Foundation for Fisher-information-based derivations of physical laws

B. Roy Frieden
Optical Sciences Center, University of Arizona, Tucson, Arizona 85721

W. J. Cocke
Steward Observatory, University of Arizona, Tucson, Arizona 85721

(Received 11 January 1996)

The principle of extreme physical information (EPI) places physics on an information-theoretic footing. It yields many of the fundamental wave equations of physics (of Klein-Gordon, Dirac, etc.). EPI is based upon the measurement of particle four-vectors $\boldsymbol{\theta}_n$ with random errors $\mathbf{x}_n$, $n=1,\ldots,N$. This leads to an additive Fisher information form whose extremization provides the derivations. However, the model measurement procedure was, in the past, overly restrictive: the real and imaginary parts $q_n(\mathbf{x})$, $n=1,\ldots,N$ of the probability amplitude functions for the fluctuations were required to have nonoverlapping support regions. Here we rederive the requisite information form without this restriction. Our model measurement procedure is simply the efficient collection of four-data, that is, N independent measurements of four-coordinates of particles of the field. In effect, each measurement defines a different degree of freedom $q_n(\mathbf{x})$ of the scenario. [S1063-651X(96)02707-9]

PACS number(s): 05.40.+j, 03.65.Bz, 89.70.+c

INTRODUCTION

Consider the following measurement scenario [1]. An observer wants to define with optimum accuracy four-vectors $\boldsymbol{\theta}_n$ (say, on position), $n=1,\ldots,N$, for one or more particles. Thus he measures the $\boldsymbol{\theta}_n$. Measurements are necessarily imperfect The imperfect data must imply a finite amount of information I, by some suitable measure. To satisfy his goal, the observer forms optimum estimates of the $\boldsymbol{\theta}_n$ from the measurements, in the presence of the finite scalar I. The purpose of this paper is to *define* the information I that is appropriate to the measurement scenario.

The physics of the scenario lies in the fluctuations (called $\mathbf{x}$ below) of the data from the ideal values $\boldsymbol{\theta}_n$. The preceding epistemological setting gives rise to a zero-sum mathematical game of information maximization between the observer and nature [1]. The payoff point of the game establishes the probability density function (PDF) on the required fluctuations $\mathbf{x}$.

The game arises as follows. Designate J as the physical manifestation of data information I [2]. The zero-sum nature of the game follows from the fact that the observer and nature form a closed system, so that any information δI gained by the observer is at the expense δJ of nature (the physical phenomenon under study). Thus $\delta I=\delta J$ or, equivalently, $\delta(I-J)=0$ so that $I-J=$extrem.

This extremization problem also implies a game in which both the observer and nature ''want'' to maximize their information states: the observer, because he seeks maximal knowledge of the physical scenario; and nature, because the observer is part of nature and, hence, must be mirroring nature's ''attitude.'' (Nature's attitude leads to a situation of maximal disorder for the observer, as is discussed next.)

As we saw, the payoff point of the game is defined by a variational problem of information distance $(I-J)\equiv K$ extremization. K is also called the ''physical information.'' *Extremization of K via the Euler-Lagrange equation gives the wave equation appropriate to the particle measured*. (The solution also minimizes [3] information I, which means maximally broad wave functions and hence maximal disorder in the space-time predictability of the particles. This agrees with the second law.) In this way, the wave equations of Schrödinger, Dirac, Klein-Gordon, Helmhotz [1,3,4], etc., have been found to derive from an idealized measurement procedure.

All such derivations are based upon use of an additive form [Eqs. (19) or (20) below] for information I. Previously [1,4], this form was derived under the assumption of a *single* data four measurement. We also assumed that the underlying probability density function on $\mathbf{x}$ consists of a sequence of functions [the amplitude functions $q_n(\mathbf{x})$ defined in Eq. (18) below] *that do not have overlapping support regions*. The latter restriction is overly restrictive. It would be good to avoid making it.

We show here that this restriction can be lifted. As will be seen, the solution lies in replacing the previous scenario of a single four measurement with one of N *independent* four measurements.

DERIVATION OF INFORMATION FORM

In the measurement scenario, N *four*-vectors [4] of any physical nature (positions or potentials, etc.) are observed,

$$\mathbf{y}_n=\boldsymbol{\theta}_n+\mathbf{x}_n, \quad n=1,\ldots,N. \tag{1}$$

Vector $\mathbf{y}_n$ denotes the nth four measurement of the nth four-parameter vector $\boldsymbol{\theta}_n$ in the presence of the nth error four-fluctuation $\mathbf{x}_n$.

In accord with the observer's aim, the data $\mathbf{y}_n$ are to be collected efficiently, i.e., independently. This can be accomplished by two different experimental procedures: (a) N independent experiments upon a single particle, measuring its $\boldsymbol{\theta}_n$ at each repetition of the experiment; or (b) one experi-

ment 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. In scenario (b), the particles must be assumed to have $\boldsymbol{\theta}_n$ values that are sufficiently separated to give the required independence in the data $\mathbf{y}_n$.

Each component of every four-vector is, in general, either purely real or purely imaginary [5].

It is convenient to define ''grand'' vectors $\boldsymbol{\theta}$, $\mathbf{y}$, $d\mathbf{y}$ over all n as

$$\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N),$$

$$\mathbf{y}=(\mathbf{y}_1,\ldots,\mathbf{y}_N),$$

$$d\mathbf{y}=d\mathbf{y}_1\cdots d\mathbf{y}_N. \tag{2}$$

One can delineate two general reasons for taking data $\mathbf{y}$: (i) when the goal is to estimate a *function* of ideal positions (say) $\boldsymbol{\theta}$, e.g., the electromagnetic four-potential $\mathbf{A}(\boldsymbol{\theta}_n)$; or (ii) when the goal is to estimate the $\boldsymbol{\theta}$ *per se*, as when the $\boldsymbol{\theta}$ are the four positions of a material particle. For either scenario (i), (ii), we want to evaluate the information that resides in all the data $\mathbf{y}$.

Optimum, unbiased estimates

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

$$\hat{\boldsymbol{\theta}}_n\equiv\hat{\boldsymbol{\theta}}_n(\mathbf{y}) \tag{3}$$

of each four-parameter $\boldsymbol{\theta}_n$ is 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, but this will usually not be optimum. One class of optimum estimators is ''maximum likelihood'' estimators [6].

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

$$\langle\hat{\boldsymbol{\theta}}_n(\mathbf{y})\rangle\equiv\int d\mathbf{y}\,\hat{\boldsymbol{\theta}}_n(\mathbf{y})p(\mathbf{y}|\boldsymbol{\theta})=\boldsymbol{\theta}_n, \tag{4}$$

where $p(\mathbf{y}|\boldsymbol{\theta})$ is the conditional probability of all data $\mathbf{y}$ in the presence of all parameters $\boldsymbol{\theta}$. Equation (4) says that, although a given estimate will generally be in error, on the average it will be correct. How *small* the error may be, is next established. This introduces the vital concept of information.

Cramer-Rao inequality

We temporarily suppress index n and focus attention on the four components of any one (n fixed) foursome of *scalar* values θ_ν, y_ν, x_ν, $\nu=0,1,2,3$. The mean-square errors from the true values θ_ν are

$$e_\nu^2\equiv\int d\mathbf{y}[\hat{\theta}_\nu(y)-\theta_\nu]^2p(\mathbf{y}|\boldsymbol{\theta}). \tag{5}$$

It is known [6] that each mean-square error obeys complementarily with an ''information'' quantity I_ν,

$$e_\nu^2 I_\nu\geqslant 1, \tag{6}$$

where

$$I_\nu\equiv\int d\mathbf{y}\left[\frac{\partial\ln p(\mathbf{y}|\boldsymbol{\theta})}{\partial\theta_\nu}\right]^2p(\mathbf{y}|\boldsymbol{\theta}). \tag{7}$$

Equations (6) and (7) comprise the ''Cramer-Rao'' inequality. They hold for either real *or imaginary* components θ_ν; see [5]. When equality is attained in (6), the minimum possible error e_ν^2 is attained. Then the estimator is called ''efficient.'' The I_ν thus comprise a *vector* of informations.

Stam's information

We are now in a position to decide how to construct a single scalar information quantity I out of the vector of informations I_ν. Regaining subscripts n and summing on Eq. (6) gives

$$\sum_n\sum_\nu 1/e_{n\nu}^2\leqslant\sum_n\sum_\nu I_{n\nu}. \tag{8}$$

The left-hand sum of ''intrinsic accuracies'' (as termed by Fisher) equates to Stam's proposed [7] information measure

$$I_S\equiv\sum_n\sum_\nu 1/e_{n\nu}^2\leqslant\sum_n\sum_\nu I_{n\nu} \tag{9}$$

by Eq. (8). (We parenthetically note that Stam's information, in depending explicitly upon the error variances, ignores all possible error cross correlations. But it is easily shown that, for our additive error case (1), where the data $\mathbf{y}_n$ are independent and the estimators are unbiased (4), all error cross correlations are zero.) We adapt Stam's information to our purposes.

The right-hand side of Eq., (9) is a kind of ''channel capacity'' C of the problem: when efficient estimators are used $I_s=C$. Therefore, as in standard communication theory, we adapt C as the measure of system information performance. Then Eq. (9) becomes

$$I\equiv C=\sum_n\int d\mathbf{y}\;p(\mathbf{y}|\boldsymbol{\theta})\sum_\nu\left(\frac{\partial\ln p(\mathbf{y}|\boldsymbol{\theta})}{\partial\theta_{n\nu}}\right)^2 \tag{10}$$

in view of (7). This is the trace of the Fisher information matrix [6]. The information form further simplifies, as follows.

Independent data and additivity of the information

As mentioned before, the data are collected independently. Then the joint probability of all the data separates into

$$p(\mathbf{y}|\boldsymbol{\theta})=\prod_{n=1}p_n(\mathbf{y}_n|\boldsymbol{\theta})=\prod_{n=1}p_n(\mathbf{y}_n|\boldsymbol{\theta}_n). \tag{11}$$

This is a product of marginal laws. The latter equality follows since, by Eq. (1), $\boldsymbol{\theta}_m$ has no influence on $\mathbf{y}_n$, $m\neq n$. Taking the logarithm of Eq. (11) and differentiating then gives

$$\frac{\partial \ln p(\mathbf{y}|\boldsymbol{\theta})}{\partial \theta_{n\mu}} = \frac{1}{p_n}\frac{\partial p_n}{\partial \theta_{n\mu}}, \quad p_n \equiv p_n(\mathbf{y}_n|\boldsymbol{\theta}_n). \tag{12}$$

Substitution of Eqs. (11) and (12) into Eq. (10) gives

$$I = \sum_n \int d\mathbf{y} \prod_m p_m(\mathbf{y}_m|\boldsymbol{\theta}_m) \sum_\nu \frac{1}{p_n^2}\left(\frac{\partial p_n}{\partial \theta_{n\nu}}\right)^2, \tag{13}$$

$$= \sum_n \int d\mathbf{y}_n p_n(\mathbf{y}_n|\boldsymbol{\theta}_n) \sum_\nu \frac{1}{p_n^2}\left(\frac{\partial p_n}{\partial \theta_{n\nu}}\right)^2 \tag{14}$$

after integrating out $d\mathbf{y}_m$ for terms in $m \neq n$, using normalization of each probability p_m. After an obvious cancellation, we get

$$I = \sum_n \int d\mathbf{y}_n \frac{1}{p_n} \sum_\nu \left(\frac{\partial p_n}{\partial \theta_{n\nu}}\right)^2. \tag{15}$$

Gallilean invariance

In the preceding, the parameters $\boldsymbol{\theta}_n$ are assumed to be unknown and *fixed*, as was implied by the notation $(\mathbf{y}_n|\boldsymbol{\theta}_n)$. Then, because of the additive nature of the random components x_n in Eq. (1), it must be that fluctuations in $\mathbf{y}_n$ follow those of $\mathbf{x}_n$. Hence [8]

$$p_n(\mathbf{y}_n|\boldsymbol{\theta}_n) = p_{X_n}(\mathbf{y}_n - \boldsymbol{\theta}_n|\boldsymbol{\theta}_n) = p_{X_n}(\mathbf{y}_n - \boldsymbol{\theta}_n) = p_{X_n}(\mathbf{x}_n),$$

$$\mathbf{x}_n \equiv \mathbf{y}_n - \boldsymbol{\theta}_n, \tag{16}$$

assuming Galilean invarince. Then the $p_{X_n}(\mathbf{x}_n)$ are independent of absolute origins $\boldsymbol{\theta}_n$. Substituting the $P_{X_n}(\mathbf{x}_n)$ into Eq. (15) and changing the integration variables to x_n, gives

$$I = \sum_n \int d\mathbf{x}_n \frac{1}{p_{X_n}(\mathbf{x}_n)} \sum_\nu \left(\frac{\partial p_{X_n}(\mathbf{x}_n)}{\partial x_{n\nu}}\right)^2. \tag{17}$$

Observing the disappearance of absolute origins [8] $\boldsymbol{\theta}_n$ from the expression, the information likewise obeys Galilean invariance.

Use of probability amplitudes

Equation (17) further simplifies if we introduce real probability ''amplitudes'' $q_n(\mathbf{x}_n)$,

$$I = 4\sum_n \int d\mathbf{x}_n \sum_\nu \left(\frac{\partial q_n}{\partial x_{n\nu}}\right)^2, \quad p_{X_n}(\mathbf{x}_n) \equiv q_n^2(\mathbf{x}_n). \tag{18}$$

The subscript n of $\mathbf{x}$ can now be suppressed, since each $\mathbf{x}_n$ ranges over the same values. Then Eq. (18) becomes

$$I = 4\int d\mathbf{x} \sum_n \boldsymbol{\nabla} q_n \cdot \boldsymbol{\nabla} q_n,$$

$$\mathbf{x} = (x_0, \ldots, x_3), \quad d\mathbf{x} \equiv |dx_0| dx_1 dx_2 dx_3,$$

$$\boldsymbol{\nabla} \equiv \partial/\partial x_\nu, \quad \nu = 0,1,2,3. \tag{19}$$

Derivation of this equation was the aim of the paper. This is the form of I that was used in all extreme physical information (EPI) based derivations of physical laws [1,3,4,9].

Semicovariant form for the information

It is interesting to go one step further, using the fact that $\mathbf{x}$ is a four-vector, so that its first component is linear in the imaginary unit i. Then the first term in the sum in (19) is negative, and using covariant notation, (19) becomes

$$I = 4\int d^4x q_{n,\lambda} q_n{}^{,\lambda}. \tag{20}$$

We see that the derivative indices λ in this equation form a covariant pair. Moreover, since index n is merely a measurement number, and not (yet) indicative of a vector component (as it becomes in the applications of EPI), (20) is formally covariant. However, once the vector connection is made, the equation becomes noncovariant in index n. Nevertheless, the Euler-Lagrange solution that follows from the use of (20) in EPI *is* generally covariant in index n [1,4], as well as in coordinates $\mathbf{x}$. Evidentally, the fact that the *derivatives* in (20) occur covariantly is sufficient to yield covariant solutions to the EPI principle.

NET PROBABILITY $p(\mathbf{x})$ FOR A PARTICLE

The measurement scenario also makes a prediction on the overall PDF $p(\mathbf{x})$ for a single particle. The single particle case was scenario (a) as previously defined. Hence imagine one particle to be repeatedly measured. Then we may drop subscript n in $\mathbf{x}_n$ in Eqs. (16) and (18), which now give

$$p_n(\mathbf{y}_n|\boldsymbol{\theta}_n) = p_X(\mathbf{x}|\boldsymbol{\theta}_n) = q_n(\mathbf{x})^2. \tag{21}$$

Here we want the net $p(\mathbf{x})$ for all possible $\boldsymbol{\theta}_n$, in contrast to Eqs. (16) and (18) which express it as conditional upon N specific values $\boldsymbol{\theta}_n$. To eliminate the dependence upon $\boldsymbol{\theta}_n$ requires a Bayesian viewpoint [6], whereby a probability law for the ''prior'' parameters $\boldsymbol{\theta}_n$ is to be assigned. The parameters are fixed by the initial conditions of the experiment. With the lack of any prior information on how the physical system is constrained, the initial conditions must be assumed to be random, such that the $\boldsymbol{\theta}_n$ are equally probable,

$$P(\boldsymbol{\theta}_n) \equiv P_n = \frac{1}{N} \tag{22}$$

by normalization. This may be regarded as an ''equal weights'' or maximum ignorance property, analogous to that of quantum mechanics.

Equations (21) and (22) may be combined, via the partition law of statistics [10], to give the net PDF on $\mathbf{x}$ as

$$p(\mathbf{x}) = \sum_n p_X(\mathbf{x}|\boldsymbol{\theta}_n) P_n = \frac{1}{N} \sum_n q_n^2. \tag{23}$$

We specialize, next, to the case of the relativistic electron. Define complex amplitudes as [1,4]

$$\psi_n \equiv \frac{1}{\sqrt{N}} (q_{2n-1} + i q_{2n}), \quad i = \sqrt{-1}, \quad n = 1,\ldots,N/2. \tag{24}$$

Fluctuations $\mathbf{x}$ are now, in particular, those of the space-time coordinates of the electron. Using Eq. (24) and then Eq. (23) gives

$$\sum_{n=1}^{N/2} \psi_n^* \psi_n = \frac{1}{N} \sum_n q_n^2 = p(\mathbf{x}). \tag{25}$$

Hence the familiar dependence (25) of $p(\mathbf{x})$ upon $\psi_n(\mathbf{x})$ is a straightforward expression of the partition law of statistics [10]. By (25), the $\psi_n(\mathbf{x})$ also have the significance of being complex *probability* amplitudes. The Born assumption to this effect does not have to be made. A further property of the $\psi_n(\mathbf{x})$ defined in (24) is that they are found, via EPI [1,4], to obey the Dirac equation, which is of course the correct result.

DISCUSSION

The key information I form (19) for the use of EPI has been shown to derive from a realistic measurement procedure—the independent and efficient collection of four-vector data. The assumption that $I = C$, the system channel capacity, is justified by the success of EPI in deriving physical laws [1,3,4,9]. By our formulation, each measurement $\mathbf{y}_n$ provides a degree of freedom $q_n(\mathbf{x})$ in the information sum (19) and in the PDF (25). As examples, $N = 8$ measurements define the quantum mechanics of the electron [1]; while $N = 1$ defines classical Maxwell-Boltzmann statistics [9]. This agrees nicely with the EPI view that ''smart'' measurement (measurement followed by optimum estimation) elicits physical law.

Galilean invariance effect (16), (17) is built into the theory. In the special case where the $\mathbf{x}$ are space-time coordinates, the Galilean invariance becomes relativistic invariance as well. Finally, when information (19) is used in EPI, as supplemented by definition (24) of complex amplitudes, both the complex Dirac equation and the usual formula (25) for the PDF of the electron result.

[1] B. R. Frieden and B. H. Soffer, Phys. Rev. E **52**, 2274 (1995).
[2] An analogous point of view was taken by L. Brillouin, who showed that thermodynamic entropy and Shannon information are equivalent for a closed system. By this equivalence, the total entropy plus information in the system remains fixed. The information gain (in bits) of an observer is exactly balanced by the natural entropy change (in cal/K) of the physical system. See L. Brillouin, *Science and Information Theory* (Academic, New York, 1962), pp. 168, 232.
[3] B. R. Frieden, Am. J. Physics **57**, 1004 (1989).
[4] B. R. Frieden, in *Advances in Imaging and Electron Physics*, edited by P. W. Hawkes (Academic, Orlando, 1994), Vol. 90, pp. 123–204.
[5] Ref. [4], pp. 131 and 132.
[6] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I* (Wiley, New York, 1968).
[7] A. J. Stam, Inf. Control **2**, 101 (1959).
[8] See Ref. [4], pp. 135 and 136.
[9] B. R. Frieden, Phys. Rev. A **41**, 4265 (1990).
[10] B. R. Frieden, *Probability, Statistical Optics and Data Testing*, 2nd ed. (Springer-Verlag, New York, 1991), p. 25.