Appearance of Gibbs states in quantum-state tomography

Jochen Rau[*]

Institute for Theoretical Physics, University of Ulm, Albert-Einstein-Allee 11, 89069 Ulm, Germany
and Department of Engineering, RheinMain University of Applied Sciences, Am Brückweg 26, 65428 Rüsselsheim, Germany

(Received 8 April 2014; revised manuscript received 2 October 2014; published 10 December 2014)

I investigate the extent to which the description of quantum systems by Gibbs states can be justified purely on the basis of tomographic data, without recourse to theoretical concepts such as infinite ensembles, environments, or information or to the systems' dynamics. I show that the use of Gibbs states amounts to a relevance hypothesis, which I spell out in detail. This hypothesis can be subjected to statistical hypothesis testing and hence assessed on the basis of the experimental data.

DOI: 10.1103/PhysRevA.90.062114 PACS number(s): 03.65.Wj, 05.30.Ch

I. INTRODUCTION

To describe the static or dynamic properties of a macroscopic quantum system, typically only a few observables $\{G_a\}$ are deemed relevant—for example, the system's constants of the motion (if static), slow observables (if dynamic), or observables pertaining to some subsystem of interest. In statistical mechanics the system is then assigned that quantum state which, while reproducing the observed values $\{g_a\}$ of the relevant observables, maximizes the *von Neumann entropy*,

$$S[\mu] := -\mathrm{tr}(\mu \ln \mu); \tag{1}$$

i.e., its state—which I denote μ_g—is determined by the maximization

$$\mu_g := \arg \max_{\mu \in g} S[\mu], \tag{2}$$

where $\mu \in g$ is short for the constraints $\langle G_a \rangle_\mu = g_a \forall a$. It has the *Gibbs form*

$$\mu_g \propto \exp(-\lambda^a G_a), \tag{3}$$

with Lagrange parameters $\{\lambda^a\}$ and a constant of proportionality (the inverse of the partition function) chosen to ensure state normalization, $\mathrm{tr}\mu_g = 1$. For ease of notation I adopt here the Einstein convention that identical upper and lower indices are to be summed over. In the special case where only the system's energy is relevant, the set $\{G_a\}$ contains just the Hamiltonian, and the associated Lagrange parameter is the inverse temperature; the Gibbs state is then a canonical state. While they first arose in the context of statistical mechanics, Gibbs states nowadays play an important role also on smaller scales. For instance, they have been employed successfully in nanoscale thermodynamics [1–3], high-energy physics [4], and incomplete quantum-state tomography [5–11].

Why entropy maximization, and hence the use of the Gibbs form, should be the proper paradigm for constructing the quantum state has been the subject of much debate. The classic textbook argument in statistical mechanics relies on an idealization, the thermodynamic limit: The system of interest is viewed as but one member of a fictitious infinite ensemble of identically prepared systems. If the global state of this fictitious ensemble is constrained by sharp values (not expectation values) for the totals of the relevant observables, then the reduced state of any single member of the ensemble has the Gibbs form [12]. Recent research suggests that one can do without such fictitious ensembles and derive the Gibbs form just as well directly from a few generic assumptions, as long as the state in question pertains to a subsystem coupled to a sufficiently large environment [13,14]. Another popular argument invokes the intimate connection between entropy and information: By maximizing the entropy, Gibbs states discard, to a maximal extent, all information (and thus retain no spurious bias) as to irrelevant degrees of freedom; so they carry information solely about the relevant ones. This insight is at the heart of the information-theoretic approach to statistical mechanics [15–18]. Yet another line of reasoning, going back to Boltzmann's H theorem [19], brings into play the system's effective dynamics on some coarse-grained level of description, in particular, its tendency to increase entropy [20–23]. Such arguments rely on the existence of disparate time scales in the system [24]. Finally, some authors in both the statistics [25,26] and the physics [27] communities have argued (for the classical case only) that the maximum entropy paradigm is mandated by logical consistency; but this point of view remains controversial [28,29].

In the present paper I wish to add a different perspective. State construction via maximum entropy is a special instance of a much broader task: estimating a quantum state on the basis of imperfect data. Experimental data are in fact never perfect, not even for simple systems; because the investigated samples have a finite size, measurement devices have limited accuracy, and possibly—as is the case in statistical mechanics—the observables measured are not informationally complete. So in practice, data *never* specify a unique quantum state. Rather, among the many states compatible with the data one must infer the most probable one, using suitable statistical estimation techniques. Such techniques have become an indispensable tool for data analysis in modern quantum physics experiments and are called *quantum-state tomography* [30,31].

If the maximum entropy paradigm may thus be subsumed under the broader framework of quantum-state tomography, then perhaps the latter can shed some light on the question of when and how Gibbs states arise. Exploring the extent to which this is indeed possible is the purpose of the present paper. Consequently, I tackle the issue of Gibbs states solely with the help of statistical methods from quantum-state tomography,

[*] jochen.rau@q-info.org; www.q-info.org

and nothing else; in particular, without any recourse to the thermodynamic limit, environments, the concept of information, or dynamics.

The remainder of the present paper is organized as follows. In Sec. II, I review some basic concepts of quantum-state tomography. In Sec. III, I focus on the situation where the experimental data come in the form of sample means of some informationally incomplete set of observables. I show that in this case one can apply the quantum Sanov theorem to find the asymptotics of the pertinent likelihood function. The subsequent section, IV, is then crucial for the understanding of Gibbs states: I argue that the use of the Gibbs form is tantamount to a statistical "relevance hypothesis," for which I give a precise mathematical formulation. In Sec. V, I discuss how the likelihood of this hypothesis and of possible rival hypotheses may be assessed in the light of experimental data. Finally, in Sec. VI, I conclude with a brief summary and a few additional remarks.

II. QUANTUM-STATE TOMOGRAPHY

It is possible to know the precise state of an individual quantum system *after* a measurement: For instance, if a measurement of some observable returns one of its nondegenerate eigenvalues, then after the measurement, the system is known with certainty to be in the associated eigenstate. (Precise knowledge of the postmeasurement state thus hinges on precise knowledge of the observable. Strictly speaking, the latter necessitates additional measurements.) Yet it is impossible to reconstruct, based on measurements on the individual system alone, its state *before* the measurement [32]. Such a reconstruction instead requires measurements on many identically prepared copies; and even then, due to the always-finite number of copies (let alone the limited accuracy of measurement devices), the reconstruction can never be perfect. Thus in practice, measurements never yield a unique quantum state. Indeed, current experiments that implement quantum circuits or probe fundamental aspects of quantum information in many-body systems work with typical sample sizes of several hundreds or thousands, leading to statistical errors of up to 10% [33]. Under these circumstances one can only aspire to identify, among the many states compatible with the data, the *most probable* one. This requires the use of suitable statistical estimation techniques and is the subject of quantum-state tomography [30,31].

Identical preparation of copies means that these form an *exchangeable sequence* [34]. Such a sequence has finite length L, which may be chosen freely. It can be thought of as being drawn randomly from a fictitious infinite sequence of systems whose order is irrelevant (Fig. 1). Exchangeability entails two basic properties for the L-body state $\rho^{(L)}$ of the sequence: (i) it is symmetric under permutation of the constituents; and (ii) since the exchangeable sequence of length L can always be considered a subsequence of a longer, equally exchangeable sequence of length $L+1$, the state $\rho^{(L)}$ can be written as a marginal of $\rho^{(L+1)}$.

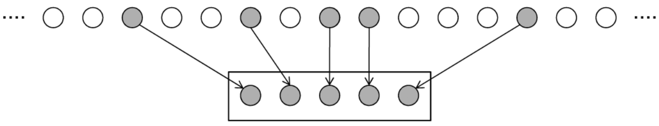

FIG. 1. Exchangeable sequence of quantum systems. One obtains a finite exchangeable sequence of L quantum systems by drawing randomly L systems from a fictitious infinite symmetric sequence.

Exchangeability is more than mere symmetry. For example, the two-body density matrix $\rho_{AB} = |\psi_{AB}\rangle\langle\psi_{AB}|$ associated with the Bell state $|\psi_{AB}\rangle = (1/\sqrt{2})(|00\rangle + |11\rangle)$ is invariant under permutation of the constituents and hence meets the symmetry criterion; yet it cannot be written as the marginal of an equally symmetric three-body state and thus fails to meet the second criterion for exchangeability. Being exchangeable is also not the same as being independent and identically distributed (i.i.d.): in general, it is $\rho^{(L)} \neq \rho^{\otimes L}$. Rather, by a quantum generalization [34] of the classical de Finetti theorem [35], the state of an exchangeable sequence can always be represented as an incoherent mixture of i.i.d. sequences $\rho^{\otimes L}$,

$$\rho^{(L)} = \int d\rho \operatorname{prob}(\rho) \rho^{\otimes L}, \tag{4}$$

with respective weights prob(ρ), where the integral is over all normalized single-particle states. Conversely, any state of this form describes an exchangeable sequence. The de Finetti representation shows that an exchangeable sequence may well exhibit classical correlations. However, it never exhibits entanglement.

Exchangeable sequences are the "raw material" of quantum-state tomography. The uncertainty about the state of an individual constituent is reflected in the density function prob(ρ); the latter may be considered (in somewhat loose terminology [34]) the probability distribution for the unknown single-constituent state. To learn more about this state, a sample of size N ($N < L$) is taken from the exchangeable sequence and a measurement performed on it, yielding data D. Afterwards the remaining $L - N$ systems (i.e., the original sequence minus the sample) still form an exchangeable sequence whose state has the above de Finetti representation; yet the probability distribution featuring in this de Finetti representation must be updated according to a quantum generalization of *Bayes' rule* [36],

$$\operatorname{prob}(\rho|D) \propto \operatorname{prob}(D|\rho^{\otimes N})\operatorname{prob}(\rho), \tag{5}$$

where the probabilities denote (from left to right) the posterior, the likelihood function, and the prior, respectively, and the constant of proportionality is independent of ρ. This Bayesian update encapsulates the process of *learning* from sample data (Fig. 2).

Upon the investigation of additional samples, Bayes' rule is iterated, leading to consecutive updates of the posterior. As more and more data accumulate—by investigating more samples or increasing their sizes—the posterior narrows until eventually its width falls below some desired error bound. Then within this error, the location of the posterior peak is the best estimate for the unknown quantum state. In the hypothetical limit of infinite sample size, informationally complete measurements, and perfectly accurate measurement devices, the posterior converges towards the likelihood function, which in turn approaches a δ function. The state estimate is then determined—to perfect accuracy—by experimental data only and becomes independent of the prior. (It is the fact that this is

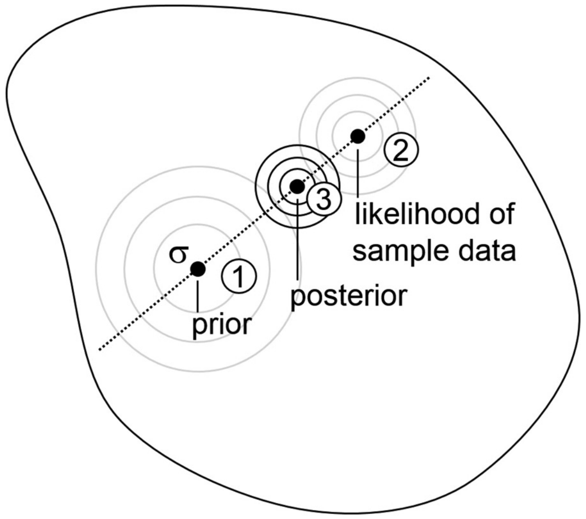

FIG. 2. Learning from sample data. (1) Before the experiment an exchangeable sequence is characterized by some prior probability distribution prob(ρ) in single-constituent state space. If, say, on theoretical grounds one expects the members of the sequence to be in a state close to σ, then this prior will be peaked around σ. It has a finite width reflecting the finite degree of confidence in this prior bias. (2) Investigation of a sample yields data D. Associated with these data is the likelihood function prob$(D|\rho^{\otimes N})$, typically peaked around some other state which might be close to, but is usually not equal to, σ. The likelihood function, too, has a finite width, reflecting the finite size N of the sample (and possibly other sources of error). (3) According to Bayes' rule, multiplying the prior by the likelihood function yields the posterior prob$(\rho|D)$. The latter is typically narrower than the prior, reflecting the growing confidence in the state estimate as experimental data accumulate. The center of the posterior has shifted from the original bias σ to a new state interpolating between σ and the center of the likelihood function.

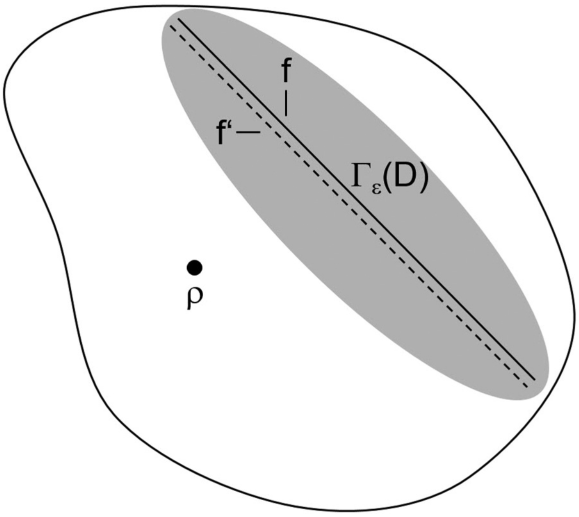

FIG. 3. Relating data to sample means. The set $\Gamma_\epsilon(D)$ contains all states compatible (up to some error ϵ) with the observed data D (shaded area). In order to encompass sample means $\{f_b\}$ this set must contain all states yielding $\langle F_b\rangle = f_b \forall b$ (solid line). In this example, $\Gamma_\epsilon(D)$ is so large that it also encompasses other values $\{f'_b\}$ for the sample means (dashed line).

possible, at least in principle, that gives operational meaning to the notion of "state.") Against this backdrop many state estimation techniques focus from the outset on the likelihood function, equating the location of its peak with the most plausible state estimate; such techniques fall into the class of maximum likelihood methods [37–39]. In contrast, methods that take into account the residual influence of the prior (which, in practice, is always present and, for small samples, may be quite significant) are termed *Bayesian* [40–42].

Strictly speaking, even in the above hypothetical limit the posterior coincides with the likelihood function only if the prior has support in the entire state space. The prior reflects any theoretical constraint or bias that one may have, prior to measurement, as to the parametric form or parameter values of the quantum state. As long as one knows nothing or little about the state *a priori*, this prior is broad and indeed has full support. But as soon as one has advance knowledge that constrains the quantum state to some region or proper submanifold of state space, the prior has support in this region or submanifold only; and so will the posterior, *regardless of the data* [43]. In Sec. IV, I argue that such *a priori* restrictions play an important role in the understanding of Gibbs states.

III. SANOV LIKELIHOOD

I suppose that the experimental data consist of a set of sample means $\{f_b\}$ gleaned from a sample of large but finite size N. These sample means may have been obtained directly by measurement of the pertinent observables $\{F_b\}$ or inferred indirectly from other data D—a possibility which is particularly relevant when the sample means pertain to observables that do not commute. In the latter case I say that the observed data D "encompass" sample means $\{f_b\}$ if and only if the set of states compatible with the data,

$$\Gamma_\epsilon(D) := \{\mu|\text{prob}(D|\mu^{\otimes N}) \geqslant 1-\epsilon\} \tag{6}$$

(up to some finite error parameter ϵ, $0<\epsilon<1$, which is independent of sample size), *contains* the set of states yielding expectation values $\langle F_b\rangle_\mu = f_b \forall b$; in short, $f \subseteq \Gamma_\epsilon(D)$. When the set $\Gamma_\epsilon(D)$ is large, the data might encompass not just $\{f_b\}$ but also different values $\{f'_b\}$ for the sample means (Fig. 3). On the other hand, if the set $\Gamma_\epsilon(D)$ contains only f, and no other f', and is, moreover, the smallest set to do so, then I say that the data *amount to* having measured the sample means f. With this understanding, the likelihood of measuring sample means f reads

$$\text{prob}_\epsilon(\{f_b\}|\rho^{\otimes N}) := \inf_D\{\text{prob}(D|\rho^{\otimes N})|f \subseteq \Gamma_\epsilon(D)\}. \tag{7}$$

Defined in the above way, the likelihood generally depends on the error parameter ϵ. For large sample sizes N, however, the infimum on the right-hand side behaves asymptotically as

$$\inf_D\{\ldots\} \sim \exp\left[-N\min_{\mu\in f} S(\mu\|\rho)\right] \tag{8}$$

and hence loses its dependence on ϵ; this follows from the quantum generalization [44–46] of the classical Sanov theorem

[47–49]. Here

$$S(\mu\|\rho) := \begin{cases} \mathrm{tr}(\mu\ln\mu - \mu\ln\rho), & \mathrm{supp}\mu \subseteq \mathrm{supp}\rho, \\ +\infty & \text{otherwise} \end{cases} \tag{9}$$

denotes the relative entropy of the two states μ and ρ [50–53]. In other words, for large N the likelihood function behaves as

$$\mathrm{prob}(\{f_b\}|\rho^{\otimes N}) \sim \exp\big[-NS\big(\mu^\rho_f\|\rho\big)\big], \tag{10}$$

with

$$\mu^\rho_f := \arg\min_{\mu\in f} S(\mu\|\rho), \tag{11}$$

independently of ϵ. Due to its close connection to the quantum Sanov theorem, I call this asymptotic likelihood the *Sanov likelihood*. The state μ^ρ_f, which, under given constraints on the expectation values $\{\langle F_b\rangle\}$, minimizes the relative entropy with respect to the "reference state" ρ, has the generalized Gibbs form [54]

$$\mu^\rho_f \propto \exp[(\ln\rho - \langle\ln\rho\rangle_\rho) - \kappa^b F_b], \tag{12}$$

with Lagrange parameters $\{\kappa^b\}$ and the constant of proportionality again chosen to ensure $\mathrm{tr}\mu^\rho_f = 1$.

There is the special case where the sample means $\{f_b\}$ are informationally complete. In this case the data determine a unique tomographic image (i.e., center of the likelihood function) μ, the sole state to yield $\langle F_b\rangle_\mu = f_b \forall b$. The quantum Sanov theorem then reduces to the quantum Stein lemma [55–57], and the asymptotic likelihood function becomes

$$\mathrm{prob}(\mu|\rho^{\otimes N}) \sim \exp[-NS(\mu\|\rho)]. \tag{13}$$

I call this the *Stein likelihood*. Thanks to a mixing rule for the relative entropy [58], the Stein likelihood satisfies

$$\mathrm{prob}(\mu|\rho^{\otimes N}) \cdot \mathrm{prob}(\mu'|\rho^{\otimes N'}) \\ \propto \mathrm{prob}\left(\frac{N}{N+N'}\mu + \frac{N'}{N+N'}\mu'|\rho^{\otimes(N+N')}\right), \tag{14}$$

with a constant of proportionality that does not depend on the state ρ. So for the purposes of Bayesian updating via Eq. (5), obtaining, first, a tomographic image μ from a sample of size N and, subsequently, a tomographic image μ' from another sample of size N' is tantamount to obtaining the weighted average of μ and μ' from the combined sample of size $N + N'$; sequential or joint processing of the data yields the same posterior. In other words, in the asymptotic limit considered here it does not matter how the system copies under investigation are grouped into samples.

IV. RELEVANCE HYPOTHESIS

The statement, "The observables $\{G_a\}$ are relevant," entails two distinct assertions: (i) The expectation values of the $\{G_a\}$ completely determine the state estimate (and all predictions following from it); and (ii) any update of this state estimate is determined by additional data pertaining to the $\{G_a\}$ only, and not by any other data. (This notion of "relevance" is similar to the notion of "consistency" invoked—in the classical case and for one special set of observables only—in Ref. [27].) In this section, I prove that the relevance hypothesis imposes on the state estimate the Gibbs form, (3).

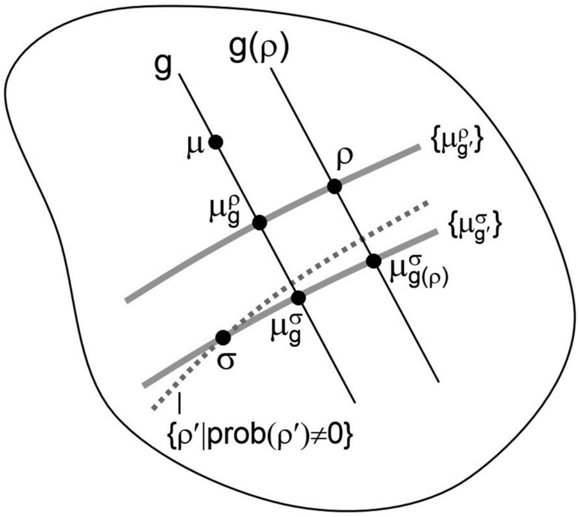

FIG. 4. States (filled circles) and sets of states (lines) featuring in the discussion of the relevance hypothesis. A tomographic measurement on a large but finite sample comes in two versions, one complete and the other incomplete. The complete version returns a unique tomographic image μ, whereas the incomplete version merely returns sample means g for the informationally incomplete set of observables $\{G_a\}$. State μ is one of the many states yielding $\langle G_a\rangle = g_a$ (left solid black line). An arbitrary state ρ yields instead expectation values $g(\rho)$, as do all other states that lie on the right solid black line. States which minimize the relative entropy with respect to ρ, while satisfying given constraints on the expectation values $\{\langle G_a\rangle\}$, form a proper submanifold of state space (upper solid gray line); for $\langle G_a\rangle = g_a$, the pertinent state is μ^ρ_g. According to Bayes' rule, the posterior state estimate depends not only on the data but also on the prior. In particular, any estimate must be among the states with a nonvanishing prior probability (dotted gray line). Let the latter include some specific state σ. States which minimize the relative entropy with respect to this σ, while satisfying constraints on the $\{\langle G_a\rangle\}$, form another submanifold (lower solid gray line); for $\langle G_a\rangle = g_a$ and $\langle G_a\rangle = g_a(\rho)$, the pertinent states are μ^σ_g and $\mu^\sigma_{g(\rho)}$, respectively. If the relevance hypothesis holds, then the latter submanifold contains all states with nonvanishing prior probability; so then the lower solid gray line in fact covers the dotted gray line.

The first assertion implies that there must exist an algorithm $\{g_a\} \mapsto \rho$ assigning to any set of expectation values $\{g_a\} \equiv \{\langle G_a\rangle\}$ a unique state ρ. This algorithm need not necessarily be the maximum entropy algorithm. More generally, when the expectation values $\{g_a\}$ are not known exactly, but only their probability distribution $\mathrm{prob}(g)$, then there must exist an algorithm assigning to this probability distribution a unique probability distribution of states, $\mathrm{prob}(g) \mapsto \mathrm{prob}(\rho)$.

I now show that the second, logically independent, assertion singles out the maximum entropy algorithm and, hence, the Gibbs form, (3). In my proof I invoke various states and sets of states which are illustrated in Fig. 4. Let a tomographic measurement on a sample of large but finite size N come in two versions, one informationally complete and the other informationally incomplete. Whereas the complete version returns a unique tomographic image μ, the incomplete version merely returns sample means $\{g_a\}$ for the observables $\{G_a\}$. The latter are consistent with the former, $g_a = \langle G_a\rangle_\mu$. If indeed the observables $\{G_a\}$ are the relevant ones, then by the second assertion, it must not make a difference which of the two data sets, complete or incomplete, is processed in the Bayesian update, (5). Both must yield the same posterior, and so it must hold that

$$\text{prob}(\mu|\rho^{\otimes N})\text{prob}(\rho) \propto \text{prob}(\{g_a\}|\rho^{\otimes N})\text{prob}(\rho), \tag{15}$$

with a constant of proportionality that does not depend on ρ. This requirement can only be met if either the prior prob(ρ) vanishes or, by Eqs. (10) (with $f = g$) and (13), the difference in relative entropies $[S(\mu\|\rho) - S(\mu_g^\rho\|\rho)]$ is independent of ρ. By the law of Pythagoras for the relative entropy [59], this difference is itself a relative entropy:

$$S(\mu\|\mu_g^\rho) = S(\mu\|\rho) - S(\mu_g^\rho\|\rho). \tag{16}$$

It ought to have the same value for all ρ that have a nonvanishing prior probability. Let σ be one specific such state with nonvanishing prior probability. Then for all other ρ, it must hold that

$$S(\mu\|\mu_g^\rho) = S(\mu\|\mu_g^\sigma) \quad \forall\, \mu,\rho : \text{prob}(\rho) \neq 0. \tag{17}$$

In the special case $\mu = \rho$ it is $g_a = \langle G_a\rangle_\rho =: g_a(\rho)$ and hence also $\mu_g^\rho = \rho$, so the left-hand side vanishes. Then so must the right-hand side, and therefore

$$\rho = \mu_{g(\rho)}^\sigma \quad \forall\, \rho : \text{prob}(\rho) \neq 0. \tag{18}$$

Regardless of the experimental data, the admissible states (prob(ρ) $\neq 0$) are restricted *a priori* to Gibbs states of the generalized form, (12), with reference state σ and $\{F_b\} = \{G_a\}$.

The reference state σ may be any state that has a nonvanishing prior probability. Among the states with nonvanishing prior probability there is usually (albeit not necessarily always) the totally mixed state. If so, it will be most convenient to choose σ to be the totally mixed state. With the totally mixed state as the reference state, minimizing the relative entropy becomes equivalent to maximizing the ordinary von Neumann entropy; and then the generalized Gibbs form, (12), reduces to the ordinary Gibbs form, (3). QED.

The relevance hypothesis has significant implications for quantum-state tomography. It affects both the location (in state space) and the accuracy of the posterior state estimate: Given the same data, different choices for the set of relevant observables generally lead to different state estimates with different degrees of confidence. This can be illustrated with the simple example of state tomography on an exchangeable sequence of qubits. The state space of a single qubit is the Bloch sphere, with the totally mixed state at its origin. Let an incomplete tomographic measurement probe the Pauli spin component X, yielding sample mean $\bar{x}$. Associated with these data is a likelihood function on the Bloch sphere. It is broad on the two-dimensional plane containing states that yield $\langle X\rangle = \bar{x}$ and narrowly peaked in the direction perpendicular to this plane, the latter width decreasing with increasing sample size. Considering solely this likelihood function would lead to a maximum likelihood state estimate equal or close to $\rho_{\text{ml}} = (1 + \bar{x}X)/2$. However, as discussed in Sec. II, one must take into account also the prior; in particular, whether the prior has support in the entire Bloch sphere or in some subspace only.

For the present example I consider three cases: (i) all observables are relevant in the sense defined above; (ii) only X is relevant (say, because the physical qubit is a spin in a ferromagnet which is strongly anisotropic in the x direction); and (iii) only Y is relevant. Whereas in the first case the prior has support in the entire Bloch sphere, in the other two cases it has support only in the x or y axis, respectively. On the respective support, in the absence of further information, the prior is some broad symmetric distribution around the origin of the Bloch sphere. Multiplying the respective priors by the likelihood function yields the respective posteriors. These posteriors vary strongly from case to case; they are depicted schematically in Fig. 5.

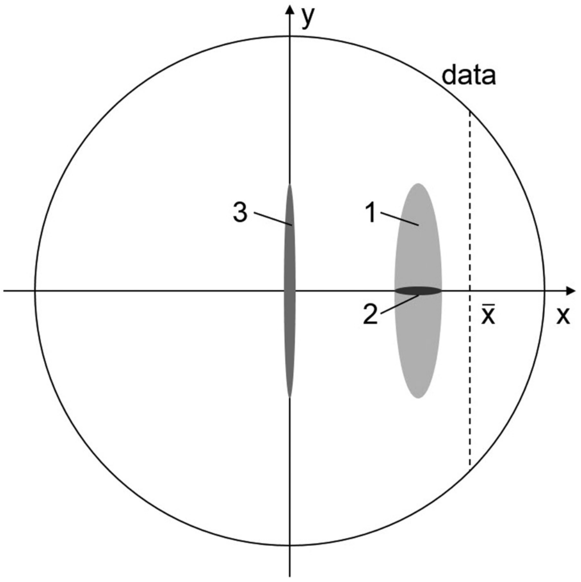

FIG. 5. Two-dimensional section ($z = 0$) of the Bloch sphere. The dashed vertical line at $x = \bar{x}$ indicates states yielding $\langle X\rangle = \bar{x}$, the observed sample mean. Shaded areas indicate the location and width of the posterior when the relevant observables comprise (1) all observables, (2) only X, or (3) only Y. In the first two cases the resultant state estimate (center of the posterior) lies somewhere between the initial bias (totally mixed state) and the data, the precise location depending on the size of the sample. The two cases differ in the degree of confidence regarding the unmeasured observable Y. In the third case the posterior equals the prior because the data carry no information about the then relevant observable Y.

V. MODEL SELECTION

In the preceding section I have shown how the relevance hypothesis constrains state estimates *a priori* to the Gibbs form, (3), and how this affects quantum-state estimation. One may wonder what in turn justifies the relevance hypothesis, and to which set of observables it should apply.

First, it is important to note that the relevant observables do not necessarily coincide with the observables which are being measured in an experiment; an observable is not relevant simply because one happens to measure it. In the above example the observable X was measured. And if the physical qubit was a spin in a strongly anisotropic ferromagnet with the preferred direction along the x axis, then indeed, the observable X would also be the relevant one. But if the preferred direction of the ferromagnet was along the y axis, the relevant observable would be Y rather than X—even though X was measured. Rather than the experimental setup, the relevance hypothesis reflects prior knowledge about the *physics* of the system. As is familiar from statistical mechanics, the choice of relevant observables is usually linked to time scales: If the system is in equilibrium, then the relevant observables comprise the system's constants of the motion; or else, provided that the system's degrees of freedom evolve on disparate time scales, they comprise the slow observables [24].

The above warning notwithstanding, identifying measured with relevant observables is precisely what is being done—implicitly—in maximum entropy quantum-state estimation: State estimates are effectively constrained to the Gibbs form with the measured observables in the exponent [5–11]. Actually, this identification is often a useful shortcut (albeit not usually spelled out as such) because, for an observable to be measurable in practice, it must vary slowly; and as long as the system exhibits a clear hierarchy of time scales, "slow" means indeed "relevant."

Yet when one investigates a hitherto unknown substance, a hierarchy of time scales or other clues as to the relevant observables are not available *a priori;* nor is there any assurance that the shortcut "measured = relevant" is warranted. In this situation one can only formulate conjectures as to the set of relevant observables. The relevance hypothesis then becomes truly a *hypothesis* in the statistical sense, subject to experimental scrutiny and possibly refutation. There might be several competing hypotheses, perhaps even including the hypothesis that the system cannot be described by Gibbs states at all—that is, unless the set of relevant observables is extended to become informationally complete, in which case *all* observables would be relevant. Every proposal as to the set of relevant observables constitutes a statistical model, in the sense that state estimates are constrained to a certain parametric form (Gibbs form) with a certain number of adjustable parameters (Lagrange parameters). Choosing among rival proposals in the light of the experimental data then becomes an instance of statistical model selection. It is this scenario on which I focus in the present section.

First, some general remarks about statistical model selection may be in order. In general, rival statistical models for the same experimental data differ in the number and type of adjustable parameters. On the one hand, a model ought to be in good agreement with the data, which is best achieved with a large number of adjustable parameters; yet on the other hand, excessive complexity must be avoided ("Occam's razor"). The purpose of model selection is to render this trade-off quantitative. How this works in practice can be illustrated with a simple textbook example [60,61]. Let A be a simple model without an adjustable parameter and B a more complex model with one adjustable parameter λ. Which model is to be preferred on the basis of data D will be determined by the ratio of their respective posterior probabilities. Due to Bayes' rule, this ratio is given by

$$\frac{\text{prob}(A|D)}{\text{prob}(B|D)} = \frac{\text{prob}(D|A)}{\text{prob}(D|B)}\,\frac{\text{prob}(A)}{\text{prob}(B)}. \tag{19}$$

As long as one does not have any strong *a priori* preference for either of the models the right-hand side will be dominated by the first factor, the ratio of likelihoods. By the law of total probability, the likelihood function of model B reads

$$\text{prob}(D|B) = \int d\lambda \text{prob}(D|\lambda,B)\text{prob}(\lambda|B). \tag{20}$$

Let λ_0 be the value of the adjustable parameter that yields the best fit with the experimental data. Then the first factor in the integral, considered as a function of λ, will be peaked around a maximum at λ_0; a typical shape is a Gaussian

$$\text{prob}(D|\lambda,B) = \text{prob}(D|\lambda_0,B)\ \exp\left[-\frac{(\lambda-\lambda_0)^2}{2(\delta\lambda)^2}\right] \tag{21}$$

of some width $\delta\lambda$. This width indicates the accuracy to which the parameter λ is known after processing the experimental data. In contrast, the second factor in the integral describes the distribution of λ *prior* to processing the data; this *a priori* distribution has a larger width, $\Delta\lambda > \delta\lambda$. Provided that the best fit λ_0 lies within the *a priori* expected range, the ratio of likelihoods will scale as

$$\frac{\text{prob}(D|A)}{\text{prob}(D|B)} \sim \frac{\text{prob}(D|A)}{\text{prob}(D|\lambda_0,B)}\,\frac{\Delta\lambda}{\delta\lambda}. \tag{22}$$

The first ratio on the right-hand side is typically <1, favoring the more complex model B, because thanks to its adjustable parameter, B can achieve a better fit with the data. In contrast, the second ratio $(\Delta\lambda/\delta\lambda)$ is >1, favoring the simpler model A; this is the quantitative manifestation of Occam's razor. It is thus the relative value of these two ratios which will tip the balance in favor of one specific model.

The same logic applies to the identification of the most plausible set of relevant observables on the basis of experimental data. Details of the pertinent statistical analysis have been spelled out by the author (in a different context) in previous publications for two basic scenarios. In the first scenario [42], different hypotheses as to the set of relevant observables are formulated *prior* to experiment, and subsequently the experiment is performed on a single sample only. In the second scenario [58], which resembles more closely the way an unknown substance is investigated in practice, measurements are performed first, before formulating any hypotheses. Moreover, measurements are performed not just on a single sample but on multiple samples of the same unknown substance. These samples need not—in fact, ought not—be in the same state; yet it is hypothesized that for all these samples the *same* set of observables is relevant. Their respective states should therefore all lie in the same submanifold of Gibbs states, possibly with varying values for the associated Lagrange parameters. It is this second scenario on which I focus here.

For instance, the samples might all have been brought into contact with heat baths at different temperatures. Then one hypothesis might say that they are now all in canonical states, the Hamiltonian (identical for all samples) being the sole relevant observable, with only the temperature varying across samples. Another hypothesis might say that beyond the Hamiltonian there are further constants of the motion (again, identical for all samples) which need to be added to the set of relevant observables, increasing the number of adjustable Lagrange parameters. And a third, extreme hypothesis might claim that the samples have not thermalized fully and that, hence, a Gibbs form is not justified and *all* observables remain relevant. As in the textbook example above, increasing thus the

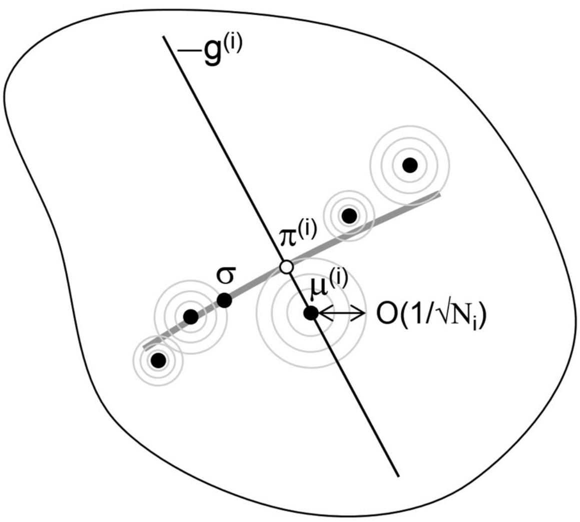

FIG. 6. Modeling the data with Gibbs states. Filled circles represent tomographic images for various samples of the same substance. Gray concentric circles around them symbolize the associated likelihood functions, whose widths are an indicator of the measurement uncertainties. These uncertainties vary by sample and typically scale as $1/\sqrt{N_i}$. The relevance hypothesis, which is to be tested, claims that all data can be modeled on a joint manifold of Gibbs states (thick gray line) with reference state σ. If so, the Gibbs state closest to a tomographic image $\mu^{(i)}$ is state $\pi^{(i)}$ (open circle). Both states yield, for the relevant observables, the same expectation values $\{g_a^{(i)}\}$; i.e., they both belong to the set of states yielding $\langle G_a\rangle = g_a^{(i)}$ (solid black line).

number of adjustable parameters will lead to a successively better fit with the experimental data, yet at the expense of simplicity. Again, goodness-of-fit has to be traded off against simplicity in a quantitative fashion. The optimal trade-off yields the most plausible set of relevant observables.

Here I briefly summarize the key findings of the quantitative analysis; details are reported in Ref. [58]. Various differently prepared samples are subjected to the same tomography, which, if not complete, must encompass at least all candidates for relevancy. Let the ith sample have a large but finite size N_i, and let tomography on this sample render the tomographic image $\mu^{(i)}$. (If the tomography is not complete, the image $\mu^{(i)}$ is constructed via ordinary maximum entropy state estimation.) Assuming that the finite sample size is the principal source of noise, error bars on the measurements are of the order $1/\sqrt{N_i}$. As before, let σ be a reference state (usually the totally mixed state) with nonvanishing prior probability, and let $\mathcal{G}$ denote the hypothesis that for all samples the p observables $\{G_a\}_{a=1}^p$ are the relevant ones in the sense defined above. Given this hypothesis, states are constrained *a priori* to Gibbs states of the generalized form μ_g^σ. For a tomographic image $\mu^{(i)}$ the closest such state is $\pi^{(i)} := \mu^\sigma_{g^{(i)}}$, where $g_a^{(i)} := \langle G_a\rangle_{\mu^{(i)}}$ (Fig. 6). In terms of these variables, and under certain reasonable additional assumptions spelled out in Ref. [58], the log-likelihood of the hypothesis $\mathcal{G}$ behaves asymptotically as

$$\ln \text{prob}(D|\mathcal{G}) \sim -\sum_i N_i S(\mu^{(i)}\|\pi^{(i)}) - \frac{p}{2}\sum_i \ln N_i \qquad (23)$$

modulo additive constants that do not depend on the choice of relevant observables. Here D is short for the totality of experimental data. As long as there is no strong *a priori* preference for a specific set of relevant observables, the difference in the log-likelihoods of rival hypotheses dominates their relative posterior probabilities.

The above formula for the log-likelihood reflects in a quantitative fashion the expected trade-off between goodness-of-fit and simplicity. On the right-hand side there are two contributions, both with a negative sign and thus "penalizing"—in terms of likelihood—hypothesis $\mathcal{G}$. The first contribution penalizes a bad fit to the data: The farther away the Gibbs states $\pi^{(i)}$ are from the original tomographic images $\mu^{(i)}$, the higher the relative entropies $S(\mu^{(i)}\|\pi^{(i)})$ and, hence, the larger the penalty. To avoid this penalty, the set of relevant observables should be chosen sufficiently large. In contrast, the second contribution embodies Occam's razor, penalizing excessive complexity: The larger the number p of relevant observables, the larger the penalty. In order to avoid the latter penalty, the set of relevant observables should be kept as small as possible. So in line with our general considerations, one must trade off these two penalties in order to find the most likely set of relevant observables.

In sum, the relevance of observables reflects underlying physics and is independent of their being measured. For a hitherto unknown substance, whose underlying physics is yet to be explored, the relevant observables must be inferred from the data. This is achieved via the above statistical analysis. The analysis yields the set of relevant observables that is most likely, which, if informationally incomplete, in turn implies a corresponding Gibbs form. In this sense, the use of Gibbs states is corroborated solely by the data.

VI. CONCLUSIONS

In the preceding sections I have explored the extent to which the use of Gibbs states can be understood with the help of methods from quantum-state tomography. These methods apply to systems that are finite and isolated; hence they require neither the limit $N \to \infty$ nor auxiliary concepts such as infinite ensembles or large environments. (However, I did assume that sample sizes are large enough to justify the use of the asymptotic Sanov likelihood to good accuracy.) Moreover, they refrain from invoking information-theoretical arguments or exploiting the system's dynamics.

Without any knowledge of the system's dynamics or other clues as to the relevant observables, of course, it is impossible to derive a Gibbs form from first principles. Rather, in this situation the Gibbs form constitutes a statistical *hypothesis* that can be supported or refuted only by data. In Sec. IV, I gave a precise formulation of the relevance hypothesis which is behind the use of the Gibbs form. And in Sec. V, I outlined the statistical tools needed to test this hypothesis in the light of experimental data.

Taken together, the statistical methods discussed in this paper comprise a toolbox which can be used to ascertain (i) whether a hitherto unknown substance, for which, in particular, the dynamics and the constants of the motion are not known, can be described by a Gibbs state at all; and (ii) if it can, which observables are most likely the relevant ones. The pertinent statistical analysis is based entirely on the tomographic data gleaned from a collection of differently prepared, finite samples. It focuses on the relative likelihoods rather than the posterior probabilities of rival hypotheses. The former are a good proxy for the latter as long as there is no prior knowledge and, hence, no *a priori* bias in favor of any particular hypothesis.

Once the set of relevant observables, valid for the entire collection of samples, has been established, there remains the statistical task of estimating the values of the pertinent Lagrange parameters for any given *individual* sample. This is an interesting subject in itself, which has been dealt with elsewhere [41,42].

ACKNOWLEDGMENTS

I thank the anonymous referee for helpful comments. This work was supported by the EU Integrating Project SIQS, the EU STREP EQUAM, and the ERC Synergy grant BioQ.

[1] A. E. Allahverdyan, R. Balian, and T. M. Nieuwenhuizen, J. Mod. Opt. **51**, 2703 (2004).

[2] A. E. Allahverdyan, R. Balian, and T. M. Nieuwenhuizen, Europhys. Lett. **67**, 565 (2004).

[3] D. Janzing, J. Stat. Phys. **122**, 531 (2006).

[4] F. Becattini, P. Castorina, J. Manninen, and H. Satz, Eur. Phys. J. C **56**, 493 (2008).

[5] V. Buzek, G. Drobny, G. Adam, R. Derka, and P. L. Knight, J. Mod. Opt. **44**, 2607 (1997).

[6] V. Buzek, R. Derka, G. Adam, and P. L. Knight, Ann. Phys. **266**, 454 (1998).

[7] V. Buzek, G. Drobny, R. Derka, G. Adam, and H. Wiedemann, Chaos Sol. Frac. **10**, 981 (1999).

[8] V. Buzek and G. Drobny, J. Mod. Opt. **47**, 2823 (2000).

[9] V. Buzek, Lect. Notes Phys. **649**, 189 (2004).

[10] G. Drobny and V. Buzek, Phys. Rev. A **65**, 053410 (2002).

[11] S. Deléglise, I. Dotsenko, C. Sayrin, J. Bernu, M. Brune, J.-M. Raimond, and S. Haroche, Nature **455**, 510 (2008).

[12] R. Balian and N. L. Balazs, Ann. Phys. **179**, 97 (1987).

[13] S. Goldstein, J. L. Lebowitz, R. Tumulka, and N. Zanghì, Phys. Rev. Lett. **96**, 050403 (2006).

[14] S. Popescu, A. J. Short, and A. Winter, Nature Phys. **2**, 754 (2006).

[15] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).

[16] E. T. Jaynes, Phys. Rev. **108**, 171 (1957).

[17] A. Katz, *Principles of Statistical Mechanics: The Information Theory Approach* (W. H. Freeman, New York, 1967).

[18] R. Baierlein, *Atoms and Information Theory* (W. H. Freeman, New York, 1971).

[19] L. Boltzmann, Wiener Berichte **66**, 275 (1872).

[20] P. Reimann, Phys. Rev. Lett. **101**, 190403 (2008).

[21] N. Linden, S. Popescu, A. J. Short, and A. Winter, Phys. Rev. E **79**, 061103 (2009).

[22] A. J. Short, New J. Phys. **13**, 053009 (2011).

[23] A. Riera, C. Gogolin, and J. Eisert, Phys. Rev. Lett. **108**, 080402 (2012).

[24] J. Rau and B. Müller, Phys. Rep. **272**, 1 (1996).

[25] J. E. Shore and R. W. Johnson, IEEE Trans. Inf. Theory **26**, 26 (1980).

[26] J. Skilling, in *Maximum Entropy and Bayesian Methods*, edited by G. J. Erickson and C. R. Smith (Kluwer Academic, Dordrecht, 1988), pp. 173–187.

[27] Y. Tikochinsky, N. Z. Tishby, and R. D. Levine, Phys. Rev. Lett. **52**, 1357 (1984).

[28] J. Uffink, Stud. Hist. Phil. Mod. Phys. **26**, 223 (1995).

[29] J. Uffink, Stud. Hist. Phil. Mod. Phys. **27**, 47 (1996).

[30] G. M. D'Ariano, M. G. A. Paris, and M. F. Sacchi, Lect. Notes Phys. **649**, 7 (2004).

[31] K. Banaszek, M. Cramer, and D. Gross, New J. Phys. **15**, 125020 (2013).

[32] J. B. Hartle, Am. J. Phys. **36**, 704 (1968).

[33] H. Häffner, W. Hänsel, C. F. Roos, J. Benhelm, D. Chek-al-kar, M. Chwalla, T. Körber, U. D. Rapol, M. Riebe, P. O. Schmidt, C. Becher, O. Gühne, W. Dür, and R. Blatt, Nature **438**, 643 (2005).

[34] C. M. Caves, C. A. Fuchs, and R. Schack, J. Math. Phys. **43**, 4537 (2002).

[35] B. de Finetti, *Theory of Probability* (Wiley, New York, 1990).

[36] R. Schack, T. A. Brun, and C. M. Caves, Phys. Rev. A **64**, 014305 (2001).

[37] Z. Hradil, Phys. Rev. A **55**, R1561 (1997).

[38] D. F. V. James, P. G. Kwiat, W. J. Munro, and A. G. White, Phys. Rev. A **64**, 052312 (2001).

[39] Z. Hradil, J. Rehacek, J. Fiurasek, and M. Jezek, Lect. Notes Phys. **649**, 59 (2004).

[40] K. M. R. Audenaert and S. Scheel, New J. Phys. **11**, 023028 (2009).

[41] J. Rau, Phys. Rev. A **82**, 012104 (2010).

[42] J. Rau, Phys. Rev. A **84**, 012101 (2011).

[43] C. A. Fuchs and R. Schack, AIP Conf. Proc. **1101**, 255 (2009).

[44] M. Hayashi, J. Phys. A: Math. Gen. **35**, 10759 (2002).

[45] I. Bjelaković, J.-D. Deuschel, T. Krüger, R. Seiler, R. Siegmund-Schultze, and A. Szkoła, Commun. Math. Phys. **260**, 659 (2005).

[46] K. Audenaert, M. Nussbaum, A. Szkoła, and F. Verstraete, Commun. Math. Phys. **279**, 251 (2008).

[47] I. N. Sanov, Mat. Sb. (N.S.) **42**(84), 11 (1957).

[48] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley, New York, 2006).

[49] J. D. Deuschel and D. W. Stroock, *Large Deviations* (Academic Press, New York, 1989).

[50] W. Ochs, Rep. Math. Phys. **9**, 135 (1976).

[51] A. Wehrl, Rev. Mod. Phys. **50**, 221 (1978).

[52] M. J. Donald, Commun. Math. Phys. **105**, 13 (1986).

[53] V. Vedral, Rev. Mod. Phys. **74**, 197 (2002).

[54] M. B. Ruskai, Rep. Math. Phys. **26**, 143 (1988).

[55] F. Hiai and D. Petz, Commun. Math. Phys. **143**, 99 (1991).

[56] T. Ogawa and H. Nagaoka, IEEE Trans. Inf. Theory **46**, 2428 (2000).

[57] F. G. S. L. Brandão and M. B. Plenio, Commun. Math. Phys. **295**, 791 (2010).

[58] J. Rau, Phys. Rev. A **84**, 052101 (2011).

[59] D. Petz, *Quantum Information Theory and Quantum Statistics* (Springer, Berlin, 2008).

[60] D. S. Sivia, W. I. F. David, K. S. Knight, and S. F. Gull, Physica D **66**, 234 (1993).

[61] D. S. Sivia, *Data Analysis: A Bayesian Tutorial* (Oxford University Press, Oxford, 1996).