Properties of Classical and Quantum Jensen-Shannon Divergence

Jop Briët* and Peter Harremoës†
Centrum voor Wiskunde en Informatica, Amsterdam

Flemming Topsøe‡
Department of Mathematical Sciences, University of Copenhagen

Abstract

The Jensen-Shannon divergence (JSD) is a symmetrized and smoothed version of the all important divergence measure of informaton theory, the Kullback-Leibler divergence. It defines a true metric – precisely, it is the square of a metric. We prove a stronger result for a new family of divergence measures based on the Tsallis entropy, that includes the JSD. Furthermore we elaborate on details of geometric properties of the JSD. Analogously, the *quantum* Jensen-Shannon divergence (QJSD) is a symmetrized version of the quantum relative entropy that has recently been considered as a distance measure for quantum states. We prove for a new family of distance measures for states, including the QJSD, that each member is the square of a metric for all qubits, strengthening recent results by Lamberti et al. We also discuss geometric properties of the QJSD. In analogy to Lin's generalization of the JSD, we also define the *general* QJSD for a weighting of any number of states and discuss interpretations of both quantities.

1 Introduction

The general Jensen-Shannon divergence is a divergence measure for a weighted set of probability distributions which has recently gained interest among physicists and statisticians. It has useful interpretations in information theory and attractive features as a function: it is symmetric, bounded and always defined. For evenly weighted pairs of distributions, Endres and Schindelin proved that it defines the square of a metric [1], which we call the *transmission metric* (d_T).

We define a family of divergence measures based on Tsallis entropies, which includes the JSD and prove that each member defines the square of a metric that with elegant geometric properties.

Divergence measures also play an important role in quantum information theory. In this case, their purpose is to define a distance measure for quantum states. Similar to the classical version, we generalize the JSD to a family of quantum divergence measures based on quantum Tsallis entropies (defined in Section 2.3), which we put in the context of quantum channel capacity and a coding scheme called *indeterminate length quantum encoding* [2]. In contrast to most other distance measures for states, this family is also symmetric, bounded and always defined.

*jop.briet@cwi.nl. 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.

†P.Harremoes@cwi.nl. Supported by the Villum Kann Rasmussen Foundation, by Danish Natural Science Research Council, by INTAS (project 00-738) and by the European Pascal Network.

‡topsoe@math.ku.dk. Has been supported by the Danish Natural Science Resarch Council and by INTAS, project 00-738.

The JSD itself has also been studied in the context of a distance measure for even mixtures of two quantum states [3, 4]. However, the quantum Jensen-Shannon divergence (QJSD) was specifically tailored for this purpose [5, 4]. The most interesting question regarding this quantity is if it also defines the square of a metric. Lamberti et. al proved that this is indeed the case for pure-states, but whether or not this holds in general is still unknown.

Our main result regarding our family of quantum divergence measures is that for two-dimensional mixed states (qubits) and pure-states of arbitrary dimension, each member defines the square of a metric has nice geometric properties.

Our technique is to evoke a theorem by Schoenberg that links the metric and geometric properties to the condition of negative semidefiniteness. We prove that all members of our families of (classical and quantum) divergence measures indeed satisfy this condition, in the quantum case with restriction to qubit- and pure-states. Though we concentrate on the discrete case, well-known limiting properties imply that the results related to classical divergence – and this includes Jensen-Shannon divergence – carry over to the general case of probability measures on arbitrary Borel spaces.

2 Preliminaries and notation

Throughout we use the notation introduced in this section. We provide a concise overview of the concepts of quantum theory used in this paper, for extensive introductions we refer to [6].

2.1 Classical information theoretic quantities

We write $[n]$ for the set $\{1, 2, \dots, n\}$. The set of probability distributions supported by $\mathbb{N}$ is denoted by $M_+^1(\mathbb{N})$ and those supported by $[n]$ is denoted by $M_+^1(n)$. We associate point probabilities $(p_1, p_2, \dots, p_i, \dots)$ and $(q_1, q_2, \dots, q_i, \dots)$ with probability distributions P and Q respectively. The Tsallis entropy of order $\alpha \neq 1$, Shannon entropy and Kullback-Leibler divergence are given by

$$S_a(P) := \frac{1 - \sum_i p_i^\alpha}{\alpha - 1},$$

$$H(P) := -\sum_i p_i \log p_i \tag{1}$$

and

$$D(P\|Q) := \sum_i p_i \log \frac{p_i}{q_i} \tag{2}$$

respectively. Here log denotes the natural logarithm. It is important to note that $S_a(P) \to H(P)$ for $\alpha \to 1$.

2.2 Quantum theory

States. A d-dimensional complex Hilbert space, denoted by $\mathcal{H}_d$, is a 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 operator (a trace-1 positive semidefinite linear operator) acting on the Hilbert space. The set of density operators on a Hilbert space $\mathcal{H}$ is denoted by $\mathcal{B}_+^1(\mathcal{H})$.[1] Rank-1 density operators are called a *pure-states*.

[1]This deviates from the notation used in [4]. We do this for the sake of consistency with regard to the notation for probability distributions.

Two-dimensional Hilbert spaces are called *qubits*. Because the eigenvalues of a density operator 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. This interpretation is not possible when such a common basis does not exist. States thus generalize probability distributions. Two states ρ and σ have a set of common eigenvectors if and only if and only if they commute: $\rho\sigma = \sigma\rho$.

Measurements. Information about the state of a physical system can only be obtained by performing a *measurement*. The most general measurement with k outcomes is a set $A := \{A_1, \dots, A_k\}$ of k positive semidefinite matrices that satisfy $\sum_{i=1}^{k} A_i = I$ and correspond to k measurement outcomes. 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.

2.3 Quantum information theoretic quantities

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

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

$$S(\rho) := -\mathrm{Tr}(\rho \log \rho), \tag{3}$$

and

$$S(\rho\|\sigma) := \mathrm{Tr}\rho \log \rho - \mathrm{Tr}\rho \log \sigma \tag{4}$$

respectively. Here, it also holds that $S_\alpha(\rho) \to S(\rho)$ for $\alpha \to 1$.

3 Divergence measures

3.1 The general Jensen-Shannon divergence

Consider a mixture $\overline{P} = \sum_\nu \alpha_\nu P_\nu$. Concavity of the Shannon entropy implies that

$$H\left(\sum_\nu \alpha_\nu P_\nu\right) \geq \sum_\nu \alpha_\nu H(P_\nu),$$

according to Jensen's inequality. 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 linear. The difference is called the *general Jensen-Shannon divergence* and, allowing an abuse of notation, is denoted $\mathrm{JSD}(\sum_\nu \alpha_\nu P_\nu)$. We have

$$H\left(\sum_\nu \alpha_\nu P_\nu\right) - \sum_\nu \alpha_\nu H(P_\nu) = \sum_\nu \alpha_\nu D(P_\nu\|\overline{P}) \tag{5}$$

and therefore

$$\mathrm{JSD}\left(\sum_\nu \alpha_\nu P_\nu\right) = \sum_\nu \alpha_\nu D(P_\nu\|\overline{P}). \tag{6}$$

In the general case where entropies may be infinite the last extession can be used, but we will focus on the situation where the distributions are on a finite set so that all entropies are finite and in this case can use the left-hand side (5).

Like Shannon entropy, the Tsallis entropy is a concave function of P so we shall define the Jensen-Tsallis divergence of order α by the formula

$$\mathrm{JSD}_\alpha\left(\sum_\nu \alpha_\nu P_\nu\right) = S_\alpha\left(\sum_\nu \alpha_\nu P_\nu\right) - \sum_\nu \alpha_\nu S_\alpha(P_\nu)$$

Similarly, if ρ_ν are states on a Hilbert space we define

$$\mathrm{QJSD}\left(\sum_\nu \alpha_\nu \rho_\nu\right) = \sum_\nu \alpha_\nu S(\rho_\nu \| \Sigma), \tag{7}$$

where $\Sigma = \sum_\nu \alpha_\nu \rho_\nu$. For states on finite dimensional Hilbert spaces we have

$$\mathrm{QJSD}\left(\sum_\nu \alpha_\nu \rho_\nu\right) = S\left(\sum_\nu \alpha_\nu \rho_\nu\right) - \sum_\nu \alpha_\nu S(\rho_\nu).$$

The quantum Jensen-Tsallis divergence of order α is defined by

$$\mathrm{QJSD}_\alpha\left(\sum_\nu \alpha_\nu \rho_\nu\right) = S_\alpha\left(\sum_\nu \alpha_\nu \rho_\nu\right) - \sum_\nu \alpha_\nu S_\alpha(\rho_\nu).$$

We refer to quantities (6) and (7) as the general JSD_1 and the general QJSD_1 respectively.

Interpretations of the general JSD. A *discrete memoryless channel* is defined as a system with input alphabet X, output alphabet Y and conditional probabilities $p(y|x)$ that give the probability that $y \in Y$ is received when $x \in X$ was sent (see for example [7]). For an input distribution π over letters $x \in X$ that are sent over a discrete memoryless channel with $Y = X$ and conditional distributions $P_x := \{(y, p(y|x)) \mid y \in Y\}$, we have that $\mathrm{JSD}_1(\sum_{x\in X} \pi(x) P_x)$ in fact gives the transmission rate.

Another interpretation relates to the model which we shall refer to as the *switching model*, previously presented in [8]. In this model, a source generates an infinite sequence of letters $x_1 x_2 \ldots$, such that each letter is selected independently of previous letters, according to distribution P_ν with probability α_ν.

Consider an observer who knows the P_ν's as well as the α_ν's but does not know which distribution is used at any particular time. The observer wants to design a code such that the expected *redundancy* is minimized. By redundancy we have the following in mind:

An *ideal observer* always knows which distribution was used for the letter selection at the source, and will use this insight to choose, at each instant, a code adapted to the distribution used for the selection. Hence, the ideal observer will use on average $H(P_\nu)$ nits for his observations (here we use units based on the natural logarithm instead of bits as a unit of information). A typical observer, however, will not know when this or that distribution is used at the source and has to choose one and the same code, say κ, all the time. If Q is the probability distribution for which κ is the optimal encoding scheme, then the code length used for the letter with index i is $-\log q_i$. In this case we say that κ *corresponds to* Q. If the code κ, used by the actual observer corresponds to Q, then the average amount of nits used by the observer is $\sum_i p_{\nu,i} \log \frac{1}{q_i}$ when P_ν is the actual distribution. The redundancy is the difference between this number and $H(P_\nu)$. Hence, the redundancy is $D(P_\nu \| Q)$ and the average redundancy is $R(Q) := \sum_\nu \alpha_\nu D(P_\nu \| Q)$.

Thus, the observer should choose the distribution Q as the basis for coding which minimizes average redundancy $R(Q)$. In order to identify the associated argmin-distribution, we refer to the so called *compensation identity* which states that for $\overline{P} = \sum_\nu \alpha_\nu P_\nu$, the equality

$$\sum_\nu \alpha_\nu D(P_\nu \| Q) = \sum_\nu \alpha_\nu D(P_\nu \| \overline{P}) + D(\overline{P} \| Q) \tag{8}$$

holds for any distribution Q. It follows immediately from this identity that $Q = \overline{P}$ is the unique argmin-distribution sought and that $\mathrm{JSD}_1(\sum_\nu \alpha_\nu P_\nu)$ is the corresponding minimum value. Therefore, the general Jensen-Shannon divergence can also be interpreted as *minimum redundancy* for the switching model.

By (8), for any fixed Q, divergence $D(\cdot \| Q)$ is a convex function:

$$D\left(\sum_\nu \alpha_\nu P_\nu \| Q\right) \leq \sum_\nu \alpha_\nu D(P_\nu \| Q)\,. \tag{9}$$

Furthermore, we realize that if we apply the same strategy of definition as the one applied to the entropy function and consider the "Jensen-type" divergence, looking at the difference between the right hand and the left hand side in (9), we are back to the quantity (6), we started with and that this is independent of the distribution Q we choose to take as our reference.

Interpretations of the general Quantum JSD. Suppose now that we have a channel with input and output alphabets X and Y with $Y = X$, and input distribution π over $x \in X$, where the letters are encoded in quantum states ρ_x before they are sent to the receiver (see for example [9]). The receiver decodes the message by performing a measurement $\{M_y \mid y \in Y\}$ on the state he or she obtained after the transmission. As proved by Holevo [11], the maximum achievable transmission rate (the channel capacity) is given by $\mathrm{QJSD}_1\left(\sum_{x \in X} \pi(x)\rho_x\right)$.

Analogously to the classical switching model, we can think of a situation where a source generates an infinite sequence of quantum states $\rho_1, \rho_2, \ldots$ with each state selected independently of the previous ones, according to distribution P_ν with probability α_ν. A typical observer does not know when which distribution is used but wants to encode the received states using as few qubits as possible. An ensemble $\{(\rho_1, p_{\nu,1}), (\rho_2, p_{\nu,2}), \ldots\}$ can be represented by the density operator $\rho_\nu = \sum_i p_{\nu,i}\rho_i$. Schumacher [12] showed that the mean number of qubits necessary to encode the states in such an ensemble is $S(\rho_\nu)$, which equals $H(P_\nu)$ if all ρ_i are projectors onto orthogonal subspaces. Schumacher and Westmoreland [13] proposed an encoding scheme called *indeterminate length quantum coding*, where the optimal encoding of a state ρ_ν indeed requires $S(\rho_\nu)$ qubits. In this scheme, an optimal encoding τ for a state σ would require $S(\rho_\nu) + S(\rho_\nu \| \sigma)$ qubits to encode the state ρ_ν. Hence, when τ is used as the indeterminate length quantum encoding, the average redundancy is $R(\sigma) := \sum_\nu \alpha_\nu S(\rho_\nu \| \sigma)$. Again, the typical observer is faced with the task of seeking a state σ as basis for encoding for which $R(\sigma)$ is minimized. Let $\rho = \sum_\nu \alpha_\nu \rho_\nu$. From Donald's identity [14]:

$$\sum_\nu \alpha_\nu S(\rho_\nu \| \sigma) = \sum_\nu \alpha_\nu S(\rho_\nu \| \rho) + S(\rho \| \sigma),$$

it follows that $\sigma = \rho$ is the argmin-state that the typical observer should code for, and that $\mathrm{QJSD}_1(\sum_\nu \alpha_\nu \rho_\nu)$ is the minimum redundancy.

3.2 The Jensen-Shannon divergence

For even mixtures of two distributions the general JSD_1 has been discussed before. Implicitly, the quantity (5) was introduced by Wong and You [15]. Then Lin and Wong derived some simple properties [16, 17]. More recently, Topsøe derived further identities and inequalities [18]. For other recent references with a focus on statistical applications, see El-Yaniv et al [19].

We consider a family of divergence measures based on the Tsallis entropy and introduce the notation $\mathrm{JSD}_\alpha(P,Q)$ for $\mathrm{JSD}_\alpha(\frac{1}{2}P+\frac{1}{2}Q)$. That is,

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

For even mixtures of two states the QJSD_1 was defined in [5], to which we refer for a list of its basic properties. We generalize this quantity the same way as JSD_1 and write $\mathrm{QJSD}_\alpha(\rho,\sigma)$ for $\mathrm{QJSD}_\alpha(\frac{1}{2}\rho+\frac{1}{2}\sigma)$. That is,

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

For the remainder of this paper we shall work with the specific divergence measures just introduced and refer to (10) and (11) simply as the *Jensen-Tsallis divergence of order* α (JSD_α) and *quantum Jensen-Tsallis divergence of order* α (QJSD_α) respectively.

4 Metric properties

We begin by recalling the properties that a function should posses for it to be a metric. For a set X, a function $d : X \times X \to \mathbb{R}$ is a metric if for every $x, y, z \in X$:

1. $d(x,y) \geq 0$ with equality if and only if $x = y$
2. d is symmetric: $d(x,y) = d(y,x)$
3. d satisfies the triangle inequality: $d(x,y) + d(x,z) \geq d(y,z)$.

If d only satisfies properties 1 and 2 then it is called a distance. Endres and Schindelin proved that JSD_1 is in fact the square of a metric [1]. The QJSD_1 was proved to be the square of a metric for pure-states by Lamberti et al. [4], who conjectured this to be true in general. Here we show the stronger result that JSD_α is in fact the square of a metric which can be isometrically embedded in a real Hilbert space, and that the same holds for QJSD_α for two-dimensional mixed states and pure-states of any dimension.

We prove these results by a method which is quite different from the methods used by the previous authors.

Proposition 1. *Let X be a set and $K : X \times X \to \mathbb{R}$ be a distance and for $x_1, x_2, x_3 \in X$ and $c_1, c_2, c_3 \in \mathbb{R}$ such that $c_1 + c_2 + c_3 = 0$, satisfies*

$$\sum_{1 \leq i,j \leq 3} c_i c_j K(x_i, x_j) \leq 0. \tag{12}$$

Then K defines the square of a metric.

Proof: First, let us agree that a triple (α, β, γ) of non-negative numbers are *squared distances* if $\alpha^{1/2} \leq \beta^{1/2} + \delta^{1/2}$, $\beta^{1/2} \leq \delta^{1/2} + \alpha^{1/2}$ and $\delta^{1/2} \leq \alpha^{1/2} + \beta^{1/2}$. Assuming that $\gamma = \max\{\alpha, \beta, \gamma\}$, this amounts to the condition $\gamma^{1/2} \leq \alpha^{1/2} + \beta^{1/2}$ which, after some simple algebra can be transformed to the equivalent inequality

$$\alpha^2 + \beta^2 + \gamma^2 \leq 2\alpha\beta + 2\beta\gamma + 2\gamma\alpha\,. \tag{13}$$

Assume that (12) holds and let $\alpha = K(x_1, x_2)$, $\beta = K(x_2, x_3)$ and $\gamma = K(x_3, x_1)$. We realize that what we have to prove is that (α, β, γ) are squared distances. To prove this we exploit (12) with $c_1 = 1$, $c_2 = t$, $c_3 = -t-1$ where t is a real parameter. This gives us the inequality

$$\alpha t + \beta t(-t-1) + \gamma(-t-1) \leq 0\,.$$

It is then a matter of simple algebra to deduce (13) from the non-positivity of this second order polynomial. ∎

Theorem 2. *Let $\mathcal{H}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and associated norm $\|\cdot\|_2$. Then for all finite sets of real numbers $(c_i)_{i \leq n}$ and all finite sets $(x_i)_{i \leq n}$ of points in $\mathcal{H}$, the implication*

$$\sum_{i=1}^{n} c_i = 0 \implies \sum_{1 \leq i,j \leq n} c_i c_j \|x_i - x_j\|_2^2 \leq 0 \tag{14}$$

holds.

Proof: Note that if we expand (14) in terms of the inner product, we get

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

∎

The importance of the implication in (14) justifies the following formal definition.

Definition 1 (Negative semidefiniteness)**.** *Let X be a set and $K : X \times X \to \mathbb{R}$ a mapping. Then K is said to be* negative semidefninite *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, the implication*

$$\sum_{i=1}^{n} c_i = 0 \implies \sum_{i,j} c_i c_j K(x_i, x_j) \leq 0 \tag{15}$$

holds.

Conversely, we define *positive semidefiniteness*:

Definition 2 (Positive semidefiniteness)**.** *Let X be a set and $K : X \times X \to \mathbb{R}$ a mapping. Then K is said to be* positive semidefinite *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, we*

$$\sum_{i,j} c_i c_j K(x_i, x_j) \geq 0 \tag{16}$$

holds.

Note that in the last definition, there is no restriction on the sum of the c's. The basis of our results in this section is that the converse of Theorem 2 also holds, as proved by Schoenberg in 1938 [20]:

Theorem 3 (Schoenberg's theorem)**.** *Let X be a set and $K : X \times X \to \mathbb{R}$ a negative semidefinite mapping (a kernel) such that, for $(x, y) \in X \times X$, $K(x, y) = 0$ if and only if $x = y$, $K(x, y) = K(y, x)$, $K(x, y) \geq 0$. Then $(X, \sqrt{K})$ is a metric space which can be embedded isometrically as a subspace of a real Hilbert space.*

The kernels we are going to discuss are defined on convex sets and the following definition shall be useful.

Definition 3 (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 and only if the kernel $X \times X \to \mathbb{R}$ given by $(x, y) \to \phi\left(\frac{x+y}{2}\right)$ is positive semidefinite.*

4.1 Metric properties of the JSD

First, using Theorem 3, we prove the following.

Theorem 4. *For all $\alpha \geq 1$ the kernel JSD_α is negative semidefinite on $M_+^1(\mathbb{N}) \times M_+^1(\mathbb{N})$. Therefore, there exists a subset $\overline{\mathcal{H}}_{JSD} \subseteq \overline{\mathcal{H}}$ of a real separable Hilbert space $\overline{\mathcal{H}}$ and a one-to-one bijection Φ between $M_+^1(\mathbb{N})$ and $\overline{\mathcal{H}}_{\mathrm{JSD}}$ such that, for all $(P, Q) \in M_+^1(\mathbb{N}) \times M_+^1(\mathbb{N})$, $\mathrm{JSD}_\alpha(P, Q) = \|\Phi(P) - \Phi(Q)\|^2$ with $\|\cdot\|$ denoting the norm in $\overline{\mathcal{H}}$.*

This theorem implies that the same holds for QJSD_α for sets of commuting elements in $\mathcal{B}_+^1(\mathcal{H})$. Theorem 4 can be proved in a basically elementary way, using facts and tricks from harmonic analysis, as found in, for example, [20].

Proof of Theorem 4: We need to prove that JSD_α is negative semidefinite on $M_+^1(\mathbb{N}) \times M_+^1(\mathbb{N})$. The implication of the theorem then follows from Schoenberg's theorem. Let $(c_i)_{i \leq n}$ be the set from the definition of negative semidefiniteness. First assume that $1 < \alpha < 2$. For two probability distributions P and Q, we have

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

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

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

is negative semidefinite. From this decomposition into a sum over point probabilities it follows that we need to prove that $x \curvearrowright x^\alpha$ is exponentially convex. For this to run smoothly, we need the following representation formula for the power functions involved[2]:

$$-x^{\alpha+1} = \frac{\alpha(1+\alpha)}{\Gamma(1-\alpha)} \int_0^\infty \frac{1 - e^{-tx} - tx}{t^{\alpha+2}} dt, \tag{17}$$

[2]This follows by integration in the formula

$$x^\beta = \frac{\beta}{\Gamma(1-\beta)} \int_0^\infty \frac{1 - e^{-tx}}{t^{\beta+1}} dt$$

which, in turn, can be easily checked by differentiation. As usual Γ denotes the *gamma function*: $\Gamma(a) = \int_0^\infty t^{\beta-1}e^{-t}dt$; $\beta > 0$.

Equation (17) shows that for fixed $0 < \beta < 1$, $x \curvearrowright -x^{\beta+1}$ can be obtained as the limit of linear combinations with positive coefficients of functions of the type $x \curvearrowright 1 - e^{-tx} - tx$. Each such function is exponentially convex since $x \curvearrowright 1 - tx$ is, 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 (17) is straightforward. ■

Precisely how one can make the Hilbert space embedding is described by Fuglede [21]. Motivated by the interpretations of the general JSD_1, we call the metric defined by the square root of JSD_1 the *transmission metric*, and denote it by d_T.

4.2 Metric properties of the QJSD for qubits

Using the same approach as above, we prove:

Theorem 5. *The kernel* QJSD_α *is negative semidefinite on* $\mathcal{B}_+^1(\mathcal{H}_2) \times \mathcal{B}_+^1(\mathcal{H}_2)$. *Therefore, there exists a subset* $\overline{\mathcal{H}}_{\mathrm{QJSD}} \subseteq \overline{\mathcal{H}}$ *of a real separable Hilbert space* $\overline{\mathcal{H}}$ *and a one-to-one bijection* Φ *between* $\mathcal{B}_+^1(\mathcal{H}_2)$ *and* $\overline{\mathcal{H}}_{\mathrm{QJSD}}$ *such that, for all* $(\rho,\sigma) \in \mathcal{B}_+^1(\mathcal{H}_2) \times \mathcal{B}_+^1(\mathcal{H}_2)$, $\mathrm{QJSD}_\alpha(\rho,\sigma) = \|\Phi(\rho) - \Phi(\sigma)\|^2$ *with* $\|\cdot\|$ *denoting the norm in* $\overline{\mathcal{H}}$.

Proof: Using the same techniques as in the commutative case, what has to be shown is that for $\rho \in \mathcal{B}_+^1(\mathcal{H}_2)$, $\rho \curvearrowright \mathrm{Tr}\left(\exp\left(-t\rho\right)\right)$ is exponentially convex. Since ρ acts on a two-dimensional Hilbert space it has only two eigenvalues, λ_+ and λ_-, that satisfy

$$\begin{aligned} \lambda_+ + \lambda_- &= 1 \\ \lambda_+^2 + \lambda_-^2 &= \mathrm{Tr}\left(\rho^2\right). \end{aligned}$$

A straightforward calculation gives

$$\lambda_{+/-} = \frac{1}{2} \pm \frac{\left(2\mathrm{Tr}\left(\rho^2\right) - 1\right)^{1/2}}{2}. \tag{18}$$

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

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

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 ([20, Slight reformulation of Theorem 1.12 in this book])**.** *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 operators is defined as the trace-norm $\|\cdot\|_1$ of their difference. The trace-distance is a Hilbert-space metric and therefore by Theorem 2, we have that $(\rho,\sigma) \curvearrowright \|\rho - \sigma\|_1^2$ is negative semidefinite. Since

$$\|\rho - \sigma\|_1^2 = \mathrm{Tr}(\rho - \sigma)^2 = 2(\mathrm{Tr}\rho^2 + \mathrm{Tr}\sigma^2) - \mathrm{Tr}(\rho + \sigma)^2$$

this implies that $\mathrm{Tr}\left((\rho + \sigma)^2\right)$ is positive semidefinte and thus that $2\mathrm{Tr}\left(\rho^2\right) - 1$ is exponentially convex. ■

4.3 Metric properties of the QJSD for pure-states

Third, we prove that QJSD_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 any $d \geq 1$, the kernel* QJSD *is negative semidefinite on $P(\mathcal{H}_d) \times P(\mathcal{H}_d)$. Therefore, there exists a subset $\overline{\mathcal{H}}_{\mathrm{QJSD}} \subseteq \overline{\mathcal{H}}$ of a real separable Hilbert space $\overline{\mathcal{H}}$ and a one-to-one bijection Φ between $P(\mathcal{H}_d)$ and $\overline{\mathcal{H}}_{\mathrm{QJSD}}$ such that, for all $(\rho, \sigma) \in P(\mathcal{H}_d) \times P(\mathcal{H}_d)$, $\mathrm{QJSD}_\alpha(\rho, \sigma) = \|\Phi(\rho) - \Phi(\sigma)\|^2$ with $\|\cdot\|$ denoting the norm in $\overline{\mathcal{H}}$.*

Proof: Again using the same method as in Section 4.1, we have to prove that $\rho \curvearrowright \mathrm{Tr}\,(\exp(-t\rho))$ is exponentially convex. For $\rho, \sigma \in P(\mathcal{H}_d)$ such that $\rho \neq \sigma$, the matrix $\frac{\rho+\sigma}{2}$ has two non-zero eigenvalues, λ_+ and λ_-, which can be calculated in the same way as above. In this case (18) reduces to

$$\lambda_\pm = \frac{1}{2} \pm \frac{1}{2}\left(\mathrm{Tr}(\rho \cdot \sigma)\right)^{1/2}.$$

When we plug this into $\mathrm{Tr}(\exp(-t(\rho + \sigma)))$, we get

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

where the $(n-2)$ term comes from the fact that $n-2$ of the eigenvalues are zero. We need to prove that $(\rho, \sigma) \curvearrowright (\mathrm{Tr}(\rho \cdot \sigma))^k$ is positive semidefinite for all integers $k \geq 0$. Using Theorem 6, we only need to prove it for $k = 1$. Again appealing to the trace distance, we have

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

Since this is negative semidefinite, the result follows. ■

5 Geometric properties

Theorems 4, 5 and 7 indicate that interesting geometric properties are associated with JSD_1 and QJSD_1.

5.1 Induced topology of $(M_+^1(\mathbb{N}), d_T)$

Recall that $\mathrm{JSD}_1 = d_T^2$. We focus on the fact that $(M_+^1(\mathbb{N}), d_T)$ is a metric space and answer the natural question concerning identification of the induced topology.

Denoting total variation by V, the following inequalities hold for any pair of distributions $P, Q \in M_+^1(\mathbb{N})$:

$$\frac{1}{8}V^2(P,Q) \leq \mathrm{JSD}_1(P,Q) \leq \frac{1}{2}\log 2 \cdot V(P,Q)\,. \tag{19}$$

This gives the following theorem:

Theorem 8. *Consider the space $M_+^1(\mathbb{N})$ with the metric d_T. This metric is a complete, bounded metric and the induced topology is that of convergence in total variation.*

These inequalities were proved in [18]. However, the inequalities can be strengthened significantly by determining, for each $n \in \mathbb{N}$, the exact range of the "$V/\,\mathrm{JSD}_1$-map" defined by $(P, Q) \curvearrowright (V(P,Q), \mathrm{JSD}_1(P,Q))$ with $P, Q \in M_+^1(n)$.

5.2 $V/\,\mathrm{JSD}_1$-diagrams

Let A and B be two quantities which depend on one or more parameters, say $A = A(\alpha)$, $B = B(\alpha)$, possibly with $\alpha = (\alpha_1, \cdots, \alpha_\nu)$ running over some multi-dimensional parameter set. By the *A/B-diagram* we mean the range of the map $\alpha \curvearrowright (A(\alpha),\, B(\alpha))$. In [22] and [23] examples and techniques for the determination of such diagrams can be found. Determining the A/B-diagram often leads to more precise inequalities relating A and B than one would obtain by a more direct approach.

In this section we shall determine the $V/\,\mathrm{JSD}_1$-diagrams Δ_n for all $n \in \mathbb{N}$, where

$$\Delta_n = \{(V(P,Q), \mathrm{JSD}_1(P,Q)) \mid (P,Q) \in M_+^1(n) \times M_+^1(n)\}\,. \tag{20}$$

We first note that Δ_1 is trivial ($= \{(0,0)\}$). The technique we apply for $n \geq 2$ is a direct variational analysis, which is simpler than the variational principles which were applied in the sources quoted.

Characterize Δ_n by its *upper* and *lower functions* given, for $V \in [0,2]$, by

$$U_n(V) = \sup\{\mathrm{JSD}_1(P,Q) \mid P,Q \in M_+^1(n), V(P,Q) = V\}\,, \tag{21}$$

$$L_n(V) = \inf\{\mathrm{JSD}_1(P,Q) \mid P,Q \in M_+^1(n), V(P,Q) = V\}\,. \tag{22}$$

Clearly, $U_n(0) = L_n(0) = 0$ and $U_n(2) = L_n(2) = 2\log 2$. So we from now on assume that $0 < V < 2$.

A natural compactness argument shows that the extrema in (21) and (22) are actually attained. Using the variational argument we have in mind, we determine $U_n(V)$ and $L_n(V)$ simultaneously.

The point of departure is an *extremal pair* $(P,Q) \in M_+^1(n) \times M_+^1(n)$ by which we mean that for $V(P,Q) = V$, either $\mathrm{JSD}_1(P,Q) = U_n(V)$ or $\mathrm{JSD}_1(P,Q) = L_n(V)$ holds. Recall that the point probabilities in P and Q are denoted as $P = (p_1, \cdots, p_n)$ and $Q = (q_1, \cdots, q_n)$. Define a $2 \times n$ matrix $A := \binom{P}{Q} = \binom{p_1 \,\cdots\, p_n}{q_1 \,\cdots\, q_n}$, such that that the number of *zero-columns* $\binom{0}{0}$ of A, is maximal. In other words, no other matrix $\binom{P'}{Q'}$ with $P', Q' \in M_+^1(n)$ such that $V(P',Q') = V$ and $\mathrm{JSD}_1(P',Q') = \mathrm{JSD}_1(P,Q)$ has a higher number of zero-columns than A.

For the variational arguments used below, we introduce a parameter ε and consider distributions P_ε and Q_ε with point probabilities $p_i = p_i(\varepsilon)$ and $q_i = q_i(\varepsilon)$, whose derivates are given by $p_i' = \frac{d}{d\varepsilon} p_i(\varepsilon)$ and $q_i' = \frac{d}{d\varepsilon} q_i(\varepsilon)$. It is straightforward to verify that

$$\frac{d}{d\varepsilon}\,\mathrm{JSD}_1(P_\varepsilon, Q_\varepsilon) = \sum_{i=1}^{n} (\log p_i \cdot p_i' + \log q_i \cdot q_i' - \log(p_i + q_i) \cdot (p_i' + q_i'))\,. \tag{23}$$

Observation 1. All non-zero columns of A are linearly independent.

That is, there exists no $i \in [n]$ such that $\binom{p_i}{q_i} = \sum_{j \in [n]\setminus\{i\}} \alpha_j \binom{p_j}{q_j}$ for $\alpha_j \in \mathbb{R}$. To see this, assume the contrary and replace $\binom{p_i}{q_i}$ by $\binom{0}{0}$ and $\binom{p_j}{q_j}$ by $(1+\alpha_j)\binom{p_j}{q_j}$ for every $j \in [n]\setminus\{i\}$ to obtain $\binom{P'}{Q'}$ which has one more zero column than A, but $V(P',Q') = V$ and $\mathrm{JSD}_1(P',Q') = \mathrm{JSD}_1(P,Q)$. This contradiction gives the result.

Observation 2. For no $i,j \in [n]$ such that $i \neq j$ can both $p_i > q_i$ and $p_j > q_j$ hold.

To see this, consider $i,j \in [n]$ such that $i \neq j$ and assume that $p_i > q_i$ and $p_j > q_j$. Then, for $|\varepsilon|$ sufficiently small consider $(P_\varepsilon, Q_\varepsilon)$ defined by $Q_\varepsilon = Q$ (from the extremal pair) and by changing p_i to $p_i+\varepsilon$ and p_j to $p_j-\varepsilon$. As $V(P_\varepsilon, Q_\varepsilon)=V$ for $|\varepsilon|$ sufficiently small, we must have $\frac{d}{d\varepsilon}\,\mathrm{JSD}_1(P_\varepsilon, Q_\varepsilon)=0$ for $\varepsilon=0$. By (23) this gives

$$(\log p_i - \log(p_i+q_i)) + (-\log p_j + \log(p_j+q_j)) = 0\,.$$

It follows that

$$\frac{p_i}{q_i} = \frac{p_j}{q_j}\,,$$

thus $\binom{p_i}{q_i}$ and $\binom{p_j}{q_j}$ are proportional. This contradicts the first observation and we conclude that the second observation is correct.

As P and Q appear symmetrically in our problem, the second observation also shows that for no $i \neq j$ do $q_i > p_i$ and $q_j > p_j$ both hold. It now follows that the number of columns $\binom{p_i}{q_i}$ with $p_i \neq q_i$ is at most 2. As this number cannot be 0 or 1 (since $P \neq Q$), it must be exactly 2. We can therefore assume that $\binom{p_1}{q_1}$ and $\binom{p_2}{q_2}$ are the two columns in question and that $p_1 > q_1$ and $p_2 < q_2$. By Observation 1, there is at most one column $\binom{p_i}{q_i}$ with $p_i = q_i$. We have now shown that A must have the following form, assuming that the zero-columns come last:

$$A = \begin{pmatrix} p_1 & p_2 & p_3 & 0 & \cdots & 0 \\ q_1 & q_2 & p_3 & 0 & \cdots & 0 \end{pmatrix}.$$

This is convenient for our last observation:

Observation 3. If, for $i, j \in [n]$ such that $i \neq j$, all numbers p_i, q_i, p_j and q_j lie in the open interval $]0,1[$, then $p_i p_j = q_i q_j$.

To prove this, consider P_ε, Q_ε with $|\varepsilon|$ sufficiently small and with point probabilities as for P and Q except for the following four changes:

$$p_i(\varepsilon) = p_i+\varepsilon\,,\ p_j(\varepsilon) = p_j-\varepsilon\,,\ q_i(\varepsilon) = q_i+\varepsilon\,,\ q_j(\varepsilon) = q_j-\varepsilon\,.$$

Again, as for the second observation, the total variation is unchanged and $\frac{d}{d\varepsilon}\,\mathrm{JSD}_1(P_\varepsilon, Q_\varepsilon)$ must vanish for $\varepsilon=0$. Hence by (23),

$$(\log p_i + \log q_i - 2\log(p_i+q_i)) + (-\log p_j - \log q_j + 2\log(p_j+q_j)) = 0\,.$$

From this we conclude that

$$(1+x_i)(1+\frac{1}{x_i}) = (1+x_j)(1+\frac{1}{x_j})$$

where $x_i = \frac{p_i}{q_i}$, $x_j = \frac{p_j}{q_j}$. As $x_i = x_j$ is not feasible in view of Observation 1, the only other solution is $x_i = \frac{1}{x_j}$, which gives the desired conclusion.

With these three observations it is now easy to conclude the analysis.

Case 1: $p_3 = 0$**.** In this case A must be of the form

$$A = \begin{pmatrix} p & 1-p & 0 & \cdots & 0 \\ 1-\frac{V}{2} & 1-p+\frac{V}{2} & 0 & \cdots & 0 \end{pmatrix}$$

with $\frac{V}{2} \le p \le 1$. If $\frac{V}{2} < p < 1$ we conclude from Observation 3 that $p(1-p) = \left(p-\frac{V}{2}\right)\left(1-p+\frac{V}{2}\right)$ or

$$\frac{1}{4} - \left(p-\frac{1}{2}\right)^2 = \frac{1}{4} - \left(p-\frac{V}{2}-\frac{1}{2}\right)^2\,,$$

hence $|p-\frac{1}{2}| = |p-\frac{1}{2}-\frac{V}{2}|$ and we see that $p=\frac{1}{2}+\frac{V}{4}$. Together with the two "end point possibilities" $\big(p=\frac{V}{2}$ or $p=1\big)$ this gives three possibilities:

$$A = \begin{pmatrix} \frac{2+V}{4} & \frac{2-V}{4} & 0 & \cdots & 0 \\ \frac{2-V}{4} & \frac{2+V}{4} & 0 & \cdots & 0 \end{pmatrix} \tag{24}$$

$$A = \begin{pmatrix} \frac{V}{2} & \frac{2-V}{2} & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \end{pmatrix} \tag{25}$$

and then a further possibility (corresponding to $p = 1$) which is, however, symmetric to (25) and may therefore be discarded.

Case 2: $p_3 > 0$. Note that this requires $n \geq 3$. If, in this case $q_1 > 0$, then Observation 3 applies to column 1 and 3 and we conclude that $p_1 p_3 = q_1 p_3$, i.e. $p_1 = q_1$. However, in view of Observation 1, this is not possible. Therefore, $q_1 = 0$ must hold. Similarly, $p_2 = 0$ must hold. This leaves us with A of the form

$$A = \begin{pmatrix} P \\ Q \end{pmatrix} = \begin{pmatrix} \frac{V}{2} & 0 & \frac{2-V}{2} \\ 0 & \frac{V}{2} & \frac{2-V}{2} \end{pmatrix}. \tag{26}$$

It remains to calculate the Jensen-Shannon divergence in the three extremal cases isolated:

$$\text{Case } a: \qquad \mathrm{JSD}_1(P,Q) = \frac{1}{2}\big(1+\frac{V}{2}\big)\log\big(1+\frac{V}{2}\big) + \frac{1}{2}\big(1-\frac{V}{2}\big)\log\big(1-\frac{V}{2}\big) \tag{27}$$

$$\text{Case } b: \qquad \mathrm{JSD}_1(P,Q) = \log 2 - \frac{1}{2}\big(2-\frac{V}{2}\big)\log\big(2-\frac{V}{2}\big) + \frac{1}{2}\big(1-\frac{V}{2}\big)\log\big(1-\frac{V}{2}\big) \tag{28}$$

$$\text{Case } c: \qquad \mathrm{JSD}_1(P,Q) = \frac{\log 2}{2}\cdot V\,. \tag{29}$$

The functions in (27)-(29) are related as follows:

$$\frac{1}{2}\big(1+\frac{V}{2}\big)\log\big(1+\frac{V}{2}\big) + \frac{1}{2}\big(1-\frac{V}{2}\big)\log\big(1-\frac{V}{2}\big) < \\ \log 2 - \frac{1}{2}\big(2-\frac{V}{2}\big)\log\big(2-\frac{V}{2}\big) + \frac{1}{2}\big(1-\frac{V}{2}\big)\log\big(1-\frac{V}{2}\big) < \\ \frac{\log 2}{2}\cdot V \tag{30}$$

This follows by standard considerations. Of course, for $V = 0$ or $V = 2$, equality holds in (30). Recalling that (29) requires $n \geq 3$ we can now conclude as follows:

Theorem 9. *For all $n \geq 2$, the lower function L_n coincides with the common lower function L given by*

$$L(V) = \frac{1}{2}\,(1+\frac{V}{2})\log(1+\frac{V}{2}) + \frac{1}{2}\,(1-\frac{V}{2})\log(1-\frac{V}{2})\,. \tag{31}$$

For all $n \geq 3$, the upper function U_n coincides with the common upper function U given by

$$U(V) = \frac{1}{2}\,\log 2\cdot V \tag{32}$$

and for $n = 2$, we have

$$U_2(V) = \log 2 - \frac{1}{2}\,(2-\frac{V}{2})\log(2-\frac{V}{2}) + \frac{1}{2}\,(1-\frac{V}{2})\log(1-\frac{V}{2})\,. \tag{33}$$

With just a little more effort, we can determine the exact form of the V/JSD_1-diagrams Δ_n.

Theorem 10. *For $n = 2$, the V/JSD_1-diagram Δ_n is the compact region in the plane determined by the Jordan curve composed by the two curves given by (32), run from $V = 0$ to $V = 2$, followed by (31), run from $V = 2$ to $V = 1$.*

For $n \geq 3$, the diagrams Δ_n coincide with the common V/JSD_1-diagram constructed as above but with the curve given by (33) as the upper curve.

Proof: Assume first that $n \geq 3$. By Theorem 5 we know that Δ_n is contained in the compact domain described. That there are no "holes" in Δ_n follows by considering P_t, Q_t given by

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

for $t \in [0,1]$, which gives a homotopy from the lower bounding curve to the upper bounding curve (i.e. a continuous deformation of the lower curve into the upper). The case $n = 2$ is handled in a similar way. ■

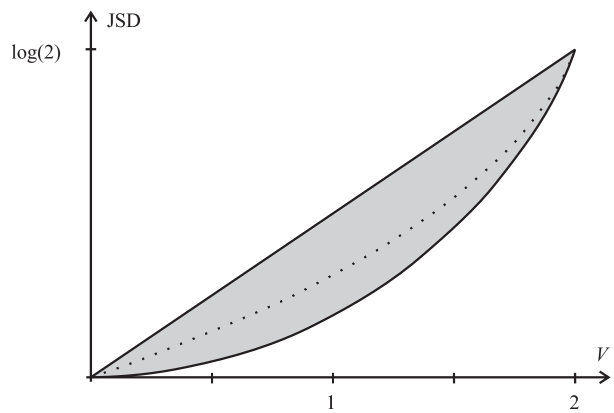

Fig. 1. V/JSD_1-diagram for $n \geq 3$ (the shaded region) and for $n = 2$ (the region obtained by replacing the upper bounding curve by the dotted curve)

In Figure 1 we have depicted the V/JSD_1-diagram for $n \geq 3$. By expansion of $L(V)$ given by (31), one obtains the inequality

$$\operatorname{JSD}_1(P,Q) \geq \sum_{n=1}^{\infty} \frac{1}{2n(2n-1)} \left(\frac{V}{2}\right)^{2n}, \tag{34}$$

which was proved by a different method in [18, Theorem 5].

5.3 Bounds on the QJSD

Analogous to the bounds given in Theorem 8, we can bound QJSD_1. We use the following well-known fact.

Fact 1 (Lindblad-Uhlmann)**.** *Let $\rho, \sigma \in \mathcal{B}_+^1(\mathcal{H})$ be two states on some Hilbert space $\mathcal{H}$ and $\mathcal{M} := \{M_i \mid i = 1, 2, \dots\}$ be a measurement on $\mathcal{H}$. Then $S(\rho\|\sigma) \geq D(P_\mathcal{M}\|Q_\mathcal{M})$, where $P_\mathcal{M}$ and $Q_\mathcal{M}$ are probability distributions with point probabilities $P_\mathcal{M}(i) = \operatorname{Tr}(M_i\rho)$ and $Q_\mathcal{M}(i) = \operatorname{Tr}(M_i\sigma)$ respectively.*

Theorem 11 (Nielsen & Chuang, [6] Theorem 9.1)**.** *Let $\rho, \sigma \in \mathcal{B}^1_+(\mathcal{H})$ be two states on some Hilbert space $\mathcal{H}$ and $\mathcal{M} := \{M_i \mid i = 1, 2, \dots\}$ be a measurement on $\mathcal{H}$. Then $\|\rho - \sigma\|_1 = \max_{\mathcal{M}} V(P_{\mathcal{M}}, Q_{\mathcal{M}})$, where $P_{\mathcal{M}}$ and $Q_{\mathcal{M}}$ are probability distributions with point probabilities $P_{\mathcal{M}}(i) = \mathrm{Tr}(M_i\rho)$ and $Q_{\mathcal{M}}(i) = \mathrm{Tr}(M_i\sigma)$ respectively.*

Theorem 12. *For all states $\rho, \sigma \in \mathcal{B}^1_+(\mathcal{H})$, we have*

$$\mathrm{QJSD}_1(\rho, \sigma) \geq L(\|\rho - \sigma\|_1)\,,$$

where L is the function given by (31).

Proof: In the same way as the proof of Theorem III.1 in [24], we make 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 11 we have $\|\rho - \sigma\|_1 = V(P_{\mathcal{M}}, Q_{\mathcal{M}})$. Fact 1 gives us

$$\begin{aligned} \mathrm{QJSD}_1(\rho, \sigma) &\geq \frac{1}{2} D\left(P_{\mathcal{M}}, \frac{P_{\mathcal{M}} + Q_{\mathcal{M}}}{2}\right) + \frac{1}{2} D\left(Q_{\mathcal{M}}, \frac{P_{\mathcal{M}} + Q_{\mathcal{M}}}{2}\right) \\ &= \mathrm{JSD}_1(P_{\mathcal{M}}, Q_{\mathcal{M}}). \end{aligned}$$

The result now follows from Theorem 9 and equation (34). ■

6 Further bounds

If Φ is a Markov kernel then the data reduction inequality for divergence implies that

$$d(\Phi(P), \Phi(Q)) \leq d(P, Q).$$

For product measures this implies

$$\begin{aligned} d(P_1 \times P_2, Q_1 \times Q_2) &\leq d(P_1 \times P_2, Q_1 \times P_2) + d(Q_1 \times P_2, Q_1 \times Q_2) \\ &\leq d(P_1, Q_1) + d(P_2, Q_2). \end{aligned}$$

The inequality

$$d(P_1 \times P_2, Q_1 \times Q_2) \geq \frac{1}{2}\left(d(P_1, Q_1) + d(P_2, Q_2)\right)$$

is a trivial consequence of the data reduction inequality.

7 Conclusions and open problems

Using the Tsallis entropy we defined generalizations of the (general) Jensen-Shannon divergence and its quantum analogue. JSD_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 QJSD_α restricted to qubit- and pure-states. Both these results were derived by evoking Schoenberg's theorem and showing that these quantities are negative semidefinite.

Whether QJSD_1 is a metric for all states remains unknown. However, based on a large amount of numerical evidence, we conjecture the kernel $A \curvearrowright \mathrm{Tr}(e^A)$ to be exponentially convex for Hermitian matrices A. Proving this would imply that QJSD_α is negative semidefinite, and hence the square of a metric that can be embedded in a real Hilbert space.

References

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

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

[3] 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.

[4] P. W. Lamberti, A. P. Majtey, A. Borras, M. Casas, and A. Plastino. Metric character of the quantum Jensen-Shannon divergence. *Physical Review A (Atomic, Molecular, and Optical Physics)*, 77(5):052311, 2008.

[5] A. P. Majtey, P. W. Lamberti, and D. P. Prato. Jensen-Shannon divergence as a measure of distinguishability between mixedquantum states. *Physical Review A (Atomic, Molecular, and Optical Physics)*, 72(5):052310, 2005.

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

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

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

[9] P. Hausladen, R. Jozsa, B. Schumacher, M. Westmoreland, and W. Wootters. Classical information capacity of a quantum channel. *Physical Review A*, 54:1869, 1996.

[10] L. B. Levitin. Information Theory for Quantum Systems. *Information, Complexity and Control in Quantum systems*, 1987. Edited by A. Blaquière, S. Diner and G. Lochak.

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

[12] B. Schumacher. Quantum coding. *Physical Review A*, 51:2738–2747, April 1995.

[13] 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.

[14] M. J. Donald. *Math. Proc. Cam. Phil. Soc.*, 101:363, 1987.

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

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

[17] J. Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Infor. Theory*, 37:145–151, 1991.

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

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

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

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

[22] P. Harremoës and F. Topsøe. Maximum entropy fundamentals. *Entropy*, 3(3):191–226, Sept. 2001.

[23] A. Fedotov, P. Harremoës, and F. Topsøe. Refinements of Pinsker's inequality. *IEEE Trans. Inform. Theory*, 49(6):1491–1498, June 2003.

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