Inferring the Gibbs state of a small quantum system

Jochen Rau*

Institut für Theoretische Physik, Johann Wolfgang Goethe-Universität, Max-von-Laue-Strasse 1, D-60438 Frankfurt am Main, Germany

(Received 22 August 2010; published 1 July 2011)

Gibbs states are familiar from statistical mechanics, yet their use is not limited to that domain. For instance, they also feature in the maximum entropy reconstruction of quantum states from incomplete measurement data. Outside the macroscopic realm, however, estimating a Gibbs state is a nontrivial inference task, due to two complicating factors: the proper set of relevant observables might not be evident *a priori*; and whenever data are gathered from a small sample only, the best estimate for the Lagrange parameters is invariably affected by the experimenter's prior bias. I show how the two issues can be tackled with the help of Bayesian model selection and Bayesian interpolation, respectively, and illustrate the use of these Bayesian techniques with a number of simple examples.

DOI: 10.1103/PhysRevA.84.012101 PACS number(s): 03.65.Wj, 02.50.Cw, 02.50.Tt, 05.30.Ch

I. INTRODUCTION

Quantum states are not accessible to direct observation, and so do not constitute *per se* a physical reality. Rather, they provide a convenient mathematical summary of an agent's expectations as to the outcomes of future experiments [1]. Such expectations are formed both on the basis of past measurement data and on the basis of any prior knowledge (say, about specific symmetries) that the agent may have. In practice, the available experimental data are often far from perfect: measurement devices work with limited accuracy; sample sizes are finite; and the set of observables measured might not be informationally complete. Under such circumstances, a quantum state represents merely a *model*, and hence a hypothesis, which is subject to testing, debate, and modification. The more complex the physical system under study, and the sketchier the available data, the more this model will be informed by the agent's prior knowledge.

Prior knowledge may be of two types: (i) the expectation, often based on symmetry considerations, that the quantum state has a certain *parametric form*; and (ii) given a parametric form (including free-form as a special case), a bias as to its *parameter values*. Making proper use of such prior knowledge can lead to significant gains in the efficiency and accuracy of quantum-state tomography, i.e., the reconstruction of a quantum state from imperfect data. One recent example where prior knowledge about the parametric form has been exploited to great advantage is the polynomial scheme for reconstructing near matrix product states [2]. The second type of prior knowledge, on the other hand, has been used in recent Bayesian modifications to the conventional maximum likelihood tomography scheme [3–5].

One parametric form that occupies a special place in physics is that of a *Gibbs state*. Such a state maximizes the entropy, or more generally, minimizes the relative entropy with respect to some reference state, under given constraints on some selected set of expectation values. Gibbs states are familiar from statistical mechanics where, in both the classical [6,7] and quantum [8] cases, the principle of maximum entropy has long been recognized as the appropriate prescription for constructing the macrostate. The common justification of this principle rests on a number of assumptions: (i) the system under consideration may be viewed as one constituent of a larger ensemble of identically prepared systems, whose size approaches infinity (the "thermodynamic limit"); (ii) there are constraints pertaining to the global state of this fictitious infinite ensemble, in the form of sharp values for the totals of certain "relevant" observables; and (iii) it is clear *a priori* which observables should be deemed "relevant." While the first two assumptions are of a purely statistical nature, the last implicitly invokes the system's dynamics. In equilibrium statistical mechanics, the relevant observables are taken to be the system's constants of the motion; whereas in nonequilibrium transport theory, they comprise those degrees of freedom that vary slowly [9].

Yet Gibbs states play an important role even in realms where the above assumptions are not justified. For instance, application of thermodynamics has been extended to nanoscale quantum systems, and in particular, to the study of work extraction from such systems [10–13]. In high-energy physics, hadronization in e^+e^- collisions is described with thermal distributions, even though the number of hadrons produced in one collision is hardly more than a handful [14]. The equilibrium states of completely integrable many-body quantum systems need not be thermal but can be described very well with generalized Gibbs distributions [15]. And finally, in incomplete quantum-state tomography, Gibbs models feature in reconstruction schemes based on maximum entropy [16–20], or in case there is an initial bias toward some nonuniform reference state, on the principle of minimum relative entropy [21]; such schemes have proven remarkably successful in practice [22,23]. In all these examples, the measurement data, and hence any derived constraints, no longer pertain to fictitious infinite ensembles but to finite samples, which can be quite small. And often, one can no longer easily discern slow from fast degrees of freedom. In such case the choice of relevant observables is no longer obvious, and in particular, need not necessarily be related to the set of observables that are being measured.

When prior knowledge suggests that a small physical system ought to be described by a Gibbs state, yet the proper set of relevant observables is not evident *a priori*, the choice of the latter becomes a matter of statistical inference. Competing

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

theories might propose different sets of relevant observables; and the task is then to decide rationally between them on the basis of rather sketchy data. Typically, this inference task involves a trade-off between goodness-of-fit on the one hand, favoring a large number of relevant observables; and simplicity on the other ("Occam's razor"), favoring a number that is as small as possible. The appropriate framework for deciding such a trade-off is *Bayesian model selection* [24]. Adapting this framework to the task of finding the optimal set of relevant observables in a Gibbs model, be it classical or quantum, is one central objective of the present paper.

Once the set of relevant observables is fixed, the next inference task is the estimation of the associated Lagrange parameters. Whenever the sample size is small, the estimate must take into account not only experimental data but also prior expectations. The fact that the prior bias invariably exerts an influence on the parameter estimate is readily seen in a trivial example: If five tosses of a coin yield "heads" five times, one is not yet ready to abandon one's prior bias toward a more or less fair coin; only as evidence to the contrary accumulates, does this belief gradually erode. The challenge, then, is to find the relative weights to be attributed to prior bias and data. Again, the appropriate tools are furnished by Bayesian theory, namely *Bayesian interpolation* [25] in combination with the *evidence procedure* [5]. Putting these tools to use for the estimation of Lagrange parameters in a Gibbs model, is the second main objective of the present paper.

The paper is organized as follows. In Sec. II, I will start with some preliminaries about the χ^2 distribution, the entropy concentration theorem, and the concept of statistical significance, which will be needed in subsequent arguments. Then I shall turn to the two inference problems outlined above, albeit in reverse order. In Sec. III, I will assume that the relevant observables of a Gibbs model are given and show how its Lagrange parameters can be estimated in a way that accounts for both prior knowledge and measured data. The estimation procedure will yield not just the optimal values for the Lagrange parameters but also the associated error bars. In Sec. IV, I shall consider the issue of the proper set of relevant observables and show that Bayesian model selection provides a rational framework for choosing between rival proposals. I will illustrate this method with two examples, the classical analysis of Wolf's die (Sec. V) and the quantum problem of deciding between an Ising and a Heisenberg description of an assembly of qubits (Sec. VI). In Sec. VII, I shall conclude with a brief summary.

There are a number of appendices in which I collect technical definitions and results that might not be familiar to readers of this journal, yet whose inclusion in the main body of the text would render the flow of exposition unnecessarily cumbersome. Specifically, in Appendix A, I shall introduce the notion of a level of description; in Appendix B, the notions of coarse graining, relevant part of a state, and generalized Gibbs states; and in Appendix C, the definition and basic geometry of a Gibbs manifold, which includes as a special case (discussed in Appendix D) the geometry of the Bloch sphere. In Appendix E, I will introduce the concept of an entropic distribution on the Gibbs manifold; and in Appendix F, I will consider the meaning of the Gaussian approximation and of the thermodynamic limit. Finally, in Appendix G, I shall connect the general framework of Gibbs models to the familiar terminology and basic relations of thermodynamics.

II. STATISTICAL SIGNIFICANCE

In a generic experiment, one measures some selected set of observables, which span the *experimental level of description* $\mathcal{F}$, on N identically prepared copies of a physical system, yielding sample means f. Given a reference state σ, these experimental data can always be represented as a Gibbs model $\mu \in \pi^{\sigma}_{\mathcal{F}}(\mathcal{S})$, with Lagrange parameters adjusted such as to reproduce the observed sample means, $f(\mu) = f$. This correspondence $f \leftrightarrow \mu$ is one-to-one. On the same Gibbs manifold, let ρ denote a theoretical model yielding expectation values $f(\rho)$; these generally differ from the observed sample means. As long as the difference is small, the relative entropy between data and theoretical model is approximately quadratic in the differentials δf. According to Eq. (C9), it is

$$2NS(\mu\|\rho) \approx \chi^2(\mu\|\rho), \tag{1}$$

with

$$\chi^2(\mu\|\rho) := N \sum_{ab} (C^{-1})^{ab} \delta f_a \delta f_b. \tag{2}$$

For definitions of the mathematical objects used here $(\mathcal{F}, \pi^{\sigma}_{\mathcal{F}}(\mathcal{S}), C^{-1})$, see Appendices A through C.

Since the theoretical model may contain parameters that have been fitted to the data, the differentials δf might not be all independent; the number k of independent differentials is generally smaller than the dimension of the Gibbs manifold. Given a (possibly fitted) theoretical model, the likelihood that the k remaining independent degrees of freedom yield a χ^2 in the interval $[x, x + dx]$ is determined for large N by the probability density function

$$\mathrm{PDF}(x|k) = 2^{-k/2}\Gamma(k/2)^{-1} x^{k/2-1} \exp(-x/2), \tag{3}$$

known as the χ^2 distribution [26]. In this distribution, the exponential factor stems from the quantum Stein lemma (B4); the power factor from the k-dimensional volume element; and the numerical factors ensure proper normalization. The χ^2 distribution is peaked at $\chi^2_{\max} = k - 2$ (for $k > 2$) and has expectation value and variance

$$\langle \chi^2 \rangle = k, \quad \mathrm{var}(\chi^2) = 2k, \tag{4}$$

respectively. For large arguments ($x \gg k \ln k$), it has an exponential tail that is independent of k,

$$\mathrm{PDF}(x|k) \sim \exp(-x/2). \tag{5}$$

The above distribution of χ^2, and hence of relative entropy, implies the *entropy concentration theorem* [27]: As the sample size N increases, the relative entropy between data and theoretical model is predicted to become more and more concentrated (with a width of order $1/N$) around a smaller and smaller expectation value (also of order $1/N$). Relative entropy being approximately quadratic in the coordinate differentials δf, this implies as a corollary that deviations between measured sample means and theoretical expectation values are expected to scale as $O(1/\sqrt{N})$. The entropy concentration theorem can thus be employed to assess quickly the statistical significance

of experimental deviations from theoretical predictions: As long as the relative entropy between data and theoretical model is of order $1/N$, deviations likely fall within the range of statistical fluctuations; yet as soon as their relative entropy exceeds this limit, deviations become significant and may indicate the need to revise the theoretical model.

This simple entropy test is closely related to the χ^2 test in conventional statistics. Theoretical models are typically rejected whenever at the observed χ^2 the cumulative distribution function exceeds a predefined bound, whose value in turn depends on the confidence level required. The χ^2 test then points to the need to revise the theoretical model and thus triggers creative thinking about possible alternatives.

III. ESTIMATING LAGRANGE PARAMETERS

I consider the situation where it is assumed from the outset that a physical system ought to be described by some Gibbs model $\omega \in \pi^\sigma_\mathcal{G}(\mathcal{S})$, with given reference state σ and level of description $\mathcal{G}$, yet unknown parameter values. The initial uncertainty about the parameter values is reflected in a prior probability distribution $\text{prob}(\omega|\sigma,\mathcal{G})$ over the Gibbs manifold. Subsequently, on a sample of size N, one measures some set of sample means f. The associated experimental level of description $\mathcal{F}$ may or may not coincide with the theoretical level of description $\mathcal{G}$. In light of the observed sample means and the prior distribution over the Gibbs manifold, one wants to infer the most plausible estimate for ω.

After collecting the experimental data, the probability distribution over the Gibbs manifold must be updated according to the *Bayes rule* [28]:

$$\text{prob}(\omega|f,N,\mathcal{F};\sigma,\mathcal{G}) \propto \text{prob}(f|N,\omega,\mathcal{F})\,\text{prob}(\omega|\sigma,\mathcal{G}). \tag{6}$$

The first factor on the right-hand side is the *likelihood* of observing the sample means f, given ω; it is

$$\text{prob}(f|N,\omega,\mathcal{F}) \propto \int_{\mathcal{S}|_f} d\rho\,\text{prob}(\rho|N,\omega), \tag{7}$$

with the integration ranging over the submanifold $\mathcal{S}|_f$ of all states that satisfy the constraints $f(\rho) = f$, and normalized according to Eq. (C10):

$$\int \prod_b df_b \sqrt{\det C^{-1}}\,\text{prob}(f|N,\omega,\mathcal{F}) = 1. \tag{8}$$

For large N, by virtue of the quantum Stein lemma (B4) and the law of Pythagoras (F6), the likelihood can be written as

$$\text{prob}(f|N,\omega,\mathcal{F}) \propto \exp[-NS(\mu\|\omega)], \tag{9}$$

where $\mu \in \pi^\omega_\mathcal{F}(\mathcal{S})$ is the unique Gibbs model associated with the measured f and reference state ω. In other words, on the Gibbs manifold $\pi^\omega_\mathcal{F}(\mathcal{S})$, the Gibbs model μ representing experimental data is distributed entropically around ω, $\mu \sim \text{E}(N,\omega,\mathcal{F})$ (see Appendix E).

The second factor on the right hand side of the Bayes rule [Eq. (6)] is the *prior*. In principle, it can take any form; there is no constraint as to the prior knowledge that an agent may have. But there are both conceptual and practical reasons to assume it to be entropic, too, $\omega \sim \text{E}(\alpha,\sigma,\mathcal{G})$, with its peak at some initial bias σ, and the parameter α characterizing the agent's degree of confidence as to this initial bias. Conceptually, if σ is the only prior knowledge available, it is reasonable to demand of a prior that it be peaked at and symmetric around this bias; that it be form-invariant under coarse graining; and that upon composition of systems, it be noncommittal as to any correlations between the systems. As I discuss in more detail in Appendix E, these requirements are indeed satisfied by entropic distributions. Moreover, an entropic prior is particularly convenient because it is (approximately) conjugate to the likelihood (9). Thanks to this property, any measurement that is informationally complete with respect to the unknown model parameters, $\mathcal{F} \supset \mathcal{G}$, yields a Bayesian update which (in the Gaussian approximation) is again entropic and which differs from the prior only by a change of parameters: $(\alpha,\sigma) \to (\alpha',\sigma')$.

Assuming an entropic prior and making the Gaussian approximation, the Bayes rule yields for $\mathcal{F} \supset \mathcal{G}$ the posterior (F8), and hence

$$\text{prob}(\omega|f,N,\mathcal{F};\sigma,\mathcal{G}) \propto \text{prob}(\omega|\alpha+N,\rho,\mathcal{G}); \tag{10}$$

whereas for $\mathcal{F} \subset \mathcal{G}$, it yields the posterior (F9), and hence

$$\begin{aligned}&\text{prob}(\omega|f,N,\mathcal{F};\sigma,\mathcal{G})\\ &\quad\propto \text{prob}\big(\pi^\sigma_\mathcal{F}(\omega)|\alpha+N,\rho,\mathcal{F}\big)\text{prob}\big(\pi^\rho_{\neg_{\mathcal{G},\rho}\mathcal{F}}(\omega)|\alpha,\rho,\neg_{\mathcal{G},\rho}\mathcal{F}\big).\end{aligned} \tag{11}$$

Both posteriors are peaked at the model

$$\rho \propto \exp\left[\frac{\alpha}{\alpha+N}\ln\sigma + \frac{N}{\alpha+N}\ln\pi^\sigma_{\mathcal{F}\cap\mathcal{G}}(\mu)\right], \tag{12}$$

which constitutes thus the most plausible posterior estimate for ω.

The posterior estimate for ω interpolates between initial bias and data, and depending on the relative sizes of α and N, may attribute more weight to one or the other; this is an example of *Bayesian interpolation* [25]. In the extreme case where the prior is sharply peaked while sample sizes are small, $N \ll \alpha$, parameter estimation will be dominated by the prior, and one is therefore advised to stick to the initial bias, $\rho \approx \sigma$; while in the opposite case where the prior is broad and sample sizes are big, $N \gg \alpha$, parameter estimation will be dominated by the likelihood function, and the best estimate for the model is close to the *maximum likelihood* estimate $\rho \approx \pi^\sigma_{\mathcal{F}\cap\mathcal{G}}(\mu)$. Attached to the estimate are error bars of the order $O(1/\sqrt{\alpha+N})$ as to those model parameters that have been measured, and in case $\mathcal{F} \subset \mathcal{G}$, $O(1/\sqrt{\alpha})$ as to those that have not.

The above estimation procedure preserves the Gibbs form, in the following sense. Whenever the prior bias σ is a generalized Gibbs state, with some level of description that encloses $\mathcal{F} \cap \mathcal{G}$, then for arbitrary values of α and N, the posterior estimate retains this parametric form:

$$\sigma \in \pi^\tau_\mathcal{H}(\mathcal{S})\,,\ \mathcal{H} \supset \mathcal{F}\cap\mathcal{G} \Rightarrow \rho \in \pi^\tau_\mathcal{H}(\mathcal{S}); \tag{13}$$

in particular, it is $\rho \in \pi^\sigma_{\mathcal{F}\cap\mathcal{G}}(\mathcal{S})$. If on the Gibbs manifold $\pi^\tau_\mathcal{H}(\mathcal{S})$ the prior bias has Lagrange parameters $\lambda(\sigma)$, and the maximum likelihood estimate has Lagrange parameters $\lambda(\pi^\sigma_{\mathcal{F}\cap\mathcal{G}}(\mu))$, then

the Lagrange parameters of the posterior estimate [Eq. (12)] are given by linear interpolation:

$$\lambda(\rho) = \frac{\alpha}{\alpha + N}\lambda(\sigma) + \frac{N}{\alpha + N}\lambda\big(\pi^{\sigma}_{\mathcal{F}\cap\mathcal{G}}(\mu)\big). \tag{14}$$

The posterior estimate depends critically on the parameter α, which so far has been left unspecified. Provided the experimental and theoretical levels of description are sufficiently detailed and the experiment reveals a significant deviation from the initial bias (only then does the need arise to update this bias),

$$\chi^2\big(\pi^{\sigma}_{\mathcal{F}\cap\mathcal{G}}(\mu)\|\sigma\big) > \dim \pi^{\sigma}_{\mathcal{F}\cap\mathcal{G}}(\mathcal{S}) \gg 1, \tag{15}$$

the optimal value for α can be estimated to a good approximation with the help of the evidence procedure [5]. The procedure yields the interpolation parameter

$$\alpha/(\alpha + N) \approx \dim \pi^{\sigma}_{\mathcal{F}\cap\mathcal{G}}(\mathcal{S})/\chi^2\big(\pi^{\sigma}_{\mathcal{F}\cap\mathcal{G}}(\mu)\|\sigma\big). \tag{16}$$

As long as $\mathcal{F} \subset \mathcal{G}$, this estimate depends on the experimental level of description $\mathcal{F}$ only, and not on the larger theoretical level $\mathcal{G}$ employed for the Gibbs model ω.

To illustrate the above framework, I consider the following simple example. A source emits a physical system (say, a molecule) that can be in its ground state or in one of 24 excited states; this spectrum may or may not be degenerate. Prior theoretical considerations suggest that the source is thermal, and hence, that the occupation probabilities $\{p_i\}, i = 0 \ldots 24$, of the energy levels follow a canonical distribution. There is uncertainty about the temperature, but according to initial estimates, it is expected to be around 100 K. As regards the system's state, therefore, the initial bias is a canonical state $\sigma \propto \exp(-\beta H)$, with level of description $\mathcal{H} = \mathrm{span}\{1, H\}$, uniform reference state, and Lagrange parameter $\beta(\sigma) \approx (100\ \mathrm{K})^{-1}$. Then one performs $N = 12{,}000$ runs of the experiment, and in each run, measures the actual occupation of the energy levels. One finds that the measured distribution of relative frequencies $\{f_i\}$ differs from the expected $\{p_i\}$. The observed mean energy, $\sum_i f_i E_i$, corresponds to a temperature 110 K rather than 100 K; and in addition, the shape of the observed distribution may or may not deviate from the canonical form. Altogether, one finds that the data differ from prior expectation by a distance, say,

$$\chi^2\big(\pi^{\sigma}_{\mathcal{F}}(\mu)\|\sigma\big) \approx 2N \sum_{i=0}^{24} f_i \ln(f_i/p_i) \approx 96. \tag{17}$$

If one uses the full experimental level of description to reconstruct the state, $\mathcal{G} = \mathcal{F}$, this deviation is significant enough, and the number of independent sample means ($\dim \pi^{\sigma}_{\mathcal{F}}(\mathcal{S}) = 24$) is sufficiently large, to satisfy both conditions in Eq. (15). To high accuracy, then, the evidence procedure yields the interpolation parameter $\alpha/(\alpha + N) \approx 1/4$. On the other hand, if one insists that the system be modeled by a canonical distribution, $\mathcal{G} = \mathcal{H}$, it is $\dim \pi^{\sigma}_{\mathcal{F}\cap\mathcal{G}}(\mathcal{S}) = 1$ and the evidence procedure must be taken with a pinch of salt. Assuming that it yields again the interpolation parameter 1/4, the posterior estimate for the system's inverse temperature is then neither the initial $(100\ \mathrm{K})^{-1}$ nor the observed $(110\ \mathrm{K})^{-1}$, but approximately the interpolation (14), which in this example gives $\beta(\rho) \approx (107.3\ \mathrm{K})^{-1}$.

IV. COMPARING LEVELS OF DESCRIPTION

Up to this point, the level of description of the theoretical model, and hence the Gibbs manifold $\pi^{\sigma}_{\mathcal{G}}(\mathcal{S})$ from which a model was to be selected, have been assumed to be given *a priori*. Now they will become themselves subject to statistical inference.

If a model is to have explanatory value, its number of parameters must be strictly smaller than the number of data points; and so its level of description must be a proper subspace of the space spanned by the measured observables, $\mathcal{G} \subset \mathcal{F}$. In fact, in the spirit of Occam's razor, one would always prefer simpler models over more complicated ones; yet when this is taken too far, the fit with the data might deteriorate. Striking the right balance between simplicity and goodness-of-fit, and determining thus the optimal level of description, constitutes a nontrivial inference task. In this section, I shall discuss how the Bayesian framework for model selection can guide the proper choice of the level of description. If presented with two rival proposals for the level of description, this framework allows one to evaluate their relative degree of plausibility in the light of experimental data and prior expectations.

If a χ^2 analysis has revealed that observed deviations from model predictions are statistically significant, one might consider moving to a more accurate model by expanding the level of description, $\mathcal{G} \to \mathcal{H}$, with $\mathcal{G} \subset \mathcal{H} \subset \mathcal{F}$. Provided the priors on the respective Gibbs manifolds $\pi^{\sigma}_{\mathcal{G}}(\mathcal{S})$ and $\pi^{\sigma}_{\mathcal{H}}(\mathcal{S})$ are both entropic around the same initial bias σ, the relative plausibility of the two levels of description is given by the Bayes rule:

$$\frac{\mathrm{prob}(\mathcal{G}|\mu,N,\mathcal{F};\alpha,\sigma)}{\mathrm{prob}(\mathcal{H}|\mu,N,\mathcal{F};\alpha,\sigma)} = \frac{\mathrm{prob}(\mathcal{G})}{\mathrm{prob}(\mathcal{H})}\,\frac{\mathrm{prob}(\mu|N,\mathcal{F};\alpha,\sigma,\mathcal{G})}{\mathrm{prob}(\mu|N,\mathcal{F};\alpha,\sigma,\mathcal{H})}. \tag{18}$$

Here, $\mathcal{F}$ denotes the experimental level of description, N the sample size, and $\mu \in \pi^{\sigma}_{\mathcal{F}}(\mathcal{S})$ the Gibbs model associated with the measured data. For simplicity, the parameter α, which characterizes the degree of confidence as to the initial bias, is assumed to be identical for both entropic priors.

The first factor on the right-hand side is the ratio of prior preferences, which, to be fair, is often taken to be of order 1. The second factor can be calculated via marginalization,

$$\begin{aligned}&\mathrm{prob}(\mu|N,\mathcal{F};\alpha,\sigma,\mathcal{G})\\ &\quad= \int_{\pi^{\sigma}_{\mathcal{G}}(\mathcal{S})} d\omega\, \mathrm{prob}(\mu|N,\omega,\mathcal{F})\,\mathrm{prob}(\omega|\alpha,\sigma,\mathcal{G}),\end{aligned} \tag{19}$$

and likewise for $\mathcal{H}$. In the Gaussian approximation, the integrand is given by Eq. (F8), which in the regime $N \gg \alpha$, with $\rho \approx \pi^{\sigma}_{\mathcal{G}}(\mu)$, yields

$$\begin{aligned}&\mathrm{prob}(\mu|N,\mathcal{F};\alpha,\sigma,\mathcal{G})\\ &\quad\approx\ \mathrm{prob}\big(\mu|N,\pi^{\sigma}_{\mathcal{G}}(\mu),\neg_{\mathcal{F},\pi^{\sigma}_{\mathcal{G}}(\mu)}\mathcal{G}\big)\,\mathrm{prob}\big(\pi^{\sigma}_{\mathcal{G}}(\mu)|\alpha,\sigma,\mathcal{G}\big).\end{aligned} \tag{20}$$

The ratio is then

$$\frac{\mathrm{prob}(\mu|N,\mathcal{F};\alpha,\sigma,\mathcal{G})}{\mathrm{prob}(\mu|N,\mathcal{F};\alpha,\sigma,\mathcal{H})} \approx \frac{N^{s/2}}{\alpha^{s/2}}\,\frac{\exp\big[-NS\big(\pi^{\sigma}_{\mathcal{H}}(\mu)\|\pi^{\sigma}_{\mathcal{G}}(\mu)\big)\big]}{\exp\big[-\alpha S\big(\pi^{\sigma}_{\mathcal{H}}(\mu)\|\pi^{\sigma}_{\mathcal{G}}(\mu)\big)\big]}, \tag{21}$$

where $s := (\dim \mathcal{H} - \dim \mathcal{G})$ denotes the number of additional model parameters introduced in the expansion $\mathcal{G} \to \mathcal{H}$. The power factors $N^{s/2}$ and $\alpha^{s/2}$ stem from the normalization factors of likelihood and prior, respectively, which do depend on the dimension of the theoretical level of description.

Bayesian model selection is thus effectively driven by two factors [24,25]: (i) a ratio of exponentials (of which, in the regime $N \gg \alpha$, the denominator can often be approximated by 1) favoring the finer-grained model with better fit; and (ii) the "Occam factor" $(N/\alpha)^{s/2}$, which favors the simpler model. It is the trade-off between the exponentials on the one hand, and the Occam factor on the other, which typically determines whether or not the level of description should be expanded. If their product is much larger than 1, one better stay with the original, coarser-grained description. In contrast, if it is much less than 1, one is advised to switch to the finer-grained description. And if it is of the order 1, the analysis remains inconclusive, and more data must be collected.

It is important to note that the trade-off decision is not based on experimental data alone. Rather, it depends also on the initial bias σ and on the parameter α. The initial bias constitutes one's starting hypothesis for the model, prior to performing any measurements, and is usually based entirely on symmetry and other theoretical considerations; whereas α quantifies the associated degree of confidence. Both σ and α reflect prior expectations of the agent who conducts the experiment, and so in principle, carry aspects which remain irreducibly subjective. In practice, however, rational agents typically agree on the symmetries of the system under study, and hence on a unique initial bias to mirror these symmetries. In fact, in many cases the initial bias is just equidistribution, $\sigma = 1/d$, being maximally noncommittal in the absence of any empirical data. The parameter α, on the other hand, can often be estimated *a posteriori* with the help of the evidence procedure.

For large N, the estimate (16) for α becomes independent of N (at fixed relative entropy), and hence the asymptotic behavior of the ratio (21) is governed entirely by its numerator. Models can then be selected according to the simple rule of thumb

$$\chi^2\big(\pi^\sigma_{\mathcal{H}}(\mu)\|\pi^\sigma_{\mathcal{G}}(\mu)\big)/s \begin{cases} \ll \ln N & : \text{keep } \mathcal{G} \\ \sim \ln N & : \text{inconclusive} \\ \gg \ln N & : \text{expand } \mathcal{G} \to \mathcal{H} \end{cases} . \quad (22)$$

Loosely speaking, whenever the gain in accuracy per additional parameter stays below the threshold $\ln N$, one better stick to the simpler model. Only when this threshold is exceeded, is one advised to move to the finer-grained model with better fit. The threshold is higher than the threshold for mere statistical significance; if $1 < \chi^2/s < \ln N$ then the potential accuracy gain is significant, yet a refinement of the model is still not recommended.

The selection rule can be readily applied to the example introduced at the end of the previous Section. In case the shape of the observed distribution deviates from the canonical form, one may wonder whether it might be better to use the full experimental level of description $\mathcal{F}$ to reconstruct the state, rather than insisting on the canonical level $\mathcal{H}$. This would add $s = 23$ parameters to the model. However, such a massive increase in the number of parameters is not warranted by the corresponding gain in accuracy. Indeed, it is $\chi^2(\pi^\sigma_{\mathcal{F}}(\mu)\|\pi^\sigma_{\mathcal{H}}(\mu))/s \leqslant \chi^2(\pi^\sigma_{\mathcal{F}}(\mu)\|\sigma)/s \approx 4.2$, which, while possibly significant, remains well below the threshold $\ln N \approx 9.4$.

While Bayesian model selection is a useful quantitative tool to guide the search for the proper level of description, it does not amount to an algorithm leading uniquely to *the* optimal level of description. The number of possible levels of description is infinite; and while the above framework may help choose between any two of them, it cannot replace the creative act of coming up with suitable candidates [25]. This creative part is beyond the realm of pure probability and must involve additional physical considerations, such as the study of symmetries, conservation laws, and time scales.

V. WOLF'S DIE

To warm up for the interesting quantum case, I shall illustrate the use of the above mathematical tools in a famous classical example, Jaynes's analysis of Wolf's die data [27]. Rudolph Wolf (1816–1893), a Swiss astronomer, had performed a number of random experiments, presumably to check the validity of statistical theory. In one of these experiments, a die was tossed $N = 20{,}000$ times in a way that precluded any systematic favoring of any face over any other. The prior expectation was a perfect die, $\sigma = 1/6$. However, the observed relative frequencies $\{f_i\}$ deviated from this expectation; their measured values are shown in Table I. A quick analysis reveals that $1/\sqrt{N} \sim 0.007$, so several deviations Δ_i are outside the typical range. More precisely, the observed $\chi^2(\mu\|\sigma) \approx 271$ lies in the exponential tail far beyond its expected value. The probability density for such a large χ^2 is extremely small, $\mathrm{PDF}(271|5) \sim 10^{-56}$, pointing to the presence of systematic defects of the die.

To reflect the presumed nature of the die's imperfections, one may consider a multitude of different levels of description. Three specific examples are (i) the simplest level of description, $\mathcal{O} = \mathrm{span}\{1\}$, corresponding to a Gibbs manifold $\pi^\sigma_{\mathcal{O}}(\mathcal{S})$ that consists of the single state σ only, where one stubbornly sticks to the initial bias; (ii) at the opposite extreme, the most accurate level of description $\mathcal{F}$, where one denies the existence of any simple explanation for the observed deviations and just introduces as many model parameters as data points; and (iii) an intermediate level of description $\mathcal{G}$, with two observables characterizing the two most likely imperfections. These are, according to Jaynes:

TABLE I. Wolf's die data: frequency distribution f and its deviation Δ from the uniform distribution.

i	f_i	Δ_i
1	0.16230	−0.00437
2	0.17245	+0.00578
3	0.14485	−0.02182
4	0.14205	−0.02462
5	0.18175	+0.01508
6	0.19660	+0.02993

(i) a shift of the center of gravity due to the mass of ivory excavated from the spots, which being proportional to the number of spots on any side, should make the "observable"

$$G_1^i := i - 3.5 \tag{23}$$

have a nonzero average. Indeed, the measured sample mean is $g_1(\mu) = 0.0983 \neq 0$; and

(ii) errors in trying to machine a perfect cube, which will tend to make one dimension (the last side cut) slightly different from the other two. It is clear from the data that Wolf's die gave a lower frequency for the faces (3,4); and therefore that the (3–4) dimension was greater than the (1–6) or (2–5) ones. The effect of this is that the "observable"

$$G_2^i := \begin{cases} 1, & i = 1,2,5,6 \\ -2, & i = 3,4 \end{cases} \tag{24}$$

has a nonzero average. Indeed, $g_2(\mu) = 0.1393 \neq 0$.
If this intermediate level of description turned out to be the most plausible, it would provide a genuine explanation, rather than merely a description, of the observed data.

The sample size is large enough to warrant the use of the rule of thumb (22). Successive refinements $\mathcal{O} \to \mathcal{G} \to \mathcal{F}$ of the level of description entail additional model parameters and accuracy gains as summarized in Table II. Only the first refinement, $\mathcal{O} \to \mathcal{G}$, delivers an accuracy gain per additional model parameter that exceeds the threshold $\ln N \approx 10$. In contrast, the second refinement $\mathcal{G} \to \mathcal{F}$, albeit delivering a further accuracy gain that is statistically significant, does not pass this threshold. In case the intermediate level of description $\mathcal{G}$ was not available, and hence there was a choice only between the "trivial" level of description $\mathcal{O}$ and the "perfect fit" level of description $\mathcal{F}$, the latter would be more plausible. Sticking stubbornly to the initial bias is the least plausible of the three options.

If presented with the choice between the three levels of description outlined above, therefore, statistical analysis reveals the intermediate, "explanatory" level of description to be the most plausible. This is not to say, however, that this is indeed the best level of description; one might come up with many more alternative proposals, which would all have to be compared with one another. Moreover, even if the above intermediate level of description were confirmed as the winner, statistical analysis would only yield its relative degree of plausibility and would never provide certainty about its being the "true" level of description. Statistical analysis cannot replace the creative act of designing levels of description which, as in the example above, are not only supported by the data but also well motivated physically.

TABLE II. Wolf's die data: number of additional model parameters and accuracy gain associated with expansions of the level of description.

Refinement	s	χ^2	χ^2/s
$\mathcal{O} \to \mathcal{G}$	2	262	131
$\mathcal{G} \to \mathcal{F}$	3	9	3
$\mathcal{O} \to \mathcal{F}$	5	271	54

VI. ISING VS. HEISENBERG

Conceptually, Bayesian model selection for quantum systems proceeds in the same way as in the classical case. The quantumness of the problem enters through the different geometry of the Gibbs manifold. As the simplest example, I shall study an exchangeable assembly of qubits; there, the geometry to consider is that of the Bloch sphere.

Initially, nothing is known about the qubits, so the prior bias σ is uniform. Then measurements on a sample of N qubits reveal an average Bloch vector of length r, with an orientation $\hat{n}$ that is tilted by a small angle $\delta\theta$ against the z axis. The Bloch-vector-length r is considerably larger than zero, so a new model, which is different from the uniform initial bias, is called for. There might be good physical reasons to expect that the system under consideration is strongly anisotropic in the z direction, suggesting a level of description $\mathcal{I}$ (for "Ising") comprising the z component of Pauli spin only, $\mathcal{I} = \text{span}\{1,\sigma_z\}$. In view of the observed tilting angle, however, there might be controversy about this, and a rival proposal ("Heisenberg") might claim that the level of description should rather include the full Pauli vector, $\mathcal{H} = \text{span}\{1,\sigma_x,\sigma_y,\sigma_z\}$.

To weigh these alternatives in light of the data, one must evaluate

$$\chi^2\big(\pi_{\mathcal{H}}^{\sigma}(\mu)\|\pi_{\mathcal{I}}^{\sigma}(\mu)\big)/s = NC_{\theta\theta}^{-1}\delta\theta^2/2, \tag{25}$$

where $C_{\theta\theta}^{-1}$ denotes the polar component of the entropy-induced metric tensor (D9) on the Bloch sphere. For instance, for $N = 20{,}000$, $r = 0.73$ and a tilting angle of 1 degree, $\delta\theta = 2\pi/360$, it is $C_{\theta\theta}^{-1} \approx 0.678$ and $\chi^2/s \approx 2.1$. Despite a significant gain in accuracy, this does not exceed the threshold $\ln N \approx 10$, and hence Bayesian model selection favors the simpler anisotropic model. At an angle of 2 degrees, χ^2/s grows to approximately 8.3; and this being close to the threshold, the analysis remains largely inconclusive. Finally, for a tilting angle of 3 degrees, the accuracy gain per additional model parameter attains a value well beyond the threshold, $\chi^2/s \approx 18.6$, tipping the balance in favor of the more detailed level of description.

Had one measured a Bloch-vector-length $r = 0.995$ instead of 0.73, the balance would have tipped in favor of the expanded level of description already at a critical angle of 1 degree, rather than 2 degrees. In general, the more the measured state approaches purity, the more sensitive the choice of level of description becomes to minor directional aberrations from the preferred axis.

VII. CONCLUSIONS

Outside the macroscopic domain, estimating a Gibbs state is a nontrivial inference task, due to two complicating factors. First, for lack of a clear hierarchy of time scales, the proper set of relevant observables might not be evident *a priori* but subject to statistical inference. Second, whenever experimental data are gathered from a small sample only, the best estimate for the Lagrange parameters is invariably affected by the experimenter's prior bias. Both issues can be tackled with the help of Bayesian techniques, suitably adapted to the problem at hand: Bayesian model selection, Bayesian interpolation, and the evidence procedure.

The results presented in this paper may have ramifications in a variety of areas. For the study of thermal properties of a microscopic system (e.g., a tiny probe taken from a larger system that is presumed to be thermal or the debris from a single collision experiment), the framework presented here allows one to decide rationally between rival theories about the proper set of relevant observables and, subsequently, to find the best estimate for the associated Lagrange parameters. For incomplete quantum-state tomography, the results imply Bayesian corrections to the conventional maximum entropy scheme; these corrections become important whenever sample sizes are small. Moreover, the approach presented here yields not just estimates for the Lagrange parameters but also the attached error bars. Finally, on a conceptual level, the framework allows for a careful consideration of the thermodynamic limit and so may shed new light on the long-standing debate about the generality of, or possible limitations of, the maximum entropy paradigm in statistical mechanics [29–33].

I see three avenues for further research. First, it will be interesting to see how the Bayesian corrections to conventional state reconstruction schemes play out in practice. A simple example has been discussed (in the context of the evidence procedure) in Ref. [5]; more examples and application to real-world experimental data will be the subject of further work. Second, while the model selection framework used here allows one to assess different proposals for the set of relevant observables, it does not provide a direct route to the optimal such set. Doing so requires an extension of Bayesian reasoning from the space of states to the space of levels of description, which will be tackled in future work. Finally, I consider it worthwhile to study in more detail the asymptotic behavior of the schemes presented here, in an effort to understand better the emergence of orthodox theory in the macroscopic limit.

ACKNOWLEDGMENTS

I thank Gernot Alber and Joe Renes for stimulating discussions.

APPENDIX A: LEVEL OF DESCRIPTION

Any real linear combination of observables is again an observable. The observables of a physical system thus constitute a real vector space. This vector space can be endowed with a positive definite scalar product, the *canonical correlation function*

$$\langle X;Y\rangle_\sigma := \int_0^1 d\nu\, \mathrm{tr}(\sigma^\nu X \sigma^{1-\nu} Y), \tag{A1}$$

with respect to a reference state σ; so it is in fact a Hilbert space. Within this real Hilbert space of observables, the (typically small) set of observables $\{G_a\}$, which are deemed relevant for the problem at hand, together with the unit operator, span a proper subspace:

$$\mathcal{G} := \mathrm{span}\{1, G_a\}. \tag{A2}$$

This subspace is termed the *level of description* [9].

Levels of description might be related by coarse graining or complementation. A level of description $\mathcal{G}$ is "coarser" than another level of description $\mathcal{F}$, $\mathcal{G} \subset \mathcal{F}$, if the former is a subspace of the latter. The coarse graining relation $\subset$ induces a partial ordering of the levels of description, with unique minimal element $\mathcal{O} := \mathrm{span}\{1\}$ and maximal element $\mathcal{A}$, the total Hilbert space of observables. The level of description $\mathcal{G}$ is "complementary" to $\mathcal{F}$, $\mathcal{G} = \neg_{\mathcal{A},\sigma}\mathcal{F}$, if observables from both levels of description together span the entire space of observables, and if in the reference state σ the two levels are uncorrelated,

$$\begin{aligned} \mathcal{G} = \neg_{\mathcal{A},\sigma}\mathcal{F} :\Leftrightarrow \mathrm{span}\{1, G_a, F_b\} = \mathcal{A}, \\ \langle \delta X; \delta Y\rangle_\sigma = 0 \;\forall X \in \mathcal{G}, Y \in \mathcal{F}, \end{aligned} \tag{A3}$$

with $\delta X := X - \langle X\rangle_\sigma$. Complementation reverses the direction of coarse graining,

$$\mathcal{G} \subset \mathcal{F} \Rightarrow \neg_{\mathcal{A},\sigma}\mathcal{F} \subset \neg_{\mathcal{A},\sigma}\mathcal{G}; \tag{A4}$$

and when applied twice, it returns the original level of description,

$$\neg_{\mathcal{A},\sigma}\neg_{\mathcal{A},\sigma}\mathcal{G} = \mathcal{G}. \tag{A5}$$

The properties of coarse graining and complementation are reminiscient of those of logical implication and negation. In this sense, one may say that the space of observables gives rise to a minimal logical structure.

The intersection and closed hull of two levels of description are denoted by $\mathcal{G} \cap \mathcal{F}$ and $\mathcal{G} \cup \mathcal{F}$, respectively. In line with the logical structure mentioned above, the operations $\cap, \cup$ share some properties with the Boolean "and" and "or" operations such as commutativity, associativity, and reversal under complementation; but the analogy is not perfect, since in contrast to classical Boolean logic, they violate distributivity. If the levels of description pertain to two different physical systems A and B, then it is $\mathcal{G}^A \cap \mathcal{F}^B = \mathcal{O}^{AB}$, and

$$\mathcal{G}^A \cup \mathcal{F}^B = \mathrm{span}\left\{1^A \otimes 1^B, G_a^A \otimes 1^B, 1^A \otimes F_b^B\right\}. \tag{A6}$$

Another way to concatenate the two constituent levels of description is by means of the tensor product:

$$\mathcal{G}^A \otimes \mathcal{F}^B := \mathrm{span}\left\{1^A \otimes 1^B, G_a^A \otimes 1^B, 1^A \otimes F_b^B, G_a^A \otimes F_b^B\right\}. \tag{A7}$$

APPENDIX B: RELEVANT PART OF A STATE

For an arbitrary state ρ, its *relevant part* with respect to a level of description $\mathcal{G}$ and reference state σ is the unique state $\pi_\mathcal{G}^\sigma(\rho)$ which for all observables in the level of description yields the same expectation values as ρ, yet within this constraint is as close as possible to the reference state. The distance to the reference state is measured in terms of the *relative entropy* [34–37]:

$$S(\rho\|\sigma) := \mathrm{tr}(\rho \ln \rho - \rho \ln \sigma). \tag{B1}$$

The relevant part is thus determined by the minimization

$$S\left(\pi_\mathcal{G}^\sigma(\rho)\|\sigma\right) = \min_{g(\rho')=g(\rho)} S(\rho'\|\sigma), \tag{B2}$$

where I employed $g(\rho')$ as a shorthand notation for the set $\{\langle G_a\rangle_{\rho'}\}$.

That the relative entropy is the appropriate measure for the distance between two states, follows from the *quantum Stein*

lemma [38–40]. According to this lemma, given a finite sample of size N taken from an i.i.d. source of states σ, the probability that tomography on this sample will erroneously reveal some different state ρ,

$$\begin{aligned}&\mathrm{prob}_{1-\epsilon}(\rho|N,\sigma)\\&\quad:=\inf_{\Gamma}\{\mathrm{prob}(\Gamma|\sigma^{\otimes N})\big|\,\mathrm{prob}(\Gamma|\rho^{\otimes N})\geqslant 1-\epsilon\},\end{aligned}\tag{B3}$$

decreases asymptotically as

$$\mathrm{prob}_{1-\epsilon}(\rho|N,\sigma)\sim\exp[-NS(\rho\|\sigma)],\tag{B4}$$

regardless of the specific value of the error parameter ϵ ($0<\epsilon<1$). The Γ featuring in the above definition are propositions (projection operators) about the sample, which asymptotically, i.e., to within an error probability ϵ that does not depend on sample size, are compatible with the sample being in the state $\rho^{\otimes N}$. Taking the infimum over Γ picks the proposition that is most confined, and hence discriminates best between σ and ρ. The coefficient in the exponent is the relative entropy between the two states, which is thus recognized as the proper measure of their distinguishability [37].

The relevant part of a state has the generalized *Gibbs form* [41]

$$\pi^{\sigma}_{\mathcal{G}}(\rho)=Z(\lambda)^{-1}\exp\left[(\ln\sigma-\langle\ln\sigma\rangle_{\sigma})-\sum_a\lambda^a G_a\right],\tag{B5}$$

with the *partition function*

$$Z(\lambda):=\mathrm{tr}\left\{\exp\left[(\ln\sigma-\langle\ln\sigma\rangle_{\sigma})-\sum_a\lambda^a G_a\right]\right\}\tag{B6}$$

ensuring state normalization and the *Lagrange parameters* $\{\lambda^a\}$ adjusted such that $g[\pi^{\sigma}_{\mathcal{G}}(\rho)]=g(\rho)$. Among all states of the above generalized Gibbs form, the relevant part of ρ is that which comes closest to ρ in terms of relative entropy,

$$S\big(\rho\|\pi^{\sigma}_{\mathcal{G}}(\rho)\big)=\min_{\rho'}S\big(\rho\|\pi^{\sigma}_{\mathcal{G}}(\rho')\big).\tag{B7}$$

The reference state is often, but not always, the uniform distribution; if so, the Gibbs state acquires the more familiar form

$$\pi_{\mathcal{G}}(\rho)=Z(\lambda)^{-1}\exp\left[-\sum_a\lambda^a G_a\right]\tag{B8}$$

(with superscript σ omitted), which maximizes the von Neumann entropy $S[\rho]:=-\mathrm{tr}(\rho\ln\rho)$ under the given constraints.

Since the relevant part of a state retains information solely about selected degrees of freedom (the observables contained in the level of description), while discarding information about the rest, the map $\pi^{\sigma}_{\mathcal{G}}:\rho\to\pi^{\sigma}_{\mathcal{G}}(\rho)$ may be regarded as a *coarse graining* operation. Indeed, this operation bears some resemblance to a projection operator: it is idempotent,

$$\pi^{\tau}_{\mathcal{G}}\circ\pi^{\sigma}_{\mathcal{G}}=\pi^{\tau}_{\mathcal{G}}\tag{B9}$$

(even for $\tau\neq\sigma$); successive coarse grainings with smaller and smaller levels of description are equivalent to a one-step coarse graining with the smallest level of description,

$$\mathcal{G}\subset\mathcal{F}\;\Leftrightarrow\;\pi^{\sigma}_{\mathcal{G}}\circ\pi^{\sigma}_{\mathcal{F}}=\pi^{\sigma}_{\mathcal{G}};\tag{B10}$$

and it is covariant under unitary transformations,

$$\pi^{U\sigma U^{\dagger}}_{U\mathcal{G}U^{\dagger}}(U\rho U^{\dagger})=U\pi^{\sigma}_{\mathcal{G}}(\rho)U^{\dagger}.\tag{B11}$$

In contrast to a true projection operator, however, the coarse graining map is in general not linear. In case of a uniform reference state, the coarse graining map is the (possibly nonlinear) dual of the Kawasaki-Gunton projector, a projection superoperator acting on the space of observables [42].

APPENDIX C: GIBBS MANIFOLD

Let $\mathcal{S}$ denote the set of all normalized mixed states of a given physical system. This set constitutes a differentiable manifold of dimension (d^2-1), where d is the Hilbert space dimension. In this manifold, states of the generalized Gibbs form (B5) constitute a submanifold $\pi^{\sigma}_{\mathcal{G}}(\mathcal{S})$; I call it the *Gibbs manifold* associated with level of description $\mathcal{G}$ and reference state σ. A point on this Gibbs manifold, and hence a specific state of generalized Gibbs form, is a *Gibbs model*. The Gibbs manifold has dimension

$$\dim\pi^{\sigma}_{\mathcal{G}}(\mathcal{S})=\dim\mathcal{G}-1,\tag{C1}$$

which equals the number of relevant observables $\{G_a\}$ as long as these are linearly independent. Coordinates on the manifold may be the Lagrange parameters $\{\lambda^a\}$ or the expectation values $\{g_a\}$, or any set of $(\dim\mathcal{G}-1)$ independent functions thereof. Lagrange parameter coordinates are related to expectation value coordinates via

$$g_a=-\partial(\ln Z)/\partial\lambda^a.\tag{C2}$$

Upon infinitesimal variation of the Lagrange parameters, the expectation value of an arbitrary observable A changes by

$$d\langle A\rangle=-\sum_a\langle\delta G_a;A\rangle d\lambda^a,\tag{C3}$$

with $\delta G_a:=G_a-g_a$, and the expectation values and the canonical correlation function evaluated in the model with coordinates $\{\lambda^a\}$. A special case is the variation of the relevant expecation values,

$$dg_b=-\sum_a d\lambda^a C_{ab},\tag{C4}$$

where the coefficients

$$C_{ab}:=\langle\delta G_a;\delta G_b\rangle=\frac{\partial^2}{\partial\lambda^a\partial\lambda^b}\ln Z,\tag{C5}$$

form the *correlation matrix*. As the canonical correlation function has all properties of a positive definite scalar product in the space of observables, the correlation matrix is symmetric and positive.

The Gibbs manifold is endowed with a natural Riemannian metric and volume element, induced by the relative entropy [43,44]. As one would expect from a proper distance measure, the relative entropy between two states is always positive,

$$S(\rho\|\rho')\geqslant 0,\tag{C6}$$

with equality if and only if $\rho = \rho'$; and even though it is in general not symmetric, $S(\rho\|\rho') \neq S(\rho'\|\rho)$, it is approximately so for nearby states:

$$S(\rho\|\rho + \delta\rho) \sim O((\delta\rho)^2). \tag{C7}$$

The relative entropy between two points (ω,ω') on the same Gibbs manifold is

$$S(\omega\|\omega') = \sum_a (\lambda'^a - \lambda^a) g_a + (\ln Z' - \ln Z), \tag{C8}$$

which for nearby states is approximately quadratic in the coordinate differentials,

$$\begin{aligned} S(\omega\|\omega + \delta\omega) &\approx (1/2) \sum_{ab} C_{ab} \delta\lambda^a \delta\lambda^b \\ &\approx (1/2) \sum_{ab} (C^{-1})^{ab} \delta g_a \delta g_b. \end{aligned} \tag{C9}$$

The correlation matrix C or its inverse C^{-1}, respectively, may thus be regarded as a metric tensor on the Gibbs manifold. Associated with this metric is the volume element

$$\int_{\pi^\sigma_{\mathcal{G}}(\mathcal{S})} d\omega = \int \prod_a d\lambda^a \sqrt{\det C} = \int \prod_a dg_a \sqrt{\det C^{-1}}. \tag{C10}$$

Given some coarser level of description $\mathcal{H}$, $\mathcal{H} \subset \mathcal{G}$, the Gibbs manifold $\pi^\sigma_{\mathcal{G}}(\mathcal{S})$ can be viewed as a fiber bundle, with the reduced Gibbs manifold $\pi^\sigma_{\mathcal{H}}(\mathcal{S})$ as its base and the coarse graining map $\pi^\sigma_{\mathcal{H}}$ as the bundle projection:

$$\pi^\sigma_{\mathcal{H}} : \pi^\sigma_{\mathcal{G}}(\mathcal{S}) \ni \omega \to \zeta \in \pi^\sigma_{\mathcal{H}}(\mathcal{S}). \tag{C11}$$

The fiber over ζ is the submanifold of Gibbs models satisfying the constraint $h(\omega) = h(\zeta)$,

$$\pi^\sigma_{\mathcal{G}} \circ (\pi^\sigma_{\mathcal{H}})^{-1}(\zeta) = \pi^\sigma_{\mathcal{G}}(\mathcal{S})|_{h(\zeta)}. \tag{C12}$$

It is then possible to factorize volume elements of the original Gibbs manifold into those of its fiber and base,

$$\int_{\pi^\sigma_{\mathcal{G}}(\mathcal{S})} d\omega = \int_{\pi^\sigma_{\mathcal{H}}(\mathcal{S})} d\zeta \int_{\pi^\sigma_{\mathcal{G}}(\mathcal{S})|_{h(\zeta)}} d\omega. \tag{C13}$$

APPENDIX D: GEOMETRY OF THE BLOCH SPHERE

Any normalized mixed state of a single qubit can be written as

$$\rho = (1/2)(1 + \langle\vec{\sigma}\rangle_\rho \cdot \vec{\sigma}), \tag{D1}$$

with $\vec{\sigma}$ defined as the vector of Pauli matrices, $\vec{\sigma} := (\sigma_x,\sigma_y,\sigma_z)$. The expectation value of the latter is the *Bloch vector*; it has the spatial direction $\hat{n}$ and length r:

$$\langle\vec{\sigma}\rangle_\rho = r\hat{n}. \tag{D2}$$

The Pauli matrices being informationally complete, the above state can always be brought into the Gibbs form

$$\rho = Z(\vec{\lambda})^{-1} \exp(-\vec{\lambda} \cdot \vec{\sigma}), \tag{D3}$$

with Lagrange parameters

$$\vec{\lambda} = -(\tanh^{-1} r)\hat{n} \tag{D4}$$

and partition function

$$Z(\vec{\lambda}) = 2\cosh|\vec{\lambda}| = 2/\sqrt{1 - r^2}. \tag{D5}$$

The relative entropy between two arbitrary qubit states is

$$\begin{aligned} S(\rho\|\rho') = {} & r \tanh^{-1} r - r \tanh^{-1} r'(\hat{n} \cdot \hat{n}') \\ & + (1/2) \ln[(1 - r^2)/(1 - r'^2)], \end{aligned} \tag{D6}$$

which for nearby states becomes approximately

$$S(\rho\|\rho') \approx (1/2)\left[C^{-1}_{rr}\delta r^2 + C^{-1}_{\theta\theta}\delta\theta^2 + C^{-1}_{\phi\phi}\delta\phi^2\right]. \tag{D7}$$

Here (r,θ,ϕ) are the spherical coordinates of the Bloch vector as defined by

$$\langle\sigma_x\rangle = r\sin\theta\cos\phi, \quad \langle\sigma_y\rangle = r\sin\theta\sin\phi, \quad \langle\sigma_z\rangle = r\cos\theta, \tag{D8}$$

with $r \in [0,1]$, $\theta \in [0,\pi]$, and $\phi \in [0,2\pi)$; and C^{-1} denotes the entropy-induced metric tensor (inverse of the correlation matrix) on the Bloch sphere. In spherical coordinates this metric tensor is diagonal,

$$C^{-1} = \mathrm{diag}(1/(1 - r^2),\ r\tanh^{-1} r,\ r\tanh^{-1} r \sin^2\theta), \tag{D9}$$

but differs from the ordinary metric $\mathrm{diag}(1,r^2,r^2\sin^2\theta)$. Consequently, the associated volume element

$$\sqrt{\det C^{-1}} = r\tanh^{-1} r \sin\theta/\sqrt{1 - r^2}, \tag{D10}$$

too, differs from its ordinary counterpart, especially near the surface of the Bloch sphere:

$$\frac{\sqrt{\det C^{-1}}}{r^2\sin\theta} = \frac{\tanh^{-1} r}{r\sqrt{1 - r^2}} \begin{cases} \approx 1 & : r \ll 1 \\ \to \infty & : r \to 1 \end{cases}. \tag{D11}$$

Distinguishable quantum states are thus not spread uniformly throughout the Bloch sphere, as one might expect classically, but are concentrated on or near its surface.

APPENDIX E: ENTROPIC DISTRIBUTION

The coordinates of a Gibbs model $\omega \in \pi^\sigma_{\mathcal{G}}(\mathcal{S})$, and hence its location on the Gibbs manifold, might not be precisely known but have some probability distribution. Such a distribution over the Gibbs manifold is *entropic*, $\omega \sim \mathrm{E}(\alpha,\sigma,\mathcal{G})$, if it has the form

$$\mathrm{prob}(\omega|\alpha,\sigma,\mathcal{G}) \propto \begin{cases} \exp[-\alpha S(\omega\|\sigma)] & : \omega \in \pi^\sigma_{\mathcal{G}}(\mathcal{S}) \\ 0 & : \text{else} \end{cases}, \tag{E1}$$

with $\alpha > 0$ and a factor of proportionality that does not depend on ω. For large α, this is approximately a Gaussian on $\pi^\sigma_{\mathcal{G}}(\mathcal{S})$ of width $1/\sqrt{\alpha}$ around the reference state σ.

The entropic distribution has a number of important properties. (i) If ω is entropically distributed, then so is $U\omega U^\dagger$ for any unitary U, with co-transformed reference state and level of description,

$$\mathrm{prob}(U\omega U^\dagger|\alpha,U\sigma U^\dagger,U\mathcal{G}U^\dagger) = \mathrm{prob}(\omega|\alpha,\sigma,\mathcal{G}). \tag{E2}$$

(ii) Coarse graining $\mathcal{G} \to \mathcal{H} \subset \mathcal{G}$ leaves relative probabilities invariant,

$$\mathrm{prob}\big(\pi^\sigma_{\mathcal{H}}(\omega)|\alpha,\sigma,\mathcal{H}\big) \propto \mathrm{prob}\big(\pi^\sigma_{\mathcal{H}}(\omega)|\alpha,\sigma,\mathcal{G}\big), \tag{E3}$$

with a factor of proportionality that is independent of ω. (iii) If the reference state is uncorrelated, then the entropic distribution does not introduce any bias toward spurious correlations:

$$\begin{aligned}&\mathrm{prob}(\omega^{AB}|\alpha,\sigma^A\otimes\sigma^B,\mathcal{G}^A\otimes\mathcal{F}^B)\\&\quad\leqslant \mathrm{prob}(\omega^A\otimes\omega^B|\alpha,\sigma^A\otimes\sigma^B,\mathcal{G}^A\otimes\mathcal{F}^B),\end{aligned}\tag{E4}$$

where ω^A,ω^B are the respective reductions of ω^{AB}. And (iv) for uncorrelated states, the probability factorizes:

$$\begin{aligned}&\mathrm{prob}(\omega^A\otimes\omega^B|\alpha,\sigma^A\otimes\sigma^B,\mathcal{G}^A\otimes\mathcal{F}^B)\\&\quad\propto \mathrm{prob}(\omega^A|\alpha,\sigma^A,\mathcal{G}^A)\,\mathrm{prob}(\omega^B|\alpha,\sigma^B,\mathcal{F}^B).\end{aligned}\tag{E5}$$

If the reference state σ is uniform, then the entropic distribution is in fact the *only* probability distribution with the above four properties, both in the classical [45] and in the quantum case [46]. In contrast, for arbitrary σ the uniqueness of the entropic distribution has been shown in the classical case only [47]; but I conjecture that this result, too, should carry over to the quantum case.

Finally, the product of two entropic distributions is again entropic,

$$\begin{aligned}&\mathrm{prob}(\omega|N,\mu,\mathcal{G})\,\mathrm{prob}(\omega|\alpha,\sigma,\mathcal{G})\\&\quad\propto \mathrm{prob}(\omega|\alpha+N,\rho(\mu,\sigma;t),\mathcal{G}),\end{aligned}\tag{E6}$$

provided they are defined over the same Gibbs manifold, $\pi^\mu_{\mathcal{G}}(\mathcal{S})=\pi^\sigma_{\mathcal{G}}(\mathcal{S})$. Here $\rho(\mu,\sigma;t)\in\pi^\sigma_{\mathcal{G}}(\mathcal{S})$ denotes the interpolated reference state,

$$\rho(\mu,\sigma;t)\propto\exp[(1-t)\ln\mu+t\ln\sigma],\tag{E7}$$

with $t:=\alpha/(\alpha+N)$.

APPENDIX F: GAUSSIAN APPROXIMATION AND THERMODYNAMIC LIMIT

In many practical applications, the states under consideration are all concentrated inside some small region of state space. It is then often justified to make the *Gaussian approximation*, in which relative entropies are quadratic in the coordinate differentials. In this approximation, the relative entropy is symmetric:

$$S(\rho\|\omega)\approx S(\omega\|\rho).\tag{F1}$$

Furthermore, the normalization factor of the entropic distribution becomes

$$\int_{\pi^\sigma_{\mathcal{G}}(\mathcal{S})}d\omega\,\exp[-\alpha S(\omega\|\sigma)]\approx(2\pi/\alpha)^{\dim\pi^\sigma_{\mathcal{G}}(\mathcal{S})/2},\tag{F2}$$

and is thus independent not only of ω but also of σ. As a consequence, the entropic distribution becomes invariant under exchange of ω and σ:

$$\mathrm{prob}(\omega|\alpha,\sigma,\mathcal{G})\approx\mathrm{prob}(\sigma|\alpha,\omega,\mathcal{G}).\tag{F3}$$

When the Gibbs manifold $\pi^\sigma_{\mathcal{G}}(\mathcal{S})$ is considered as a fiber bundle (see Appendix C), with some reduced manifold $\pi^\sigma_{\mathcal{H}}(\mathcal{S})$, $\mathcal{H}\subset\mathcal{G}$, as its base, then in the Gaussian approximation the fiber over $\zeta\in\pi^\sigma_{\mathcal{H}}(\mathcal{S})$ is given by

$$\pi^\sigma_{\mathcal{G}}\circ\left(\pi^\sigma_{\mathcal{H}}\right)^{-1}(\zeta)\approx\pi^{\zeta}_{\neg_{\mathcal{G},\zeta}\mathcal{H}}(\mathcal{S}).\tag{F4}$$

Moreover, for any $\omega\in\pi^\sigma_{\mathcal{G}}(\mathcal{S})$, the four states

$$\begin{array}{ccc}\pi^\sigma_{\mathcal{H}}(\omega) & \text{———} & \omega\\ | & & |\\ \sigma & \text{———} & \pi^\sigma_{\neg_{\mathcal{G},\sigma}\mathcal{H}}(\omega)\end{array}\tag{F5}$$

form a rectangle as shown, with opposite sides having approximately equal length as measured by the relative entropy. Together with the (exact) *law of Pythagoras* [48] for the relative entropy,

$$S(\omega\|\sigma)=S\big(\omega\|\pi^\sigma_{\mathcal{H}}(\omega)\big)+S\big(\pi^\sigma_{\mathcal{H}}(\omega)\|\sigma\big),\tag{F6}$$

these properties imply that the entropic distribution factorizes into separate distributions over fiber and base:

$$\begin{aligned}\mathrm{prob}(\omega|\alpha,\sigma,\mathcal{G})\approx{}&\mathrm{prob}\big(\pi^\sigma_{\neg_{\mathcal{G},\sigma}\mathcal{H}}(\omega)|\alpha,\sigma,\neg_{\mathcal{G},\sigma}\mathcal{H}\big)\\&\times\mathrm{prob}\big(\pi^\sigma_{\mathcal{H}}(\omega)|\alpha,\sigma,\mathcal{H}\big).\end{aligned}\tag{F7}$$

If a model μ is entropically distributed, $\mu\sim\mathrm{E}(N,\omega,\mathcal{F})$, around a reference state ω which is itself entropically distributed, $\omega\sim\mathrm{E}(\alpha,\sigma,\mathcal{G})$, then in the Gaussian approximation and for $\mathcal{F}\supset\mathcal{G}$, the product of these entropic distributions is

$$\begin{aligned}&\mathrm{prob}(\mu|N,\omega,\mathcal{F})\,\mathrm{prob}(\omega|\alpha,\sigma,\mathcal{G})\\&\quad\propto\mathrm{prob}(\mu|N,\rho,\mathcal{F})\mathrm{prob}(\omega|\alpha+N,\rho,\mathcal{G})\,\mathrm{prob}(\rho|\alpha,\sigma,\mathcal{F}\cap\mathcal{G}),\end{aligned}\tag{F8}$$

with a factor of proportionality that is independent of both μ and ω, and with $\rho\in\pi^\sigma_{\mathcal{F}\cap\mathcal{G}}(\mathcal{S})$ short for the interpolated state $\rho(\pi^\sigma_{\mathcal{F}\cap\mathcal{G}}(\mu),\sigma;t)$ as defined in Eq. (E7). In the case $\mathcal{F}\subset\mathcal{G}$, the product is approximately

$$\begin{aligned}&\mathrm{prob}(\mu|N,\omega,\mathcal{F})\,\mathrm{prob}(\omega|\alpha,\sigma,\mathcal{G})\\&\quad\propto\mathrm{prob}\big(\pi^\sigma_{\mathcal{F}}(\mu)|N,\rho,\mathcal{F}\big)\,\mathrm{prob}\big(\pi^\sigma_{\mathcal{F}}(\omega)|\alpha+N,\rho,\mathcal{F}\big)\\&\quad\times\mathrm{prob}\big(\pi^\rho_{\neg_{\mathcal{G},\rho}\mathcal{F}}(\omega)|\alpha,\rho,\neg_{\mathcal{G},\rho}\mathcal{F}\big)\,\mathrm{prob}(\rho|\alpha,\sigma,\mathcal{F}\cap\mathcal{G}).\end{aligned}\tag{F9}$$

In the *thermodynamic limit*, the parameter N (but not the parameter α) approaches infinity, $N\to\infty$. The interpolation parameter t then approaches zero, $t\to0$, and as a consequence, $\rho\to\pi^\sigma_{\mathcal{F}\cap\mathcal{G}}(\mu)$. In this limit, the above product approaches asymptotically, for $\mathcal{F}\supset\mathcal{G}$,

$$\begin{aligned}&\mathrm{prob}(\mu|N,\omega,\mathcal{F})\,\mathrm{prob}(\omega|\alpha,\sigma,\mathcal{G})\\&\quad\sim\delta_{\neg_{\mathcal{F},\sigma}\mathcal{G}}(\mu-\rho)\delta_{\mathcal{G}}(\omega-\rho)\,\mathrm{prob}(\rho|\alpha,\sigma,\mathcal{F}\cap\mathcal{G}),\end{aligned}\tag{F10}$$

where $\delta_{\neg_{\mathcal{F},\sigma}\mathcal{G}}$ and $\delta_{\mathcal{G}}$ are multidimensional delta functions on the Gibbs manifolds $\pi^\rho_{\neg_{\mathcal{F},\sigma}\mathcal{G}}(\mathcal{S})$ and $\pi^\rho_{\mathcal{G}}(\mathcal{S})$, respectively; and for $\mathcal{F}\subset\mathcal{G}$,

$$\begin{aligned}&\mathrm{prob}(\mu|N,\omega,\mathcal{F})\,\mathrm{prob}(\omega|\alpha,\sigma,\mathcal{G})\\&\quad\sim\delta_{\mathcal{F}}\big(\pi^\sigma_{\mathcal{F}}(\omega)-\rho\big)\,\mathrm{prob}\big(\pi^\rho_{\neg_{\mathcal{G},\rho}\mathcal{F}}(\omega)|\alpha,\rho,\neg_{\mathcal{G},\rho}\mathcal{F}\big)\\&\quad\times\mathrm{prob}(\rho|\alpha,\sigma,\mathcal{F}\cap\mathcal{G}).\end{aligned}\tag{F11}$$

APPENDIX G: THERMODYNAMIC RELATIONS

Much of conventional thermodynamics amounts to exploring the differential geometry of the Gibbs manifold, and in particular, transforming its coordinates to that set of variables which is best suited for the problem at hand. Contained in this

set is usually the *thermodynamic entropy*, which for a Gibbs model $\omega \in \pi^{\sigma}_{\mathcal{G}}(\mathcal{S})$ is defined as

$$S := -S(\omega\|\sigma) - \langle \ln \sigma \rangle_{\sigma}. \tag{G1}$$

(I employ natural units with $k_B = 1$.) If the reference state is uniform, this reduces to the more familiar expression $S = -\langle \ln \omega \rangle_{\omega}$. The thermodynamic entropy is related to the Lagrange parameters and expectation values via

$$S = \ln Z + \sum_a \lambda^a g_a, \tag{G2}$$

with differential

$$dS = \sum_a \lambda^a dg_a. \tag{G3}$$

In addition to the Lagrange parameters $\{\lambda^a\}$, the partition function Z might depend on further parameters $\{\xi^b\}$; these might be, say, parameters that determine the choice of Hilbert space (e.g., a fixed spatial volume or particle number) or control parameters on which the operators $\{G_a\}$ depend (e.g., an external field). Associated with these parameters $\{\xi^b\}$ are then further variables

$$\kappa_b := \partial(\ln Z)/\partial\xi^b. \tag{G4}$$

Taking these into account, the entropy differential reads

$$dS = \sum_a \lambda^a dg_a + \sum_b \kappa_b d\xi^b. \tag{G5}$$

Being the only assured constant of the motion, the *internal energy* U features always as a variable in conventional thermodynamics, with the *inverse temperature* β as its conjugate. Depending on whether the energy is given on average or as a sharp constraint, the pair (U,β) may be of the type (g,λ) ("canonical ensemble") or (ξ,κ) ("microcanonical ensemble"), respectively. In both cases, defining the *temperature*

$$T := 1/\beta \tag{G6}$$

and new variables

$$w^a := -\lambda^a/\beta, \quad X_b := -\kappa_b/\beta, \tag{G7}$$

one obtains from Eq. (G5) the *first law of thermodynamics*,

$$dU = \underbrace{T\,dS}_{\delta Q} + \underbrace{\sum_a w^a dg_a + \sum_b X_b d\xi^b}_{\delta W}. \tag{G8}$$

Here, the differential δQ denotes *heat*, and δW denotes *work*. Some common choices for the pairs (g,w) and (ξ,X) are listed in Table III.

TABLE III. Common examples of thermodynamic variables. In cases where two alternative pairings are given, the proper choice depends on the specific situation: For instance, $(\vec{M},\vec{B})$ should be used if the magnetization is an (approximate) constant of the motion and given on average, whereas $(\vec{B},-\vec{M})$ should be employed if the magnetic field is an external control parameter for the Hamiltonian.

(g,w)	(ξ,X)	Names
$(\vec{p},\vec{v})$		Momentum, velocity
$(\vec{L},\vec{\omega})$		Angular momentum, angular velocity
(N,μ)	(N,μ)	Particle number, chemical potential
$(\vec{M},\vec{B})$	$(\vec{B},-\vec{M})$	Magnetic field, magnetization
$(\vec{P},\vec{E})$	$(\vec{E},-\vec{P})$	Electric field, electric polarization
	$(V,-p)$	Volume, pressure

The internal energy U is an example of a *thermodynamic potential*. Other examples are the *free energy*

$$F := U - TS \tag{G9}$$

and the *grand potential*

$$A := U - TS - \sum_a w^a g_a, \tag{G10}$$

with respective differentials

$$dF = -SdT + \sum_a w^a dg_a + \sum_b X_b d\xi^b \tag{G11}$$

and

$$dA = -SdT - \sum_a g_a dw^a + \sum_b X_b d\xi^b. \tag{G12}$$

The latter implies, e.g., $S = -(\partial A/\partial T)_{w,\xi}$, where the subscripts denote the variables to be kept fixed when taking the partial derivative.

The grand potential is linked directly to the partition function,

$$A(T,w^a,\xi^b) = -T \ln Z(T,w^a,\xi^b), \tag{G13}$$

which in turn can be calculated microscopically. An important part of statistical mechanics is determining the partition function and hence the grand potential and subsequently relating the latter, via suitable coordinate transformations on the Gibbs manifold, to the other thermodynamic variables of interest.

[1] C. M. Caves, C. A. Fuchs, and R. Schack, Phys. Rev. A **65**, 022305 (2002).
[2] M. Cramer, M. B. Plenio, S. T. Flammia, R. Somma, D. Gross, S. D. Bartlett, O. Landon-Cardinal, D. Poulin, and Y.-K. Liu, Nature Comm. **1**, 149 (2010).
[3] K. M. R. Audenaert and S. Scheel, New J. Phys. **11**, 023028 (2009).
[4] R. Blume-Kohout, Phys. Rev. Lett. **105**, 200504 (2010).
[5] J. Rau, Phys. Rev. A **82**, 012104 (2010).
[6] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).

[7] E. T. Jaynes, Phys. Rev. **108**, 171 (1957).
[8] R. Balian and N. L. Balazs, Ann. Phys. **179**, 97 (1987).
[9] J. Rau and B. Müller, Phys. Rep. **272**, 1 (1996).
[10] A. E. Allahverdyan, R. Balian, and T. M. Nieuwenhuizen, J. Mod. Opt. **51**, 2703 (2004).
[11] A. E. Allahverdyan, R. Balian, and T. M. Nieuwenhuizen, Europhys. Lett. **67**, 565 (2004).
[12] D. Janzing, J. Stat. Phys. **122**, 531 (2006).
[13] O. C. O. Dahlsten, R. Renner, E. Rieper, and V. Vedral, New J. Phys. **13**, 053015 (2011).
[14] F. Becattini, P. Castorina, J. Manninen, and H. Satz, Eur. Phys. J. C **56**, 493 (2008).
[15] M. Rigol, V. Dunjko, V. Yurovsky, and M. Olshanii, Phys. Rev. Lett. **98**, 050405 (2007).
[16] V. Buzek, G. Drobny, G. Adam, R. Derka, and P. L. Knight, J. Mod. Opt. **44**, 2607 (1997).
[17] V. Buzek, R. Derka, G. Adam, and P. L. Knight, Ann. Phys. **266**, 454 (1998).
[18] V. Buzek, G. Drobny, R. Derka, G. Adam, and H. Wiedemann, Chaos Solitions Fractals **10**, 981 (1999).
[19] V. Buzek and G. Drobny, J. Mod. Opt. **47**, 2823 (2000).
[20] V. Buzek, Lect. Notes Phys. **649**, 189 (2004).
[21] S. Olivares and M. G. A. Paris, Phys. Rev. A **76**, 042120 (2007).
[22] G. Drobny and V. Buzek, Phys. Rev. A **65**, 053410 (2002).
[23] S. Deléglise, I. Dotsenko, C. Sayrin, J. Bernu, M. Brune, J.-M. Raimond, and S. Haroche, Nature **455**, 510 (2008).
[24] D. S. Sivia, W. I. F. David, K. S. Knight, and S. F. Gull, Physica D **66**, 234 (1993).
[25] D. J. C. MacKay, Neural Comp. **4**, 415 (1992).
[26] E. T. Jaynes, *Probability Theory: The Logic of Science*, edited by G. L. Bretthorst (Cambridge University Press, 2003).
[27] E. T. Jaynes, in *Papers on Probability, Statistics and Statistical Physics*, edited by R. D. Rosenkrantz (Kluwer Academic Publishers, 1989), pp. 315–336.
[28] R. Schack, T. A. Brun, and C. M. Caves, Phys. Rev. A **64**, 014305 (2001).
[29] J. E. Shore and R. W. Johnson, IEEE Trans. Inf. Theory **26**, 26 (1980).
[30] Y. Tikochinsky, N. Z. Tishby, and R. D. Levine, Phys. Rev. Lett. **52**, 1357 (1984).
[31] J. Skilling, in *Maximum Entropy and Bayesian Methods*, edited by G. J. Erickson and C. R. Smith (Kluwer Academic, 1988), pp. 173–187.
[32] J. Uffink, Stud. Hist. Phil. Mod. Phys. **26**, 223 (1995).
[33] J. Uffink, Stud. Hist. Phil. Mod. Phys. **27**, 47 (1996).
[34] W. Ochs, Rep. Math. Phys. **9**, 135 (1976).
[35] A. Wehrl, Rev. Mod. Phys. **50**, 221 (1978).
[36] M. J. Donald, Commun. Math. Phys. **105**, 13 (1986).
[37] V. Vedral, Rev. Mod. Phys. **74**, 197 (2002).
[38] F. Hiai and D. Petz, Commun. Math. Phys. **143**, 99 (1991).
[39] T. Ogawa and H. Nagaoka, IEEE Trans. Inf. Theory **46**, 2428 (2000).
[40] F. G. S. L. Brandão and M. B. Plenio, Commun. Math. Phys. **295**, 791 (2010).
[41] M. B. Ruskai, Rep. Math. Phys. **26**, 143 (1988).
[42] K. Kawasaki and J. D. Gunton, Phys. Rev. A **8**, 2048 (1973).
[43] R. Balian, Y. Alhassid, and H. Reinhardt, Phys. Rep. **131**, 1 (1986).
[44] J. Skilling, in *Maximum Entropy and Bayesian Methods*, edited by J. Skilling (Kluwer Academic, 1989), pp. 45–52.
[45] J. Aczél, B. Forte, and C. T. Ng, Adv. Appl. Probab. **6**, 131 (1974).
[46] W. Ochs, Rep. Math. Phys. **8**, 109 (1975).
[47] W. Ochs, Rep. Math. Phys. **9**, 331 (1976).
[48] D. Petz, *Quantum Information Theory and Quantum Statistics* (Springer, 2008).