Open access • Journal Article • DOI:10.1063/1.4901039

Quantum skew divergence — Source link

Koenraad M.R. Audenaert

Institutions: Royal Holloway, University of London

Published on: 11 Nov 2014 - Journal of Mathematical Physics (American Institute of Physics)

Topics: Jensen–Shannon divergence, Quantum relative entropy, Quantum algorithm, Quantum discord and Kullback–Leibler divergence

Quantum Skew Divergence

Koenraad M.R. Audenaert[1,2]

[1] *Department of Mathematics, Royal Holloway University of London, Egham TW20 0EX, U.K.*

[2] *Department of Physics and Astronomy, University of Ghent, S9, Krijgslaan 281, B-9000 Ghent, Belgium*

Electronic address: koenraad.audenaert@rhul.ac.uk

October 23, 2014

Abstract

In this paper we study the quantum generalisation of the skew divergence, which is a dissimilarity measure between distributions introduced by L. Lee in the context of natural language processing. We provide an in-depth study of the quantum skew divergence, including its relation to other state distinguishability measures. Finally, we present a number of important applications: new continuity inequalities for the quantum Jensen-Shannon divergence and the Holevo information, and a new and short proof of Bravyi's Small Incremental Mixing conjecture.

1 Introduction

The quantum relative entropy of two density operators ρ and σ, denoted $S(\rho||\sigma) = \mathrm{Tr}\,\rho(\log\rho - \log\sigma)$, was introduced by Umegaki [34] in 1962. Since the 90's it gained in prominence, especially in the quantum information theory community, when Hiai and Petz [15] showed that Umegaki's formula provided the proper quantum generalisation of the classical Kullback-Leibler divergence $\mathrm{KL}(p||q)$ of two probability distributions, as an operational measure of dissimilarity between quantum states. A lot of research has been spent exploring its mathematical and physical properties. Despite having many universally useful features, the relative entropy exhibits certain properties that in some applications may be considered as drawbacks. In particular, the relative entropy is not a distance measure in the mathematical sense of the word: it is asymmetric with respect to interchanging arguments, $S(\rho||\sigma) \neq S(\sigma||\rho)$, and it does not satisfy a triangle inequality. Moreover, the relative entropy is infinite whenever the support of σ is not contained in the support of ρ. This makes the relative entropy completely unsuitable as a distance measure between pure states, for example. We will refer to this feature as the 'infinity problem'.

Over the years, several modifications to the relative entropy have been proposed. Some of the better known modifications are the Quantum Jensen-Shannon divergence [13, 14], and the closely related Holevo information or Holevo χ [16, 25] (even though this is not usually considered as a modification of the relative entropy in the QIT community because it serves entirely different purposes).

In the present paper we introduce another modification of the quantum relative entropy, which we call the *quantum skew divergence*. We have coined this term [1] because of its close similarity to the already existing classical concept of skew divergence of two probability distributions, which was introduced by Lee [17, 18] in the context of natural language processing to overcome the infinity problem for the Kullback-Leibler divergence. As no confusion will arise we will henceforth refer to the quantum skew divergence as skew divergence (SD) for short. It is not to be confused with the Wigner-Yanase-Dyson skew information and related notions, to which it bears no obvious resemblance.

The skew divergence is essentially the relative entropy but with 'skewed' second argument. That is, the second argument σ is replaced by the convex combination $\alpha\rho+(1-\alpha)\sigma$, where α is a scalar $(0 < \alpha < 1)$ which we call the *skewing parameter*. As one of its basic properties we will show that $S(\rho||\alpha\rho + (1-\alpha)\sigma)$ is no longer infinite but is bounded above by $-\log\alpha$, and we define the skew divergence as the skewed relative entropy divided by this factor $-\log\alpha$:

$$\mathrm{S}_\alpha(\rho||\sigma) := \frac{1}{-\log\alpha} S(\rho||\alpha\rho + (1-\alpha)\sigma).$$

Hence, S_α always takes values between 0 and 1. It is to be noted that Lee's skew divergence does not have this normalisation factor.

This paper can be subdivided roughly in two parts: the first part is a theoretical study of the properties of the skew divergence, and the second part is on applications. The first part consists of six sections. After some preliminaries (Section 2), in Section 3 we give precise definitions for the skew divergence and state and prove its basic properties.

Sections 6 and 7 are devoted to the more complicated continuity properties of the quantum skew divergence. These are properties that have no counterparts for the relative entropy, as a direct consequence of the infinity problem. First, we show that continuity holds in the sense that states that are close in trace norm distance are also close when measured by the SD (Section 6). Secondly, we show that the SD is also continuous with respect to perturbations of each of its arguments (Section 7). The proofs of these statements rely on some technical results about the derivatives of the operator logarithm, and this is presented in Sections 4 and 5.

In the second part of this paper we consider applications of the quantum skew divergence. In Section 8 we give a simple proof of the so-called Small Incremental Mixing Conjecture that was postulated by Bravyi [8] and recently proven by Van Acoleyen [33]. Our proof yields a better proportionality constant (2 instead of 9) and may yield additional insight into the more general 'mixing problem' proposed by Lieb and Vershynina [22].

The second application (Section 9) is as a dissimilarity measure between quantum states, being the original purpose for introducing the skew divergence. Here we give a detailed overview of the relative entropy's drawbacks and of the various proposals that have been made in the literature and how the skew divergence fits in.

In Section 10 we note the close connection between SD and the generalised quantum Jensen-Shannon divergence (QJS), i.e. the Holevo information. By exploiting the sharp continuity estimates for the SD derived in this paper, we obtain new continuity-type bounds for the QJS and the Holevo information that in many cases improve on existing estimates from the literature.

2 Preliminaries

First, let us recall the definition of the quantum relative entropy [26, 28, 36]. For quantum states ρ and σ, both positive,

$$S(\rho||\sigma) := \mathrm{Tr}\,\rho(\log\rho - \log\sigma). \tag{1}$$

For non-normalised positive operators A and B, one defines more generally

$$S(A||B) := \mathrm{Tr}\,A(\log A - \log B) - \mathrm{Tr}(A - B). \tag{2}$$

For positive scalars a and b, we will also write

$$S(a|b) := a(\log a - \log b) - (a - b). \tag{3}$$

Strictly speaking, when σ (or B) is not invertible, the quanutm relative entropy is no longer defined. However, when the supports of ρ and σ satisfy the condition $\mathrm{supp}\,\rho \subseteq \mathrm{supp}\,\sigma$ one customarily adopts the convention that '$0^+ \log 0^+ = 0$' and redefines the relative entropy as

$$\begin{aligned} S(\rho||\sigma) &:= S(\rho|_\sigma||\sigma|_\sigma), \\ S(A||B) &:= S(A|_B||B|_B), \end{aligned}$$

where the symbol $A|_B$ denotes the restriction of A to the support of B. When $\mathrm{supp}\,\rho \not\subseteq \mathrm{supp}\,\sigma$ this redefinition is not possible and one says that the relative entropy is infinite, leading to the infinity problem mentioned in the introduction.

Another important distance measure between density operators is the trace norm distance:

$$T(\rho,\sigma) := \frac{1}{2}||\rho - \sigma||_1,$$

where $||.||_1$ denotes the trace norm,

$$||X||_1 := \mathrm{Tr}\,|X| = \mathrm{Tr}(X^*X)^{1/2}.$$

For any self-adjoint operator X, let X_+ and X_- denote the positive part $X_+ = (X + |X|)/2$ and negative part $X_- = (|X| - X)/2$; both parts are positive semidefinite (note that the negative part is positive for the same reason that the imaginary part of a complex number is real). Then another expression for the trace norm distance is

$$T(\rho,\sigma) = \mathrm{Tr}(\rho - \sigma)_+ = \mathrm{Tr}(\rho - \sigma)_-.$$

3 Quantum Skew Divergence

In this section we give a rigorous definition of the quantum generalisation of the skew divergence (SD) and state and prove its basic properties.

The quantum skew divergence is based on the functional $S(\rho||\alpha\rho+(1-\alpha)\sigma)$, or $S(A||\alpha A+(1-\alpha)B)$ in the non-normalised case, where α is a scalar, with $0 < \alpha < 1$. Since, for all such α, $\mathrm{supp}(A) \subseteq \mathrm{supp}(A + B) = \mathrm{supp}(\alpha A + (1-\alpha)B)$, no problem of infinities arises. Henceforth, we will always write $S(A||\alpha A + (1-\alpha)B)$, whether $A, B > 0$ or $A, B \geq 0$. In the latter case this is to mean $S(A|_{A+B}||(\alpha A + (1-\alpha)B)|_{A+B})$.

Definition 1. *For fixed $\alpha \in (0,1)$, the* quantum α-skew divergence *between states ρ and σ is defined as*

$$\mathrm{S}_\alpha(\rho||\sigma) := \frac{1}{-\log(\alpha)}\, S(\rho||\alpha\rho + (1-\alpha)\sigma). \tag{4}$$

Likewise, for non-normalised operators $A, B \geq 0$,

$$\mathrm{S}_\alpha(A||B) := \frac{1}{-\log(\alpha)}\, S(A||\alpha A + (1-\alpha)B). \tag{5}$$

We call α the *skewing parameter*.

The reason for incorporating the scale factor $1/(-\log\alpha)$ is to normalise the range of the SD to the interval $[0,1]$.

Theorem 1. *For all states ρ and σ and $0 < \alpha < 1$,*

$$0 \leq \mathrm{S}_\alpha(\rho||\sigma) \leq 1,$$

and $\mathrm{S}_\alpha(\rho||\sigma) = 1$ if and only if $\rho \perp \sigma$.

Recall that two quantum states are mutually orthogonal, denoted $\rho \perp \sigma$, iff $\operatorname{Tr}\rho\sigma = 0$.

Proof. Let $\tau = \alpha\rho + (1-\alpha)\sigma$. By operator monotonicity of the logarithm, we have

$$\log(\tau) = \log(\alpha\rho + (1-\alpha)\sigma) \geq \log(\alpha\rho),$$

and, therefore,

$$\begin{aligned} S(\rho||\tau) &= \operatorname{Tr}\rho(\log\rho - \log\tau) \\ &\leq \operatorname{Tr}\rho(\log\rho - \log(\alpha\rho)) \\ &= -\log\alpha. \end{aligned}$$

Thus, $S(\rho||\tau)$ is bounded above by $-\log\alpha$, which is finite for $0 < \alpha < 1$. It therefore makes perfect sense to normalise $S(\rho||\tau)$ by dividing it by $-\log\alpha$, producing a quantity that is always between 0 and 1.

The equality case was proven in [2]. □

The definition of the skew divergence for non-normalised operators is also applicable to non-negative scalars. To distinguish the scalar case more clearly from the matrix case we will use the symbol $\mathrm{S}_\alpha(b|c)$ for scalars; we have

$$\mathrm{S}_\alpha(b|c) = \frac{b(\log b - \log(\alpha b + (1-\alpha)c)) - (1-\alpha)(b-c)}{-\log\alpha}. \tag{6}$$

As we do not restrict the arguments of the SD to be normalised states, the following *scaling identities* can be useful.

Theorem 2. *For $0 < \alpha < 1$, operators $X, Y \geq 0$, and positive scalars b, c,*

$$\mathrm{S}_\alpha(bX||bY) = b\,\mathrm{S}_\alpha(X||Y) \tag{7}$$

$$\mathrm{S}_\alpha(bX||cX) = \mathrm{S}_\alpha(b|c)\operatorname{Tr}X. \tag{8}$$

This is easy to prove by simple calculation.

The quantum skew divergence inherits many desirable properties from the quantum relative entropy:

Theorem 3. *For $0 < \alpha < 1$, states ρ, σ, any unitary matrix U and any completely positive trace-preserving (CPTP) map Φ,*

1. Positivity: $\mathrm{S}_\alpha(\rho||\sigma) \geq 0$, *and* $\mathrm{S}_\alpha(\rho||\sigma) = 0$ *if and only if* $\rho = \sigma$;
2. Unitary invariance: $\mathrm{S}_\alpha(U\rho U^*||U\sigma U^*) = \mathrm{S}_\alpha(\rho||\sigma)$;
3. Contractivity: $\mathrm{S}_\alpha(\Phi(\rho)||\Phi(\sigma)) \leq \mathrm{S}_\alpha(\rho||\sigma)$;
4. Joint convexity: *the map* $(\rho, \sigma) \mapsto \mathrm{S}_\alpha(\rho||\sigma)$ *is jointly convex.*

The proof is again straightforward. Note that these are the same properties that the quantum Jensen-Shannon divergence obeys [24].

4 The Operator Logarithm and its Derivatives

The following integral representation of the logarithm lies at the basis of much of the subsequent treatment. For $x > 0$, we have

$$\log x = \int_0^\infty ds \left(\frac{1}{1+s} - \frac{1}{x+s} \right). \tag{9}$$

Using functional calculus, this definition can be extended to the operator logarithm. For $A > 0$,

$$\log A = \int_0^\infty ds \left(\frac{1}{1+s}\mathbb{I} - (A + s\mathbb{I})^{-1} \right). \tag{10}$$

From this representation follow representations of the first and second derivatives of the operator logarithm.

4.1 First Derivative

Following [21], let us define for $A > 0$ the linear map $\Delta \to \mathcal{T}_A(\Delta)$ for self-adjoint Δ as the Fréchet derivative of the operator logarithm:

$$\mathcal{T}_A(\Delta) := \left. \frac{d}{dt} \right|_{t=0} \log(A + t\Delta). \tag{11}$$

From integral representation (9) we get an integral representation for $\mathcal{T}_A$ as well:

$$\mathcal{T}_A(\Delta) = \int_0^\infty ds \; (A + s\mathbb{I})^{-1} \Delta (A + s\mathbb{I})^{-1}. \tag{12}$$

Here we have used the fact that

$$\frac{d}{dt}(A + t\Delta)^{-1} = -(A + t\Delta)^{-1}\Delta(A + t\Delta)^{-1}.$$

Being a positive linear combination of conjugations it follows that, for any $A > 0$, $\mathcal{T}_A$ is a completely positive map. In particular, it preserves the positive semidefinite order; that is, if $X \leq Y$, then $\mathcal{T}_A(X) \leq \mathcal{T}_A(Y)$. Also, $X > 0$ implies $\mathcal{T}_A(X) > 0$.

Lemma 1. *For $A > 0$ and $\Delta = \Delta^*$, and scalars $a > 0$ and δ,*

$$\mathcal{T}_{aA}(\delta\Delta) = \frac{\delta}{a}\mathcal{T}_A(\Delta). \tag{13}$$

Furthermore,

$$\mathcal{T}_A(A) = \mathbb{I}. \tag{14}$$

Proof. For the first identity:

$$\begin{aligned}\mathcal{T}_{aA}(\delta\Delta) &= \frac{d}{dt}\Big|_{t=0} \log(aA + t\delta\Delta) = \frac{d}{dt}\Big|_{t=0} \log(A + t(\delta/a)\Delta) \\ &= \mathcal{T}_A((\delta/a)\Delta) = (\delta/a)\mathcal{T}_A(\Delta).\end{aligned}$$

The second identity follows similarly from the fact that $\log(A + tA) = (1 + t)\mathbb{I} + \log A$, and the term $\log A$ drops out after differentiating. □

Hence, for scalar arguments we have

$$\mathcal{T}_a(\delta) = \delta/a. \tag{15}$$

Lemma 2. *For $A, B \geq 0$ with $A + B > 0$*

$$\mathcal{T}_{A+B}(A) \leq \mathbb{I}. \tag{16}$$

Proof. Since $B \geq 0$, we have $A + B \geq A$ and because $\mathcal{T}_{A+B}$ preserves the positive semidefinite ordering, $\mathcal{T}_{A+B}(A) \leq \mathcal{T}_{A+B}(A + B) = \mathbb{I}$. □

4.2 The metric $M_A(B, C)$

The sesquilinear form

$$M_A(B, C) := \langle B^*, \mathcal{T}_A(C)\rangle = \operatorname{Tr} B^*\mathcal{T}_A(C) \tag{17}$$

which is defined for $A > 0$, is a *metric*: it is self-adjoint ($M_A(B, C) = \overline{M_A(C, B)}$), positive semidefinite ($M_A(B, B) \geq 0$ for any B), with $M_A(B, B) = 0$ iff $B = 0$, and $M_A(B, B)$ is continuous in B for any A. As the metric is contractive under completely positive trace-preserving (CPTP) maps Φ,

$$M_{\Phi(A)}(\Phi(B), \Phi(B)) \leq M_A(B, B)$$

for any $A > 0$ and any B, it is a *monotone metric* [20, 27]. Lieb has shown that the map $(A, B) \mapsto M_A(B, B)$, for $A > 0$ and any B, is jointly convex in A and B ([21], Theorem 3).

M satisfies the following limit property:

Lemma 3. *Let $A, B, C \geq 0$ with $B + C > 0$ and $\operatorname{supp} A \subseteq \operatorname{supp} B$. Then*

$$\lim_{\epsilon\to 0} M_{B+\epsilon C}(A, A) = M_{B|_B}(A|_B, A|_B).$$

Proof. Let P be the projector on $\operatorname{supp} B$ and Q the projector on the orthogonal complement of $\operatorname{supp} B$. Consider the 2×2 partitioning induced by P and Q:

$$A \to \begin{pmatrix} PAP^* & PAQ^* \\ QAP^* & QAQ^* \end{pmatrix},$$

and similarly for all other operators. Because of the conditions on the supports, we have $PAQ^* = QAP^* = QAQ^* = 0$ and $PBQ^* = QBP^* = QBQ^* = 0$. Hence,

$$\begin{aligned}
&\operatorname{Tr} A\mathcal{T}_{B+\epsilon C}(A) \\
&= \int_0^\infty ds\ \operatorname{Tr} A(B+\epsilon C+s)^{-1}A(B+\epsilon C+s)^{-1} \\
&= \int_0^\infty ds\ \operatorname{Tr}(PAP^*)\ (P(B+\epsilon C+s)^{-1}P^*)\ (PAP^*)\ (P(B+\epsilon C+s)^{-1}P^*).
\end{aligned}$$

Using Schur complements, we can find the explicit expression

$$\begin{aligned}
&P(B+\epsilon C+s)^{-1}P^* \\
&\quad = \left((PBP^*+\epsilon PCP^*+s) - \epsilon^2 PCQ^*(\epsilon QCQ^*+s)^{-1}QCP^*\right)^{-1}.
\end{aligned}$$

In the limit $\epsilon \to 0$, this simplifies as

$$\lim_{\epsilon\to 0} P(B+\epsilon C+s)^{-1}P^* = (PBP^*+s)^{-1},$$

since all operator blocks appearing here are invertible. Therefore,

$$\begin{aligned}
\lim_{\epsilon\to 0} \operatorname{Tr} A\mathcal{T}_{B+\epsilon C}(A) &= \int_0^\infty ds\ \operatorname{Tr}(PAP^*)\ (PBP^*+s)^{-1}\ (PAP^*)\ (PBP^*+s)^{-1} \\
&= \operatorname{Tr} A|_B \mathcal{T}_{B|_B}(A|_B).
\end{aligned}$$

□

4.3 Second Derivative

Having defined the linear operator $\mathcal{T}$ via the first derivative of the logarithm, we can also define a quadratic operator $\mathcal{R}$ via the second derivative [21]. For $A > 0$ and Δ self-adjoint, let

$$\mathcal{R}_A(\Delta) := -\frac{d^2}{dt^2}\bigg|_{t=0} \log(A+t\Delta). \tag{18}$$

A simple calculation using the integral representation of the first derivative yields the integral representation

$$\mathcal{R}_A(\Delta) = 2\int_0^\infty ds\ (A+s\mathbb{I})^{-1}\Delta(A+s\mathbb{I})^{-1}\Delta(A+s\mathbb{I})^{-1}. \tag{19}$$

One can similarly define a bilinear form, for $A > 0$ and self-adjoint Δ_1 and Δ_2:

$$\begin{aligned}
\mathcal{R}_A(\Delta_1,\Delta_2) &:= -\frac{d^2}{dt_1 dt_2}\bigg|_{t_1=t_2=0} \log(A+t_1\Delta_1+t_2\Delta_2) &(20)\\
&= \int_0^\infty ds\ (A+s\mathbb{I})^{-1}\Delta_1(A+s\mathbb{I})^{-1}\Delta_2(A+s\mathbb{I})^{-1} \\
&\quad + \int_0^\infty ds\ (A+s\mathbb{I})^{-1}\Delta_2(A+s\mathbb{I})^{-1}\Delta_1(A+s\mathbb{I})^{-1}. &(21)
\end{aligned}$$

Clearly,

$$\begin{aligned} \mathcal{R}_A(\Delta,\Delta) &= \mathcal{R}_A(\Delta), &(22)\\ \mathcal{R}_A(\Delta_1,\Delta_2) &= \mathcal{R}_A(\Delta_2,\Delta_1), &(23)\\ \operatorname{Tr}\Delta_0\mathcal{R}_A(\Delta_1,\Delta_2) &= \operatorname{Tr}\Delta_2\mathcal{R}_A(\Delta_0,\Delta_1). &(24)\end{aligned}$$

It is readily checked that for scalar a and δ we have

$$\mathcal{R}_{aA}(\delta\Delta) = (\delta/a)^2\mathcal{R}_A(\Delta) \tag{25}$$

and

$$\mathcal{R}_a(\delta) = (\delta/a)^2. \tag{26}$$

Lemma 4. *For $A > 0$ and $\Delta = \Delta^*$*

$$\mathcal{R}_A(A,\Delta) = \mathcal{T}_A(\Delta).$$

Hence

$$\mathcal{R}_A(A) = \mathbb{I}.$$

Proof.

$$\begin{aligned} \mathcal{R}_A(A,\Delta) &= -\frac{d^2}{dt_1dt_2}\bigg|_{t_1=t_2=0}\log(A+t_1A+t_2\Delta)\\ &= -\frac{d^2}{dt_1dt_2}\bigg|_{t_1=t_2=0}\log(1+t_1)\mathbb{I}+\log(A+t_2/(1+t_1)\Delta)\\ &= -\frac{d}{dt_1}\bigg|_{t_1=0}\frac{d}{dt_2}\bigg|_{t_2=0}\log(A+t_2/(1+t_1)\Delta).\end{aligned}$$

The derivative w.r.t. t_2 is, with $u = t_2/(1+t_1)$,

$$\begin{aligned} \frac{d}{dt_2}\bigg|_{t_2=0}\log(A+t_2/(1+t_1)\Delta) &= \frac{d}{du}\bigg|_{u=0}\log(A+u\Delta)\frac{1}{1+t_1}\\ &= \mathcal{T}_A(\Delta)\frac{1}{1+t_1}.\end{aligned}$$

Therefore,

$$\mathcal{R}_A(A,\Delta) = -\frac{d}{dt_1}\bigg|_{t_1=0}\mathcal{T}_A(\Delta)\frac{1}{1+t_1} = \mathcal{T}_A(\Delta).$$

□

Lemma 5. *For $A, B \geq 0$, with $A + B > 0$,*

$$\mathcal{R}_{A+B}(A) \leq \mathbb{I}. \tag{27}$$

Proof. Due to the bilinearity of $\mathcal{R}_A(\Delta_1, \Delta_2)$ and Lemma 4, we have

$$\begin{aligned}\mathcal{R}_{A+B}(A) &= \mathcal{R}_{A+B}(A+B-B) = \mathcal{R}_{A+B}(A+B-B, A+B-B)\\ &= \mathcal{R}_{A+B}(A+B, A+B) + \mathcal{R}_{A+B}(-B, A+B)\\ &\quad + \mathcal{R}_{A+B}(A+B, -B) + \mathcal{R}_{A+B}(-B, -B)\\ &= \mathbb{I} - 2\mathcal{T}_{A+B}(B) + \mathcal{R}_{A+B}(B).\end{aligned}$$

The third term can be bounded in terms of the second. Since $A+B+s\mathbb{I} \geq B$, for any $s \geq 0$, we have $(A+B+s\mathbb{I})^{-1} \leq B^{-1}$ and $B(A+B+s\mathbb{I})^{-1}B \leq B$. Therefore,

$$\begin{aligned}\mathcal{R}_{A+B}(B) &= 2\int_0^\infty ds\ (A+B+s\mathbb{I})^{-1}\ B(A+B+s\mathbb{I})^{-1}B\ (A+B+s\mathbb{I})^{-1}\\ &\leq 2\int_0^\infty ds\ (A+B+s\mathbb{I})^{-1}\ B\ (A+B+s\mathbb{I})^{-1}\\ &= 2\mathcal{T}_{A+B}(B).\end{aligned}$$

We finally get

$$\mathcal{R}_{A+B}(A) \leq \mathbb{I} - 2\mathcal{T}_{A+B}(B) + 2\mathcal{T}_{A+B}(B) = \mathbb{I}.$$

□

5 A Continuity Inequality for the metric M

In this section we will prove the following technical inequality for the metric M, which will be used heavily in the proofs of the continuity properties of the quantum skew divergence.

Theorem 4. *For $A, B, C \geq 0$ with $A+B > 0$, and with $a = \operatorname{Tr} A$, $c = \operatorname{Tr} C$,*

$$0 \leq M_{A+B}(A, A) - M_{A+B+C}(A, A) \leq M_a(a, a) - M_{a+c}(a, a) \tag{28}$$

or, explicitly,

$$0 \leq \operatorname{Tr} A\mathcal{T}_{A+B}(A) - \operatorname{Tr} A\mathcal{T}_{A+B+C}(A) \leq a - \frac{a^2}{a+c}. \tag{29}$$

To prove this theorem, we need the following lemma:

Lemma 6. *Let $f(t)$ be a real-valued convex function on $[0,1]$. If, moreover, $f(0) \leq 0$ and $f(0) \leq f'(0)$, then $\forall t \in [0,1], f(0) \leq (1-t)f(t)$.*

Proof. Since $f(0) \leq 0$, for all $t \in [0,1]$ we have $f(0) \leq f(0)(1-t) \leq f'(0)(1-t)$. Multiplying both sides by t and adding $(1-t)f(0)$ gives $f(0) \leq t(1-t)f'(0) + (1-t)f(0) = (1-t)(f(0) + tf'(0))$. By convexity of f, $f(0) + tf'(0)$ is a lower bound on $f(t)$, and the inequality of the lemma follows. □

Proof of Theorem 4. The first inequality in (29) easily follows from the fact that $x \mapsto 1/x$ is operator monotone decreasing together with the identity

$$\operatorname{Tr} X\mathcal{T}_A(X) = \int_0^\infty ds\ \operatorname{Tr}(X^{1/2}(A+s\mathbb{I})^{-1}X^{1/2})^2,$$

and monotonicity of the function $X \to \operatorname{Tr} X^2$.

The second inequality involves more work. Let us thereto consider two positive density operators ρ and σ, an operator $G \geq \rho$, and the function

$$\begin{aligned} f(t) &= -1 + \frac{d}{ds}\bigg|_{s=0} \operatorname{Tr} \rho \log(G + s\rho + t(\sigma - G)) \\ &= \operatorname{Tr} \rho \mathcal{T}_{t\sigma+(1-t)G}(\rho) - 1. \end{aligned}$$

We start by showing that $(1-t)f(t) \geq f(0)$ for $0 \leq t \leq 1$.
Firstly, $f(0) = \operatorname{Tr} \rho \mathcal{T}_G(\rho) - 1$. Since $\mathcal{T}_G(\rho) \leq \mathcal{T}_G(G) = \mathbb{I}$, we have $f(0) \leq 0$.
Secondly, the derivative $f'(0)$ is given by

$$\begin{aligned} f'(0) &= \frac{d^2}{dsdt}\bigg|_{s=t=0} \operatorname{Tr} \rho \log(G + s\rho + t(\sigma - G)) \\ &= \operatorname{Tr} \rho \mathcal{R}_G(\rho, G - \sigma) = \operatorname{Tr} \rho \mathcal{R}_G(\rho, G) - \operatorname{Tr} \rho \mathcal{R}_G(\rho, \sigma). \end{aligned}$$

The first term can be rewritten as

$$\operatorname{Tr} \rho \mathcal{R}_G(G, \rho) = \operatorname{Tr} \rho \mathcal{T}_G(\rho),$$

by Lemma 4. Because $G \geq \rho$, the second term can be bounded using Lemma 5 as

$$\operatorname{Tr} \rho \mathcal{R}_G(\rho, \sigma) = \operatorname{Tr} \sigma \mathcal{R}_G(\rho) \leq \operatorname{Tr} \sigma = 1.$$

We therefore obtain

$$\operatorname{Tr} \rho \mathcal{R}_G(\rho, G - \sigma) \geq \operatorname{Tr} \rho \mathcal{T}_G(\rho) - 1,$$

which proves that $f'(0) \geq f(0)$.

By Lieb's convexity theorem, the map $G \mapsto \operatorname{Tr} \rho \mathcal{T}_G(\rho)$ is convex, hence $f(t)$ is convex.

All three conditions of Lemma 6 are therefore satisfied, so that $(1-t)f(t) \geq f(0)$, for $0 \leq t \leq 1$.

Now let $a > 0$, $c \geq 0$, and $G = \rho + B/a$, with $B \geq 0$; this choice indeed satisfies the condition $G \geq \rho$. With this substitution, we get

$$f(t) = \operatorname{Tr} \rho \mathcal{T}_{(1-t)\rho + \frac{1}{a}(1-t)B + t\sigma}(\rho) - 1.$$

In particular, with the choice $t = c/(a+c)$,

$$\begin{aligned} (1-t)f(t) &= \frac{a}{a+c}\left(\operatorname{Tr} \rho \mathcal{T}_{\frac{a}{a+c}\rho + \frac{1}{a+c}B + \frac{c}{a+c}\sigma}(\rho) - 1\right) \\ &= \operatorname{Tr} \rho \mathcal{T}_{\rho + \frac{1}{a}B + \frac{c}{a}\sigma}(\rho) - \frac{a}{a+c} \\ f(0) &= \operatorname{Tr} \rho \mathcal{T}_{\rho + \frac{1}{a}B}(\rho) - 1. \end{aligned}$$

The inequality $(1-t)f(t) \geq f(0)$ therefore gives (after multiplying by a) Multiplying by a yields

$$\operatorname{Tr} a\rho \mathcal{T}_{a\rho + B + c\sigma}(a\rho) - \frac{a^2}{a+c} \geq \operatorname{Tr} a\rho \mathcal{T}_{a\rho + B}(a\rho) - a.$$

or, after rearranging terms,

$$\operatorname{Tr} a\rho \mathcal{T}_{a\rho + B}(a\rho) - \operatorname{Tr} a\rho \mathcal{T}_{a\rho + B + c\sigma}(a\rho) \leq a - \frac{a^2}{a+c}.$$

Setting $A = a\rho$ and $C = c\sigma$ we obtain the second inequality of (29). □

6 Quantum Skew Divergence as Integral of the Metric M

The reason for considering the metric M in such detail as we have done, is that the quantum skew divergence can be conveniently written as an integral of M. To this purpose, let us introduce the following quantity based on M, which can be seen as a differential version of the SD:

Definition 2. *Let $A, B \geq 0$ such that $A + B > 0$. For $0 < \alpha < 1$, define*

$$\mathcal{D}_\alpha(A||B) := \alpha(1-\alpha) M_{\alpha A+(1-\alpha)B}(A-B, A-B). \tag{30}$$

For $\alpha = 0, 1$, define $\mathcal{D}_\alpha$ to be identically zero.

For general $A, B \geq 0$ not satisfying the condition $A + B > 0$, define $\mathcal{D}_\alpha(A||B)$ as $\mathcal{D}_\alpha(A|_{A+B}||B|_{A+B})$.

Other explicit formulas for $\mathcal{D}_\alpha$ are:

$$\begin{aligned} \mathcal{D}_\alpha(A||B) &= \alpha(\operatorname{Tr} A\mathcal{T}_{\alpha A+(1-\alpha)B}(A-B) - \operatorname{Tr}(A-B)) && (31)\\ &= \frac{\alpha}{1-\alpha}\operatorname{Tr} A\mathcal{T}_{\alpha A+(1-\alpha)B}(A) - \frac{\alpha}{1-\alpha}\operatorname{Tr} A - \alpha\operatorname{Tr}(A-B). && (32) \end{aligned}$$

These formulas follow from (30) by expressing $A - B$ as

$$A - B = \frac{1}{1-\alpha}(A - (\alpha A + (1-\alpha)B))$$

and exploiting the identities $\mathcal{T}_X(X) = \mathbb{I}$ and $\operatorname{Tr} A\mathcal{T}_X(B) = \operatorname{Tr} B\mathcal{T}_X(A)$.

We denote $\mathcal{D}_\alpha$ for scalar arguments by $\mathcal{D}_\alpha(b|c)$. Explicit formulas are

$$\begin{aligned} \mathcal{D}_\alpha(b|c) &= \alpha(1-\alpha)\frac{(b-c)^2}{\alpha b + (1-\alpha)c} && (33)\\ &= \frac{\alpha}{1-\alpha}\left(\frac{b^2}{\alpha b + (1-\alpha)c} - b\right) - \alpha(b-c). && (34) \end{aligned}$$

In particular,

$$\mathcal{D}_\alpha(b|0) = (1-\alpha)b, \qquad \mathcal{D}_\alpha(0|c) = \alpha c. \tag{35}$$

From the properties of M, it follows that $\mathcal{D}_\alpha$ is positive and contractive under CPTP maps. For example, with $a = \operatorname{Tr} A$ and $b = \operatorname{Tr} B$, we have:

$$\mathcal{D}_\alpha(A||B) \leq \mathcal{D}_\alpha(a|b). \tag{36}$$

Clearly, $\mathcal{D}_\alpha$ is unitarily invariant: for any unitary U, $\mathcal{D}_\alpha(UAU^*||UBU^*) = \mathcal{D}_\alpha(A||B)$.

A very useful property of $\mathcal{D}_\alpha$ is the following *symmetry property.*

Theorem 5. *For $A, B \geq 0$, and $0 < \alpha < 1$,*

$$\mathcal{D}_\alpha(A||B) = \mathcal{D}_{1-\alpha}(B||A). \tag{37}$$

Proof. This follows immediately from formula (30). □

We will now show how the quantum skew divergence is related to $\mathcal{D}_\alpha$. It is well-known that the quantum relative entropy $S(A||B)$ is differentiable w.r.t. A and B whenever $A, B > 0$. Hence, for $A, B > 0$, the function $\alpha \mapsto S(A||\alpha A + (1-\alpha)B)$ is differentiable over the open interval $(0, 1)$. For $A, B \geq 0$ this is no longer true as the relative entropy is in general only lower semicontinuous [36]. However, if one restricts A and B to the support of $A + B$, the function $\alpha \mapsto S(A||\alpha A + (1-\alpha)B)$ is still differentiable for $A, B \geq 0$. Because of this, the following connection between $\mathcal{D}_\alpha$ and S_α emerges:

Lemma 7. *For $A, B \geq 0$ and $0 < \alpha < 1$,*

$$\begin{aligned} \mathcal{D}_\alpha(A||B) &= \frac{d}{d(-\log\alpha)} S(A||\alpha A + (1-\alpha)B) && (38) \\ &= -\alpha \frac{d}{d\alpha} S(A||\alpha A + (1-\alpha)B). && (39) \end{aligned}$$

Conversely, S_α can be obtained from $\mathcal{D}_\alpha$ by a simple *averaging procedure.*

Theorem 6. *For operators $A, B \geq 0$ and $0 < \alpha < 1$,*

$$\mathrm{S}_\alpha(A||B) = \frac{1}{-\log\alpha} \int_0^{-\log\alpha} \mathcal{D}_{\alpha'}(A||B)\, d(-\log\alpha'). \tag{40}$$

Proof. Define the function $f(\alpha) = S(A||\alpha A + (1-\alpha)B)$. By the substitution $b = -\log\alpha$, we can write

$$\begin{aligned} \mathrm{S}_\alpha(A||B) &= \frac{1}{b} f(\exp(-b)) \\ \mathcal{D}_\alpha(A||B) &= \frac{d}{db} f(\exp(-b)). \end{aligned}$$

Therefore, as for $b = 0$, $f(\exp(-b)) = f(1) = S(A||A) = 0$,

$$\begin{aligned} \mathrm{S}_\alpha(A||B) &= \frac{1}{b} \int_0^b \frac{d}{db} f(\exp(-b))\, db \\ &= \frac{1}{-\log\alpha} \int_0^{-\log\alpha} \mathcal{D}_{\alpha'}(A||B)\, d(-\log\alpha'), \end{aligned}$$

which is indeed an average w.r.t. $-\log\alpha$. □

This is an important fact, because whenever one has an equality or inequality involving several instances of $\mathcal{D}_\alpha$ with the same value of α, one can immediately obtain the corresponding (in)equality for S_α by averaging over a suitable range of $-\log\alpha$.

To end this section, we use the averaging technique to derive sharp inequalities relating $\mathrm{S}_\alpha(\rho, \sigma)$ to the trace norm distance $T(\rho, \sigma)$. We will encounter further applications of this technique in the proofs of Proposition 1 and Theorems 11 and 12.

The quantity $\mathcal{D}_\alpha$ is related to one of the so-called *quantum χ^2-divergences* introduced by Temme *et al* [32], namely the one induced by the logarithm. This logarithmic quantum χ^2-divergence is defined for $A, B > 0$ as

$$\chi^2_{\log}(A, B) := M_B(A - B, A - B) = \mathrm{Tr}(A - B)\mathcal{T}_B(A - B).$$

A short calculation reveals that

$$\mathcal{D}_\alpha(A||B) = \frac{\alpha}{1-\alpha}\chi^2_{\log}(A, \alpha A + (1-\alpha)B). \tag{41}$$

This means that certain properties that were proven in [32] for the quantum χ^2-divergences carry over to $\mathcal{D}_\alpha$. One such property is the following lower bound on $\mathcal{D}_\alpha$ in terms of the trace norm distance $T(\rho, \sigma)$:

Theorem 7. *For all density operators ρ and σ and any $0 < \alpha < 1$,*

$$\mathcal{D}_\alpha(\rho||\sigma) \geq 4\alpha(1-\alpha)T(\rho, \sigma)^2. \tag{42}$$

Proof. This follows from Lemma 5 in [32] according to which $\chi^2(\rho, \sigma) \geq ||\rho - \sigma||_1^2$. With the substitution $\sigma \to \tau := \alpha\rho + (1-\alpha)\sigma$ and noting that $\rho - \tau = (1-\alpha)(\rho - \sigma)$, the inequality follows. □

We can also furnish an upper bound on $\mathcal{D}_\alpha$ in terms of the trace norm distance.

Theorem 8. *For density operators $\rho, \sigma \geq 0$ and $0 < \alpha < 1$,*

$$\mathcal{D}_\alpha(\rho||\sigma) \leq T(\rho, \sigma). \tag{43}$$

Proof. From formula (31) and the basic properties of $\mathcal{T}$,

$$\begin{aligned}
\mathcal{D}_\alpha(\rho||\sigma) &= \alpha \operatorname{Tr} \rho \mathcal{T}_{\alpha\rho+(1-\alpha)\sigma}(\rho - \sigma) \\
&= \alpha \operatorname{Tr}(\rho - \sigma)\mathcal{T}_{\alpha\rho+(1-\alpha)\sigma}(\rho) \\
&\leq \alpha \operatorname{Tr}(\rho - \sigma)_+ \mathcal{T}_{\alpha\rho+(1-\alpha)\sigma}(\rho) \\
&\leq \operatorname{Tr}(\rho - \sigma)_+ \mathcal{T}_{\alpha\rho+(1-\alpha)\sigma}(\alpha\rho + (1-\alpha)\sigma) \\
&= \operatorname{Tr}(\rho - \sigma)_+ = T(\rho, \sigma).
\end{aligned}$$

□

Using the averaging procedure, Theorem 6, we immediately get the promised relations for S_α:

Theorem 9. *For density operators $\rho, \sigma \geq 0$ and $0 < \alpha < 1$,*

$$\frac{2(1-\alpha)^2}{-\log(\alpha)} T(\rho, \sigma)^2 \leq \mathrm{S}_\alpha(\rho||\sigma) \leq T(\rho, \sigma). \tag{44}$$

To prove the lower bound we note that using (40) the factor $4\alpha(1-\alpha)$ averages to $2(1-\alpha)^2/(-\log(\alpha))$.

The upper bound shows that two states that are close in trace norm distance are also close in terms of S_α. Despite the very simple form of the upper bound, it is the strongest one possible. Equality can be obtained for any value of $t = T(\rho, \sigma)$ for states in dimension 3 (and higher), for example by choosing $\rho = \operatorname{Diag}(t, 0, 1-t)$ and $\sigma = \operatorname{Diag}(0, t, 1-t)$.

7 Continuity Properties of the Quantum Skew Divergence

The inequalities of Theorem 4 lead to several inequalities for $\mathcal{D}_\alpha$, which in turn lead to inequalities for the quantum skew divergence.

Theorem 10. *For $A, B, C \geq 0$ and $0 < \alpha < 1$, with $a = \operatorname{Tr} A$ and $c = \operatorname{Tr} C$,*

$$\begin{aligned} -\alpha c = -\mathcal{D}_\alpha(0|c) &\leq \mathcal{D}_\alpha(A||B) - \mathcal{D}_\alpha(A||B+C) \\ &\leq \mathcal{D}_\alpha(a|0) - \mathcal{D}_\alpha(a|c). \end{aligned} \tag{45}$$

$$\begin{aligned} 0 &\leq \mathcal{D}_\alpha(B||A+B) - \mathcal{D}_\alpha(B+C||A+B+C) \\ &\leq \mathcal{D}_\alpha(0|a) - \mathcal{D}_\alpha(c|a+c). \end{aligned} \tag{46}$$

Proof. Consider first the case $A, B, C > 0$ of inequalities (45). These follow from Theorem 4 and expressions (32) and (34). We have

$$\begin{aligned} &\mathcal{D}_\alpha(A||B) - \mathcal{D}_\alpha(A||B+C) \\ &= \frac{\alpha}{1-\alpha}\left(M_{\alpha A+(1-\alpha)B}(A,A) - M_{\alpha A+(1-\alpha)(B+C)}(A,A)\right) - \alpha \operatorname{Tr} C \\ &= \frac{1}{1-\alpha}\left(M_{A+\frac{1-\alpha}{\alpha}B}(A,A) - M_{A+\frac{1-\alpha}{\alpha}B+\frac{1-\alpha}{\alpha}C}(A,A)\right) - \alpha \operatorname{Tr} C. \end{aligned}$$

The first term is now of the form that allows Theorem 4 to be invoked and (45) follows immediately.

To treat the case $A, B, C \geq 0$ we use Lemma 3 to bring both terms on a 'common denominator' as far as supports are concerned. Whereas $\mathcal{D}_\alpha(A||B)$ is defined as $\mathcal{D}_\alpha(A|_{A+B}||B|_{A+B})$, and in the second term the operators are restricted to the potentially larger subspace $\operatorname{supp}(A+B+C)$, we can write

$$\mathcal{D}_\alpha(A||B) - \mathcal{D}_\alpha(A||B+C) = \lim_{\epsilon \to 0} \mathcal{D}_\alpha(A||B+\epsilon C) - \mathcal{D}_\alpha(A||B+C),$$

in which the operators in both terms are now restricted to the support of $A+B+C$, allowing to use the positive case, as before.

To prove the second set of inequalities (46) we use the expression (30) and the substitution $A' = (1-\alpha)A$ (so $\operatorname{Tr} A' = (1-\alpha)\operatorname{Tr} A$):

$$\begin{aligned} &\mathcal{D}_\alpha(B||A+B) - \mathcal{D}_\alpha(B+C||A+B+C) \\ &= \alpha(1-\alpha)\left(M_{\alpha B+(1-\alpha)(A+B)}(A,A) - M_{\alpha(B+C)+(1-\alpha)(A+B+C)}(A,A)\right) \\ &= \frac{\alpha}{1-\alpha}\left(M_{A'+B}(A',A') - M_{A'+B+C}(A',A')\right), \end{aligned}$$

which is again of the form required by Theorem 4. □

Equality in the lower bounds of (45) and (46) is attained for $A = \operatorname{Diag}(a, 0)$, $B = \operatorname{Diag}(b_1, b_2)$ and $C = \operatorname{Diag}(0, c)$. Equality in the upper bounds is attained for scalar A, B, C. Thus, the given bounds are the best possible among all bounds that are only based on a, c and α.

Theorem 10 immediately yields:

Proposition 1. *For operators $A, B, C \geq 0$, with $a = \operatorname{Tr} A$ and $c = \operatorname{Tr} C$,*

$$-\operatorname{S}_\alpha(0|c) \leq \operatorname{S}_\alpha(A||A+B) - \operatorname{S}_\alpha(A||A+B+C) \leq -\operatorname{S}_\alpha(a|a+c) \tag{47}$$

$$-S(0|c) \leq S(A||A+B) - S(A||A+B+C) \leq -S(a|a+c). \tag{48}$$

$$
\begin{aligned}
0 &\leq \mathrm{S}_\alpha(B||A+B) - \mathrm{S}_\alpha(B+C||A+B+C) \\
&\leq \mathrm{S}_\alpha(0|a) - \mathrm{S}_\alpha(c|a+c) \qquad (49) \\
0 &\leq S(B||A+B) - S(B+C||A+B+C) \\
&\leq S(0|a) - S(c|a+c). \qquad (50)
\end{aligned}
$$

Proof. Inequalities (47) follow by averaging those of (45) and noting that $\mathrm{S}_\alpha(a|a) = 0$. Inequalities (49) follow by averaging those of (46).

Then note that

$$
\begin{aligned}
(-\log\alpha)\,\mathrm{S}_\alpha(A||A+B) &= S(A||\alpha A + (1-\alpha)(A+B)) \\
&= S(A||A+(1-\alpha)B).
\end{aligned}
$$

Doing this for all the terms in (47) and absorbing the factors $(1-\alpha)$ in B, C and c yields (48). A similar procedure yields (50) from (49). □

From Theorem 10, it is easy to derive quantitative continuity properties for S_α. The following theorem gives bounds on the change of S_α (and $\mathcal{D}_\alpha$) when either of its arguments changes, as expressed by the trace distance. Here we restrict to density operators (trace equal to 1).

Theorem 11. *Let $0 < \alpha < 1$.*

For density operators ρ, σ_1, σ_2 such that $T(\sigma_1, \sigma_2) = t$,

$$
\begin{aligned}
|\mathcal{D}_\alpha(\rho||\sigma_1) - \mathcal{D}_\alpha(\rho||\sigma_2)| &\leq \mathcal{D}_\alpha(1|0) - \mathcal{D}_\alpha(1|t) + \mathcal{D}_\alpha(0|t) \qquad (51) \\
|\mathcal{D}_\alpha(\sigma_1||\rho) - \mathcal{D}_\alpha(\sigma_2||\rho)| &\leq \mathcal{D}_\alpha(0|1) - \mathcal{D}_\alpha(t|1) + \mathcal{D}_\alpha(t|0) \qquad (52)
\end{aligned}
$$

and

$$
\begin{aligned}
|\,\mathrm{S}_\alpha(\rho||\sigma_1) - \mathrm{S}_\alpha(\rho||\sigma_2)| &\leq \mathrm{S}_\alpha(1|0) - \mathrm{S}_\alpha(1|t) + \mathrm{S}_\alpha(0|t) \qquad (53) \\
|\,\mathrm{S}_\alpha(\sigma_1||\rho) - \mathrm{S}_\alpha(\sigma_2||\rho)| &\leq \mathrm{S}_\alpha(0|1) - \mathrm{S}_\alpha(t|1) + \mathrm{S}_\alpha(t|0). \qquad (54)
\end{aligned}
$$

Proof. Let $A, B_1, B_2 \geq 0$. A successive application of the first and then the second inequality of (45) yields

$$
\begin{aligned}
&\mathcal{D}_\alpha(A||B_1) - \mathcal{D}_\alpha(A||B_2) \\
&\quad= \mathcal{D}_\alpha(A||B_1) - \mathcal{D}_\alpha(A||B_1 + (B_2-B_1)_+ - (B_2-B_1)_-) \\
&\quad\leq \mathcal{D}_\alpha(A||B_1) - \mathcal{D}_\alpha(A||B_1 + (B_2-B_1)_+) + \mathcal{D}_\alpha(0|\operatorname{Tr}(B_2-B_1)_-) \\
&\quad\leq \mathcal{D}_\alpha(\operatorname{Tr}A|0) - \mathcal{D}_\alpha(\operatorname{Tr}A|\operatorname{Tr}(B_2-B_1)_+) + \mathcal{D}_\alpha(0|\operatorname{Tr}(B_2-B_1)_-).
\end{aligned}
$$

Specialising to $A = \rho$ and $B_i = \sigma_i$, with $\operatorname{Tr}(\sigma_2-\sigma_1)_+ = \operatorname{Tr}(\sigma_2-\sigma_1)_- =: t$, we get (51). Inequality (52) follows immediately from (51) by the symmetry of $\mathcal{D}_\alpha$ (Theorem 5). Using the averaging procedure we get the same inequalities with $\mathcal{D}_\alpha$ replaced by S_α, giving (53) and (54). Due to the symmetry under exchanging σ_1 and σ_2 we can add an absolute value sign to the left-hand side of all these inequalities. □

Remarks.

1. It can be checked that the right-hand side of inequality (53) is a concave and monotonously increasing function of t for any $0 < \alpha < 1$.

2. It is also easily verified that equality is achieved in (53) for $\rho \perp \sigma_1$ and $\sigma_2 = t\rho + (1-t)\sigma_1$.

3. Unlike in Proposition 1, this approach does not lead to corresponding inequalities for the relative entropy proper, S, as no such inequalities can exist. Indeed, no matter how small t is, one can always find states ρ, σ_1 and σ_2 such that $|S(\rho||\sigma_1) - S(\rho||\sigma_2)|$ is unbounded; take, for example, $\rho = \sigma_2$ and σ_1 such that $\operatorname{supp}\rho$ is not a subspace of $\operatorname{supp}\sigma_1$.

8 The Small Incremental Mixing Conjecture

Consider an ensemble of time-dependent states, $\mathcal{E}(t) = \{(p_j, \rho_j(t))\}_{j=1}^n$, where each state $\rho_j(t)$ evolves under the influence of a Hamiltonian H_j; that is, $\rho_j(t) = U_j(t)\rho_j U_j(t)^*$, where $U_j(t) = \exp(itH_j)$. Let $\rho_0(t)$ be the ensemble averaged state, $\rho_0(t) = \sum_{j=1}^n p_j\rho_j(t)$. We will drop the time argument to indicate the state at time 0, $\rho_j := \rho_j(0)$.

The *mixing rate* $\Lambda(\mathcal{E})$ of this ensemble is defined as

$$\Lambda(\mathcal{E}) := \frac{d}{dt}\bigg|_{t=0} S(\rho_0(t)).$$

Bravyi conjectured in [8] the following upper bound on the mixing rate for binary ensembles ($n = 2$):

$$\Lambda(\mathcal{E}) \leq c\, h_2(p)\, ||H_1 - H_2||,$$

where c is a dimension- and state-independent constant, and $h_2(p)$ is the Shannon entropy of the distribution $(p, 1-p)$. He called this the *Small Incremental Mixing* (SIM) conjecture. Lieb and Vershynina considered this conjecture in [22] and inquired whether this bound could also be valid for larger ensembles ($n > 2$); that is, whether

$$\Lambda(\mathcal{E}) \leq c\, H(\mathbf{p}),$$

where $H(\mathbf{p})$ is the Shannon entropy of the ensemble's probability vector, and all the Hamiltonians satisfy $||H_j|| \leq 1$.

Bravyi's SIM conjecture was proven very recently by Van Acoleyen *et al* [33], with a value for the constant $c = 9$. More details about the physical relevance of this conjecture (now a theorem), in particular to entanglement generating rates and entanglement area laws, can be found in [8, 22, 33].

In this Section we provide an entirely different proof, and obtain a sharper form of the inequality, with constant $c = 2$. Our approach is based on the observation that the mixing rate can be expressed in terms of S_α. Without loss of generality we can put $H_1 = 0$ and replace H_2 by H, so that $U_1(t) = \mathbb{I}$, $U_2(t) = U(t)$ and $\rho_1(t) = \rho_1$. Because the entropy of the

signal states $\rho_j(t)$ does not change under unitary evolution, we have

$$\begin{aligned}
& S(\rho_0(t)) - S(\rho_0) \\
&= \left(S(\rho_0(t)) - \sum_j p_j S(\rho_j(t)) \right) - \left(S(\rho_0) - \sum_j p_j S(\rho_j) \right) \\
&= \sum_j p_j \left(S(\rho_j(t)||\rho_0(t)) - S(\rho_j||\rho_0) \right) \\
&= -p_1 \log(p_1)(\mathrm{S}_{p_1}(\rho_1||\rho_2(t)) - \mathrm{S}_{p_1}(\rho_1||\rho_2)) \\
&\quad -p_2 \log(p_2)(\mathrm{S}_{p_2}(\rho_2(t)||\rho_1) - \mathrm{S}_{p_2}(\rho_2||\rho_1)) \\
&= -p_1 \log(p_1)(\mathrm{S}_{p_1}(\rho_1||U(t)\rho_2 U^*(t)) - \mathrm{S}_{p_1}(\rho_1||\rho_2)) \\
&\quad -p_2 \log(p_2)(\mathrm{S}_{p_2}(\rho_2||U^*(t)\rho_1 U(t)) - \mathrm{S}_{p_2}(\rho_2||\rho_1)). \qquad (55)
\end{aligned}$$

In the last line we have exploited unitary invariance of S_α.

A natural first attempt is to try inequality (53) of Theorem 11 (with $s = 0$).

$$\begin{aligned}
S(\rho_0(t)) - S(\rho_0) &\leq -\sum_j p_j \log(p_j)(\mathrm{S}_{p_j}(1|0) - \mathrm{S}_{p_j}(1|t_j) + \mathrm{S}_{p_j}(0|t_j)) \\
&\leq -\sum_j p_j \log(p_j) \frac{1-p_j}{-p_j \log(p_j)} t_j,
\end{aligned}$$

where $t_1 = T(U(t)\rho_2 U^*(t), \rho_2)$ and $t_2 = T(U^*(t)\rho_1 U(t), \rho_1)$. This requires estimating the trace norm distances t_j but it can already be seen that we will obtain a bound that is too weak, due to the occurrence of the factor $(1-p_j)/(-p_j \log(p_j))$, which can become arbitrarily large for small p_j.

The following theorem is a substantial sharpening of inequality (53) for the special case that σ_1 and σ_2 are unitarily equivalent.

Theorem 12. *For states ρ and σ, for $0 < \alpha < 1$, and $U = \exp(iH)$,*

$$\mathrm{S}_\alpha(\rho||U\sigma U^*) - \mathrm{S}_\alpha(\rho||\sigma) \leq 2||H||. \qquad (56)$$

This is the key result leading to our proof of the SIM conjecture.

The proof of this theorem relies on the following simple estimate of the trace norm distance between two unitarily equivalent states.

Lemma 8. *For a state ρ subject to a unitary evolution $U(t) = \exp(itH)$,*

$$T(U(t)\rho U^*(t), \rho) \leq t\,||H||. \qquad (57)$$

Proof. Let $\rho' = U(t)\rho U^*(t)$. For infinitesimal dt, $U = \mathbb{I} + i\,dt\,H$ and $U\rho U^* = \rho + i\,dt\,[H, \rho]$. Thus $||\rho' - \rho||_1 = dt\;||\,[H, \rho]\,||_1 \leq dt\;2||H||\;||\rho||_1$, where we used the triangle inequality for the trace norm, and Hölder's inequality. Integrating over t and using the triangle inequality once more shows that this is also true for finite t. □

Proof of Theorem 12. Rather than working with S_α, we consider $\mathcal{D}_\alpha$ because its symmetry property is essential. For all density operators ρ, σ_1 and σ_2, and $0 < \alpha < 1$, with $\tau = T(\sigma_1, \sigma_2)$, inequality (51) reads

$$\begin{aligned}
\mathcal{D}_\alpha(\rho||\sigma_1) - \mathcal{D}_\alpha(\rho||\sigma_2) &\leq \mathcal{D}_\alpha(1|0) - \mathcal{D}_\alpha(1|\tau) + \mathcal{D}_\alpha(0|\tau) \\
&= \frac{\tau}{\alpha + (1-\alpha)\tau} \leq \frac{\tau}{\alpha}.
\end{aligned}$$

In particular, for $\sigma_2 = \sigma$ and $\sigma_1 = U\sigma U^*$, with $U = \exp(iH)$,

$$\mathcal{D}_\alpha(\rho||U\sigma U^*) - \mathcal{D}_\alpha(\rho||\sigma) \leq \frac{1}{a} T(U\sigma U^*, \sigma) \leq \frac{1}{\alpha}||H||,$$

where we also have used Lemma 8.

From the symmetry property of $\mathcal{D}_\alpha$, Theorem 5, it follows that the inequality also holds when replacing α in the right-hand side by $1-\alpha$. Indeed,

$$\begin{aligned} \mathcal{D}_\alpha(\rho||U\sigma U^*) - \mathcal{D}_\alpha(\rho||\sigma) &= \mathcal{D}_{1-\alpha}(U\sigma U^*||\rho) - \mathcal{D}_{1-\alpha}(\sigma||\rho) \\ &= \mathcal{D}_{1-\alpha}(\sigma||U^*\rho U) - \mathcal{D}_{1-\alpha}(\sigma||\rho) \\ &\leq \frac{1}{1-\alpha} T(U^*\rho U, \rho) \leq \frac{1}{1-\alpha}||H||. \end{aligned}$$

Hence, combining the two inequalities yields

$$\mathcal{D}_\alpha(\rho||U\sigma U^*) - \mathcal{D}_\alpha(\rho||\sigma) \leq \min\left(\frac{1}{\alpha}, \frac{1}{1-\alpha}\right)||H|| \leq 2||H||.$$

Using the averaging procedure then yields the inequality of the theorem. □

Theorem 13 (Small Incremental Mixing)**.** *Within the setup described above,*

$$S(\rho_0(t)) - S(\rho_0) \leq 2t\ h(p_1, p_2)||H||. \tag{58}$$

Proof. To each term of (55) we apply Theorem 12 to estimate the differences between the S_α and get

$$S(\rho_0(t)) - S(\rho_0) \leq -\sum_{j=1}^{2} p_j \log(p_j)\ 2t||H|| = 2t\ h(p_1, p_2)\ ||H||.$$

□

9 Quantum Skew Divergence as a State Distinguishability Measure

The quantum relative entropy (QRE) between two quantum states ρ and σ, $S(\rho||\sigma) = \mathrm{Tr}\,\rho(\log\rho - \log\sigma)$, is a non-commutative generalisation of the Kullback-Leibler divergence (KLD) $\mathrm{KL}(p||q)$ between probability distributions p and q, and is widely used as a measure of dissimilarity of quantum states [26].

Both the KLD and the QRE exhibit a number of features that arise naturally from their underlying mathematical model and that may be useful in certain circumstances. However, these features also imply that neither the KLD nor the QRE is a proper distance measure in the mathematical sense. First of all, the KLD and QRE are asymmetric in their arguments. This alone already precludes their use as a distance measure, and prompted the terminology KL 'divergence', rather than KL 'distance'. Secondly, neither obeys the triangle inequality. A third feature, and the one considered in this paper, is that the KLD is infinite whenever for some i, the probability $q(i)$ is zero when $p(i)$ is not. Likewise, $S(\rho||\sigma)$ is infinite when the support of ρ is not contained in the support of σ. In particular, this renders the relative

entropy useless as a useful distance measure between pure states, since it is infinite for pure ρ and σ, unless ρ and σ are exactly equal (in which case it always gives 0). It is therefore unable to tell by how much two distinct pure states are dissimilar.

It is illustrative to see how this feature comes about in one of the more important operational interpretations of the KLD and QRE, namely in the context of asymmetric hypothesis testing. Let the null hypothesis H_0 be that a random variable X is drawn from the distribution p; the alternative hypothesis H_1, that it is drawn from distribution q. A test is to be designed that optimally discriminates between the two. Two types of error are relevant: a type I error (false positive) is when the test selects H_1 when in fact H_0 is true; a type II error (false negative) is when the test selects H_0 when H_1 is true. The probability of a type I error is usually denoted by α, and the probability of a type II error by β. These probabilities cannot usually both be made zero, but they can be made to both tend to 0 exponentially fast when N, the number of samples of X looked at by the test, tends to infinity. One can then define the corresponding error rates, α_R and β_R, as the limits $\alpha_R = -\lim_{N\to\infty}(1/N)\log\alpha_N$ and $\beta_R = -\lim_{N\to\infty}(1/N)\log\beta_N$. These rates quantify how fast α_N and β_N tend to 0 with N.

The KLD can be given a clear operational meaning in this context, as the best possible rate β_R when α_N (not α_R) is to be kept below a certain value ϵ (a value which, surprisingly, does not ultimately enter in the value of the optimal β_R). It is now not hard to see why the KLD should be infinite when, for some i, $q(i)$ is zero but $p(i)$ is not. In this case an optimal test should only look at outcome i. If this outcome occurs, even if only once, this immediately rules out the alternative hypothesis. The number of samples required to find outcome i amongst them (which depends on $p(i)$) is finite, therefore the rate β_R is infinite. In other words, the infinity of the KLD represents the fact that "the theory 'All crows are black' can be refuted by the single observation of a white crow".

Whereas the emergence of this feature of the KLD (and the QRE) seems quite natural, it may not always be that desirable. Firstly, the unboundedness of the KLD may be a source of numerical instability in applications. Secondly, the extreme focus on zeros of q (zero eigenvalues of σ, respectively) implies a complete disregard of other discriminating information. As stated before, the QRE can only tell distinctness of pure states, but not by how much. Thirdly, in applications where q is an *empirical* distribution, the weight put on events with $q(i) = 0$ is totally inappropriate: in empirical distributions this corresponds to unseen events, not to impossible ones. This is a serious concern in applications such as natural language processing [17], where the events are occurrences of word combinations in a large (but not infinitely large) corpus of text, and in which many genuine but rare word combinations do not occur at all; consider, for example, the total number of occurrences of the word combination "relative entropy" in the combined issues of the New York Times. Similar concerns can be raised in the quantum case, when σ is a reconstructed quantum state obtained from quantum state tomography experiments. When maximum likelihood reconstruction of nearly pure states produces reconstructed states with one or more zero eigenvalues, these zeroes should not be interpreted as zero probabilities. How to properly deal with these empirical quantum states is known in the tomography literature as the 'zero-eigenvalue problem' [7]. A final problem is of a theoretical nature: because KLD and QRE can become infinite, it is much harder (and less natural) to obtain good upper bounds on these quantities in terms of other distance measures. Invariably, some information about the smallest eigenvalues of ρ and σ have to be supplied to allow even the existence of such bounds (see, e.g. [3, 4]).

Several solutions have been put forward to overcome the problems associated with this infinity feature, in the classical case and in the quantum case, in the form of modifications of the KLD (QRE). In the classical case, one of the first to discuss several of these modifications in detail was Lin [23]. In addition to the infinity problem, he also considered the asymmetry issue. He introduced the following dissimilarity measures based on the KLD, which he called the *K-divergence* and *L-divergence*, respectively:

$$\begin{aligned} K(p||q) &= S(p||(p+q)/2) && (59)\\ L(p,q) &= K(p||q) + K(q||p) && (60)\\ &= 2H((p+q)/2) - H(p) - H(q). && (61) \end{aligned}$$

Here, $H(p)$ is the Shannon entropy of a distribution, which for the discrete case reads $H(p) = -\sum_i p(i)\log p(i)$. Lin also considered a generalisation of the L-divergence with different weights, which he called the *Jensen-Shannon divergence*:

$$\mathrm{JS}^{\alpha}(p,q) = H(\alpha p + (1-\alpha)q) - \alpha H(p) - (1-\alpha)H(q). \quad (62)$$

Lin pointed out that the K divergence is a special case of the Csiszár f-divergences with the function f given by $f(x) = x\log(2x/(1+x))$ [10].

In [17], Lee introduced a generalisation of Lin's K-divergence that incorporates different weights,

$$s_{\alpha}(p||q) = S(p||\alpha q + (1-\alpha)p), \quad (63)$$

a quantity which she called the *α-skew divergence*. In contrast to Lin's, whose motivations were mainly theoretical and driven by the lack of good upper bounds on the KL divergence, Lee's proposal was fuelled by a practical application in natural language processing: the estimation and subsequent use of probabilities of unseen word combinations [17, 18]. Here, the asymmetry of the KLD had to be maintained but its inordinate focus on zero-probabilities had to be alleviated. Lee proposed a 'smoothing' of the q distribution with p by mixing a small amount of p into q (she used $\alpha = 0.99$), in order to shift the focus to events that are seen under both distributions.

In the quantum case, the first attempt to overcome the infinity problem of the QRE was undertaken by Lendi, Farhadmotamed and van Wonderen [19], who proposed to mix both ρ and σ with the maximally mixed state. They introduced the *regularised relative entropy* as

$$R(\rho||\sigma) = c_d\ S\left(\frac{\rho+\mathbb{I}_d}{1+d}\middle|\middle|\frac{\sigma+\mathbb{I}_d}{1+d}\right),$$

where d is the dimension of state space, and c_d is a normalisation constant. It is clear that this procedure only works for finite-dimensional states. One might also consider mixing both states with a smaller amount of the maximally mixed state, for example as a quantum generalisation of Laplace's rule of succession for empirical distributions, by which 1 is added to the frequencies of all outcomes, in order to properly account for unseen events.

Another possibility, also applicable to the infinite dimensional case, is to apply a smoothing process. One can define the *smooth relative entropy* between states ρ and σ as the infimum of the ordinary relative entropy between ρ and another (unnormalised) state τ, where τ is constrained to be ϵ-close to σ in trace norm distance:

$$S_{\epsilon}(\rho||\sigma) = \inf_{\tau}\left\{S(\rho||\tau) : \tau \geq 0, \mathrm{Tr}\,\tau \leq 1, ||\tau-\sigma||_1 \leq \epsilon\right\}. \quad (64)$$

This form of smoothing has already been applied to Renyi entropies and min- and max-relative entropy [11, 29, 35], giving rise to a quantity with an operational interpretation. However, the process can equally well be applied to ordinary relative entropy.

By far the most popular modification of the QRE in the quantum case is the *quantum Jensen-Shannon divergence* (QJSD) [9, 13, 14, 24, 30], which has the additional feature of being symmetric in its arguments. It comes in several flavours: for pairs of states and equal weights, we have the 'vanilla' style:

$$\begin{aligned} \mathrm{QJS}(\rho,\sigma) &= S(\rho||\frac{1}{2}\rho+\frac{1}{2}\sigma)+S(\sigma||\frac{1}{2}\rho+\frac{1}{2}\sigma) && (65)\\ &= S((\rho+\sigma)/2)-(S(\rho)+S(\sigma))/2. && (66)\end{aligned}$$

Here $S(\rho)$ is the von Neumann entropy $S(\rho) = -\operatorname{Tr}\rho\log\rho$. The latter formula allows for a straightforward generalisation to general statistical weights, and to more than two states:

$$\mathrm{QJS}^{(\pi_1,\ldots,\pi_n)}(\rho_1,\ldots,\rho_n) = S(\sum_{i=1}^n \pi_i\rho_i) - \sum_{i=1}^n \pi_i S(\rho_i). \tag{67}$$

In the context of quantum channels, this quantity is also known as the *Holevo* χ of an ensemble $\{(\rho_i,\pi_i)\}_{i=1}^n$.

It seems that in the quantum case, Lee's α-skew divergence has not been studied before. It was highly rewarding to discover the many interesting properties of the skew divergence, not to mention the applications presented in this paper.

The skew divergence is closely related to other distinguishability measures. Firstly, it can be seen as a simplified version of smoothed relative entropy: to calculate the latter a minimisation problem over states τ has to be solved. However, there is a simple canonical choice for τ that achieves the same purpose of regularisation but without having to find the exact minimiser. Namely, we can take that τ that lies on the m-geodesic (mixing geodesic) from ρ to σ; i.e. $\tau = \alpha\rho + (1-\alpha)\sigma$. Note that it is not a good idea to choose an e-geodesic (exponential geodesic) here as this once again leads to infinities. In so doing we obtain exactly the skew divergence with $\alpha = \epsilon/||\rho-\sigma||_1$. For that reason, the skew divergence can be a useful approximation for the smoothed relative entropy. Further study will be devoted to the question how good this approximation may be.

The skew divergence is also the non-symmetric distinguishability measure underpinning the quantum Jensen-Shannon divergence. It is therefore not surprising that mathematical results for the skew divergence lead to useful mathematical results for the QJSD and the Holevo χ. This is the topic of the next and final section.

10 Inequalities for the Quantum Jensen-Shannon Divergence and Holevo Information

The Quantum Jensen-Shannon Divergence (QJS) of n states ρ_i, with weights p_i, is formally equal to the Holevo information χ, of the quantum ensemble $\mathcal{E} = \{(\rho_i,p_i)\}_{i=1}^n$, and is defined as

$$\mathrm{QJS}^{(p_1,\ldots,p_n)}(\rho_1,\ldots,\rho_n) = \chi(\mathcal{E}) = S\Big(\sum_i p_i\rho_i\Big) - \sum_i p_i S(\rho_i). \tag{68}$$

We will denote by $\mathbf{p}$ the probability vector $\mathbf{p} = (p_1, \ldots, p_n)$. Let the averaged state of the ensemble be denoted by $\rho_0 := \sum_i p_i \rho_i$. It will also be useful to define the *complementary states*

$$\overline{\rho}_i := \frac{\rho_0 - p_i \rho_i}{1 - p_i} = \frac{\sum_{j, j \neq i} p_j \rho_j}{1 - p_i}.$$

The Holevo χ can be rewritten in terms of quantum skew divergences as follows:

$$\chi(\mathcal{E}) \;=\; \sum_i p_i S(\rho_i || \rho_0) = -\sum_i p_i \log(p_i)\, \mathrm{S}_{p_i}(\rho_i || \overline{\rho}_i). \tag{69}$$

From this representation and the bounds on the skew divergence follow several bounds for χ that improve on existing bounds in the literature.

Let $t_{ij} = T(\rho_i, \rho_j) = ||\rho_i - \rho_j||_1/2$, the trace distance between signal states ρ_i and ρ_j. Also, let $t = \max_{i,j} t_{ij}$. From the bound of Theorem 9, $\mathrm{S}_\alpha(\rho||\sigma) \leq T(\rho, \sigma)$, and the convexity of T in each of its arguments, we immediately obtain

$$\begin{aligned}
\chi(\mathcal{E}) \;&=\; -\sum_i p_i \log(p_i)\, \mathrm{S}_{p_i}(\rho_i || \overline{\rho}_i) \\
&\leq\; -\sum_i p_i \log(p_i) T(\rho_i, \overline{\rho}_i) \\
&\leq\; -\sum_i p_i \log(p_i) \sum_{j \neq i} \frac{p_j}{1 - p_i} t_{ij} \qquad (70) \\
&\leq\; H(\mathbf{p})\, t. \qquad (71)
\end{aligned}$$

In the last line, $H(\mathbf{p}) := -\sum_i p_i \log(p_i)$ is the Shannon entropy of the ensemble's probability vector. Hence we have shown:

Theorem 14. *Let $\mathcal{E}$ be the ensemble $\mathcal{E} = \{(p_i, \rho_i)\}_{i=1}^n$ with corresponding probability vector $\mathbf{p} = (p_i)_{i=1}^n$. Let t be the largest of the trace distances $t_{ij} = T(\rho_i, \rho_j) = ||\rho_i - \rho_j||_1/2$. Then*

$$\chi(\mathcal{E}) \leq H(\mathbf{p})\, t.$$

This bound combines the well-known bound $\chi(\mathcal{E}) \leq H(\mathbf{p})$ (see, e.g. [28], Th. 3.7), with the bound $\chi(\mathcal{E}) \leq \log(n)\, t$ of Theorem 14 in [9] (only proven there for $n = 2$ but clearly true in general), and therefore improves on both.

For binary ensembles, Roga [30] proves the following bound on $\chi(\mathcal{E})$ in terms of the Uhlmann fidelity F between the two signal states (see also [13] for extensions to more than 2 states):

$$\chi(\mathcal{E}) \leq S(\sigma), \quad \sigma = \begin{pmatrix} p & \sqrt{p(1-p)}F \\ \sqrt{p(1-p)}F & 1-p \end{pmatrix}, \tag{72}$$

where $F = F(\rho_1, \rho_2) = \mathrm{Tr}\, \sqrt{\sqrt{\rho_1}\rho_2\sqrt{\rho_1}}$. A numerical investigation showed that this gives a bound that is sometimes lower in value than (71), which is in terms of the trace distance, and sometimes higher. However, when replacing t by its upper bound $\sqrt{1 - F^2}$ in (71), Roga's bound (72) is always better. Which bound to choose of course also depends on ease of use and generality.

Now consider two ensembles $\mathcal{E}$ and $\mathcal{E}'$ with the same probabilities p_i, but different signal states ρ_i and ρ_i', respectively. Let $t_i = ||\rho_i - \rho_i'||_1/2$ be the trace distance between corresponding signal states. We wish to obtain a bound on $|\chi(\mathcal{E}) - \chi(\mathcal{E}')|$ in terms of the t_i. A

naïve way to do so would be to use Fannes' continuity bound on the von Neumann entropy [12]. However, this would lead to a bound that is dimension dependent. Here we show how the two continuity inequalities of the skew divergence (Theorem 11) can be used to obtain a dimension-independent bound.

Define ρ_0', $\overline{\rho}_i'$ analogously as above, $t_0 = T(\rho_0, \rho_0')$ and $\bar{t}_i = T(\overline{\rho}_i, \overline{\rho}_i')$. The distances $\bar{t}_i$ can be bounded in terms of the t_j as

$$\bar{t}_i \leq \frac{\sum_{j:j\neq i} p_j t_j}{1-p_i} \leq \max_{j:j\neq i} t_j. \tag{73}$$

To simplify the formulas, we will express everything in terms of the largest t_j, which we denote by t.

First consider the difference between terms

$$\begin{aligned}
&\mathrm{S}_{p_i}(\rho_i||\overline{\rho}_i) - \mathrm{S}_{p_i}(\rho_i'||\overline{\rho}_i') \\
&\quad = \mathrm{S}_{p_i}(\rho_i||\overline{\rho}_i) - \mathrm{S}_{p_i}(\rho_i'||\overline{\rho}_i) + \mathrm{S}_{p_i}(\rho_i'||\overline{\rho}_i) - \mathrm{S}_{p_i}(\rho_i'||\overline{\rho}_i') \\
&\quad \leq \mathrm{S}_{p_i}(0|1) - \mathrm{S}_{p_i}(t_i|1) + \mathrm{S}_{p_i}(t_i|0) + \mathrm{S}_{p_i}(1|0) - \mathrm{S}_{p_i}(1|\bar{t}_i) + \mathrm{S}_{p_i}(0|\bar{t}_i) \\
&\quad \leq \mathrm{S}_{p_i}(0|1) - \mathrm{S}_{p_i}(t|1) + \mathrm{S}_{p_i}(t|0) + \mathrm{S}_{p_i}(1|0) - \mathrm{S}_{p_i}(1|t) + \mathrm{S}_{p_i}(0|t) \\
&\quad = \frac{1}{-\log(p_i)} \left(t \log \frac{p_i t + 1 - p_i}{p_i t} + \log \frac{p_i + (1-p_i)t}{p_i} \right).
\end{aligned}$$

Summing over all terms then yields

$$|\chi(\mathcal{E}) - \chi(\mathcal{E}')| \leq \sum_i p_i t \log\left(1 + \frac{1-p_i}{p_i}\,\frac{1}{t}\right) + \sum_i p_i \log\left(1 + \frac{1-p_i}{p_i} t\right). \tag{74}$$

The probabilities p_i can be eliminated by exploiting concavity of the logarithm, giving the promised dimension-independent bound:

Theorem 15. *Let $\mathcal{E}$ and $\mathcal{E}'$ be two ensembles of n quantum states with the same probabilities p_i, but with different states ρ_i and ρ_i', respectively. Let t be the largest of $t_i := T(\rho_i, \rho_i') = ||\rho_i - \rho_i'||_1/2$. Then*

$$|\chi(\mathcal{E}) - \chi(\mathcal{E}')| \leq t \log(1 + (n-1)/t) + \log(1 + (n-1)t). \tag{75}$$

For small t, this bound is approximated well by $(\log(n-1) + (n-1) - \log t)t$.

Acknowledgments

A substantial part of this work was done at the Institut Mittag-Leffler, Djursholm (Sweden), during an extended stay at its Fall 2010 Semester on Quantum Information Theory. I also acknowledge conversations with R. Werner, J. Oppenheim, B. Nachtergaele, M-B. Ruskai and M. Shirokov.

Thanks to Tobias Osborne for bringing Bravyi's problem to my attention and to Karel, Michaël and Frank for sharing their preprint [33].

This work has been supported in part by an Odysseus grant from the Flemish Fund for Scientific Research (FWO).

References

[1] A preliminary version of this work has already been presented at TQC-2011, Madrid [2], but as we were then unaware of Lee's work the quantity came with another name, namely 'telescopic relative entropy'. We now feel that 'quantum skew divergence' is a more informative name.

[2] K.M.R. Audenaert, "Telescopic Relative Entropy", in "Theory of Quantum Computation, Communication, and Cryptography – 6th Conference, TQC 2011, Madrid, Spain, May 24-26, 2011, Revised Selected Papers", Bacon, Martin-Delgado and Roetteler (eds.), Springer Lecture Notes in Computer Science **6745**, 39–52 (2014). (arXiv:1102.3040).

[3] K.M.R. Audenaert, "On the asymmetry of the relative entropy", J. Math. Phys. **54**, 073506 (2013).

[4] K.M.R. Audenaert and J. Eisert, "Continuity bounds on the quantum relative entropy - II", J. Math. Phys. **52**, 112201 (2011).

[5] K.M.R. Audenaert, M. Nussbaum, A. Szkoła and F. Verstraete, Commun. Math. Phys. **279**, 251–283 (2008).

[6] R. Bhatia, *Matrix Analysis*, Springer (1997).

[7] R. Blume-Kohout, "Optimal, reliable estimation of quantum states", New J. Phys. **12**, 043034 (2010).

[8] S. Bravyi, "Upper bounds on entangling rates of bipartite Hamiltonians", Phys. Rev. A **76**, 052319 (2007).

[9] J. Briët and P. Harremoës, "Properties of Classical and Quantum Jensen-Shannon Divergence", Phys. Rev. A **79**, 052311 (2009).

[10] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations", Studia Sci. Math. Hungar. **2**, 299–318 (1967).

[11] N. Datta, "Min- and Max-Relative Entropies and a New Entanglement Monotone," IEEE Trans. Information Theory **55**, 2816–2826 (2009).

[12] M. Fannes, "A continuity property of the entropy density for spin lattice systems", Commun. Math. Phys. **31**, 291–294 (1973).

[13] M. Fannes, F. de Melo, W. Roga and K. Życzkowski, "Matrices of fidelities for ensembles of quantum states and the Holevo quantity", Quantum Inf. Comp. **12**(5-6), 472–489 (2012).

[14] B. Fuglede and F. Topsøe, "Jensen-Shannon Divergence and Hilbert space embedding", Proceedings of the 2004 IEEE International Symposium on Information Theory, art. 31 (2004).

[15] F. Hiai and D. Petz, "The proper formula for relative entropy and its asymptotics in quantum probability", Commun. Math. Phys. **143**, 99–114 (1991).

[16] A.S. Holevo, "Capacity of a quantum communications channel", Problems of Inf. Transm. **5**(4), 247–253 (1979).

[17] L. Lee, "Measures of Distributional Similarity", Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 25–32 (1999).

[18] L. Lee, "On the Effectiveness of the Skew Divergence for Statistical Language Analysis", Artificial Intelligence and Statistics 2001, 65–72 (2001).

[19] K. Lendi, F. Farhadmotamed and A.J. van Wonderen, "Regularization of quantum relative entropy in finite dimensions and application to entropy production", J. Stat. Phys. **92**(5/6), 1115–1135 (1998).

[20] A. Lesniewski and M.B. Ruskai, "Monotone Riemannian metrics and relative entropy on noncommutative probability spaces", J. Math. Phys. **40**(11), 5702–5724 (1999).

[21] E.H. Lieb, "Convex Trace Functions and the Wigner-Yanase-Dyson Conjecture", Adv. Math. **11**, 267–288 (1973).

[22] E.H. Lieb and A. Vershynina, "Upper bounds on mixing rates", Quantum Inf. Comp. **13**, 0986–0994 (2013).

[23] J. Lin, "Divergence measures based on the Shannon entropy", IEEE Trans. Inf. Th. **IT-37**(1), 145–151 (1991).

[24] A.P. Majtey, P.W. Lamberti and D.P. Prato, "Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states", Phys. Rev. A **72**, 052310 (2005).

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

[26] M. Ohya and D. Petz, *Quantum entropy and its use*, Springer (1993).

[27] D. Petz, "Monotone metrics on matrix spaces", Linear Algebra Appl. **244**, 81–96 (1996).

[28] D. Petz, *Quantum Information Theory and Quantum Statistics*, Springer, Berlin (2008).

[29] R. Renner, "Security of quantum key distribution", PhD thesis, ETH Zurich, arXiv:quant-ph/0512258 (2005).

[30] W. Roga, M. Fannes and K. Życzkowski, "Universal bounds for the Holevo quantity, coherent information and the Jensen-Shannon divergence", Phys. Rev. Lett. **105**, 040505 (2010).

[31] R.L. Schilling, R. Song and Z. Vondraček, *Bernstein functions, theory and applications*, de Gruyter Studies in Mathematics 37, Walter de Gruyter, Berlin (2010).

[32] K. Temme, M.J. Kastoryano, M.B. Ruskai, M.M. Wolf and F. Verstraete, "The χ^2-divergence and mixing times of quantum Markov processes", J. Math. Phys. **51**(12), 122201 (2011).

[33] K. Van Acoleyen, M. Mariën and F. Verstraete, "Entanglement rates and area laws", Phys. Rev. Lett. **111**, 170501 (2013).

[34] H. Umegaki, "Conditional expectation in an operator algebra, IV (entropy and information)", Kodai Math. Sem. Rep. **14**, 59–85 (1962).

[35] A. Vitanov, F. Dupuis, M. Tomamichel and R. Renner, "Chain rules for smooth min- and max-entropies", IEEE Trans. Inf. Th. **IT-59**, 2603–2612 (2013).

[36] A. Wehrl, "General properties of entropy", Rev. Mod. Phys. **50**(2), 221–259 (1978).