STRONG CONVEXITY OF SANDWICHED ENTROPIES AND RELATED OPTIMIZATION PROBLEMS

RAJENDRA BHATIA, TANVI JAIN, AND YONGDO LIM

Abstract. We present several theorems on strict and strong convexity, and higher order differential formulae for sandwiched quasi-relative entropy (a parametrised version of the classical fidelity). These are crucial for establishing global linear convergence of the gradient projection algorithm for optimisation problems for these functions. The case of the classical fidelity is of special interest for the multimarginal optimal transport problem (the n -coupling problem) for Gaussian measures.

1. Introduction

Let $\mathbb{P}$ be the space of $n \times n$ complex positive definite matrices. An element A of $\mathbb{P}$ with $\operatorname{tr} A = 1$ is called a *density matrix* or a *state*. Many of the statements in this paper are of special interest for density matrices though we do not make that restriction. The *fidelity* between two elements A and B of $\mathbb{P}$ is defined by

$$F(A,B) = \operatorname{tr}\left(A^{\frac{1}{2}} B A^{\frac{1}{2}}\right)^{\frac{1}{2}}. \tag{1}$$

Fidelity plays an important role in quantum information theory and quantum computation, and it has deep connections with quantum entanglement, quantum chaos, and quantum phase transitions. See [34, 35]. Although fidelity by itself is not a metric, it has played a role as a measure of the closeness of two states. It occurs also in another context. There is a metric on $\mathbb{P}$ defined as

$$d(A,B) = \left[\frac{\operatorname{tr}(A+B)}{2} - \operatorname{tr}\left(A^{\frac{1}{2}} B A^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]^{\frac{1}{2}} \tag{2}$$

which is called the *Bures distance* in the literature on quantum information and the *Wasserstein metric* in statistics and the theory of optimal transport. See [15, 18, 22, 25, 30].

The multimarginal optimal transport problem (alternatively, the coupling problem) involves solving the minimization problem: given $A_1, \dots, A_m$ in $\mathbb{P}$ and weights $w_1, \dots, w_m$, find

$$\min_{X \in \mathbb{P}} \sum_{j=1}^{m} w_j d^2(X, A_j). \tag{3}$$

Date: August 17, 2018.

2010 *Mathematics Subject Classification.* 15A90, 15B48, 47A63, 49A51, 49Q20, 49D45, 81P16, 81P45, 94A17.

Key words and phrases. Positive definite matrix, multimarginal optimal transport, fidelity, sandwiched quasi-relative entropy, strong convexity, gradient projection algorithm.

This minimization problem coincides with the least squares problem of Gaussian measures for the Wasserstein distance between probability measures with finite second moment on $\mathbb{R}^n$. See [1, 20, 22, 25, 30, 32]. The concavity and strict concavity of the function

$$f(X) = \operatorname{tr}\left(A^{\frac{1}{2}} X A^{\frac{1}{2}}\right)^{\frac{1}{2}} \tag{4}$$

on $\mathbb{P}$ play a very crucial role in the proofs of existence and uniqueness of the solution to (3). See [12].

In some recent works a parameterized version of fidelity defined as

$$F_t(A,B) = \operatorname{tr}\left(A^{\frac{1-t}{2t}} B A^{\frac{1-t}{2t}}\right)^t, \qquad t \in (0,\infty) \tag{5}$$

has been studied. See [19, 37]. The usual fidelity (1) is the special case $t = 1/2$. In [37] $F_t(A,B)$ is called the *sandwiched quasi-relative entropy*. Using this the *sandwiched Rényi relative entropy* is defined as

$$D_t(B \parallel A) = \frac{1}{t-1} \log F_t(A,B), \quad t \in (0,\infty) \setminus \{1\}. \tag{6}$$

This is a variant of the traditional *relative Rényi entropy* defined as

$$D'_t(B \parallel A) = \frac{1}{t-1} \log \operatorname{tr}\left(A^{1-t} B^t\right). \tag{7}$$

Among other things, it is known [28] that

$$\lim_{t\to\infty} D_t(B \parallel A) = \| \log A^{-\frac{1}{2}} B A^{-\frac{1}{2}} \| \tag{8}$$

and

$$\lim_{t\to 1} D_t(B \parallel A) = \frac{1}{\operatorname{tr} B} \operatorname{tr}\left[B(\log B - \log A)\right], \tag{9}$$

where $|| \cdot ||$ is the operator norm

$$\|A\| = \sup_{\|x\|=1} \|Ax\|,$$

which for a positive semidefinite matrix A is equal to $\lambda_1(A)$, the largest eigenvalue of A. It turns out [4] that the expression in (8) coincides with

$$d_T(A,B) := \max\{\log \lambda_1(AB^{-1}), \log \lambda_1(BA^{-1})\}$$

and is closely related to the max-relative entropy $D_{\max}(A\|B) := \log \lambda_1(AB^{-1})$ in the context of quantum information theory [17]. We note that d_T is known as the *Thompson metric* on $\mathbb{P}$ and is a complete metric invariant under inversion and congruence transformations [33, 29], and the expression in (9) is the *relative entropy*, first introduced by Umegaki.

The entity (6) was introduced by Müller-Lennert et al in [28] and by Wilde et al in [37]. Several of its properties were established in these papers and some others conjectured. Since then these have been established in various papers. In particular, we draw attention to the paper [19] by Frank and Lieb. In [37] Wilde, Winter and Yang have employed $D_t(A \parallel B)$ to prove theorems on the capacity of entanglement-breaking channels. Differentiability, monotonicity and convexity properties of F_t and D_t are a major theme in all these papers.

In this paper we study some related, though slightly different, convexity problems. Let $f:\mathbb{P}\to\mathbb{R}$ be a smooth function. Let $\nabla f(X)$ and $\nabla^2 f(X)$ denote the gradient and the Hessian of f. See [13] for gradient and Hessian of scalar valued functions. Suppose f is strictly convex. The *Bregman distance* associated with f is the function $\mathcal{D}_f:\mathbb{P}\times\mathbb{P}\to\mathbb{R}$ defined as

$$\mathcal{D}_f(Y,X)=f(Y)-f(X)-\langle\nabla f(X),Y-X\rangle, \tag{10}$$

where $\langle X,Y\rangle=\operatorname{tr}(XY)$ on $\mathbb{H}$, the space of $n\times n$ complex Hermitian matrices. The convexity of f ensures that $\mathcal{D}_f(Y,X)\geq 0$, and strict convexity ensures that it is zero if and only if $X=Y$. Let $\mathbb{K}$ be a compact convex subset of $\mathbb{P}$. We say that f is k*-strongly convex* on $\mathbb{K}$ (with $k>0$) if for all $X,Y\in\mathbb{K}$

$$\mathcal{D}_f(Y,X)\geq\frac{k}{2}\|X-Y\|_2^2. \tag{11}$$

Here $\|A\|_2=(\operatorname{tr}A^*A)^{\frac{1}{2}}$ is the Hilbert-Schmidt norm. The condition (11) says

$$f(Y)\geq f(X)+\langle\nabla f(X),Y-X\rangle+\frac{k}{2}\|X-Y\|_2^2. \tag{12}$$

So f is k-strongly convex on $\mathbb{K}$ if and only if

$$\nabla^2 f(X)\geq kI, \tag{13}$$

for all $X\in\mathbb{K}$. On the other hand, we say that f is k*-smooth* on $\mathbb{K}$ if ∇f is k-Lipschitz; i.e.,

$$\|\nabla f(X)-\nabla f(Y)\|_2\leq k\|X-Y\|_2, \tag{14}$$

for all $X,Y\in\mathbb{K}$. This condition is equivalent to

$$\nabla^2 f(X)\leq k\ I, \tag{15}$$

for all $X\in\mathbb{K}$.

The two constants k in (13) and (15) play a fundamental role in the design and convergence analysis of optimisation algorithms. We refer the reader to Chapter 9 of the standard text [13]. Here it is also pointed out that these constants " are known only in rare cases". The main new result in this paper is the following.

Theorem 1. *Let $f:\mathbb{P}\to\mathbb{R}_+$ be the function*

$$f(X)=\operatorname{tr}\left(A^{\frac{1-t}{2t}}XA^{\frac{1-t}{2t}}\right)^t \tag{16}$$

where $A\in\mathbb{P}$ and $0<t<1$. Let $\mathbb{K}$ be a compact convex subset of $\mathbb{P}$. Let α,β be positive numbers such that $\alpha I\leq Y\leq\beta I$ for all $Y\in\mathbb{K}\cup\{A\}$. Then for all $X\in\mathbb{K}$

$$t(1-t)\alpha^{1-t}\beta^{t-2}\leq-\nabla^2 f(X)\leq t(1-t)\beta^{1-t}\alpha^{t-2}. \tag{17}$$

In other words, the function $-f$ is k_1-strongly convex and k_2-smooth on $\mathbb{K}$ with k_1,k_2 given by the two extreme sides of (17). The condition number of an operator A is defined as

$$\operatorname{cond}(A)=\|A\|\|A^{-1}\|.$$

As a corollary to Theorem 1 we have:

Corollary 2. *Let* f *be the function defined in* (16). *Then for all* $X \in \mathbb{K}$

$$\operatorname{cond}\left(\nabla^2 f(X)\right) \leq \left(\frac{\beta}{\alpha}\right)^{3-2t}.$$

Now suppose $A_j, 1 \leq j \leq m$ are positive definite matrices, and let $\alpha I \leq A_j \leq \beta I$ for all j. It is known [1] that the minimization problem (3) has a unique solution X and $\alpha I \leq X \leq \beta I$. The objective function in (3) is

$$\varphi(X) = \sum_{j=1}^{m} w_j \left[\frac{\operatorname{tr}(A_j + X)}{2} - \operatorname{tr}\left(A_j{}^{\frac{1}{2}} X A_j{}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right].$$

The first term in the square brackets above is linear in X, and its second derivative is zero. Our theorem shows that

$$\frac{1}{4}\frac{\alpha^{1/2}}{\beta^{3/2}} \leq \nabla^2\varphi(X) \leq \frac{1}{4}\frac{\beta^{1/2}}{\alpha^{3/2}}.$$

The condition number of $\nabla^2\varphi(X)$ is bounded by $\left(\frac{\beta}{\alpha}\right)^2$. We generalize this result into the setting of sandwiched quasi-relative entropy $F_t(A, B), 0 < t < 1$. Let

$$\varphi_t(X) = \sum_{j=1}^{m} w_j \left[\operatorname{tr}((1-t)A_j + tX) - \operatorname{tr}\left(A_j{}^{\frac{1-t}{2t}} X A_j{}^{\frac{1-t}{2t}}\right)^t\right].$$

Corollary 3. *The function* $\varphi_t : \mathbb{P} \to \mathbb{R}_+$ *is strictly convex and has a unique minimizer. Moreover, it is* $t(1-t)\beta^{1-t}\alpha^{t-2}$ *-smooth and* $t(1-t)\beta^{t-2}\alpha^{1-t}$ *-strongly convex.*

Theorem 1 is about second order derivatives of the fidelity function. The classical fidelity case is t = 1/2, and the results are new even for that case. Our methods lead to several interesting observations for the first and higher order derivatives as well. These are of independent interest and are given in Section 2 of the paper. Section 3 includes a proof of Theorem 1. A proof of Corollary 3 and the standard gradient projection method where this can be put to use are obtained in Section 4.

2. Derivative Computations

Let f be a smooth map from $\mathbb{P}$ into the positive half-line $\mathbb{R}_+ = [0, \infty)$. We denote by $Df(X)$ the (Fréchet) derivative of f at X, and by $\nabla f(X)$ the gradient of f at X. $Df(X)$ is a linear map from the space $\mathbb{H}$ of $n \times n$ Hermitian matrices into $\mathbb{R}$, and its action is given by

$$Df(X)(Y) = \frac{d}{dt}\bigg|_{t=0} f(X + tY).$$

$\nabla f(X)$ is an element of $\mathbb{H}$ and is related to $Df(X)$ by the equation

$$Df(X)(Y) = \langle \nabla f(X), Y\rangle = \operatorname{tr}(\nabla f(X)Y).$$

Of interest here are special kinds of functions. Let f be a smooth map from $\mathbb{R}_+$ into itself and let f also denote the map this induces from $\mathbb{P}$

into itself. Let $\hat{f}(A) = \mathrm{tr} f(A)$. As expected, convexity properties of f are inherited by $\hat{f}$. In some situations it may be useful to consider functions other than the trace. Let Φ be a symmetric gauge function on $\mathbb{R}^n$, i.e., a norm on $\mathbb{R}^n$ which is invariant under sign changes and permutations of the components, and let $\|\cdot\|_\Phi$ be the corresponding unitarily invariant norm on the space $\mathbb{M}(n)$ of $n \times n$ matrices. See Chapter IV of [8]. If $s(A) = (s_1(A), \ldots, s_n(A))$ is the n-tuple of singular values of A, then

$$\|A\|_\Phi = \Phi(s(A)) = \Phi(s_1(A), \ldots, s_n(A)).$$

Every symmetric gauge function is *monotone*; i.e., if x and y are two vectors with $0 \leq x \leq y$ for all j, then $\Phi(x) \leq \Phi(y)$. We say that Φ is *strictly monotone* if $\Phi(x) < \Phi(y)$ whenever $0 \leq x_j \leq y_j$ for all j and $x_j < y_j$ for at least one j. For example, the symmetric gauge functions $\Phi(x) = \left(\sum_{j=1}^n |x_j|^p\right)^{1/p}$ are strictly monotone for $1 \leq p < \infty$.

Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be two n-tuples of nonnegative numbers. Let $x_1^\downarrow \geq x_2^\downarrow \geq \ldots \geq x_n^\downarrow$ be the decreasing rearrangement of $x_1, \ldots, x_n$. If for all $1 \leq k \leq n$

$$\sum_{j=1}^k x_j^\downarrow \leq \sum_{j=1}^k y_j^\downarrow,$$

we say that x is *weakly majorised* by y. If, in addition to (2) we also have

$$\sum_{j=1}^n x_j^\downarrow = \sum_{j=1}^n y_j^\downarrow,$$

we say x is *majorised* by y, and write this as $x \prec y$. See Chapter II of [8] for facts on majorization need here.

Lemma 4. *Let x, y be two vectors with nonnegative coordinates that are not permutations of each other. Suppose $x \prec y$. Then for every strictly convex function f on nonnegative reals and every strictly monotone symmetric gauge function Φ, we have*

$$\Phi(f(x_1), \ldots, f(x_n)) < \Phi(f(y_1), \ldots, f(y_n)).$$

Proof. If $x \prec y$, then x can be expressed as a convex combination

$$x = \sum a_\sigma y_\sigma,$$

where σ varies over all permutations on n symbols, and y_σ denotes the vector $(y_{\sigma(1)}, \ldots, y_{\sigma(n)})$. If x and y are not permutations of each other, there are at least two distinct terms in this convex combination. Since f is convex,

$$f(x_j) \leq \sum a_\sigma f(y_{\sigma(j)})$$

for all j. If f is strictly convex, then this inequality is strict for some j. The statement of the lemma then follows from the properties of Φ. ■

Theorem 5. *Let f be a function from $\mathbb{R}_+$ into itself, and let $\|\cdot\|_\Phi$ be a unitarily invariant norm on $\mathbb{M}(n)$. Let $\hat{f}_\Phi$ be the map from $\mathbb{P}$ into $\mathbb{R}_+$ defined by*

$$\hat{f}_\Phi(A) = \|f(A)\|_\Phi.$$

If f is convex, then so is $\hat{f}_\Phi$. Further, if f is strictly convex and Φ is strictly monotone, then $\hat{f}_\Phi$ is strictly convex.

Proof. Let $A, B \in \mathbb{P}$, and let $C = (1/2)(A+B)$. Let $\{\lambda_j(C)\}$ denote the decreasingly ordered eigenvalues of C, and let $\{u_j\}$ be the corresponding orthonormal set of eigenvectors. Then

$$\begin{aligned} \|f(C)\|_\Phi &= \Phi\left(\lambda_1(f(C)), \ldots, \lambda_n(f(C))\right) \\ &= \Phi\left(f(\lambda_1(C)), \ldots, f(\lambda_n(C))\right) \\ &= \Phi\left(f(\langle u_1, Cu_1\rangle), \ldots, f(\langle u_n, Cu_n\rangle)\right). \end{aligned}$$

Since f is convex,

$$\begin{aligned} f\left(\langle u_j, Cu_j\rangle\right) &= f\left(\frac{\langle u_j, Au_j\rangle + \langle u_j, Bu_j\rangle}{2}\right) \\ &\le \frac{1}{2}\left[f(\langle u_j, Au_j\rangle) + f(\langle u_j, Bu_j\rangle)\right]. \end{aligned} \tag{18}$$

Every symmetric gauge function is monotone and convex. So, the relations above give

$$\begin{aligned} \|f(C)\|_\Phi &\le \frac{1}{2}\Phi\left(f(\langle u_1, Au_1\rangle), \ldots, f(\langle u_n, Au_n\rangle)\right) \\ &\quad + \frac{1}{2}\Phi\left(f(\langle u_1, Bu_1\rangle), \ldots, f(\langle u_n, Bu_n\rangle)\right). \end{aligned} \tag{19}$$

Since f is convex, by Problem IX. 8. 14 in [8] we see that

$$f\left(\langle u_j, Au_j\rangle\right) \le \langle u_j, f(A)u_j\rangle.$$

By the Schur majorisation theorem (Exercise II. 1.2 in [8]) the n-tuple $\{\langle u_j, f(A)u_j\rangle\}$ is majorised by the eigenvalue n-tuple $\{\lambda_j(f(A))\}$. Every symmetric gauge function is monotone with respect to majorisation (" isotone" in the terminology used on page 41 of [8]). Combining these observations we see that

$$\begin{aligned} \Phi\left(f(\langle u_1, Au_1\rangle), \ldots, f(\langle u_n, Au_n\rangle)\right) &\le \Phi\left(\lambda_1(f(A)), \ldots, \lambda_n(f(A))\right) \\ &= \|f(A)\|_\Phi. \end{aligned}$$

The same argument applies to B in place of A. Hence

$$\|f(C)\|_\Phi \le \frac{1}{2}\|f(A)\|_\Phi + \frac{1}{2}\|f(B)\|_\Phi. \tag{20}$$

This shows that $\hat{f}_\Phi$ is convex if f is convex. Now, suppose f is strictly convex and Φ is strictly monotone. Let $A \ne B$. There are two possibilities: (i) There exists a j such that $\langle u_j, Au_j\rangle \ne \langle u_j, Bu_j\rangle$. Then for this j, the inequality (18) is strict and hence the inequality (19) is also strict. The argument above then shows the inequality (20) is strict. (ii) If $\langle u_j, Au_j\rangle = \langle u_j, Bu_j\rangle$ for all j, then in the orthonormal basis $\{u_1, \ldots, u_n\}$, C is diagonal, and the diagonals of A and B are equal. This means that $\operatorname{diag}(C) = \operatorname{diag}(A) = \operatorname{diag}(B)$.

Since $A \neq B$, neither A nor B is diagonal. By Schur's majorization theorem (See (II. 14) of [8])

$$\operatorname{diag}(A) \prec \lambda(A),$$

where $\lambda(A)$ is the vector whose components are the eigenvalues of A. Since A is not diagonal, $\operatorname{diag}(A)$ is not a permutation of $\lambda(A)$ (because $||A||_2 = ||\lambda(A)||_2$). It follows from Lemma 4 that

$$\|f(\operatorname{diag}(A))\|_\Phi < \|f(A)\|_\Phi.$$

The same argument applies to B. Since $C = \operatorname{diag}(A)$, this shows the inequality (20) is strict. This proves the last statement of the theorem. ■

The sum of singular values is a strictly monotone unitarily invariant norm. So, the function

$$\hat{f}(A) = \operatorname{tr} f(A)$$

is (strictly) convex if f is (strictly) convex. In addition, using the linearity of the trace function we can see that $\hat{f}(A)$ is (strictly) concave if f is (strictly) concave. This is a well-known fact. See [16].

Corollary 6. *The function $f(X) = \operatorname{tr} X^t$ on positive definite matrices is strictly concave if $0 < t < 1$ and strictly convex if $1 < t < \infty$, or if $t < 0$.*

Lemma 7. *Let f be a smooth function on $\mathbb{R}_+$ and let $\hat{f}$ be the function on $\mathbb{P}$ defined as $\hat{f}(X) = \operatorname{tr} f(X)$. Then for all $X \in \mathbb{P}$ and $Y \in \mathbb{H}$,*

$$D\hat{f}(X)(Y) = \operatorname{tr}\left(f'(X)Y\right).$$

Proof. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of X and let $L_f(X)$ be the Loewner matrix

$$L_f(X) = \left[\frac{f(\lambda_i) - f(\lambda_j)}{\lambda_i - \lambda_j}\right].$$

The difference quotient in this expression is the ij th entry of $L_f(X)$, and it is understood that this is equal to $f'(\lambda_i)$ if $\lambda_i = \lambda_j$. By the Daleckii-Krein formula (Theorem V. 3.3 in [8]) the derivative $Df(X)$ is given by

$$Df(X)(Y) = L_f(X) \circ Y,$$

where $\circ$ stands for the Hadamard product (entrywise product) of two matrices taken in an orthonormal basis in which X is diagonal. Combining this with the linear functional tr, we get

$$D\hat{f}(X)(Y) = \operatorname{tr}\left(L_f(X) \circ Y\right) = \operatorname{tr}\left(f'(X)Y\right).$$

■

To state the next proposition we need the notion of the weighted geometric mean of two positive definite matrices. This is defined as

$$A \#_t B = A^{\frac{1}{2}}(A^{-\frac{1}{2}}BA^{-\frac{1}{2}})^t A^{\frac{1}{2}}, \quad 0 \leq t \leq 1. \tag{21}$$

This is a smooth curve joining A and B, and is a geodesic with respect to the Riemannian distance

$$\delta(A, B) = \|\log A^{-\frac{1}{2}}BA^{-\frac{1}{2}}\|_2,$$

on $\mathbb{P}$. See Chapter 6 of [9]. The right hand side of (21) is meaningful for all $t \in \mathbb{R}$, and we continue to use the notation $A\#_t B$ for it.

Proposition 8. *Let A be any element of $\mathbb{P}$ and let $t \in \mathbb{R}$. Let $h : \mathbb{P} \to \mathbb{R}_+$ be the map $h(X) = \operatorname{tr}\left(A^{\frac{1}{2}} X A^{\frac{1}{2}}\right)^t$. Then*

$$Dh(X)(Y) = t \operatorname{tr}\left(A\#_{1-t}X^{-1}\right) Y; \tag{22}$$

i.e.,

$$\nabla h(X) = t\left(A\#_{1-t}X^{-1}\right). \tag{23}$$

Proof. Let $k(X) = \operatorname{tr}X^t$. Then by Lemma 7, $Dk(X)(Y) = t\operatorname{tr}X^{t-1}Y$. By the chain rule

$$\begin{aligned} Dh(X)(Y) &= Dk(A^{\frac{1}{2}}XA^{\frac{1}{2}})(A^{\frac{1}{2}}YA^{\frac{1}{2}}) \\ &= t \operatorname{tr}(A^{\frac{1}{2}}XA^{\frac{1}{2}})^{t-1}A^{\frac{1}{2}}YA^{\frac{1}{2}} \\ &= t \operatorname{tr}A^{\frac{1}{2}}\left(A^{-\frac{1}{2}}X^{-1}A^{-\frac{1}{2}}\right)^{1-t} A^{\frac{1}{2}}Y \\ &= t \operatorname{tr}\left(A\#_{1-t}X^{-1}\right)Y. \end{aligned}$$

∎

Extremal representations for the fidelity $F(A, B)$ are useful in deriving various relations. See [30, 12]. Our next theorem gives such representations for $F_t(A, B)$. Some of these have been derived in [19] and [7].

Theorem 9. *Let A, B be any two elements of $\mathbb{P}$ and let $0 < t < 1$. Then*

(i) $F_t(A,B) = \min\limits_{X \in \mathbb{P}} \operatorname{tr}\left[(1-t)\left(A^{\frac{t-1}{2t}}XA^{\frac{t-1}{2t}}\right)^{\frac{t}{t-1}} + tXB\right]$.

(ii) $F_t(A,B) = \min\limits_{X \in \mathbb{P}} \left[\operatorname{tr}(A^{\frac{t-1}{2t}}XA^{\frac{t-1}{2t}})^{\frac{t}{t-1}}\right]^{1-t}\left[\operatorname{tr}XB\right]^t$.

(iii) $F_t(A,B) = \min\limits_{X \in \mathbb{P}} \operatorname{tr}\left[tA^{\frac{1-t}{t}}X + (1-t)\left(B^{-\frac{1}{2}}XB^{-\frac{1}{2}}\right)^{\frac{t}{t-1}}\right]$.

(iv) $F_t(A,B) = \min\limits_{X \in \mathbb{P}} \left[\operatorname{tr}A^{\frac{1-t}{t}}X\right]^t\left[\operatorname{tr}(B^{-\frac{1}{2}}XB^{-\frac{1}{2}})^{\frac{t}{t-1}}\right]^{1-t}$.

Proof. The representations (i) and (ii) have been derived and used in [19]. We will give here proofs of (iii) and (iv). The same ideas can be used to give proofs of (i) and (ii), which are different from the ones given in [19].

(iii) By Corollary 6, the function

$$f(X) = \operatorname{tr}\left[tA^{\frac{1-t}{t}}X + (1-t)\left(B^{-\frac{1}{2}}XB^{-\frac{1}{2}}\right)^{\frac{t}{t-1}}\right]$$

is strictly convex for $0 < t < 1$. Using Proposition 8 we see that

$$\nabla f(X) = t\left(A^{\frac{1-t}{t}} - B^{-1}\#_{\frac{1}{1-t}}X^{-1}\right).$$

So $\nabla f(X_0) = 0$ if and only if

$$A^{\frac{1-t}{t}} = B^{-1}\#_{\frac{1}{1-t}}X_0^{-1}. \tag{24}$$

Now, if $C = Y^{-1}\#_\alpha X^{-1}$, then from the definition (21) one can see that $X = Y\#_{\frac{1}{\alpha}} C^{-1}$. So, from (24) we see that $\nabla f(X_0) = 0$ if and only if

$$X_0 = B\#_{1-t} A^{\frac{t-1}{t}} = A^{\frac{t-1}{t}} \#_t B.$$

A little calculation shows that

$$\mathrm{tr} A^{\frac{1-t}{t}} X_0 = \mathrm{tr}\left(B^{-\frac{1}{2}} X_0 B^{-\frac{1}{2}}\right)^{\frac{t}{t-1}} = \mathrm{tr}\left(B^{\frac{1}{2}} A^{\frac{1-t}{t}} B^{\frac{1}{2}}\right)^t.$$

For any two positive matrices P and Q,

$$\mathrm{tr}\left(P^{\frac{1}{2}} Q P^{\frac{1}{2}}\right)^t = \mathrm{tr} Q^{\frac{1}{2}} P^{\frac{1}{2}} \left(P^{\frac{1}{2}} Q P^{\frac{1}{2}}\right)^t P^{-\frac{1}{2}} Q^{-\frac{1}{2}} = \mathrm{tr}\left(Q^{\frac{1}{2}} P Q^{\frac{1}{2}}\right)^t.$$

Hence

$$\mathrm{tr} A^{\frac{1-t}{t}} X_0 = \mathrm{tr}\left(B^{-\frac{1}{2}} X_0 B^{-\frac{1}{2}}\right)^{\frac{t}{t-1}} = \mathrm{tr}\left(A^{\frac{1-t}{2t}} B A^{\frac{1-t}{2t}}\right)^t = F_t(A,B).$$

We have shown that X_0 is the unique minimizer for the problem (iii) and the minimum value is equal to $F_t(A,B)$.

(iv) For an $n \times n$ matrix X, let $|X|$ be the absolute value of X defined as $|X| = (X^*X)^{\frac{1}{2}}$. Let p, q, r be positive numbers with $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$. By the matrix version of Hölder's inequality (Exercise IV. 2.7 in [8])

$$\mathrm{tr}|ST|^r \le \left(\mathrm{tr}|S|^p\right)^{\frac{r}{p}} \left(\mathrm{tr}|T|^q\right)^{\frac{r}{q}},$$

Note that

$$\begin{aligned} F_t(A,B) &= \mathrm{tr}\left(A^{\frac{1-t}{2t}} B A^{\frac{1-t}{2t}}\right)^t \\ &= \mathrm{tr}\left(B^{\frac{1}{2}} A^{\frac{1-t}{t}} B^{\frac{1}{2}}\right)^t \\ &= \mathrm{tr}\left(B^{\frac{1}{2}} X^{-\frac{1}{2}} X^{\frac{1}{2}} A^{\frac{1-t}{2t}} A^{\frac{1-t}{2t}} X^{\frac{1}{2}} X^{-\frac{1}{2}} B^{\frac{1}{2}}\right)^t. \end{aligned}$$

Taking $S = A^{\frac{1-t}{2t}} X^{\frac{1}{2}}$, $T = X^{-\frac{1}{2}} B^{\frac{1}{2}}$, $p = 1$ and $q = \frac{t}{1-t}$ in Hölder's inequality we get

$$\begin{aligned} F_t(A,B) &\le \left[\mathrm{tr}\left(X^{\frac{1}{2}} A^{\frac{1-t}{t}} X^{\frac{1}{2}}\right)\right]^t \left[\mathrm{tr}\left(B^{\frac{1}{2}} X^{-1} B^{\frac{1}{2}}\right)^{\frac{t}{1-t}}\right]^{1-t} \\ &= \left[\mathrm{tr} A^{\frac{1-t}{t}} X\right]^t \left[\mathrm{tr}\left(B^{-\frac{1}{2}} X B^{-\frac{1}{2}}\right)^{\frac{t}{t-1}}\right]^{1-t}. \end{aligned}$$

We have seen in the proof of (iii) that when $X = X_0 = A^{\frac{t-1}{t}} \#_t B$, then each of the expressions inside the square brackets on the right hand side is equal to $F_t(A,B)$. This proves (iv). ■

For given $A, B \in \mathbb{P}$ let

$$\gamma(t) = \left(A^{\frac{1-t}{2t}} B A^{\frac{1-t}{2t}}\right)^t \qquad 0 \le t \le 1. \tag{25}$$

The value $\gamma(0)$ is given by the following proposition, first established in [5].

Proposition 10. *For all* $A, B \in \mathbb{P}$ *we have*

$$\lim_{t\to 0^+} \left(A^{\frac{1-t}{2t}} B A^{\frac{1-t}{2t}}\right)^t = A.$$

Proof. Let α, β be positive numbers such that $\alpha I \leq B \leq \beta I$. Then

$$\alpha A^{\frac{1-t}{t}} \leq A^{\frac{1-t}{2t}} B A^{\frac{1-t}{2t}} \leq \beta A^{\frac{1-t}{t}},$$

and hence for $0 < t < 1$,

$$\alpha^t A^{1-t} \leq \left(A^{\frac{1-t}{2t}} B A^{\frac{1-t}{2t}}\right)^t \leq \beta^t A^{1-t}.$$

Taking the limit as $t \to 0$, we see that

$$A \leq \lim_{t\to 0^+} \left(A^{\frac{1-t}{2t}} B A^{\frac{1-t}{2t}}\right)^t \leq A.$$

This proves the proposition. ∎

Thus $\gamma(t), 0 \leq t \leq 1$, is a differentiable curve joining A and B. It is of interest to compare this with two other curves: the Riemannian geodesic (21) and the straight line segment. In this direction we have

Theorem 11. *For* $0 < t < 1$

$$\mathrm{tr} A \#_t B \leq \mathrm{tr} A^{1-t} B^t \leq \mathrm{tr} \left(A^{\frac{1-t}{2t}} B A^{\frac{1-t}{2t}}\right)^t \leq \mathrm{tr}\left[(1-t)A + tB\right]. \tag{26}$$

The first inequality in (26) is known; see e.g., [11]. The second inequality follows from the Lieb-Thirring inequality [27], and this has been recorded in the papers [19, 28, 37]. The last inequality follows from Theorem 9 (i) upon choosing $X = I$.

Our next proposition gives a formula for the derivative, with respect to t, of $F_t(A, B)$. This result has been obtained earlier as Proposition 15 in [28] and as the main ingredient in the proof of Proposition 11 in [37]. Our proof is different.

Proposition 12. *Let* A, B *be positive definite matrices and let* $\varphi : \mathbb{R} \to \mathbb{P}$ *be the function*

$$\varphi(t) = A^{\frac{1-t}{2t}} B A^{\frac{1-t}{2t}}.$$

Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ *be the function*

$$F(t) = \mathrm{tr} \varphi(t)^t = F_t(A, B).$$

Then

$$F'(t) = \mathrm{tr}\left[\varphi(t)^t \left(\log \varphi(t) - \frac{1}{t} \log A\right)\right]. \tag{27}$$

In particular,

$$F'(1) = \mathrm{tr}\left[B(\log B - \log A)\right]. \tag{28}$$

Proof. We have

$$\varphi(t) = A^{\frac{1}{2t}} \left(A^{-\frac{1}{2}} B A^{-\frac{1}{2}}\right) A^{\frac{1}{2t}}.$$

Differentiation gives

$$\varphi'(t) = -\frac{1}{2t^2}\left((\log A)\varphi(t) + \varphi(t)\log A\right).$$

Let $h : \mathbb{R}_+ \to \mathbb{H}$ be the map $h(t) = t \log \varphi(t)$. Then

$$\begin{aligned} h'(t) &= \log \varphi(t) + t\varphi(t)^{-1}\varphi'(t) \\ &= \log \varphi(t) - \frac{1}{2t}\varphi(t)^{-1}\left[(\log A)\varphi(t) + \varphi(t)\log A\right]. \end{aligned}$$

Our function $F(t) = \mathrm{tr} e^{h(t)}$. Hence

$$\begin{aligned} F'(t) &= \mathrm{tr}\left(e^{h(t)} h'(t)\right) \\ &= \mathrm{tr}\left(\varphi(t)\log\varphi(t)\right) - \frac{1}{2t}\mathrm{tr}\left[\varphi(t)^{t-1}\left((\log A)\varphi(t) + \varphi(t)\log A\right)\right] \\ &= \mathrm{tr}\left(\varphi(t)^t \log\varphi(t)\right) - \frac{1}{t}\mathrm{tr}\left(\varphi(t)^t \log A\right). \end{aligned}$$

This proves (27). ■

Using L'Hopital's rule and (28) we obtain the relation (9).

3. Higher derivatives and strong convexity

We now turn to the proof of Theorem 1. For $0 < t < 1$, let μ be the measure on $(0, \infty)$ defined by

$$d\mu(\lambda) = \frac{\sin t\pi}{\pi}\lambda^{t-1} d\lambda.$$

Then for all $x > 0$ we have

$$x^{t-1} = \int_0^\infty \frac{1}{\lambda + x} d\mu(\lambda). \tag{29}$$

See (V.4) in [8]. Differentiating both sides with respect to x, we obtain

$$(1-t)x^{t-2} = \int_0^\infty \frac{1}{(\lambda + x)^2} d\mu(\lambda). \tag{30}$$

Let $h : \mathbb{P} \to \mathbb{R}$ be the function

$$h(X) = -\frac{1}{t}\mathrm{tr} X^t.$$

By Lemma 7, the derivative of h is given by

$$Dh(X)(Y) = -\mathrm{tr} X^{t-1} Y. \tag{31}$$

Let $g(X) = X^{t-1}$. Then the second derivative $D^2 h(X)$ is the symmetric bilinear function

$$D^2 h(X)(Y, Z) = -\mathrm{tr}\left(Dg(X)(Z)\right) Y.$$

Using the integral representation (29) we see that

$$Dg(X)(Z) = -\int_0^\infty (\lambda + X)^{-1} Z (\lambda + X)^{-1} d\mu(\lambda),$$

and hence

$$D^2 h(X)(Y, Z) = \operatorname{tr} \int_0^\infty (\lambda + X)^{-1} Z (\lambda + X)^{-1} Y \, d\mu(\lambda). \tag{32}$$

In the notation of gradients

$$D^2 h(X)(Y, Z) = \langle \nabla^2 h(X)(Y), Z \rangle.$$

So, we can write (32) also as

$$\nabla^2 h(X)(Y) = \int_0^\infty (\lambda + X)^{-1} Y (\lambda + X)^{-1} \, d\mu(\lambda). \tag{33}$$

In passing, we note that this shows $\nabla^2 h(X)$ is a completely positive linear map on the space $\mathbb{H}$ of Hermitian matrices.

Now let A be any positive matrix and let

$$\tilde{h}(X) = h(A^{\frac{1}{2}} X A^{\frac{1}{2}}) = -\frac{1}{t} \operatorname{tr} \left(A^{\frac{1}{2}} X A^{\frac{1}{2}} \right)^t. \tag{34}$$

Then

$$D\tilde{h}(X)(Y) = Dh(A^{\frac{1}{2}} X A^{\frac{1}{2}})(A^{\frac{1}{2}} Y A^{\frac{1}{2}})$$

and

$$D^2 \tilde{h}(X)(Y, Z) = D^2 h(A^{\frac{1}{2}} X A^{\frac{1}{2}})(A^{\frac{1}{2}} Y A^{\frac{1}{2}}, A^{\frac{1}{2}} Z A^{\frac{1}{2}}).$$

Hence, from (32)

$$D^2 \tilde{h}(X)(Y, Z) = \operatorname{tr} \int_0^\infty \left(\lambda + A^{\frac{1}{2}} X A^{\frac{1}{2}} \right)^{-1} A^{\frac{1}{2}} Z A^{\frac{1}{2}} \left(\lambda + A^{\frac{1}{2}} X A^{\frac{1}{2}} \right)^{-1} A^{\frac{1}{2}} Y A^{\frac{1}{2}} d\mu(\lambda).$$

Using the identity

$$\left(\lambda + A^{\frac{1}{2}} X A^{\frac{1}{2}} \right)^{-1} = \left(A^{1/2} (\lambda A^{-1} + X) A^{1/2} \right)^{-1} = A^{-\frac{1}{2}} \left(\lambda A^{-1} + X \right)^{-1} A^{-\frac{1}{2}},$$

we obtain

$$D^2 \tilde{h}(X)(Y, Z) = \operatorname{tr} \int_0^\infty \left(\lambda A^{-1} + X \right)^{-1} Y \left(\lambda A^{-1} + X \right)^{-1} Z \, d\mu(\lambda).$$

In other words

$$\nabla^2 \tilde{h}(X)(Y) = \int_0^\infty \left(\lambda A^{-1} + X \right)^{-1} Y \left(\lambda A^{-1} + X \right)^{-1} d\mu(\lambda). \tag{35}$$

Let $C_\lambda = (\lambda A^{-1} + X)^{-1}$ and let Γ_{C_λ} be the map on the space of matrices defined as $\Gamma_{C_\lambda}(Y) = C_\lambda Y C_\lambda$. The eigenvalues of Γ_{C_λ} are the products of the eigenvalues of C_λ. The expression (35) can be rewritten as

$$\nabla^2 \tilde{h}(X)(Y) = \int_0^\infty \Gamma_{C_\lambda}(Y) d\mu(\lambda).$$

By the extremal principle for eigenvalues

$$\frac{\langle \Gamma_{C_\lambda}(Y), Y \rangle}{\langle Y, Y \rangle} \geq \lambda_{\min}(\Gamma_{C_\lambda}) = \lambda_{\min}(C_\lambda)^2.$$

Now let α and β be positive reals with $\alpha \le \beta$ and suppose that $\alpha I \le X \le \beta I$. Then for all $A \in \mathbb{P}$

$$\left(\frac{\lambda}{\lambda_{\min}(A)} + \beta\right)^{-1} \le C_\lambda \le \left(\frac{\lambda}{\lambda_{\max}(A)} + \alpha\right)^{-1}.$$

Using the last three relations above, we get

$$\begin{aligned}\frac{\langle \nabla^2 \tilde{h}(X)(Y), Y\rangle}{\langle Y, Y\rangle} &\ge \int_0^\infty \left(\frac{\lambda}{\lambda_{\min}(A)} + \beta\right)^{-2} d\mu(\lambda) \\ &= \lambda_{\min}(A)^2 \int_0^\infty \frac{1}{(\lambda + \beta\lambda_{\min}(A))^2}\, d\mu(\lambda) \\ &= (1-t)\beta^{t-2}\lambda_{\min}(A)^t,\end{aligned}$$

the last equality being a consequence of (30). This shows that

$$\nabla^2 \tilde{h}(X) \ge (1-t)\beta^{t-2}\lambda_{\min}(A)^t, \tag{36}$$

for all $A \in \mathbb{P}$ and $\alpha I \le X \le \beta I$.

Finally, let f be the function defined by (16). Then $-f$ is the function obtained from $\tilde{h}$ by multiplying it by t and replacing A by $A^{\frac{1-t}{t}}$. Hence, (36) leads to the inequality

$$-\nabla^2 f(X) \ge t(1-t)\beta^{t-2}\lambda_{\min}(A)^{1-t}, \tag{37}$$

for all $\alpha I \le X \le \beta I$. So, if we assume $\lambda_{\min}(A) \ge \alpha$, then we obtain the first inequality in (17).

The second inequality in (17) has an analogous proof. ■

Our method can be used to calculate higher derivatives of any order, and to estimate their norms. For example, we can show that

$$\|\nabla^3 f(X)\| \le t(1-t)(2-t)\beta^{1-t}\alpha^{t-3},$$

from which it follows that

$$\|\nabla^2 f(X) - \nabla^2 f(Y)\|_2 \le t(1-t)(2-t)\beta^{1-t}\alpha^{t-3}\|X - Y\|_2.$$

4. Gradient Projection Algorithm

Let $A_1, \dots, A_m \in \mathbb{P}$. For $0 < t < 1$ define the function φ_t on $\mathbb{P}$ as

$$\varphi_t(X) = \sum_{j=1}^m w_j \left[\operatorname{tr}((1-t)A_j + tX) - \operatorname{tr}\left(A_j^{\frac{1-t}{2t}} X A_j^{\frac{1-t}{2t}}\right)^t\right]$$

We consider the optimization problem

$$\min_{X \in \mathbb{P}} \varphi_t(X) \tag{38}$$

on the convex cone $\mathbb{P}$. The multimarginal optimal transport problem of Gaussian measures ([1, 20, 21]) is the special case $t = 1/2$. Let α and β be positive numbers such that

$$\alpha I \le A_j \le \beta I, \quad j = 1, \dots, m.$$

We note that the optimal values of α and β are

$$\min_{1\le j\le m} \lambda_{\min}(A_j) \quad \text{and} \quad \max_{1\le j\le m} \lambda_{\max}(A_j)$$

respectively. By the results obtained in the previous section, we see that φ_t is $t(1-t)\beta^{1-t}\alpha^{t-2}$ -smooth and $t(1-t)\beta^{t-2}\alpha^{1-t}$ -strongly convex, and the condition number of $\nabla^2\varphi_t(X)$ is bounded by $\left(\frac{\beta}{\alpha}\right)^{3-2t}$. By Proposition 8 φ_t is strictly convex with

$$\begin{aligned} D\varphi_t(X)(Y) &= t\sum_{j=1}^m w_j \mathrm{tr}\left[I-(A_j^{\frac{1-t}{t}}\#_{1-t}X^{-1})\right]Y \\ &= t\ \mathrm{tr}\left[I-\sum_{j=1}^m w_j(A_j^{\frac{1-t}{t}}\#_{1-t}X^{-1})\right]Y. \end{aligned}$$

In terms of the gradient

$$\nabla\varphi_t(X) = t\ \left[I-\sum_{j=1}^m w_j(A_j^{\frac{1-t}{t}}\#_{1-t}X^{-1})\right].$$

To prove the existence and uniqueness of the minimization problem (38), it is enough to show that the equation $\nabla\varphi_t(X)=0$ has a positive definite solution. This is equivalent to the nonlinear matrix equation

$$X=\sum_{j=1}^m w_j X^{1/2}\left(X^{-1}\#_t A_j^{\frac{1-t}{t}}\right)X^{1/2}=\sum_{j=1}^m w_j\left(X^{1/2}A_j^{\frac{1-t}{t}}X^{1/2}\right)^t.$$

Let $F:\mathbb{P}\to\mathbb{P}$ be the map defined by

$$F(X)=\sum_{j=1}^m w_j\left(X^{1/2}A_j^{\frac{1-t}{t}}X^{1/2}\right)^t.$$

If all A_j, $1\le j\le m$, and X are bounded from below by αI and from above by βI, then

$$X^{1/2}A_j^{\frac{1-t}{t}}X^{1/2} \quad\le\quad X^{1/2}(\beta^{\frac{1-t}{t}}I)X^{1/2}\le\beta^{\frac{1-t}{t}}X\le\beta^{\frac{1-t}{t}}\beta I\le\beta^{1/t}I$$

and hence $F(X)\le\sum_{j=1}^m w_j\beta I=\beta I$. Similarly $F(X)\ge\alpha I$. This shows that F is a self-map on the compact and convex interval $[\alpha I,\beta I]:=\{X>0:\alpha I\le X\le\beta\}$. By Brouwer's fixed point theorem, F has a fixed point. This settles the problem of existence and uniqueness of the minimizer in (38).

Now we apply the classical gradient projection method for (constrained) strongly convex functions. Let

$$\begin{aligned} X_{k+1} &= [X_k-\eta\nabla f(X_k)]_+ \\ &= \left[X_k-t\eta I+t\eta\sum_{j=1}^n w_j\left(A_j^{\frac{1-t}{t}}\#_{1-t}X_k^{-1}\right)\right]_+ \end{aligned}$$

where $X_0\in[\alpha I,\beta I]$ and $[\cdot]_+$ denotes the projection to $[\alpha I,\beta I]$ and $0<\eta<\frac{2}{\beta_*}$. Since φ_t is $\beta_*:=t(1-t)\beta^{1-t}\alpha^{t-2}$ -smooth and $\alpha_*:=t(1-t)\beta^{t-2}\alpha^{1-t}$ -strongly convex, the iteration converges to the unique minimizer X_* with linear convergence rate

$$\|X_{k+1} - X_*\|_2 \leq q^k \|X_k - X_*\|_2 \tag{39}$$

where

$$q = \max\{|1 - \eta\alpha_*|, |1 - \eta\beta_*|\}.$$

Or, with $\eta = 1/\beta_*$,

$$\|X_{k+1} - X_*\|_2^2 \leq e^{-\frac{k\alpha_*}{\beta_*}} \|X_1 - X_*\|_2 = e^{-k(\alpha/\beta)^{3-2t}} \|X_1 - X_*\|_2^2.$$

See (Theorem 3.10, [14]). A gradient-based optimization method with sublinear convergence for $t = 1/2$ has recently appeared in [26].

5. Appendix

Some of the inequalities in (26) have much stronger versions, and these are related to recurring themes in matrix analysis and mathematical physics. See e.g., [3, 8, 9, 11, 16, 27].

Let x, y be two n-vectors with nonnegative components. Let $x_1^\downarrow \geq \cdots \geq x_n^\downarrow$ be the components of x arranged in decreasing order. We say that x is *weakly log majorised* by y, in symbols $x \prec_{\mathrm{wlog}} y$, if for $1 \leq k \leq n$

$$\prod_{j=1}^{k} x_j^\downarrow \leq \prod_{j=1}^{k} y_j^\downarrow. \tag{40}$$

If in addition

$$\prod_{j=1}^{n} x_j^\downarrow = \prod_{j=1}^{n} y_j^\downarrow,$$

then we say that x is *log majorised* by y, and write this as $x \prec_{\log} y$. We write $x \leq y$ if $x_j^\downarrow \leq y_j^\downarrow$ for all $j = 1, \ldots, n$.

Let X be any $n \times n$ matrix and let $\lambda(X) = (\lambda_1(X), \ldots, \lambda_n(X))$ and $s(X) = (s_1(X), \ldots, s_n(X))$ be the n-tuples whose components are the eigenvalues and the singular values of X, respectively. A famous inequality of H. Weyl says that

$$(|\lambda_1(X)|, \ldots, |\lambda_n(X)|) \prec_{\log} s(X). \tag{41}$$

(See [8] p. 43.)

Now let A and B be positive definite matrices and let $0 < t < 1$. It has been shown in [11] that

$$\lambda(A\#_t B) \prec_{\log} \lambda(A^{1-t}B^t). \tag{42}$$

A matrix version of Young's inequality proved by T. Ando [3] says that

$$s(A^{1-t}B^t) \leq \lambda((1-t)A + tB). \tag{43}$$

Combining (42) and (43) we have the chain

$$\begin{aligned} \lambda(A\#_t B) &\prec_{\log} \lambda(A^{1-t}B^t) \prec_{\log} s(A^{1-t}B^t) \\ &\leq \lambda((1-t)A + tB). \end{aligned} \tag{44}$$

The inequality (26) raises the intriguing question of how the eigenvalue tuple $\lambda\left(A^{\frac{1-t}{2t}}BA^{\frac{1-t}{2t}}\right)^t$ fits into this chain. To answer this we recall the Araki-Lieb-Thirring inequalities which say that if X and Y are positive definite matrices, then

$$\lambda(X^tY^tX^t) \prec_{\log} \lambda(XYX)^t, \text{ for } 0 \le t \le 1, \tag{45}$$

and

$$\lambda(XYX)^t \prec_{\log} \lambda(X^tY^tX^t), \text{ for } t \ge 1. \tag{46}$$

See the proof of Theorem IX.2.10 in [8]. Using the first of these inequalities, we see that for $0 \le t \le 1$,

$$\lambda(A^{1-t}B^t) = \lambda\left(A^{\frac{1-t}{2}}B^tA^{\frac{1-t}{2}}\right) \prec_{\log} \lambda\left(A^{\frac{1-t}{2t}}BA^{\frac{1-t}{2t}}\right)^t. \tag{47}$$

Now suppose $\frac{1}{2} \le t \le 1$. Then from (46) we obtain

$$\lambda\left(A^{\frac{1-t}{2t}}BA^{\frac{1-t}{2t}}\right)^{2t} \prec_{\log} \lambda\left(A^{1-t}B^{2t}A^{1-t}\right).$$

Taking square roots of both sides, we get

$$\lambda\left(A^{\frac{1-t}{2t}}BA^{\frac{1-t}{2t}}\right)^t \prec_{\log} s(A^{1-t}B^t). \tag{48}$$

Combining (44), (47) and (48) we have

$$\begin{aligned} \lambda(A\#_tB) &\prec_{\log} \lambda(A^{1-t}B^t) \\ &\prec_{\log} \lambda\left(A^{\frac{1-t}{2t}}BA^{\frac{1-t}{2t}}\right)^t \\ &\prec_{\log} s(A^{1-t}B^t) \\ &\le \lambda((1-t)A+tB), \end{aligned} \tag{49}$$

for $\frac{1}{2} \le t \le 1$.

On the other hand if $0 \le t \le \frac{1}{2}$, then from (45) we obtain

$$\lambda\left(A^{1-t}B^{2t}A^{1-t}\right) \prec_{\log} \lambda\left(A^{\frac{1-t}{2t}}BA^{\frac{1-t}{2t}}\right)^{2t}.$$

Taking square roots of both sides we get

$$s(A^{1-t}B^t) \prec_{\log} \lambda\left(A^{\frac{1-t}{2t}}BA^{\frac{1-t}{2t}}\right)^t.$$

So for $0 \le t \le \frac{1}{2}$, we have

$$\begin{aligned} \lambda(A\#_tB) &\prec_{\log} \lambda(A^{1-t}B^t) \\ &\prec_{\log} s(A^{1-t}B^t) \\ &\prec_{\log} \lambda\left(A^{\frac{1-t}{2t}}BA^{\frac{1-t}{2t}}\right)^t. \end{aligned} \tag{50}$$

To complete this chain in the same way as (49) it remains to answer whether for $0 \le t \le \frac{1}{2}$, $\lambda\left(A^{\frac{1-t}{2t}}BA^{\frac{1-t}{2t}}\right)^t$ is dominated by $\lambda((1-t)A+tB)$.

Acknowledgement. We thank the anonymous referee and Professor Fumio Hiai for a careful reading of the manuscript. The work of R. Bhatia is supported by a J. C. Bose National Fellowship and of Y. Lim is supported by the National Research Foundation of Korea (NRF) grant founded by the Korea government (MEST) (No. 2015R1A3A2031159) and 2016R1A5A1008055.

References

[1] M. Agueh and G. Carlier, *Barycenters in the Wasserstein space*, SIAM J. Math. Anal. Appl. 43 (2011), 904-924.

[2] P. M. Alberti, *A note on the transition probability over C^* -algebras*, Lett. Math. Phys. 7 (1983), 25-32.

[3] T. Ando, *Matrix Young inequalities*, Oper. Theory Adv. Appl., Birkhouser, Basel, 75 (1995) 33-38.

[4] E. Andruchow, G. Corach and D. Stojanoff, *Geometrical significance of the Löwner-Heinz inequality*, Proc. Amer. Math. Soc. 128 (2000), 1031-1037.

[5] K. Audenaert and N. Datta, *$\alpha - z$ Rényi relative entropies*, J. Mathematical Physics 56 (2015), 022202.

[6] I. Bengtsson and K. Zyczkowski, *Geometry of Quantum States: An Introduction to Quantum Entanglement*, Cambridge University Press, 2006.

[7] M. Berta, O. Fawzi and M. Tomamichel, *On variational expressions for quantum relative entropies*, Letters in Mathematical Physics, 107 (2017), 2239-2265.

[8] R. Bhatia, *Matrix Analysis,* Springer, 1997.

[9] R. Bhatia, *Positive Definite Matrices,* Princeton University Press, 2007.

[10] R. Bhatia, *The Riemannian mean of positive matrices*, in Matrix Information Geometry, eds. F. Nielsen and R. Bhatia, Springer, (2013), 35-51.

[11] R. Bhatia and P. Grover, *Norm inequalities related to the matrix geometric mean*, Linear Algebra Appl. 437 (2012), 726-733.

[12] R. Bhatia, T. Jain and Y. Lim, *On the Bures-Wasserstein distance between positive definite matrices*, to appear in Expositiones Mathematicae.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[14] S. Bubeck, *Convex Optimization: Algorithms and Complexity*, Foundations and Trends in Machine Learning 8 (2015), 231-357.

[15] D. Bures, *An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w^* -algebras*, Trans. Am. Math. Soc. **135** (1969), 199-212.

[16] E. Carlen, *Trace Inequalities and Quantum Entropy: An Introductory Course,* in Entropy and the Quantum, 73-140, Contemp. Math. 529, Amer. Math. Soc., 2010.

[17] N. Datta, *Min-and max-relative entropies and a new entanglement monotone*, IEEE Transactions on Information Theory 55 (2009), 2816-2826.

[18] D. Dowson and B. Landau, *The Fréchet distance between multivariate normal distributions*, J. Math. Anal. 12 (1982), 450-455.

[19] R. Frank and E. Lieb, *Monotonicity of a relative Rényi entropy*, Journal of Mathematical Physics **54** (2013), 122201.

[20] W. Gangbo and R. J. McCann, *The geometry of optimal transportation*, Acta Math. **177** (1996), 113-161.

[21] W. Gangbo and A. Swiech, *Optimal maps for the multidimensional Monge-Kantrovich problem*, Commu. Pure and Applied Math. **51** (1998), 23-45.

[22] C. Givens and R. Shortt, *A class of Wasserstein metrics for probability distributions*, Michigan Math. J. 31 (1984), 231-240.

[23] M. Hayashi, *Quantum Information*, Springer, 2006.

[24] A. Jencová, *Geodesic distances on density matrices*, J. Math. Phys. 45 (2004), 1787-1794.

[25] M. Knott and C. Smith, *On the optimal mapping of distributions*, J. Optim. Theory Appl. 43 (1984), 39-49.

[26] S. Kum, Y. Lim and S. Yun, *Gradient projection methods for the n-coupling problem*, submitted.

[27] E. Lieb and W. Thirring, *Inequalities for the moments of the eigenvalues of the Schrodinger Hamiltonian and their relation to Sobolev inequalities*, in Studies in Mathematical Physics, edited by E. Lieb, B. Simon, and A. Wightman (Princeton University, Princeton, NJ, 1976), 269-303.

[28] M. Müller-Lennert, F. Dupuis, O. Szehr, S. Fehr, M. Tomamichel, *On quantum Renyi entropies: a new definition and some properties*, Journal of Mathematical Physics, **54** (2013), 122203

[29] R. D. Nussbaum, *Finsler structures for the part metric and Hilbert's projective metric and applications to ordinary differential equations*, Differential and Integral Equations **7** (1994), 1649-1707.

[30] I. Olkin and F. Pukelsheim, *The distance between two random vectors with given dispersion matrices*, Linear Algebra Appl. 48 (1982), 257-263.

[31] W. Pusz and S. L. Woronowicz, *Functional calculus for sesquilinear forms and the purification map*, Rep. Math. Phys. 8 (1975), 159-170.

[32] A. Takatsu, *Wasserstein geometry of Gaussian measures*, Osaka J. Math. 48 (2011), 1005-1026.

[33] A. C. Thompson, *On certain contraction mappings in a partially ordered vector space*, Proc. Amer. Math. Soc. **14** (1963), 438–443.

[34] A. Uhlmann, *Density operators as an arena for differential geometry*, Rep. Math. Phys. 33 (1993), 255-263.

[35] A. Uhlmann, *Transition probability (fidelity) and its relatives*, Found. Phys. 41 (2011), 288-298.

[36] C. Villani, *Optimal Transport: Old and New*, Springer, 2008.

[37] M. Wilde, A. Winter and D. Yang, *Strong converse for the classical capacity of entanglement-breaking and Hadamardchannels via a sandwiched Renyi relative entropy*, Communications in Mathematical Physics **331** (2014), 593-622

Ashoka University, Sonepat, Haryana, 131029, India
E-mail address: `rajendra.bhatia@ashoka.edu.in`

Indian Statistical Institute, New Delhi 110016, India
E-mail address: `tanvi@isid.ac.in`

Department of Mathematics, Sungkyunkwan University, Suwon 440-746, Korea
E-mail address: `ylim@skku.edu`