The Strong Convexity of von Neumann's Entropy

Yao-Liang Yu
Department of Computing Science
University of Alberta, Canada
yaoliang@cs.ualberta.ca

June 27, 2013

Abstract

The purpose of this note is to give an (almost) self-contained proof of the strong convexity of von Neumann's entropy.

1 Preliminary

Let us begin with some definitions. Fix two normed spaces $(\mathcal{X}, \|\cdot\|)$ and $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$. We use $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ to denote the space of all continuous linear operators from $\mathcal{X}$ to $\mathcal{Y}$, equipped with the usual operator norm. We also use the abbreviation $\mathcal{X}' := \mathcal{L}(\mathcal{X}, \mathbb{R})$, and the dual pairing $\langle x; f\rangle := f(x)$ for $x \in \mathcal{X}, f \in \mathcal{X}'$.

Definition 1 (Fréchet Derivative) *Let $\mathcal{O}$ be an open set of $\mathcal{X}$. The Fréchet derivative of $f : \mathcal{O} \to \mathcal{Y}$ at $x \in \mathcal{O}$ is the (continuous)* linear *operator $g \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ such that*

$$\|f(x + w) - f(x) - g(w)\|_{\mathcal{Y}} = o(\|w\|). \tag{1}$$

If f is differentiable in each point of $\mathcal{O}$, the derivatives collectively induce $f' : \mathcal{O} \to \mathcal{L}(\mathcal{X}, \mathcal{Y})$. When $\mathcal{Y} = \mathbb{R}$, we have $f' : \mathcal{O} \to \mathcal{X}'$ and consequently the more familiar notation $[f'(x)](w) = \langle w; f'(x)\rangle$. One can also iterate the definition to define even higher order derivatives. Essentially we need only the first order.

Importantly, we observe that by definition Fréchet derivative is a topological property. Therefore, if two norms $\|\cdot\|_1$ and $|||\cdot|||_1$ are equivalent on $\mathcal{X}$, and two norms $\|\cdot\|_2$ and $|||\cdot|||_2$ are equivalent on $\mathcal{Y}$, then we obtain the same Fréchet derivative of f under both $(X, \|\cdot\|_1) \to (Y, \|\cdot\|_2)$ and $(X, |||\cdot|||_1) \to (Y, |||\cdot|||_2)$. Moreover, the induced norms on $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ are also equivalent. Iterating the argument we see that derivatives of all orders are the same under both $(X, \|\cdot\|_1) \to (Y, \|\cdot\|_2)$ and $(X, |||\cdot|||_1) \to (Y, |||\cdot|||_2)$. If the norm we are interested in is equivalent to a Hilbertian norm, then we can calculate the derivatives under the Hilbertian setting, which is much more convenient.

Definition 2 (Strong Convexity) *The function $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is strongly convex iff there exists $\sigma > 0$ such that for all $x, w \in \mathcal{X}, \lambda \in]0, 1[$, we have*

$$\lambda f(x) + (1 - \lambda)f(w) \geq f(\lambda x + (1 - \lambda)w) + \frac{\sigma}{2}\lambda(1 - \lambda)\|x - w\|^2. \tag{2}$$

When the above inequality holds with $\sigma = 0$, we call f convex.

It is not hard to see that there exists the largest possible σ so that (2) is true (assuming $f \not\equiv \infty$). This is usually referred to as the modulus of convexity and denoted as $\sigma(f)$. It is easy to verify that

$$\forall \alpha > 0,\ \sigma(\alpha f) = \alpha\sigma(f), \qquad \sigma(f + g) \geq \sigma(f) + \sigma(g), \tag{3}$$

i.e., $\sigma(\cdot)$ is a positive homogeneous concave function. Moreover,

$$\sigma(\sup_\alpha f_\alpha) \geq \inf_\alpha \sigma(f_\alpha), \tag{4}$$

thus pointwise supremum of σ-strongly convex functions is σ-strongly convex. It is clear that by definition strong convexity is a topological property[1], although the moduli of convexity is not.

Interestingly, there is a simple relation between strong convexity and convexity in the Hilbertian setting:

Proposition 1 *Suppose the norm $\|\cdot\|$ on $\mathcal{X}$ is Hilbertian, then $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is σ-strongly convex iff $f - \frac{\sigma}{2}\|\cdot\|^2$ is convex.*

Proof: Simply verify Definition 2 with the aid of the parallelogram law:

$$\|x + w\|^2 + \|x - w\|^2 = 2(\|x\|^2 + \|w\|^2). \tag{5}$$

■

Proposition 1 no longer holds when we equip $\mathcal{X}$ with a non-Hilbertian norm; in fact, a strong converse is true, but we need a characterization of inner product spaces first.

Definition 3 *The modulus of convexity of a normed space is defined for every $\epsilon \in [0, 2]$ as:*

$$\delta(\epsilon) = \inf_{\|x\|=\|w\|=1, \|x-w\|=\epsilon} 1 - \|\tfrac{x+w}{2}\|. \tag{6}$$

Clearly, $\delta(0) = 0$, and $\delta(2) = 1$. Intuitively, δ characterizes how round the unit ball of $\|\cdot\|$ is.

Proposition 2 *$\frac{1}{2}\|\cdot\|^2$ is 1-strongly convex (*w.r.t. *the norm $\|\cdot\|$) iff $\|\cdot\|$ is Hilbertian.*

Proof: The if part follows immediately from Proposition 1. The only if part follows from Day [1947], or Chelidze [2009]. ■

Example 1 *We now show that Proposition 1 does not extend to any other norm. Firstly, for any abstract norm $\|\cdot\|$, clearly $\frac{\sigma}{2}\|\cdot\|^2 - \frac{\sigma}{2}\|\cdot\|^2 \equiv 0$ is convex, but according to Proposition 2, $\frac{\sigma}{2}\|\cdot\|^2$ is σ-strongly convex iff $\|\cdot\|$ is Hilbertian.*

Conversely, consider an abstract norm $\|\cdot\|$ on $\mathbb{R}^2$, with its unit ball C. Let B be the unique ellipsoid of maximum volume in C, i.e., *John's ellipsoid. Note that B touches C at four points or more. Let $\|\cdot\|_2$ be the Hilbertian norm on $\mathbb{R}^2$ whose unit ball is B. Since $B \subseteq C$, $\|x\|_2 \geq \|x\|$ for all $x \in \mathbb{R}^2$, with equality at say $w \neq \pm z$. According to Theorem 3 below we know $f = \frac{1}{2}\|\cdot\|_2^2$ is 1-strongly convex* w.r.t. *the norm $\|\cdot\|$. However, we show that $\frac{1}{2}\|\cdot\|_2^2 - \frac{1}{2}\|\cdot\|^2$ is convex iff $\|\cdot\| = \|\cdot\|_2$. Indeed, from the definition of convexity:*

$$\|\tfrac{x+y}{2}\|_2^2 - \|\tfrac{x+y}{2}\|^2 \leq \tfrac{1}{2}[\|x\|_2^2 + \|y\|_2^2 - \|x\|^2 - \|y\|^2] \tag{7}$$

$$\iff 2[\|x\|^2 + \|y\|^2] \leq \|x - y\|_2^2 + \|x + y\|^2. \tag{8}$$

Plug in $x = w, y = z$ we have

$$2[\|w\|_2^2 + \|z\|_2^2] = 2[\|w\|^2 + \|z\|^2] \leq \|w - z\|_2^2 + \|w + z\|^2,$$

implying $\|w + z\| \geq \|w + z\|_2$, but by construction $\|w + z\| \leq \|w + z\|_2$, hence $\|w + z\| = \|w + z\|_2$, i.e., *B and C agree on the middle ray of w and z. Iterating the argument we know B and C agree on a dense subset of rays. By continuity we must have $B = C$,* i.e., *$\|\cdot\| = \|\cdot\|_2$. The result immediately extends to any space $\mathcal{X}$ with dimension bigger than 2, since we can always restrict to a two-dimensional subspace.*

Example 2 *Not all norms whose square is strongly convex. Take, for instance, the ℓ_1 norm:*

$$\frac{\frac{1}{2}\|e_1\|_1^2 + \frac{1}{2}\|e_2\|_1^2}{2} = \frac{1}{2} = \frac{1}{2}\|\frac{e_1 + e_2}{2}\|_1^2.$$

Next we present some rules for checking strong convexity.

[1]Contrarily, convexity (*i.e.* $\sigma = 0$) itself does not even require a topology! From this one is hinted that strong convexity of f is not only about f, it also reveals something about the topology on $\mathcal{X}$.

Theorem 1 (Zero Order Rule) *The convex function* $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ *is* σ*-strongly convex on* core(dom f) *iff for all* $x, w \in$ core(dom f)*,*

$$f(x) \geq f(w) + f'(w; x - w) + \frac{\sigma}{2} \|x - w\|^2. \tag{9}$$

Proof: Note first that on core(dom f), the directional derivative is finite and sublinear. By (2), we have

$$f(x) - f(w) \geq \frac{f(w + \lambda(x - w)) - f(w)}{\lambda} + \frac{\sigma}{2}(1 - \lambda)\|x - w\|^2.$$

Letting $\lambda \downarrow 0$, we obtain (9).

Conversely, let $z = \lambda x + (1 - \lambda)w$, (9) yields

$$\begin{aligned} f(x) &\geq f(z) + (1 - \lambda)f'(z; x - w) + \frac{\sigma}{2}(1 - \lambda)^2 \|x - w\|^2 \\ f(w) &\geq f(z) + \lambda f'(z; w - x) + \frac{\sigma}{2}\lambda^2 \|x - w\|^2. \end{aligned}$$

Taking convex combination, we have

$$\lambda f(x) + (1 - \lambda)f(w) \geq f(z) + \lambda(1 - \lambda)(f'(z; x - w) + f'(z; w - x)) + \frac{\sigma}{2}\lambda(1 - \lambda)\|x - w\|^2,$$

which leads to σ-strong convexity due to sublinearity of the directional derivative. ■

When dom f is open and f is continuous, $f'(x; d) = \max\{\langle d; x^* \rangle : x^* \in \partial f(x)\}$. Therefore we can replace the directional derivative with any subdifferential.

Theorem 2 (First Order Rule) *Let* $\mathcal{O}$ *be an open convex set of* $\mathcal{X}$*. The Gateâux differentiable function* $f : \mathcal{O} \to \mathbb{R}$ *is* σ*-strongly convex iff for all* $x, w \in \mathcal{O}$*,*

$$\langle x - w; f'(x) - f'(w) \rangle \geq \sigma \|x - w\|^2. \tag{10}$$

Note that we have used the fact that f is real-valued so that $f' \in \mathcal{X}'$ (hence the notation $\langle w; f'(x) \rangle$).

Proof: By (9), we have

$$\begin{aligned} f(x) &\geq f(w) + \langle x - w; f'(w) \rangle + \frac{\sigma}{2}\|x - w\|^2, \\ f(w) &\geq f(x) - \langle x - w; f'(x) \rangle + \frac{\sigma}{2}\|x - w\|^2. \end{aligned}$$

Adding them we obtain (10).

Conversely, define $h(\lambda) := \frac{\sigma}{2}\lambda^2 \|x - w\|^2$ and $g(\lambda) := f(w + \lambda(x - w)) - \lambda \langle x - w; f'(w) \rangle$. Then by (10) $g'(\lambda) \geq h'(\lambda)$. By the mean value theorem, we have $g(1) - g(0) \geq h(1) - h(0)$, which is (9). ■

Theorem 3 (Second Order Rule) *Let* $\mathcal{O}$ *be an open convex set of* $\mathcal{X}$*. The twice Fréchet differentiable function* $f : \mathcal{O} \to \mathbb{R}$ *is* σ*-strongly convex iff* $\forall x \in \mathcal{O}, w \in \mathcal{X}$

$$\langle w; [f''(x)](w) \rangle \geq \sigma \|w\|^2, \tag{11}$$

or equivalently

$$\inf_{x \in \mathcal{O}, w \in \mathcal{X} : \|w\| = 1} \langle w; [f''(x)](w) \rangle \geq \sigma. \tag{12}$$

Proof: Due to homogeneity of (11), we can assume w.l.o.g. that $x + w \in \mathcal{O}$. By (10) we have

$$\langle w; f'(x + w) - f'(x) \rangle \geq \sigma \|w\|^2.$$

Since f is twice Fréchet differentiable, $\langle w; f'(x + w) - f'(x) \rangle = \langle w; [f''(x)]w + o(\|w\|) \rangle$. Letting $\|w\| \to 0$ yields (11).

Conversely, let $h(\lambda) := \sigma\lambda\|x-w\|^2$ and $g(\lambda) = \langle x-w; f'(w+\lambda(x-w))\rangle$. Then by (11) $g'(\lambda) \geq h'(\lambda)$. By the mean value theorem, we have $h(1)-h(0) \leq g(1)-g(0)$, which is (10). ■

Functions defined on the vector space $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$, consisting of all continuous linear operators from the Hilbert space $\mathcal{H}_1$ to another Hilbert space $\mathcal{H}_2$, deserve some special attention. Equip $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ with a norm $\|\cdot\|$ (not necessarily the operator norm, which is Hilbertian in this case), and consider the function $f : \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2) \to \mathbb{R} \cup \{\infty\}$.

Definition 4 (Unitary Invariance) *$f : \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2) \to \mathbb{R} \cup \{\infty\}$ is called unitarily invariant iff for all $L \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$, isometries $U : \mathcal{H}_2 \to \mathcal{H}_2$ and $V : \mathcal{H}_1 \to \mathcal{H}_1$, we have $f(L) = f(ULV)$. In other words, f only depends on the singular values of its input.*

If we equip $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ with the operator norm $\|\cdot\|_{\mathsf{op}}$, then it becomes very easy to check strong convexity, thanks to the next theorem:

Theorem 4 ([]) *content...*

2 Pinsker's Inequality

Consider the normed space $\mathcal{X} = \mathbb{R}^n, \|\cdot\| = \ell_1^n$ and denote the (open) unit ball (restricted to the positive orthant) $\mathcal{B}_+^n := \{x \in \mathbb{R}_{++}^n : \|x\|_1 < 1\}$. Recall that the unnormalized (negative) entropy is defined as

$$\forall x \in \mathbb{R}_+^n, \quad h(x) := \sum_i x_i \log x_i - x_i, \tag{13}$$

under the convention[2] $0 \log 0 := 0$. For illustration purpose, let us first prove

Theorem 5 *The unnormalized entropy $h : \mathcal{B}_+^n \to \mathbb{R}$ has moduli of convexity 1 with respect to the ℓ_1^n norm.*

Proof: Not surprisingly, we are going to apply Theorem 3. As noted before, the Fréchet derivative is a topological property, so we can calculate it by changing the norm $\|\cdot\|$ to ℓ_2^n. After some *usual* elementary differentiation we arrive at

$$\inf_{x\in\mathcal{B}_+^n, y\in\mathbb{R}^n:\|y\|_1=1} \sum_{i=1}^n \frac{y_i^2}{x_i} \geq \inf_{x\in\mathcal{B}_+^n, y\in\mathbb{R}^n:\|y\|_1=1} \sum_{i=1}^n \frac{y_i^2}{x_i} \cdot \sum_{i=1}^n x_i \geq \inf_{x\in\mathcal{B}_+^n, y\in\mathbb{R}^n:\|y\|_1=1} \left(\sum_{i=1}^n |y_i|\right)^2 = 1. \tag{14}$$

Clearly the lower bound cannot be improved. ■

Let us denote $\bar{\mathcal{B}}_+^n := \{x \in \mathbb{R}_+^n : \|x\|_1 \leq 1\}$. Under the usual convention $0 \log 0 := 0$, we have

Corollary 1 *The unnormalized entropy $h : \bar{\mathcal{B}}_+^n \to \mathbb{R}$ has moduli of convexity 1 with respect to the ℓ_1^n norm.*

Proof: Apply Theorem 5, with a limiting argument, to (2). ■

Corollary 2 (Pinsker's Inequality) *For all $x, y \in \mathcal{P}_1^n := \{x \in \mathbb{R}_+^n : \|x\|_1 = 1\}$, the Kullback-Leibler divergence*

$$d(x\|y) := \sum_i x_i \log x_i - x_i \log y_i \geq \tfrac{1}{2}\|x-y\|_1^2. \tag{15}$$

Proof: Clearly strong convexity is preserved under restricting the (effective) domain. Apply Corollary 1 and Theorem 1. ■

Note that Pinsker's inequality, mostly known in information theory, usually comes with an extra factor $\frac{1}{\ln 2}$, which is simply due to scaling because information theory prefers $\log_2$ while we use $\log_e$.

[2] Clearly h is continuous under this convention.

3 Matrix Pinsker's Inequality

Now let us move on to the matrix case. Let $\mathcal{X} = \mathbb{S}^n$, the space of all Hermitian $n \times n$ matrices, and equip the trace norm $\|\cdot\| = \|\cdot\|_{\text{tr}}$ (*i.e.* the sum of the singular values). Denote $\mathbb{S}^n_+ \subseteq \mathbb{S}^n$ as the cone of all $n \times n$ positive semidefinite matrices and $\mathbb{S}^n_{++}$ its interior. Clearly for $X \in \mathbb{S}^n_+, \|X\|_{\text{tr}} = \text{tr}(X)$.

Define similarly as before $\mathcal{B}^n_+ := \{X \in \mathbb{S}^n_{++} : \text{tr}(X) < 1\}, \bar{\mathcal{B}}^n_+ := \{X \in \mathbb{S}^n_+ : \text{tr}(X) \leq 1\}$ and $\mathcal{P}^n_{\text{tr}} := \{X \in \mathbb{S}^n_+ : \text{tr}(X) = 1\}$. The following "generalization"[3] of the unnormalized (negative) entropy is due to von Neumann [1927]:

$$\forall X \in \mathbb{S}^n_+, \quad H(X) := \text{tr}(X \log X - X), \tag{16}$$

where the convention $0 \log 0 := 0$ is again adopted. Our main goal is to prove

Theorem 6 *The unnormalized entropy $H : \mathcal{B}^n_+ \to \mathbb{R}$ has moduli of convexity 1 under the trace norm.*

Before delving into the proof, let us lay down the immediate consequences (as we saw previously):

Corollary 3 *The unnormalized entropy $H : \bar{\mathcal{B}}^n_+ \to \mathbb{R}$ has moduli of convexity 1 under the trace norm.*

Corollary 4 *For all $X, Y \in \mathcal{P}^n_{\text{tr}}$, $D(X\|Y) := \text{tr}(X \log X - X \log Y) \geq \frac{1}{2}\|X - Y\|^2_{\text{tr}}$.*

Not surprisingly, we are going to apply again Theorem 3 to prove Theorem 6. For this, we need to compute the derivatives.

As a gentle start, we claim that $H'(X) = \langle \cdot ; \log X \rangle$, which may or may not be trivial depending on the reader's background. We will not present the proof but mention only that H is a (real-valued) spectral function (*i.e.* depends only on the spectrum of its input), whose differential rules are well-understood, see for example [Borwein and Lewis, 2006, page 105].

Next, we need to differentiate $H'(X)$. By changing the trace norm to the Frobenius norm (so that $(\mathbb{S}^n)' \cong \mathbb{S}^n$) we may identify the (first order) Fréchet derivative H' with the function $\log : \mathbb{S}^n_{++} \to \mathbb{S}^n$. We need the following lemma to proceed:

Lemma 1 *Let $f \in C^1(I)$ where I is an open interval of $\mathbb{R}$. For $X \in \mathbb{S}^n$ whose spectrum lies in I, we have*

$$\forall W \in \mathbb{S}^n, \quad [f'(X)](W) = U \left[f^{[1]}(\Lambda) \circ (U^\top W U) \right] U^\top, \tag{17}$$

where U is some unitary matrix that diagonalizes X, i.e. $X = U \Lambda U^\top$ *with $\Lambda_{ii} = \lambda_i \in \mathbb{R}$; $f^{[1]}(\Lambda) \in \mathbb{S}^n$ with its ij-th element being*

$$\begin{cases} \frac{f(\lambda_i) - f(\lambda_j)}{\lambda_i - \lambda_j}, & \text{if } \lambda_i \neq \lambda_j \\ f'(\lambda_i), & \text{otherwise} \end{cases} ; \tag{18}$$

and $\circ$ denotes the Hadamard (elementwise) product.

This lemma, however complicated it might appear at a first glance, is actually quite natural. Its proof is also (in some sense) straightforward: Start with polynomials and then extend by continuity to all C^1 functions[4]. For those who are determined, consult, say [Bhatia, 1997, page 124] for a complete proof.

The rest is computation (what is not?). Apply Lemma 1:

$$\langle W; [f'(X)](W) \rangle = \text{tr}\left(U \left[\log^{[1]}(\Lambda) \circ (U^\top W U) \right] U^\top W \right) = \text{tr}\left(\left[\log^{[1]}(\Lambda) \circ (U^\top W U) \right] U^\top W U \right).$$

Recall that in the vector case (corresponding to $U^\top W U$ diagonal), we calculated the infimum by applying simply the Cauchy-Schwarz inequality. The matrix case seems genuinely harder, but not much if we have the following integral representation:

$$\forall \lambda \in \mathcal{B}^n_+, \quad \frac{\log \lambda_i - \log \lambda_j}{\lambda_i - \lambda_j} = \int_0^\infty \frac{1}{(t + \lambda_i)(t + \lambda_j)} \text{d}t, \tag{19}$$

[3]The reason to put a quotation mark can be understood by checking the year when von Neumann published his result!

[4]We have deliberately not mentioned what do we mean by the logarithm of a matrix so that if one has no difficulty in arriving here, s/he should recognize this standard proof technique in functional calculus.

where the left hand side is interpreted as $\frac{1}{\lambda_i}$ in case $\lambda_i = \lambda_j$. This last interpretation should always be kept in mind from now on (so that we do not have to cumbersomely split the sum). Therefore

$$\begin{aligned}\langle W; [f'(X)](W)\rangle &= \sum_{ij} \frac{\log\lambda_i - \log\lambda_j}{\lambda_i - \lambda_j}(U^\top W U)_{ij}^2 \\ &\geq \int_0^\infty \sum_{ij} \frac{(U^\top W U)_{ij}^2}{(t+\lambda_i)(t+\lambda_j)}\mathrm{d}t \\ &= \int_0^\infty \operatorname{tr}\left(\frac{1}{t+X}W\frac{1}{t+X}W\right)\mathrm{d}t.\end{aligned}$$

where we note that all terms involved are nongegative (so that there is no worry about integrability).

Fix t and W. Consider the function $g(X) := \operatorname{tr}\left(\frac{1}{t+X}W\frac{1}{t+X}W\right) = \mathsf{vec}(W)(\frac{1}{t+X}\otimes\frac{1}{t+X})\mathsf{vec}(W)$. We make two important observations: 1) g is convex on $\mathbb{S}^n_{++}$, which follows from the (operator) convexity of the map $A \mapsto A^{-1}\otimes A^{-1}$ on $\mathbb{S}^n_{++}$ [5]; 2) $g(V^\top X V) = g(X)$ for any *unitary* V that commutes with W.

We can now continue our computation. Diagonalize $W = V\Delta V^\top$ and define the **random** *diagonal* matrix $\mathbf{\Gamma}$ whose diagonal entries are sampled independently from $\{1, -1\}$ with equal odds. Note that $V\mathbf{\Gamma}V^\top$ commutes with W hence $g(V\mathbf{\Gamma}V^\top X V\mathbf{\Gamma}V^\top) = g(X)$. Therefore due to the convexity of g,

$$\mathsf{E}\left(g(V\mathbf{\Gamma}V^\top X V\mathbf{\Gamma}V^\top)\right) \geq g(\mathsf{E}(V\mathbf{\Gamma}V^\top X V\mathbf{\Gamma}V^\top)) = g(V\operatorname{diag}(V^\top X V)V^\top),$$

where $\operatorname{diag} : \mathbb{S}^n \to \mathbb{S}^n$ is the diagonal operator, *i.e.* zeroing out all off-diagonal entries. Therefore

$$\begin{aligned}\inf_{X\in\mathcal{B}^n_+, W\in\mathbb{S}^n:\|W\|_{\mathrm{tr}}=1}\langle W; [f'(X)](W)\rangle &= \inf_{X\in\mathcal{B}^n_+, W\in\mathbb{S}^n:\|W\|_{\mathrm{tr}}=1}\operatorname{tr}\left(\left[\log^{[1]}(\Lambda)\circ(U^\top W U)\right]U^\top W U\right) \\ &\geq \inf_{X\in\mathcal{B}^n_+, W\in\mathbb{S}^n:\|W\|_{\mathrm{tr}}=1}\int_0^\infty \operatorname{tr}\left(\frac{1}{t+X}W\frac{1}{t+X}W\right)\mathrm{d}t \\ &\geq \inf_{X\in\mathcal{B}^n_+, W\in\mathbb{S}^n:\|W\|_{\mathrm{tr}}=1}\int_0^\infty \operatorname{tr}\left(V\frac{1}{t+\operatorname{diag}(V^\top X V)}V^\top W V\frac{1}{t+\operatorname{diag}(V^\top X V)}V^\top W\right)\mathrm{d}t \\ &= \inf_{X\in\mathcal{B}^n_+, W\in\mathbb{S}^n:\|W\|_{\mathrm{tr}}=1}\int_0^\infty \sum_{i=1}^n \frac{\Delta_{ii}^2}{(t+(V^\top X V)_{ii})^2}\mathrm{d}t \\ &= \inf_{X\in\mathcal{B}^n_+, W\in\mathbb{S}^n:\|W\|_{\mathrm{tr}}=1}\sum_{i=1}^n \frac{\Delta_{ii}^2}{(V^\top X V)_{ii}} \\ &\geq \inf_{X\in\mathcal{B}^n_+, W\in\mathbb{S}^n:\|W\|_{\mathrm{tr}}=1}\frac{1}{\sum_{i=1}^n (V^\top X V)_{ii}} \\ &= \inf_{X\in\mathcal{B}^n_+, W\in\mathbb{S}^n:\|W\|_{\mathrm{tr}}=1}\frac{1}{\operatorname{tr}(V^\top X V)} \\ &= 1,\end{aligned}$$

where the last inequality is due to Cauchy-Schwarz (recall the diagonalization $W = V\Delta V^\top$). Clearly the lower bound cannot be improved (since it cannot even be improved in the vector case). We remark that our proof is inspired by [Ball et al., 1994], whose main result will be reproduced in the next section.

We mention another extremely "short" proof of Theorem 6, first appearing in [Hiai et al., 1981]. The proof is based on the following neat lemma:

Lemma 2 (Umegaki [1962], Lindblad [1975]) *Let $\Phi : \mathbb{S}^n \to \mathbb{S}^m$ be any completely positive trace-preserving linear operator.*

$$\forall X, Y \in \mathbb{S}^n_+, \quad D(\Phi(X)\|\Phi(Y)) \leq D(X\|Y). \tag{20}$$

[5] We did not find an appropriate reference for this fact so we supply a proof: For $A, B \in \mathbb{S}^n_{++}$, define their arithmetic mean $A\%B := \frac{A+B}{2}$, geometric mean $A\#B := A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{1/2}$ and harmonic mean $A?B := \left(\frac{A^{-1}+B^{-1}}{2}\right)^{-1}$. The arithmetic-geometric-harmonic mean inequality can be again verified by diagonalizing A and then B. Therefore $(A?B)\otimes(A?B) \preceq (A\#B)\otimes(A\#B) = (A\otimes A)\#(B\otimes B) \preceq (A\otimes A)\%(B\otimes B)$, whence follows the convexity of $A \mapsto A^{-1}\otimes A^{-1}$. We have used the fact that $A \succeq B \succeq 0 \implies (A\otimes A) \succeq (B\otimes B)$ since $(A+B)\otimes(A-B) + (A-B)\otimes(A+B) \succeq 0$.

We observe that it is enough to prove Lemma 2 for density matrices (*i.e.* those with unit trace), due to a simple scaling argument (and the assumption that Φ is trace-preserving). The rest of the proof amounts to decomposing the completely positive (whatever that means) linear map Φ into the average of elementary ones and then applying Jensen's inequality since the quantum divergence D is known to be jointly convex (which itself is a highly nontrivial fact!).

Equipped with Lemma 2, we can prove Theorem 6 with ease. Take some *unitary* matrix U that diagonalizes $X-Y$, *i.e.* $X-Y = U\Lambda U^\top$ where Λ is diagonal and *nonnegative*. Consider $\tilde{X} := \mathrm{diag}(U^\top XU), \tilde{Y} := \mathrm{diag}(U^\top YU)$ where the operator diag simply zeros out all off-diagonal entries. Importantly, $\mathrm{diag} : \mathbb{S}^n \to \mathbb{S}^n$ is completely positive and trace-preserving. Therefore

$$D(X\|Y) = D(U^\top XU\|U^\top YU) \geq D(\tilde{X}\|\tilde{Y}) \geq \tfrac{1}{2}\|\tilde{X} - \tilde{Y}\|_1^2 = \tfrac{1}{2}\|X - Y\|_{\mathrm{tr}}^2,$$

where the first inequality follows from Lemma 2 and the second is due to Theorem 5.

Let us remark that the first proof we presented for Theorem 6 is a generic procedure while the second proof seems only customized for the von Neumann entropy. This point is best illustrated with our next example.

4 $\frac{1}{2}\|\cdot\|_p^2$ is $(p-1)$-strongly convex w.r.t $\|\cdot\|_p$

We will apply Theorem 3. Take the derivative of $\frac{1}{2}\|\cdot\|_p^2$ (where w.l.o.g. assume $\mathbf{x} > 0$):

$$\partial_i = x_i^{p-1}(x_1^p + \cdots + x_d^p)^{2/p-1} \tag{21}$$

$$\partial_{ij} = (2-p)x_i^{p-1}x_j^{p-1}(x_1^p + \cdots + x_d^p)^{2/p-2} + \mathbf{1}_{i=j}\cdot(p-1)x_i^{p-2}(x_1^p + \cdots + x_d^p)^{2/p-1} \tag{22}$$

Thus for $\|\mathbf{y}\|_p = 1$:

$$\left\langle \mathbf{y}\cdot\partial^2(\tfrac{1}{2}\|\mathbf{x}\|_p^2), \mathbf{y}\right\rangle = (2-p)\left(\sum_i x_i^p\right)^{\frac{2}{p}-2}\left(\sum_i x_i^{p-1}y_i\right)^2 + (p-1)\left(\sum_i x_i^p\right)^{\frac{2}{p}-1}\sum_i x_i^{p-2}y_i^2. \tag{23}$$

Using Hölder's inequality, and noting that $\frac{p}{2} + \frac{2-p}{2} = 1$:

$$\left(\sum_i x_i^p\right)^{\frac{2}{p}-1}\sum_i x_i^{p-2}y_i^2 \geq \left(\sum_i (x_i^{p-2}y_i^2)^{p/2}(x_i^p)^{(2-p)/2}\right)^{2/p} = \left(\sum_i y_i^p\right)^{2/p} = 1. \tag{24}$$

Plugging back to (23) we obtain

$$\left\langle \mathbf{y}\cdot\partial^2(\tfrac{1}{2}\|\mathbf{x}\|_p^2), \mathbf{y}\right\rangle \geq (2-p)\left(\sum_i x_i^p\right)^{\frac{2}{p}-2}\left(\sum_i x_i^{p-1}y_i\right)^2 + p - 1 \geq p-1. \tag{25}$$

The bound is tight, since by choosing the signs (and magnitudes) of $\mathbf{y}$ properly we can drive $\sum_i x_i^{p-1}y_i$ to 0.

References

Keith Ball, Eric A. Carlen, and Elliott H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones Mathematicae*, 115:463–482, 1994.

Rajendra Bhatia. *Matrix Analysis*. Springer, 1997.

Jonathan M. Borwein and Adrian S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2nd edition, 2006.

George Z. Chelidze. On Nordlanders conjecture in the three-dimensional case. *Arkiv för Matematik*, 47(2): 267–272, 2009.

Mahlon M. Day. Some characterizations of inner-product spaces. *Transactions of the American Mathematical Society*, 62(2):320–337, 1947.

Fumio Hiai, Masanori Ohya, and Makoto Tsukada. Sufficiency, KMS condition and relative entropy in von Neumann algebras. *Pacific Journal of Mathematics*, 96(1):147–151, 1981.

Göran Lindblad. Completely positive maps and entropy inequalities. *Commun. Math. Phys.*, 40:147–151, 1975.

Hisaharu Umegaki. Conditional expectation in an operator algebra, iv. *Kodai Math. Sem. Rep.*, 14(2):59–85, 1962.

John von Neumann. Thermodynamik quantenmechanischer gesamtheiten. *Göttingen Nachr.*, pages 273–291, 1927.