

# PHYSICS FROM FISHER INFORMATION

A Unification

This book defines and develops a unifying principle of physics, that of ‘extreme physical information’. The information in question is, perhaps surprisingly, not Shannon or Boltzmann entropy but, rather, Fisher information, a simple concept little known to physicists.

Both statistical and physical properties of Fisher information are developed. This information is shown to be a physical measure of disorder, sharing with entropy the property of monotonic change with time. The information concept is applied ‘phenomenally’ to derive most known physics, from statistical mechanics and thermodynamics to quantum mechanics, the Einstein field equations, and quantum gravity. Many new physical relations and concepts are developed including new definitions of disorder, time and temperature. The information principle is based upon a new theory of measurement, one which incorporates the observer into the phenomenon that he/she observes. The ‘request’ for data creates the law that, ultimately, gives rise to the data. The observer creates his or her local reality.

This fascinating work will be of great interest to students and researchers from all areas of physics with an interest in new ways of looking at the subject.



PHYSICS FROM FISHER  
INFORMATION

A Unification

B. ROY FRIEDEN

*Optical Sciences Center, The University of Arizona*

 **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 <http://www.cup.cam.ac.uk>  
40 West 20th Street, New York, NY 10011-4211, USA <http://www.cup.org>  
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

& Cambridge University Press 1998

This book is in copyright. Subject to statutory exception  
and to the provisions of relevant collective licensing agreements,  
no reproduction of any part may take place without  
the written permission of Cambridge University Press.

First published 1998  
Reprinted 1999

Printed in the United Kingdom at the University Press, Cambridge

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

*A catalogue record for this book is available from the British Library*

*Library of Congress Cataloguing in Publication data*

Frieden, B. Roy, 1936–  
Physics from Fisher information : a unification / B. Roy Frieden.  
p. cm.  
Includes index.  
ISBN 0-521-63167-X (hardbound)  
1. Physical measurements. 2. Information theory. 3. Physics-  
-Methodology. I. Title.  
QC39.F75 1998  
530.8-dc21 98-20461 CIP

ISBN 0 521 63167 X hardback

To my wife Sarah  
and to my children Mark, Amy and Miriam



# Contents

0	Introduction	<i>page</i> 1
	0.1 Aim of the book	1
	0.2 Level of approach	4
	0.3 Calculus of variations	5
	0.4 Dirac delta function	19
1	What is Fisher information?	22
	1.1 On Lagrangians	23
	1.2 Classical measurement theory	26
	1.3 Comparisons of Fisher information with Shannon's form of entropy	29
	1.4 Relation of $I$ to Kullback–Leibler entropy	31
	1.5 Amplitude form of $I$	33
	1.6 Efficient estimators	33
	1.7 Fisher $I$ as a measure of system disorder	35
	1.8 Fisher $I$ as an entropy	36
2	Fisher information in a vector world	51
	2.1 Classical measurement of four-vectors	51
	2.2 Optimum, unbiased estimates	53
	2.3 Stam's information	54
	2.4 Physical modifications	55
	2.5 Net probability $p(x)$ for a particle	59
	2.6 The two types of 'data' used in the theory	60
	2.7 Alternative scenarios for the channel capacity $I$	60
	2.8 Multiparameter $I$ -theorem	62
3	Extreme physical information	63
	3.1 Covariance, and the 'bound' information $J$	63
	3.2 The equivalence of entropy and Shannon information	65
	3.3 System information model	69

3.4	Principle of extreme physical information (EPI)	71
3.5	Derivation of Lorentz group of transformations	84
3.6	Gauge covariance property	89
3.7	Field dynamics from information	90
3.8	An optical measurement device	92
3.9	EPI as a state of knowledge	106
3.10	EPI as a physical process	107
3.11	On applications of EPI	109
4	Derivation of relativistic quantum mechanics	112
4.1	Derivation of Klein–Gordon equation	112
4.2	Derivation of Dirac equation	122
4.3	Uncertainty principles	127
4.4	Overview	130
5	Classical electrodynamics	134
5.1	Derivation of vector wave equation	134
5.2	Maxwell’s equations	156
5.3	Overview	157
6	The Einstein field equation of general relativity	161
6.1	Motivation	161
6.2	Tensor manipulations: an introduction	162
6.3	Derivation of the weak-field wave equation	165
6.4	Einstein field equation and equations of motion	177
6.5	Overview	177
7	Classical statistical physics	179
7.1	Goals	179
7.2	Covariant EPI problem	179
7.3	Boltzmann probability law	181
7.4	Maxwell–Boltzmann velocity law	187
7.5	Fisher information as a bound to entropy increase	194
7.6	Overview	203
8	Power spectral $1=f$ noise	206
8.1	The persistence of $1=f$ noise	206
8.2	Temporal evolution of tone amplitude	208
8.3	Use of EPI principle	210
8.4	Overview	215
9	Physical constants and the $1=x$ probability law	216
9.1	Introduction	216
9.2	Can the constants be viewed as random numbers?	218
9.3	Use of EPI to find the PDF on the constants	218
9.4	Statistical properties of the $1=x$ law	226



9.5	What histogram of numbers do the constants actually obey?	230
9.6	Overview	233
10	Constrained-likelihood quantum measurement theory	235
10.1	Introduction	235
10.2	Measured coordinates	236
10.3	Likelihood law	238
10.4	Instrument noise properties	239
10.5	Final log-likelihood form	239
10.6	EPI variational principle with measurements	240
10.7	Klein–Gordon equation with measurements	240
10.8	On the Dirac equation with measurements	241
10.9	Schroedinger wave equation with measurements	242
10.10	Overview	249
11	Research topics	254
11.1	Scope	254
11.2	Quantum gravity	254
11.3	Nearly incompressible turbulence	265
12	Summing up	273
Appendix A	Solutions common to entropy and Fisher $I$ -extremization	283
Appendix B	Cramer–Rao inequalities for vector data	287
Appendix C	Cramer–Rao inequality for an imaginary parameter	291
Appendix D	Simplified derivation of the Schroedinger wave equation	294
Appendix E	Factorization of the Klein–Gordon information	296
Appendix F	Evaluation of certain integrals	301
Appendix G	Schroedinger wave equation as a non-relativistic limit	303
Appendix H	Non-uniqueness of potential $\mathbf{A}$ for finite boundaries	305
References		307
Index		312

# 1

## What is Fisher information?

Knowledge of Fisher information is not part of the educational background of most physicists. Why should a physicist bother to learn about this concept? Surely the (related) concept of entropy is sufficient to describe the degree of disorder of a given phenomenon. These important questions may be answered as follows.

- (a) The point made about entropy is true, but does not go far enough. Why not seek a measure of disorder whose variation *derives* the phenomenon? The concept of entropy cannot do this, for reasons discussed in Sec. 1.3. Fisher information will turn out to be the appropriate measure of disorder for this purpose.
- (b) Why should a physicist bother to learn this concept? Aside from the partial answer in (a): (i) Fisher information is a *simple* and intuitive concept. As theories go, it is quite elementary. To understand it does not require mathematics beyond differential equations. Even no prior knowledge of statistics is needed: this is easy enough to learn ‘on the fly’. The derivation of the defining property of Fisher information, in Sec. 1.2.3, is readily understood. (ii) The subject has very little specialized jargon or notation. The beginner does not need a glossary of terms and symbols to aid in its understanding. (iii) Most importantly, once understood, the concept gives strong payoff – one might call it ‘phenomen-all’ – in scope of application. It’s simply worth learning.

Fisher information has two basic roles to play in theory. First, it is a measure of the ability to estimate a parameter; this makes it a cornerstone of the statistical field of study called parameter estimation. Second, it is a measure of the state of disorder of a system or phenomenon. As will be seen, this makes it a cornerstone of physical theory.

Before starting the study of Fisher information, we take a temporary detour into a subject that will provide some immediate physical motivation for it.



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 if ‘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 remains a mystery: Schroedinger called it ‘incomprehensible’ (Schroedinger, 1926).

*Historical note:* As will become evident below, *Schroedinger’s mysterious Lagrangian term was simply Fisher’s data information.* May we presume from this that Schroedinger 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.

What we will show is that, in general, the squared gradient represents a phenomenon that is natural to all fields, i.e., *information*. In particular, it is the amount of Fisher information residing in a variety of data called *intrinsic data*.

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 Mech.	$\frac{1}{2}m\left(\frac{\partial q}{\partial t}\right)^2 - V$
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 Eq.	$-\nabla\psi\cdot\nabla\psi^* - \dots$
Schrödinger W. E.	$-\frac{\hbar^2}{2m}\nabla\psi\cdot\nabla\psi^* - \dots$
Klein–Gordon Eq.	$-\frac{\hbar^2}{2m}\nabla\psi\cdot\nabla\psi^* - \dots$
Elastic W. E.	$\frac{1}{2}\rho\dot{q}^2 - \dots$
Electromagnetic Eqs.	$4\sum_{n=1}^4\Box q_n\cdot\Box q_n - \dots$
Dirac Eqs.	$-\frac{\hbar^2}{2m}\nabla\psi\cdot\nabla\psi^* - \dots = 0$
General Relativity (Eqs. of motion)	$\sum_{m,n=1}^4 g_{mn}(q(\tau))\frac{\partial q_m}{\partial\tau}\frac{\partial q_n}{\partial\tau}$ <div style="text-align: center;"> <math>\uparrow</math>                      metric tensor                 </div>
Boltzmann Law	$4\left(\frac{\partial q(E)}{\partial E}\right)^2 - \dots, p(E) \equiv q^2(E)$
Maxwell–Boltzmann Law	$4\left(\frac{\partial q(v)}{\partial v}\right)^2 - \dots, p(v) \equiv q^2(v)$
Lorentz Transformation (special relativity)	$\partial_i q_n \partial_i q_n$ (invariance of integral)
Helmholtz W. E.	$-\nabla\psi\cdot\nabla\psi^* - \dots$

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 value  $\theta$ , and let there be  $N$  data values  $y_1, \dots, y_N \equiv \mathbf{y}$  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  $\mathbf{y} = \theta + \mathbf{x}$ , where the  $x_1, \dots, x_N \equiv \mathbf{x}$  are added noise values. The data are used in an estimation principle to form an estimate of  $\theta$  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  $\theta$  than is any one of the data observables.

The noise  $\mathbf{x}$  is assumed to be *intrinsic* to the parameter  $\theta$  under measurement. For example,  $\theta$  and  $\mathbf{x}$  might be, respectively, the ideal position and quantum fluctuations of a particle. Data  $\mathbf{y}$  are, correspondingly, called *intrinsic data*. No additional noise effects, such as noise of detection, are assumed present here. (We later allow for such additional noise in Sec. 3.8 and Chap. 10.) The system consisting of quantities  $\mathbf{y}$ ,  $\theta$ ,  $\mathbf{x}$  is a *closed*, or physically isolated, one.

### 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)

$$e^2 I \geq 1, \quad (1.1)$$

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

$$I = \int dx p'^2(x)/p(x), \quad p' \equiv dp/dx. \quad (1.2)$$

(Throughout the book, integration limits are infinite unless otherwise specified.) Quantity  $p(x)$  denotes the probability density function for the noise value

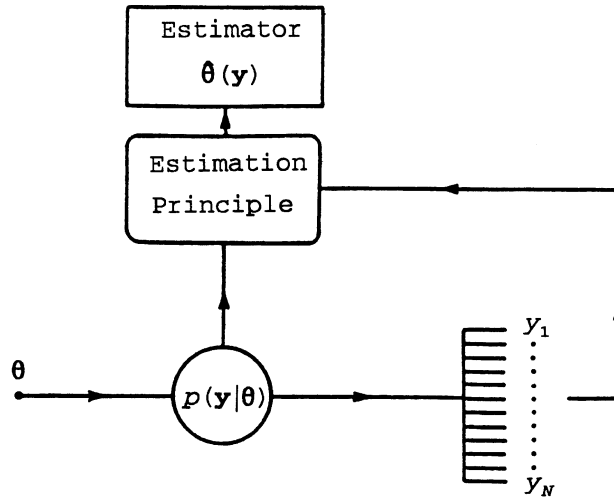


Fig. 1.1. The parameter estimation problem of classical statistics. An unknown but fixed parameter value  $\theta$  causes intrinsic data  $\mathbf{y}$  through random sampling of a likelihood law  $p(\mathbf{y}|\theta)$ . Then, the random likelihood law and the data are used to form the estimator  $\hat{\theta}(\mathbf{y})$  via an estimation principle. (Reprinted from Frieden, 1991, by permission of Springer-Verlag Publishing Co.)

x. If  $p(x)$  is Gaussian, then  $I = 1/\sigma^2$  with  $\sigma^2$  the variance (see derivation in Sec. 8.3.1).

Eq. (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 at 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  $\theta$ . 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’. Eqs. (1.1) and (1.2) derive quite easily, shown next.

### 1.2.3 Derivation

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

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

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

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

Use the identity

$$\frac{\partial p}{\partial \theta} = p \frac{\partial \ln p}{\partial \theta} \quad (1.5)$$

and the fact that  $p$  obeys normalization. Then Eq. (1.4) becomes

$$\int d\mathbf{y} (\hat{\theta} - \theta) \frac{\partial \ln p}{\partial \theta} p = 1. \quad (1.6)$$

Factor the integrand as

$$\int d\mathbf{y} \left[ \frac{\partial \ln p}{\partial \theta} \sqrt{p} \right] [(\hat{\theta} - \theta) \sqrt{p}] = 1. \quad (1.7)$$

Square the equation. Then the Schwarz inequality gives

$$\left[ \int d\mathbf{y} \left( \frac{\partial \ln p}{\partial \theta} \right)^2 p \right] \left[ \int d\mathbf{y} (\hat{\theta} - \theta)^2 p \right] \geq 1. \quad (1.8)$$

The left-most factor is defined to be the Fisher information  $I$ ,

$$I \equiv I(\theta) \equiv \int d\mathbf{y} \left( \frac{\partial \ln p}{\partial \theta} \right)^2 p, \quad p \equiv p(\mathbf{y}|\theta), \quad (1.9)$$

while the second factor exactly defines the mean-squared error  $e^2$ ,

$$e^2 \equiv \int d\mathbf{y} [\hat{\theta}(\mathbf{y}) - \theta]^2 p. \quad (1.10)$$

This proves Eq. (1.1).

It is noted that  $I = I(\theta)$  in Eq. (1.9), i.e., in general  $I$  depends upon the (fixed) value of parameter  $\theta$ . But note the following important exception to this rule.

#### 1.2.4 Important case of shift invariance

Suppose that there is only  $N = 1$  data value taken so that  $p(\mathbf{y}|\theta) = p(y|\theta)$ . Also, suppose that the PDF obeys a property

$$p(y|\theta) = p(y - \theta). \quad (1.11)$$

This means that the fluctuations in  $y$  from  $\theta$  are invariant to the size of  $\theta$ , a



kind of shift invariance. (This becomes an expression of *Galilean invariance* when random variables  $y$  and  $\theta$  are 3-vectors instead.) Using condition (1.11) and identity (1.5) in Eq. (1.9) gives

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

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

### 1.3 Comparisons of Fisher information with Shannon's form of entropy

A related quantity to  $I$  is the Shannon entropy (Shannon, 1948)  $H$  (called Shannon 'information' in this book). This has the form

$$H \equiv - \int dx p(x) \ln p(x). \quad (1.13)$$

Like  $I$ ,  $H$  is a functional of an underlying probability density function (PDF)  $p(x)$ . Historically,  $I$  predates the Shannon form by about 25 years (1922 vs. 1948). There are some known relations connecting the two information concepts (Stam, 1959; Blachman, 1965; Frieden, 1991) but these are not germane to our purposes.  $H$  can be, but is not always, the thermodynamic, Boltzmann entropy.

The analytic properties of the two information measures are quite different. Thus, whereas  $H$  is a *global* measure of smoothness in  $p(x)$ ,  $I$  is a *local* measure. Hence, when extremized through variation of  $p(x)$ , Fisher's form gives a differential equation while Shannon's always gives directly the same form of solution, an exponential function. These are shown next.

#### 1.3.1 Global vs. local nature

For our purposes, it is useful to work with a discrete form of Eq. (1.13),

$$H = -\Delta x \sum_n p(x_n) \ln p(x_n) \equiv \delta H, \Delta x \rightarrow 0. \quad (1.14)$$

(Notation  $\delta H$  emphasizes that Eq. (1.14) represents an *increment* in information.) Of course, the sum in Eq. (1.14) may be taken in any order. Graphically, this means that if the curve  $p(x_n)$  undergoes a rearrangement of its points  $(x_n, p(x_n))$ , although the shape of the curve will drastically change the value of

$H$  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),

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

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  $\Delta x \rightarrow 0$ , and so apply to the continuous probability density  $p(x)$  as well.

This global vs. local property has an interesting ramification. 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 Chap. 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,b). 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 Chap. 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 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.

### 1.3.2 Additivity properties

It is of further interest to compare  $I$  and  $H$  in the special case of mutually isolated systems, which give rise to independent data. As is well-known, the entropy  $H$  obeys additivity in this case. Indeed, many people have been led to believe that, because  $H$  has this property, it is *the only* functional of a probability law that obeys additivity. In fact, information  $I$  obeys additivity as well. This will be shown in Sec. 1.8.11.

### 1.4 Relation of $I$ to Kullback–Leibler entropy

The Kullback–Leibler entropy  $G$  (Kullback, 1959) is a functional of (now) two PDFs  $p(x)$  and  $r(x)$ ,

$$G \equiv - \int dx p(x) \ln [p(x)/r(x)]. \quad (1.16)$$

This is also called the ‘cross entropy’ or ‘relative entropy’ between  $p(x)$  and a reference PDF  $r(x)$ . Note that if  $r(x)$  is a constant,  $G$  becomes essentially the entropy  $H$ . Also,  $G = 0$  if  $p(x) = r(x)$ . Thus,  $G$  is often used as a measure of the ‘distance’ between two PDFs  $p(x)$  and  $r(x)$ . Also, in a multidimensional case  $x \rightarrow (x, y)$  the information  $G$  can be used to define the mutual information of Shannon (1948).

Now we show that the Fisher information  $I$  relates to  $G$ . Using Eq. (1.15), with  $x_{n+1} = x_n + \Delta x$ ,  $I$  may be expressed as

$$I = \Delta x^{-1} \sum_n p(x_n) \left[ \frac{p(x_n + \Delta x)}{p(x_n)} - 1 \right]^2 \quad (1.17)$$

in the limit  $\Delta x \rightarrow 0$ . Now quantity  $p(x_n + \Delta x)/p(x_n)$  is close to *unity* since  $\Delta x$  is small. Therefore, the [·] quantity in (1.17),

$$p(x_n + \Delta x)/p(x_n) - 1 \equiv \nu, \quad (1.18)$$

is small. Now for small  $\nu$  the expansion

$$\ln(1 + \nu) = \nu - \nu^2/2 \quad (1.19)$$

holds, or equivalently,

$$\nu^2 = 2[\nu - \ln(1 + \nu)]. \quad (1.20)$$

Then by Eqs. (1.18) and (1.20), Eq. (1.17) becomes

$$\begin{aligned} I = & -2\Delta x^{-1} \sum_n p(x_n) \ln \frac{p(x_n + \Delta x)}{p(x_n)} \\ & + 2\Delta x^{-1} \sum_n p(x_n + \Delta x) - 2\Delta x^{-1} \sum_n p(x_n). \end{aligned} \quad (1.21)$$

But each of the two far-right sums is  $\Delta x^{-1}$ , by normalization, so that their difference cancels, leaving

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

$$\rightarrow - (2/\Delta x^2) \int dx p(x) \ln \frac{p(x + \Delta x)}{p(x)} \quad (1.22b)$$

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

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

#### 1.4.1 Historical note

Vstovsky (1995) first proved the converse of the preceding, that  $I$  is an approximation to  $G$ . However, the expansion contains lower-order terms as well, in distinction to the effect in (1.21) where our lower-order terms cancel out exactly.

#### 1.4.2 Exercise

One notes that the form (1.22b) is indeterminate 0/0 in the limit  $\Delta x \rightarrow 0$ . 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

Eq. (1.22b) 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:

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

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

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

called '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  $\Delta x$ , 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  $p(x) \rightarrow 0$  at certain  $x$ ?) A way out is to work with a real ‘amplitude’ function  $q(x)$ ,

$$p(x) = q^2(x). \quad (1.23)$$

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

$$I = 4 \int dx q'^2(x). \quad (1.24)$$

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 independent 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)

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

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  $B(\mathbf{y})$ ,

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

Here the continuous index  $\mathbf{y}$  defines the  $y$ 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,

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

(Note that function  $k(\theta)$  remains constant since the parameter  $\theta$  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),

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

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

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

$$e_{\min}^2 = 1/I. \quad (1.28)$$

The estimator  $\hat{\theta}(\mathbf{y})$  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  $p(\mathbf{y}|\theta) = A \prod_n \exp [-(y_n - \theta)^2/2\sigma^2]$ ,  $A = \text{const.}$ , and (b) the exponential law  $p(\mathbf{y}|\theta) = \prod_n e^{-y_n/\theta}/\theta$ ,  $y_n \geq 0$ . On the other hand, with  $N = 1$ , (c) a PDF of the form

$$p(y|\theta) = A \sin^2(y - \theta), \quad A = \text{const.}, \quad |y - \theta| \leq \pi$$

does not satisfy (1.27). Note that this PDF arises when the position  $\theta$  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  $\theta$  (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 Chap. 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 uniform or ‘unbiased’ probability density function  $p(x)$ . Such a curve is shown in Fig. 1.2b. The curve has small gradient content (if it is physically meaningful, i.e., is piecewise continuous). Simply stated, *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. 1.2a. 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.

On the other hand, the ability to measure disorder is usually associated with the word ‘entropy’. For example, the Shannon entropy  $H$  is known to measure the degree of disorder of a system. (Example: By direct use of Eq. (1.13), if  $p(x)$  is normal with variance  $\sigma^2$  then  $H = \ln \sigma + \ln \sqrt{2\pi e}$ . This shows that  $H$  monotonically increases with the ‘width’  $\sigma$  of the PDF, i.e., with the degree of disorder in the system.)

Since we found that  $I$  likewise measures the disorder of a system this suggests that  $I$  ought to likewise be regarded as an ‘entropy’. However, the entropy  $H$  has another important property: When  $H$  is, as well, the *Boltzmann* entropy, it obeys the Second Law of thermodynamics, increasing monotonically *with time*,

$$\frac{dH(t)}{dt} \geq 0. \quad (1.29)$$

Does  $I$ , then, also change monotonically with time? A particular scenario suggests that this is so.

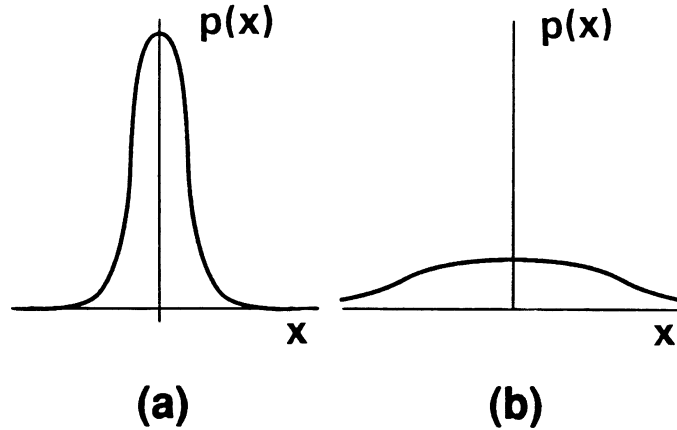


Fig. 1.2. Degree of disorder measured by  $I$  values. In (a), random variable  $x$  shows relatively low disorder and large  $I$  (gradient content). In (b),  $x$  shows high disorder and small  $I$ . (Reprinted from Frieden and Soffer, 1995.)

### 1.8 Fisher $I$ as an entropy

We next show that, for a particular isolated system,  $I$  monotonically decreases with time (Frieden, 1990). All measurements and PDF laws are now taken at a specified time  $t$ , so that the system PDF now has the form  $p(x|t)$ , i.e., the probability of reading  $(x, x + dx)$  conditional upon a time  $(t, t + dt)$ .

#### 1.8.1 Paradigm of the broken urn

Consider a scenario where many particles fill a small urn. Imagine these to be ideal, point masses that collide elastically and that are not in an exterior force field. We want a smart measurement of their mean horizontal position  $\theta$ . Accordingly, a particle at horizontal position  $y$  is observed,  $y = \theta + x$ , where  $x$  is a random fluctuation from  $\theta$ . Define the mean-square error  $e^2(t) = \langle [\theta - \hat{\theta}(y)]^2 \rangle$  due to repeatedly forming estimates  $\hat{\theta}(y)$  of  $\theta$  within a small time interval  $(t, t + dt)$ . How should error  $e$  vary with  $t$ ?

Initially, at  $t = 0$ , the particles are within the small urn. Hence, any observed value  $y$  should be near to  $\theta$ ; then, any good estimate  $\hat{\theta}(y)$  will likewise be close to  $\theta$ , and resultingly  $e^2(0)$  will be small. Next, the walls of the container are broken, so that the particles are free to randomly move away. They will follow, of course, the random walk process which is called Brownian motion (Papoulis, 1965).

Consider a later time interval  $(t, t + dt)$ . For Brownian motion, the PDF  $p(x|t)$  is Gaussian with a variance  $\sigma^2 = Dt$ ,  $D = \text{const.}$ ,  $D \geq 0$ . For a Gaus-



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’

Eq. (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’

$$\frac{dI(t)}{dt} \leq 0. \quad (1.30)$$

### 1.8.3 Proof

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

$$I(t) = -2 \lim_{\Delta x \rightarrow 0} \Delta x^{-2} \int dx p \ln(p_{\Delta x}/p) \quad (1.31)$$

$$p \equiv p(x|t), \quad p_{\Delta x} \equiv p(x + \Delta x|t).$$

Under certain physical conditions, e.g., ‘detailed balance’, short-term correlation, shift-invariant statistics (Gardiner, 1985; Reif, 1965; Risken, 1984)  $p$  obeys a *Fokker–Planck* differential equation

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

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 (1984) shows that two PDFs, such as  $p$  and  $p_{\Delta x}$ , that obey the Fokker–Planck equation have a cross-entropy

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

that obeys an  $H$ -theorem (1.29),

$$\frac{dG(t)}{dt} \geq 0. \quad (1.34)$$

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

(as in Chaps. 4–11) than does the  $H$ -theorem, such conditions must exist; however, they are yet to be found.

It should be remarked that the  $I$ -theorem was first proven (Plastino and Plastino, 1996) from the direct defining form (1.2) for  $I$  (i.e., *without* recourse to the cross-entropy form (1.22b)).

#### 1.8.4 Ramification to definition of time

The  $I$ -theorem (1.30) is an extremely important result. It states that the Fisher information of a physical system can only decrease (or remain constant) in time. Combining this with Eq. (1.28) indicates that  $e_{\min}^2$  must increase, so that even in the presence of efficient estimation *the quality of estimates must decrease with time*. This seems to be a reasonable alternative statement of the Second Law of thermodynamics. If, by the Second Law, the disorder of a system (as measured by the Boltzmann entropy) must increase, then the disorder of any measuring system must increase as well. This must degrade its use as a measuring instrument, causing the error  $e_{\min}^2$  to increase. On this basis, one could estimate the age of an instrument by simply observing how well it measures.

Thus,  $I$  is a measure of physical disorder that has its mathematical roots in estimation theory. By the same token, one may regard the Boltzmann entropy  $H$  to be a measure of physical disorder that has its mathematical roots in communication theory (Shannon, 1948). Communication theory plays a complementary role to estimation theory: the former describes how well messages can be *transmitted*, in the presence of given errors in the channel (system noise properties); the latter describes how accurately messages may be *estimated*, also in the presence of given errors in the channel.

If  $I$  really is a physical measure of system disorder, it ought to somehow relate to temperature, pressure, and all other extrinsic parameters of thermodynamics. This is, in fact, the subject of Secs. 1.8.5–1.8.7.

Next, consider the concept of the flow of thermodynamic time (Zeh, 1992; Halliwell *et al.*, 1994). This concept is intimately tied to that of the Boltzmann entropy: an increase in the latter *defines* the positive passage of Boltzmann time. The  $I$ -theorem suggests an alternative definition to Boltzmann time: a decrease in  $I$  defines an increase in ‘Fisher time’. However, whether the two times always agree is an open question. In numerical experiments on randomly perturbed PDFs (Frieden, 1990), usually the resulting perturbations  $\delta I$  went down when  $\delta H$  went up, i.e., both measures agreed that disorder (and time) increased. They also usually agreed on decreases of disorder. However, there were disagreements about 1% of the time.

### 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 is attained, a situation

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

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  $\theta, \theta'$  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

$$I(E) + I'(E') = \text{Min}. \quad (1.36)$$

On the other hand, energy is conserved, so that

$$E + E' \equiv C, \quad (1.37)$$

$C = \text{constant}$ . (Notice that this is a deterministic relation between the two ideal parameter values, and 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

$$I(E) + I'(C - E) = \text{Min}. \quad (1.38)$$

We now define a generalized ‘Fisher temperature’  $T_\theta$  as

$$\frac{1}{T_\theta} \equiv -k_\theta \frac{\partial I}{\partial \theta}. \quad (1.39)$$

Notice that  $\theta$  is any parameter under measurement. Hence, there is a Fisher ‘temperature’ associated with any parameter to be measured. From (1.39),  $T_\theta$  simply measures the sensitivity of information level  $I$  to a change in system parameter  $\theta$ . The constant  $k_\theta$  gives each  $T_\theta$  value the same units. A relation between the two temperatures  $T$  and  $T_\theta$  is found below for a perfect gas.

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

$$\frac{\partial I}{\partial E} + \frac{\partial I'}{\partial E'}(-1) = 0 \text{ or } T_E = T_{E'} \quad (1.40)$$

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  $\theta$ . This occurs in close to a shift invariant case (1.11), (1.12) 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  $\theta = E$  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, 1991) 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  $\chi^2$  (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  $\bar{p}$  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)

$$\bar{p} = \frac{F dx}{A dx} = -\frac{\partial E}{\partial V} \quad (1.41)$$

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

$$\bar{p} = -\frac{\partial E}{\partial I} \frac{\partial I}{\partial V} = k_E T_E \frac{\partial I}{\partial V}, \quad (1.42)$$

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  $\bar{p}$ ,  $dI$  and  $T_E$  can be functions of the position  $\mathbf{r}$  of volume  $dV$  within a gas. Multiplying (1.42) by  $dV$  gives

$$\bar{p}(\mathbf{r})dV = k_E T_E(\mathbf{r})dI(\mathbf{r}) \quad (1.43)$$

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

$$\bar{p}V = k_E T_E I. \quad (1.44)$$

Note that  $I = \int dI(\mathbf{r})$  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  $\bar{p}$  upon  $V$  and  $T_E$  is of the same form as the known equation of state of the gas

$$\bar{p}V = MkT, \quad (1.45)$$

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

$$\left( \frac{kT}{k_E T_E} \right) = I/M, \quad (1.46)$$

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. At this point in time, the question remains an open one.

### 1.8.8 Ramification to derivations of physical laws

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

$$I(t) = 4 \int dx q'^2(x|t) \rightarrow \text{Min}. \quad (1.47)$$

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):

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

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

The kernel functions  $f_k(x)$ , constraint exponent  $\alpha$  and data values  $F_k$  are assumed known. The multipliers  $\lambda_k$  are found such that the constraint equations (1.48b) are obeyed. See also Huber (1981).

The most difficult step in this 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 Chap. 3 and applied in Chaps. 4–9 and 11 – 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 aims. For this purpose, we may use *Legendre’s condition* for a minimum (Sec. 0.3.3): Let  $\mathcal{L}$  denote the integrand (or Lagrangian) of the total integral to be extremized. In our scalar case, if

$$\frac{\partial^2 \mathcal{L}}{\partial q'^2} > 0 \quad (1.49)$$

the solution will be a minimum. From Eqs. (1.47) and (1.48a), our Lagrangian is

$$\mathcal{L} = 4q'^2 + \sum_k \lambda_k q^\alpha f_k. \quad (1.50)$$

Using this in Eq. (1.49) gives

$$\frac{\partial^2 \mathcal{L}}{\partial q'^2} = +8, \quad (1.51)$$

showing that a minimum is indeed attained.

The foregoing assumed that coordinate  $x$  is real and a scalar. However, most

scenarios will involve *multiple* coordinates due to Lorentz covariance requirements. One or more of these are purely imaginary. For example, in Chap. 4 we use a space coordinate  $x$  that is purely imaginary. The same analysis as the preceding shows that in this case the second derivative (1.51) is negative, so that a maximum is instead attained in this coordinate. However, others of the coordinates are real and, hence, tend to give a minimum. Obviously Legendre's condition cannot give a unique answer in this scenario, and other criteria for the determination must be used. The question of maximum or minimum under such general conditions remains an open question.

Note that these results apply to all physical applications of our variational principle (as formed in Chap. 3). These applications only differ in their effective constraint terms, none of which contains terms in  $q'^2$ .

When  $I$  is minimized by the variational technique,  $q(x|t)$  tends to be maximally smooth (Sect. 1.7). We saw that this describes a situation of maximum disorder. The Second Law of thermodynamics causes, as well, increased disorder. In this behavior, then, the  $I$ -theorem (1.30) acts like the Second Law.

### 1.8.9 Is the $I$ -theorem equivalent to the $H$ -theorem?

We showed at Eq. (1.34) that if a PDF  $p(x)$  and its *infinitesimally* shifted version  $p(x + \Delta x)$  both obey the Fokker–Planck equation, then the  $I$ -theorem follows. On the other hand, Eq. (1.34) with  $\Delta x$  *finite* is an expression of the  $H$ -theorem. Hence, the two theorems have a common pedigree, so to speak. Are, then, the two theorems equivalent? In fact they are not equivalent because the  $I$ -theorem is a *limiting form* (as  $\Delta x \rightarrow 0$ ) of Risken's  $H$ -theorem. Taking the limit introduces the derivative  $p'$  of the PDF into the integrand of what was  $H$ , transforming it into the form Eq. (1.2) of  $I$ . The presence of this derivative in  $I$ , and its absence in  $H$ , has strong mathematical implications. One example is as follows.

The equilibrium solutions for  $p(x)$  that are obtained by extremizing Fisher  $I$  (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 the form of Eq. (1.32), the time rate of change of  $p$  only depends 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 (Chap. 8), a

law famous for exhibiting long-term memory. Also, the relativistic electron obeys an equation of continuity of flow  $\partial p/\partial t = c\nabla \cdot (\boldsymbol{\psi}^*[\alpha]\boldsymbol{\psi})$ ,  $p \equiv \boldsymbol{\psi}^*\boldsymbol{\psi}$  (Schiff, 1955), where all quantities are defined in Chap. 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 (Chap. 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  $\theta$  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  $(x_n, n = 1, \dots, N) \equiv \mathbf{x}$ . The  $\mathbf{x}$  are random because of the Brownian motion. Assuming nonrelativistic speeds, the intrinsic data  $(y_n, n = 1, \dots, N) \equiv \mathbf{y}$  obey simply

$$\mathbf{y} = \theta + \mathbf{x}. \quad (1.52)$$

Assume that the urns are physically isolated from one another by the use of barriers  $B$  (see Fig. 1.3), so that there is no interaction between particles from different urns. Then the data  $\mathbf{y}$  are independent. This causes the likelihood law to break up into a product of factors (Frieden, 1991)



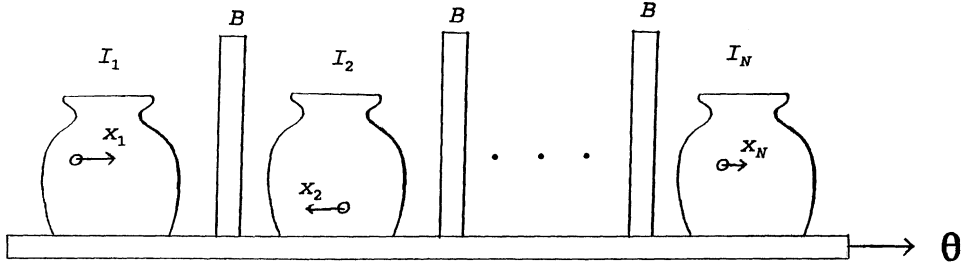


Fig. 1.3.  $N$  urns are moving at a common speed  $\theta$ . Each contains particles in Brownian motion. Barriers  $B$  physically isolate the urns. Measurements  $y_n = \theta + x_n$  of particle velocities are made, one to an urn. Each  $y_n$  gives rise to an information amount  $I_n$ . The total information  $I$  over all the data  $\mathbf{y}$  is the sum of the individual informations  $I_n$ ,  $n = 1, \dots, N$  from the urns.

$$p(\mathbf{y}|\theta) = \prod_{n=1}^N p_n(y_n|\theta), \quad (1.53)$$

where  $p_n$  is the likelihood law for the  $n$ th observation.

We may now compute the information  $I$  in the independent data  $\mathbf{y}$ . By Eq. (1.53)

$$\begin{aligned} \ln p &= \sum_n \ln p_n, \quad p \equiv p(\mathbf{y}|\theta), \quad p_n \equiv p_n(y_n|\theta), \\ \text{so that } \frac{\partial \ln p}{\partial \theta} &= \sum_n \frac{1}{p_n} \frac{\partial p_n}{\partial \theta}. \end{aligned} \quad (1.54)$$

Squaring the latter gives

$$\left( \frac{\partial \ln p}{\partial \theta} \right)^2 = \sum_{\substack{mn \\ m \neq n}} \frac{1}{p_m} \frac{1}{p_n} \frac{\partial p_m}{\partial \theta} \frac{\partial p_n}{\partial \theta} + \sum_n \frac{1}{p_n^2} \left( \frac{\partial p_n}{\partial \theta} \right)^2 \quad (1.55)$$

where the last sum is for indices  $m = n$ . Then the defining Eq. (1.9) for  $I$  gives, with the substitutions Eqs. (1.53), (1.55),

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

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,

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

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

$$\sum_n \int dy_n \frac{\partial p_n}{\partial \theta} \quad (1.58)$$

with a corresponding one in index  $m$ . But

$$\int dy_n \frac{\partial p_n}{\partial \theta} = \frac{\partial}{\partial \theta} \int dy_n p_n = \frac{\partial}{\partial \theta} 1 = 0 \quad (1.59)$$

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

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

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

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

$$I = \sum_{n=1}^N I_n \quad (1.61)$$

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

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

under the independence condition Eq. (1.53) it gives

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

### 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 to uniquely imply entropy as the measure. Undoubtedly many other measures satisfy additivity as well.

### 1.8.13 $I = \text{Min. from statistical mechanics viewpoint}$

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  $(x_n, n = 1, \dots, N)$  with a constant, small spacing  $\varepsilon$ . 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  $n = 1, \dots, N$ . There are  $M$  particles in all, with  $M$  very large. What is the joint probability  $P(m_1, \dots, m_n)$ ?

Each image event is a possible position  $x_n$ , of which there are  $N$ . Therefore the image events comprise an  $N$ -ary events sequence. This obeys a multinomial probability law (Frieden, 1991) of order  $N$ ,

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

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

The ideal source  $S$  for the experiment is a very small aperture that is located on-axis. This situation would give rise to ideal (prior) probabilities  $r(x_n)$ ,  $n = 1, \dots, N$ . 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  $\Delta x$  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,

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

By the law of large numbers (Frieden, 1991), since  $M$  is large the probabilities  $p(x_n)$  agree with the occurrences  $m_n$ , by the simple rule

$$m_n = Mp(x_n). \quad (1.66)$$

(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

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

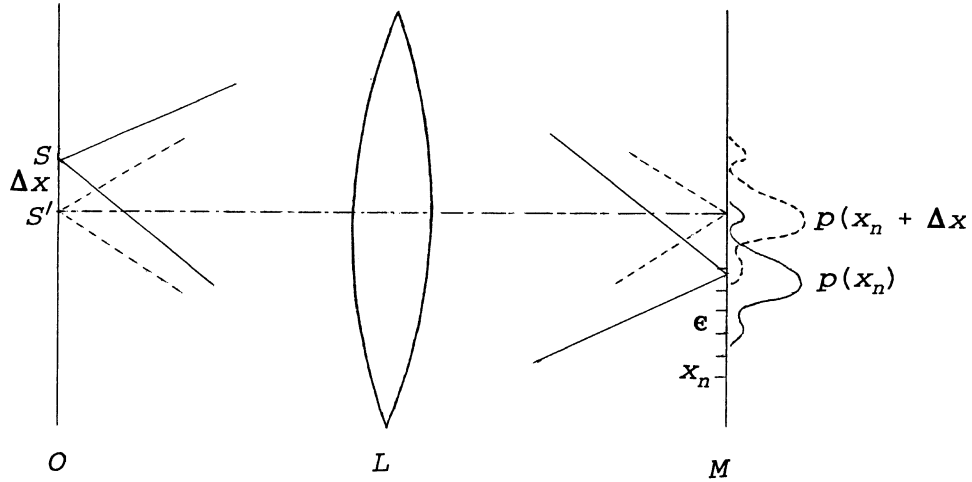


Fig. 1.4. A statistical mechanics view of Fisher information. The ideal point source position  $S'$  gives rise to the ideal PDF  $p(x_n + \Delta x)$ , while the actual point source position  $S$  gives rise to the empirical PDF  $p(x_n)$ . Maximizing the logarithm of the probability of the latter PDF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

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

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

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  $\varepsilon$  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

$$\ln P \approx \int dx p(x) \ln \frac{p(x + \Delta x)}{p(x)} = \text{Max}. \quad (1.69)$$

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

$$I[p(x)] \equiv I = \text{Min.}, \quad (1.70)$$

agreeing with Eq. (1.47).

This approach can be generalized. Regardless of the physical nature of coordinate  $x$ , there will always be uncertainty  $\Delta x$  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’  $\Delta x$  of finite extent.

The two derivations may be further compared on the basis of effective ‘resolution lengths’. In Sec. 1.8.3 the limit  $\Delta x \rightarrow 0$  is rigorously taken, since  $\Delta x$  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  $\Delta x$ . By contrast, in the current derivation  $\Delta x$  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  $\Delta x$ , 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  $\Delta x$* . For example,  $\Delta x$  results as the Compton length in the EPI derivation of quantum mechanics (Chap. 4). 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 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  $\Delta x$ . Such considerations of course lead one to thoughts of quantum gravity (Misner *et al.*, 1973, p. 1193); see also Chap. 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, defining the limits of its range of validity.

Likewise, the electromagnetic wave equation (Chap. 5) would break down at scales  $\Delta x$  finer than the vacuum fluctuation length given by Eq. (5.4); suggesting a transition to a quantum electrodynamics. And the classical gravitational field equation (Chap. 6) would break down at scales finer than the Planck length, Eq. (6.22); again suggesting a transition to quantum gravity (see preceding paragraph). The magnitudes of these ultimate resolution lengths are, in fact, predicted by the EPI approach; see Chaps. 5, 6 and Sec. 3.4.15.

However, this kind of reasoning can be repeated to endlessly finer and finer scales. Thus, the theory of quantum gravity that is derived by EPI in Chap. 11 must break down beyond a finest scale  $\Delta x$  defined by the (presumed)

approximate nature of  $I$  at that scale. The length  $\Delta x$  would be much smaller than the Planck length. This, in turn, suggests the need for a *new* ‘mechanics’ that would hold at finer scales than the Planck length, etc., to ever-finer scales. The same reasoning applies to electromagnetic and gravitational theories. Perhaps all three theories would converge to a common theory at a finest resolution length to be determined, at which ‘point’ the endless subdivision process would terminate.

On the other hand, if one does not take this derivation seriously then, based upon the derivation of Sec. (1.8.2), the wave equations of quantum mechanics are valid *down to all scales*, and a transition to quantum gravity is apparently not needed. Which of the two views to take is, at this time, unknown.

The statistical mechanics approach of this section is based partly on work by Shirai (1998).

#### **1.8.14 Multiple PDF cases**

In all of the preceding, there was one, scalar parameter  $\theta$  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  $E$  (cf. Chap. 7).

Of course, however, nature is not that simple. There are physical phenomena that require *multiple* PDFs  $p_n(\mathbf{x})$ ,  $n = 1, \dots, N$  or amplitude functions  $q_n(\mathbf{x})$ ,  $n = 1, \dots, N$  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  $q_n(\mathbf{x})$ ,  $n = 1, \dots, 8$ , corresponding to the real and imaginary parts of the 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.