entropy

MDPI

Article

Entropic Updating of Probabilities and Density Matrices

Kevin Vanslette

Department of Physics, University at Albany (SUNY), Albany, NY 12222, USA; kvanslette@albany.edu

Received: 2 November 2017; Accepted: 2 December 2017; Published: 4 December 2017

Abstract: We find that the standard relative entropy and the Umegaki entropy are designed for the purpose of inferentially updating probabilities and density matrices, respectively. From the same set of inferentially guided design criteria, both of the previously stated entropies are derived in parallel. This formulates a quantum maximum entropy method for the purpose of inferring density matrices in the absence of complete information.

Keywords: probability theory; entropy; quantum relative entropy; quantum information; quantum mechanics; inference

1. Introduction

We *design* an inferential updating procedure for probability distributions and density matrices such that inductive inferences may be made. The inferential updating tools found in this derivation take the form of the standard and quantum relative entropy functionals, and thus we find the functionals are *designed* for the purpose of updating probability distributions and density matrices, respectively. Previously formulated design derivations which found the entropy to be a tool for inference originally required five *design criteria* (DC) [1–3], this was reduced to four in [4–6], and then down to three in [7]. We reduced the number of required DC down to two while also providing the first *design* derivation of the quantum relative entropy—*using the same design criteria and inferential principles in both instances*.

The designed quantum relative entropy takes the form of Umegaki's quantum relative entropy, and thus it has the "proper asymptotic form of the relative entropy in quantum (mechanics)" [8–10]. Recently, Wilming, etc. [11] gave an axiomatic characterization of the quantum relative entropy that "uniquely determines the quantum relative entropy". Our derivation differs from their's, again in that we *design* the quantum relative entropy for a purpose, but also that our DCs are imposed on what turns out to be the functional derivative of the quantum relative entropy rather than on the quantum relative entropy itself. The use of a quantum entropy for the purpose of inference has a large history: Jaynes [12,13] invented the notion of the quantum maximum entropy method [14], while it was perpetuated by [15–22] and many others. However, we find the quantum *relative* entropy to be the suitable entropy for updating density matrices, rather than the von Neuman entropy [23], as is suggested in [24]. I believe the present article provides the desired motivation for why the appropriate quantum relative entropy for updating density matrices, from prior to posterior, should be logarithmic in form while also providing a solution for updating non-uniform prior density matrices [24]. The relevant results of these papers may be found using the quantum relative entropy with suitably chosen prior density matrices.

It should be noted that because the relative entropies were reached by design, they may be interpreted as such, "the relative entropies are tools for updating", which means we no longer need to attach an interpretation *ex post facto*—as a measure of disorder or amount of missing information. In this sense, the relative entropies were built for the purpose of saturating their own interpretation [4,7], and, therefore, the quantum relative entropy *is the tool designed for updating density matrices*.

This article takes an inferential approach to probabilities and density matrices that is expected to be notionally consistent with the Bayesian derivations of Quantum Mechanics, such as Entropic Dynamics [7,25–27], as well as Bayesian interpretations of Quantum Mechanics, such as QBism [28]. The quantum maximum entropy method is, however, expected to be useful independent of one's interpretation of Quantum Mechanics because the entropy is designed at the level of density matrices rather than being formulated from arguments about the "inner workings" of Quantum Mechanics. This inferential approach is, at the very least, *verbally convenient* so we will continue writing in this language.

A few applications of the quantum maximum entropy method are given in an another article [29]. By maximizing the quantum relative entropy with respect to a "data constraint" and the appropriate prior density matrix, the Quantum Bayes Rule [30–34] (a positive-operator valued measure (POVM) measurement and collapse) is derived. The quantum maximum entropy method can reproduce the density matrices in [35,36] that are cited as "Quantum Bayes Rules", but the required constraints are difficult to motivate; however, it is expected that the results of this paper may be useful for further understanding Machine Learning techniques that involve the quantum relative entropy [37]. The Quantum Bayes Rule derivation in [29] is analogous to the standard Bayes Rule derivation from the relative entropy given in [38], as was suggested to be possible in [24]. This article provides the foundation for [29], and thus, the quantum maximum entropy method unifies a few topics in Quantum Information and Quantum Measurement through entropic inference.

As is described in this article and in [29], the quantum maximum entropy method is able to provide solutions even if the constraints and prior density matrix in question do not all mutually commute. This might be useful for subjects as far reaching as [39], which seeks to use Quantum Theory as a basis for building models for cognition. The immediate correspondence is that the quantum maximum entropy method might provide a solution toward addressing the empirical evidence for noncommutative cognition, which is how one's cognition changes when addressing questions in permuted order [39]. A simpler model for noncommutative cognition may also be possible by applying sequential updates via the standard maximum entropy method with their order permuted. Sequential updating does not, in general, give the same resultant probability distribution when the updating order is permuted—this is argued to be a feature of the standard maximum entropy method [40]. Similarly, sequential updating in the quantum maximum entropy method also has this feature, but it should be noted that the noncommutativity of sequential updating is different in principle than simultaneously updating with respect to expectation values of noncommuting operators.

The remainder of the paper is organized as follows: first, we will discuss some universally applicable principles of inference and motivate the design of an entropy function able to rank probability distributions. This entropy function will be designed such that it is consistent with inference by applying a few reasonable design criteria, which are guided by the aforementioned principles of inference. Using the same principles of inference and design criteria, we find the form of the quantum relative entropy suitable for inference. The solution to an example of updating 2 × 2 prior density matrices with respect to expectation values over spin matrices that do not commute with the prior via the quantum maximum entropy method is given in the Appendix B. We end with concluding remarks (I thank the reviewers for providing several useful references in this section).

2. The Design of Entropic Inference

Inference is the appropriate updating of probability distributions when new information is received. Bayes rule and Jeffrey's rule are both equipped to handle information in the form of data; however, the updating of a probability distribution due to the knowledge of an expectation value was realized by Jaynes [12–14] through the method of maximum entropy. The two methods for inference were thought to be devoid of one another until the work of [38,40], which showed Bayes Rule and Jeffrey's Rule to be consistent with the method of maximum entropy when the expectation values were in the form of data [38,40]. In the spirit of the derivation we will carry on as if the maximum entropy method were not known and show how it may be derived as an application of inference.

Given a probability distribution $\varphi(x)$ over a general set of propositions $x \in X$, it is self evident that if new information is learned, we are entitled to assign a new probability distribution $\rho(x)$ that somehow reflects this new information while also respecting our prior probability distribution $\varphi(x)$. The main question we must address is: "Given some information, to what posterior probability distribution $\rho(x)$ should we update our prior probability distribution $\varphi(x)$?", that is,

$$\varphi(x) \xrightarrow{*} \rho(x)?$$

This specifies the problem of inductive inference. Since "information" has many colloquial, yet potentially conflicting, definitions, we remove potential confusion by defining **information** operationally $(*)$ as the *rationale* that causes a probability distribution to change (inspired by and adapted from [7]). Directly from [7]:

> Our goal is to design a method that allows a systematic search for the preferred posterior distribution. The central idea, first proposed in [4], is disarmingly simple: to select the posterior, first rank all candidate distributions in increasing *order of preference* and then pick the distribution that ranks the highest. Irrespective of what it is that makes one distribution preferable over another (we will get to that soon enough), it is clear that any ranking according to preference must be transitive: if distribution ρ_1 is preferred over distribution ρ_2, and ρ_2 is preferred over ρ_3, then ρ_1 is preferred over ρ_3. Such transitive rankings are implemented by assigning to each $\rho(x)$ a real number $S[\rho]$, which is called the entropy of ρ, in such a way that if ρ_1 is preferred over ρ_2, then $S[\rho_1] > S[\rho_2]$. The selected distribution (one or possibly many, for there may be several equally preferred distributions) is that which maximizes the entropy functional.

Because we wish to update from prior distributions φ to posterior distributions ρ by ranking, the entropy functional $S[\rho, \varphi]$ is a real function of both φ and ρ. In the absence of new information, there is no available *rationale* to prefer any ρ to the original φ, and thereby the relative entropy should be designed such that the selected posterior is equal to the prior φ (in the absence of new information). The prior information encoded in $\varphi(x)$ is valuable and we should not change it unless we are informed otherwise. Due to our definition of information, and our desire for objectivity, we state the predominate guiding principle for inductive inference:

> The Principle of Minimal Updating (PMU):
> *A probability distribution should only be updated to the extent required by the new information.*

This simple statement provides the foundation for inference [7]. If the updating of probability distributions is to be done objectively, then possibilities should not be needlessly ruled out or suppressed. Being informationally stingy, that we should only update probability distributions when the information requires it, pushes inductive inference toward objectivity. Thus, using the PMU helps formulate a pragmatic (and objective) procedure for making inferences using (informationally) subjective probability distributions [41].

This method of inference is only as universal and general as its ability to apply *equally well* to *any* specific inference problem. The notion of "specificity" is the notion of statistical independence; a special case is only special in that it is separable from other special cases. The notion that systems may be "sufficiently independent" plays a central and deep-seated role in science and the idea that some things can be neglected and that not everything matters, is implemented by imposing criteria that tells us how to handle independent systems [7]. Ironically, the universally *shared* property by all specific inference problems is their ability to be *independent* of one another—they share independence. Thus, a universal inference scheme based on the PMU permits:

Properties of Independence (PI):
Subdomain Independence: When information is received about one set of propositions, it should not affect or change the state of knowledge (probability distribution) of the other propositions (else information was also received about them too);

And,

Subsystem Independence: When two systems are a priori believed to be independent and we only receive information about one, then the state of knowledge of the other system remains unchanged.

The PIs are special cases of the PMU that ultimately take the form of *design criteria* in this design derivation. The process of constraining the form of $S[\rho,\varphi]$ by imposing design criteria may be viewed as the process of *eliminative induction*, and after sufficient constraining, a single form for the entropy remains. Thus, the justification behind the surviving entropy is not that it leads to demonstrably correct inferences, but, rather, that all other candidate entropies demonstrably fail to perform as desired [7]. Rather than the *design criteria* instructing one how to update, they instruct in what instances one should *not* update. That is, rather than justifying one way to skin a cat over another, we tell you when *not* to skin it, which is operationally unique—namely you don't do it—luckily enough for the cat.

The Design Criteria and the Standard Relative Entropy

The following *design criteria* (DC), guided by the PMU, are imposed and formulate the standard relative entropy as a tool for inference. The form of this presentation is inspired by [7].

DC1: Subdomain Independence

We keep DC1 from [7] and review it below. DC1 imposes the first instance of when one should not update—the Subdomain PI. Suppose the information to be processed does *not* refer to a particular subdomain $\mathcal{D}$ of the space $\mathcal{X}$ of xs. In the absence of new information about $\mathcal{D}$, the PMU insists we do not change our minds about probabilities that are conditional on $\mathcal{D}$. Thus, we design the inference method so that $\varphi(x|\mathcal{D})$, the prior probability of x conditional on $x \in \mathcal{D}$, is not updated and therefore the selected conditional posterior is

$$P(x|\mathcal{D}) = \varphi(x|\mathcal{D}). \tag{1}$$

(The notation will be as follows: we denote priors by φ, candidate posteriors by lower case ρ, and the selected posterior by upper case P.) We emphasize the point is not that we make the unwarranted assumption that keeping $\varphi(x|\mathcal{D})$ unchanged is guaranteed to lead to correct inferences. It need not; induction is risky. The point is, rather, that, in the absence of any evidence to the contrary, there is no reason to change our minds and the prior information takes priority.

DC1 Implementation

Consider the set of microstates $x_i \in \mathcal{X}$ belonging to either of two non-overlapping domains $\mathcal{D}$ or its compliment $\mathcal{D}'$, such that $\mathcal{X} = \mathcal{D} \cup \mathcal{D}'$ and $\varnothing = \mathcal{D} \cap \mathcal{D}'$. For convenience, let $\rho(x_i) = \rho_i$. Consider the following constraints:

$$\rho(\mathcal{D}) = \sum_{i\in\mathcal{D}} \rho_i \quad \text{and} \quad \rho(\mathcal{D}') = \sum_{i\in\mathcal{D}'} \rho_i, \tag{2}$$

such that $\rho(\mathcal{D}) + \rho(\mathcal{D}') = 1$, and the following "local" expectation value constraints over $\mathcal{D}$ and $\mathcal{D}'$,

$$\langle A \rangle = \sum_{i\in\mathcal{D}} \rho_i A_i \quad \text{and} \quad \langle A' \rangle = \sum_{i\in\mathcal{D}'} \rho_i A'_i, \tag{3}$$

where $A = A(x)$ is a scalar function of x and $A_i \equiv A(x_i)$. As we are searching for the candidate distribution which maximizes S while obeying (2) and (3), we maximize the entropy $S \equiv S[\rho,\varphi]$ with respect to these expectation value constraints using the Lagrange multiplier method,

$$0 = \delta\Big(S - \lambda[\rho(\mathcal{D}) - \sum_{i\in\mathcal{D}} \rho_i] - \mu[\langle A\rangle - \sum_{i\in\mathcal{D}} \rho_i A_i]$$
$$-\lambda'[\rho(\mathcal{D}') - \sum_{i\in\mathcal{D}'} \rho_i] - \mu'[\langle A'\rangle - \sum_{i\in\mathcal{D}'} \rho_i A_i]\Big),$$

and, thus, the entropy is maximized when the following differential relationships hold:

$$\frac{\delta S}{\delta \rho_i} = \lambda + \mu A_i \quad \forall\, i \in \mathcal{D}, \tag{4}$$
$$\frac{\delta S}{\delta \rho_i} = \lambda' + \mu' A'_i \quad \forall\, i \in \mathcal{D}'. \tag{5}$$

Equations (2)–(5), are $n+4$ equations we must solve to find the four Lagrange multipliers $\{\lambda, \lambda', \mu, \mu'\}$ and the n probability values $\{\rho_i\}$ associated to the n microstates $\{x_i\}$. If the subdomain constraint DC1 is imposed in the most restrictive case, then it will hold in general. The most restrictive case requires splitting $\mathcal{X}$ into a set of $\{\mathcal{D}_i\}$ domains such that each $\mathcal{D}_i$ singularly includes one microstate x_i. This gives,

$$\frac{\delta S}{\delta \rho_i} = \lambda_i + \mu_i A_i \quad \text{in each } \mathcal{D}_i. \tag{6}$$

Because the entropy $S = S[\rho_1, \rho_2, ...; \varphi_1, \varphi_2, ...]$ is a functional over the probability of each microstate's posterior and prior distribution, its variational derivative is also a function of said probabilities in general,

$$\frac{\delta S}{\delta \rho_i} \equiv \phi_i(\rho_1, \rho_2, ...; \varphi_1, \varphi_2, ...) = \lambda_i + \mu_i A_i \quad \text{for each } (i, \mathcal{D}_i). \tag{7}$$

DC1 is imposed by constraining the form of $\phi_i(\rho_1, \rho_2, ...; \varphi_1, \varphi_2, ...) = \phi_i(\rho_i; \varphi_1, \varphi_2, ...)$ to ensure that changes in $A_i \to A_i + \delta A_i$ have no influence over the value of ρ_j in domain $\mathcal{D}_j$, through ϕ_i, for $i \neq j$. If there is no new information about propositions in $\mathcal{D}_j$, its distribution should remain equal to φ_j by the PMU. We further restrict ϕ_i such that an arbitrary variation of $\varphi_j \to \varphi_j + \delta\varphi_j$ (a change in the prior state of knowledge of the microstate j) has no effect on ρ_i for $i \neq j$ and therefore DC1 imposes $\phi_i = \phi_i(\rho_i, \varphi_i)$, as is guided by the PMU. At this point, it is easy to generalize the analysis to continuous microstates such that the indices become continuous $i \to x$, sums become integrals, and discrete probabilities become probability densities $\rho_i \to \rho(x)$.

Remark

We are designing the entropy for the purpose of ranking posterior probability distributions (for the purpose of inference); however, the highest ranked distribution is found by setting the variational derivative of $S[\rho, \varphi]$ equal to the variations of the expectation value constraints by the Lagrange multiplier method,

$$\frac{\delta S}{\delta \rho(x)} = \lambda + \sum_i \mu_i A_i(x). \tag{8}$$

Therefore, the real quantity of interest is $\frac{\delta S}{\delta \rho(x)}$ rather than the specific form of $S[\rho, \varphi]$. *All* forms of $S[\rho, \varphi]$ that give the correct form of $\frac{\delta S}{\delta \rho(x)}$ are *equally valid* for the purpose of inference. Thus, every design criteria may be made on the variational derivative of the entropy rather than the entropy itself, which we do. When maximizing the entropy, for convenience, we will let,

$$\frac{\delta S}{\delta \rho(x)} \equiv \phi_x(\rho(x), \varphi(x)), \tag{9}$$

and further use the shorthand $\phi_x(\rho, \varphi) \equiv \phi_x(\rho(x), \varphi(x))$, in all cases.

DC1′: *In the absence of new information, our new state of knowledge $\rho(x)$ is equal to the old state of knowledge $\varphi(x)$.*

This is a special case of DC1, and is implemented differently than in [7]. The PMU is in principle a statement about informational honestly—that is, one should not "jump to conclusions" in light of new information and in the absence of new information, one should not change their state of knowledge. If no new information is given, the prior probability distribution $\varphi(x)$ does not change, that is, the posterior probability distribution $\rho(x) = \varphi(x)$ is equal to the prior probability. If we maximizing the entropy without applying constraints,

$$\frac{\delta S}{\delta \rho(x)} = 0, \tag{10}$$

then DC1′ imposes the following condition:

$$\frac{\delta S}{\delta \rho(x)} = \phi_x(\rho, \varphi) = \phi_x(\varphi, \varphi) = 0, \tag{11}$$

for all x in this case. This special case of the DC1 and the PMU turns out to be incredibly constraining as we will see over the course of DC2.

Comment

If the variable x is continuous, DC1 requires that when information refers to points infinitely close but just outside the domain $\mathcal{D}$, that it will have no influence on probabilities conditional on $\mathcal{D}$ [7]. This may seem surprising as it may lead to updated probability distributions that are discontinuous. Is this a problem? No.

In certain situations (e.g., physics) we might have explicit reasons to believe that conditions of continuity or differentiability should be imposed and this information might be given to us in a variety of ways. The crucial point, however—and this is a point that we keep and will keep reiterating—is that unless such information is explicitly given, we should not assume it. If the new information leads to discontinuities, so be it.

DC2: Subsystem Independence

DC2 imposes the second instance of when one should not update—the Subsystem PI. We emphasize that DC2 *is not a consistency requirement*. The argument we deploy is *not* that both the prior *and* the new information tells us the systems are independent, in which case consistency requires that it should not matter whether the systems are treated jointly or separately. Rather, DC2 refers to a situation where the new information does not say whether the systems are independent or not, but information is given about each subsystem. The updating is being *designed* so that the independence reflected in the prior is maintained in the posterior by default via the PMU and the second clause of the PIs [7].

The point is not that when we have no evidence for correlations we draw the firm conclusion that the systems must necessarily be independent. They could indeed have turned out to be correlated and then our inferences would be wrong. Again, induction involves risk. The point is rather that if the joint prior reflects independence and the new evidence is silent on the matter of correlations, then the prior independence takes precedence. As before, in this case subdomain independence, the probability distribution should not be updated unless the information requires it [7].

DC2 Implementation

Consider a composite system, $x = (x_1, x_2) \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$. Assume that all prior evidence led us to believe the subsystems are independent. This belief is reflected in the prior distribution: if the individual system priors are $\varphi_1(x_1)$ and $\varphi_2(x_2)$, then the prior for the whole system is their product

$\varphi_1(x_1)\varphi_2(x_2)$. Further suppose that new information is acquired such that $\varphi_1(x_1)$ would by itself be updated to $P_1(x_1)$ and that $\varphi_2(x_2)$ would be itself be updated to $P_2(x_2)$. By design, the implementation of DC2 constrains the entropy functional such that, in this case, the joint product prior $\varphi_1(x_1)\varphi_2(x_2)$ updates to the selected product posterior $P_1(x_1)P_2(x_2)$ [7].

The argument below is considerably simplified if we expand the space of probabilities to include distributions that are not necessarily normalized. This does not represent any limitation because a normalization constraint may always be applied. We consider a few special cases below:

Case 1: We receive the extremely constraining information that the posterior distribution for system 1 is completely specified to be $P_1(x_1)$ while we receive no information at all about system 2. We treat the two systems jointly. Maximize the joint entropy $S[\rho(x_1,x_2),\varphi(x_1)\varphi(x_2)]$ subject to the following constraints on the $\rho(x_1,x_2)$:

$$\int dx_2\,\rho(x_1,x_2) = P_1(x_1)\,. \tag{12}$$

Notice that the probability of each $x_1 \in \mathcal{X}_1$ within $\rho(x_1,x_2)$ is being constrained to $P_1(x_1)$ in the marginal. We therefore need a one Lagrange multiplier $\lambda_1(x_1)$ for each $x_1 \in \mathcal{X}_1$ to tie each value of $\int dx_2\,\rho(x_1,x_2)$ to $P_1(x_1)$. Maximizing the entropy with respect to this constraint is,

$$\delta\left[S - \int dx_1\lambda_1(x_1)\left(\int dx_2\,\rho(x_1,x_2) - P_1(x_1)\right)\right] = 0\,, \tag{13}$$

which requires that

$$\lambda_1(x_1) = \phi_{x_1x_2}\left(\rho(x_1,x_2),\varphi_1(x_1)\varphi_2(x_2)\right)\,, \tag{14}$$

for arbitrary variations of $\rho(x_1,x_2)$. By design, DC2 is implemented by requiring $\varphi_1\varphi_2 \to P_1\varphi_2$ in this case, therefore,

$$\lambda_1(x_1) = \phi_{x_1x_2}\left(P_1(x_1)\varphi_2(x_2),\varphi_1(x_1)\varphi_2(x_2)\right)\,. \tag{15}$$

This equation must hold for all choices of x_2 and all choices of the prior $\varphi_2(x_2)$ as $\lambda_1(x_1)$ is independent of x_2. Suppose we had chosen a different prior $\varphi_2'(x_2) = \varphi_2(x_2) + \delta\varphi_2(x_2)$ that disagrees with $\varphi_2(x_2)$. For all x_2 and $\delta\varphi_2(x_2)$, the multiplier $\lambda_1(x_1)$ remains unchanged as it constrains the independent $\rho(x_1) \to P_1(x_1)$. This means that any dependence that the right-hand side might potentially have had on x_2 and on the prior $\varphi_2(x_2)$ *must cancel out*. This means that

$$\phi_{x_1x_2}\left(P_1(x_1)\varphi_2(x_2),\varphi_1(x_1)\varphi_2(x_2)\right) = f_{x_1}(P_1(x_1),\varphi_1(x_1)). \tag{16}$$

Since φ_2 is arbitrary in f, suppose further that we choose a constant prior set equal to one, $\varphi_2(x_2) = 1$, therefore

$$f_{x_1}(P_1(x_1),\varphi_1(x_1)) = \phi_{x_1x_2}\left(P_1(x_1)*1,\varphi_1(x_1)*1\right) = \phi_{x_1}\left(P_1(x_1),\varphi_1(x_1)\right) \tag{17}$$

in general. This gives

$$\lambda_1(x_1) = \phi_{x_1}\left(P_1(x_1),\varphi_1(x_1)\right). \tag{18}$$

The left-hand side does not depend on x_2, and therefore neither does the right-hand side. An argument exchanging systems 1 and 2 gives a similar result.

Case 1—Conclusion: When the system 2 is not updated the dependence on φ_2 and x_2 drops out,

$$\phi_{x_1x_2}\left(P_1(x_1)\varphi_2(x_2),\varphi_1(x_1)\varphi_2(x_2)\right) = \phi_{x_1}\left(P_1(x_1),\varphi_1(x_1)\right), \tag{19}$$

and vice-versa when system 1 is not updated,

$$\phi_{x_1x_2}\left(\varphi_1(x_1)P_2(x_2),\varphi_1(x_1)\varphi_2(x_2)\right) = \phi_{x_2}\left(P_2(x_2),\varphi_2(x_2)\right). \tag{20}$$

As we seek the general functional form of $\phi_{x_1x_2}$, and because the x_2 dependence drops out of (19) and the x_1 dependence drops out of (20) for arbitrary φ_1, φ_2 and $\varphi_{12} = \varphi_1\varphi_2$, the explicit coordinate dependence in ϕ consequently drops out of both such that,

$$\phi_{x_1x_2} \to \phi, \tag{21}$$

as $\phi = \phi(\rho(x), \varphi(x))$ must only depend on coordinates through the probability distributions themselves. (As a double check, explicit coordinate dependence was included in the following computations but inevitably dropped out due to the form the functional equations and DC1′. By the argument above, and for simplicity, we drop the explicit coordinate dependence in ϕ here.)

Case 2: Now consider a different special case in which the marginal posterior distributions for systems 1 and 2 are both completely specified to be $P_1(x_1)$ and $P_2(x_2)$, respectively. Maximize the joint entropy $S[\rho(x_1, x_2), \varphi(x_1)\varphi(x_2)]$ subject to the following constraints on the $\rho(x_1, x_2)$,

$$\int dx_2\, \rho(x_1, x_2) = P_1(x_1) \quad \text{and} \quad \int dx_1\, \rho(x_1, x_2) = P_2(x_2)\,. \tag{22}$$

Again, this is one constraint for each value of x_1 and one constraint for each value of x_2, which, therefore, require the separate multipliers $\mu_1(x_1)$ and $\mu_2(x_2)$. Maximizing S with respect to these constraints is then,

$$\begin{aligned} 0 \;=\; & \delta\left[S - \int dx_1 \mu_1(x_1) \left(\int dx_2\, \rho(x_1, x_2) - P_1(x_1)\right)\right. \\ & \left. - \int dx_2 \mu_2(x_2) \left(\int dx_1\, \rho(x_1, x_2) - P_2(x_2)\right)\right]\,, \end{aligned} \tag{23}$$

leading to

$$\mu_1(x_1) + \mu_2(x_2) = \phi\left(\rho(x_1, x_2), \varphi_1(x_1)\varphi_2(x_2)\right). \tag{24}$$

The updating is being designed so that $\varphi_1\varphi_2 \to P_1P_2$, as the independent subsystems are being updated based on expectation values which are silent about correlations. DC2 thus imposes,

$$\mu_1(x_1) + \mu_2(x_2) = \phi\left(P_1(x_1)P_2(x_2), \varphi_1(x_1)\varphi_2(x_2)\right). \tag{25}$$

Write (25) as,

$$\mu_1(x_1) = \phi\left(P_1(x_1)P_2(x_2), \varphi_1(x_1)\varphi_2(x_2)\right) - \mu_2(x_2). \tag{26}$$

The left-hand side is independent of x_2 so we can perform a trick similar to that we used before. Suppose we had chosen a different *constraint* $P_2'(x_2)$ that differs from $P_2(x_2)$ and a new prior $\varphi_2'(x_2)$ that differs from $\varphi_2(x_2)$ except at the value $\bar{x}_2$. At the value $\bar{x}_2$,the multiplier $\mu_1(x_1)$ remains unchanged for all $P_2'(x_2)$, $\varphi_2'(x_2)$, and thus x_2. This means that any dependence that the right-hand side might potentially have had on x_2 and on the choice of $P_2(x_2)$, $\varphi_2'(x_2)$ must cancel out, leaving $\mu_1(x_1)$ unchanged. That is, the Lagrange multiplier $\mu(x_2)$ "pushes out" these dependences such that

$$\phi\left(P_1(x_1)P_2(x_2), \varphi_1(x_1)\varphi_2(x_2)\right) - \mu_2(x_2) = g(P_1(x_1), \varphi_1(x_1)). \tag{27}$$

Because $g(P_1(x_1), \varphi_1(x_1))$ is independent of arbitrary variations of $P_2(x_2)$ and $\varphi_2(x_2)$ on the left hand side (LHS) above—it is satisfied equally well for all choices. The form of $g = \phi(P_1(x_1), q_1(x_1))$ is apparent if $P_2(x_2) = \varphi_2(x_2) = 1$ as $\mu_2(x_2) = 0$ similar to Case 1 as well as DC1′. Therefore, the Lagrange multiplier is

$$\mu_1(x_1) = \phi\left(P_1(x_1), \varphi_1(x_1)\right). \tag{28}$$

A similar analysis carried out for $\mu_2(x_2)$ leads to

$$\mu_2(x_2) = \phi\left(P_2(x_2), \varphi_2(x_2)\right). \tag{29}$$

Case 2—Conclusion: Substituting back into (25) gives us a functional equation for ϕ ,

$$\phi\left(P_1 P_2, \varphi_1 \varphi_2\right) = \phi\left(P_1, \varphi_1\right) + \phi\left(P_2, \varphi_2\right). \tag{30}$$

The general solution for this functional equation is derived in the Appendix A.3, and is

$$\phi(\rho, \varphi) = a_1 \ln(\rho(x)) + a_2 \ln(\varphi(x)), \tag{31}$$

where a_1, a_2 are constants. The constants are fixed by using DC1′. Letting $\rho_1(x_1) = \varphi_1(x_1) = \varphi_1$ gives $\phi(\varphi, \varphi) = 0$ by DC1′, and, therefore,

$$\phi(\varphi, \varphi) = (a_1 + a_2)\ln(\varphi) = 0, \tag{32}$$

so we are forced to conclude $a_1 = -a_2$ for arbitrary φ. Letting $a_1 \equiv A = -|A|$ such that we are really maximizing the entropy (although this is purely aesthetic) gives the general form of ϕ to be

$$\phi(\rho, \varphi) = -|A| \ln\left(\frac{\rho(x)}{\varphi(x)}\right). \tag{33}$$

As long as $A \neq 0$, the value of A is arbitrary as it always can be absorbed into the Lagrange multipliers. The general form of the entropy designed for the purpose of inference of ρ is found by integrating ϕ, and, therefore,

$$S(\rho(x), \varphi(x)) = -|A| \int dx \left(\rho(x) \ln\left(\frac{\rho(x)}{\varphi(x)}\right) - \rho(x)\right) + C[\varphi]. \tag{34}$$

The constant in ρ, $C[\varphi]$, will always drop out when varying ρ. The apparent extra term ($|A| \int \rho(x) dx$) from integration cannot be dropped while simultaneously satisfying DC1′, which requires $\rho(x) = \varphi(x)$ in the absence of constraints or when there is no change to one's information. In previous versions where the integration term ($|A| \int \rho(x) dx$) is dropped, one obtains solutions like $\rho(x) = e^{-1}\varphi(x)$ (independent of whether $\varphi(x)$ was previously normalized or not) in the absence of new information. Obviously, this factor can be taken care of by normalization, and, in this way, both forms of the entropy are equally valid; however, this form of the entropy better adheres to the PMU through DC1′. Given that we may regularly impose normalization, we may drop the extra $\int \rho(x) dx$ term and $C[\varphi]$. For convenience then, (34) becomes

$$S(\rho(x), \varphi(x)) \rightarrow S^*(\rho(x), \varphi(x)) = -|A| \int dx\, \rho(x) \ln\left(\frac{\rho(x)}{\varphi(x)}\right), \tag{35}$$

which is a special case when the normalization constraint is being applied. Given normalization is applied, the same selected posterior $\rho(x)$ maximizes both $S(\rho(x), \varphi(x))$ and $S^*(\rho(x), \varphi(x))$, and the star notation may be dropped.

Remarks

It can be seen that the relative entropy is invariant under coordinate transformations. This implies that a system of coordinates carry no information and it is the "character" of the probability distributions that are being ranked against one another rather than the specific set of propositions or microstates they describe.

The general solution to the maximum entropy procedure with respect to N linear constraints in ρ, $\langle A_i(x)\rangle$, and normalization gives a canonical-like selected posterior probability distribution,

$$\rho(x) = \varphi(x) \exp\Big(\sum_i \alpha_i A_i(x)\Big). \tag{36}$$

The positive constant $|A|$ may always be absorbed into the Lagrange multipliers so we may let it equal unity without loss of generality. DC1′ is fully realized when we maximize with respect to a constraint on $\rho(x)$ that is already held by $\varphi(x)$, such as $\langle x^2\rangle = \int x^2\rho(x)\,dx$, which happens to have the same value as $\langle x^2\rangle_\varphi = \int x^2\varphi(x)\,dx$, then its Lagrange multiplier is forcibly zero $\alpha_1 = 0$ (as can be seen in (36) using (34)), in agreement with Jaynes. This gives the expected result $\rho(x) = \varphi(x)$ as there is no new information. Our design has arrived at a refined maximum entropy method [12] as a universal probability updating procedure [38].

3. The Design of the Quantum Relative Entropy

In the last section, we assumed that the universe of discourse (the set of relevant propositions or microstates) $\mathcal{X} = \mathcal{A} \times \mathcal{B} \times ...$ was known. In quantum physics, things are a bit more ambiguous because many probability distributions, or many experiments, can be associated with a given density matrix. In this sense, it is helpful to think of density matrices as "placeholders" for probability distributions rather than a probability distributions themselves. As any probability distribution from a given density matrix, $\rho(\cdot) = \mathrm{Tr}(|\cdot\rangle\langle\cdot|\hat{\rho})$, may be ranked using the standard relative entropy, it is unclear why we would chose one universe of discourse over another. In lieu of this, such that one universe of discourse is not given preferential treatment, we consider ranking entire density matrices against one another. Probability distributions of interest may be found from the selected posterior density matrix. This moves our universe of discourse from sets of propositions $\mathcal{X} \to \mathcal{H}$ to Hilbert space(s).

When the objects of study are quantum systems, we desire an objective procedure to update from a prior density matrix $\hat{\varphi}$ to a posterior density matrix $\hat{\rho}$. We will apply the same intuition for ranking probability distributions (Section 2) and implement the PMU, PI, and design criteria to the ranking of density matrices. We therefore find the quantum relative entropy $S(\hat{\rho},\hat{\varphi})$ to be designed for the purpose of inferentially updating density matrices.

3.1. Designing the Quantum Relative Entropy

In this section, we design the quantum relative entropy using the same inferentially guided *design criteria* as were used in the standard relative entropy.

DC1: Subdomain Independence

The goal is to design a function $S(\hat{\rho},\hat{\varphi})$ that is able to rank density matrices. This insists that $S(\hat{\rho},\hat{\varphi})$ be a real scalar valued function of the posterior $\hat{\rho}$, and prior $\hat{\varphi}$ density matrices, which we will call the quantum relative entropy or simply the entropy. An arbitrary variation of the entropy with respect to $\hat{\rho}$ is,

$$\delta\, S(\hat{\rho},\hat{\varphi}) = \sum_{ij}\frac{\delta S(\hat{\rho},\hat{\varphi})}{\delta\rho_{ij}}\delta\rho_{ij} = \sum_{ij}\Big(\frac{\delta S(\hat{\rho},\hat{\varphi})}{\delta\hat{\rho}}\Big)_{ij}\delta(\hat{\rho})_{ij} = \sum_{ij}\Big(\frac{\delta S(\hat{\rho},\hat{\varphi})}{\delta\hat{\rho}^T}\Big)_{ji}\delta(\hat{\rho})_{ij} = \mathrm{Tr}\Big(\frac{\delta S(\hat{\rho},\hat{\varphi})}{\delta\hat{\rho}^T}\delta\hat{\rho}\Big), \tag{37}$$

where Tr(...) is the trace. We wish to maximize this entropy with respect to expectation value constraints, such as $\langle A\rangle = \mathrm{Tr}(\hat{A}\hat{\rho})$ on $\hat{\rho}$. Using the Lagrange multiplier method to maximize the entropy with respect to $\langle A\rangle$ and normalization, and setting the variation equal to zero,

$$\delta\Big(S(\hat{\rho},\hat{\varphi}) - \lambda[\mathrm{Tr}(\hat{\rho}) - 1] - \alpha[\mathrm{Tr}(\hat{A}\hat{\rho}) - \langle A\rangle]\Big) = 0, \tag{38}$$

where λ and α are the Lagrange multipliers for the respective constraints. Because $S(\hat{\rho}, \hat{\varphi})$ is a real number, we inevitably require δS to be real, but without imposing this directly, we find that requiring δS to be real requires $\hat{\rho}$, $\hat{A}$ to be Hermitian. At this point, it is simpler to allow for arbitrary variations of $\hat{\rho}$ such that,

$$\mathrm{Tr}\Big(\Big(\frac{\delta S(\hat{\rho}, \hat{\varphi})}{\delta \hat{\rho}^T} - \lambda \hat{1} - \alpha \hat{A}\Big)\delta \hat{\rho}\Big) = 0. \tag{39}$$

For these arbitrary variations, the variational derivative of S must satisfy,

$$\frac{\delta S(\hat{\rho}, \hat{\varphi})}{\delta \hat{\rho}^T} = \lambda \hat{1} + \alpha \hat{A} \tag{40}$$

at the maximum. As in the remark earlier, *all* forms of S that give the correct form of $\frac{\delta S(\hat{\rho}, \hat{\varphi})}{\delta \hat{\rho}^T}$ under variation are *equally valid* for the purpose of inference. For notational convenience, we let

$$\frac{\delta S(\hat{\rho}, \hat{\varphi})}{\delta \hat{\rho}^T} \equiv \phi(\hat{\rho}, \hat{\varphi}), \tag{41}$$

which is a matrix valued function of the posterior and prior density matrices. The form of $\phi(\hat{\rho}, \hat{\varphi})$ is already "local" in $\hat{\rho}$ (the variational derivative is with respect to the whole density matrix), so we don't need to constrain it further as we did in the original DC1.

DC1′: *In the absence of new information, the new state $\hat{\rho}$ is equal to the old state $\hat{\varphi}$*

Applied to the ranking of density matrices, in the absence of new information, the density matrix $\hat{\varphi}$ should not change, that is, the posterior density matrix $\hat{\rho} = \hat{\varphi}$ is equal to the prior density matrix. Maximizing the entropy without applying any constraints gives,

$$\frac{\delta S(\hat{\rho}, \hat{\varphi})}{\delta \hat{\rho}^T} = \hat{0}, \tag{42}$$

and, therefore, DC1′ imposes the following condition in this case:

$$\frac{\delta S(\hat{\rho}, \hat{\varphi})}{\delta \hat{\rho}^T} = \phi(\hat{\rho}, \hat{\varphi}) = \phi(\hat{\varphi}, \hat{\varphi}) = \hat{0}. \tag{43}$$

As in the original DC1′, if $\hat{\varphi}$ is known to obey some expectation value $\langle \hat{A} \rangle$, and then if one goes out of their way to constrain $\hat{\rho}$ to that expectation value and nothing else, it follows from the PMU that $\hat{\rho} = \hat{\varphi}$, as no information has been gained. This is not imposed directly but can be verified later.

DC2: Subsystem Independence

The discussion of DC2 is the same as the standard relative entropy DC2—it is not a consistency requirement, and the updating is *designed* so that the independence reflected in the prior is maintained in the posterior by default via the PMU when the information provided is silent about correlations.

DC2 Implementation

Consider a composite system living in the Hilbert space $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$. Assume that all prior evidence led us to believe the systems were independent. This is reflected in the prior density matrix: if the individual system priors are $\hat{\varphi}_1$ and $\hat{\varphi}_2$, then the joint prior for the whole system is $\hat{\varphi}_1 \otimes \hat{\varphi}_2$. Further suppose that new information is acquired such that $\hat{\varphi}_1$ would itself be updated to $\hat{\rho}_1$ and that $\hat{\varphi}_2$ would be itself be updated to $\hat{\rho}_2$. By design, the implementation of DC2 constrains the entropy functional such that in this case, the joint product prior density matrix $\hat{\varphi}_1 \otimes \hat{\varphi}_2$ updates to the product posterior $\hat{\rho}_1 \otimes \hat{\rho}_2$ so that inferences about one do not affect inferences about the other.

The argument below is considerably simplified if we expand the space of density matrices to include density matrices that are not necessarily normalized. This does not represent any limitation because normalization can always be easily achieved as one additional constraint. We consider a few special cases below:

Case 1: We receive the extremely constraining information that the posterior distribution for system 1 is completely specified to be $\hat{\rho}_1$ while we receive no information about system 2 at all. We treat the two systems jointly. Maximize the joint entropy $S[\hat{\rho}_{12}, \hat{\varphi}_1 \otimes \hat{\varphi}_2]$, subject to the following constraints on the $\hat{\rho}_{12}$,

$$\mathrm{Tr}_2(\hat{\rho}_{12}) = \hat{\rho}_1. \tag{44}$$

Notice all of the N^2 elements in $\mathcal{H}_1$ of $\hat{\rho}_{12}$ are being constrained. We therefore need a Lagrange multiplier which spans $\mathcal{H}_1$ and therefore it is a square matrix $\hat{\lambda}_1$. This is readily seen by observing the component form expressions of the Lagrange multipliers $(\hat{\lambda}_1)_{ij} = \lambda_{ij}$. Maximizing the entropy with respect to this $\mathcal{H}_2$ independent constraint is

$$0 = \delta\Big(S - \sum_{ij} \lambda_{ij}\Big(\mathrm{Tr}_2(\hat{\rho}_{1,2}) - \hat{\rho}_1\Big)_{ij}\Big), \tag{45}$$

but reexpressing this with its transpose $(\hat{\lambda}_1)_{ij} = (\hat{\lambda}_1^T)_{ji}$, gives

$$0 = \delta\Big(S - \mathrm{Tr}_1(\hat{\lambda}_1[\mathrm{Tr}_2(\hat{\rho}_{1,2}) - \hat{\rho}_1])\Big), \tag{46}$$

where we have relabeled $\hat{\lambda}_1^T \rightarrow \hat{\lambda}_1$, for convenience, as the name of the Lagrange multipliers are arbitrary. For arbitrary variations of $\hat{\rho}_{12}$, we therefore have

$$\hat{\lambda}_1 \otimes \hat{1}_2 = \phi\left(\hat{\rho}_{12}, \hat{\varphi}_1 \otimes \hat{\varphi}_2\right). \tag{47}$$

DC2 is implemented by requiring $\hat{\varphi}_1 \otimes \hat{\varphi}_2 \rightarrow \hat{\rho}_1 \otimes \hat{\varphi}_2$, such that the function ϕ is designed to reflect subsystem independence in this case; therefore, we have

$$\hat{\lambda}_1 \otimes \hat{1}_2 = \phi\left(\hat{\rho}_1 \otimes \hat{\varphi}_2, \hat{\varphi}_1 \otimes \hat{\varphi}_2\right). \tag{48}$$

Had we chosen a different prior $\hat{\varphi}_2' = \hat{\varphi}_2 + \delta\hat{\varphi}_2$, for all $\delta\hat{\varphi}_2$ the LHS $\hat{\lambda}_1 \otimes \hat{1}_2$ remains unchanged given that ϕ is independent of scalar functions (I would like to thank M. Krumm for pointing this out.) of $\hat{\varphi}_2$, as those could be lumped into $\hat{\lambda}_1$ while keeping $\hat{\rho}_1$ fixed. The potential dependence on scalar functions of $\hat{\varphi}_2$ can be removed by imposing DC2 in a subsystem independent situation where $\hat{\rho}_1'$ in ϕ need not be fixed under variations of $\hat{\varphi}_2$. The resulting equation in such a situation, for instance maximizing the entropy of an independent joint prior with respect to $\mathrm{Tr}(\hat{A}_1 \otimes \hat{1}_2 \cdot \hat{\rho}_{12}) = \langle A \rangle$, facilitated by a scalar Lagrange multiplier λ, and after imposing DC2,

$$\lambda \hat{A}_1 \otimes \hat{1}_2 = \phi\left(\hat{\rho}_1' \otimes \hat{\varphi}_2, \hat{\varphi}_1 \otimes \hat{\varphi}_2\right). \tag{49}$$

For subsystem independence to be imposed here, $\hat{\rho}_1'$ must be independent of variations in $\hat{\varphi}_2$, and, therefore, in a general subsystem independent case, ϕ is independent of scalar functions of $\hat{\varphi}_2$. This means that any dependence that the right-hand side of (48) might potentially have had on $\hat{\varphi}_2$ *must drop out*, meaning,

$$\phi\left(\hat{\rho}_1 \otimes \hat{\varphi}_2, \hat{\varphi}_1 \otimes \hat{\varphi}_2\right) = f(\hat{\rho}_1, \hat{\varphi}_1) \otimes \hat{1}_2. \tag{50}$$

Since $\hat{\varphi}_2$ is arbitrary, suppose further that we choose a unit prior, $\hat{\varphi}_2 = \hat{1}_2$, and note that $\hat{\rho}_1 \otimes \hat{1}_2$ and $\hat{\varphi}_1 \otimes \hat{1}_2$ are block diagonal in $\mathcal{H}_2$. Because the LHS is block diagonal in $\mathcal{H}_2$,

$$f(\hat{\rho}_1, \hat{\varphi}_1) \otimes \hat{1}_2 = \phi\left(\hat{\rho}_1 \otimes \hat{1}_2, \hat{\varphi}_1 \otimes \hat{1}_2\right). \tag{51}$$

The RHS is block diagonal in $\mathcal{H}_2$ and, because the function ϕ is understood to be a power series expansion in its arguments,

$$f(\hat{\rho}_1, \hat{\varphi}_1) \otimes \hat{1}_2 = \phi\left(\hat{\rho}_1 \otimes \hat{1}_2, \hat{\varphi}_1 \otimes \hat{1}_2\right) = \phi\left(\hat{\rho}_1, \hat{\varphi}_1\right) \otimes \hat{1}_2. \tag{52}$$

This gives

$$\hat{\lambda}_1 \otimes \hat{1}_2 = \phi\left(\hat{\rho}_1, \hat{\varphi}_1\right) \otimes \hat{1}_2, \tag{53}$$

and, therefore, the $\hat{1}_2$ factors out and $\hat{\lambda}_1 = \phi\left(\hat{\rho}_1, \hat{\varphi}_1\right)$. A similar argument exchanging systems 1 and 2 shows $\hat{\lambda}_2 = \phi\left(\hat{\rho}_2, \hat{\varphi}_2\right)$.

Case 1—Conclusion: The analysis leads us to conclude that when the system 2 is not updated, the dependence on $\hat{\varphi}_2$ drops out,

$$\phi\left(\hat{\rho}_1 \otimes \hat{\varphi}_2, \hat{\varphi}_1 \otimes \hat{\varphi}_2\right) = \phi\left(\hat{\rho}_1, \hat{\varphi}_1\right) \otimes \hat{1}_2, \tag{54}$$

and, similarly,

$$\phi\left(\hat{\varphi}_1 \otimes \hat{\rho}_2, \hat{\varphi}_1 \otimes \hat{\varphi}_2\right) = \hat{1}_1 \otimes \phi\left(\hat{\rho}_2, \hat{\varphi}_2\right). \tag{55}$$

Case 2: Now consider a different special case in which the marginal posterior distributions for systems 1 and 2 are both completely specified to be $\hat{\rho}_1$ and $\hat{\rho}_2$, respectively. Maximize the joint entropy, $S[\hat{\rho}_{12}, \hat{\varphi}_1 \otimes \hat{\varphi}_2]$, subject to the following constraints on the $\hat{\rho}_{12}$,

$$\mathrm{Tr}_2(\hat{\rho}_{12}) = \hat{\rho}_1 \quad \text{and} \quad \mathrm{Tr}_1(\hat{\rho}_{12}) = \hat{\rho}_2, \tag{56}$$

where $\mathrm{Tr}_i(\ldots)$ is the partial trace function, which a trace over the vectors in over

$\mathcal{H}_i$. Here, each expectation value constrains the entire space $\mathcal{H}_i$, where $\hat{\rho}_i$ lives. The Lagrange multipliers must span their respective spaces, so we implement the constraint with the Lagrange multiplier operator $\hat{\mu}_i$, then,

$$0 = \delta\Big(S - \mathrm{Tr}_1(\hat{\mu}_1[\mathrm{Tr}_2(\hat{\rho}_{12}) - \hat{\rho}_1]) - \mathrm{Tr}_2(\hat{\mu}_2[\mathrm{Tr}_1(\hat{\rho}_{12}) - \hat{\rho}_2])\Big). \tag{57}$$

For arbitrary variations of $\hat{\rho}_{12}$, we have

$$\hat{\mu}_1 \otimes \hat{1}_2 + \hat{1}_1 \otimes \hat{\mu}_2 = \phi\left(\hat{\rho}_{12}, \hat{\varphi}_1 \otimes \hat{\varphi}_2\right). \tag{58}$$

By design, DC2 is implemented by requiring $\hat{\varphi}_1 \otimes \hat{\varphi}_2 \to \hat{\rho}_1 \otimes \hat{\rho}_2$ in this case; therefore, we have

$$\hat{\mu}_1 \otimes \hat{1}_2 + \hat{1}_1 \otimes \hat{\mu}_2 = \phi\left(\hat{\rho}_1 \otimes \hat{\rho}_2, \hat{\varphi}_1 \otimes \hat{\varphi}_2\right). \tag{59}$$

Write (59) as

$$\hat{\mu}_1 \otimes \hat{1}_2 = \phi\left(\hat{\rho}_1 \otimes \hat{\rho}_2, \hat{\varphi}_1 \otimes \hat{\varphi}_2\right) - \hat{1}_1 \otimes \hat{\mu}_2\,. \tag{60}$$

The LHS is independent of changes that might occur in $\mathcal{H}_2$ on the RHS of (60). This means that any variation of $\hat{\rho}_2$ and $\hat{\varphi}_2$ must be "pushed out" by $\hat{\mu}_2$—it removes the dependence of $\hat{\rho}_2$ and $\hat{\varphi}_2$ in ϕ. Any dependence that the RHS might potentially have had on $\hat{\rho}_2$, $\hat{\varphi}_2$ must cancel out in a general subsystem independent case, leaving $\hat{\mu}_1$ unchanged. Consequently,

$$\phi\left(\hat{\rho}_1 \otimes \hat{\rho}_2, \hat{\varphi}_1 \otimes \hat{\varphi}_2\right) - \hat{1}_1 \otimes \hat{\mu}_2 = g(\hat{\rho}_1, \hat{\varphi}_1) \otimes \hat{1}_2. \tag{61}$$

Because $g(\hat{\rho}_1, \hat{\varphi}_1)$ is independent of arbitrary variations of $\hat{\rho}_2$ and $\hat{\varphi}_2$ on the LHS above—it is satisfied equally well for all choices. The form of $g(\hat{\rho}_1, \hat{\varphi}_1)$ reduces to the form of $f(\hat{\rho}_1, \hat{\varphi}_1)$ from Case 1 when $\hat{\rho}_2 = \hat{\varphi}_2 = \hat{1}_2$ and, similarly, DC1′ gives $\hat{\mu}_2 = 0$. Therefore, the Lagrange multiplier is

$$\hat{\mu}_1 \otimes \hat{1}_2 = \phi(\hat{\rho}_1, \hat{\varphi}_1) \otimes \hat{1}_2. \tag{62}$$

A similar analysis is carried out for $\hat{\mu}_2$ leading to

$$\hat{1}_1 \otimes \hat{\mu}_2 = \hat{1}_1 \otimes \phi(\hat{\rho}_2, \hat{\varphi}_2). \tag{63}$$

Case 2—Conclusion: Substituting back into (59) gives us a functional equation for ϕ ,

$$\phi(\hat{\rho}_1 \otimes \hat{\rho}_2, \hat{\varphi}_1 \otimes \hat{\varphi}_2) = \phi(\hat{\rho}_1, \hat{\varphi}_1) \otimes \hat{1}_2 + \hat{1}_1 \otimes \phi(\hat{\rho}_2, \hat{\varphi}_2), \tag{64}$$

which is

$$\phi(\hat{\rho}_1 \otimes \hat{\rho}_2, \hat{\varphi}_1 \otimes \hat{\varphi}_2) = \phi(\hat{\rho}_1 \otimes \hat{1}_2, \hat{\varphi}_1 \otimes \hat{1}_2) + \phi(\hat{1}_1 \otimes \hat{\rho}_2, \hat{1}_1 \otimes \hat{\varphi}_2). \tag{65}$$

The general solution to this matrix valued functional equation is derived in Appendix A.5 and is

$$\phi(\hat{\rho}, \hat{\varphi}) = \tilde{A} \ln(\hat{\rho}) + \tilde{B} \ln(\hat{\varphi}), \tag{66}$$

where tilde $\tilde{A}$ is a "super-operator" having constant coefficients and twice the number of indicies as $\hat{\rho}$ and $\hat{\varphi}$ as discussed in the Appendix (i.e., $\left(\tilde{A} \ln(\hat{\rho})\right)_{ij} = \sum_{k\ell} A_{ijk\ell}(\log(\hat{\rho}))_{k\ell}$ and similarly for $\tilde{B} \ln(\hat{\varphi})$). DC1′ imposes

$$\phi(\hat{\varphi}, \hat{\varphi}) = \tilde{A} \ln(\hat{\varphi}) + \tilde{B} \ln(\hat{\varphi}) = \hat{0}, \tag{67}$$

which is satisfied in general when $\tilde{A} = -\tilde{B}$, and, now,

$$\phi(\hat{\rho}, \hat{\varphi}) = \tilde{A} \left(\ln(\hat{\rho}) - \ln(\hat{\varphi})\right). \tag{68}$$

We may fix the constant $\tilde{A}$ by substituting our solution into the RHS of Equation (64), which is equal to the RHS of Equation (65),

$$\left(\tilde{A}_1 \left(\ln(\hat{\rho}_1) - \ln(\hat{\varphi}_1)\right)\right) \otimes \hat{1}_2 + \hat{1}_1 \otimes \left(\tilde{A}_2 \left(\ln(\hat{\rho}_2) - \ln(\hat{\varphi}_2)\right)\right)$$

$$= \tilde{A}_{12} \left(\ln(\hat{\rho}_1 \otimes \hat{1}_2) - \ln(\hat{\varphi}_1 \otimes \hat{1}_2)\right) + \tilde{A}_{12} \left(\ln(\hat{1}_1 \otimes \hat{\rho}_2) - \ln(\hat{1}_1 \otimes \hat{\varphi}_2)\right), \tag{69}$$

where $\tilde{A}_{12}$ acts on the joint space of 1 and 2 and $\tilde{A}_1$, $\tilde{A}_2$ acts on single subspaces 1 or 2, respectively. Using the well known log tensor product identity in this case (The proof is demonstrated by taking the log of $\hat{\rho}_1 \otimes \hat{1}_2 \equiv \exp(\hat{\rho}_1') \otimes \hat{1}_2 = \exp(\hat{\rho}_1' \otimes \hat{1}_2)$ and substituting $\hat{\rho}_1' = \log(\hat{\rho}_1)$.), $\ln(\hat{\rho}_1 \otimes \hat{1}_2) = \ln(\hat{\rho}_1) \otimes \hat{1}_2$, the RHS of Equation (69) becomes

$$= \tilde{A}_{12} \left(\ln(\hat{\rho}_1) \otimes \hat{1}_2 - \ln(\hat{\varphi}_1) \otimes \hat{1}_2\right) + \tilde{A}_{12} \left(\hat{1}_1 \otimes \ln(\hat{\rho}_2) - \hat{1}_1 \otimes \ln(\hat{\varphi}_2)\right). \tag{70}$$

Note that arbitrarily letting $\hat{\rho}_2 = \hat{\varphi}_2$ gives

$$\left(\tilde{A}_1 \left(\ln(\hat{\rho}_1) - \ln(\hat{\varphi}_1)\right)\right) \otimes \hat{1}_2 = \tilde{A}_{12} \left(\ln(\hat{\rho}_1) \otimes \hat{1}_2 - \ln(\hat{\varphi}_1) \otimes \hat{1}_2\right), \tag{71}$$

or arbitrarily letting $\hat{\rho}_1 = \hat{\varphi}_1$ gives

$$\hat{1}_1 \otimes \Big(\tilde{A}_2 \left(\ln(\hat{\rho}_2) - \ln(\hat{\varphi}_2) \right) \Big) = \tilde{A}_{12} \left(\hat{1}_1 \otimes \ln(\hat{\rho}_2) - \hat{1}_1 \otimes \ln(\hat{\varphi}_2) \right). \tag{72}$$

As $\tilde{A}_{12}$, $\tilde{A}_1$, and $\tilde{A}_2$ are constant tensors, inspecting the above equalities determines the form of the tensor to be $\tilde{A} = A\,\overset{\sim}{1}$ where A is a scalar constant and $\overset{\sim}{1}$ is the super-operator identity over the appropriate (joint) Hilbert space.

Because our goal is to maximize the entropy function, we let the arbitrary constant $A = -|A|$ and distribute $\overset{\sim}{1}$ identically, which gives the final functional form,

$$\phi(\hat{\rho}, \hat{\varphi}) = -|A| \Big(\ln(\hat{\rho}) - \ln(\hat{\varphi}) \Big). \tag{73}$$

"Integrating" ϕ gives a general form for the quantum relative entropy,

$$S(\hat{\rho}, \hat{\varphi}) = -|A|\mathrm{Tr}(\hat{\rho}\log\hat{\rho} - \hat{\rho}\log\hat{\varphi} - \hat{\rho}) + C[\hat{\varphi}] = -|A|S_U(\hat{\rho}, \hat{\varphi}) + |A|\mathrm{Tr}(\hat{\rho}) + C[\hat{\varphi}], \tag{74}$$

where $S_U(\hat{\rho}, \hat{\varphi})$ is Umegaki's form of the relative entropy [42–44], the extra $|A|\mathrm{Tr}(\hat{\rho})$ from integration is an artifact present for the preservation of DC1′, and $C[\hat{\varphi}]$ is a constant in the sense that it drops out under arbitrary variations of $\hat{\rho}$. This entropy leads to the same inferences as Umegaki's form of the entropy with an added bonus that $\hat{\rho} = \hat{\varphi}$ in the absence of constraints or changes in information—rather than $\hat{\rho} = e^{-1}\hat{\varphi}$, which would be given by maximizing Umegaki's form of the entropy. In this sense, the extra $|A|\mathrm{Tr}(\hat{\rho})$ only improves the inference process as it more readily adheres to the PMU though DC1′; however, now, because $S_U \geq 0$, we have $S(\hat{\rho}, \hat{\varphi}) \leq \mathrm{Tr}(\hat{\rho}) + C[\hat{\varphi}]$, which provides little nuisance. In the spirit of this derivation, we will keep the $\mathrm{Tr}(\hat{\rho})$ term there, but, for all practical purposes of inference, as long as there is a normalization constraint, it plays no role, and we find (letting $|A| = 1$ and $C[\hat{\varphi}] = 0$),

$$S(\hat{\rho}, \hat{\varphi}) \rightarrow S^*(\hat{\rho}, \hat{\varphi}) = -S_U(\hat{\rho}, \hat{\varphi}) = -\mathrm{Tr}(\hat{\rho}\log\hat{\rho} - \hat{\rho}\log\hat{\varphi}), \tag{75}$$

Umegaki's form of the relative entropy. $S^*(\hat{\rho}, \hat{\varphi})$ is an equally valid entropy because, given normalization is applied, the same selected posterior $\hat{\rho}$ maximizes both $S(\hat{\rho}, \hat{\varphi})$ and $S^*(\hat{\rho}, \hat{\varphi})$.

3.2. Remarks

Due to the universality and the equal application of the PMU by using the same design criteria for both the standard and quantum case, the quantum relative entropy reduces to the standard relative entropy when $[\hat{\rho}, \hat{\varphi}] = 0$ or when the experiment being preformed $\hat{\rho} \rightarrow \rho(a) = \mathrm{Tr}(\hat{\rho}|a\rangle\langle a|)$ is known. The quantum relative entropy we derive has the correct asymptotic form of the standard relative entropy in the sense of [8–10]. Further connections will be illustrated in a follow up article that is concerned with direct applications of the quantum relative entropy. Because two entropies are derived in parallel, we expect the well-known inferential results and consequences of the relative entropy to have a quantum relative entropy representation.

Maximizing the quantum relative entropy with respect to some constraints $\langle \hat{A}_i \rangle$, where $\{\hat{A}_i\}$ are a set of arbitrary Hermitian operators, and normalization $\langle \hat{1} \rangle = 1$, gives the following general solution for the posterior density matrix:

$$\hat{\rho} = \exp\left(\alpha_0 \hat{1} + \sum_i \alpha_i \hat{A}_i + \ln(\hat{\varphi}) \right) = \frac{1}{Z}\exp\left(\sum_i \alpha_i \hat{A}_i + \ln(\hat{\varphi}) \right) \equiv \frac{1}{Z}\exp\left(\hat{C} \right), \tag{76}$$

where α_i are the Lagrange multipliers of the respective constraints and normalization may be factored out of the exponential in general because the identity commutes universally. If $\hat{\varphi} \propto \hat{1}$, it is well known that the analysis arrives at the same expression for $\hat{\rho}$ after normalization, as it would if the von Neumann entropy were used, and thus one can find expressions for thermalized quantum states $\hat{\rho} = \frac{1}{Z} e^{-\beta \hat{H}}$. The remaining problem is to solve for the N Lagrange multipliers using their N associated expectation value constraints. In principle, their solution is found by computing Z and using standard methods from Statistical Mechanics,

$$\langle \hat{A}_i \rangle = -\frac{\partial}{\partial \alpha_i} \ln(Z), \tag{77}$$

and inverting to find $\alpha_i = \alpha_i(\langle \hat{A}_i \rangle)$, which has a unique solution due to the joint concavity (convexity depending on the sign convention) of the quantum relative entropy [8,9] when the constraints are linear in $\hat{\rho}$. The simple proof that (77) is monotonic in α, and therefore invertible, is that its derivative $\frac{\partial}{\partial \alpha} \langle \hat{A}_i \rangle = \langle \hat{A}_i^2 \rangle - \langle \hat{A}_i \rangle^2 \geq 0$. Between the Zassenhaus formula [45]

$$e^{t(\hat{A}+\hat{B})} = e^{t\hat{A}} e^{t\hat{B}} e^{-\frac{t^2}{2}[\hat{A},\hat{B}]} e^{\frac{t^3}{6}(2[\hat{B},[\hat{A},\hat{B}]]+[\hat{A},[\hat{A},\hat{B}]])} \ldots, \tag{78}$$

and Horn's inequality [46–48], the solutions to (77) lack a certain calculational elegance because it is difficult to express the eigenvalues of $\hat{C} = \log(\hat{\varphi}) + \sum \alpha_i \hat{A}_i$ (in the exponential) in simple terms of the eigenvalues of the $\hat{A}_i$'s and $\hat{\varphi}$, in general, when the matrices do not commute. The solution requires solving the eigenvalue problem for $\hat{C}$, such the the exponential of $\hat{C}$ may be taken and evaluated in terms of the eigenvalues of the $\alpha_i \hat{A}_i$s and the prior density matrix $\hat{\varphi}$. A pedagogical exercise is starting with a prior that is a mixture of spin-z up and down $\hat{\varphi} = a|+\rangle\langle+| + b|-\rangle\langle-|$ $(a, b \neq 0)$, maximizing the quantum relative entropy with respect to an expectation of a general Hermitian operator with which the prior density matrix does not commute. This example for spin is given in the Appendix B.

4. Conclusions

This approach emphasizes the notion that entropy is a tool for performing inference and downplays counter-notional issues that arise if one interprets entropy as a measure of disorder, a measure of distinguishability, or an amount of missing information [7]. Because the same design criteria, guided by the PMU, are applied equally well to the design of a relative and quantum relative entropy, we find that both the relative and quantum relative entropy are designed for the purpose of inference. Because the quantum relative entropy is the functional that fits the requirements of a tool designed for the inference of density matrices, we now know what it is and how to use it—formulating an inferential quantum maximum entropy method. This article provides the foundation for [29], which, in particular, derives the Quantum Bayes Rule and collapse as special cases of the quantum maximum entropy method, as was craved in [24], analogous to [38,40]'s treatment for deriving Bayes Rule using the standard maximum entropy method. The quantum maximum entropy method thereby unifies a few topics in Quantum Information and Quantum Measurement through entropic inference.

Acknowledgments: I must give ample acknowledgment to Ariel Caticha who suggested the problem of justifying the form of the quantum relative entropy as a criterion for ranking of density matrices. He cleared up several difficulties by suggesting that design constraints be applied to the variational derivative of the entropy rather than the entropy itself. In addition, he provided substantial improvements to the method for imposing DC2 that led to the functional equations for the variational derivatives ($\phi_{12} = \phi_1 + \phi_2$)—with more rigor than in earlier versions of this article. His time and guidance are all greatly appreciated—thanks, Ariel. I would also like to thank M. Krumm, the reviewers, as well as our information physics group at UAlbany for our many intriguing discussions about probability, inference, and quantum mechanics.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

The Appendix loosely follows the relevant sections in [49], and then uses the methods reviewed to solve the relevant functional equations for ϕ. The last section is an example of the quantum maximum entropy method applied to a mixed spin state.

Appendix A.1. Simple Functional Equations

From [49] pages 31–44.

Theorem A1. *If Cauchy's functional equation*

$$f(x+y) = f(x) + f(y) \tag{A1}$$

is satisfied for all real x, y, and if the function $f(x)$ is (a) continuous at a point, (b) nonnegative for small positive x's, or (c) bounded in an interval, then,

$$f(x) = cx \tag{A2}$$

is the solution to (A1) for all real x. If (A1) is assumed only over all positive x, y, then under the same conditions, (A2) holds for all positive x.

Proof. The most natural assumption for our purposes is that $f(x)$ is continuous at a point (which later extends to continuity all points as given by Darboux [50]). Cauchy solved the functional equation by induction. In particular, Equation (A1) implies,

$$f(\sum_i x_i) = \sum_i f(x_i), \tag{A3}$$

and if we let each $x_i = x$ as a special case to determine f, we find

$$f(nx) = nf(x). \tag{A4}$$

We may let $nx = mt$ such that

$$f(x) = f(\frac{m}{n}t) = \frac{m}{n}f(t). \tag{A5}$$

Letting $\lim_{t\to 1} f(t) = f(1) = c$ gives

$$f(\frac{m}{n}) = \frac{m}{n}f(1) = \frac{m}{n}c, \tag{A6}$$

and, because for $t = 1$, $x = \frac{m}{n}$ above, we have

$$f(x) = cx, \tag{A7}$$

which is the general solution of the linear functional equation. In principle, c can be complex. The importance of Cauchy's solution is that it can be used to give general solutions to the following Cauchy equations:

$$\begin{aligned} f(x+y) &= f(x)f(y), && \text{(A8)}\\ f(xy) &= f(x)+f(y), && \text{(A9)}\\ f(xy) &= f(x)f(y), && \text{(A10)} \end{aligned}$$

by preforming consistent substitution until they are the same form as (A1), as given by Cauchy. We will briefly discuss the first two. □

Theorem A2. *The general solution of $f(x+y) = f(x)f(y)$ is $f(x) = e^{cx}$ for all real or for all positive x, y that are continuous at one point and, in addition to the exponential solution, the solution $f(0) = 1$ and $f(x) = 0$ for $(x > 0)$ are in these classes of functions.*

The first functional $f(x+y)=f(x)f(y)$ is solved by first noting that it is strictly positive for real x, y, $f(x)$, which can be shown by considering $x=y$,

$$f(2x)=f(x)^2>0. \tag{A11}$$

If there exists $f(x_0)=0$, then it follows that $f(x)=f((x-x_0)+x_0)=0$, a trivial solution, hence the reason why the possibility of being equal to zero is excluded above. Given $f(x)$ is nowhere zero, we are justified in taking the natural logarithm $\ln(x)$, due to its positivity $f(x)>0$. This gives,

$$\ln(f(x+y))=\ln(f(x))+\ln(f(y)), \tag{A12}$$

and letting $g(x)=\ln(f(x))$ gives,

$$g(x+y)=g(x)+g(y), \tag{A13}$$

which is Cauchy's linear equation, and thus has the solution $g(x)=cx$. Because $g(x)=\ln(f(x))$, one finds in general that $f(x)=e^{cx}$.

Theorem A3. *If the functional equation $f(xy)=f(x)+f(y)$ is valid for all positive x,y then its general solution is $f(x)=c\ln(x)$ given it is continuous at a point. If $x=0$ (or $y=0$) are valid, then the general solution is $f(x)=0$. If all real x,y are valid except 0, then the general solution is $f(x)=c\ln(|x|)$.*

In particular, we are interested in the functional equation $f(xy)=f(x)+f(y)$ when x,y are positive. In this case, we can again follow Cauchy and substitute $x=e^u$ and $y=e^v$ to get,

$$f(e^ue^v)=f(e^u)+f(e^v), \tag{A14}$$

and letting $g(u)=f(e^u)$ gives $g(u+v)=g(u)+g(v)$. Again, the solution is $g(u)=cu$ and, therefore, the general solution is $f(x)=c\ln(x)$ when we substitute for u. If x could equal 0, then $f(0)=f(x)+f(0)$, which has the trivial solution $f(x)=0$. The general solution for $x\neq 0$, $y\neq 0$ and x,y positive is therefore $f(x)=c\ln(x)$.

Appendix A.2. Functional Equations with Multiple Arguments

From [49] pages 213–217. Consider the functional equation,

$$F(x_1+y_1,x_2+y_2,...,x_n+y_n)=F(x_1,x_2,...,x_n)+F(y_1,y_2,...,y_n), \tag{A15}$$

which is a generalization of Cauchy's linear functional Equation (A1) to several arguments. Letting $x_2=x_3=...=x_n=y_2=y_3=...=y_n=0$ gives

$$F(x_1+y_1,0,...,0)=F(x_1,0,...,0)+F(y_1,0,...,0), \tag{A16}$$

which is the Cauchy linear functional equation having solution $F(x_1,0,...,0)=c_1x_1$, where $F(x_1,0,...,0)$ is assumed to be continuous or at least measurable majorant. Similarly,

$$F(0,...,0,x_k,0,...,0)=c_kx_k, \tag{A17}$$

and if you consider

$$F(x_1+0,0+y_2,0,...,0)=F(x_1,0,...,0)+F(0,y_2,0,...,0)=c_1x_1+c_2y_2, \tag{A18}$$

and, as y_2 is arbitrary, we could have let $y_2=x_2$ such that in general

$$F(x_1,x_2,...,x_n)=\sum c_ix_i, \tag{A19}$$

formulating the general solution.

Appendix A.3. Relative Entropy

We are interested in the following functional equation:

$$\phi(\rho_1\rho_2, \varphi_1\varphi_2) = \phi(\rho_1, \varphi_1) + \phi(\rho_2, \varphi_2). \tag{A20}$$

This is an equation of the form,

$$F(x_1y_1, x_2y_2) = F(x_1, x_2) + F(y_1, y_2), \tag{A21}$$

where $x_1 = \rho(x_1)$, $y_1 = \rho(x_2)$, $x_2 = \varphi(x_1)$, and $y_2 = \varphi(x_2)$. First, assume all q and p are greater than zero. Then, substitute: $x_i = e^{x_i'}$ and $y_i = e^{y_i'}$ and let $F'(x_1', x_2') = F(e^{x_1'}, e^{x_2'})$ and so on such that

$$F'(x_1' + y_1', x_2' + y_2') = F'(x_1', x_2') + F'(y_1', y_2'), \tag{A22}$$

which is of the form of (A15). The general solution for F is therefore

$$F'(x_1' + y_1', x_2' + y_2') = a_1(x_1' + y_1') + a_2(x_2' + y_2') = a_1 \ln(x_1y_1) + a_2 \ln(x_2y_2) = F(x_1y_1, x_2y_2), \tag{A23}$$

which means the general solution for ϕ is

$$\phi(\rho_1, \varphi_1) \quad = \quad a_1 \ln(\rho(x_1)) + a_2 \ln(\varphi(x_1)). \tag{A24}$$

In such a case, when $\varphi(x_0) = 0$ for some value $x_0 \in \mathcal{X}$, we may let $\varphi(x_0) = \epsilon$, where ϵ is as close to zero as we could possibly want—the trivial general solution $\phi = 0$ is saturated by the special case when $\rho = \varphi$ from DC1′. Here, we return to the text.

Appendix A.4. Matrix Functional Equations

(This derivation is implied in [49] pages 347–349). First, consider a Cauchy matrix functional equation,

$$f(\hat{X} + \hat{Y}) = f(\hat{X}) + f(\hat{Y}), \tag{A25}$$

where $\hat{X}$ and $\hat{Y}$ are $n \times n$ square matrices. Rewriting the matrix functional equation in terms of its components gives

$$f_{ij}(x_{11} + y_{11}, x_{12} + y_{12}, ..., x_{nn} + y_{nn}) = f_{ij}(x_{11}, x_{12}, ..., x_{nn}) + f_{ij}(y_{11}, y_{12}, ..., y_{nn}) \tag{A26}$$

and is now in the form of (A15), and, therefore, the solution is

$$f_{ij}(x_{11}, x_{12}, ..., x_{nn}) = \sum_{\ell,k=0}^{n} c_{ij\ell k} x_{\ell k} \tag{A27}$$

for $i, j = 1, ..., n$. We find it convenient to introduce super indices, $A = (i, j)$ and $B = (\ell, k)$ such that the component equation becomes

$$f_A = \sum_B c_{AB} x_B, \tag{A28}$$

and resembles the solution for the linear transformation of a vector from [49]. In general, we will be discussing matrices $\hat{X} = \hat{X}_1 \otimes \hat{X}_2 \otimes ... \otimes \hat{X}_N$ which stem from tensor products of density matrices. In this situation, $\hat{X}$ can be thought of as $2N$ index tensor or a $z \times z$ matrix where $z = \prod_i^N n_i$ is the product of the ranks of the matrices in the tensor product or even as a vector of length z^2. In such a case, we may abuse the super index notation where A and B lump together the appropriate number of indices such that (A28) is the form of the solution for the components in general. The matrix form of the general solution is

$$f(\hat{X}) = \widetilde{C}\hat{X}, \tag{A29}$$

where $\widetilde{C}$ is a constant super-operator having components c_{AB}.

Appendix A.5. Quantum Relative Entropy

The functional equation of interest is

$$\phi\Big(\hat{\rho}_1 \otimes \hat{\rho}_2, \hat{\varphi}_1 \otimes \hat{\varphi}_2\Big) = \phi\Big(\hat{\rho}_1 \otimes \hat{1}_2, \hat{\varphi}_1 \otimes \hat{1}_2\Big) + \phi\Big(\hat{1}_1 \otimes \hat{\rho}_2, \hat{1}_1 \otimes \hat{\varphi}_2\Big). \tag{A30}$$

These density matrices are Hermitian, positive semi-definite, have positive eigenvalues, and are not equal to $\hat{0}$. Because every invertible matrix can be expressed as the exponential of some other matrix, we can substitute $\hat{\rho}_1 = e^{\hat{\rho}'_1}$, and so on for all four density matrices giving,

$$\phi\Big(e^{\hat{\rho}'_1} \otimes e^{\hat{\rho}'_2}, e^{\hat{\varphi}'_1} \otimes e^{\hat{\varphi}'_2}\Big) = \phi\Big(e^{\hat{\rho}'_1} \otimes \hat{1}_2, e^{\hat{\varphi}'_1} \otimes \hat{1}_2\Big) + \phi\Big(\hat{1}_1 \otimes e^{\hat{\rho}'_2}, \hat{1}_1 \otimes e^{\hat{\varphi}'_2}\Big). \tag{A31}$$

Now, we use the following identities for Hermitian matrices:

$$e^{\hat{\rho}'_1} \otimes e^{\hat{\rho}'_2} = e^{\hat{\rho}'_1 \otimes \hat{1}_2 + \hat{1}_1 \otimes \hat{\rho}'_2} \tag{A32}$$

and

$$e^{\hat{\rho}'_1} \otimes \hat{1_2} = e^{\hat{\rho}'_1 \otimes \hat{1}_2}, \tag{A33}$$

to recast the functional equation as,

$$\phi\Big(e^{\hat{\rho}'_1 \otimes \hat{1}_2 + \hat{1}_1 \otimes \hat{\rho}'_2}, e^{\hat{\varphi}'_1 \otimes \hat{1}_2 + \hat{1}_1 \otimes \hat{\varphi}'_2}\Big) = \phi\Big(e^{\hat{\rho}'_1 \otimes \hat{1}_2}, e^{\hat{\varphi}'_1 \otimes \hat{1}_2}\Big) + \phi\Big(e^{\hat{1}_1 \otimes \hat{\rho}'_2}, e^{\hat{1}_1 \otimes \hat{\varphi}'_2}\Big). \tag{A34}$$

Letting $G(\hat{\rho}'_1 \otimes \hat{1}_2, \hat{\varphi}'_1 \otimes \hat{1}_2) = \phi\Big(e^{\hat{\rho}'_1 \otimes \hat{1}_2}, e^{\hat{\varphi}'_1 \otimes \hat{1}_2}\Big)$, and the like, gives

$$G(\hat{\rho}'_1 \otimes \hat{1}_2 + \hat{1}_1 \otimes \hat{\rho}'_2, \hat{\varphi}'_1 \otimes \hat{1}_2 + \hat{1}_1 \otimes \hat{\varphi}'_2) = G(\hat{\rho}'_1 \otimes \hat{1}_2, \hat{\varphi}'_1 \otimes \hat{1}_2) + G(\hat{1}_1 \otimes \hat{\rho}'_2, \hat{1}_1 \otimes \hat{\varphi}'_2). \tag{A35}$$

This functional equation is of the form

$$G(\hat{X}'_1 + \hat{Y}'_1, \hat{X}'_2 + \hat{Y}'_2) = G(\hat{X}'_1, \hat{X}'_2) + G(\hat{Y}'_1, \hat{Y}'_2), \tag{A36}$$

which has the general solution

$$G(\hat{X}', \hat{Y}') = \widetilde{A}\ \hat{X}' + \widetilde{B}\hat{Y}', \tag{A37}$$

analogous to (A19), and finally, in general,

$$\phi(\hat{\rho}, \hat{\varphi}) = \widetilde{A}\ \ln(\hat{\rho}) + \widetilde{B}\ln(\hat{\varphi}), \tag{A38}$$

where $\widetilde{A}$, $\widetilde{B}$ are super-operators having constant coefficients. Here, we return to the text.

Appendix B. Spin Example

Consider an arbitrarily mixed prior (in the spin-z basis for convenience) with $a, b \neq 0$,

$$\hat{\varphi} = a|+\rangle\langle+| + b|-\rangle\langle-| \tag{A39}$$

and a general Hermitian matrix in the spin-1/2 Hilbert space,

$$c_\mu \hat{\sigma}^\mu = c_1 \hat{1} + c_x \hat{\sigma}_x + c_y \hat{\sigma}_x + c_z \hat{\sigma}_z \tag{A40}$$

$$= (c_1 + c_z)|+\rangle\langle+| + (c_x - ic_y)|+\rangle\langle-| + (c_x + ic_y)|-\rangle\langle+| + (c_1 - c_z)|-\rangle\langle-|, \tag{A41}$$

having a known expectation value,

$$\mathrm{Tr}(\hat{\rho} c_\mu \hat{\sigma}^\mu) = c. \tag{A42}$$

Maximizing the entropy with respect to this general expectation value and normalization is:

$$0 = \Big(\delta S - \lambda[\mathrm{Tr}(\hat{\rho}) - 1] - \alpha(\mathrm{Tr}(\hat{\rho} c_\mu \hat{\sigma}^\mu) - c)\Big), \tag{A43}$$

which after varying gives the solution,

$$\hat{\rho} = \frac{1}{Z} \exp(\alpha c_\mu \hat{\sigma}^\mu + \log(\hat{\varphi})). \tag{A44}$$

Letting

$$\hat{C} = \alpha c_\mu \hat{\sigma}^\mu + \log(\hat{\varphi}) \tag{A45}$$

gives

$$\begin{aligned} \hat{\rho} = \frac{1}{Z} e^{\hat{C}} = U e^{U^{-1} \hat{C} U} U^{-1} = \frac{1}{Z} U e^{\hat{\lambda}} U^{-1} \\ = \frac{e^{\lambda_+}}{Z} U|\lambda_+\rangle\langle\lambda_+|U^{-1} + \frac{e^{\lambda_-}}{Z} U|\lambda_-\rangle\langle\lambda_-|U^{-1}, \end{aligned} \tag{A46}$$

where $\hat{\lambda}$ is the diagonalized matrix of $\hat{C}$ having real eigenvalues. They are

$$\lambda_\pm = \lambda \pm \delta\lambda, \tag{A47}$$

due to the quadratic formula, where explicitly:

$$\lambda = \alpha c_1 + \frac{1}{2} \log(ab), \tag{A48}$$

and

$$\delta\lambda = \frac{1}{2} \sqrt{\Big(2\alpha c_z + \log(\frac{a}{b})\Big)^2 + 4\alpha^2 (c_x^2 + c_y^2)}. \tag{A49}$$

Because $\lambda_\pm$ and a, b, c_1, c_x, c_y, c_z are real, $\delta\lambda$ is real and ≥ 0. The normalization constraint specifies the Lagrange multiplier Z,

$$1 = \mathrm{Tr}(\hat{\rho}) = \frac{e^{\lambda_+} + e^{\lambda_-}}{Z}, \tag{A50}$$

so $Z = e^{\lambda_+} + e^{\lambda_-} = 2e^{\lambda}\cosh(\delta\lambda)$. The expectation value constraint specifies the Lagrange multiplier α,

$$c = \mathrm{Tr}(\hat{\rho} c_\mu \sigma^\mu) = \frac{\partial}{\partial\alpha}\log(Z) = c_1 + \tanh(\delta\lambda)\frac{\partial}{\partial\alpha}\delta\lambda, \tag{A51}$$

which becomes

$$c = c_1 + \frac{\tanh(\delta\lambda)}{2\delta\lambda}\Big(2\alpha(c_x^2 + c_y^2 + c_z^2) + c_z\log(\frac{a}{b})\Big),$$

or

$$c = c_1 + \tanh\left(\frac{1}{2}\sqrt{\Big(2\alpha c_z + \log(\frac{a}{b})\Big)^2 + 4\alpha^2(c_x^2 + c_y^2)}\right)\frac{2\alpha(c_x^2 + c_y^2 + c_z^2) + c_z\log(\frac{a}{b})}{\sqrt{\Big(2\alpha c_z + \log(\frac{a}{b})\Big)^2 + 4\alpha^2(c_x^2 + c_y^2)}}. \tag{A52}$$

This equation is monotonic in α and therefore it is uniquely specified by the value of c. Ultimately, this is a consequence from the concavity of the entropy. The specific proof of (A52)'s monotonicity is below:

Proof. For $\hat{\rho}$ to be Hermitian, $\hat{C}$ is Hermitian and $\delta\lambda = \frac{1}{2}\sqrt{f(\alpha)}$ is real—furthermore, because $\delta\lambda$ is real $f(\alpha) \geq 0$ and thus $\delta\lambda \geq 0$. Because $f(\alpha)$ is quadratic in α and positive, it may be written in vertex form,

$$f(\alpha) = a(\alpha - h)^2 + k, \tag{A53}$$

where $a > 0$, $k \geq 0$, and (h,k) are the (x,y) coordinates of the minimum of $f(\alpha)$. Notice that the form of (A52) is

$$F(\alpha) = \frac{\tanh(\frac{1}{2}\sqrt{f(\alpha)})}{\sqrt{f(\alpha)}} \times \frac{\partial f(\alpha)}{\partial\alpha}. \tag{A54}$$

Making the change of variables $\alpha' = \alpha - h$ centers the function such that $f(\alpha') = f(-\alpha')$ is symmetric about $\alpha' = 0$. We can then write

$$F(\alpha') = \frac{\tanh(\frac{1}{2}\sqrt{f(\alpha')})}{\sqrt{f(\alpha')}} \times 2a\alpha', \tag{A55}$$

where the derivative has been computed. Because $f(\alpha')$ is a positive, symmetric, and monotonically increasing on the (symmetric) half-plane (for α' greater than or less that zero), $S(\alpha') \equiv \frac{\tanh(\frac{1}{2}\sqrt{f(\alpha')})}{\sqrt{f(\alpha')}}$ is also positive and symmetric, but it is unclear whether $S(\alpha)$ is strictly monotonic in the half-plane or not. We may restate

$$F(\alpha') = S(\alpha') \times 2a\alpha'. \tag{A56}$$

We are now in a convenient position to preform the derivate test for monotonic functions:

$$\begin{aligned}
\frac{\partial}{\partial\alpha'}F(\alpha') &= 2aS(\alpha') + 2a\alpha'\frac{\partial}{\partial\alpha'}S(\alpha') \\
&= 2aS(\alpha')\Big(1 - \frac{a\alpha'^2}{a\alpha'^2 + k}\Big) + a\frac{a\alpha'^2}{a\alpha'^2 + k}\left(1 - \tanh^2(\frac{1}{2}\sqrt{a\alpha'^2 + k})\right) \\
&\geq 2aS(\alpha')\Big(1 - \frac{a(\alpha')^2}{a\alpha'^2 + k}\Big) \geq 0
\end{aligned} \tag{A57}$$

because $a, k, S(\alpha')$, and therefore $\frac{a\alpha'^2}{a\alpha'^2+k}$ are all > 0. The function of interest $F(\alpha')$ is therefore monotonic for all α', and therefore it is monotonic for all α, completing the proof that there exists a unique real Lagrange multiplier α in (A52).

Although (A52) is monotonic in α, it is seemingly a transcendental equation. This can be solved graphically for the given values c, c_1, c_x, c_y, c_z, i.e., given the Hermitian matrix and its expectation value are specified. Equation (A52) and the eigenvalues take a simpler form when $a = b = \frac{1}{2}$ because, in this instance, $\hat{\varphi} \propto \hat{1}$ and commutes universally so it may be factored out of the exponential in (A44). □

References

1. Shore, J.E.; Johnson, R.W. Axiomatic derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy. *IEEE Trans. Inf. Theory* **1980**, *26*, 26–37.
2. Shore, J.E.; Johnson, R.W. Properties of Cross-Entropy Minimization. *IEEE Trans. Inf. Theory* **1981**, *27*, 472–482.
3. Csiszár, I. Why least squares and maximum entropy: An axiomatic approach to inference for linear inverse problems. *Ann. Stat.* **1991**, *19*, 2032.
4. Skilling, J. The Axioms of Maximum Entropy. In *Maximum-Entropy and Bayesian Methods in Science and Engineering*; Erickson, G.J., Smith, C.R., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1988.
5. Skilling, J. Classic Maximum Entropy. In *Maximum-Entropy and Bayesian Methods in Science and Engineering*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1988.
6. Skilling, J. Quantified Maximum Entropy. In *Maximum-Entropy and Bayesian Methods in Science and Engineering*; Fougére, P.F., Ed.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1990.
7. Caticha, A. Entropic Inference and the Foundations of Physics (Monograph Commissioned by the 11th Brazilian Meeting on Bayesian Statistics—EBEB-2012). Available online: http://www.albany.edu/physics/ACaticha-EIFP-book.pdf (accessed on 30 November 2017).
8. Hiai, F.; Petz, D. The Proper Formula for Relative Entropy and its Asymptotics in Quantum Probability. *Commun. Math. Phys.* **1991**, *143*, 99–114.
9. Petz, D. Characterization of the Relative Entropy of States of Matrix Algebras. *Acta Math. Hung.* **1992**, *59*, 449–455.
10. Ohya, M.; Petz, D. *Quantum Entropy and Its Use*; Springer: New York, NY, USA, 1993; ISBN 0-387-54881-5.
11. Wilming, H.; Gallego, R.; Eisert, J. Axiomatic Characterization of the Quantum Relative Entropy and Free Energy. *Entropy* **2017**, *19*, 241.
12. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630.
13. Jaynes, E.T. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, UK, 2003.
14. Jaynes, E.T. Information Theory and Statistical Mechanics II. *Phys. Rev.* **1957**, *108*, 171–190.
15. Balian, R.; Vénéroni, M. Incomplete descriptions, relevant information, and entropy production in collision processes. *Ann. Phys.* **1987**, *174*, 229–224.
16. Balian, R.; Balazs, N.L. Equiprobability, inference and entropy in quantum theory. *Ann. Phys.* **1987**, *179*, 97–144.
17. Balian, R. Justification of the Maximum Entropy Criterion in Quantum Mechanics. In *Maximum Entropy and Bayesian Methods*; Skilling, J., Ed.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1989; pp. 123–129.
18. Balian, R. On the principles of quantum mechanics. *Am. J. Phys.* **1989**, *57*, 1019–1027.
19. Balian, R. Gain of information in a quantum measurement. *Eur. J. Phys.* **1989**, *10*, 208–213
20. Balian, R. Incomplete descriptions and relevant entropies. *Am. J. Phys.* **1999**, *67*, 1078–1090.
21. Blankenbecler, R.; Partovi, H. Uncertainty, Entropy, and the Statistical Mechanics of Microscopic Systems. *Phys. Rev. Lett.* **1985**, *54*, 373–376.
22. Blankenbecler, R.; Partovi, H. Quantum Density Matrix and Entropic Uncertainty. In Proceedings of the Fifth Workshop on Maximum Entropy and Bayesian Methods in Applied Statistics, Laramie, WY, USA, 5–8 August 1985.
23. Von Neumann, J. *Mathematische Grundlagen der Quantenmechanik*; Springer: Berlin, Germany, 1932. English Translation: *Mathematical Foundations of Quantum Mechanics*; Princeton University Press: Princeton, NY, USA, 1983.

24. Ali, S.A.; Cafaro, C.; Giffin, A.; Lupo, C.; Mancini, S. On a Differential Geometric Viewpoint of Jaynes' Maxent Method and its Quantum Extension. *AIP Conf. Proc.* **2012**, *1443*, 120–128.
25. Caticha, A. Entropic Dynamics: Quantum Mechanics from Entropy and Information Geometry. Available online: https://arxiv.org/abs/1711.02538 (accessed on 30 November 2017).
26. Reginatto, M.; Hall, M.J.W. Quantum-classical interactions and measurement: A consistent description using statistical ensembles on configuration space. *J. Phys. Conf. Ser.* **2009**, *174*, 012038.
27. Reginatto, M.; Hall, M.J.W. Information geometry, dynamics and discrete quantum mechanics. *AIP Conf. Proc.* **2013**, *1553*, 246–253.
28. Caves, C.; Fuchs, C.; Schack, R. Quantum probabilities as Bayesian probabilities. *Phys. Rev. A* **2002**, *65*, 022305.
29. Vanslette, K. The Quantum Bayes Rule and Generalizations from the Quantum Maximum Entropy Method. Available online: https://arxiv.org/abs/1710.10949 (accessed on 30 November 2017).
30. Schack, R.; Brun, T.; Caves, C. Quantum Bayes rule. *Phys. Rev. A* **2001**, *64*, 014305.
31. Korotkov, A. Continuous quantum measurement of a double dot. *Phys. Rev. B* **1999**, *60*, 5737–5742.
32. Korotkov, A. Selective quantum evolution of a qubit state due to continuous measurement. *Phys. Rev. B* **2000**, *63*, 115403.
33. Jordan, A.; Korotkov, A. Qubit feedback and control with kicked quantum nondemolition measurements: A quantum Bayesian analysis. *Phys. Rev. B* **2006**, *74*, 085307.
34. Hellmann, F.; Kamiński, W.; Kostecki, P. Quantum collapse rules from the maximum relative entropy principle. *New J. Phys.* **2016**, *18*, 013022.
35. Warmuth, M. A Bayes Rule for Density Matrices. In *Advances in Neural Information Processing Systems 18, Proceedings of the Neural Information Processing Systems Conference, Montréal, QC, Canada, 7–12 December 2005*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2015.
36. Warmuth, M.; Kuzmin, D. A Bayesian Probability Calculus for Density Matrices. *Mach. Learn.* **2010**, *78*, 63–101.
37. Tsuda, K. Machine learning with quantum relative entropy. *J. Phys. Conf. Ser.* **2009**, *143*, 012021.
38. Giffin, A.; Caticha, A. Updating Probabilities. Presented at the 26th International Workshop on Bayesian Inference and Maximum Entropy Methods (MaxEnt 2006), Paris, France, 8–13 July 2006.
39. Wang, Z.; Busemeyer, J.; Atmanspacher, H.; Pothos, E. The Potential of Using Quantum Theory to Build Models of Cognition. *Top. Cogn. Sci.* **2013**, *5*, 672–688.
40. Giffin, A. Maximum Entropy: The Universal Method for Inference. Ph.D. Thesis, University at Albany (SUNY), Albany, NY, USA, 2008.
41. Caticha, A. Toward an Informational Pragmatic Realism. *Minds Mach.* **2014**, *24*, 37–70.
42. Umegaki, H. Conditional expectation in an operator algebra, IV (entropy and information). *Ködai Math. Sem. Rep.* **1962**, *14*, 59–85.
43. Uhlmann, A. Relative entropy and the Wigner-Yanase-Dyson-Lieb concavity in an interpolation theory. *Commun. Math. Phys.* **1997**, *54*, 21–32.
44. Schumacher, B.; Westmoreland, M. Relative entropy in quantum information theory. In Proceedings of the AMS Special Session on Quantum Information and Computation, Washington, DC, USA, 19–21 January 2000.
45. Suzuki, M. On the Convergence of Exponential Operators—The Zassenhaus Formula, BCH Formula and Systematic Approximants. *Commun. Math. Phys.* **1977**, *57*, 193–200.
46. Horn, A. Eigenvalues of sums of Hermitian matrices. *Pac. J. Math.* **1962**, *12*, 225–241.
47. Bhatia, R. Linear Algebra to Quantum Cohomology: The Story of Alfred Horn's Inequalities. *Am. Math. Mon.* **2001**, *108*, 289–318.
48. Knutson, A.; Tao, T. Honeycombs and Sums of Hermitian Matrices. *Not. AMS* **2001**, *48*, 175–186.
49. Aczél, J. *Lectures on Functional Equations and Their Applications*; Academic Press Inc.: New York, NY, USA, 1966; Volume 19, pp. 31–44, 141–145, 213–217, 301–302, 347–349.
50. Darboux, G. Sur le théorème fondamental de la géométrie projective. *Math. Ann.* **1880**, *17*, 55–61.

© 2017 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).