Mach Learn (2010) 78: 63–101
DOI 10.1007/s10994-009-5133-7

Bayesian generalized probability calculus for density matrices

Manfred K. Warmuth · Dima Kuzmin

Received: 29 December 2008 / Revised: 12 May 2009 / Accepted: 2 June 2009 /
Published online: 23 July 2009

Abstract One of the main concepts in quantum physics is a density matrix, which is a symmetric positive definite matrix of trace one. Finite probability distributions can be seen as a special case when the density matrix is restricted to be diagonal.

We develop a probability calculus based on these more general distributions that includes definitions of joints, conditionals and formulas that relate these, including analogs of the Theorem of Total Probability and various Bayes rules for the calculation of posterior density matrices. The resulting calculus parallels the familiar "conventional" probability calculus and always retains the latter as a special case when all matrices are diagonal. We motivate both the conventional and the generalized Bayes rule with a minimum relative entropy principle, where the Kullbach-Leibler version gives the conventional Bayes rule and Umegaki's quantum relative entropy the new Bayes rule for density matrices.

Whereas the conventional Bayesian methods maintain uncertainty about which model has the highest data likelihood, the generalization maintains uncertainty about which unit direction has the largest variance. Surprisingly the bounds also generalize: as in the conventional setting we upper bound the negative log likelihood of the data by the negative log likelihood of the MAP estimator.

Keywords Generalized probability · Probability calculus · Density matrix · Quantum Bayes rule

Editor: Nicolo Cesa-Bianchi.

Supported by NSF grant IIS 0325363. Some of this work was done while visiting National ICT Australia in Canberra.

M.K. Warmuth (✉) · D. Kuzmin
UC California—Santa Cruz, Santa Cruz, CA 95064, USA
e-mail: manfred@cse.ucsc.edu

D. Kuzmin
e-mail: dimakuzmin@google.com

1 Introduction

The main notion of a "mixture state" used in quantum physics is a density matrix. States are unit vectors $\boldsymbol{u}$ ($\|\boldsymbol{u}\|_2 = 1$). For the sake of simplicity we assume in this paper that the underlying vector space is $\mathbb{R}^n$ (for finite n). Each state $\boldsymbol{u}$ (unit column vector in $\mathbb{R}^n$) is associated with a *dyad* $\boldsymbol{u}\boldsymbol{u}^\top \in \mathbb{R}^{n\times n}$. The dyad $\boldsymbol{u}\boldsymbol{u}^\top$ may be seen as a one-dimensional projection matrix which projects any vector onto direction $\boldsymbol{u}$. These dyads are the elementary events of a *generalized probability space*. It is useful to keep the corresponding "conventional" probability space in mind, which consists of a finite set of size n. The n points are the elementary events and a probability distribution may be seen as a mixture over the n points, i.e. such a probability distribution is specified by n real numbers that are bigger than zero and add to one. In the generalized case there are infinitely many dyads even if the dimension n is finite.[1]

Density matrices generalize finite probability distributions. They can be defined as mixtures of dyads $\boldsymbol{W} = \sum_i \omega_i \boldsymbol{w}_i \boldsymbol{w}_i^\top$ where the mixture coefficients ω_i are non-negative and sum to one. There may be an arbitrary number of components in the mixture. However, any n dimensional density matrix can be decomposed into a mixture of n orthogonal *eigendyads*, one for each eigenvector (see Fig. 1). Mixtures of dyads are always symmetric[2] and positive definite. A density matrix $\boldsymbol{W}$ can be depicted as an ellipse which is an affine transformation of the unit ball: $\{\boldsymbol{W}\boldsymbol{u} : \|\boldsymbol{u}\|_2 = 1\}$ (see Fig. 2). A dyad is a degenerate ellipse with a single axis in direction $\pm\boldsymbol{u}$ that has radius one (Fig. 1). Note that dyads have trace one:

$$\operatorname{tr}(\boldsymbol{u}\boldsymbol{u}^\top) = \operatorname{tr}(\boldsymbol{u}^\top\boldsymbol{u}) = \|\boldsymbol{u}\|_2^2 = 1.$$

Therefore, density matrices also have trace one.

A density matrix $\boldsymbol{W}$ assigns generalized probability $\operatorname{tr}(\boldsymbol{W}\boldsymbol{u}\boldsymbol{u}^\top)$ to each unit vector $\boldsymbol{u}$ and its associated dyad $\boldsymbol{u}\boldsymbol{u}^\top$ (see Fig. 2). This probability is independent of how $\boldsymbol{W}$ is expressed as a mixture and can be rewritten as $\boldsymbol{u}^\top\boldsymbol{W}\boldsymbol{u}$. Note that if the symmetric positive definite matrix $\boldsymbol{A}$ is viewed as a covariance matrix of a random cost vector $\boldsymbol{c}$, then $\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}$ is the variance of the cost along direction $\boldsymbol{u}$, i.e. the variance of $\boldsymbol{c}\cdot\boldsymbol{u}$.

If $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ is a probability vector, then the n-dimensional matrix $\operatorname{diag}(\boldsymbol{\alpha})$ with vector $\boldsymbol{\alpha}$ as its diagonal is a density matrix. Note that $\operatorname{diag}(\boldsymbol{\alpha}) = \sum_i \alpha_i \boldsymbol{e}_i \boldsymbol{e}_i^\top$, where the $\boldsymbol{e}_i$ are the standard basis vectors. Thus conventional probability distributions are special density matrices where the eigensystem is restricted to be the identity matrix. In this paper we develop a Bayesian style analysis for the case when the eigensystem is allowed to be arbitrary.

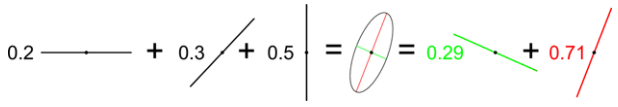


Fig. 1 Two different dyad mixtures that lead to the same density matrix: $0.2\binom{1}{0}(1\ 0) + 0.3\binom{\sqrt{2}/2}{\sqrt{2}/2} \times (\frac{\sqrt{2}}{2}\ \frac{\sqrt{2}}{2}) + 0.5\binom{0}{1}(0\ 1) = \binom{0.35\ 0.15}{0.15\ 0.65} = 0.29\binom{-0.92}{0.38}(-0.92\ 0.38) + 0.71\binom{0.38}{0.92}(0.38\ 0.92)$. Matrices are depicted as ellipses and dyads are degenerate single axis ellipses

[1] The machinery for infinite dimensional vector spaces is available. However, in this paper we start with the simplest finite dimensional setting.

[2] In quantum physics complex numbers are used instead of reals. In that case "symmetric" is replaced by "hermitian" and all our formulas hold for that case as well.

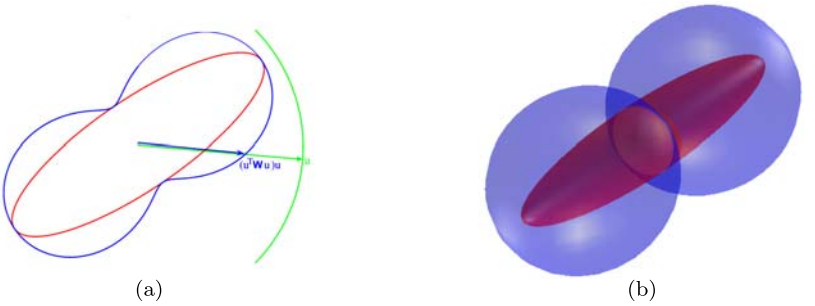


Fig. 2 (Color online) Figure (**a**) depicts a *red ellipse* $\{\boldsymbol{W}\boldsymbol{u} : \|\boldsymbol{u}\|_2 = 1\}$ for some density matrix $\boldsymbol{W}$. *The green curve* shows part of the unit ball. *The blue figure-eight* is a plot of the generalized probabilities in direction $\boldsymbol{u}$, i.e. $\mathrm{tr}(\boldsymbol{W}\boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{u}$. Figure (**b**) plots a 3-dimensional density matrix (*red ellipsoid*) and its associated generalized probability surface (*in blue*)

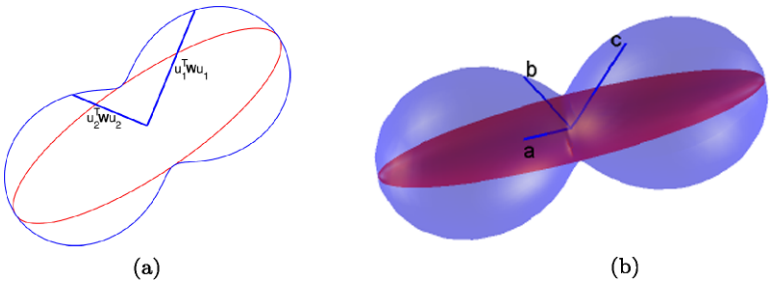


Fig. 3 (Color online) For a set of orthogonal directions $\boldsymbol{u}_i$ and a density matrix $\boldsymbol{W}$, the sum of generalized probabilities $\mathrm{tr}(\boldsymbol{W}\boldsymbol{u}_i\boldsymbol{u}_i^\top)$ over the set is one. Figure (**a**) shows this for 2-dimensional case: *the red ellipse* is a density matrix $\boldsymbol{W}$, *the blue figure-eight* is a plot of the generalized probability $\mathrm{tr}(\boldsymbol{W}\boldsymbol{u}\boldsymbol{u}^\top)$ around the circle, and for any two orthogonal vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, $\mathrm{tr}(\boldsymbol{W}\boldsymbol{u}_1\boldsymbol{u}_1^\top) + \mathrm{tr}(\boldsymbol{W}\boldsymbol{u}_2\boldsymbol{u}_2^\top) = 1$. Figure (**b**) shows the three-dimensional case: for any three orthogonal directions $\boldsymbol{u}_1$, $\boldsymbol{u}_2$ and $\boldsymbol{u}_3$, the probabilities a, b and c of the three associated dyads sum to one

Perhaps the simplest case to see that something unusual is going on is the uniform density matrix, i.e. $\frac{1}{n}$ times identity $\boldsymbol{I}$. This density matrix assigns probability $\frac{1}{n}$ to every unit vector, even though there are infinitely many of them. However, note that the sum of generalized probabilities of any set of n orthogonal dyads is $n\frac{1}{n} = 1$. As a matter of fact for any density matrix $\boldsymbol{W}$ and any set of n orthogonal directions $\boldsymbol{u}_i$, the total generalized probability is one (see Fig. 3)

$$\sum_{i=1}^{n} \mathrm{tr}(\boldsymbol{W}\boldsymbol{u}_i\boldsymbol{u}_i^\top) = \mathrm{tr}\Bigg(\boldsymbol{W}\underbrace{\sum_i \boldsymbol{u}_i\boldsymbol{u}_i^\top}_{\boldsymbol{I}}\Bigg) = \mathrm{tr}(\boldsymbol{W}) = 1. \tag{1.1}$$

This means that while in the conventional case probabilities are additive over the points in the set, in the generalized case probabilities are additive over orthogonal sets of dyads.

In this paper we use density matrices as generalized priors and develop a unifying Bayesian probability calculus for density matrices with rules for translating between joints and conditionals. All formulas retain the conventional case as the special case when the matrices are diagonal. In previous work (Warmuth 2005) we derived a generalized Bayes rule based on the minimum relative entropy principle, but no satisfactory probabilistic interpre-

tation was given for this rule. This Bayes rule fits nicely into our new calculus and we can interpret it using the notion of generalized probability introduced above.

For any fixed orthonormal system $\boldsymbol{u}_i$, one can use the dyads $\boldsymbol{u}_i\boldsymbol{u}_i^\top$ as elementary events of a conventional probability space. As already discussed, any density matrix can be seen as assigning conventional probabilities to these events that sum to one. Thus if the orthonormal system is fixed, generalized probability space is reduced to conventional probability space over the vectors in the chosen system. Our approach is fundamentally different in that we use density matrices to maintain uncertainty over all orthonormal systems. Our conditional density matrices are part of the probabilistic system specified by a generalized joint probability distribution. In particular, our conditioning method leads to generalizations of the theorem of total probability that involve density matrices.

In Tsuda et al. (2005) various on-line learning updates were generalized from vector parameters to matrix parameters. Following (Kivinen and Warmuth 1997), the updates were derived by minimizing the loss on the current instance plus a divergence to the last parameter. In this paper we use the same method for deriving a Bayes rule for density matrices, which becomes the foundation of our generalized probability calculus. When the parameters are probability vectors over the set of models, then the "conventional" Bayes rule can be derived using the relative entropy as the divergence (e.g. Zellner 1998; Kivinen and Warmuth 1999; Singh et al. 2003). Analogously, we now use the quantum relative entropy, introduced by Umegaki, to derive the generalized Bayes rule.

The new rule uses matrix logarithms and exponentials to avoid the fact that symmetric positive definite matrices are not closed under the matrix product. The rule is strikingly similar to the conventional Bayes rule and retains the latter as a special case when the matrices are diagonal. Various cancellations occur when the conventional Bayes rule is applied iteratively and as we shall see, similar cancellations happen with the new rule (see Sect. 9.2). The conventional Bayes rule may be seen as a soft maximum calculation and the new rule as a soft calculation of the eigenvector with the largest eigenvalue (see Figs. 7 and 8). In Figs. 9 and 10 we plot the projections of posterior onto the eigendirections of the fixed data-likelihood matrix $\boldsymbol{D}(y|\mathbb{M})$. The projection onto the eigendirection of the largest eigenvalue is a sigmoid like function.

The mathematics applied in this paper are most commonly used in quantum physics. For example, the assignment of generalized probabilities $\mathrm{tr}(\boldsymbol{W}\boldsymbol{u}\boldsymbol{u}^\top)$, can be seen as the outcome of a quantum measurement of a system in mixture state $\boldsymbol{W}$ being acted upon by a measurement apparatus described by the dyad $\boldsymbol{u}\boldsymbol{u}^\top$. It is tempting to call the new rule the "quantum Bayes rule". However, we currently do not have a quantum physical interpretation of this rule. In particular, the state collapse following a measurement does not explicitly appear in our calculus, also our Bayes rule can not be described as a unitary evolution of the prior state. The term "quantum Bayes rule" also has been claimed before in Schack et al. (2001), where they derive a rule that describes uncertainty information about unobserved quantum measurements of a composite system as a density matrix.

Our work is most closely related to a paper by Cerf and Adami (1999), where, in the context of quantum information theory, a formula was proposed for the conditional density matrix that uses the matrix exponential and matrix logarithm. This special formula appears in our calculus and is now put in a more general context. We hope to transfer many techniques developed in Bayesian Statistics based on the conventional Bayes rule to the context of generalized probabilities.

The use of the quantum relative entropy as a regularizer is reminiscent of the work on quantum state estimation (Olivares and Paris 2007; Bužek et al. 1999). Curiously, the update rules produced in that work do not make use of the matrix logarithm and matrix exponentials

as our new Bayes rule does. Conceivably our calculus can be applied to the state estimation problem.

The paper is organized as follows. Section 2 recalls the relevant matrix algebra facts. Section 3 introduces density matrices and generalized probability distributions and states Gleason's theorem that establishes an equivalence between them. Then, in Sect. 4 we introduce a generalization $\odot$ of the matrix product that is commutative and preserves positive definiteness. This $\odot$ operation is central to our calculus. Section 5 introduces generalized joint distributions. Section 6 discusses marginalizing the joints. Next, in Sect. 7 we give formulas for conditional density matrices. Section 8 presents generalizations of the Theorem of Total Probability. In Sect. 9 we present the founding piece of this work, the Bayes rule for density matrices, its derivation and various properties. We also discuss how the new Bayes rule for density matrices is in some sense the conventional Bayes rule in an optimally chosen eigensystem. Section 10 summarizes all the rules in our calculus and their justifications. In the conclusion section we discuss again how our new calculus relates to quantum physics and possible generalizations of it.

2 Facts on matrices and basic notation

In this paper generalized probability distributions, conditionals and data likelihoods are represented as symmetric positive definite matrices. We will now discuss some relevant matrix algebra facts.

The basic fact that we use a lot is the eigendecomposition of symmetric matrices:

$$\boldsymbol{S} = \boldsymbol{\mathcal{S}}\boldsymbol{\sigma}\boldsymbol{\mathcal{S}}^\top = \sum_{i=1}^{n} \sigma_i \boldsymbol{s}_i \boldsymbol{s}_i^\top.$$

This says that every such matrix can be written as a product of an orthogonal matrix of eigenvectors $\boldsymbol{\mathcal{S}}$ times a diagonal matrix of eigenvalues $\boldsymbol{\gamma}$ times $\boldsymbol{\mathcal{S}}^\top$. Alternatively it can be written as mixture of eigendyads formed from the eigenvectors where the eigenvalues act as mixture coefficients.

Any symmetric positive definite[3] matrix $\boldsymbol{C}$ can be seen as a covariance matrix of some random cost vector $\boldsymbol{c} \in \mathbb{R}^n$, i.e. $\boldsymbol{C} = \boldsymbol{E}((\boldsymbol{c} - \boldsymbol{E}(\boldsymbol{c})(\boldsymbol{c} - \boldsymbol{E}(\boldsymbol{c}))^\top)$. A covariance matrix $\boldsymbol{C}$ can be depicted as an ellipse $\{\boldsymbol{C}\boldsymbol{u} : \|\boldsymbol{u}\|_2 = 1\}$ centered at the origin, where the eigenvectors form the principal axes and the eigenvalues are the radii of the axes (see Fig. 2).

Note that a covariance matrix $\boldsymbol{C}$ is diagonal if the components of the cost vector are independent. The variance of the cost vector $\boldsymbol{c}$ along a vector $\boldsymbol{u}$, that is the variance of the dot product $\boldsymbol{c}^\top \boldsymbol{u}$, has the form

$$\begin{aligned}
\mathbb{V}(\boldsymbol{c}^\top \boldsymbol{u}) &= \boldsymbol{E}\left((\boldsymbol{c}^\top \boldsymbol{u} - \boldsymbol{E}(\boldsymbol{c}^\top \boldsymbol{u}))^2\right) \\
&= \boldsymbol{E}\left(((\boldsymbol{c}^\top - \boldsymbol{E}(\boldsymbol{c}^\top))\boldsymbol{u})^\top((\boldsymbol{c}^\top - \boldsymbol{E}(\boldsymbol{c}^\top))\boldsymbol{u})\right) \\
&= \boldsymbol{E}\left(\boldsymbol{u}^\top(\boldsymbol{c} - \boldsymbol{E}(\boldsymbol{c}))(\boldsymbol{c} - \boldsymbol{E}(\boldsymbol{c}))^\top)\boldsymbol{u}\right) \\
&= \boldsymbol{u}^\top \boldsymbol{C}\boldsymbol{u}.
\end{aligned}$$

[3] We use the convention that positive definite matrices have non-negative eigenvalues and *strictly* positive definite matrices have positive eigenvalues.

The variance along an eigenvector of the covariance matrix is the corresponding eigenvalue. Using this interpretation, the matrix $\boldsymbol{C}$ may be seen as a mapping from the unit ball to $\mathbb{R}_{\geq 0}$, i.e. unit vector $\boldsymbol{u}$ is mapped to $\boldsymbol{u}^\top \boldsymbol{C}\boldsymbol{u}$. Figure 2 depicts the resulting figure-8-like plots in 2 and 3 dimensions. A second interpretation of the scalar $\boldsymbol{u}^\top \boldsymbol{C}\boldsymbol{u}$ is the square length of $\boldsymbol{u}$ w.r.t. the basis $\sqrt{\boldsymbol{C}}$, that is $\boldsymbol{u}^\top \boldsymbol{C}\boldsymbol{u} = \boldsymbol{u}^\top \sqrt{\boldsymbol{C}}\sqrt{\boldsymbol{C}}\boldsymbol{u} = \|\sqrt{\boldsymbol{C}}\boldsymbol{u}\|_2^2$.

The trace $\operatorname{tr}(\boldsymbol{E})$ of an arbitrary square matrix $\boldsymbol{E}$ is the sum of its diagonal elements $\boldsymbol{E}_{ii}$. It is a linear operator. Recall that $\operatorname{tr}(\boldsymbol{E}\boldsymbol{F}) = \operatorname{tr}(\boldsymbol{F}\boldsymbol{E})$ for any matrices $\boldsymbol{E} \in \mathbb{R}^{n\times m}$, $\boldsymbol{F} \in \mathbb{R}^{m\times n}$. Also, for symmetric square matrices, $\operatorname{tr}(\boldsymbol{S}\boldsymbol{T}) = \sum_{i,j} S_{ij}T_{ij}$, thus trace can be seen as a dot product between matrices. The trace has a useful cycling property: for arbitrary matrices $\boldsymbol{E}, \boldsymbol{F}, \boldsymbol{G}$ with compatible dimensions $\operatorname{tr}(\boldsymbol{E}\boldsymbol{F}\boldsymbol{G}) = \operatorname{tr}(\boldsymbol{F}\boldsymbol{G}\boldsymbol{E}) = \operatorname{tr}(\boldsymbol{G}\boldsymbol{E}\boldsymbol{F})$. From this follows that trace is *rotation invariant* in the sense that for any orthogonal matrix $\boldsymbol{\mathcal{U}}$, $\operatorname{tr}(\boldsymbol{\mathcal{U}}\boldsymbol{E}\boldsymbol{\mathcal{U}}^\top) = \operatorname{tr}(\boldsymbol{\mathcal{U}}^\top\boldsymbol{\mathcal{U}}\boldsymbol{E}) = \operatorname{tr}(\boldsymbol{E})$. If $\boldsymbol{S}$ is symmetric, setting $\boldsymbol{\mathcal{U}}$ to be the eigensystem of $\boldsymbol{S}$ results in the observation that trace is equal to the sum of eigenvalues of a matrix. Also, for any orthogonal system[4] $\boldsymbol{u}_i$,

$$\operatorname{tr}(\boldsymbol{S}) = \operatorname{tr}\Bigg(\underbrace{\sum_{i=1}^{n} \boldsymbol{u}_i\boldsymbol{u}_i^\top}_{\boldsymbol{I}}\,\boldsymbol{S}\Bigg) = \sum_{i=1}^{n} \boldsymbol{u}_i^\top \boldsymbol{S}\boldsymbol{u}_i.$$

Therefore if $\boldsymbol{S}$ is symmetric positive definite, then $\operatorname{tr}(\boldsymbol{S})$ is the total variance along any set of orthogonal directions. Recall that density matrices have trace one and therefore in this case this total variance is always one (see Fig. 3).

The matrix exponential $\exp(\boldsymbol{S})$ of the symmetric matrix $\boldsymbol{S} = \sum_i \sigma_i \boldsymbol{s}_i\boldsymbol{s}_i^\top$ is computed by exponentiating the eigenvalues and leaving the eigenvectors unchanged: $\exp(\boldsymbol{S}) = \sum_i \exp(\sigma_i)\boldsymbol{s}_i\boldsymbol{s}_i^\top$. The matrix logarithm $\log(\boldsymbol{A})$ is defined similarly but now $\boldsymbol{A}$ must be strictly positive definite. Clearly, the two functions are inverses of each other. It is important to remember that $\exp(\boldsymbol{S}+\boldsymbol{T}) = \exp(\boldsymbol{S})\exp(\boldsymbol{T})$ only holds if $\boldsymbol{S}$ and $\boldsymbol{T}$ commute i.e. $\boldsymbol{S}\boldsymbol{T} = \boldsymbol{T}\boldsymbol{S}$.[5] However, the following trace inequality, known as the Golden-Thompson inequality[6] (Bhatia 1997), always holds:

$$\operatorname{tr}(\exp(\boldsymbol{S})\exp(\boldsymbol{T})) \geq \operatorname{tr}(\exp(\boldsymbol{S}+\boldsymbol{T})) \quad \text{for symmetric } \boldsymbol{S} \text{ and } \boldsymbol{T}, \tag{2.1}$$

where equality holds iff both symmetric matrices commute.

3 Generalized probability distributions and density matrices

In quantum physics a dyad $\boldsymbol{u}\boldsymbol{u}^\top$ represents a pure state and density matrices are mixture states. As we shall see density matrices can be interpreted as generalized probability distributions over the set of dyads. Note that in this paper we want to address the statistics community and use linear algebra notation instead of Dirac notation. Any probability vector $(P(M_i))$ can be represented as a diagonal matrix $\operatorname{diag}(P(M_i)) = \sum_i P(M_i)\boldsymbol{e}_i\boldsymbol{e}_i^\top$, where $\boldsymbol{e}_i$ denotes the ith standard basis vector. This means that conventional probability vectors are special density matrices where the eigenvectors are fixed to be the standard basis vectors.

[4] A set of unit vectors $\boldsymbol{u}_i$ is orthogonal iff $\sum_i \boldsymbol{u}_i\boldsymbol{u}_i^\top = \boldsymbol{I}$.

[5] This occurs iff the two symmetric matrices have the same eigensystem.

[6] Note that the Golden-Thompson inequality does not generalize to three matrices, i.e. there exist symmetric $\boldsymbol{S}, \boldsymbol{T}, \boldsymbol{U}$, s.t. $\operatorname{tr}(\exp(\boldsymbol{S})\exp(\boldsymbol{T})\exp(\boldsymbol{U})) \not\geq \operatorname{tr}(\exp(\boldsymbol{S}+\boldsymbol{T}+\boldsymbol{U}))$.

For the sake of simplicity we assume that our vector space is $\mathbb{R}^n$. However, everything discussed in this section holds for separable finite or infinite dimensional real and complex Hilbert spaces.

A function $\mu(\boldsymbol{u})$ from unit vectors $\boldsymbol{u}$ in $\mathbb{R}^n$ to $\mathbb{R}$ is called a *generalized probability distributions* if the following two conditions hold:

- $\forall \boldsymbol{u}, 0 \leq \mu(\boldsymbol{u}) \leq 1$.
- If $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ form an orthonormal system for $\mathbb{R}^n$, then $\sum \mu(\boldsymbol{u}_i) = 1$.

Gleason's Theorem states that there is a one-to-one correspondence between generalized probability distributions and density matrices[7] in $\mathbb{R}^{n\times n}$:

Theorem 1 (Gleason 1957) *Let $n \geq 3$.*[8] *Then any generalized probability distribution μ on $\mathbb{R}^n$ has the form $\mu(\boldsymbol{u}) = \operatorname{tr}(\boldsymbol{W}\boldsymbol{u}\boldsymbol{u}^\top)$, for a uniquely defined density matrix $\boldsymbol{W}$.*

It is easy to see that every density matrix defines a generalized probability distribution. The other direction, is highly non-trivial.[9] As discussed in the introduction, the dyads $\boldsymbol{u}\boldsymbol{u}^\top$ function as elementary events. One may ask what corresponds to arbitrary events and how probabilities are defined for them. In the conventional case, an event is a subset of the domain which can be represented as a vector in $\{0,1\}^n$. In the generalized setting, an event is a symmetric positive definite matrix $\boldsymbol{P}$ with eigenvalues in $\{0,1\}$. Each such matrix $\boldsymbol{P}$ with eigendecomposition $\sum_{i=1}^k \boldsymbol{p}_i \boldsymbol{p}_i^\top$ is a projection matrix for a subspace of $\mathbb{R}^n$ and its probability w.r.t. a distribution $\boldsymbol{W}$ is defined as the sum of the probabilities of the elementary events $\boldsymbol{p}_i \boldsymbol{p}_i^\top$ comprising $\boldsymbol{P}$:

$$\operatorname{tr}(\boldsymbol{W}\boldsymbol{P}) = \sum_{i=1}^{k} \operatorname{tr}(\boldsymbol{W}\boldsymbol{p}_i\boldsymbol{p}_i^\top).$$

Interpreting $\boldsymbol{P}$ as a covariance matrix of some random variable, we can also expand $\boldsymbol{W}$ and sum the variance along its eigendirections $\boldsymbol{w}_i$ weighted by the eigenvalues ω_i which are probabilities:

$$\operatorname{tr}(\boldsymbol{W}\boldsymbol{P}) = \operatorname{tr}\left(\sum_{i=1}^{n} \omega_i \boldsymbol{w}_i \boldsymbol{w}_i^\top \boldsymbol{P}\right) = \sum_{i=1}^{n} \underbrace{\overbrace{\omega_i}^{\text{probability}} \overbrace{\boldsymbol{w}_i^\top \boldsymbol{P} \boldsymbol{w}_i}^{\text{variance}}}_{\text{expected variance}}. \tag{3.1}$$

Random variables are defined in an analogous way. In the conventional case a random variable associates a real value with each point. Now a random variable is an arbitrary symmetric matrix $\boldsymbol{S}$. Such matrices have arbitrary real numbers as their eigenvalues and trace $\operatorname{tr}(\boldsymbol{W}\boldsymbol{S})$ when $\boldsymbol{S}$ is expanded becomes the expectation of the random variable w.r.t. den-

[7] The core of the original proof of Gleason's Theorem was for $\mathbb{R}^3$ (Gleason 1957), and he then generalized the proof to separable real and complex Hilbert spaces of dimension $n \geq 3$.

[8] A slightly different version of this theorem that is based on "effects" instead of dyads holds for dimension 2 as well (Caves et al. 2004).

[9] However, if dyads are replaced by "effects" then the proofs are much simpler (Caves et al. 2004).

sity $\boldsymbol{W}$:

$$\operatorname{tr}(\boldsymbol{W}\boldsymbol{S}) = \operatorname{tr}\left(\boldsymbol{W}\sum_i \sigma_i \boldsymbol{s}_i \boldsymbol{s}_i^\top\right) = \underbrace{\sum_i \overbrace{\sigma_i}^{\text{outcome}} \overbrace{\boldsymbol{s}_i^\top \boldsymbol{W}\boldsymbol{s}_i}^{\text{probability}}}_{\text{expected outcome}}. \tag{3.2}$$

As discussed before, the conventional case of the expectation calculation is always retained as a special case when all the matrices are diagonal (i.e. fixed eigensystem $\boldsymbol{I}$). In quantum physics the expectation calculation $\operatorname{tr}(\boldsymbol{W}\boldsymbol{S})$ has the following interpretation: an instrument is represented by a hermitian matrix $\boldsymbol{S}$ and $\operatorname{tr}(\boldsymbol{W}\boldsymbol{S})$ is the expected value of a *quantum measurement* of the mixed state $\boldsymbol{W}$ with instrument $\boldsymbol{S}$. The eigenvalues σ_i of the instrument represent the possible numerical measurement outcomes. Each one of those outcomes is observed with probability $\boldsymbol{s}_i^\top \boldsymbol{W}\boldsymbol{s}_i$, where $\boldsymbol{s}_i$ is the associated eigenvector of the instrument matrix $\boldsymbol{S}$.

In real quantum systems the measurement causes the mixtures state $\boldsymbol{W}$ to *collapse* into one of the orthogonal states $\{\boldsymbol{s}_1\boldsymbol{s}_1^\top, \ldots, \boldsymbol{s}_n, \boldsymbol{s}_n^\top\}$: the successor state is $\boldsymbol{s}_i\boldsymbol{s}_i^\top$ with probability $\boldsymbol{s}_i^\top \boldsymbol{W}\boldsymbol{s}_i$:

$$\boldsymbol{W} \overset{\substack{\text{measurement}\\ \text{collapse}}}{\longrightarrow} \underbrace{\sum_i \overbrace{\boldsymbol{s}_i^\top \boldsymbol{W}\boldsymbol{s}_i}^{\text{probability}}\, \boldsymbol{s}_i\boldsymbol{s}_i^\top}_{\text{expected state}}.$$

As we shall see, the expected measurement calculations play an important part in our calculus. However our update rules for density matrices (such as our Bayes rule) do not explicitly include a collapse in the above sense.

Note that some of the equations above hold for arbitrary decompositions into a linear combination of dyads of any size. For example (3.1), holds for any decomposition $\boldsymbol{W} = \sum_i \omega_i \boldsymbol{w}_i \boldsymbol{w}_i^\top$, i.e. the ω_i may be negative, the $\boldsymbol{w}_i$ may be non-orthogonal, and the size of the decomposition may be larger than n. If the ω_i are non-negative, then they form a probability vector. Similarly, (3.2) also holds for any decomposition $\boldsymbol{S} = \sum_i \sigma_i \boldsymbol{s}_i \boldsymbol{s}_i^\top$. However, quantum measurements are always based on an orthogonal system. Furthermore, orthogonal systems are special in that the orthogonal decomposition of a density matrix $\boldsymbol{W} = \sum_i \omega_i \boldsymbol{w}_i \boldsymbol{w}_i^\top$ attains the minimum of the entropy $\sum_i -\omega_i \ln \omega_i$ over all possible decompositions of $\boldsymbol{W}$ (inequality (11.86) in Nielsen and Chuang 2000).

A question that naturally arises is whether we can model the generalized probability distributions defined above with a conventional probability space. In other words, is there a conventional probability space and two mappings: one that maps density matrices to conventional probability distributions and the other mapping dyads to events of this probability space. The requirement on these two mappings is that the conventional probability calculations using the images of density matrices and dyads under these mappings satisfy the definition of the generalized probability distributions given above. Essentially, it is known that conventional probability spaces cannot satisfactorily model generalized probabilities, but the details are rather involved. This topic has received considerable attention in the quantum physics community and we refer readers to Holevo (2001) for an extended discussion of impossibility results. Here we only give one simple attempt to model density matrices with a conventional probability space and show that the two natural mappings fail to satisfy the requirements.

A natural interpretation of a density matrix is to view it as a parameterized density over the unit sphere. We claim that if $\mu(\boldsymbol{u})$ is the uniform density on the sphere, then for any symmetric positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ of trace n, $\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u}\mu(\boldsymbol{u})$ is also a conventional probability density on the sphere:

$$\begin{aligned}\int \boldsymbol{u}^\top \overbrace{\sum_i \alpha_i \boldsymbol{a}_i \boldsymbol{a}_i^\top}^{\boldsymbol{A}} \boldsymbol{u}\mu(\boldsymbol{u})d\boldsymbol{u} &= \sum_i \alpha_i \int (\boldsymbol{u}^\top \boldsymbol{a}_i)^2 \mu(\boldsymbol{u})d\boldsymbol{u} \\ &= \sum_i \alpha_i \int (\boldsymbol{u}^\top (1,0,\ldots,0))^2 \mu(\boldsymbol{u})d\boldsymbol{u} \\ &= \operatorname{tr}(\boldsymbol{A}) \int (u_1)^2 \mu(\boldsymbol{u})d\boldsymbol{u} \\ &= \underbrace{\frac{\operatorname{tr}(\boldsymbol{A})}{n}}_{1} \int \underbrace{\sum_i u_i^2}_{1}\ \mu(\boldsymbol{u})d\boldsymbol{u} = 1.\end{aligned}$$

In the second equality we used the fact that $\mu(\boldsymbol{u})d\boldsymbol{u}$ is uniform and therefore the integral of $(\boldsymbol{u}^\top \boldsymbol{a}_i)^2$ is the same as the integral of the squared dot product of $\boldsymbol{u}$ with the unit vector $(1,0,\ldots,0)^\top$. In the last equality we again used symmetry: The integral of the square of one component equals the average of the squares of all components.

We modeled density matrices as conventional probability densities over the sphere. Now the natural mapping from dyads to events in the conventional probability space (the sphere) maps $\boldsymbol{u}\boldsymbol{u}^\top$ to $\{\boldsymbol{u}, -\boldsymbol{u}\}$. However the probability of the latter sets of size 2 is zero with respect to the conventional probabilities densities we defined on the sphere. In particular the probability on any n orthogonal dyads does not sum to one.

4 Commutative matrix product operation

It is well known that the product of two symmetric positive definite matrices might be neither symmetric nor positive definite (see Fig. 4). In this section we define a commutative "product" operation between symmetric positive definite matrices that does result in a symmetric positive definite matrix. Our first definition of this operation requires the two matrices to be strictly positive definite. We then extend the definition to arbitrary symmetric positive definite matrices and prove many properties of this product.

For two symmetric and strictly positive definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we first define the $\odot$ as:

$$\boldsymbol{A} \odot \boldsymbol{B} := \exp\Big(\log \overbrace{\overbrace{\boldsymbol{A}}^{\text{sym.pos.def.}}}^{\text{sym.}} + \log \overbrace{\overbrace{\boldsymbol{B}}^{\text{sym.pos.def.}}}^{\text{sym.}}\Big), \tag{4.1}$$

where here the exponential and logarithm are matrix functions. The matrix log of both matrices produces symmetric matrices which are closed under addition and the matrix exponential of the sum returns a symmetric positive definite matrix. See Fig. 4 for a comparison of matrix product and $\odot$.

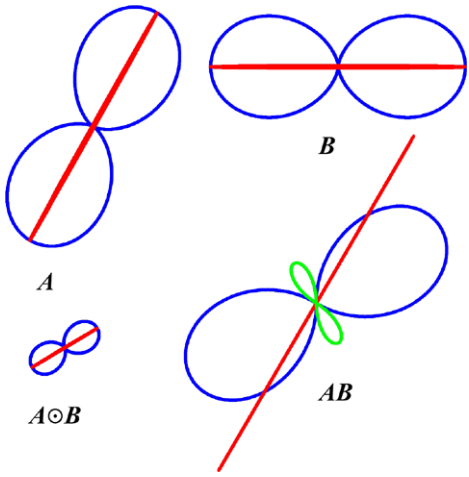


Fig. 4 The matrix product of two positive definite matrices does not preserve positive definiteness. For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ we plot their ellipses $\boldsymbol{Au}$, $\boldsymbol{Bu}$ and figure eights $\mathrm{tr}(\boldsymbol{Auu}^\top)\boldsymbol{u}$, $\mathrm{tr}(\boldsymbol{Buu}^\top)\boldsymbol{u}$ (for unit $\boldsymbol{u}$). Both ellipses are very thin, i.e. the ratio between the two eigenvalues of each matrix is 100. We also plot the ellipse $\boldsymbol{ABu}$ and the curve $\mathrm{tr}(\boldsymbol{ABuu}^\top)\boldsymbol{u}$. The latter curve consists of two figure eights, the larger one constitutes the part where the trace is positive and the smaller and skinnier one is the part where the trace is negative. This means that $\boldsymbol{AB}$ is not positive definite any more. The product is also not symmetric because the min/max value of $\mathrm{tr}(\boldsymbol{ABuu}^\top)$ does not correspond to the axes of the ellipse. Finally, the corresponding plots for $\boldsymbol{A} \odot \boldsymbol{B}$ indicate that this matrix is symmetric and positive definite

Note that we expressed the operation $\odot$ between symmetric strictly positive definite matrices as a $+$ operation between symmetric matrices. Similarly, for any two arbitrary symmetric matrices $\boldsymbol{S}$ and $\boldsymbol{T}$,

$$\boldsymbol{S} + \boldsymbol{T} = \log(\exp(\boldsymbol{S}) \odot \exp(\boldsymbol{T})).$$

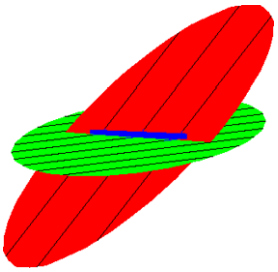
The operation $\odot$ was used in Alexa (2002) to define a "product" between two linear transformations that is commutative. In this paper we use $\odot$ to define conditional density matrices and generalizations of the Bayes rule. A similar path was followed by Cerf and Adami (1999) for defining conditional density matrices of composite systems. We also give a motivation for the operation based on the minimum relative entropy principle (as was done in the conference paper Warmuth 2005) and our probability calculus includes the formula of Cerf and Adami (1999) for composite systems as a special case.

Note that the formula for $\odot$ in (4.1) is not defined if some of the eigenvalues of $\boldsymbol{A}$ or $\boldsymbol{B}$ are zero. We now rewrite the operation using the Lie-Trotter formula and then extend it to arbitrary positive definite matrices. The Lie-Trotter formula (see e.g. Bhatia 1997) is the following equation:

$$\exp(\boldsymbol{E} + \boldsymbol{F}) = \lim_{n\to\infty} (\exp(\boldsymbol{E}/n)\exp(\boldsymbol{F}/n))^n, \quad \text{any square matrices } \boldsymbol{E}, \boldsymbol{F}.$$

By choosing $\boldsymbol{E} = \log \boldsymbol{A}$ and $\boldsymbol{F} = \log \boldsymbol{B}$, for symmetric and strictly positive definite $\boldsymbol{A}$ and $\boldsymbol{B}$, we obtain:

$$\exp(\log \boldsymbol{A} + \log \boldsymbol{B}) = \lim_{n\to\infty} (\boldsymbol{A}^{1/n}\boldsymbol{B}^{1/n})^n.$$



diag($\boldsymbol{A}$)	diag($\boldsymbol{B}$)	diag($\boldsymbol{AB}$)
0	0	0
a	0	0
0	b	0
a	b	ab

Fig. 5 (Color online) When the ellipses $\boldsymbol{A}$ and $\boldsymbol{B}$ don't have the same span, then $\boldsymbol{A} \odot \boldsymbol{B}$ lies in the intersection of both spans. In the depicted case the intersection is a degenerate ellipse of dimension one (*blue line*). This generalizes the following intersection property of the matrix product when $\boldsymbol{A}$ and $\boldsymbol{B}$ are both diagonal (here of dimension four): $(\boldsymbol{AB})_{i,i} \neq 0$ iff $\boldsymbol{A}_{i,i} \neq 0$ and $\boldsymbol{B}_{i,i} \neq 0$

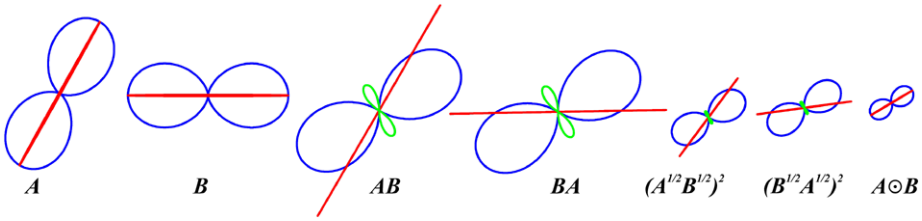


Fig. 6 The behavior of the limit formula for $\odot$ operation. We can see that the additional figure eights indicating negative definiteness are smaller for $(\boldsymbol{A}^{1/2}\boldsymbol{B}^{1/2})^2$ than for $\boldsymbol{AB}$. As n increases, the additional figure eights shrink further and $\lim_{n\to\infty}(\boldsymbol{A}^{1/n}\boldsymbol{B}^{1/n})^n = \boldsymbol{A} \odot \boldsymbol{B}$ becomes positive definite. Also, $\boldsymbol{AB}$ and $\boldsymbol{BA}$ are fairly different from one another. The matrices $(\boldsymbol{A}^{1/2}\boldsymbol{B}^{1/2})^2$ and $(\boldsymbol{B}^{1/2}\boldsymbol{A}^{1/2})^2$ are already more similar and the difference between the two multiplication orders decreases with n until in the limit $\boldsymbol{A} \odot \boldsymbol{B} = \boldsymbol{B} \odot \boldsymbol{A}$

As n increases, $(\boldsymbol{A}^{1/n}\boldsymbol{B}^{1/n})^n$ gets closer and closer to being positive definite and symmetric. The first couple iterations of the limit formula are plotted in Fig. 6. See Alexa (2002) for additional plots. Notice that the limit is defined even when $\boldsymbol{A}$ and $\boldsymbol{B}$ have zero eigenvalues. We therefore extend the definition of $\odot$ to arbitrary symmetric positive definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$\boldsymbol{A} \odot \boldsymbol{B} := \lim_{n\to\infty} (\boldsymbol{A}^{1/n}\boldsymbol{B}^{1/n})^n. \tag{4.2}$$

From now on we use the above exended definition of $\odot$. Numerous properties of this operation are given below.

Theorem 2 *For any symmetric positive definite matrices* $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ *the following holds*:

OP1 *Intersection property*:

$$\text{range}(\boldsymbol{A} \odot \boldsymbol{B}) = \text{range}(\boldsymbol{A}) \cap \text{range}(\boldsymbol{B}).$$

where the range of a matrix is the linear subspace spanned by the columns of the matrix. This property generalizes the intersection properties for products of diagonal matrices (*which model conventional probability distributions*): *the product of two diagonal matrices with the characteristic vectors of two subsets as diagonals gives a diagonal matrix formed from the characteristic vector of the intersection* (*see Fig.* 5).

OP2 *Let* $\boldsymbol{R_A}$ *be a matrix whose columns form an orthonormal basis for the range of* $\boldsymbol{A}$, *i.e.* $\boldsymbol{R_A} \in \mathbb{R}^{n\times k}$ *and* $\boldsymbol{R_A^\top R_A} = \boldsymbol{I}_k$, *where* k *is the dimensionality of the range of* $\boldsymbol{A}$. *Define* $\boldsymbol{R_B}$ *analogously. In a similar fashion* $\boldsymbol{R_{A\cap B}}$ *will contain the basis for the intersection of ranges. Let* $\log^+$ *denote the modified matrix logarithm that takes the log of non-zero eigenvalues but leaves the zero eigenvalues unchanged. This operation can be also defined by the following formula:*[10]

$$\log^+ \boldsymbol{A} = \boldsymbol{R_A} \log(\boldsymbol{R_A^\top A R_A}) \boldsymbol{R_A^\top}. \tag{4.3}$$

With this notation, $\odot$ *can be written as*

$$\boldsymbol{A} \odot \boldsymbol{B} = \boldsymbol{R_{A\cap B}} \exp(\boldsymbol{R_{A\cap B}^\top}(\log^+ \boldsymbol{A} + \log^+ \boldsymbol{B}) \boldsymbol{R_{A\cap B}}) \boldsymbol{R_{A\cap B}^\top}. \tag{4.4}$$

OP3 $\boldsymbol{A} \odot \boldsymbol{B} = \boldsymbol{AB}$ *if* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *commute.*

OP4 $\odot$ *is commutative, i.e.* $\boldsymbol{A} \odot \boldsymbol{B} = \boldsymbol{B} \odot \boldsymbol{A}$.

OP5 *The identity matrix is the neutral element, i.e.* $\boldsymbol{A} \odot \boldsymbol{I} = \boldsymbol{A}$.

OP6 $(c\boldsymbol{A}) \odot \boldsymbol{B} = c(\boldsymbol{A} \odot \boldsymbol{B})$, *for any scalar* $c > 0$.

OP7 $\boldsymbol{A} \odot \boldsymbol{A}^{-1} = \boldsymbol{I}$ *for invertible* $\boldsymbol{A}$. *Also,* $\boldsymbol{A} \odot \boldsymbol{A}^{+} = \boldsymbol{P_A}$, *where* $\boldsymbol{A}^+$ *denotes the pseudoinverse and* $\boldsymbol{P_A}$ *is the projection matrix*[11] *for* $\operatorname{range}(\boldsymbol{A})$.

OP8 $\odot$ *is associative, i.e.* $(\boldsymbol{A} \odot \boldsymbol{B}) \odot \boldsymbol{C} = \boldsymbol{A} \odot (\boldsymbol{B} \odot \boldsymbol{C})$.

OP9 *Monotonic convergence of the limit defining* $\odot$:

$$\forall n \geq 1 : \operatorname{tr}(\boldsymbol{A}^{1/(n+1)} \boldsymbol{B}^{1/(n+1)})^{n+1} \leq \operatorname{tr}(\boldsymbol{A}^{1/n} \boldsymbol{B}^{1/n})^n.$$

OP10 $\operatorname{tr}(\boldsymbol{A} \odot \boldsymbol{B}) \leq \operatorname{tr}(\boldsymbol{AB})$, *where equality holds iff* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *commute. In particular, for any unit* $\boldsymbol{u}$ $\operatorname{tr}(\boldsymbol{A} \odot \boldsymbol{uu}^\top) = \operatorname{tr}(\boldsymbol{Auu}^\top)$ *iff* $\boldsymbol{u}$ *is an eigenvector of* $\boldsymbol{A}$.

OP11 *For any unit direction* $\boldsymbol{u} \in \operatorname{range}(\boldsymbol{A})$, $\boldsymbol{A} \odot \boldsymbol{uu}^\top = e^{\boldsymbol{u}^\top (\log^+ \boldsymbol{A})\boldsymbol{u}} \boldsymbol{uu}^\top$.

OP12 *For any unit direction* $\boldsymbol{u}$ *and eigendecomposition* $\sum_i \alpha_i \boldsymbol{a}_i \boldsymbol{a}_i^\top$ *of a strictly positive definite matrix* $\boldsymbol{A}$,

$$\operatorname{tr}(\boldsymbol{Auu}^\top) = \sum_i (\boldsymbol{u}^\top \boldsymbol{a}_i)^2 \alpha_i, \quad \textit{and} \quad \operatorname{tr}(\boldsymbol{A} \odot \boldsymbol{uu}^\top) = \prod_i \alpha_i^{(\boldsymbol{u}^\top \boldsymbol{a}_i)^2},$$

i.e. the matrix product corresponds to an arithmetic average and the $\odot$ *product to a geometric average of the eigenvalues of* $\boldsymbol{A}$.

OP13 $\det(\boldsymbol{A} \odot \boldsymbol{B}) = \det(\boldsymbol{A}) \det(\boldsymbol{B})$, *which is the same as for the normal matrix product.*

OP14 *For any orthogonal system* $\boldsymbol{u}_i$, *we have* $\prod_i \operatorname{tr}(\boldsymbol{A} \odot \boldsymbol{u}_i \boldsymbol{u}_i^\top) = \det(\boldsymbol{A})$.

OP15 *For any unit direction* $\boldsymbol{u}$, $\operatorname{tr}((\boldsymbol{A} \odot \boldsymbol{B}) \odot \boldsymbol{uu}^\top) = \operatorname{tr}(\boldsymbol{A} \odot \boldsymbol{uu}^\top) \operatorname{tr}(\boldsymbol{B} \odot \boldsymbol{uu}^\top)$.

OP16 *For any unit direction* $\boldsymbol{u} \in \operatorname{range}\boldsymbol{A}$, $\operatorname{tr}(\boldsymbol{A}^+ \odot \boldsymbol{uu}^\top) = \frac{1}{\operatorname{tr}(\boldsymbol{A} \odot \boldsymbol{uu}^\top)}$, *where* $\boldsymbol{A}^+$ *denotes the pseudoinverse.*

Proof Properties OP1 and OP2 follow from results in Kato (1978) or Theorem 1.2 of Simon (1979). Here we only prove that $\operatorname{range}(\boldsymbol{A} \odot \boldsymbol{B}) \subseteq \operatorname{range}(\boldsymbol{A}) \cap \operatorname{range}(\boldsymbol{B})$. We can split the limit

[10] Note that when the rank k of $\boldsymbol{A}$ is zero, then one still can define the projections in a consistent manner. In this case $\boldsymbol{R_A}$ is of dimension $n \times 0$, and the matrices $\boldsymbol{R_A^\top R_A}$ and $\log(\boldsymbol{R_A^\top E R_A})$ are of dimension 0×0 for any $\boldsymbol{E} \in \mathbb{R}^{n\times n}$. Also it is natural to define $\boldsymbol{R_A E R_A^\top}$ as the $n \times n$ zero matrix $\boldsymbol{0}$. With this definition, the r.h.s. of (4.3) is $\boldsymbol{0}$ when $\boldsymbol{A}$ is $\boldsymbol{0}$.

[11] Note that $\boldsymbol{P_A} = \boldsymbol{R_A R_A^\top}$.

defining $\odot$ as follows:

$$\boldsymbol{A} \odot \boldsymbol{B} = \lim_{n\to\infty} (\boldsymbol{A}^{1/n}\boldsymbol{B}^{1/n})^n = \lim_{n\to\infty} \boldsymbol{A}^{1/n} \lim_{n\to\infty} \boldsymbol{B}^{1/n}(\boldsymbol{A}^{1/n}\boldsymbol{B}^{1/n})^{n-1}. \tag{4.5}$$

Here we used the property that $\lim \boldsymbol{E}_n\boldsymbol{F}_n = \lim \boldsymbol{E}_n \lim \boldsymbol{F}_n$ if all the limits exist. This follows from the corresponding sum and product properties of scalar limits and the fact that entries of a product matrix are finite sums of products.

It is easy to see that $\lim_{n\to\infty} \boldsymbol{A}^{1/n} = \boldsymbol{P}_{\boldsymbol{A}}$ because the matrix power for a symmetric matrix corresponds to taking powers of the eigenvalues and n-th roots converge to either zero or one. Thus the limit is a matrix whose eigenvalues are 0 or 1, which is a projection matrix. By plugging

$$\lim_{n\to\infty} \boldsymbol{A}^{1/n} = \boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{P}_{\boldsymbol{A}} \lim_{n\to\infty} \boldsymbol{A}^{1/n}$$

into (4.5) we get

$$\boldsymbol{A} \odot \boldsymbol{B} = \boldsymbol{P}_{\boldsymbol{A}} \lim_{n\to\infty} \boldsymbol{A}^{1/n} \lim_{n\to\infty} \boldsymbol{B}^{1/n}(\boldsymbol{A}^{1/n}\boldsymbol{B}^{1/n})^{n-1} = \boldsymbol{P}_{\boldsymbol{A}}(\boldsymbol{A} \odot \boldsymbol{B}).$$

This implies that $\text{range}(\boldsymbol{A} \odot \boldsymbol{B}) \subseteq \text{range}(\boldsymbol{A})$. Similarly we can prove $\boldsymbol{A} \odot \boldsymbol{B} = (\boldsymbol{A} \odot \boldsymbol{B})\boldsymbol{P}_{\boldsymbol{B}}$, which implies that $\text{range}(\boldsymbol{A} \odot \boldsymbol{B}) \subseteq \text{range}(\boldsymbol{B})$, and therefore $\text{range}(\boldsymbol{A} \odot \boldsymbol{B}) \subseteq \text{range}(\boldsymbol{A}) \cap \text{range}(\boldsymbol{B})$.

Property OP3 can be seen from the definition of $\odot$ via the limit formula (4.2): when $\boldsymbol{A}$ and $\boldsymbol{B}$ commute, then the n copies of $\boldsymbol{A}^{1/n}$ in $(\boldsymbol{A}^{1/n}\boldsymbol{B}^{1/n})^n$ can be gathered into $\boldsymbol{A}$ and similarly for $\boldsymbol{B}^{1/n}$.

Properties 4–7 easily follow from the formula (4.4) for $\odot$.

Property OP8. For strictly positive definite $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ associativity reduces to the associativity of addition in the log domain. To show it in general we use the representation (4.4) of $\odot$ via the $\log^+$ operation. Let $\boldsymbol{R} = \boldsymbol{R}_{(\boldsymbol{A}\cap\boldsymbol{B})\cap\boldsymbol{C}} = \boldsymbol{R}_{\boldsymbol{A}\cap(\boldsymbol{B}\cap\boldsymbol{C})}$. Then:

$$(\boldsymbol{A} \odot \boldsymbol{B}) \odot \boldsymbol{C} = \boldsymbol{R}\exp(\boldsymbol{R}^\top(\log^+(\boldsymbol{A} \odot \boldsymbol{B}) + \log^+ \boldsymbol{C})\boldsymbol{R})\boldsymbol{R}^\top.$$

Now we rewrite $\log^+(\boldsymbol{A} \odot \boldsymbol{B})$ using (4.3):

$$\log^+(\boldsymbol{A} \odot \boldsymbol{B}) = \boldsymbol{R}_{\boldsymbol{A}\cap\boldsymbol{B}} \log(\boldsymbol{R}_{\boldsymbol{A}\cap\boldsymbol{B}}^\top(\boldsymbol{A} \odot \boldsymbol{B})\boldsymbol{R}_{\boldsymbol{A}\cap\boldsymbol{B}})\boldsymbol{R}_{\boldsymbol{A}\cap\boldsymbol{B}}^\top.$$

Substituting expression (4.4) for $\boldsymbol{A} \odot \boldsymbol{B}$ into the above and using $\boldsymbol{R}_{\boldsymbol{A}\cap\boldsymbol{B}}^\top\boldsymbol{R}_{\boldsymbol{A}\cap\boldsymbol{B}} = \boldsymbol{I}_k$ we get:

$$\log^+(\boldsymbol{A} \odot \boldsymbol{B}) = \underbrace{\boldsymbol{R}_{\boldsymbol{A}\cap\boldsymbol{B}}\boldsymbol{R}_{\boldsymbol{A}\cap\boldsymbol{B}}^\top}_{\boldsymbol{P}_{\boldsymbol{A}\cap\boldsymbol{B}}}(\log^+ \boldsymbol{A} + \log^+ \boldsymbol{B}) \underbrace{\boldsymbol{R}_{\boldsymbol{A}\cap\boldsymbol{B}}\boldsymbol{R}_{\boldsymbol{A}\cap\boldsymbol{B}}^\top}_{\boldsymbol{P}_{\boldsymbol{A}\cap\boldsymbol{B}}}. \tag{4.6}$$

Here $\boldsymbol{P}_{\boldsymbol{A}\cap\boldsymbol{B}} = \boldsymbol{R}_{\boldsymbol{A}\cap\boldsymbol{B}}\boldsymbol{R}_{\boldsymbol{A}\cap\boldsymbol{B}}^\top$ is the projection matrix onto the subspace $\text{range}(\boldsymbol{A}) \cap \text{range}(\boldsymbol{B})$. All the basis vectors of $\boldsymbol{A} \cap \boldsymbol{B} \cap \boldsymbol{C}$ obviously lie in the larger subspace as well, thus the projection leaves them unchanged and we get $\boldsymbol{P}_{\boldsymbol{A}\cap\boldsymbol{B}}\boldsymbol{R} = \boldsymbol{R}$, $\boldsymbol{R}^\top\boldsymbol{P}_{\boldsymbol{A}\cap\boldsymbol{B}} = \boldsymbol{R}^\top$. Thus:

$$(\boldsymbol{A} \odot \boldsymbol{B}) \odot \boldsymbol{C} = \boldsymbol{R}\exp(\boldsymbol{R}^\top(\log^+ \boldsymbol{A} + \log^+ \boldsymbol{B} + \log^+ \boldsymbol{C})\boldsymbol{R})\boldsymbol{R}^\top.$$

The same expression can be obtained for $\boldsymbol{A} \odot (\boldsymbol{B} \odot \boldsymbol{C})$, thus establishing associativity.

Property OP9. By Fact 8.10.9 of Bernstein (2005), we have that for any positive definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and $n \geq 1$:

$$\text{tr}(\boldsymbol{A}^n\boldsymbol{B}^n)^{n+1} \leq \text{tr}(\boldsymbol{A}^{n+1}\boldsymbol{B}^{n+1})^n.$$

Now by substituting $\boldsymbol{A} = \boldsymbol{A}^{1/(n(n+1))}$, $\boldsymbol{B} = \boldsymbol{B}^{1/(n(n+1))}$ the monotonicity property OP9 immediately follows.

Property OP10. When $\boldsymbol{A}$ and $\boldsymbol{B}$ are strictly positive definite, this inequality is an instantiation of the Golden-Thompson inequality (2.1). For arbitrary positive definite matrices, the property follows from the previous monotonicity property OP9. Note that there are symmetric positive definite matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ s.t. $\operatorname{tr}(\boldsymbol{A} \odot \boldsymbol{B} \odot \boldsymbol{C}) \not\leq \operatorname{tr}(\boldsymbol{ABC})$.

Property OP11. We use the expression for $\odot$ operation given in (4.4). Since $\boldsymbol{u} \in \operatorname{range}(\boldsymbol{A})$, the basis of the intersection space is $\boldsymbol{u}$ itself:

$$\boldsymbol{u}\boldsymbol{u}^\top \odot \boldsymbol{A} = \boldsymbol{u} \exp(\boldsymbol{u}^\top (\log^+ \boldsymbol{u}\boldsymbol{u}^\top + \log^+ \boldsymbol{A})\boldsymbol{u})\boldsymbol{u}^\top.$$

Note that $\log^+ \boldsymbol{u}\boldsymbol{u}^\top = \boldsymbol{0}$ and that the expression inside the exponential is a scalar. The desired property immediately follows by moving this scalar to the front.

Property OP12. The expression for the trace of the matrix product is the expected measurement interpretation (3.2) discussed in Sect. 3. Note that $(\boldsymbol{u}^\top \boldsymbol{a}_i)^2$ is a probability vector and in this expression $\boldsymbol{u}\boldsymbol{u}^\top$ can be replaced by any density matrix.

For the second trace $\operatorname{tr}(\boldsymbol{A} \odot \boldsymbol{u}\boldsymbol{u}^\top)$, we can rewrite it using OP11 and eigendecomposition of $\boldsymbol{A}$ as follows:

$$\operatorname{tr}(\boldsymbol{A} \odot \boldsymbol{u}\boldsymbol{u}^\top) = e^{\boldsymbol{u}^\top \log \boldsymbol{A}\boldsymbol{u}} = e^{\sum_i (\boldsymbol{a}_i \cdot \boldsymbol{u})^2 \log \alpha_i} = \prod_i \alpha_i^{(\boldsymbol{u}^\top \boldsymbol{a}_i)^2},$$

which is a weighted geometric average of α_i with weights $(\boldsymbol{u}^\top \boldsymbol{a}_i)^2$.

Property OP13. Since $\det(\boldsymbol{EF}) = \det(\boldsymbol{E}) \det(\boldsymbol{F})$, and for symmetric matrices $\boldsymbol{S}$ and $r \in \mathbb{R}$, $\det(\boldsymbol{S}^r) = \det(\boldsymbol{S})^r$, we have $\det((\boldsymbol{A}^{1/n} \boldsymbol{B}^{1/n})^n) = \det(\boldsymbol{A}) \det(\boldsymbol{B})$ for all $n \in \mathbb{N}$. By property (4.2), the limit of the l.h.s. of the last equality becomes $\boldsymbol{A} \odot \boldsymbol{B}$ and this proves the property.

Property OP14. If $\boldsymbol{A}$ is not full rank, then $\det(\boldsymbol{A})$ is zero. In that case, there will be some $\boldsymbol{u}_i$ that is not in the range of $\boldsymbol{A}$. For that $\boldsymbol{u}_i$, $\operatorname{tr}(\boldsymbol{A} \odot \boldsymbol{u}_i \boldsymbol{u}_i^\top) = 0$, making the whole product zero. When $\boldsymbol{A}$ has full rank, we rewrite the product as follows:

$$\begin{aligned} \prod_i \operatorname{tr}(\boldsymbol{A} \odot \boldsymbol{u}_i \boldsymbol{u}_i^\top) &\overset{\text{OP11}}{=} \prod_i e^{\operatorname{tr}(\log \boldsymbol{A} \boldsymbol{u}_i \boldsymbol{u}_i^\top)} \\ &= e^{\operatorname{tr}(\log \boldsymbol{A} \sum \boldsymbol{u}_i \boldsymbol{u}_i^\top)} = e^{\operatorname{tr}(\log \boldsymbol{A})} = \prod_i \alpha_i = \det(\boldsymbol{A}). \end{aligned}$$

Property OP15. If $\boldsymbol{u} \notin \operatorname{range}(\boldsymbol{A}) \cap \operatorname{range}(\boldsymbol{B}) \overset{\text{OP1}}{=} \operatorname{range}(\boldsymbol{A} \odot \boldsymbol{B})$, then the property trivially holds because $\operatorname{tr}((\boldsymbol{A} \odot \boldsymbol{B}) \odot \boldsymbol{u}\boldsymbol{u}^\top)$ and either $\operatorname{tr}(\boldsymbol{A} \odot \boldsymbol{u}\boldsymbol{u}^\top)$ or $\operatorname{tr}(\boldsymbol{B} \odot \boldsymbol{u}\boldsymbol{u}^\top)$ are zero. When $\boldsymbol{u} \in \operatorname{range}(\boldsymbol{A}) \cap \operatorname{range}(\boldsymbol{B})$, then the property essentially follows from $e^{a+b} = e^a e^b$:

$$\begin{aligned} &\operatorname{tr}((\boldsymbol{A} \odot \boldsymbol{B}) \odot \boldsymbol{u}\boldsymbol{u}^\top) \\ &\overset{\text{OP11}}{=} e^{\boldsymbol{u}^\top \log^+(\boldsymbol{A} \odot \boldsymbol{B})\boldsymbol{u}} \overset{(4.6)}{=} e^{\boldsymbol{u}^\top \boldsymbol{P}_{\boldsymbol{A} \cap \boldsymbol{B}} (\log^+ \boldsymbol{A} + \log^+ \boldsymbol{B}) \boldsymbol{P}_{\boldsymbol{A} \cap \boldsymbol{B}} \boldsymbol{u}} = e^{\boldsymbol{u}^\top (\log^+ \boldsymbol{A} + \log^+ \boldsymbol{B})\boldsymbol{u}} \\ &= e^{\boldsymbol{u}^\top \log^+ \boldsymbol{A}\boldsymbol{u}} e^{\boldsymbol{u}^\top \log^+ \boldsymbol{B}\boldsymbol{u}} = \operatorname{tr}(\boldsymbol{A} \odot \boldsymbol{u}\boldsymbol{u}^\top) \operatorname{tr}(\boldsymbol{B} \odot \boldsymbol{u}\boldsymbol{u}^\top). \end{aligned}$$

Needless to say Property OP15 does not hold if $\boldsymbol{u}\boldsymbol{u}^\top$ is replaced by a mixture of dyads.

Property OP16. Trivially follows from OP11. □

We will now discuss some of the properties further. In particular, we will show a simple example that demonstrates that the upper bound OP10 can be quite loose when both matrices

are dyads. In this case the inequality becomes:

$$\operatorname{tr}(\boldsymbol{u}\boldsymbol{u}^\top \odot \boldsymbol{v}\boldsymbol{v}^\top) \leq \operatorname{tr}(\boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{v}\boldsymbol{v}^\top) = (\boldsymbol{u}\cdot\boldsymbol{v})^2.$$

The right hand side can be made arbitrarily close to one by choosing almost parallel $\boldsymbol{u}$ and $\boldsymbol{v}$. The left side is zero in this case, which can be seen by analyzing the intersection of the ranges. Dyads are rank one matrices and their ranges are lines through the origin. The intersection of two such lines is either only the origin or the line itself. Thus, by Property OP1 it follows that $\boldsymbol{u}\boldsymbol{u}^\top \odot \boldsymbol{v}\boldsymbol{v}^\top = \boldsymbol{0}$, unless $\boldsymbol{u} = \pm\boldsymbol{v}$. This can also be seen from the limit expression in (4.2):

$$\begin{aligned}\boldsymbol{u}\boldsymbol{u}^\top \odot \boldsymbol{v}\boldsymbol{v}^\top &= \lim_{n\to\infty} ((\boldsymbol{u}\boldsymbol{u}^\top)^{\frac{1}{n}} (\boldsymbol{v}\boldsymbol{v}^\top)^{\frac{1}{n}})^n = \lim_{n\to\infty} (\boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{v}\boldsymbol{v}^\top)^n \\ &= \Big(\lim_{n\to\infty} (\boldsymbol{u}\cdot\boldsymbol{v})^{2n-1}\Big)\boldsymbol{u}\boldsymbol{v}^\top = \boldsymbol{0}, \quad \text{unless } \boldsymbol{u} = \pm\boldsymbol{v}.\end{aligned}$$

Where the last equality holds because $|\boldsymbol{u}\cdot\boldsymbol{v}| < 1$, when $\boldsymbol{u} \neq \pm\boldsymbol{v}$.

Note that the expression (4.4) for $\odot$ based on $\log^+$ gives us a convenient method for computing the operation even when the matrices have some zero eigenvalues. The modified matrix logarithm $\log^+$ is easily computed via (4.3). The matrix $\boldsymbol{R}_{\boldsymbol{A}}$ containing the orthonormal basis for range of $\boldsymbol{A}$ can be computed using Gram-Schmidt orthogonalization procedure or the QR-decomposition. To compute the basis for the intersection of range($\boldsymbol{A}$) and range($\boldsymbol{B}$), we express the intersection i.t.o. the union and the orthogonal complement $^\perp$ of a space:

$$\operatorname{range}(\boldsymbol{A}) \cap \operatorname{range}(\boldsymbol{B}) = \big(\operatorname{range}(\boldsymbol{A})^\perp \cup \operatorname{range}(\boldsymbol{B})^\perp\big)^\perp.$$

For any matrix $\boldsymbol{E}$, an orthonormal basis for range$(\boldsymbol{E})^\perp$ can be obtained by completing an orthonormal basis for range($\boldsymbol{E}$) to an orthonormal basis for the whole space. The additional basis vectors needed are the basis for range$(\boldsymbol{E})^\perp$. Also, if we have two matrices $\boldsymbol{E}$ and $\boldsymbol{F}$, we can get the range for the union of their ranges just by putting all columns of $\boldsymbol{E}$ and $\boldsymbol{F}$ together into a bigger matrix $\boldsymbol{G} = (\boldsymbol{E}, \boldsymbol{F})$. Clearly, range($\boldsymbol{G}$) = range($\boldsymbol{E}$) $\cup$ range($\boldsymbol{F}$). Piecing all of this together gives an implementation of the $\odot$ operation.

5 Joint distributions

A density matrix defines a generalized probability distribution over the dyads from one space. However we need to consider several spaces and joint distributions over them. In the conventional case A, B denote finite sets $\{a_1, \ldots, a_{n_A}\}$, $\{b_1, \ldots, b_{n_B}\}$, $(P(a_i))$, $(P(b_j))$ probability vectors over these sets and $(P(a_i, b_j))$ is an $n_A \times n_B$ dimensional matrix of probabilities for the tuple set $A \times B$. In the generalized case, $\mathbb{A}, \mathbb{B}$ denote real finite dimensional vector spaces of dimension $n_{\mathbb{A}}, n_{\mathbb{B}}$ and $\boldsymbol{D}(\mathbb{A})$, $\boldsymbol{D}(\mathbb{B})$ are the density matrices defining the generalized probability distributions over these spaces. The *joint space* $(\mathbb{A}, \mathbb{B})$ is the tensor product[12] between the spaces $\mathbb{A}$ and $\mathbb{B}$, which is of dimension $n_{\mathbb{A}} n_{\mathbb{B}}$. The joint distribution is specified by a density matrix over this joint space, denoted by $\boldsymbol{D}(\mathbb{A}, \mathbb{B})$.

[12] See Bhatia (1997) for a formal definition of tensor product between vector spaces. For us, the tensor product of $\mathbb{R}^{n_{\mathbb{A}}}$ and $\mathbb{R}^{n_{\mathbb{B}}}$ is $\mathbb{R}^{n_{\mathbb{A}} n_{\mathbb{B}}}$.

We let $\boldsymbol{D}(\boldsymbol{a})$, $\boldsymbol{D}(\boldsymbol{b})$ denote the probabilities assigned to dyads $\boldsymbol{a}\boldsymbol{a}^\top, \boldsymbol{b}\boldsymbol{b}^\top$ from the spaces $\mathbb{A}, \mathbb{B}$ by the density matrices $\boldsymbol{D}(\mathbb{A})$, $\boldsymbol{D}(\mathbb{B})$, respectively:

$$\boldsymbol{D}(\boldsymbol{a}) := \mathrm{tr}(\boldsymbol{D}(\mathbb{A})\boldsymbol{a}\boldsymbol{a}^\top), \qquad \boldsymbol{D}(\boldsymbol{b}) := \mathrm{tr}(\boldsymbol{D}(\mathbb{B})\boldsymbol{b}\boldsymbol{b}^\top). \tag{MJ1}$$

The conventional probability distributions can be seen as diagonal density matrices. A probability distribution $(P(a_i))$ on the set A is the density matrix $\mathrm{diag}((P(a_i)))$. Also $P(a_j) = \boldsymbol{e}_j^\top \mathrm{diag}((P(a_i)))\boldsymbol{e}_j$.

To introduce the joint probability $\boldsymbol{D}(\boldsymbol{a}, \boldsymbol{b})$ we need the Kronecker matrix product. Given two matrices $\boldsymbol{E}$ and $\boldsymbol{F}$ with dimensions $n \times m$ and $p \times q$, their Kronecker product (also known as direct product or tensor product) $\boldsymbol{E} \otimes \boldsymbol{F}$ is a matrix with dimensions $np \times mq$ which in block form is given as:

$$\begin{pmatrix} e_{11}\boldsymbol{F} & e_{12}\boldsymbol{F} & \dots & e_{1m}\boldsymbol{F} \\ e_{21}\boldsymbol{F} & e_{22}\boldsymbol{F} & \dots & e_{2m}\boldsymbol{F} \\ \dots & \dots & \dots & \dots \\ e_{n1}\boldsymbol{F} & e_{n2}\boldsymbol{F} & \dots & e_{nm}\boldsymbol{F} \end{pmatrix}.$$

The Kronecker product has the following useful properties:

KP1 $(\boldsymbol{E} \otimes \boldsymbol{F})^\top = \boldsymbol{E}^\top \otimes \boldsymbol{F}^\top$.

KP2 $(\boldsymbol{E} \otimes \boldsymbol{F})(\boldsymbol{G} \otimes \boldsymbol{H}) = \boldsymbol{E}\boldsymbol{G} \otimes \boldsymbol{F}\boldsymbol{H}$ if the dimensions are appropriate.

KP3 $\mathrm{tr}(\boldsymbol{E} \otimes \boldsymbol{F}) = \mathrm{tr}(\boldsymbol{E})\mathrm{tr}(\boldsymbol{F})$.

KP4 If symmetric matrix $\boldsymbol{S}$ has eigenvalues σ_i and eigenvectors $\boldsymbol{s}_i$ and symmetric matrix $\boldsymbol{T}$ has eigenvalues τ_j and eigenvectors $\boldsymbol{t}_j$, then $\boldsymbol{S} \otimes \boldsymbol{T}$ has eigenvalues $\sigma_i\tau_j$ and eigenvectors $\boldsymbol{s}_i \otimes \boldsymbol{t}_j$.

KP5 For symmetric positive definite matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$, $(\boldsymbol{A} \otimes \boldsymbol{B}) \odot (\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{A} \odot \boldsymbol{C}) \otimes (\boldsymbol{B} \odot \boldsymbol{D})$.

The first four properties are standard. The last property follows from the limit definition (4.2) of the $\odot$ operation.

$$\begin{aligned}(\boldsymbol{A} \otimes \boldsymbol{B}) \odot (\boldsymbol{C} \otimes \boldsymbol{D}) &= \lim_{n\to\infty} \left((\boldsymbol{A} \otimes \boldsymbol{B})^{\frac{1}{n}} (\boldsymbol{C} \otimes \boldsymbol{D})^{\frac{1}{n}}\right)^n \\ &= \lim_{n\to\infty} \left((\boldsymbol{A}^{\frac{1}{n}} \boldsymbol{C}^{\frac{1}{n}}) \otimes (\boldsymbol{B}^{\frac{1}{n}} \boldsymbol{D}^{\frac{1}{n}})\right)^n \\ &= \left(\lim_{n\to\infty} (\boldsymbol{A}^{\frac{1}{n}} \boldsymbol{B}^{\frac{1}{n}})^n\right) \otimes \left(\lim_{n\to\infty} (\boldsymbol{C}^{\frac{1}{n}} \boldsymbol{D}^{\frac{1}{n}})^n\right).\end{aligned}$$

The last transition which moved the limit inside the Kronecker product, follows from the fact that the elements of the Kronecker product matrix are just pairwise products of elements from the two matrices. And when all limits exist, a limit of a product of two number sequences is a product of limits.

Now the joint probability $\boldsymbol{D}(\boldsymbol{a}, \boldsymbol{b})$ becomes the probability assigned by density matrix $\boldsymbol{D}(\mathbb{A}, \mathbb{B})$ to the jointly specified dyad $(\boldsymbol{a} \otimes \boldsymbol{b})(\boldsymbol{a} \otimes \boldsymbol{b})^\top$:

$$\boldsymbol{D}(\boldsymbol{a}, \boldsymbol{b}) := \mathrm{tr}(\boldsymbol{D}(\mathbb{A}, \mathbb{B})(\boldsymbol{a} \otimes \boldsymbol{b})(\boldsymbol{a} \otimes \boldsymbol{b})^\top) = \mathrm{tr}(\boldsymbol{D}(\mathbb{A}, \mathbb{B})(\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{b}\boldsymbol{b}^\top)). \tag{MJ3}$$

Note that in the conventional case a joint probability between two sets A and B is defined over all pairs of points from A and B. This corresponds to the so-called "separable case"

in the generalized setting when the joint density $\boldsymbol{D}(\mathbb{A},\mathbb{B})$ can be expressed as $\sum_i \sigma_i \boldsymbol{a}_i \boldsymbol{a}_i^\top \otimes \boldsymbol{b}_i \boldsymbol{b}_i^\top$, where the $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are states of spaces $\mathbb{A}$ and $\mathbb{B}$, respectively. However, in the more general case, there are elementary events in the joint space $(\mathbb{A},\mathbb{B})$ that don't decompose into elementary events of the spaces $\mathbb{A}$ and $\mathbb{B}$, i.e. there are dyads in the joint space that are not of the form $\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{b}\boldsymbol{b}^\top$, where $\boldsymbol{a}\boldsymbol{a}^\top$ and $\boldsymbol{b}\boldsymbol{b}^\top$ are dyads of $\mathbb{A}$ and $\mathbb{B}$, respectively. In quantum physics the inseparability of a density matrix is called "entanglement".

6 Marginalization of the joint via partial traces

We would like to be able to perform marginalization operations on our joint density matrix $\boldsymbol{D}(\mathbb{A},\mathbb{B})$, i.e. obtain the density matrix $\boldsymbol{D}(\mathbb{A})$ from the joint matrix. In the conventional case the marginalization was performed by summing out one of the variables by summing the rows or the columns of the matrix specifying the joint probability distribution. For density matrices, the analogous operation is the *partial trace* (see e.g. Nielsen and Chuang 2000).

The partial trace is a generalization of normal matrix trace. It typically produces a matrix instead of a number and can be used to retrieve the (scaled) factor matrices from a Kronecker product. We denote the partial trace with $\mathrm{tr}_{\mathbb{A}}$, where $\mathbb{A}$ specifies the space to be "summed out". Suppose $\boldsymbol{G}$ is a matrix over the space $\mathbb{A}\otimes\mathbb{B}$ and $\mathbb{A}$ has dimension n and $\mathbb{B}$ dimension m. Thus $\boldsymbol{G}$ has dimension $nm \times nm$ and can be written in block form as a $n \times n$ matrix of $m \times m$ matrices $\boldsymbol{G}_{ij}$:

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} & \dots & \boldsymbol{G}_{1n} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22} & \dots & \boldsymbol{G}_{2n} \\ \dots & \dots & \dots & \dots \\ \boldsymbol{G}_{n1} & \boldsymbol{G}_{n2} & \dots & \boldsymbol{G}_{nn} \end{pmatrix}.$$

Here we suppose that space $\mathbb{A}$ is $\mathbb{R}^n$ and space $\mathbb{B}$ is $\mathbb{R}^m$. Then the two partial traces of this matrix are given by:

$$\underbrace{\mathrm{tr}_{\mathbb{A}}(\boldsymbol{G})}_{m\times m} = \boldsymbol{G}_{11} + \boldsymbol{G}_{22} + \cdots + \boldsymbol{G}_{nn},$$

$$\underbrace{\mathrm{tr}_{\mathbb{B}}(\boldsymbol{G})}_{n\times n} = \begin{pmatrix} \mathrm{tr}(\boldsymbol{G}_{11}) & \mathrm{tr}(\boldsymbol{G}_{12}) & \dots & \mathrm{tr}(\boldsymbol{G}_{1n}) \\ \mathrm{tr}(\boldsymbol{G}_{21}) & \mathrm{tr}(\boldsymbol{G}_{22}) & \dots & \mathrm{tr}(\boldsymbol{G}_{2n}) \\ \dots & \dots & \dots & \dots \\ \mathrm{tr}(\boldsymbol{G}_{n1}) & \mathrm{tr}(\boldsymbol{G}_{n2}) & \dots & \mathrm{tr}(\boldsymbol{G}_{nn}) \end{pmatrix}.$$

In multilinear algebra partial traces are known as tensor contractions and can of course be generalized to the tensor product of more than two spaces. The partial trace is a linear operator and we now give some other useful properties:

PT1 $\mathrm{tr}_{\mathbb{A}}(\boldsymbol{E}\otimes\boldsymbol{F}) = \mathrm{tr}(\boldsymbol{E})\boldsymbol{F},\ \mathrm{tr}_{\mathbb{B}}(\boldsymbol{E}\otimes\boldsymbol{F}) = \mathrm{tr}(\boldsymbol{F})\boldsymbol{E}.$
PT2 $\mathrm{tr}(\boldsymbol{G}) = \mathrm{tr}(\mathrm{tr}_{\mathbb{A}}(\boldsymbol{G})) = \mathrm{tr}(\mathrm{tr}_{\mathbb{B}}(\boldsymbol{G})).$
PT3 $\mathrm{tr}_{\mathbb{A}}(\boldsymbol{G}(\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{F})) = \mathrm{tr}_{\mathbb{A}}(\boldsymbol{G})\boldsymbol{F},\ \mathrm{tr}_{\mathbb{A}}((\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{F})\boldsymbol{G}) = \boldsymbol{F}\mathrm{tr}_{\mathbb{A}}(\boldsymbol{G}).$
PT4 $\mathrm{tr}(\boldsymbol{G}(\boldsymbol{E}\otimes\boldsymbol{F})) = \mathrm{tr}(\mathrm{tr}_{\mathbb{B}}(\boldsymbol{G}(\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{F}))\boldsymbol{E}).$

The first three properties are straightforward and the last one follows from the others as follows:

$$\operatorname{tr}(\boldsymbol{G}(\boldsymbol{E}\otimes\boldsymbol{F})) \overset{\text{KP2}}{=} \operatorname{tr}(\boldsymbol{G}(\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{F})(\boldsymbol{E}\otimes\boldsymbol{I}_{\mathbb{B}})) = \operatorname{tr}((\boldsymbol{E}\otimes\boldsymbol{I}_{\mathbb{B}})\boldsymbol{G}(\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{F})) \overset{\text{PT2}}{=} \operatorname{tr}(\operatorname{tr}_{\mathbb{B}}((\boldsymbol{E}\otimes\boldsymbol{I}_{\mathbb{B}})\boldsymbol{G}(\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{F}))) \overset{\text{PT3}}{=} \operatorname{tr}(\boldsymbol{E}\operatorname{tr}_{\mathbb{B}}(\boldsymbol{G}(\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{F}))).$$

We use the partial trace to define marginals as follows:

$$\boldsymbol{D}(\mathbb{A}) := \operatorname{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})), \qquad \boldsymbol{D}(\mathbb{B}) := \operatorname{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})). \tag{MJ2}$$

The following lemma shows that $\boldsymbol{D}(\mathbb{A})$ and $\boldsymbol{D}(\mathbb{B})$ defined this way are again density matrices.

Lemma 1 *Partial trace of a density matrix is also a density matrix.*

Proof Symmetry is obvious. Trace one follows from Property PT2 of the partial trace:

$$\operatorname{tr}(\boldsymbol{D}(\mathbb{A})) = \operatorname{tr}(\operatorname{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A},\mathbb{B}))) = \operatorname{tr}(\boldsymbol{D}(\mathbb{A},\mathbb{B})) = 1.$$

Positive definiteness follows by a similar argument:

$$\begin{aligned}\boldsymbol{a}^{\top}\boldsymbol{D}(\mathbb{A})\boldsymbol{a} &= \operatorname{tr}(\boldsymbol{D}(\mathbb{A})\boldsymbol{a}\boldsymbol{a}^{\top}) \overset{\text{PT3}}{=} \operatorname{tr}(\operatorname{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{a}\boldsymbol{a}^{\top}\otimes\boldsymbol{I}_{\mathbb{B}})))\\ &\overset{\text{PT2}}{=} \operatorname{tr}\Bigg(\boldsymbol{D}(\mathbb{A},\mathbb{B})\Bigg(\boldsymbol{a}\boldsymbol{a}^{\top}\otimes\underbrace{\sum_i \boldsymbol{b}_i\boldsymbol{b}_i^{\top}}_{\boldsymbol{I}_{\mathbb{B}}}\Bigg)\Bigg)\\ &= \sum_i \operatorname{tr}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{a}\boldsymbol{a}^{\top}\otimes\boldsymbol{b}_i\boldsymbol{b}_i^{\top}))\\ &= \sum_i (\boldsymbol{a}\otimes\boldsymbol{b}_i)^{\top}\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{a}\otimes\boldsymbol{b}_i) \geq 0.\end{aligned}$$

□

Partial traces also allow us to define objects of the type $\boldsymbol{D}(\mathbb{A},\boldsymbol{b})$. In the conventional case this corresponds to taking one row or column out of the joint probability table. In the generalized case we want the following property to be satisfied:

$$\operatorname{tr}(\boldsymbol{D}(\mathbb{A},\boldsymbol{b})\boldsymbol{a}\boldsymbol{a}^{\top}) = \boldsymbol{D}(\boldsymbol{a},\boldsymbol{b}). \tag{MJ5}$$

This is accomplished by defining $\boldsymbol{D}(\mathbb{A},\boldsymbol{b})$ via the following formula:

$$\boldsymbol{D}(\mathbb{A},\boldsymbol{b}) := \operatorname{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_A\otimes\boldsymbol{b}\boldsymbol{b}^{\top})). \tag{MJ4}$$

Property MJ5 now follows from partial trace Property PT4. We can also see that trace of $\boldsymbol{D}(\boldsymbol{A},\boldsymbol{b})$ gives us the probability $\boldsymbol{D}(\boldsymbol{b})$:

$$\operatorname{tr}(\boldsymbol{D}(\mathbb{A},\boldsymbol{b})) \overset{\text{MJ4}}{=} \operatorname{tr}(\operatorname{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_A \otimes \boldsymbol{b}\boldsymbol{b}^\top))) \overset{\text{PT2}}{=} \operatorname{tr}(\operatorname{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_A \otimes \boldsymbol{b}\boldsymbol{b}^\top)))$$
$$\overset{\text{PT3}}{=} \operatorname{tr}(\operatorname{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{A},\mathbb{B}))\boldsymbol{b}\boldsymbol{b}^\top) \overset{\text{MJ2}}{=} \operatorname{tr}(\boldsymbol{D}(\mathbb{B})\boldsymbol{b}\boldsymbol{b}^\top) = \boldsymbol{D}(\boldsymbol{b}). \tag{6.1}$$

A brief note on matrix properties of $\boldsymbol{D}(\mathbb{A},\boldsymbol{b})$. We just saw that its trace is $\boldsymbol{D}(\boldsymbol{b})$ which is between zero and one. Since it satisfies Property (MJ5), it is positive definite as well. Symmetry is also easily verified.

Note that for any orthogonal system $\boldsymbol{b}_i$ of $\mathbb{B}$,

$$\boldsymbol{D}(\mathbb{A}) = \sum_i \boldsymbol{D}(\mathbb{A},\boldsymbol{b}_i).$$

This can be seen as follows.

$$\begin{aligned}\boldsymbol{D}(\mathbb{A}) &\overset{\text{MJ2}}{=} \operatorname{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})) = \operatorname{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{I}_{\mathbb{B}})) \\ &= \operatorname{tr}_{\mathbb{B}}\Bigg(\boldsymbol{D}(\mathbb{A},\mathbb{B})\Bigg(\boldsymbol{I}_{\mathbb{A}} \otimes \underbrace{\sum_i \boldsymbol{b}_i\boldsymbol{b}_i^\top}_{\boldsymbol{I}_{\mathbb{B}}}\Bigg)\Bigg) = \sum_i \operatorname{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{b}_i\boldsymbol{b}_i^\top)) \\ &\overset{\text{MJ4}}{=} \sum_i \boldsymbol{D}(\mathbb{A},\boldsymbol{b}_i).\end{aligned}$$

The conventional definition of independence also naturally generalizes: $\boldsymbol{D}(\mathbb{A})$ is independent of $\boldsymbol{D}(\mathbb{B})$ if the joint density matrix decomposes: $\boldsymbol{D}(\mathbb{A},\mathbb{B}) = \boldsymbol{D}(\mathbb{A}) \otimes \boldsymbol{D}(\mathbb{B})$. It is easy to see that in this case we have $\boldsymbol{D}(\boldsymbol{a},\boldsymbol{b}) = \boldsymbol{D}(\boldsymbol{a})\boldsymbol{D}(\boldsymbol{b})$ for all $\boldsymbol{a},\boldsymbol{b}$:

$$\begin{aligned}\boldsymbol{D}(\boldsymbol{a},\boldsymbol{b}) &= \operatorname{tr}((\boldsymbol{D}(\mathbb{A}) \otimes \boldsymbol{D}(\mathbb{B}))(\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{b}\boldsymbol{b}^\top)) \overset{\text{KP2}}{=} \operatorname{tr}((\boldsymbol{D}(\mathbb{A})\boldsymbol{a}\boldsymbol{a}^\top) \otimes (\boldsymbol{D}(\mathbb{B})\boldsymbol{b}\boldsymbol{b}^\top)) \\ &\overset{\text{KP3}}{=} \operatorname{tr}(\boldsymbol{D}(\mathbb{A})\boldsymbol{a}\boldsymbol{a}^\top)\operatorname{tr}(\boldsymbol{D}(\mathbb{B})\boldsymbol{b}\boldsymbol{b}^\top) = \boldsymbol{D}(\boldsymbol{a})\boldsymbol{D}(\boldsymbol{b}).\end{aligned}$$

7 Conditional probabilities

The topic of conditional probabilities in this generalized setting contains many subtleties. First we will give the defining formulas for conditional density matrices and then discuss some of the issues.

CP1 $\boldsymbol{D}(\mathbb{A}|\mathbb{B}) := \boldsymbol{D}(\mathbb{A},\mathbb{B}) \odot (\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B}))^{-1}$ (formula (4) of Cerf and Adami 1999 expressed with the $\odot$ operation). This formula requires $\boldsymbol{D}(\mathbb{B})$ to be invertible. In the conventional case, this corresponds to the conditional probabilities being undefined if the event conditioned on has probability zero.

CP2 $\boldsymbol{D}(\mathbb{A}|\boldsymbol{b}) := \frac{\boldsymbol{D}(\mathbb{A},\boldsymbol{b})}{\boldsymbol{D}(\boldsymbol{b})}$.

CP3 $\boldsymbol{D}(\boldsymbol{a}|\mathbb{B}) := \boldsymbol{D}(\boldsymbol{a},\mathbb{B}) \odot \boldsymbol{D}(\mathbb{B})^{-1}$.

CP4 $\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b}) := \frac{\boldsymbol{D}(\boldsymbol{a},\boldsymbol{b})}{\boldsymbol{D}(\boldsymbol{b})}$. This basic conditional probability is a straightforward generalization of the conventional case. It also has a quantum-mechanical interpretation. See Appendix for details.

Note that CP1 has the form: density matrix $\odot$ inverse of a normalization. We can also reexpress the other definitions in this unified form:

$$\text{CP}'2\quad \boldsymbol{D}(\mathbb{A}|\boldsymbol{b}) = \underbrace{\text{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_{\mathbb{A}}\otimes \boldsymbol{b}\boldsymbol{b}^{\top}))}_{\boldsymbol{D}(\mathbb{A},\boldsymbol{b})} \odot \underbrace{\text{tr}_{\mathbb{B}}((\boldsymbol{I}_{\mathbb{A}}\otimes \boldsymbol{D}(\mathbb{B}))(\boldsymbol{I}_{\mathbb{A}}\otimes \boldsymbol{b}\boldsymbol{b}^{\top}))^{-1}}_{\frac{1}{\boldsymbol{D}(\boldsymbol{b})}\boldsymbol{I}_{\mathbb{A}}}}.$$

$$\text{CP}'3\quad \boldsymbol{D}(\boldsymbol{a}|\mathbb{B}) = \underbrace{\text{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{a}\boldsymbol{a}^{\top}\otimes \boldsymbol{I}_{\mathbb{B}}))}_{\boldsymbol{D}(\boldsymbol{a},\mathbb{B})} \odot \underbrace{\text{tr}_{\mathbb{A}}((\boldsymbol{I}_{\mathbb{A}}\otimes \boldsymbol{D}(\mathbb{B}))(\boldsymbol{a}\boldsymbol{a}^{\top}\otimes \boldsymbol{I}_{\mathbb{B}}))^{-1}}_{\boldsymbol{D}(\mathbb{B})^{-1}}.$$

$$\text{CP}'4\quad \boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b}) = \underbrace{\text{tr}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{a}\boldsymbol{a}^{\top}\otimes \boldsymbol{b}\boldsymbol{b}^{\top}))}_{\boldsymbol{D}(\boldsymbol{a},\boldsymbol{b})} \odot \underbrace{\text{tr}((\boldsymbol{I}_{\mathbb{A}}\otimes \boldsymbol{D}(\mathbb{B}))(\boldsymbol{a}\boldsymbol{a}^{\top}\otimes \boldsymbol{b}\boldsymbol{b}^{\top}))^{-1}}_{\frac{1}{\boldsymbol{D}(\boldsymbol{b})}}.$$

We say that the joint density $\boldsymbol{D}(\mathbb{A},\mathbb{B})$ is *decoupled* if its eigendecomposition has the form: $\boldsymbol{D}(\mathbb{A},\mathbb{B}) = (\boldsymbol{\mathcal{W}}_{\mathbb{A}}\otimes\boldsymbol{\mathcal{W}}_{\mathbb{B}})\,\text{diag}(\boldsymbol{\omega})(\boldsymbol{\mathcal{W}}_{\mathbb{A}}\otimes\boldsymbol{\mathcal{W}}_{\mathbb{B}})^{\top}$. Note that $\boldsymbol{\mathcal{W}}_{\mathbb{A}}\otimes\boldsymbol{\mathcal{W}}_{\mathbb{B}}$ is orthogonal iff both $\boldsymbol{\mathcal{W}}_{\mathbb{A}}$ and $\boldsymbol{\mathcal{W}}_{\mathbb{B}}$ are orthogonal. As we shall see later, dealing with conditionals is often simpler in the decoupled case. We first prove an upper bound for $\text{tr}(\boldsymbol{D}(\mathbb{A}|\mathbb{B}))$ that is tight iff the joint is decoupled.

Lemma 2 *The following inequality holds*:

$$\text{tr}(\boldsymbol{D}(\mathbb{A}|\mathbb{B})) \leq n_{\mathbb{B}},$$

where $n_{\mathbb{B}}$ is the dimensionality of space $\mathbb{B}$. Furthermore, $\text{tr}(\boldsymbol{D}(\mathbb{A}|\mathbb{B})) = n_{\mathbb{B}}$ *if and only if the joint $\boldsymbol{D}(\mathbb{A},\mathbb{B})$ is decoupled.*

Proof The inequality is shown using properties of $\odot$ and partial traces:

$$\begin{aligned}\text{tr}(\boldsymbol{D}(\mathbb{A}|\mathbb{B})) &\overset{\text{CP1}}{=} \text{tr}(\boldsymbol{D}(\mathbb{A},\mathbb{B}) \odot (\boldsymbol{I}_{\mathbb{A}}\otimes \boldsymbol{D}(\mathbb{B})^{-1})) \overset{\text{OP10}}{\leq} \text{tr}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_{\mathbb{A}}\otimes \boldsymbol{D}(\mathbb{B})^{-1}))\\ &\overset{\text{PT2}}{=} \text{tr}(\text{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_{\mathbb{A}}\otimes \boldsymbol{D}(\mathbb{B})^{-1}))) \overset{\text{PT3}}{=} \text{tr}(\underbrace{\text{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{A},\mathbb{B}))}_{\boldsymbol{D}(\mathbb{B})}\boldsymbol{D}(\mathbb{B})^{-1})\\ &= \text{tr}(\boldsymbol{I}_{\mathbb{B}}) = n_{\mathbb{B}}.\end{aligned}$$

Remember that equality in Property OP10 of $\odot$ only occurs when the two matrices commute. Two matrices commute iff their eigensystems are the same. This gives us the condition that the eigensystem of $\boldsymbol{D}(\mathbb{A},\mathbb{B})$ must be the same as the eigensystem of $\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{D}(\mathbb{B})^{-1}$. The latter eigensystem is clearly decoupled. Thus for equality to hold it is *necessary* that the eigensystem of $\boldsymbol{D}(\mathbb{A},\mathbb{B})$ be decoupled.

Now we will argue that it is also *sufficient*. Let the joint density matrix have eigensystem $\boldsymbol{D}(\mathbb{A},\mathbb{B}) = (\boldsymbol{\mathcal{W}}_{\mathbb{A}}\otimes\boldsymbol{\mathcal{W}}_{\mathbb{B}})\,\text{diag}(\boldsymbol{\omega})(\boldsymbol{\mathcal{W}}_{\mathbb{A}}^{\top}\otimes\boldsymbol{\mathcal{W}}_{\mathbb{B}}^{\top})$. $\boldsymbol{I}_{\mathbb{A}}$ commutes with any matrix on space $\mathbb{A}$. Therefore it suffices to show that the marginal $\boldsymbol{D}(\mathbb{B})$ in this case has eigensystem $\boldsymbol{\mathcal{W}}_{\mathbb{B}}$. The decoupled eigensystem matrix $\boldsymbol{\mathcal{W}}_{\mathbb{A}}\otimes\boldsymbol{\mathcal{W}}_{\mathbb{B}}$ has the following list of $n_{\mathbb{A}}n_{\mathbb{B}}$ columns:

$$\begin{aligned}\boldsymbol{\mathcal{W}}_{\mathbb{A}}\otimes\boldsymbol{\mathcal{W}}_{\mathbb{B}} = \Big(&\boldsymbol{w}_{\mathbb{A}}^{1}\otimes\boldsymbol{w}_{\mathbb{B}}^{1}, \boldsymbol{w}_{\mathbb{A}}^{1}\otimes\boldsymbol{w}_{\mathbb{B}}^{2}, \ldots, \boldsymbol{w}_{\mathbb{A}}^{1}\otimes\boldsymbol{w}_{\mathbb{B}}^{n_{\mathbb{B}}},\\ &\boldsymbol{w}_{\mathbb{A}}^{2}\otimes\boldsymbol{w}_{\mathbb{B}}^{1}, \boldsymbol{w}_{\mathbb{A}}^{2}\otimes\boldsymbol{w}_{\mathbb{B}}^{2}, \ldots, \boldsymbol{w}_{\mathbb{A}}^{2}\otimes\boldsymbol{w}_{\mathbb{B}}^{n_{\mathbb{B}}},\\ &\ldots,\ldots,\ldots,\ldots,\ldots,\ldots,\ldots,\ldots,\ldots,\ldots,\\ &\boldsymbol{w}_{\mathbb{A}}^{n_{\mathbb{A}}}\otimes\boldsymbol{w}_{\mathbb{B}}^{1}, \boldsymbol{w}_{\mathbb{A}}^{n_{\mathbb{A}}}\otimes\boldsymbol{w}_{\mathbb{B}}^{2}\ldots, \boldsymbol{w}_{\mathbb{A}}^{n_{\mathbb{A}}}\otimes\boldsymbol{w}_{\mathbb{B}}^{n_{\mathbb{B}}}\Big).\end{aligned}$$

In correspondence with this structure we adopt a double indexing scheme for the eigenvalues $\omega_{i,j}$ of the joint matrix $\boldsymbol{D}(\mathbb{A},\mathbb{B})$, where $\omega_{i,j}$ is the eigenvalue associated with eigenvector

$\boldsymbol{w}_{\mathbb{A}}^i \otimes \boldsymbol{w}_{\mathbb{B}}^j$. The index i runs from 1 to $n_{\mathbb{A}}$, and j runs to $n_{\mathbb{B}}$. Now the eigendecomposition can be written as:

$$\boldsymbol{D}(\mathbb{A},\mathbb{B}) = \sum_{i,j} \omega_{i,j} (\boldsymbol{w}_{\mathbb{A}}^i (\boldsymbol{w}_{\mathbb{A}}^i)^\top \otimes \boldsymbol{w}_{\mathbb{B}}^j (\boldsymbol{w}_{\mathbb{B}}^j)^\top).$$

Partial trace is a linear operator and $\mathrm{tr}_{\mathbb{A}}(\boldsymbol{w}_{\mathbb{A}}^i (\boldsymbol{w}_{\mathbb{A}}^i)^\top \otimes \boldsymbol{w}_{\mathbb{B}}^j (\boldsymbol{w}_{\mathbb{B}}^j)^\top) \overset{\text{PT1}}{=} \boldsymbol{w}_{\mathbb{B}}^j (\boldsymbol{w}_{\mathbb{B}}^j)^\top$. Therefore:

$$\boldsymbol{D}(\mathbb{B}) = \mathrm{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})) = \sum_j \omega_j \boldsymbol{w}_{\mathbb{B}}^j (\boldsymbol{w}_{\mathbb{B}}^j)^\top,$$

where $\omega_j = \sum_i \omega_{i,j}$. Thus we produced the eigendecomposition of the marginal $\boldsymbol{D}(\mathbb{B})$ and it indeed has eigensystem $\boldsymbol{\mathcal{W}}_{\mathbb{B}}$. □

Let us briefly discuss the connection and difference between our notion of decoupled joints and the notion of entanglement that appears in quantum physics. Recall that entanglement, as we mentioned at the end of Sect. 5 corresponds to the fact that there are dyads $\boldsymbol{c}\boldsymbol{c}^\top$ in the joint space $(\mathbb{A},\mathbb{B})$ that can't be written as $\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{b}\boldsymbol{b}^\top$ for any two dyads $\boldsymbol{a}\boldsymbol{a}^\top$ and $\boldsymbol{b}\boldsymbol{b}^\top$ in $\mathbb{A}$ and $\mathbb{B}$. This notion carries over to mixed states or density matrices. In quantum physics, a joint density matrix $\boldsymbol{D}(\mathbb{A},\mathbb{B})$ is called *separable* (or non-entangled) if it can be expressed as $(\boldsymbol{\mathcal{W}}_{\mathbb{A}} \otimes \boldsymbol{\mathcal{W}}_{\mathbb{B}})\,\mathrm{diag}(\boldsymbol{\omega})(\boldsymbol{\mathcal{W}}_{\mathbb{A}} \otimes \boldsymbol{\mathcal{W}}_{\mathbb{B}})^\top$. The crucial difference between the definitions of separable and decoupled matrices is that in the separable case, $\boldsymbol{\mathcal{W}}_{\mathbb{A}}$ and $\boldsymbol{\mathcal{W}}_{\mathbb{B}}$ don't have to be orthogonal. Every decoupled matrix is separable, but there are separable density matrices that are not decoupled. The question of deciding whether a given matrix is separable is known to be very difficult, whereas the question of being decoupled is easily decided by e.g. the condition of the above lemma. One of the reasons for which Cerf and Adami (1999) introduced a conditional density matrix via Rule CP1 was to give a necessary condition for the separability of a joint density matrix.

To complete the rules for conditional density matrices, we would need rules that allow us to marginalize the conditionals, e.g. for going from $\boldsymbol{D}(\mathbb{A}|\mathbb{B})$ to $\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b})$. One obvious consequence of our definitions is the marginalization rule for $\boldsymbol{D}(\mathbb{A}|\boldsymbol{b})$:

$$\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b}) \overset{\text{CP4}}{=} \frac{\boldsymbol{D}(\boldsymbol{a},\boldsymbol{b})}{\boldsymbol{D}(\boldsymbol{b})} \overset{\text{MJ5}}{=} \frac{\mathrm{tr}(\boldsymbol{D}(\mathbb{A},\boldsymbol{b})\boldsymbol{a}\boldsymbol{a}^\top)}{\boldsymbol{D}(\boldsymbol{b})} \overset{\text{CP2}}{=} \mathrm{tr}(\boldsymbol{D}(\mathbb{A}|\boldsymbol{b})\boldsymbol{a}\boldsymbol{a}^\top). \tag{MC4}$$

There don't seem to be any other simple marginalization rules for $\boldsymbol{D}(\mathbb{A}|\mathbb{B})$ and $\boldsymbol{D}(\boldsymbol{a}|\mathbb{B})$ that hold for arbitrary joints. However, when the joint is decoupled, then the following additional marginalization rule for $\boldsymbol{D}(\mathbb{A}|\mathbb{B})$ is valid:

Lemma 3 *For all decoupled joints* $\boldsymbol{D}(\mathbb{A},\mathbb{B})$,

$$\boldsymbol{D}(\boldsymbol{a}|\mathbb{B}) = \mathrm{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{A}|\mathbb{B})(\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{I}_{\mathbb{B}})).$$

Proof We will compute both sides of the equation and show them to be identical. We begin by writing down the decomposition of the decoupled joint from Lemma 2:

$$\boldsymbol{D}(\mathbb{A},\mathbb{B}) = \sum_{i,j} \omega_{i,j} (\boldsymbol{w}_{\mathbb{A}}^i (\boldsymbol{w}_{\mathbb{A}}^i)^\top \otimes \boldsymbol{w}_{\mathbb{B}}^j (\boldsymbol{w}_{\mathbb{B}}^j)^\top). \tag{7.1}$$

Additionally, in the same lemma, the following form for $\boldsymbol{D}(\mathbb{B})$ was established in this case:

$$\boldsymbol{D}(\mathbb{B}) = \mathrm{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})) = \sum_j \omega_j \boldsymbol{w}_{\mathbb{B}}^j (\boldsymbol{w}_{\mathbb{B}}^j)^\top, \tag{7.2}$$

where $\omega_j = \sum_i \omega_{i,j}$. According to CP3, $\boldsymbol{D}(\boldsymbol{a}|\mathbb{B}) = \boldsymbol{D}(\boldsymbol{a},\mathbb{B}) \odot \boldsymbol{D}(\mathbb{B})^{-1}$, therefore we will need to compute $\boldsymbol{D}(\boldsymbol{a},\mathbb{B})$:

$$\boldsymbol{D}(\boldsymbol{a},\mathbb{B}) \overset{\text{MJ4}}{=} \mathrm{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{I}_{\mathbb{B}})) \overset{(7.1)}{=} \sum_{i,j} \omega_{i,j} \mathrm{tr}_{\mathbb{A}}((\boldsymbol{w}^i_{\mathbb{A}}(\boldsymbol{w}^i_{\mathbb{A}})^\top \otimes \boldsymbol{w}^j_{\mathbb{B}}(\boldsymbol{w}^j_{\mathbb{B}})^\top)(\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{I}_{\mathbb{B}}))$$

$$\overset{\text{KP2}}{=} \sum_{i,j} \omega_{i,j} \mathrm{tr}_{\mathbb{A}}((\boldsymbol{w}^i_{\mathbb{A}}(\boldsymbol{w}^i_{\mathbb{A}})^\top)\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{w}^j_{\mathbb{B}}(\boldsymbol{w}^j_{\mathbb{B}})^\top) \overset{\text{PT1}}{=} \sum_{i,j} \omega_{i,j} (\boldsymbol{w}^i_{\mathbb{A}} \cdot \boldsymbol{a})^2 \boldsymbol{w}^j_{\mathbb{B}}(\boldsymbol{w}^j_{\mathbb{B}})^\top.$$

Together with (7.2), this gives:

$$\boldsymbol{D}(\boldsymbol{a}|\mathbb{B}) = \sum_{i,j} \frac{\omega_{i,j}(\boldsymbol{w}^i_{\mathbb{A}} \cdot \boldsymbol{a})^2}{\omega_j} \boldsymbol{w}^j_{\mathbb{B}}(\boldsymbol{w}^j_{\mathbb{B}})^\top.$$

Now, we proceed to the right side of the equation in the lemma. Substituting (7.1) and (7.2) into the formula for $\boldsymbol{D}(\mathbb{A}|\mathbb{B})$ we obtain:

$$\boldsymbol{D}(\mathbb{A}|\mathbb{B}) \overset{\text{CP1}}{=} \boldsymbol{D}(\mathbb{A},\mathbb{B}) \odot (\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B})^{-1}) = \sum_{i,j} \frac{\omega_{i,j}}{\omega_j} (\boldsymbol{w}^i_{\mathbb{A}}(\boldsymbol{w}^i_{\mathbb{A}})^\top \otimes \boldsymbol{w}^j_{\mathbb{B}}(\boldsymbol{w}^j_{\mathbb{B}})^\top). \tag{7.3}$$

Using linearity of the partial trace we compute the right side as follows:

$$\mathrm{tr}_{\mathbb{A}}(\boldsymbol{D}(\boldsymbol{A}|\boldsymbol{B})(\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{I}_{\mathbb{B}})) \overset{(7.3)}{=} \sum_{i,j} \frac{\omega_{i,j}}{\omega_j} \mathrm{tr}_{\mathbb{A}}((\boldsymbol{w}^i_{\mathbb{A}}(\boldsymbol{w}^i_{\mathbb{A}})^\top \otimes \boldsymbol{w}^j_{\mathbb{B}}(\boldsymbol{w}^j_{\mathbb{B}})^\top)(\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{I}_{\mathbb{B}}))$$

$$\overset{\text{KP2}}{=} \sum_{i,j} \frac{\omega_{i,j}}{\omega_j} \mathrm{tr}_{\mathbb{A}}((\boldsymbol{w}^i_{\mathbb{A}}(\boldsymbol{w}^i_{\mathbb{A}})^\top)\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{w}^j_{\mathbb{B}}(\boldsymbol{w}^j_{\mathbb{B}})^\top)$$

$$\overset{\text{PT1}}{=} \sum_{i,j} \frac{\omega_{i,j}(\boldsymbol{w}^i_{\mathbb{A}} \cdot \boldsymbol{a})^2}{\omega_j} \boldsymbol{w}^j_{\mathbb{B}}(\boldsymbol{w}^j_{\mathbb{B}})^\top.$$

□

As discussed, obtaining $\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b}) = \frac{\mathrm{tr}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{b}\boldsymbol{b}^\top))}{\mathrm{tr}(\boldsymbol{D}(\mathbb{B})\boldsymbol{b}\boldsymbol{b}^\top)}$ from $\boldsymbol{D}(\mathbb{A}|\mathbb{B})$ is non-trivial. In particular, there are cases where

$$\mathrm{tr}(\boldsymbol{D}(\mathbb{A}|\mathbb{B})(\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{b}\boldsymbol{b}^\top)) \neq D(\boldsymbol{a}|\boldsymbol{b}),$$

even when $\boldsymbol{D}(\mathbb{A},\mathbb{B})$ is decoupled and $\boldsymbol{a}$ and $\boldsymbol{b}$ are not eigenvectors of $\boldsymbol{D}(\mathbb{A})$ and $\boldsymbol{D}(\mathbb{B})$, respectively. Curiously enough, if we replace the matrix product with $\odot$, then we always have

$$\begin{aligned}
&\mathrm{tr}(\boldsymbol{D}(\mathbb{A}|\mathbb{B}) \odot (\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{b}\boldsymbol{b}^\top)) \\
&\overset{\text{CP1}}{=} \mathrm{tr}((\boldsymbol{D}(\mathbb{A},\mathbb{B}) \odot (\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B})^{-1})) \odot (\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{b}\boldsymbol{b}^\top)) \\
&\overset{\text{OP15}}{=} \mathrm{tr}(\boldsymbol{D}(\mathbb{A},\mathbb{B}) \odot (\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{b}\boldsymbol{b}^\top))\mathrm{tr}((\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B})^{-1}) \odot (\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{b}\boldsymbol{b}^\top)) \\
&\overset{\text{OP16}}{=} \frac{\mathrm{tr}(\boldsymbol{D}(\mathbb{A},\mathbb{B}) \odot (\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{b}\boldsymbol{b}^\top))}{\mathrm{tr}(\boldsymbol{D}(\mathbb{B}) \odot \boldsymbol{b}\boldsymbol{b}^\top)}.
\end{aligned}$$

Let us now recall the conditionals in the conventional probability theory. The full conditional table $P(A|B)$ lists conditional probabilities of all pairs of elementary events $P(a_i|b_j)$.

This table has the obvious properties: The sum of all entries is n_B and the sum of any column is 1, i.e. $\sum_i P(a_i|b_j) = \sum_i \frac{P(a_i,b_j)}{P(b_j)} = \frac{P(b_j)}{P(b_j)} = 1$. Thus a conditional table is a column-stochastic matrix and for any such matrix we can construct a joint that has that matrix as its conditional table. For example we can take arbitrary probability vector p and multiply the i-th column of $P(A|B)$ by p_i, now the sum of each column is p_i and thus the sum of all entries is 1 and we have a valid joint. Note that this implies that many different joints have the same conditional table.

The decoupled case behaves as the conventional case, i.e. many joints correspond to the same conditional. A decoupled joint and conditional always have the same eigensystem and going from the joint to the conditional is similar to the conventional case (see (7.1–7.3) for details).

However, for non-decoupled joint density matrices, i.e. when $\mathrm{tr}(\boldsymbol{D}(\mathbb{A}|\mathbb{B})) < n_{\mathbb{B}}$ (Lemma 2), the situation is quite different. For example, the eigenvalues of $\boldsymbol{D}(\mathbb{A}|\mathbb{B})$ can now be bigger than 1 (Cerf and Adami 1999). Also based on numerical experiments, we conjecture that in the non-decoupled case, the mapping between $\boldsymbol{D}(\mathbb{A},\mathbb{B})$ and $\boldsymbol{D}(\mathbb{A}|\mathbb{B})$ is invertible, i.e. unlike the conventional case, there is only one joint that gives rise to a given conditional matrix. In other words we conjecture that in the non-decoupled case it suffices to specify the conditional $\boldsymbol{D}(\mathbb{A}|\mathbb{B})$.

More specifically, we claim that the following EM-like algorithm converges to $\boldsymbol{D}(\mathbb{B})$ and then $\boldsymbol{D}(\mathbb{A},\mathbb{B}) \overset{\text{CP1}}{=} \boldsymbol{D}(\mathbb{A}|\mathbb{B}) \odot \boldsymbol{D}(\mathbb{B})$: $\boldsymbol{W}_0$ is initialized to $\boldsymbol{I}_{\mathbb{B}}/n_{\mathbb{B}}$ and the estimate $\boldsymbol{W}_{t+1}$ for $\boldsymbol{D}(\mathbb{B})$ is computed from $\boldsymbol{D}(\mathbb{A}|\mathbb{B})$ and the previous estimate $\boldsymbol{W}_t$ as

$$\boldsymbol{W}_{t+1} = \frac{\mathrm{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{A}|\mathbb{B}) \odot (\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{W}_t))}{\mathrm{tr}(\boldsymbol{D}(\mathbb{A}|\mathbb{B}) \odot (\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{W}_t))}.$$

8 Theorems of total probability

The Theorem of Total Probability is an important calculation in conventional probability theory. It expresses probability of some event a as an expected conditional probability of the elementary events b_i that form a partition of the probability space B:

$$P(a) = \sum_i P(a|b_i)P(b_i).$$

TP1 For any orthogonal system $\boldsymbol{b}_i$ of $\mathbb{B}$, $\boldsymbol{D}(\boldsymbol{a}) = \sum_i \boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b}_i)\boldsymbol{D}(\boldsymbol{b}_i)$.
TP2 $\boldsymbol{D}(\boldsymbol{a}) = \mathrm{tr}(\boldsymbol{D}(\boldsymbol{a}|\mathbb{B}) \odot \boldsymbol{D}(\mathbb{B}))$.
TP3 $\boldsymbol{D}(\mathbb{A}) = \mathrm{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A}|\mathbb{B}) \odot (\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B})))$.

The first formula can be shown as follows:

$$\boldsymbol{D}(\boldsymbol{a}) = \mathrm{tr}(\boldsymbol{D}(\boldsymbol{a},\mathbb{B})) = \sum_i \boldsymbol{b}_i^{\top} \boldsymbol{D}(\boldsymbol{a},\mathbb{B})\boldsymbol{b}_i = \sum_i \boldsymbol{D}(\boldsymbol{a},\boldsymbol{b}_i) = \sum_i \boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b}_i)\boldsymbol{D}(\boldsymbol{b}_i).$$

To derive the second apply $\odot \boldsymbol{D}(\mathbb{B})$ to both sides of CP3, take trace of both sides and use (6.1). The proof of the third property follows the same outline but uses CP1 and MJ2.

Conventional versions of the last two properties are obtained when the density and conditional matrices are diagonal. Note that in general these generalizations of the Theorem of Total Probability do not "decouple", i.e. you cannot write them as a sum of products of conditional and marginal probabilities. However, using the Property OP10 of $\odot$ operation we

can establish upper bounds on probability of $\boldsymbol{D}(\boldsymbol{a})$ in terms of "decoupled" sums that look like the conventional versions of the Theorem of Total Probability. If $\boldsymbol{D}(\mathbb{B}) = \sum_i \omega_i \boldsymbol{w}_i \boldsymbol{w}_i^\top$ and $\boldsymbol{D}(\boldsymbol{a}|\mathbb{B}) = \sum_i \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ are eigendecompositions of the corresponding matrices, then

$$\begin{aligned}\boldsymbol{D}(\boldsymbol{a}) &= \operatorname{tr}(\boldsymbol{D}(\boldsymbol{a}|\mathbb{B}) \odot \boldsymbol{D}(\mathbb{B})) \leq \operatorname{tr}(\boldsymbol{D}(\boldsymbol{a}|\mathbb{B})\boldsymbol{D}(\mathbb{B})) \\ &= \underbrace{\sum_i \overbrace{\omega_i}^{\text{probability}} \overbrace{\boldsymbol{w}_i^\top \boldsymbol{D}(\boldsymbol{a}|\mathbb{B})\boldsymbol{w}_i}^{\text{variance}}}_{\text{expected variance}} \\ &= \underbrace{\sum_i \overbrace{\boldsymbol{u}_i^\top \boldsymbol{D}(\mathbb{B})\boldsymbol{u}_i}^{\text{probability } \boldsymbol{D}(\boldsymbol{u}_i)} \overbrace{\lambda_i}^{\text{outcome}}}_{\text{expected measurement}}. \end{aligned} \tag{8.1}$$

The first version of the upper bound corresponds to using the eigendecomposition of $\boldsymbol{D}(\mathbb{B})$ and can be interpreted as an expected variance calculation with $\boldsymbol{D}(\boldsymbol{a}|\mathbb{B})$ as the covariance matrix. The second version expands $\boldsymbol{D}(\boldsymbol{a}|\mathbb{B})$ and corresponds to a quantum measurement of system in state $\boldsymbol{D}(\mathbb{B})$ with instrument specified by $\boldsymbol{D}(\boldsymbol{a}|\mathbb{B})$. Letting $p(b_i)$ equal ω_i or $\boldsymbol{u}_i^\top \boldsymbol{D}(\mathbb{B})\boldsymbol{u}_i$ and letting $p(a_i|b_i)$ equal $\boldsymbol{w}_i^\top \boldsymbol{D}(\boldsymbol{a}|\mathbb{B})\boldsymbol{w}_i$ or λ_i, we see the correspondence of these upper bounds to the conventional Theorem of Total Probability. The equality only occurs when $\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{B})$ and $\boldsymbol{D}(\boldsymbol{B})$ commute.

9 Bayes rules

In the conventional setup we assume that a model M_i is chosen with prior probability $P(M_i)$. The model then generates the data y with probability $P(y|M_i)$, i.e.

$$\begin{aligned}P(y) &= \sum_i P(M_i)P(y|M_i) \\ &= \operatorname{tr}(\operatorname{diag}((P(M_i))\operatorname{diag}((P(y|M_i))).\end{aligned}$$

The reason why we expressed $P(y)$ as a trace of two diagonal matrices will become apparent in a moment.

The generalized setup is completely analogous. There is an underlying joint space $(\mathbb{M}, \mathbb{Y})$ between the model space $\mathbb{M}$ and the data space $\mathbb{Y}$. The prior is specified by a density matrix $\boldsymbol{D}(\mathbb{M})$. The data is a unit direction $\boldsymbol{y}$ in $\mathbb{Y}$ space that is generated by the density $\boldsymbol{D}(\mathbb{Y})$. The probability $\boldsymbol{D}(\boldsymbol{y})$ can be expressed i.t.o. the prior $\boldsymbol{D}(\mathbb{M})$ and data likelihood $\boldsymbol{D}(\boldsymbol{y}|\mathbb{M})$ using TP2:

$$\boldsymbol{D}(\boldsymbol{y}) = \operatorname{tr}(\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})).$$

Note that in the conventional case we first chose a model based on the prior and then generated data based on the chosen model. In the generalized case we do not know how to decouple the action on the prior from the choice of the data when conditioned on the prior.

Let us first recall the conventional Bayes rule and rewrite it in matrix notation:

$$P(M_i|y) = \frac{P(M_i)P(y|M_i)}{P(y)}, \quad \text{where } P(y) = \sum_j P(M_j)P(y|M_j),$$
$$\operatorname{diag}(P(M_i|y)) = \frac{\operatorname{diag}(P(M_i))\operatorname{diag}(P(y|M_i))}{\operatorname{tr}(\operatorname{diag}(P(M_i))\operatorname{diag}(P(y|M_i)))}. \tag{9.1}$$

We now present and discuss the analogous Bayes rule for the generalized setting. At the end of this section we present a list of all Bayes rules.

In the generalized Bayes rule we cannot simply multiply the prior density matrix with the data likelihood matrix. This is because a product of two symmetric positive definite matrices can be neither symmetric nor positive definite (see Fig. 4). Instead, we replace the matrix multiplication with $\odot$ operation:

$$\boldsymbol{D}(\mathbb{M}|\boldsymbol{y}) = \frac{\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})}{\boldsymbol{D}(\boldsymbol{y})}, \quad \text{where } \boldsymbol{D}(\boldsymbol{y}) = \operatorname{tr}(\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})). \tag{9.2}$$

Normalizing by the trace ensures that the trace of the posterior density matrix is one. In both the conventional as well as the new Bayes rule above, the normalization constant is the likelihood of the data. When the matrices $\boldsymbol{D}(\mathbb{M})$ and $\boldsymbol{D}(\boldsymbol{y}|\mathbb{M})$ have the same eigensystem, then $\odot$ becomes the matrix multiplication. In the following subsections we derive the above Bayes rules from the minimum relative entropy principle. For the conventional Bayes rule the standard relative entropy between probability vectors is used, whereas the generalized Bayes rule and the crucial $\odot$ operation is motivated by the quantum relative entropy between density matrices due to Umegaki (see e.g. Nielsen and Chuang 2000).

We visualize the conventional Bayes rule in Fig. 7. Repeated application of the rule with the same likelihood makes the posteriors increasingly concentrated on the point with maximum data likelihood $P(y|M_i)$. Therefore this rule can be interpreted as a soft max-likelihood calculation. Figure 8 demonstrates the generalized Bayes rule. There the posterior gradually moves towards the eigenvector belonging to the largest eigenvalue of the data likelihood matrix $\boldsymbol{D}(\boldsymbol{y}|\mathbb{M})$. Thus the new rule can be interpreted as a soft calculation of the eigenvector with maximum eigenvalue.

In Fig. 11 we depict a sequence of updates with the new Bayes rule when the data likelihood matrix is different in each iteration. Observe that based on the relative lengths of the axes (eigenvalues) and the directions of the axes (eigenvectors) in the ellipse describing the current data likelihood matrix, the posterior adjusts its axis lenghts and directions.

Other Bayes rules for our calculus are listed below. They all express one conditional in terms of the corresponding reverse conditional.

BR1 $\boldsymbol{D}(\mathbb{B}|\mathbb{A}) = (\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B})) \odot \boldsymbol{D}(\mathbb{A}|\mathbb{B}) \odot (\boldsymbol{D}(\mathbb{A}) \otimes \boldsymbol{I}_{\mathbb{B}})^{-1}$, where $\boldsymbol{D}(\mathbb{A}) = \operatorname{tr}_{\mathbb{B}}((\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B})) \odot \boldsymbol{D}(\mathbb{A}|\mathbb{B}))$.

BR2 $\boldsymbol{D}(\boldsymbol{b}|\mathbb{A}) = \boldsymbol{D}(\boldsymbol{b})\boldsymbol{D}(\mathbb{A}|\boldsymbol{b}) \odot \boldsymbol{D}(\mathbb{A})^{-1}$, where $\boldsymbol{D}(\mathbb{A}) \overset{\text{TP3}}{=} \operatorname{tr}_{\mathbb{B}}((\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B})) \odot \boldsymbol{D}(\mathbb{A}|\mathbb{B}))$.

BR3 $\boldsymbol{D}(\mathbb{B}|\boldsymbol{a}) = \frac{\boldsymbol{D}(\mathbb{B}) \odot \boldsymbol{D}(\boldsymbol{a}|\mathbb{B})}{\boldsymbol{D}(\boldsymbol{a})}$, where $\boldsymbol{D}(\boldsymbol{a}) \overset{\text{TP2}}{=} \operatorname{tr}(\boldsymbol{D}(\mathbb{B}) \odot \boldsymbol{D}(\boldsymbol{a}|\mathbb{B}))$.
This is the Bayes rule derived in Warmuth (2005) that was discussed above.

BR4 $\boldsymbol{D}(\boldsymbol{b}|\boldsymbol{a}) = \frac{\boldsymbol{D}(\boldsymbol{b})\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b})}{\boldsymbol{D}(\boldsymbol{a})}$, where $\boldsymbol{D}(\boldsymbol{a}) \overset{\text{TP1}}{=} \sum_i \boldsymbol{D}(\boldsymbol{b}_i)\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b}_i)$.
The summation in the normalization factor proceeds over any orthogonal system $\boldsymbol{b}_i$.

All these Bayes rules can be easily derived as follows: first express the conditional on the left i.t.o. the joint by applying the definitions of conditional probability from Sect. 7; then

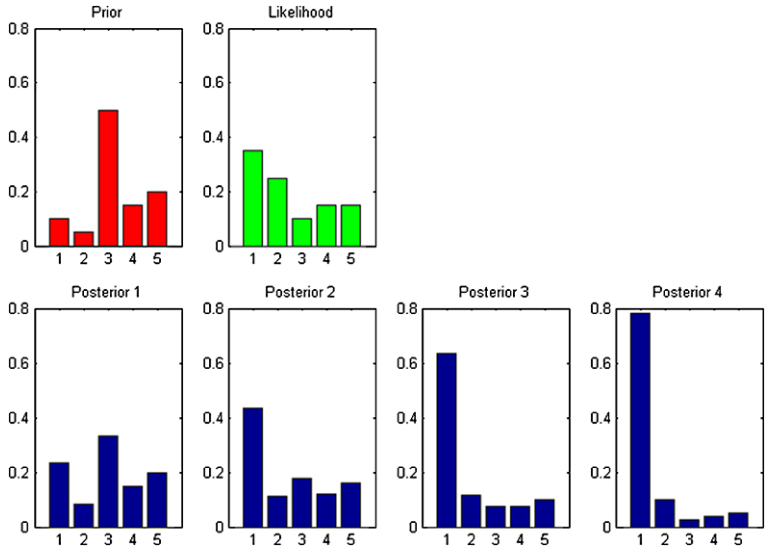


Fig. 7 (Color online) We apply the conventional Bayes rule 4 times, using the same data likelihood vector $P(y|M_i)$ and making the current posterior the new prior. At first, the posteriors are close to the initial prior but eventually the posteriors focus their weight on $\arg\max_i P(y|M_i)$. The conventional Bayes rule may be seen as a soft maximum calculation. The initial prior is in *red*, the likelihood is in *green* and posteriors are in *blue*

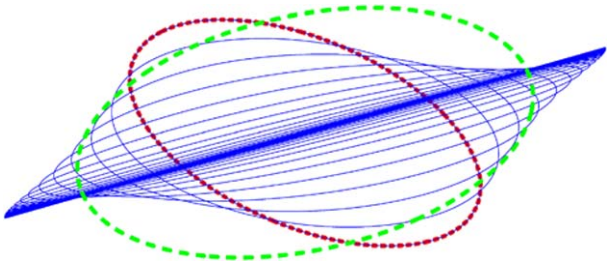


Fig. 8 (Color online) We depict several iterations of the generalized Bayes rule. The *red ellipse* depicts the prior $\boldsymbol{D}(\mathbb{M})$, the *green ellipse* depicts the data likelihood matrix $\boldsymbol{D}(\boldsymbol{y}|\mathbb{M})$, which is kept fixed on successive iterations, and the *blue ellipses* depict posteriors $\boldsymbol{D}(\mathbb{M}|\boldsymbol{y})$. The posterior density matrices gradually move away from the prior and focus on the longest axis of the covariance matrix. The generalized Bayes rule can be seen as a soft calculation of eigenvector with largest eigenvalue

apply these definitions again for expressing the joint in terms of the reverse conditional. For example,

$$\boldsymbol{D}(\mathbb{B}|\boldsymbol{a}) \overset{\text{CP2}}{=} \frac{\boldsymbol{D}(\mathbb{B},\boldsymbol{a})}{\boldsymbol{D}(\boldsymbol{a})} \overset{\text{CP3}}{=} \frac{\boldsymbol{D}(\mathbb{B}) \odot \boldsymbol{D}(\boldsymbol{a}|\mathbb{B})}{\boldsymbol{D}(\boldsymbol{a})}.$$

As was mentioned above, the new Bayes rule can be seen as a soft maximum eigenvalue calculation. We will now give an example that shows that its impossible to track the maximum eigenvalue without changing the eigensystem. First, suppose that we have a diagonal density matrix $\boldsymbol{W} = \sum_i \omega_i \boldsymbol{e}_i \boldsymbol{e}_i^\top$ and another diagonal matrix $\boldsymbol{S} = \sum_i \sigma_i \boldsymbol{e}_i \boldsymbol{e}_i^\top$. Then $\operatorname{tr}(\boldsymbol{W}\boldsymbol{S}) = \sum_i \omega_i \sigma_i$ and this means that by changing ω_i we can easily focus on the

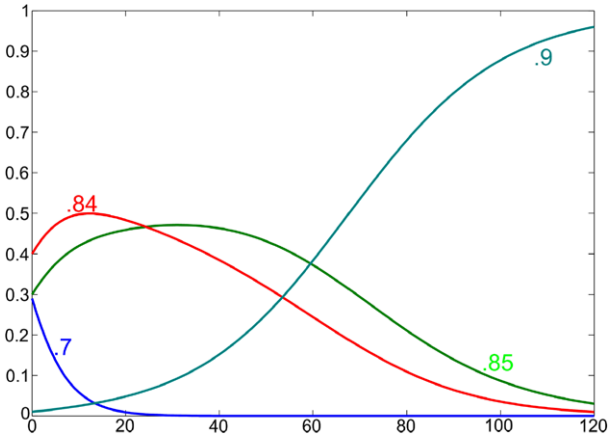


Fig. 9 We plot many iterations of the conventional Bayes rule when the same data likelihood $(P(y|M_i)) = (.7, .84, .85, .9)$ is used in each iteration and the prior is $(P(M_i)) = (.29, .4, .3, .01)$. For each of the four models we plot the posterior probability as a function of the iteration number. Initially the posterior curve with likelihood .85 overtakes the curve with likelihood .84, but eventually the curve with likelihood .9 takes over both. Note that the curve with the largest data likelihood looks like a sigmoid and the one with smallest like a reverse sigmoid

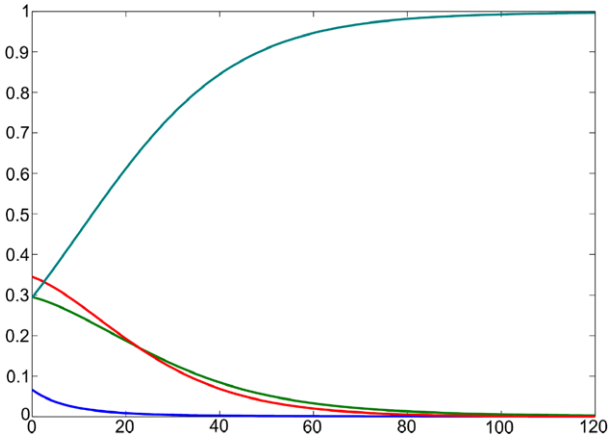


Fig. 10 We plot many iterations of the generalized Bayes rule when the same data likelihood matrix $\boldsymbol{D}(y|\mathbb{M})$ is used in each iteration. As the prior $\boldsymbol{D}(\mathbb{M})$ we choose the diagonalized prior $\operatorname{diag}((P(M_i))$ of Fig. 9 on the left and as the likelihood $\boldsymbol{D}(y|\mathbb{M})$ we choose $\boldsymbol{U}\operatorname{diag}((P(y|M_i))\boldsymbol{U}^T$, where the eigensystem $\boldsymbol{U}$ is a random rotation matrix. Let $\boldsymbol{D}(\mathbb{M}|t)$ denote the posterior at iteration t when the fixed $\boldsymbol{D}(y|\mathbb{M})$ is used in all iterations. The curves are the projections of this posterior onto the four eigendirections of $\boldsymbol{D}(y|\mathbb{M})$ as a function of t, i.e. $\boldsymbol{u}_i^\top \boldsymbol{D}(\mathbb{M}|t)\boldsymbol{u}_i$, where $\boldsymbol{u}_i$ are the columns of $\boldsymbol{U}$. The above plot is qualitatively similar to the left plot. The curve corresponding to the largest eigenvalue of the data likelihood is again a partial sigmoid

high σ_i. Now suppose $\boldsymbol{W}$ is diagonal as before, but $\boldsymbol{S}$ has the Hadamard matrix eigensystem. Hadamard matrices $\boldsymbol{H}$ are square $n \times n$ matrices that have ± 1 elements and satisfy the condition $\boldsymbol{H}\boldsymbol{H}^\top = n\boldsymbol{I}$. Thus $\frac{\boldsymbol{H}}{\sqrt{n}}$ is an orthogonal matrix. Let $\boldsymbol{h}_i$ be the columns of this orthogonal matrix derived from a Hadamard matrix and let $\boldsymbol{S} = \sum_i \sigma_i \boldsymbol{h}_i \boldsymbol{h}_i^\top$. Entries of $\boldsymbol{h}_i$

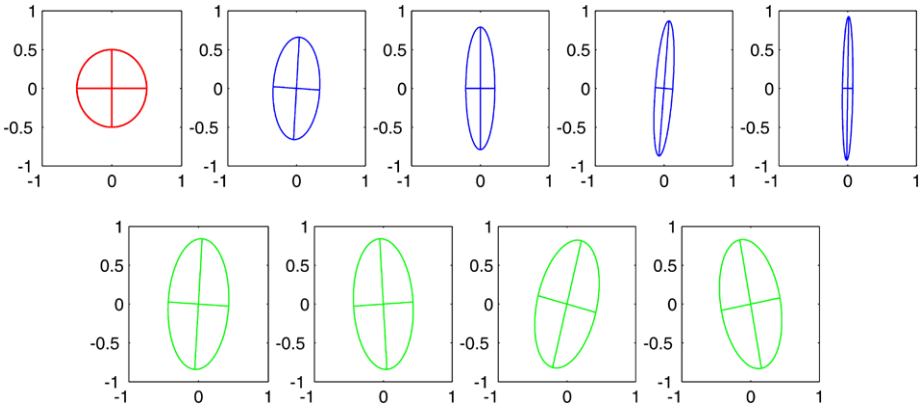


Fig. 11 (Color online) Sequence of Bayes updates with the new Bayes rule (9.2): from *left* to *right*, the prior is in *red*; the first data likelihood matrix is below in *green*; the first posterior is above in *blue*, and so forth

are $\pm\frac{1}{\sqrt{n}}$, therefore $\operatorname{tr}(\boldsymbol{e}_i\boldsymbol{e}_i^\top \boldsymbol{h}_j\boldsymbol{h}_j^\top) = \frac{1}{n}$. Computing the trace we obtain:

$$\operatorname{tr}(\boldsymbol{W}\boldsymbol{S}) = \sum_{i,j} \sigma_i \tau_j \operatorname{tr}(\boldsymbol{e}_i\boldsymbol{e}_i^\top \boldsymbol{h}_j\boldsymbol{h}_j^\top) = \frac{1}{n}\operatorname{tr}(\boldsymbol{W})\operatorname{tr}(\boldsymbol{S}) = \frac{\operatorname{tr}(\boldsymbol{S})}{n}.$$

This means that *any* diagonal density matrix $\boldsymbol{W}$ only "sees" the average of eigenvalues of $\boldsymbol{S}$ and is unable to focus on the highest eigenvalue.

9.1 Deriving the conventional and generalized Bayes rule

In this section we show how to derive the conventional Bayes rule (9.1) and the generalized Bayes rule for density matrices (9.2) by minimizing a tradeoff between a relative entropy and an expected log likelihood. For two probability vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, the relative entropy is defined as $\Delta(\boldsymbol{x}, \boldsymbol{y}) := \sum_i x_i \log \frac{x_i}{y_i}$. We use the convention that $0 \log 0 := 0$ which is justified by $\lim_{x\to 0} x \log x = 0$. It is well known that $\Delta(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ and that $\Delta(\boldsymbol{x}, \boldsymbol{y}) = 0$ iff $\boldsymbol{x} = \boldsymbol{y}$.

Theorem 3 *Let the prior $(P(M_i))$ be any probability vector and the data likelihood $(P(y|M_i))$ be any non-negative vector of the same dimension. Then*

$$-\log P(y) = \inf_{(\omega_i)\ \text{prob.vec.}} \Delta\big((\omega_i), (P(M_i))\big) - \sum_i \omega_i \log P(y|M_i),$$

and $\boldsymbol{\omega} = (P(M_i)P(y|M_i)/P(y))$ is the unique optimum solution.

Proof Let the support of a vector $\boldsymbol{x}$ be the set of all indices $1 \leq i \leq n$ s.t. $x_i \neq 0$ and denote this set as $\mathrm{s}(\boldsymbol{x})$. For any probability vector (ω_i), such that $\mathrm{s}((\omega_i)) \subseteq \mathrm{s}(P(M_i)) \cap \mathrm{s}(P(y|M_i))$, we have

$$-\log P(y) = \underbrace{\sum_i \omega_i \log \frac{\omega_i}{P(M_i)}}_{\Delta((\omega_i),(P(M_i)))} - \sum_i \omega_i \log P(y|M_i) - \underbrace{\sum_i \omega_i \log \frac{\omega_i}{P(M_i)P(y|M_i)/P(y)}}_{\Delta((\omega_i),(P(M_i)P(y|M_i)/P(y)))}.$$

The precondition on the support of (ω_i) assures that all three sums above are finite because it avoids the case $\omega_i \log 0$, when $\omega_i > 0$. Since the l.h.s. is a constant,

$$\inf_{\substack{(\omega_i) \text{ prob.vec.} \\ \mathrm{s}((\omega_i)) \subseteq \mathrm{s}(P(M_i)) \cap \mathrm{s}(P(y|M_i))}} \Delta\left((\omega_i), (P(M_i))\right) - \sum_i \omega_i P(y|M_i)$$

$$= \sup_{\substack{(\omega_i) \text{ prob.vec.} \\ \mathrm{s}((\omega_i)) \subseteq \mathrm{s}(P(M_i)) \cap \mathrm{s}(P(y|M_i))}} -\Delta\left((\omega_i), (P(M_i)P(y|M_i)/P(y))\right).$$

The sup clearly has $\boldsymbol{\omega} = (P(M_i)P(y|M_i)/P(y))$ as its unique solution and the inf remains unchanged if the condition on the support of (ω_i) is dropped. This gives us the statement of the theorem. □

This theorem can also be proven using differentiation (see e.g. Zellner 1998; Kivinen and Warmuth 1997; Singh et al. 2003). For the density matrix case this was done in Warmuth (2005), Tsuda et al. (2005). We now prove the corresponding theorem for density matrices in a different way. For two density matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, the quantum relative entropy is defined as $\Delta(\boldsymbol{A}, \boldsymbol{B}) := \mathrm{tr}(\boldsymbol{A}(\log \boldsymbol{A} - \log \boldsymbol{B}))$. There is a potential problem when some of the eigenvalues of the matrices are zero. However, we will now reason that this definition is justified under the assumption $0 \log 0 = 0$ and $\Delta(\boldsymbol{A}, \boldsymbol{B})$ is bounded iff $\mathrm{range}(\boldsymbol{A}) \subseteq \mathrm{range}(\boldsymbol{B})$.

The first term $\mathrm{tr}(\boldsymbol{A} \log \boldsymbol{A})$ becomes $\sum_i \alpha_i \log \alpha_i$, where the α_i are the eigenvalues of $\boldsymbol{A}$. This term is always finite. If $\boldsymbol{B}$ is eigendecomposed as $\sum_i \beta_i \boldsymbol{b}_i \boldsymbol{b}_i^\top$, then the second term $\mathrm{tr}(\boldsymbol{A} \log \boldsymbol{B})$ can be rewritten as $\sum_i \boldsymbol{b}_i^\top \boldsymbol{A} \boldsymbol{b}_i \log \beta_i$. If $\mathrm{range}(\boldsymbol{A}) \subseteq \mathrm{range}(\boldsymbol{B})$, then $\mathrm{range}(\boldsymbol{B})^\perp \subseteq \mathrm{range}(\boldsymbol{A})^\perp$, where $\perp$ denotes the orthogonal complement space. If $\beta_i = 0$, then $\boldsymbol{b}_i \in \mathrm{range}(\boldsymbol{B})^\perp$ and under our assumption on $\mathrm{range}(\boldsymbol{A})$ this also means that $\boldsymbol{b}_i^\top \boldsymbol{A} \boldsymbol{b}_i = 0$. Therefore, for all i, s.t. $\beta_i = 0$, the summand $\boldsymbol{b}_i^\top \boldsymbol{A} \boldsymbol{b}_i \log \beta_i$ has the form $0 \log 0 = 0$. If on the other hand, $\mathrm{range}(\boldsymbol{A}) \not\subseteq \mathrm{range}(\boldsymbol{B})$, this also means $\mathrm{range}(\boldsymbol{B})^\perp \not\subseteq \mathrm{range}(\boldsymbol{A})^\perp$. The eigenvectors $\boldsymbol{b}_i$ with zero eigenvalues form a basis for $\mathrm{range}(\boldsymbol{B})^\perp$ and therefore there exists some $\boldsymbol{b}_i$ s.t. $\boldsymbol{b}_i^\top \boldsymbol{A} \boldsymbol{b}_i \neq 0$. This gives a summand of the form $x \log 0$, with $x \neq 0$, and this is infinite. Notice that this discussion also means that

$$\mathrm{tr}(\boldsymbol{A} \log \boldsymbol{B}) = \begin{cases} \mathrm{tr}(\boldsymbol{A} \log^+ \boldsymbol{B}) & \text{when } \mathrm{range}(\boldsymbol{A}) \subseteq \mathrm{range}(\boldsymbol{B}), \\ -\infty & \text{otherwise.} \end{cases} \tag{9.3}$$

As before the function $\Delta(\boldsymbol{A}, \boldsymbol{B})$ is non-negative and equal zero iff both arguments agree (e.g. Nielsen and Chuang 2000).

Theorem 4 *Let the prior $\boldsymbol{D}(\mathbb{M})$ be any density matrix and data likelihood $\boldsymbol{D}(\boldsymbol{y}|\mathbb{M})$ be any symmetric positive definite matrix of the same dimension. Then*

$$-\log \boldsymbol{D}(\boldsymbol{y}) = \inf_{\boldsymbol{W} \text{ dens.mat.}} \Delta(\boldsymbol{W}, \boldsymbol{D}(\mathbb{M})) - \mathrm{tr}(\boldsymbol{W} \log \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})),$$

and $\boldsymbol{W} = \frac{\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})}{D(\boldsymbol{y})}$ is the unique optimum solution.

Proof For any density matrix $\boldsymbol{W}$ s.t. $\mathrm{range}(\boldsymbol{W}) \subseteq \mathrm{range}(\boldsymbol{D}(\mathbb{M})) \cap \mathrm{range}(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M}))$, we have

$$-\log D(\boldsymbol{y}) = \underbrace{\mathrm{tr}(\boldsymbol{W}(\log \boldsymbol{W} - \log \boldsymbol{D}(\mathbb{M}))}_{\Delta(\boldsymbol{W}, \boldsymbol{D}(\mathbb{M}))} - \mathrm{tr}(\boldsymbol{W}(\log \boldsymbol{D}(\boldsymbol{y}|\mathbb{M}))$$
$$- \mathrm{tr}(\boldsymbol{W}(\log \boldsymbol{W} - (\log \boldsymbol{D}(\mathbb{M}) + \log \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})/\boldsymbol{D}(\boldsymbol{y})))).$$

Since $\text{range}(\boldsymbol{W}) \subseteq \text{range}(\boldsymbol{D}(\mathbb{M})) \cap \text{range}(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M}))$, $\text{tr}(\boldsymbol{W} \log \boldsymbol{D}(\mathbb{M}))$ and $\text{tr}(\boldsymbol{W} \log \boldsymbol{D}(\boldsymbol{y}|\mathbb{M}))$ are both finite. Assuming that for any symmetric positive definite matrices $\boldsymbol{W}$, $\boldsymbol{A}$ and $\boldsymbol{B}$

$$\text{tr}(\boldsymbol{W}(\log \boldsymbol{A} + \log \boldsymbol{B}) = \text{tr}(\boldsymbol{W} \log(\boldsymbol{A} \odot \boldsymbol{B})), \text{ when } \text{range}(\boldsymbol{W}) \subseteq \text{range}(\boldsymbol{A}) \cap \text{range}(\boldsymbol{B}), \quad (9.4)$$

the above equality would become

$$-\log D(\boldsymbol{y}) = \Delta(\boldsymbol{W}, \boldsymbol{D}(\mathbb{M})) - \text{tr}(\boldsymbol{W} \log \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})) - \Delta(\boldsymbol{W}, (\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}|\mathbb{M}))/\boldsymbol{D}(\boldsymbol{y})).$$

Since the l.h.s. is a constant,

$$\begin{aligned} &\inf_{\substack{\boldsymbol{W} \text{ dens.mat.} \\ \text{range}(\boldsymbol{W}) \subseteq \text{range}(\boldsymbol{D}(\mathbb{M})) \cap \text{range}(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M}))}} \Delta(\boldsymbol{W}, \boldsymbol{D}(\mathbb{M})) - \text{tr}(\boldsymbol{W} \log \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})) \\ = &\sup_{\substack{\boldsymbol{W} \text{ dens.mat.} \\ \text{range}(\boldsymbol{W}) \subseteq \text{range}(\boldsymbol{D}(\mathbb{M})) \cap \text{range}(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M}))}} -\Delta(\boldsymbol{W}, (\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}|\mathbb{M}))/\boldsymbol{D}(\boldsymbol{y})). \end{aligned}$$

The sup clearly has the unique solution $\frac{\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})}{\boldsymbol{D}(\boldsymbol{y})}$ and the inf remains unchanged if the condition on the range of $\boldsymbol{W}$ is dropped. This gives us the statement of the theorem.

We still need to show (9.4). Since $\text{range}(\boldsymbol{W}) \subseteq \text{range}(\boldsymbol{A}) \cap \text{range}(\boldsymbol{B}) \overset{\text{OP1}}{=} \text{range}(\boldsymbol{A} \odot \boldsymbol{B})$,

$$\begin{aligned} \text{tr}(\boldsymbol{W} \log(\boldsymbol{A} \odot \boldsymbol{B})) &\overset{(9.3)}{=} \text{tr}(\boldsymbol{W} \log^+(\boldsymbol{A} \odot \boldsymbol{B})) \\ &\overset{(4.6)}{=} \text{tr}(\boldsymbol{W} \boldsymbol{P}_{\boldsymbol{A} \cap \boldsymbol{B}} (\log^+ \boldsymbol{A} + \log^+ \boldsymbol{B}) \boldsymbol{P}_{\boldsymbol{A} \cap \boldsymbol{B}}) \\ &= \text{tr}(\underbrace{\boldsymbol{P}_{\boldsymbol{A} \cap \boldsymbol{B}} \boldsymbol{W} \boldsymbol{P}_{\boldsymbol{A} \cap \boldsymbol{B}}}_{\boldsymbol{W}} (\log^+ \boldsymbol{A} + \log^+ \boldsymbol{B})) \\ &\overset{(9.3)}{=} \text{tr}(\boldsymbol{W}(\log \boldsymbol{A} + \log \boldsymbol{B})). \end{aligned}$$

□

We conclude with a discussion of the relationship between the conventional Bayes rule for probability vectors and the generalized Bayes rule for density matrices. Density matrices are determined by a probability vector of eigenvalues as well as an orthogonal eigensystem. An orthogonal system $\boldsymbol{w}_i$ turns the prior density matrix $\boldsymbol{D}(\mathbb{M})$ into the probability vector $(\text{tr}(\boldsymbol{D}(\mathbb{M}) \boldsymbol{w}_i \boldsymbol{w}_i^\top))$, which we call a *pinching* of $\boldsymbol{D}(\mathbb{M})$. Similarly the pinching of the data likelihood matrix $\boldsymbol{D}(\boldsymbol{y}|\mathbb{M})$ is the vector $(\text{tr}(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M}) \boldsymbol{w}_i \boldsymbol{w}_i^\top)) \in [0, 1]^n$. The idea is to express our Bayes rule for density matrices as the conventional Bayes rule for the pinched priors and likelihoods w.r.t. a certain eigensystem. That is, we want to be able to say that the generalized Bayes rule is the conventional Bayes rule for the "best" pinching.

The above outline is essentially true, but we need to pinch in the log domain. With (9.3), Property OP11 can be extended to

$$\text{tr}(\boldsymbol{A} \odot \boldsymbol{u}\boldsymbol{u}^\top) = e^{\boldsymbol{u}^\top \log \boldsymbol{A} \boldsymbol{u}}, \quad \text{for } \textit{any} \text{ unit } \boldsymbol{u} \text{ and symmetric positive definite matrix } \boldsymbol{A}. \quad (9.5)$$

We call $\text{tr}(\boldsymbol{A} \odot \boldsymbol{w}_i \boldsymbol{w}_i^\top)$ a *remote pinching* of $\boldsymbol{A}$. Since its components satisfy $\text{tr}(\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{w}_i \boldsymbol{w}_i^\top) \overset{\text{OP10}}{\leq} \text{tr}(\boldsymbol{D}(\mathbb{M}) \boldsymbol{w}_i \boldsymbol{w}_i^\top)$, the remote pinchings of $\boldsymbol{D}(\mathbb{M})$ must be normalized to form a probability vector.

We can rewrite the argument of the optimization problem for the generalized Bayes rule based on the eigendecomposition $\boldsymbol{\mathcal{W}} \boldsymbol{\omega} \boldsymbol{\mathcal{W}}^\top$ of the density matrix $\boldsymbol{W}$:

$$\begin{aligned} &\Delta(\boldsymbol{W}, \boldsymbol{D}(\mathbb{M})) - \text{tr}(\boldsymbol{W} \log \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})) \\ &\quad = \text{tr}(\boldsymbol{\omega} \boldsymbol{\mathcal{W}}^\top (\log \boldsymbol{W} - \log \boldsymbol{D}(\mathbb{M})) \boldsymbol{\mathcal{W}}) - \text{tr}(\boldsymbol{\omega} \boldsymbol{\mathcal{W}}^\top (\log \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})) \boldsymbol{\mathcal{W}}) \end{aligned}$$

$$
\begin{aligned}
&= \sum_i \omega_i (\log \omega_i - \boldsymbol{w}_i^\top (\log \boldsymbol{D}(\mathbb{M}))\boldsymbol{w}_i) - \sum_i \omega_i \boldsymbol{w}_i^\top (\log \boldsymbol{D}(\boldsymbol{y}|\mathbb{M}))\boldsymbol{w}_i \\
&\overset{(9.5)}{=} \sum_i \omega_i (\log \omega_i - \log \operatorname{tr}(\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{w}_i \boldsymbol{w}_i^\top)) - \sum_i \omega_i \log \operatorname{tr}(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M}) \odot \boldsymbol{w}_i \boldsymbol{w}_i^\top) \\
&= \Delta\left((\boldsymbol{\omega}_i), \left(\operatorname{tr}(\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{w}_i \boldsymbol{w}_i^\top)/Z_{\mathcal{W}}\right)\right) - \sum_i \omega_i \log \operatorname{tr}(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M}) \odot \boldsymbol{w}_i \boldsymbol{w}_i^\top) - \log Z_{\mathcal{W}},
\end{aligned}
$$

where the normalization $Z_{\mathcal{W}} = \sum_j \operatorname{tr}(\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{w}_j \boldsymbol{w}_j^\top)$ does not depend on the eigenvalues. By Theorem 3, the above is minimized w.r.t. $\boldsymbol{\omega}$ when $\boldsymbol{\omega} = (P_{\mathcal{W}}(M_i) P_{\mathcal{W}}(y|M_i)/P_{\mathcal{W}}(y))$, where $P_{\mathcal{W}}(M_i) := \operatorname{tr}(\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{w}_i \boldsymbol{w}_i^\top)/Z_{\mathcal{W}}$ is the normalized remote pinching of the prior and $P_{\mathcal{W}}(y|M_i) := \operatorname{tr}(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M}) \odot \boldsymbol{w}_i \boldsymbol{w}_i^\top)$ is the remote pinching of the data likelihood matrix. With this optimum choice of $\boldsymbol{\omega}$, the minimization problem of the generalized Bayes rule simplifies to

$$
\begin{aligned}
&\inf_{\mathcal{W}\mathcal{W}^\top = \boldsymbol{I}} -\log P_{\mathcal{W}}(y) - \log Z_{\mathcal{W}} \\
&\quad = \inf_{\mathcal{W}\mathcal{W}^\top = \boldsymbol{I}} -\log\left(\sum_i \operatorname{tr}(\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{w}_i \boldsymbol{w}_i^\top) \operatorname{tr}(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M}) \odot \boldsymbol{w}_i \boldsymbol{w}_i^\top)\right) \\
&\quad \overset{\text{OP15}}{=} \inf_{\mathcal{W}\mathcal{W}^\top = \boldsymbol{I}} -\log\left(\sum_i \operatorname{tr}((\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}|\mathbb{M}) \odot \boldsymbol{w}_i \boldsymbol{w}_i^\top)\right) \\
&\quad \overset{\text{OP10}}{\geq} \inf_{\mathcal{W}\mathcal{W}^\top = \boldsymbol{I}} -\log\left(\sum_i \operatorname{tr}((\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}|\mathbb{M}) \boldsymbol{w}_i \boldsymbol{w}_i^\top)\right) \\
&\quad = -\log \operatorname{tr}(\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})).
\end{aligned}
$$

The above inequality is tight iff $\mathcal{W}$ is an eigensystem of $\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})$. We conclude that the optimization problem for the generalized Bayes rule is optimized when $\mathcal{W}$ is an eigensystem of $\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})$ and the vector of eigenvalues $\boldsymbol{\omega}$ is conventional posterior derived from the normalized remote pinchings of the prior and the remote pinchings of the data likelihood.

9.2 Chaining of the Bayes rule

The conventional Bayes rule can be applied iteratively to a sequence of data and various cancellations occur. For the sake of simplicity we only consider two data points y_1, y_2:

$$
P(M_i|y_2, y_1) = \frac{P(M_i|y_1)P(y_2|M_i, y_1)}{P(y_2|y_1)} = \frac{P(M_i)P(y_1|M_i)P(y_2|M_i, y_1)}{P(y_2|y_1)P(y_1)}.
$$

The normalization can be rewritten as:

$$P(y_2|y_1)P(y_1) = \Bigg(\sum_i \underbrace{P(M_i|y_1)}_{\text{using (9.1)}} P(y_2|M_i, y_1)\Bigg)\Bigg(\sum_i P(M_i)P(y_1|M_i)\Bigg)$$
$$= \sum_i P(M_i)P(y_1|M_i)P(y_2|M_i, y_1) = P(y_2, y_1). \tag{9.6}$$

Analogously, by essentially applying the generalized Bayes rule (9.2) two times we get:

$$\boldsymbol{D}(\mathbb{M}|\boldsymbol{y}_2, \boldsymbol{y}_1) = \frac{\boldsymbol{D}(\mathbb{M}|\boldsymbol{y}_1) \odot \boldsymbol{D}(\boldsymbol{y}_2|\mathbb{M}, \boldsymbol{y}_1)}{\boldsymbol{D}(\boldsymbol{y}_2|\boldsymbol{y}_1)} = \frac{\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}_1|\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}_2|\mathbb{M}, \boldsymbol{y}_1)}{\boldsymbol{D}(\boldsymbol{y}_2|\boldsymbol{y}_1)\boldsymbol{D}(\boldsymbol{y}_1)}.$$

As in the diagonal case (9.6), the normalization can be rewritten into one term (by applying TP2 twice and then the generalized Bayes rule (9.2)):

$$\begin{aligned}\boldsymbol{D}(\boldsymbol{y}_2|\boldsymbol{y}_1)\boldsymbol{D}(\boldsymbol{y}_1) &= \operatorname{tr}(\boldsymbol{D}(\mathbb{M}|\boldsymbol{y}_1) \odot \boldsymbol{D}(\boldsymbol{y}_2|\mathbb{M}, \boldsymbol{y}_1))\operatorname{tr}(\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}_1|\mathbb{M})) \\ &= \operatorname{tr}\left(\frac{\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}_1|\mathbb{M})}{\operatorname{tr}(\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}_1|\mathbb{M}))} \odot \boldsymbol{D}(\boldsymbol{y}_2|\mathbb{M}, \boldsymbol{y}_1)\right)\operatorname{tr}(\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}_1|\mathbb{M}) \\ &= \operatorname{tr}(\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}_1|\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}_2|\mathbb{M}, \boldsymbol{y}_1)) = \boldsymbol{D}(\boldsymbol{y}_1, \boldsymbol{y}_2).\end{aligned}$$

Finally as in (8.1), we can upper bound the data probability $\boldsymbol{D}(\boldsymbol{y}_1, \boldsymbol{y}_2)$ in terms of the product of the expected variances for the two trials:

$$\begin{aligned}\boldsymbol{D}(\boldsymbol{y}_2, \boldsymbol{y}_1) &= \operatorname{tr}(\boldsymbol{D}(\mathbb{M}|\boldsymbol{y}_1) \odot \boldsymbol{D}(\boldsymbol{y}_2|\mathbb{M}, \boldsymbol{y}_1))\operatorname{tr}(\boldsymbol{D}(\mathbb{M}) \odot \boldsymbol{D}(\boldsymbol{y}_1|\mathbb{M})) \\ &\leq \operatorname{tr}(\boldsymbol{D}(\mathbb{M}|\boldsymbol{y}_1)\boldsymbol{D}(\boldsymbol{y}_2|\mathbb{M}, \boldsymbol{y}_1))\operatorname{tr}(\boldsymbol{D}(\mathbb{M})\boldsymbol{D}(\boldsymbol{y}_1|\mathbb{M})).\end{aligned}$$

9.3 Bounds

Recall the following conventional bound for the negative log-likelihood of the data i.t.o. the negative log-likelihood of the MAP estimator:

$$\begin{aligned}-\log P(y) &= -\log \sum_i P(y|M_i)P(M_i) \\ &\leq \min_i(-\log P(y|M_i) - \log P(M_i)).\end{aligned} \tag{9.7}$$

We will give analogous bound for density matrices. For this we need the following inequality: For any unit vector $\boldsymbol{m}$ and symmetric positive definite matrix $\boldsymbol{A}$:

$$-\log \boldsymbol{m}^\top \boldsymbol{A}\boldsymbol{m} \overset{\text{OP10}}{\leq} -\log \operatorname{tr}(\boldsymbol{A} \odot \boldsymbol{m}\boldsymbol{m}^\top) \overset{(9.5)}{=} -\boldsymbol{m}^\top(\log \boldsymbol{A})\boldsymbol{m}. \tag{9.8}$$

Using the fact that $\operatorname{tr}(\boldsymbol{A}) \geq \boldsymbol{m}^\top \boldsymbol{A}\boldsymbol{m}$, we can now prove an analogous MAP bound for the generalized probabilities:

$$\begin{aligned}-\log \boldsymbol{D}(\boldsymbol{y}) &= -\log \operatorname{tr}(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M}) \odot \boldsymbol{D}(\mathbb{M})) \\ &\leq \min_{\boldsymbol{m}}(-\log \boldsymbol{m}^\top(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M}) \odot \boldsymbol{D}(\mathbb{M}))\boldsymbol{m})\end{aligned}$$

$$\overset{(9.8)}{\leq} \min_{\boldsymbol{m}}(-\boldsymbol{m}^\top \log(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M}) \odot \boldsymbol{D}(\mathbb{M}))\boldsymbol{m})$$

$$\leq \min_{\boldsymbol{m}}(-\boldsymbol{m}^\top \log \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})\boldsymbol{m} - \boldsymbol{m}^\top \log \boldsymbol{D}(\mathbb{M})\boldsymbol{m}).$$

The last inequality becomes (9.4), when $\boldsymbol{m} \in \text{range}(\boldsymbol{D}(\mathbb{M})) \cap \text{range}(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M}))$. Otherwise, it holds trivially because $-\boldsymbol{m}^\top \log(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M}) \odot \boldsymbol{D}(\mathbb{M}))\boldsymbol{m} = +\infty$.

Intuitively, there are two domains: the probability domain and the log probability domain. The conventional bound (9.7) can also be written in the probability domain:

$$P(y) \geq \max_i P(M_i)P(y|M_i).$$

However for the generalized probability case, there does not seem to be a simple similar inequality in the probability domain. Throughout the paper we always notice that the matrix operations need to be done in the log domain.

In the conventional case $P(y)$ is also upper bounded by $\max_i P(y|M_i)$. For the generalized case, the analogous formula is the following, where μ_i and $\boldsymbol{m}_i$ are the eigenvalues/vectors of $\boldsymbol{D}(\mathbb{M})$ and $\boldsymbol{m}$ any unit direction:

$$\begin{aligned}
\boldsymbol{D}(\boldsymbol{y}) &= \text{tr}(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M}) \odot \boldsymbol{D}(\mathbb{M})) \\
&\leq \text{tr}(\boldsymbol{D}(\boldsymbol{y}|\mathbb{M})\boldsymbol{D}(\mathbb{M})) \\
&= \sum_i \mu_i \boldsymbol{m}_i^\top \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})\boldsymbol{m}_i \\
&\leq \max_i \boldsymbol{m}_i^\top \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})\boldsymbol{m}_i \\
&\leq \max_{\boldsymbol{m}} \boldsymbol{m}^\top \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})\boldsymbol{m}.
\end{aligned}$$

10 Summary of the probability calculus for density matrices

In this section we give a summary of all the rules of our calculus. The definitions are indicated with := and at the end we summarize the justification for our choice of definitions. Table 1 shows connections between different objects and the formulas that relate them.

10.1 Marginalization rules for joints of Sects. 5 and 6

MJ1 $\boldsymbol{D}(\boldsymbol{a}) := \text{tr}(\boldsymbol{D}(\mathbb{A})\boldsymbol{a}\boldsymbol{a}^\top) = \boldsymbol{a}^\top \boldsymbol{D}(\mathbb{A})\boldsymbol{a}$.
MJ2 $\boldsymbol{D}(\mathbb{A}) := \text{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A}, \mathbb{B}))$.
MJ3 $\boldsymbol{D}(\boldsymbol{a}, \boldsymbol{b}) := \text{tr}(\boldsymbol{D}(\mathbb{A}, \mathbb{B})(\boldsymbol{a} \otimes \boldsymbol{b})(\boldsymbol{a} \otimes \boldsymbol{b})^\top) = \text{tr}(\boldsymbol{D}(\mathbb{A}, \mathbb{B})(\boldsymbol{a}\boldsymbol{a}^\top \otimes \boldsymbol{b}\boldsymbol{b}^\top))$.
MJ4 $\boldsymbol{D}(\mathbb{A}, \boldsymbol{b}) := \text{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A}, \mathbb{B})(\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{b}\boldsymbol{b}^\top))$.
MJ5 $\boldsymbol{D}(\boldsymbol{a}, \boldsymbol{b}) = \text{tr}(\boldsymbol{D}(\mathbb{A}, \boldsymbol{b})\boldsymbol{a}\boldsymbol{a}^\top)$.

10.2 Conditional probability rules of Sect. 7

CP1 $\boldsymbol{D}(\mathbb{A}|\mathbb{B}) := \boldsymbol{D}(\mathbb{A}, \mathbb{B}) \odot (\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B}))^{-1}$.
CP2 $\boldsymbol{D}(\mathbb{A}|\boldsymbol{b}) := \frac{\boldsymbol{D}(\mathbb{A}, \boldsymbol{b})}{\text{tr}(\boldsymbol{D}(\mathbb{A}, \boldsymbol{b}))}$.
CP3 $\boldsymbol{D}(\boldsymbol{a}|\mathbb{B}) := \boldsymbol{D}(\boldsymbol{a}, \mathbb{B}) \odot \boldsymbol{D}(\mathbb{B})^{-1}$.
CP4 $\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b}) := \frac{\boldsymbol{D}(\boldsymbol{a}, \boldsymbol{b})}{\boldsymbol{D}(\boldsymbol{b})}$.

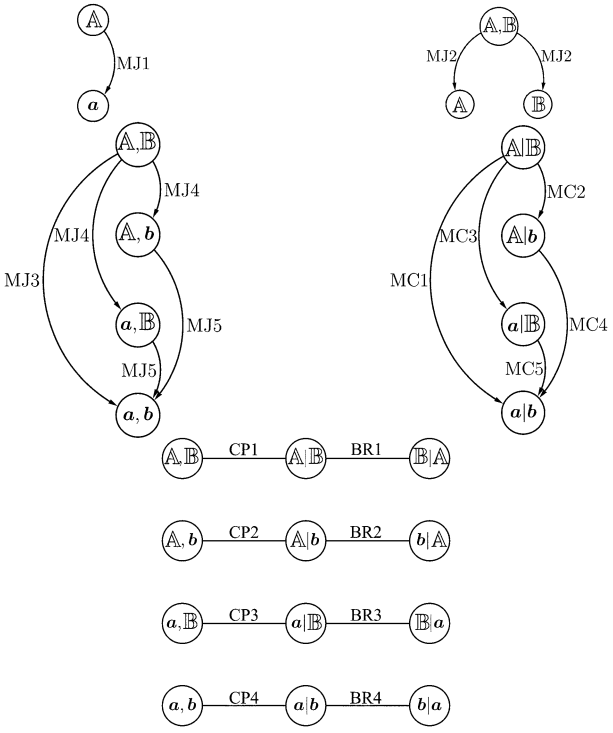


Table 1 A series of charts summarizing the different relationships for joints and conditionals. Each edge references the formula stating the relationship. For symmetric cases only one formula is given and the corresponding edges in the chart will have the same label

CP1 has the form: density matrix $\odot$ inverse of a normalization. Below we reexpress the other definitions in this unified form:

CP′2 $\boldsymbol{D}(\mathbb{A}|\boldsymbol{b}) = \mathrm{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{b}\boldsymbol{b}^{\top})) \odot \mathrm{tr}_{\mathbb{B}}((\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B}))(\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{b}\boldsymbol{b}^{\top}))^{-1}$.
CP′3 $\boldsymbol{D}(\boldsymbol{a}|\mathbb{B}) = \mathrm{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{a}\boldsymbol{a}^{\top} \otimes \boldsymbol{I}_{\mathbb{B}})) \odot \mathrm{tr}_{\mathbb{A}}((\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B}))(\boldsymbol{a}\boldsymbol{a}^{\top} \otimes \boldsymbol{I}_{\mathbb{B}}))^{-1}$.
CP′4 $\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b}) = \mathrm{tr}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{a}\boldsymbol{a}^{\top} \otimes \boldsymbol{b}\boldsymbol{b}^{\top})) \odot \mathrm{tr}((\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B}))(\boldsymbol{a}\boldsymbol{a}^{\top} \otimes \boldsymbol{b}\boldsymbol{b}^{\top}))^{-1}$.

10.3 Marginalization rules for conditionals of Sect. 7

MC1 $\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b}) = \frac{\mathrm{tr}(\boldsymbol{D}(\mathbb{A}|\mathbb{B}) \odot (\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B}))(\boldsymbol{a}\boldsymbol{a}^{\top} \otimes \boldsymbol{b}\boldsymbol{b}^{\top}))}{\mathrm{tr}(\boldsymbol{D}(\mathbb{B})\boldsymbol{b}\boldsymbol{b}^{\top})}$.
MC2 $\boldsymbol{D}(\mathbb{A}|\boldsymbol{b}) = \frac{\mathrm{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A}|\mathbb{B}) \odot (\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B}))(\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{b}\boldsymbol{b}^{\top}))}{\mathrm{tr}(\boldsymbol{D}(\mathbb{B})\boldsymbol{b}\boldsymbol{b}^{\top})}$.
MC3 $\boldsymbol{D}(\boldsymbol{a}|\mathbb{B}) = \mathrm{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{A}|\mathbb{B}) \odot (\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B}))(\boldsymbol{a}\boldsymbol{a}^{\top} \otimes \boldsymbol{I}_{\mathbb{B}})) \odot \boldsymbol{D}(\mathbb{B})^{-1}$.
MC4 $\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b}) = \mathrm{tr}(\boldsymbol{D}(\mathbb{A}|\boldsymbol{b})\boldsymbol{a}\boldsymbol{a}^{\top})$.
MC5 $\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b}) = \frac{\mathrm{tr}((\boldsymbol{D}(\boldsymbol{a}|\mathbb{B}) \odot \boldsymbol{D}(\mathbb{B}))\boldsymbol{b}\boldsymbol{b}^{\top})}{\mathrm{tr}(\boldsymbol{D}(\mathbb{B})\boldsymbol{b}\boldsymbol{b}^{\top})}$.

All the rules here except for MC4 require additional information for marginalization, which was not necessary in the conventional case. See discussion of marginalization of conditionals in Sect. 7.

10.4 Theorems of total probability of Sect. 8

TP1 $\boldsymbol{D}(\boldsymbol{a}) = \sum_i \boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b}_i)\boldsymbol{D}(\boldsymbol{b}_i)$ for any orthogonal system $\boldsymbol{b}_i$ of space $\mathbb{B}$.
TP2 $\boldsymbol{D}(\mathbb{A}) = \sum_i \boldsymbol{D}(\mathbb{A}, \boldsymbol{b}_i)$ for any orthogonal system $\boldsymbol{b}_i$ of space $\mathbb{B}$.
TP3 $\boldsymbol{D}(\mathbb{A}) = \mathrm{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A}|\mathbb{B}) \odot (\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B})))$.

10.5 Bayes rules of Sect. 9

BR1 $\boldsymbol{D}(\mathbb{B}|\mathbb{A}) = (\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B})) \odot \boldsymbol{D}(\mathbb{A}|\mathbb{B}) \odot (\boldsymbol{D}(\mathbb{A}) \otimes \boldsymbol{I}_{\mathbb{B}})^{-1}$, where $\boldsymbol{D}(\mathbb{A}) = \mathrm{tr}_{\mathbb{B}}((\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B})) \odot \boldsymbol{D}(\mathbb{A}|\mathbb{B}))$.
BR2 $\boldsymbol{D}(\boldsymbol{a}|\mathbb{B}) = \boldsymbol{D}(\boldsymbol{a})\boldsymbol{D}(\mathbb{B}|\boldsymbol{a}) \odot \boldsymbol{D}(\mathbb{B})^{-1}$, where $\boldsymbol{D}(\mathbb{B}) = \mathrm{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{B}|\mathbb{A}) \odot (\boldsymbol{D}(\mathbb{A}) \otimes \boldsymbol{I}_{\mathbb{B}}))$.
BR3 $\boldsymbol{D}(\mathbb{B}|\boldsymbol{a}) = \frac{\boldsymbol{D}(\mathbb{B}) \odot \boldsymbol{D}(\boldsymbol{a}|\mathbb{B})}{\boldsymbol{D}(\boldsymbol{a})}$, where $\boldsymbol{D}(\boldsymbol{a}) = \mathrm{tr}(\boldsymbol{D}(\mathbb{B}) \odot \boldsymbol{D}(\boldsymbol{a}|\mathbb{B}))$.
BR4 $\boldsymbol{D}(\boldsymbol{b}|\boldsymbol{a}) = \frac{\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b})\boldsymbol{D}(\boldsymbol{b})}{\boldsymbol{D}(\boldsymbol{a})}$, where $\boldsymbol{D}(\boldsymbol{a}) = \sum_i \boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b}_i)\boldsymbol{D}(\boldsymbol{b}_i)$ and the summation is over any orthogonal system $\boldsymbol{b}_i$.

10.6 Summary of justifications for the definitions

Note that only the rules MJ1–4 and CP rules are definitions. Everything else in our calculus can be derived from these. MJ1 is justified by Gleason's Theorem as discussed in Sect. 3. Gleason's Theorem also justifies MJ3, where the Kronecker product provides the natural way to specify a joint unit (see discussion in Sect. 5). MJ2 is standard in quantum physics and $\mathrm{tr}_{\mathbb{B}}(\boldsymbol{D}(\mathbb{A}, \mathbb{B}))$ was shown to be a density matrix in Lemma 1. The rule is also compatible with the conventional case as well as with the natural generalization of independence discussed in Sect. 6. MJ4 is the natural definition of $\boldsymbol{D}(\mathbb{A}, \boldsymbol{b})$ that satisfies MJ5 and is compatible with the conventional case.

We will outline how CP2 can be motivated as a quantum relative entropy projection. For positive definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we extend the definition of quantum relative entropy as follows: $\Delta(\boldsymbol{A}, \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}(\log \boldsymbol{A} - \log \boldsymbol{B}) + \boldsymbol{B} - \boldsymbol{A})$. Note that this "unnormalized" relative entropy, coincides with the standard one when $\boldsymbol{A}$ and $\boldsymbol{B}$ have trace one. Now CP2 is motivated as

$$\boldsymbol{D}(\mathbb{A}|\boldsymbol{b}) = \operatorname*{arg\,inf}_{\boldsymbol{W}\ \text{dens. mat.}} \Delta(\boldsymbol{W}, \boldsymbol{D}(\mathbb{A}, \boldsymbol{b})).$$

CP3 is motivated analogous to the generalized Bayes rule (see Sect. 9):

$$\boldsymbol{D}(\boldsymbol{a}, \boldsymbol{B}) = \operatorname*{arg\,inf}_{\boldsymbol{W}} \Delta(\boldsymbol{W}, \boldsymbol{D}(\mathbb{B})) - \mathrm{tr}(\boldsymbol{W} \log \boldsymbol{D}(\boldsymbol{a}|\mathbb{B})).$$

CP1 can be motivated in a similar fashion, but now the variable is over the joint space $(\mathbb{A}, \mathbb{B})$:

$$\boldsymbol{D}(\boldsymbol{A}, \boldsymbol{B}) = \operatorname*{arg\,inf}_{\boldsymbol{W}\ \text{dens.mat.}} \Delta(\boldsymbol{W}, \boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B})) - \mathrm{tr}(\boldsymbol{W} \log \boldsymbol{D}(\mathbb{A}|\mathbb{B})).$$

CP1 also was previously used in Cerf and Adami (1999) to allow a suitable definition of conditional quantum entropy. Finally, the last rule CP4 was chosen in analogy to the conventional case. It also has an interpretation as two successive quantum measurements (see Appendix).

Historically, we first justified the generalized Bayes rule BR3 based on the minimum relative entropy principle (see Warmuth 2005 and Sect. 9). After that we chose definitions CP1–CP4 to be compatible with this generalized Bayes rule.

11 Conclusions

Density matrices are central to quantum physics. We utilize many mathematical techniques from that field to develop a Bayesian probability calculus for density matrices. Intuitively, the new calculus will be useful when the data likelihood $\boldsymbol{D}(\boldsymbol{y}|\mathbb{M})$ has non-zero off-diagonal elements, i.e. information about which components are correlated or anti-correlated. The main new operation $\boldsymbol{A} \odot \boldsymbol{B}$ first takes logs of the matrices adds the logs and finally exponentiates. Any straightforward implementation of the $\odot$ operation requires the eigendecompositions of the matrices, which are expensive to obtain. Throughout our work we notice that the log domain seems to be more important in the matrix case.

Interestingly enough the $\odot$ operation has also been employed in computer graphics for combining affine transformation (Alexa 2002). Also the simulation of quantum computations based on the Lie Trotter Formula (Nielsen and Chuang 2000, Chap. 4.7) can be interpreted as applying the $\odot$ operation to unitary matrices and not to symmetric positive definite matrices as we do in this paper.

The main update in quantum physic is a unitary evolution of the current density matrix $\boldsymbol{A}$, i.e. $\boldsymbol{A} := \boldsymbol{U}\boldsymbol{A}\boldsymbol{U}^\top$, where $\boldsymbol{U}$ is unitary. For example, the main differential equation for density matrices in quantum physics is the following version of the Schrödinger Equation (Feynman 1972):

$$\frac{\partial \boldsymbol{D}(\mathbb{M}|t)}{\partial t} = i(\boldsymbol{H}\boldsymbol{D}(\mathbb{M}|t) - \boldsymbol{D}(\mathbb{M}|t)\boldsymbol{H}), \quad \text{where } \boldsymbol{H} \text{ is skew Hermitian.}$$

The solution has the form

$$\boldsymbol{D}(\mathbb{M}|t) = \exp(-it\boldsymbol{H})\boldsymbol{D}(\mathbb{M}|0)\exp(it\boldsymbol{H}),$$

where $\boldsymbol{D}(\mathbb{M}|0)$ is the initial density matrix. Since $it\boldsymbol{H}$ is skew Hermitian, both exponentials are unitary. Thus the above update represents a unitary transformation of the initial density matrix $\boldsymbol{D}(\mathbb{M}|0)$. Such transformations leave the eigenvalues unchanged and only affect the eigensystem. In contrast our generalized Bayes rule updates both the eigenvalues and eigenvectors, and the conventional Bayes rule can be seen as only updating the eigenvalues while keeping the eigenvectors fixed. Therefore the Bayes rules are decidedly not unitary updates.

For the sake of completeness we now express the Bayes rules also as solutions to differential equations. In the conventional case, the differential equations are ($1 \leq i \leq n$):

$$\frac{\partial \log P(M_i|t)}{\partial t} = \log P(y|M_i) - \sum_j P(M_j|t)\log P(y|M_j).$$

The solution is

$$P(M_i|t) = \frac{P(M_i|0)P(y|M_i)^t}{\sum_j P(M_j|0)P(y|M_j)^t}.$$

If we take the value $P(M_i|0)$ as the prior $P(M_i)$ then the expression for $P(M_i|1)$ becomes the conventional Bayes rule (9.1). There is a similar differential equation for the generalized Bayes rule (for the sake of simplicity we assume that the prior $\boldsymbol{D}(\mathbb{M})$ and data likelihood matrix $\boldsymbol{D}(\boldsymbol{y}|\mathbb{M})$ are strictly positive definite):

$$\frac{\partial \log \boldsymbol{D}(\mathbb{M}|t)}{\partial t} = \log \boldsymbol{D}(\boldsymbol{y}|\mathbb{M}) - \operatorname{tr}(\boldsymbol{D}(\mathbb{M}|t)\log \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})).$$

The solution has the form

$$\boldsymbol{D}(\mathbb{M}|t) = \frac{\exp(\log \boldsymbol{D}(\mathbb{M}|0) + t \log \boldsymbol{D}(\boldsymbol{y}|\mathbb{M}))}{\operatorname{tr}(\exp(\log \boldsymbol{D}(\mathbb{M}|0) + t \log \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})))} \overset{(4.1)}{=} \frac{\boldsymbol{D}(\mathbb{M}|0) \odot \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})^t}{\operatorname{tr}(\boldsymbol{D}(\mathbb{M}|0) \odot \boldsymbol{D}(\boldsymbol{y}|\mathbb{M})^t)}.$$

If we set $\boldsymbol{D}(\mathbb{M}|0)$ to the prior $\boldsymbol{D}(\mathbb{M})$, then the expression for $\boldsymbol{D}(\mathbb{M}|1)$ becomes the generalized Bayes rule (9.2). Notice again that the differential equations emphasize the log domain and that the $\odot$ operation appears in the solution.

At this point we have no convincing application for the new probability calculus. However, a similar methodology was used to derive and prove bounds for parameter updates of density matrices that led to a version of Boosting (Tsuda et al. 2005) where the distribution over the examples is replaced by a density matrix, an online variance minimization algorithm where the parameter space is the unit ball (Warmuth and Kuzmin 2006), and an on-line algorithm for Principal Component Analysis (Warmuth and Kuzmin 2008).

In this paper our parameters expressing the uncertainty are symmetric positive definite matrices. However using essentially the EG± transformation (Kivinen and Warmuth 1997), it has been shown recently that inference can be done with arbitrarily shaped matrices (Warmuth 2007). This leaves the strong possibility that the calculus developed here will generalize to arbitrary shaped matrices as well. In that case the elementary events are "asymmetric dyads" $\boldsymbol{u}\boldsymbol{v}^\top$ and the underlying decomposition is the SVD decomposition.

The new calculus seems to be rich enough to bring out some of the interesting phenomena of quantum physics, such as superposition and entanglement. Maybe the new calculus can be used to maintain "uncertainty" in quantum computation.

On a more technical note, we conjecture that for all non-decoupled joints $\boldsymbol{D}(\mathbb{A}, \mathbb{B})$ there is a one-to-one mapping to the conditionals $\boldsymbol{D}(\mathbb{A}|\mathbb{B})$, and the EM-like algorithm given in Sect. 7 converges to $\boldsymbol{D}(\mathbb{B})$, s.t. $\boldsymbol{D}(\mathbb{A}, \mathbb{B}) = \boldsymbol{D}(\mathbb{A}|\mathbb{B}) \odot (\boldsymbol{I}_{\mathbb{A}} \otimes \boldsymbol{D}(\mathbb{B}))$.

Finally, we will reason in a simple case that generalized probability space is more "connected" and a clever algorithm might be able to exploit this. Assume zero is encoded as the distribution $(1, 0)$ and one as the distribution $(0, 1)$. Moving from the zero distribution to the one distributions can be done by lowering the probability of the first component and increasing the probability of the second. As density matrices, zero and one would be $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, respectively. Note that the eigensystem for both matrices is the identity matrix and there is now a second way to go from zero to one that keeps the eigenvalues/probabilities fixed but swaps the eigenvectors:

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Acknowledgements We are grateful to Allen Van Gelder for making us aware of Alexa (2002). Also many thanks to Torsten Ehrhardt who first proved to us the range intersection property OP1 and the $\log^+$ formula OP2 for the $\odot$ operation.

Appendix: Quantum-mechanical interpretation of conditional probability $\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b})$

We will now show how to interpret the conditional probability $\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b})$ in terms of two quantum measurements. The two measurements will be performed one after another on the joint density $\boldsymbol{D}(\mathbb{A},\mathbb{B})$ and $\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b})$ will be a probability of outcome 1 for the second measurement given the first measurement had outcome 1. First, we measure $\boldsymbol{D}(\mathbb{A},\mathbb{B})$ with event $\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{b}\boldsymbol{b}^{\top}$. Assume that we get outcome 1. Using the generalization of collapse rule for events (see e.g. Nielsen and Chuang 2000), the successor density matrix can be computed as follows:

$$\widehat{\boldsymbol{D}}(\mathbb{A},\mathbb{B})=\frac{(\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{b}\boldsymbol{b}^{\top})\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{b}\boldsymbol{b}^{\top})}{\operatorname{tr}((\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{b}\boldsymbol{b}^{\top})\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{b}\boldsymbol{b}^{\top}))}.$$

The second measurement consists of measuring the updated joint with event $\boldsymbol{a}\boldsymbol{a}^{\top}\otimes\boldsymbol{I}_{\mathbb{B}}$. Now the probability for getting outcome 1 is computed as:

$$\begin{aligned}\operatorname{tr}(\widehat{\boldsymbol{D}}(\mathbb{A},\mathbb{B})(\boldsymbol{a}\boldsymbol{a}^{\top}\otimes\boldsymbol{I}_{\mathbb{B}})) &= \frac{\operatorname{tr}((\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{b}\boldsymbol{b}^{\top})\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{b}\boldsymbol{b}^{\top})(\boldsymbol{a}\boldsymbol{a}^{\top}\otimes\boldsymbol{I}_{\mathbb{B}}))}{\operatorname{tr}((\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{b}\boldsymbol{b}^{\top})\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{b}\boldsymbol{b}^{\top}))}\\ &\overset{\text{KP2+cycle}}{=}\frac{\operatorname{tr}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{a}\boldsymbol{a}^{\top}\otimes\boldsymbol{b}\boldsymbol{b}^{\top}))}{\operatorname{tr}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{b}\boldsymbol{b}^{\top}))}=\frac{\boldsymbol{D}(\boldsymbol{a},\boldsymbol{b})}{\operatorname{tr}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{b}\boldsymbol{b}^{\top}))}.\end{aligned}$$

The denominator can be simplified using partial trace properties:

$$\operatorname{tr}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{b}\boldsymbol{b}^{\top}))\overset{\text{PT2}}{=}\operatorname{tr}(\operatorname{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{A},\mathbb{B})(\boldsymbol{I}_{\mathbb{A}}\otimes\boldsymbol{b}\boldsymbol{b}^{\top})))\overset{\text{PT3}}{=}\operatorname{tr}(\underbrace{\operatorname{tr}_{\mathbb{A}}(\boldsymbol{D}(\mathbb{A},\mathbb{B}))}_{\boldsymbol{D}(\mathbb{B})}\boldsymbol{b}\boldsymbol{b}^{\top})=\boldsymbol{D}(\boldsymbol{b}).$$

Therefore the probability of outcome 1 on the second measurement (given the first outcome was 1) is:

$$\operatorname{tr}(\widehat{\boldsymbol{D}}(\mathbb{A},\mathbb{B})(\boldsymbol{a}\boldsymbol{a}^{\top}\otimes\boldsymbol{I}_{\mathbb{B}}))=\frac{\boldsymbol{D}(\boldsymbol{a},\boldsymbol{b})}{\boldsymbol{D}(\boldsymbol{b})}\overset{\text{CP4}}{=}\boldsymbol{D}(\boldsymbol{a}|\boldsymbol{b}).$$

References

Alexa, M. (2002). Linear combination of transformations. In *SIGGRAPH'02: Proceedings of the 29th annual conference on computer graphics and interactive techniques* (pp. 380–387). New York: ACM Press.

Bernstein, D. S. (2005). *Matrix mathematics: theory, facts, and formulas with application to linear systems theory*. Princeton: Princeton University Press.

Bhatia, R. (1997). *Matrix analysis*. Berlin: Springer.

Bužek, V., Drobný, G., Derka, R., Adam, G., & Wiedemann, H. (1999). Quantum state reconstruction from incomplete data. *Chaos Solitons Fractals*, *10*, 981–1074.

Caves, C. M., Fuchs, C. A., Manne, K. K., & Renes, J. M. (2004). Gleason-type derivations of the quantum probability rule for generalized measurements. *Foundations of Physics*, *34*, 193–209.

Cerf, N. J., & Adami, C. (1999). Quantum extension of conditional probability. *Physical Review A*, *60*(2), 893–897.

Feynman, R. P. (1972). *Statistical mechanics: a set of lectures*. Reading: Addison-Wesley.

Gleason, A. (1957). Measures on the closed subspaces of a Hilbert space. *Indiana University Mathematics Journal*, *6*, 885–893.

Holevo, A. S. (2001). *Lecture notes in physics. Monographs*: *Vol. 67. Statistical structure of quantum theory*, Berlin, New York: Springer.

Kato, T. (1978). Trotter's product formula for an arbitrary pair of self-adjoint contraction semigroups. *Topics in Functional Analysis (Advances in Mathematics—Supplementary Studies)*, *3*, 185–195.

Kivinen, J., & Warmuth, M. K. (1997). Additive versus exponentiated gradient updates for linear prediction. *Information and Computation*, *132*(1), 1–64.

Kivinen, J., & Warmuth, M. K. (1999). Averaging expert predictions. In *Lecture notes in artificial intelligence: Vol. 1572. Computational learning theory, 4th European conference (EuroCOLT'99), Nordkirchen, Germany, March 29–31, 1999, Proceedings* (pp. 153–167). Berlin: Springer.

Nielsen, M. A., & Chuang, I. L. (2000). *Quantum computation and quantum information*. Cambridge: Cambridge University Press.

Olivares, S., & Paris, M. G. A. Quantum estimation via the minimum Kullback entropy principle. *Physical Review A*, 76, 2007.

Schack, R., Brun, T. A., & Caves, C. M. (2001). Quantum Bayes rule. *Physical Review A*, *64*, 014305.

Simon, B. (1979). *Functional integration and quantum physics*. San Diego: Academic Press.

Singh, R., Warmuth, M. K., Raj, B., & Lamere, P. (2003). Classification with free energy at raised temperatures. In *Proc. of EUROSPEECH 2003*, *September 2003* (pp. 1773–1776)

Tsuda, K., Raätsch, G., & Warmuth, M. K. (2005). Matrix exponentiated gradient updates for on-line learning and Bregman projections. *Journal of Machine Learning Research*, *6*, 995–1018.

Warmuth, M. K. (2005). Bayes rule for density matrices. In *Advances in neural information processing systems 18 (NIPS'05)*. Cambridge: MIT Press.

Warmuth, M. K. (2007). Winnowing subspaces. In *Proceedings of the 24th international conference on machine learning (ICML'07)*. New York: ACM.

Warmuth, M. K., & Kuzmin, D. (2006). Online variance minimization. In *Proceedings of the 19th annual conference on learning theory (COLT'06), Pittsburg, June 2006*. New York: Springer.

Warmuth, M. K., & Kuzmin, D. (2008). Randomized PCA algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, *9*, 2217–2250.

Zellner, A. (1998). Optimal information processing and Bayes's theorem. *The American Statistician*, *42*(4), 278–284.