Principle of maximum Fisher information from Hardy's axioms applied to statistical systems

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

Robert A. Gatenby
Mathematical Oncology and Radiology, Moffitt Cancer Center, Tampa, Florida 33612, USA
(Received 13 August 2013; published 28 October 2013)

Consider a finite-sized, multidimensional system in parameter state **a**. The system is either at statistical equilibrium or general nonequilibrium, and may obey either classical or quantum physics. L. Hardy's mathematical axioms provide a basis for the physics obeyed by any such system. One axiom is that the number N of distinguishable states **a** in the system obeys $N = \max$. This assumes that N is known as deterministic prior knowledge. However, most observed systems suffer statistical fluctuations, for which N is therefore only known approximately. Then what happens if the scope of the axiom $N = \max$ is extended to include such observed systems? It is found that the state **a** of the system must obey a principle of maximum Fisher information, $I = I_{\max}$. This is important because many physical laws have been derived, *assuming* as a working hypothesis that $I = I_{\max}$. These derivations include uses of the principle of *extreme physical information* (EPI). Examples of such derivations were of the De Broglie wave hypothesis, quantum wave equations, Maxwell's equations, new laws of biology (e.g., of Coulomb force-directed cell development and of *in situ* cancer growth), and new laws of economic fluctuation and investment. That the principle $I = I_{\max}$ itself *derives* from suitably extended Hardy axioms thereby eliminates its need to be *assumed* in these derivations. Thus, uses of $I = I_{\max}$ and EPI express physics at its most fundamental level, its *axiomatic basis* in math.

DOI: 10.1103/PhysRevE.88.042144 PACS number(s): 05.70.Ln, 03.65.Ta, 05.40.−a, 02.10.−v

I. BACKGROUND: ON HARDY'S AXIOMS

It is long known that all classical physics follows the ZFC (Zermelo-Fraenkel-Choice) axioms of mathematical set theory. Quantum mechanics does not, however, obey these axioms. Nevertheless quantum mechanics does follow from many alternative sets of axioms, e.g., approaches by von Neumann, Mackey, Jauch and Lande, and Hardy [1]. Here we concern ourselves with the particular set of five axioms due to Hardy [1]. For brevity, we do not discuss all five of the axioms, only those relevant to our analysis. These are Eq. (2) and axiom 3 listed below Eq. (6).

The *motivation* for this study is as follows. The probability $p(\mathbf{x})$ or amplitude $\Psi(\mathbf{x})$ laws governing physical systems have been derived [2–20] on the basis of a unifying hypothesis or *assumption*: that the Fisher information [21,22] in each obeys $I = \max \equiv I_{\max}$, subject to a physical constraint. The derivations often take the form of a principle of extreme physical information (EPI), as discussed below. Can the assumption $I = I_{\max}$ be justified?

One of Hardy's axioms is Eq. (2), that the number N of distinguishable states of the system obeys $N = \max$. However, this value of N is assumed as deterministically known prior knowledge. But, most observed systems suffer random fluctuations, so that any *observed* value of N is intrinsically random and subject to error. On the grounds that acknowledging such random error might lead to a method of determining its probability law, we assume the following.

(1) Hardy's axiom $N = \max$ may be applied to *observed,* noise-prone systems.

(2) This observable value of N obeys the Hardy subsystem separation property $N = N_A N_B$, as enumerated in Sec. I B below.

We show that under these assumptions the state **a** of the system has the property that its Fisher $I = I_{\max}$. This, then, will justify past uses of the latter principle and EPI to derive physical laws governing the observed system. In a nutshell, the aim of this paper is to provide axiomatic justification for such past derivations of physical laws.

A. Estimating a system state: Distinction from the Hardy scenario

The whole derivation devolves upon determining the number N of distinguishable states of the system. To determine N requires knowledge of *how well* the parameters defining the state may be distinguished. This, in turn, asks how well they may be estimated. Hence, the problem becomes one of parameter estimation. Consider, then, an n-dimensional system characterized by a vector (in quantum systems, a Hilbert-) state $\mathbf{a} \equiv a_j, j = 1, \ldots, n$. Each a_j is a dimensional component or "channel," j of the fixed state **a.** State **a** can be, e.g., that of boson polarization, fermion spin, or position, depending on application. Each a_j is thereby a degree of freedom of the fixed state **a**. Each a_j, as a trigonometric projection, obeys $0 \leqslant a_j \leqslant l$ *on the continuum* within an n-dimensional cube with a common side l in each dimension j.

It is important to distinguish our system scenario from that in the introductions to Ref. [1]. There, as an idealized example, a_j is the spin in *a binary problem* where the particle can either have spin up or spin down. In that case there are straightforwardly $N = 2$ distinguishable states of the system. This is known as prior knowledge of the observed phenomenon. We instead take the experimentalist's view where only their continuously variable projections are known as data. Such continuity is required because we shall be using a Fisher-information approach to parameter estimation.

This*requires*the parameters to be differentiable and, hence, continuous.

1. Random nature of system

Our major distinctions from the Hardy scenario result from allowing particle state **a** to realistically obey random fluctuations $\mathbf{x}_j$. These, moreover, obey an unknown probability amplitude law. For example, in quantum mechanics, the amplitude law governing positional fluctuations $\mathbf{x}_j$ of a nonrelativistic particle from a state position **a** obeys *the Schrödinger wave equation*. It is desired to estimate that law. To aid in its estimation, the dimensional projections a_j of **a** are repeatedly measured, a total of k times, either in identically prepared experiments, or k times in the same experiment. (An example of the latter is where the state **a** is that of spin, and it is measured at k positions in the system.)

The measurements are the $\mathbf{y}_j \equiv y_{ji}, i = 1,\ldots,k$. These are generally imperfect versions of a_j, obeying

$$\mathbf{y}_j \equiv a_j + \mathbf{x}_j, \quad \text{or} \quad y_{ji} = a_j + x_{ji}, \quad \text{with} \quad 0 \leqslant a_j \leqslant l \quad \text{and} \quad j = 1,\ldots,n, \quad i = 1,\ldots,k. \tag{1}$$

The unknown state **a** thereby lies within an n-dimensional "box" of length l. Also, the data $\mathbf{y}_j$ suffer errors $\mathbf{x}_j \equiv x_{ji}$, random fluctuations on the continuum that are *characteristic of the physics of the particle*. It is assumed that *no additional fluctuations due to, say, noise of detection enter in*. The point of the measurement experiment is to determine the physics of the fluctuations $\mathbf{x}_j$, which are considered to define the system of state **a**. Thus, there are, in total, nk measurements of n scalar parameters a_j.

2. Generalizing Hardy's axiom N = max

Let two states **a** of the system be *distinguishable* if an observer can tell which is present when viewing a single copy **y** of the system. Thus, how many states **a** of the system are distinguishable? Call this N. It is intuitive that the less random are the measurement errors **x** the larger should be N. Hardy postulates, using an idealized spin experiment [1] where in effect fluctuations $\mathbf{x}_j = 0$, that in all physical cases, whether classical or quantum,

$$N = \max. \tag{2}$$

Instead, in our scenario Eq. (1), the observables are data $\mathbf{y}_j$ and these are *generally imperfect*versions of a_j due to fluctuations $\mathbf{x}_j$. We wish to define the physics of the fluctuations $\mathbf{x}_j$. Because of these fluctuations in the data, system property (2) is not immediately apparent. Instead,*we regard it as a corollary of the Hardy axioms*.

The value of N can only be estimated, on the basis of the data obeying Eqs. (1). Intuitively, N should increase as $l/\delta x$ (quantified below) where δx is some measure of the uncertainty in fluctuations **x** and l is the total length of the box interval. Thus, because of property (2) and the fixed nature of l, uncertainty δx is assumed to be minimal. Intuitively this suggests maximum information I in the data, as turns out to be the case at Eq. (15).

For simplicity of notation, denote all $\mathbf{y}_j,\mathbf{x}_j,a_j$, $j = 1,\ldots,n$ as **x**,**y**,**a**. Let the **y** occurring in the presence of the state **a** obey a fixed likelihood density law $p(\mathbf{y}|\mathbf{a})$. This law defines the physics of the data-forming system. The law is most conveniently found within the context of *estimating the unknown state* **a** *of the system*.

3. System fluctuations x_j as additive noise

The $\mathbf{x}_j \equiv x_{ji}$, $i = 1,\ldots,k$ are random, statistically independent fluctuations from the unknown state values a_j and are assumed independent of values a_j. Then likelihood law $p(\mathbf{y}|\mathbf{a}) = p_X(\mathbf{x})$. As mentioned below Eq. (1), the physics of the system lies in its fluctuations $\mathbf{x}_j$. Thus, each fluctuation vector $\mathbf{x}_j$ follows a generally different probability law $p_j(\mathbf{x}_j)$, $j = 1,\ldots,n$. Observing data $\mathbf{y}_j$ allows one to estimate the system state **a**, but also (as will be seen) to estimate the probabilities $p_j(\mathbf{x}_j)$ defining these input channels. This is taken up in Sec. IV. Meanwhile, we return to the problem of determining state **a**.

By Eq. (1), the k numbers in each dimensional data vector $\mathbf{y}_j$ suffer from k corresponding random fluctuations $\mathbf{x}_j = x_{ji}, i = 1,\ldots,k$ from the ideal state value a_j. It is assumed, for simplicity, that fluctuations $\mathbf{x}_j$ are independent of the corresponding parameter component a_j. This describes a shift-invariant system. Of course, physical effects commonly obey shift invariance [2,22]; this means, e.g., that the physics of a given system does not change as it is moved from one laboratory room to another. (Of course, the potential source is assumed to rigidly move with the rest of the system.) However, we first proceed as far as possible (until Sec. III A) *without* the assumption of shift invariance.

4. Discrete, or continuous, nature of the x_j

The state fluctuation values **x** are required by Hardy to be either (i) finite in number or (ii) countably infinite in number. These correspond, respectively, to either a (i) discrete coordinate system $\mathbf{x}_j \equiv x_{ji} = i_j\Delta x$ with $i_j = 0,1,\ldots,l/\Delta x$, with Δx the common finite spacing of positions in all dimensions, or a (ii) continuous coordinate system of **x** on interval $(0,l)$, where $\Delta x \to dx$, a differential. Both alternatives (i) and (ii) are included within the analysis by leaving unspecified the nature of the expectations $<>$ indicated in Eqs. (3), *et seq.*, defining the Fisher information. They could be either integrals over probability densities or sums over absolute probability laws. In the former case the $p_j(\mathbf{x}_j)$ and $p(\mathbf{y}|\mathbf{a})$ are regarded as absolute probabilities; in the latter they are regarded as probability *densities*.

With choice (i) of Δx finite, given also the finite size l of the cube the $l/\Delta x$ coordinates for each $\mathbf{x}_j$ are finite in number [choice (i) of Hardy]. With choice (ii) there are an infinity of fluctuation values $\mathbf{x}_j$. However, as results below show, in either case (i) or case (ii) the resulting N is generally finite, but can be *countably infinite* [choice (ii) of Hardy]. This would arise when one or more of the probability density functions $p_j(\mathbf{x}_j)$ contain regions of very high slope, as near edges or impulses. This causes the Fisher information I to approach infinity, implying that determinism is being approached.

B. Hardy's requirement $N = N_A N_B$

Consider subsystems A and B of the system, e.g., event states in two different dimensions. This separability property

$N = N_A N_B$ [see below Eq. (6)] is axiom 3 of Hardy. This axiom and Eq. (2) are the usual properties of accessible microstates in a statistical mechanical system at equilibrium, although neither Hardy nor we assume an equilibrium state.

Thus, Eq. (2) and axiom 3 are assumed to hold for systems in general states of nonequilibrium. In fact, these are actually *the only facets of Hardy's approach* that are used to prove our required results that (i) the system's Fisher information obeys $I = I_{\max}$ and that (ii) this holds for systems in general states of nonequilibrium. See, e.g., past applications of $I = I_{\max}$ to general nonequilibrium thermodynamics in [2,4] and to living cells in [11–14] (which must be in nonequilibrium states in order *to be* alive).

By Hardy's approach, Eq. (2) and axiom 3 also hold whether the system obeys classical or quantum physics.

II. COMPUTING N

The number N of *observably*distinguishable states of the system is next shown to relate to its level of Fisher information.

A. Fisher information

Suppose that the system is "coarse grained," defined as any perturbation of a system that causes it to lose local structural detail. This takes place during a small time interval Δt. The physical requirement is that this coarse graining must, by perturbation of the state, cause it to generally lose information, i.e., to obey $\delta I \leqslant 0$ for $\Delta t > 0$. On this basis a measure of "order" was derived as well, in Ref. [15] for the case of a single scalar dimension x and in [16] for the most general case of a vector coordinate $\mathbf{x} = (x_1, \ldots, x_n)$ of general dimension n. The order measure has analogous properties to the measures of Kolmogoroff [23] and Chaitin [24].

Coarse graining has the further property of demarking the transition from a quantum to a classical universe [25]. Although Fisher's classic information I defined by Eq. (9) has been used to derive quantum mechanics [2] a version of Fisher information specifically designed to apply to quantum scenarios is being developed as well [26,27].

B. Fisher information matrix

The principal aim of this paper is to show that extending Hardy's postulate $N = \max$ to a scenario of noisy data, where distinguishable states are not obvious, gives rise to a principle of maximization of the Fisher information. First we relate N to the Fisher information matrix. The latter is defined [2,22] as the $n \times n$ matrix $[I]$ of elements

$$I_{ij} \equiv \left\langle \frac{\partial \ln p(\mathbf{y} \mid \mathbf{a})}{\partial a_i} \frac{\partial \ln p(\mathbf{y} \mid \mathbf{a})}{\partial a_j} \right\rangle, \quad i,j = 1, \ldots, n \text{ independently.} \tag{3}$$

In Eq. (3), the brackets $\langle\ \rangle$ denote an expectation with respect to likelihood law $p(\mathbf{y}|\mathbf{a})$. Let the matrix $[I]$ have*an inverse*$[I]^{-1}$. Denote the latter's ijth element as I^{ij} (note the superscripts). The Cramer-Rao inequality states that the mean-squared error ε_j^2 in determining a_j has as its lower bound the jth diagonal element of $[I]^{-1}$,

$$\varepsilon_j^2 \geqslant I^{jj}, \quad \text{or} \quad 1/\varepsilon_j \leqslant (I^{jj})^{-1/2}. \tag{4}$$

In fact, the lower bound errors $\varepsilon_j^2 \equiv \varepsilon_{j\,\min}^2 = I^{jj}$ can be *achieved* in systems whose laws $p(\mathbf{y}|\mathbf{a})$ obey an "efficiency" condition [22].

C. Alternative measures of error

As will be seen, this specifically second-order measure of error $\varepsilon_j \equiv \langle |\mathbf{x}|^2 \rangle^{1/2}$ may be used to define the distinguishability of neighboring state values a_j. However, certain systems might instead require third- or higher-order error measures. Nevertheless, all such higher-order error measures are generally larger than ε_j, as shown by *Lyapunov's inequalities* [28],

$$\varepsilon_j \equiv \sqrt{\varepsilon_j^2} \equiv \left\langle |\mathbf{x}|_j^2 \right\rangle^{1/2} \leqslant \left\langle |\mathbf{x}|_j^3 \right\rangle^{1/3} \leqslant \left\langle |\mathbf{x}|_j^4 \right\rangle^{1/4} \leqslant \cdots. \tag{5}$$

These will be shown in Eqs. (15) and (16) to give the same result $I = I_{\max}$ as does the use of second-order measure ε_j.

D. Number N of distinguishable states

1. Quantum entanglement

As a counterexample, our scenario of multiple measurements $i = 1, \ldots, k$ of a in each dimension j is capable of encountering *quantum entanglement*. This where measurement pairs y_{ji} and $y_{ji'}$ are at least partially dependent for $i \neq i'$. The *intrinsic* fluctuations x_{ji} and $x_{ji'}$ then have the same property of entanglement. This can be entanglement in states of position, or momentum, spin, polarization, etc. An example is where, in the usual EPR experiment, the daughter 1 particle spin component state a_{fi} is measured on the earth and the daughter 2 state $a_{ji'}$ is simultaneously measured on the moon. Their measurements $y_{ji}, y_{ji'}$ are found to always point in opposite directions. This rules out contribution to N_j of the two measurements pointing *in the same* direction, thereby reducing N_j by 2 (states up-up and down-down). Thus, taking entanglement cases into account over all dimensions j can only reduce the number N of distinguishable states. In fact [29], "a set of entangled states can be completely indistinguishable, that is, it is not possible to correctly identify even one state with a non-vanishing probability; whereas a set containing at least one product state is always conclusively distinguishable (with non-zero probability)."

Ignoring, for the moment, the possibility of quantum entanglement (see above), the maximum possible number of distinguishable component state positions a_j in projection j obeys simply $N_j = l/\varepsilon_j$ [30]. Therefore, allowing for possible quantum entanglement can only reduce this total, giving $N_j \leqslant l/\varepsilon_j$.

2. Resulting upper bound to N

Each of the l/ε_j states characterizing a dimension j can occur simultaneously with any one of the $l/\varepsilon_{j'}$ states of *any other* dimension $j' \neq j$. Therefore, the total number N of distinguishable states $\mathbf{a}$ over *all* dimensions obeys a simple product,

$$N \leqslant \frac{l}{\varepsilon_1} \cdots \frac{l}{\varepsilon_n}, \tag{6}$$

over the n dimensions [29]. This product result also utilizes the Hardy axiom 3, that a composite system consisting of subsystems A and B satisfies $N = N_A N_B$.

Note that the inequality sign in Eq. (6) is replaced by an equality sign for classical systems (where entanglement does not occur and hence cannot reduce the number of distinguishable states). However, this will not affect the results below, beginning with Eq. (7). The latter is already an inequality because of inequality (4), which has nothing to do with entanglement. Hence, the final result $I = I_{\max}$ will hold for either classical or quantum systems.

Using inequalities (5) in Eq. (6) shows that N computed on the basis of third- or higher-order error measures is smaller than N computed on the basis of the second-order error ε_j. This will ultimately show [below Eq. (16)] that the Fisher $I = \max$ even when such *higher-order* error measures are used.

III. MAXIMIZING *N*

By Eq. (4), Eq. (6) is equivalent to

$$N \equiv N_m \leqslant l^n (I^{11} \cdots I^{nn})^{-1/2}. \tag{7}$$

The subscript m anticipates that N_m will be maximized.

A. Utilizing the independence of fluctuations x

As mentioned in Sec. I A3, all fluctuations $\mathbf{x}$ are independent of one another and of the states $\mathbf{a}$, implying a shift-invariant system.

In this scenario the information obeys simple additivity [31]

$$I = \sum_{j=1}^{n} I_j \tag{8}$$

over the n dimensions. Also, the matrix $[I]$ defined by Eq. (3) becomes purely diagonal, with elements

$$I_{jj} \equiv I_j \equiv \left\langle \left(\frac{\partial \ln p(\mathbf{y} \mid \mathbf{a})}{\partial a_j} \right)^2 \right\rangle. \tag{9}$$

Finally, the inverse matrix $[I]^{-1}$ is likewise diagonal, with elements I^{jj} that are the reciprocals of corresponding elements I_j of Eq. (9),

$$I^{jj} \equiv I^j = 1/I_j = \left\langle \left(\frac{\partial \ln p(\mathbf{y} \mid \mathbf{a})}{\partial a_j} \right)^2 \right\rangle^{-1}. \tag{10}$$

Then Eqs. (7) and (9) give

$$N_m \leqslant l^n \sqrt{I_1 I_2 \cdots I_n} = l^n \prod_{j,i} \left\langle \left(\frac{\partial \ln p(\mathbf{y} \mid \mathbf{a})}{\partial a_j} \right)^2 \right\rangle^{1/2}. \tag{11}$$

Taking a logarithm gives

$$\ln N_m \leqslant n \ln l + \frac{1}{2} \sum_{j=1}^{n} \ln I_j. \tag{12}$$

Axiom (2) states that the left-hand side of Eq. (12) must be maximized. However, by the nature of the inequality in (12) the right-hand side is even larger. Then, how large can it be?

B. Use of Lagrange undetermined multipliers

Since the individual I_j in Eq. (12) must obey the constraint Eq. (8) they are not all free to vary in attaining the value of N_m. This may be accommodated by amending requirement (12) via use of undetermined Lagrange multipliers,

$$\ln N_m = n \ln l + \frac{1}{2} \sum_{j=1}^{n} \ln I_j + \lambda \left(\sum_{j=1}^{n} I_j - I \right) = \max. \tag{13}$$

This has a single undetermined multiplier λ. Differentiating (13) with respect to any I_j and setting the result equal to zero gives an interim solution that $I_j = \text{const.} \equiv I_0$ for all $j = 1, \ldots, n$. Also, $\lambda = -n/(2I) = \text{const.}$ since $I = \text{const.}$ The size I_0 of the common information value I_j is found by requiring it to obey constraint (8). This gives the value $I_0 = I/n = I_j$ for all $j = 1, \ldots, n$. Using this in Eq. (13) and taking the antilog gives a value of

$$N \equiv N_m = \left(\frac{l^2}{n} I \right)^{n/2}. \tag{14}$$

Thus, since N is maximized by axiom (2), likewise,

$$I = I_{\max}. \tag{15}$$

Thus, an n-dimensional system with independent data will obey Hardy's axiom (2) if its total information I is maximized. Showing this was a major aim of the paper.

Principle (15), supplemented by constraint information, has in the past been used to derive laws of statistical mechanics [2–4,17], biology [2,3,5,6,11–14,20], and economics [2,3,18,19].

C. Use of third- or higher-order error measures

We noted below Eq. (6) that an alternative error measure $\langle |\mathbf{x}|_j^3 \rangle^{1/3}$, $\langle |\mathbf{x}|_j^4 \rangle^{1/4}$or higher order produces *smaller* values of N than does the specifically *second-order* meaure ε_j^2. Call these respective N values $N_m^{(3)}, N_m^{(4)}, \ldots$. For example, consider a case $N_m^{(3)}$. Then by Eqs. (5) and (6), $N_m^{(3)} \leqslant N_m$. Next, by the Hardy axiom Eq. (2) $N_m^{(3)}$ is to be maximized. Then the problem Eq. (13) is replaced with the problem

$$\begin{aligned} \max = \ln N_m^{(3)} &\leqslant \ln N_m \\ &= n \ln l + \frac{1}{2} \sum_{j=1}^{n} \ln I_j + \lambda \left(\sum_{j=1}^{n} I_j - I \right). \end{aligned} \tag{16}$$

The outer equality is identical with the maximization problem Eq. (13). Hence, it has the same solutions (14) and (15). Thus, $I = I_{\max}$ results by use of any error measure of order 2 or higher.

D. Verifying that the extremized *N* is a maximum

However, seeking the solution to Eq. (13) by variation of the I_j guarantees that it achieves an extreme value in $\ln N_m$, but not necessarily its maximum one. The extremum could, e.g., have instead been a minimum. For $\ln N_m$ to be maximized its second derivative must obey $[\partial^2 (\ln N_m)/\partial I_j^2] < 0$. The first derivative of (13) gives $\partial (\ln N_m)/\partial I_j = 1/(2I_j)$, so the second derivative gives $-1/(2I_j^2) < 0$, negative as required. Thus, $N_m = \max$ as required.

IV. SIGNIFICANCE TO PAST I-BASED DERIVATIONS OF PHYSICS

As mentioned at the outset, in general, the system can be coherent, obeying a complex vector amplitude function $\Psi(\mathbf{x})$ such that

$$p(\mathbf{x}) = \mathbf{\Psi}^*(\mathbf{x}) \cdot \mathbf{\Psi}(\mathbf{x}) = |\mathbf{\Psi}(\mathbf{x})|^2. \quad (17)$$

The physics obeyed by many systems specified by either scalar $p(\mathbf{x})$ or vector $\mathbf{\Psi}(\mathbf{x})$ laws have been found [2–20] by the use of relation (15) *as a hypothesis*. Here, by comparison, we have shown that (15) is generally valid, i.e., consistent with Hardy's axioms. Thus, it is no longer merely a hypothesis.

A. Need for prior knowledge-based constraints

But, what value of I will result from its maximization? Mathematically, the maximum possible value of I is infinity [by Eq. (9), occurring if edge- or impulselike details exist in p, causing one or more slope values $\partial \ln p/\partial a_j$ to approach infinity] However, infinite values of I could, by the Cramer-Rao inequality (4) in an "efficient" scenario [see the sentence following (4)], allow zero error, i.e., perfect determinism, to be present. This is contrary to the assumption that the system is statistical. We of course wish to avoid such inconsistent uses of the principle $I = I_{\max}$. To avoid the causative "runaway" value for I requires *prior knowledge of a constraint on* I that limits it to finite values. Such a one was used in problem (13), where the λ term was added on for precisely that purpose. (As an actual application to biology, principle (15) was used to derive the law of natural selection [20].) However, does the constraint term have a *physical* meaning?

B. Physical meaning of constraint term

Thus, the principle we now seek that avoids predictions of infinite slope, and determinism, must take the form

$$I - J = \text{extremum, where } J = I_{\max} \quad (18)$$

and the latter is of *finite* value. This is called the principle of 'extreme physical information' or EPI. The extremum achieved is nearly always a minimum. The solution $p(\mathbf{y} \mid \mathbf{a}) = p(\mathbf{x})$ or $\Psi(\mathbf{y}|\mathbf{x}) = \Psi(\mathbf{x})$ to (18) defines the physics obeyed by the system [2–5,7,10,15,17,19,32]. These probability and amplitude laws are the mathematical solutions to the second-order Euler-Lagrange differential equation solutions to principle (18). Examples are the Klein-Gordon or Schrodinger wave equation [2,32].

In principle (18), J is the *physical* representation, or source, of the *observed* information I. This can take the form of squared particle momentum, or charge and/or energy flux, depending upon application. Also, since J is the source of I, it represents the *maximum possible value* $I_{\max}$ of I. This works mathematically because the extremum in Eq. (18) is nearly always a minimum [2]. So with $I - J = \min$, necessarily $I \to J = I_{\max}$ in value. Thus, I is forced *toward its maximum value*$I_{\max}$, as required by principle (15). In particular, any solution for I attains some*fraction* of $I_{\max}$ between 1/2 and 1, depending on application (e.g., value 1/2 in classical physics and 1 in quantum physics). Thus, mathematically, the finite nature of source term J acts as a constraint that limits the maximized value of I. Thus, it plays the analogous role to the λ-dependent term in principle (16).

As we saw, principle (18) holds for a classical or quantum system in any state of nonequilibrium.

V. DISCUSSION

In a quantum system, the amount of information encoded is limited by the dimension of its Hilbert space, i.e., its number of perfectly distinguishable quantum states **a**. In past work [33], principle (2) that $N = \max$ was shown to hold, in particular, for a binary quantum system of unit vector states, provided the *Shannon* mutual information between the (random) quantum state and the results of its measurements is maximized. Thus, the Shannon replaced our use here of the Fisher information. This Shannon result was derived in cases where the angle between the two states in a two-dimensional vector basis space is distributed with uniform randomness over its allowable range.

The issue of *why* the concept of N, the maximum number of distinguishable states, should be used as a founding axiom of physics is of interest. It actually ties in with Godel's two incompleteness theorems. According to these any finite axiomatic theory must be incomplete in scope (i.e., admit of undecidable theorems). However, it must at least be self-*consistent*. In fact, the concept of a quantum holographic universe [34] is *consistent with* a maximum N [of size $N \sim \exp(10^{10})$]. Likewise, in statistical mechanics, the maximum *consistent* event **a** is the one that can occur in the*maximum* number of ways N [35].

There are a fixed total nk of outcomes resulting from the measurements described in Eq. (1). Then the result (15) that $I = I_{\max}$ also means that the amount of information per measurement obeys $I/nk = \max$. In turn, this may be interpreted as meaning that any observed system is "maximally free" [36]; in the sense that a specification of its statistical state **a** that is sufficient for predicting the probabilities of outcome of *future* measurements **y** will require the *maximum amount of information* I per experimental outcome. In this sense, the experiment is minimally predictive, thereby exhibiting "free will." Moreover, as we saw, $I = I_{\max}$ holds for classical or quantum systems. This raises the possibility that the molecular neurons of the brain, regardless of whether they obey classical or quantum [37] physics, exhibit free will. This might translate, in turn, into *conscious* free will [38]. However, the evidence for this is not yet compelling.

Our principal use of Hardy's axioms was the postulate Eq. (2), *which holds (a) whether the system is classical or quantum and also (b) whether it is in any state of equilibrium or nonequilibrium.* Property (a) must also relate to the fact that I is a limiting form [2,39] of the density function $\rho(\mathbf{x},\mathbf{x}')$ of the system, which represents the system in the presence of only partial knowledge of its phase properties. This describes a mesoscopic system, one partway between being classical and quantum. Thus, our main results, Eqs. (15) and (18), which state that $I = I_{\max}$, likewise hold under these general conditions (a) and (b). Then these justify past uses of EPI principle (18) to derive [2–20] diverse laws of statistical physics, including those of cell biology, whose living cells *must be in nonequilibrium states*.

In summary, the EPI principle (18) and information maximization principle (15) are no longer merely convenient hypotheses for deriving statistical laws of physics. Rather, they are *a priori* known to be valid for all current physics on its most basic level, that of its underpinnings in Hardy's mathematical axioms. This fills an important gap in all past applications [2–20] of the principle of maximum I to physics, biology, and econophysics, which *assumed*, as a working hypothesis, that $I = I_{\max}$. It also justifies future uses of principles (15) and (18).

ACKNOWLEDGMENTS

The authors acknowledge support from the National Cancer Institute under Grant No. 1U54CA143970-01. One of us (B.R.F.) thanks Professor J. O. Kessler for suggesting that a principle of maximum Fisher information might conceivably be implied by L. Hardy's axioms.

[1] L. Hardy, arXiv:quant-ph/0101012; also, Stud. Hist. Philos. Mod. Phys. **34**, 381 (2003).
[2] B. R. Frieden, S*cience from Fisher Information*, 2nd ed. (Cambridge University Press, Cambridge, UK, 2004).
[3] B. R. Frieden and R. A. Gatenby (eds.), *Exploratory Data Analysis Using Fisher Information* (Springer, London, 2007).
[4] B. R.Frieden, A. Plastino, A. R. Plastino, and B. H. Soffer, Phys. Rev. E **66**, 046128 (2002).
[5] B. R. Frieden and R. A. Gatenby, Phys. Rev. E **72**, 036101 (2005).
[6] R. A. Gatenby and B. R. Frieden, J. Math. Med. Biol. **25**, 267 (2008).
[7] B. R. Frieden and B. H. Soffer, Physica A **388**, 1315 (2009).
[8] B. R. Frieden, A. Plastino, and A. R. Plastino, Physica A **391**, 410 (2012).
[9] B. R. Frieden and M. Petri, Phys. Rev. E **86**, 032102 (2012).
[10] R. Carroll, *On the Quantum Potential* (Arima, Suffolk, UK, 2007).
[11] R. A. Gatenby and B. R. Frieden, PLoS ONE **5**, e12084 (2010).
[12] B. R. Frieden and R. A. Gatenby, PLoS ONE **6**, 22085 (2011).
[13] R. A. Gatenby and B. R. Frieden, J. Phys. Chem. Biophys. **1**, 102 (2011).
[14] B. R. Frieden and R. A. Gatenby, Front. Biosci. E **5**, 1017 (2013).
[15] B. R. Frieden and R. J. Hawkins, Phys. Rev. E **82**, 066117 (2010).
[16] B. R. Frieden and R. A. Gatenby, Phys. Rev. E **84**, 011128 (2011).
[17] B. R. Frieden, A. Plastino, A. R. Plastino, and B. H. Soffer, Phys. Rev. E **60**, 48 (1999).
[18] R. J. Hawkins and B. R. Frieden, Phys. Lett. A **322**, 126 (2004).
[19] R. J. Hawkins, M. Aoki, and B. R. Frieden, Physica A **389**, 3565 (2010).
[20] S. A. Frank, J. Evol. Biol. **22**, 231 (2009).
[21] R. A. Fisher, Phil. Trans. R. Soc., A **222**, 309 (1922).
[22] B. R. Frieden, *Probability, Statistical Optics and Data Testing*, 3rd ed. (Springer-Verlag, Berlin, 2001); H. L. Van Trees, *Detection, Estimation and Modulation Theory* (Wiley, New York, 1968), Part 1, pp. 79–81.
[23] A. N. Kolmogoroff, Probl. Inf. Transm. **1**, 1 (1965).
[24] G. J. Chaitin, J. Assoc. Comput. Mach. **16**, 407 (1969).
[25] J. J. Halliwell, Contemp. Phys. **46**, 93 (2005).
[26] S.-L. Luo, Proc. Am. Math. Soc. **132**, 885 (2003); also Commun. Theor. Phys. **47**, 597 (2007).
[27] P. Gibilisco, D. Imparato, and T. Isola, Linear Algebra Appl. **428**, 1706 (2008); also P. Gibilisco, F. Hiai, and D. Petz, IEEE Trans. Inf. Theory **55**, 439 (2009).
[28] V. K. Rohatgi, *An Introduction to Probability Theory and Mathematical Statistics* (Wiley, New York, 1976).
[29] S. Bandyopadhyayar, Phys. Rev. A **81**, 022327 (2010).
[30] M. R. Foreman and P. Torok, Phys. Rev. A **82**, 043835 (2010); The number of distinguishable states in, specifically, polarization space is defined to be the ratio of the volume of uncertainty in polarization before a measurement to the volume of uncertainty after a measurement. In our problem, in a given dimension j the former uncertainty is l, the latter ϵ_j.
[31] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, UK, 2001).
[32] B. R. Frieden and B. H. Soffer, Phys. Rev. E **52**, 2274 (1995).
[33] L. B. Levitin, T. Toffoli, and Z. Walton, Int. J. Theor. Phys. **44**, 219 (2005).
[34] C. J. Hogan, Phys. Rev. D **66**, 023521 (2002).
[35] F. Reif, *Fundamentals of Statistical and Thermal Physics* (McGraw-Hill, New York, 1965).
[36] L. Smolin, arXiv:1205.3707 [quant-ph].
[37] W. R. Loewenstein, *Physics in Mind: A Quantum View of the Brain* (Basic Books, New York, 2013).
[38] S. R. Hameroff and R. Penrose, J. Conscious. Stud. **3**, 36 (1996).
[39] B. R. Frieden, Phys. Rev. A **66**, 022107 (2002).