A sharp continuity estimate for the von Neumann entropy

2007 J. Phys. A: Math. Theor. 40 8127

(http://iopscience.iop.org/1751-8121/40/28/S18)

J. Phys. A: Math. Theor. **40** (2007) 8127–8136 doi:10.1088/1751-8113/40/28/S18

A sharp continuity estimate for the von Neumann entropy

Koenraad M R Audenaert

Institute for Mathematical Sciences, Imperial College London, 53 Princes Gate, London SW7 2PG, UK

E-mail: k.audenaert@imperial.ac.uk

Received 23 October 2006, in final form 8 December 2006
Published 27 June 2007
Online at stacks.iop.org/JPhysA/40/8127

Abstract
We derive an inequality relating the entropy difference between two quantum states to their trace norm distance, sharpening a well-known inequality due to Fannes. In our inequality, equality can be attained for every prescribed value of the trace norm distance.

PACS number: 3.65.Hk

(Some figures in this article are in colour only in the electronic version)

1. Introduction

The initial motivation of the present paper was given by a purely pedagogical issue: given the ubiquity of powerful computers on nearly every desk, one should be able to quickly illustrate (rather than prove) the validity of many basic inequalities. In quantum mechanics, and in quantum information theory in particular, perhaps the best known inequality is the eponymous continuity inequality (4) for the von Neumann entropy, discovered by Fannes. This inequality gives an upper bound on the absolute value of the difference between the von Neumann entropies of two finite-dimensional quantum states, in terms of their trace norm distance (3).

The inequality can be easily illustrated using a computer, as it deals with finite-dimensional quantum states and each of its constituents can be calculated efficiently. One has to generate random pairs of states, calculate both the trace norm distance and the absolute value of the difference of their von Neumann entropies, and produce a scatter plot of these two quantities. Adding to that a graph of the upper bound, one should see a cloud of points lying below the latter graph. Indeed, only a few minutes of work is required to produce plots akin to those of figures 1–3.

Now one directly sees that the bound is indeed an upper bound, but also that the bound is not sharp. There are no points on the graph, or even near it. Although this is certainly not a problem for the originally intended use of the bound—proving a continuity property of

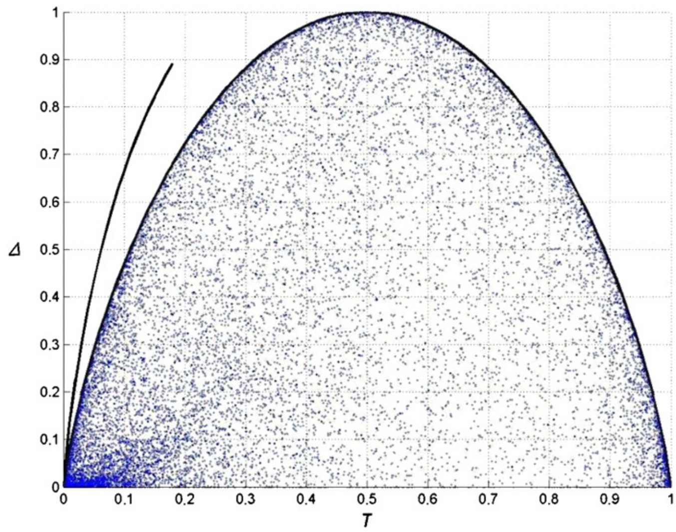

Figure 1. Scatter plot of 20 000 randomly generated pairs (ρ, σ) of qubit states $(d = 2)$; shown is the trace norm distance $T = \|\rho - \sigma\|_1/2$ versus the difference $\Delta = |S(\rho) - S(\sigma)|$ of the vN entropies. The upper curve in the interval $0 \leqslant T \leqslant 1/(2e)$ represents the Fannes bound (4). The lower curve represents our sharp bound (6) and is seen to follow the boundary of the set of scatter points tightly.

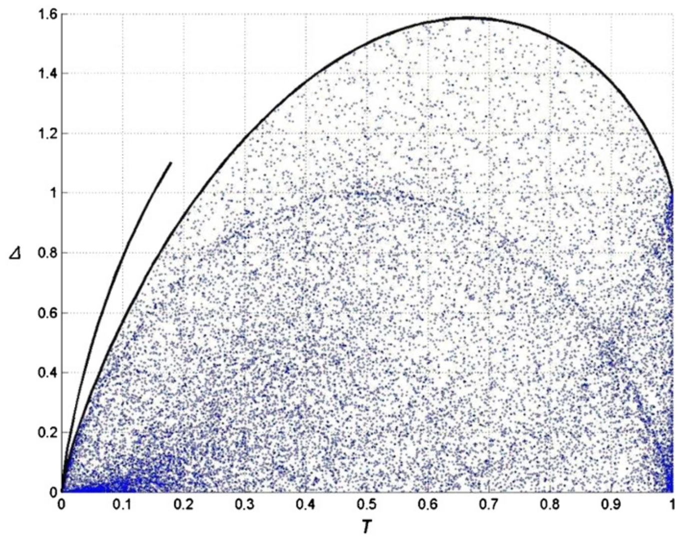

Figure 2. Same as figure 1, but for qutrits $(d = 3)$.

the von Neumann entropy—nevertheless, like the present author, one could be compelled to find a better bound; a sharp bound, that exactly describes the upper boundary of the cloud of randomly generated points.

In [5], the author, together with Eisert, did exactly this for the relative entropy, which is in a sense a quantity derived from the von Neumann entropy. In the present paper, the same is done for the von Neumann entropy itself. The present paper could therefore be considered

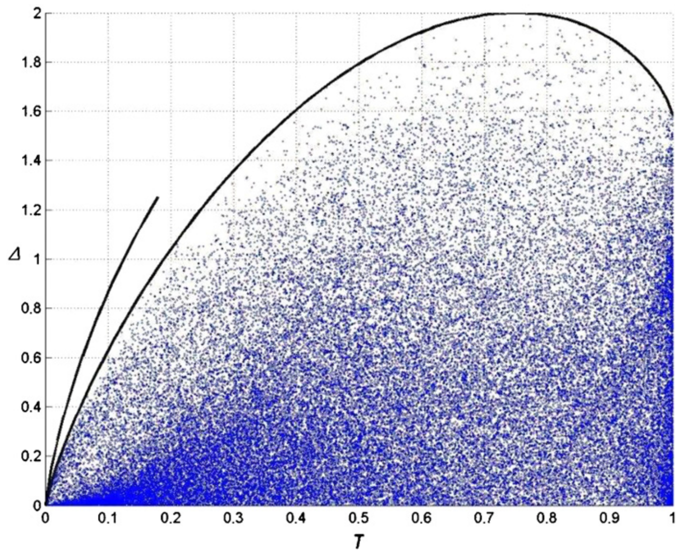

Figure 3. Same as figure 1, but for four-dimensional quantum systems ($d = 4$).

the 'prequel' of [5]. The outcome is a new, sharp bound, of the same type as Fannes' one, and, rather surprisingly, of the same complexity.

Obviously, by restricting to commuting states, one recovers a continuity bound for the Shannon entropy. This bound is known in information theory as the ℓ_1 bound on entropy (see e.g. [6], theorem 16.3.2). Moreover, the proof of the quantum version proceeds by first reducing the problem to the classical case (section 2.1 in our paper), to the extent that if one finds a sharp bound for the ℓ_1 bound this immediately yields a sharp bound for the von Neumann entropy as well. This is indeed what we do in this paper.

As mentioned, there are no real benefits in the new bound w.r.t. proving continuity of the von Neumann entropy. However, in recent times, new usage of such a bound has been found, e.g. in entanglement theory. For this modern usage our bound has the important benefit that it is actually easier to use, because it is valid over the whole range of possible values of the trace norm distance, unlike Fannes' one, which only holds for trace norm distances less than $1/e$ and has to be modified for larger ones. Furthermore, it is the sharpest bound possible and improves on the older one. The only added cost of the new bound goes in its proof, which is much longer.

Before stating the main result, let us first introduce some notations. The acronyms LHS and RHS are short for left-hand side and right-hand side. To denote Hermitian conjugate, we follow mathematical conventions and use the asterisk rather than the dagger. The notation $\mathrm{Diag}(x, y, z, \ldots)$ denotes the diagonal matrix with diagonal elements $x, y, z, \ldots$, and $\mathrm{Eig}^{\downarrow}(A)$ denotes the vector of eigenvalues of a Hermitian matrix A, sorted in non-increasing order. Following information-theoretical convention, we use base-2 logarithms, denoted by $\log_2$. The natural logarithm will be denoted by ln. The von Neumann (vN) entropy, when expressed in units of qubits, is then defined as

$$S(\rho) := -\mathrm{Tr}[\rho \log_2 \rho]. \tag{1}$$

For classical probability distributions, this reduces to the Shannon entropy

$$H(p) := -\sum_i p_i \log_2 p_i, \tag{2}$$

where p is a probability vector. We will occasionally indulge in overloaded usage of the symbol H and define $H(x) := -x\log_2 x$ for non-negative scalars x. Thus the relation $H(p) = \sum_i H(p_i)$ holds.

We use the following definition for trace norm distance:

$$T(\rho, \sigma) = \|\rho - \sigma\|_1/2, \tag{3}$$

including the factor $1/2$ to have T between 0 and 1.

The original inequality for the continuity of the vN entropy, as proven by Fannes [2, 4], reads

$$|S(\rho) - S(\sigma)| \leqslant 2T\log_2(d) - 2T\log_2(2T), \tag{4}$$

which is valid for $0 \leqslant T \leqslant 1/2e$. For larger T one can use the weaker inequality

$$|S(\rho) - S(\sigma)| \leqslant 2T\log_2(d) + 1/(e\ln(2)). \tag{5}$$

Our main result is a sharpening of this pair of inequalities. Since, according to our definition, T lies between 0 and 1, the vector $(T, 1-T)$ can be regarded as a probability vector, and its Shannon entropy $H((T, 1-T))$ formally exists.

Theorem 1. *For all d-dimensional states ρ, σ such that their trace norm distance is given by T,*

$$|S(\rho) - S(\sigma)| \leqslant T\log_2(d-1) + H((T, 1-T)). \tag{6}$$

By construction of this bound, there is no sharper bound than this one that exploits knowledge of T and d only.

To show that sharpness holds for any value of T and d, we just note that the following pair of (commuting) states achieves the bound:

$$\rho = \mathrm{Diag}(1-T, T/(d-1), \ldots, T/(d-1)) \tag{7}$$

$$\sigma = \mathrm{Diag}(1, 0, \ldots, 0). \tag{8}$$

In other notations:

$$\sigma = |0\rangle\langle 0| \tag{9}$$

$$\rho = \frac{Td}{d-1}\frac{\mathbb{1}_d}{d} + \left(1 - \frac{Td}{d-1}\right)|0\rangle\langle 0|. \tag{10}$$

Note that the coefficient of $|0\rangle\langle 0|$ in ρ may be negative. A simple calculation then yields that their trace norm distance is T, and their entropy difference is $T\log_2(d-1) + H((T, 1-T))$. We once again stress that Fannes' original bound is not sharp: there are no pairs of states saturating Fannes' bound except in the trivial case when they are identical ($T = 0$).

2. Proof

The remainder of this paper will be devoted to the proof of our inequality. Because of its complexity, we will proceed in several stages.

2.1. Reduction to classical case

The first step of the proof is to reduce the statement to the commuting (classical) case. Since S is unitarily invariant, $S(\rho)$ only depends on the eigenvalues of ρ. Let us denote the eigenvalue decompositions of ρ and σ by $\rho = V\,\mathrm{Diag}(\Lambda_\rho)V^*$ and $\sigma = W\,\mathrm{Diag}(\Lambda_\sigma)W^*$; here, $\Lambda_\rho = \mathrm{Eig}^\downarrow(\rho)$. The LHS of (6) then becomes $|H(\Lambda_\rho) - H(\Lambda_\sigma)|$, and the trace norm distance, which is the only ingredient of the RHS that depends on the states, is given by $T = \|\mathrm{Diag}(\Lambda_\rho) - U\,\mathrm{Diag}(\Lambda_\sigma)U^*\|_1/2$, where $U = V^*W$.

Let us now fix the eigenvalues of ρ and σ; the only degree of freedom is then in the unitary matrix U, which only appears in the RHS. The LHS is thus fixed, while the RHS can be varied. Referring to the figures, this amounts to looking at cross-sections of the plot along the horizontal lines. To prove correctness of the bound (6) we have to look at the points of minimal (leftmost) and maximal (rightmost) trace norm distance. Inequality IV.62 in [1], which essentially seems to be due to Mirsky [3], reads

$$|||\,\mathrm{Eig}^\downarrow(A) - \mathrm{Eig}^\downarrow(B)||| \leqslant |||A - B|||$$
$$|||A - B||| \leqslant |||\,\mathrm{Eig}^\downarrow(A) - \mathrm{Eig}^\uparrow(B)|||,$$

for all Hermitian A and B and all unitarily invariant norms. In particular, we get that the extremal values of $T = \|\mathrm{Diag}(\Lambda_\rho) - U\,\mathrm{Diag}(\Lambda_\sigma)U^*\|_1/2$, when varying U, are obtained for U equal to certain permutation matrices. More precisely, the minimal value is obtained for $U = \mathbb{1}$, and the maximal value for U the permutation matrix that totally reverses the diagonal entries.

This shows, in particular, that the boundary of the 'point cloud' can be found for diagonal ρ and σ, i.e. for commuting states.

2.2. Proof strategy

In the following we can therefore restrict to the commuting case and only look at (discrete) probability distributions and their Shannon entropies. To highlight the classical nature of the remainder of the proof, we will replace the states ρ and σ by d-dimensional probability vectors p and q. We have to show that the following inequality holds:

$$|H(p) - H(q)| \leqslant T\log_2(d-1) + H((T, 1-T)), \tag{11}$$

where T is now

$$T := (1/2)\sum_{i=1}^{d} |p_i - q_i|. \tag{12}$$

We will do this in a constructive way, by fixing T and looking for pairs p, q that maximize the LHS. The maximal value of the LHS thus obtained then will be a sharp upper bound by construction.

At this point it is interesting to mention that simple things do not work. For example, it is not obvious that $|H(p) - H(q)|$ should be maximal for p 'pure', because this quantity is neither convex nor concave, and furthermore is to be maximized over the rather complicated set of all (p, q) such that $(1/2)\sum_{i=1}^{d}|p_i - q_i| = T$, $p_i \geqslant 0$, $q_i \geqslant 0$ and $\sum_i p_i = \sum_i q_i = 1$ hold.

Let us introduce the symbol $\delta := p - q$. Since p and q are probability vectors, the δ_i are real numbers adding up to 0. We can decompose δ in a positive and negative part, which we denote by δ^+ and δ^-. Thus we have $\delta = \delta^+ - \delta^-$. Both parts consist of non-negative reals and their elementwise product $\delta_i^+\delta_i^-$ is 0. The constraint (12) then translates to $\sum_i \delta_i^+ = T$ and $\sum_i \delta_i^- = T$.

In the following, we will shift attention to the quantity $H(q) - H(p)$ (without taking absolute values) and try to find its global minimum. Subsequently taking the absolute value then yields the maximum of $|H(q) - H(p)|$.

2.3. The case $d = 2$

When $d = 2$, we automatically get that δ must be given by $\delta = (+T, -T)$. The quantity to be minimized is then

$$H(q) - H(p) = H((p_1 + T, 1 - p_1 - T)) - H((p_1, 1 - p_1)),$$

where p_1 is the first entry of p. As this quantity is obtained by setting $d = 2$ in (24) below, we need not spend more time on this special case. The reader interested in this case only is advised to proceed to the end of subsection 2.7.

2.4. Optimal δ^+

We will prove here that the optimal δ^+ is 'rank 1'; that is, it has just one non-zero entry, which then is given by T. W.l.o.g., since nothing has been claimed yet about p or q themselves, we can put this non-zero entry on the first position. Furthermore, δ^- can then take non-zero values on all positions except the first one.

Letting p_1 be the first entry of p, p and q must then be of the form

$$p = (p_1, (1 - p_1)r) \tag{13}$$

$$q = (p_1 + T, (1 - p_1)r - Ts), \tag{14}$$

where r and s are $(d - 1)$-dimensional probability vectors, with the restrictions

$$p_1 + T \leqslant 1 \tag{15}$$

$$(1 - p_1)r - Ts \geqslant 0. \tag{16}$$

Here, Ts is just δ^-. The value of $H(q) - H(p)$ corresponding to this is given by

$$H(q) - H(p) = H(p_1 + T) - H(p_1) + H((1 - p_1)r - Ts) - H((1 - p_1)r). \tag{17}$$

The remaining minimization over r, s and p_1 will be performed in the subsequent stages.

Proof. Let us now prove that the optimal δ^+ must indeed be rank 1. So we put $q = p + \delta^+ - \delta^-$ and fix p and δ^-, under the restrictions $p - \delta^- \geqslant 0$. The restrictions on δ^+ are, as mentioned before, $\delta^+ \geqslant 0$, $\delta^+\delta^- = 0$, and $\sum_i \delta_i^+ = T$. Hence, δ^+ is restricted to a convex set. If δ^- is 0 on the positions 1 to k, say, then the extremal points of this convex set are given by $Te^1, Te^2, \ldots, Te^k$. Now the optimality of one of these extremal points follows because $H(q) - H(p) = H(p + \delta^+ - \delta^-) - H(p)$ is concave in δ^+ (since H is concave), and it is well-known that concave functions reach their global minimum over a convex set in one (or more) of the extremal points of that set. □

2.5. Optimal $(1 - p_1)r - Ts$

Next, we minimize $H(q) - H(p)$ over r and s, which are general $(d - 1)$-dimensional probability vectors. By (17), we have to minimize $H((1 - p_1)r - Ts) - H((1 - p_1)r)$. The only extra condition on r and s is $(1 - p_1)r - Ts \geqslant 0$. We will show that minimality is achieved when $(1 - p_1)r - Ts$ is rank 1.

Proof. Given that r and s are probability vectors and that the condition $(1-p_1)r - Ts \geqslant 0$ is satisfied, $((1-p_1)r - Ts)/(1-p_1-T)$ is also a probability vector, which we will denote by η. Thus $(1-p_1)r - Ts = (1-p_1-T)\eta$. Conversely, for any pair of probability vectors s and η, $r' := ((1-p_1-T)\eta + Ts)/(1-p_1)$ is a probability vector satisfying $(1-p_1)r' - Ts \geqslant 0$. Therefore, we can do the substitution $(1-p_1)r - Ts = (1-p_1-T)\eta$ and forget about r altogether. Thus we are down to minimizing

$$H((1-p_1-T)\eta) - H((1-p_1-T)\eta + Ts)$$

over all probability vectors η and s.

Now note that for all $x, y \geqslant 0$, $H(x) - H(x+y)$ is concave and monotonously increasing in x. Indeed, the first derivative w.r.t. x is $\log(1+y/x) \geqslant 0$ and the second derivative is $-y/(x(x+y)) \leqslant 0$. Thus, as in the previous stage, we can conclude that $H((1-p_1-T)\eta) - H((1-p_1-T)\eta + Ts)$ is minimal for an extremal η. Since we have not yet decided on s, we will put w.l.o.g. $\eta = e^1$. □

With this optimal value for η, and putting

$$s = (s_1, (1-s_1)\phi)$$

(with ϕ a $(d-2)$-dimensional probability vector), we get

$$\begin{aligned} &H((1-p_1-T)\eta) - H((1-p_1-T)\eta + Ts) \\ &\quad = H(1-p_1-T) - H(1-p_1-T(1-s_1)) - H(T(1-s_1)\phi). \end{aligned}$$

The remaining minimization over s now consists of first minimizing over ϕ, and then over s_1.

The minimization over ϕ is easy, because it only involves the term $H(T(1-s_1)\phi)$, without any constraint other than that ϕ be a probability vector. This term achieves its maximum when $\phi = (1, 1, \ldots, 1)/(d-2)$, the uniform distribution, and the maximum value is $T(1-s_1)\log_2(d-2) + H(T(1-s_1))$.

We are now left with a minimization over s_1 of the function

$$H(1-p_1-T) - H(1-p_1-T(1-s_1)) - T(1-s_1)\log_2(d-2) - H(T(1-s_1)). \quad (18)$$

We will tackle this minimization in the next stage.

2.6. Optimal s_1

In terms of s_1, (18) is the sum of a linear term,

$$H(1-p_1-T) - T(1-s_1)\log_2(d-2),$$

and the nonlinear term

$$-H(1-p_1-T(1-s_1)) - H(T(1-s_1)).$$

This term is of the form $-H(y-x) - H(x)$, with $0 \leqslant x \leqslant y$, and is therefore convex in s_1. The only constraint on s_1 is that it be in the interval $[0, 1]$.

We therefore have find the local minimum of (18); by convexity of the function, we are guaranteed there is only one. If this minimum is inside the feasible interval $0 \leqslant s_1 \leqslant 1$, then this gives the answer; if it is outside it, then the minimum of the constrained minimization is either 0 or 1, depending on the location of the local minimum.

The derivative of (18) w.r.t. s_1 is

$$T(\log_2(d-2) + \log_2(1-p_1-T(1-s_1)) - \log_2(T(1-s_1))).$$

For $T > 0$, this is 0 when

$$(d-2)(1-p_1-T(1-s_1)) = T(1-s_1),$$

that is, when

$$T(1-s_1) = \frac{(d-2)(1-p_1)}{d-1}.$$

Recall that from the restriction $p_1 \leqslant 1-T$ follows $T \leqslant 1-p_1$. As the LHS lies between 0 and T, we have to consider two cases.

Case (i). If $0 < T < (d-2)(1-p_1)/(d-1)$, the local optimum cannot be achieved, and we have to take the nearest point, which is where $T(1-s_1) = T$, i.e. $s_1 = 0$. Then the minimum of (18) is given by

$$-T\log_2(d-2) - H(T). \tag{19}$$

Case (ii). If $(d-2)(1-p_1)/(d-1) \leqslant T \leqslant 1-p_1$, the local optimum is a feasible point, and we can put $T(1-s_1) = (d-2)(1-p_1)/(d-1)$. For the minimum of (18) this gives

$$H(1-p_1-T) - H\left(\frac{1-p_1}{d-1}\right) - \frac{(d-2)(1-p_1)}{d-1}\log_2(d-2) - H\left(\frac{(d-2)(1-p_1)}{d-1}\right). \tag{20}$$

2.7. Optimal p_1

For the final step of the procedure, we have to find the p_1 that minimizes the complete expression of the minimum of $H(q) - H(p)$ that we have found so far, under the restriction $0 \leqslant p_1 \leqslant 1-T$. We have to consider the two cases from the previous stage.

Case (i). If $T < (d-2)(1-p_1)/(d-1)$, that is, $0 \leqslant p_1 \leqslant 1-(d-1)T/(d-2)$, we need to minimize

$$H(p_1+T) - H(p_1) - T\log_2(d-2) - H(T). \tag{21}$$

This case only occurs when $T \leqslant (d-2)/(d-1)$. By a previously obtained result, the function $x \mapsto H(x+y) - H(x)$ is monotonously decreasing in x (and convex). Its minimum therefore occurs for the largest possible value of p_1, which in this case is $p_1 = 1-(d-1)T/(d-2)$. This gives as minimal value

$$H(1-T/(d-2)) - H(1-(d-1)T/(d-2)) - T\log_2(d-2) - H(T). \tag{22}$$

Case (ii). If $(d-2)(1-p_1)/(d-1) \leqslant T \leqslant 1-p_1$, that is, $1-(d-1)T/(d-2) \leqslant p_1 \leqslant 1-T$, we need to minimize

$$\begin{aligned} &H(p_1+T) - H(p_1) + H(1-p_1-T) - H\left(\frac{1-p_1}{d-1}\right) \\ &\qquad - \frac{(d-2)(1-p_1)}{d-1}\log_2(d-2) - H\left(\frac{(d-2)(1-p_1)}{d-1}\right). \end{aligned} \tag{23}$$

The derivative of (23) w.r.t. p_1 equals the logarithm of

$$\frac{(d-1)p_1(1-p_1-T)}{(1-p_1)(p_1+T)}.$$

This expression obviously decreases with T, and for the minimal allowed value $T = (d-2)(1-p_1)/(d-1)$ it is given by

$$\frac{(d-2)(1-p_1)}{d-2+p_1},$$

which is easily seen to be below 1; its logarithm is therefore negative. Consequentially, the derivative of (23) is negative over the range under consideration. We conclude that (23) is minimal for the maximal allowed p_1, which is $p_1 = 1 - T$.

This gives as minimal value for $H(q) - H(p)$,

$$H(1) - H(1-T) + H(0) - H\left(\frac{T}{d-1}\right) - \frac{(d-2)T}{d-1}\log_2(d-2) - H\left(\frac{(d-2)T}{d-1}\right),$$

which simplifies to

$$-(T\log_2(d-1) + H(T) + H(1-T)). \tag{24}$$

The final step is now to take the minimum of the two cases (22) and (24), the former one only being valid for $T \leqslant (d-2)/(d-1)$. From the fact that $H(x+y) - H(x)$ is monotonously decreasing in x one deduces the relation $H(1-a) + H(1-b) \geqslant H(1-a-b)$, for $0 \leqslant a, b$ and $a + b \leqslant 1$. The terms $H(1 - T/(d-2)) - H(1 - (d-1)T/(d-2))$ in (22) are therefore larger than the term $H(1-T)$ in (24). Furthermore, $-T\log_2(d-2)$ is larger than $-T\log_2(d-1)$. Hence, (24) is always smaller than (22).

Taking absolute values and noting that (24) is always negative then finally yields inequality (11).

Acknowledgments

The author is grateful for the hospitality of the Max Planck Institute for Quantum Optics (in Garching bei München) where this work was initiated. This work is supported by The Leverhulme Trust (grant F/07 058/U), by EU Integrated Project QAP, and by the Institute of Mathematical Sciences, Imperial College London, and is part of the QIP-IRC (www.qipirc.org) supported by EPSRC (GR/S82176/0). Humble thanks to A Retzker for proofreading. Thanks also to Norbert Schuch, Frank Verstraete, Michael Wolf and Tobias Osborne for stimulating discussions and for sharing enjoyment in the beauty of inequality (6), and to D Petz for informing us about Csiszár's proof.

Appendix. Sharp continuity for Tsallis and Renyi entropies

After the first draft of the paper appeared, the author was informed by Dénes Petz that the classical version of the sharp bound (i.e. its information-theoretic counterpart) was already known to Csiszár, who derived it from Fano's inequality using an ingenious trick. However, the proof we presented here is algebraic in nature rather than information-theoretic and is therefore more general. For example, it can be adapted to Tsallis entropies in a straightforward way.

The Tsallis entropies are given by $\mathrm{Tr}\,\rho^\alpha$, where α can be any real number. The proof of (6) goes through line by line (suitably modifying the expressions) for the Tsallis entropies provided $0 < \alpha < 1$. One therefore finds that for these entropies the extremal cases are again

$$\begin{aligned}\hat\rho &= \mathrm{Diag}(1-T, T/(d-1), \ldots, T/(d-1))\\ \hat\sigma &= \mathrm{Diag}(1, 0, \ldots, 0),\end{aligned} \tag{A.1}$$

giving as upper bound on the Tsallis entropy difference, for $0 < \alpha < 1$:

$$|\mathrm{Tr}\,\rho^\alpha - \mathrm{Tr}\,\sigma^\alpha| \leqslant (1-T)^\alpha - 1 + (d-1)^{1-\alpha}T^\alpha. \tag{A.2}$$

This bound also seems to be valid for $\alpha > 1$, but we have not proven this yet. For negative values of α the Tsallis entropy difference is of course unbounded.

In the case $0 < \alpha < 1$, a short argument allows us to conclude that the $\hat{\rho}$ and $\hat{\sigma}$ of (A.1) are also extreme cases for the Rényi entropies, which are given by $\log(\mathrm{Tr}\,\rho^\alpha)/(1-\alpha)$. Then we get

$$|\log \mathrm{Tr}\,\rho^\alpha - \log \mathrm{Tr}\,\sigma^\alpha| \leqslant \log[(1-T)^\alpha + (d-1)^{1-\alpha}T^\alpha]. \tag{A.3}$$

For $\alpha > 1$ this is no longer the case, due to the fact that for these α the Rényi entropy is neither convex nor concave, and the upper bound in this case is a more complicated expression, which we have not been able to find yet.

The argument goes as follows. Let y, z be real scalars; then $y \geqslant z \geqslant 1$ implies $y - z + 1 \geqslant y/z$. Indeed, from $y \geqslant z \geqslant 1$ follows $y(z-1) \geqslant z(z-1)$, or $yz - z^2 + z \geqslant y$, so that upon division by $z > 0$, we get $y - z + 1 \geqslant y/z$.

Now consider a third scalar $x \geqslant 1$, and still assume $y \geqslant z \geqslant 1$. Then we see that $x - 1 \geqslant y - z$ implies $x \geqslant y/z$, or upon taking the logarithm, $\log x \geqslant \log y - \log z$.

In particular, take $x = \mathrm{Tr}\,\hat{\sigma}^\alpha$, $y = \mathrm{Tr}\,\sigma^\alpha$ and $z = \mathrm{Tr}\,\rho^\alpha$, for ρ and σ such that $y \geqslant z$. These quantities are all $\geqslant 1$ since $0 < \alpha < 1$. Since $\mathrm{Tr}\,\hat{\rho}^\alpha = 1$, we find that

$$\mathrm{Tr}\,\sigma^\alpha - \mathrm{Tr}\,\rho^\alpha \leqslant \mathrm{Tr}\,\hat{\sigma}^\alpha - \mathrm{Tr}\,\hat{\rho}^\alpha$$

implies

$$\log \mathrm{Tr}\,\sigma^\alpha - \log \mathrm{Tr}\,\rho^\alpha \leqslant \log \mathrm{Tr}\,\hat{\sigma}^\alpha - \log \mathrm{Tr}\,\hat{\rho}^\alpha,$$

which is just the statement that optimality of $\hat{\rho}$ and $\hat{\sigma}$ in the Tsallis entropy case directly implies their optimality in the Rényi entropy case.

References

[1] Bhatia R 1997 *Matrix Analysis* (Heidelberg: Springer)
[2] Fannes M 1973 *Commun. Math. Phys.* **31** 291
[3] Mirsky L 1960 *Q. J. Math. Oxford Ser.* (2) **11** 50–9
[4] Nielsen M A and Chuang I L 2000 *Quantum Computation and Quantum Information* (Cambridge: Cambridge University Press)
[5] Audenaert K M R and Eisert J 2005 *J. Math. Phys.* **46** 102104
[6] Cover T M and Thomas J A 1991 *Elements of Information Theory* (New York: Wiley)