X}_{p+q} . The probability of a domain $\mathcal{A} \subset \mathbf{X}_{p+q}$ is computed as

$$P(\mathcal{A}) = \int_{\mathcal{A}} dv(\mathbf{r}, \mathbf{s}) \bar{f}(\mathbf{r}, \mathbf{s}) \quad , \quad (2.149)$$

where $dv(\mathbf{r}, \mathbf{s}) = \sqrt{\det \mathbf{g}} ds^1 \wedge \cdots \wedge ds^q \wedge dr^1 \wedge \cdots \wedge dr^p$. Explicitly,

$$P(\mathcal{A}) = \int_{\mathcal{A}} ds^1 \wedge \cdots \wedge ds^q \wedge dr^1 \wedge \cdots \wedge dr^p \sqrt{\det \mathbf{g}} f(\mathbf{r}, \mathbf{s}) \quad . \quad (2.150)$$

As the (infinitesimal) probability of the ‘crescent’ around the hypersurface ‘H’ in figure 2.14 is

$$dP_q(\mathbf{s}) = ds^1 \wedge \cdots \wedge ds^q \int_{\text{all values of } \mathbf{r}} dr^1 \wedge \cdots \wedge dr^p \sqrt{\det \mathbf{g}} f(\mathbf{r}, \mathbf{s}) \quad , \quad (2.151)$$

we can introduce the definition

$$\bar{f}_s(\mathbf{s}) = \int_{\text{all values of } \mathbf{r}} dr^1 \wedge \cdots \wedge dr^p \sqrt{\det \mathbf{g}} f(\mathbf{r}, \mathbf{s}) \quad , \quad (2.152)$$

to have $dP_q(\mathbf{s}) = ds^1 \wedge \cdots \wedge ds^q \bar{f}_s(\mathbf{s})$. When the parameters \mathbf{s} are formally seen as coordinates over some (yet undefined) space, the probability of a region \mathcal{B} of this space is, by definition of $\bar{f}_s(\mathbf{s})$, computed as

$$P(\mathcal{B}) = \int_{\mathcal{B}} ds^1 \wedge \cdots \wedge ds^q \bar{f}_s(\mathbf{s}) \quad , \quad (2.153)$$

this showing that $\bar{f}_s(\mathbf{s})$ can be interpreted as a probability density over \mathbf{s} , that, by construction, corresponds to the integrated probability over the hypersurface defined by a constant value of the parameters \mathbf{s} (see figure 2.14 again). The expression 2.153 is the typical one for evaluating finite probabilities from a probability density (see equations 2.55–2.56); for this reason we shall call $\bar{f}_s(\mathbf{s})$ the *marginal probability density* (for the variables \mathbf{s}).

This is the most one can do given only the elements of the problem, i.e., a probability distribution over a space and a continuous family of hypersurfaces. Note that we have been able to introduce a probability density over the variables \mathbf{s} , but not a volumetric probability, that can only be defined over a well defined space.

Once we understand that we can only define a probability density $\bar{f}_s(\mathbf{s})$ (and not a volumetric probability) we can rewrite equation 2.152 as

$$\bar{f}_s(\mathbf{s}) = \int_{\text{all values of } \mathbf{r}} dr^1 \wedge \cdots \wedge dr^p \bar{f}(\mathbf{r}, \mathbf{s}) \quad , \quad (2.154)$$

where

$$\bar{f}(\mathbf{r}, \mathbf{s}) = \sqrt{\det \mathbf{g}} f(\mathbf{r}, \mathbf{s}) \quad (2.155)$$

is the probability density representing (in the coordinates $\{\mathbf{r}, \mathbf{s}\}$) the initial probability distribution over the space \mathbf{X}_{p+q} .

The elements used in the definition of the marginal probability density $\bar{f}_s(\mathbf{s})$ are: (i) a probability distribution over a $(p+q)$ -dimensional metric space \mathbf{X}_{p+q} , and (ii) a continuous family of p -dimensional hypersurfaces characterized by some q parameters $\mathbf{s} = \{s^1, \dots, s^q\}$. This is independent of any coordinate system over \mathbf{X}_{p+q} . It remains that the q parameters \mathbf{s} can be considered as q coordinates over \mathbf{X}_{p+q} that can be completed, in an arbitrary manner, by p more coordinates $\mathbf{r} = \{r^1, \dots, r^p\}$ in order to have a complete coordinate system $\mathbf{x} = \{\mathbf{r}, \mathbf{s}\}$ over \mathbf{X}_{p+q} . That the probability density $\bar{f}_s(\mathbf{s})$ is independent of the choice of the coordinates \mathbf{r} is seen by considering equation 2.152. For any fixed value of \mathbf{s} (i.e., on a given p -dimensional submanifold), the term $\sqrt{\det \mathbf{g}} dr^1 \wedge \dots \wedge dr^p$ is just the expression of the volume element on the submanifold, that, by definition, is an invariant, as is the volumetric probability f . Therefore, the integral sum in equation 2.152 shall keep its value invariant under any change of the coordinates \mathbf{r} .

In many applications, the continuous family of p -dimensional hypersurfaces is not introduced per se. Rather, one has a given coordinate system \mathbf{x} over \mathbf{X}_{p+q} that is, for some reason, splitted into p coordinates \mathbf{r} and q coordinates \mathbf{s} . These coordinates define different coordinate hypersurfaces over \mathbf{X}_{p+q} , and, among them, the p -dimensional hypersurfaces defined by constant values of the coordinates \mathbf{s} . Then, the definition of marginal probability density given above applies.

NOTE COME BACK HERE AFTER ANALYZING POISSON.

In this particular situation, the metric properties of the space need not to be taken into account, and the two equations 2.153–2.154, than only invoke probability densities can be used.

2.5.2 Marginal Volumetric Probability

Consider now the special situation where the $(p+q)$ -dimensional space \mathbf{X}_{p+q} is defined as the Cartesian product of two spaces, $\mathbf{X} = \mathbf{R} \times \mathbf{S}$, with respective dimensions p and q . The notion of Cartesian product of two metric manifolds has been introduced in section 2.4.2.2.

Note: recall here equations 2.122–2.123:

$$ds^2 = ds_{\mathbf{r}}^2 + ds_{\mathbf{s}}^2 \quad . \quad (2.156)$$

This implies that the metric $\mathbf{g}(\mathbf{x}) = \mathbf{g}(\mathbf{r}, \mathbf{s})$ has the partitioned form

$$\mathbf{g}(\mathbf{r}, \mathbf{s}) = \begin{pmatrix} \mathbf{g}_{\mathbf{r}}(\mathbf{r}) & 0 \\ 0 & \mathbf{g}_{\mathbf{s}}(\mathbf{s}) \end{pmatrix} \quad . \quad (2.157)$$

In particular, over the $p+q$ -dimensional manifold \mathbf{X} one then has the induced volume element

$$dv(\mathbf{r}, \mathbf{s}) = dv_{\mathbf{r}}(\mathbf{r}) dv_{\mathbf{s}}(\mathbf{s}) \quad , \quad (2.158)$$

where the ‘marginal’ volume elements $dv_{\mathbf{r}}(\mathbf{r})$ and $dv_{\mathbf{s}}(\mathbf{s})$ are those given in equations 2.119.

Consider now a probability distribution over \mathbf{X} , characterized by a volumetric probability $f(\mathbf{x}) = f(\mathbf{r}, \mathbf{s})$. It is *not* assumed that this volumetric probability factors as a product of a volumetric probability over \mathbf{R} by a volumetric probability over \mathbf{S} . Assuming that this probability is normalizable, we can write the equivalent expressions

$$P(\mathbf{X}) = \int_{\mathbf{x} \in \mathbf{X}} dv(\mathbf{x}) f(\mathbf{x}) = \int_{\mathbf{r} \in \mathbf{R}} dv_{\mathbf{r}}(\mathbf{r}) \left(\int_{\mathbf{s} \in \mathbf{S}} dv_{\mathbf{s}}(\mathbf{s}) f(\mathbf{r}, \mathbf{s}) \right) = \int_{\mathbf{s} \in \mathbf{S}} dv_{\mathbf{s}}(\mathbf{s}) \left(\int_{\mathbf{r} \in \mathbf{R}} dv_{\mathbf{r}}(\mathbf{r}) f(\mathbf{r}, \mathbf{s}) \right) \quad . \quad (2.159)$$

Defining the two *marginal volumetric probabilities*

$$\boxed{f_{\mathbf{r}}(\mathbf{r}) = \int_{\mathbf{s} \in \mathbf{S}} dv_{\mathbf{s}}(\mathbf{s}) f(\mathbf{r}, \mathbf{s}) \quad ; \quad f_{\mathbf{s}}(\mathbf{s}) = \int_{\mathbf{r} \in \mathbf{R}} dv_{\mathbf{r}}(\mathbf{r}) f(\mathbf{r}, \mathbf{s})} \quad (2.160)$$

this can be written

$$P(\mathbf{X}) = \int_{\mathbf{x} \in \mathbf{X}} dv(\mathbf{x}) f(\mathbf{x}) = \int_{\mathbf{r} \in \mathbf{R}} dv_{\mathbf{r}}(\mathbf{r}) f_{\mathbf{r}}(\mathbf{r}) = \int_{\mathbf{s} \in \mathbf{S}} dv_{\mathbf{s}}(\mathbf{s}) f_{\mathbf{s}}(\mathbf{s}) \quad . \quad (2.161)$$

It is clear that the marginal volumetric probability $f_{\mathbf{r}}(\mathbf{r})$ defines a probability over \mathbf{R} , while the marginal volumetric probability $f_{\mathbf{s}}(\mathbf{s})$ defines a probability over \mathbf{S} .

2.5.3 Interpretation of Marginal Volumetric Probability

These definitions can be intuitively interpreted as follows. Assume that there is a volumetric probability $f(\mathbf{x}) = f(\mathbf{r}, \mathbf{s})$ defined over a space \mathbf{X} that is the Cartesian product of two spaces \mathbf{R} and \mathbf{S} , in the sense just explained.

A sampling of the (probability distribution over \mathbf{X} associated to the) ‘joint’ volumetric probability f would produce points (of \mathbf{X})

$$\mathbf{x}_1 = (\mathbf{r}_1, \mathbf{s}_1) \quad , \quad \mathbf{x}_2 = (\mathbf{r}_2, \mathbf{s}_2) \quad , \quad \mathbf{x}_3 = (\mathbf{r}_3, \mathbf{s}_3) \quad , \quad \dots \quad . \quad (2.162)$$

Then,

- the points (of \mathbf{R}) $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \dots$ are samples of the (probability distribution over \mathbf{R} associated to the) marginal volumetric probability $f_{\mathbf{r}}$; and
- the points (of \mathbf{S}) $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \dots$ are samples of the (probability distribution over \mathbf{S} associated to the) marginal volumetric probability $f_{\mathbf{s}}$.

Thus, if when working with a Cartesian product of two manifolds $\mathbf{X} = \mathbf{R} \times \mathbf{S}$, and facing a ‘joint’ volumetric probability $f(\mathbf{r}, \mathbf{s})$, one is only interested in the probability properties induced by $f(\mathbf{r}, \mathbf{s})$ over \mathbf{R} (respectively over \mathbf{S}) one only needs to consider the marginal volumetric probability $f_{\mathbf{r}}(\mathbf{r})$ (respectively $f_{\mathbf{s}}(\mathbf{s})$). This, of course, implies that one is *not* interested in the possible dependences between the variables \mathbf{r} and the variables \mathbf{s} .

2.5.4 Bayes Theorem

Let us continue to work in the special situation where the n -dimensional space \mathbf{X} is defined as the Cartesian product of two spaces, $\mathbf{X} = \mathbf{R} \times \mathbf{S}$, with respective dimensions p and q , with $n = p + q$. Given a ‘joint’ volumetric probability $f(\mathbf{r}, \mathbf{s})$ over \mathbf{X}_n , we have defined two marginal volumetric probabilities $f_{\mathbf{r}}(\mathbf{r})$ and $f_{\mathbf{s}}(\mathbf{s})$ using equations 2.160.

We have also written, for any fixed value of \mathbf{s} (equation 2.137 dropping the index ‘0’)

$$f(\mathbf{r}|\mathbf{s}) = \frac{f(\mathbf{r}, \mathbf{s})}{\int_{\mathbf{R}} dv_{\mathbf{r}}(\mathbf{r}) f(\mathbf{r}, \mathbf{s})} \quad , \quad = \quad f(\mathbf{r}|\mathbf{s}) = \frac{f(\mathbf{r}, \mathbf{s})}{f_{\mathbf{s}}(\mathbf{s})} \quad , \quad (2.163)$$

where, in the second equality we have used the definition of marginal volumetric probability. It follows

$$f(\mathbf{r}, \mathbf{s}) = f(\mathbf{r}|\mathbf{s}) f_{\mathbf{s}}(\mathbf{s}) \quad , \quad (2.164)$$

equation that can be read as saying bla, bla, bla ... Similarly,

$$f(\mathbf{r}, \mathbf{s}) = f(\mathbf{s}|\mathbf{r}) f_{\mathbf{r}}(\mathbf{r}) \quad , \quad (2.165)$$

and comparing these two equations we deduce the well known *Bayes theorem*

$$\boxed{f(\mathbf{r}|\mathbf{s}) = \frac{f(\mathbf{s}|\mathbf{r}) f_{\mathbf{r}}(\mathbf{r})}{f_{\mathbf{s}}(\mathbf{s})} \quad ,} \quad (2.166)$$

equation that can be read as saying bla, bla, bla ...

Note: explain here again that the assumption that the metric of the space \mathbf{X} takes the form expressed in equation 2.157 is fundamental.

2.5.5 Independent Probability Distributions

Assume again that there is a volumetric probability $f(\mathbf{x}) = f(\mathbf{r}, \mathbf{s})$ defined over a space \mathbf{X} that is the Cartesian product of two spaces \mathbf{R} and \mathbf{S} , in the sense being considered. Then, one may define the marginal volumetric probabilities $f_{\mathbf{r}}(\mathbf{r})$ and $f_{\mathbf{s}}(\mathbf{s})$ defined by equation 2.160.

If it happens that the ‘joint’ volumetric probability $f(\mathbf{r}, \mathbf{s})$ is just the product of the two marginal probability distributions,

$$\boxed{f(\mathbf{r}, \mathbf{s}) = f_{\mathbf{r}}(\mathbf{r}) f_{\mathbf{s}}(\mathbf{s}) \quad ,} \quad (2.167)$$

it is said that the probability distributions over \mathbf{R} and \mathbf{S} (as characterized by the marginal volumetric probabilities $f_{\mathbf{r}}(\mathbf{r})$ and $f_{\mathbf{s}}(\mathbf{s})$) are *independent*.

Note: the comparison of this definition with equations 2.164–2.165 shows that, in this case,

$$f(\mathbf{r}|\mathbf{s}) = f_{\mathbf{r}}(\mathbf{r}) \quad ; \quad f(\mathbf{s}|\mathbf{r}) = f_{\mathbf{s}}(\mathbf{s}) \quad , \quad (2.168)$$

from where the ‘independence’ notion can be understood (note: explain this).

NOTE: REFRESH THE EXAMPLE BELOW.

Example 2.10 *Over the surface of the unit sphere, using geographical coordinates, we have the two displacement elements*

$$ds_{\varphi}(\varphi, \lambda) = \cos \lambda d\varphi \quad ; \quad ds_{\lambda}(\varphi, \lambda) = d\lambda \quad , \quad (2.169)$$

with the associated surface element (as the coordinates are orthogonal) $ds(\varphi, \lambda) = \cos \lambda d\varphi d\lambda$. Consider a (2D) volumetric probability $f(\varphi, \lambda)$ over the surface of the sphere, normed under the usual condition

$$\int_{\text{surface}} ds(\varphi, \lambda) f(\varphi, \lambda) = \int_{-\pi}^{+\pi} d\varphi \int_{-\pi/2}^{+\pi/2} d\lambda \cos \lambda f(\varphi, \lambda) = \int_{-\pi/2}^{+\pi/2} d\lambda \cos \lambda \int_{-\pi}^{+\pi} d\varphi f(\varphi, \lambda) = 1 \quad . \quad (2.170)$$

One may define the partial integrations

$$\eta_\varphi(\varphi) = \int_{-\pi/2}^{+\pi/2} d\lambda \cos \lambda f(\varphi, \lambda) \quad ; \quad \eta_\lambda(\lambda) = \int_{-\pi}^{+\pi} d\varphi f(\varphi, \lambda) \quad , \quad (2.171)$$

so that the probability of a sector between two meridians and of an annulus between two parallels are respectively computed as

$$P(\varphi_1 < \varphi < \varphi_2) = \int_{\varphi_1}^{\varphi_2} d\varphi \eta_\varphi(\varphi) \quad ; \quad P(\lambda_1 < \lambda < \lambda_2) = \int_{\lambda_1}^{\lambda_2} d\lambda \cos \lambda \eta_\lambda(\lambda) \quad , \quad (2.172)$$

but the terms $d\varphi$ and $\cos \lambda d\lambda$ appearing in these two expressions are not the displacement elements on the sphere's surface (equation 2.169). The functions $\eta_\varphi(\varphi)$ and $\eta_\lambda(\lambda)$ should not be mistaken as marginal volumetric probabilities: as the surface of the sphere is not the Cartesian product of two 1D spaces, marginal volumetric probabilities are not defined. [**End of example.**]

2.6 Transport of Probabilities

2.6.0.1 The Problem

We are contemplating:

- a p -dimensional metric space \mathbb{R}_p , with coordinates $\mathbf{r} = \{r^\alpha\}$, and a metric matrix that, in these coordinates, is \mathbf{g}_r ;
- a q -dimensional metric space \mathbb{S}_q , with coordinates $\mathbf{s} = \{s^i\}$, and a metric matrix that, in these coordinates, is \mathbf{g}_s ;
- an application $\mathbf{s} = \boldsymbol{\sigma}(\mathbf{r})$ from \mathbb{R}_p into \mathbb{S}_q .

To any volumetric probability $f_r(\mathbf{r})$ over \mathbb{R}_p , the application

$$\mathbf{s} = \boldsymbol{\sigma}(\mathbf{r}) \quad (2.173)$$

associates a unique volumetric probability $f_s(\mathbf{s})$ over \mathbb{S}_q . To intuitively understand this, consider a large collection of samples of $f_r(\mathbf{r})$, say $\{\mathbf{r}_1, \mathbf{r}_2, \dots\}$. To each of these points in \mathbb{R}_p we can associate a unique point in \mathbb{S}_q , via $\mathbf{s} = \boldsymbol{\sigma}(\mathbf{r})$, so we have a large collection of points $\{\mathbf{s}_1, \mathbf{s}_2, \dots\}$ in \mathbb{S}_q . Of which volumetric probability $f_s(\mathbf{s})$ are these points samples?

Although the major inference problems considered in this book (conditional probability, product of probabilities, etc.) are only defined when the considered spaces are metric, this problem of transport of probabilities makes perfect sense even if the spaces do not have a metric. For this reason, one could set the problem of transportation of a probability distribution in terms of probability densities, instead of volumetric probabilities. I prefer to use the metric concepts and language, but shall also give below the equivalent formulas for those who may choose to work with volumetric probabilities.

In what follows, \mathbf{S} denotes the matrix of partial derivatives

$$S^i_\alpha = \frac{\partial s^i}{\partial r^\alpha} \quad (2.174)$$

Note: write somewhere what follows:

As we have represented by \mathbf{g}_r the metric in the space \mathbb{R}_p , the volume element is given by the usual expression

$$\boxed{dv_r(\mathbf{r}) = \sqrt{\det \mathbf{g}_r(\mathbf{r})} dr^1 \wedge \dots \wedge dr^p,} \quad (2.175)$$

the volume of a finite region \mathcal{A} being computed via

$$V(\mathcal{A}) = \int_{\mathcal{A}} dv_r(\mathbf{r}) = \int_{\mathcal{A}} dr^1 \wedge \dots \wedge dr^p \sqrt{\det \mathbf{g}_r(\mathbf{r})} \quad (2.176)$$

2.6.0.2 Case $p \leq q$

When $p \leq q$, the p -dimensional manifold \mathbf{R}_p is mapped, via $\mathbf{s} = \mathbf{s}(\mathbf{r})$, into a p -dimensional submanifold of \mathbf{S}_q , say \mathbf{S}_p (see figure 2.16). In that submanifold we can use as coordinates the coordinates induced from the coordinates \mathbf{r} of \mathbf{R}_p via $\mathbf{s} = \mathbf{s}(\mathbf{r})$. So, now, the coordinates \mathbf{r} define, at the same time, a point of \mathbf{R}_p and a point of $\mathbf{S}_p \subset \mathbf{S}_q$ (if the points of \mathbf{S}_q are covered more than once by the application $\mathbf{s} = \mathbf{s}(\mathbf{r})$, then, let us assume that we work inside a subdomain of \mathbf{R}_p where the problem does not exist). Note: I should mention here figure 2.19.

The application $\mathbf{s} = \mathbf{s}(\mathbf{r})$ maps the p -dimensional volume element $dv_{\mathbf{r}}$ on \mathbf{R}_p into a p -dimensional volume element $dv_{\mathbf{s}}$ on the submanifold \mathbf{S}_p of \mathbf{S}_q . Let us characterize it.

The distance element between two points in \mathbf{S}_q is $ds^2 = (\mathbf{g}_{\mathbf{s}})_{ij} ds^i ds^j$. If, in fact, those are points of \mathbf{S}_p , then we can write $ds^i = S^i_{\alpha} dr^{\alpha}$, to obtain $ds^2 = G_{\alpha\beta} dr^{\alpha} dr^{\beta}$, where $\mathbf{G} = \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S}$ (remember that we can use \mathbf{r} as coordinates over the submanifold \mathbf{S}_p of \mathbf{S}_q). The p -dimensional volume element obtained on \mathbf{S}_p by transportation of the volume element $dv_{\mathbf{r}}$ of \mathbf{R}_p (via $\mathbf{s} = \mathbf{s}(\mathbf{r})$) is

$$\boxed{dv_{\mathbf{s}}(\mathbf{r}) = \sqrt{\det \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S}} dr^1 \wedge \cdots \wedge dr^p}, \quad (2.177)$$

where $\mathbf{g}_{\mathbf{s}} = \mathbf{g}_{\mathbf{s}}(\mathbf{s}(\mathbf{r}))$ and $\mathbf{S} = \mathbf{S}(\mathbf{r})$. The volume of a finite region \mathcal{A} of \mathbf{S}_p is computed via

$$V(\mathcal{A}) = \int_{\mathcal{A}} dv_{\mathbf{s}}(\mathbf{r}) = \int_{\mathcal{A}} dr^1 \wedge \cdots \wedge dr^p \sqrt{\det \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S}}. \quad (2.178)$$

Note that by comparing equations 2.175 and 2.177 we obtain the ratio of the volumes,

$$\frac{dv_{\mathbf{s}}}{dv_{\mathbf{r}}} = \frac{\sqrt{\det \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S}}}{\sqrt{\det \mathbf{g}_{\mathbf{r}}}}. \quad (2.179)$$

We have seen that bla, bla, bla, and we have the same coordinates over the two spaces, and bla, bla, bla, and to a common capacity element $dr^1 \wedge \cdots \wedge dr^p$ corresponds the two volume elements

$$\begin{aligned} dv_{\mathbf{r}}(\mathbf{r}) &= \sqrt{\det \mathbf{g}_{\mathbf{r}}} dr^1 \wedge \cdots \wedge dr^p \\ dv_{\mathbf{s}}(\mathbf{r}) &= \sqrt{\det \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S}} dr^1 \wedge \cdots \wedge dr^p. \end{aligned} \quad (2.180)$$

When there are volumetric probabilities $f_{\mathbf{r}}(\mathbf{r})$ and $f_{\mathbf{s}}(\mathbf{r})$, they are defined so as to have

$$\begin{aligned} dP_{\mathbf{r}} &= f_{\mathbf{r}} dv_{\mathbf{r}} \\ dP_{\mathbf{s}} &= f_{\mathbf{s}} dv_{\mathbf{s}}. \end{aligned} \quad (2.181)$$

We say the the volumetric probability $f_{\mathbf{s}}$ has been ‘transported’ from $f_{\mathbf{r}}$ if the two probabilities associated to the two volumes defined by the common capacity element $dr^1 \wedge \cdots \wedge dr^p$ are identical, i.e., if $dP_{\mathbf{r}} = dP_{\mathbf{s}}$. It follows the relation $f_{\mathbf{s}} = \frac{dv_{\mathbf{r}}}{dv_{\mathbf{s}}} f_{\mathbf{r}}$, i.e.,

$$\boxed{f_{\mathbf{s}} = \frac{\sqrt{\det \mathbf{g}_{\mathbf{r}}}}{\sqrt{\det \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S}}} f_{\mathbf{r}}}, \quad (2.182)$$

or, more explicitly,

$$f_{\mathbf{s}}(\mathbf{r}) = \frac{\sqrt{\det \mathbf{g}_{\mathbf{r}}(\mathbf{r})}}{\sqrt{\det \mathbf{S}^t(\mathbf{r}) \mathbf{g}_{\mathbf{s}}(\mathbf{s}(\mathbf{r})) \mathbf{S}(\mathbf{r})}} f_{\mathbf{r}}(\mathbf{r}). \quad (2.183)$$

The matrix \mathbf{S} of partial derivatives has dimension $(q \times p)$, and unless $p = q$, it is not a squared matrix. This implies that, in general, $\det \mathbf{S}^t \mathbf{g}_s \mathbf{S} \neq \det(\mathbf{S}^t \mathbf{S}) \det \mathbf{g}_s$.

While the probability of a domain \mathcal{A} of \mathbb{R}_p is to be evaluated as

$$P_{\mathbf{r}}(\mathcal{A}) = \int_{\mathbf{r} \in \mathcal{A}} dr^1 \wedge \cdots \wedge dr^p \sqrt{\det \mathbf{g}_{\mathbf{r}}} f_{\mathbf{r}}(\mathbf{r}) \quad , \quad (2.184)$$

the probability of its image $\mathbf{s}(\mathcal{A})$ (that, by definition is identical to the probability of \mathcal{A}), is to be evaluated as

$$P_{\mathbf{s}}(\mathbf{s}(\mathcal{A})) = P_{\mathbf{r}}(\mathcal{A}) = \int_{\mathbf{r} \in \mathcal{A}} dr^1 \wedge \cdots \wedge dr^p \sqrt{\det \mathbf{S}^t \mathbf{g}_s \mathbf{S}} f_{\mathbf{s}}(\mathbf{r}) \quad . \quad (2.185)$$

Of course, one could introduce the probability densities

$$\bar{f}_{\mathbf{r}}(\mathbf{r}) = \sqrt{\det \mathbf{g}_{\mathbf{r}}} f_{\mathbf{r}}(\mathbf{r}) \quad ; \quad \bar{f}_{\mathbf{s}}(\mathbf{r}) = \sqrt{\det \mathbf{S}^t \mathbf{g}_s \mathbf{S}} f_{\mathbf{s}}(\mathbf{r}) \quad . \quad (2.186)$$

Using them, the integrations 2.184–2.185 would formally simplify into

$$P_{\mathbf{r}} = \int_{\mathbf{r} \in \mathcal{A}} dr^1 \wedge \cdots \wedge dr^p \bar{f}_{\mathbf{r}}(\mathbf{r}) \quad , \quad ; \quad P_{\mathbf{s}} = \int_{\mathbf{r} \in \mathcal{A}} dr^1 \wedge \cdots \wedge dr^p \bar{f}_{\mathbf{s}}(\mathbf{r}) \quad , \quad (2.187)$$

while the relation 2.183 would trivialize into

$$\bar{f}_{\mathbf{s}}(\mathbf{r}) = \bar{f}_{\mathbf{r}}(\mathbf{r}) \quad . \quad (2.188)$$

There is no harm in using equations 2.187–2.188 in analytical developments (I have already mentioned that for numerical integrations is much better to use volume elements, rather than capacity elements, and volumetric probabilities rather than probability densities), provided one remembers that the volume of a domain \mathcal{A} of \mathbb{R}_p is to be evaluated as

$$V(\mathcal{A}) = \int_{\mathbf{r} \in \mathcal{A}} dr^1 \wedge \cdots \wedge dr^p \sqrt{\det \mathbf{g}_{\mathbf{r}}} \quad , \quad (2.189)$$

while the volume of its image $\mathbf{s}(\mathcal{A})$ (of course, different from that of \mathcal{A}) is to be evaluated as

$$V(\mathbf{s}(\mathcal{A})) = \int_{\mathbf{r} \in \mathcal{A}} dr^1 \wedge \cdots \wedge dr^p \sqrt{\det \mathbf{S}^t \mathbf{g}_s \mathbf{S}} \quad . \quad (2.190)$$

2.6.0.3 Case $p \geq q$

Let us now consider the case $p \geq q$, i.e., when the ‘starting space’ has larger (or equal) dimension than the ‘arrival space’.

Let us begin by choosing over \mathbb{R}_p a new system of coordinates specially adapted to the problem. Remember that we are using Latin indices for the coordinates s^i , where $1 \leq i \leq q$, and Greek indices for the coordinates r^α , where $1 \leq \alpha \leq p$. We pass from the p coordinates \mathbf{r} to the new p coordinates

$$\begin{aligned} s^i &= s^i(\mathbf{r}) & ; & & (1 \leq i \leq q) \\ t^A &= t^A(\mathbf{r}) & ; & & (q+1 \leq A \leq p) \quad , \end{aligned} \quad (2.191)$$

Figure 2.16: Transporting a volume element from a p -dimensional space \mathbb{R}_p into a q -dimensional space \mathbf{S}_q , via an expression $\mathbf{s} = \mathbf{s}(\mathbf{r})$. Left: $1 = p < q = 2$; in this case, we start with a p -dimensional volume in \mathbb{R}_p and arrive at \mathbf{S}_q with a volume of same dimension (equations 2.175 and 2.177). Right: $2 = p > q = 1$: in this case the we start with a p -dimensional volume in \mathbb{R}_p but arrive at \mathbf{S}_q with a q -dimensional volume, i.e., a volume of lower dimension (equations 2.176 and ??).

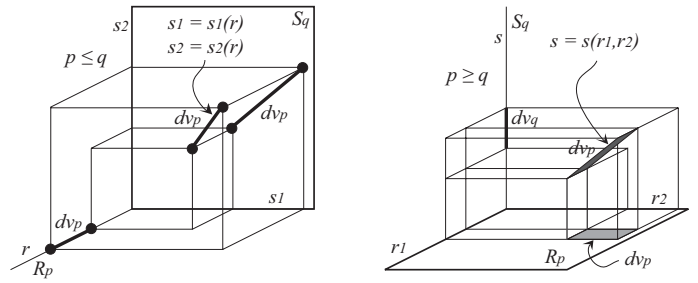
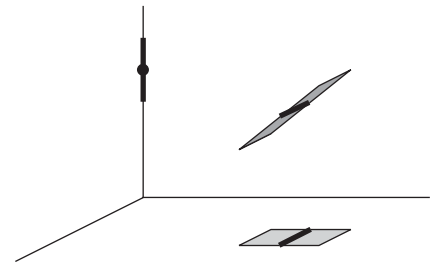


Figure 2.17: Detail of figure 2.16, showing a domain of \mathbb{R}_p that maps into a single point of \mathbf{S}_q .



where the functions σ^i are the same as those appearing in equation 2.173 (i.e., the q coordinates \mathbf{s} of \mathbf{S}_q are used as q of the p coordinates of \mathbb{R}_p), and where the functions τ^A are arbitrary (one could, for instance, choose $t^A = r^A$, for $q + 1 \leq A \leq p$).

It may well happen that the coordinates $\{\mathbf{s}, \mathbf{t}\}$ are only regular inside distinct regions of \mathbb{R}_p . Let us work inside one such region, letting the ad-hoc management of the more general situation be just suggested in figure 2.19.

We need to express the metric tensor in the new coordinates, and, for this, we must introduce the (Jacobian) matrix \mathbf{K} of partial derivatives

$$\{K^\alpha_\beta\} = \left\{ \begin{matrix} S^i_\beta \\ T^A_\beta \end{matrix} \right\} = \left\{ \begin{matrix} \partial s^i / \partial r^\beta \\ \partial t^A / \partial r^\beta \end{matrix} \right\} , \quad (2.192)$$

and its inverse

$$\mathbf{L} = \mathbf{K}^{-1} . \quad (2.193)$$

Using \mathbf{L} , the matrix representing the metric tensor of the space \mathbb{R}_p in the new coordinates is (see, for instance, equation 1.23)

$$\mathbf{G} = \mathbf{L}^t \mathbf{g}_r \mathbf{L} , \quad (2.194)$$

while, in terms of the matrix \mathbf{K} , equivalently,

$$\mathbf{G}^{-1} = \mathbf{K} \mathbf{g}_r^{-1} \mathbf{K}^t . \quad (2.195)$$

Note: I have to say here that, as the matrices \mathbf{K} and \mathbf{L} are invertible,

$$\sqrt{\det \mathbf{G}} = L \sqrt{\det \mathbf{g}_r} = \frac{1}{K} \sqrt{\det \mathbf{g}_r} , \quad (2.196)$$

where

$$L = \sqrt{\mathbf{L}^t \mathbf{L}} \quad ; \quad K = \sqrt{\mathbf{K} \mathbf{K}^t} \quad . \quad (2.197)$$

[Note: Emphasize here that when only the determinant of the metric appears, and not the full metric, this means that we only need a volume element over the space, not a distance element. Important to solve the problem of ‘relative weights’.]

The definition of volumetric probability that we have used makes it an invariant. The relation between a volumetric probability $f_{\mathbf{r}}(\mathbf{r})$, expressed in the coordinates \mathbf{r} and the equivalent volumetric probability $f(\mathbf{s}, \mathbf{t})$, expressed in the coordinates $\{\mathbf{s}, \mathbf{t}\}$, is, simply,

$$f_{\mathbf{r}}(\mathbf{r}) = f(\mathbf{s}(\mathbf{r}), \mathbf{t}(\mathbf{r})) \quad . \quad (2.198)$$

In the coordinates \mathbf{r} , the probability of a region of \mathbf{R}_p is computed as

$$P_{\mathbf{r}} = \int dr^1 \wedge \cdots \wedge dr^p \sqrt{\det \mathbf{g}_{\mathbf{r}}} f_{\mathbf{r}}(\mathbf{r}) \quad . \quad (2.199)$$

In the coordinates $\{\mathbf{s}, \mathbf{t}\}$, it is computed using the equation

$$P_{\mathbf{r}} = \int ds^1 \wedge \cdots \wedge ds^q \int dt^{q+1} \wedge \cdots \wedge dt^p \sqrt{\det \mathbf{G}} f(\mathbf{s}, \mathbf{t}) \quad , \quad (2.200)$$

where \mathbf{G} is given by equation 2.194. While this expression defines the probability of an arbitrary region of \mathbf{R}_p , the expression

$$P_{\mathbf{s}} = \int ds^1 \wedge \cdots \wedge ds^q \int_{\text{all } \mathbf{t}} dt^{q+1} \wedge \cdots \wedge dt^p \sqrt{\det \mathbf{G}} f(\mathbf{s}, \mathbf{t}) \quad , \quad (2.201)$$

where the first sum is taken over an arbitrary domain of the coordinates \mathbf{s} , but the second sum is now taken for all possible values of the coordinates \mathbf{t} , corresponds to the probability of a region of \mathbf{S}_q (as the coordinates \mathbf{s} are not only some of the coordinates of \mathbf{R}_p , but are also the coordinates over \mathbf{S}_q).

As a volumetric probability $f_{\mathbf{s}}(\mathbf{s})$ over \mathbf{S}_q is to be integrated via

$$P_{\mathbf{s}} = \int ds^1 \wedge \cdots \wedge ds^q \sqrt{\det \mathbf{g}_{\mathbf{s}}} f_{\mathbf{s}}(\mathbf{s}) \quad , \quad (2.202)$$

then, by comparison with equation 2.201, we deduce that the expression representing the volumetric probability we wished to characterize is

$$f_{\mathbf{s}}(\mathbf{s}) = \frac{1}{\sqrt{\det \mathbf{g}_{\mathbf{s}}}} \int_{\text{all } \mathbf{t}} dt^{q+1} \wedge \cdots \wedge dt^p \sqrt{\det \mathbf{G}} f(\mathbf{s}, \mathbf{t}) \quad , \quad (2.203)$$

and our problem, essentially, solved. Note that, here, the volumetric probability appears with the variables \mathbf{s} and \mathbf{t} , while the original volumetric probability was $f_{\mathbf{r}}(\mathbf{r})$. Although the two expressions are linked through equation 2.198, this is not enough to actually have the expression of $f(\mathbf{s}, \mathbf{t})$. This requires that we solve the change of variables 2.191, to obtain the relations

$$\mathbf{r} = \mathbf{r}(\mathbf{s}, \mathbf{t}) \quad , \quad (2.204)$$

so we can write

$$f(\mathbf{s}, \mathbf{t}) = f_{\mathbf{r}}(\mathbf{r}(\mathbf{s}, \mathbf{t})) \quad . \quad (2.205)$$

Explicitly, using equation 2.196, the volumetric probability $f_{\mathbf{s}}(\mathbf{s})$ can be written

$$f_{\mathbf{s}}(\mathbf{s}) = \frac{1}{\sqrt{\det \mathbf{g}_{\mathbf{s}}}} \int_{\text{all } \mathbf{t}} dt^{q+1} \wedge \dots \wedge dt^p \underbrace{\frac{\sqrt{\det \mathbf{g}_{\mathbf{r}}(\mathbf{r})}}{K(\mathbf{r})} f_{\mathbf{r}}(\mathbf{r})}_{\mathbf{r}=\mathbf{r}(\mathbf{s}, \mathbf{t})} \quad . \quad (2.206)$$

As the probability densities associated to the volumetric probabilities $f_{\mathbf{s}}$ and $f_{\mathbf{r}}$ are

$$\bar{f}_{\mathbf{s}} = \sqrt{\det \mathbf{g}_{\mathbf{s}}} f_{\mathbf{s}} \quad ; \quad \bar{f}_{\mathbf{r}} = \sqrt{\det \mathbf{g}_{\mathbf{r}}} f_{\mathbf{r}} \quad , \quad (2.207)$$

equation 2.206 can also be written

$$\bar{f}_{\mathbf{s}}(\mathbf{s}) = \int_{\text{all } \mathbf{t}} dt^{q+1} \wedge \dots \wedge dt^p \frac{1}{K(\mathbf{r}(\mathbf{s}, \mathbf{t}))} \bar{f}_{\mathbf{r}}(\mathbf{r}(\mathbf{s}, \mathbf{t})) \quad , \quad (2.208)$$

an expression that is independent of the metrics in the spaces \mathbf{R}_p and \mathbf{S}_q .

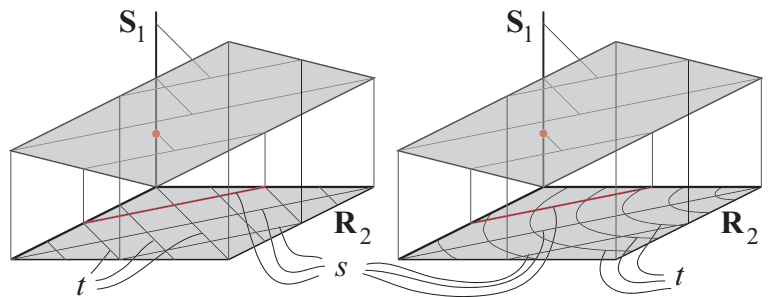
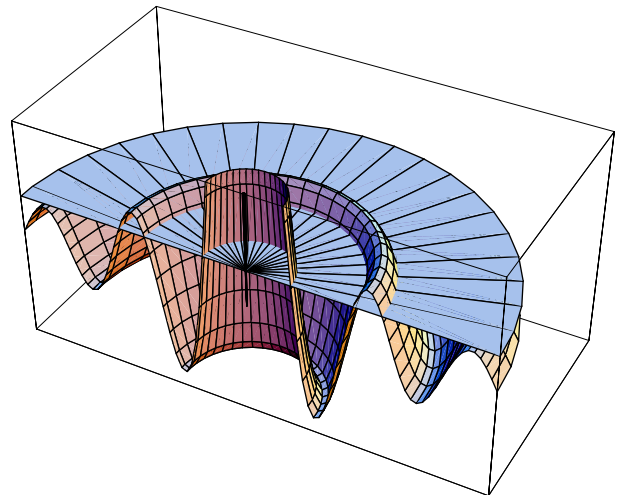


Figure 2.18: Lines that map into a same value of \mathbf{s} . Two different choices for the variables \mathbf{t} .

Figure 2.19: Consider that we have a mapping from the Euclidean plane, with polar coordinates $\mathbf{r} = \{\rho, \varphi\}$, into a one-dimensional space with a metric coordinate s (in this illustration, $s = s(\rho, \varphi) = \sin \rho / \rho$). When transporting a probability from the plane into the ‘vertical axis’, for a given value of $s = s_0$ we have, first, to obtain the set of discrete values ρ_n giving the same s_0 , and, for each of these values, we have to perform the integration for $-\pi < \varphi \leq +\pi$ corresponding to that indicated in equations ??–2.206.



Example 2.11 A one-dimensional material medium with an initial length X is deformed into a second state, where its length is Y . The strain that has affected the medium, denoted ε , is defined as

$$\varepsilon = \log \frac{Y}{X} . \quad (2.209)$$

A measurement of X and Y provides the information represented by a volumetric probability $f_{\mathbf{r}}(Y, X)$. This induces an information on the actual value of the strain, that shall be represented by a volumetric probability $f_{\mathbf{s}}(\varepsilon)$. The problem is to express $f_{\mathbf{s}}(\varepsilon)$ using as ‘inputs’ the definition 2.209 and the volumetric probability $f_{\mathbf{r}}(Y, X)$. Let us introduce the two-dimensional ‘data’ space \mathbf{R}_2 , over which the quantities X and Y are coordinates. The lengths X and Y being Jeffreys quantities (see discussion in section XXX), we have, in the space \mathbf{R}_2 , the distance element $ds_{\mathbf{r}}^2 = \left(\frac{dY}{Y}\right)^2 + \left(\frac{dX}{X}\right)^2$, associated to the metric matrix

$$\mathbf{g}_{\mathbf{r}} = \begin{pmatrix} \frac{1}{Y^2} & 0 \\ 0 & \frac{1}{X^2} \end{pmatrix} . \quad (2.210)$$

This, in particular, gives

$$\sqrt{\det \mathbf{g}_{\mathbf{r}}} = \frac{1}{YX} , \quad (2.211)$$

so the (2D) volume element over \mathbf{R}_2 is $dv_{\mathbf{r}} = \frac{dY \wedge dX}{YX}$, and any volumetric probability $f_{\mathbf{r}}(Y, X)$ over \mathbf{R}_2 is to be integrated via

$$P_{\mathbf{r}} = \int dY \wedge dX \frac{1}{YX} f_{\mathbf{r}}(Y, X) , \quad (2.212)$$

over the appropriate bounds. In particular, a volumetric probability $f_{\mathbf{r}}(Y, X)$ is normalized if the integral over $(0 < Y < \infty ; 0 < X < \infty)$ equals one. Let us also introduce the one-dimensional ‘space of deformations’ \mathbf{S}_1 , over which the quantity ε is the chosen coordinate (one could as well chose the exponential of ε , or twice the strain as coordinate). The strain being an ordinary Cartesian coordinate, we have, in the space of deformations \mathbf{S}_1 the distance element $ds_{\mathbf{s}}^2 = d\varepsilon^2$, associated to the trivial metric matrix $\mathbf{g}_{\mathbf{s}} = (1)$. Therefore,

$$\sqrt{\det \mathbf{g}_{\mathbf{s}}} = 1 . \quad (2.213)$$

The (1D) volume element over \mathbf{S}_1 is $dv_{\mathbf{s}} = d\varepsilon$, and any volumetric probability $f_{\mathbf{s}}(\varepsilon)$ over \mathbf{S}_1 is to be integrated via

$$P_{\mathbf{s}} = \int d\varepsilon f_{\mathbf{s}}(\varepsilon) , \quad (2.214)$$

over given bounds. A volumetric probability $f_{\mathbf{s}}(\varepsilon)$ is normalized by the condition that the integral over $(-\infty < \varepsilon < +\infty)$ equals one. As suggested in the general theory, we must change the coordinates in \mathbf{R}_2 using as part of the coordinates those of \mathbf{S}_1 , i.e., here, using the strain ε . Then, arbitrarily, select X as second coordinate, so we pass in \mathbf{R}_2 from the coordinates $\{Y, X\}$ to the coordinates $\{\varepsilon, X\}$. Then, the Jacobian matrix defined in equation 2.192 is

$$\mathbf{K} = \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} = \begin{pmatrix} \partial\varepsilon/\partial Y & \partial\varepsilon/\partial X \\ \partial X/\partial Y & \partial X/\partial X \end{pmatrix} = \begin{pmatrix} 1/Y & -1/X \\ 0 & 1 \end{pmatrix} , \quad (2.215)$$

and we obtain, using the metric 2.210,

$$\sqrt{\det \mathbf{K} \mathbf{g}_r^{-1} \mathbf{K}^t} = X \quad . \quad (2.216)$$

Noting that the expression 2.209 can trivially be solved for Y as

$$Y = X \exp \varepsilon \quad , \quad (2.217)$$

everything is ready now to attack the problem. If a measurement of X and Y has produced the information represented by the volumetric probability $f_r(Y, X)$, this transports into a volumetric probability $f_s(\varepsilon)$ that is given by equation 2.206. Using the particular expressions 2.213, 2.216 and 2.217 this gives

$$f_s(\varepsilon) = \int_0^\infty dX \frac{1}{X} f_r(X \exp \varepsilon, X) \quad . \quad (2.218)$$

[End of example.]

Example 2.12 In the context of the previous example, assume that the measurement of the two lengths X and Y has provided an information on their actual values that: (i) has independent uncertainties and (ii) is Gaussian (which, as indicated in section 2.8.4, means that the dependence of the volumetric probability on the Jeffreys quantities X and Y is expressed by the lognormal function). Then we have

$$f_X(X) = \frac{1}{\sqrt{2\pi} s_X} \exp \left(-\frac{1}{2 s_X^2} \left(\log \frac{X}{X_0} \right)^2 \right) \quad , \quad (2.219)$$

$$f_Y(Y) = \frac{1}{\sqrt{2\pi} s_Y} \exp \left(-\frac{1}{2 s_Y^2} \left(\log \frac{Y}{Y_0} \right)^2 \right) \quad (2.220)$$

and

$$f_r(Y, X) = f_Y(Y) f_X(X) \quad . \quad (2.221)$$

The volumetric probability for X is centered at point X_0 , with standard deviation s_X , and the volumetric probability for Y is centered at point Y_0 , with standard deviation s_Y (see section 2.7 for a precise —invariant— definition of standard deviation). In this simple example, the integration in equation 2.218 can be performed analytically, and one obtains a Gaussian probability distribution for the strain, represented by the normal function

$$f_s(\varepsilon) = \frac{1}{\sqrt{2\pi} s_\varepsilon} \exp \left(-\frac{(\varepsilon - \varepsilon_0)^2}{2 s_\varepsilon^2} \right) \quad , \quad (2.222)$$

where ε_0 , the center of the probability distribution for the strain, equals the logarithm of the ratio of the centers of the probability distributions for the lengths,

$$\varepsilon_0 = \log \frac{Y_0}{X_0} \quad , \quad (2.223)$$

and where s_ε^2 , the variance of the probability distribution for the strain, equals the sum of the variances of the probability distributions for the lengths,

$$s_\varepsilon^2 = s_X^2 + s_Y^2 \quad . \quad (2.224)$$

[End of example.]

2.6.0.4 Case $p = q$

The two cases examined above, $p \leq q$ and $p \geq q$, both contain the case $p = q$, but let us, to avoid possible misunderstandings, treat the case explicitly here.

In the case $p \leq q$, we have chosen to use over the subspace \mathbf{S}_p , image of \mathbf{S}_q through $\mathbf{s} = \boldsymbol{\sigma}(\mathbf{r})$, the image of the coordinates \mathbf{r} of \mathbf{R}_p , and, in these coordinates, we have found the expression 2.182

$$f_{\mathbf{s}}(\mathbf{r}) = \frac{\sqrt{\det \mathbf{g}_{\mathbf{r}}}}{\sqrt{\det \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S}}} f_{\mathbf{r}}(\mathbf{r}) \quad , \quad (2.225)$$

that is directly valid here. As the matrix \mathbf{S} is a squared matrix, we could further write

$$f_{\mathbf{s}}(\mathbf{r}) = \frac{\sqrt{\det \mathbf{g}_{\mathbf{r}}}}{\sqrt{\det \mathbf{g}_{\mathbf{s}}}} \frac{1}{S} f_{\mathbf{r}}(\mathbf{r}) \quad , \quad (2.226)$$

where $S = \det \mathbf{S}$.

In the case $p \geq q$, we have used over \mathbf{S}_q its own coordinates. Expression 2.206 drastically simplifies when $p = q$ (there are no variables \mathbf{t}), to give

$$f_{\mathbf{s}}(\mathbf{s}) = \frac{1}{\sqrt{\det \mathbf{g}_{\mathbf{s}}}} \frac{1}{\sqrt{\det \mathbf{S} \mathbf{g}_{\mathbf{r}}^{-1} \mathbf{S}^t}} f_{\mathbf{r}}(\mathbf{r}(\mathbf{s})) \quad , \quad (2.227)$$

or, as the matrix \mathbf{S} is now squared,

$$f_{\mathbf{s}}(\mathbf{s}) = \frac{\sqrt{\det \mathbf{g}_{\mathbf{r}}}}{\sqrt{\det \mathbf{g}_{\mathbf{s}}}} \frac{1}{S} f_{\mathbf{r}}(\mathbf{r}(\mathbf{s})) \quad . \quad (2.228)$$

This is, of course, the same expression than that in 2.226: we know that volumetric probabilities are invariant, and have the same value, at a given point, irrespectively of the coordinates being used.

Note that the expression $\mathbf{s} = \boldsymbol{\sigma}(\mathbf{r})$ is *not* defining a change of variables inside a given space: we have two different spaces, a \mathbf{R}_p with coordinates \mathbf{r} and a metric matrix $\mathbf{g}_{\mathbf{r}}$, and a \mathbf{S}_q with coordinates \mathbf{s} and metric matrix $\mathbf{g}_{\mathbf{s}}$. These two metric are totally independent, and the application $\mathbf{s} = \boldsymbol{\sigma}(\mathbf{r})$ is mapping points from \mathbf{R}_p into \mathbf{S}_q . If we were contemplating a change of variables inside a given space, then the metric matrices, instead of being independently given, would be related in the usual way tensors relate under a change of variables, $(\mathbf{g}_{\mathbf{r}})_{\alpha\beta} = \frac{\partial s^i}{\partial r^\alpha} (\mathbf{g}_{\mathbf{s}})_{ij} \frac{\partial s^j}{\partial r^\beta}$ i.e., for short,

$$\mathbf{g}_{\mathbf{r}} = \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S} \quad (\text{if we were considering a change of variables}) \quad . \quad (2.229)$$

In particular, then, $\sqrt{\det \mathbf{g}_{\mathbf{r}}} = \sqrt{\det \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S}}$, i.e., as the matrix \mathbf{S} is $(p \times p)$,

$$\sqrt{\det \mathbf{g}_{\mathbf{r}}} = S \sqrt{\det \mathbf{g}_{\mathbf{s}}} \quad (\text{if we were considering a change of variables}) \quad , \quad (2.230)$$

where S is the Jacobian determinant, $S = \det \mathbf{S}$. Then, the two equations 2.226–2.228 would simply give

$$f_{\mathbf{s}}(\mathbf{s}) = f_{\mathbf{r}}(\mathbf{r}) \quad , \quad (2.231)$$

expressing the invariance of a volumetric probability under a change of variables (equation 2.65). Of course, we are *not* considering this situation: equations 2.226–2.228 represent a transport of a probability distribution between two spaces, not a change of variables inside a given space.

2.6.0.5 Transportation into the manifold $\mathbf{s} = \mathbf{s}(\mathbf{r})$ itself.

Note: say here that we use in the space \mathbf{X}_{p+q} the ‘induced metric’.

We have seen that bla, bla, bla, and we have the same coordinates over the two spaces, and bla, bla, bla, and to a common capacity element $dr^1 \wedge \cdots \wedge dr^p$ corresponds the two volume elements

$$\begin{aligned} dv_{\mathbf{r}}(\mathbf{r}) &= \sqrt{\det \mathbf{g}_{\mathbf{r}}} dr^1 \wedge \cdots \wedge dr^p \\ dv_{\mathbf{x}}(\mathbf{r}) &= \sqrt{\det (\mathbf{g}_{\mathbf{r}} + \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S})} dr^1 \wedge \cdots \wedge dr^p \quad . \end{aligned} \quad (2.232)$$

When there are volumetric probabilities $f_{\mathbf{r}}(\mathbf{r})$ and $f_{\mathbf{x}}(\mathbf{r})$, they are defined so as to have

$$\begin{aligned} dP_{\mathbf{r}} &= f_{\mathbf{r}} dv_{\mathbf{r}} \\ dP_{\mathbf{x}} &= f_{\mathbf{x}} dv_{\mathbf{x}} \quad . \end{aligned} \quad (2.233)$$

We say the the volumetric probability $f_{\mathbf{x}}$ and has been ‘transported’ from $f_{\mathbf{r}}$ if the two probabilities associated to the two volumes defined by the common capacity element $dr^1 \wedge \cdots \wedge dr^p$ are identical, i.e., if

$$dP_{\mathbf{r}} = dP_{\mathbf{x}} \quad . \quad (2.234)$$

It follows the relation

$$f_{\mathbf{x}} dv_{\mathbf{x}} = f_{\mathbf{r}} dv_{\mathbf{r}} \quad (2.235)$$

i.e.,

$$\boxed{f_{\mathbf{x}} \sqrt{\det (\mathbf{g}_{\mathbf{r}} + \mathbf{S}^t \mathbf{g}_{\mathbf{s}} \mathbf{S})} = f_{\mathbf{r}} \sqrt{\det \mathbf{g}_{\mathbf{r}}} \quad .} \quad (2.236)$$

2.7 Central Estimators and Dispersion Estimators

2.7.1 Introduction

Let \mathbf{X} be an n -dimensional manifold, and let $\mathcal{P}, \mathcal{Q}, \dots$ represent points of \mathbf{X} . The manifold is assumed to have a metric defined over it, i.e., the distance between any two points \mathcal{P} and \mathcal{Q} is defined, and denoted $D(\mathcal{Q}, \mathcal{P})$. Of course, $D(\mathcal{Q}, \mathcal{P}) = D(\mathcal{P}, \mathcal{Q})$.

A normalized probability distribution P is defined over \mathbf{X} , represented by the volumetric probability f . The probability of $\mathcal{A} \subset \mathbf{X}$ is obtained, using the notations of equation 2.49, as

$$P(\mathcal{A}) = \int_{\mathcal{P} \in \mathcal{A}} dV(\mathcal{P}) f(\mathcal{P}) \quad . \quad (2.237)$$

If $\psi(\mathcal{P})$ is a scalar (invariant) function defined over \mathbf{X} , its *average value* is denoted $\langle \psi \rangle$, and is defined as

$$\langle \psi \rangle = \int_{\mathcal{P} \in \mathbf{X}} dV(\mathcal{P}) f(\mathcal{P}) \psi(\mathcal{P}) \quad . \quad (2.238)$$

This clearly corresponds to the intuitive notion of ‘average’.

2.7.2 Center and Radius of a Probability Distribution

Let p be a real number in the range $1 \leq p < \infty$. To any point \mathcal{P} we can associate the quantity (having the dimension of a length)

$$\sigma_p(\mathcal{P}) = \left(\int_{\mathcal{Q} \in \mathbf{X}} dV(\mathcal{Q}) f(\mathcal{Q}) D(\mathcal{Q}, \mathcal{P})^p \right)^{\frac{1}{p}} \quad . \quad (2.239)$$

Definition 2.1 *The point⁹ where $\sigma_p(\mathcal{P})$ attains its minimum value is called the L_p -norm center of the probability distribution $f(\mathcal{P})$, and it is denoted \mathcal{P}_p .*

Definition 2.2 *The minimum value of $\sigma_p(\mathcal{P})$ is called the L_p -norm radius of the probability distribution $f(\mathcal{P})$, and it is denoted σ_p .*

The interpretation of these definitions is simple. Take, for instance $p = 1$. Comparing the two equations 2.238–2.239, we see that, for a fixed point \mathcal{P} , the quantity $\sigma_1(\mathcal{P})$ corresponds to the average of the distances from the point \mathcal{P} to all the points. The point \mathcal{P} that minimizes this average distance is ‘at the center’ of the distribution (in the L_1 -norm sense). For $p = 2$, it is the average of the squared distances that is minimized, etc.

The following terminology shall be used:

- \mathcal{P}_1 is called the *median*, and σ_1 is called the *mean deviation*;
- \mathcal{P}_2 is called the *barycenter* (or the *center*, or the *mean*), and σ_2 is called the *standard deviation* (while its square is called the *variance*);

⁹If there is more than one point where $\sigma_p(\mathcal{P})$ attains its minimum value, any such point is called a center (in the L_p -norm sense) of the probability distribution $f(\mathcal{P})$.

- \mathcal{P}_∞ is called¹⁰ the *circumcenter*, and σ_∞ is called the *circumradius*.

Calling \mathcal{P}_∞ and σ_∞ respectively the ‘circumcenter’ and the ‘circumradius’ seems justified when considering, in the Euclidean plane, a volumetric probability that is constant inside a triangle, and zero outside. The ‘circumcenter’ of the probability distribution is then the circumcenter of the triangle, in the usual geometrical sense, and the ‘circumradius’ of the probability distribution is the radius of the circumscribed circle¹¹. More generally, the circumcenter of a probability distribution is always at the point that minimizes the maximum distance to all other points, and the circumradius of the probability distribution is this ‘minimax’ distance.

Example 2.13 Consider a one-dimensional space \mathcal{N} , with a coordinate ω , such that the distance between the point ν_1 and the point ν_2 is

$$D(\nu_2, \nu_1) = \left| \log \frac{\nu_2}{\nu_1} \right|. \quad (2.240)$$

As suggested in XXX, the space \mathcal{N} could be the space of musical notes, and ν the frequency of a note. Then, this distance is just (up to a multiplicative factor) the usual distance between notes, as given by the number of ‘octaves’. Consider a normalized volumetric probability $f(\nu)$, and let us be interested in the L_2 -norm criteria. For $p = 2$, equation 2.239 can be written

$$(\sigma_2(\mu))^2 = \int_0^\infty ds(\nu) f(\nu) \left(\log \frac{\nu}{\mu} \right)^2, \quad (2.241)$$

The L_2 -norm center of the probability distribution, i.e., the value ν_2 at which $\sigma_2(\mu)$ is minimum, is easily found¹² to be

$$\nu_2 = \nu_0 \exp \left(\int_0^\infty ds(\nu) f(\nu) \log \frac{\nu}{\nu_0} \right), \quad (2.242)$$

where ν_0 is an arbitrary constant (in fact, and by virtue of the properties of the log-exp functions, the value ν_2 is independent of this constant). This mean value ν_2 corresponds to what in statistical theory is called the ‘geometric mean’. The variance of the distribution, i.e., the value of the expression 2.241 at its minimum, is

$$(\sigma_2)^2 = \int_0^\infty ds(\nu) f(\nu) \left(\log \frac{\nu}{\nu_2} \right)^2. \quad (2.243)$$

The distance element associated to the distance in equation 2.240 is, clearly, $ds(\nu) = d\nu/\nu$, and the probability density associated to $f(\nu)$ is $\bar{f}(\nu) = f(\nu)/\nu$, so, in terms of the probability density $\bar{f}(\nu)$, equation 2.242 becomes

$$\nu_2 = \nu_0 \exp \left(\int_0^\infty d\nu \bar{f}(\nu) \log \frac{\nu}{\nu_0} \right), \quad (2.244)$$

¹⁰The L_∞ -norm center and radius are defined as the limit $p \rightarrow \infty$ of the L_p -norm center and radius.

¹¹The circumscribed circle is the circle that contains the three vertices of the triangle. Its center (called circumcenter) is at the the point where the perpendicular bisectors of the sides cross.

¹²For the minimization of the function $\sigma_2(\mu)$ is equivalent to the minimization of $(\sigma_2(\mu))^2$, and this gives the condition $\int ds(\nu) f(\nu) \log(\nu/\mu) = 0$. For any constant ν_0 , this is equivalent to $\int ds(\nu) f(\nu) (\log(\nu/\nu_0) - \log(\mu/\nu_0)) = 0$, i.e., $\log(\mu/\nu_0) = \int ds(\nu) f(\nu) \log(\nu/\nu_0)$, from where the result follows. The constant ν_0 is necessary in these equations for reasons of physical dimensions (only the logarithm of adimensional quantities is defined).

while equation 2.243 becomes

$$(\sigma_2)^2 = \int_0^{\infty} d\nu \bar{f}(\nu) \left(\log \frac{\nu}{\nu_2} \right)^2 . \quad (2.245)$$

The reader shall easily verify that if instead of the variable ν , one chooses to use the logarithmic variable $\nu^* = \log(\nu/\nu_0)$, where ν_0 is an arbitrary constant (perhaps the same as above), then instead of the six expressions 2.240–2.245 we would have obtained, respectively,

$$\begin{aligned} s(\nu_2^*, \nu_1^*) &= |\nu_2^* - \nu_1^*| \\ (\sigma_2(\mu^*))^2 &= \int_{-\infty}^{+\infty} ds(\nu^*) f(\nu^*) (\nu^* - \mu^*)^2 \\ \nu_2^* &= \int_{-\infty}^{+\infty} ds(\nu^*) f(\nu^*) \nu^* \\ (\sigma_2)^2 &= \int_{-\infty}^{+\infty} ds(\nu^*) f(\nu^*) (\nu^* - \nu_2^*)^2 \end{aligned} \quad (2.246)$$

$$\boxed{\nu_2^* = \int_{-\infty}^{+\infty} d\nu^* \bar{f}(\nu^*) \nu^*} \quad (2.247)$$

and

$$\boxed{(\sigma_2)^2 = \int_{-\infty}^{+\infty} d\nu^* \bar{f}(\nu^*) (\nu^* - \nu_2^*)^2 ,} \quad (2.248)$$

with, for this logarithmic variable, $ds(\nu^*) = d\nu^*$ and $f(\nu^*) = \bar{f}(\nu^*)$. The two last expressions are the ordinary equations used to define the mean and the variance in elementary texts. [**End of example.**]

Example 2.14 Consider a one-dimensional space, with a coordinate χ , the distance between two points χ_1 and χ_2 being denoted $D(\chi_2, \chi_1)$. Then, the associated length element is $d\ell(\chi) = D(\chi + d\chi, \chi)$. Finally, consider a (1D) volumetric probability $f(\chi)$, and let us be interested in the L_1 -norm case. Assume that χ runs from a minimum value χ_{\min} to a maximum value χ_{\max} (both could be infinite). For $p = 1$, equation 2.239 can be written

$$\sigma_1(\chi) = \int d\ell(\chi') f(\chi') D(\chi', \chi) . \quad (2.249)$$

Denoting χ_1 be the median, i.e., the point the point where $\sigma_1(\chi)$ is minimum), one easily¹³ finds that χ_1 is characterized by the property that it separates the line into two regions of equal probability, i.e.,

$$\boxed{\int_{\chi_{\min}}^{\chi_1} d\ell(\chi) f(\chi) = \int_{\chi_1}^{\chi_{\max}} d\ell(\chi) f(\chi) ,} \quad (2.250)$$

¹³In fact, the property 2.250 of the median being intrinsic (independent of any coordinate system), we can limit ourselves to demonstrate it using a special ‘Cartesian’ coordinate, where $d\ell(x) = dx$, and $D(x_1, x_2) = |x_2 - x_1|$, where the property is easy to demonstrate (and well known).

expression that can readily be used for an actual computation of the median, and which corresponds to its elementary definition. The mean deviation is then given by

$$\sigma_1 = \int_{\chi_{\min}}^{\chi_{\max}} d\ell(\chi) f(\chi) D(\chi, \chi_1) \quad . \quad (2.251)$$

[End of example.]

Example 2.15 Consider the same situation as in the previous example, but let us become interested in the L_∞ -norm case. Let χ_{\min} and χ_{\max} the minimum and the maximum values of χ for which $f(\chi) \neq 0$. It can be shown that the circumcenter of the probability distribution is the point χ_∞ that separates the interval $\{\chi_{\min}, \chi_{\max}\}$ in two intervals of equal length, i.e., satisfying the condition

$$D(\chi, \chi_{\min}) = D(\chi_{\max}, \chi) \quad , \quad (2.252)$$

and that the circumradius is

$$\sigma_\infty = \frac{D(\chi_{\max}, \chi_{\min})}{2} \quad . \quad (2.253)$$

[End of example.]

Example 2.16 Consider, in the Euclidean n -dimensional space \mathcal{E}_n , with Cartesian coordinates $\mathbf{x} = \{x^1, \dots, x^n\}$, a normalized volumetric probability $f(\mathbf{x})$, and let us be interested in the L_2 -norm case. For $p = 2$, equation 2.239 can be written, using obvious notations,

$$(\sigma_2(\mathbf{y}))^2 = \int d\mathbf{x} f(\mathbf{x}) \|\mathbf{x} - \mathbf{y}\|^2 \quad . \quad (2.254)$$

Let \mathbf{x}_2 denote the mean of the probability distribution, i.e., the point where $\sigma_2(\mathbf{y})$ is minimum (or, equivalently, where $(\sigma_2(\mathbf{y}))^2$ is minimum). The condition of minimum (the vanishing of the derivatives) gives $\int d\mathbf{x} f(\mathbf{x}) (\mathbf{x} - \mathbf{x}_2) = 0$, i.e.,

$$\boxed{\mathbf{x}_2 = \int d\mathbf{x} f(\mathbf{x}) \mathbf{x} \quad ,} \quad (2.255)$$

which is an elementary definition of mean. The variance of the probability distribution is then

$$(\sigma_2)^2 = \int d\mathbf{x} f(\mathbf{x}) \|\mathbf{x} - \mathbf{x}_2\|^2 \quad . \quad (2.256)$$

In the context of this example, we can define the covariance tensor

$$\boxed{\mathbf{C} = \int d\mathbf{x} f(\mathbf{x}) (\mathbf{x} - \mathbf{x}_2) \otimes (\mathbf{x} - \mathbf{x}_2) \quad .} \quad (2.257)$$

Note that equation 2.255 and equation 2.257 can be written, using indices, as

$$x_2^i = \int dx^1 \wedge \dots \wedge dx^n f(x^1, \dots, x^n) x^i \quad , \quad (2.258)$$

and

$$C^{ij} = \int dx^1 \wedge \dots \wedge dx^n f(x^1, \dots, x^n) (x^i - x_2^i) (x^j - x_2^j) \quad . \quad (2.259)$$

[End of example.]

2.8 Appendixes

2.8.1 Appendix: Conditional Probability Density

Note to the reader: this section can be skipped, unless one is particularly interested in probability densities.

In view of equation 2.100, the *conditional probability density* (over the submanifold \mathbf{X}_p) is to be defined as

$$\bar{f}_p(\mathbf{r}) = \bar{g}_p(\mathbf{r}) f_p(\mathbf{r}) \quad (2.260)$$

i.e.,

$$\bar{f}_p(\mathbf{r}) = \eta_{\mathbf{r}} \sqrt{\det \mathbf{g}_p(\mathbf{r})} f_p(\mathbf{r}) \quad , \quad (2.261)$$

so the probability of a region \mathcal{A}_p of the submanifold is given by

$$P(\mathcal{A}_p) = \int_{\mathbf{r} \in \mathbf{X}_p} dv_p(\mathbf{r}) \bar{f}_p(\mathbf{r}) \quad , \quad (2.262)$$

where $dv_p(\mathbf{r}) = dr^1 \wedge \dots \wedge dr^p$.

We must now express $\bar{f}_p(\mathbf{r})$ in terms of $\bar{f}(\mathbf{r}, \mathbf{s})$. First, from equations 2.95 and 2.261 we obtain

$$\bar{f}_p(\mathbf{r}) = \eta_{\mathbf{r}} \sqrt{\det \mathbf{g}_p(\mathbf{r})} \frac{f(\mathbf{r}, \mathbf{s}(\mathbf{r}))}{\int_{\mathbf{r} \in \mathbf{X}_p} dv_p(\mathbf{r}) f(\mathbf{r}, \mathbf{s}(\mathbf{r}))} \quad . \quad (2.263)$$

As $f(\mathbf{r}, \mathbf{s}) = \bar{f}(\mathbf{r}, \mathbf{s}) / (\eta \sqrt{\det \mathbf{g}})$ (equation 2.54),

$$\bar{f}_p(\mathbf{r}) = \eta_{\mathbf{r}} \sqrt{\det \mathbf{g}_p(\mathbf{r})} \frac{\bar{f}(\mathbf{r}, \mathbf{s}(\mathbf{r})) / \sqrt{\det \mathbf{g}}}{\int_{\mathbf{r} \in \mathbf{X}_p} dv_p(\mathbf{r}) \bar{f}(\mathbf{r}, \mathbf{s}(\mathbf{r})) / \sqrt{\det \mathbf{g}}} \quad . \quad (2.264)$$

Finally, using 2.97, and expliciting $\mathbf{g}_p(\mathbf{r})$,

$$\bar{f}_p(\mathbf{r}) = \frac{\frac{\sqrt{\det(\mathbf{g}_{rr} + \mathbf{g}_{rs} \mathbf{S} + \mathbf{S}^T \mathbf{g}_{sr} + \mathbf{S}^T \mathbf{g}_{ss} \mathbf{S})}}{\sqrt{\det \mathbf{g}}} \bar{f}(\mathbf{r}, \mathbf{s}(\mathbf{r}))}{\int_{\mathbf{r} \in \mathbf{X}_p} dr^1 \wedge \dots \wedge dr^p \frac{\sqrt{\det(\mathbf{g}_{rr} + \mathbf{g}_{rs} \mathbf{S} + \mathbf{S}^T \mathbf{g}_{sr} + \mathbf{S}^T \mathbf{g}_{ss} \mathbf{S})}}{\sqrt{\det \mathbf{g}}} \bar{f}(\mathbf{r}, \mathbf{s}(\mathbf{r}))} \quad . \quad (2.265)$$

Again, it is understood here that all the ‘matrices’ are taken at the point $(\mathbf{r}, \mathbf{s}(\mathbf{r}))$.

This expression does not coincide the the conditional probability defined given in usual texts (even when the manifold is defined by the condition $\mathbf{s} = \mathbf{s}_0 = \text{const.}$). This is because we contemplate here the ‘metric’ or ‘orthogonal’ limit to the manifold (in the sense of figure 2.9), while usual texts just consider the ‘vertical limit’. Of course, I take this approach here because I think it is essential for consistent applications of the notion of conditional probability. The best known expression of this problem is the so called ‘Borel Paradox’ that we analyze in section 2.8.10.

Example 2.17 *If we face the case where the space \mathbf{X} is the Cartesian product of two spaces $\mathbf{R} \times \mathbf{S}$, with $\mathbf{g}_{uv} = \mathbf{g}_{vu} = 0$, $\mathbf{g}_{rr} = \mathbf{g}_r(\mathbf{r})$ and $\mathbf{g}_{ss} = \mathbf{g}_s(\mathbf{s})$, then $\det \mathbf{g}(\mathbf{r}, \mathbf{s}) = \det \mathbf{g}_r(\mathbf{r}) \det \mathbf{g}_s(\mathbf{s})$, and the conditional probability density of equation 2.265 becomes,*

$$\bar{f}_p(\mathbf{r}) = \frac{\frac{\sqrt{\det(\mathbf{g}_r(\mathbf{r}) + \mathbf{S}^T(\mathbf{r}) \mathbf{g}_s(\mathbf{s}(\mathbf{r})) \mathbf{S}(\mathbf{r}))}}{\sqrt{\det \mathbf{g}_r(\mathbf{r})} \sqrt{\det \mathbf{g}_s(\mathbf{s}(\mathbf{r}))}} \bar{f}(\mathbf{r}, \mathbf{s}(\mathbf{r}))}{\int_{\mathbf{r} \in \mathbf{X}_p} dr^1 \wedge \dots \wedge dr^p \frac{\sqrt{\det(\mathbf{g}_r(\mathbf{r}) + \mathbf{S}^T(\mathbf{r}) \mathbf{g}_s(\mathbf{s}(\mathbf{r})) \mathbf{S}(\mathbf{r}))}}{\sqrt{\det \mathbf{g}_r(\mathbf{r})} \sqrt{\det \mathbf{g}_s(\mathbf{s}(\mathbf{r}))}} \bar{f}(\mathbf{r}, \mathbf{s}(\mathbf{r}))} . \quad (2.266)$$

[End of example.]

Example 2.18 *If, in addition to the condition of the previous example, the hypersurface is defined by a constant value of \mathbf{s} , say $\mathbf{s} = \mathbf{s}_0$, then, the probability density becomes*

$$\bar{f}_p(\mathbf{r}) = \frac{\bar{f}(\mathbf{r}, \mathbf{s}_0)}{\int_{\mathbf{r} \in \mathbf{X}_p} dr^1 \wedge \dots \wedge dr^p \bar{f}(\mathbf{r}, \mathbf{s}_0)} . \quad (2.267)$$

[End of example.]

Example 2.19 *In the situation of the previous example, let us rewrite equation 2.267 dropping the index $_0$ from \mathbf{s}_0 , and use the notations*

$$\bar{f}_{r|s}(\mathbf{r}|\mathbf{s}) = \frac{\bar{f}(\mathbf{r}, \mathbf{s})}{\bar{f}_s(\mathbf{s})} , \quad ; \quad \bar{f}_s(\mathbf{s}) = \int_{\mathbf{r} \in \mathbf{X}_p} dr^1 \wedge \dots \wedge dr^p \bar{f}(\mathbf{r}, \mathbf{s}) . \quad (2.268)$$

We could redo all the computations to define the conditional for \mathbf{s} , given a fixed value \mathbf{v} , but it is clear by simple analogy that we obtain, in this case,

$$\bar{f}_{s|r}(\mathbf{s}|\mathbf{r}) = \frac{\bar{f}(\mathbf{r}, \mathbf{s})}{\bar{f}_r(\mathbf{r})} , \quad ; \quad \bar{f}_r(\mathbf{r}) = \int_{\mathbf{r} \in \mathbf{X}_q} ds^1 \wedge \dots \wedge ds^q \bar{f}(\mathbf{r}, \mathbf{s}) . \quad (2.269)$$

Solving in these two equations for $\bar{f}(\mathbf{r}, \mathbf{s})$ gives the ‘Bayes theorem’

$$\bar{f}_{s|r}(\mathbf{s}|\mathbf{r}) = \frac{\bar{f}_{r|s}(\mathbf{r}|\mathbf{s}) \bar{f}_s(\mathbf{s})}{\bar{f}_r(\mathbf{r})} . \quad (2.270)$$

*Note that this theorem is valid only if we work in the Cartesian product of two spaces. In particular, we must have $\mathbf{g}_{ss}(\mathbf{r}, \mathbf{s}) = \mathbf{g}_s(\mathbf{s})$. Working, for instance, at the surface of the sphere with geographical coordinates $(\mathbf{r}, \mathbf{s}) = (r, s) = (\varphi, \lambda)$ this condition is **not** fulfilled, as $g_\varphi = \cos \lambda$ is a function of λ : the surface of the sphere is not the Cartesian product of two 1D spaces. As we shall later see, this enters in the discussion of the so-called ‘Borel paradox’ (there is no paradox, if we do things properly). [End of example.]*

2.8.2 Appendix: Marginal Probability Density

In the context of section 2.5.2, where a manifold \mathbf{X} is built through the Cartesian product $\mathbf{R} \times \mathbf{S}$ of two manifolds, and given a ‘joint’ volumetric probability $f(\mathbf{r}, \mathbf{s})$, the marginal volumetric probability $f_{\mathbf{r}}(\mathbf{r})$ is defined as (see equation 2.160)

$$f_{\mathbf{r}}(\mathbf{r}) = \int_{\mathbf{s} \in \mathbf{S}} dv_{\mathbf{s}}(\mathbf{s}) f(\mathbf{r}, \mathbf{s}) \quad . \quad (2.271)$$

Let us find the equivalent expression using probability densities instead of volumetric probabilities.

Here below, following our usual conventions, the following notations

$$\bar{g}(\mathbf{r}, \mathbf{s}) = \sqrt{\det \mathbf{g}(\mathbf{r}, \mathbf{s})} \quad ; \quad \bar{g}_{\mathbf{r}}(\mathbf{r}) = \sqrt{\det \mathbf{g}_{\mathbf{r}}(\mathbf{r})} \quad ; \quad \bar{g}_{\mathbf{s}}(\mathbf{s}) = \sqrt{\det \mathbf{g}_{\mathbf{s}}(\mathbf{s})} \quad (2.272)$$

are introduced. First, we may use the relation

$$f(\mathbf{r}, \mathbf{s}) = \frac{\bar{f}(\mathbf{r}, \mathbf{s})}{\bar{g}(\mathbf{r}, \mathbf{s})} \quad (2.273)$$

linking the volumetric probability $f(\mathbf{r}, \mathbf{s})$ and the probability density $\bar{f}(\mathbf{r}, \mathbf{s})$. Here, \mathbf{g} is the metric of the manifold \mathbf{X} , that has been assumed to have a partitioned form (equation 2.123). Then, $f(\mathbf{r}, \mathbf{s}) = \bar{f}(\mathbf{r}, \mathbf{s}) / (\bar{g}_{\mathbf{r}}(\mathbf{r}) \bar{g}_{\mathbf{s}}(\mathbf{s}))$, and equation 2.271 becomes

$$f_{\mathbf{r}}(\mathbf{r}) = \frac{1}{\bar{g}_{\mathbf{r}}(\mathbf{r})} \int_{\mathbf{s} \in \mathbf{S}} dv_{\mathbf{s}}(\mathbf{s}) \frac{\bar{f}(\mathbf{r}, \mathbf{s})}{\bar{g}_{\mathbf{s}}(\mathbf{s})} \quad . \quad (2.274)$$

As the volume element $dv_{\mathbf{s}}(\mathbf{s})$ is related to the capacity element $d\bar{v}_{\mathbf{s}}(\mathbf{s}) = ds^1 \wedge ds^2 \wedge \dots$ via the relation

$$dv_{\mathbf{s}}(\mathbf{s}) = \bar{g}_{\mathbf{s}}(\mathbf{s}) d\bar{v}_{\mathbf{s}}(\mathbf{s}) \quad , \quad (2.275)$$

we can write

$$f_{\mathbf{r}}(\mathbf{r}) = \frac{1}{\bar{g}_{\mathbf{r}}(\mathbf{r})} \int_{\mathbf{s} \in \mathbf{S}} d\bar{v}_{\mathbf{s}}(\mathbf{s}) \bar{f}(\mathbf{r}, \mathbf{s}) \quad , \quad (2.276)$$

i.e.,

$$\bar{g}_{\mathbf{r}}(\mathbf{r}) f_{\mathbf{r}}(\mathbf{r}) = \int_{\mathbf{s} \in \mathbf{S}} d\bar{v}_{\mathbf{s}}(\mathbf{s}) \bar{f}(\mathbf{r}, \mathbf{s}) \quad . \quad (2.277)$$

We recognize, at the left-hand side, the usual definition of a probability density as the product of a volumetric probability by the volume density, so we can introduce the *marginal probability density*

$$\bar{f}_{\mathbf{r}}(\mathbf{r}) = \bar{g}_{\mathbf{r}}(\mathbf{r}) f_{\mathbf{r}}(\mathbf{r}) \quad . \quad (2.278)$$

Then, equation 2.277 becomes

$$\boxed{\bar{f}_{\mathbf{r}}(\mathbf{r}) = \int_{\mathbf{s} \in \mathbf{S}} d\bar{v}_{\mathbf{s}}(\mathbf{s}) \bar{f}(\mathbf{r}, \mathbf{s}) \quad ,} \quad (2.279)$$

expression that could be taken as a direct definition of the marginal probability density $\bar{f}_{\mathbf{r}}(\mathbf{r})$ in terms of the ‘joint’ probability density $\bar{f}(\mathbf{r}, \mathbf{s})$.

Note that this expression is formally identical to 2.271. This contrasts with the expression of a conditional probability density (equation 2.265) that is formally very different from the expression of a conditional volumetric probability (equation 2.95).

2.8.3 Appendix: Replacement Gymnastics

In an n -dimensional manifold with coordinates \mathbf{x} , the volume element $dv_{\mathbf{x}}(\mathbf{x})$, is related to the the capacity element $d\underline{v}_{\mathbf{x}}(\mathbf{x}) = dx^1 \wedge \cdots \wedge dx^n$ via the volume density $\bar{g}_{\mathbf{x}}(\mathbf{x}) = \sqrt{\det \mathbf{g}_{\mathbf{x}}(\mathbf{x})}$,

$$dv_{\mathbf{x}}(\mathbf{x}) = \bar{g}_{\mathbf{x}}(\mathbf{x}) d\underline{v}_{\mathbf{x}}(\mathbf{x}) \quad , \quad (2.280)$$

while the relation between a volumetric probability $f_{\mathbf{x}}(\mathbf{x})$ and the associated probability density $\bar{f}_{\mathbf{x}}(\mathbf{x})$ is

$$\bar{f}_{\mathbf{x}}(\mathbf{x}) = \bar{g}_{\mathbf{x}}(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) \quad . \quad (2.281)$$

In a change of variables $\mathbf{x} \rightleftharpoons \mathbf{y}$, while the capacity element changes according to

$$d\underline{v}_{\mathbf{x}}(\mathbf{x}) = X(\mathbf{y}) d\underline{v}_{\mathbf{y}}(\mathbf{y}) \quad , \quad (2.282)$$

where the Jacobian determinant X is the determinant of the matrix $\{X^i_j\} = \{\partial x^i / \partial y^j\}$, the probability density changes according to

$$\bar{f}_{\mathbf{x}}(\mathbf{x}) = \frac{1}{X(\mathbf{y})} \bar{f}_{\mathbf{y}}(\mathbf{y}) \quad . \quad (2.283)$$

In the variables \mathbf{y} , the relation between a volumetric probability $f_{\mathbf{y}}(\mathbf{y})$ and the associated probability density $\bar{f}_{\mathbf{y}}(\mathbf{y})$ is

$$\bar{f}_{\mathbf{y}}(\mathbf{y}) = \bar{g}_{\mathbf{y}}(\mathbf{y}) f_{\mathbf{y}}(\mathbf{y}) \quad , \quad (2.284)$$

where $\bar{g}_{\mathbf{y}}(\mathbf{y}) = \sqrt{\det \mathbf{g}_{\mathbf{y}}(\mathbf{y})}$ is the volume density in the coordinates \mathbf{y} . Finally, the volume element $dv_{\mathbf{y}}(\mathbf{y})$, is related to the the capacity element $d\underline{v}_{\mathbf{y}}(\mathbf{y}) = dy^1 \wedge \cdots \wedge dy^n$ through

$$dv_{\mathbf{y}}(\mathbf{y}) = \bar{g}_{\mathbf{y}}(\mathbf{y}) d\underline{v}_{\mathbf{y}}(\mathbf{y}) \quad . \quad (2.285)$$

Using these relations in turn, we can obtain the following circle of equivalent equations:

$$\begin{aligned} P(\mathcal{A}) &= \int_{\mathcal{P} \in \mathcal{A}} dV(\mathcal{P}) f(\mathcal{P}) = \int_{\mathbf{x} \in \mathcal{A}} dv_{\mathbf{x}}(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) \\ &= \int_{\mathbf{x} \in \mathcal{A}} d\underline{v}_{\mathbf{x}}(\mathbf{x}) \bar{g}_{\mathbf{x}}(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}) \\ &= \int_{\mathbf{x} \in \mathcal{A}} d\underline{v}_{\mathbf{x}}(\mathbf{x}) \bar{f}_{\mathbf{x}}(\mathbf{x}) \\ &= \int_{\mathbf{y} \in \mathcal{A}} (X(\mathbf{y}) d\underline{v}_{\mathbf{y}}) \left(\frac{1}{X(\mathbf{y})} \bar{f}_{\mathbf{y}}(\mathbf{y}) \right) \\ &= \int_{\mathbf{y} \in \mathcal{A}} d\underline{v}_{\mathbf{y}}(\mathbf{y}) \bar{f}_{\mathbf{y}}(\mathbf{y}) \\ &= \int_{\mathbf{y} \in \mathcal{A}} d\underline{v}_{\mathbf{y}}(\mathbf{y}) \bar{g}_{\mathbf{y}}(\mathbf{y}) f_{\mathbf{y}}(\mathbf{y}) \\ &= \int_{\mathbf{y} \in \mathcal{A}} dv_{\mathbf{y}}(\mathbf{y}) f_{\mathbf{y}}(\mathbf{y}) = \int_{\mathcal{P} \in \mathcal{A}} dV(\mathcal{P}) f(\mathcal{P}) = P(\mathcal{A}) \quad . \end{aligned} \quad (2.286)$$

Each one of them may be useful in different circumstances. The student should be able to easily move from one equation to the next.

Example 2.20 In the example Cartesian-geographical, the equations above give, respectively (using the index \mathbf{r} for the geographical coordinates),

$$dv_{\mathbf{x}}(x, y, z) = dx \wedge dy \wedge dz \quad (2.287)$$

$$\bar{f}_{\mathbf{x}}(x, y, z) = f_{\mathbf{x}}(x, y, z) \quad (2.288)$$

$$dx \wedge dy \wedge dz = r^2 \cos \lambda \, dr \wedge d\varphi \wedge d\lambda \quad (2.289)$$

$$\bar{f}_{\mathbf{x}}(x, y, z) = \frac{1}{r^2 \cos \lambda} \bar{f}_{\mathbf{r}}(r, \varphi, \lambda) \quad (2.290)$$

$$\bar{f}_{\mathbf{r}}(r, \varphi, \lambda) = r^2 \cos \lambda \, f_{\mathbf{r}}(r, \varphi, \lambda) \quad (2.291)$$

$$dv_{\mathbf{r}}(r, \varphi, \lambda) = r^2 \cos \lambda \, dr \wedge d\varphi \wedge d\lambda \quad , \quad (2.292)$$

to obtain the circle of equations,

$$\begin{aligned} P(\mathcal{A}) &= \int_{\mathcal{P} \in \mathcal{A}} dV(\mathcal{P}) f(\mathcal{P}) = \int_{\{x,y,z\} \in \mathcal{A}} dv_{\mathbf{x}}(x, y, z) f_{\mathbf{x}}(x, y, z) \\ &= \int_{\{x,y,z\} \in \mathcal{A}} dx \wedge dy \wedge dz f_{\mathbf{x}}(x, y, z) \\ &= \int_{\{x,y,z\} \in \mathcal{A}} dx \wedge dy \wedge dz \bar{f}_{\mathbf{x}}(x, y, z) \\ &= \int_{\{r,\varphi,\lambda\} \in \mathcal{A}} (r^2 \cos \lambda \, dr \wedge d\varphi \wedge d\lambda) \left(\frac{1}{r^2 \cos \lambda} \bar{f}_{\mathbf{r}}(r, \varphi, \lambda) \right) \\ &= \int_{\{r,\varphi,\lambda\} \in \mathcal{A}} dr \wedge d\varphi \wedge d\lambda \bar{f}_{\mathbf{r}}(r, \varphi, \lambda) \\ &= \int_{\{r,\varphi,\lambda\} \in \mathcal{A}} dr \wedge d\varphi \wedge d\lambda \, r^2 \cos \lambda \, f_{\mathbf{r}}(r, \varphi, \lambda) \\ &= \int_{\{r,\varphi,\lambda\} \in \mathcal{A}} dv_{\mathbf{r}}(r, \varphi, \lambda) f_{\mathbf{r}}(r, \varphi, \lambda) = \int_{\mathcal{P} \in \mathcal{A}} dV(\mathcal{P}) f(\mathcal{P}) = P(\mathcal{A}) \quad . \end{aligned} \quad (2.293)$$

Note that the Cartesian system of coordinates is special: scalar densities, scalar capacities and invariant scalars coincide. **[End of example.]**

2.8.4 Appendix: The Gaussian Probability Distribution

2.8.4.1 One Dimensional Spaces

Let \mathbf{X} be a one-dimensional metric line with points $\mathcal{P}, \mathcal{Q}, \dots$, and let $s(\mathcal{Q}, \mathcal{P})$ denote the displacement from point \mathcal{P} to point \mathcal{Q} , the distance or ‘length’ between the two points being the absolute value of the displacement, $L(\mathcal{Q}, \mathcal{P}) = L(\mathcal{P}, \mathcal{Q}) = |s(\mathcal{Q}, \mathcal{P})|$. Given any particular point \mathcal{P} on the line, it is assumed that the line extends to infinite distances from \mathcal{P} in the two senses. The one-dimensional Gaussian probability distribution is defined by the volumetric probability

$$f(\mathcal{P}; \mathcal{P}_0; \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} s^2(\mathcal{P}, \mathcal{P}_0)\right) \quad , \quad (2.294)$$

and it follows from the general definition of volumetric probability, that the probability of the interval between any two points \mathcal{P}_1 and \mathcal{P}_2 is

$$P = \int_{\mathcal{P}_1}^{\mathcal{P}_2} dL(\mathcal{P}) f(\mathcal{P}; \mathcal{P}_0; \sigma) \quad , \quad (2.295)$$

where dL denotes the elementary length element. The following properties are easy to demonstrate:

- the probability of the whole line equals one (i.e., the volumetric probability $f(\mathcal{P}; \mathcal{P}_0; \sigma)$ is normalized);
- the mean of $f(\mathcal{P}; \mathcal{P}_0; \sigma)$ is the point \mathcal{P}_0 ;
- the standard deviation of $f(\mathcal{P}; \mathcal{P}_0; \sigma)$ equals σ .

Example 2.21 Consider a coordinate X such that the displacement between two points is $s_X(X', X) = \log(X'/X)$. Then, the Gaussian distribution 2.294 takes the form

$$f_X(X; X_0, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} \left(\log \frac{X}{X_0}\right)^2\right) \quad , \quad (2.296)$$

where X_0 is the mean and σ the standard deviation. As, here, $ds(X) = dX/X$, the probability of an interval is

$$P(X_1 \leq X \leq X_2) = \int_{X_1}^{X_2} \frac{dX}{X} f_X(X; X_0, \sigma) \quad , \quad (2.297)$$

and we have the normalization

$$\int_0^\infty \frac{dX}{X} f_X(X; X_0, \sigma) = 1 \quad . \quad (2.298)$$

This expression of the Gaussian probability distribution, written in terms on the variable X , is called the lognormal law. I suggest that the information on the parameter X represented by the volumetric probability 2.296 should be expressed by a notation like¹⁴

$$\log \frac{X}{X_0} = \pm \sigma \quad , \quad (2.299)$$

¹⁴Equivalently, one may write $X = X_0 \exp(\pm\sigma)$, or $X = X_0 \div \Sigma$, where $\Sigma = \exp \sigma$.

that is the exact equivalent of the notation used in equation 2.303 below. Defining the difference $\delta X = X - X_0$ one converts this equation into $\log(1 + \delta X/X_0)$, whose first order approximation is $\delta X/X_0 = \pm\sigma$. This shows that σ corresponds to what is usually called the ‘relative uncertainty’. I do not recommend this terminology, as, with the definitions used in this book (see section 2.7), σ is the actual standard deviation of the quantity X . [End of example.]

Exercise: write the equivalent of the three expressions 2.296–2.298 using, instead of the variable X , the variables $U = 1/X$ or $Y = X^n$.

Example 2.22 Consider a coordinate x such that the displacement between two points is $s_x(x', x) = x' - x$. Then, the Gaussian distribution 2.294 takes the form

$$f_x(x; x_0, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - x_0)^2\right) \quad , \quad (2.300)$$

where x_0 is the mean and σ the standard deviation. As, here, $ds(x) = dx$, the probability of an interval is

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} dx f_x(x; x_0, \sigma) \quad , \quad (2.301)$$

and we have the normalization

$$\int_{-\infty}^{+\infty} dx f_x(x; x_0, \sigma) = 1 \quad . \quad (2.302)$$

This expression of the Gaussian probability distribution, written in terms on the variable x , is called the normal law. The information on the parameter x represented by the volumetric probability 2.300 is commonly expressed by a notation like¹⁵

$$x = x_0 \pm \sigma \quad . \quad (2.303)$$

[End of example.]

Example 2.23 It is easy to verify that through the change of variable

$$x = \log \frac{X}{K} \quad , \quad (2.304)$$

where K is an arbitrary constant, the equations of the example 2.21 become those of the example 2.22, and vice-versa. In this case, the quantity x has no physical dimensions (this is, of course, a possibility, but not a necessity, for the quantity x in example 2.22). [End of example.]

The Gaussian probability distribution is represented in figure 2.20. Note that there is no need to make different plots for the normal and the lognormal volumetric probabilities. When one is interested in a wide range of values, a logarithmic version of the vertical axis may be necessary (see figure 2.21).

¹⁵ More concise notations are also used. As an example, the expression $x = 1\,234.567\,89\text{ m} \pm 0.000\,11\text{ m}$ (here, ‘m’ represents the physical unit ‘meter’) is sometimes written $x = (1\,234.567\,89 \pm 0.000\,11)\text{ m}$ or even $x = 1\,234.567\,89(11)\text{ m}$.

Figure 2.20: A representation of the Gaussian probability distribution, where the example of a temperature T is used. Reading the scale at the top, we associate to each value of the temperature T the value $h(T)$ of a lognormal volumetric probability. Reading the scale at the bottom, we associate to every value of the logarithmic temperature t the value $g(t)$ of a normal volumetric probability. There is no need to make a special plot where the lognormal volumetric probability $h(T)$ would not be represented ‘in a logarithmic axis’, as this strongly distorts the beautiful Gaussian bell (see figures 2.22 and 2.23). In the figure represented here, one standard deviation corresponds to one unit of t , so the whole range represented equals 8σ .

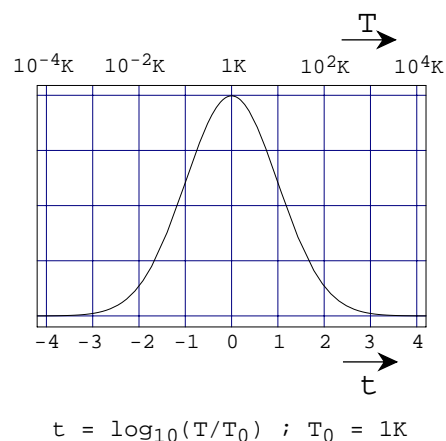


Figure 2.21: A representation of the normal volumetric probability using a logarithmic vertical axis (here, a base 10 logarithm of the volumetric probability, relative to its maximum value). While the representation in figure 2.20 is not practical is one is interested in values of t outside the interval with endpoints situated at $\pm 3\sigma$ of the center, this representation allows the examination of the statistics concerning as many decades as we may wish. Here, the whole range represented equals more than 20 standard deviations.

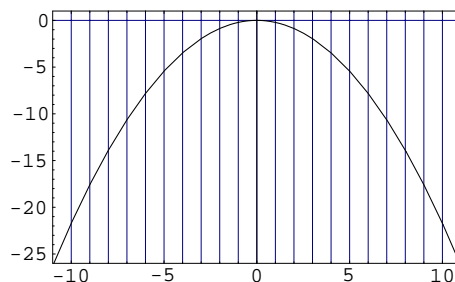


Figure 2.22: Left: the lognormal volumetric probability $h(X)$. Right: the lognormal probability density $\bar{h}(X)$. Distributions centered at 1, with standard deviations respectively equal to 0.1, 0.2, 0.4, 0.8, 1.6 and 3.2.

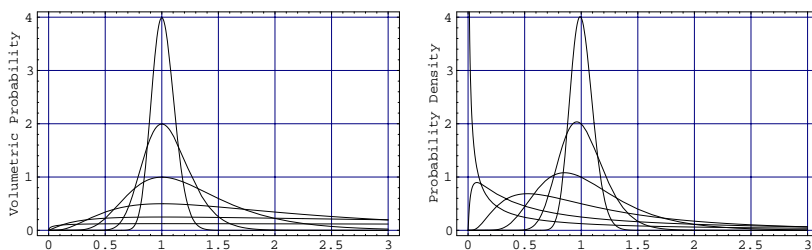
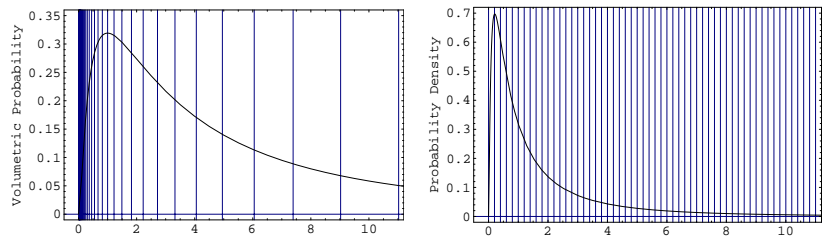


Figure 2.23 gives the interpretation of these functions in terms of histograms. By definition of volumetric probability, an histogram should be made dividing the interval under study in segments of same length $ds(X) = dX/Y$, as opposed to the definition of probability density, where the interval should be divided in segments of equal ‘variable increment’ dX . We clearly see, at the right of the figure the impracticality of making the histogram corresponding to the probability density: while the right part of the histogram oversamples the variable, the left part undersamples it. The histogram suggested at the left samples the variable homogeneously, but this only means that we are using constant steps of the logarithmic quantity x associated to the positive quantity X . Better, then, to directly use the representation suggested in figure 2.20 or in figure 2.21. We have then a double conclusion: (i) the lognormal probability density (at the right in figures 2.22 and 2.23) does not correspond to any practical histogram; it is generally uninteresting. (ii) the lognormal volumetric probability (at the left in figures 2.22 and 2.23) does correspond to a practical histogram, but is better handled when the associated normal volumetric probability is used instead (figure 2.20 or figure 2.21). In short: *lognormal functions should never be used*.

Figure 2.23: A typical Gaussian distribution, with central point 1 and standard deviation $5/4$, represented here, using a Jeffreys (positive) quantity, by the lognormal volumetric probability (left) and the lognormal probability density (right).



2.8.4.2 Multi Dimensional Spaces

In dimension grater than one, the spaces may have curvature. But the multidimensional Gaussian distribution is useful in flat spaces (i.e., Euclidean spaces) only. Although it is possible to give general expressions for arbitrary coordinate systems, let us simplify the exposition, and assume that we are using rectilinear coordinates. The squared distance between points \mathbf{x}_1 and \mathbf{x}_2 is then given by the ‘sum of squares’

$$D^2(\mathbf{x}_2, \mathbf{x}_1) = (\mathbf{x}_2 - \mathbf{x}_1)^t \mathbf{g} (\mathbf{x}_2 - \mathbf{x}_1) \quad , \quad (2.305)$$

where the metric tensor \mathbf{g} is constant (because the assumption of an Euclidean space with rectilinear coordinates). The volume element is, then,

$$dv(\mathbf{x}) = \sqrt{\det \mathbf{g}} \, dx^1 \wedge \cdots \wedge dx^n \quad , \quad (2.306)$$

where, again, $\sqrt{\det \mathbf{g}}$ is a constant.

Let $f(\mathbf{x})$ be a volumetric probability over the space. By definition, the probability of a region \mathcal{A} is

$$P(\mathcal{A}) = \int_{\mathcal{A}} dv(\mathbf{x}) f(\mathbf{x}) \quad , \quad (2.307)$$

i.e.,

$$P(\mathcal{A}) = \sqrt{\det \mathbf{g}} \int dx^1 \wedge \cdots \wedge dx^n f(\mathbf{x}) \quad . \quad (2.308)$$

The *multidimensional Gaussian volumetric probability* is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \frac{\sqrt{\det \mathbf{G}}}{\sqrt{\det \mathbf{g}}} \exp \left(-\frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^t \mathbf{G} (\mathbf{x} - \mathbf{x}_0) \right) \quad . \quad (2.309)$$

The following properties are slight generalizations of well known results concerning the multidimensional Gaussian:

- $f(\mathbf{x})$ is normed, i.e., $\int dv(\mathbf{x}) f(\mathbf{x}) = 1$;
- the mean of $f(\mathbf{x})$ is \mathbf{x}_0 ;
- the covariance matrix of $f(\mathbf{x})$ is¹⁶ $\mathbf{C} = \mathbf{G}^{-1}$.

Note that when in an Euclidean space with metric \mathbf{g} one defines a Gaussian with covariance \mathbf{C} , one may use the inverse of the covariance matrix, $\mathbf{G} = \mathbf{C}^{-1}$, as a supplementary metric over the space.

¹⁶Remember that the general definition of covariance gives here $C^{ij} = \int dv(\mathbf{x})(x^i - x_0^i)(x^j - x_0^j) f(\mathbf{x})$, so this property is not as obvious as it may seem.

2.8.5 Appendix: The Laplacian Probability Distribution

2.8.5.1 Appendix: One Dimensional Laplacian

Let \mathbb{X} be a metric manifold with points $\mathcal{P}, \mathcal{Q} \dots$, and let $s(\mathcal{P}, \mathcal{Q}) = s(\mathcal{Q}, \mathcal{P})$ denote the distance between two points \mathcal{P} and \mathcal{Q} . The Gaussian probability distribution is represented by the volumetric probability

$$f(\mathcal{P}) = k \exp\left(-\frac{1}{\sigma} s(\mathcal{P}, \mathcal{Q})\right) . \quad (2.310)$$

[Note: Elaborate this.]

2.8.6 Appendix: Exponential Distribution

Note: I have to verify that the following terminology has been introduced. By $\mathbf{s}(\mathcal{P}, \mathcal{P}_0)$ we denote the geodesic line arriving at \mathcal{P} , with origin at \mathcal{P}_0 . If the space is 1D, we write $s(\mathcal{P}, \mathcal{P}_0)$, and call this a *displacement*. Then, the distance is $D(\mathcal{P}, \mathcal{P}_0) = |s(\mathcal{P}, \mathcal{P}_0)|$. In an n D Euclidean space, using Cartesian coordinates, we write $\mathbf{s}(\mathbf{x}, \mathbf{x}_0) = \mathbf{x} - \mathbf{x}_0$, and call this the *displacement vector*.

2.8.6.1 Definition

Consider a one-dimensional space, and denote $s(\mathcal{Q}, \mathcal{P})$, the displacement from point \mathcal{P} to point \mathcal{Q} . The *exponential distribution* has the (1D) volumetric probability

$$f(\mathcal{P}; \mathcal{P}_0) = \alpha \exp(-\alpha s(\mathcal{P}, \mathcal{P}_0)) \quad ; \quad \alpha \geq 0 \quad , \quad (2.311)$$

where \mathcal{P}_0 is some fixed point. This volumetric probability is normed via $\int ds(\mathcal{P}) f(\mathcal{P}, \mathcal{P}_0) = 1$, where the sum concerns the half-interval at the right or at the left of point \mathcal{P}_0 , depending on the orientation chosen (see examples 2.24 and 2.25).

Example 2.24 Consider a coordinate X such that the displacement between two points is $s_X(X', X) = \log(X'/X)$. Then, the exponential distribution 2.311 takes the form $f_X(X; X_0) = k \exp(-\alpha \log(X/X_0))$, i.e.,

$$f_X(X) = \alpha \left(\frac{X}{X_0} \right)^{-\alpha} \quad ; \quad \alpha \geq 0 \quad . \quad (2.312)$$

As, here, $ds(X) = dX/X$, the probability of an interval is $P(X_1 \leq X \leq X_2) = \int_{X_1}^{X_2} \frac{dX}{X} f_X(X)$. The volumetric probability $f_X(X)$ has been normed using

$$\int_{X_0}^{\infty} \frac{dX}{X} f_X(X) = 1 \quad . \quad (2.313)$$

This form of the exponential distribution is usually called the Pareto law. The cumulative probability function is

$$g_X(X) = \int_{X_0}^X \frac{dX'}{X'} f_X(X') = 1 - \left(\frac{X}{X_0} \right)^{-\alpha} \quad . \quad (2.314)$$

It is negative for $X < X_0$, zero for $X = X_0$, and positive for $X > X_0$. The power α of the ‘power law’ 2.312 may be any real number, but in most examples concerning the physical, biological or economical sciences, it is of the form $\alpha = p/q$, with p and q being small positive integers¹⁷. With a variable $U = 1/X$, equation 2.317 becomes

$$f_U(U) = k' U^\alpha \quad ; \quad \alpha \geq 0 \quad , \quad (2.315)$$

¹⁷In most problems, the variables seem to be chosen in such a way that $\alpha = 2/3$. This is the case for the probability distributions of Earthquakes as a function of their energy (Gutenberg-Richter law, see figure 2.25), or of the probability distribution of meteorites hitting the Earth as a function of their volume (see figure 2.28).

the probability on an interval is $P(U_1 \leq U \leq U_2) = \int_{U_1}^{U_2} \frac{dU}{U} f_U(U)$, and one typically uses the norming condition $\int_0^{U_0} \frac{dU}{U} f_U(U) = 1$, where U_0 is some selected point. Using a variable $Y = X^n$, one arrives at the volumetric probability

$$f_Y(Y) = k' Y^{-\beta} \quad ; \quad \beta = \frac{\alpha}{n} \geq 0 \quad . \quad (2.316)$$

Example 2.25 Consider a coordinate x such that the displacement between two points is $s_x(x', x) = x' - x$. Then, the exponential distribution 2.311 takes the form

$$\boxed{f_x(x) = \alpha \exp(-\alpha(x - x_0)) \quad ; \quad \alpha \geq 0 \quad .} \quad (2.317)$$

As, here, $ds(s) = ds$, the probability of an interval is $P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} dx f_x(x)$, and $f_x(x)$ is normed by

$$\int_{x_0}^{+\infty} dx f_x(x) = 1 \quad . \quad (2.318)$$

With a variable $u = -x$, equation 2.317 becomes

$$f_u(u) = \alpha \exp(\alpha(u - u_0)) \quad ; \quad \alpha \geq 0 \quad , \quad (2.319)$$

and the norming condition is $\int_{-\infty}^{u_0} du f_u(u) = 1$. For the plotting of these volumetric probabilities, sometimes a logarithmic 'vertical axis' is used, as suggested in figure 2.24. Note that via a logarithmic change of variables $x = \log(X/K)$ (where K is some constant) this example is identical to the example 2.24. The two volumetric probabilities 2.312 and 2.317 represent the same exponential distribution.

Note: mention here figure 2.24.

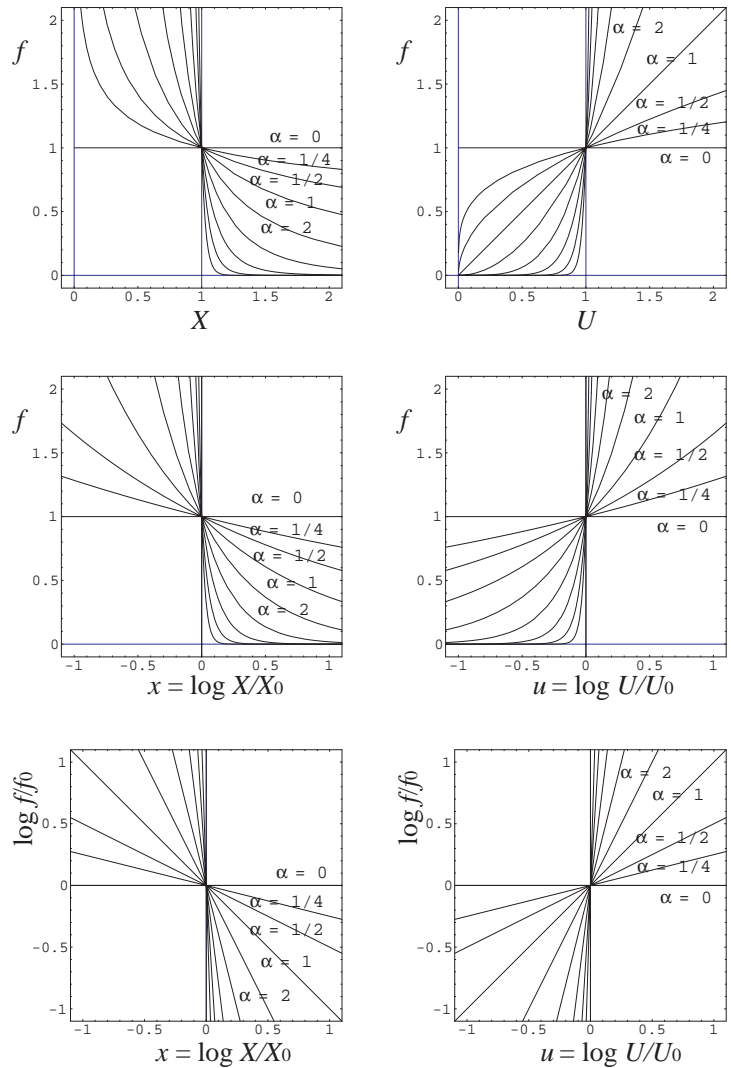


Figure 2.24: Plots of exponential distribution for different definitions of the variables. Top: The power functions $f_X(X) = 1/X^{-\alpha}$, and $f_U(U) = 1/U^\alpha$. Middle: Using logarithmic variables x and u , one has the exponential functions $f_x(x) = \exp(-\alpha x)$ and $f_u(u) = \exp(\alpha u)$. Bottom: the ordinate is also represented using a logarithmic variable, this giving the typical log-log linear functions.

2.8.6.2 Example: Distribution of Earthquakes

The historically first example of power law distribution is the distribution of energies of Earthquakes (the famous Gutenberg-Richter law).

An earthquake can be characterized by the seismic energy generated, E , or by the moment corresponding to the dislocation, that I denote here¹⁸ M . As a rough approximation, the moment is given by the product $M = \nu \ell S$, where ν is the elastic shear modulus of the medium, ℓ the average displacement between the two sides of the fault, and S is the faults' surface (Aki and Richards, 1980).

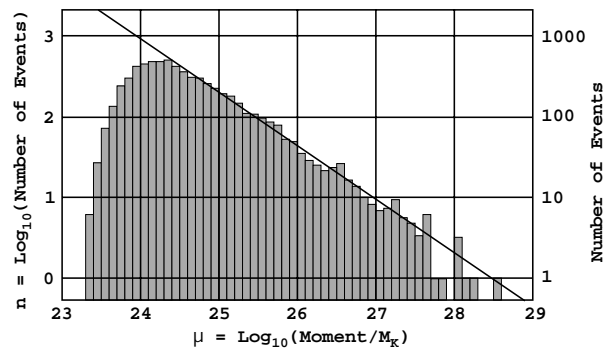
Figure 2.25 shows the distribution of earthquakes in the Earth. As the same logarithmic base (of 10) has been chosen in both axes, the slope of the line approximating the histogram (which is quite close to $-2/3$) directly leads to the power of the power law (Pareto) distribution. The volumetric probability $f(M)$ representing the distribution of earthquakes in the Earth is

$$f(M) = \frac{k}{M^{2/3}}, \quad (2.320)$$

where k is a constant. Kanamori (1977) pointed that the moment and the seismic energy liberated are roughly proportional: $M \approx 2.0 \cdot 10^4 E$ (energy and moment have the same physical dimensions). This implies that the volumetric probability as a function of the energy has the same form as for the moment:

$$g(E) = \frac{k'}{E^{2/3}}. \quad (2.321)$$

Figure 2.25: Histogram of the number of earthquakes (in base 10 logarithmic scale) recorded by the global seismological networks in a period of xxx years, as a function of the logarithmic seismic moment (adapted from Lay and Wallace, 1995). More precisely, the quantity in the horizontal axis is $\mu = \log_{10}(M/M_K)$, where M is the seismic moment, and $M_K = 10^7 \text{ J} = 1 \text{ erg}$ is a constant, whose value is arbitrarily taken equal the unit of moment (and of energy) in the cgs system of units. [note: Ask for the permission to publish this figure.]



2.8.6.3 Example: Shapes at the Surface of the Earth.

Note: mention here figure 2.26.

2.8.6.4 Example: Size of oil fields

Note: mention here figure 2.27.

¹⁸It is traditionally denoted M_0 .

Figure 2.26: Wessel and Smith (1996) have compiled a high-resolution shoreline data, and have processed it to suppress erratic points and crossing segments. The shorelines are closed polygons, and they are classified in 4 levels: ocean boundaries, lake boundaries, islands-in-lake boundaries and pond-in-island-in-lake boundaries. The 180,496 polygons they encountered had the size distribution shown at the right (the approximate numbers are in the quoted paper, the exact numbers were kindly sent to me by Wessel). A line of slope is $-2/3$ is suggested in the figure.

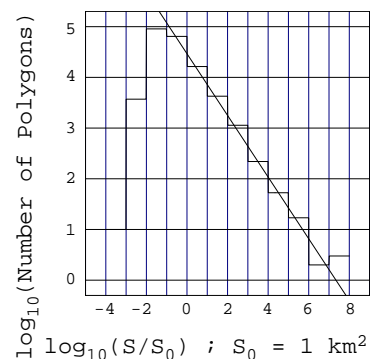
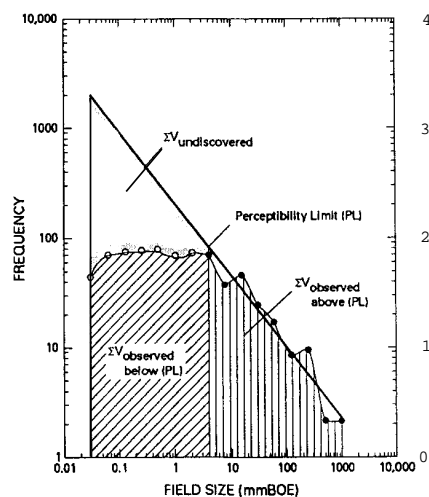


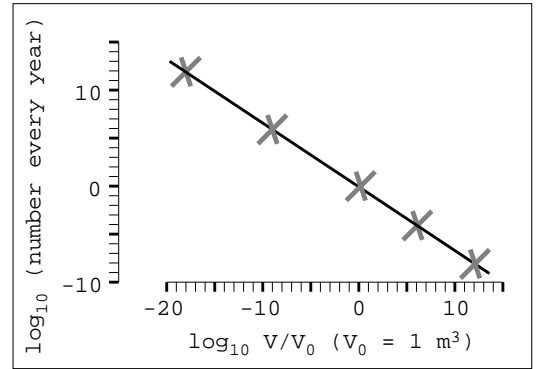
Figure 2.27: Histogram of the sizes of oil fields in a region of Texas. The horizontal axis corresponds, with a logarithmic scale, to the ‘millions of Barrels of Oil Equivalent’ (mmBOE). Extracted from chapter 2 (The fractal size and spatial distribution of hydrocarbon accumulation, by Christopher C. Barton and Christopher H. Scholz) of the book “Fractals in petroleum geology and Earth processes”, edited by Christopher C. Barton and Paul R. La Pointe, Plenum Press, New York and London, 1995. [note: ask for the permission to publish this figure]. The slope of the straight line is $-2/3$, comparable to the value found with the data of Wessel & Smith (figure 2.26).



2.8.6.5 Example: Meteorites

Note: mention here figure 2.28.

Figure 2.28: The approximate number of meteorites falling on Earth every year is distributed as follows: 10^{12} meteorites with a diameter of 10^{-3} mm, 10^6 with a diameter 1 mm, 1 with a diameter 1 m, 10^{-4} with a diameter 100 m, and 10^{-8} with a diameter 10 km. The statement is loopy, and I have extracted it from the general press. It is nevertheless clear that a log-log plot of this ‘histogram’ gives a linear trend with a slope equal to -2 . Rather, transforming the diameter D into volume $V = D^3$ (which is proportional to mass), gives the ‘histogram’ at the right, with a slope of $-2/3$.



2.8.7 Appendix: Spherical Gaussian Distribution

The simplest probabilistic distribution over the circle and over the surface of the sphere are the von Mises and the Fisher probability distributions, respectively.

2.8.7.1 The von Mises Distribution

As already mentioned in example 2.5, and demonstrated in section 2.8.7.3 here below, the conditional volumetric probability induced over the unit circle by a 2D Gaussian is

$$f(\lambda) = k \exp\left(\frac{\sin \lambda}{\sigma^2}\right) . \quad (2.322)$$

The constant k is to be fixed by the normalization condition $\int_0^{2\pi} d\varphi f(\varphi) = 1$, this giving

$$k = \frac{1}{2\pi I_0(1/\sigma^2)} , \quad (2.323)$$

where $I_0(\cdot)$ is the modified Bessel function of order zero.

Figure 2.29: The circular (von Mises) distribution corresponds to the intersection of a 2D Gaussian by a circle passing by the center of the Gaussian. Here, the unit circle has been represented, and two Gaussians with standard deviations $\sigma = 1$ (left) and $\sigma = 1/2$ (right). In fact, this is my preferred representation of the von Mises distribution, rather than the conventional functional display of figure 2.30.

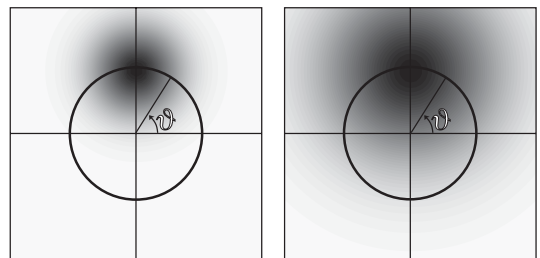
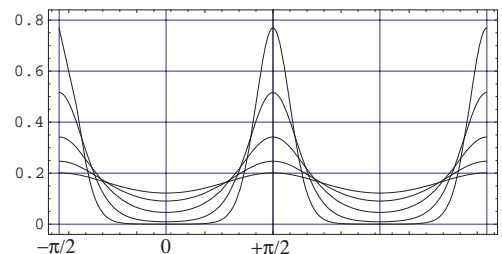


Figure 2.30: The circular (von Mises) distribution, drawn for two full periods, centered at zero, and with values of σ equal to $2, \sqrt{2}, 1, 1/\sqrt{2}, 1/2$ (from smooth to sharp).



2.8.7.2 The Fisher Probability Distribution

Note: mention here Fisher (1953).

As already mentioned in example 2.5, and demonstrated in section 2.8.7.3 here below, the conditional volumetric probability induced over the surface of a sphere by a 3D Gaussian is, using geographical coordinates

$$f(\varphi, \lambda) = k \exp\left(\frac{\sin \lambda}{\sigma^2}\right) . \quad (2.324)$$

We can normalize this volumetric probability by

$$\int dS(\varphi, \lambda) f(\varphi, \lambda) = 1, \quad (2.325)$$

with $dS(\varphi, \lambda) = \cos \lambda d\varphi d\lambda$. This gives

$$k = \frac{1}{4\pi\chi(1/\sigma^2)}, \quad (2.326)$$

where

$$\chi(x) = \frac{\sinh(x)}{x}. \quad (2.327)$$

2.8.7.3 Appendix: Fisher from Gaussian (Demonstration)

Let us demonstrate here that the Fisher probability distribution is obtained as the conditional of a Gaussian probability distribution over a sphere. As the demonstration is independent of the dimension of the space, let us take an space with n dimensions, where the (generalized) geographical coordinates are

$$\begin{aligned} x^1 &= r \cos \lambda \cos \lambda_2 \cos \lambda_3 \cos \lambda_4 \dots \cos \lambda_{n-2} \cos \lambda_{n-1} \\ x^2 &= r \cos \lambda \cos \lambda_2 \cos \lambda_3 \cos \lambda_4 \dots \cos \lambda_{n-2} \sin \lambda_{n-1} \\ &\dots = \dots \\ x^{n-2} &= r \cos \lambda \cos \lambda_2 \sin \lambda_3 \\ x^{n-1} &= r \cos \lambda \sin \lambda_2 \\ x^n &= r \sin \lambda \end{aligned} \quad (2.328)$$

We shall consider the unit sphere at the origin, and an isotropic Gaussian probability distribution with standard deviation σ , with its center along the x^n axis, at position $x^n = 1$.

The Gaussian volumetric probability, when expressed as a function of the Cartesian coordinates is

$$f_x(x^1, \dots, x^n) = k \exp\left(-\frac{1}{2\sigma^2} \left((x^1)^2 + (x^2)^2 + \dots + (x^{n-1})^2 + (x^n - 1)^2\right)\right). \quad (2.329)$$

As the volumetric probability is an invariant, to express it using the geographical coordinates we just need to use the replacements 2.328, to obtain

$$f_r(r, \lambda, \lambda', \dots) = k \exp\left(-\frac{1}{2\sigma^2} \left(r^2 \cos^2 \lambda + (r \sin \lambda - 1)^2\right)\right), \quad (2.330)$$

i.e.,

$$f_r(r, \lambda, \lambda', \dots) = k \exp\left(-\frac{1}{2\sigma^2} (r^2 + 1 - 2r \sin \lambda)\right). \quad (2.331)$$

The condition to be on the sphere is just

$$r = 1, \quad (2.332)$$

so that the conditional volumetric probability, as given in equation 2.95, is just obtained (up to a multiplicative constant) by setting $r = 1$ in equation 2.331,

$$f(\lambda, \lambda', \dots) = k' \exp\left(\frac{\sin \lambda - 1}{\sigma^2}\right) , \quad (2.333)$$

i.e., absorbing the constant $\exp(1/\sigma^2)$,

$$f(\lambda, \lambda', \dots) = k'' \exp\left(\frac{\sin \lambda}{\sigma^2}\right) . \quad (2.334)$$

This volumetric probability corresponds to the n -dimensional version of the Fisher distribution. Its expression is identical in all dimensions, only the norming constant depends on the dimension of the space.

2.8.8 Appendix: Probability Distributions for Tensors

In this appendix we consider a symmetric second rank tensor, like the stress tensor $\boldsymbol{\sigma}$ of continuum mechanics.

A symmetric tensor, $\sigma_{ij} = \sigma_{ji}$, has only six degrees of freedom, while it has nine components. It is important, for the development that follows, to agree in a proper definition of a set of ‘independent components’. This can be done, for instance, by defining the following six-dimensional basis for symmetric tensors

$$\mathbf{e}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} ; \quad \mathbf{e}_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} ; \quad \mathbf{e}_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (2.335)$$

$$\mathbf{e}_4 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} ; \quad \mathbf{e}_5 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} ; \quad \mathbf{e}_6 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} . \quad (2.336)$$

Then, any symmetric tensor can be written as

$$\boldsymbol{\sigma} = s^\alpha \mathbf{e}_\alpha , \quad (2.337)$$

and the six values s^α are the six ‘independent components’ of the tensor, in terms of which the tensor writes

$$\boldsymbol{\sigma} = \begin{pmatrix} s^1 & s^6/\sqrt{2} & s^5/\sqrt{2} \\ s^6/\sqrt{2} & s^2 & s^4/\sqrt{2} \\ s^5/\sqrt{2} & s^4/\sqrt{2} & s^3 \end{pmatrix} . \quad (2.338)$$

The only natural definition of distance between two tensors is the norm of their difference, so we can write

$$D(\boldsymbol{\sigma}_2, \boldsymbol{\sigma}_1) = \| \boldsymbol{\sigma}_2 - \boldsymbol{\sigma}_1 \| , \quad (2.339)$$

where the norm of a tensor $\boldsymbol{\sigma}$ is¹⁹

$$\| \boldsymbol{\sigma} \| = \sqrt{\sigma_{ij} \sigma^{ji}} . \quad (2.340)$$

The basis in equation 2.336 is normed with respect to this norm²⁰. In terms of the independent components in expression 2.338 the norm of a tensor simply becomes

$$\| \boldsymbol{\sigma} \| = \sqrt{(s^1)^2 + (s^2)^2 + (s^3)^2 + (s^4)^2 + (s^5)^2 + (s^6)^2} , \quad (2.341)$$

this showing that the six components s^α play the role of Cartesian coordinates of this 6D space of tensors.

A Gaussian volumetric probability in this space has then, obviously, the form

$$f_{\mathbf{s}}(\mathbf{s}) = k \exp \left(- \frac{\sum_{\alpha=1}^6 (s^\alpha - s_0^\alpha)^2}{2 \rho^2} \right) , \quad (2.342)$$

¹⁹Of course, as, here, $\sigma_{ij} = \sigma_{ji}$ one can also write $\| \boldsymbol{\sigma} \| = \sqrt{\sigma_{ij} \sigma^{ij}}$, but this expression is only valid for symmetric tensors, while the expression 2.340 is generally valid.

²⁰It is also orthonormed, with the obvious definition of scalar product from which this norm derives.

or, more generally,

$$f_{\mathbf{s}}(\mathbf{s}) = k \exp \left(-\frac{1}{2\rho^2} (s^\alpha - s_0^\alpha) W_{\alpha\beta} (s^\beta - s_0^\beta) \right) . \quad (2.343)$$

It is easy to find probabilistic models for tensors, when we choose as coordinates the independent components of the tensor, as this Gaussian example suggests. But a symmetric second rank tensor may also be described using its three eigenvalues $\{\lambda_1, \lambda_2, \lambda_3\}$ and the three Euler angles $\{\psi, \theta, \varphi\}$ defining the eigenvector's directions

$$\begin{pmatrix} s^1 & s^6/\sqrt{2} & s^5/\sqrt{2} \\ s^6/\sqrt{2} & s^2 & s^4/\sqrt{2} \\ s^5/\sqrt{2} & s^4/\sqrt{2} & s^3 \end{pmatrix} = \mathbf{R}(\psi) \mathbf{R}(\theta) \mathbf{R}(\varphi) \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \mathbf{R}(\varphi)^T \mathbf{R}(\theta)^T \mathbf{R}(\psi)^T , \quad (2.344)$$

where \mathbf{R} denotes the usual rotation matrix. Some care is required when using the coordinates $\{\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi\}$.

To write a Gaussian volumetric probability in terms on eigenvectors and eigendirections only requires, of course, to insert in the $f_{\mathbf{s}}(\mathbf{s})$ of equation 2.343 the expression 2.344 giving the tensor components as a function of the eigenvectors and eigendirections (we consider volumetric probabilities—that are invariant—and not probability densities—that would require an extra multiplication by the Jacobian determinant of the transformation—),

$$f(\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi) = f_{\mathbf{s}}(s^1, s^2, s^3, s^4, s^5, s^6) . \quad (2.345)$$

But then, of course, we still need how to integrate in the space using these new coordinates, in order to evaluate probabilities.

Before facing this problem, let us remark that it is the replacement in equation 2.343 of the components s^α in terms of the eigenvalues and eigendirections of the tensor that shall express a Gaussian probability distribution in terms of the variables $\{\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi\}$. Using a function a would ‘look Gaussian’ in the variables $\{\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi\}$ would not correspond to a Gaussian probability distribution, in the sense of section 2.8.4.

The Jacobian determinant of the transformation $\{s^1, s^2, s^3, s^4, s^5, s^6\} \rightleftharpoons \{\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi\}$ can be obtained using a direct computation, that gives²¹

$$\left| \frac{\partial(s^1, s^2, s^3, s^4, s^5, s^6)}{\partial(\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi)} \right| = (\lambda_1 - \lambda_2) (\lambda_2 - \lambda_3) (\lambda_3 - \lambda_1) \sin \theta . \quad (2.346)$$

The capacity elements in the two systems of coordinates are

$$\begin{aligned} d\underline{v}_{\mathbf{s}}(s^1, s^2, s^3, s^4, s^5, s^6) &= ds^1 \wedge ds^2 \wedge ds^3 \wedge ds^4 \wedge ds^5 \wedge ds^6 \\ d\underline{v}(\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi) &= d\lambda_1 \wedge d\lambda_2 \wedge d\lambda_3 \wedge d\psi \wedge d\theta \wedge d\varphi . \end{aligned} \quad (2.347)$$

As the coordinates $\{s^\alpha\}$ are Cartesian, the volume element of the space is numerically identical to the capacity element,

$$dv_{\mathbf{s}}(s^1, s^2, s^3, s^4, s^5, s^6) = d\underline{v}_{\mathbf{s}}(s^1, s^2, s^3, s^4, s^5, s^6) , \quad (2.348)$$

²¹If instead of the 3 Euler angles, we take 3 rotations around the three coordinate axes, the $\sin \theta$ here above becomes replaced by the cosine of the second angle. This is consistent with the formula by Xu and Grafarend (1997).

but in the coordinates $\{\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi\}$ the volume element and the capacity are related via the Jacobian determinant in equation 2.346,

$$dv(\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi) = (\lambda_1 - \lambda_2)(\lambda_2 - \lambda_3)(\lambda_3 - \lambda_1) \sin \theta \underline{d}v(\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi) \quad . \quad (2.349)$$

Then, while the evaluation of a probability in the variables $\{s^1, s^2, s^3, s^4, s^5, s^6\}$ should be done via

$$\begin{aligned} P &= \int dv_{\mathbf{s}}(s^1, s^2, s^3, s^4, s^5, s^6) f_{\mathbf{s}}(s^1, s^2, s^3, s^4, s^5, s^6) \\ &= \int ds^1 \wedge ds^2 \wedge ds^3 \wedge ds^4 \wedge ds^5 \wedge ds^6 f_{\mathbf{s}}(s^1, s^2, s^3, s^4, s^5, s^6) \quad , \end{aligned} \quad (2.350)$$

in the variables $\{\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi\}$ it should be done via

$$\begin{aligned} P &= \int dv(\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi) f(\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi) \\ &= \int d\lambda_1 \wedge d\lambda_2 \wedge d\lambda_3 \wedge d\psi \wedge d\theta \wedge d\varphi \quad \times \\ &\quad \times (\lambda_1 - \lambda_2)(\lambda_2 - \lambda_3)(\lambda_3 - \lambda_1) \sin \theta f(\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi) \quad . \end{aligned} \quad (2.351)$$

To conclude this appendix, we may remark that the homogeneous probability distribution (defined as the one who is ‘proportional to the volume distribution’) is obtained by taking both $f_{\mathbf{s}}(s^1, s^2, s^3, s^4, s^5, s^6)$ and $f(\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi)$ as constants.

[Note: I should explain somewhere that there is a complication when, instead of considering ‘a tensor like the stress tensor’ one consider a positive tensor (like an electric permittivity tensor). The treatment above applies approximately to the logarithm of such a tensor.]

2.8.9 Appendix: Determinant of a Partitioned Matrix

Using well known properties of matrix algebra (e.g., Lütkepohl, 1996), the determinant of a partitioned matrix can be expressed as

$$\det \begin{pmatrix} \mathbf{g}^{rr} & \mathbf{g}^{rs} \\ \mathbf{g}^{sr} & \mathbf{g}^{ss} \end{pmatrix} = \det \mathbf{g}^{rr} \det \left(\mathbf{g}^{ss} - \mathbf{g}^{sr} \mathbf{g}^{rr^{-1}} \mathbf{g}^{rs} \right) . \quad (2.352)$$

2.8.10 Appendix: The Borel ‘Paradox’

[Note: This appendix has to be updated.]

A description of the paradox is given, for instance, by Kolmogorov (1933), in his *Foundations of the Theory of Probability* (see figure 2.31).

Figure 2.31: 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”. His explanation is not profound: instead of discussing the behaviour of a conditional probability density under a change of variables, it concerns the interpretation of a probability density over the sphere when using spherical coordinates. I do not agree with the conclusion (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 \mathfrak{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 of set A . Let us now choose two diametrically opposite points for our poles, so that each meridian circle will be uniquely defined by the longitude ψ , $0 \leq \psi < \pi$. Since ψ varies from 0 only to π , — in other words, we are considering *complete* meridian circles (and not merely semicircles) — the latitude θ 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 θ , $-\pi \leq \theta < +\pi$, for a given longitude ψ .

It is easy to calculate that

$$P_{\psi}\{\theta_1 \leq \theta < \theta_2\} = \frac{1}{4} \int_{\theta_1}^{\theta_2} |\cos \theta| d\theta.$$

The probability distribution of θ for a given ψ is not uniform.

If we assume that the conditional probability distribution of θ “with the hypothesis that ξ 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 can obtain a probability distribution for θ 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 $\bar{f}(u, v)$ 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 α , 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 $\bar{f}(u, v)$ 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 $\bar{g}(\alpha)$ is the conditional probability density induced on the circle by the probability density $\bar{f}(u, v)$, 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 α , the conditional probability density over the circle must be

$$\bar{g}(\alpha) = \frac{1}{2\pi}, \quad (2.353)$$

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 $\{\lambda, \varphi\}$ are used, where λ is the latitude (rather than the colatitude θ), so the domains of definition of the variables are

$$-\pi/2 < \lambda \leq +\pi/2 \quad ; \quad -\pi < \varphi \leq +\pi \quad . \quad (2.354)$$

As the surface element is $dS(\lambda, \varphi) = \cos \lambda d\lambda d\varphi$, the homogeneous probability distribution over the surface of the sphere is represented, in spherical coordinates, by the probability density

$$\bar{f}(\lambda, \varphi) = \frac{1}{4\pi} \cos \lambda, \quad (2.355)$$

and we satisfy the normalization condition

$$\int_{-\pi/2}^{+\pi/2} d\lambda \int_{-\pi}^{+\pi} d\varphi \bar{f}(\lambda, \varphi) = 1 \quad . \quad (2.356)$$

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, 4π), so the probability density in equation 2.355 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 ?? 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),

$$ds^2 = d\lambda^2 + \cos^2 \lambda d\varphi^2, \quad (2.357)$$

which gives the metric components

$$g_{\lambda\lambda}(\lambda, \varphi) = 1 \quad ; \quad g_{\varphi\varphi}(\lambda, \varphi) = \cos^2 \lambda \quad (2.358)$$

and the surface element

$$dS(\lambda, \varphi) = \cos \lambda d\lambda d\varphi \quad . \quad (2.359)$$

Letting $\bar{f}(\lambda, \varphi)$ 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 ??,

$$\bar{f}_\lambda(\lambda|\varphi = \varphi_0) = k \frac{\bar{f}(\lambda, \varphi_0)}{\sqrt{g_{\varphi\varphi}(\lambda, \varphi_0)}} \quad . \quad (2.360)$$

In our case, using the second of equations 2.358

$$\bar{f}_\lambda(\lambda|\varphi = \varphi_0) = k \frac{\bar{f}(\lambda, \varphi_0)}{\cos \lambda} \quad , \quad (2.361)$$

or, in normalized version,

$$\bar{f}_\lambda(\lambda|\varphi = \varphi_0) = \frac{\bar{f}(\lambda, \varphi_0)/\cos \lambda}{\int_{-\pi/2}^{+\pi/2} d\lambda \bar{f}(\lambda, \varphi_0)/\cos \lambda} \quad . \quad (2.362)$$

If the original probability density $\bar{f}(\lambda, \varphi)$ represents an homogeneous probability, then it must be proportional to the surface element dS (equation 2.359), so, in normalized form, the homogeneous probability density is

$$\bar{f}(\lambda, \varphi) = \frac{1}{4\pi} \cos \lambda \quad . \quad (2.363)$$

Then, equation 2.361 gives

$$\bar{f}_\lambda(\lambda|\varphi = \varphi_0) = \frac{1}{\pi} \quad . \quad (2.364)$$

We see that this conditional probability density is constant²².

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 2.360, the conditional probability density is ‘defined’ by

$$\bar{f}_\lambda(\lambda|\varphi = \varphi_0) = k \bar{f}(\lambda, \varphi_0) = \frac{\bar{f}(\lambda, \varphi_0)}{\int_{-\pi/2}^{+\pi/2} d\lambda \bar{f}(\lambda, \varphi_0)/\cos \lambda} \quad \text{wrong definition} \quad , \quad (2.365)$$

this leading, in the considered case, to the conditional probability density

$$\bar{f}_\lambda(\lambda|\varphi = \varphi_0) = \frac{\cos \lambda}{2} \quad \text{wrong result} \quad . \quad (2.366)$$

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 2.31). 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*

²²This constant value is $1/\pi$ if we consider half a meridian, or it is $1/2\pi$ if we consider a whole meridian.

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 be 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 text.

2.8.11 Appendix: Axioms for the Sum and the Product

2.8.11.1 The Sum

I guess that the two defining axioms for the union of two probabilities are

$$P(\mathcal{A}) = 0 \quad \text{AND} \quad Q(\mathcal{A}) = 0 \quad \implies \quad (P \cup Q)(\mathcal{A}) = 0 \quad (2.367)$$

and

$$P(\mathcal{A}) \neq 0 \quad \text{OR} \quad Q(\mathcal{A}) \neq 0 \quad \implies \quad (P \cup Q)(\mathcal{A}) \neq 0 \quad . \quad (2.368)$$

But the last property is equivalent to its negation,

$$P(\mathcal{A}) = 0 \quad \text{AND} \quad Q(\mathcal{A}) = 0 \quad \iff \quad (P \cup Q)(\mathcal{A}) = 0 \quad , \quad (2.369)$$

and this can be reunited with the first property, to give the single axiom

$$\boxed{P(\mathcal{A}) = 0 \quad \text{AND} \quad Q(\mathcal{A}) = 0 \quad \iff \quad (P \cup Q)(\mathcal{A}) = 0 \quad .} \quad (2.370)$$

2.8.11.2 The product

We only have the axiom

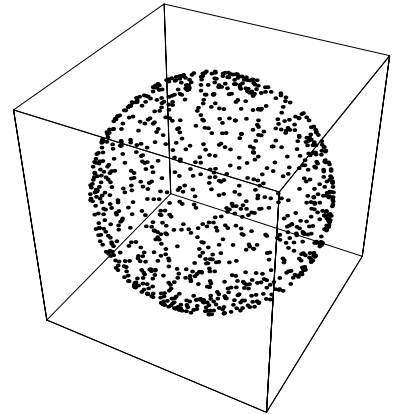
$$P(\mathcal{A}) = 0 \quad \text{OR} \quad Q(\mathcal{A}) = 0 \quad \implies \quad (P \cap Q)(\mathcal{A}) = 0 \quad . \quad (2.371)$$

and, of course, its (equivalent) negation

$$P(\mathcal{A}) \neq 0 \quad \text{AND} \quad Q(\mathcal{A}) \neq 0 \quad \iff \quad (P \cap Q)(\mathcal{A}) \neq 0 \quad (2.372)$$

2.8.12 Appendix: Random Points on the Surface of the Sphere

Figure 2.32: 1000 random points on the surface of the sphere.



Note: Figure 2.32 has been generated using the following Mathematica code:

```
spc[t_,p_,r_:1] := r {Sqrt[1-t^2] Cos[p], Sqrt[1-t^2] Sin[p], t}
```

```
Show[Graphics3D[Table[Point[spc[Random[Real,{-1,1}],  
Random[Real,{0,2Pi}]]],{1000}]]]
```

Figure 2.33: A geodesic dome dividing the surface of the sphere into regions with approximately the same area.

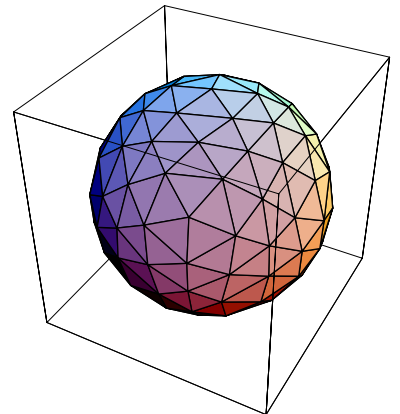
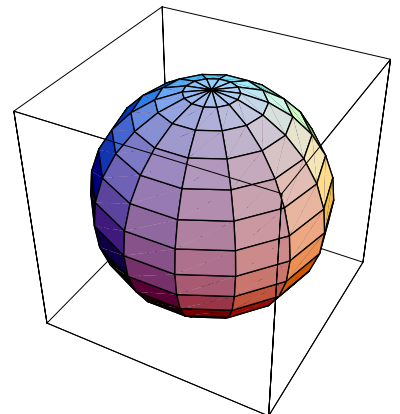


Figure 2.34: The coordinate division of the surface of the sphere.



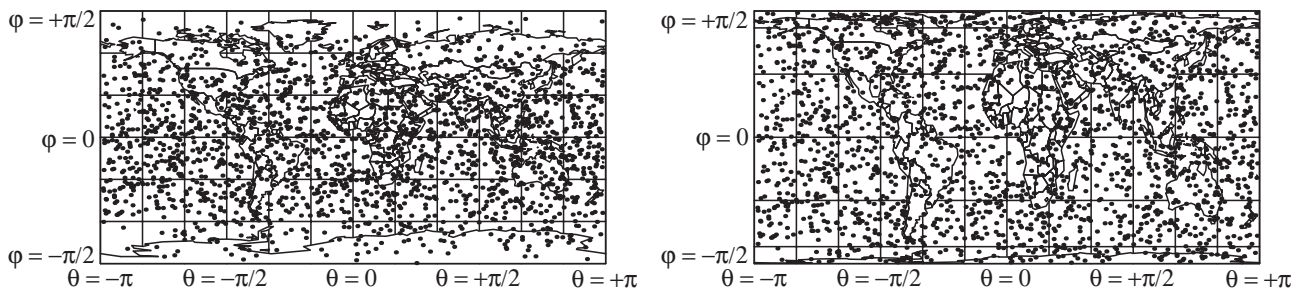


Figure 2.35: Map representation of a random homogeneous distribution of points at the surface of the sphere. At the left, the naïve division of the surface of the sphere using constant increments of the coordinates. At the right, the cylindrical equal-area projection. Counting the points in each ‘rectangle’ gives, at the left, the probability density of points. At the right, the volumetric probability.

2.8.13 Appendix: Histograms for the Volumetric Mass of Rocks

Figure 2.36: Histogram of the volumetric mass for the 557 minerals listed in the *Handbook of Physical Properties of rocks* (Johnson and Olhoeft, 1984). A logarithmic axis is used that represents the variable $u = \log_{10}(\rho/K)$, with $K = 1 \text{ g/cm}^3$. Superposed to the histogram is the normal function with mean 0.60 and standard deviation 0.23. The vertical lines correspond to successive deviations multiples of the standard deviation. See the lognormal function in 2.37).

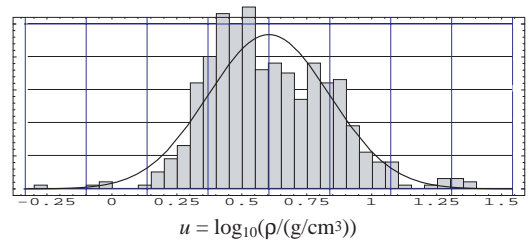


Figure 2.37: A naïve version of the histogram in figure 2.36, using an axis labeled in volumetric mass.

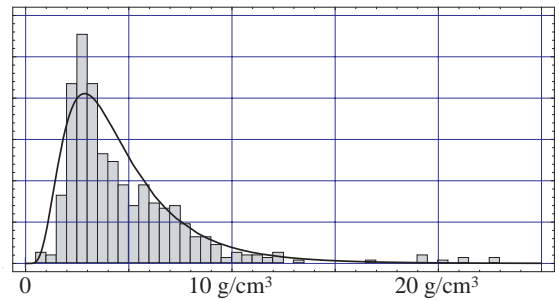
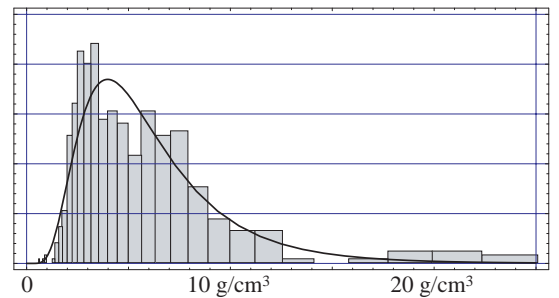


Figure 2.38: A third version of the histogram, obtained using intervals of constant length $\delta\rho/\rho$.



Chapter 3

Monte Carlo Sampling Methods

Note: write here a small introduction to the chapter.

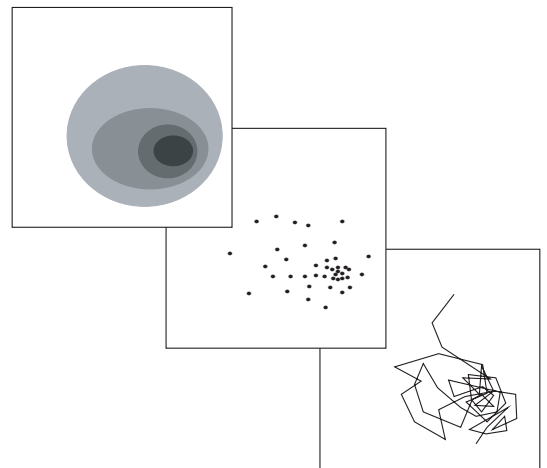
3.1 Introduction

When a probability distribution has been defined, we have to face the problem of how to ‘use’ it. The definition of some ‘central estimators’ (like the mean or the median) and some ‘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), then we can directly represent the associated probability density¹. This is trivial in one or two dimensions. It is easy in three dimensions, using, for instance, virtual reality software. 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 3.1. 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 left². 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 multidimensional, each ‘point’ may corresponds 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 nice drawing with plenty of colors. Then a collection of ‘points’ is a collections of such drawings. Our experience shows that, given such 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 3.1: 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 structure?* To answer this, we can make an histogram of the depth of the given geological

¹Or, best, the associated volumetric probability.

²There are two ways for making an histogram. If the space is divided in cells with constant coordinate differences dx^1, dx^2, \dots , then the limit converges to the probability density. If, instead, the space is divided in cells of constant volume dV , then the limit converges to the volumetric probability.

structure over the collection of random models, and the histogram *is* the answer to the question. *Which 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 to a large number of randomly generated models in order to intuitively apprehend the basic properties of the probability distribution, followed by precise computations of the probability 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 3.1, 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 core³. At present, Monte Carlo methods are typically implemented on digital computers, and are based on the pseudorandom generation of numbers⁴. 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 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. That this is not only important, but in fact vital in high dimensional spaces, can be seen in figure 3.2, where the failure of a plain Monte Carlo sampling (one that samples the space uniformly) in high dimensional spaces is made clear.

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

3.2 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 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.

³Note: Cite here the example of Buffon, and a couple of other simple examples.

⁴I.e., series of numbers that appear random if tested with any reasonable statistical test. Note: cite here some references (Press, etc.).

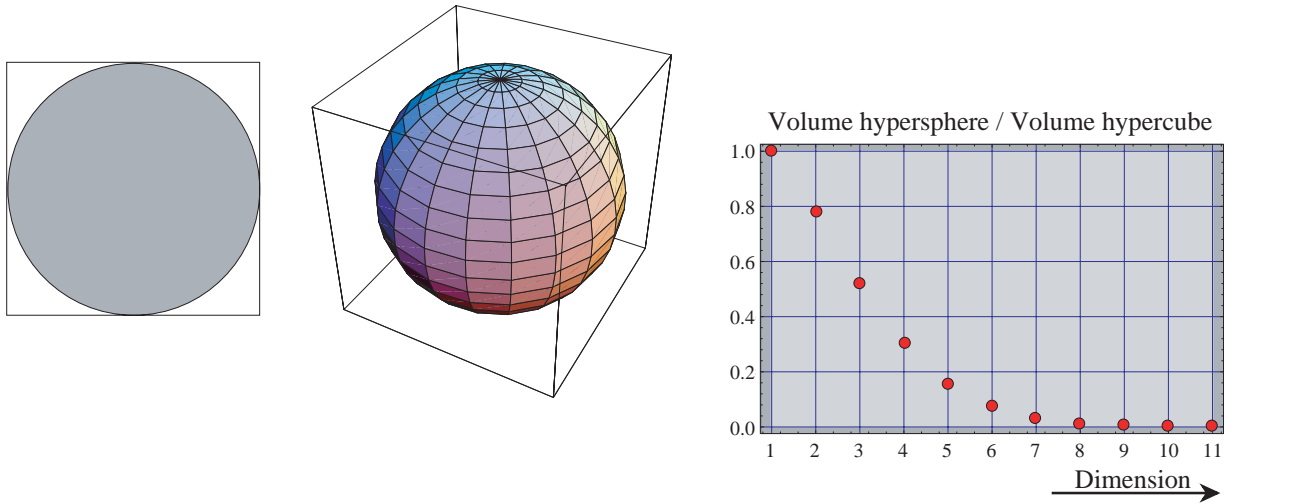


Figure 3.2: Consider a square and the inscribed circle. If the circle’s surface is πR^2 , that of the square is $(2R)^2$. If we generate a random point inside the square, with homogeneous probability distribution, the probability of hitting the circle equals the ratio of the surfaces, i.e., $P = \pi/4$. We can do the same in 3D, but, in this case, the ratio of volumes is $P = \pi/6$: the probability of hitting the target is smaller in 3D than in 2D. This probability tends dramatically to zero when the dimension of the space increases. For instance, in dimension 100, the probability of hitting the hypersphere inscribed in the hypercube is $P = 1.9 \cdot 10^{-70}$, what means that it is practically impossible to hit the target ‘by chance’. The formulas at the top give the volume of an hypersphere of radius R in a space of dimension $2n$ or $2n + 1$ (the formula is not the same for spaces with even or odd dimension), and the volume of an hypercube with sides of length $2R$. The graph at the bottom shows the evolution, as a function of the dimension of the space, of the ratio between the volume of the hypersphere and the volume of the hypercube. In large dimension, the hypersphere fills a negligible amount of the hypercube.

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.

Before going into more complex sampling situations, we should mention that there exist methods for sampling probability densities that can be described by an explicit mathematical expressions. Information on some of the most important of these methods can be found in appendix 3.10.3.

Sampling in cases where only point values of the probability density are available upon request can be done by means of Monte Carlo algorithms based on *random walks*. In the following, we shall describe the essential properties of random walks performing the so-called *importance sampling*.

3.3 Modification of Random Walks

Assume here that we can start with a random walk that samples some probability density $f(\mathbf{x})$, and have the goal of having a random walk that samples the probability density

$$h(\mathbf{x}) = k f(\mathbf{x}) \frac{g(\mathbf{x})}{\mu(\mathbf{x})} . \quad (3.1)$$

Call \mathbf{x}_i the ‘current point’. With this current point as starting point, run one step of the random walk that unimpeded would sample the probability density $f(\mathbf{x})$, and generate a ‘test point’ \mathbf{x}_{test} . Compute the value

$$q_{\text{test}} = \frac{g(\mathbf{x}_{\text{test}})}{\mu(\mathbf{x}_{\text{test}})} . \quad (3.2)$$

If that value is ‘high enough’, let that point ‘survive’. If q_{test} is not ‘high enough’, discard this point and generate another one (making another step of the random walk sampling the prior probability density $f(\mathbf{x})$), using again the ‘current point’ x_i as starting point).

There are many criteria for deciding when a point should survive or should be discarded, all of them resulting in a collection of ‘surviving points’ that are samples of the target probability density $h(\mathbf{x})$. For instance, if we know the maximum possible value of the ratio $g(\mathbf{x})/\mu(\mathbf{x})$, say q_{max} , then define

$$P_{\text{test}} = \frac{q_{\text{test}}}{q_{\text{max}}} , \quad (3.3)$$

and give the point x_{test} the probability P_{test} of survival (note that $0 < P_{\text{test}} < 1$). It is intuitively obvious why the random walk modified using such a criterion produces a random walk that actually samples the probability density $h(\mathbf{x})$ defined by equation 3.1.

Among the many criteria that can be used, the by far most efficient is the *Metropolis criterion*, the criterion behind the *Metropolis Algorithm* (Metropolis et al. 1953). In the following we shall describe this algorithm in some detail.

3.4 The Metropolis Rule

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 . If that ‘proposed transition’ $\mathbf{x}_i \leftarrow \mathbf{x}_j$ is 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 :

$$P = \frac{g(\mathbf{x}_i)/\mu(\mathbf{x}_i)}{g(\mathbf{x}_j)/\mu(\mathbf{x}_j)} . \quad (3.4)$$

Then we have the following

Theorem 3.1 *The random walker samples the conjunction $h(\mathbf{x})$ of the probability densities $f(\mathbf{x})$ and $g(\mathbf{x})$*

$$h(\mathbf{x}) = k f(\mathbf{x}) \frac{g(\mathbf{x})}{\mu(\mathbf{x})} = k \frac{f(\mathbf{x}) g(\mathbf{x})}{\mu(\mathbf{x})} \quad (3.5)$$

(see appendix 3.10.2 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 un-normalizable probability densities will not cause any trouble in the application of the above algorithm.

The algorithm above is reminiscent (see appendix 3.10.2) of the Metropolis algorithm (Metropolis et al., 1953), originally designed to sample the Gibbs-Boltzmann distribution⁵. Accordingly, we will refer to the above acceptance rule as the *Metropolis rule*.

3.5 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 of $f_1(\mathbf{x})$, to propose a new point \mathbf{x}_i ,

⁵To 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 $p_{ij}^{\text{acc}} = \exp(-(E(\mathbf{x}_i) - E(\mathbf{x}_j))/T)$.

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

$$h(\mathbf{x}) = k f_1(\mathbf{x}) \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})} \dots \frac{f_n(\mathbf{x})}{\mu(\mathbf{x})} . \quad (3.6)$$

(see appendix XXX for a demonstration).

3.6 Initiating a Random Walk

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

$$h(\mathbf{x}) = k f_1(\mathbf{x}) \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})} \frac{f_3(\mathbf{x})}{\mu(\mathbf{x})} \dots , \quad (3.7)$$

and let us examine three common situations.

3.6.0.0.1 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 $f_2(\mathbf{x})/\mu(\mathbf{x})$, $f_3(\mathbf{x})/\mu(\mathbf{x}) \dots$, using the cascaded Metropolis rule, to obtain a random walk that samples $h(\mathbf{x})$.

3.6.0.0.2 We start with a random walk that samples $\mu(\mathbf{x})$: We can write equation 3.7 as

$$h(\mathbf{x}) = k \left(\left(\left(\mu(\mathbf{x}) \frac{f_1(\mathbf{x})}{\mu(\mathbf{x})} \right) \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})} \right) \dots \right) . \quad (3.8)$$

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 $f_1(\mathbf{x})/\mu(\mathbf{x})$.

3.6.0.0.3 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 probability distribution $\psi(\mathbf{x})$, we can write equation 3.7 on the form

$$h(\mathbf{x}) = k \left(\left(\left(\left(\psi(\mathbf{x}) \frac{\mu(\mathbf{x})}{\psi(\mathbf{x})} \right) \frac{f_1(\mathbf{x})}{\mu(\mathbf{x})} \right) \frac{f_2(\mathbf{x})}{\mu(\mathbf{x})} \right) \dots \right) . \quad (3.9)$$

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.

3.7 Designing Primeval Walks

What the Metropolis algorithm does is to modify some initial walk, in cascade, to produce a final random walk that samples the target probability distribution. The initial walk, that is designed ab initio, i.e., independently of the Metropolis algorithm (or any similar algorithm), may be called the *primeval walk*. We shall see below some examples where primeval walks are designed that sample the homogeneous probability distribution $\mu(\mathbf{x})$, or directly the probability density $f(\mathbf{x})$ (see equation 3.7). If we do not know how to do this, then we have to resort to using a primeval walk that samples the arbitrary function $\psi(\mathbf{x})$ mentioned above.

Example 3.1 Consider the homogeneous probability density on the 2D surface of a sphere of radius R , $\mu(\vartheta, \varphi) = \frac{R^2}{4\pi} \cos \vartheta$, where we use geographical coordinates. This distribution can be sampled by generating a value of ϑ using the probability density $\frac{1}{2\pi} \cos \vartheta$, and then a value of φ using a constant probability density. Alternatively, one could use a purely geometrical approach. [End of example.]

Example 3.2 If instead of the surface of a sphere, we have some spheroid, with spheroidal coordinates $\{\vartheta, \varphi\}$, the homogeneous probability density will have some expression $\mu(\vartheta, \varphi)$, that will not be identical to that corresponding to a sphere (see example 3.1). We may then use the function $\psi(\vartheta, \varphi) = \frac{R^2}{4\pi} \cos \vartheta$, i.e., we may start with the same primeval walk as in example 3.1, using, in the Metropolis rule, the ‘corrective step’ mentioned in section 3.6, and depending on the values $\mu(\vartheta, \varphi)/\psi(\vartheta, \varphi)$. [End of example.]

Example 3.3 If x is a one-dimensional Cartesian quantity, i.e., if the associated homogeneous probability density is constant, then, it is trivial to designate a random walk that samples it. If x is the ‘current point’, choose randomly a real number e with an arbitrary probability density that is symmetric around zero, and jump to $x + e$. The iteration of this rule produces a random walk that samples the homogeneous probability density for a Cartesian parameter, $\mu(x) = k$. [End of example.]

Example 3.4 Consider the homogeneous probability density for a temperature, $\mu(T) = 1/T$, as an example of a Jeffreys parameter. This distribution can be sampled by the following procedure. If T is the ‘current point’, choose randomly a real number e with an arbitrary probability density that is symmetric around zero, let be $Q = \exp e$, and jump⁶ to QT . The

⁶Note that if $Q > 1$, the algorithm ‘goes to the right’, while if $Q < 1$, it ‘goes to the left’.

iteration of this rule produces⁷ a random walk that samples the homogeneous probability density for a Jeffreys parameter, $\mu(T) = 1/T$. [End of example.]

Example 3.5 Consider a random walk that, when it is at point \mathbf{x}_j , chooses another point \mathbf{x}_i with a probability density $f(\mathbf{x}) = U(\mathbf{x}|\mathbf{x}_j)$, satisfying

$$U(\mathbf{x}|\mathbf{y}) = U(\mathbf{y}|\mathbf{x}) \quad . \quad (3.10)$$

Then, the random walk samples the constant probability density $f(\mathbf{x}) = k$ (see appendix ?? for a proof). [End of example.]

The reader should be warned that although the Metropolis rule would allow to use a primeval walk sampling a probability density $\psi(\mathbf{x})$ that may be quite different from the homogeneous probability density $\mu(\mathbf{x})$, this may be quite inefficient. One should not, in general, use the random walk defined in example 3.5 as a general primeval walk.

3.8 Multistep Iterations

An algorithm will converge to a unique equilibrium distribution if the random walk is irreducible. Often, it is convenient to split up an iteration in a number of steps, having their own transition probability densities, and their own transition probabilities. A typical example is a random walk in an N -dimensional Euclidian space where we are interested in dividing an iteration of the random walk into N steps, where the n -th move of the random walker is in a direction parallel to the n -th axis.

The question is now: if we want to form an iteration consisting of a series of steps, can we give a sufficient condition to be satisfied by each step such that the complete iteration has the desired convergence properties?

It is easy to see that if the individual steps in an iteration all have the same probability density $p(\mathbf{x})$ as their equilibrium probability density (not necessarily unique), then the complete iteration also has $p(\mathbf{x})$ as an equilibrium probability density. This follows from the fact that the equilibrium probability density is an eigenfunction with eigenvalue 1 for the integral operators corresponding to each of the step transition probability densities. Then it is also an eigenfunction with eigenvalue 1, and hence an equilibrium probability density, for the integral operator corresponding to the transition probability density for the complete iteration.

If this transition probability density is to be the unique equilibrium probability density for the complete iteration, then random walk must be irreducible. That is, it must be possible to go from any point to any other point by performing iterations consisting of the specified steps.

If the steps of an iteration satisfy these sufficient conditions, there is also another way of defining an iteration with the desired, unique equilibrium density. Instead of performing an iteration as a series of steps, it is possible to define the iteration as consisting of one of the steps, chosen randomly (with any distribution having nonzero probabilities) among the possible steps. In this case, the transition probability density for the iteration is equal to a linear combination of the transition probability densities for the individual steps. The coefficient of the transition probability density for a given step is the probability that this step is selected. Since the

⁷It is easy to see why. Let $t = \log T/T_0$. Then $f(T) = 1/T$ transforms into $g(t) = \text{const}$. This example is then just the 'exponentiated version' of example 3.3.

desired probability density is an equilibrium probability density (eigenfunction with eigenvalue 1) for the integral operators corresponding to each of the step transition probability matrices, and since the sum of all the coefficients in the linear combination is equal to 1, it is also an equilibrium probability density for the integral operator corresponding to the transition probability density for the complete iteration. This equilibrium probability density is unique, since it is possible, following the given steps, to go from any point to any other point in the space.

Of course, a step of an iteration can, in the same way, be built from substeps, and in this way acquire the same (not necessarily unique) equilibrium probability density as the substeps.

3.9 Choosing Random Directions and Step Lengths

A random walk is an iterative process where, when we stay at some ‘current point’, we may jump to a neighboring point. We must decide two things, the direction of the jump and its step length. Let us examine the two problems in turn.

3.9.1 Choosing Random Directions

When the number of dimensions is small, a ‘direction’ in a space is something simple. This is not so when we work in large-dimensional spaces. Consider, for instance, the problem of choosing a direction in a space of functions. Of course, a space where each point is a function is infinite-dimensional, and we work here with finite-dimensional spaces, but we may just assume that we have discretized the functions using a large number of points, say 10 000 or 10 000 000 points.

If we are ‘at the origin’ of the space, i.e., at point $\{0, 0, \dots\}$ representing a function that is everywhere zero, we may decide to choose a direction pointing towards smooth functions, or fractal functions, gaussian-like functions, functions having zero mean value, L_1 functions, L_2 functions, functions having a small number of large jumps, etc. This freedom of choice, typical of large-dimensional problems, has to be carefully analyzed, and it is indispensable to take advantage of it when designing random walks.

Assume that we are able to design a primeval random walk that samples the probability density $f(\mathbf{x})$, and we wish to modify it considering the values $g(\mathbf{x})/\mu(\mathbf{x})$, using the Metropolis rule (or any equivalent rule), in order to obtain a random walk that samples

$$h(\mathbf{x}) = k f(\mathbf{x}) \frac{g(\mathbf{x})}{\mu(\mathbf{x})} . \quad (3.11)$$

We can design many primeval random walks that sample $f(\mathbf{x})$. Using Metropolis modification of a random walk, we will always obtain a random walk that samples $h(\mathbf{x})$. A well designed primeval random walk will ‘present’ to the Metropolis criterion test points \mathbf{x}_{test} that have a large probability of being accepted (i.e., that have a large value of $g(\mathbf{x})_{\text{test}}/\mu(\mathbf{x})_{\text{test}}$). A poorly designed primeval random walk will test points with a low probability of being accepted. Then, the algorithm is very slow in producing accepted points. Although high acceptance probability can always be obtained with very small step lengths (if the probability density to be sampled is smooth), we need to discover directions that give high acceptance ratios even for large step lengths.

3.9.2 Choosing Step Lengths

Numerical algorithms are usually forced to compromise between some conflicting wishes. For instance, a gradient-based minimization algorithm has to select a finite step length along the direction of steepest descent. The larger the step length, the smaller may be the number of iterations required to reach the minimum, but if the step length is chosen too large, we may lose efficiency; we can even increase the value of the target function, instead of diminishing it.

The random walks contemplated here faces exactly the same situation. The direction of the move is not deterministically calculated, but is chosen randomly, with the common-sense constraint discussed in the previous section. But once a direction has been decided, the size of the jump along this direction, that has to be submitted to the Metropolis criterion, has to be ‘as large as possible’, but not too large. Again, the ‘Metropolis theorem’ guarantees that the final random walk will sample the target probability distribution, but the better we are in choosing the step length, the more efficient the algorithm will be.

In practice, a neighborhood size giving an acceptance rate of 30% – 60% (for the final, posterior sampler) can be recommended.

3.10 Appendixes

3.10.1 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 $p(\mathbf{x}_i)\mathbf{d}\mathbf{x}_i$ particles in the neighborhood of point \mathbf{x}_i . The simplest equilibrium flows are *symmetric*, that is, they satisfy

$$F(\mathbf{x}_i, \mathbf{x}_j) = F(\mathbf{x}_j, \mathbf{x}_i) \quad (3.12)$$

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 $p(\mathbf{x}_j)\mathbf{d}\mathbf{x}_j$ 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 $p(\mathbf{x}_j)\mathbf{d}\mathbf{x}_j$ 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 $p(\mathbf{x}_j)\mathbf{d}\mathbf{x}_j$. 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 $p(\mathbf{x}_j)\mathbf{d}\mathbf{x}_j$.

3.10.2 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 $P(\mathbf{x}_i | \mathbf{x}_j)$ 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 $P(\mathbf{x}_i | \mathbf{x}_j)$ 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

$$\int_{\mathcal{X}} P(\mathbf{x}_i | \mathbf{x}_j) d\mathbf{x}_i = 1. \quad (3.13)$$

For convenience we shall assume that $P(\mathbf{x}_i | \mathbf{x}_j)$ 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 $P(\mathbf{x}_i | \mathbf{x}_j)$. 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 $P(\mathbf{x}_i | \mathbf{x}_j)$. Then, $p(\mathbf{x})$ is an eigenfunction with eigenvalue 1 of the linear integral operator with kernel $P(\mathbf{x}_i | \mathbf{x}_j)$:

$$\int_{\mathcal{X}} P(\mathbf{x}_i | \mathbf{x}_j) p(\mathbf{x}_j) d\mathbf{x}_j = p(\mathbf{x}_i). \quad (3.14)$$

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 $P(\mathbf{x}_i | \mathbf{x}_j)$. 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 (**Note: Find appropriate reference...**).

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 $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \dots$ obtained by a random walk where $P(\mathbf{x}_i | \mathbf{x}_j)$ 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 $P(\mathbf{x}_i | \mathbf{x}_j)$, then

$$F(\mathbf{x}_i, \mathbf{x}_j) = P(\mathbf{x}_i | \mathbf{x}_j) p(\mathbf{x}_j) \quad (3.15)$$

represents the probability density of transitions from \mathbf{x}_j to \mathbf{x}_i : while $P(\mathbf{x}_i | \mathbf{x}_j)$ 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 $p(\mathbf{x}_i)d\mathbf{x}_i$:

$$\int_{\mathcal{X}} F(\mathbf{x}_i, \mathbf{x}_j)d\mathbf{x}_j = \int_{\mathcal{X}} F(\mathbf{x}_k, \mathbf{x}_i)d\mathbf{x}_k = p(\mathbf{x}_i) \quad (3.16)$$

Consider a random walk with transition probability density $P(\mathbf{x}_i | \mathbf{x}_j)$ 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

$$\varphi(\mathbf{x}_i, \mathbf{x}_j) = F(\mathbf{x}_i, \mathbf{x}_j)\psi(\mathbf{x}_i, \mathbf{x}_j) \quad (3.17)$$

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

$$\begin{aligned} P(\mathbf{x}_i, \mathbf{x}_j)^{\text{modified}} &= F(\mathbf{x}_i, \mathbf{x}_j) \frac{\psi(\mathbf{x}_i, \mathbf{x}_j)}{p(\mathbf{x}_j)q(\mathbf{x}_j)} \\ &= P(\mathbf{x}_i | \mathbf{x}_j) \frac{\psi(\mathbf{x}_i, \mathbf{x}_j)}{q(\mathbf{x}_j)}, \end{aligned}$$

which is the product of the original transition probability density, and a new probability — the acceptance probability

$$P_{ij}^{\text{acc}} = \frac{\psi(\mathbf{x}_i, \mathbf{x}_j)}{q(\mathbf{x}_j)}. \quad (3.18)$$

If we choose to multiply $F(\mathbf{x}_i, \mathbf{x}_j)$ with the symmetric flow density

$$\psi_{ij} = \text{Min}(q(\mathbf{x}_i), q(\mathbf{x}_j)), \quad (3.19)$$

we obtain the Metropolis acceptance probability

$$P_{ij}^{\text{metrop}} = \text{Min} \left(1, \frac{q(\mathbf{x}_i)}{q(\mathbf{x}_j)} \right), \quad (3.20)$$

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

$$\psi_{ij} = \text{Min} (q(\mathbf{x}_i), q(\mathbf{x}_j)). \quad (3.21)$$

3.10.3 Appendix: Sampling Explicitly Given Probability Densities

Three methods for sampling explicitly known probability densities are important, and they are given by the following three theorems (formulated for a probability density over a 1-dimensional space):

Theorem 1. Let p be an everywhere nonzero probability density with distribution function P , given by

$$P(s) = \int_{-\infty}^s p(s)ds, \quad (3.22)$$

and let r be a random number chosen uniformly at random between 0 and 1. Then the random number x generated through the formula

$$x = P^{-1}(r) \quad (3.23)$$

has probability density p .

Theorem 2. Let p be a nonzero probability density defined on the interval $I = [a, b]$ for which there exists a positive number M , such that

$$p(x) \leq M. \quad (3.24)$$

and let r and u be two random numbers chosen uniformly at random from the intervals $[0, 1]$ and I , respectively. If u survives the test

$$r \leq \frac{p(u)}{M} \quad (3.25)$$

it is a sample of the probability density p .

More special, yet useful, is the following way of generating Gaussian random numbers:

Theorem 3. Let r_1 and r_2 be random numbers chosen uniformly at random between 0 and 1. Then the random numbers x_1 and x_2 generated through the formulas

$$\begin{aligned} x_1 &= \sqrt{-2 \ln r_2} \cos(2\pi r_1) \\ x_2 &= \sqrt{-2 \ln r_2} \sin(2\pi r_1) \end{aligned}$$

are independent and Gaussian distributed with zero mean and unit variance.

These theorems are straightforward to use in practice. The proofs are left to the reader as an exercise.

Chapter 4

Homogeneous Probability Distributions

4.1 Parameters

To describe a physical system (a planet, an elastic sample, etc.) we use physical *quantities* (temperature and mass density at some given points, total mass, surface color, etc.). We examine here the situation where the total number of physical quantities is finite. The limitation to a finite number of quantities may seem essential to some (in inverse problems, the school of thought developed by Backus and Gilbert) and accessory to others (like the authors of this text). When we consider a function (for instance, a temperature profile as a function of depth), we assume that the function has been discretized in sufficient detail. By ‘sufficient’ we mean that a limit has practically been attained where the computation of the finite probability of any event becomes practically independent of any further refinement of the discretization of the function¹.

In this section, $\{x^1, x^2 \dots x^n\}$ represents a set of n physical quantities, for which we will assume to have a probability distribution defined. The quantities $\{x^1, x^2 \dots x^n\}$ are assumed to take real values (with, generally, some physical dimensions).

Example 4.1 *We may consider, for instance, (i) the mass of a particle, (ii) the temperature at the center of the Earth, (iii) the value of the fine-structure constant, etc. [End of example.]*

Assuming that we have a set of *real* quantities excludes the possibility that we may have a quantity that takes only discrete values, like $\text{spin} \in \{ +1/2, -1/2 \}$, or even a nonnumerical variable, like $\text{organism} \in \{ \text{plant}, \text{animal} \}$. This is not essential, and the formulas given here could easily be generalized to the case where we have both, discrete and continuous probabilities. But, as discrete probability distributions have obvious definitions of marginal and conditional probability distributions, we do not wish to review them here. On the contrary, probabilities over continuous manifolds have specific problems (change of variables, limits, etc.) that demand our attention.

¹A random function is a function that, at each point, is a random variable. A random function is completely characterized if, for whatever choice of n points we may make, we are able to exhibit the joint n -dimensional probability distribution for the n random variables, and this for any value of n . If the considered random function has some degree of smoothness, there is a limit in the value of n such that any finite probability computed using the actual random function is practically identical to the same probability computed from an n -dimensional discretization of the random function. For an excellent introductory text on random functions, see Pugachev (1965).

[Note: Explain here that we shall use the language of ‘manifolds’.]

[Note: Explain here that ‘space’ is used as synonymous of ‘manifold’]

In the following, we will meet two distinct categories of uncertain ‘parameters’. The first category consists of *physical quantities* whose ‘actual values’ are not exactly known but cannot be analyzed by generating many realizations of the parameter values in a repetitive experiment. An obvious example of such a parameter is the radius of the earth’s core (say r). If $f(r)$ is a probability density over r , we will never say that r is a ‘random variable’; we will rather say that we have a probability density defined over a ‘physical quantity’. The second category of parameters are bona fide ‘random variables’, for which we can obtain histograms through repeated experiments. Such ‘random variables’ do not play any major role in this article.

Although in mathematical texts there is a difference in notation between a parameter and a particular value of the parameter (for instance, by denoting them X and x respectively), we choose here to simplify the notation and use expressions like ‘let $x = x_0$ be a particular value of the parameter x .’

Note: I have to talk about the **commensurability of distances**,

$$ds^2 = ds_{\mathbf{r}}^2 + ds_{\mathbf{s}}^2 \quad , \quad (4.1)$$

every time I have to define the Cartesian product of two spaces each with its own metric.

4.2 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 $\{x, y, z\}$, then, the volume element in the space would be

$$dV(x, y, z) = dx dy dz \quad . \quad (4.2)$$

Should we choose to use geographical coordinates, then the volume element would be

$$dV(r, \vartheta, \varphi) = r^2 \cos \vartheta dr d\vartheta d\varphi \quad . \quad (4.3)$$

Question: what would be, in this parameter space, a *homogeneous probability distribution* of points? Answer: a probability distribution assigning to each region of the space a probability proportional to the volume of the region.

Then, question: 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:

$$f(x, y, z) = k \quad . \quad (4.4)$$

- If we use geographical coordinates $\{r, \vartheta, \varphi\}$, as we have $dV(r, \vartheta, \varphi) = r^2 \cos \vartheta dr d\vartheta d\varphi$, the probability density representing the homogeneous probability distribution is (see example 2.3)

$$g(r, \vartheta, \varphi) = k r^2 \cos \vartheta \quad . \quad (4.5)$$

- 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

$$h(u, v, w) = k v(u, v, w) \quad . \quad (4.6)$$

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 $V(\mathcal{A})$.

We can arrive at some conclusions from this example, that are of general validity. First, the homogeneous probability distribution is represented by a constant probability density **only** if we use Cartesian (or rectilinear) coordinates. Two other conclusions can be stated as two (equivalent) rules:

Rule 4.1 *The probability density representing the homogeneous probability distribution is easily obtained if the expression of the volume element $dV(u_1, u_2, \dots) = v(u_1, u_2, \dots) du_1 du_2 \dots$ of the space is known, as it is then given by $h(u_1, u_2, \dots) = k v(u_1, u_2, \dots)$, where k is a proportionality constant (that may have physical dimensions).*

Rule 4.2 If there is a metric $g_{ij}(u_1, u_2, \dots)$ in the space, then, as mentioned above, the volume element is given by $dV(u_1, u_2, \dots) = \sqrt{\det \mathbf{g}(u_1, u_2, \dots)} du_1 du_2 \dots$, i.e., we have $v(u_1, u_2, \dots) = \sqrt{\det \mathbf{g}(u_1, u_2, \dots)}$. The probability density representing the homogeneous probability distribution is, then, $h(u_1, u_2, \dots) = k \sqrt{\det \mathbf{g}(u_1, u_2, \dots)}$.

Rule 4.3 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 (equation ??).

Indeed, in the expression above, $g(r, \vartheta, \varphi) = k r^2 \cos \vartheta$, 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, etc. One of the properties of the parameters we have in mind is that they occur in pairs of mutually reciprocal parameters:

$$\begin{array}{llll}
 \text{Period} & T = 1/\nu & ; & \text{Frequency} & \nu = 1/T \\
 \text{Resistivity} & \rho = 1/\sigma & ; & \text{Conductivity} & \rho = 1/\sigma \\
 \text{Temperature} & T = 1/(k\beta) & ; & \text{Thermodynamic parameter} & \beta = 1/(kT) \\
 \text{Mass density} & \rho = 1/\ell & ; & \text{Lightness} & \ell = 1/\rho \\
 \text{Compressibility} & \gamma = 1/\kappa & ; & \text{Bulk modulus (uncompressibility)} & \kappa = 1/\gamma \quad .
 \end{array}$$

When physical theories are elaborated, 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 ν 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 σ_{ij} to strain ε_{ij} can be expressed as $\sigma_{ij} = c_{ij}{}^{kl} \varepsilon_{kl}$, thus introducing the stiffness tensor c_{ijkl} , or as $\varepsilon_{ij} = d_{ij}{}^{kl} \sigma_{kl}$, thus introducing the compliance tensor d_{ijkl} , 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 ρ_1 and ρ_2 . Correspondingly, their conductivities are $\sigma_1 = 1/\rho_1$ and $\sigma_2 = 1/\rho_2$. How should we define the ‘distance’ between 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

$$D(S_1, S_2) = \left| \log \frac{\rho_2}{\rho_1} \right| = \left| \log \frac{\sigma_2}{\sigma_1} \right| . \quad (4.7)$$

This definition (i) treats symmetrically the two equivalent parameters ρ and σ 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 ‘sensible’ definition of distance between the two samples S_1 and S_2 .

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

$$dL(x) = \frac{dx}{x} . \quad (4.8)$$

This being a ‘one-dimensional volume’ we can apply now the rule 4.1 above, to get the expression of the homogeneous probability density for such a positive parameter:

$$f(x) = \frac{k}{x} . \quad (4.9)$$

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

$$g(y) = \frac{k}{y} . \quad (4.10)$$

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

$$x^* = \log \frac{x}{x_0} \quad ; \quad y^* = \log \frac{y}{y_0} , \quad (4.11)$$

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

$$f'(x^*) = k \quad ; \quad f'(y^*) = k . \quad (4.12)$$

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 4.12, together with the Jacobian rule, that allows full understanding of the (homogeneous) probability densities 4.9–4.10.

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

Rule 4.5 *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 (by definition of Newtonian time) $f(u) = k$. 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 4.6 *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 λ and μ , the bulk modulus κ , the Poisson ratio σ , etc. A simple inspection of the theoretical range of variation of these parameters shows that the first Lamé parameter λ and the Poisson ratio σ may take negative values, so they are certainly not Jeffreys parameters. In contrast, Hooke's law $\sigma_{ij} = c_{ijkl} \varepsilon^{kl}$, defining a linearity between stress σ_{ij} and strain ε_{ij} , defines the positive definite stiffness tensor c_{ijkl} or, if we write $\varepsilon_{ij} = d_{ijkl} \sigma^{kl}$, defines its inverse, the compliance tensor d_{ijkl} . The two reciprocal tensors c_{ijkl} and d_{ijkl} are 'Jeffreys tensors'. This is a notion that would take too long to develop here, but we can give the following rule:

Rule 4.7 *The eigenvalues of a Jeffreys tensor are Jeffreys quantities².*

As the two (different) eigenvalues of the stiffness tensor c_{ijkl} are $\lambda_\kappa = 3\kappa$ (with multiplicity 1) and $\lambda_\mu = 2\mu$ (with multiplicity 5), we see that the incompressibility modulus κ and the shear modulus μ are Jeffreys parameters³ (as are any parameter proportional to them, or any power of them, including the inverses). If for some reason, instead of working with κ and μ , we wish to work with other elastic parameters, like for instance the Young modulus Y and the Poisson ratio σ , then the homogeneous probability distribution must be found using the Jacobian of the transformation between (Y, σ) and (κ, μ) . This is done in appendix 4.3.2.

Some probability densities have conspicuous 'dispersion parameters', like the σ 's in the normal probability density $f(x) = k \exp(-(x - x_0)^2 / (2\sigma^2))$, in the lognormal probability $g(X) = k \exp(-(\log X/X_0)^2 / (2\sigma^2))$ or in the Fisher probability density $h(\vartheta, \varphi) = k \cos \vartheta \exp(\sin \vartheta / \sigma^2)$. A consistent probability model requires that when the dispersion parameter σ tends to infinity, the probability density tends to the homogeneous probability distribution. For instance, in the three examples just given, $f(x) \rightarrow k$, $g(X) \rightarrow k/X$ and $h(\vartheta, \varphi) \rightarrow k \cos \vartheta$, 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 4.8 *A probability density is only consistent if it tends to the homogeneous probability density when its dispersion parameters tend to infinity.*

As an example, using the normal probability density $f(x) = k \exp(-(x - x_0)^2 / (2\sigma^2))$, 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 ??.

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

²This 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.

³The definition of the elastic constants was made before the tensorial structure of the theory was understood. Seismologists, today, should never introduce, at a theoretical level, parameters like the first Lamé coefficient λ or the Poisson ratio. Instead they should use κ and μ (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 $\lambda_\mu = 2\mu$, that we propose to call the *eigen-bulk-modulus* and the *eigen-shear-modulus*.

enthusiastic Bayesians do not call it ‘homogeneous’ but ‘noninformative’. We do not agree with this. The homogeneous probability distribution is as informative as any other distribution, it is just the homogeneous one⁴.

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. We emphasize the need of defining a distance in the parameter space, from which the notion of homogeneity will follow. In this, we slightly depart from the original work by Jeffreys and Jaynes.

⁴Note that 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

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

This means, in particular, that we can 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 $\mu(\mathbf{x})$. For the homogeneous probability distribution ‘fills the space’. Then, one may take the convention to measure the information content of any probability density $f(\mathbf{x})$ with respect to the homogeneous probability density:

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

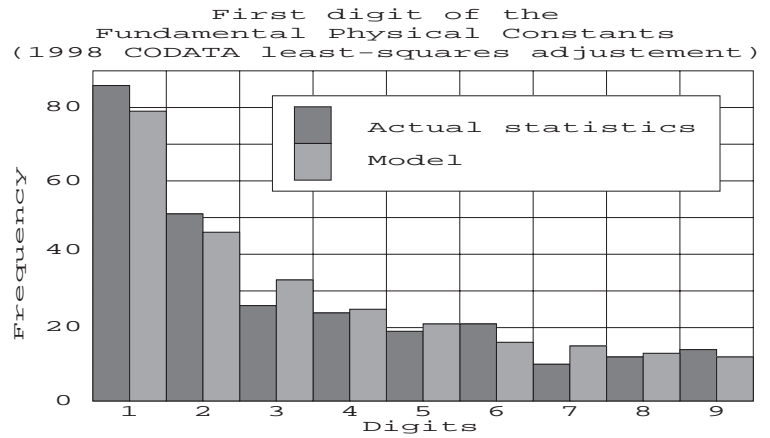
The homogeneous probability density is then ‘noninformative’, $I(\mu) = I(\mu|\mu) = 0$, but this is just by definition.

4.3 Appendixes

4.3.1 Appendix: First Digit of the Fundamental Physical Constants

Note: mention here figure 4.1, and explain. Say that the negative numbers of the table are ‘false negatives’. Figure 4.3 statistics of surfaces and populations of States and Islands.

Figure 4.1: Statistics of the first digit in the table of Fundamental Physical Constants (1998 CODATA least-squares adjustment; Mohr and Taylor, 2001). I have indiscriminately taken all the constants of the table (263 in total). The ‘model’ corresponds to the prediction that the relative frequency of digit n in a base K system of numeration is $\log_K(n+1)/n$. Here, $K = 10$.



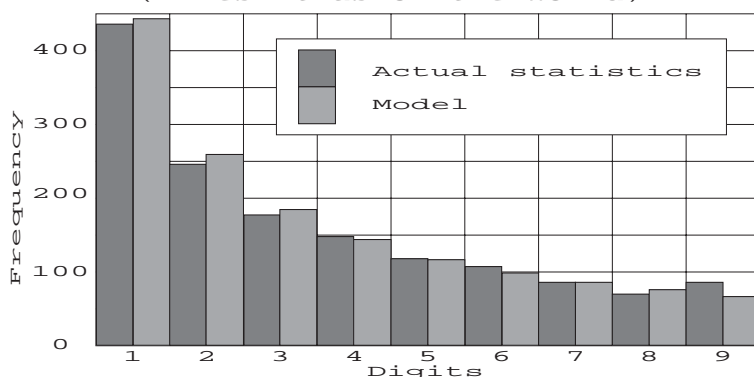
STATES, TERRITORIES & PRINCIPAL ISLANDS OF THE WORLD

Figure 4.2: The beginning of the list of the States, Territories and Principal Islands of the World, in the Times Atlas of the World (Times Books, 1983), with the first digit of the surfaces and populations highlighted. The statistics of this first digit in shown in figure 4.3.

Name [Plate] and Description	Sq. km	Sq. miles	Population
Abu Dhabi , see <i>United Arab Emirates</i>			
Afghanistan [31] Capital: <i>Kabul</i>	636,267	245,664	15,551,358* 1979
Ajman , see <i>United Arab Emirates</i>			
Åland [51] Self-governing Island Territory of Finland	1,505	581	22,000 1981
Albania [83] Capital: <i>Tirana (Tiranë)</i>	28,748	11,097	2,590,600 1979
Aleutian Islands [113] Territory of U.S.A.	17,666	6,821	6,730* 1980
Algeria [88] Capital: <i>Algiers (Alger)</i>	2,381,745	919,354	18,250,000 1979
American Samoa [10] Unincorporated Territory of U.S.A.	197	76	30,600 1977
Andorra [75] Capital: <i>Andorra la Vella</i>	465	180	35,460 1981
Angola [91] Capital: <i>Luanda</i>	1,246,700	481,226	6,920,000 1981
...

Figure 4.3: Statistics of the first digit in the table of the surfaces (both in squared kilometers and squared miles) and populations of the States, Territories and Principal Islands of the World, as printed in the first few pages of the Times Atlas of the World (Times Books, 1983). As for figure 4.1, the ‘model’ corresponds to the prediction that the relative frequency of digit n is $\log_{10}(n + 1)/n$.

Surfaces and Populations of the States, Territories and Principal Islands (Times Atlas of the World)



4.3.2 Appendix: Homogeneous Probability for Elastic Parameters

In this appendix, we start from the assumption that the incompressibility modulus and the shear modulus are Jeffreys parameters (they are the eigenvalues of the stiffness tensor c_{ijkl}), 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 - Transverse wave velocity } .

4.3.2.1 Uncompressibility Modulus and Shear Modulus

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

$$\kappa^* = \log \frac{\kappa}{\kappa_0} \quad ; \quad \mu^* = \log \frac{\mu}{\mu_0} \quad , \quad (4.13)$$

where κ_0 and μ_0 are two arbitrary constants. The homogeneous probability density is just constant for these parameters (a constant that we set arbitrarily to one)

$$f_{\kappa^* \mu^*}(\kappa^*, \mu^*) = 1 \quad . \quad (4.14)$$

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

$$f_{\kappa \mu}(\kappa, \mu) = \frac{1}{\kappa \mu} \quad . \quad (4.15)$$

This $1/x$ form of the probability density remains invariant if we take any power of κ and of μ . In particular, if instead of using the incompressibility κ 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 4.14 there is the Euclidean definition of distance

$$ds^2 = (d\kappa^*)^2 + (d\mu^*)^2 \quad , \quad (4.16)$$

that corresponds, in the variables (κ, μ) , to

$$ds^2 = \left(\frac{d\kappa}{\kappa} \right)^2 + \left(\frac{d\mu}{\mu} \right)^2 \quad , \quad (4.17)$$

i.e., to the metric

$$\begin{pmatrix} g_{\kappa\kappa} & g_{\kappa\mu} \\ g_{\mu\kappa} & g_{\mu\mu} \end{pmatrix} = \begin{pmatrix} 1/\kappa^2 & 0 \\ 0 & 1/\mu^2 \end{pmatrix} \quad . \quad (4.18)$$

4.3.2.2 Young Modulus and Poisson Ratio

The Young modulus Y and the Poisson ratio σ can be expressed as a function of the incompressibility modulus and the shear modulus as

$$Y = \frac{9 \kappa \mu}{3\kappa + \mu} \quad ; \quad \sigma = \frac{1}{2} \frac{3\kappa - 2\mu}{3\kappa + \mu} \quad (4.19)$$

or, reciprocally,

$$\kappa = \frac{Y}{3(1-2\sigma)} \quad ; \quad \mu = \frac{Y}{2(1+\sigma)} \quad . \quad (4.20)$$

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

$$J = \frac{Y}{2(1+\sigma)^2(1-2\sigma)^2} \quad , \quad (4.21)$$

and the Jacobian rule transforms the probability density 4.15 into

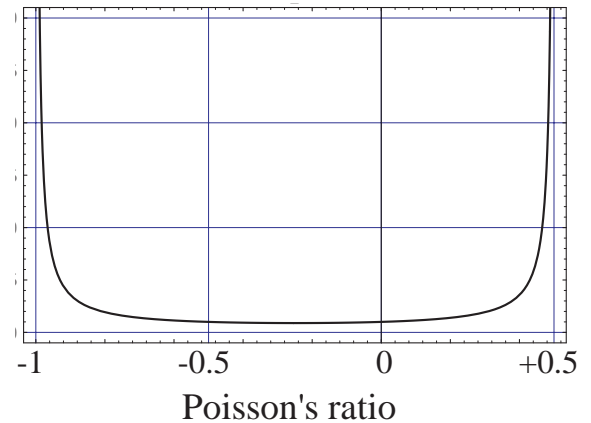
$$f_{Y\sigma}(Y, \sigma) = \frac{1}{\kappa\mu} J = \frac{3}{Y(1+\sigma)(1-2\sigma)} \quad , \quad (4.22)$$

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

$$g(\sigma) = \frac{3}{Y(1+\sigma)(1-2\sigma)} \quad (4.23)$$

for the Poisson ratio. This probability density is represented in figure 4.4. From the definition of σ 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 σ is negative (e.g., Yeganeh-Haeri et al., 1992).

Figure 4.4: 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 4.4 corresponds to a homogeneous distribution. If we have many samples of elastic materials, and if their logarithmic uncompressibility modulus κ^* and their logarithmic shear modulus μ^* have a constant probability density (what *is* the definition of homogeneous distribution of elastic materials), then, σ will be distributed according to the $g(\sigma)$ of the figure.

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

$$g_{IJ} = \Lambda_I^i \Lambda_J^j g_{ij} = \frac{\partial x^i}{\partial x^I} \frac{\partial x^j}{\partial x^J} g_{ij} \quad . \quad (4.24)$$

The metric 4.17 then transforms into

$$\begin{pmatrix} g_{YY} & g_{Y\sigma} \\ g_{\sigma Y} & g_{\sigma\sigma} \end{pmatrix} = \begin{pmatrix} \frac{2}{Y^2} & \frac{2}{(1-2\sigma)Y} - \frac{1}{(1+\sigma)Y} \\ \frac{2}{(1-2\sigma)Y} - \frac{1}{(1+\sigma)Y} & \frac{4}{(1-2\sigma)^2} + \frac{1}{(1+\sigma)^2} \end{pmatrix} . \quad (4.25)$$

The surface element is

$$dS_{Y\sigma}(Y, \sigma) = \sqrt{\det g} dY d\sigma = \frac{3 dY d\sigma}{Y(1+\sigma)(1-2\sigma)} , \quad (4.26)$$

a result from which expression 4.22 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 4.4). In fact, the Poisson ratio σ depends only on the ratio κ/μ (incompressibility modulus over shear modulus), as we have

$$\frac{1+\sigma}{1-2\sigma} = \frac{3}{2} \frac{\kappa}{\mu} . \quad (4.27)$$

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*.

4.3.2.3 Longitudinal and Transverse Wave Velocities

Equation 4.15 gives the probability density representing the homogeneous homogeneous probability distribution of elastic media, when parameterized by the uncompressibility modulus and the shear modulus:

$$f_{\kappa\mu}(\kappa, \mu) = \frac{1}{\kappa \mu} . \quad (4.28)$$

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

$$f_{\kappa\mu\rho}(\kappa, \mu, \rho) = \frac{1}{\kappa \mu \rho} . \quad (4.29)$$

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 α and the shear wave velocity β are related to the uncompressibility modulus κ and the shear modulus μ through

$$\alpha = \sqrt{\frac{\kappa + 4\mu/3}{\rho}} ; \quad \beta = \sqrt{\frac{\mu}{\rho}} , \quad (4.30)$$

and a direct use of the Jacobian rule transforms the probability density 4.29 into

$$f_{\alpha\beta\rho}(\alpha, \beta, \rho) = \frac{1}{\rho \alpha \beta \left(\frac{3}{4} - \frac{\beta^2}{\alpha^2}\right)} . \quad (4.31)$$

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

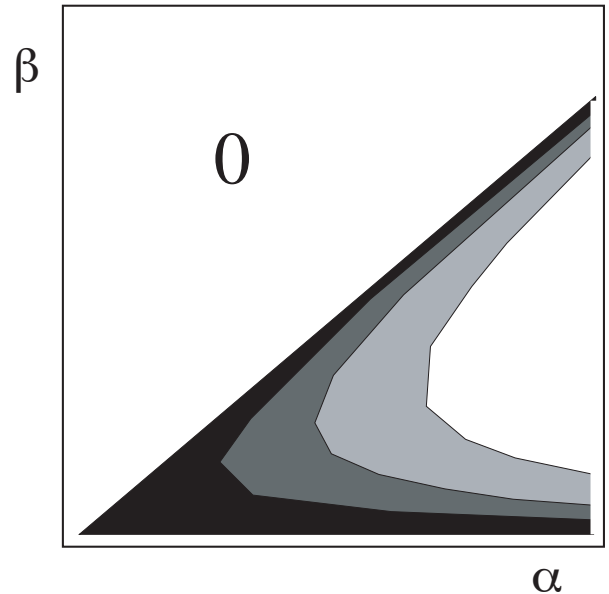
$$\alpha > \frac{2}{\sqrt{3}}\beta . \quad (4.32)$$

Let us focus for a moment on the homogeneous probability density for the two wave velocities (α, β) existing in an elastic solid (disregard here the mass density ρ). We have

$$f_{\alpha\beta}(\alpha, \beta) = \frac{1}{\alpha \beta \left(\frac{3}{4} - \frac{\beta^2}{\alpha^2}\right)} . \quad (4.33)$$

It is displayed in figure 4.5.

Figure 4.5: The joint homogeneous probability density for the velocities (α, β) 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 α and β is of the form $1/x$. For we have to compute

$$f_{\alpha}(\alpha) = \int_0^{\sqrt{3}\alpha/2} d\beta f(\alpha, \beta) \quad (4.34)$$

and

$$f_{\beta}(\beta) = \int_{2\beta/\sqrt{3}}^{+\infty} d\alpha f(\alpha, \beta) \quad (4.35)$$

(the bounds of integration can easily be understood by a look at figure 4.5). These integrals can be evaluated as

$$f_{\alpha}(\alpha) = \lim_{\varepsilon \rightarrow 0} \int_{\sqrt{\varepsilon} \sqrt{3} \alpha/2}^{\sqrt{1-\varepsilon} \sqrt{3} \alpha/2} d\beta f(\alpha, \beta) = \lim_{\varepsilon \rightarrow 0} \left(\frac{4}{3} \log \frac{1-\varepsilon}{\varepsilon} \right) \frac{1}{\alpha} \quad (4.36)$$

and

$$f_{\beta}(\beta) = \lim_{\varepsilon \rightarrow 0} \int_{\sqrt{1+\varepsilon} 2\beta/\sqrt{3}}^{2\beta/(\sqrt{\varepsilon} \sqrt{3})} d\alpha f(\alpha, \beta) = \lim_{\varepsilon \rightarrow 0} \left(\frac{2}{3} \log \frac{1/\varepsilon - 1}{\varepsilon} \right) \frac{1}{\beta} \quad (4.37)$$

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

$$f_{\alpha}(\alpha) = \frac{1}{\alpha} \quad (4.38)$$

and

$$f_{\beta}(\beta) = \frac{1}{\beta} \quad (4.39)$$

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

$$f_{n\eta\rho}(n, \eta, \rho) = \frac{1}{\rho n \eta \left(\frac{3}{4} - \frac{n^2}{\eta^2} \right)}. \quad (4.40)$$

4.3.3 Appendix: 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:

$$f(\sigma_{xx}, \sigma_{yy}, \sigma_{zz}, \sigma_{xy}, \sigma_{yz}, \sigma_{zx}) = k \quad . \quad (4.41)$$

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

$$\{\sigma_{xx}, \sigma_{yy}, \sigma_{zz}, \sigma_{xy}, \sigma_{yz}, \sigma_{zx}\} \rightleftharpoons \{\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi\} \quad (4.42)$$

is

$$\left| \frac{\partial(\sigma_{xx}, \sigma_{yy}, \sigma_{zz}, \sigma_{xy}, \sigma_{yz}, \sigma_{zx})}{\partial(\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi)} \right| = (\lambda_1 - \lambda_2)(\lambda_2 - \lambda_3)(\lambda_3 - \lambda_1) \sin \theta \quad , \quad (4.43)$$

the homogeneous probability density 4.41 transforms into

$$g(\lambda_1, \lambda_2, \lambda_3, \psi, \theta, \varphi) = k (\lambda_1 - \lambda_2)(\lambda_2 - \lambda_3)(\lambda_3 - \lambda_1) \sin \theta \quad . \quad (4.44)$$

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 4.8, 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 4.44 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.

Chapter 5

Basic Measurements

Note: Complete and expand what follows:

I 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 underplaying the role assigned by ISO to the particular Gaussian model for uncertainties (see section 5.8) and by not assuming that the uncertainties are ‘small’.

5.1 Terminology

Note: Introduce here the ISO terminology for analyzing uncertainties in measurements.

Note: Note say that we are interested in volumetric probabilities, not ‘uncertainties’.

Note: For the time being, this section is written in telegraphical style. It will, obviously, be rewritten.

Measurand: Particular quantity subject to measurement. It is the *input* to the measuring instrument.

Input may be a length; output may be an electric tension. They may not be the same physical quantity.

For instance, the input of a seismometer is a displacement, the output is a voltage. At a given time, the voltage is a convolution of the past input by a transfer function.

5.2 Old text: Measuring physical parameters

To define the experimental procedure that will lead to a “measurement” we need to conceptualize the objects of the “universe”: do we have point particles or a continuous medium? Any instrument that we can build will have finite accuracy, as any manufacture is imperfect. Also, during the measurement act, the instrument will always be submitted to unwanted solicitations (like uncontrolled vibrations).

This is why, even if the experimenter postulates the existence of a well defined, “true value”, of the measured parameter, she/he will never be able to measure it exactly. Careful modeling of experimental uncertainties is not easy. Sometimes, the result of a measurement of a parameter p is presented as $p = p_0 \pm \sigma$, where the interpretation of σ may be diverse. For instance, the experimenter may imagine a bell-shaped probability density around p_0 representing her/his state of information “on the true value of the parameter”. The constant σ can be the standard deviation (or mean deviation, or other estimator of dispersion) of the probability density used to model the experimental uncertainty.

In part, the shape of this probability density may come from histograms of observed or expected fluctuations. In part, it will come from a subjective estimation of the defects of the unique pieces of the instrument. We postulate here that the result of any measurement can, in all generality, be described by defining a probability density over the measured parameter, representing the information brought by the experiment on the “true”, unknowable, value of the parameter. The official guidelines for expressing uncertainty in measurement, as given by the International Organization for Standardization (ISO) and the National Institute of Standards and Technology¹, although stressing the special notion of standard deviation, are consistent with the possible use of general probability distributions to express the result of a measurement, as advocated here.

Any shape of the density function is not acceptable. For instance, the use of a Gaussian density to represent the result of a measurement of a positive quantity (like an electric resistivity) would give a finite probability for negative values of the variable, which is inconsistent (a lognormal probability density, on the contrary, could be acceptable).

In the event of an “infinitely bad measurement” (like when, for instance, an unexpected event prevents, in fact, any meaningful measure) the result of the measurement should be described using the null information probability density introduced above. In fact, when the density function used to represent the result of a measurement has a parameter σ describing the “width” of the function, it is the limit of the density function for $\sigma \rightarrow \infty$ that should represent a measurement of infinitely bad quality. This is consistent, for instance, with the use of a lognormal probability density for a parameter like an electric resistivity r , as the limit of the lognormal for $\sigma \rightarrow \infty$ is the $1/r$ function, which is the right choice of noninformative probability density for r .

Another example of possible probability density to represent the result of a measurement of a parameter p is to take the noninformative probability density for $p_1 < p < p_2$ and zero outside. This fixes strict bounds for possible values of the parameter, and tends to the noninformative probability density when the bounds tend to infinity.

The point of view proposed here will be consistent with the the use of “theoretical param-

¹Guide to the expression of uncertainty in measurement, International Organization of Standardization (ISO), Switzerland, 1993. B.N. Taylor and C.E. Kuyatt, 1994, Guidelines for evaluating and expressing the uncertainty of NIST measurement results, NIST technical note 1297.

eter correlations” as proposed in section ??, so that there is no difference, from our point of view, between a “simple measurement” and a measurement using physical theories, including, perhaps, sophisticated inverse methods.

5.3 From ISO

The International Organization for Standardization (ISO) has published (ISO, 1993) a “Guide to the expression of uncertainty in measurement”, which is the result of a joint work with the BIPM², the IEC³ and the OIML⁴. The recommendations of the Guide have also been adopted by the U.S. National Institute of Standards and Technology (Taylor and Kuyatt, 1994).

These recommendations have the advantage of being widely accepted (in addition of being legal). It is therefore important to see into which extent the approach proposed in this book to describe the result of a measurement is consistent with that proposed by ISO.

5.3.1 Proposed vocabulary to be used in metrology

In the definitions that follow, the use of parentheses around certain words of some terms means that the words may be omitted if this is unlikely to cause confusion.

5.3.1.1 (measurable) quantity:

attribute of a phenomenon, body or substance that may be distinguished qualitatively and determined quantitatively.

5.3.1.2 value (of a quantity):

magnitude of a particular quantity generally expressed as a unit of measurement multiplied by a number.

5.3.1.3 true value (of a quantity):

definition not reproduced here.

Comments from the ISO guide: The term “true value of a measurand” or of a quantity (often truncated to “true value”) is avoided in this guide because the word “true” is viewed as redundant. “Measurand” means “particular quantity subject to measurement”, hence “value of a measurand” means “value of a particular quantity subject to measurement”. Since “particular quantity” is generally understood to mean a definite or specified quantity, the adjective “true” in “true value of a measurand” (or in “true value of a quantity”) is unnecessary — the “true” value of the measurand (or quantity) is simply the value of the measurand (or quantity). In addition, as indicated in the discussion above, a unique “true” value is only an idealized concept.

My comments: I have not reproduced the definition of the term “true value” because i) I do not understand it, and ii) it does not seem consistent with the comment above (that I understand perfectly).

5.3.1.4 measurement:

set of operations having the object of determining a value of a quantity.

²Bureau International des Poids et Mesures

³International Electrotechnical Commission

⁴International Organization of Legal Metrology

My comments: I do not agree. The object of a measurement is not to determine “a value” of a quantity, but, rather, to obtain a “state of information” on the (true) value of a quantity. The proposed definition is acceptable only in the particular case when the information obtained in the measurement can be represented by a probability density that, being practically monomodal, can be well described by a central estimator (the “determined value” of the quantity) and an estimator of dispersion (the “uncertainty” of the measurement).

5.3.1.5 measurand:

particular quantity subject to measurement.

Comments from the ISO guide: The specification of a measurand may require statements about quantities such as time, temperature and pressure.

5.3.1.6 influence quantity:

quantity that is not the measurand but that affects the result of the measurement.

5.3.1.7 result of a measurement:

value attributed to a measurand, obtained by measurement.

My comments: see comments in “measurement”.

5.3.1.8 uncertainty (of measurement):

parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand.

Comments from the ISO guide: The word “uncertainty” means doubt, and thus in its broadest sense “uncertainty of measurement” means doubt about the validity of the result of a measurement. Because of the lack of different words for this general concept of uncertainty and the specific quantities that provide quantitative measures of the concept, for example, the standard deviation, it is necessary to use the word “uncertainty” in these two different senses.

More comments from the ISO guide: The definition of uncertainty of measurement is an operational one that focuses on the measurement result and its evaluated uncertainty. However, it is not inconsistent with other concepts of uncertainty of measurement, such as i) a measure of the possible error in the estimated value of the measurand as provided by the result of a measurement; ii) an estimate characterizing the range of values within which the true value of a measurand lies. Although these two traditional concepts are valid as ideals, they focus on *unknowable* quantities: the “error” of the result of a measurement and the “true value” of the measurand (in contrast to its estimated value), respectively.

Still more comments from the ISO guide: Uncertainty of measurement comprises, in general, many components. Some of these components may be evaluated from the statistical distribution of the results of series of measurements and can be characterized by experimental standard deviations. The other components, which can also be characterized by standard deviations, are evaluated from assumed probability distributions based on experience or other information.

My comments: I could almost agree with this definition, but would rather say that as the result of a measurement is a probability density, the uncertainty, as a parameter, is any estimator of dispersion associated to the probability density. I was pleasantly surprised to discover that the ISO guidelines accept probability distributions coming from subjective knowledge as

an essential part of the description of the results of a measurement. One could fear that normal statistical practices, that exclude Bayesian (subjective) reasoning, were exclusively adopted. I am personally inclined (as this book demonstrates) to push the other way, and reject the notion of “statistical distribution of results of series of measurements”: the maximum generality is obtained when each individual measurement is used, and the statistical well known rules for combining individual measurement “results” will appear by themselves when working properly at the elementary level. At most, the rules proposed by statistical texts are a way for (approximately) short-circuiting some of the steps of the inference methods proposed in this book.

5.3.2 Some basic concepts

Note: what follows is important for the chapter on “physical theories” too.

In practice, the required specification or definition of the measurand is dictated by the required accuracy of [the] measurement. The measurand should be defined with sufficient completeness with respect to the required accuracy so that for all practical purposes associated with the measurement its value is unique. It is in this sense that the expression “value of the measurand” is used in this Guide.

Example: If the length of a nominally one-metre long steel bar is to be determined to micrometre accuracy, its specification should include the temperature and pressure at which the length is defined. This the measurand should be specified as, for example, the length of the bar at 35.00 °C and 101 325 Pa (plus any other defining parameters deemed necessary, such as the way the bar is to be supported). However, if the length is to be determined to only millimetre accuracy, its specification would not require a defining temperature or pressure or a value for any other defining parameter.

Note: Incomplete definition of the measurand can give rise to a component of uncertainty sufficiently large that it must be included in the evaluation of the uncertainty of the measurement result.

Note: The first step in making a measurement is to specify the measurand — the quantity to be measured; the measurand cannot be specified by a value but only by a description of a quantity. However, in principle, a measurand cannot be *completely* described without an infinite amount of information. Thus, to the extent that it leaves room for interpretation, incomplete definition of the measurand introduces into the uncertainty of the result of a measurement a component of uncertainty that may or may not be significant relative to the accuracy required of the measurement.

Note: At some level, every measurand has [...] an “intrinsic” uncertainty that can in principle be estimated in some way. This is the minimum uncertainty with which a measurand can be determined, and every measurement that achieves such an uncertainty may be viewed as the best possible measurement of the measurand. To obtain a value of the quantity in question having a smaller uncertainty requires that the measurand be more completely defined.

[...]

The uncertainty in the result of a measurement generally consists of several components which may be grouped into two categories according to the way in which their numerical value is estimated:

- A. those which are evaluated by statistical methods,

- B. those which are evaluated by other means.

[...] a type A standard uncertainty is obtained from a probability density function derived from an observed frequency distribution, while a type B standard uncertainty is obtained from an assumed probability density function based on the degree of belief that an event will occur (often called subjective probability). Both approaches employ recognized interpretations of probability.

[...]

In practice, there are many possible sources of uncertainty in a measurement, including

- incomplete definition of the measurand;
- imperfect realization of the definition of the measurand;
- nonrepresentative sampling — the sample measured may not represent the defined measurand;
- inadequate knowledge of the effects of environmental conditions on the measurement or imperfect measurement of environmental conditions;
- personal bias in reading analogue instruments;
- finite instrument resolution or discrimination threshold;
- inexact values of measurement standards and reference materials;
- inexact values of constants and other parameters obtained from external sources and used in the data-reduction algorithm;
- approximations and assumptions incorporated in the measurement method and procedure;
- variations in repeated observations of the measurand under apparently identical conditions.

[...]

5.3.2.1 The need for type B evaluations.

If a measurement laboratory had limitless time and resources, it could conduct an exhaustive statistical investigation of every conceivable cause of uncertainty, for example, by using many different makes and kinds of instruments, different methods of measurement, different applications of the method, and different approximations in its theoretical models of the measurement. The uncertainties associated with all of these causes could then be evaluated by the statistical analysis of series of observations and the uncertainty of each cause would be characterized by a statistically evaluated standard deviation. In other words, all of the uncertainty components would be obtained from type A evaluations. Since such an investigation is not an economic practicality, many uncertainty components must be evaluated by whatever other means is practical.

5.3.2.2 Single observation, calibrated instruments.

If an input estimate has been obtained from a single observation with a particular instrument that has been calibrated against a standard of small uncertainty, the uncertainty of the estimate is mainly one of repeatability. The variance of repeated measurements by the instrument may have been obtained on an earlier occasion, not necessarily at precisely the same value of the reading but near enough to be useful, and it may be possible to assume the variance to be applicable to the input value in question. If no such information is available, an estimate must be made based on the nature of the measuring apparatus or instrument, the known variances of other instruments of similar construction, etc.

5.3.2.3 Single observation, verified instruments

Not all measuring instruments are accompanied by a calibration certificate or a calibration curve. Most instruments, however, are constructed to a written standard and verified, either by the manufacturer or by an independent authority, to conform to that standard. Usually the standard contains metrological requirements, often in the form of “maximum permissible errors”, to which the instrument is required to conform. The compliance of the instrument with these requirements is determined by comparison with a reference instrument whose maximum allowed uncertainty is usually specified in the standard. This uncertainty is then a component of the uncertainty of the verified instrument.

If nothing is known about the characteristic error curve of the verified instrument it must be assumed that there is an equal probability that the error has any value within the permitted limits, this is, a rectangular probability distribution. However, certain types of instruments have characteristic curves such that the errors are, for example, likely always to be positive in part of the measuring range and negative in other parts. Sometimes such information can be deduced from a study of the written standard.

5.4 The Ideal Output of a Measuring Instrument

Note: mention here figures 5.1 and 5.2.

Figure 5.1: Instrument built to measure pithches of musical notes. Due to unavoidable measuring noises, a measurement is never infinitely accurate. Figure 5.2 suggests an ideal instrument output.

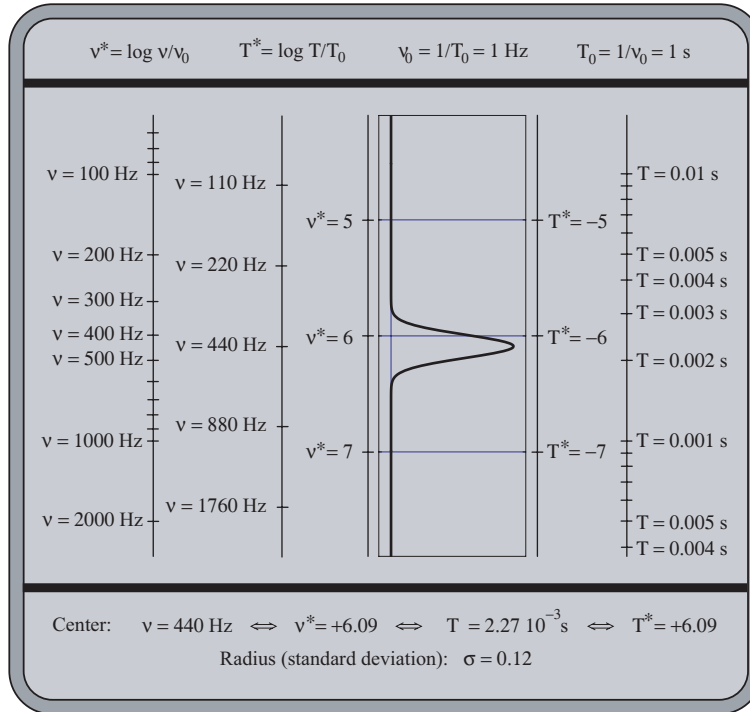
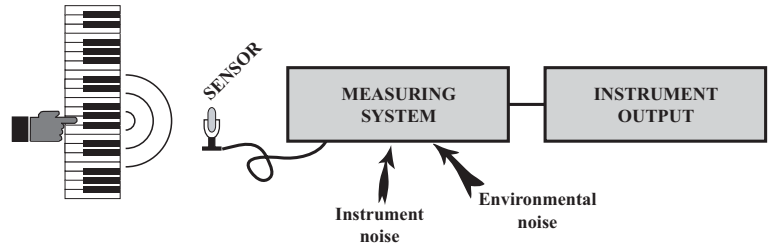
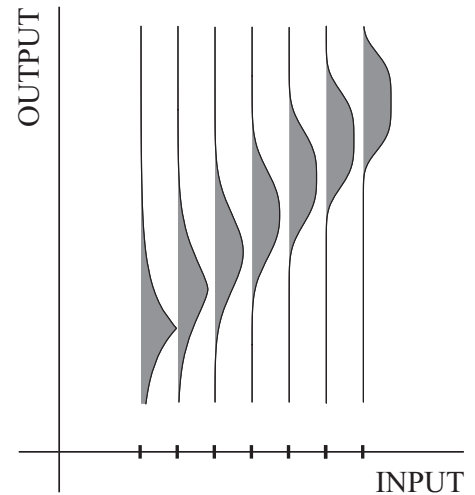


Figure 5.2: The ideal output of a measuring instrument (in this example, measuring frequencies-periods). The curve in the middle corresponds to the volumetric probability describing the information brought by the measurement (on ‘the measurand’). Five different scales are shown (in a real instrument, the user would just select one of the scales). Here, the logarithmic scales correspond to the natural logarithms that a physicist should prefer, but engineers could select scales using decimal logarithms. Note that all the scales are ‘linear’ (with respect to the natural distance in the frequency-period space [see section XXX]): I do not recommend the use of a scale where the frequencies (or the periods) would ‘look linear’.

5.5 Output as Conditional Probability Density

As suggested by figure 5.3, an ‘measuring instrument’ is specified when the conditional volumetric probability $f(y|x)$ for the output y , given the input x is given.

Figure 5.3: The input (or measurand) and the output of a measuring instrument. The output is never an actual value, but a probability distribution, in fact, a conditional volumetric probability $f(y|x)$ for the output y , given the input x .



5.6 A Little Bit of Theory

We want to measure a given property of an object, say the quantity x . Assume that the object has been randomly selected from a set of objects, so that the ‘prior’ probability for the quantity x is $f_x(x)$.

Then, the conditional ...

Then, Bayes theorem ...

5.7 Example: Instrument Specification

[Note: This example is to be put somewhere, I don't know yet where.]

It is unfortunate that ordinary measuring instruments tend to just display some ‘observed value’, the ‘measurement uncertainty’ tending to be hidden inside some written documentation. Awaiting the day when measuring instruments directly display a probability distribution for the measurand, let us contemplate the simple situation where the maker of an instrument, say a frequencymeter, writes something like the following.

This frequencymeter can operate, with high accuracy, in the range $10^2 \text{ Hz} < \nu < 10^9 \text{ Hz}$. When very far from this range, one may face uncontrollable uncertainties. Inside (or close to) this range, the measurement uncertainty is, with a good approximation, independent of the value of the measured frequency. When the instrument displays the value ν_0 , this means that the (1D) volumetric probability for the measurand is

$$\begin{cases} \text{if } \log \frac{\nu}{\nu_0} \leq -\sigma & \text{then } f(\nu) = 0 \\ \text{if } -\sigma < \log \frac{\nu}{\nu_0} < +2\sigma & \text{then } f(\nu) = \frac{2}{9\sigma^2} \left(2\sigma - \log \frac{\nu}{\nu_0} \right) \\ \text{if } +2\sigma \leq \log \frac{\nu}{\nu_0} & \text{then } f(\nu) = 0 \end{cases}, \quad (5.1)$$

where $\sigma = 10^{-4}$. This volumetric probability is displayed at the top of figure 5.4. Using the logarithmic frequency as coordinate, this is an asymmetric triangle.

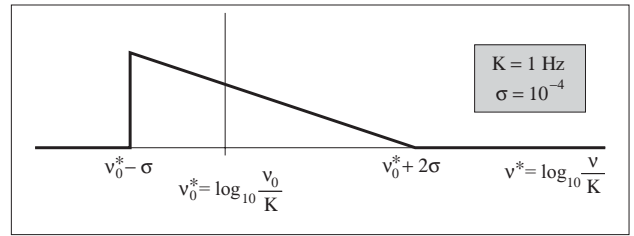
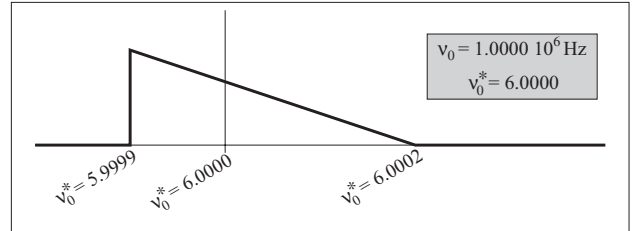


Figure 5.4: Figure for ‘instrument specification’. Note: write this caption.



5.8 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 certainty, described with a volumetric probability over the data parameter space, that we denote here $\rho_{\mathbf{d}}(\mathbf{d})$. 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 appreciation (like for picking arrival times) are also relevant sources of uncertainty.

Example 5.1 Non-analytic volumetric probability *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*

$$t = t_0 \pm \sigma \quad . \quad (5.2)$$

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 5.5)? 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 volumetric probability. Figure 5.6 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 volumetric probability 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. Using, for instance, the Mathematica (registered trademark) computer language we may define the volumetric probability $f(t)$ as

$$f[t_] := (\text{If}[t_1 < t < t_2, a, c] \text{If}[t_3 < t < t_4, b, c]) \quad .$$

Here, a and b are the 'levels' of the two steps, and c is the 'background' volumetric probability. [End of example.]

Figure 5.5: What has an experimenter in mind when she/he describes the result of a measurement by something like $t = t_0 \pm \sigma$?

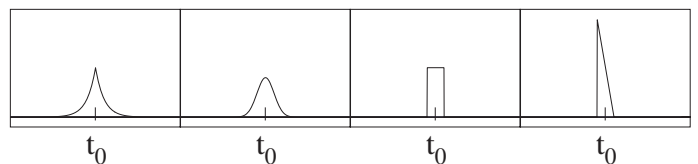
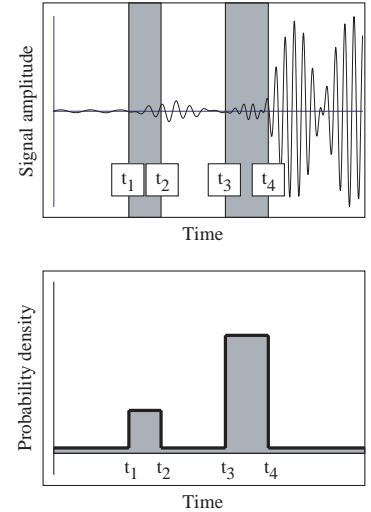


Figure 5.6: A seismologist tries to measure the arrival time of a seismic wave at a seismic station, by ‘reading’ the seismogram at the top of the figure. The seismologist may find quite likely that the arrival time of the wave is between times t_3 and t_4 , and believe that what is before t_3 is just 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 volumetric probability, as the one suggested at the bottom of the figure. Typically, the actual form of each peak of the volumetric probability is not crucial (here, box-car functions are chosen), but the position of the peaks is important. Rather than assigning a zero volumetric probability 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 5.2 *The Gaussian model for uncertainties. The simplest probabilistic model that can be used to describe experimental uncertainties is the Gaussian model*

$$\rho_{\mathcal{D}}(\mathbf{d}) = k \exp\left(-\frac{1}{2}(\mathbf{d} - \mathbf{d}_{\text{obs}})^T \mathbf{C}_D^{-1} (\mathbf{d} - \mathbf{d}_{\text{obs}})\right) . \quad (5.3)$$

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

$$\rho_{\mathcal{D}}(\mathbf{d}) = k \exp\left(-\frac{1}{2} \sum_i \left(\frac{d^i - d_{\text{obs}}^i}{\sigma^i}\right)^2\right) , \quad (5.4)$$

where the σ^i are the ‘standard deviations’. [End of example.]

Example 5.3 *The Generalized Gaussian model for uncertainties. An alternative to the Gaussian model, is to use the Laplacian (double exponential) model for uncertainties,*

$$\rho_{\mathcal{D}}(\mathbf{d}) = k \exp\left(-\sum_i \frac{|d^i - d_{\text{obs}}^i|}{\sigma^i}\right) . \quad (5.5)$$

While the Gaussian model leads to least-squares related methods, this Laplacian model leads to absolute-values methods (see section 8.2.6), well known for producing robust⁵ results. More generally, there is the L_p model of uncertainties

$$\rho_p(\mathbf{d}) = k \exp\left(-\frac{1}{p} \sum_i \frac{|d^i - d_{\text{obs}}^i|^p}{(\sigma_p)^p}\right) \quad (5.6)$$

(see figure 5.7). [End of example.]

⁵A numerical method is called robust if it is not sensitive to a small number of large errors.

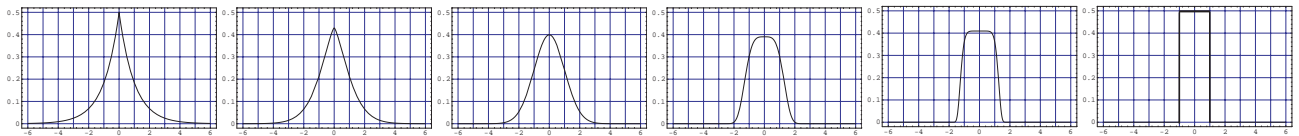


Figure 5.7: Generalized Gaussian for values of the parameter $p = 1, \sqrt{2}, 2, 4, 8$ and ∞ .

5.9 Appendixes

5.9.1 Appendix: Operational Definitions can not be Infinitely Accurate

Note: refer here to figure 5.8, and explain that “the length” of a real object (as opposed to a mathematically defined object) can only be defined by specifying the measuring instrument. There are different notions of length associated to a given object. For instance, figure 5.8 suggests that the length of a piece of wood is larger when defined by the use of a calliper⁶ than when defined by the use of a ruler⁷, because a calliper tends to measure the distance between extremal points, while an observer using a ruler tends to average the rugosities at the wood ends.

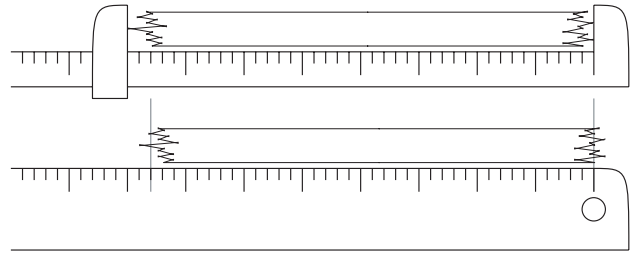


Figure 5.8: Different definitions of the length of an object.

⁶Calliper: an instrument for measuring diameters (as of logs or trees) consisting of a graduated beam and at right angles to it a fixed arm and a movable arm. From the Digital Webster.

⁷Ruler: a smooth-edged strip (as of wood or metal) that is usu. marked off in units (as inches) and is used as a straightedge or for measuring. From the Digital Webster.

5.9.2 Appendix: The International System of Units (SI)

Note: make here a small introduction about the usefulness of a unified system of units.

The rest of this appendix is a reproduction (with permission) of a text published by Robert A. Nelson in the August 1996, issue of *Physics Today*, pages 15–16. ROBERT NELSON is the author of the booklet *SI: The International System of Units*, 2nd ed. (American Association of Physics Teachers, College Park, Maryland, 1982). He is Program Director for Commercial Space at Veda Incorporated in Alexandria, Virginia and teaches in the Department of Aerospace Engineering at the University of Maryland.

Note: ASK FOR THE PERMISSION TO REPRODUCE!!!

Note: The accent in “ampère” is valid in French; check if it is valid in English.

5.9.2.1 Guide for Metric Practice, by Robert A. Nelson

The modernized metric system is known as the *Système International d’Unités* (International System of Units), with the international abbreviation SI. It is founded on seven base units, listed in table 1, that by convention are regarded as dimensionally independent. All other units are derived units, formed coherently by multiplying and dividing units within the system without numerical factors. Examples of derived units, including some with special names, are listed in table 2. The expression of multiples and submultiples of SI units is facilitated through the use of the prefixes listed in table 3.

Quantity	Unit	
	Name	Symbol
length	meter	m
mass	kilogram	kg
time	second	s
electric current	ampère	A
thermodynamic temperature	kelvin	K
amount of substance	mole	mol
luminous intensity	candela	cd

SI obtains its international authority from the Meter Convention, signed in Paris by the delegates of 17 countries, including the United States, on 20 May 1875, and amended in 1921. Today 48 states are members. The treaty established the *Conférence Générale des Poids et Mesures* (General Conference on Weights and Measures) as the formal diplomatic body responsible for ratification of the new proposals related to metric units. The scientific decisions are made by the *Comité International des Poids et Mesures* (International Committee for Weights and Measures). It is assisted by the advise of eight Consultative Committees specializing in particular areas of metrology. The activities of the national standards laboratories are coordinated by the *Bureau International des Poids et Mesures* (International Bureau of Weights and Measures), whose headquarters is at the Pavillon de Breuteuil in Sèvres, France, and which is under the supervision of the CIPM. The SI was established by the 11th CGPM in 1960, when the metric unit definitions, symbols and terminology were extensively revised and simplified.⁸

⁸For history of the metric system and SI units, see R.A. Nelson, *Phys. Teach.* **19**; 596 (1981).

TABLE 2. Examples of SI derived units

Quantity	Unit		Equivalent
	Special name	Symbol	
plane angle	radian	rad	$m/m = 1$
solid angle	steradian	sr	$m^2/m^2 = 1$
speed, velocity			m/s
acceleration			m/s^2
angular velocity			$rad/s = 1$
angular acceleration			rad/s^2
frequency	hertz	Hz	s^{-1}
force	newton	N	$kg \cdot m/s^2$
pressure, stress	pascal	Pa	N/m^2
work, energy, heat	joule	J	$N \cdot m$, $kg \cdot m^2/s^2$
impulse, momentum			$N \cdot s$, $kg \cdot m/s$
power	watt	W	J/s
electric charge	coulomb	C	$A \cdot s$
electric potential, emf	volt	V	J/C , W/A
resistance	ohm	Ω	V/A
conductance	siemens	S	A/V , Ω^{-1}
magnetic flux	weber	Wb	$V \cdot s$
inductance	henry	H	Wb/A
capacitance	farad	F	C/V
electric field strength			V/m , N/C
magnetic flux density	tesla	T	Wb/m^2 , $N/(A \cdot m)$
electric displacement			C/m^2
magnetic field strength			A/m
Celsius temperature	degree Celsius	$^{\circ}C$	K
luminous flux	lumen	lm	$cd \cdot sr$
illuminance	lux	lx	lm/m^2
radioactivity	becquerel	Bq	s^{-1}

TABLE 3. SI prefixes

Factor	Prefix	Symbol	Factor	Prefix	Symbol
10^{24}	yotta	Y	10^{-1}	deci	d
10^{21}	zetta	Z	10^{-2}	centi	c
10^{18}	exa	E	10^{-3}	milli	m
10^{15}	peta	P	10^{-6}	micro	μ
10^{12}	tera	T	10^{-9}	nano	n
10^9	giga	G	10^{-12}	pico	p
10^6	mega	M	10^{-15}	femto	f
10^3	kilo	k	10^{-18}	atto	a
10^2	hecto	h	10^{-21}	zepto	z
10^1	deka	da	10^{-24}	yocto	y

The BIPM, with the guidance of the Consultative Committee for Units and approval of the CIPM, periodically publishes a document⁹ that summarizes the historical decisions of the CGPM and the CIPM and gives some conventions for metric practice. In addition, Technical Committee 12 of the International Organization for Standardization has prepared recommendations concerning the practical use of the SI¹⁰. Some other recommendations have been given by the Commission for Symbols, Units, Nomenclature, Atomic Masses and Fundamental Constants of the International Union of Pure and Applied Physics¹¹. The National Institute of Standards and Technology has published a practical guide for the use of the SI¹². The Institute of Electrical and Electronics Engineers has developed a metric practise manual¹³ that has been recognized by the American National Standards Institute and has been adopted by the US Department of Defense. The American Society for Testing and Materials has prepared a similar manual¹⁴. The Secretary of Commerce, through NIST, has also issued recommendations for US metric practise¹⁵ as provided under the Metric Conversion Act of 1975 and the Omnibus Trade and Competitiveness Act of 1988.

In October 1995 the 20th CGPM, at the recommendations of the CCU and CIPM, eliminated the “supplementary units” radian and steradian as a special class of derived units having dimension 1 (so-called dimensionless derived units). Thus the SI now consists of only two classes of units, base units and derived units, with the radian and steradian included among the derived units as shown in table 2.

5.9.2.2 Style conventions

Letter symbols include quantity symbols and unit symbols. Symbols for physical quantities are set in italic (sloping) type, while symbols for units are set in roman (upright) type (for example, $F = 15 \text{ N}$).

Symbols for unit names derived from proper names have the first letter capitalized — otherwise unit symbols are lower case — but the unit names themselves are not capitalized (for example, tesla, T; meter, m). A unit symbol is a mathematical entity (not an abbreviation) and is usually denoted by the first letter of the unit name (for example, the symbol for gram is g, not gm; the symbol for second is s, not sec), with some exceptions (for example, mol, cd and Hz). The unit symbol is not followed by a period, and plurals of unit symbols are not followed by an “s” (for example, 3 Kg, not 3 Kgs. or 3 Kgs).

⁹Bureau International des Poids et Mesures, *Le système International d'unités (SI)*, 6th ed., BIPM Sèvres, France (1991); US ed.: *The International System of Units (SI)*, B.N. Taylor, ed., Natl. Inst. Stand. Technol. Spec. Pub. 330, US Govt. Printing Office, Washington, D.C. (1991).

¹⁰International Organization for Standardization, *Quantities and Units*, ISO Standards Handbook, 3rd ed., ISO, Geneva (1993). This is a compilation of individual standards ISO 31-0 to 31-13 and ISO 1000, available from Am. Natl. Stand. Inst., New York.

¹¹E. R. Cohen, P. Giacomo, eds., *Physica* **146A**, 1 (1987). Reprinted as *symbols, Units, Nomenclature and Fundamental Constants in Physics* (1987 revision), document IUPAP-25 (SUNAMCO 87-1).

¹²B. N. Taylor, *Guide for the Use of the International System of Units*, Natl. Inst. Stand. Technol. Spec. Pub. 811, US Govt. Printing Office, Washington, D.C. (1995).

¹³Inst. of Electrical and Electronics Engineers, *American National Standard for Metric Practise*, ANSI/IEEE Std. 268-1992, IEEE, New York (1992).

¹⁴Am. Soc. for Testing and Materials, *Standard Practise for Use of the International System of Units (SI) (The Modernized Metric System)*, ASTM E 380-93, ASTM, Philadelphia (1993).

¹⁵“Metric System of Measurement; Interpretation of the International System of Units for the United States,” Fed. Register **55** (245), 52 242 (20 December 1990).

The word “degree” and its symbol, °, are omitted from the unit of thermodynamic temperature T (that is, one uses kelvin or K, not degree Kelvin or °K). However, they are retained in the unit of Celsius temperature t , defined as $t \equiv T - T_0$, where $T_0 = 273.15$ K exactly (that is, degree Celsius, °C).

Symbols for prefixes representing 10^6 or greater are capitalized; all others are lower case. There is no space between the prefix and the unit. Compound prefixes are to be avoided (for example, pF, not $\mu\mu\text{F}$). An exponent applies to the whole unit including its prefix (for example, $\text{cm}^3 = 10^{-6} \text{ m}^3$). When a unit multiple or submultiple is written out in full, the prefix should be written in full, beginning with a lower-case letter (for example, megahertz, not Megahertz or Mhertz). The kilogram is the only base unit whose name, for historical reasons, contains a prefix; names of multiples and submultiples of the kilogram and their symbols are formed by attaching prefixes to the word “gram” and the symbol “g”.

Multiplication of units is indicated by inserting a raised dot or by leaving a space between the units (for example, N·m or N m). Division may be indicated by the use of the solidus, a horizontal fraction bar or a negative exponent (for example, m/s, $\frac{\text{m}}{\text{s}}$ or $\text{m}\cdot\text{s}^{-1}$) but repeated use of the solidus is not permitted (for example, m/s^2 , not $\text{m}/\text{s}/\text{s}$). To avoid possible misinterpretation when more than one unit appears in the denominator, the preferred practise is to use parentheses or negative exponents (for example, $\text{W}/(\text{m}^2\cdot\text{K}^4)$ or $\text{W}\cdot\text{m}^{-2}\cdot\text{K}^{-4}$). The unit expression may include a prefixed unit (for example, kJ/mol, W/cm^2).

Unit names should not be mixed with symbols for mathematical operations. (For example, one should write “meter per second” but not “meter/second” or “meter second⁻¹”). When spelling out the product of two units, a space is recommended (although a hyphen is permissible), but one should never use a centered dot. (Write, for example, “newton meter” or “newton-meter”, but not “newton·meter”).

Three-digit groups in numbers with more than four digits are separated by thin spaces instead of commas (for example, 229 792 458, not 299,792,458) to avoid confusion with the decimal marker in European literature. This spacing convention is also used to the right of the decimal marker. The numerical value and unit symbol must be separated by a space, even when used as an adjective (for example, 35 mm, not 35mm or 35-mm). A zero should be placed in front of the decimal marker in decimal fractions (for example, 0.3 J, not .3 J). The prefix of a unit should be chosen so that the numerical value will be within a practical range, usually between 0.1 and 1000 (for example, 200 kN, 0.5 mA).¹⁶

5.9.2.3 Non-SI units

An important function of the SI is to discourage the proliferation of unnecessary units. However, it is recognized that some units outside the SI are so well established that their use is to be permitted. Units in use with the SI are listed in table 4. As exceptions to the rules, the symbols °, ’ and ” for units of plane angle are not preceded by a space, and the symbol for liter, L, is capitalized to avoid confusion between the letter l and the number 1. Certain units whose values are obtained experimentally, listed in table 5, are also accepted for use in special fields.

¹⁶This footnote is from A. Tarantola, not from R. Nelson: Remark that “Three-digit groups (. . .) are separated by *thin spaces*”. In L^AT_EX document preparation system, for instance, a thin space is obtained by “\,”. I also use *thin spaces* to separate the numerical value and unit symbol (for example, 35 mm, not 35 mm), but I do not know if this is an explicit specification.

TABLE 4. Units in use with the SI

Quantity	Unit		
	Name	Symbol	Definition
time	minute	min	1 min = 60 s
	hour	h	1 h = 60 min = 3600 s
	day	d	1 d = 24 h = 86 400 s
plane angle	degree	°	1° = ($\pi/180$) rad
	minute	'	1' = (1/60)° = ($\pi/10\,800$) rad
	second	"	1" = (1/60)' = ($\pi/648\,800$) rad
volume	liter	L	1 L = 1 dm ³ = 10 ⁻³ m ³
mass	metric ton	t	1 t = 1000 kg
land area	hectare	ha	1 ha = 1 hm ² = 10 ⁴ m ²

TABLE 5. Units whose values are obtained experimentally

Quantity	Unit		
	Name	Symbol	Value
energy	electron volt	eV	$1.602\,177\,33(49) \times 10^{-19}$ J
mass	unified atomic mass unit	u	$1.660\,540\,2(10) \times 10^{-27}$ kg

Chapter 6

Inference Problems of the First Kind (Sum of Probabilities)

Note: Say here the we consider here the Problem of Making Histograms.

6.1 Experimental Histograms

[Note: This is a provisional text, to be expanded.]

Consider an n -dimensional manifold, with a volume element dv , and a probability distribution defined over it, represented by the (normalized) volumetric probability f . Although this is not necessary, let us simplify the exposition by assuming that some coordinates have been chosen over the manifold. Then, the probability distribution is represented by the volumetric probability function $f(\mathbf{x})$, and the volume distribution by the volume element function $dv(\mathbf{x})$.

Some process, mathematical or physical, produces points $\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_K$ that are samples of the probability distribution. Assume that we don't know $f(\mathbf{x})$, and that we wish to obtain a reasonable estimation of it, by measuring the coordinates of the points $\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_K$.

As any physical measure has some experimental uncertainties, the measure of the coordinates of the point \mathcal{P}_1 shall not produce some values \mathbf{x}_1 but, rather, an information about the coordinates of the point, that we can represent by the volumetric probability $f_1(\mathbf{x})$.

Let, then, $f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_K(\mathbf{x})$ be the (normalized) volumetric probabilities obtained when measuring the coordinates of the points $\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_K$.

When we have a large enough number of points, i.e., when K is large enough¹ we can start having some information about the probability distribution $f(\mathbf{x})$ itself.

Which volumetric probability $f(\mathbf{x})$ shall we choose to represent our information? Of course, the one that satisfies the postulates used in section 2.3 to define the 'sum' of probabilities.

We then arrive to the volumetric probability

$$f(\mathbf{x}) = \frac{1}{K} \sum_{i=1}^K f_i(\mathbf{x}) \quad . \quad (6.1)$$

This is the equivalent, but in a slightly more sophisticated manner, to 'making an histogram of the observed points'.

Example 6.1 *A seismologist has analyzed for many years the seismicity of a quite active region of the Earth. For every earthquake, using the arrival times of the seismic waves at some observatories, she/he has estimated its epicentral (geographic) coordinates $\{\varphi, \lambda\}$, obtaining the (2D) volumetric probabilities $f_1(\varphi, \lambda), f_2(\varphi, \lambda), \dots, f_K(\varphi, \lambda)$. If the next earthquake has to be a standard earthquake, the best estimate we have for the probability distribution of its epicentral coordinates (in the absence of any supplementary information) is that represented by the volumetric probability $f(\varphi, \lambda) = \frac{1}{K} \sum_{i=1}^K f_i(\varphi, \lambda)$. [End of example.]*

As suggested in chapter 2, let us write the volume element of the space as

$$dv(\mathbf{x}) = \bar{g}(\mathbf{x}) d\underline{v}(\mathbf{x}) \quad , \quad (6.2)$$

where $\bar{g}(\mathbf{x})$ and $d\underline{v}(\mathbf{x})$ are respectively the volume density and the capacity element of the space in the coordinates \mathbf{x} .

By definition of probability density (see section 2.2.3), the relation between a volumetric probability $h(\mathbf{x})$ and the associated probability density $\bar{h}(\mathbf{x})$ is

$$\bar{h}(\mathbf{x}) = \bar{g}(\mathbf{x}) h(\mathbf{x}) \quad . \quad (6.3)$$

¹How large is large enough? This depends, of course, on the relative radiuses of f and of the f_i , on the number of dimensions of the space, and on the relative degree of smoothness of the probability distributions.

Equation 6.1 can obviously also be written as

$$f(\mathbf{x}) = \frac{1}{K} \sum_{i=1}^K f_i(\mathbf{x}) \quad , \quad (6.4)$$

where, now only probability densities are invoked.

6.2 Sampling a Sum

Note: explain here that if we wish to obtain a sample of the volumetric probability

$$f(\mathbf{x}) = \frac{1}{K} \sum_{i=1}^K f_i(\mathbf{x}) \quad , \quad (6.5)$$

we can:

- first, select at random, with equal probability, a value i in the interval $1 \leq i \leq K$;
- then, obtain a sample of $f_i(\mathbf{x})$.

6.3 Further Work to be Done

Note: I have to prove here the following conjecture.

Consider a metric coordinate x over a one-dimensional metric space. Let $f(x)$ be a (1D) volumetric probability over the space, and let x_1, x_2, \dots be samples of it.

When trying to measure the coordinate x with a given instrument, assume that ‘the reading’ of the instrument is a value x' that is a sample of a volumetric probability $g(x'; x; \sigma)$ centered at x and with standard deviation σ . Given the reading x' , then the volumetric probability for the measurand is

$$h(x) = h(x; x', \sigma) = \text{WRITE THIS} \quad . \quad (6.6)$$

The readings have been x_1, x_2, \dots .

Then,

$$F(x) = k \sum_i h_i(x) = k \sum_i h(x; x_i; \sigma) = k \sum_i g(x_i; x; \sigma) \quad . \quad (6.7)$$

And I conjecture that the relation between the original $f(x)$ and our estimation $F(x)$ is

$$F(x) = \int dx' g(x', x, \sigma) f(x') \quad . \quad (6.8)$$

This, in fact, is a convolution.

Chapter 7

Inference Problems of the Second Kind (Product of Probabilities)

Note: write an introduction here.

7.1 The ‘Shipwrecked Person’ Problem

Note: this example is to be developed. For the time being this is just a copy of example 2.4

Let \mathbf{S} represent the surface of the Earth, using geographical coordinates (longitude φ and latitude λ). 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 (2D) volumetric probability $f(\varphi, \lambda)$, and an independent, simultaneous estimation of the position by another airplane navigator gives a probability distribution corresponding to the volumetric probability $g(\varphi, \lambda)$. How the two volumetric probabilities $f(\varphi, \lambda)$ and $g(\varphi, \lambda)$ should be ‘combined’ to obtain a ‘resulting’ volumetric probability? The answer is given by the ‘product’ of the two volumetric probabilities densities:

$$(f \cdot g)(\varphi, \lambda) = \frac{f(\varphi, \lambda) g(\varphi, \lambda)}{\int_{\mathbf{S}} dS(\varphi, \lambda) f(\varphi, \lambda) g(\varphi, \lambda)} . \quad (7.1)$$

7.2 Physical Laws as Probabilistic Correlations

7.2.1 Physical Laws

Are we forced to introduce uncertainties in physical laws to be used as ‘thicknesses’ of a mathematical function $\mathbf{d} = \mathbf{g}(\mathbf{m})$ via a metric in the space?

In fact, actual theories are always approximate and they have some ‘uncertainty bars’ associated to them (see an example in section 7.2.2). The conditional volumetric probability has to be seen as a way of taking a limit when the uncertainty bars tend to zero. Then the sort of limit defining the conditional probability density is imposed by the form of the ‘theoretical uncertainty bars’. Rather than basing inversion theory on an expression like 8.16, it is better to introduce explicitly the theoretical uncertainties, and take any ‘small uncertainty limit’ afterwards. Let us do this.

Assume 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 $\vartheta(\mathbf{m}, \mathbf{d})$. Then, the conjunction of the ‘a priori and experimental information’ contained in $\rho(\mathbf{m}, \mathbf{d})$ and the ‘theoretical information’ contained in $\vartheta(\mathbf{m}, \mathbf{d})$ can be combined using the conjunction operation defined by equation ??, to give

$$\sigma(\mathbf{m}, \mathbf{d}) = k \frac{\rho(\mathbf{m}, \mathbf{d}) \vartheta(\mathbf{m}, \mathbf{d})}{\mu(\mathbf{m}, \mathbf{d})} , \quad (7.2)$$

where $\mu(\mathbf{m}, \mathbf{d})$ is the homogeneous probability density. The implications of this equation will be examined later.

7.2.2 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,

$$x(t) = x_0 + v_0 t + \frac{1}{2} g t^2 . \quad (7.3)$$

Of course, for 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 σ_x ;

- the initial velocity of the apple is random, with a Gaussian distribution centered at v_0 , and with standard deviation σ_v ;

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

$$\vartheta(x|t) = \frac{1}{\sqrt{2\pi} \sqrt{\sigma_x^2 + \sigma_v^2 t^2}} \exp\left(-\frac{1}{2} \frac{(x - (x_0 + v_0 t + \frac{1}{2} g t^2))^2}{\sigma_x^2 + \sigma_v^2 t^2}\right). \quad (7.4)$$

This is obviously a conditional probability density for x , given t . If we select the time t randomly with homogeneous probability distribution (i.e., if we assume that the marginal probability density for t is constant), then the joint probability density for x and t is

$$\vartheta(x, t) = k \vartheta(x|t) \quad (7.5)$$

where k is a constant, and where $\vartheta(x|t)$ is that in equation 7.4. This probability density is represented in figure 7.1, together with the two marginals, and the conditional probability density at three different times is represented in figure 7.2.

Figure 7.1: 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 equation 7.5, with $x_0 = 0$, $v_0 = 1 \text{ m/s}$, $\sigma_x = 1 \text{ m}$, $\sigma_v = 1 \text{ m/s}$ and $g = 9.91 \text{ m/s}^2$. 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’).

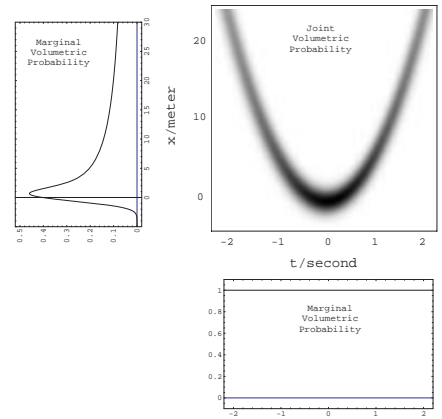
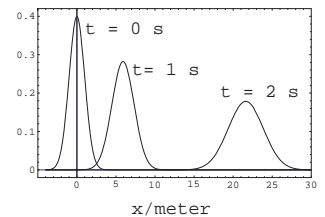


Figure 7.2: Three conditional volumetric probabilities from the joint distribution of the previous figure at times $t = 0$, $t = 1 \text{ s}$ and $t = 2 \text{ s}$. The width increases with time because of the uncertainty in the initial velocity.



7.2.3 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 (\mathbf{M}, \mathbf{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 (\mathbf{M}, \mathbf{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

$$\rho(\mathbf{m}, \mathbf{d}) \quad (\text{experimental information}) . \quad (7.6)$$

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 problems¹. Let $\vartheta(\mathbf{m}, \mathbf{d})$ be the probability density representing this ‘theoretical information’:

$$\vartheta(\mathbf{m}, \mathbf{d}) \quad (\text{theoretical information}) . \quad (7.7)$$

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 4):

$$\mu(\mathbf{m}, \mathbf{d}) \quad (\text{homogeneous probability distribution}) . \quad (7.8)$$

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

$$\sigma(\mathbf{m}, \mathbf{d}) = k \frac{\rho(\mathbf{m}, \mathbf{d}) \vartheta(\mathbf{m}, \mathbf{d})}{\mu(\mathbf{m}, \mathbf{d})} \quad (\text{conjunction of informations}) , \quad (7.9)$$

with marginal probability densities²

$$\sigma_m(\mathbf{m}) = \int_{\mathbf{D}} d\mathbf{d} \sigma(\mathbf{m}, \mathbf{d}) \quad ; \quad \sigma_d(\mathbf{d}) = \int_{\mathbf{M}} d\mathbf{m} \sigma(\mathbf{m}, \mathbf{d}) . \quad (7.10)$$

Example 7.1 *We may assume that the physical correlations between the parameters \mathbf{m} and \mathbf{d} are of the form*

$$\vartheta(\mathbf{m}, \mathbf{d}) = \vartheta_{D|M}(\mathbf{d}|\mathbf{m}) \vartheta_M(\mathbf{m}) , \quad (7.11)$$

this expressing that a ‘physical theory’ gives, on the one hand, the conditional probability for \mathbf{d} , given \mathbf{m} , and on the other hand, the marginal probability density for \mathbf{m} . [End of example.]

¹Remember 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.

²As explained in section ??, the definition of marginal probability density is only intrinsic if the total space is the Cartesian product of the two spaces, i.e., when $(\mathbf{M}, \mathbf{D}) = \mathbf{M} \times \mathbf{D}$.

Example 7.2 Many applications concern the special situation where we have

$$\mu(\mathbf{m}, \mathbf{d}) = \mu_m(\mathbf{m}) \mu_d(\mathbf{d}) \quad ; \quad \rho(\mathbf{m}, \mathbf{d}) = \rho_m(\mathbf{m}) \rho_d(\mathbf{d}) \quad . \quad (7.12)$$

In this case, equations 7.9–7.10 give

$$\sigma_m(\mathbf{m}) = k \frac{\rho_m(\mathbf{m})}{\mu_m(\mathbf{m})} \int_{\mathbf{D}} d\mathbf{d} \frac{\rho_d(\mathbf{d}) \vartheta(\mathbf{m}, \mathbf{d})}{\mu_d(\mathbf{d})} \quad ; \quad (7.13)$$

and

$$\sigma_d(\mathbf{d}) = k \frac{\rho_d(\mathbf{d})}{\mu_d(\mathbf{d})} \int_{\mathbf{M}} d\mathbf{m} \frac{\rho_m(\mathbf{m}) \vartheta(\mathbf{m}, \mathbf{d})}{\mu_m(\mathbf{m})} \quad . \quad (7.14)$$

If equation 7.11 holds, then

$$\sigma_m(\mathbf{m}) = k \rho_m(\mathbf{m}) \frac{\vartheta_m(\mathbf{m})}{\mu_m(\mathbf{m})} \int_{\mathbf{D}} d\mathbf{d} \frac{\rho_d(\mathbf{d}) \vartheta_{D|M}(\mathbf{d} | \mathbf{m})}{\mu_d(\mathbf{d})} \quad (7.15)$$

and

$$\sigma_d(\mathbf{d}) = k \frac{\rho_d(\mathbf{d})}{\mu_d(\mathbf{d})} \int_{\mathbf{M}} d\mathbf{m} \rho_m(\mathbf{m}) \vartheta_{D|M}(\mathbf{d} | \mathbf{m}) \frac{\vartheta_m(\mathbf{m})}{\mu_m(\mathbf{m})} \quad . \quad (7.16)$$

Finally, if the simplification $\vartheta_M(\mathbf{m}) = \mu_m(\mathbf{m})$ arises (this usually holds only if nonlinearities are weak³), then,

$$\sigma_m(\mathbf{m}) = k \rho_m(\mathbf{m}) \int_{\mathbf{D}} d\mathbf{d} \frac{\rho_d(\mathbf{d}) \vartheta(\mathbf{d} | \mathbf{m})}{\mu_d(\mathbf{d})} \quad (7.17)$$

and

$$\sigma_d(\mathbf{d}) = k \frac{\rho_d(\mathbf{d})}{\mu_d(\mathbf{d})} \int_{\mathbf{M}} d\mathbf{m} \rho_m(\mathbf{m}) \vartheta(\mathbf{d} | \mathbf{m}) \quad . \quad (7.18)$$

[End of example.]

Example 7.3 Let us reproduce here equation 7.17,

$$\sigma_m(\mathbf{m}) = k \rho_m(\mathbf{m}) \int_{\mathbf{D}} d\mathbf{d} \frac{\rho_d(\mathbf{d}) \vartheta(\mathbf{d} | \mathbf{m})}{\mu_d(\mathbf{d})} \quad . \quad (7.19)$$

Assume that observational uncertainties are Gaussian,

$$\rho_d(\mathbf{d}) = k \exp \left(-\frac{1}{2} (\mathbf{d} - \mathbf{d}_{\text{obs}})^t \mathbf{C}_D^{-1} (\mathbf{d} - \mathbf{d}_{\text{obs}}) \right) \quad . \quad (7.20)$$

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:

$$\vartheta(\mathbf{d} | \mathbf{m}) = k \exp \left(-\frac{1}{2} (\mathbf{d} - \mathbf{f}(\mathbf{m}))^t \mathbf{C}_T^{-1} (\mathbf{d} - \mathbf{f}(\mathbf{m})) \right) \quad . \quad (7.21)$$

³Note: some explanation is needed here.

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 7.19 can be analytically evaluated, and gives

$$\int_{\mathbf{D}} d\mathbf{d} \frac{\rho_d(\mathbf{d}) \vartheta(\mathbf{d}|\mathbf{m})}{\mu_d(\mathbf{d})} = k \exp\left(-\frac{1}{2}(\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^t (\mathbf{C}_D + \mathbf{C}_T)^{-1} (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}})\right). \quad (7.22)$$

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 7.4 In the ‘Galilean law’ example developed in section 7.2.1, 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 7.9, gives

$$\sigma(x, t) = k \frac{\rho(x, t) \vartheta(x, t)}{\mu(x, t)}, \quad (7.23)$$

where, as the coordinates (x, t) are ‘Cartesian’, $\mu(x, t) = k$. Taking the explicit expression given for $\vartheta(x, t)$ in equations 7.24–7.5,

$$\vartheta(x, t) = \frac{1}{\sqrt{2\pi} \sqrt{\sigma_x^2 + \sigma_v^2 t^2}} \exp\left(-\frac{1}{2} \frac{(x - (x_0 + v_0 t + \frac{1}{2} g t^2))^2}{\sigma_x^2 + \sigma_v^2 t^2}\right), \quad (7.24)$$

and assuming the Gaussian form⁴ for $\rho(x, t)$,

$$\rho(x, t) = \rho_x(x) \rho_t(t) = k \exp\left(-\frac{1}{2} \frac{(x - x_{\text{obs}})^2}{\Sigma_x^2}\right) \exp\left(-\frac{1}{2} \frac{(t - t_{\text{obs}})^2}{\Sigma_t^2}\right), \quad (7.25)$$

we obtain the combined probability density

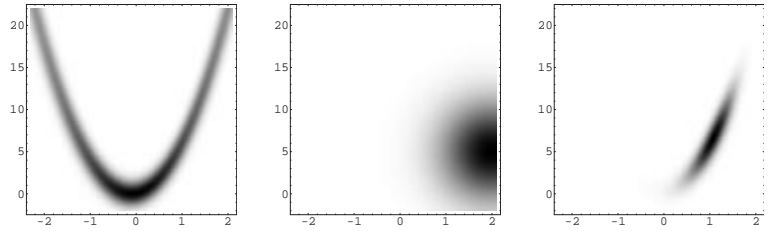
$$\sigma(x, t) = \frac{k}{\sqrt{\sigma_x^2 + \sigma_v^2 t^2}} \exp\left(-\frac{1}{2} \left(\frac{(x - x_{\text{obs}})^2}{\Sigma_x^2} + \frac{(t - t_{\text{obs}})^2}{\Sigma_t^2} + \frac{(x - (x_0 + v_0 t + \frac{1}{2} g t^2))^2}{\sigma_x^2 + \sigma_v^2 t^2} \right)\right). \quad (7.26)$$

Figure 7.3 illustrates the three probability densities $\vartheta(x, t)$, $\rho(x, t)$ and $\sigma(x, t)$. **[End of example.]**

Note: explain here that $\delta(\mathbf{d} - \mathbf{f}(\mathbf{m}))$, as it concerns a *difference* in the data space (rather than a distance), it is not a mathematically nice object.

⁴Note that taking the limit of $\vartheta(x, t)$ or of $\rho(x, t)$ for infinite variances we obtain $\mu(x, t)$, as we should.

Figure 7.3: Note: this is a provisional figure. It was made with the numerical values mentioned in figure 7.1 with, in addition, $x_{\text{obs}} = 5.0 \text{ m}$, $\Sigma_x = 4.0 \text{ m}$, $t_{\text{obs}} = 2.0 \text{ s}$ and $\Sigma_t = 0.75 \text{ s}$.



Example 7.5 Note: consider here equation 7.11 and let us formally take

$$\vartheta(\mathbf{d}|\mathbf{m}) = \delta(\mathbf{d} - \mathbf{f}(\mathbf{m}))$$

$$\vartheta_M(\mathbf{m}) = k \sqrt{\det(\mathbf{g}_{mm} + \mathbf{g}_{md} \mathbf{F} + \mathbf{F}^T \mathbf{g}_{dm} + \mathbf{F}^T \mathbf{g}_{dd} \mathbf{F})} \Big|_{\mathbf{d}=\mathbf{f}(\mathbf{m})}. \quad (7.27)$$

[Note: Explain this choice for $\vartheta_M(\mathbf{m})$...] Then we arrive at

$$\sigma_{\mathbf{m}}(\mathbf{m}) = k \rho(\mathbf{m}, \mathbf{f}(\mathbf{m})) \frac{\sqrt{\det(\mathbf{g}_{mm} + \mathbf{g}_{md} \mathbf{F} + \mathbf{F}^T \mathbf{g}_{dm} + \mathbf{F}^T \mathbf{g}_{dd} \mathbf{F})}}{\mu(\mathbf{m}, \mathbf{d})} \Big|_{\mathbf{d}=\mathbf{f}(\mathbf{m})}. \quad (7.28)$$

If $\mu(\mathbf{m}, \mathbf{d}) = k \sqrt{\det \mathbf{g}(\mathbf{m}, \mathbf{d})}$ (i.e., if we use the same metric to represent theoretical uncertainties as we used to define the homogeneous probability distributions), this equation is identical to equation 8.16, obtained using the equation $\mathbf{d} = \mathbf{f}(\mathbf{m})$ to define a conditional probability. [End of example.]

The previous example is important because it shows that the formulation using an ‘exact physical law’ can be found as a particular case of this, more general, approach were physical correlations are represented probabilistically.

Chapter 8

Inference Problems of the Third Kind (Conditional Probabilities)

Note: Say here the we consider here two problems: (i) ‘adjusting measurements’ to a physical theory and (ii) resolution of Inverse problems.

These two problems are mathematically very similar, and are essentially solved using either the notion of ‘conditional probability’ or the notion of ‘product of probabilities’ (see chapter 2).

Note: what follows comes from an old text:

A so-called ‘inverse problem’ usually consists in a sort quite complex measurement, sometimes a gigantic measurement, involving years of observations and thousands of instruments. Any measurement is indirect (we may weigh a mass by observing the displacement of the cursor of a balance), and as such, a possibly nontrivial analysis of uncertainties must be done.

Any good 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 any measurement involves, at least, two different sources of uncertainties: those that we estimate using statistical methods, and those that we estimate 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.

8.1 Adjusting Measurements to a Physical Theory

When a particle of mass m is submitted to a force F , one has

$$F = m \frac{d}{dt} \frac{v}{\sqrt{1 - v^2/c^2}} . \quad (8.1)$$

Assuming initial conditions of rest (at a time arbitrarily set to 0), the trajectory of the particle is

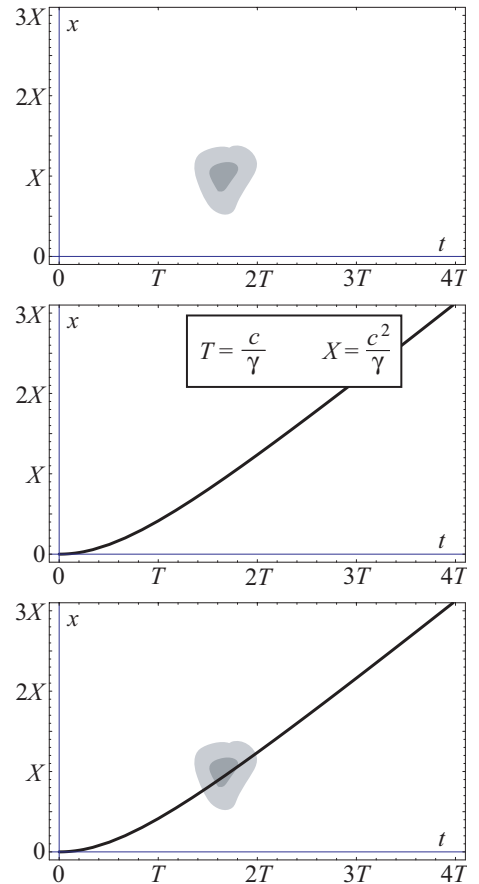
$$x(t) = \frac{c^2}{\gamma} \left(\sqrt{1 + \left(\frac{\gamma t}{c} \right)^2} - 1 \right) , \quad (8.2)$$

where

$$\gamma = \frac{F}{m} . \quad (8.3)$$

Note: introduce here the problem set in the caption of figure 8.1. Say, in particular that we have a measurement whose results are represented by the volumetric probability $f(t, x)$.

Figure 8.1: In the space-time of special relativity, we have measured the space-time coordinates of an event, and obtained the volumetric probability $f(t, x)$ displayed in the figure at the top. We then learn that that event happened on the trajectory of a particle with mass m submitted to a constant force F (equation 8.2). This trajectory is represented in the figure at the middle. It is clear that thanks to the theory, we can ameliorate the knowledge of the coordinates of the event, by considering the conditional volumetric probability induced on the trajectory. See text for details.



The problem here, is clearly a problem of conditional probability, and it makes sense because we do have a metric over our 2D space, the Minkowskian metric

$$ds^2 = dt^2 - \frac{1}{c^2} dx^2 . \quad (8.4)$$

With respect to the notations in section 2.4.2.2, we have here $\mathbf{r} = r = t$, and $\mathbf{s} = s = x$, and the relation $\mathbf{s} = \mathbf{s}(\mathbf{r})$ is, here, the relation $x = x(t)$ given by equation 8.2.

As we have, here, $\sqrt{\det(\mathbf{g}_r + \mathbf{S}^t \mathbf{g}_s \mathbf{S})} = c/\sqrt{1 + (\gamma t/c)^2}$, A direct use of equation 2.127 gives the (1D) volumetric probability over the time variable

$$\boxed{f_t(t) = \frac{k}{\sqrt{1 + (\gamma t/c)^2}} f(t, x)|_{x=x(t)} \quad ,} \quad (8.5)$$

where k is the normalization constant ensuring that

$$\int_0^\infty dt f_t(t) = 1 \quad , \quad (8.6)$$

and where $x = x(t)$ is a short-hand notation for the relation 8.2.

Note: I have now to transport this volumetric probability over the time axis into a volumetric probability over the x axis, using the transport of probabilities introduced in section 2.6.

Note: I have to convince the reader here that we can not give an intrinsic definition of this problem inside the Galilean physics, as there is no space-time metric. **This is very important, and enforces my decision to use a metric definition of the conditional volumetric probabilities.**

8.2 Inverse Problems

[*Note: Complete and expands what follows.*]

In the so called ‘inverse problems’, values of the parameters describing physical systems are estimated, using as data some indirect measurements. A consistent formulation of inverse problems can be made using the concepts of probability theory. Data and attached uncertainties, (a possibly vague) a priori information on model parameters, and a physical theory relating the model parameters to the observations are the fundamental elements of any inverse problem. While the most general solution of the inverse problem requires extensive use of Monte Carlo methods, special hypothesis (e.g., Gaussian uncertainties) allow, in some cases, to solve part of the problem analytically (e.g., using the method of least squares).

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 (the solution is not unique). In the seismic example, two different Earth models may predict the same seismograms¹, the finite bandwidth of our data sets 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 accepted. I only like this name moderately, as I see the problem more as a problem of ‘conjunction of states of information’ (theoretical, experimental and a priori 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 observation in a realistic situation where the parameters describing the Earth model are not ‘given’, but only known approximately.

In fact, I 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’.

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

8.2.1 Model Parameters and Observable Parameters

Although the separation of all the variables of a problem in two groups 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 $\mathbf{m} = \{m^1, m^2, \dots, m^{\text{NM}}\}$ 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 may make some indirect measurement 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 $\mathbf{d} = \{d^1, d^2, \dots, d^{\text{ND}}\}$.

We assume that we have a physical theory that solves 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

$$d^i = g^i(m^1, m^2, \dots, m^{\text{NM}}) \quad ; \quad i = 1, 2, \dots, \text{ND} \quad , \quad (8.7)$$

or, for short,

$$\mathbf{d} = \mathbf{f}(\mathbf{m}) \quad . \quad (8.8)$$

In fact, it is this expression that separates the whole set of our parameters into the subsets \mathbf{d} and \mathbf{m} , as sometimes there is no difference of 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 some seismic observatories, and we need to put in \mathbf{m} the coordinates of the observatories —as these are parameters that are needed to compute the travel times—, although we estimate arrival times of waves as well as coordinates of the observatories using similar types of measurements.

8.2.2 A Priori 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 *a priori 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 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 volumetric probability that we denote as $\rho_{\mathbf{m}}(\mathbf{m})$.

If an explicit expression for the volumetric probability $\rho_{\mathbf{m}}(\mathbf{m})$ is known, then it can be used in analytical developments. But such an explicit expression is, by no means, necessary.

All that is needed is a set of probabilistic rules that allows us to generate samples of $\rho_{\mathbf{m}}(\mathbf{m})$ in the model space (random samples distributed according to $\rho_{\mathbf{m}}(\mathbf{m})$).

Example 8.1 Gaussian a priori Information.

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 volumetric probability is not the normal, but the lognormal). But if we have chosen the right parameters (for instance, taking the logarithms of all Jeffreys parameters), it may happen that the Gaussian probabilistic model is acceptable. We then have

$$\rho_{\mathbf{m}}(\mathbf{m}) = k \exp\left(-\frac{1}{2}(\mathbf{m} - \mathbf{m}_{\text{prior}})^T \mathbf{C}_{\text{prior}}^{-1}(\mathbf{m} - \mathbf{m}_{\text{prior}})\right) . \quad (8.9)$$

When this Gaussian volumetric probability is used, $\mathbf{m}_{\text{prior}}$, the center of the Gaussian is called the ‘a priori model’ while $\mathbf{C}_{\text{prior}}$ is called the ‘a priori covariance matrix’. The name ‘a priori model’ is dangerous, as for large dimensional problems, the average model may not be a good representative of the models that can be obtained as samples of the distribution (see figure 8.27 as an example). Other usual sources of prior information are the ranges and distribution of media properties in the rocks, or probabilities for the localization of media discontinuities. If the information refers to marginals of the model parameters, and is not including the description of relations across model parameters, the prior volumetric probability reduces to a product of univariate densities, $\rho_{\mathbf{m}}(\mathbf{m}) = \prod_i \rho_i(m_i)$. The next example illustrates this case. [End of example.]

Example 8.2 Prior Information for a 1D Mass Density Model

We 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 8.2, 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, $\mathbf{m} = \{\ell_1, \ell_2, \dots, \ell_{NL}, \rho_1, \rho_2, \dots, \rho_{NL}\}$. The marginal prior probability densities for the layer thicknesses are all assumed to be identical and of the form (exponential volumetric probability)

$$f(\ell) = \frac{1}{\ell_0} \exp\left(-\frac{\ell}{\ell_0}\right) , \quad (8.10)$$

where the constant ℓ_0 has the value $\ell_0 = 4$ km (see the left of figure 8.2), while all the marginal prior probability densities for the mass density are also assumed to be identical, and of the form (lognormal volumetric probability)

$$g(\rho) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{1}{2\sigma^2} \left(\log \frac{\rho}{\rho_0}\right)^2\right) , \quad (8.11)$$

where $\rho_0 = 3.98$ g/cm³ and $\sigma = 0.58$ (see the right of figure 8.2). 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 a priori volumetric probability in this problem is the product of a priori probability densities (equations 8.10 and 8.11) for each parameter,

$$\rho_{\mathbf{m}}(\mathbf{m}) = \rho_{\mathbf{m}}(\ell_1, \ell_2, \dots, \ell_{NL}, \rho_1, \rho_2, \dots, \rho_{NL}) = k \prod_i^{NL} f(\rho_i) g(\rho_i) \quad . \quad (8.12)$$

Figure 8.3 shows (pseudo) random models generated according to this probability distribution. Of course, the explicit expression 8.12 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 8.10 and 8.11. **[End of example.]**

Figure 8.2: At left, the probability density for the layer thickness. At right, the probability density for the density of mass.

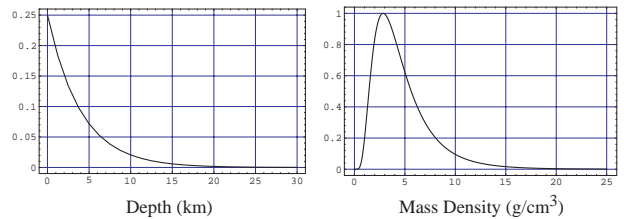
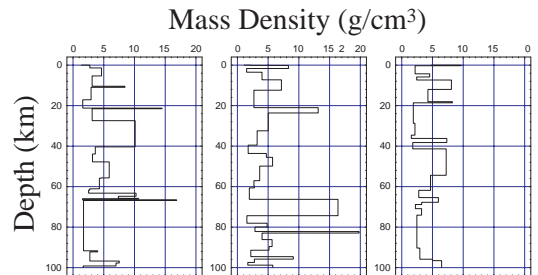


Figure 8.3: Three random Earth models generated according to the a priori probability density in the model space.



8.2.3 Measurements and Experimental Uncertainties

Note: the text that was here has been moved to section 5.8.

8.2.4 Joint ‘Prior’ Probability Distribution in the (\mathbf{M}, \mathbf{D}) Space

We have just seen that the a priori information on model parameters can be described by a volumetric probability in the model space, $\rho_{\mathbf{m}}(\mathbf{m})$, and that the result of measurements can be described by a volumetric probability in the data space $\rho_d(\mathbf{d})$. As by ‘a priori’ information on model parameters we mean information obtained independently from the measurements, we can multiply these two volumetric probabilities (see section 2.5.5 on Independent Probability Distributions) to define a joint volumetric probability in the $\mathbf{X} = (\mathbf{M}, \mathbf{D})$ space.

$$\rho(\mathbf{x}) = \rho(\mathbf{m}, \mathbf{d}) = \rho_{\mathbf{m}}(\mathbf{m}) \rho_d(\mathbf{d}) \quad . \quad (8.13)$$

Although we have introduced $\rho_{\mathbf{m}}(\mathbf{m})$ and $\rho_d(\mathbf{d})$ separately, and we have suggested to build a probability distribution in the (\mathbf{M}, \mathbf{D}) space by the multiplication 8.13, 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 (\mathbf{M}, \mathbf{D}) space, represented by the volumetric probability $\rho(\mathbf{x}) = \rho(\mathbf{m}, \mathbf{d})$ and let us contemplate equation 8.13 as just a special case.

8.2.5 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

$$\mathbf{d} = \mathbf{d}(\mathbf{m}) \quad . \quad (8.14)$$

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 $\mathbf{m} \rightarrow \mathbf{d} = \mathbf{d}(\mathbf{m})$.

At this point, given the volumetric probability $\rho(\mathbf{m}, \mathbf{d})$ and given the relation $\mathbf{d} = \mathbf{d}(\mathbf{m})$, one may wish to define the associated conditional volumetric probability. But we have emphasized in chapter 2 that there is no way to define a conditional volumetric probability given only an equation like $\mathbf{d} = \mathbf{d}(\mathbf{m})$: we must, in addition, specify a metric in the (\mathbf{M}, \mathbf{D}) space², that we may denote here by

$$\mathbf{g}(\mathbf{m}, \mathbf{d}) = \begin{pmatrix} \mathbf{g}_m(\mathbf{m}) & \mathbf{0} \\ \mathbf{0} & \mathbf{g}_d(\mathbf{d}) \end{pmatrix} \quad , \quad (8.15)$$

where, to simplify the exposition, I assume the special case where the metric partitions into a metric $\mathbf{g}_m(\mathbf{m})$ in the model space \mathbf{M} and a metric $\mathbf{g}_d(\mathbf{d})$ in the data space \mathbf{D} .

8.2.6 Inverse Problems

In the $\mathbf{X} = (\mathbf{M}, \mathbf{D})$ space, we have the volumetric probability $\rho(\mathbf{m}, \mathbf{d})$, and we have the hypersurface defined by the relation $\mathbf{d} = \mathbf{d}(\mathbf{m})$. We can ‘combine’ these two kinds of information by using the conditional volumetric probability deduced from $\rho(\mathbf{m}, \mathbf{d})$ on the hypersurface $\mathbf{d} = \mathbf{d}(\mathbf{m})$ (see equation 2.127)

$$\sigma_m(\mathbf{m}) = k \rho(\mathbf{m}, \mathbf{d}(\mathbf{m})) \frac{\sqrt{\det(\mathbf{g}_m + \mathbf{D}^T \mathbf{g}_d \mathbf{D})}}{\sqrt{\det \mathbf{g}_m}} \quad , \quad (8.16)$$

where $\mathbf{D} = \mathbf{D}(\mathbf{m})$ is the matrix of partial derivatives, with components $D^i_\alpha = \partial d^i / \partial m^\alpha$, where $\mathbf{g}_m = \mathbf{g}_m(\mathbf{m})$ and where $\mathbf{g}_d = \mathbf{g}_d(\mathbf{d}(\mathbf{m}))$.

The probability of a finite domain \mathcal{A} of the model space is then to be evaluated as

$$P(\mathcal{A}) = \int_{\mathcal{A}} dm^1 \wedge \cdots \wedge dm^{NM} \sqrt{\det \mathbf{g}_m} \sigma_m(\mathbf{m}) \quad . \quad (8.17)$$

Example 8.3 *In the particular case where*

$$\rho(\mathbf{m}, \mathbf{d}) = \rho_m(\mathbf{m}) \rho_d(\mathbf{d}) \quad , \quad (8.18)$$

²Or, at least in the vicinity of the submanifold $\mathbf{d} = \mathbf{d}(\mathbf{m})$.

equation 8.16 becomes

$$\sigma_{\mathbf{m}}(\mathbf{m}) = k \rho_{\mathbf{m}}(\mathbf{m}) \rho_{\mathbf{d}}(\mathbf{d}(\mathbf{m})) \frac{\sqrt{\det(\mathbf{g}_{\mathbf{m}} + \mathbf{D}^T \mathbf{g}_{\mathbf{d}} \mathbf{D})}}{\sqrt{\det \mathbf{g}_{\mathbf{m}}}} \quad , \quad (8.19)$$

where, again, $\mathbf{D} = \mathbf{D}(\mathbf{m})$, $\mathbf{g}_{\mathbf{m}} = \mathbf{g}_{\mathbf{m}}(\mathbf{m})$ and $\mathbf{g}_{\mathbf{d}} = \mathbf{g}_{\mathbf{d}}(\mathbf{d}(\mathbf{m}))$. [End of example.]

Example 8.4 The conditional volumetric probability has been defined by taking an ‘orthogonal limit’. Should one have some reason to prefer the ‘vertical limit’, it can be obtained here by formally taking the limit $\mathbf{g}_{\mathbf{d}} \rightarrow \mathbf{0}$. Then, equation 8.19 simplifies into

$$\sigma_{\mathbf{m}}(\mathbf{m}) = k \rho_{\mathbf{m}}(\mathbf{m}) \rho_{\mathbf{d}}(\mathbf{d}(\mathbf{m})) \quad , \quad (8.20)$$

where the partial derivatives \mathbf{D} don’t appear. [End of example.]

Example 8.5 Gaussian Case. Let us examine here how equation 8.19 simplifies when assuming that the ‘input’ probability densities are Gaussian, and that the weight matrices (inverse of the covariance matrices) are the metric matrices (note: explain this, and give here the argument that the accuracy of a theory is, ultimately, the accuracy of the experiments used to control it):

$$\rho_{\mathbf{m}}(\mathbf{m}) = k \exp \left(-\frac{1}{2} (\mathbf{m} - \mathbf{m}_{\text{prior}})^t \mathbf{g}_{\mathbf{m}} (\mathbf{m} - \mathbf{m}_{\text{prior}}) \right) \quad (8.21)$$

$$\rho_{\mathbf{d}}(\mathbf{d}) = k \exp \left(-\frac{1}{2} (\mathbf{d} - \mathbf{d}_{\text{obs}})^t \mathbf{g}_{\mathbf{d}} (\mathbf{d} - \mathbf{d}_{\text{obs}}) \right) \quad . \quad (8.22)$$

Equation 8.19 then gives

$$\begin{aligned} \sigma_{\mathbf{m}}(\mathbf{m}) &= k \frac{\sqrt{\det(\mathbf{g}_{\mathbf{m}} + \mathbf{D}^t(\mathbf{m}) \mathbf{g}_{\mathbf{d}} \mathbf{D}(\mathbf{m}))}}{\sqrt{\det \mathbf{g}_{\mathbf{m}}}} \times \\ &\times \exp \left(-\frac{1}{2} \left((\mathbf{m} - \mathbf{m}_{\text{prior}})^t \mathbf{g}_{\mathbf{m}} (\mathbf{m} - \mathbf{m}_{\text{prior}}) + (\mathbf{d}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^t \mathbf{g}_{\mathbf{d}} (\mathbf{d}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) \right) \right) \end{aligned} \quad (8.23)$$

(the constant factor $\sqrt{\det \mathbf{g}_{\mathbf{m}}}$ has been left for subsequent simplifications). Defining the misfit

$$S(\mathbf{m}) = -\log \frac{\sigma_{\mathbf{m}}(\mathbf{m})}{\sigma_0} \quad , \quad (8.24)$$

where σ_0 is an arbitrary value of $\sigma_{\mathbf{m}}(\mathbf{m})$, gives, up to an additive constant,

$$S(\mathbf{m}) = S_1(\mathbf{m}) - S_2(\mathbf{m}) \quad , \quad (8.25)$$

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

$$2 S_1(\mathbf{m}) = (\mathbf{m} - \mathbf{m}_{\text{prior}})^t \mathbf{g}_{\mathbf{m}} (\mathbf{m} - \mathbf{m}_{\text{prior}}) + (\mathbf{d}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^t \mathbf{g}_{\mathbf{d}} (\mathbf{d}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) \quad (8.26)$$

and where (as $\log \sqrt{\mathbf{A}} = \frac{1}{2} \log \mathbf{A}$)

$$2 S_2(\mathbf{m}) = \log \frac{\det(\mathbf{g}_{\mathbf{m}} + \mathbf{D}^t(\mathbf{m}) \mathbf{g}_{\mathbf{d}} \mathbf{D}(\mathbf{m}))}{\det \mathbf{g}_{\mathbf{m}}} \quad . \quad (8.27)$$

The maximum likelihood point is defined as the point where the volumetric probability is maximum³. If $\hat{\gamma}$ denotes the gradient of the misfit,

$$\hat{\gamma}_\alpha = \frac{\partial S}{\partial m^\alpha} \quad , \quad (8.28)$$

then, the steepest ascent direction is the vector γ defined through

$$\mathbf{g}_m \gamma = \hat{\gamma} \quad . \quad (8.29)$$

The algorithm

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \epsilon_k \gamma_k \quad , \quad (8.30)$$

where ϵ_k is an ad-hoc, well chosen number, called the algorithm of steepest descent, converges to the maximum likelihood point (or, at least, to a local maximum). To ensure convergence, it is sufficient to use a descent direction, not necessarily the steepest one. This, in practice, allows two simplifications: (i) compute only an approximation to the gradient, (ii) use physical intuition to define directions that are better (for finite jumps) than the locally steepest one. In many applications, it is the gradient of $S_1(\mathbf{m})$ that is computed, not that of $S(\mathbf{m}) = S_1(\mathbf{m}) + S_2(\mathbf{m})$, and this gradient is approximated by dropping the derivatives of $\mathbf{D}(\mathbf{m})$ (i.e., second derivatives of $\mathbf{d}(\mathbf{m})$). One then has $\hat{\gamma}_k \approx \mathbf{g}_m(\mathbf{m}_k - \mathbf{m}_{\text{prior}}) + \mathbf{D}_k^t \mathbf{g}_d(\mathbf{d}_k - \mathbf{d}_{\text{obs}})$, where $\mathbf{D}_k = \mathbf{D}(\mathbf{m}_k)$ and $\mathbf{d}_k = \mathbf{d}(\mathbf{m}_k)$. Using the relation between gradient and steepest descent (equation 8.29) this gives⁴

$$\mathbf{g}_m \gamma_k \approx \mathbf{g}_m(\mathbf{m}_k - \mathbf{m}_{\text{prior}}) + \mathbf{D}_k^t \mathbf{g}_d(\mathbf{d}_k - \mathbf{d}_{\text{obs}}) \quad . \quad (8.31)$$

The two equations 8.30–8.31 encapsulate the algorithm of steepest descent. Once the algorithm has converged, if the volumetric probability $\sigma_m(\mathbf{m})$ is approximated by a Gaussian centered on the maximum likelihood point, then, the weight matrix (inverse of the covariance matrix) of the Gaussian is (see equation 2.144)

$$\mathbf{g}'_m = \mathbf{g}_m + \mathbf{D}^t \mathbf{g}_d \mathbf{D} \quad . \quad (8.32)$$

[End of example.]

Example 8.6 If the ‘relation solving the forward problem’ $\mathbf{d} = \mathbf{d}(\mathbf{m})$ happens to be a linear relation,

$$\mathbf{d} = \mathbf{D} \mathbf{m} \quad , \quad (8.33)$$

then the volumetric probability $\sigma_m(\mathbf{m})$ in equation 8.23 becomes⁵

$$\sigma_m(\mathbf{m}) = \quad (8.34)$$

³Unfortunately, many authors define, inconsistently, the maximum likelihood point as the point where the probability density is maximum.

⁴Of course, one could equivalently write $\gamma_k \approx (\mathbf{m}_k - \mathbf{m}_{\text{prior}}) + \mathbf{g}_m^{-1} \mathbf{D}_k^t \mathbf{g}_d(\mathbf{d}_k - \mathbf{d}_{\text{obs}})$, but, numerically, it is usually much better to solve a linear system than to evaluate the inverse of a matrix. This may be important in large-dimensional spaces.

⁵The last multiplicative factor in equation 8.23 is a constant that can be integrated into the constant k .

$$k \exp \left(-\frac{1}{2} \left((\mathbf{m} - \mathbf{m}_{\text{prior}})^t \mathbf{g}_{\mathbf{m}} (\mathbf{m} - \mathbf{m}_{\text{prior}}) + (\mathbf{D} \mathbf{m} - \mathbf{d}_{\text{obs}})^t \mathbf{g}_{\mathbf{d}} (\mathbf{D} \mathbf{m} - \mathbf{d}_{\text{obs}}) \right) \right) .$$

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

$$\sigma_{\mathbf{m}}(\mathbf{m}) = k \exp \left(-\frac{1}{2} (\mathbf{m} - \mathbf{m}')^t \mathbf{g}'_{\mathbf{m}} (\mathbf{m} - \mathbf{m}') \right) , \quad (8.35)$$

this implying that $\sigma_{\mathbf{m}}(\mathbf{m})$ is a Gaussian volumetric probability. The values \mathbf{m}' and $\mathbf{g}'_{\mathbf{m}}$ of the center and the weight matrix (inverse of the covariance matrix), respectively, of the Gaussian representing the a posteriori information in the model space, can be computed using certain matrix identities (see, for instance, Tarantola, 1987, problem 1.19). For the weight matrix, this gives

$$\mathbf{g}'_{\mathbf{m}} = \mathbf{g}_{\mathbf{m}} + \mathbf{D}^t \mathbf{g}_{\mathbf{d}} \mathbf{D} \quad (8.36)$$

and the central point \mathbf{m}' is obtained via

$$\mathbf{g}'_{\mathbf{m}} (\mathbf{m}' - \mathbf{m}_{\text{prior}}) = \mathbf{D}^t \mathbf{g}_{\mathbf{d}} (\mathbf{d}_{\text{obs}} - \mathbf{D} \mathbf{m}_{\text{prior}}) . \quad (8.37)$$

Let us introduce the covariance matrices

$$\mathbf{C}_{\mathbf{m}} = \mathbf{g}_{\mathbf{m}}^{-1} \quad ; \quad \mathbf{C}'_{\mathbf{m}} = \mathbf{g}'_{\mathbf{m}}^{-1} \quad ; \quad \mathbf{C}_{\mathbf{d}} = \mathbf{g}_{\mathbf{d}}^{-1} . \quad (8.38)$$

An equation equivalent to 8.36 is

$$\mathbf{C}'_{\mathbf{m}} = \mathbf{C}_{\mathbf{m}} - \mathbf{C}_{\mathbf{m}} \mathbf{D}^t (\mathbf{D} \mathbf{C}_{\mathbf{m}} \mathbf{D}^t + \mathbf{C}_{\mathbf{d}})^{-1} \mathbf{D} \mathbf{C}_{\mathbf{m}} , \quad (8.39)$$

while an equation equivalent to 8.37 is

$$\mathbf{m}' - \mathbf{m}_{\text{prior}} = \mathbf{C}_{\mathbf{m}} \mathbf{D}^t (\mathbf{D} \mathbf{C}_{\mathbf{m}} \mathbf{D}^t + \mathbf{C}_{\mathbf{d}})^{-1} (\mathbf{d}_{\text{obs}} - \mathbf{D} \mathbf{m}_{\text{prior}}) . \quad (8.40)$$

[End of example.]

Example 8.7 If, in the context of the previous example, we do not have any a priori information on the model parameters, then $\mathbf{C}_{\mathbf{M}} \rightarrow \infty \mathbf{I}$, i.e., $\mathbf{g}_{\mathbf{m}} \rightarrow \mathbf{0}$. In this case,

$$\mathbf{g}'_{\mathbf{m}} = \mathbf{D}^t \mathbf{g}_{\mathbf{d}} \mathbf{D} , \quad (8.41)$$

and equation 8.37 simplifies to

$$\mathbf{m}' = (\mathbf{D}^t \mathbf{g}_{\mathbf{d}} \mathbf{D})^{-1} (\mathbf{D}^t \mathbf{g}_{\mathbf{d}} \mathbf{d}_{\text{obs}}) . \quad (8.42)$$

[End of example.]

Example 8.8 In the context of the previous example, let us explore the very special circumstance where we have the same number of 'data' and 'unknowns', i.e., the case where the matrix \mathbf{D} is a square matrix. Assume that the matrix is regular, so its inverse exists. It is easy to see that equation 8.42 then becomes

$$\mathbf{m}' = \mathbf{D}^{-1} \mathbf{d}_{\text{obs}} . \quad (8.43)$$

We see that in this special case, \mathbf{m}' is just the Cramer solution of the linear equation $\mathbf{d}_{\text{obs}} = \mathbf{D} \mathbf{m}'$. [End of example.]

The formulas in the examples above give expressions that contain analytic parts. What we write as $\mathbf{d} = \mathbf{d}(\mathbf{m})$ may sometimes correspond to an explicit expression; sometimes it may correspond to the solution of an implicit equation⁶. Should $\mathbf{d} = \mathbf{d}(\mathbf{m})$ be an explicit expression, and should the ‘prior probability densities’ $\rho_{\mathbf{m}}(\mathbf{m})$ and $\rho_{\mathbf{d}}(\mathbf{d})$ (or the joint $\rho(\mathbf{m}, \mathbf{d})$) also be given by explicit expressions (like as when we have Gaussian probability densities), then the formulas of this section would give explicit expressions for the posterior volumetric probability $\sigma_{\mathbf{m}}(\mathbf{m})$.

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

Once the volumetric probability $\sigma_{\mathbf{m}}(\mathbf{m})$ has been defined, there are different ways of ‘using’ it.

If the ‘model space’ \mathbf{M} has a small number of dimensions (say between one and four), the values of $\sigma_{\mathbf{m}}(\mathbf{m})$ can be computed at every point of a grid and a graphical representation of $\sigma_{\mathbf{m}}(\mathbf{m})$ 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 $\sigma_{\mathbf{m}}(\mathbf{m})$ are known at all significant points, these estimators can also be computed. This point of view is emphasized in section ???. If the ‘model space’ \mathbf{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 $\sigma_{\mathbf{m}}(\mathbf{m})$. We discuss the application of Monte Carlo methods to inverse problems in 8.3.6. Finally, the optimization techniques are discussed in section 8.3.7.

⁶Practically, it may correspond to the output of some ‘black box’ solving the ‘forward problem’.

8.3 Appendixes

8.3.1 Appendix: Short Bibliographical Review

For 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 over-determined problem, Hadamard (1902, 1932) introduced the notion of “ill-posed problem”, that can be seen as an underdetermined problem.

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.

The first uses of the 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 quite beautiful, its initial predominance over the more ‘brute force’ (but more powerful) Monte Carlo theory was 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).

Interesting contributions to the theory were 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).

The 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’. In an interesting paper, Rietsch (1977) made a nontrivial use of the notion of ‘noninformative’ (homogeneous, in our terminology) a priori distribution for positive parameters.

Jackson (1979) made an explicit introduction of a priori information in the context of linear inverse problems, approach that was generalized by Tarantola and Valette (1982) to nonlinear problems.

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) is easy to read. Finally, Parker (1994) exposes his view of the general theory of linear problems.

From time to time, some authors try to resuscitate the Laplacian ‘least absolute criterion’ (and this is good). Claerbout and Muir (1973), for instance, show that the use of the ℓ_1 -norm can accommodate for erratic data, and Djikpéssé and Tarantola (1999) used the ℓ_1 -norm in a large scale inverse problem, involving seismic waveforms (a seismic reflection experiment).

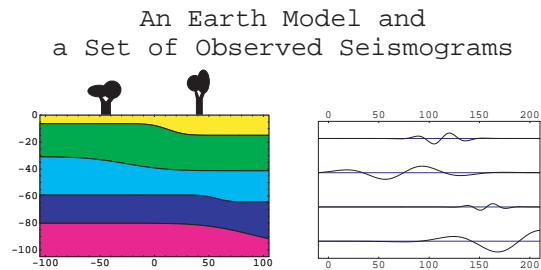
Recently, the interest in Monte Carlo methods, for the solution of Inverse Problems, has been increasing. Mosegaard and Tarantola (1995) proposed a generalization of the Metropolis algorithm for analysis of general inverse problems, introducing explicitly a priori probability

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).

8.3.2 Appendix: 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 8.4 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 8.4: 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 a 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: Thickness 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 always the case if 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 $\mathbf{m} = \{m^1, m^2, \dots, m^M\}$ to describe an Earth model, where M may be a small number (for simple 1D models) or a large number (by the millions or billions for complex 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 loose 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 insensitivity 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 8.4. For our needs, the particular structure of a seismogram set is not interesting, and we will simply represent such a set using the notation $\mathbf{d} = \{d^1, d^2, \dots, d^N\}$, 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 description of the recording instrument, etc.

As mentioned elsewhere [note: where?] there are many reasons for which a ‘theory’ is not an exact functional relationship but, rather, a conditional volumetric probability $\vartheta(\mathbf{d}|\mathbf{m})$. [note: explain this better.] 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 7.2.2. Then, for instance, using a Gaussian model, we may write

$$\vartheta(\mathbf{d}|\mathbf{m}) = k \exp\left(-\frac{1}{2}(\mathbf{d} - \mathbf{f}(\mathbf{m}))^T \mathbf{C}_T^{-1}(\mathbf{d} - \mathbf{f}(\mathbf{m}))\right), \quad (8.44)$$

where the uncertainty 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, like 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 8.5 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 [note: say where], it may always be represented by a probability distribution over the model parameter space, corresponding to a volumetric probability $\rho_m(\mathbf{m})$. 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,

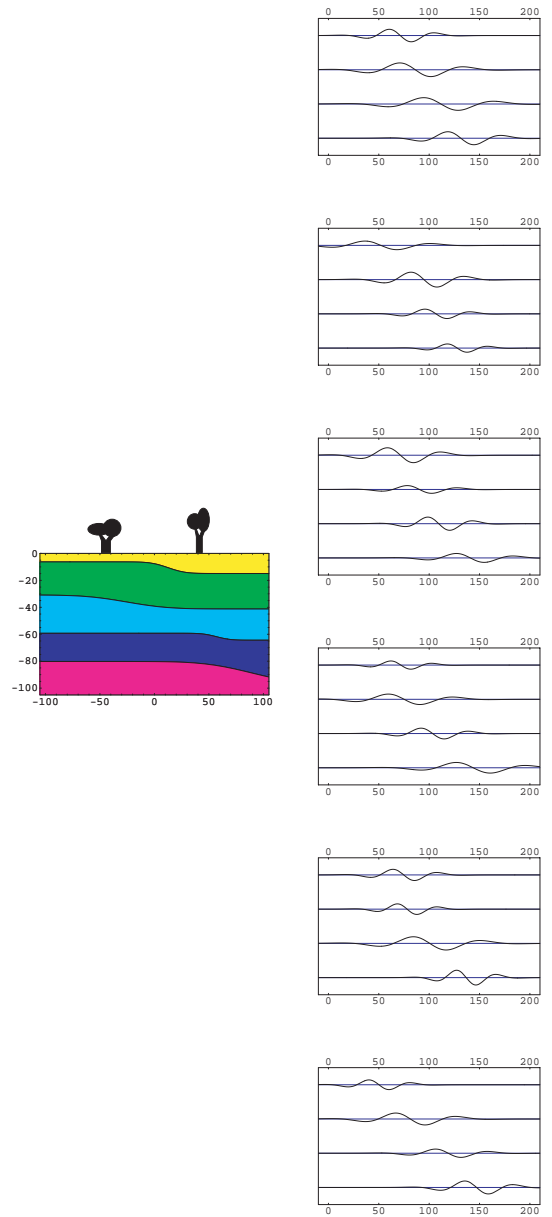
$$\rho_m(\mathbf{m}) = k \exp\left(-\frac{1}{2}(\mathbf{m} - \mathbf{m}_{\text{prior}})^T \mathbf{C}_m^{-1}(\mathbf{m} - \mathbf{m}_{\text{prior}})\right). \quad (8.45)$$

Other probability models (Laplace, Pareto, etc.) may, or course, be used. In more realistic situations, the a 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

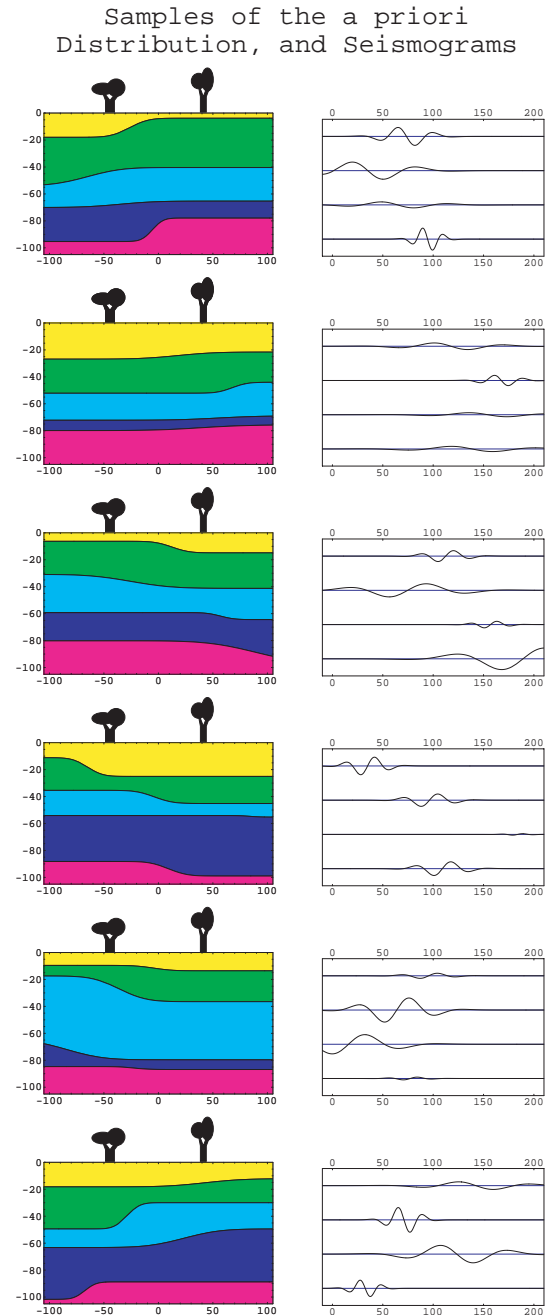
Theoretical Sets of Seismograms
(inside theoretical
uncertainties)

Figure 8.5: 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 sampled, producing the sets of seismograms shown here.



probability $2/3$, a layer with a thickness larger than 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 π and rigidity μ 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 8.6: Samples of the a priori distribution of Earth models, each accompanied by the predicted set of seismograms. A set of rules, some deterministic, 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 right, are as different as the models are different. These 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 8.6 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, even 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 [note: say where], a probability distribution in the data space, corresponding to a volumetric probability $\rho_d(\mathbf{d})$.

The simplest examples of probability distribution in the data space are obtained when using simple probability models. For instance, the assumption of Gaussian uncertainties would give

$$\rho_d(\mathbf{d}) = k \exp\left(-\frac{1}{2}(\mathbf{d} - \mathbf{d}_{\text{obs}})^T \mathbf{C}_d^{-1}(\mathbf{d} - \mathbf{d}_{\text{obs}})\right), \quad (8.46)$$

where \mathbf{d}_{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.

Actual experimental uncertainties are quite difficult to model. [note: develop here this notion, and explain, here or somewhere, what is ‘noise’ in a data set (unmodeled signal)].

[Note: explain that figure 8.7 represents a few samples of ‘data points’ generated according to $\rho_d(\mathbf{d})$.]

Note: explain here that from

$$\begin{aligned} \sigma(\mathbf{d}, \mathbf{m}) &= k \rho(\mathbf{d}, \mathbf{m}) \vartheta(\mathbf{d}, \mathbf{m}) \\ \rho(\mathbf{d}, \mathbf{m}) &= \rho_d(\mathbf{d}) \rho_m(\mathbf{m}) \\ \vartheta(\mathbf{d}, \mathbf{m}) &= \vartheta(\mathbf{d}|\mathbf{m}) \vartheta_m(\mathbf{m}) = k \vartheta(\mathbf{d}|\mathbf{m}) \end{aligned} \quad (8.47)$$

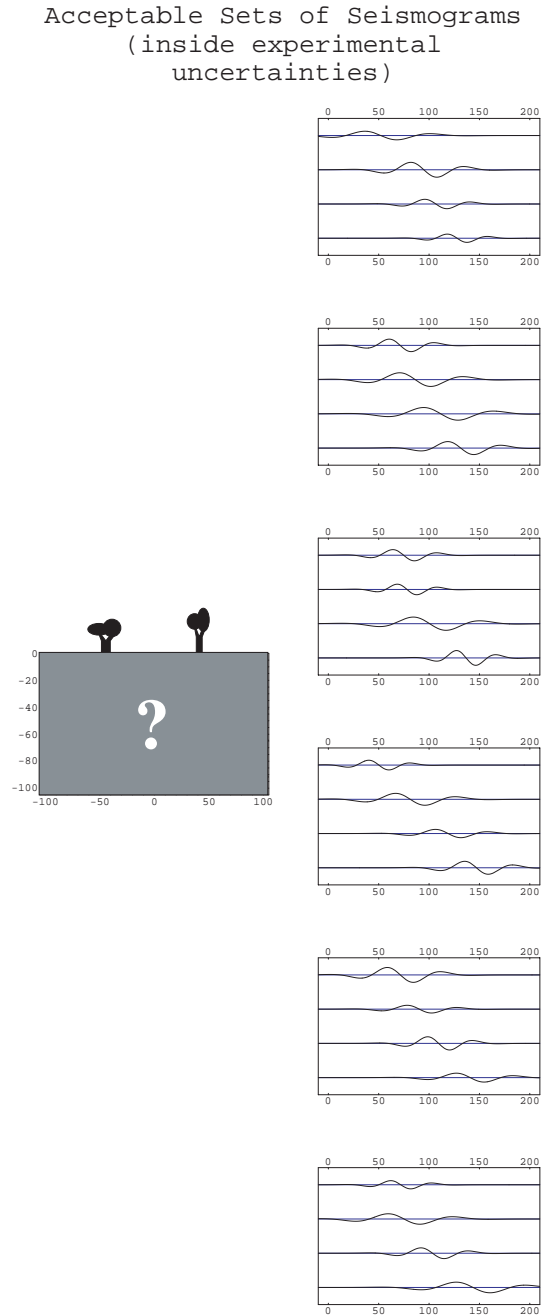
it follows

$$\sigma(\mathbf{d}, \mathbf{m}) = k \rho_d(\mathbf{d}) \vartheta(\mathbf{d}|\mathbf{m}) \rho_m(\mathbf{m}). \quad (8.48)$$

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 8.7: 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 ‘observed set of seismograms’, pseudorandom realizations of the probability distribution in the data space are displayed here.



[Note: explain here that figure 8.8 shows some samples of the a posteriori probability distribution.]

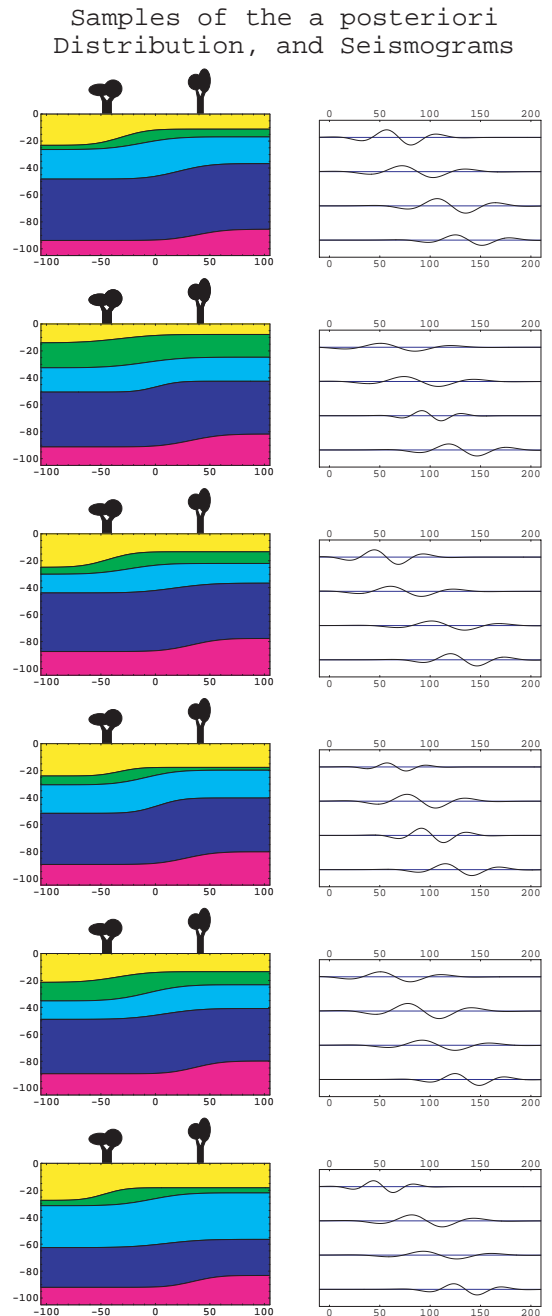


Figure 8.8: Samples of the a posteriori distribution of Earth models, each accompanied by the predicted set of seismograms. Note that, contrary to what happens with the a priori samples, all the models presented here have ‘left dipping interfaces’. The second layer is quite thin. Etc.

[Note: the marginal for \mathbf{m} corresponds to the same ‘movie’, just looking to the models, and disregarding the data sets. Reciprocally, the marginal for $\mathbf{d} \dots$]

‘Things’ can be considerably simplified if uncertainties in the theory can be neglected (i.e., if the ‘theory’ is assumed to be exact):

$$\vartheta(\mathbf{d}|\mathbf{m}) = \delta(\mathbf{d} - \mathbf{f}(\mathbf{m})) . \quad (8.49)$$

Then, the marginal for \mathbf{m} , $\sigma_m(\mathbf{m}) = \int dV_d(d) \sigma(\mathbf{d}, \mathbf{m})$, is using 8.48,

$$\sigma_m(\mathbf{m}) = k \rho_m(\mathbf{m}) \rho_d(\mathbf{f}(\mathbf{m})) . \quad (8.50)$$

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.

[Note: explain that both algorithms require the resolution of the ‘forward problem’.]

[Note: explain that the initial point can not be completely arbitrary.]

[Note: the validity of the algorithm with the conditional probability inside has not been demonstrated.]

[Note: develop these notions.]

8.3.3 Appendix: Probabilistic Estimation of Earthquake Locations

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. This 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).

8.3.3.1 A Priori Information on Model Parameters

The ‘unknowns’ of the problem are the hypocentral coordinates of an Earthquake⁷ $\{X, Z\}$, 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 its origin time. This a priori information is assumed to be represented using the probability density

$$\rho_m(X, Z, T) \quad . \quad (8.51)$$

Because we use Cartesian coordinates and Newtonian time, the homogeneous probability density is just a constant,

$$\mu_m(X, Y, T) = k \quad . \quad (8.52)$$

For consistency, we must assume (rule 4.8) that the limit of $\rho_m(X, Z, T)$ for infinite ‘dispersions’ is $\mu_m(X, Z, T)$.

Example 8.9 *We assume that the a priori probability density for (X, Z) is constant inside the region $0 < X < 60$ km, $0 < Z < 50$ km, and that the (unnormalizable) probability density for T is constant. [End of example.]*

8.3.3.2 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

$$\rho_d(t^1, t^2, t^3, t^4) \quad (8.53)$$

over the ‘data space’. As these are Newtonian times, the associated homogeneous probability density is constant:

$$\mu_d(t^1, t^2, t^3, t^4) = k \quad . \quad (8.54)$$

For consistency, we must assume (rule 4.8) that the limit of $\rho_d(t^1, t^2, t^3, t^4)$ for infinite ‘dispersions’ is $\mu_d(t^1, t^2, t^3, t^4)$.

⁷To simplify, here, we consider a 2D flat model of the Earth, and use Cartesian coordinates.

Example 8.10 *Assuming Gaussian, independent uncertainties, we have*

$$\begin{aligned} \rho_d(t^1, t^2, t^3, t^4) &= k \exp\left(-\frac{1}{2} \frac{(t^1 - t_{\text{obs}}^1)^2}{\sigma_1^2}\right) \exp\left(-\frac{1}{2} \frac{(t^2 - t_{\text{obs}}^2)^2}{\sigma_2^2}\right) \\ &\times \exp\left(-\frac{1}{2} \frac{(t^3 - t_{\text{obs}}^3)^2}{\sigma_3^2}\right) \exp\left(-\frac{1}{2} \frac{(t^4 - t_{\text{obs}}^4)^2}{\sigma_4^2}\right) . \end{aligned} \quad (8.55)$$

[End of example.]

8.3.3.3 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 :

$$t^i = f^i(X, Z, T) . \quad (8.56)$$

Example 8.11 *Assuming that the velocity of the medium is constant, equal to v ,*

$$t_{\text{cal}}^1 = T + \frac{\sqrt{(X - x^i)^2 + (Z - z^i)^2}}{v} . \quad (8.57)$$

8.3.3.4 Solution of the Inverse Problem

Note: explain here that ‘putting all this together’,

$$\sigma_m(X, Z, T) = k \rho_m(X, Z, T) \rho_d(t^1, t^2, t^3, t^4) \Big|_{t^i=f^i(X, Z, T)} . \quad (8.58)$$

8.3.3.5 Numerical Implementation

To show how simple is to implement an estimation of the hypocentral coordinates using the solution given by equation 8.58, 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 suppressed in what 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 8.10. Explicitly:

```
t10BS = 30.3;
s1 = 0.1;
t20BS = 29.4;
s2 = 0.2;
t30BS = 28.6;
s3 = 0.1;
t40BS = 28.3;
s4 = 0.1;
```

```
rho1[t1_] := Exp[ - (1/2) (t1 - t10BS)^2/s1^2 ]
rho2[t2_] := Exp[ - (1/2) (t2 - t20BS)^2/s2^2 ]
rho3[t3_] := Exp[ - (1/2) (t3 - t30BS)^2/s3^2 ]
rho4[t4_] := Exp[ - (1/2) (t4 - t40BS)^2/s4^2 ]
```

```
rho[t1_,t2_,t3_,t4_] := rho1[t1] rho2[t2] rho3[t3] rho4[t4]
```

Although an arbitrarily complex velocity model could be considered here, let us take, for solving the forward problem, the simple model in example 8.11:

```
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 8.58:

```
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 8.9, 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 probability 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 functions⁸. 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]
```

⁸Typing `sigmaXZ[X,Z]` presents the result.

The result is represented in figure 8.9 (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 σ , 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 σ) 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 everything 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} ]
```

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 $0 < X < 60$ km, $0 < Z < 50$ km but limiting the integration domain to that area (see example 8.9). 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 8.10. The software was not very stable in producing the results of the numerical integration.

Figure 8.9: The probability density for the location of the hypocenter. Its asymmetric shape is quite typical, as seismic observatories tend to be asymmetrically placed.

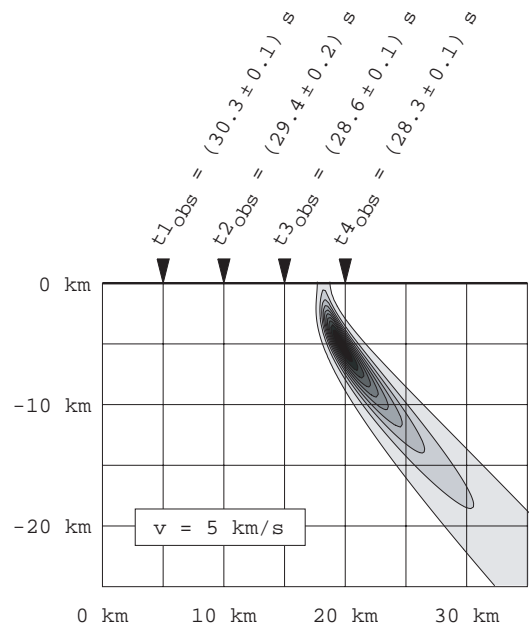
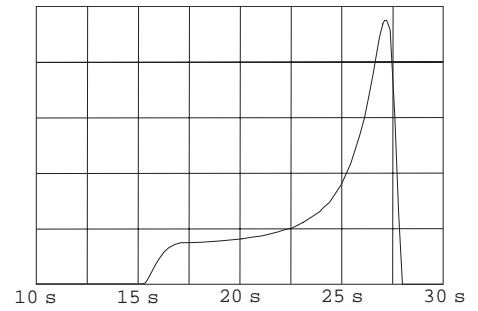


Figure 8.10: The marginal probability density for the origin time. The asymmetry seen in the probability density in figure 8.9, where the decay of probability is slow downwards, translates here in significant probabilities for early times. The sharp decay of the probability density for $t < 17s$ does not come from the values of the arrival times, but from the a priori information that the hypocenters must be above the depth $Z = -50$ km .



8.3.3.6 An Example of Bimodal Probability Density for an Arrival Time.

As an exercise, the reader could reformulate the problem replacing the assumption of Gaussian uncertainties in the arrival times by multimodal probability densities. For instance, figure 5.6 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

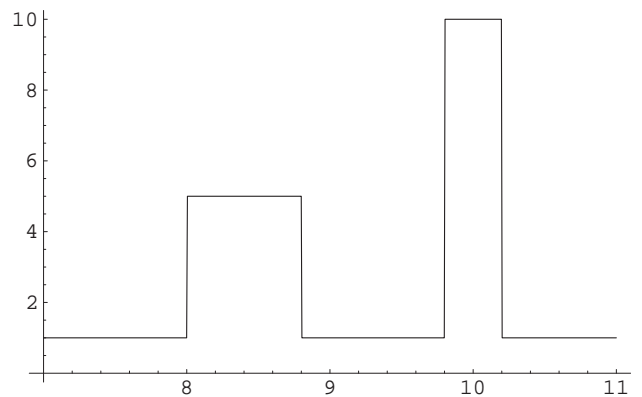
```
rho[t_] := (If[8.0<t<8.8,5,1] If[9.8<t<10.2,10,1])
```

defines a probability density that, when plotted using the command

```
Plot[ rho[t] ,{t,7,11} ]
```

produces the result displayed in figure 8.11.

Figure 8.11: In figure 5.6 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.



8.3.4 Appendix: Functional Inverse Problems

8.3.4.1 Introduction

As mentioned in section 2.2, 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 function⁹. 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 functions¹⁰, 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 (char-

⁹The Erf function, or error function, is the primitive of a Gaussian. It is a simple example of a ‘sigmoidal’ function.

¹⁰Grossly speaking, a function $f(x)$ belongs to L_2 if $\|f\| = (\int dx f(x)^2)^{1/2}$ is finite. A function $f(x)$ belongs to L_p if $\|f\| = (\int dx |f(x)|^p)^{1/p}$ is finite. The limit for $p \rightarrow \infty$ corresponds to the l_∞ space.

acterized 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’.

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).

8.3.4.2 The Functional Spaces Under Investigation

A seismologist may consider a (three-component) seismogram

$$\mathbf{u} = \{ u^i(t) \ ; \ i = 1, 2, 3 \ ; \ t_0 \leq t \leq t_1 \} \quad , \quad (8.59)$$

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

$$\| \mathbf{u} \|^2 = \int_{t_0}^{t_1} dt u_i(t) u^i(t) \quad , \quad (8.60)$$

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

$$\boldsymbol{\rho} = \{ \rho(z) \ ; \ z_0 \leq z \leq z_1 \} \quad , \quad (8.61)$$

but it will generally not make sense to define

$$\| \boldsymbol{\rho} \|^2 = \int_{z_0}^{z_1} dz \rho(z)^2 \quad (\text{bad definition}) \quad . \quad (8.62)$$

For the resistivity-versus-depth profile is equivalent to the conductivity-versus-depth profile

$$\boldsymbol{\sigma} = \{ \sigma(z) \ ; \ z_0 \leq z \leq z_1 \} \quad , \quad (8.63)$$

where, for any z , $\rho(z) \sigma(z) = 1$, and the definition of the norm

$$\| \boldsymbol{\sigma} \|^2 = \int_{z_0}^{z_1} dz \sigma(z)^2 \quad (\text{bad definition}) \quad , \quad (8.64)$$

would not be consistent with that of the norm $\| \boldsymbol{\rho} \|$ (we do not have, in general, any reason to assume that $\sigma(z)$ could 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 ρ_0 and σ_0 are arbitrary constants) allow the only sensible definition of norm

$$\| \mathbf{r} \|^2 = \| \mathbf{s} \|^2 = \int_{z_0}^{z_1} dz r(z)^2 = \int_{z_0}^{z_1} dz s(z)^2 \quad (\text{good definition}) \quad , \quad (8.65)$$

or, in terms of ρ and σ ,

$$\| \boldsymbol{\rho} \|^2 = \| \boldsymbol{\sigma} \|^2 = \int_{z_0}^{z_1} dz \left(\log \frac{\rho(z)}{\rho_0} \right)^2 = \int_{z_0}^{z_1} dz \left(\log \frac{\sigma(z)}{\sigma_0} \right)^2 \quad (\text{good definition}) \quad , \quad (8.66)$$

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.

8.3.4.3 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

$$\langle \boldsymbol{\sigma}, \boldsymbol{\varepsilon} \rangle = \int dt \int dV(\mathbf{x}) \sigma_{ij}(\mathbf{x}, t) \varepsilon^{ij}(\mathbf{x}, t), \quad (8.67)$$

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 called ‘the dual space’, but this is just a matter of convention.

The product 8.67 is one example of *duality product*, where one element of the primal space and one element of the dual space are ‘multiplied’ 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 8.67; and that this association defines a linear application. But this rough definition of duality doesn’t help readers to understand the actual mathematical structure.

8.3.4.4 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 number¹¹ 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

$$\widehat{\mathbf{f}} = \mathbf{C}^{-1} \mathbf{f}. \quad (8.68)$$

¹¹It is usually a real number, but it may have physical dimensions.

The scalar product of two elements \mathbf{f}_1 and \mathbf{f}_2 of \mathcal{F} is then defined as

$$(\mathbf{f}_1, \mathbf{f}_2) = \langle \widehat{\mathbf{f}}_1, \mathbf{f}_2 \rangle = \langle \mathbf{C}^{-1} \mathbf{f}_1, \mathbf{f}_2 \rangle \quad (8.69)$$

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 $\widehat{\mathcal{F}} = \mathbf{C}^{-1} \mathcal{F}$ or, equivalently, $\mathcal{F} = \mathbf{C} \widehat{\mathcal{F}}$. The explicit expression of the equation

$$\mathbf{f} = \mathbf{C} \widehat{\mathbf{f}} \quad (8.70)$$

is

$$f(x) = \int dx' C(x, x') \widehat{f}(x'). \quad (8.71)$$

Let us denote \mathbf{W} the inverse of the covariance operator,

$$\mathbf{W} = \mathbf{C}^{-1}, \quad (8.72)$$

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*, satisfies

$$\int dx' C(x, x') W(x', x'') = \int dx' W(x, x') C(x', x'') = \delta(x - x''), \quad (8.73)$$

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).

Equations 8.70–8.71 can equivalently be written

$$\widehat{\mathbf{f}} = \mathbf{W} \mathbf{f} \quad (8.74)$$

and

$$\widehat{f}(x) = \int dx' W(x, x') f(x'). \quad (8.75)$$

If the duality product between $\widehat{\mathbf{f}}_1$ and \mathbf{f}_2 is written

$$\langle \widehat{\mathbf{f}}_1, \mathbf{f}_2 \rangle = \int dx \widehat{f}_1(x) f_2(x), \quad (8.76)$$

the scalar product, as defined by equation 8.69, becomes

$$(\mathbf{f}_1, \mathbf{f}_2) = \langle \widehat{\mathbf{f}}_1, \mathbf{f}_2 \rangle = \langle \mathbf{C}^{-1} \mathbf{f}_1, \mathbf{f}_2 \rangle = \langle \mathbf{W} \mathbf{f}_1, \mathbf{f}_2 \rangle$$

$$\begin{aligned}
&= \int dx \left(\int dx' W(x, x') f_1(x') \right) f_2(x) \\
&= \int dx \int dx' f_1(x) W(x, x') f_2(x') .
\end{aligned} \tag{8.77}$$

The norm of \mathbf{f} , denoted $\|\mathbf{f}\|$ and defined as

$$\|\mathbf{f}\|^2 = (\mathbf{f}, \mathbf{f}) , \tag{8.78}$$

is expressed, in this example, as

$$\|\mathbf{f}\|^2 = \int dx \int dx' f(x) W(x, x') f(x') . \tag{8.79}$$

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 $\hat{f}(x)$ is a random realization of a Gaussian white noise with zero mean, then, the function $f(x)$ defined by equation 8.71 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, $\hat{\mathcal{F}}$, is the space of all the realizations of a Gaussian white noise.

Example 8.12 Consider the covariance operator \mathbf{C} , with covariance function $C(x, x')$,

$$\mathbf{f} = \mathbf{C}\hat{\mathbf{f}} \quad \Longleftrightarrow \quad f(x) = \int_{-\infty}^{+\infty} dx' C(x, x') \hat{f}(x') , \tag{8.80}$$

in the special case where the covariance function is the exponential function,

$$C(x, x') = \sigma^2 \exp\left(-\frac{|x - x'|}{X}\right) , \tag{8.81}$$

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

$$\hat{\mathbf{f}} = \mathbf{C}^{-1}\mathbf{f} \quad \Longleftrightarrow \quad \hat{f}(t) = \frac{1}{2\sigma^2} \left(\frac{1}{X} f(x) - X \ddot{f}(x) \right) , \tag{8.82}$$

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 $\hat{f}(x)$ 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:

$$\|\mathbf{f}\|^2 = \langle \hat{\mathbf{f}}, \mathbf{f} \rangle = \frac{1}{2\sigma^2} \left(\frac{1}{X} \int_{-\infty}^{+\infty} dx f^2(x) + X \int_{-\infty}^{+\infty} dx \dot{f}^2(x) \right) . \tag{8.83}$$

This is the usual norm in the so-called Sobolev space H^1 . [End of example.]

8.3.4.5 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

$$\mathbf{G} : \mathcal{E} \rightarrow \mathcal{F} . \quad (8.84)$$

If $\mathbf{e} \in \mathcal{E}$ and $\mathbf{f} \in \mathcal{F}$, then we write

$$\mathbf{f} = \mathbf{G} \mathbf{e} . \quad (8.85)$$

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

$$\mathbf{G}^T : \widehat{\mathcal{F}} \rightarrow \widehat{\mathcal{E}} \quad (8.86)$$

$$\forall \mathbf{e} \in \mathcal{E} \quad ; \quad \forall \widehat{\mathbf{f}} \in \widehat{\mathcal{F}} \quad : \quad \langle \widehat{\mathbf{f}}, \mathbf{G} \mathbf{e} \rangle_{\mathcal{F}} = \langle \mathbf{G}^T \widehat{\mathbf{f}}, \mathbf{e} \rangle_{\mathcal{E}} . \quad (8.87)$$

Example 8.13 The Transposed of a Matrix. *Let us consider a discrete situation where*

$$\mathbf{f} = \mathbf{G} \mathbf{e} \quad \Longleftrightarrow \quad f_i = \sum_{\alpha} G_{i\alpha} e_{\alpha} . \quad (8.88)$$

In this circumstance, the duality products in each space will read

$$\langle \widehat{\mathbf{f}}, \mathbf{f} \rangle_{\mathcal{F}} = \sum_i \widehat{f}_i f_i \quad ; \quad \langle \widehat{\mathbf{e}}, \mathbf{e} \rangle_{\mathcal{E}} = \sum_{\alpha} \widehat{e}_{\alpha} e_{\alpha} . \quad (8.89)$$

The linear operator \mathbf{H} is the transposed of \mathbf{G} if for any $\widehat{\mathbf{f}}$ and for any \mathbf{e} (equation 8.87),

$$\langle \widehat{\mathbf{f}}, \mathbf{G} \mathbf{e} \rangle_{\mathcal{F}} = \langle \mathbf{H} \widehat{\mathbf{f}}, \mathbf{e} \rangle_{\mathcal{E}} , \quad (8.90)$$

i.e., if

$$\sum_i \widehat{f}_i (\mathbf{G} \mathbf{e})_i = \sum_{\alpha} (\mathbf{H} \widehat{\mathbf{f}})_{\alpha} e_{\alpha} \quad (8.91)$$

or, explicitly,

$$\sum_i \widehat{f}_i \left(\sum_{\alpha} G_{i\alpha} e_{\alpha} \right) = \sum_{\alpha} \left(\sum_i H_{\alpha i} \widehat{f}_i \right) e_{\alpha} . \quad (8.92)$$

The condition can be written

$$\sum_i \sum_{\alpha} \widehat{f}_i G_{i\alpha} e_{\alpha} = \sum_i \sum_{\alpha} \widehat{f}_i H_{\alpha i} e_{\alpha} , \quad (8.93)$$

and it is clear that this true for any $\hat{\mathbf{f}}$ and for any \mathbf{e} iff

$$H_{\alpha i} = G_{i\alpha} , \quad (8.94)$$

i.e., if the matrix representing \mathbf{H} is the transposed (in the elementary matricial sense) of the matrix representing \mathbf{G} :

$$\mathbf{H} = \mathbf{G}^T . \quad (8.95)$$

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 8.14 The Transposed of the Derivative Operator. Let us consider a situation where

$$\mathbf{v} = \mathbf{D} \mathbf{x} \quad \Longleftrightarrow \quad v(t) = \frac{dx}{dt}(t) , \quad (8.96)$$

i.e., where the linear operator \mathbf{D} is the derivative operator. In this circumstance, the duality products in each space will typically read

$$\langle \hat{\mathbf{v}} , \mathbf{v} \rangle_{\mathcal{V}} = \int_{t_1}^{t_2} dt \hat{v}(t) v(t) \quad ; \quad \langle \hat{\mathbf{x}} , \mathbf{x} \rangle_{\mathcal{X}} = \int_{t_1}^{t_2} dt \hat{x}(t) x(t) . \quad (8.97)$$

If the linear operator \mathbf{D}^T has to be the transposed of \mathbf{D} , for any $\hat{\mathbf{v}}$ and for any \mathbf{x} we must have (equation 8.87)

$$\langle \hat{\mathbf{v}} , \mathbf{D} \mathbf{x} \rangle_{\mathcal{V}} = \langle \mathbf{D}^T \hat{\mathbf{v}} , \mathbf{x} \rangle_{\mathcal{X}} . \quad (8.98)$$

[End of example.]

Let us demonstrate that the derivative operator is an antisymmetric operator i.e., that

$$\mathbf{D}^T = -\mathbf{D} . \quad (8.99)$$

To demonstrate this, we will need to make a restrictive condition, interesting to analyze.

Using 8.99, equation 8.98 writes

$$\int_{t_1}^{t_2} dt \hat{v}(t) (\mathbf{D} \mathbf{x})(t) = - \int_{t_1}^{t_2} dt (\mathbf{D} \hat{\mathbf{v}})(t) x(t) \quad (8.100)$$

i.e.,

$$\int_{t_1}^{t_2} dt \hat{v}(t) \frac{dx}{dt}(t) + \int_{t_1}^{t_2} dt \frac{d\hat{v}}{dt}(t) x(t) = 0 . \quad (8.101)$$

We have to check if this equation holds for any $x(t)$ and any $v(t)$.

The condition is equivalent to

$$\int_{t_1}^{t_2} dt \left(\hat{v}(t) \frac{dx}{dt}(t) + \frac{d\hat{v}}{dt}(t) x(t) \right) = 0 , \quad (8.102)$$

i.e., to

$$\int_{t_1}^{t_2} dt \frac{d}{dt} (\widehat{v}(t) x(t)) = 0 , \quad (8.103)$$

or, using the elementary properties of the integral, to

$$\widehat{v}(t_2) x(t_2) + \widehat{v}(t_1) x(t_1) = 0 . \quad (8.104)$$

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 8.104 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:

$$x(t_1) = 0 \quad ; \quad \widehat{v}(t_2) = 0 . \quad (8.105)$$

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)

$$\gamma(t) = \frac{dx^2}{dt^2}(t) . \quad (8.106)$$

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 boundary condition

$$\widehat{\gamma}(t_2) \frac{dx}{dt}(t_2) - \frac{d\widehat{\gamma}}{dt}(t_2) x(t_2) = \widehat{\gamma}(t_1) \frac{dx}{dt}(t_1) - \frac{d\widehat{\gamma}}{dt}(t_1) x(t_1) . \quad (8.107)$$

A typical example where this condition is satisfied is when we have

$$x(t_1) = 0 \quad ; \quad \frac{dx}{dt}(t_1) = 0 \quad ; \quad \widehat{\gamma}(t_2) = 0 \quad ; \quad \frac{d\widehat{\gamma}}{dt}(t_2) = 0 , \quad (8.108)$$

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 8.3.4.7 below.

As an exercise, the reader may try to understand why the quite obvious property

$$\left(\frac{\partial}{\partial x^i} \right)^T = - \left(\frac{\partial}{\partial x^i} \right) \quad (8.109)$$

corresponds, in fact, to the properties

$$\mathbf{grad}^T = -\mathbf{div} \quad ; \quad \mathbf{div}^T = -\mathbf{grad} \quad (8.110)$$

(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 demonstrate that the operator representing the acoustic wave equation is symmetric. Starting from¹²

$$\mathbf{L} = \frac{1}{\kappa(\mathbf{x})} \frac{\partial^2}{\partial t^2} - \operatorname{div} \frac{1}{\rho(\mathbf{x})} \mathbf{grad} , \quad (8.111)$$

we have

$$\begin{aligned} \mathbf{L}^T &= \left(\frac{1}{\kappa(\mathbf{x})} \frac{\partial^2}{\partial t^2} - \operatorname{div} \frac{1}{\rho(\mathbf{x})} \mathbf{grad} \right)^T \\ &= \left(\frac{1}{\kappa(\mathbf{x})} \frac{\partial^2}{\partial t^2} \right)^T - \left(\operatorname{div} \frac{1}{\rho(\mathbf{x})} \mathbf{grad} \right)^T . \end{aligned} \quad (8.112)$$

Using the property $(\mathbf{A} \mathbf{B})^T = \mathbf{B}^T \mathbf{A}^T$, we arrive at

$$\mathbf{L}^T = \left(\frac{\partial^2}{\partial t^2} \right)^T \left(\frac{1}{\kappa(\mathbf{x})} \right)^T - (\mathbf{grad})^T \left(\frac{1}{\rho(\mathbf{x})} \right)^T (\operatorname{div})^T . \quad (8.113)$$

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) $\mathbf{grad}^T = -\operatorname{div}$ and $\operatorname{div}^T = -\mathbf{grad}$. We then have

$$\mathbf{L}^T = \frac{\partial^2}{\partial t^2} \frac{1}{\kappa(\mathbf{x})} - \operatorname{div} \frac{1}{\rho(\mathbf{x})} \mathbf{grad} , \quad (8.114)$$

and, as the incompressibility κ is assumed to be independent on time,

$$\mathbf{L}^T = \frac{1}{\kappa(\mathbf{x})} \frac{\partial^2}{\partial t^2} - \operatorname{div} \frac{1}{\rho(\mathbf{x})} \mathbf{grad} = \mathbf{L} , \quad (8.115)$$

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) demonstrates 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 ordinary attenuation: there is no difficulty in the implementation of \mathbf{L}^T .

Example 8.15 The Kernel of the Transposed Operator *If the explicit expression of the equation*

$$\mathbf{f} = \mathbf{G} \mathbf{e} \quad (8.116)$$

¹²Here, 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

$$f(t) = \int dt G(t, x) e(t) , \quad (8.117)$$

where $G(t, x)$ is an ordinary function¹³, 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 8.116 belong. An equation like

$$\widehat{\mathbf{e}} = \mathbf{G}^T \widehat{\mathbf{f}} \quad (8.118)$$

will correspond, explicitly, to

$$\widehat{e}(t) = \int dx G^T(x, t) \widehat{f}(t) . \quad (8.119)$$

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

$$G^T(x, t) = G(t, x) . \quad (8.120)$$

We see that the kernels of \mathbf{G} and of \mathbf{G}^T are, in fact, identical, via a simple ‘transposition’ of the variables.

8.3.4.6 The Adjoint Operator

Let \mathbf{G} be a linear operator mapping an space \mathcal{E} into an space \mathcal{F} :

$$\mathbf{G} : \mathcal{E} \rightarrow \mathcal{F} . \quad (8.121)$$

If $\mathbf{e} \in \mathcal{E}$ and $\mathbf{f} \in \mathcal{F}$, then we write

$$\mathbf{f} = \mathbf{G} \mathbf{e} . \quad (8.122)$$

Assume that both, \mathcal{E} and \mathcal{F} are furnished with an scalar product each (see section 8.3.4.4), 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

$$\mathbf{G}^* : \mathcal{F} \rightarrow \mathcal{E} \quad (8.123)$$

$$\forall \mathbf{e} \in \mathcal{E} ; \quad \forall \mathbf{f} \in \mathcal{F} : \quad (\mathbf{f}, \mathbf{G} \mathbf{e})_{\mathcal{F}} = (\mathbf{G}^* \mathbf{f}, \mathbf{e})_{\mathcal{E}} . \quad (8.124)$$

¹³If $G(t, x)$ is a distribution (like the derivative of a Dirac’s delta) then equation 8.116 may be a disguised expression for a differential operator.

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

$$\begin{aligned} (\mathbf{e}_1, \mathbf{e}_2)_{\mathcal{E}} &= \langle \mathbf{C}_{\mathcal{E}}^{-1} \mathbf{e}_2, \mathbf{e}_1 \rangle_{\mathcal{E}} \\ (\mathbf{f}_1, \mathbf{f}_2)_{\mathcal{F}} &= \langle \mathbf{C}_{\mathcal{F}}^{-1} \mathbf{f}_2, \mathbf{f}_1 \rangle_{\mathcal{F}} . \end{aligned} \quad (8.125)$$

Then, equation 8.124 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 $\widehat{\mathbf{f}} = \mathbf{C}_{\mathcal{F}}^{-1} \mathbf{f}$,

$$\langle \widehat{\mathbf{f}}, \mathbf{G} \mathbf{e} \rangle_{\mathcal{F}} = \langle \mathbf{C}_{\mathcal{E}}^{-1} \mathbf{G}^* \mathbf{C}_{\mathcal{F}} \widehat{\mathbf{f}}, \mathbf{e} \rangle_{\mathcal{E}} . \quad (8.126)$$

The comparison with equation 8.124 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

$$\mathbf{G}^* = \mathbf{C}_{\mathcal{E}} \mathbf{G}^T \mathbf{C}_{\mathcal{F}}^{-1} . \quad (8.127)$$

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 defined 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 8.127 is to be interpreted.

8.3.4.7 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’

$$\frac{1}{\kappa(\mathbf{x})} \frac{\partial^2 p}{\partial t^2}(\mathbf{x}, t) - \operatorname{div} \left(\frac{1}{\rho(\mathbf{x})} \mathbf{grad} p(\mathbf{x}, t) \right) = S(\mathbf{x}, t) . \quad (8.128)$$

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

$$\frac{1}{\kappa} \frac{\partial^2 p}{\partial t^2} - \operatorname{div} \left(\frac{1}{\rho} \mathbf{grad} p \right) = S . \quad (8.129)$$

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,

$$\mathbf{L} \mathbf{p} = \mathbf{S} , \quad (8.130)$$

where \mathbf{L} is the second order differential operator defined through equation 8.129. 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 ,

$$\begin{aligned} p(\mathbf{x}, t_1) &= 0 & ; & & \mathbf{x} \in V \\ \dot{p}(\mathbf{x}, t_1) &= 0 & ; & & \mathbf{x} \in V \\ p(\mathbf{x}, t) &= 0 & ; & & \mathbf{x} \in S ; t \in (t_1, t_2) \end{aligned} \quad (8.131)$$

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} :

$$\mathbf{G} = \mathbf{L}^{-1} \quad (8.132)$$

We can then write

$$\mathbf{L}\mathbf{p} = \mathbf{S} \quad \Longleftrightarrow \quad \mathbf{p} = \mathbf{G}\mathbf{S} \quad (8.133)$$

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

$$p(\mathbf{x}, t) = \int_V dV(\mathbf{x}') \int_{t_1}^{t_2} dt' G(\mathbf{x}, t; \mathbf{x}', t') S(\mathbf{x}', t') \quad (8.134)$$

It is easy to demonstrate¹⁴ 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 8.14 above).

Using the method outlined in section 8.14, the boundary conditions dual to those in equations 8.131 are

$$\begin{aligned} p(\mathbf{x}, t_2) &= 0 & ; & & \mathbf{x} \in V \\ \dot{p}(\mathbf{x}, t_2) &= 0 & ; & & \mathbf{x} \in V \\ p(\mathbf{x}, t) &= 0 & ; & & \mathbf{x} \in S ; t \in (t_1, t_2) \end{aligned} \quad (8.135)$$

i.e., we have *final* conditions of rest instead of initial conditions of rest (and the same surface condition). We have to understand that while the equation $\mathbf{L}\mathbf{p} = \mathbf{S}$ is associated to the boundary conditions 8.131, equations like

$$\mathbf{L}^T \widehat{\mathbf{p}} = \widehat{\mathbf{S}} \quad ; \quad \widehat{\mathbf{p}} = \mathbf{G}^T \widehat{\mathbf{S}} \quad (8.136)$$

are associated to the dual boundary conditions 8.135 (the hats here mean that the transpose operator operates in the dual spaces (see section 8.3.4.3). This being understood, we can write $\mathbf{L}^T = \mathbf{L}$ and $\mathbf{G}^T = \mathbf{G}$, and rewrite equations 8.136 as

$$\mathbf{L} \widehat{\mathbf{p}} = \widehat{\mathbf{S}} \quad ; \quad \widehat{\mathbf{p}} = \mathbf{G} \widehat{\mathbf{S}} \quad (8.137)$$

¹⁴This comes from the property that the derivative operator is antisymmetric, (so that the second derivative is a symmetric operator) and from the properties $\mathbf{grad}^T = -\mathbf{div}$ and $\mathbf{div}^T = -\mathbf{grad}$, mentioned in section 8.14.

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

$$\widehat{p}(\mathbf{x}, t) = \int dV(\mathbf{x}') \int_{t_1}^{t_2} dt' G^T(\mathbf{x}, t; \mathbf{x}', t') \widehat{S}(\mathbf{x}', t') . \quad (8.138)$$

Some text is missing here. Some text is missing here. Some text is missing here. Some text is missing here. Some text is missing here. Some text is missing here. Some text is missing here. Some text is missing here. Some text is missing here. Some text is missing here. Some text is missing here.

8.3.4.8 Born Approximation for the Acoustic Wave Equation

Let us start from equation 8.129, using the same notations:

$$\frac{1}{\kappa} \frac{\partial^2 p}{\partial t^2} - \operatorname{div} \left(\frac{1}{\rho} \mathbf{grad} p \right) = S . \quad (8.139)$$

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, κ and ρ 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 $\{\kappa, \rho\} \rightarrow \mathbf{p} = \psi(\kappa, \rho)$, i.e., an application that associates to each medium $\{\kappa, \rho\}$ 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 κ_0 and ρ_0 , i.e., $\mathbf{p}_0 = \psi(\kappa_0, \rho_0)$, and let \mathbf{p} be the pressure field propagating in the medium defined by κ and ρ , i.e., $\mathbf{p} = \psi(\kappa, \rho)$. Clearly, if κ and ρ are close (in a sense to be defined) to κ_0 and ρ_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 κ and ρ could be perturbed, but I simplify the discussion here by considering only perturbations in the incompressibility κ . The reader may easily obtain the general case.

The pressure inside an elastic fluid medium is (note: check if this sign is consistent with the sign given to the stress tensor elsewhere in the book)

$$p = -\frac{1}{3} \sigma_k^k . \quad (8.140)$$

So defined, the pressure 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.

Note: check what follows. Perhaps it is better to assume that the pressure P is positive quantity, and to define

$$p = P_0 \log \frac{P}{P_0} , \quad (8.141)$$

where P_0 is the ‘ambient pressure’. For small pressure perturbations, we have

$$p = P_0 \log \left(1 + \frac{(P - P_0)}{P_0} \right) \approx P - P_0 . \quad (8.142)$$

The incompressibility 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 $\kappa - \kappa_0$, but we have seen that this is not the right way, as the resulting approximation will depend on the fact that we are using incompressibility $\kappa(\mathbf{x})$ instead of compressibility $\gamma(\mathbf{x}) = 1/\kappa(\mathbf{x})$.

At this point we may introduce the logarithmic parameters, and proceed trivially (note: explain why this is important). The logarithmic incompressibilities for the reference medium and for the perturbed medium are

$$\kappa_0^* = \log \frac{\kappa_0}{K} \quad ; \quad \kappa^* = \log \frac{\kappa}{K} \quad , \quad (8.143)$$

where K and R are arbitrary constants (having the right physical dimension). Reciprocally,

$$\kappa_0 = K \exp \kappa_0^* \quad ; \quad \kappa = K \exp \kappa^* \quad . \quad (8.144)$$

In particular, we have

$$\kappa = \kappa_0 \exp(\delta\kappa^*) \quad , \quad (8.145)$$

where

$$\delta\kappa^* = \kappa^* - \kappa_0^* = \log \frac{\kappa}{\kappa_0} \quad . \quad (8.146)$$

Note that we have here a perturbation $\delta\kappa^*$ of a logarithmic (Cartesian) quantity, not of the positive (Jeffreys) one. We also write

$$p = p_0 + \delta p \quad . \quad (8.147)$$

The reference solution satisfies

$$\frac{1}{\kappa_0} \frac{\partial^2 p_0}{\partial t^2} - \operatorname{div} \left(\frac{1}{\rho_0} \mathbf{grad} p_0 \right) = S \quad , \quad (8.148)$$

while the perturbed solution satisfies

$$\frac{1}{\kappa} \frac{\partial^2 p}{\partial t^2} - \operatorname{div} \left(\frac{1}{\rho_0} \mathbf{grad} p \right) = S \quad . \quad (8.149)$$

In this equation, κ can be replaced by the expression 8.145, and p by the expression 8.147. Using then the first order approximation $\exp(-\delta\kappa^*) = 1 - \delta\kappa^*$ leads to

$$\left(\frac{1}{\kappa_0} - \frac{\delta\kappa^*}{\kappa_0} \right) \left(\frac{\partial^2 p_0}{\partial t^2} + \frac{\partial^2 \delta p}{\partial t^2} \right) - \operatorname{div} \left(\frac{1}{\rho_0} (\mathbf{grad} p_0 + \mathbf{grad} \delta p) \right) = S \quad . \quad (8.150)$$

Some of the terms in this equation correspond to the terms in the reference equation 8.148, and can be simplified. Keeping only first order terms then leads to

$$\frac{1}{\kappa_0} \frac{\partial^2 \delta p}{\partial t^2} - \operatorname{div} \left(\frac{1}{\rho_0} \mathbf{grad} \delta p \right) = \frac{\delta\kappa^*}{\kappa_0} \frac{\partial^2 p_0}{\partial t^2} \quad . \quad (8.151)$$

Explicitly, replacing $\delta p = p - p_0$ and $\delta \kappa^* = \log \kappa / \kappa_0$, gives

$$\frac{1}{\kappa_0} \frac{\partial^2 (p - p_0)}{\partial t^2} - \operatorname{div} \left(\frac{1}{\rho_0} \mathbf{grad} (p - p_0) \right) = \frac{1}{\kappa_0} \log \frac{\kappa}{\kappa_0} \frac{\partial^2 p_0}{\partial t^2}. \quad (8.152)$$

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 incompressibility* κ_0 , and is generated by the ‘Born secondary source’

$$S_{\text{Born}} = \frac{1}{\kappa_0} \log \frac{\kappa}{\kappa_0} \frac{\partial^2 p_0}{\partial t^2}. \quad (8.153)$$

Should we have made the development using the compressibility $\gamma = 1/\kappa$ instead of the incompressibility, we would have arrived at the secondary source

$$S_{\text{Born}} = \gamma_0 \log \frac{\gamma}{\gamma_0} \frac{\partial^2 p_0}{\partial t^2} \quad (8.154)$$

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 κ would have lead to the Born secondary source

$$S_\kappa = \frac{\delta \kappa}{\kappa^2} \frac{\partial^2 p_0}{\partial t^2} = \frac{\kappa - \kappa_0}{\kappa^2} \frac{\partial^2 p_0}{\partial t^2} \quad (8.155)$$

while an additive perturbation $\gamma = \gamma_0 + \delta \gamma$ of the positive parameter $\gamma = 1/\kappa$ would have lead to the Born secondary source

$$S_\gamma = -\delta \gamma \frac{\partial^2 p_0}{\partial t^2} = (\gamma - \gamma_0) \frac{\partial^2 p_0}{\partial t^2}, \quad (8.156)$$

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 8.155 or 8.156 quite bad approximations, while the logarithmic expressions in the right Born source (equation 8.153 or 8.154) may remain good.

8.3.4.9 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 incompressibility is perturbed by the amount $\delta \kappa^*(\mathbf{x})$ as (linear tangent application)

$$\delta \mathbf{p} = \mathbf{F} \delta \kappa^*, \quad (8.157)$$

or, introducing the kernel of the Fréchet derivative \mathbf{F} ,

$$\delta p(\mathbf{x}_i, t) = \int_V dV(\mathbf{x}') F(\mathbf{x}_i, t; \mathbf{x}') \delta \kappa^*(\mathbf{x}') . \quad (8.158)$$

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 8.151)

$$S_{\text{Born}}(\mathbf{x}, t) = \frac{\delta \kappa^*(\mathbf{x})}{\kappa_0(\mathbf{x})} \ddot{p}_0(\mathbf{x}, t) . \quad (8.159)$$

Then, using the Green function,

$$\begin{aligned} \delta p(\mathbf{x}_i, t) &= \int_V dV(\mathbf{x}') \int_{t_2}^{t_1} dt' G(\mathbf{x}_i, t; \mathbf{x}', t') S_{\text{Born}}(\mathbf{x}', t') \\ &= \int_V dV(\mathbf{x}') \int_{t_2}^{t_1} dt' G(\mathbf{x}_i, t; \mathbf{x}', t') \frac{\delta \kappa^*(\mathbf{x}')}{\kappa_0(\mathbf{x}')} \ddot{p}_0(\mathbf{x}', t') . \end{aligned} \quad (8.160)$$

The last expression can be rearranged into the form used in equation 8.158, this showing that $F(\mathbf{x}_i, t; \mathbf{x}', t')$ is given by

$$F(\mathbf{x}_i, t; \mathbf{x}') = \frac{1}{\kappa_0(\mathbf{x}')} \int_{t_2}^{t_1} dt' G(\mathbf{x}_i, t; \mathbf{x}', t') \ddot{p}_0(\mathbf{x}', t') \quad (8.161)$$

This is the kernel of the Fréchet derivative of the data with respect to the parameter $\kappa^*(\mathbf{x})$.

8.3.4.10 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 expression like

$$\delta \widehat{\boldsymbol{\kappa}}^* = \mathbf{F}^T \delta \widehat{\mathbf{p}} ? \quad (8.162)$$

Denoting by $F^T(\mathbf{x}'; \mathbf{x}_i, t)$ the kernel of \mathbf{F}^T , such an expression writes

$$\delta \widehat{\boldsymbol{\kappa}}^* = \sum_i \int_{t_2}^{t_1} dt F^T(\mathbf{x}'; \mathbf{x}_i, t) \delta \widehat{p}(\mathbf{x}_i, t) , \quad (8.163)$$

but we know that the kernel of the transpose operator equals the kernel of the original operator, with variables transposed (note: say where this has been demonstrated), so that we can write this equation as

$$\delta \widehat{\boldsymbol{\kappa}}^* = \sum_i \int_{t_2}^{t_1} dt F(\mathbf{x}_i, t; \mathbf{x}') \delta \widehat{p}(\mathbf{x}_i, t) , \quad (8.164)$$

where $F(\mathbf{x}_i, t; \mathbf{x}')$ is the kernel given in equation 8.161. Replacing the kernel by its expression gives

$$\delta \widehat{\boldsymbol{\kappa}}^*(\mathbf{x}') = \sum_i \int_{t_2}^{t_1} dt \frac{1}{\kappa_0(\mathbf{x}')} \int_{t_2}^{t_1} dt' G(\mathbf{x}_i, t; \mathbf{x}', t') \ddot{p}_0(\mathbf{x}', t') \delta \widehat{p}(\mathbf{x}_i, t) , \quad (8.165)$$

and this can be rearranged into (note that primed and nonprimed variables have been exchanged)

$$\delta\widehat{\kappa}^*(\mathbf{x}) = \frac{1}{\kappa_0(\mathbf{x})} \int_{t_2}^{t_1} dt \psi(\mathbf{x}, t) \ddot{p}_0(\mathbf{x}, t) , \quad (8.166)$$

where

$$\psi(\mathbf{x}, t) = \sum_i \int_{t_2}^{t_1} dt' G(\mathbf{x}_i, t'; \mathbf{x}, t) \delta\widehat{p}(\mathbf{x}_i, t') , \quad (8.167)$$

or, using the kernel of the transposed Green's operator,

$$\psi(\mathbf{x}, t) = \sum_i \int_{t_2}^{t_1} dt' G^T(\mathbf{x}, t; \mathbf{x}_i, t') \delta\widehat{p}(\mathbf{x}_i, t') . \quad (8.168)$$

(note: explain here that this means that 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).

8.3.4.11 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

$$S(\boldsymbol{\kappa}^*) = \frac{1}{2} (\| \mathbf{f}(\boldsymbol{\kappa}^*) - \mathbf{p}_{\text{obs}} \|^2 + \| \boldsymbol{\kappa}^* - \boldsymbol{\kappa}_{\text{prior}}^* \|^2) \quad (8.169)$$

$$= \frac{1}{2} (\langle \mathbf{C}_p^{-1}(\mathbf{f}(\boldsymbol{\kappa}^*) - \mathbf{p}_{\text{obs}}) , \mathbf{f}(\boldsymbol{\kappa}^*) - \mathbf{p}_{\text{obs}} \rangle + \langle \mathbf{C}_{\boldsymbol{\kappa}^*}^{-1}(\boldsymbol{\kappa}^* - \boldsymbol{\kappa}_{\text{prior}}^*) , \boldsymbol{\kappa}^* - \boldsymbol{\kappa}_{\text{prior}}^* \rangle) .$$

Using, in this functional context, the steepest descent algorithm proposed in section 8.3.7.4, we arrive at

$$\boldsymbol{\kappa}_{n+1}^* = \boldsymbol{\kappa}_n^* - \epsilon (\mathbf{C}_{\boldsymbol{\kappa}^*} \mathbf{F}_n^T \mathbf{C}_p^{-1} (\mathbf{p}_n - \mathbf{p}_{\text{obs}}) + (\boldsymbol{\kappa}_n^* - \boldsymbol{\kappa}_{\text{prior}}^*)) , \quad (8.170)$$

where $\mathbf{p}_n = \mathbf{f}(\boldsymbol{\kappa}_n^*)$ and where \mathbf{F}_n^T is the transposed operator 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).

8.3.5 Appendix: Nonlinear Inversion of Waveforms (by Charara & Barnes)

[Note: I plan to convince Marwan and Christophe to contribute to our book by writing this section (on a work that, unfortunately, has never been published).]

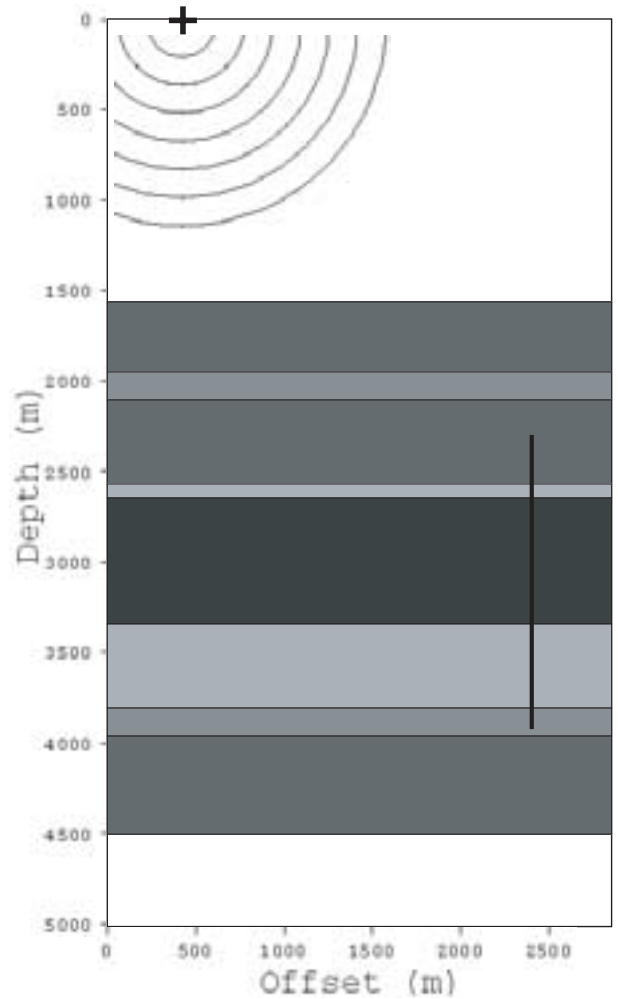


Figure 8.12: Geometry.

VSP WEST filtered real data X

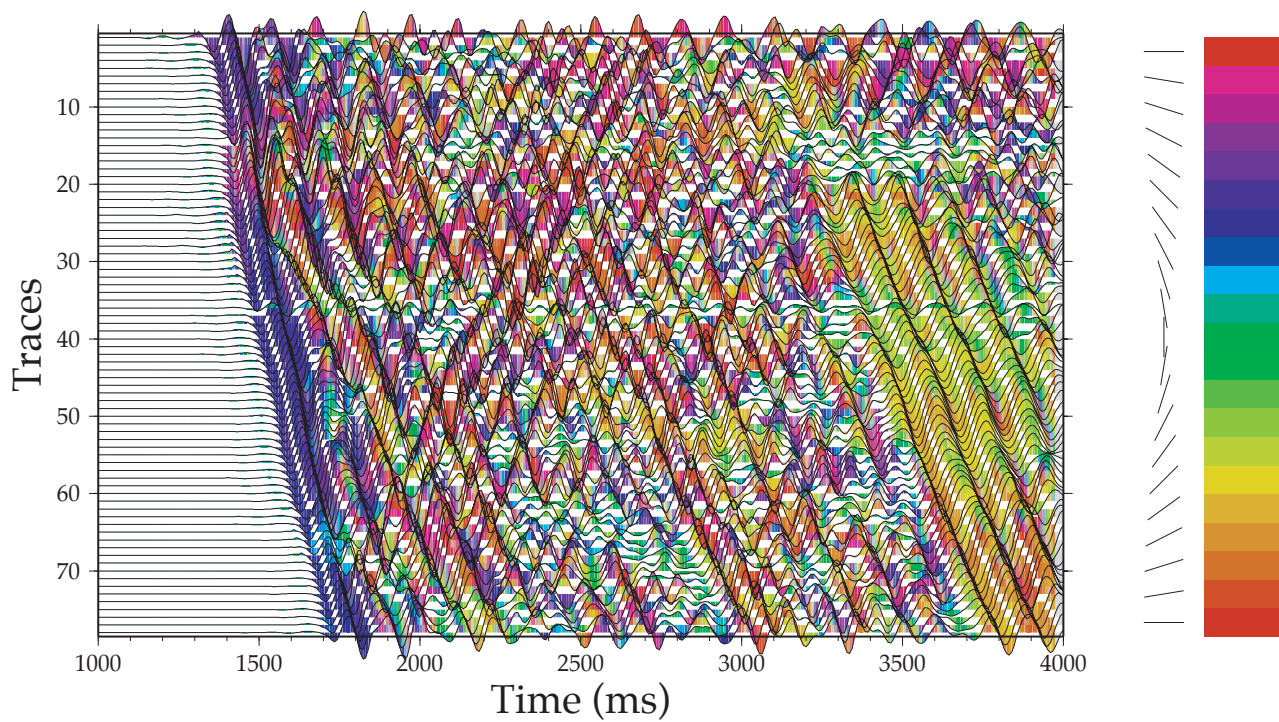


Figure 8.13: Observed seismograms. X component.

VSP WEST filtered real data Z

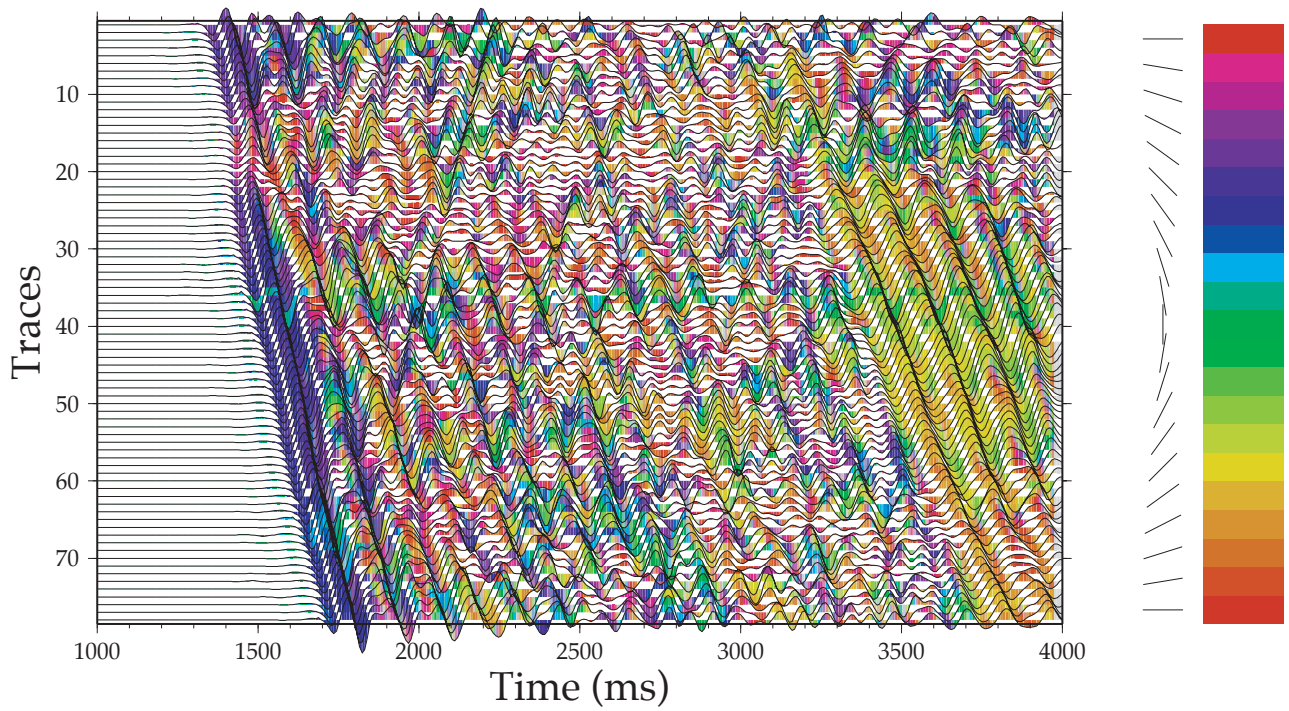


Figure 8.14: Observed seismograms. Z component.

P Velocity

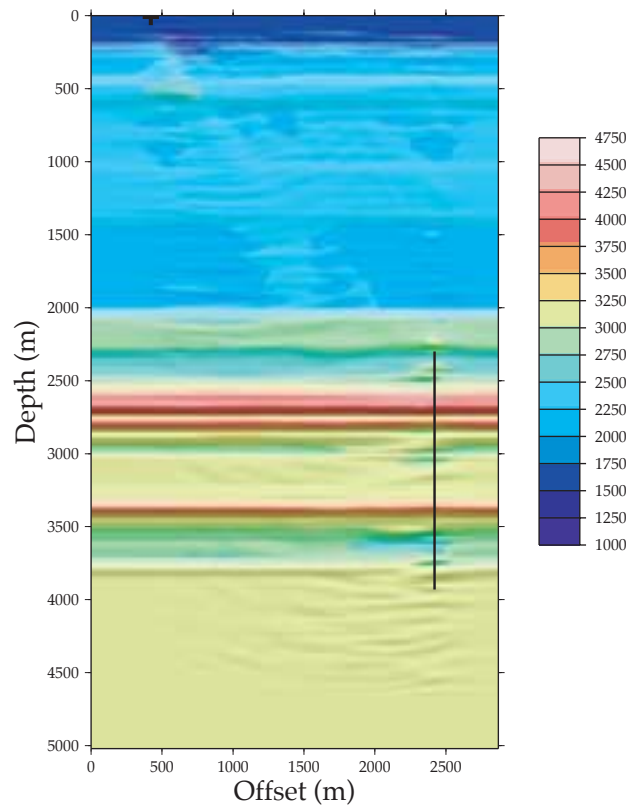


Figure 8.15: Model. VP.

Figure 8.16: Model. VS.

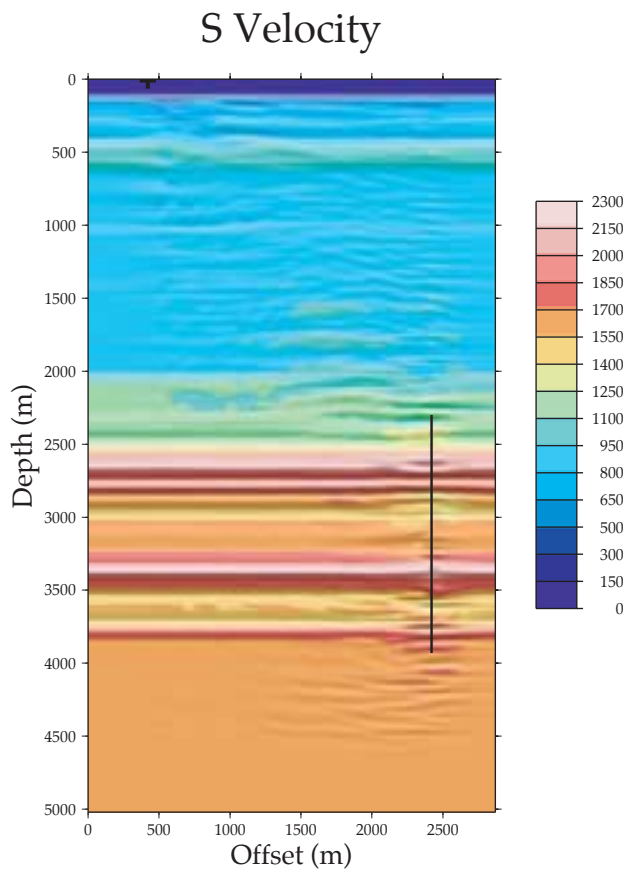


Figure 8.17: Model. RHO.

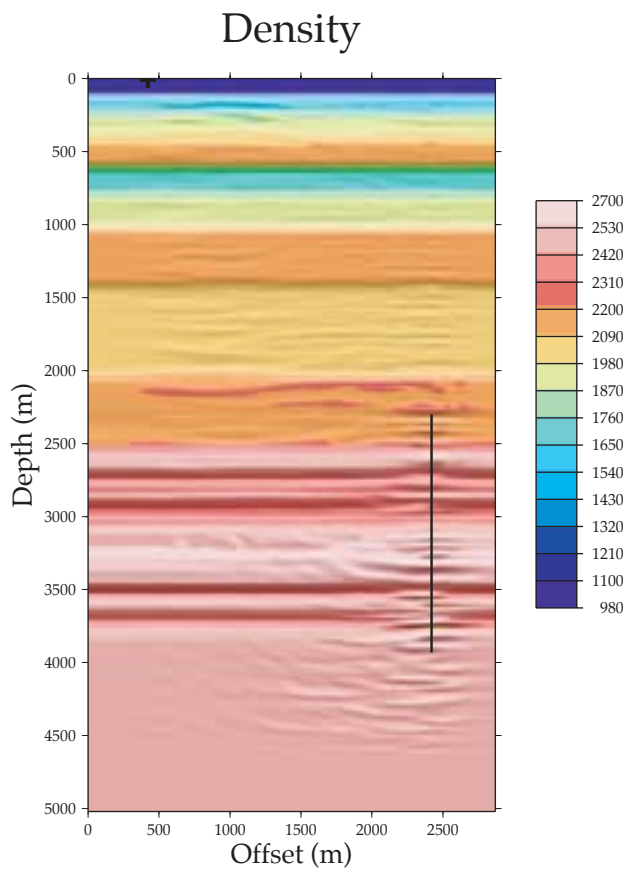
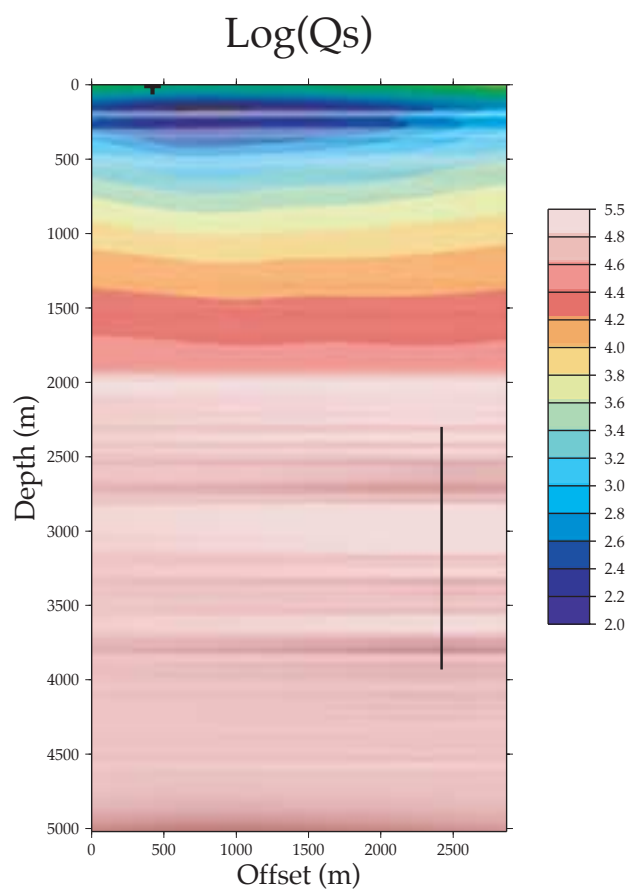


Figure 8.18: Model. Q.



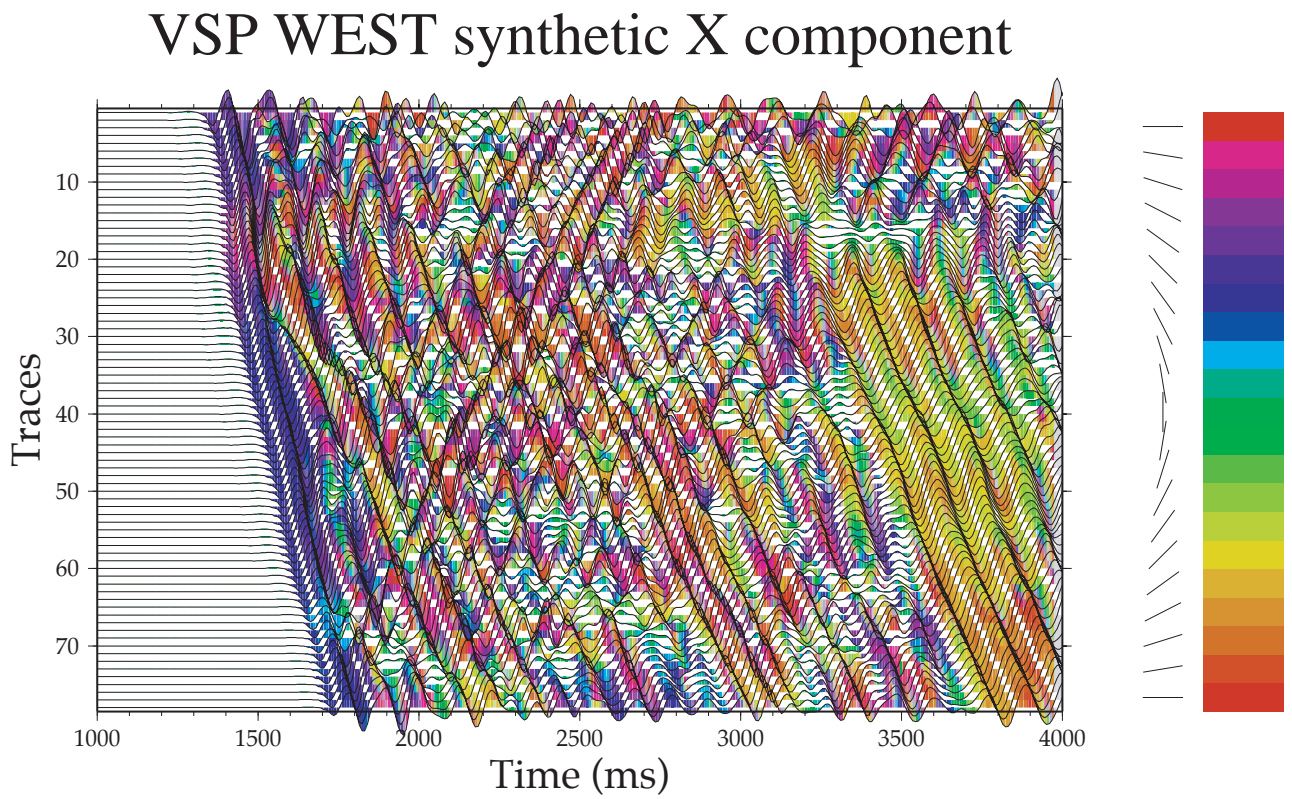


Figure 8.19: Calculated seismograms. X component.

VSP WEST synthetic Z component

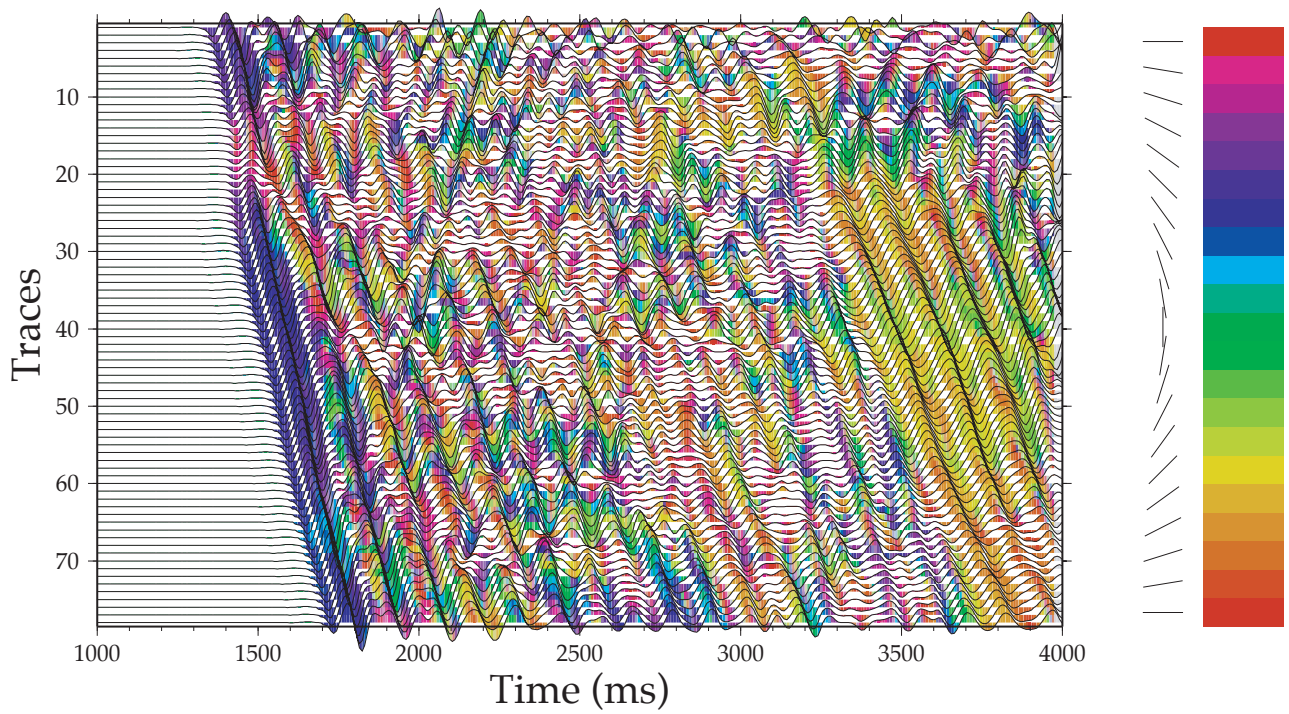


Figure 8.20: Calculated seismograms. Z component.

VSP WEST residuals X component

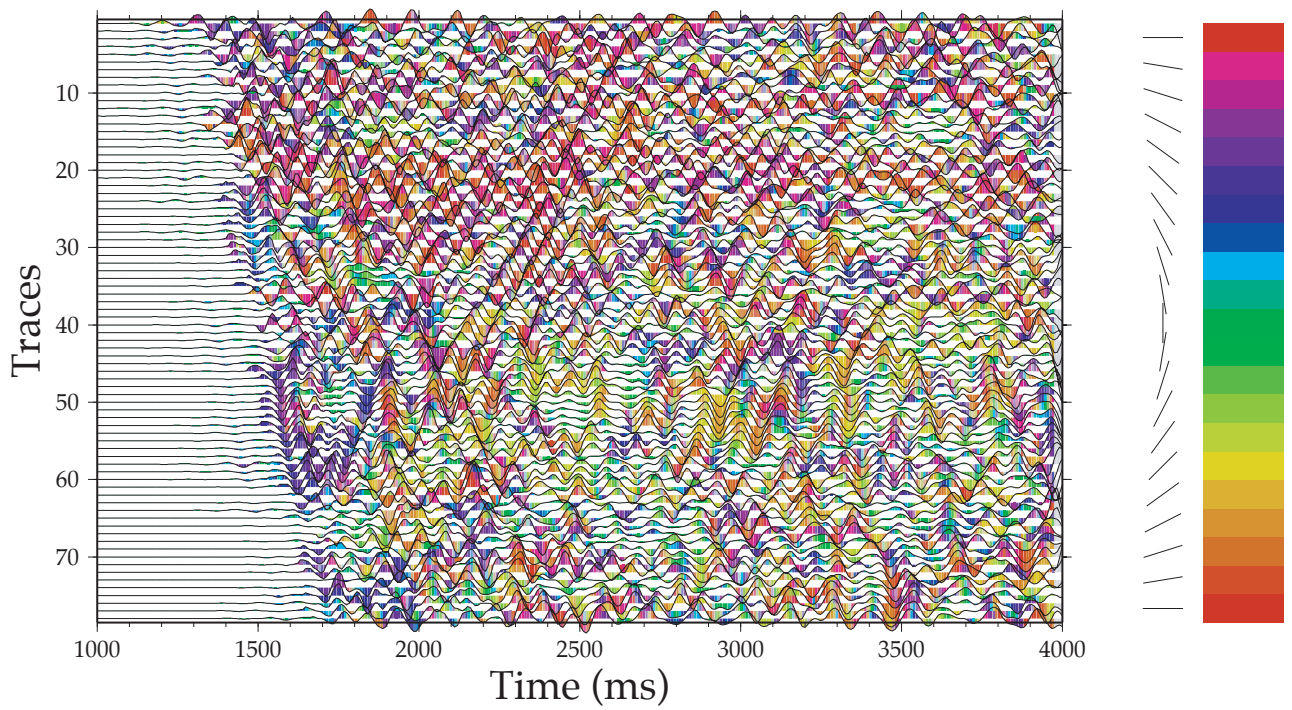


Figure 8.21: Residuals seismograms. X component.

VSP WEST residuals Z component

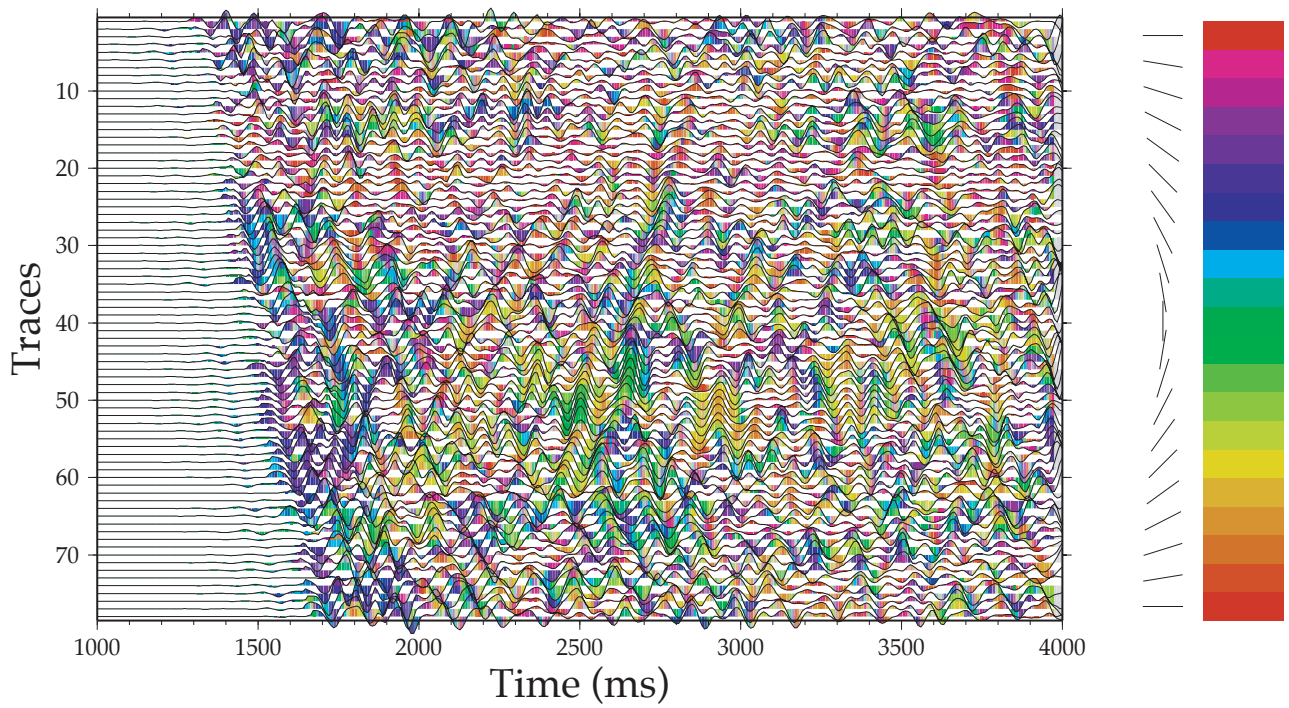


Figure 8.22: Residuals seismograms. Z component.

8.3.6 Appendix: Using Monte Carlo Methods

[Note: Write a small introduction here].

8.3.6.1 Basic Equations

The starting point could be the general equation 7.9,

$$\sigma(\mathbf{m}, \mathbf{d}) = k \frac{\rho(\mathbf{m}, \mathbf{d}) \vartheta(\mathbf{m}, \mathbf{d})}{\mu(\mathbf{m}, \mathbf{d})} , \quad (8.171)$$

combining the ‘a priori’ information $\rho(\mathbf{m}, \mathbf{d})$ with the ‘theoretical’ information $\vartheta(\mathbf{m}, \mathbf{d})$. We have seen in section 3 that if we are able to design a random walk that samples $\rho(\mathbf{m}, \mathbf{d})$, then, the Metropolis rule can be used to obtain a random walk that samples $\sigma(\mathbf{m}, \mathbf{d})$. We have also seen that if we are not able to design a (primeval) random walk that samples $\rho(\mathbf{m}, \mathbf{d})$, then we can start using a random walk that samples the homogeneous probability density $\mu(\mathbf{m}, \mathbf{d})$, of even an arbitrary¹⁵ probability density $\psi(\mathbf{m}, \mathbf{d})$.

This point of view, is very general, but more practical algorithms are obtained when we particularize.

Let us consider, for instance, the explicit expression (equation ??) for $\sigma_m(\mathbf{m})$ given in section 8.2.6:

$$\sigma_m(\mathbf{m}) = k \rho_m(\mathbf{m}) \frac{\phi(\mathbf{m})}{\mu_m(\mathbf{m})} . \quad (8.172)$$

where

$$\phi(\mathbf{m}) = \left(\frac{\rho_d(\mathbf{d})}{\mu_d(\mathbf{d})} \sqrt{\det(\mathbf{g}_m(\mathbf{m}) + \mathbf{F}^T(\mathbf{m}) \mathbf{g}_d(\mathbf{d}) \mathbf{F}(\mathbf{m}))} \right) \Big|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} . \quad (8.173)$$

In this expression the matrix of partial derivatives $\mathbf{F} = \mathbf{F}(\mathbf{m})$, with components $D_{i\alpha} = \partial d_i / \partial m_\alpha$, 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 8.171, the effect is automatically accounted for, and there is no need to explicitly consider the partial derivatives.

In any case, equation 8.172 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 $\mathbf{F}^T(\mathbf{m}) \mathbf{g}_d(\mathbf{d}) \mathbf{F}(\mathbf{m}) \ll \mathbf{g}_m(\mathbf{m})$ and, then,

$$\phi(\mathbf{m}) = \mu_m(\mathbf{m}) \frac{\rho_d(\mathbf{d})}{\mu_d(\mathbf{d})} \Big|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} , \quad (8.174)$$

and equation 8.172 becomes

$$\sigma_m(\mathbf{m}) = k \rho_m(\mathbf{m}) L(\mathbf{m}) , \quad (8.175)$$

¹⁵Although, hopefully, not too different from $\mu(\mathbf{m}, \mathbf{d})$.

where

$$L(\mathbf{m}) = \frac{\rho_d(\mathbf{d})}{\mu_d(\mathbf{d})} \Big|_{\mathbf{d}=\mathbf{f}(\mathbf{m})} . \quad (8.176)$$

This expression is also ready for use using 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” $L(\mathbf{m})$ (in fact, a volumetric probability) to calculate acceptance probabilities.

8.3.6.2 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 $\mu(\mathbf{m})$. In some cases this is easy, but in other cases only an algorithm (a *primeval random walk*) that samples a probability density $\psi(\mathbf{m}) \neq \mu(\mathbf{m})$ is available. Then the Metropolis Rule can be used to modify $\psi(\mathbf{m})$ into $\mu(\mathbf{m})$. This way of generating samples from $\mu(\mathbf{m})$ is efficient if $\psi(\mathbf{m})$ is close to $\mu(\mathbf{m})$, otherwise it may be very inefficient. Methods for designing primeval random walks are found in Section 3.4.

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

8.3.6.3 Sampling the Prior Probability Distribution

The first step in the Monte Carlo analysis is to switch off the comparison between computed and observed data, thereby generating samples of the a priori probability density. This allows us 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”. The more models present in this movie, the more accurate representation of the prior probability density.

If we are interested in smooth Earth models (knowing, e.g., that only smooth properties are resolved by the data), a smooth movie can be produced simply by smoothing the individual models of the original movie.

8.3.6.4 Sampling the Posterior Probability Distribution

If we now switch on the comparison between computed and observed data using, e.g., the Metropolis Rule, the random walk sampling the prior distribution is modified into a walk sampling the posterior distribution. Again, smoothed versions of this “posterior movie” can be generated by smoothing the individual models in the original, posterior movie.

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. 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

$$R_f = \int_{\mathcal{A}} d\mathbf{m} f(\mathbf{m}) \sigma(\mathbf{m}) \quad (8.177)$$

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,

$$\mathcal{A} = \{\mathbf{m} \mid \text{a given range of parameters in } \mathbf{m} \text{ is } \textit{cyclic}\}. \quad (8.178)$$

In the special case when $\mathcal{A} = \mathcal{M}$ is the entire model space, and $f(\mathbf{m}) = m_i$, the R_f in eq. (8.177) 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 (8.177) analytically because we have no analytical expression for $\sigma(\mathbf{m})$. However, from the samples of the posterior movie $\mathbf{m}_1, \dots, \mathbf{m}_N$ we can approximate R_f by the simple average:

$$R_f \approx \sum_{\{n \mid \mathbf{m}_n \in \mathcal{A}\}} f(\mathbf{m}_n). \quad (8.179)$$

8.3.7 Appendix: Using Optimization 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 highly-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. *[Note: I rephrased this sentence]* 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).

8.3.7.1 Maximum Likelihood Point

Let us consider a space \mathcal{X} , with a notion of volume element dV defined. If some coordinates $\mathbf{x} \equiv \{x^1, x^2, \dots, x^n\}$ 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 ΔV , we can search for the point \mathbf{x}_{ML} such that the probability dP of the small volume, when centered around \mathbf{x}_{ML} , gets a maximum. In the limit $\Delta V \rightarrow 0$ this defines the *maximum likelihood point*. The maximum likelihood point may be unique (if the probability distribution is monomodal), may be degenerated (if the probability distribution is ‘roof-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. *[Note: Rephrase the following sentence...]* For our definition imposes that what must be maximum is the ratio of the probability density by the function $v(\mathbf{x})$ defining the volume element:

$$\mathbf{x} = \mathbf{x}_{ML} \iff F(\mathbf{x}) = \frac{f(\mathbf{x})}{v(\mathbf{x})} \text{ maximum} . \quad (8.180)$$

We recognize in the ratio $F(\mathbf{x}) = f(\mathbf{x})/v(\mathbf{x})$ the volumetric probability associated to the probability density $f(\mathbf{x})$ (see equation ??). As the homogeneous probability density is $\mu(\mathbf{x}) = k v(\mathbf{x})$ (see rule 4.2), we can equivalently define the maximum likelihood point by the condition

$$\mathbf{x} = \mathbf{x}_{ML} \iff \frac{f(\mathbf{x})}{\mu(\mathbf{x})} \text{ maximum} . \quad (8.181)$$

The point at which a probability density has its maximum is not \mathbf{x}_{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 8.181, that is an intrinsically defined point: no matter which coordinates we use in the computation we always obtain the same point of the space.

8.3.7.2 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

$$S(\mathbf{x}) = -\log \frac{f(\mathbf{x})}{\mu(\mathbf{x})} . \quad (8.182)$$

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 $S(\mathbf{x})$:

$$\mathbf{x} = \mathbf{x}_{ML} \iff S(\mathbf{x}) \text{ minimum} . \quad (8.183)$$

Example 8.16 The conjunction $\sigma(\mathbf{x})$ of two probability densities $\rho(\mathbf{x})$ and $\vartheta(\mathbf{x})$ was defined (equation ??) as

$$\sigma(\mathbf{x}) = p \frac{\rho(\mathbf{x}) \vartheta(\mathbf{x})}{\mu(\mathbf{x})} . \quad (8.184)$$

Then,

$$S(\mathbf{x}) = S_\rho(\mathbf{x}) + S_\vartheta(\mathbf{x}) , \quad (8.185)$$

where

$$S_\rho(\mathbf{x}) = -\log \frac{\rho(\mathbf{x})}{\mu(\mathbf{x})} ; \quad S_\vartheta(\mathbf{x}) = -\log \frac{\vartheta(\mathbf{x})}{\mu(\mathbf{x})} . \quad (8.186)$$

[End of example.]

Example 8.17 In the context of Gaussian distributions, we have found the probability density (see example ??)

$$\sigma_m(\mathbf{m}) = \quad (8.187)$$

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

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

$$2S(\mathbf{m}) = (\mathbf{m} - \mathbf{m}_{\text{prior}})^t \mathbf{C}_M^{-1} (\mathbf{m} - \mathbf{m}_{\text{prior}}) + (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^t \mathbf{C}_D^{-1} (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) . \quad (8.188)$$

The reader should remember that this misfit function is valid only for weakly nonlinear problems (see examples 8.5 and ??). The maximum likelihood model here is the one that minimizes the sum of squares 8.188. This corresponds to the least squares criterion. [End of example.]

Example 8.18 *In the context of Laplacian distributions, we have found the probability density (see example ??)*

$$\sigma_m(\mathbf{m}) = k \exp \left(- \left(\sum_{\alpha} \frac{|m^{\alpha} - m_{\text{prior}}^{\alpha}|}{\sigma_{\alpha}} + \sum_i \frac{|f^i(\mathbf{m}) - d_{\text{obs}}^i|}{\sigma_i} \right) \right) . \quad (8.189)$$

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

$$S(\mathbf{m}) = \sum_{\alpha} \frac{|m^{\alpha} - m_{\text{prior}}^{\alpha}|}{\sigma_{\alpha}} + \sum_i \frac{|f^i(\mathbf{m}) - d_{\text{obs}}^i|}{\sigma_i} . \quad (8.190)$$

The reader should remember that this misfit function is valid only for weakly nonlinear problems. The maximum likelihood model here is the one that minimizes the sum of least absolute values 8.190. This correpponds to the least absolute values criterion. [End of example.]

8.3.7.3 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 function¹⁶ $S(P, T)$ over a pressure P and a temperature T . Any sensible definition of the gradient of S will lead to an expression like

$$\text{grad } S = \begin{pmatrix} \frac{\partial S}{\partial P} \\ \frac{\partial S}{\partial T} \end{pmatrix} \quad (8.191)$$

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 application that is tangent to $S(\mathbf{x})$ at \mathbf{x}_0 . [Note: Rephrase the following sentence...]* This definition of gradient is consistent with the more elementary one, based on the use of the first order development

$$S(\mathbf{x}_0 + \delta\mathbf{x}) = S(\mathbf{x}_0) + \widehat{\gamma}_0^T \delta\mathbf{x} + \dots \quad (8.192)$$

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

$$(\widehat{\gamma}_0)_p = \frac{\partial S}{\partial x^p}(\mathbf{x}_0) . \quad (8.193)$$

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

¹⁶We take this example because typical misfit functions are adimensional, but the argument has general validity.

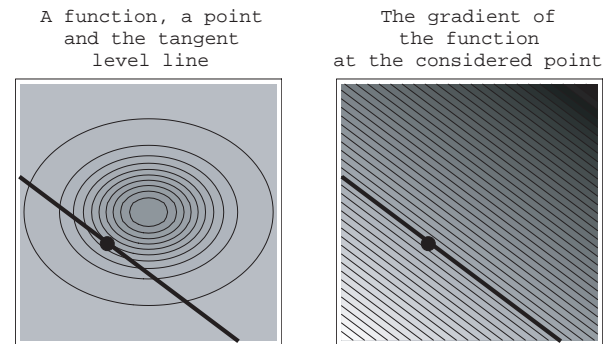
gradient (at a given point) of a function defined over a given space \mathcal{E}) is an element of the dual of the space. To obtain a direction in \mathcal{E} , we must pass from the dual to the primal space. As usual, it is the metric of the space that maps the dual of the space into the space itself. So if \mathbf{g} is the metric of the space where $S(\mathbf{x})$ is defined, and if $\hat{\gamma}$ is the gradient of S at a given point, the *direction of steepest ascent* is

$$\boldsymbol{\gamma} = \mathbf{g}^{-1} \hat{\boldsymbol{\gamma}} \quad . \quad (8.194)$$

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 $S(\mathbf{x})$ gets its maximum value.

Example 8.19 *Figure 8.23 represents the level lines of a scalar function $S(u, v)$ in a 2D space. A particular point has been selected. What is the gradient of the function at the given point? As suggested in the main text, it is not an arrow ‘perpendicular’ to the level lines of the function at the considered point, as the notion of perpendicularity will depend on a metric not yet specified (and unnecessary to define the gradient). The gradient must be seen as ‘the linear function that is tangent to $S(u, v)$ at the considered point’. If $S(u, v)$ has been represented by its level lines, then the gradient may also be represented by its level lines (right of the figure). We see that the condition, in fact, is that the level lines of the gradient are tangent to the level lines of the original function (at the considered point). Contrary to the notion of perpendicularity, the notion of tangency is metric-independent. [End of example.]*

Figure 8.23: The gradient of a function has not to be seen as a vector orthogonal to the level lines, but as a form parallel to them (see text.)



Example 8.20 *In the context of least squares, we consider a misfit function $S(\mathbf{m})$ and a covariance matrix \mathbf{C}_M . If $\hat{\boldsymbol{\gamma}}_0$ 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*

$$\boldsymbol{\gamma}_0 = \mathbf{C}_M \hat{\boldsymbol{\gamma}}_0 \quad . \quad (8.195)$$

[End of example.]

Example 8.21 *If the misfit function $S(P, T)$ depends on a pressure P and on a temperature T , the gradient of S is, as mentioned above (equation 8.191),*

$$\hat{\boldsymbol{\gamma}} = \begin{pmatrix} \frac{\partial S}{\partial P} \\ \frac{\partial S}{\partial T} \end{pmatrix} \quad . \quad (8.196)$$

As the quantities P and T are Jeffreys quantities, associated to the metric $ds^2 = \left(\frac{dP}{P}\right)^2 + \left(\frac{dT}{T}\right)^2$, the direction of steepest ascent is¹⁷

$$\boldsymbol{\gamma} = \begin{pmatrix} P^2 \frac{\partial S}{\partial P} \\ T^2 \frac{\partial S}{\partial T} \end{pmatrix} . \quad (8.197)$$

[End of example.]

8.3.7.4 The Steepest Descent Method

Consider that we have a probability distribution defined over an n -dimensional space \mathcal{X} . Having chosen some coordinates $\mathbf{x} \equiv \{x^1, x^2, \dots, x^n\}$ over the space, the probability distribution is represented by the probability density $f(\mathbf{x})$ whose homogeneous limit (in the sense developed in section 4) is $\mu(\mathbf{x})$. We wish to calculate the coordinates \mathbf{x}_{ML} of the maximum likelihood point. By definition (equation 8.181),

$$\mathbf{x} = \mathbf{x}_{ML} \iff \frac{f(\mathbf{x})}{\mu(\mathbf{x})} \text{ maximum} , \quad (8.198)$$

i.e.,

$$\mathbf{x} = \mathbf{x}_{ML} \iff S(\mathbf{x}) \text{ minimum} , \quad (8.199)$$

where $S(\mathbf{x})$ is the misfit (equation 8.182)

$$S(\mathbf{x}) = -k \log \frac{f(\mathbf{x})}{\mu(\mathbf{x})} . \quad (8.200)$$

Let us denote by $\widehat{\boldsymbol{\gamma}}(\mathbf{x}_k)$ the gradient of $S(\mathbf{x})$ at point \mathbf{x}_k , i.e. (equation 8.193),

$$(\widehat{\boldsymbol{\gamma}}_0)_p = \frac{\partial S}{\partial x^p}(\mathbf{x}_0) . \quad (8.201)$$

We have seen above that $\widehat{\boldsymbol{\gamma}}(\mathbf{x})$ is not to be interpreted as a direction in the space \mathcal{X} , but a direction in the dual space. The gradient can be converted into a direction using some metric $\mathbf{g}(\mathbf{x})$ over \mathcal{X} . In simple situations the metric \mathbf{g} will be that 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 8.195)

$$\boldsymbol{\gamma}(\mathbf{x}_k) = \mathbf{g}^{-1}(\mathbf{x}_k) \widehat{\boldsymbol{\gamma}}(\mathbf{x}_k) . \quad (8.202)$$

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,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \varepsilon_k \mathbf{g}_k^{-1} \widehat{\boldsymbol{\gamma}}_k , \quad (8.203)$$

¹⁷We have here $\begin{pmatrix} g_{PP} & g_{PT} \\ g_{TP} & g_{TT} \end{pmatrix} = \begin{pmatrix} 1/P^2 & 0 \\ 0 & 1/T^2 \end{pmatrix}$.

where ε_k is an ad-hoc (real, positive) value adjusted to force the algorithm to converge rapidly (if ε_k is chosen too small the convergence may be too slow; if it is 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 8.203. These algorithms are not consistent. Even the physical dimensionality of the equation is not assured. The authors of this article have traced some ‘numerical’ problems in existing computer implementations of steepest descent algorithms to this neglect of the metric.

Example 8.22 *In the context of example 8.17, where the misfit function $S(\mathbf{m})$ is given by*

$$2S(\mathbf{m}) = (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^t \mathbf{C}_D^{-1} (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) + (\mathbf{m} - \mathbf{m}_{\text{prior}})^t \mathbf{C}_M^{-1} (\mathbf{m} - \mathbf{m}_{\text{prior}}) \quad , \quad (8.204)$$

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

$$\hat{\boldsymbol{\gamma}}(\mathbf{m}) = \mathbf{F}^t(\mathbf{m}) \mathbf{C}_D^{-1} (\mathbf{f}(\mathbf{m}) - \mathbf{d}_{\text{obs}}) + \mathbf{C}_M^{-1} (\mathbf{m} - \mathbf{m}_{\text{prior}}) \quad , \quad (8.205)$$

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

$$F^{i\alpha} = \frac{\partial f^i}{\partial m^\alpha} \quad . \quad (8.206)$$

An example of computation of partial derivatives is given in appendix ??. [End of example.]

Example 8.23 *In the context of example 8.22 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 8.205, the steepest descent algorithm 8.203 becomes*

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \varepsilon_k (\mathbf{C}_M \mathbf{F}_k^t \mathbf{C}_D^{-1} (\mathbf{f}_k - \mathbf{d}_{\text{obs}}) + (\mathbf{m}_k - \mathbf{m}_{\text{prior}})) \quad , \quad (8.207)$$

where $\mathbf{F}_k \equiv \mathbf{F}(\mathbf{m}_k)$ and $\mathbf{f}_k \equiv \mathbf{f}(\mathbf{m}_k)$. The real positive quantities ε_k can be fixed, after some trial and error, by accurate linear search, or by using a linearized approximation¹⁸. [End of example.]

Example 8.24 *In the context of example 8.22 the model space \mathcal{M} has a less obvious metric, namely that defined by the inverse of the ‘a posteriori’ covariance operator, $\mathbf{g} = \tilde{\mathbf{C}}_M^{-1}$. Note: Explain here that the ‘best current estimator’ of $\tilde{\mathbf{C}}_M$ is*

$$\tilde{\mathbf{C}}_M \approx \left(\mathbf{F}_k^t \mathbf{C}_D^{-1} \mathbf{F}_k + \mathbf{C}_M^{-1} \right)^{-1} \quad . \quad (8.208)$$

Using this metric and the gradient given by equation 8.205, the steepest descent algorithm 8.203 becomes

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \varepsilon_k \left(\mathbf{F}_k^t \mathbf{C}_D^{-1} \mathbf{F}_k + \mathbf{C}_M^{-1} \right)^{-1} \left(\mathbf{F}_k^t \mathbf{C}_D^{-1} (\mathbf{f}_k - \mathbf{d}_{\text{obs}}) + \mathbf{C}_M^{-1} (\mathbf{m}_k - \mathbf{m}_{\text{prior}}) \right) \quad , \quad (8.209)$$

¹⁸As 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 ε_k gives
$$\varepsilon_k = \frac{\boldsymbol{\gamma}_k^t \mathbf{C}_M^{-1} \boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_k^t (\mathbf{F}_k^t \mathbf{C}_D^{-1} \mathbf{F}_k + \mathbf{C}_M^{-1}) \boldsymbol{\gamma}_k} \quad .$$

where $\mathbf{F}_k \equiv \mathbf{F}(\mathbf{m}_k)$ and $\mathbf{f}_k \equiv \mathbf{f}(\mathbf{m}_k)$. The real positive quantities ε_k can be fixed, after some trial and error, by accurate linear search, or by using a linearized approximation that simply gives¹⁹ $\varepsilon_k \approx 1$. [End of example.]

The algorithm 8.209 is usually called a ‘quasi-Newton algorithm’. [Note: Rephrase the following sentence...] This is a misname, as a Newton method applied to the minimization of the misfit function $S(\mathbf{m})$ would be a method using the second derivatives of $S(\mathbf{m})$, and thus the derivatives $H_{\alpha\beta}^i = \frac{\partial^2 f^i}{\partial m^\alpha \partial m^\beta}$, that are not computed (or not estimated) when using this algorithm. It is just a steepest descent algorithm with a nontrivial definition of metric in the working space. In this sense it belongs to the wider class of ‘variable metric methods’, not discussed in this article.

Example 8.25 In the context of example 8.18, where the misfit function $S(\mathbf{m})$ is given by

$$S(\mathbf{m}) = \sum_i \frac{|f^i(\mathbf{m}) - d_{\text{obs}}^i|}{\sigma_i} + \sum_\alpha \frac{|m^\alpha - m_{\text{prior}}^\alpha|}{\sigma_\alpha} \quad , \quad (8.210)$$

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

$$\hat{\gamma}_\alpha = \sum_i F^{i\alpha} \frac{1}{\sigma_i} \text{sign}(f^i - d_{\text{obs}}^i) + \frac{1}{\sigma_\alpha} \text{sign}(m^\alpha - m_{\text{prior}}^\alpha) \quad , \quad (8.211)$$

where $F^{i\alpha} = \partial f^i / \partial m^\alpha$. We can now choose in the model space the ad-hoc metric defined as the inverse of the ‘covariance matrix’ formed by the square of the mean deviations σ_i and σ_α (interpreted as if they were variances). Using this metric, the direction of steepest ascent associated to the gradient in 8.211, is

$$\gamma_\alpha = \sum_i F^{i\alpha} \sigma_i \text{sign}(f^i - d_{\text{obs}}^i) + \sigma_\alpha \text{sign}(m^\alpha - m_{\text{prior}}^\alpha) \quad . \quad (8.212)$$

The steepest descent algorithm can now be applied:

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \varepsilon_k \boldsymbol{\gamma}_k \quad . \quad (8.213)$$

The real positive quantities ε_k can be fixed after some trial and error or by accurate linear search. [End of example.]

An expression like 8.210 defines a sort of deformed polyhedron, and to solve this sort of minimization problems the linear programming techniques are often advocated (e.g., Claerbout and Muir, 1973). We have found that for problems involving many dimensions the crude steepest descent method defined by equations 8.212–8.213 performs extremely well. For instance, in Djikpéssé and Tarantola (1999) a large-sized problem of waveform fitting is solved using this algorithm. It is well known that the sum of absolute values 8.210 provides a more robust²⁰ criterion than the sum of squares 8.204. If one fears that the data set to be used is corrupted by some unexpected errors, the least-absolute values criterion should be preferred to the least squares criterion²¹.

¹⁹While a sensible estimation of the optimal values of the real positive quantities ε_k is crucial for the algorithm 8.207, they can, in many usual circumstances, be dropped from the algorithm 8.209.

²⁰A method is ‘robust’ if its output is not sensible to a small number of large errors in the inputs.

²¹Of course, it would be much better to develop a realistic model of the uncertainties, and use the more general probabilistic methods developed above, but if those models are not available, then the least absolute values criterion is a valuable criterion.

8.3.7.5 Estimation of A Posteriori 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

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \varepsilon_k \left(\mathbf{C}_M \mathbf{F}_k^t \mathbf{C}_D^{-1} (\mathbf{f}_k - \mathbf{d}_{\text{obs}}) + (\mathbf{m}_k - \mathbf{m}_{\text{prior}}) \right) \quad , \quad (8.214)$$

(equation 8.207) or, equivalently, by the iterative algorithm

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \varepsilon_k \left(\mathbf{F}_k^t \mathbf{C}_D^{-1} \mathbf{F}_k + \mathbf{C}_M^{-1} \right)^{-1} \left(\mathbf{F}_k^t \mathbf{C}_D^{-1} (\mathbf{f}_k - \mathbf{d}_{\text{obs}}) + \mathbf{C}_M^{-1} (\mathbf{m}_k - \mathbf{m}_{\text{prior}}) \right) \quad (8.215)$$

(equation 8.209). The covariance of the tangent gaussian is

$$\tilde{\mathbf{C}}_M \approx \left(\mathbf{F}_\infty^t \mathbf{C}_D^{-1} \mathbf{F}_\infty + \mathbf{C}_M^{-1} \right)^{-1} \quad , \quad (8.216)$$

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

[note: Emphasize here the importance of $\tilde{\mathbf{C}}_M$].

8.3.7.6 Some Comments on the Use of Deterministic Methods

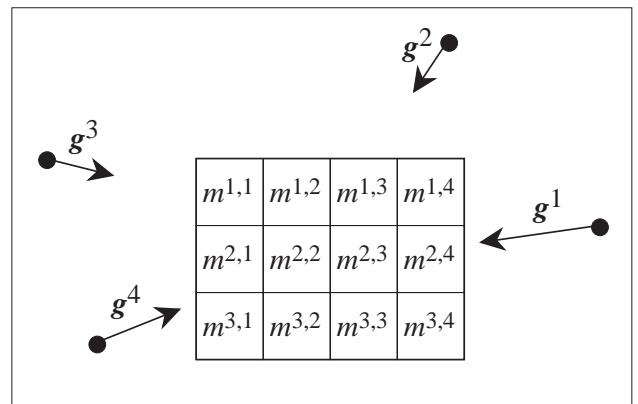
8.3.7.6.1 About the Use of the Term ‘Matrix’ [note: Warning, old text to be updated.]

Contrary to the next chapter, where the model parameter space and the data space may be functional spaces, I assume here that we have discrete spaces, with a finite number of dimensions. [Note: What is ‘indicial’ ?] Then, it makes sense to use the indicial notation

$$\mathbf{d} = \{d^i\} \quad , \quad i \in \mathcal{I}_D \quad ; \quad \mathbf{m} = \{m^\alpha\} \quad , \quad i \in \mathcal{I}_M \quad , \quad (8.217)$$

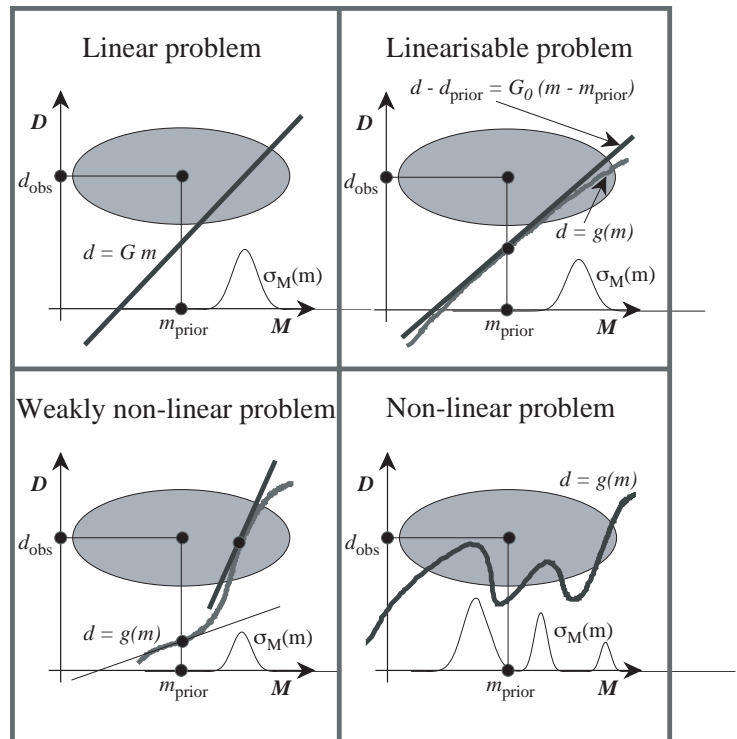
where \mathcal{I}_D and \mathcal{I}_M are two index sets, for the data and the model parameters respectively. In the simplest case, the indices are simple integers, $\mathcal{I}_D = \{1, 2, 3, \dots\}$, and $\mathcal{I}_M = \{1, 2, 3, \dots\}$, but this is not necessarily true. For instance, figure 8.24 suggests a 2D problem where we compute the gravitational field from a distribution of masses. Then, the index α is better understood as consisting on a pair of integers.

Figure 8.24: A simple example where the index in $\mathbf{m} = \{m^\alpha\}$ is not necessarily an integer. In this case, where we are interested in predicting the gravitational field \mathbf{g} generated by a 2-D distribution of mass, the index α is better understood as consisting on a pair of integers. Here, for instance, $m^{A,B}$ means the total mass in the block at row A and column B.



8.3.7.6.2 Linear, Weakly Nonlinear and Nonlinear Problems There are different degrees of nonlinearity. Figure 8.25 illustrates the four domains of nonlinearity allowing the use of the different optimisation algorithms. This figure symbolically represents the model space in the abscissa axis, and the data space in the ordinates axis. The gray oval represents the information coming in part from a priori information on the model parameters and coming in part from the data observations²². It is the function $\rho(\mathbf{d}, \mathbf{m}) = \rho_d(\mathbf{d}) \rho_m(\mathbf{m})$ seen elsewhere (note: say where).

Figure 8.25: Illustration of the four domains of non-linearity allowing the use of the different optimization algorithms. The model space is symbolically represented in the abscissa axis, and the data space in the ordinates axis. The gray oval represents the information coming in part from a priori information on the model parameters and coming in part from the data observations. What is important is not some intrinsic nonlinearity of the function relating model parameters to data, but how linear the function is *inside the domain of significant probability*.



To fix ideas, the oval suggests here a Gaussian probability, but the sorting of problems we are about to make as a function of their nonlinearity will not depend fundamentally on this.

First, there are some strictly linear problems. For instance, in the example illustrated by figure 8.24, the gravitational field \mathbf{g} depends linearly on the masses inside the blocks²³

²²The gray oval is the product of the probability density over the model space, representing the a priori information, times the probability density over the data space representing the experimental results.

²³The gravitational field at point \mathbf{x}_0 generated by a distribution of volumetric mass $\rho(\mathbf{x})$ is given by

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

When the volumetric mass is constant inside some predefined (2-D) volumes, as suggested in figure 8.24, this gives

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

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

Strictly linear problems are illustrated at the top left of figure 8.25. The linear relationship between data and model parameters, $\mathbf{d} = \mathbf{G} \mathbf{m}$, is represented by a straight line. The a priori probability density $\rho(\mathbf{d}, \mathbf{m})$ “induces”, on this straight line, the a posteriori probability density (warning: this notation corresponds to volumetric probabilities) $\sigma(\mathbf{d}, \mathbf{m})$ whose “projection” over the model space gives the a posteriori probability density over the model parameter space, $\sigma_m(\mathbf{m})$. Should the a priori probability densities be Gaussian, then the a posteriori probability distribution would also be Gaussian: this is the simplest situation (in such problems, as we will later see (section xxx), the problem reduces to find the mean and the covariance of the a posteriori Gaussian).

Quasi-linear problems are illustrated at the bottom-left of figure 8.25. If the relationship linking the observable data \mathbf{d} to the model parameters \mathbf{m} ,

$$\mathbf{d} = \mathbf{g}(\mathbf{m}), \quad (8.218)$$

is approximately linear *inside the domain of significant a priori probability* (i.e., inside the gray oval of the figure), then the a posteriori probability is as simple as the a priori probability. For instance, if the a priori probability is Gaussian the a posteriori probability 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” $\mathbf{m}_{\text{prior}}$) (note: explain clearly somewhere in this section that “a priori model” is a language abuse for the “mean a priori model”), 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}_∞ at which the “quasi-Gaussian” $\sigma_m(\mathbf{m})$ is maximum. The linearizations made in order to arrive to \mathbf{m}_∞ are not, so far, an approximation: the point \mathbf{m}_∞ 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,

$$\mathbf{d} - \mathbf{g}(\mathbf{m}_\infty) = \mathbf{G}_\infty (\mathbf{m} - \mathbf{m}_\infty), \quad (8.219)$$

allows to obtain a good approximation of the a posteriori uncertainties. For instance, if the a priori probability is Gaussian this will give the covariance of the “tangent Gaussian”.

Between linear and quasi-linear problem there are the “linearizable problems”. The scheme at the top-right of figure 8.25 shows the case where the linearization of the function $\mathbf{d} = \mathbf{g}(\mathbf{m})$ around the a priori model,

$$\mathbf{d} - \mathbf{g}(\mathbf{m}_{\text{prior}}) = \mathbf{G}_{\text{prior}} (\mathbf{m} - \mathbf{m}_{\text{prior}}), \quad (8.220)$$

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 8.25.

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).

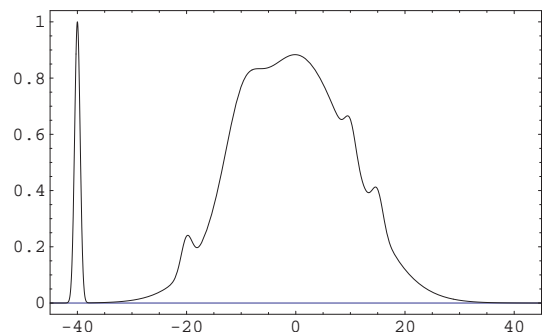
In this case, even if the a priori probability is simple, the a posteriori probability can be quite complicated. For instance, it can be multimodal. [Note: *Rephrase the following sentence...*] These problems are, in general, quite complex to solve, and only the Monte Carlo methods described in the previous chapter are sufficiently general.

If full Monte Carlo methods cannot be used, because they are too expensive, then one can mix some random part (for instance, to choose the starting point) and some 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 (note: explain this better). Repeating these computations for different starting points one can arrive at a good idea of the a posteriori probability in the model space.

8.3.7.6.3 The Maximum Likelihood Model The *most likely model* is, by definition, that at which the volumetric probability $\sigma_\beta(\mathbf{m})$ 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 8.26). 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, in all rigor, corresponds to an infinite number of random variables). Figure XXX, 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 (specimen) 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.

Figure 8.26: 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 at the left of this probability distribution.



8.3.7.6.4 The Interpretation of ‘The Least Squares Solution’ Note: explain here that when working with a large number of dimensions, the center of a Gaussian is a bad representer of the possible realizations of the Gaussian.

Mention somewhere that \mathbf{m}_{post} is not the ‘posterior model’, but the center of the a posteriori Gaussian, and explain that for multidimensional problems, the center of a Gaussian is not representative of a random realisation of the Gaussian.

[note: Mention somewhere that one should not compute the inverse of the matrices, but solve the associated linear system.]

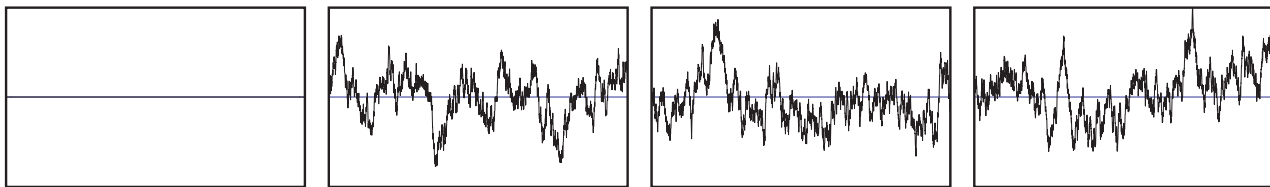


Figure 8.27: At the right, three random realizations of a Gaussian random function with zero mean and (approximately) exponential correlation function. The most likely function, i.e., the center of the Gaussian, is shown at the left. We see that the most likely function is not a representative of the probability distribution.

Chapter 9

Inference Problems of the Fourth Kind (Transport of Probabilities)

Note: Say here the we consider here two problems: (i) the measure of physical quantities — through a direct use of their definition— and (ii) the prediction of observations. It is, of course, our goal to pay attention to the uncertainties involved.

These two problems are mathematically very similar, and are essentially solved using the notion od ‘transport of probabilities’ introduced in chapter 2.

9.1 Measure of Physical Quantities

Note: we develop here a problem that is fundamental in metrology: when a quantity \mathbf{s} is defined as a function of some other quantity \mathbf{r} , through $\mathbf{s} = \mathbf{s}(\mathbf{r})$, and we measure \mathbf{r} , we must ‘transport’ the information we have obtained on \mathbf{r} into information on \mathbf{s} .

Note: give the main ideas here.

The method is illustrated in section 9.1.1 where the Poisson ratio of a solid is evaluated, using its definition in terms of stresses and deformations.

It is also illustrated in appendix 9.3.1, in an example of mass calibration.

9.1.1 Example: Measure of Poisson’s Ratio

9.1.1.1 Hooke’s Law in Isotropic Media

For an elastic medium, in the limit of infinitesimal strains (Hooke’s law),

$$\sigma_{ij} = c_{ijkl} \varepsilon^{kl} \quad , \quad (9.1)$$

where c_{ijkl} is the *stiffness tensor*. If the elastic medium is isotropic,

$$c_{ijkl} = \frac{\lambda_\kappa}{3} g_{ij} g_{kl} + \frac{\lambda_\mu}{2} (g_{ik} g_{jl} + g_{il} g_{jk} - \frac{2}{3} g_{ij} g_{kl}) \quad , \quad (9.2)$$

where λ_κ (with multiplicity one) and λ_μ (with multiplicity five) are the two eigenvalues of the stiffness tensor c_{ijkl} . They are related to the common uncompressibility modulus κ and shear modulus μ through

$$\kappa = \lambda_\kappa/3 \quad ; \quad \mu = \lambda_\mu/2 \quad . \quad (9.3)$$

The Hooke’s law 9.1 can, alternatively, be written

$$\varepsilon_{ij} = d_{ijkl} \sigma^{kl} \quad , \quad (9.4)$$

where d_{ijkl} , the inverse of the stiffness tensor, is called the *compliance tensor*. If the elastic medium is isotropic,

$$d_{ijkl} = \frac{\gamma}{3} g_{ij} g_{kl} + \frac{\varphi}{2} (g_{ik} g_{jl} + g_{il} g_{jk} - \frac{2}{3} g_{ij} g_{kl}) \quad , \quad (9.5)$$

where γ (with multiplicity one) and φ (with multiplicity five) are the two eigenvalues of d_{ijkl} . These are, of course, the inverse of the eigenvalues of c_{ijkl} :

$$\gamma = \frac{1}{\lambda_\kappa} = \frac{1}{3\kappa} \quad ; \quad \varphi = \frac{1}{\lambda_\mu} = \frac{1}{2\mu} \quad . \quad (9.6)$$

From now on, I shall call γ the *eigencompressibility* or, if there is no risk of confusion with $1/\kappa$, the compressibility. The quantity φ shall be called the *eigen-shearability* or, if there is no risk of confusion with $1/\mu$, the shearability.

With the isotropic stiffness tensor of equation 9.2, the Hooke’s law 9.1 becomes

$$\sigma_{ij} = \frac{\lambda_\kappa}{3} g_{ij} \varepsilon_k^k + \lambda_\mu (\varepsilon_{ij} - \frac{1}{3} g_{ij} \varepsilon_k^k) \quad , \quad (9.7)$$

or, equivalently, with the isotropic compliance tensor of equation 9.5, the Hooke’s law 9.4 becomes

$$\varepsilon_{ij} = \frac{\gamma}{3} g_{ij} \sigma_k^k + \varphi (\sigma_{ij} - \frac{1}{3} g_{ij} \sigma_k^k) \quad . \quad (9.8)$$

9.1.1.2 Definition of the Poisson's Ratio

Consider the experimental arrangement of figure 9.1, where an elastic medium is submitted to the (homogeneous) uniaxial stress (using Cartesian coordinates)

$$\sigma_{xx} = \sigma_{yy} = \sigma_{xy} = \sigma_{yz} = \sigma_{zx} = 0 \quad ; \quad \sigma_{zz} \neq 0 \quad . \quad (9.9)$$

Then, the Hooke's law 9.4 predicts the strain

$$\begin{aligned} \varepsilon_{xx} &= \varepsilon_{yy} = \frac{1}{3}(\gamma - \varphi)\sigma_{zz} \\ \varepsilon_{zz} &= \frac{1}{3}(\gamma + 2\varphi)\sigma_{zz} \\ \sigma_{xy} &= \sigma_{yz} = \sigma_{zx} = 0 \quad . \end{aligned} \quad (9.10)$$

The *Young modulus* Y and the *Poisson ratio* ν are defined as

$$Y = \frac{\sigma_{zz}}{\varepsilon_{zz}} \quad ; \quad \nu = -\frac{\varepsilon_{xx}}{\varepsilon_{zz}} = -\frac{\varepsilon_{yy}}{\varepsilon_{zz}} \quad , \quad (9.11)$$

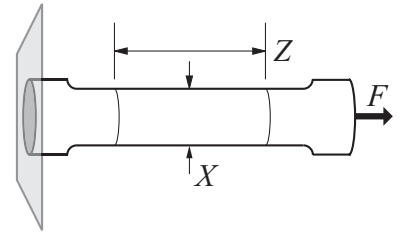
and equation 9.10 gives

$$Y = \frac{3}{2\varphi + \gamma} \quad ; \quad \nu = \frac{\varphi - \gamma}{2\varphi + \gamma} \quad , \quad (9.12)$$

with reciprocal relations

$$\gamma = \frac{1 - 2\nu}{Y} \quad ; \quad \varphi = \frac{1 + \nu}{Y} \quad . \quad (9.13)$$

Figure 9.1: A possible experimental setup for measuring the Young modulus and the Poisson ratio of an elastic medium. The measurement of the force F of the 'bar length' Z and of the bar diameter X allows to estimate the two elastic parameters. Details below.



Note that when γ and φ take values inside their natural range

$$0 < \gamma < \infty \quad ; \quad 0 < \varphi < \infty \quad , \quad (9.14)$$

the variation of Y and ν is

$$0 < Y < \infty \quad ; \quad -1 < \nu < +1/2 \quad . \quad (9.15)$$

Although most materials have positive values of the Poisson ratio ν , there are materials where it is negative (see figures 9.2 and 9.3)

The Poisson ratio has mainly a historical interest. Note that a simple function of it would have given a bona fide Jeffreys quantity,

$$J = \frac{1 + \nu}{1 - 2\nu} = \frac{\lambda_{\kappa}}{\lambda_{\mu}} \quad , \quad (9.16)$$

with the natural domain of variation $0 < J < \infty$.

Figure 9.2: An example of a 2D elastic structure with a positive value of the Poisson ratio. When imposing a stretching in one direction (the ‘horizontal’ here), the elastic structure reacts contracting in the perpendicular direction.

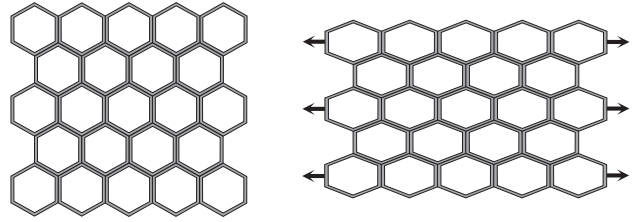
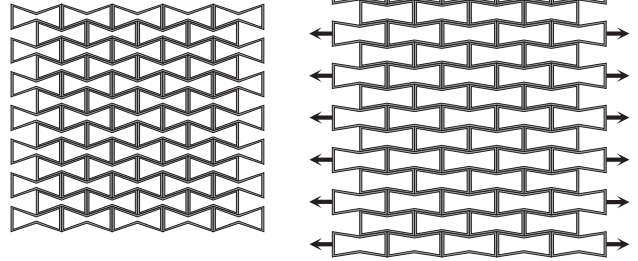


Figure 9.3: An example of a 2D elastic structure with a negative value of the Poisson ratio. When imposing a stretching in one direction (the ‘horizontal’ here), the elastic structure reacts also stretching in the perpendicular direction.



9.1.1.3 The Parameters

Although one may be interested in the Young modulus Y and the Poisson ratio ν , we may choose to measure the compressibility $\gamma = 1/\lambda_\kappa$ and the shearability $\varphi = 1/\lambda_\mu$. Any information we may need on Y and ν can be obtained, as usual, through the change of variables.

From the two first equations in expression 9.10 it follows that the relation between the elastic parameters γ and φ , the stress and the strains is

$$\gamma = \frac{\varepsilon_{zz} + 2\varepsilon_{xx}}{\sigma_{zz}} \quad ; \quad \varphi = \frac{\varepsilon_{zz} - \varepsilon_{xx}}{\sigma_{zz}} \quad . \quad (9.17)$$

As the uniaxial stress is generated by a force F applied to one of the ends of the bar (and the reaction force of the support),

$$\sigma_{zz} = \frac{F}{s} \quad , \quad (9.18)$$

where s , the section of the bar, is

$$s = \frac{\pi X^2}{4} \quad . \quad (9.19)$$

The most general definition of strain (that does not assume the strains to be small) is

$$\varepsilon_{xx} = \log \frac{X}{X_0} \quad ; \quad \varepsilon_{zz} = \log \frac{Z}{Z_0} \quad , \quad (9.20)$$

where X_0 and Z_0 are the initial lengths (see figure 9.1) and X and Z are the final lengths. We have then the final relation

$$\gamma = \frac{\pi X^2 (\log Z/Z_0 + 2 \log X/X_0)}{4 F} \quad ; \quad \varphi = \frac{\pi X^2 (\log Z/Z_0 - \log X/X_0)}{4 F} \quad . \quad (9.21)$$

When necessary, these two expressions shall be written

$$\gamma = \gamma(X_0, Z_0, X, Z, F) \quad ; \quad \varphi = \varphi(X_0, Z_0, X, Z, F) \quad . \quad (9.22)$$

We shall later need to extract from these relations the two parameters X_0 and Z_0 :

$$X_0 = X \exp\left(-\frac{4F(\gamma - \varphi)}{3\pi X^2}\right) \quad ; \quad Z_0 = Z \exp\left(-\frac{4F(\gamma + 2\varphi)}{3\pi X^2}\right) \quad , \quad (9.23)$$

expressions that, when necessary, shall be written

$$X_0 = X_0(\gamma, \varphi, X, Z, F) \quad ; \quad Z_0 = Z_0(\gamma, \varphi, X, Z, F) \quad . \quad (9.24)$$

9.1.1.4 The Partial Derivatives

In what follows, let us use the notation

$$\mathbf{r} = \{X_0, Z_0, X, Z, F\} \quad ; \quad \mathbf{s} = \{\gamma, \varphi\} \quad , \quad (9.25)$$

so the relation 9.21 may be written

$$\mathbf{s} = \mathbf{s}(\mathbf{r}) \quad . \quad (9.26)$$

We need to complete the set of two variables \mathbf{s} to have a set of five variables, as suggested in section 2.6.0.3. The simplest choice is

$$\mathbf{t} = \{X, Z, F\} \quad (9.27)$$

as supplementary variables. We can then introduce the matrix of partial derivatives

$$\mathbf{K} = \begin{pmatrix} \partial\gamma/\partial X_0 & \partial\gamma/\partial Z_0 & \partial\gamma/\partial X & \partial\gamma/\partial Z & \partial\gamma/\partial F \\ \partial\varphi/\partial X_0 & \partial\varphi/\partial Z_0 & \partial\varphi/\partial X & \partial\varphi/\partial Z & \partial\varphi/\partial F \\ \partial X/\partial X_0 & \partial X/\partial Z_0 & \partial X/\partial X & \partial X/\partial Z & \partial X/\partial F \\ \partial Z/\partial X_0 & \partial Z/\partial Z_0 & \partial Z/\partial X & \partial Z/\partial Z & \partial Z/\partial F \\ \partial F/\partial X_0 & \partial F/\partial Z_0 & \partial F/\partial X & \partial F/\partial Z & \partial F/\partial F \end{pmatrix} \quad , \quad (9.28)$$

to easily obtain

$$K = \sqrt{\det \mathbf{K} \mathbf{K}^t} = \frac{3\pi^2 X^4}{16 F^2 X_0 Z_0} \quad . \quad (9.29)$$

9.1.1.5 The Measurement Space and the Measurand Space

We measure the five quantities $\mathbf{r} = \{X_0, Z_0, X, Z, F\}$ in order to evaluate the two quantities $\mathbf{s} = \{\gamma, \varphi\}$. Let us denote by \mathbf{R}_5 the five-dimensional *measurement space*, over which $\mathbf{r} = \{X_0, Z_0, X, Z, F\}$ shall be considered coordinates. The distance element over the measurement space is [note: explain why]

$$ds^2 = \frac{1}{a^2} \left(\left(\frac{dX_0}{X_0} \right)^2 + \left(\frac{dX}{X} \right)^2 + \left(\frac{dZ_0}{Z_0} \right)^2 + \left(\frac{dZ}{Z} \right)^2 \right) + \frac{dF^2}{b^2} \quad , \quad (9.30)$$

where a and b represent arbitrary ‘weights’. We then have the metric determinant

$$\sqrt{\det \mathbf{g}_r} = \frac{k}{X_0 Z_0 X Z} \quad , \quad (9.31)$$

where the constant $k = 1/(a^5 b)$ shall not play any important role in what follows (it will spontaneously disappear).

Similarly, let us denote by \mathbf{S}_2 the two-dimensional *measurand space*, over which $\mathbf{s} = \{\gamma, \varphi\}$ shall be considered coordinates. The distance element over the measurand space is [note: explain why]

$$ds^2 = \frac{1}{c^2} \left(\left(\frac{d\gamma}{\gamma} \right)^2 + \left(\frac{d\varphi}{\varphi} \right)^2 \right) \quad , \tag{9.32}$$

where c represents an arbitrary ‘weight’. The metric matrix is, therefore,

$$\mathbf{g}_r = \frac{1}{c^2} \begin{pmatrix} 1/\gamma^2 & 0 \\ 0 & 1/\varphi^2 \end{pmatrix} \quad , \tag{9.33}$$

and this gives the metric determinant

$$\sqrt{\det \mathbf{g}_s} = \frac{k'}{\gamma \varphi} \quad , \tag{9.34}$$

where the constant $k' = 1/(c^2)$ shall not play any important role in what follows (it will spontaneously disappear).

9.1.1.6 The Measurement

We measure $\{X_0, Z_0, X, Z, F\}$ and describe the result of our measurement via a volumetric probability

$$f_r(X_0, Z_0, X, Z, F) \quad . \tag{9.35}$$

[Note: Explain this.]

9.1.1.7 Transportation of the Probability Distribution

Equation 2.206 applies here directly, and gives the transported volumetric probability over the measurand space. Using the present notations, this gives

$$f_s(\gamma, \varphi) = \frac{1}{\sqrt{\det \mathbf{g}_s}} \int_0^\infty dX \int_0^\infty dZ \int_{-\infty}^{+\infty} dF \underbrace{\frac{\sqrt{\det \mathbf{g}_r}}{K} f_r(X_0, Z_0, X, Z, F)}_{X_0=X_0(\gamma, \varphi, X, Z, F) ; Z_0=Z_0(\gamma, \varphi, X, Z, F)} \quad , \tag{9.36}$$

where the functions $X_0 = X_0(\gamma, \varphi, X, Z, F)$ and $Z_0 = Z_0(\gamma, \varphi, X, Z, F)$ are those expressed by equations 9.23–9.24. More explicitly, using the result for the Jacobian determinant K given by equation 9.29, and the two metric determinants given by equations 9.31 and 9.34,

$f_s(\gamma, \varphi) = \frac{k}{k'} \frac{16}{3 \pi^2} \gamma \varphi \int_0^\infty \frac{dX}{X} \int_0^\infty \frac{dZ}{Z} \int_{-\infty}^{+\infty} \frac{dF}{X^4} \underbrace{f_r(X_0, Z_0, X, Z, F)}_{X_0=X_0(\gamma, \varphi, X, Z, F) ; Z_0=Z_0(\gamma, \varphi, X, Z, F)} \quad .$

(9.37)

The two associated marginal volumetric probabilities are, then,

$$f_\gamma(\gamma) = \int_0^\infty \frac{d\varphi}{\varphi} f_s(\gamma, \varphi) \quad (9.38)$$

and

$$f_\varphi(\varphi) = \int_0^\infty \frac{d\gamma}{\gamma} f_s(\gamma, \varphi) \quad (9.39)$$

To represent these volumetric probabilities I prefer to use the ‘Cartesian parameters’ of the problem [note: explain]. Here, the logarithmic parameters

$$\gamma^* = \log \frac{\gamma}{\gamma_0} \quad \varphi^* = \log \frac{\varphi}{\varphi_0} \quad , \quad (9.40)$$

where γ_0 and φ_0 are two arbitrary constants having the dimension of a compliance are Cartesian coordinates over the 2D space of elastic (isotropic) media. For the distance element of equation 9.32 becomes

$$c^2 ds^2 = (d\gamma^*)^2 + (d\varphi^*)^2 \quad , \quad (9.41)$$

typical of Cartesian coordinates in Euclidean spaces. As volumetric probabilities are invariant quantities, the new volumetric probability function, say $g_s(\gamma^*, \varphi^*)$, is simply given by

$$g_s(\gamma^*, \varphi^*) = f_s(\gamma, \varphi) |_{\gamma = \gamma_0 \exp \gamma^* ; \varphi = \varphi_0 \exp \varphi^*} \quad (9.42)$$

To be complete, let us mention that equations 9.37–9.39 define volumetric probabilities; should we wish to evaluate probability densities,

$$\bar{f}_s(\gamma, \varphi) = \frac{f_s(\gamma, \varphi)}{\gamma \varphi} \quad ; \quad \bar{f}_\gamma(\gamma) = \frac{f_\gamma(\gamma)}{\gamma} \quad ; \quad \bar{f}_\varphi(\varphi) = \frac{f_\varphi(\varphi)}{\varphi} \quad , \quad (9.43)$$

then

$$\bar{f}_s(\gamma, \varphi) = \frac{k}{k'} \frac{16}{3\pi^2} \int_0^\infty \frac{dX}{X} \int_0^\infty \frac{dZ}{Z} \int_{-\infty}^{+\infty} \frac{dF}{X^4} \underbrace{f_r(X_0, Z_0, X, Z, F)}_{X_0=X_0(\gamma, \varphi, X, Z, F) ; Z_0=Z_0(\gamma, \varphi, X, Z, F)} \quad , \quad (9.44)$$

$$\bar{f}_\gamma(\gamma) = \int_0^\infty d\varphi \bar{f}_s(\gamma, \varphi) \quad \text{and} \quad \bar{f}_\varphi(\varphi) = \int_0^\infty d\gamma \bar{f}_s(\gamma, \varphi) \quad (9.45)$$

9.1.1.8 Numerical Illustration

Note: to do things properly, the constants k and k' of equations 9.31 and 9.34 should appear here, as they measures distances. They should all simplify and disappear.

Let us use the notations $N(u, u_0, s)$ and $L(U, U_0, s)$ respectively for the normal and the lognormal functions

$$N(u, u_0, s) = k \exp\left(-\frac{(u - u_0)^2}{2s^2}\right) \quad ; \quad L(U, U_0, s) = k \exp\left(-\frac{1}{2s^2} \left(\log \frac{U}{U_0}\right)^2\right) \quad (9.46)$$

Assume that the result of the measurement of the quantities X_0 , Z_0 (initial diameter and length of the bar), X , Z (final diameter and length of the bar), and the force F , has given an information that can be represented by a five-dimensional Gaussian volumetric probability with independent uncertainties

$$f_r(X, X_0, Z, Z_0, F) = L(X_0, X_0^{\text{obs}}, s_{X_0}) L(Z_0, Z_0^{\text{obs}}, s_{Z_0}) L(X, X^{\text{obs}}, s_X) L(Z, Z^{\text{obs}}, s_Z) N(F, F^{\text{obs}}, s_F) \quad (9.47)$$

with the numerical values

$$\begin{aligned} X_0^{\text{obs}} &= 1.000 \text{ m} & ; & \quad s_{X_0} = 0.015 \\ Z_0^{\text{obs}} &= 1.000 \text{ m} & ; & \quad s_{Z_0} = 0.015 \\ X^{\text{obs}} &= 0.975 \text{ m} & ; & \quad s_X = 0.015 \\ Z^{\text{obs}} &= 1.105 \text{ m} & ; & \quad s_Z = 0.015 \\ F^{\text{obs}} &= 9.81 \text{ kg m/s}^2 & ; & \quad s_F \approx 0 \end{aligned}$$

This is the volumetric probability that appears at the right of equation 9.37. To simplify the example I have assumed that the uncertainty on the force F is much smaller than the other uncertainties, so, in fact, F can be treated as a constant. With the small uncertainties chosen, the lognormal functions in 9.47 look much like a normal one. Figure 9.4 displays the four (marginal) one-dimensional lognormal functions. To illustrate how the uncertainties in the measurement of the lengths propagate into uncertainties in the elastic parameters, I have chosen the quite unrealistic example where the uncertainties in X and X_0 overlap: it is likely that the diameter of the rod has decreased (so the Poisson ratio is positive) but the probability that it has increased (negative Poisson ratio) is significant. In fact, as we shall see, the measurement don't even exclude the virtuality of negative elastic parameters γ and φ (this possibility being excluded by the elastic theory).

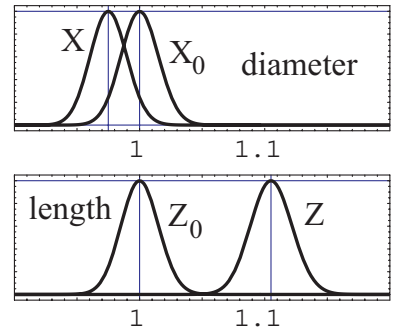


Figure 9.4: The four 1D marginal volumetric probabilities for the initial and final lengths. Note that the uncertainties in X and X_0 overlap: it is likely that the diameter of the rod has decreased (so the Poisson ratio is positive) but the probability that it has increased (negative Poisson ratio) is significant.

Figure 9.5 represents the volumetric probability $f_s(\gamma, \varphi)$ defined by equations 9.37 and 9.42. It represents the information that the measurements of the length has given on the elastic parameters γ and φ . [Note: Explain this better.] [Note: Explain that negative values of γ and φ are excluded 'by hand'].

The two associated marginal volumetric probabilities are defined in equations 9.38–9.39, and are represented in figure 9.6.

Note: mention here figure 9.7.

9.1.1.9 Translation into the Young Modulus and Poisson Ratio Language

To obtain the expression of the metric in the coordinates $\{Y, \nu\}$ one can use the partial derivatives of the old coordinates with respect to the new coordinates, and equation 1.23.

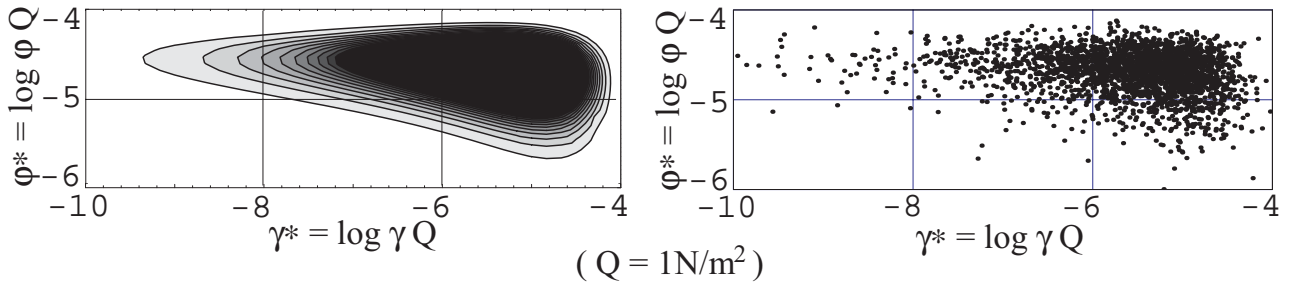


Figure 9.5: The (2D) volumetric probability for the compressibility γ and the shearability ϕ , as induced from the measurement results. At the left a direct representation of the volumetric probability defined by equation 9.37 and 9.42. At the right, a Monte Carlo simulation of the measurement (see section XXX). Here, natural logarithms are used, and $Q = 1 \text{ N/m}^2$. Of the 3000 points used, 9 fell at the left and 7 below the domain plotted, and are not represented. The zone of nonvanishing probability extends over all the space, and only the level lines automatically proposed by the plotting software have been used.

Figure 9.6: The marginal (1D) volumetric probabilities defined by equations 9.38–9.39.

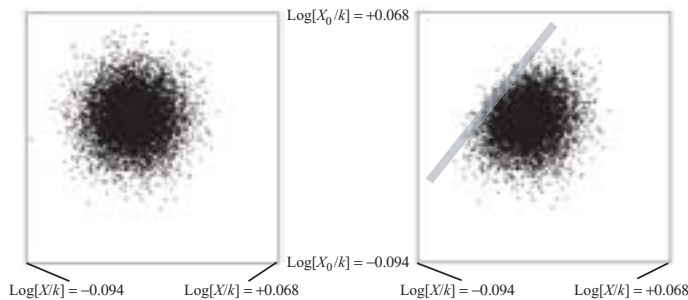
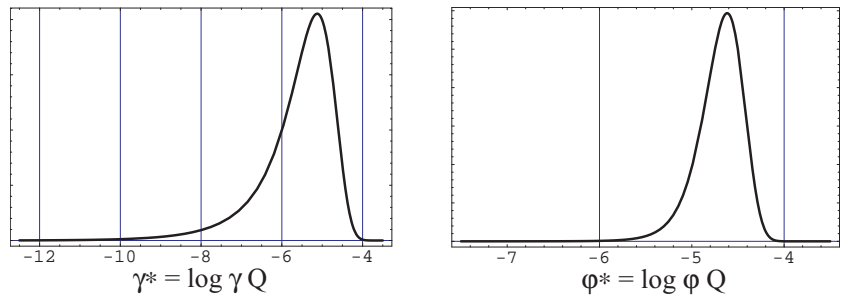


Figure 9.7: The marginal probability distributions for the lengths X and X_0 . At the left, a Monte Carlo sampling of the probability distribution for X as X_0 defined by equation 9.47 (the values Z and Z_0 are also sampled, but are not shown). At the right, the same Monte Carlo sampling, but where only the points that correspond, through equation 9.21, to positive values of γ and ϕ (and, thus, acceptable by the theory of elastic media). Note that many of the points ‘behind’ the diagonal bar have been suppressed.

Then, the metric matrix in equation 9.33, written in the coordinates $\{\gamma, \varphi\}$ becomes

$$\begin{pmatrix} g_{YY} & g_{Y\nu} \\ g_{\nu Y} & g_{\nu\nu} \end{pmatrix} = \begin{pmatrix} \frac{2}{Y^2} & \frac{2}{Y(1-2\nu)} - \frac{1}{Y(1+\nu)} \\ \frac{2}{Y(1-2\nu)} - \frac{1}{Y(1+\nu)} & \frac{4}{(1-2\nu)^2} + \frac{1}{(1+\nu)^2} \end{pmatrix}, \tag{9.48}$$

with the metric determinant being given as

$$\sqrt{\det g} = \frac{3}{Y(1+\nu)(1-2\nu)}. \tag{9.49}$$

To obtain the equivalent of the volumetric probability $f_s(\gamma, \varphi)$ in terms of the Young modulus Y and the Poisson ratio ν we just need to perform the change of variables (remember that volumetric probabilities are invariant under a change of variables), so the volumetric probability $f_s(\gamma, \varphi)$ transforms into a volumetric probability $\psi(Y, \nu)$ that is given by (see relations 9.13)

$$q(Y, \nu) = f_s(\gamma, \varphi)|_{\gamma=\frac{1-2\nu}{Y}, \nu=\frac{1+\nu}{Y}}. \tag{9.50}$$

To evaluate the probability of a domain we have to integrate, in view of equation 9.49, as

$$P(Y_1 < Y < Y_2, \nu_1 < \nu < \nu_2) = \int_{Y_1}^{Y_2} dY \int_{\nu_1}^{\nu_2} d\nu \frac{3}{Y(1+\nu)(1-2\nu)} q(Y, \nu). \tag{9.51}$$

This being said, the question now is: how should we represent the volumetric probability $q(Y, \nu)$? A direct, naïve plot, using Y as an abscissa and ν as ordinate is possible, and only needs the use of equation 9.50 (as the probability density $f_s(\gamma, \varphi)$ has already been evaluated). But let us first use a subtler approach.

We have seen that the quantities γ^* and φ^* (logarithmic compressibility and logarithmic shearability) are Cartesian quantities in the 2D space of linear elastic media. My preferred choice for visualizing $q(Y, \nu)$ is a direct representation of the ‘new coordinates’ on a metrically correct representation, i.e., to superimpose in figure 9.5, where the coordinates γ^* and φ^* were used, the new coordinates $\{Y, \nu\}$ (the change of variables being defined by equations 9.12–9.13). This gives the representation displayed in figure 9.8.

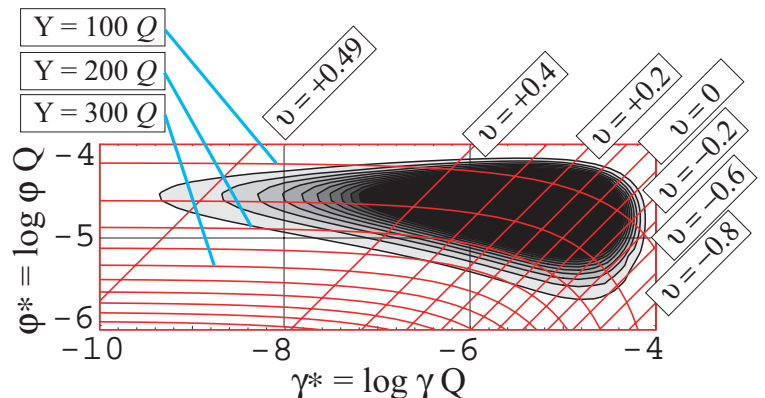
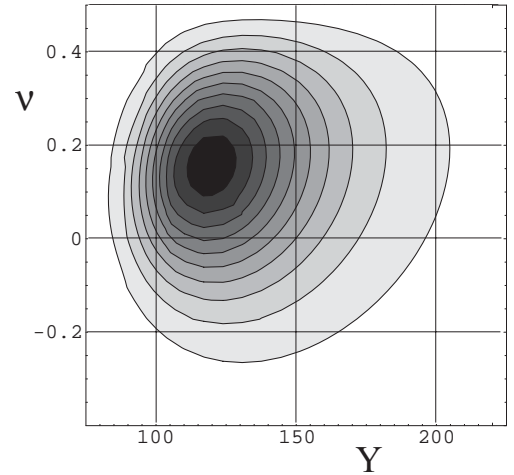


Figure 9.8: The metrically correct representation of the volumetric probability $q(Y, \nu)$, obtained by just superimposing on the figure 9.5 the new coordinates $\{Y, \nu\}$. As above, $Q = 1 \text{ N/m}^2$.

As this is not the conventional way of plotting probability distributions, let us also examine the more conventional plot of $q(Y, \nu)$ in figure 9.9. One may observe, in particular, the ‘round’ character of the ‘level lines’ in this plot, due to the fact that the experiment was specially

Figure 9.9: The volumetric probability for the Young modulus Y and the Poisson ratio ν , deduced, using a change of variables, from the volumetric probability on γ and φ represented in figure 9.5(see equation 9.50).



designed to have a good (and independent) resolution of the Young modulus and the Poisson ratio.

As the metric matrix is not diagonal in the coordinates $\{Y, \nu\}$, one can not define marginal volumetric probabilities, but marginal probability densities only (see section 2.5). We can start by introducing the probability density $\bar{q}(Y, \nu) = \sqrt{\det g} q(Y, \nu)$, i.e.,

$$\bar{q}(Y, \nu) = \frac{3 q(Y, \nu)}{Y (1 + \nu) (1 - 2\nu)} . \quad (9.52)$$

Then, the marginal probability density for the Young modulus is $\bar{q}_Y(Y) = \int_{-1}^{+1/2} d\nu \bar{q}(Y, \nu)$, i.e.,

$$\bar{q}_Y(Y) = \frac{3}{Y} \int_{-1}^{+1/2} d\nu \frac{q(Y, \nu)}{(1 + \nu) (1 - 2\nu)} , \quad (9.53)$$

and the marginal probability density for the Poisson ratio is $\bar{q}_\nu(\nu) = \int_0^\infty dY \bar{q}(Y, \nu)$, i.e.,

$$\bar{q}_\nu(\nu) = \frac{3}{(1 + \nu) (1 - 2\nu)} \int_0^\infty dY \frac{q(Y, \nu)}{Y} . \quad (9.54)$$

Then, the can evaluate probabilities like

$$P(Y_1 < Y < Y_2) = \int_{Y_1}^{Y_2} dY \bar{q}_Y(Y) \quad ; \quad P(\nu_1 < \nu < \nu_2) = \int_{\nu_1}^{\nu_2} d\nu \bar{q}_\nu(\nu) . \quad (9.55)$$

As an example, the marginal probability density for the Poisson ratio, $\bar{q}_\nu(\nu)$, is plotted in figure 9.10.

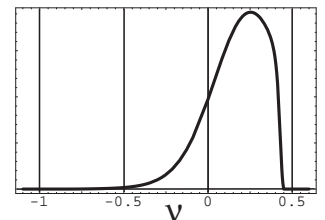


Figure 9.10: The marginal probability density for the Poisson ratio ν (equation 9.54).

9.1.1.10 Direct Evaluation Using Young Modulus and Poisson Ratio

Rather than deducing the volumetric probability for $\{Y, \nu\}$ from that of $\{\gamma, \varphi\}$, we could redo all the computations using directly $\{Y, \nu\}$ as parameters, the only major difference is that the metric matrix 9.48 replaces that in equation 9.33. I leave this as an exercise for the reader.

9.2 Prediction of Observations

This is the typical prediction problem in physics: any serious physical theory is to able to make predictions (that may be confronted to experiments). An engineer, for instance, may wish to predict the load at which a given bridge may collapse, or an astrophysicist may wish to predict the flux of neutrinos from the Sun. In these situations, the parameters defining the system (the bridge or the Sun) may be known with some uncertainties, and these uncertainties shall reflect as an uncertainty on the prediction.

Note: I could use here a notation like

$$\mathbf{d} = \mathbf{d}(\mathbf{p}) \tag{9.56}$$

or like

$$\mathbf{d} = \mathbf{d}(\mathbf{m}) \quad . \tag{9.57}$$

9.3 Appendixes

9.3.1 Appendix: Mass Calibration

Note: I take this problem from *Measurement Uncertainty and the Propagation of Distributions*, by Cox and Harris, 10-th International Metrology Congress, 2001.

When two bodies, with masses m_W and m_R , equilibrate in a balance that operates in air of density a , one has (taking into account Archimedes' buoyancy),

$$\left(1 - \frac{a}{\rho_W}\right) m_W = \left(1 - \frac{a}{\rho_R}\right) m_R \quad , \quad (9.58)$$

where ρ_W and ρ_R are the two volumetric masses of the bodies.

Given a body with mass m , and volumetric mass ρ , it is a common practice in metrology to define its 'conventional mass', denoted m_0 , as the mass of a (hypothetical) body of conventional density $\rho_0 = 8000 \text{ kg/m}^3$ in air of conventional density $a_0 = 1.2 \text{ kg/m}^3$. The equation above then gives the relation

$$\left(1 - \frac{a_0}{\rho_0}\right) m_0 = \left(1 - \frac{a_0}{\rho}\right) m \quad . \quad (9.59)$$

In terms of conventional masses, equation 9.58 becomes

$$\frac{\rho_W - a}{\rho_W - a_0} m_{W,0} = \frac{\rho_R - a}{\rho_R - a_0} m_{R,0} \quad . \quad (9.60)$$

To evaluate the mass $m_{W,0}$ of a body one puts a mass $m_{R,0}$ in the other arm, and selects the (typically small) mass $\delta m_{R,0}$ (with same volumetric mass as $m_{R,0}$) that equilibrates the balance. Replacing $m_{R,0}$ by $m_{R,0} + \delta m_{R,0}$ in the equation above, and solving for $m_{W,0}$ gives

$$m_{W,0} = \frac{(\rho_R - a)(\rho_W - a_0)}{(\rho_W - a)(\rho_R - a_0)} (m_{R,0} + \delta m_{R,0}) \quad . \quad (9.61)$$

The knowledge of the five quantities $\{m_{R,0}, \delta m_{R,0}, a, \rho_W, \rho_R\}$ allows, via equation 9.61, to evaluate $m_{W,0}$. Assume that a measure of these five quantities has provided the information represented by the probability density $f(m_{R,0}, \delta m_{R,0}, a, \rho_W, \rho_R)$. Which is the probability density induced over the quantity $m_{W,0}$ by equation 9.61?

This is just a special case of the transport of probabilities considered in section 2.6.0.3, so we can directly apply here the results of the section. In the five-dimensional 'measurement space' over which the variables $\{m_{R,0}, \delta m_{R,0}, a, \rho_W, \rho_R\}$ can be considered as coordinates, we can change to the variables $\{m_{W,0}, \delta m_{R,0}, a, \rho_W, \rho_R\}$, this defining the matrix \mathbf{K} of partial derivatives (see equation 2.192). One easily arrives at the simple result

$$\sqrt{\det \mathbf{K} \mathbf{K}^t} = \frac{(\rho_R - a)(\rho_W - a_0)}{(\rho_W - a)(\rho_R - a_0)} \quad . \quad (9.62)$$

Because of the change of variables used, we shall also need to express $m_{W,0}$ as a function of $\{m_{R,0}, \delta m_{R,0}, a, \rho_W, \rho_R\}$. From equation 9.61 one immediately obtains

$$m_{R,0} = \frac{(\rho_W - a)(\rho_R - a_0)}{(\rho_R - a)(\rho_W - a_0)} m_{W,0} - \delta m_{R,0} \quad . \quad (9.63)$$

Equation 2.206 gives the probability density for $m_{W,0}$:

$$g(m_{W,0}) = \int d\delta m_{R,0} \int da \int d\rho_W \int d\rho_R \frac{(\rho_W - a)(\rho_R - a_0)}{(\rho_R - a)(\rho_W - a_0)} f(m_{R,0}, \delta m_{R,0}, a, \rho_W, \rho_R) , \quad (9.64)$$

where in $f(m_{R,0}, \delta m_{R,0}, a, \rho_W, \rho_R)$ one has to replace the variable $m_{R,0}$ by its expression as a function of the other five variables, as given by equation 9.63.

Given the probability density $f(m_{R,0}, \delta m_{R,0}, a, \rho_W, \rho_R)$ representing the information obtained though the measurement act, one can try an analytic integration (provided the probability density f has an analytical expression, or it can be approximated by one). More generally, the probability density f can be sampled using the Monte Carlo methods described in section XXX.

This is, in fact, quite trivial here. Let us denote $\mathbf{r} = \{m_{R,0}, \delta m_{R,0}, a, \rho_W, \rho_R\}$ and $s = m_{W,0}$. Then the relation 9.61 can be written formally as $s = s(\mathbf{r})$. One just needs to sample $f(\mathbf{r})$ to obtain points $\mathbf{r}_1, \mathbf{r}_2, \dots$. The points $s_1 = s(\mathbf{r}_1), s_2 = s(\mathbf{r}_2), \dots$ are samples of $g(s)$ (because of the very definition of the notion of transport of probabilities).

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.
- Aki, K., and Richards, P.G., 1980, *Quantitative seismology*, (2 volumes), Freeman and Co.
- Andresen, B., Hoffmann, K. H., Mosegaard, K., Nulton, J. D., Pedersen, J. M., and Salamon, P., On lumped models for thermodynamic properties of simulated annealing problems, *Journal de Physique*, **49**, 1485–1492, 1988.
- 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.
- Bamberger, A., Chavent, G, Hemon, Ch., and Lailly, P., 1982. Inversion of normal incidence seismograms, *Geophysics*, 47, 757-770.
- Ben-Menahem, A., and Singh, S.J., 1981. *Seismic waves and sources*, Springer Verlag.
- Bender, C.M., and Orszag, S.A., 1978. *Advanced mathematical methods for scientists and engineers*, McGraw-Hill.
- Borel, É., 1967, *Probabilités, erreurs*, 14^e éd., Paris.
- Borel, É., dir., 1924–1952, *Traité du calcul des probabilités et de ses applications*, 4 t., Gauthier Villars, Paris.
- Cary, P.W., and C.H. Chapman, Automatic 1-D waveform inversion of marine seismic refraction data, *Geophys. J. R. Astron. Soc.*, **93**, 527–546, 1988.
- Claerbout, J.F., 1971. Toward a unified theory of reflector mapping, *Geophysics*, 36, 467-481.
- Claerbout, J.F., 1976. *Fundamentals of Geophysical data processing*, McGraw Hill.
- Claerbout, J.F., 1985. *Imaging the Earth's interior*, Blackwell Science Publishers.

- Claerbout, J.F., and Muir, F., 1973. Robust modelling with erratic data, *Geophysics*, 38, 5, 826-844.
- 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.
- Davidon, W.C., 1959. Variable metric method for minimization, AEC Res. and Dev., Report ANL-5990 (revised).
- Devaney, A.J., 1984. Geophysical diffraction tomography, *IEEE trans. Geos. remote sensing*, Vol. GE-22, No. 1.
- Djikpéssé, H.A. and Tarantola, A., 1999, Multiparameter ℓ_1 norm waveform fitting: Interpretation of Gulf of Mexico reflection seismograms, *Geophysics*, Vol. 64, No. 4, 1023-1035.
- Evrard, G., 1995, La recherche des paramètres des modèles standard de la cosmologie vue comme un problème inverse, Thèse de Doctorat, Univ. Montpellier.
- Evrard, G., 1966, Objective prior for cosmological parameters, *Proc. of the Maximum Entropy and Bayesian Methods 1995 workshop*, K. Hanson and R. Silver (eds), Kluwer.
- Evrard, G. and P. Coles, 1995. Getting the measure of the flatness problem, *Classical and quantum gravity*, Vol. 12, No. 10, pp. L93-L97.
- Feller, W., *An introduction to probability theory and its applications*, Wiley, N.Y., 1971 (or 1970?).
- 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 Coelestium*.
- Gauthier, O., Virieux, J., and Tarantola, A., 1986. Two-dimensional inversion of seismic waveforms: numerical results, *Geophysics*, 51, 1387-1403.
- 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., Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *Inst. Elect. Electron. Eng. Trans. on pattern analysis and machine intelligence*, **PAMI-6**, 721-741, 1984.
- Goldberg, D.E., *Genetic algorithms in search, optimization, and machine learning* (Addison-Wesley, 1989).
- Hadamard, J., 1902, Sur les problèmes 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.
- Hammersley, J. M., and Handscomb, D.C., Monte Carlo Methods, in *Monographs on Statistics and Applied Probability*, Cox, D. R., and Hinkley, D. V.(eds.), Chapman and Hall, 1964.
- Herman, G.T., 1980. Image reconstruction from projections, the fundamentals of computerized tomography, Academic Press.
- Holland, J.H., *Adaptation in Natural and Artificial Systems*, University of Michigan Press, 1975.

- Ikelle, L.T., Diet, J.P., and Tarantola, A., 1986. Linearized inversion of multi offset seismic reflection data in the f - \mathbf{k} domain, *Geophysics*, **51**, 1266-1276.
- 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.
- Jannane, M., Beydoun, W., Crase, E., Cao Di, Koren, Z., Landa, E., Mendes, M., Pica, A., Noble, M., Röth, G., Singh, S., Snieder, R., Tarantola, A., Trézéguet, D., and Xie, M., Wavelengths of earth structures that can be resolved from seismic reflected data. *Geophysics*, **54**, 906–910, 1988.
- Jaynes, E.T., Prior probabilities, *IEEE Transactions on systems, science, and cybernetics*, Vol. SSC-4, No. 3, 227–241, 1968.
- Jaynes, E.T., 1995, Probability theory: the logic of science, Available on Internet (ftp: bayes.wustl.edu)
- Jaynes, E.T., Where do we go from here?, in Smith, C. R., and Grandy, W. T., Jr., Eds., *Maximum-entropy and Bayesian methods in inverse problems*, Reidel, 1985.
- Jeffreys, H., 1939, Theory of probability, Clarendon Press, Oxford. Reprinted in 1961 by Oxford University Press. **Here he introduces the positive parameters.**
- Johnson, G.R. and Olhoeft, G.R., Density of rocks and minerals, in: CRC Handbook of Physical Properties of rocks, Vol. III, ed: R.S. Carmichael, CRC, Boca Raton, Florida, USA, 1984.
- Journel, A. and Huijbregts, Ch., *Mining Geostatistics*, Academic Press, 1978.
- Kalos, M.H. & Whitlock, P.A., *Monte Carlo methods*, Wiley, N.Y., 1986.
- Kandel, A., 1986, Fuzzy mathematical techniques with applications, Addison-Wesley.
- Keilis-Borok, V.J., and Yanovskaya, T.B., Inverse problems in seismology (structural review), *Geophys. J. R. astr. Soc.*, **13**, 223–234, 1967.
- 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.* (in press).
- Khintchine, A.I., 1969, Introduction à la théorie des probabilités (Elementarnoe vvedenie v teoriyu veroyatnostej), trad. M. Gilliard, 3^e ed., Paris; en anglais: An elementary introduction to the theory of probability, avec B.V., Gnedenko, New York, 1962.
- Kirkpatrick, S., Gelatt, C.D., Jr., and Vecchi, M.P., Optimization by Simulated Annealing, *Science*, **220**, 671–680, 1983.
- Kolmogorov, A.N., 1933, Grundbegriffe der Wahrscheinlichkeitsrechnung, Springer, Berlin; Engl. trans.: Foundations of the theory of probability, Chelsea, New York, 1950.
- Koren, Z., Mosegaard, K., Landa, E., Thore, P., and Tarantola, A., Monte Carlo estimation and resolution analysis of seismic background velocities, *J. Geophys. Res.*, **96**, B12, 20,289–20,299 (1991).
- Kullback, S., 1967, The two concepts of information, *J. Amer. Statist. Assoc.*, **62**, 685–686.
- Landa, E., Beydoun, W., and Tarantola, A., Reference velocity model estimation from prestack waveforms: coherency optimization by simulated annealing, *Geophysics*, **54**, 984–990, 1989.
- Lehtinen, M.S., Päivärinta, L., and Somersalo, E., 1989, Linear inverse problems for generalized random variables, *Inverse Problems*, **5**, 599–612.
- Lions, J.L., 1968. Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles, Dunod, Paris. English translation: Optimal control of systems governed by partial differential equations, Springer, 1971.

- Lütkepohl, H., 1996, Handbook of Matrices, John Wiley & Sons.
- Marroquin, J., Mitter, S., and Poggio, T., 1987, Probabilistic solution of ill-posed problems in computational vision, *Journal of the American Statistical Association*, **82**, 76–89.
- Mehrabadi, M.M., and S.C. Cowin, 1990, Eigentensors of linear anisotropic elastic materials, *Q. J. Mech. appl. Math.*, **43**, 15–41.
- Mehta, M.L., 1967, Random matrices and the statistical theory of energy levels, Academic Press, New York and London.
- Menke, W., 1984, Geophysical data analysis: discrete inverse theory, Academic Press.
- Metropolis, N., and Ulam, S.M., The Monte Carlo Method, *J. Amer. Statist. Assoc.*, 44, 335–341, 1949.
- 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.
- Miller, K.S., 1964, Multidimensional Gaussian distributions, John Wiley and Sons, New York.
- Minster, J.B. and Jordan, T.M., 1978, Present-day plate motions, *J. Geophys. Res.*, **83**, 5331–5354.
- Mohr, P.J., and B.N. Taylor, 2001, The Fundamental Physical Constants, *Physics Today*, Vol. 54, No. 8, BG6–BG13.
- Moritz, H., 1980. Advanced physical geodesy, Herbert Wichmann Verlag, Karlsruhe, Abacus Press, Tunbridge Wells, Kent.
- Morse, P.M., and Feshbach, H., 1953. Methods of theoretical physics, McGraw Hill.
- 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.
- Mosegaard, K. and Vestergaard, P.D., A simulated annealing approach to seismic model optimization with sparse prior information, *Geophysical Prospecting* , **39**, 599–611, 1991.
- Nercessian, Al., Hirn, Al., and Tarantola, Al., 1984. Three-dimensional seismic transmission prospecting of the Mont-Dore volcano, France, *Geophys. J.R. astr. Soc.*, **76**, 307-315.
- Nolet, G., 1985. Solving or resolving inadequate and noisy tomographic systems, *J. Comp. Phys.*, **61**, 463-482.
- Nulton, J.D., and Salamon, P., 1988, Statistical mechanics of combinatorial optimization: *Physical Review A*, **37**, 1351-1356.
- Parker, R.L., 1975. The theory of ideal bodies for gravity interpretation, *Geophys. J. R. astron. Soc.*, **42**, 315-334.
- Parker, R.L., 1977. Understanding inverse theory, *Ann. Rev. Earth Plan. Sci.*, **5**, 35-64.
- Parker, R.L., 1994, Geophysical Inverse Theory, Princeton University Press.
- Pedersen, J.B., and Knudsen, O., Variability of estimated binding parameters, *Biophys. Chemistry*, **36**, 167–176 , 1990.
- Pica, A., Diet, J.P., and Tarantola, A., 1990, Nonlinear inversion of seismic reflection data in a laterally medium, *Geophysics*, Vol. 55, No. 3, pp 284–292.
- Polack, E. et Ribière, G., 1969. Note sur la convergence de méthodes de directions conjuguées, *Revue Fr. Inf. Rech. Oper.*, 16-R1, 35-43.

- Popper, K., Objective knowledge, Oxford, 1972. Trad. franç.: La logique de la découverte scientifique, Payot, Paris, 1978.
- 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.
- Press, F., An introduction to Earth structure and seismotectonics, *Proceedings of the International School of Physics Enrico Fermi*, Course L, Mantle and Core in Planetary Physics, J. Coulomb and M. Caputo (editors), Academic Press, 1971.
- Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T., *Numerical Recipes*, Cambridge, 1986.
- Pugachev, V.S., *Theory of random functions and its application to control problems*, Pergamon, 1965.
- Rényi, A., 1966, Calcul des probabilités, Dunod, Paris.
- Rényi, A., 1970, Probability theory, Elsevier, New York.
- Rietsch, E., The maximum entropy approach to inverse problems, *J. Geophys.*, **42**, 489–506, 1977.
- Rothman, D.H., Nonlinear inversion, statistical mechanics, and residual statics estimation, *Geophysics*, **50**, 2797–2807, 1985.
- Rothman, D.H., Automatic estimation of large residual static corrections, *Geophysics*, **51**, 332–346, 1986.
- 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.
- Scales, J., 1996, Uncertainties in seismic inverse calculations, *in*: Inverse methods, Interdisciplinary elements of methodology, computation, and applications, Eds.: B.H. Jacobsen, K. Mosegaard and P. Sibani, Springer, Berlin, p. 79–97.
- Shannon, C.E., 1948, A mathematical theory of communication, Bell System Tech. J., **27**, 379–423.
- Simon, J.L., 1995, Resampling: the new statistics, Resampling stats Inc., Arlington, VA, USA.
- Stein, S.R., 1985, Frequency and time — their measure and characterization, *in*: Precision frequency control, Vol. 2, edited by E.A. Gerber and A. Ballato, Academic Press, New York, pp. 191–232 and pp. 399–416.
- Tarantola, A., 1984. Linearized inversion of seismic reflection data, *Geophysical Prospecting*, **32**, 998–1015.
- Tarantola, A., 1984. Inversion of seismic reflection data in the acoustic approximation, *Geophysics*, **49**, 1259–1266.
- Tarantola, A., 1984. The seismic reflection inverse problem, *in*: Inverse problems of Acoustic and Elastic Waves, edited by: F. Santosa, Y.-H. Pao, W. Symes, and Ch. Holland, SIAM, Philadelphia.
- 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.
- Tarantola, A., 1987. Inversion of travel time and seismic waveforms, *in*: Seismic tomography, edited by G. Nolet, Reidel.

- Tarantola, A., 1990, Probabilistic foundations of Inverse Theory, in: *Geophysical Tomography*, Desaubies, Y., Tarantola, A., and Zinn-Justin, J., (eds.), North Holland.
- Tarantola, A., Jobert, G., Trézéguet, D., and Denelle, E., 1987. The inversion of seismic waveforms can either be performed by time or by depth extrapolation, submitted to *Geophysics*.
- Tarantola, A. and Nercessian, A., 1984. Three-dimensional inversion without blocks, *Geophys. J. R. astr. Soc.*, 76, 299-306.
- 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.
- Taylor, S.J., 1966, *Introduction to measure and integration*, Cambridge Univ. Press.
- Taylor, A.E., and Lay, D.C., 1980. *Introduction to functional analysis*, Wiley.
- Taylor, B.N., and C.E. Kuyatt, 1994, Guidelines for evaluating and expressing the uncertainty of NIST measurement results, NIST Technical note 1297.
- Watson, G.A., 1980. *Approximation theory and numerical methods*, Wiley.
- Weinberg, S., 1972, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*, John Wiley & Sons.
- Wiggins, R.A., 1969, Monte Carlo Inversion of Body-Wave Observations, *J. Geoph. Res.*, Vol. 74, No. 12, 3171-3181.
- Wiggins, R.A., 1972, 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.
- Winogradzki, J., 1979, *Calcul Tensoriel (I)*, Masson.
- Winogradzki, J., 1987, *Calcul Tensoriel (II)*, Masson.
- Xu, P. and Grafarend, E., 1997, Statistics and geometry of the eigenspectra of 3-D second-rank symmetric random tensors, *Geophys. J. Int.* **127**, 744-756.
- Xu, P., 1999, Spectral theory of constrained second-rank symmetric random tensors, *Geophys. J. Int.* **138**, 1-24.
- Yeganeh-Haeri, A., Weidner, D.J. and Parise, J.B., Elasticity of α -cristobalite: a silicon dioxide with a negative Poisson ratio, *Science*, **257**, 650-652, 1992.