Properties of Classical and Quantum Jensen-Shannon Divergence

Jop Briët* and Peter Harremoës†
Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands
(Dated: April 15, 2009)

Jensen-Shannon divergence (JD) is a symmetrized and smoothed version of the most important divergence measure of information theory, Kullback divergence. As opposed to Kullback divergence it determines in a very direct way a metric; indeed, it is the square of a metric. We consider a family of divergence measures (JD_α for $\alpha > 0$), the *Jensen divergences of order* α, which generalize JD as $\mathrm{JD}_1 = \mathrm{JD}$. Using a result of Schoenberg, we prove that JD_α is the square of a metric for $\alpha \in (0, 2]$, and that the resulting metric space of probability distributions can be isometrically embedded in a real Hilbert space. *Quantum Jensen-Shannon divergence* (QJD) is a symmetrized and smoothed version of quantum relative entropy and can be extended to a family of *quantum Jensen divergences of order* α (QJD_α). We strengthen results by Lamberti et al. by proving that for qubits and pure states, $\mathrm{QJD}_\alpha^{1/2}$ is a metric space which can be isometrically embedded in a real Hilbert space when $\alpha \in (0, 2]$. In analogy with Burbea and Rao's generalization of JD, we also define *general* QJD by associating a Jensen-type quantity to any weighted family of states. Appropriate interpretations of quantities introduced are discussed and bounds are derived in terms of the total variation and trace distance.

PACS numbers: 89.70.Cf, 03.67.-a

I. INTRODUCTION

For two probability distributions $P = (p_1, \ldots, p_n)$ and $Q = (q_1, \ldots, q_n)$ on a finite alphabet of size $n \geq 2$, Jensen-Shannon divergence (JD) is a measure of divergence between P and Q. It measures the deviation between the Shannon entropy of the mixture $(P+Q)/2$ and the mixture of the entropies, and is given by

$$\mathrm{JD}(P, Q) = H\left(\frac{P+Q}{2}\right) - \frac{1}{2}\big(H(P) + H(Q)\big). \qquad (1)$$

Attractive features of this function are that it is everywhere defined, bounded, symmetric and only vanishes when $P = Q$. Endres and Schindelin [1] proved that it is the square of a metric, which we call the *transmission metric* ($\mathrm{d_T}$). This result implies, for example, that Banach's fixed point theorem holds for the space of probability distributions endowed with the metric $\mathrm{d_T}$. A natural way to extend Jensen-Shannon divergence is to consider a mixture of k probability distributions $P_1, \ldots, P_k$, with weights $\pi_1, \ldots, \pi_k$, respectively. With $\pi = (\pi_1, \ldots, \pi_k)$, we can then define the *general Jensen divergence* as

$$\mathrm{JD}^\pi(P_1, \ldots, P_k) = H\left(\sum_{i=1}^k \pi_i P_i\right) - \sum_{i=1}^k \pi_i H(P_i).$$

This was already considered by Gallager [2] in 1968, who proved that, for fixed π, this is a convex function in $(P_1, \cdots, P_k)$. Further identities and inequalities were derived by Lin and Wong [3, 4], and Topsøe [5]. It has found a variety of important applications: Sibson [6] showed that it has applications in biology and cluster analysis, Wong and You [7] used it as a measure of distance between random graphs, and recently, Rosso et al. used it to quantify the deterministic vs. the stochastic part of a time series [8]. For its statistical applications we refer to El-Yaniv et al. [9] and references therein.

Burbea and Rao [10] introduced another level of generalization, based on more general entropy functions. For an interval I in $\mathbb{R}$ and a function $\phi : I \to \mathbb{R}$, they define the *ϕ-entropy* of $x \in I^n$ (where I^n denotes the Cartesian product of n copies of I) as

$$H_\phi(x) = -\sum_{i=1}^n \phi(x_i).$$

Based on this, they define the *generalized mutual information measure* as

$$\mathrm{JD}_\phi^\pi(P_1, \ldots, P_k) = H_\phi\left(\sum_{i=1}^k \pi_i P_i\right) - \sum_{i=1}^k \pi_i H_\phi(P_i),$$

for which they established some strong convexity properties. If $k = 2$, $I = [0, 1]$ and ϕ is the function $x \to \frac{1}{\alpha-1}(x^\alpha - x)$, then H_ϕ defines the *entropy of order* α. In this case, Burbea and Rao proved that JD_ϕ^π is convex for all π, if and only if $\alpha \in [1, 2]$, except if $n = 2$ when convexity holds if and only if $\alpha \in [1, 2]$ or $\alpha \in [3, 11/3]$.

We focus on the functions JD_ϕ^π, where $k \geq 2$, $I = [0, 1]$ and ϕ defines entropy of order α. For ease of notation we write these as JD_α^π if $k \geq 2$ and as JD_α if $k = 2$ and $\pi = (1/2, 1/2)$.

Shannon entropy is *additive* in the sense that the entropy of independent random variables, defined as the entropy of their joint distribution, is the sum of their

*Electronic address: jop.briet@cwi.nl.
†Electronic address: P.Harremoes@cwi.nl.

individual entropies. Like Shannon entropy Rényi of order α entropy is additive but in general Rényi entropy is not convex [26]. The power entropy of order α is a monotone function of Rényi entropy but, contrary to Rényi entropy, it is a concave function which is what we are interested in. The study of power entropy dates back to J.H. Havrda and F. Charvat [27]. Since then it was rediscovered independently several times [11, 12, 28], but we have chosen the more neutral term entropy of order α rather than calling it Havrda-Chervat-Lindhardt-Nielsen-Aczél-Dar'oczy-Tsallis entropy. Entropy of order α is not additive (unless $\alpha = 1$). This is one of the reasons why this function is used by physicists in attempts to model long range interaction in statistical mechanics, cf. Tsallis [11] and followers (can be traced from a bibliography maintained by Tsallis).

Martins et al. [13–16] give non-extensive (i.e. non-additive) generalizations of JD based on entropies of order α and an extension of the concept of convexity to what they call *q-convexity*. For these functions they extend Burbea and Rao's results in terms of q-convexity.

Distance measures between quantum states, which generalize probability distributions, are of great interest to the field of quantum information theory [17–21]. They play a central role in state discrimination and in quantifying entanglement. For example, the quantum relative entropy of two states ρ_1 and ρ_2, given by $S(\rho_1\|\rho_2) = -\mathrm{Tr}\rho_1(\ln\rho_1 - \ln\rho_2)$, is a commonly used distance measure. (For a review of its basic properties and applications see [22]). However, it is not symmetric and does not obey the triangle inequality. As an alternative, Lamberti et al. [21, 23, 24] proposed to use the (classical) JD as a distance function for quantum states, but also introduced a quantum version based on the von Neumann entropy, which we denote by QJD. Like its classical variant, it is everywhere defined, bounded, symmetric and zero only when the inputs are two identical quantum states. They prove that it is a metric on the set of pure quantum states and that it is close to the Wootter's distance and its generalization introduced by Braunstein and Caves [18]. Whether the metric property holds in general is unknown.

As an analogue to $\mathrm{JD}^{\pi}_{\alpha}$ for quantum states, we introduce the *general quantum Jensen divergence of order* α ($\mathrm{QJD}^{\pi}_{\alpha}$). In the limit $\alpha \to 1$ we obtain the "von Neumann version":

$$\mathrm{QJD}^{\pi}(\rho_1, \ldots, \rho_k) = S\left(\sum_{i=1}^{k} \pi_i \rho_i\right) - \sum_{i=1}^{k} \pi_i S(\rho_i),$$

where $S(\rho) = -\mathrm{Tr}\rho\ln\rho$ is the von Neumann entropy. For $k = 2$ and $\pi = (1/2, 1/2)$ one obtains the *quantum Jensen divergence of order* α (QJD_{α}), which generalizes QJD as $\lim_{\alpha\to 1} \mathrm{QJD}_{\alpha} = \mathrm{QJD}$.

1. Our results.

We extend the results of Endres and Schindelin, concerning the metric property of JD, and those of Lamberti et al., concerning the metric property of QJD, as follows:

- Denoting the set of probability distributions on a set X by $M^1_+(X)$, we prove that for $\alpha \in (0, 2]$, the pair $\left(M^1_+(X), \mathrm{JD}^{1/2}_{\alpha}\right)$ is a metric space which can be isometrically embedded in a real separable Hilbert space.
- Denoting the set of quantum states on qubits (2-dimensional Hilbert spaces) by $\mathcal{B}^1_+(\mathcal{H}_2)$ and the set of pure-states on d-dimensional Hilbert spaces by $\mathcal{P}(\mathcal{H}_d)$, we prove that for $\alpha \in (0, 2]$, the pairs $\left(\mathcal{B}^1_+(\mathcal{H}_2), \mathrm{QJD}^{1/2}_{\alpha}\right)$ and $\left(\mathcal{P}(\mathcal{H}_d), \mathrm{QJD}^{1/2}_{\alpha}\right)$ are metric spaces which can be isometrically embedded in a real separable Hilbert space.
- We show that these results do *not* extend to the cases $\alpha \in (2, 3)$ and $\alpha \in (\frac{7}{2}, \infty)$. More precisely, we show that, for $\alpha \in (2, 3)$, neither JD_{α} nor QJD_{α} can be the square of a metric, and for $\alpha \in (\frac{7}{2}, \infty)$, isometric embedding in a real Hilbert space is impossible (though the metric property may still hold).

2. Techniques.

To prove our positive results, we evoke a theorem by Schoenberg which links Hilbert-space embeddability of a metric space (X, d) to the property of *negative definiteness* (defined in Section IV). We prove that for $\alpha \in (0, 2]$, JD_{α} satisfies this condition for every set of probability distributions, and that QJD_{α} satisfies this condition for every set of qubits or pure-states.

A. Interpretations of JD^{π} and QJD^{π}

1. Channel capacity.

A *discrete memoryless channel* is a system with input and output alphabets X and Y respectively, and conditional probabilities $p(y|x)$ for the probability that $y \in Y$ is received when $x \in X$ is sent. For a discrete memoryless channel with $|X| = k$, input distribution π over X and conditional distributions $P_x(y) = p(y|x)$, we have that $\mathrm{JD}^{\pi}\left(P_{x_1}, \ldots, P_{x_k}\right)$ in fact gives the transmission rate. (See for example [25].) Inspired by this fact, we call the metric defined by the square root of JD the *transmission metric* and denote it by $\mathrm{d_T}$.

A *quantum channel* has classical input alphabet X, and an encoding of every element $x \in X$ into a quantum state ρ_x. A receiver decodes a message by performing a

measurement with $|Y|$ possible outcomes, on the state he or she obtained. For a quantum channel with $|X| = k$, input distribution π over X, and encoded elements ρ_x, Holevo's Theorem [29] says that the maximum transmission rate of classical information (the classical channel capacity) is at most $\mathrm{QJD}^{\pi}(\rho_{x_1}, \ldots, \rho_{x_k})$. Holevo [30], and Schumacher and Westmoreland [31] proved that this bound is also asymptotically achievable.

2. *Data compression and side information.*

Let $X = [k]$ be an input alphabet and for each $i \in X$ let P_i be a distribution over output alphabet Y with $|Y| = n$. Consider a setting where a sender uses a weighting π over X, and a receiver who has to compress the received output data losslessly. We call the receiver's knowledge of which distribution P_i is used at any time the *side information*, and difference between the average number of nats (units based on the natural logarithm instead of bits) used for the encoding when the side information is known, and when it is not known, the *redundancy*. In [32], this setting is referred to as the *switching model*.

If the receiver always knows which input distribution is used, then for each distribution P_i, he or she can apply the optimal compression encoding given by $H(P_i)$. Hence, if the receiver has access to the side information, the average number of nats, that the optimal compression encoding uses is given by $\sum_{i=1}^{k} \pi_i H(P_i)$.

However, if the receiver does not know when which input distribution is used, he or she always has to use the same encoding. We say that a compression encoding C *corresponds to* an input distribution Q, if C is optimal for Q (i.e., the number of nats used is $H(Q)$). If the sender transmits an infinite sequence of letters $y_1 y_2 \cdots$, picked according to distribution P_i, and the receiver compresses it using an encoding C which corresponds to distribution Q, then the average number of used nats is given by $\sum_{j=1}^{n} P_i(y_j) \ln \frac{1}{Q(y_j)}$.

Hence, with the weighting $\pi_1, \ldots, \pi_k$, we get the redundancy

$$\begin{aligned} R(Q) &:= \sum_{i=1}^{k} \left(\pi_i H(P_i) - \sum_{j=1}^{n} \pi_i P_i(y_j) \ln \frac{1}{Q(y_j)} \right) \\ &= \sum_{i=1}^{k} \pi_i D(P_i \| Q)\,, \end{aligned}$$

a weighted average of Kullback divergences between the P_i's and Q. The *compensation identity* states that for $\overline{P} = \sum_{i=1}^{k} \pi_i P_i$, the equality

$$\sum_{i=1}^{k} \pi_i D(P_i \| Q) = \sum_{i=1}^{k} \pi_i D(P_i \| \overline{P}) + D(\overline{P} \| Q) \qquad (2)$$

holds for any distribution Q, cf. [33, 34].

It follows immediately that $Q = \overline{P}$ is the unique argmin-distribution for $R(Q)$, and that $\mathrm{JD}^{\pi}(P_1, \ldots, P_k)$ is the corresponding minimum value.

Analogously in a quantum setting, let $X = [k]$ be an input alphabet, and for each $i \in X$ let ρ_i be a state on an output Hilbert space $\mathcal{H}_Y$. We can think of a sender who uses the weighting π of distributions X, but a receiver who has to compress the states on $\mathcal{H}_Y$ using as few *qubits* as possible.

Schumacher [35] showed that the mean number of qubits necessary to encode a state ρ_i is given by $S(\rho_i)$. Later, Schumacher and Westmoreland [36] introduced a quantum encoding scheme, in which an encoding C_Q that is optimal (i.e., requires the least number of qubits) for a state σ requires on average $S(\rho_i) + S(\rho_i \| \sigma)$ qubits to encode ρ_i. Hence, when the receiver uses C_Q as the encoding, the mean redundancy is $R(\sigma) := \sum_{i=1}^{k} \pi_i S(\rho_i \| \sigma)$. Let $\bar{\rho} = \sum_{i=1}^{k} \pi_i \rho_i$. The quantum analogue of (2) is given by Donald's identity [37]:

$$\sum_{i=1}^{k} \pi_i S(\rho_i \| \sigma) = \sum_{i=1}^{k} \pi_i S(\rho_i \| \bar{\rho}) + S(\bar{\rho} \| \sigma),$$

from which it follows that $\sigma = \bar{\rho}$ is the argmin-state that the receiver should code for, and that $\mathrm{QJD}^{\pi}(\rho_1, \ldots, \rho_k)$ is the minimum redundancy.

II. PRELIMINARIES AND NOTATION

In this section we fix notation to be used throughout the paper. We also provide a concise overview of those concepts from quantum theory which we need. For an extensive introduction we refer to [38].

A. Classical information theoretic quantities

We write $[n]$ for the set $\{1, 2, \ldots, n\}$. The set of probability distributions supported by $\mathbb{N}$ is denoted by $M_+^1(\mathbb{N})$ and the set supported by $[n]$ is denoted by $M_+^1(n)$. We associate with probability distributions $P, Q \in M_+^1(n)$ point probabilities $(p_1, \ldots, p_n)$ and $(q_1, \ldots, q_n)$, respectively. Entropy of order $\alpha \neq 1$, Shannon entropy and Kullback divergence are given by

$$S_\alpha(P) := \frac{1 - \sum_{i=1}^{n} p_i^{\alpha}}{\alpha - 1},$$

$$H(P) := -\sum_{i=1}^{n} p_i \ln p_i \qquad (3)$$

and

$$D(P \| Q) := \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}, \qquad (4)$$

respectively. Note that $\lim_{\alpha\to 1^+} S_\alpha(P) = H(P)$. For two-point probability distributions $P = (p, 1-p)$ we let $s_\alpha(p)$ denote $S_\alpha(p, 1-p)$.

B. Quantum theory

1. States.

The d-dimensional complex Hilbert space, denoted by $\mathcal{H}_d$, is the space composed of all d-dimensional complex vectors, endowed with the standard inner product. A physical system is mathematically represented by a Hilbert space. Our knowledge about a physical system is expressed by its *state*, which in turn is represented by a *density matrix* (a trace-1 positive matrix) acting on the Hilbert space. The set of density matrices on a Hilbert space $\mathcal{H}$ is denoted by $\mathcal{B}_+^1(\mathcal{H})$ [51]. Rank-1 density matrices are called *pure-states*. Systems described by two-dimensional Hilbert spaces are called *qubits*. As the eigenvalues of a density matrix are always positive real numbers that sum to one, a state can be interpreted as a probability distribution over pure-states. Hence, sets of states with a complete set of common eigenvectors can be interpreted as probability distributions on the same set of pure-states. States thus generalize probability distributions. This interpretation is not possible when a common basis does not exist. Two states ρ and σ have a set of common eigenvectors if and only if they commute; i.e. $\rho\sigma = \sigma\rho$.

2. Measurements.

Information about a physical system can be obtained by performing a *measurement* on its state. The most general measurement with k outcomes is described by k positive matrices $A_1, \ldots, A_k$, which satisfy $\sum_{i=1}^k A_i = I$. This is a special case of the more general concept of a *positive operator valued measure* (POVM, see for example [38]). The probability that a measurement A of a system in state ρ yields the i'th outcome is $\mathrm{Tr}(A_i\rho)$. Hence, the measurement yields a random variable $A(\rho)$ with $\Pr[A(\rho) = \lambda_i] = \mathrm{Tr}(A_i\rho)$. Naturally, the measurement operators and quantum states should act on the same Hilbert space.

C. Quantum information theoretic quantities

For states $\rho, \sigma \in \mathcal{B}_+^1(\mathcal{H})$, we use the quantum version of entropy of order α, von Neumann entropy and quantum relative entropy, given by

$$S_\alpha(\rho) := \frac{1 - \mathrm{Tr}(\rho^\alpha)}{\alpha - 1},$$

$$S(\rho) := -\mathrm{Tr}(\rho \ln \rho) \tag{5}$$

and

$$S(\rho\|\sigma) := \mathrm{Tr}\rho\ln\rho - \mathrm{Tr}\rho\ln\sigma, \tag{6}$$

respectively. Note that $\lim_{\alpha\to 1^+} S_\alpha(\rho) = S(\rho)$. We refer to [39] for a discussion of quantum relative entropy.

III. DIVERGENCE MEASURES

A. The general Jensen divergence

Let us consider a mixture of k probability distributions $P_1, \ldots, P_k$ with weights $\pi_1, \ldots, \pi_k$ and let $\overline{P} = \sum_{i=1}^k \pi_i P_i$. Jensen's inequality and concavity of Shannon entropy implies that

$$H\left(\sum_{i=1}^k \pi_i P_i\right) \geq \sum_{i=1}^k \pi_i H(P_i).$$

When entropies are finite, we can subtract the right-hand side from the left-hand side and use this as a measure of how much Shannon entropy deviates from being affine. This difference is called the *general Jensen-Shannon divergence* and we denote it by $\mathrm{JD}^\pi(P_1, \ldots, P_k)$, where $\pi = (\pi_1, \ldots, \pi_k)$. One finds that

$$H\left(\sum_{i-1}^k \pi_i P_i\right) - \sum_{i=1}^k \pi_i H(P_i) = \sum_{i=1}^k \pi_i D(P_i\|\overline{P}) \tag{7}$$

and therefore

$$\mathrm{JD}^\pi(P_1, \ldots, P_k) = \sum_{i=1}^k \pi_i D(P_i\|\overline{P}). \tag{8}$$

In the general case when entropies may be infinite the last expression can be used, but we will focus on the situation where the distributions are over a finite set and in this case we can use the left-hand side of (7).

Jensen divergence of order α is defined by the formula

$$\mathrm{JD}_\alpha^\pi(P_1, \ldots, P_k) = S_\alpha\left(\sum_{i=1}^k \pi_i P_i\right) - \sum_{i=1}^k \pi_i S_\alpha(P_i).$$

Similarly, if $\rho_1, \ldots, \rho_k$ are states on a Hilbert space we define

$$\mathrm{QJD}^\pi(\rho_1, \ldots, \rho_k) = \sum_{i=1}^k \pi_i S(\rho_i\|\overline{\rho}), \tag{9}$$

where $\overline{\rho} = \sum_{i=1}^k \pi_i\rho_i$. For states on a finite dimensional

Hilbert space we have

$$\mathrm{QJD}^{\pi}(\rho_1,\ldots,\rho_k) = S\left(\sum_{i=1}^{k}\pi_i\rho_i\right) - \sum_{i=1}^{k}\pi_i S(\rho_i).$$

The quantum Jensen divergence of order α is defined by

$$\mathrm{QJD}^{\pi}_{\alpha}(\rho_1,\ldots,\rho_k) = S_\alpha\left(\sum_{i=1}^{k}\pi_i\rho_i\right) - \sum_{i=1}^{k}\pi_i S_\alpha(\rho_i).$$

B. The Jensen divergence

For even mixtures of two distributions, we introduce the notation $\mathrm{JD}_\alpha(P,Q)$ for $\mathrm{JD}_\alpha(\frac{1}{2}P+\frac{1}{2}Q)$. That is,

$$\mathrm{JD}_\alpha(P,Q) := S_\alpha\left(\frac{P+Q}{2}\right) - \frac{1}{2}S_\alpha(P) - \frac{1}{2}S_\alpha(Q). \quad (10)$$

For even mixtures of two states the QJD was defined in [23], to which we refer for some of its basic properties. We consider the order α version of this and write $\mathrm{QJD}_\alpha(\rho,\sigma)$ for $\mathrm{QJD}_\alpha(\frac{1}{2}\rho+\frac{1}{2}\sigma)$. That is,

$$\mathrm{QJD}_\alpha(\rho,\sigma) := S_\alpha\left(\frac{\rho+\sigma}{2}\right) - \frac{1}{2}S_\alpha(\rho) - \frac{1}{2}S_\alpha(\sigma). \quad (11)$$

We refer to (10) and (11) simply as *Jensen divergence of order* α (JD_α) and *quantum Jensen divergence of order* α (QJD_α) respectively.

IV. METRIC PROPERTIES

In this section we borrow most of the notational conventions and definitions from Deza and Laurent [40]. We refer to this book, to Berg, Christensen and Ressel [48], and to Blumenthal [41] for extensive introductions to the used results. Like Berg, Christensen and Ressel [48] we shall use the expressions "positive and negative definite" for what most textbook would call "positive and negative semi-definite".

Definition 1. *For a set X, a function $d : X \times X \to \mathbb{R}$ is called a* distance *if for every $x, y \in X$:*

1. *$d(x,y) \geq 0$ with equality if $x = y$.*
2. *d is symmetric: $d(x,y) = d(y,x)$.*

   The pair (X,d) is then called a distance space. *If in addition to 1 and 2, for every triple $x, y, z \in X$, the function d satisfies*
3. *$d(x,y) + d(x,z) \geq d(y,z)$* (the triangle inequality), then d is called a *pseudometric* and (X,d) a *pseudometric space*. If also, $d(x,y) = 0$ holds if *and only if* $x = y$, then we speak of a *metric* and a *metric space*.

Our techniques to prove our embeddability results for JD_α and QJD_α are somewhat indirect. To provide some intuition, we briefly mention the following facts. Only Definition 1, Proposition 1 and Theorem 3 are needed for our proofs.

Work of Cayley and Menger gives a characterization of ℓ_2 embeddability of a distance space in terms of Cayley-Menger determinants. Given a finite distance space (X,d), the Cayley-Menger matrix $\mathrm{CM}(X,d)$ is given in terms of the matrix $D_{ij} = d(x_i,x_j)$, for $x_i, x_j \in X$, and the all-ones vector e:

$$\mathrm{CM}(X,d) := \begin{pmatrix} D & e \\ e^T & 0 \end{pmatrix}.$$

Menger proved the following relation between ℓ_2 embeddability and the determinant of $\mathrm{CM}(X,d)$.

Proposition 1 ([42])**.** *Let (X,d) be a finite distance space. Then $(X, d^{1/2})$ is ℓ_2 embeddable if and only if for every $Y \subseteq X$, we have $(-1)^{|Y|}\det \mathrm{CM}(Y,d) \geq 0$.*

As an example, consider a distance space with $|X| = 3$. If we set $a := d(x_1,x_2)^{1/2}$, $b := d(x_1,x_3)^{1/2}$ and $d(x_2,x_3)^{1/2}$, then we obtain

$$-\det \mathrm{CM}(X,d) = (a+b+c)(a-b-c)(-a+b-c)(-a-b+c). \quad (12)$$

On the one hand, this at least zero if d is a pseudometric, and hence pseudometric spaces on three points are ℓ_2 embeddable. On the other hand, up to a factor $1/16$, the right-hand-side of (12) is the square of Heron's formula for the area of a triangle with edge-lengths a, b and c. In general, Cayley-Menger determinants give the formulas needed to calculate the squared hypervolumes of higher dimensional simplices. Menger's result can thus be interpreted as saying that a distance space $(X, d^{1/2})$ is ℓ_2 embeddable if and only if every subset is a simplex with real hypervolume.

Returning to our example with $|X| = 3$, we also have the following implication.

Proposition 2. *Let $(\{x_1,x_1,x_3\}, d)$ be a distance space. Assume that for every $c_1, c_2, c_3 \in \mathbb{R}$ such that $c_1 + c_2 + c_3 = 0$, the distance function d satisfies*

$$\sum_{i,j} c_i c_j d(x_i,x_j) \leq 0, \quad (13)$$

where the summation is over all pairs $i, j \in \{1,2,3\}$. Then $(\{x_1,x_1,x_3\}, d^{1/2})$ is ℓ_2 embeddable.

Proof: Let $a := d(x_1,x_2)^{1/2}$, $b := d(x_1,x_3)^{1/2}$ and $c := d(x_2,x_3)^{1/2}$. We first show that (13) implies that (12) is nonnegative. To this end, set $c_1 = 1$, $c_2 = t$, $c_3 = -t-1$ where t is a real parameter. Then, if (13) holds, we get the inequality

$$a^2 t + b^2 t(-t-1) + c^2(-t-1) \leq 0\,.$$

The nonnegativity of (12) follows from the fact that this inequality holds if and only if the discriminant of this second order polynomial is at least zero. The result now follows from Proposition 1. ∎

The basis of our positive results in this section is that, due to Schoenberg [43, 44], a more general version of Proposition 2 also holds. To state it concisely, we first define *negative definiteness*.

Definition 2 (Negative definiteness)**.** *Let (X, d) be a distance space. Then d is said to be* negative defninite *if and only if for all finite sets $(c_i)_{i\leq n}$ of real numbers such that $\sum_{i=1}^{n} c_i = 0$, and all corresponding finite sets $(x_i)_{i\leq n}$ of points in X, it holds that*

$$\sum_{i,j} c_i c_j d(x_i, x_j) \leq 0. \tag{14}$$

In this case, (X, d) is said to be a distance space of negative type*.*

The following theorem follows as a corollary of Schoenberg's theorem.

Theorem 3. *Let (X, d) be a distance space. Then $\left(X, d^{1/2}\right)$ can be isometrically embedded in a real separable Hilbert space if and only if (X, d) is of negative type.*

Note that if isometric embedding in a Hilbert space is possible, then the space must be a metric space. We define *positive definiteness* as follows.

Definition 3 (Positive definiteness)**.** *Let X be a set and $f : X \times X \to \mathbb{R}$ a mapping. Then f is said to be* positive definite *if and only if for all finite sets $(c_i)_{i\leq n}$ of real numbers and all corresponding finite sets $(x_i)_{i\leq n}$ of points in X, it holds that*

$$\sum_{i,j} c_i c_j f(x_i, x_j) \geq 0. \tag{15}$$

Because we are concerned with functions defined on convex sets, the following definition shall be useful.

Definition 4 (Exponential convexity)**.** *Let X be a convex set and $\phi : X \to \mathbb{R}$ a mapping. Then ϕ is said to be* exponentially convex *if the function $X \times X \to \mathbb{R}$ given by $(x, y) \to \phi\left(\frac{x+y}{2}\right)$ is positive definite.*

Normally exponential convexity is defined as positive definiteness of $\phi(x + y)$ (as is done in for instance [45]), but the definition given here allows the function ϕ only to be defined on a convex set.

A. Metric properties of JD_α

With Theorem 3 we prove the following for Jensen divergence of order α.

Theorem 4. *For $\alpha \in (0, 2]$, the space $\left(M_+^1(\mathbb{N}), \mathrm{JD}_\alpha^{1/2}\right)$ can be isometrically embedded in a real separable Hilbert space.*

Note that Theorem 4 implies that the same holds for QJD_α for sets of commuting quantum states.

We use the following lemma to prove that JD_α is negative definite for $\alpha \in (0, 2]$. Theorem 4 then follows from this and Theorem 3.

Lemma 1. *For $\alpha \in (0, 1)$, we have*

$$x^\alpha = \frac{1}{\Gamma(-\alpha)} \int_0^\infty \frac{e^{-xt} - 1}{t^{\alpha+1}} dt,$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ is the Gamma function. For $\alpha \in (1, 2)$, we have

$$x^\alpha = \frac{1}{\Gamma(-\alpha)} \int_0^\infty \frac{e^{-xt} - (1 - xt)}{t^{\alpha+1}} dt.$$

Proof: Let $\gamma \in (-1, 0)$. From the definition of the Gamma function, we have the following equality:

$$z^\gamma = z^\gamma \frac{1}{\Gamma(-\gamma)} \int_0^\infty r^{-(\gamma+1)} e^{-r} dr.$$

By substituting $r = tz$ we get

$$z^\gamma = \frac{1}{\Gamma(-\gamma)} \int_0^\infty \frac{e^{-zt}}{t^{\gamma+1}} dt.$$

Let $\beta \in (0, 1)$ such that $\beta = \gamma + 1$. Integrating z^γ for z from zero to y and multiplying by $\gamma + 1$ gives,

$$\begin{aligned} y^\beta &= (\gamma + 1) \int_0^y z^\gamma dz \\ &= \frac{1}{\Gamma(-\beta)} \int_0^\infty \frac{e^{-yt} - 1}{t^{\beta+1}} dt. \end{aligned}$$

Now let $\alpha \in (1, 2)$ such that $\alpha = \beta + 1$. Integrating y^β and multiplying by $\beta + 1$ gives the result.

$$\begin{aligned} x^\alpha &= (\beta + 1) \int_0^x y^\beta dy \\ &= \frac{1}{\Gamma(-\alpha)} \int_0^\infty \frac{e^{-xt} - (1 - xt)}{t^{\alpha+1}} dt. \end{aligned}$$

∎

Lemma 2. *For $\alpha \in (0, 2]$, the distance space $\left(M_+^1, \mathrm{JD}_\alpha\right)$ is of negative type.*

Proof: Let $(c_i)_{i\leq n}$ be a set of real numbers such that $\sum_{i=1}^{n} c_i = 0$. For two probability distributions P and Q, we have

$$\mathrm{JD}_\alpha(P, Q) = S_\alpha\left(\frac{P + Q}{2}\right) - \frac{1}{2} S_\alpha(P) - \frac{1}{2} S_\alpha(Q).$$

Observe that for any real valued, single-variable function f, we have $\sum_{i,j} c_i c_j f(x_i) = 0$. Hence, we only need to prove that the function

$$S_\alpha\left(\frac{P+Q}{2}\right) = \frac{1}{\alpha-1} - \frac{1}{(\alpha-1)}\sum_i \left(\frac{p_i+q_i}{2}\right)^\alpha$$

is negative definite for all $\alpha \in (0,2]$. From this decomposition of S_α into a sum over point probabilities it follows that we need to show that $x \curvearrowright x^\alpha$ is exponentially convex. Lemma 1 shows that for fixed $0 < \alpha < 1$ and fixed $1 < \alpha < 2$, the mapping $x \curvearrowright -x^\alpha$ can be obtained as the limit of linear combinations with positive coefficients of functions of the type $x \curvearrowright 1 - e^{-tx}$ and $x \curvearrowright 1 - e^{-tx} - tx$ respectively. Each such function is exponentially convex since the linear terms are, and for non-negative real numbers $x_1, \ldots, x_n$,

$$\sum_{i,j} c_i c_j (-e^{-t(x_i+x_j)}) = -\left(\sum_{i=1}^n c_i e^{-tx_i}\right)^2 \leq 0.$$

The case $\alpha = 1$ follows by continuity. The case $\alpha = 2$ also follows by continuity, but a direct proof without Lemma 1 is straightforward. ∎

Proof of Theorem 4: Follows directly from Lemma 2 and Theorem 3. ∎

A constructive proof of Theorem 4 for JD_1 (JD) is given by Fuglede [46, 47], who uses an embedding into a subset of a real Hilbert space defined by a logarithmic spiral.

B. Metric properties of QJD_α for qubits

Using the same approach as above, we prove the following for quantum Jensen divergence of order α and states on two-dimensional Hilbert spaces.

Theorem 5. *For $\alpha \in (0,2]$, the space*

$$\left(\mathcal{B}^1_+(\mathcal{H}_2), \mathrm{QJD}^{1/2}_\alpha\right)$$

can be isometrically embedded in a real separable Hilbert space.

This is established by the following lemmas and Theorem 3.

Lemma 3. *Let $(V, \langle\cdot|\cdot\rangle)$ be a real Hilbert space with norm $\|\cdot\|_2 = \langle\cdot|\cdot\rangle^{1/2}$. Then, $\left(V, \|\cdot\|_2^2\right)$ is a distance space of negative type.*

Proof: The result follows immediately if we expand the distance function $\|\cdot\|_2^2$ in terms of the inner product:

$$\begin{aligned}
&\sum_{i,j} c_i c_j \langle x_i - x_j, x_i - x_j\rangle \\
&\quad = \sum_{i,j} c_i c_j \left(\|x_i\|_2^2 + \|x_j\|_2^2 - 2\langle x_i, x_j\rangle\right) \\
&\quad = 2\sum_i c_i \sum_j c_j \|x_j\|_2^2 - 2\sum_{i,j} c_i c_j \langle x_i, x_j\rangle \\
&\quad = 0 - 2\sum_{i,j} c_i c_j \langle x_i, x_j\rangle \\
&\quad = -2\Big\|\sum_i c_i x_i\Big\|_2^2 \leq 0.
\end{aligned}$$

∎

Lemma 4. *The distance space $\left(\mathcal{B}^1_+(\mathcal{H}_2), \mathrm{QJD}_\alpha\right), \alpha \in (0,2]$ is of negative type.*

Proof: Using the same techniques as in the proof of Theorem 4, and the fact that Lemma 1 also holds when x is a matrix, what has to be shown is that for $\rho \in \mathcal{B}^1_+(\mathcal{H}_2)$, the function $\rho \curvearrowright \mathrm{Tr}\left(\exp\left(-t\rho\right)\right)$ is exponentially convex. Since ρ acts on a two-dimensional Hilbert space, it has only two eigenvalues, λ_+ and λ_-, that satisfy $\lambda_+ + \lambda_- = 1$ and $\lambda_+^2 + \lambda_-^2 = \mathrm{Tr}\left(\rho^2\right)$. A straightforward calculation gives

$$\lambda_{+/-} = \frac{1}{2} \pm \frac{\left(2\mathrm{Tr}\left(\rho^2\right) - 1\right)^{1/2}}{2}. \tag{16}$$

Plugging this into $\mathrm{Tr}\left(\exp(-t\rho)\right)$ gives

$$\begin{aligned}
\mathrm{Tr}\left(e^{-t\rho}\right) &= 2e^{-t/2}\cosh\left(\frac{t}{2}\left(2\mathrm{Tr}\left(\rho^2\right) - 1\right)^{1/2}\right) \\
&= 2e^{-t/2}\sum_{k=0}^\infty \frac{t^{2k}}{(2k)!4^k}\left(2\mathrm{Tr}\left(\rho^2\right) - 1\right)^k,
\end{aligned}$$

where the second equality follows form the Taylor expansion of hyperbolic cosine. The task can thus be reduced to proving that $\left(2\mathrm{Tr}\left(\rho^2\right) - 1\right)^k$ is exponentially convex for all $k \geq 0$. For this we can use the following theorem:

Theorem 6 ([48, Slight reformulation of Theorem 1.12])**.** *Let $\phi_1, \phi_2 : X \curvearrowright \mathbb{C}$ be exponentially convex functions. Then $\phi_1 \cdot \phi_2$ is exponentially convex too.*

This implies that proving it for $k = 1$ suffices. The trace distance of two density matrices is defined as the Hilbert-Schmidt norm $\|\cdot\|_2$ of their difference. Since the Hilbert-Schmidt norm is a Hilbert-space metric, Lemma 3 implies that $(\rho_1, \rho_2) \curvearrowright \|\rho_1 - \rho_2\|_1^2$ is negative definite and the equality

$$\|\rho_1 - \rho_2\|_2^2 = \mathrm{Tr}(\rho_1 - \rho_2)^2 = 2(\mathrm{Tr}\rho_1^2 + \mathrm{Tr}\rho_2^2) - \mathrm{Tr}\big((\rho_1 + \rho_2)^2\big)$$

implies that the function $\mathrm{Tr}\left((\rho_1 + \rho_2)^2\right)$ is positive definite. From this it follows that the function $2\mathrm{Tr}\left(\rho^2\right) - 1$

is exponentially convex. ■

Proof of Theorem 5: Follows directly from Lemma 4 and Theorem 3. ■

C. Metric properties of QJD_α for pure-states

Here we prove that QJD_α is the square of a metric when restricted to pairs of pure-states. For a Hilbert space of dimension d we denote the set of pure-states as $P(\mathcal{H}_d)$.

Theorem 7. *For $\alpha \in (0,2]$, the space $\left(P(\mathcal{H}_d), \mathrm{QJD}_\alpha^{1/2}\right)$ can be isometrically embedded in a real separable Hilbert space.*

Lemma 5. *The distance space $(P(\mathcal{H}_d), \mathrm{QJD}_\alpha), \alpha \in (0,2]$ is of negative type.*

Proof: Using the same techniques as in Theorem 4, we have to prove that for $\rho \in P(\mathcal{H}_d)$, the function $\rho \curvearrowright \mathrm{Tr}\big(\exp(-t\rho)\big)$ is exponentially convex. For $\rho_1, \rho_2 \in P(\mathcal{H}_d)$ such that $\rho_1 \neq \rho_2$, the matrix $\frac{\rho_1+\rho_2}{2}$ has two non-zero eigenvalues, λ_+ and λ_-, which can be calculated in the same way as above. In this case (16) reduces to

$$\lambda_\pm = \frac{1}{2} \pm \frac{1}{2}\big(\mathrm{Tr}(\rho_1 \cdot \rho_2)\big)^{1/2}.$$

When we plug this into $\mathrm{Tr}\big(\exp(-t(\rho_1+\rho_2))\big)$, we get

$$\begin{aligned}\mathrm{Tr}\left(e^{-2t\left(\frac{\rho_1+\rho_2}{2}\right)}\right) &= (n-2) \\ &\quad +2e^{-t}\cosh\left(t\left(\mathrm{Tr}(\rho_1\cdot\rho_2)\right)^{1/2}\right) \\ &= (n-2) \\ &\quad +2e^{-t}\sum_{k=0}^{\infty}\frac{t^{2k}\left(\mathrm{Tr}(\rho_1\cdot\rho_2)\right)^k}{(2k)!},\end{aligned}$$

where the $(n-2)$ term comes from the fact that $n-2$ of the eigenvalues are zero. We need to prove that $(\rho_1,\rho_2) \curvearrowright \left(\mathrm{Tr}(\rho_1\cdot\rho_2)\right)^k$ is positive definite for all integers $k \geq 0$. But Theorem 6 implies that we only need to prove it for $k=1$. Appealing to the trace distance, we have

$$\|\rho_1-\rho_2\|_1^2 = \mathrm{Tr}\rho_1^2 + \mathrm{Tr}\rho_2^2 - 2\mathrm{Tr}(\rho_1\cdot\rho_2),$$

Since, by Lemma 3, this is negative definite, the result follows. ■

Proof of Theorem 7: Follows directly from Lemma 5 and Theorem 3. ■

D. Counter examples

1. *Metric space counter example for $\alpha \in (2,3)$.*

To see that JD_α, and hence QJD_α, is not the square of a metric for all α we check the triangle inequality for the three probability vectors $P=(0,1), Q=(1/2,1/2)$ and $R=(1,0)$. We have

$$\begin{aligned}\mathrm{JD}_\alpha(P,Q) &= \mathrm{JD}_\alpha(Q,R) \\ &= S_\alpha(1/4,3/4) - \frac{S_\alpha(1/2,1/2)}{2}\end{aligned}$$

and

$$\mathrm{JD}_\alpha(P,R) = S_\alpha(1/2,1/2).$$

The triangle inequality is equivalent to the inequality

$$\begin{aligned}0 &\geq -2\,\mathrm{JD}_\alpha(P,Q) - 2\,\mathrm{JD}_\alpha(Q,R) + \mathrm{JD}_\alpha(P,R) \\ &= -4\left(S_\alpha(1/4,3/4) - \frac{S_\alpha(1/2,1/2)}{2}\right) + S_\alpha(1/2,1/2) \\ &= 3S_\alpha(1/2,1/2) - 4S_\alpha(1/4,3/4) \\ &= 3\frac{1-2(1/2)^\alpha}{\alpha-1} - 4\frac{1-(1/4)^\alpha-(3/4)^\alpha}{\alpha-1} \\ &= \frac{4(1/4)^\alpha + 4(3/4)^\alpha - 6(1/2)^\alpha - 1}{\alpha-1}.\end{aligned}$$

We make the substitution $x=(1/2)^\alpha$ and assume $\alpha > 1$ so the inequality is equivalent to

$$4x^2 + 4x^{\frac{\ln 4-\ln 3}{\ln 2}} - 6x - 1 \leq 0.$$

Define the function

$$f(x) = 4x^2 + 4x^{2-\frac{\ln 3}{\ln 2}} - 6x - 1.$$

Then its first and second derivatives are given by

$$\begin{aligned}f'(x) &= 8x + 4\left(2-\frac{\ln 3}{\ln 2}\right)x^{1-\frac{\ln 3}{\ln 2}} - 6 \\ f''(x) &= 8 + 4\left(2-\frac{\ln 3}{\ln 2}\right)\left(1-\frac{\ln 3}{\ln 2}\right)x^{-\frac{\ln 3}{\ln 2}}\end{aligned}$$

and we see that $f''(x)=0$ has exactly one solution. Therefore f has exactly one infliction point and the equation $f(x)=0$ has at most three solutions. Therefore the equation

$$4(1/4)^\alpha + 4(3/4)^\alpha - 6(1/2)^\alpha - 1 = 0$$

has at most three solutions. It is straightforward to check that $\alpha=1$, $\alpha=2$ and $\alpha=3$ are solutions, so these are the only ones. Therefore the sign of

$$\frac{4(1/4)^\alpha + 4(3/4)^\alpha - 6(1/2)^\alpha - 1}{\alpha-1}$$

is constant in the interval $(2,3)$ and plugging in any number will show that it is negative in this interval. Hence JD_α cannot be a square of a metric for $\alpha \in (2,3)$.

2. Counter examples for Hilbert space embeddability for $\alpha \in \left(\frac{7}{2}, \infty\right)$.

In the previous paragraph we showed that JD_α and QJD_α are not the squares of metric functions for $\alpha \in (2,3)$. Hence, for α in this interval, Hilbert space embeddings are not possible. Here we prove a weaker result for $\alpha \in (\frac{7}{2}, \infty)$, using the Cayley-Menger determinant.

Theorem 8. *The space*

$$\left(\mathcal{B}_+^1(\mathcal{H}_d), (\mathrm{JD}_\alpha)^{\frac{1}{2}}\right)$$

is not *Hilbert space embeddable for α in the interval* $\left(\frac{7}{2}, \infty\right)$.

Note that this does not exclude the possibility that JD_α is the square of a metric and that the same result holds for QJD_α,

Proof: Consider the four distributions

$$\left(\frac{1}{2} - 3\varepsilon, \frac{1}{2} + 3\varepsilon\right),$$
$$\left(\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon\right),$$
$$\left(\frac{1}{2} + \varepsilon, \frac{1}{2} - \varepsilon\right),$$
$$\left(\frac{1}{2} + 3\varepsilon, \frac{1}{2} - 3\varepsilon\right).$$

Then the Cayley-Menger determinant is

$$\begin{vmatrix} s_\alpha\left(\frac{1}{2}-3\varepsilon\right) & s_\alpha\left(\frac{1}{2}-2\varepsilon\right) & s_\alpha\left(\frac{1}{2}-\varepsilon\right) & s_\alpha\left(\frac{1}{2}\right) & 1 \\ s_\alpha\left(\frac{1}{2}-2\varepsilon\right) & s_\alpha\left(\frac{1}{2}-\varepsilon\right) & s_\alpha\left(\frac{1}{2}\right) & s_\alpha\left(\frac{1}{2}+\varepsilon\right) & 1 \\ s_\alpha\left(\frac{1}{2}-\varepsilon\right) & s_\alpha\left(\frac{1}{2}\right) & s_\alpha\left(\frac{1}{2}+\varepsilon\right) & s_{\alpha,2}\left(\frac{1}{2}+2\varepsilon\right) & 1 \\ s_\alpha\left(\frac{1}{2}\right) & s_\alpha\left(\frac{1}{2}+\varepsilon\right) & s_\alpha\left(\frac{1}{2}+2\varepsilon\right) & s_\alpha\left(\frac{1}{2}+3\varepsilon\right) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{vmatrix}$$

and if the four points are Hilbert space embeddable then this determinant is non-negative. The function $\varepsilon \to s_\alpha\left(\frac{1}{2}+\varepsilon\right)$ has a Taylor expansion given by

$$s_\alpha\left(\frac{1}{2}+\varepsilon\right) = s_\alpha\left(\frac{1}{2}\right) + \frac{s_\alpha''\left(\frac{1}{2}\right)}{2}\varepsilon^2 + \frac{s_\alpha^{(4)}\left(\frac{1}{2}\right)}{24}\varepsilon^4 + \frac{s_\alpha^{(6)}\left(\frac{1}{2}\right)}{720}\varepsilon^4 + \varepsilon^8 f(\varepsilon), \quad (17)$$

where f is some continuous function of ε. This can be used to get the expansion of the Cayley-Menger determinant:

$$\mathrm{CM} = \frac{1}{8}s_\alpha^{(4)}\left(\frac{1}{2}\right)\left(\left(s_\alpha^{(4)}\left(\frac{1}{2}\right)\right)^2 - s_\alpha''\left(\frac{1}{2}\right)h_\alpha^{(6)}\left(\frac{1}{2}\right)\right)\varepsilon^{12} + \varepsilon^{14}g(\varepsilon)$$

for some continuous function g [52]. We have the following formula for the even derivatives of s_α :

$$s_\alpha^{(2n)}(x) = -\alpha^{\underline{2n}}\left(x^{\alpha-2n} + (1-x)^{\alpha-2n}\right)$$

and

$$s_\alpha^{(2n)}\left(\frac{1}{2}\right) = -\alpha^{\underline{2n}}2^{2n+1-\alpha}.$$

If the Cayley-Menger determinant is positive for all small ε then

$$\left(s_\alpha^{(4)}\left(\frac{1}{2}\right)\right)^2 - s_\alpha''\left(\frac{1}{2}\right)s_\alpha^{(6)}\left(\frac{1}{2}\right) \le 0$$

or equivalently

$$\left(-\alpha^{\underline{4}}2^{5-\alpha}\right)^2 - \left(-\alpha^{\underline{2}}2^{3-\alpha}\right)\left(-\alpha^{\underline{6}}2^{7-\alpha}\right) \le 0$$

and

$$\begin{aligned} 0 &\ge \left(\alpha^{\underline{4}}\right)^2 - \left(\alpha^{\underline{2}}\right)\left(\alpha^{\underline{6}}\right) \\ &= \alpha^{\underline{2}}\alpha^{\underline{4}}\left((\alpha-2)(\alpha-3) - (\alpha-4)(\alpha-5)\right) \\ &= 4\alpha^{\underline{2}}(\alpha-2)(\alpha-3)\left(\alpha - \frac{7}{2}\right). \end{aligned}$$

Hence, the Cayley-Menger determinant is non-negative only for the intervals $[0,2]$ and $\left[3, \frac{7}{2}\right]$. ■

V. RELATION TO TOTAL VARIATION AND TRACE DISTANCE

The results of Section IV indicate that interesting geometric properties are associated with JD_α and QJD_α when $\alpha \in (0,2]$.

A. Bounds on JD_α

For $\alpha \in (0,2]$. we bound JD_α as follows:

Theorem 9. *Let P and Q be probability distributions in $M_+^1(n)$, and let*

$$v := \tfrac{1}{2}\sum_i |p_i - q_i| \in [0,2]$$

denote their total variation. Then for $\alpha \in (0,2]$, we have $L \le \mathrm{JD}_\alpha(P,Q) \le U$, where:

- *For every $n \geq 2$, L is given by*

$$L(P,Q) = s_\alpha\left(\tfrac{1}{2}\right) - s_\alpha\left(\tfrac{1}{2} + \tfrac{v}{4}\right). \tag{18}$$

- *For every $n \geq 3$, U is given by*

$$U_n(P,Q) = \frac{1}{\alpha - 1}\left(\frac{1}{2} - \frac{1}{2^\alpha}\right)\|P - Q\|_\alpha^\alpha. \tag{19}$$

- *For $n = 2$, U is given by the tighter quantity*

$$U_2(P,Q) = s_\alpha\left(\tfrac{v}{4}\right) - \frac{1}{2}S_{\alpha,2}\left(\tfrac{v}{2}\right). \tag{20}$$

Proof: We start with the lower bound. Let σ denote a permutation of the elements in $[n]$ and let $\sigma(P)$ denote the probability vector where the point probabilities have been permuted according to σ. Clearly, the function JD_α is invariant under such permutations of its arguments:

$$\mathrm{JD}_\alpha\left(\sigma(P), \sigma(Q)\right) = \mathrm{JD}_\alpha(P,Q). \tag{21}$$

Let B denote the set of permutations σ that satisfy

$$p_i \geq q_i \Leftrightarrow p_{\sigma(i)} \geq q_{\sigma(i)}$$

for all $i \in [n]$. Then, by the joint convexity of JD_α for $\alpha \in [1,2]$ (as proved in [10]), we have

$$\begin{aligned}\mathrm{JD}_\alpha(P,Q) &= \frac{1}{|B|}\sum_{\sigma\in B}\mathrm{JD}_\alpha\left(\sigma(P),\sigma(Q)\right) \\ &\geq \mathrm{JD}_\alpha\left(\frac{1}{|B|}\sum_{\sigma\in B}\sigma(P), \frac{1}{|B|}\sum_{\sigma\in B}\sigma(Q)\right).\end{aligned} \tag{22}$$

The distributions $\frac{1}{|B|}\sum_{\sigma\in B}\sigma(P)$ and $\frac{1}{|B|}\sum_{\sigma\in B}\sigma(Q)$ have the property that they are constant on two complementary sets, namely $\{i \in [n] \mid p_i \geq q_i\}$ and $\{i \in [n] \mid p_i < q_i\}$. Therefore, we may without loss of generality assume that P and Q are distributions on a two-element set. On a two-element set P and Q can be parametrized by $P = (p, 1-p)$ and $Q = (q, 1-q)$. If σ_2 denotes the transposition of the two elements then

$$v = V\left(\frac{P + \sigma_2(Q)}{2}, \frac{Q + \sigma_2(P)}{2}\right) = 2|p - q|.$$

By (21) and (22) we get

$$\begin{aligned}\mathrm{JD}_\alpha(P,Q) &\geq \mathrm{JD}_\alpha\left(\frac{P + \sigma_2(Q)}{2}, \frac{Q + \sigma_2(P)}{2}\right) \\ &= \mathrm{JD}_\alpha\left(\left(\frac{1}{2} + \frac{v}{4}, \frac{1}{2} - \frac{v}{4}\right), \left(\frac{1}{2} - \frac{v}{4}, \frac{1}{2} + \frac{v}{4}\right)\right) \\ &= s_\alpha(1/2) - s_\alpha\left(\frac{1}{2} + \frac{v}{4}\right),\end{aligned}$$

and this lower bound is attained for two distributions on a two element set. Next we derive the general upper bound. Define distribution $\widetilde{P}$ on $[n] \times [3]$ such that for every $i \in [n]$,

$$\begin{aligned}\widetilde{P}(i,1) &= \min\{p_i, q_i\}, \\ \widetilde{P}(i,2) &= \begin{cases} p_i - q_i & \text{if } p_i > q_i \\ 0 & \text{otherwise,}\end{cases} \\ \widetilde{P}(i,3) &= 0,\end{aligned}$$

and similarly define $\widetilde{Q}$ on $[n] \times [3]$ by

$$\begin{aligned}\widetilde{Q}(i,1) &= \min\{p_i.q_i\}, \\ \widetilde{Q}(i,2) &= 0, \\ \widetilde{Q}(i,3) &= \begin{cases} q_i - p_i & \text{if } q_i > p_i \\ 0 & \text{otherwise.}\end{cases}\end{aligned}$$

With these definitions we have $V(\widetilde{P}, \widetilde{Q}) = V(P,Q)$. Using the data processing inequality and the definitions of $\widetilde{P}$ and $\widetilde{Q}$ it is straightforward to verify that

$$\begin{aligned}\mathrm{JD}_\alpha(P,Q) &\leq \mathrm{JD}_\alpha(\widetilde{P}, \widetilde{Q}) \\ &= \frac{1}{\alpha - 1}\left(\frac{1}{2} - \frac{1}{2^\alpha}\right)\sum_{i=1}^n |p_i - q_i|^\alpha.\end{aligned}$$

This upper bound is attained on a three element set so we have

$$U_n(P,Q) = \frac{1}{\alpha - 1}\left(\frac{1}{2} - \frac{1}{2^\alpha}\right)\|P - Q\|_\alpha^\alpha.$$

To get a tight upper bound on a two-element set a special analysis is needed. The cases $p > q$ and $p < q$ are treated separately, but the two cases work the same way. We will therefore assume that $p > q$. On a two-element set parametrize P and Q by $P = (p, 1-p)$ and $Q = (q, 1-q)$. In this case we have the linear constraint $p - q = v/2$. For a fixed value of v, we have that JD_α is a convex function of q. Therefore the maximum is attained by an extreme point, i.e. a distribution where either p or q is either 0 or 1. Without loss of generality we may assume that $q = 0$ and that $p = v/2$. This gives

$$U_2(P,Q) = s_\alpha\left(\frac{v}{4}\right) - \frac{s_\alpha\left(\frac{v}{2}\right)}{2}.$$

■

It is now straightforward to determine the exact form of the joint range of V and JD_α.

Corollary 10. *The joint range of V and JD_α, denoted by Δ_n, is a compact region in the plane bounded by a (Jordan) curve composed of two curves: The first curve is given by* (18) *with V running from 2 to 0. For $n = 2$ the second curve is given by* (19) *with v running from 0 to 2, and for $n = 3$ the second curve is given by* (20) *with v running from 0 to 2.*

Proof: Assume first that $n \geq 3$. By Theorem 9 we know that Δ_n is contained in the compact domain described. A continuous deformation of the lower curve into the upper bounding curve (i.e. a homotopy from the lower bounding curve to the upper bounding curve) is given by P_t, Q_t for $t \in [0,1]$, where

$$\begin{pmatrix} P_t \\ Q_t \end{pmatrix}(v) = (1-t)\begin{pmatrix} \frac{2+v}{4} & \frac{2-v}{4} & 0 & \cdots & 0 \\ \frac{2-v}{4} & \frac{2+v}{4} & 0 & \cdots & 0 \end{pmatrix} + t\begin{pmatrix} 1-\frac{v}{2} & \frac{v}{2} & 0 & 0 & \cdots & 0 \\ 1-\frac{v}{2} & 0 & \frac{v}{2} & 0 & \cdots & 0 \end{pmatrix}$$

for $v \in [0,2]$. Therefore, Δ_n has no "holes". The case $n = 2$ is handled in a similar way. ∎

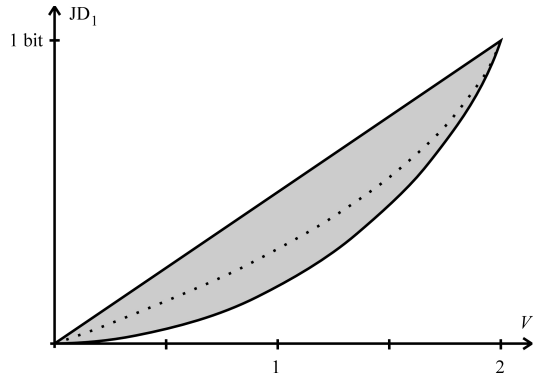

Figure 1: $V/\,\mathrm{JD}_\alpha$-diagram for $\alpha = 1$ and $n \geq 3$ (the shaded region), and for $n = 2$ (the region obtained by replacing the upper bounding curve by the dotted curve).

In Figure 1 we have depicted the $V/\,\mathrm{JD}_\alpha$-diagram for $\alpha = 1$.

The bounds (18) and (19) give us the following proposition regarding the topology induced by $(\mathrm{JD}_\alpha)^{\frac{1}{2}}$. In the limiting case $\alpha \to 1$, this was proved in [5] by a different method.

Proposition 11. *The space* $\left(M_+^1(\mathbb{N}), \mathrm{JD}_\alpha^{1/2}\right)$ *is a complete, bounded metric space for* $\alpha \in (0,2]$*, and the induced topology is that of convergence in total variation.*

Proof: By expansion of $L(P,Q)$ given by (18), in terms of the total variation v, one obtains the inequality

$$\mathrm{JD}_\alpha(P,Q) \geq \frac{1}{\alpha-1}\sum_{j=1}^{\infty}\binom{\alpha}{2j}\left(\frac{v}{2}\right)^{2j}. \tag{23}$$

Taking only the first term and bounding (19), we get

$$\begin{aligned} \frac{1}{8}V^2(P,Q) &\leq \frac{\alpha}{8}V^2(P,Q) \\ &\leq \mathrm{JD}_\alpha(P,Q) \\ &\leq \frac{1}{\alpha-1}\left(\frac{1}{2}-\frac{1}{2^\alpha}\right)\|P-Q\|_\alpha^\alpha \\ &\leq \frac{\ln 2}{2}V(P,Q)\,. \end{aligned} \tag{24}$$

∎

B. Bounds on QJD_α

With Theorem 9 we can bound QJD_α for $\alpha \in [1,2]$. We use the following two theorems.

Theorem 12 ([49], Theorem 3.9)**.** *Let* $\mathcal{H}$ *be a Hilbert space,* $\rho_1, \rho_2 \in \mathcal{B}_+^1(\mathcal{H})$ *and* $\mathcal{M} := \{M_i \mid i = 1, \ldots, n\}$ *be a measurement on* $\mathcal{H}$*. Then* $S(\rho_1\|\rho_2) \geq D(P_\mathcal{M}\|Q_\mathcal{M})$*, where* $P_\mathcal{M}, Q_\mathcal{M} \in M_+^1(n)$ *and have point probabilities* $P_\mathcal{M}(i) = \mathrm{Tr}(M_i\rho_1)$ *and* $Q_\mathcal{M}(i) = \mathrm{Tr}(M_i\rho_2)$*, respectively.*

Theorem 13 ([38], Theorem 9.1)**.** *Let* $\mathcal{H}$ *be a Hilbert space,*

$$\rho_1, \rho_2 \in \mathcal{B}_+^1(\mathcal{H})$$

and $\mathcal{M} := \{M_i \mid i = 1, \ldots, n\}$ *be a measurement on* $\mathcal{H}$*. Then* $\|\rho_1 - \rho_2\|_1 = \max_\mathcal{M} V(P_\mathcal{M}, Q_\mathcal{M})$*, where* $P_\mathcal{M}, Q_\mathcal{M} \in M_+^1(n)$ *and have point probabilities* $P_\mathcal{M}(i) = \mathrm{Tr}(M_i\rho_1)$ *and* $Q_\mathcal{M}(i) = \mathrm{Tr}(M_i\rho_2)$*, respectively.*

Theorem 14. *For* $\alpha \in (0,2]$*, for all states* $\rho_1, \rho_2 \in \mathcal{B}_+^1(\mathcal{H})$*, we have*

$$\begin{aligned} s_\alpha(\tfrac{1}{2}) - s_\alpha\left(\frac{1}{2} + \frac{\|\rho_1-\rho_2\|_1}{2}\right) &\leq \mathrm{QJD}_\alpha(\rho_1,\rho_2) \\ &\leq \frac{ln\,2}{2}\|\rho_1-\rho_2\|_1. \end{aligned}$$

Proof: The lower bound is proved in the same way as [50, Theorem III.1], by making a reduction to the case of classical probability distributions by means of measurements. Let $\mathcal{M}$ be a measurement that maximizes $V(P_\mathcal{M}, Q_\mathcal{M})$. Then from Theorem 13 we have $\|\rho_1 - \rho_2\|_1 = V(P_\mathcal{M}, Q_\mathcal{M})$. Theorem 12 gives us

$$\begin{aligned} \mathrm{QJD}_\alpha(\rho_1,\rho_2) &\geq \frac{1}{2}D\left(P_\mathcal{M}\Big\|\frac{P_\mathcal{M}+Q_\mathcal{M}}{2}\right) \\ &\quad + \frac{1}{2}D\left(Q_\mathcal{M}\Big\|\frac{P_\mathcal{M}+Q_\mathcal{M}}{2}\right) \\ &= \mathrm{JD}_\alpha(P_\mathcal{M}, Q_\mathcal{M}). \end{aligned}$$

The result now follows from Theorem 9. The upper bound is proved the same way as we proved the classical bound. Introduce a 3-dimensional Hilbert space $\mathcal{G}$ with basis vectors $|1\rangle$, $|2\rangle$ and $|3\rangle$. On $\mathcal{H} \otimes \mathcal{G}$ define the

density matrices

$$\begin{aligned}\tilde{\rho}_1 &= \frac{\rho_1+\rho_2-|\rho_1-\rho_2|}{2}\otimes|1\rangle\langle 1| \\ &\quad+\frac{\rho_1-\rho_2+|\rho_1-\rho_2|}{2}\otimes|2\rangle\langle 2|, \\ \tilde{\rho}_2 &= \frac{\rho_2+\rho_1-|\rho_2-\rho_1|}{2}\otimes|1\rangle\langle 1| \\ &\quad+\frac{\rho_2-\rho_1+|\rho_1-\rho_2|}{2}\otimes|3\rangle\langle 3|.\end{aligned}$$

Let $\mathrm{Tr}_{\mathcal{G}}$ denote the partial trace $\mathcal{B}^1_+(\mathcal{H}\otimes\mathcal{G})\to\mathcal{B}^1_+(\mathcal{H})$. Then $\mathrm{Tr}_{\mathcal{G}}(\tilde{\rho}_1)=\rho_1$ and $\mathrm{Tr}_{\mathcal{G}}(\tilde{\rho}_2)=\rho_2$. The matrices $\frac{\rho_1-\rho_2+|\rho_1-\rho_2|}{2}$ and $\frac{\rho_2-\rho_1+|\rho_1-\rho_2|}{2}$ are positive definite so

$$\begin{aligned}\|\tilde{\rho}_1-\tilde{\rho}_2\|_1 &= \mathrm{Tr}\left|\frac{\rho_1-\rho_2+|\rho_1-\rho_2|}{2}\otimes|2\rangle\langle 2|\right. \\ &\quad\left.-\frac{\rho_2-\rho_1+|\rho_1-\rho_2|}{2}\otimes|3\rangle\langle 3|\right| \\ &= \mathrm{Tr}\left(\frac{\rho_1-\rho_2+|\rho_1-\rho_2|}{2}\right) \\ &\quad+\mathrm{Tr}\left(\frac{\rho_2-\rho_1+|\rho_1-\rho_2|}{2}\right) \\ &= \mathrm{Tr}\,|\rho_1-\rho_2| = \|\rho_1-\rho_2\|_1.\end{aligned}$$

According to the "quantum data processing inequality" [49, Theorem 3.10] we have

$$\begin{aligned}&\mathrm{QJD}_\alpha(\rho_1,\rho_2)\le\mathrm{QJD}_1(\tilde{\rho}_1,\tilde{\rho}_2) \\ &\quad=\frac{1}{2}\mathrm{Tr}\left(\frac{\rho_1-\rho_2+|\rho_1-\rho_2|}{2}\otimes|2\rangle\langle 2|\right)\ln 2 \\ &\quad+\mathrm{Tr}\left(\frac{\rho_2-\rho_1+|\rho_1-\rho_2|}{2}\otimes|3\rangle\langle 3|\right)\ln 2 \\ &\qquad=\frac{\ln 2}{2}\cdot\|\rho_1-\rho_2\|_1.\end{aligned}$$

■

VI. CONCLUSIONS AND OPEN PROBLEMS

We studied generalizations of the (general) Jensen divergence and its quantum analogue. For $\alpha\in(1,2]$, JD_α was proved to be the square of a metric which can be embedded in a real Hilbert space. The same was shown to hold for QJD_α restricted to qubit states or to pure states. Both these results were derived by evoking a theorem of Schoenberg's and showing that these quantities are negative definite.

Whether $(\mathrm{QJD}_1)^{\frac{1}{2}}$ is a metric for all mixed states remains unknown. However, based on a large amount of numerical evidence, we conjecture the function $A\to\mathrm{Tr}(e^A)$ to be exponentially convex for density matrices A. Proving this would imply that QJD_α is negative definite for $\alpha\in(0,2]$, and hence the square of a metric that can be embedded in a real Hilbert space.

VII. ACKNOWLEDGEMENTS

We are greatly indebted to Flemming Topsøe. This work mainly extends his basic result jointly with Bent Fuglede presented at the conference ISIT 2004 [47], where you find the basic result on isometric embedding in Hilbert space related to JD_1. Flemming has supplied us with many valuable comments and suggestions. In particular Section V is to a large extent inspired by unpublished results of Flemming.

Jop Briët is partially supported by a Vici grant from the Netherlands Organization for Scientific Research (NWO), and by the European Commission under the Integrated Project Qubit Applications (QAP) funded by the IST directorate as Contract Number 015848. Peter Harremoës has been supported by the Villum Kann Rasmussen Foundation, by Danish Natural Science Research Council, by INTAS (project 00-738) and by the European Pascal Network.

[1] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Trans. Inf. Theory*, 49:1858–60, 2003.

[2] R. G. Gallager. *Information Theory and Reliable Communication*. Wiley and Sons, New York, 1968.

[3] J. Lin and S. K. M. Wong. A new directed divergence measure and its characterization. *Int. J. General Systems*, 17:73–81, 1990.

[4] J. Lin. Divergence Measures Based on the Shannon Entropy. *IEEE Trans Inform. Theory*, 37:145–151, 1991.

[5] F. Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inform. Theory*, 46:1602–1609, 2000.

[6] R. Sibson. Information radius. *Z. Wahrs und verw Geb.*, 14:149–160, 1969.

[7] A. K. C. Wong and M. You. Entropy and distance of random graphs with application to structural pattern-recognition. *IEEE Trans. Pattern Anal. Machine Intell.*, 7:599–609, 1985.

[8] O. A. Rosso, H. A. Larrondo, M. T. Martin, A. Plastino, and M. A. Fuentes. Distinguishing noise from chaos. *Phys. Rev. Lett.*, 99(15):154102, 2007.

[9] R. El-Yaniv, S. Fine, and N. Tishby. Agnostic classification of Markovian sequences. *NIPS, MIT-Press*, pages 465–471, 1997.

[10] J. Burbea and C. R. Rao. On the convexity of some divergence measures based on entropy functions. *IEEE Trans. Inform. Theory*, 28:489–495, 1982.

[11] C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Physics*, 52:479, 1988.

[12] J. Lindhard and V. Nielsen Studies in Dynamical Systems *Kongelige Danske Videnskabernes Selskab,*

Matematisk-Fysiske Meddelser, 38(9):1-42, 1971.

[13] M. Figueiredo A. Martins, P. Aguiarz. Tsallis kernels on measures. *IEEE Information Theory Workshop*, pages 298–302, 2008.

[14] A. Martins, P. Aguiar, and M. Figueiredo. Nonextensive Generalizations of the Jensen-Shannon Divergence. *Submitted*, Apr 2008.

[15] A. F. T. Martins, M. A. T. Figueiredo, P. M. Q. Aguiar, N. A. Smith, and E. P. Xing. Nonextensive entropic kernels. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 640–647, New York, NY, USA, 2008. ACM.

[16] P. M. Q. Aguiar N. A. Smith E. P. Xing A. F. T. Martins, M. A. T. Figueiredo. Nonextensive entropic kernels. Tech. report CMU-ML-08-106, Carnegie Mellon University, 2008.

[17] W. K. Wootters. Statistical distance and Hilbert space. *Phys. Rev. D*, 23(2):357–362, Jan 1981.

[18] S. L. Braunstein and C. M. Caves. Statistical distance and the geometry of quantum states. *Phys. Rev. Lett.*, 72(22):3439–3443, May 1994.

[19] J. Lee, M. S. Kim, and Č. Brukner. Operationally invariant measure of the distance between quantum states by complementary measurements. *Phys. Rev. Lett.*, 91(8):087902, Aug 2003.

[20] A. P. Majtey, P. W. Lamberti, M. T. Martin, and A. Plastino. Wootters' distance revisited: a new distinguishability criterium. *Eur. Phys. J. D*, 32:413–419, 2005.

[21] P. W. Lamberti, A. P. Majtey, A. Borras, M. Casas, and A. Plastino. On the metric character of the quantum Jensen-Shannon divergence. *Phys. Rev. A*, 77(5):052311, 2008.

[22] B. Schumacher and M. D. Westmoreland. Relative entropy in quantum information theory. In S. Lomonaco, editor, *Quantum Computation and Quantum Information: A Millenium Volume*. American Mathematical Society Contemporary Mathematics series, 2001.

[23] A. P. Majtey, P. W. Lamberti, and D. P. Prato. Jensen-Shannon divergence as a measure of distinguishability between mixedquantum states. *Phys. Rev. A*, 72(5):052310, 2005.

[24] P. W. Lamberti, M. Portesi, and J. Sparacino. A natural metric for quantum information theory. *arXiv:quant-ph/0807.0583v1*, Jul 2008.

[25] T. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[26] A. Rényi. On Measures of Entropy and Information. In *Proc. 4th Berkeley Symp. Math. Statist. and Prob.* 1:547–561, Univ. Calif. Press, Berkely, 1961.

[27] J. H. Havrda and F. Charvat. Quantification methods of classification processes: concepts of structural α entropy. *Kybernatica*, 3:30–35, 1967.

[28] J. Aczél and Z. Daróczy *On Measures of Information and their Characterization*. Mathematics in Science and Engineering vol. 115, Academic Press, New York, 1975.

[29] A. S. Holevo. Information theoretical aspects of quantum measurements. *Probl. Inf. Transm.*, 9:110–118, 1973.

[30] A. S. Holevo. The capacity of quantum channel with general signal states. *IEEE Trans. Inform. Theory*, 44:269–273, 1998.

[31] B. Schumacher and M. D. Westmoreland. Sending classical information via noisy quantum channels. *Phys. Rev. A*, 56(1):131–138, Jul 1997.

[32] B. Fuglede and F. Topsøe. Jensen-Shannon divergence and Hilbert space embedding. In *Proceedings 2004 International Symposium on Information Theory*, page 31, 2004.

[33] F. Topsøe. An information theoretical identity and a problem involving capacity. *Studia Scientiarum Mathematicarum Hungarica*, 2:291–292, 1967.

[34] F. Topsøe. Basic concepts, identities and inequalities – the toolkit of information theory. *Entropy*, 3(3):162–190, 2001.

[35] B. Schumacher. Quantum coding. *Phys. Rev. A*, 51:2738–2747, April 1995.

[36] B. Schumacher and M. D. Westmoreland. Indeterminate-length quantum coding. *Phys. Rev. A*, 64(4):042304, Sep 2001.

[37] M. J. Donald. Further results on the relative entropy. *Math. Proc. Cam. Phil. Soc.*, 101:363, 1987.

[38] M. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge, 2000.

[39] M. Ohya, D. Petz. *Quantum entropy and its use*. Texts and Monographs in Physics. Springer-Verlag, Berlin, 1993.

[40] M. Laurent M. Deza. *Geometry of Cuts and Metrics*. Springer-Verlag, Berlin, 1997.

[41] L. M. Blumenthal. *Theory and Applications of Distance Geometry*. Oxford University Press, London, 1953.

[42] K. Menger. *Géométrie Générale*. Gauthier-Villars, Paris, 1954.

[43] I. J. Schoenberg. Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert". *Annals of Mathematics*, 36:724–732, 1935.

[44] I. J. Schoenberg. Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.*, 44:522–536, 1938.

[45] A. E. Nussbaum. Radial exponentially convex functions. *Journal d'Analyse Mathématique*, 25(1):277–288, December 1972.

[46] B. Fuglede. Spirals in Hilbert space: with an application in information theory. *Expo. Math.*, 23:23–45, 2005.

[47] B. Fuglede and F. Topsøe. Jensen-Shannon divergence and Hilbert space embedding. In *Proceedings 2004 International Symposium on Information Theory*, page 31, 2004.

[48] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer-Verlag, New York, 1984.

[49] D. Petz. *Quantum information theory and quantum statistics*. Springer, Berlin, 2008.

[50] H. Klauck, A. Nayak, A. Ta-Shma, and D. Zuckerman. Interaction in quantum communication. *IEEE Transactions on Information Theory*, 53(6):1970–1982, 2007.

[51] This deviates from the notation used in [21]. We do this for the sake of consistency with regard to the notation for probability distributions.

[52] The calculation of the Taylor expansion involves a lot of computations but are easily performed using Maple or similar symbol manipulation program.