Science from Fisher Information

A Unification

B. Roy Frieden

SCIENCE FROM FISHER INFORMATION

A Unification

B. ROY FRIEDEN

Optical Sciences Center, The University of Arizona, Tucson, AZ

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published as *Physics from Fisher Information* 1998
Reprinted 1999 (twice),
This edition first published 2004

Printed in the United Kingdom at the University Press, Cambridge

Typeface Times 11/14pt. *System* 3b2 [KT]

A catalog record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

Frieden, B. Roy
Science from Fisher information: a unification/B. Roy Frieden/.– [2nd ed].
p. cm.
Rev. edn of: Physics from Fisher information. 1998.
Includes bibliographical references and index.

ISBN 0 521 81079 5 – ISBN 0 521 00911 1 (pbk.)
1. Physical measurements. 2. Information theory. 3. Physics–Methodology.
I. Frieden, B. Roy Physics from Fisher information. II. Title.

QC39.F75 2004
530.8–dc22 2003064021

ISBN 0 521 81079 5 hardback
ISBN 0 521 00911 1 paperback

Contents

0	Introduction	*page* 1
	0.1 Aims of the book	1
	0.2 Level of approach	5
	0.3 Calculus of variations	7
	0.4 Dirac delta function	20
1	What is Fisher information?	23
	1.1 On Lagrangians	24
	1.2 Classical measurement theory	27
	1.3 Comparisons of Fisher information with Shannon's form of entropy	35
	1.4 Relation of I to Kullback–Leibler entropy	37
	1.5 Amplitude form of I	39
	1.6 Efficient estimators	39
	1.7 Fisher I as a measure of system disorder	41
	1.8 Fisher I as an entropy	42
2	Fisher information in a vector world	58
	2.1 Classical measurement of four-vectors	58
	2.2 Optimum, unbiased estimates	60
	2.3 Stam's information	61
	2.4 Physical modifications	62
	2.5 Net probability $p(x)$ for a particle	66
	2.6 The two types of "data" used in the theory	67
	2.7 Alternative scenarios for the channel capacity I	67
	2.8 Multiparameter I-theorem	68
	2.9 Concavity, and mixing property of I	69
3	Extreme physical information	74
	3.1 Covariance, and the "bound" information J	74
	3.2 The equivalence of Boltzmann and Shannon entropies	76

3.3 System information model 80
3.4 Principle of extreme physical information (EPI) 82
3.5 Derivation of Lorentz group of transformations 98
3.6 Gauge covariance property 106
3.7 Field dynamics from information 107
3.8 An optical measurement device 110
3.9 EPI as a state of knowledge 125
3.10 EPI as a physical process 126
3.11 On applications of EPI 128
4 Derivation of relativistic quantum mechanics 131
4.1 Derivation of Klein–Gordon equation 131
4.2 Derivation of vector Dirac equation 144
4.3 Uncertainty principles 153
4.4 Overview 157
5 Classical electrodynamics 163
5.1 Derivation of vector wave equation 163
5.2 Maxwell's equations 185
5.3 Overview 187
6 The Einstein field equation of general relativity 190
6.1 Motivation 190
6.2 Tensor manipulations: an introduction 191
6.3 Derivation of the weak-field wave equation 194
6.4 Einstein field equation and equations of motion 206
6.5 Overview 207
7 Classical statistical physics 209
7.1 Goals 209
7.2 Covariant EPI problem 209
7.3 Boltzmann probability law 212
7.4 Maxwell–Boltzmann velocity law 221
7.5 Fisher information as a bound to entropy increase 230
7.6 Overview 240
8 Power spectral $1/f$ noise 243
8.1 The persistence of $1/f$ noise 243
8.2 Temporal evolution of tone amplitude 245
8.3 Use of EPI principle 247
8.4 Overview 252
9 Physical constants and the $1/x$ probability law 254
9.1 Introduction 254
9.2 Can the constants be viewed as random numbers? 256
9.3 Use of EPI to find the PDF on the constants 256

9.4 Statistical properties of the $1/x$ law 262
9.5 What histogram of numbers do the constants actually obey? 266
9.6 Overview 268
10 Constrained-likelihood quantum measurement theory 271
10.1 Introduction 271
10.2 Measured coordinates 272
10.3 Likelihood law 274
10.4 Instrument noise properties 275
10.5 Final log-likelihood form 275
10.6 EPI variational principle with measurements 276
10.7 Klein–Gordon equation with measurements 276
10.8 On the Dirac equation with measurements 277
10.9 Schrodinger wave equation with measurements 278
10.10 Overview 285
11 Research topics 290
11.1 Scope 290
11.2 Quantum gravity 290
11.3 Nearly incompressible turbulence 301
11.4 Topics in particle physics 308
11.5 On field operators in general 312
12 EPI and entangled realities: the EPR–Bohm experiment 314
12.1 EPR–Bohm experiment 315
12.2 Invariance principles 316
12.3 EPI problem 317
12.4 Use of game corollary to fix the constants A_{ab} 321
12.5 Getting the constants B_{ab}, C_{ab} by orthogonality 323
12.6 Uncertainty in angle 325
12.7 Information J as a measure of entanglement 326
12.8 Information game 327
12.9 Analogous polarization experiment 327
12.10 Discussion: Can EPI determine all unitless constants? 329
12.11 "Active" information and J 330
13 Econophysics, with Raymond J. Hawkins 333
13.1 A trade as a "measurement" 334
13.2 Intrinsic versus actual data values 336
13.3 Incorporating data values into EPI 337
13.4 Evaluating J from the standpoint of the "technical" approach to valuation 338
13.5 Net principle 339
13.6 Formal solutions 340

13.7 Applications 342
13.8 A measure of volatility 352
13.9 Discussion and summary 352
13.10 Glossary 354
13.11 Acknowledgements 355
14 Growth and transport processes 356
14.1 Introduction 356
14.2 General growth law 360
14.3 Measured parameters 363
14.4 The invariance principle 364
14.5 Change coefficients for an ecological system 366
14.6 Change coefficients of genetic growth 368
14.7 Change coefficients in laser resonator 370
14.8 Change coefficients for ideal gas 371
14.9 Change coefficients for replicating molecules 374
14.10 EPI derivation of general growth law 374
14.11 Resulting equations of growth 379
14.12 Other transport equations 380
14.13 Fisher's theorem of partial change 381
14.14 An uncertainty principle of biological growth 382
14.15 On mass extinctions 384
14.16 Entangled realities: why we don't observe them for macroscopic systems 387
14.17 A proposed mechanism for genesis 388
14.18 Discussion and summary 389
15 Cancer growth, with Robert A. Gatenby 392
15.1 Introduction 392
15.2 Cancer as a random growth process 393
15.3 Biologically free and bound information 395
15.4 Cancer growth and an EPI measurement 396
15.5 EPI approach, general considerations 399
15.6 EPI self-consistent solution 400
15.7 Determining the power by minimizing I 404
15.8 Information efficiency κ 405
15.9 Fundamental role played by Fibonacci constant 406
15.10 Predicted uncertainty in the onset time of cancer 407
15.11 Experimental verification 408
15.12 Discussion and summary 408
16 Summing up 414
Appendix A Solutions common to entropy and Fisher I-extremization 433

Appendix B	Cramer–Rao inequalities for vector data	437
Appendix C	Cramer–Rao inequality for an imaginary parameter	441
Appendix D	EPI derivations of Schrödinger wave equation, Newtonian mechanics, and classical virial theorem	445
Appendix E	Factorization of the Klein–Gordon information	452
Appendix F	Evaluation of certain integrals	457
Appendix G	Schrödinger wave equation as a non-relativistic limit	459
Appendix H	Non-uniqueness of potential **A** for finite boundaries	461
Appendix I	Four-dimensional normalization	463
Appendix J	Transfer matrix method	471
Appendix K	Numerov method	473
References		475
Index		484

0

Introduction

0.1 Aims of the book

The *primary aim* of this book is to develop a *theory of measurement* that incorporates the observer into the phenomenon under measurement. By this theory, the observer becomes both a collector of data and an activator of the phenomenon that gives rise to the data. These ideas have probably been best stated by J. A. Wheeler (1990; 1994):

> All things physical are information-theoretic in origin and this is a participatory universe ... Observer participancy gives rise to information; and information gives rise to physics.

The measurement theory that will be presented is largely, in fact, a quantification of these ideas. However, the reader might be surprised to find that the "information" that is used is not the usual Shannon or Boltzmann entropy measures, but one that is relatively unknown to physicists, that of R. A. Fisher.

The measurement theory is simply a description of how Fisher information flows from a physical source effect to a data space. It therefore applies to all scenarios where quantitative data from repeatable experiments may be collected. This describes measurement scenarios of physics but, also, of science in general. The theory of measurement is found to define an analytical procedure for deriving all laws of science. The approach is called EPI, for "extreme physical information."

The *secondary aim* of the book is to show, by example, that most existing laws of science fit within the EPI framework. That is, they can be derived by its use. (Many can of course be derived by other approaches, but, apparently, no other single approach can derive *all* of them.) In this way the EPI approach unifies science under an umbrella of measurement and information. It also leads to new insights into how the laws are interrelated and, more importantly, to new laws and to heretofore unknown *analytical expressions* for physical

Caveat 1: The usual aim of theory is to form mathematical models for physical effects. This is our aim as well. Thus, the EPI approach is limited to deriving the *mathematical expression* of physical effects. It does not form, in some way, the physical effects themselves. The latter are presumed always to exist "out there" in some fixed form.

Caveat 2: One does not get something from nothing, and EPI is no exception to this rule. Certain things must be assumed about the unknown effect. One is *knowledge of a source*. The other is knowledge of an appropriate *invariance principle*. For example, in electromagnetic theory (Chapter 5), the source is the charge-current density, and the invariance principle is the equation of continuity of charge flow. Notice that these two pieces of information do not by themselves imply electromagnetic theory. However, they do when used in tandem with EPI.

In this way, an invariance principle plays an *active* role in deriving a physical law. Note that this is the reverse of its passive role in orthodox approaches to physics, which instead regard the invariance principle as a *derived* property from a *known* law. (Noether's theorem is often used for this purpose.) This is a key distinction between the two approaches, and should be kept in mind during the derivations.

How does one know *what* invariance principle to use in describing a given scenario?

Caveat 3: Each application of EPI relies upon the user's ingenuity. EPI is not a rote procedure. It takes some imagination and resourcefulness to apply. However, experience indicates that every invariance principle that is used with EPI yields a valid physical law. The approach is exhaustive in this respect.

During the same years that quantum mechanics was being developed by Schrödinger (1926) and others, the field of classical measurement theory was being developed by R. A. Fisher (1922) and co-workers (see Fisher Box, 1978, for a personal view of his professional life). According to classical measurement theory, the quality of any measurement(s) may be specified by a form of information that has come to be called Fisher information. Since these formative years, the two fields – quantum mechanics and classical measurement theory – have enjoyed huge success in their respective domains of application. Until recent times it had been presumed that the two fields are distinct and independent.

However, the two fields actually have strong overlap. The thesis of this book is that all physical law, from the Dirac equation to the Maxwell–

Boltzmann velocity dispersion law, may be unified under the umbrella of classical measurement theory. In particular, the information aspect of classical measurement theory – Fisher information – is the key to the unification.

Fisher information is part of an overall theory of physical law called the principle of EPI. The unifying aspect of this principle will be shown by example, i.e., by application to the major fields of physics: quantum mechanics, classical electromagnetic theory, statistical mechanics, gravitational theory, etc. The defining paradigm of each such discipline is a wave equation, a field equation, or a distribution function of some sort. These will be derived by use of the EPI principle. A separate chapter is devoted to each such derivation. New effects are found, as well, by the information approach.

Such a unification is, perhaps, long overdue. Physics is often considered the science of measurement. That is, physics is a quantification of *observed* phenomena, and observed phenomena contain noise, or fluctuations. The physical paradigm equations (mentioned above) define the fluctuations or errors from ideal values that occur in such observations. That is, *the physics lies in the fluctuations*. On the other hand, classical Fisher information is a scalar measure of these very physical fluctuations. In this way, Fisher information is intrinsically tied into the laws of fluctuation that define theoretical physics.

EPI theory proposes that all physical theory results from observation: in particular, *imperfect* observation. Thus, EPI is an observer-based theory of physics. We are used to the concept of an imperfect observer in addressing quantum theory, but the imperfect observer does not seem to be terribly important to classical electromagnetic theory, for example, where it is assumed (wrongly) that fields are known exactly. The same comment can be made about the gravitational field of general relativity. What we will show is that, by admitting that any observation is imperfect, one can derive both the Maxwell equations of electromagnetic theory and the Einstein field equations of gravitational theory. The EPI view of these equations is that they are expressions of fluctuation in the values of measured field positions. Hence, the four-positions $(\boldsymbol{r}, t)$ in Maxwell's equations represent, in the EPI interpretation, random excursions from an ideal, or mean, four-position over the field.

Dispensing with the artificiality of an "ideal" observer allows us to reap many benefits for purposes of *understanding* physics. EPI is, more precisely, an expression of the "inability to know" a measured quantity. For example, EPI derives quantum mechanics from the viewpoint that an ideal position cannot be known. We have found, from teaching the material in this book, that students more easily understand quantum mechanics from this viewpoint than from the conventional viewpoint of derivative operators that somehow represent energy or momentum. Furthermore, that *the same* inability to know also leads to the

Maxwell equations when applied to that scenario is even more satisfying. It is, after all, a human desire to find common cause in the phenomena we see.

Unification is also, of course, the major aim of physics, although EPI is probably not the ultimate unification that many physicists seek. Our aim is to propose a *comprehensive* approach to deriving physical laws, based upon a new theory of measurement. Currently, the approach presumes the existence of sources and particles. EPI derives major classes of particles, but not all of them, and does not derive the sources. (See Caveat 2 preceding.) We believe, however, that EPI is a large step in the right direction. Given its successes so far, the sources and remaining particles should eventually follow from these considerations as well.

At this point we want to emphasize *what this book is not about*. This is not a book whose primary emphasis is upon the *ad hoc* construction of Lagrangians and their extremization. That is a well-plowed field. Although we often derive a physical law via the extremization of a Lagrangian integral, the information viewpoint we take leads to other types of solutions as well. Some solutions arise, for example, out of *zeroing* the integral. (See the derivation of the Dirac equation in Chapter 4.) Other laws arise out of a combination of both zeroing and extremizing the integral. Similar remarks may be made about the process by which the Lagrangians are *formed*. The zeroing and extremizing operations actually allow us to *solve for* the Lagrangians of the scenarios (see Chaps. 4–9, and 11). In this way we avoid, to a large degree, the *ad hoc* approach to Lagrange construction that is conventionally taken. This subject is discussed further in Secs. 1.1 and 1.8.8. The rationale for both zeroing and extremizing the integral is developed in Chapter 3. It is one of *information transfer* from phenomenon to data.

The layout of the book is, very briefly, as follows. The current chapter is intended to derive and exemplify mathematical techniques that the reader might not be familiar with. Chapter 1 is an introduction to the concept of Fisher information. This is for single-parameter estimation problems. Chapter 2 generalizes the concept to multidimensional estimation problems, ending with the scalar information form I that will be used thereafter in the applications Chapters 4–11. Chapter 3 introduces the concept of the "bound information" J, leading to the principle of EPI. This is derived from various points of view. Chapters 4–15 apply EPI to various measurement scenarios, in this way deriving the fundamental wave equations and distribution functions of science. Chapter 16 is a chapter-by-chapter summary of the key points made in the development. The reader in a hurry might choose to read this first, to get an idea of the scope of the approach and the phenomena covered.

0.2 Level of approach

The level of physics and mathematics that the reader is presumed to have is that of a senior undergraduate in physics. Calculus, through partial differential equations, and introductory matrix theory are presumed parts of his/her background. Some notions from elementary probability theory are also used. However, since these are intuitive in nature, the appropriate formula is usually just given, with reference to a suitable text as needed.

A cursory scan through the chapters will show that a minimal amount of prior knowledge of physical theory is actually used or needed. In fact, *this is the nature of the information approach taken* and is one of its strengths. The main physical input to each application of the approach is a simple law of invariance that is obeyed by the given phenomenon.

The overall mathematical notation that is used is that of conventional calculus, with additional matrix and vector notation as needed. Tensor notation is only used where it is a "must" – in Chaps. 6 and 11 on classical and quantum relativity, respectively. No extensive operator notation is used; this author believes that specialized notation often hinders comprehension more than it helps the student to understand theory. Sophistication *without* comprehension is definitely not our aim.

A major step of the information principle is the extremization and/or zeroing of a scalar integral. The integral has the form

$$K \equiv \int d\mathbf{x}\, \mathscr{L}[\mathbf{q}, \mathbf{q}', \mathbf{x}], \quad \mathbf{x} \equiv (x_1, \ldots, x_M), \quad d\mathbf{x} \equiv dx_1 \cdots dx_M, \quad \mathbf{q}, \mathbf{x} \text{ real},$$

$$\mathbf{q} \equiv (q_1, \ldots, q_N), \quad q_n \equiv q_n(\mathbf{x}),$$

$$\mathbf{q}'(\mathbf{x}) \equiv \partial q_1/\partial x_1, \partial q_1/\partial x_2, \ldots, \partial q_N/\partial x_M. \tag{0.1}$$

Mathematically, $K \equiv K[\mathbf{q}(\mathbf{x})]$ is a "functional," i.e., a single number that depends upon the values of one or more functions $\mathbf{q}(\mathbf{x})$ continuously over the domain of $\mathbf{x}$. Physically, K has the form of an "action" integral, whose extremization has conventionally been used to derive fundamental laws of physics (Morse and Feshbach, 1953). Statistically, we will find that K is the "physical information" of an overall system consisting of a measurer and a measured quantity. The limits of the integral are fixed and, usually, infinite. The dimension M of $\mathbf{x}$-space is usually 4 (space-time). The functions q_n of $\mathbf{x}$ are probability amplitudes, i.e., functions whose squares are probability densities. The q_n are to be found. They specify the physics of a measurement scenario. Quantity $\mathscr{L}$ is a known function of the q_n, their derivatives with respect to all the x_m, and $\mathbf{x}$. $\mathscr{L}$ is called the "Lagrangian" density (Lagrange, 1788). It also takes on the role of an information density, by our statistical interpretation.

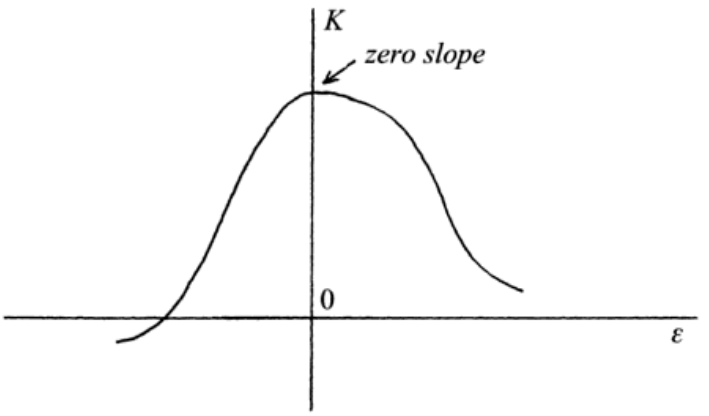

Fig. 0.2. K as a function of perturbation size parameter ε.

with ε a finite number and $\eta(x)$ any perturbing function. Any function $q_\varepsilon(x, \varepsilon)$ must pass through the endpoints so that, from Eq. (0.3),

$$\eta(a) = \eta(b) = 0. \tag{0.4}$$

Equation (0.2) is, with this representation $q_\varepsilon(x, \varepsilon)$ for $q(x)$,

$$K = \int_a^b dx\, \mathscr{L}[x, q_\varepsilon(x, \varepsilon), q'_\varepsilon(x, \varepsilon)] \equiv K(\varepsilon), \tag{0.5}$$

a function of the small parameter ε. (Once x has been integrated out, only the ε-dependence remains.)

We use ordinary calculus to find the solution. By the construction (0.3), $K(\varepsilon)$ attains the extremum value when $\varepsilon = 0$. Since an extremum value is attained there, $K(\varepsilon)$ must have zero slope at $\varepsilon = 0$ as well. That is,

$$\left.\frac{\partial K}{\partial \varepsilon}\right|_{\varepsilon=0} = 0. \tag{0.6}$$

The situation is sketched in Fig. 0.2.

We may evaluate the left-hand side of Eq. (0.6). By Eq. (0.5), $\mathscr{L}$ depends upon ε only through quantities q and q'. Therefore, differentiating Eq. (0.5) gives

$$\frac{\partial K}{\partial \varepsilon} = \int_a^b dx \left[\frac{\partial \mathscr{L}}{\partial q_\varepsilon}\frac{\partial q_\varepsilon}{\partial \varepsilon} + \frac{\partial \mathscr{L}}{\partial q'_\varepsilon}\frac{\partial q'_\varepsilon}{\partial \varepsilon}\right]. \tag{0.7}$$

The second integral is

$$\int_a^b dx\, \frac{\partial \mathscr{L}}{\partial q'_\varepsilon}\frac{\partial^2 q_\varepsilon}{\partial x\, \partial \varepsilon} = \left.\frac{\partial \mathscr{L}}{\partial q'_\varepsilon}\frac{\partial q_\varepsilon}{\partial \varepsilon}\right|_a^b - \int_a^b \frac{\partial q_\varepsilon}{\partial \varepsilon}\frac{d}{dx}\left(\frac{\partial \mathscr{L}}{\partial q'_\varepsilon}\right) dx \tag{0.8}$$

same way, whereas the above approach (0.14), (0.15) is not. Instead, the EPI approach will be used to derive the more general Einstein field equation, from which Newton's law follows as a special case (the weak-field limit). Or, see Appendix D.

The reader may well question where this particular Lagrangian came from. The answer is that it was chosen merely because it "works," i.e., leads to Newton's law of motion. It has no prior significance in its own right. This has been a well-known drawback to the use of Lagrangians. The next chapter addresses this problem in detail.

Example 2: What is the shortest path between two points in a plane? The integrated arc length between points $x = a$ and $x = b$ is

$$K = \int_a^b dx\, \mathscr{L}, \quad \mathscr{L} = \sqrt{1 + q'^2}. \tag{0.16}$$

Hence

$$\frac{\partial \mathscr{L}}{\partial q'} = \tfrac{1}{2}(1 + q'^2)^{-1/2} 2q', \quad \frac{\partial \mathscr{L}}{\partial q} = 0 \tag{0.17}$$

here, so that the Euler–Lagrange Eq. (0.13) is

$$\frac{d}{dx}\left(\frac{q'}{\sqrt{1 + q'^2}}\right) = 0. \tag{0.18}$$

The immediate solution is

$$\frac{q'}{\sqrt{1 + q'^2}} = const., \tag{0.19}$$

implying that $q' = const.$, so that $q(x) = Ax + B$, with A, $B = const.$, the equation of a straight line. Hence we have shown that the path of extreme (not necessarily shortest) distance between two fixed points in a plane is a straight line. We will show below that the extremum is a minimum, as intuition suggests.

Example 3: Maximum entropy problems (Jaynes, 1957a; 1957b) have the form

$$\int dx\, \mathscr{L} = max., \quad \mathscr{L} = -p(x) \ln p(x) + \lambda p(x) + \mu p(x) f(x) \tag{0.20}$$

with λ, μ constants and $f(x)$ a known "kernel" function. The first term in the integral defines the "entropy" of a probability density function (PDF) $p(x)$. (Notice we use the notation p in place of q here.) We will say a lot more about the concept of entropy in chapters to follow. Directly

$$\frac{\partial \mathscr{L}}{\partial p'} = 0, \quad \frac{\partial \mathscr{L}}{\partial p} = -1 - \ln p + \lambda + \mu f(x). \tag{0.21}$$

Hence the Euler–Lagrange Eq. (0.13) is

$$-1 - \ln p(x) + \lambda + \mu f(x) = 0, \quad \text{or} \quad p(x) = A \exp[\mu f(x)]. \tag{0.22}$$

The answer $p(x)$ to maximum entropy problems is always of an exponential form. We will show below that the extremum obtained is actually a maximum, as required.

Example 4 Minimum Fisher information problems (Huber, 1981) are of the form

$$\int dx\, \mathscr{L} = min., \quad \mathscr{L} = 4q'^2 + \lambda q(x) f(x) + \mu q^2(x) h(x), \tag{0.23}$$

λ, $\mu = const.$, where $f(x)$, $h(x)$ are known kernel functions. Also, the PDF $p(x) = q^2(x)$, i.e., $q(x)$ is a "probability amplitude" function. The first term in the integral defines the Fisher information. Directly

$$\frac{\partial \mathscr{L}}{\partial q'} = 8q', \quad \frac{\partial \mathscr{L}}{\partial q} = \lambda f(x) + 2\mu q(x) h(x). \tag{0.24}$$

The Euler–Lagrange Eq. (0.13) is then

$$q''(x) - (\mu/4) h(x) q(x) - (\lambda/8) f(x) = 0. \tag{0.25}$$

That is, the answer $q(x)$ is the solution to a second-order differential equation. The particular solution will depend upon the form of the kernel functions and on any imposed boundary conditions. We will show below that the extremum obtained is a minimum, as required.

In comparing the maximum entropy solution (0.22) with the minimum Fisher information solution (0.25) it is to be noted that the former has the virtue of simplicity, always being an exponential. By contrast, obtaining the Fisher information solution always requires solving a differential equation: a bit more complicated a procedure. However, for purposes of deriving physical PDF laws the Fisher answer is actually preferred: the PDFs of physics generally obey differential equations (wave equations). We will further address this issue in later chapters.

We now proceed to generalize the variational problem (0.2) by degrees. For brevity, only the solutions (Korn and Korn, 1968) will be presented.

0.3.2 Multiple-curve problems

As a generalization of problem (0.2) with its single unknown function $q(x)$, consider the problem of finding N functions $q_n(x)$, $n = 1, \ldots, N$ that satisfy

$$K = \int dx\, \mathscr{L}[x, q_1, \ldots, q_N, q_1', \ldots, q_N'] = extrem. \tag{0.26}$$

The answer to this variational problem is that the $q_n(x)$, $n = 1, \ldots, N$ must obey N Euler–Lagrange equations,

$$\frac{d}{dx}\left(\frac{\partial \mathscr{L}}{\partial q_n'}\right) = \frac{\partial \mathscr{L}}{\partial q_n}, \quad n = 1, \ldots, N. \tag{0.27}$$

In the case $N = 1$ this becomes the one-function result (0.13).

0.3.3 Condition for a minimum solution

At this point we cannot know whether the solution (0.27) to the extremum problem (0.26) gives a maximum or a minimum value for the extremum. A simple test for this purpose is as follows.

Consider the matrix of numbers $[\partial^2 \mathscr{L}/\partial q_i' \partial q_j']$. If this matrix is positive definite then the extreme value is a minimum; or, if it is negative definite, the extreme value is a maximum. This is called *Legendre's condition* for an extremum.

A particular case of interest is as follows.

0.3.4 Fisher information, multiple-component case

As will be shown in Chapter 2, the information Lagrangian is here

$$\mathscr{L} = 4 \sum_{n=1}^{N} q_n'^2. \tag{0.28}$$

Then

$$\frac{\partial \mathscr{L}}{\partial q_i'} = 8q_i', \quad \text{so that} \quad \frac{\partial^2 \mathscr{L}}{\partial q_i' \partial q_j'} = 8\delta_{ij} \tag{0.29}$$

where δ_{ij} is the Kronecker delta function. Thus the matrix $[\partial^2 \mathscr{L}/\partial q_i' \partial q_j']$ is diag $[8, \ldots, 8]$ so that all its n-row minor determinants obey

$$\det\left[\frac{\partial^2 \mathscr{L}}{\partial q_i' \partial q_j'}\right] = 8^n > 0, \; n = 1, \ldots, N. \tag{0.30}$$

Then the matrix $[\partial^2 \mathscr{L}/\partial q_i' \partial q_j']$ is positive definite (Korn and Korn, 1968, p. 420). Consequently, by Legendre's condition the extremum is a minimum.

0.3.5 Exercise

Using the Lagrangian given below Eq. (0.2), show by Legendre's condition

(Sec. 0.3.3) that the Newton's law solution (0.15) minimizes the corresponding integral K in Eq. (0.2).

0.3.6 Nature of extremum in other examples

Return to Example 2, the problem of the *minimum* path between two points. The solution $q(x) = Ax + B$ guarantees an extremum but not necessarily a minimum. Differentiating the first Eq. (0.17) gives, after some algebra,

$$\frac{\partial^2 \mathscr{L}}{\partial q'^2} = \frac{1}{(1+q'^2)^{3/2}} = \frac{1}{(1+A^2)^{3/2}} > 0, \tag{0.31}$$

signifying a minimum by Legendre's condition.

Return to Example 3, maximum entropy solutions. The exponential solution (0.22) guarantees an extreme value to the integral (0.20) but not necessarily a maximum. We attempt to use the Legendre condition. However, the Lagrangian (0.20) does not contain any dependence upon quantity $p'(x)$. Hence Legendre's rule gives the ambiguous result $\partial^2 \mathscr{L}/\partial p'^2 = 0$. This being neither positive nor negative, the nature of the extremum remains unknown.

We need to approach the problem in a different way. Temporarily replace the continuous integral (0.20) by a sum

$$K = \sum_{n=1}^{N} \Delta x\,(-p_n \ln p_n + \lambda p_n + \mu p_n f_n) = max.,$$
$$p_n \equiv p(x_n), \quad f_n \equiv f(x_n), \quad x_n \equiv n\,\Delta x, \tag{0.32}$$

where $\Delta x > 0$ is small but finite. The sum approaches the integral as $\Delta x \to 0$. K is now an ordinary function (not a functional) of the N probabilities p_n and, hence, may be extremized in each p_n using the ordinary rules of differential calculus. The nature of such an extremum may be established by observing the positive- or negative-definiteness of the second derivative matrix $[\partial^2 K/\partial p_i\,\partial p_j]$. From Eq. (0.32) we have directly $[\partial^2 K/\partial p_i\,\partial p_j] = -\Delta x\,\mathrm{diag}[1/p_1, \ldots, 1/p_N]$. Hence, all its n-row minor determinants obey

$$\det\left[\frac{\partial^2 K}{\partial p_i\,\partial p_j}\right] = -\prod_{i=1}^{n} \frac{\Delta x}{p_i} < 0 \tag{0.33}$$

since all probabilities $p_i \geqslant 0$. The matrix is negative definite, signifying a maximum as required. This result obviously holds in the (continuous) limit as $\Delta x \to 0$ through positive values.

$$\overline{\mathscr{L}} = \mathscr{L} + \sum_{nk} \lambda_{nk} q_n^{\alpha}(\mathbf{x}) f_k(\mathbf{x}). \tag{0.38}$$

That is, in Eq. (0.36), the terms in F_{nk} do not contribute to the Euler–Lagrange solution (0.34). Hence, the problem (0.36) may be re-posed more simply as

$$K + \sum_{nk} \lambda_{nk} \int d\mathbf{x}\, q_n^{\alpha}(\mathbf{x}) f_k(\mathbf{x}) = extrem. \tag{0.39}$$

That is, to incorporate constraints one merely weights and adds them to the "objective" functional K. Some examples of interest are as follows.

With K as the entropy functional, and with moment constraints, the approach (0.39) was used by Jaynes (1957a; 1957b) to estimate PDFs represented as $q^2(x)$.

Alternatively, with K as the Fisher information and with a constraint of mean kinetic energy, the approach (0.39) was used to derive the Schrödinger wave equation (Frieden, 1989) and other wave equations of physics (Frieden, 1990). Historically, the former was the author's first application of a principle of minimum Fisher information to a physical problem. The questions it raised, such as why *a priori* mean kinetic energy should be a constraint (ultimate answer: most generally it shouldn't), provoked an evolution of the theory which has culminated in this book.

0.3.9 Variational derivative, functional derivatives

The variation of a functional, the variational derivative, and the functional derivative are useful concepts that follow easily from the preceding. We shall have occasion to use the concept of the functional derivative later on. The concept of the variational derivative is also given, mainly so as to distinguish it from the functional variety.

We first define the concept of the variation of a functional. It was noted that K is a functional (Sec. 0.2). Multiply Eq. (0.12) through by a differential $d\varepsilon$. This gives

$$\left.\frac{\partial K}{\partial \varepsilon}\right|_{\varepsilon=0} d\varepsilon = \int_a^b \left[\frac{\partial \mathscr{L}}{\partial q} - \frac{d}{dx}\left(\frac{\partial \mathscr{L}}{\partial q'}\right)\right]\left(\frac{\partial q}{\partial \varepsilon}\right)\Bigg|_{\varepsilon=0} d\varepsilon\, dx. \tag{0.40}$$

We also used Eq. (0.9). Define the variation of K as

$$\delta K \equiv \left(\frac{\partial K}{\partial \varepsilon}\right)\Bigg|_{\varepsilon=0} d\varepsilon. \tag{0.41}$$

This measures the change in functional K due to a small perturbation away from the stationary solution $q(x)$. Similarly

where $q(x, \varepsilon)$ obeys Eq. (0.3). By how much does the scalar value K change if function $q(x, \varepsilon)$ is perturbed by a small amount at *each* x?

In order to use the ordinary rules of calculus we first subdivide x-space as

$$x_{n+1} = x_n + \Delta x, \quad n = 1, 2, \ldots \tag{0.46}$$

in terms of which

$$K = K[x_1, x_2, \ldots, q(x_1, \varepsilon), q(x_2, \varepsilon), \ldots] \tag{0.47}$$

(cf. Eq. (0.45)). Also, Eq. (0.3) is now discretized, to

$$q(x_n, \varepsilon) = q(x_n) + \varepsilon\eta(x_n). \tag{0.48}$$

Note that the ordinary partial derivatives $\partial K/\partial q(x_n, \varepsilon)$, $n = 1, 2, \ldots$ are well defined, by direct differentiation of Eq. (0.47).

In all of the preceding, the numbers $\eta(x_n)$ are presumed to be arbitrary but *fixed*. Then the perturbations in (0.48) are purely a function of ε. Consequently, K given by (0.47) is likewise purely a function $K(\varepsilon)$.

Next, consider the effect upon $K(\varepsilon)$ of a small change $d\varepsilon$ away from the stationary state $\varepsilon = 0$. This may simply be represented by a Taylor series in powers of $d\varepsilon$,

$$K(d\varepsilon) = dK(\varepsilon)\Big|_{\varepsilon=0} = \frac{\partial K}{\partial \varepsilon}\bigg|_{\varepsilon=0} d\varepsilon + \frac{1}{2}\frac{\partial^2 K}{\partial \varepsilon^2}\bigg|_{\varepsilon=0} d\varepsilon^2 + \cdots. \tag{0.49}$$

The coefficients of $d\varepsilon$ and $d\varepsilon^2$ are evaluated as follows.

By the chain rule of differentiation

$$\frac{\partial K}{\partial \varepsilon} = \sum_n \frac{\partial K}{\partial q(x_n)} \frac{\partial q(x_n)}{\partial \varepsilon}. \tag{0.50}$$

(For brevity, we used $q(x_n, \varepsilon) \equiv q(x_n)$.) Also,

$$\frac{\partial^2 K}{\partial \varepsilon^2} \equiv \frac{\partial}{\partial \varepsilon}\left(\frac{\partial K}{\partial \varepsilon}\right) = \sum_m \frac{\partial}{\partial q(x_m)}\left[\sum_n \frac{\partial K}{\partial q(x_n)} \frac{\partial q(x_n)}{\partial \varepsilon}\right] \frac{\partial q(x_m)}{\partial \varepsilon}$$

after re-use of (0.50),

$$= \sum_{mn} \frac{\partial^2 K}{\partial q(x_m)\, \partial q(x_n)} \frac{\partial q(x_m)}{\partial \varepsilon} \frac{\partial q(x_n)}{\partial \varepsilon} \tag{0.51}$$

after another derivative term drops out.

If we multiply Eq. (0.50) by $d\varepsilon$ and Eq. (0.51) by $d\varepsilon^2$, evaluate them at $\varepsilon = 0$, and use definitions (0.41) and (0.42), we get

$$\frac{\partial K}{\partial \varepsilon}\bigg|_{\varepsilon=0} d\varepsilon = \sum_n \frac{\partial K}{\partial q(x_n)}\, \delta q(x_n)$$

and (0.52)

$$\left.\frac{\partial^2 K}{\partial \varepsilon^2}\right|_{\varepsilon=0} d\varepsilon^2 = \sum_{mn} \frac{\partial^2 K}{\partial q(x_m)\,\partial q(x_n)}\,\delta q(x_m)\,\delta q(x_n),$$

respectively.

Following the plan, we substitute these coefficient values into Eq. (0.49). Next, multiply and divide the first sum by Δx and the second sum by Δx^2. Finally, take the continuous limit $\Delta x \to 0$. The sums approach integrals and we have

$$\left.dK(\varepsilon)\right|_{\varepsilon=0} = \int dx \left[\lim_{\Delta x\to 0} \frac{1}{\Delta x}\,\frac{\partial K}{\partial q(x_n)}\right]\delta q(x) + \frac{1}{2}\iint dx'\,dx \left[\lim_{\Delta x\to 0} \frac{1}{\Delta x^2}\,\frac{\partial^2 K}{\partial q(x_m)\,\partial q(x_n)}\right]\delta q(x')\,\delta q(x) + \cdots \tag{0.53}$$

where $x_n \to x$, $x_m \to x'$ in the limit. We demand that this take the simpler form (cf. Eq. (0.43))

$$\left.dK(\varepsilon)\right|_{\varepsilon=0} = \int dx\,\frac{\delta K}{\delta q(x)}\,\delta q(x) + \frac{1}{2}\iint dx'\,dx\,\frac{\delta^2 K}{\delta q(x')\,\delta q(x)}\,\delta q(x')\,\delta q(x) + \cdots. \tag{0.54}$$

By Eq. (0.53) this will be so if we define

$$\frac{\delta K}{\delta q(x)} \equiv \lim_{\Delta x\to 0} \frac{1}{\Delta x}\,\frac{\partial K}{\partial q(x_n)}, \qquad x_n \to x \tag{0.55}$$

and

$$\frac{\delta^2 K}{\delta q(x')\,\delta q(x)} \equiv \lim_{\Delta x\to 0} \frac{1}{\Delta x^2}\,\frac{\partial^2 K}{\partial q(x_m)\,\partial q(x_n)}, \qquad x_n \to x, \quad x_m \to x'. \tag{0.56}$$

Equation (0.55) is the first functional derivative of K in the case of a functional dependence (0.45). It answers the question "By how much will the number K change if the function $q(x)$ changes by a small amount at all x?"

Equation (0.56) defines the second mixed functional derivative of K. As noted before, we will have occasion to use this concept in Chapter 11 on quantum gravity. The dynamical equation of this phenomenon is not the usual second-order differential equation, but, rather, a second-order *functional* differential equation. The second functional derivative is with respect to (metric) functions $\mathbf{q}(\mathbf{x})$, where $\mathbf{x}$ is, now, a *four*-position. See the following.

Although the preceding derivation was for the case of a scalar coordinate x, it is easily generalized to the case of a four-vector $\mathbf{x}$ as well. (One merely

replaces all scalars x by a four-vector $\mathbf{x}$, with all subscripts as before. Of course, ε is still a scalar.) This gives a definition

$$\frac{\delta^2 K}{\delta q(\mathbf{x}')\,\delta q(\mathbf{x})} \equiv \lim_{\Delta\mathbf{x}\to 0} \frac{1}{\Delta\mathbf{x}^2} \frac{\partial^2 K}{\partial q(\mathbf{x}_m)\,\partial q(\mathbf{x}_n)}, \tag{0.57}$$

$$\mathbf{x}_n \to \mathbf{x}, \quad \mathbf{x}_m \to \mathbf{x}', \quad \Delta\mathbf{x} \equiv \Delta x\,\Delta y\,\Delta z\,c\,\Delta t,$$

where c is the speed of light and $\Delta\mathbf{x}$ is the increment volume used in four-space.

Finally, we consider problems where *many* amplitude functions $q_k(\mathbf{x}, \varepsilon)$, $k = 1, 2, \ldots$ exist. (We subscripted these with a single subscript, but results will hold for any number of subscripts as well.) In Eq. (0.45), the functional K now depends upon all of these functions. We now get the logical generalization of Eq. (0.57),

$$\frac{\delta^2 K}{\delta q_j(\mathbf{x}')\,\delta q_k(\mathbf{x})} = \lim_{\Delta\mathbf{x}\to 0} \frac{1}{\Delta\mathbf{x}^2} \frac{\partial^2 K}{\partial q_j(\mathbf{x}_m)\,\partial q_k(\mathbf{x}_n)}, \quad \mathbf{x}_n \to \mathbf{x}, \quad \mathbf{x}_m \to \mathbf{x}'. \tag{0.58}$$

0.3.10 Exercise

Show this result, using the analogous steps to Eqs. (0.45)–(0.57). *Hint*: The right-hand functions in Eq. (0.48), and ε, must now be subscripted by k. Then Eq. (0.49) becomes a power series in changes $d\varepsilon_k$ including (now) second-order mixed-product terms $d\varepsilon_k\,d\varepsilon_j$ that are *summed* over k and j. Proceeding, on this basis, through to Eq. (0.57) now logically gives the definition (0.58).

The procedure (0.49)–(0.56) may be easily extended to allow third, and all higher, orders of functional derivatives to be defined. The dots at the ends of Eqs. (0.53) and (0.54) indicate the higher-order terms, which may be easily added in.

0.3.11 Alternate form for functional derivative

Definition (0.55) is useful when the functional K has the form of a sum over discrete samples $q(x_n)$, $n = 1, 2, \ldots$. Instead, in many problems K is expressed as an action integral Eq. (0.2), where $q(x)$ is *continuously* sampled. Obviously, definition (0.55) is not directly usable for such a problem. For such continuous cases one may use, instead, the equivalent definition

$$\frac{\delta K}{\delta q(y)} \equiv \lim_{\varepsilon\to 0} \frac{K[q(x) + \varepsilon\delta(x - y)] - K[q(x)]}{\varepsilon} \tag{0.59}$$

where $\delta(x)$ is the Dirac delta function (Ryder, 1987, p. 177).

0.3.12 An observation

Suppose that two non-identical functionals $I[q(x)]$ and $J[q(x)]$ obey a relation

$$I - J = 0 \tag{0.60}$$

for some scalar function $q_1(x)$. Is there necessarily a solution $q_2(x)$ to the variational problem

$$I - J = extrem.? \tag{0.61}$$

The latter is a problem requiring quantity $I - J$ to be *stationary* for some $q_2(x)$. As we saw, its solution would have to obey the Euler–Lagrange Eq. (0.13). This is, of course, a very different requirement on $q_2(x)$ than the zero-condition (0.60). Moreover, even if $q_1(x)$ exists there is no guarantee that any solution $q_2(x)$ exists. Thus, the fact that a functional is zero for some solution does not guarantee that that, or any other, solution *extremizes* the functional. Analogously, the existence of an algebraic zero-point does not necessarily imply that an extremum condition is satisfied at either the zero-point or anywhere else.

0.4 Dirac delta function

A handy concept to use when evaluating integrals is that of the Dirac delta function $\delta(x)$. It is the continuous counterpart of the Kronecker delta function

$$\delta_{ij} = 0 \text{ for } i \neq j, \quad \sum_{j=1}^{N} \delta_{ij} = 1, \tag{0.62a}$$

$$\sum_{j=1}^{N} f_j \delta_{ij} = f_i \tag{0.62b}$$

for any i on the interval $(1, N)$. Notice that δ_{ij}, for a fixed value of i, is a function of j that is a pure "spike," i.e., zero everywhere except at the single point $i = j$ where it has a "yield" of 1.

Similarly, $\delta(x)$ obeys

$$\delta(x - a) = 0 \text{ for } x \neq a, \quad \int dx\, \delta(x - a) = 1, \tag{0.63a}$$

$$\int dx\, f(x)\delta(x - a) = f(a) \tag{0.63b}$$

for any real a and any function $f(x)$. (In these integrals and throughout the book, the limits are from $-\infty$ to ∞ unless otherwise stated.) It is useful to compare Eqs. (0.62a) with Eqs. (0.63a), and Eq. (0.62b) with Eq. (0.63b). What should function $\delta(x)$ look like?

From Eqs. (0.63a) it must be flat zero everywhere except at the point $x = a$,

$$F(k) = \frac{1}{\sqrt{2\pi}} \int dx\, f(x) \exp(-ikx), \qquad i = \sqrt{-1} \tag{0.65}$$

and its inverse relation

$$f(x) = \frac{1}{\sqrt{2\pi}} \int dk'\, F(k') \exp(ik'x). \tag{0.66}$$

Substituting Eq. (0.66) into (0.65) gives

$$\begin{aligned} F(k) &= \int dx \exp(-ikx) \frac{1}{2\pi} \int dk' F(k') \exp(ik'x) \\ &= \int dk'\, F(k') \frac{1}{2\pi} \int dx \exp[-ix(k - k')] \end{aligned} \tag{0.67}$$

after switching orders of integration. Then, by the sifting property Eq. (0.63b), it must be that

$$\frac{1}{2\pi} \int dx \exp[-ix(k - k')] = \delta(k - k'). \tag{0.68}$$

Analogous properties to Eqs. (0.63b) and (0.68) exist for *multidimensional* functions $f(\mathbf{x})$, where $\mathbf{x}$ is a vector of dimension M. Thus there is a sifting property

$$\int d\mathbf{x}\, f(\mathbf{x}) \delta(\mathbf{x} - \mathbf{a}) = f(\mathbf{a}) \tag{0.69}$$

and a Fourier representation

$$\frac{1}{(2\pi)^{M/2}} \int d\mathbf{x} \exp[-i\mathbf{x} \cdot (\mathbf{k} - \mathbf{k}')] = \delta(\mathbf{k} - \mathbf{k}'). \tag{0.70}$$

The latter is a multidimensional delta function. These relations may be derived as easily as were the corresponding scalar relations (0.63b) and (0.68).

Another relation that we will have occasion to use is

$$\delta(ax) = \frac{\delta(x)}{|a|}, \qquad a = const. \tag{0.71}$$

Other properties of the delta function may be found in Bracewell (1965).

A final relation of use is (Born and Wolf, 1959, Appendix IV)

$$\delta(x^2 - a^2) = \frac{\delta(x - a) + \delta(x + a)}{2|a|}. \tag{0.72}$$

Ronald A. Fisher, 1929, from a photograph taken in honor of his election to Fellow of the Royal Society. Sketch by the author.

1.1 On Lagrangians

The Lagrangian approach (Lagrange, 1788) to physics has been utilized now for over 200 years. It is one of the most potent and convenient tools of

theory ever invented. One well-known proponent of its use (Feynman and Hibbs, 1965) calls it "most elegant." However, an enigma of physics is the question of where its Lagrangians come from. It would be nice to justify and derive them from a prior principle, but none seems to exist. Indeed, when a Lagrangian is presented in the literature, it is often with a disclaimer, such as (Morse and Feshbach, 1953) "It usually happens that the differential equations for a given phenomenon are known first, and only later is the Lagrange function found, from which the differential equations can be obtained." Even in a case where the differential equations are *not* known, often candidate Lagrangians are first constructed, to see whether "reasonable" differential equations result.

Hence, the Lagrange function has been principally a contrivance for getting the correct answer. It is the means to an end – a differential equation – but with no significance in its own right. One of the aims of this book is to show, in fact, that Lagrangians do have prior significance. A second aim is to present *a systematic approach to deriving* Lagrangians. A third is to clarify the role of the observer in a measurement. These aims will be achieved through use of the concept of Fisher information.

R. A. Fisher (1890–1962) was a researcher whose work is not well known to physicists. He is renowned in the fields of genetics, statistics, and eugenics. Among his pivotal contributions to these fields (Fisher, 1959) are the maximum likelihood estimate, the analysis of variance, and a measure of indeterminacy now called "Fisher information." (He also found it likely that the famous geneticist Gregor Mendel contrived the "data" in his famous pea plant experiments. They were too regular to be true, statistically.) It will become apparent that his form of information has great utility in physics as well.

Table 1.1 shows a list of Lagrangians (most from Morse and Feshbach, 1953), emphasizing the common presence of a squared-gradient term. In quantum mechanics, this term represents mean kinetic energy, but why mean kinetic energy should be present is a longstanding mystery: Schrödinger called it "incomprehensible" (Schrödinger, 1926).

Historical note: As will become evident below, *Schrödinger's mysterious Lagrangian term was simply Fisher's data information.* May we presume from this that Schrödinger and Fisher, despite developing their famous theories nearly simultaneously, and with basically just the English channel between them, never communicated? If they had, it would seem that the mystery should have been quickly dispelled. This is an enigma.

In fact, Schrödinger's dilemma is a direct outgrowth of the prevailing view, both during his era and today, as to what Lagrangians physically represent. This fundamental question defines a "worldview" as well. The prevailing view was

Table 1.1. *Lagrangians for various physical phenomena. Where do these come from and, in particular, why do they all contain a squared gradient term? (Reprinted from Frieden and Soffer, 1995.)*

Phenomenon	Lagrangian
Classical mechanics	$\frac{1}{2}m\left(\frac{\partial q}{\partial t}\right)^2 - V$
Flexible string or compressible fluid	$\frac{1}{2}\rho\left[\left(\frac{\partial q}{\partial t}\right)^2 - c^2\,\nabla q \cdot \nabla q\right]$
Diffusion equation	$-\nabla\psi \cdot \nabla\psi^* - \cdots$
Schrödinger wave equation	$-\frac{\hbar^2}{2m}\nabla\psi \cdot \nabla\psi^* - \cdots$
Klein–Gordon equation	$-\frac{\hbar^2}{2m}\nabla\psi \cdot \nabla\psi^* - \cdots$
Elastic wave equation	$\frac{1}{2}\rho\dot{q}^2 - \cdots$
Electromagnetic equations	$4\sum_{n=1}^{4}\Box q_n \cdot \Box q_n - \cdots$
Dirac equations	$-\frac{\hbar^2}{2m}\nabla\psi \cdot \nabla\psi^* - \cdots = 0$
General relativity (equations of motion)	$\sum_{m,n=1}^{4} g_{mn}(q(\tau))\frac{\partial q_m}{\partial \tau}\frac{\partial q_n}{\partial \tau}$ $\uparrow$ metric tensor
Boltzmann law	$4\left(\frac{\partial q(\mathrm{E})}{\partial \mathrm{E}}\right)^2 - \cdots, \quad p(\mathrm{E}) \equiv q^2(\mathrm{E})$
Maxwell–Boltzmann law	$4\left(\frac{\partial q(v)}{\partial v}\right)^2 - \cdots, \quad p(v) \equiv q^2(v)$
Lorentz transformation (special relativity)	$\partial_i q_n\,\partial_i q_n$ (invariance of integral)
Helmholtz wave equation	$-\nabla\psi \cdot \nabla\psi^* - \cdots$

that they are *energies*, and their integrals are "action integrals." On this basis the Lagrangian for classical mechanics, shown at the top of Table 1.1, is the difference between a kinetic energy term $m(dq/dt)^2/2$ and a potential energy term V. However, consider the following counterpoint.

Lagrangians exist whose terms have *no explicit connection with energy.* Examples are those describing genetic evolution (Chapter 14), macroeconomics (Chaps. 13 and 14), and cancer growth (Chapter 15). (These Lagrangians were of course not known in Schrödinger's day.) There is no denying the law of conservation of energy, but, evidently, the concept of energy does not suffice for forming Lagrangians for *all* fields of science. Is there a concept that does?

There is no science without observation. Therefore a common denominator of all science is *measurement*. This views science from the bottom up (Sec. 0.1). On this basis the terms of the Lagrangian should describe in some way the *process* of measurement and the information flow it incurs. In fact, measurement sets in motion a flow of Fisher information (Chaps. 3, 10). On this basis the mysterious squared-gradient term of Schrödinger turns out to be the amount of Fisher information that resides in the measurement (Eq. (2.19) of Chapter 2). In particular, it is the amount of Fisher information residing in a variety of data called *intrinsic data*. The remaining terms of the Lagrangian will be seen to arise out of the information residing in the *phenomenon* that is under measurement. Thus, all Lagrangians consist entirely of two forms of Fisher information – data information and phenomenological information.

The concept of Fisher information is a natural outgrowth of classical measurement theory, as follows.

1.2 Classical measurement theory

1.2.1 The "smart" measurement

Consider the basic problem of estimating a single parameter of a system (or phenomenon) from knowledge of some measurements. See Fig. 1.1. Let the parameter have a definite, but unknown, value θ, and let there be N data values $y_1, \ldots, y_N \equiv \mathbf{y}$, in vector notation, at hand. The system is specified by a conditional probability law $p(\mathbf{y}|\theta)$ called the "likelihood law."

The data obey

$$\mathbf{y} = \theta + \mathbf{x}, \tag{1.0}$$

where the $x_1, \ldots, x_N \equiv \mathbf{x}$ are added noise values. The data are used in an estimation principle to form an estimate of θ which is an *optimal* function $\hat{\theta}(\mathbf{y})$ of all the data; e.g., the function might be the sample mean $N^{-1}\sum_n y_n$. The overall measurement procedure is "smart" in that $\hat{\theta}(\mathbf{y})$ is on average a better estimate of θ than is any one of the data observables.

The noise $\mathbf{x}$ is assumed to be *intrinsic* to the parameter θ under measure-

1.2.2 Fisher information

This information arises as a measure of the expected error in a smart measurement. Consider the class of "unbiased" estimates, obeying $\langle \hat{\theta}(\mathbf{y}) \rangle = \theta$; these are correct "on average." The mean-square error e^2 in such an estimate $\hat{\theta}$ obeys a relation (Van Trees, 1968; Cover and Thomas, 1991)

$$e^2 I \geqslant 1, \tag{1.1}$$

where I is called the Fisher "information." In a particular case of interest $N = 1$ (see below), this becomes

$$I = \int dx \, p'^2(x)/p(x), \qquad p' \equiv dp/dx. \tag{1.2}$$

(Throughout the book, integration limits are infinite unless otherwise specified.) Quantity $p(x)$ denotes the probability density function (PDF) for the noise value x. If $p(x)$ is Gaussian, then $I = 1/\sigma^2$ with σ^2 the variance (see derivation in Sec. 8.3.1).

Equation (1.1) is called the Cramer–Rao inequality. It expresses *reciprocity* between the mean-square error e^2 and the Fisher information I in the intrinsic data. Hence, it is an expression of *intrinsic* uncertainties, i.e., in the absence of outside sources of noise. It will be shown in Eq. (4.53) that the reciprocity relation goes over into the Heisenberg uncertainty principle, in the case of a single measurement of a particle position value θ. Again, this ignores the possibility of noise of detection, which would add in additional uncertainties to the relation (Arthurs and Goodman, 1988; Martens and de Muynck, 1991).

The Cramer–Rao inequality (1.1) shows that estimation quality increases (e decreases) as I increases. Therefore, I is a quality metric of the estimation procedure. This is the essential reason why I is called an "information." Equations (1.1) and (1.2) derive quite easily, as is shown next.

1.2.3 Derivation

We follow Van Trees (1968). Consider the class of estimators $\hat{\theta}(\mathbf{y})$ that are unbiased, obeying

$$\langle \hat{\theta}(\mathbf{y}) - \theta \rangle \equiv \int d\mathbf{y} \, [\hat{\theta}(\mathbf{y}) - \theta] p(\mathbf{y}|\theta) = 0. \tag{1.3}$$

PDF $p(\mathbf{y}|\theta)$ describes the fluctuations in data values $\mathbf{y}$ in the presence of the parameter value θ. PDF $p(\mathbf{y}|\theta)$ is called the "likelihood law." Differentiate Eq. (1.3) $\partial/\partial\theta$, giving

$$\int d\mathbf{y} \, (\hat{\theta} - \theta) \frac{\partial p}{\partial \theta} - \int d\mathbf{y} \, p = 0. \tag{1.4}$$

This means that the fluctuations in y from θ are invariant to the size of θ, a kind of shift invariance. (This becomes an expression of *Galilean invariance* when random variables y and θ are 3-vectors instead.) Using condition (1.11) and identity (1.5) in Eq. (1.9) gives

$$I = \int dy \left[\frac{\partial p(y-\theta)}{\partial(y-\theta)}\right]^2 \Big/ p(y-\theta), \tag{1.12a}$$

since by the chain rule $\partial/\partial\theta = (\partial/\partial(y-\theta))\,\partial(y-\theta)/\partial\theta = -\partial/\partial(y-\theta)$. Parameter θ is regarded as fixed (see above), so that a change of variable $x = y - \theta$ gives $dx = dy$. Equation (1.12a) then becomes Eq. (1.2), as required, with $p_X(x) \equiv p(x)$ as simpler notation. Note that I no longer depends upon θ. This is convenient since θ was unknown.

Shift invariance (1.11) holds quite often. Consider a scalar case $N = 1$ of Eq. (1.0) and temporarily regard x as a "noise" value. By Eq. (1.0), since θ is fixed, each time a fluctuation x occurs a corresponding y value occurs. Then the frequency of occurrence of a value of y in the presence of a fixed θ equals that for a corresponding value of x, or

$$p(y|\theta) = p_X(x|\theta) = p_X(y-\theta|\theta) \quad \text{since} \quad x \equiv y - \theta. \tag{1.12b}$$

Next, consider any effect whereby the noise fluctuation x is independent of the size of θ. By definition of independence

$$p_X(x|\theta) = p_X(x). \tag{1.12c}$$

Using this in (1.12b) with $x \equiv y - \theta$ then gives (1.11) as required.

This derivation required that the noise fluctuation x be independent of the size of θ. When does this occur physically? In fact the most fundamental physical effects obey this property. A few are considered next. In these examples, all coordinates x, y, θ are measured, as usual, from a fixed origin in the laboratory.

Suppose that a particle, of mass m and at general linear position y, is undergoing oscillatory linear motion about a *fixed* rest position θ along X. Denote its general displacement from θ as x, so that Eq. (1.0) is again obeyed. The particle is attached to one end of an elastic spring whose other end is fastened at θ. The spring exerts a restoring force $-Kx$ upon the particle, $K = const.$ As is well known, the motion of the particle is governed by Newton's second law, in the form

$$-Kx = m\,d^2x/dt^2. \tag{1.12d}$$

This says that the motion of the particle is completely described by the time dependence of x. The value of θ simply does not enter in. It results that, if the observer keeps track of the particle's trajectory values x and bins them at a constant time subdivision to form a histogram of relative occurrences $p_X(x)$,

then this histogram (or probability law) is *likewise* found to be independent of the size of θ. That is, Eq. (1.12c) is obeyed. Then by the argument below Eq. (1.12c), the required effect (1.11) is likewise obeyed.

The condition Eq. (1.11), or equivalently (1.12c), is also called one of "shift invariance." This is for the following reasons. Suppose that the origin of laboratory coordinates were shifted in X by a finite amount Δx. Then *both coordinates* y and θ are so shifted. Denote by subscript s the shifted coordinates. Thus $y_s = y + \Delta x$ and $\theta_s = \theta + \Delta x$. But then $x_s \equiv y_s - \theta_s = y - \theta \equiv x$ (by Eq. (1.11)); that is, each new value x_s of the displacement equals the old value x. Consequently, if the new *displacement* values are binned, the new probability law $p_{X_s}(x_s) = p_X(x)$, the old. The law is invariant to shift.

In many applications the invariance holds only over a finite range of shifts. This occurs, for example, for probability laws which are the "point spread functions" of optics (Born and Wolf, 1959). These are shift-invariant over only a finite area called the "isoplanatic patch." However, we shall not explicitly consider such finite-area cases in this book.

The previous argument holds as well for *any* left-hand force term in (1.12d), so long as it depends only upon x. That is, it holds so long as the force term depends only upon the particle's displacement from the center of potential located at θ. More generally, it holds for any isolated quantum mechanical system subject to such a potential. Here, ignoring the time for simplicity, a particle obeys a probability law $p_X(x) \equiv |\psi(x)|^2$, where x is again the displacement of the particle from the center of potential (or from some fixed point in the laboratory if there is no potential). The probability law is *not* of the form $|\psi(x|\theta)|^2$. The absolute position θ of the center of potential does not enter into the Schrödinger wave equation, which governs $\psi(x)$. Equivalently, irrespective of the position of the origin of coordinates in the laboratory, or indeed of the laboratory in the *universe*, the particle obeys the same Schrödinger wave equation.

Finally, we should consider cases where shift invariance *does not* hold. Possibly the most well-known example is that of the Poisson law,

$$P(y|\theta) = e^{-\theta}\frac{\theta^y}{y!}, \qquad y = 0, 1, 2, \ldots .$$

The right-hand side is visibly *not* a function of $y - \theta$, as was needed for shift invariance (Eq. (1.12b)). The Poisson law was originally used to describe the number y per year of Prussian cavalry officers kicked to death by their horses. It also describes many physical situations, e.g., the random number of photons of ambient light counted during a finite detection time, where the mean count is θ. Data y exhibit what is called "signal-dependent noise," since their

fluctuations depend upon the *absolute size* of the "signal" value θ. As an example of the signal dependence, the variance of the fluctuations equals θ.

Most of the probability laws that are derived in this book are presumed to obey shift invariance (1.11) or its multidimensional form Eq. (2.16). Notable exceptions are the PDFs derived in Chaps. 12 and 14.

1.2.5 Use of principal value in evaluating I

Certain probability laws, such as the exponential,

$$p(x) = \begin{cases} a^{-1}\exp(-x/a), & a>0, \quad x \geqslant 0 \\ 0, & x<0 \end{cases} \tag{1.12e}$$

have discontinuities in x. For (1.12e) it is the point $x = 0$. Then the slope dp/dx becomes indefinitely large in Eq. (1.2) for the information, so that the latter blows up. This must be avoided if I is to be a practical tool of the approach. Here is where the Cauchy principal value idea enters in. The point of discontinuity is simply *avoided*, by redefining the information I in (1.2) such that isolated points of discontinuity are skipped over during the integration. For example, in the case (1.12e) of the exponential law, the redefined information obeys

$$I = \lim_{\delta \to 0} \int_{0+\delta}^{\infty} dx\, p'^2(x)/p(x), \qquad \delta > 0 \tag{1.12f}$$

(cf. Eq. (1.2)). This gives the well-defined answer $I = 1/a^2$ for the law (1.12e).

In general the redefined information also still obeys the Cramer–Rao inequality, since the resulting error

$$e^2 \equiv \lim_{\delta \to 0} \int_{0+\delta}^{\infty} dx\, [\hat{\theta}(y) - \theta]^2\, p(x), \qquad y \equiv \theta + x, \qquad \delta > 0 \tag{1.12g}$$

differs from that (Eq. (1.10)) of the non-principal value approach by an isolated point of finite value. Such a point contributes negligibly to the integral.

From this point on, by information I we shall mean the *principal value* of the information.

1.2.6 Solutions p(x) of problem I = extrem.

In the case of a *flat* probability law of any width, the principal value of the information is, from Eq. (1.12f), identically $I = 0$. Does this make sense?

Consider the problem of finding the probability law $p(x)$ that *extremizes* the information functional (1.12f). By the Legendre condition of Secs. 0.3.3 and

0.3.4 the extremum is a *minimum*. What then should be the nature of the solution $p(x)$ to the problem?

The following tendencies will be found in Sec. 1.7. The more spread out and smooth the law $p(x)$ is the more random is x, and the more disordered the system is; therefore, the smaller the information I should be. Thus, solving a problem $I = min.$ should give the smoothest law $p(x)$ possible. Let us see whether using the principal value of the information gives this kind of result.

Unconstrained problem: Suppose that x is restricted to the interval (x_1, x_2). The problem is to find $p(x)$ obeying

$$I \equiv \lim_{\delta \to 0} \int_{x_1+\delta}^{x_2-\delta} dx\, p'^2(x)/p(x) = min., \quad \delta > 0. \tag{1.12h}$$

With a Lagrangian $\mathscr{L} = p'^2(x)/p(x)$, we see that $\partial\mathscr{L}/\partial p' = 2p'/p$ and $\partial\mathscr{L}/\partial p = -p'^2/p^2$. Then the Euler–Lagrange Eq. (0.13) for the solution obeys $2f' + f^2 = 0$, $f \equiv p'/p$. The solution for f is $f = 2(x+a)^{-1}$, $a = const.$ Then p is found from this to be

$$p(x) = (bx + c)^2, \quad x_1 \leqslant x \leqslant x_2, \quad b, c = const., \tag{1.12i}$$

a truncated parabola. Back substituting this result into (1.12h) gives a minimized information of size

$$I = 4b^2(x_2 - x_1). \tag{1.12j}$$

This shows that the absolute minimum value for I, $I = 0$, is attained when $b = 0$. Then by (1.12i)

$$p(x) = c^2 = const., \quad x_1 \leqslant x \leqslant x_2. \tag{1.12k}$$

Hence $p(x)$ is a rectangle function. This is indeed the smoothest law *within* the fixed interval. Hence our requirement that I monotonically decrease as $p(x)$ gets smoother is satisfied by use of the principal value definition of I. Not having had to evaluate the infinite slope values $p'(x)$ at the endpoints was essential to the calculation. To avoid such points is the reason the principal value will be implicit in all calculations of the Fisher information and Fisher channel capacity (Chapter 2) in this text.

Constrained problem: Next, consider a family of PDFs that are constrained to obey $p(0) = 0$, and to have the form of a power law, $p(x) = Cx^\gamma$ (cf. Eq. (1.12i)), $0 \leqslant x \leqslant x_2$, $C = const.$, $\gamma = const.$ These PDFs arise in the analysis of cancer growth with time (Chapter 15), where x is a value of the time. The condition $p(0) = 0$ means that the time origin is the time of inception of the cancer. The PDF (1.12i) gave an absolute minimized I value of zero. *Now how small a value of I can be attained*?

Since $p(x)$ must obey normalization and $p(0) = 0$, a constant solution $\gamma = 0$ is no longer possible, so that the absolute minimum value $I = 0$ is

H remains constant. H is then said to be a *global* measure of the behavior of $p(x_n)$.

By comparison, the discrete form of Fisher information I is, from Eq. (1.2),

$$I = \Delta x^{-1} \sum_n \frac{[p(x_{n+1}) - p(x_n)]^2}{p(x_n)}. \tag{1.15}$$

If the curve $p(x_n)$ undergoes a rearrangement of points x_n as above, discontinuities in $p(x_n)$ will now occur. Hence the local slope values $[p(x_{n+1}) - p(x_n)]/\Delta x$ will change drastically, and so the sum (1.15) will also change strongly. Since I is thereby sensitive to local rearrangement of points, it is said to have a property of *locality*.

Thus, H is a global measure, while I is a local measure, of the behavior of the curve $p(x_n)$. These properties hold in the limit $\Delta x \to 0$, and so apply to the continuous probability density $p(x)$ as well.

This global versus local property has an interesting ramification to valuating financial securities (Sec. 13.7.1). Another is as follows.

Because the integrand of I contains a squared derivative p'^2 (see Eq. (1.2)), when the integrand is used as part of a Lagrangian the resulting Euler–Lagrange equation will contain second-order derivative terms p''. Hence, a second-order differential equation results (see Eq. (0.25)). This dovetails with nature, in that the major fundamental differential equations that define probability densities or amplitudes in physics are *second-order* differential equations. Indeed, the thesis of this book is that the correct differential equations result when the information I-based EPI principle of Chapter 3 is followed.

By contrast, the integrand of H in (1.13) does not contain a derivative. Therefore, when this integrand is used as part of a Lagrangian the resulting Euler–Lagrange equation will not contain any derivatives (see Eq. (0.22)); it will be an algebraic equation, with the immediate solution that $p(x)$ has the exponential form Eq. (0.22) (Jaynes, 1957a, 1957b). This is not, then, a differential equation, and hence cannot represent a general physical scenario. The exceptions are those distributions which happen *to be* of an exponential form, as in statistical mechanics. (In these cases, I gives the correct solutions anyhow; see Chapter 7.)

It follows that, if one or the other of global measure H or local measure I is to be used in a variational principle in order to derive the physical law $p(x)$ describing a *general* scenario, the preference is given to the local measure I.

As all of the preceding discussion implies, H and I are two distinct functionals of $p(x)$. However, quite the contrary is true in comparing I with an entropy that is closely related to H, namely, the Kullback–Leibler entropy. This is discussed in Sec. 1.4.

But each of the two far-right sums is Δx^{-1}, by normalization, so that their difference cancels out, leaving

$$I = -(2/\Delta x)\sum_n p(x_n)\ln\left(\frac{p(x_n+\Delta x)}{p(x_n)}\right) \tag{1.22a}$$

$$\to -(2/\Delta x^2)\int dx\, p(x)\ln\left(\frac{p(x+\Delta x)}{p(x)}\right) \tag{1.22b}$$

$$= -(2/\Delta x^2)G[p(x),\ p(x+\Delta x)] \tag{1.22c}$$

by definition (1.16). Thus, I is proportional to the cross-entropy between the PDF $p(x)$ and a reference PDF that is its shifted version $p(x+\Delta x)$.

1.4.1 Historical note

Savage (1972) first proved the equality (1.22b). It was later independently re-proved by Vstovsky (1995).

1.4.2 Exercise

One notes that the form (1.22b) is indeterminate $0/0$ in the limit $\Delta x \to 0$. Show that one use of l'Hôpital's rule does not resolve the limit, but two does, and the limit is precisely the form (1.2) of I.

1.4.3 Fisher information as a "mother" information

Equation (1.22c) shows that I is the cross-entropy between a PDF $p(x)$ and its infinitesimally shifted version $p(x+\Delta x)$. It has been noted (Caianiello, 1992) that I more generally results as a "cross-information" between $p(x)$ and $p(x+\Delta x)$ for a host of *different* types of information measures. Some examples are as follows:

$$R_\alpha \equiv \ln\int dx\, p(x)^\alpha p(x+\Delta x)^{1-\alpha} \to -\Delta x^2\, 2^{-1}\alpha(1-\alpha)I, \tag{1.22d}$$

for $\alpha \neq 1$, where R_α is called the "Renyi information" measure (Amari, 1985); and

$$W \equiv \cos^{-1}\left[\int dx\, p^{1/2}(x)p^{1/2}(x+\Delta x)\right], \quad W^2 \to \Delta x^2\, 4^{-1} I, \tag{1.22e}$$

called the "Wootters information" measure (Wootters, 1981). To derive these results, one only has to expand the indicated function of $p(x+\Delta x)$ in the

integrand out to *second order* in Δx, and perform the indicated integrations, using the identities $\int dx\, p'(x) = 0$ and $\int dx\, p''(x) = 0$.

Hence, Fisher information is the limiting form of many different measures of information; it is a kind of "mother" information.

1.5 Amplitude form of *I*

In definition (1.2), the division by $p(x)$ is bothersome. (For example, is I undefined since necessarily $p(x) \to 0$ at certain x?) A way out is to work with a real "amplitude" function $q(x)$,

$$p(x) = q^2(x). \tag{1.23}$$

(Interestingly, probability amplitudes were used by Fisher (1943) independently of their use in quantum mechanics. The purpose was to discriminate among population classes.) Using form (1.23) in (1.2) directly gives

$$I = 4\int dx\, q'^2(x). \tag{1.24}$$

This is of a simpler form than (1.2) (no more divisions), and shows that I *simply measures the gradient content in* $q(x)$ (and hence in $p(x)$). The integrand $q'^2(x)$ in (1.24) is the origin of the squared gradients in Table 1.1 of Lagrangians, as will be seen.

Representation (1.24) for I may be computed independently of the preceding. One measure of the "distance" between an amplitude function $q(x)$ and its displaced version $q(x + \Delta x)$ is the quadratic measure (Braunstein and Caves, 1994)

$$L^2 \equiv \int dx\, [q(x + \Delta x) - q(x)]^2 \to \Delta x^2 \int dx\, q'^2(x) = \Delta x^2\, 4^{-1} I \tag{1.25}$$

after expanding out $q(x + \Delta x)$ in first-order Taylor series about point x (cf. Eqs. (1.22c–e) preceding).

1.6 Efficient estimators

Classically, the main use of information I has been as a measure of the ability to estimate a parameter. This is through the Cramer–Rao inequality (1.1), as follows.

If the equality can be realized in Eq. (1.1), then the mean-square error will go inversely with I, indicating that I determines how small (or large) the error can be in any particular scenario. The question is, then, when is the equality realized?

The left-hand side of Eq. (1.7) is actually an inner product between two "vectors" $A(\mathbf{y})$ and $B(\mathbf{y})$,

$$A(\mathbf{y}) = \frac{\partial \ln p}{\partial \theta}\sqrt{p}, \quad B(\mathbf{y}) \equiv (\hat{\theta} - \theta)\sqrt{p}. \tag{1.26a}$$

Here the continuous index $\mathbf{y}$ defines the $\mathbf{y}$th component of each such vector (in contrast to the elementary case where vector components are discrete). The inner product of two vectors A, B is always less than or equal to its value when the two vectors are *parallel*, i.e., when all their $\mathbf{y}$-components are proportional,

$$A(\mathbf{y}) = k(\theta)B(\mathbf{y}), \quad k(\theta) = const. \tag{1.26b}$$

(Note that function $k(\theta)$ remains constant since the parameter θ is, of course, constant.) Combining Eqs. (1.26a) and (1.26b) then provides a necessary condition (i) for attaining the equality in Eq. (1.1),

$$\frac{\partial \ln p(\mathbf{y}|\theta)}{\partial \theta} = k(\theta)[\hat{\theta}(\mathbf{y}) - \theta]. \tag{1.27}$$

A condition (ii) is the previously used unbiasedness assumption (1.3).

A PDF scenario where (1.27) is satisfied causes a minimized error $e^2_{\min}$ that obeys

$$e^2_{\min} = 1/I. \tag{1.28}$$

The estimator $\hat{\theta}(y)$ is then called "efficient." Notice that in this case the error varies inversely with information I, so that the latter becomes a well-defined quality metric of the measurement process.

1.6.1 Exercise

It is noted that only certain PDFs $p(\mathbf{y}|\theta)$ obey condition (1.27), among them (a) the independent normal law $p(\mathbf{y}|\theta) = A\prod_n \exp[-(y_n - \theta)^2/2\sigma^2]$, $A = const.$, and (b) the exponential law $p(\mathbf{y}|\theta) = \prod_n e^{-y_n/\theta}/\theta$, $y_n \geqslant 0$. On the other hand, with $N = 1$, (c) a PDF of the form

$$p(y|\theta) = A\sin^2(y - \theta), \quad A = const., \quad |y - \theta| \leqslant \pi$$

does not satisfy (1.27). Note that this PDF arises when the position θ of a one-dimensional quantum mechanical particle within a box is to be estimated. Hence, this fundamental measurement problem does not admit of an efficient estimate. Show these effects (a)–(c).

Also show that the estimators in (a) and (b) are unbiased, as required.

1.6.2 Exercise

If the condition (1.27) *is* obeyed, and if the estimator is unbiased, then the estimator function $\hat{\theta}(\mathbf{y})$ that *attains* efficiency is the one that maximizes the likelihood function $p(\mathbf{y}|\theta)$ through choice of θ (Van Trees, 1968). This is called the *maximum likelihood* (ML) estimator. As an example, the ML estimators for the problems (a) and (b) preceding are both the simple average of the data. Show this.

Note the simplification that occurs if one maximizes, instead of the likelihood, the *logarithm* of the likelihood. This *log-likelihood* law is also of fundamental importance to quantum measurement theory; see Chapter 10.

1.7 Fisher *I* as a measure of system disorder

We showed that information I is a quality metric of an efficient measurement procedure. Now we will find that I is also a measure of the degree of disorder of a system. *High disorder* means a lack of predictability of values x over its range, i.e., a largely uniform or "unbiased" PDF $p(x)$. Such a curve is shown in Fig. 1.2(b). The curve has small gradient content, i.e., *it is broad and smooth*. Then by (1.24) the Fisher information *I is small*.

Conversely, if a curve $p(x)$ shows bias to particular x values then it exhibits *low disorder*. See Fig. 1.2(a). Analytically, the curve will be *steeply sloped* about these x values, and so the value of *I becomes large*. The net effect is that I measures the degree of disorder of the system.

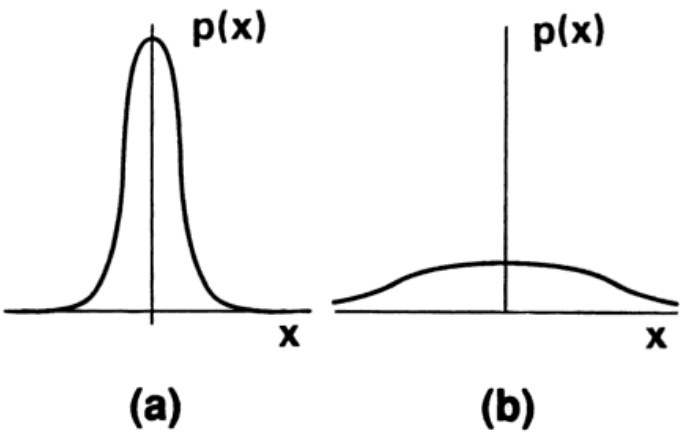

Fig. 1.2. Degree of disorder measured by I values. In (a), random variable x shows relatively low disorder and large I (gradient content). In (b), x shows high disorder and small I. (Reprinted from Frieden and Soffer, 1995.)

sian PDF, $I = 1/\sigma^2$ (see derivation in Sec. 8.3.1). Then $I = I(t) = 1/(Dt)$, or I decreases with t.

Can this result be generalized?

1.8.2 The 'I-theorem'

Equation (1.29) states that H increases monotonically with time. This result is usually called the "Boltzmann H-theorem." In fact there is a corresponding "I-theorem"

$$\frac{dI(t)}{dt} \leqslant 0. \tag{1.30}$$

1.8.3 Proof

Start with the cross-entropy representation (1.22b) of $I(t)$,

$$I(t) = -2 \lim_{\Delta x \to 0} \Delta x^{-2} \int dx\, p \ln(p_{\Delta x}/p)$$
$$p \equiv p(x|t), \qquad p_{\Delta x} \equiv p(x + \Delta x|t). \tag{1.31}$$

Under certain physical conditions, e.g., "detailed balance," short-term correlation, shift-invariant statistics (Gardiner, 1985; Reif, 1965; Risken, 1996) p obeys a *Fokker–Planck* differential equation

$$\frac{\partial p}{\partial t} = -\frac{d}{dx}[D_1(x,\, t)p] + \frac{d^2}{dx^2}[D_2(x,\, t)p], \tag{1.32}$$

where $D_1(x,\, t)$ is a drift function and $D_2(x,\, t)$ is a diffusion function. Suppose that $p_{\Delta x}$ also obeys the equation (Plastino and Plastino, 1996). Risken (1996) shows that two PDFs, such as p and $p_{\Delta x}$, that obey the Fokker–Planck equation have a cross-entropy

$$G(t) \equiv -\int dx\, p \ln(p/p_{\Delta x}) \tag{1.33}$$

that obeys an H-theorem (1.29),

$$\frac{dG(t)}{dt} \geqslant 0. \tag{1.34}$$

It follows from Eq. (1.31) that I, likewise, obeys an I-theorem (1.30). Thus, the I-theorem and the H-theorem both hold under certain physical conditions.

There also is a possibility that physical conditions exist for which one theorem holds to the exclusion of the other. From the empirical viewpoint that the I-theorem leads to the derivation of a much wider range of physical laws

1.8.5 Ramification to temperature

The Boltzmann temperature (Reif, 1965) T is defined as $1/T \equiv \partial H_B/\partial E$, where H_B is the Boltzmann entropy of an isolated system and E is its energy. Consider two systems A and A' that are in thermal contact, but are otherwise isolated, and are approaching thermal equilibrium. The Boltzmann temperature has the important property that, after thermal equilibrium has been attained, a situation

$$T = T', \quad \frac{1}{T} \equiv \frac{\partial H_B}{\partial E}, \quad \frac{1}{T'} \equiv \frac{\partial H'_B}{\partial E'} \tag{1.35}$$

of equal temperature results. Let us now look at the phenomenon from the standpoint of information I, i.e., *without* recourse to the Boltzmann entropy.

Denote the total information in system A by I, and that of system A' by I'. The parameters θ, θ' to be measured are the total energies E and E' of the two systems. The corresponding measurements are Y_E, $Y'_{E'}$. Because of the I-theorem (1.30), *both I and I' should approach minimum values as time increases*. We will show later that, since the two systems are physically separated and hence independent in their energy data Y_E, $Y'_{E'}$, the Fisher information state of the two is the sum of the two I values. Hence, the I-theorem states that, after an infinite amount of time, the information of the combined system is

$$I(E) + I'(E') = min. \tag{1.36}$$

On the other hand, energy is conserved, so that

$$E + E' \equiv C, \tag{1.37}$$

$C = const.$ (Notice that this is a deterministic relation between the two ideal parameter values, not between the data; if it held for the data, then the prior assumption of independent data would have been invalid.)

The effect of (1.37) on (1.36) is

$$I(E) + I'(C - E) = min. \tag{1.38}$$

We now define a generalized "Fisher temperature" T_θ as

$$\frac{1}{T_\theta} \equiv -k_\theta \frac{\partial I}{\partial \theta}. \tag{1.39}$$

Notice that θ is any parameter under measurement. Hence, there is a Fisher "temperature" associated with any parameter to be measured. From (1.39), T_θ simply measures the sensitivity of information level I to a change in system parameter θ. The constant k_θ gives each T_θ value the same units. A relation between the two temperatures T and T_θ is found below for a perfect gas.

Consider the case in point, $\theta = E$, $\theta' = C - \theta$. The temperature T_θ is now an energy temperature T_E. Differentiating Eq. (1.38) $\partial/\partial E$ gives

$$\frac{\partial I}{\partial E} + \frac{\partial I'}{\partial E'}(-1) = 0, \quad \text{or} \quad T_E = T'_{E'} \tag{1.40}$$

by (1.39). At equilibrium both systems attain a common Fisher energy temperature. This is analogous to the Boltzmann (conventional) result (1.35).

1.8.6 Exercise

The right-hand side of Eq. (1.39) is impractical to evaluate (although still of theoretical importance) if I is close to independent of θ. This occurs in close to a shift-invariant case (1.11) where the resulting I is close to the form (1.2). The key question is, then, whether the shift-invariance condition Eq. (1.11) holds when $\theta \equiv E$ and a measurement y_E is made. The total number N of particles comprising the system is critical here. If $N \approx 10$ or more, then (a) the PDF $p(y_E|E)$ will tend to obey the central limit theorem (Frieden, 2001) and, hence, be close to Gaussian in the shifted random variable $y_E - E$. An I results that is close to the form (1.2). At the other extreme, (b) for small N the PDF can typically be χ^2 (assuming that the $N = 1$ law is Boltzmann, i.e., exponential). Here, shift invariance would not hold. Show (a) and (b).

1.8.7 Perfect gas law

So far we have defined concepts of time and temperature on the basis of Fisher information. We now show that the perfect gas law may likewise be derived on this basis. This will also permit the (so far) unknown parameter k_E to be evaluated from known parameters of the system.

Consider an ideal gas consisting of M identical molecules confined to a volume V and kept at Fisher temperature T_E. We want to know how the pressure in the gas depends upon the extrinsic parameters V and T_E. The plan is to first compute the temporal mean pressure $\overline{p}$ within a small volume $dV = A\,dx$ of the gas and then integrate through to get the macroscopic answer.

Suppose that the pressure results from a force F that is exerted normal to area A and through the distance dx, as in the case of a moving piston. Then (Reif, 1965)

$$\overline{p} \equiv \frac{F}{A}\frac{dx}{dx} = -\frac{\partial E}{\partial V} \tag{1.41}$$

where the minus sign signifies that energy E is stored in reaction to work done by the force. Using the chain rule, Eq. (1.41) becomes

$$\overline{p} = -\frac{\partial E}{\partial I}\frac{\partial I}{\partial V} = k_E T_E \frac{\partial I}{\partial V}, \tag{1.42}$$

the latter by definition (1.39) with $\theta = E$. Here dI is the information in a data reading dy_E of the ideal energy value dE. In general, quantities $\overline{p}$, dI, and T_E can be functions of the position $\boldsymbol{r}$ of volume dV within a gas. Multiplying (1.42) by dV gives

$$\overline{p}(\boldsymbol{r})\, dV = k_E T_E(\boldsymbol{r})\, dI(\boldsymbol{r}) \tag{1.43}$$

with the $\boldsymbol{r}$-dependence now noted. Near equilibrium the gas should be well mixed and homogeneous, such that $\overline{p}$ and T are independent of position $\boldsymbol{r}$. Then Eq. (1.43) may be directly integrated to give

$$\overline{p}V = k_E T_E I. \tag{1.44}$$

Note that $I = \int dI(\boldsymbol{r})$ is simply the total information due to many independent data readings dy_E. This again states that the information adds under independent data conditions.

The dependence (1.44) of $\overline{p}$ upon V and T_E is of the same form as the known equation of state of the gas

$$\overline{p}V = MkT, \tag{1.45}$$

where k is the Boltzmann constant and T is the *ordinary* (Boltzmann) temperature. Comparing Eqs. (1.44) and (1.45), exact compliance is achieved if $k_E T_E$ is related to kT as

$$\frac{kT}{k_E T_E} = I/M, \tag{1.46}$$

the information per molecule. The latter should be a constant for a well-mixed gas.

These considerations seem to imply that thermodynamic theory may be developed completely from the standpoint of Fisher entropy, without recourse to the well-known properties of the Boltzmann entropy. In fact much progress is being made in this direction. It has been shown that Fisher information obeys the same Legendre transform property and concavity property as does entropy (Frieden *et al.*, 1999). Also, the use of Fisher information in place of entropy leads to a Schrödinger wave equation formulation of non-equilibrium thermodynamics (Frieden *et al.*, 2002a, b; also see Chapter 13). This formulation permits both quantum and thermodynamic effects to be analyzed simultaneously.

1.8.8 Ramification to derivations of physical laws

The uni-directional nature of the I-theorem (1.30) implies that, as $t \to \infty$,

$$I(t) = 4\int dx\, q'^2(x|t) \rightarrow min. \tag{1.47}$$

Here we used the shift-invariant form (1.24) of I. The minimum would be achieved through variation of the amplitude function $q(x|t)$. It is convenient, and usual, to accomplish this through use of an Euler–Lagrange equation (see Eq. (0.13)). The result would define the form of $q(x|t)$ at temporal equilibrium.

In order for this approach to be tenable, it would need to be modified by appropriate input constraint properties of $q(x|t)$, such as normalization of $p(x|t)$. Other constraints, describing the particular physical scenario, must also be tacked on. Examples are fixed values of the means of certain physical quantities (case $\alpha = 2$ below). Such constraints may be appended to principle (1.47) by using the method of Lagrange undetermined multipliers, Eq. (0.39):

$$I + \sum_{k=1}^{K_o} \lambda_k \int dx\, q^\alpha(x|t) f_k(x) = extrem., \tag{1.48a}$$

$$\int dx\, q^\alpha(x|t) f_k(x) = F_k, \quad k = 1, \ldots, \quad K_o, \alpha = const. \tag{1.48b}$$

The kernel functions $f_k(x)$, constraint exponent α, and data values F_k are assumed known. The multipliers λ_k are found such that the constraint equations (1.48b) are obeyed. This approach is taken in Chapter 13, and called the principle of minimum Fisher information (MFI). Owing to the arbitrariness of the constraints, it is Bayesian in nature, and therefore approximate. See also Huber (1981).

The most difficult step in the MFI approach is deciding what constraints to utilize (called the "input" constraints). The solution depends critically upon the choice of input constraints, and yet they cannot simply be all that are known to the user. They must be the particular subset of constraints that are *actually imposed* by nature. In general, this is difficult to know *a priori*. Our own approach – the EPI principle described in Chapter 3 and applied in subsequent chapters – is, in fact, of the Lagrange form (1.48a). However, it attempts to free the problem of the arbitrariness of the constraint terms. For this purpose, a physical rationale for the terms is utilized.

It is important to verify that a minimum (1.47) will indeed be attained in solution of the constrained variational problem. A maximum or point of inflection could conceivably result instead, defeating our present aims. For this purpose, we may use *Legendre's condition* for a minimum (Sec. 0.3.3): Let $\mathscr{L}$ denote the integrand (or Lagrangian) of the total integral to be extremized. In our scalar case, if

(called the EPI principle below) are, in general, different from those obtained by the corresponding use $H = max.$ of entropy. See Sec. 1.3.1. In fact, EPI solutions and $H = max.$ solutions agree only in statistical mechanics; this is shown in Appendix A.

It is interesting that correct solutions via EPI occur even for PDFs that do not obey the Fokker–Planck equation. By its form (1.32), the time rate of change of p depends only upon the present value of p. Hence, the process has short-term memory (see also Gardiner, 1991, p. 144). However, EPI may be used to derive the $1/f$ power spectral noise effect (Chapter 8), a law famous for exhibiting *long-term* memory. Also, the relativistic electron obeys an equation of continuity of flow $\partial p/\partial t = c\,\nabla \cdot (\psi^*[\alpha]\psi)$, $p \equiv \psi^*\psi$ (Schiff, 1955), where all quantities are defined in Chapter 4. This does not quite have the form of a Fokker–Planck Eq. (1.32) (compare right-hand sides). However, EPI may indeed be used to derive the Dirac equation of the electron (Chapter 4).

These considerations imply that the Fokker–Planck equation is a sufficient, but not necessary, condition for validity of the EPI procedure. An alternative condition of wider scope must exist. Such a one is the unitary condition to be discussed in Secs. 3.8.5 and 3.8.7.

1.8.10 Flow property

Since information I obeys an I-theorem Eq. (1.30), temperature effects Eqs. (1.39) and (1.40), and a gas law Eq. (1.44), indications are that I is every bit as "physical" an entity as is the Boltzmann entropy. This includes, in particular, a property of temporal *flow* from an information source to a sink. This property is used in our physical information model of Sec. 3.3.2.

1.8.11 Additivity property

A vital property of the information I is that of additivity: the information from mutually isolated systems adds. This is shown as follows.

Suppose that we have N copies of the urn mentioned in Sec. 1.8.1. See Fig. 1.3. As before, each urn contains particles that are undergoing Brownian motion. (This time the urns are not broken.) Each sits rigidly in place upon a table that moves with an unknown velocity θ in the X-direction, relative to the laboratory. A particle is randomly selected in each urn, and its total X-component laboratory velocity value y_n is measured. Let x_n denote the particle's *intrinsic* speed, i.e., relative to its urn, with $(x_n,\ n = 1, \ldots, N) \equiv \mathbf{x}$.

$$I = \int d\mathbf{y} \prod_k p_k \left[\sum_{\substack{mn \\ m \neq n}} \frac{1}{p_m} \frac{1}{p_n} \frac{\partial p_m}{\partial \theta} \frac{\partial p_n}{\partial \theta} + \sum_n \frac{1}{p_n^2} \left(\frac{\partial p_n}{\partial \theta} \right)^2 \right]. \tag{1.56}$$

Now use the fact that, in this equation, the probabilities p_k for $k \neq m$ or n integrate through as simply factors 1, by normalization. The remaining factors in $\prod_k p_k$ are then $p_m p_n$ for the first sum, and just p_n for the second sum. The result is, after some cancellation,

$$I = \sum_{\substack{mn \\ m \neq n}} \int\int dy_m \, dy_n \, \frac{\partial p_m}{\partial \theta} \frac{\partial p_n}{\partial \theta} + \sum_n \int dy_n \, \frac{1}{p_n} \left(\frac{\partial p_n}{\partial \theta} \right)^2. \tag{1.57}$$

This simplifies, drastically, as follows. The first sum separates into a product of a sum

$$\sum_n \int dy_n \, \frac{\partial p_n}{\partial \theta} \tag{1.58}$$

with a corresponding one in index m. But

$$\int dy_n \, \frac{\partial p_n}{\partial \theta} = \frac{\partial}{\partial \theta} \int dy_n \, p_n = \frac{\partial}{\partial \theta} 1 = 0 \tag{1.59}$$

by normalization. Hence the first sum in Eq. (1.57) is zero.

The second sum in Eq. (1.57) is, by Eq. (1.5),

$$\sum_n \int dy_n \, p_n \left(\frac{\partial}{\partial \theta} \ln p_n \right)^2 \equiv \sum_n I_n \tag{1.60}$$

by the definition Eq. (1.9) of I. Hence, we have shown that

$$I = \sum_{n=1}^{N} I_n \tag{1.61}$$

in this scenario of independent data. This is what we set out to prove.

It is well known that the Shannon entropy H obeys additivity, as well, under these conditions. That is, with

$$H = -\int d\mathbf{y} \, p(\mathbf{y}|\theta) \ln p(\mathbf{y}|\theta), \tag{1.62}$$

under the independence condition Eq. (1.53) it gives

$$H = \sum_{n=1}^{N} H_n, \qquad H_n = -\int dy_n \, p_n(y_n|\theta) \ln p_n(y_n|\theta). \tag{1.63}$$

1.8.12 Exercise

Show this. *Hint*: The proof is much simpler than the preceding. One merely uses the argument below Eq. (1.56) to collapse the multidimensional integrals into the one in y_n as needed.

One notes from all this that a requirement of *additivity* does not in itself uniquely identify the appropriate measure of disorder. It could be entropy or, as shown above, Fisher information. This is despite the identity $\ln(fg) = \ln(f) + \ln(g)$, which seems uniquely to imply entropy as the measure. Undoubtedly many other measures satisfy additivity as well.

1.8.13 I = min. from statistical mechanics viewpoint

According to a basic premise of statistical mechanics (Reif, 1965), the PDF for a system that *will occur* is the one that is maximum probable to occur.

A general image-forming system is shown in Fig. 1.4. It consists of a source S of particles – any type will do, whether electrons, photons, etc. – a focussing device L of some sort and an image plane M for receiving the particles. Plane M is subdivided into coordinate positions (x_n, $n = 1, \ldots, N$) with a constant, small spacing ε. An "image event" x_n is the receipt of a particle within the interval $(x_n, x_n + \varepsilon)$. The number m_n of image events x_n is noted, for each

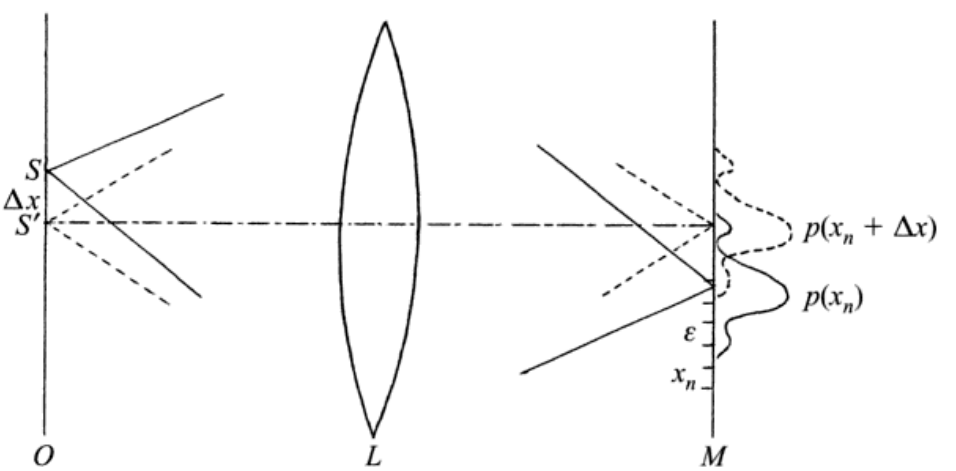

Fig. 1.4. A statistical mechanics view of Fisher information. The ideal point source position S′ gives rise to the ideal PDF $p(x_n + \Delta x)$, while the actual point source position S gives rise to the empirical PDF $p(x_n)$. Maximizing the logarithm of the probability of the latter PDF curve implies a condition of minimum Fisher information, $I[p(x_n)] = I = min$.

$n = 1, \ldots, N$. There are M particles in all, with M very large. What is the joint probability $P(m_1, \ldots, m_n)$?

Each image event is a possible position x_n, of which there are N. Therefore the image events comprise an Nary events sequence. This obeys a multinomial probability law (Frieden, 2001) of order N,

$$P(m_1, \ldots, m_n) = M! \prod_{n=1}^{N} \frac{r(x_n)^{m_n}}{m_n!}. \tag{1.64}$$

The quantities $r(x_n)$ are the "prior probabilities" of events x_n. These are considered next.

The ideal source S for the experiment would be a very small aperture that is located on-axis. This situation would give rise to ideal (prior) probabilities $r(x_n)$, $n = 1, \ldots, N$. However, in performing the experiment, we really cannot know exactly where the source is. For example, for quantum particles, there is an ultimate uncertainty in position of at least the Compton length (Sec. 4.1.17). Hence, in general, the source S will be located at a small position Δx *off-axis*. The result is that the particles will, in reality, obey a different set of probabilities $P(x_n) \neq r(x_n)$. These can be evaluated. Assuming shift invariance (Eq. (1.11)) and 1 : 1 magnification in the system,

$$p(x_n) = r(x_n - \Delta x), \quad \text{or} \quad r(x_n) = p(x_n + \Delta x). \tag{1.65}$$

By the law of large numbers (Frieden, 2001), since M is large the probabilities $p(x_n)$ agree with the occurrences m_n, by the simple rule

$$m_n = Mp(x_n). \tag{1.66}$$

(This takes the conventional, von Mises viewpoint that probabilities measure the frequency of occurrence of actual – not ideal – events (Von Mises, 1936).) Using Eqs. (1.65) and (1.66) in Eq. (1.64) and taking the logarithm gives

$$\ln P = C + \sum_n Mp(x_n) \ln p(x_n + \Delta x) - \sum_n \ln[Mp(x_n)]! \tag{1.67}$$

where C is an irrelevant constant. Since M is large we may use the Stirling approximation $\ln u! \approx u \ln u$, so that

$$\ln P \approx B + M \sum_n p(x_n) \ln\left(\frac{p(x_n + \Delta x)}{p(x_n)}\right), \tag{1.68}$$

where B is an irrelevant constant. The normalization of $p(x_n)$ was also used. Multiplying and dividing Eq. (1.68) by the fine spacing ε allows us to replace the sum by an integral. Also, since P is to be a maximum, so will be $\ln P$. The result is that Eq. (1.68) becomes

$$\ln P \approx \int dx\, p(x) \ln\left(\frac{p(x + \Delta x)}{p(x)}\right) = max. \tag{1.69}$$

after ignoring all multiplicative and additive constants. Noticing the minus sign in Eq. (1.22b), we see that Eq. (1.69) states that

$$I[p(x)] \equiv I = min., \tag{1.70}$$

agreeing with Eq. (1.47).

This approach can be generalized. Regardless of the physical nature of coordinate x, there will always be uncertainty Δx in the actual value of the origin of a PDF $p(x)$. As we saw, this uncertainty is naturally expressed as a "distance measure" I between $p(x)$ and its displaced version $p(x + \Delta x)$ (Eq. (1.69)).

It is interesting to compare this approach with the derivation of the I-theorem in Sec. 1.8.3. That was based purely on the assumption that the Fokker–Planck equation is obeyed. By comparison, here the assumptions are that (i) maximum probable PDFs actually occur (basic premise of statistical mechanics) and (ii) the system admits of an ultimate resolution "length" Δx of finite extent.

The two derivations may be further compared on the basis of effective "resolution lengths." In Sec. 1.8.3 the limit $\Delta x \to 0$ is rigorously taken, since Δx is, there, just a *mathematical* artifact (which enables I to be expressed as the cross-entropy via Eq. (1.22b)). Also, the approach by Plastino *et al.* to the I-theorem that is mentioned in that section does not even use the concept of Δx. By contrast, in the current derivation Δx is not merely of mathematical origin. It originates physically, as an ultimate resolution length and, hence, is small but intrinsically *finite*. This means that the transition from the cross-entropy on the right-hand side of Eq. (1.69) to information I via Eq. (1.22b) is, here, only an approximation.

If one takes this derivation seriously, then an important effect follows. Since I is only an approximation on the scale of Δx, the use of $I[q(x)]$ in any variational principle (such as EPI) must give solutions $q(x)$ that *lose their validity at scales finer than* Δx. For example, Δx results as the Compton length in the EPI derivation of quantum mechanics (Sec. 4.1.17). A ramification is that quantum mechanics is not represented by its famous wave equations at such scales.

This is a somewhat moot point, since then accurate observations at that scale could not be made anyhow. Nevertheless, it suggests that a different kind of mechanics ought to hold at scales finer than Δx. Such considerations of course lead one to thoughts of quantum gravity (Misner *et al.*, 1973, p. 1193); see also Chapter 11. This is a satisfying transition from a physical point of view. Also, from the statistical viewpoint, it says that EPI is a complete theory insofar as defining the limits of its range of validity.

Klein–Gordon equation (Chapter 4) and the vector wave equation of electrodynamics (Chapter 5).

1.8.15 Multiple PDF cases

In all of the preceding, there was one, scalar parameter θ to be estimated. This implies an information Eq. (1.2) that may be used to predict a single-component PDF $p(x)$ on scalar fluctuations x, as sketched in Sec. 1.8.8. Many phenomena are indeed describable by such a PDF. For example, in statistical mechanics the Boltzmann law $p(E)$ defines the single-component PDF on scalar energy fluctuations E (Chapter 7).

Of course, however, nature is not that simple. There are physical phenomena that require *multiple* PDFs $p_n(\mathbf{x})$, $n = 1, \ldots, N$ or amplitude functions $q_n(\mathbf{x})$, $n = 1, \ldots, N$ for their description. Also, the fluctuations $\mathbf{x}$ might be vector quantities (as indicated by the boldface). For example, in relativistic quantum mechanics there are four wave functions and, correspondingly, four PDFs to be determined (actually, we will find eight real wave functions $q_n(\mathbf{x})$, $n = 1, \ldots, 8$, corresponding to the real and imaginary parts of four complex wave functions). To derive a multiple-component, vector phenomenon, it turns out, requires use of the Fisher information defining the estimation quality of *multiple* vector parameters. This is the subject of the next chapter.

As before, the data are 'intrinsic' in that their fluctuations $\mathbf{x}_n$ are presumed to characterize solely the phenomenon under measurement. There is, e.g., no additional fluctuation due to noise in the instrument providing the measurements. An information measure I of fluctuations $\mathbf{x}_n$ will likewise be intrinsic to the phenomenon, as we required at the outset.

How realistic is this model? In Chapter 10 we will find that immediately before *real* measurements are made the physics of the intrinsic fluctuations $\mathbf{x}_n$ is independent of instrument noise. Indeed, how could fluctuations prior to a measurement depend upon the measuring system? Hence, ignoring the instrumental errors in this *model, prior* measurement scenario agrees with the *real*, *prior* measurement scenario. We call this scenario the 'intrinsic' data scenario.

For simplicity of notation, it is convenient to define "grand" vectors $\boldsymbol{\theta}$, $\mathbf{y}$, $d\mathbf{y}$ over all n as

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N), \quad \mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N), \quad d\mathbf{y} = d\mathbf{y}_1 \cdots d\mathbf{y}_N. \tag{2.2}$$

2.1.2 Aim of the data collection

In chapters to follow, the intrinsic data $\mathbf{y}_n$ will be analyzed presuming either of two general aims: (i) to estimate $\boldsymbol{\theta}_n$ *per se*, as when $\boldsymbol{\theta}_n$ is the nth four-position of a material particle; or (ii) to estimate a *function* of each $\boldsymbol{\theta}_n$, as when $\boldsymbol{\theta}_n$ is an ideal four-position and the electromagnetic four-potential $\mathbf{A}(\boldsymbol{\theta}_n)$ is required.

For either scenario (i), (ii), we want to form a scalar information measure that defines the quality of the $\mathbf{y}_n$. The answer should, intuitively, resemble the form of Eq. (1.2).

2.1.3 Assumption of independence

Suppose that $\boldsymbol{\theta}_n$, $\mathbf{x}_n$, and $\mathbf{y}_n$ are statistically independent. In particular, the data $\mathbf{y}_n$ are then collected efficiently. This is the usual goal in data collection. It can be accomplished by two different experimental procedures: (a) N independent repetitions of the experiment under the same initial conditions, measuring $\boldsymbol{\theta}_n$ at each; or (b) in the case of measurements upon particles, one experiment upon N particles, measuring the N different parameters $\boldsymbol{\theta}_n$ that ensue from one set of initial conditions. In scenario (a), independence is automatically satisfied. Scenario (b) tries to induce ergodicity in the data $\mathbf{y}_n$, e.g., by measuring particles that are sufficiently separated in one or more coordinates.

2.1.4 Real or imaginary coordinates

Each coordinate x_n of a four-vector $\mathbf{x}$ is, in general, either purely real or purely imaginary (Frieden and Soffer, 1995); also see Sec. 1.8.14 and Appendix C. An example of 'mixed' real and imaginary coordinates is given in Sec. 4.1.2.

2.2 Optimum, unbiased estimates

An observer's aim is generally to learn as much as possible about the parameters $\boldsymbol{\theta}$. For this purpose, an optimum estimate

$$\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\mathbf{y}) \tag{2.3}$$

of each four-parameter $\boldsymbol{\theta}_n$ may be fashioned. Each estimate is, thus, a general function of all the data. An example of such an estimator is simply $\mathbf{y}_n$, i.e. the corresponding data vector, but this will usually not be optimum. One well-known class of optimum estimators is the "maximum likelihood" estimator class discussed in Chapter 1.

2.2.1 Unbiased estimators

As with the case of "good" experimental apparatus, the estimators are assumed to be unbiased, i.e., to obey

$$\langle \hat{\boldsymbol{\theta}}_n(\mathbf{y}) \rangle \equiv \int d\mathbf{y}\, \hat{\boldsymbol{\theta}}_n(\mathbf{y}) p(\mathbf{y}|\boldsymbol{\theta}) = \boldsymbol{\theta}_n, \tag{2.4}$$

where $p(\mathbf{y}|\boldsymbol{\theta})$ is the conditional probability of all data $\mathbf{y}$ in the presence of all parameters $\boldsymbol{\theta}$. Equation (2.4) says that, although a given estimate will generally be in error, on average it will be correct. How *small* the error may be, is next established. This introduces the vital concept of information.

2.2.2 Cramer–Rao inequalities

We temporarily suppress index n and focus attention on the four components of any one (n fixed) foursome of *scalar* values $\boldsymbol{\theta}_v$, y_v, x_v, $v = 0, 1, 2, 3$. The mean-square errors from the true values $\boldsymbol{\theta}_v$ are

$$e_v^2 \equiv \int d\mathbf{y}\, [\hat{\boldsymbol{\theta}}_v(\mathbf{y}) - \boldsymbol{\theta}_v]^2 p(\mathbf{y}|\boldsymbol{\theta}). \tag{2.5}$$

Since the data are independent, each mean-square error obeys complementarity with an 'information' quantity I_v,

$$e_v^2 I_v \geqslant 1, \tag{2.6}$$

where

$$I_\nu \equiv \int d\mathbf{y} \left[\frac{\partial \ln p(\mathbf{y}|\boldsymbol{\theta})}{\partial \theta_\nu}\right]^2 p(\mathbf{y}|\boldsymbol{\theta}). \tag{2.7}$$

(See Appendix B for derivation, where $I_\nu \equiv F_{\nu\nu}$.) Equations (2.6) and (2.7) comprise Cramer–Rao inequalities for our vector quantities. They hold for either real or *imaginary* components θ_ν; see Appendix C. When equality is attained in (2.6), the minimum possible error e_ν^2 is attained. Then the estimator is called "efficient." Quantity I_ν is the νth element along the diagonal of the Fisher information matrix [F]. The I_ν thus comprise a *vector* of informations.

2.3 Stam's information

We are now in a position to decide how to construct, from the vector of informations I_ν, the single scalar information quantity I that we seek. *Regaining subscripts* n and summing on Eq. (2.6) gives

$$\sum_n \sum_\nu 1/e_{n\nu}^2 \leqslant \sum_n \sum_\nu I_{n\nu}. \tag{2.8}$$

Each term in the left-hand sum was called an "intrinsic accuracy" by Fisher. Stam (1959a) proposed using the sum as a scalar information measure I_s for a vector scenario (as here),

$$I_s \equiv \sum_n \sum_\nu 1/e_{n\nu}^2 \leqslant \sum_n \sum_\nu I_{n\nu}, \tag{2.9}$$

where the inequality is due to Eq. (2.8). Stam's information is promising, since it is a scalar quantity. We adapt it to our purposes.

2.3.1 Exercise

Stam's information I_s, in depending explicity upon the error variances, ignores all possible error cross-correlations. However, for our additive error case (2.1), where the data y_n are independent and the estimators are unbiased (2.4), all error cross-correlations are zero. Show this.

2.3.2 Trace form, channel capacity, efficiency

The right-hand side of Eq. (2.9) is seen to be an upper bound to I_s. Assume that efficient estimators are used. Then, the equality is attained in Eq. (2.6), so that each left-hand term $1/e_{n\nu}^2$ of Eq. (2.8) equals its corresponding information value $I_{n\nu}$. This means that the upper bound in Eq. (2.9) is realized. An analogous situation arises in the theory of Shannon information. There, the

channel capacity, denoted as C, denotes the maximum possible amount of information that may be passed by a channel (Reza, 1961). Hence, we likewise define a capacity C for the estimation procedure to convey Fisher information I_s about the intrinsic system,

$$I \equiv C = \sum_n \int d\mathbf{y}\, p(\mathbf{y}|\boldsymbol{\theta}) \sum_\nu \left(\frac{\partial \ln p(\mathbf{y}|\boldsymbol{\theta})}{\partial \theta_{n\nu}} \right)^2, \tag{2.10}$$

the latter due to (2.7). It is interesting that this is the trace of the Fisher information matrix (see Appendix B, Eq. (B7)). This information also satisfies our goal of measuring the intrinsic *disorder* of the phenomenon under measurement (see Sec. 2.8). It also measures its *complexity* (Sec. 2.4.1).

2.3.3 Exercise

Taking the trace of the Fisher information matrix ignores, of course, all off-diagonal elements. However, because we have assumed independent data, the off-diagonal elements are, in fact, zero. Show this.

The trace operation (sum over n) in Eq. (2.10) has many physical connotations, in particular *relativistic invariance*, as shown in Sec. 3.5.1.

2.4 Physical modifications

The channel capacity Eq. (2.10) simplifies, in steps, due to various physical aspects of the intrinsic measurement scenario.

2.4.1 Additivity of the information; a measure of complexity

Additivity was previously shown (Sec. 1.8.11) for the case of a single parameter. The generalization to a vector of parameters is now taken up. As will be seen, because of the lack of "cross talk" between different data and parameters, the proof below is a little easier.

Since the intrinsic data are collected independently (Sec. 2.1.3), the joint probability of all the data separates into a product of marginal laws

$$p(\mathbf{y}|\boldsymbol{\theta}) = \prod_n p_n(\mathbf{y}_n|\boldsymbol{\theta}) = \prod_n p_n(\mathbf{y}_n|\boldsymbol{\theta}_n). \tag{2.11}$$

The latter equality follows since by Eq. (2.1), $\boldsymbol{\theta}_m$ has no influence on $\mathbf{y}_n$, $m \neq n$. Taking the logarithm of Eq. (2.11) and differentiating then gives

$$\frac{\partial \ln p(\mathbf{y}|\boldsymbol{\theta})}{\partial \theta_{n\mu}} = \frac{1}{p_n} \frac{\partial p_n}{\partial \theta_{n\mu}}, \quad p_n \equiv p_n(\mathbf{y}_n|\boldsymbol{\theta}_n). \tag{2.12}$$

The aim of this book is to show that information is at the root of all fields of science. These fields may be generated by use of the concept of "extreme physical information" or EPI. The physical information is defined to be the loss of Fisher information resulting from observing any given scientific phenomenon. The act of observation sets off a physical process that may be modeled as a mathematical game between the observer and a "demon" characterizing the phenomenon. The currency of the game is Fisher information. The output of the game is the distribution law describing the statistics of the effect and, in particular, the acquired data. Thus, in a sense, the act of measurement creates the very law that governs the measurement. It is self-realized. This adapted edition of *Physics from Fisher Information* has been rewritten throughout in addition to the inclusion of much new material.

From reviews of the original book

"... suitable for advanced undergraduates, especially those with a strong theory background, or beginning graduate students. The book has two attractive features: the frequent discussions in basic physical terms, and the candid way in which the author describes how he has developed his thoughts ... I urge the readers of this review to take a good look at the book, which is well-written and certainly thought-provoking."
American Journal of Physics

"Frieden's information-based methods provide a stunningly clear interpretation of the laws of physics ... Unlocking the fundamental laws is impressive enough, but if this one principle really is the key to all physics, it should do more than reproduce what physicists already know. It should also reveal the secrets of unsolved mysteries."
New Scientist

"Frieden has proposed a principle that nicely incorporates the transfer of information between a measurement and the physical system that is being measured and from which, almost miraculously, all known laws of physics can be derived."
The Physics Teacher

"One suspects that the ideas it contains ... will continue to be discussed for generations to come ... It seems safe to conclude, all in all, that the unexpected union between physics and Fisher information will prove both lasting and fruitful."
SIAM News

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org
ISBN 0-521-00911-1
9 780521 009119