

## POSITIVE DEFINITE MATRICES AND THE S-DIVERGENCE

SUVRIT SRA

(Communicated by Jeremy Tyson)

**ABSTRACT.** Hermitian positive definite (hpd) matrices form a self-dual convex cone whose interior is a Riemannian manifold of nonpositive curvature. The manifold view comes with a natural distance function but the conic view does not. Thus, drawing motivation from convex optimization we introduce the *S-divergence*, a distance-like function on the cone of hpd matrices. We study basic properties of the S-divergence and explore its connections to the Riemannian distance. In particular, we show that (i) its square-root is a distance, and (ii) it exhibits numerous nonpositive-curvature-like properties.

### 1. INTRODUCTION

Hermitian positive definite (hpd) matrices form a manifold of nonpositive curvature [11, Ch.10], [6, Ch.6]. Their closure forms a self-dual convex cone central to modern convex optimization [10, 21] and numerous other areas [6]. But unlike the hpd manifold, the hpd cone does not come with a “natural” distance function.

Drawing motivation from convex optimization, we introduce a distance-like function on hpd matrices, which we name the *S-divergence*. We prove several results that uncover properties of this divergence, of which our main result is that its square-root is actually a true distance. Our other results explore geometric and analytic similarities between the S-divergence and the Riemannian distance.

**1.1. Setup.** Let  $\mathbb{H}_n$  be the set of  $n \times n$  Hermitian matrices. A matrix  $A \in \mathbb{H}_n$  is called *positive definite* if

$$(1.1) \quad \langle x, Ax \rangle > 0 \quad \text{for all } x \neq 0, \quad \text{also written as } A > 0.$$

We say  $A$  is *positive semidefinite* if  $\langle x, Ax \rangle \geq 0$  for all  $x$ , and write  $A \geq 0$ . For  $A, B \in \mathbb{H}_n$ , the inequality  $A \geq B$  means  $A - B \geq 0$ . The set of positive definite (henceforth *positive*) matrices in  $\mathbb{H}_n$  is denoted by  $\mathbb{P}_n$ . The *Frobenius norm* of a matrix  $X$  is  $\|X\|_F := \sqrt{\text{tr}(X^*X)}$ , while  $\|X\|$  denotes the operator norm. If  $f: \mathbb{R} \rightarrow \mathbb{R}$  and  $A$  has the eigendecomposition  $A = U\Lambda U^*$  with unitary  $U$ , we define  $f(A) = Uf(\Lambda)U^*$ , where  $f(\Lambda)$  is the diagonal matrix  $\text{Diag}[f(\Lambda_{11}), \dots, f(\Lambda_{nn})]$ .

The set  $\mathbb{P}_n$  is a differentiable Riemannian manifold, with the Riemannian metric given by the differential form  $ds = \|A^{-1/2}dAA^{-1/2}\|_F$ . This metric induces the *Riemannian distance* (see, e.g., [6, Ch. 6]):

$$(1.2) \quad \delta_R(X, Y) := \|\log(Y^{-1/2}XY^{-1/2})\|_F \quad \text{for } X, Y > 0.$$

---

Received by the editors March 6, 2015 and, in revised form, August 18, 2015.

2010 *Mathematics Subject Classification.* Primary 15A45, 52A99, 47B65, 65F60.

This work was done while the author was with the MPI for Intelligent Systems, Tübingen, Germany. A small fraction of this work was presented at the Neural Information Processing Systems (NIPS) Conference 2012.

The set  $\mathbb{P}_n$  is an (open) convex cone, on which we can define the following distance-like function, namely, the *S-divergence*:<sup>1</sup>

$$(1.3) \quad \delta_S^2(X, Y) := \log \det \left( \frac{X+Y}{2} \right) - \frac{1}{2} \log \det(XY) \quad \text{for } X, Y > 0.$$

This divergence complements (1.2), and it was originally proposed as a numerical alternative to (1.2) in [14]. There it was used in an application to image-search, primarily due to its computational and empirical benefits; this initial success of the S-divergence is what motivated our theoretical study in this paper.

*Main contributions.* The present paper goes substantially beyond our initial conference version [23]. The main additions are: (i) Theorems 4.1, 4.5, 4.6, 4.7, and 4.9, which establish several new geometric and analytic similarities between  $\delta_S$  and  $\delta_R$ ; (ii) the joint geodesic convexity of  $\delta_S^2$  (Proposition 4.3, Theorem 4.4); and (iii) new “conic” contraction results for  $\delta_S$  and  $\delta_R$  that uncover properties akin to those exhibited by Hilbert’s projective metric and Thompson’s part metric [20] (Proposition 4.11, Corollary 4.12, Theorem 4.14, Theorem 4.15, and Corollary 4.16).

Concurrent to our work, Chebbi and Moakher (CM) [12] have considered a family of divergences that parametrically generalize (1.3). CM actually proved  $\delta_S$  to be a distance, though only for commuting matrices; we remove this restriction to tackle both the commuting and noncommuting cases. A question closely related to  $\delta_S$  being a distance is whether the matrix  $[\det(X_i + X_j)^{-\beta}]_{i,j=1}^m$  is positive semidefinite for arbitrary  $X_1, \dots, X_m \in \mathbb{P}_n$ , every integer  $m \geq 1$ , and any scalar  $\beta \geq 0$ . We provide a complete characterization of necessary and sufficient conditions on  $\beta$  for the above matrix to be semidefinite by relating the question to a deeper result of Gindikin on symmetric spaces [17].

## 2. THE S-DIVERGENCE

Consider a differentiable strictly convex function  $f : \mathbb{R} \rightarrow \mathbb{R}$ ; then,  $f(x) \geq f(y) + f'(y)(x - y)$ , with equality if and only if  $x = y$ . The difference between the two sides of this inequality is called a *Bregman divergence*:<sup>2</sup>

$$(2.1) \quad D_f(x, y) := f(x) - f(y) - f'(y)(x - y).$$

The scalar divergence (2.1) readily extends to Hermitian matrices. Specifically, let  $f$  be differentiable and strictly convex on  $\mathbb{R}$ , and let  $X, Y \in \mathbb{H}_n$  be arbitrary. Then, the *Bregman (matrix) divergence* may be defined as

$$(2.2) \quad D_f(X, Y) := \operatorname{tr} f(X) - \operatorname{tr} f(Y) - \operatorname{tr}(f'(Y)(X - Y)).$$

It can be verified that  $D_f$  is nonnegative, strictly convex in  $X$ , and zero if and only if  $X = Y$ ; it is typically asymmetric. For example, if  $f(x) = \frac{1}{2}x^2$ , then for  $X \in \mathbb{H}_n$ ,  $\operatorname{tr} f(X) = \frac{1}{2} \operatorname{tr}(X^2)$  and (2.2) becomes  $\frac{1}{2} \|X - Y\|_F^2$ , but if  $f(x) = x \log x - x$  on  $(0, \infty)$ , then  $\operatorname{tr} f(X) = \operatorname{tr}(X \log X - X)$  and (2.2) becomes the *von Neumann divergence*  $D_{vn}(X, Y) = \operatorname{tr}(X \log X - X \log Y - X + Y)$  (which is clearly asymmetric).

<sup>1</sup>Called a divergence because although nonnegative, definite, and symmetric, it is *not* a distance.

<sup>2</sup>Over vectors, these divergences have been well studied; see, e.g., [2]. Although not distances, they often behave like squared distances, in a sense that can be made precise for certain  $f$  [13].

The asymmetry of Bregman divergences can be sometimes undesirable. Therefore, researchers have also considered symmetric divergences, among which perhaps the most popular is the *Jensen-Shannon/Bregman divergence*:

$$(2.3) \quad S_f(X, Y) := \frac{1}{2} \left( D_f\left(X, \frac{X+Y}{2}\right) + D_f\left(Y, \frac{X+Y}{2}\right) \right).$$

Divergence (2.3) may also be written in the more revealing form:

$$(2.4) \quad S_f(X, Y) = \frac{1}{2} (\operatorname{tr} f(X) + \operatorname{tr} f(Y)) - \operatorname{tr} f\left(\frac{X+Y}{2}\right).$$

The S-divergence (1.3) can be obtained from (2.4) by setting  $f(x) = -\log x$ , so that  $\operatorname{tr} f(X) = -\log \det(X)$ , the venerable barrier function for the hpd cone [21]. The S-divergence may also be viewed as the Jensen-Bregman divergence between two multivariate gaussians [15], or as the Bhattacharyya distance between them [8].

The following basic properties of  $S$  are easily verified.

**Proposition 2.1.** *Let  $\lambda(X)$  be the vector of eigenvalues of  $X$ , and  $\operatorname{Eig}(X)$  the diagonal matrix formed from  $\lambda(X)$ ; let  $A, B, C \in \mathbb{P}_n$ . Then,*

- (i)  $\delta_S(I, A) = \delta_S(I, \operatorname{Eig}(A))$ ;
- (ii)  $\delta_S(A, B) = \delta_S(P^*AP, P^*BP)$ , where  $P \in GL_n(\mathbb{C})$ ;
- (iii)  $\delta_S(A, B) = \delta_S(A^{-1}, B^{-1})$ ;
- (iv)  $\delta_S^2(A \otimes B, A \otimes C) = n\delta_S^2(B, C)$ .

### 3. THE $\delta_S$ DISTANCE

This section presents our main result: *the square-root  $\delta_S$  of the S-divergence is a distance*. Previous authors [12, 14] independently conjectured this result, and appealed to classical ideas from harmonic analysis [3, Ch. 3] to establish the result for commuting matrices. We establish the (harder) noncommutative case below.

**Theorem 3.1.** *Let  $\delta_S$  be defined by (1.3). Then,  $\delta_S$  is a metric on  $\mathbb{P}_n$ .*

To prove Theorem 3.1, we need several auxiliary results.

**Definition 3.2** ([3, Def. 1.1]). Let  $\mathcal{X}$  be a nonempty set. A function  $\psi : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is said to be *negative definite* if  $\psi(x, y) = \psi(y, x)$  for all  $x, y \in \mathcal{X}$ , and the inequality

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

holds for all integers  $n \geq 2$  and subsets  $\{x_i\}_{i=1}^n \subseteq \mathcal{X}$ ,  $\{c_i\}_{i=1}^n \subseteq \mathbb{R}$  with  $\sum_{i=1}^n c_i = 0$ .

**Theorem 3.3** ([3, Prop. 3.2, Ch. 3]). *If  $\psi : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is negative definite, then there is a Hilbert space  $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$  and a mapping  $x \mapsto \varphi(x)$  from  $\mathcal{X} \rightarrow \mathcal{H}$  such that*

$$(3.1) \quad \|\varphi(x) - \varphi(y)\|_{\mathcal{H}}^2 = \frac{1}{2} (\psi(x, x) + \psi(y, y)) - \psi(x, y).$$

*Moreover, negative definiteness of  $\psi$  is necessary for such a mapping to exist.*

Theorem 3.3 helps prove the triangle inequality for the scalar case.<sup>3</sup>

**Lemma 3.4.** *Let  $\delta_s$  be the scalar version of  $\delta_S$ , i.e.,*

$$\delta_s(x, y) := \sqrt{\log[(x+y)/(2\sqrt{xy})]}, \quad x, y > 0.$$

*Then,  $\delta_s$  is a metric on  $(0, \infty)$ .*

<sup>3</sup>Schoenberg's theorem (Theorem 3.3) for establishing the commutative case (which is essentially just the scalar case of Lemma 3.4) was also used in [12].

*Proof.* To verify that  $\psi(x, y) = \log((x + y)/2)$  is negative definite, by [3, Thm. 2.2, Ch. 3] we may equivalently show that  $e^{-\beta\psi(x, y)} = \left(\frac{x+y}{2}\right)^{-\beta}$  is positive definite for any  $\beta > 0$ . But this follows readily upon noting the integral representation

$$(3.2) \quad (x + y)^{-\beta} = \frac{1}{\Gamma(\beta)} \int_0^\infty e^{-t(x+y)} t^{\beta-1} dt = \langle f_x, f_y \rangle,$$

where  $f_x(t) = e^{-tx} t^{\frac{\beta-1}{2}} \in L_2([0, \infty))$ .  $\square$

**Corollary 3.5.** *Let  $x, y, z \in \mathbb{R}_{++}^n$  and let  $p \geq 1$ . Then,*

$$(3.3) \quad \left( \sum_i \delta_s^p(x_i, y_i) \right)^{1/p} \leq \left( \sum_i \delta_s^p(x_i, z_i) \right)^{1/p} + \left( \sum_i \delta_s^p(y_i, z_i) \right)^{1/p}.$$

**Corollary 3.6.** *Let  $X, Y, Z > 0$  be diagonal matrices. Then,*

$$(3.4) \quad \delta_S(X, Y) \leq \delta_S(X, Z) + \delta_S(Y, Z).$$

*Proof.* For diagonal  $X, Y$ ,  $\delta_S^2(X, Y) = \sum_i \delta_S^2(X_{ii}, Y_{ii})$ ; now use Corollary 3.5.  $\square$

Next, we recall an important determinantal inequality for positive matrices.

**Theorem 3.7** ([16]). *Let  $A, B > 0$ . Let  $\lambda^\downarrow(X)$  be the vector of eigenvalues of  $X$  arranged in decreasing order; define  $\lambda^\uparrow(X)$  likewise. Then,*

$$(3.5) \quad \prod_{i=1}^n (\lambda_i^\downarrow(A) + \lambda_i^\downarrow(B)) \leq \det(A + B) \leq \prod_{i=1}^n (\lambda_i^\downarrow(A) + \lambda_i^\uparrow(B)).$$

**Corollary 3.8.** *Let  $A, B > 0$ . Let  $\text{Eig}^\downarrow(X)$  denote the diagonal matrix with  $\lambda^\downarrow(X)$  as its diagonal; define  $\text{Eig}^\uparrow(X)$  likewise. Then,*

$$(3.6) \quad \delta_S(\text{Eig}^\downarrow(A), \text{Eig}^\downarrow(B)) \leq \delta_S(A, B) \leq \delta_S(\text{Eig}^\downarrow(A), \text{Eig}^\uparrow(B)).$$

*Proof.* In (3.5), divide throughout by  $2^n \sqrt{\det(A) \det(B)}$  to obtain

$$\frac{\prod_{i=1}^n (\lambda_i^\downarrow(\frac{A}{2}) + \lambda_i^\downarrow(\frac{B}{2}))}{\sqrt{\det(A) \det(B)}} \leq \frac{\det(\frac{A+B}{2})}{\sqrt{\det(A) \det(B)}} \leq \frac{\prod_{i=1}^n (\lambda_i^\downarrow(\frac{A}{2}) + \lambda_i^\uparrow(\frac{B}{2}))}{\sqrt{\det(A) \det(B)}}.$$

Since  $\det(A) = \det(\text{Eig}^\uparrow(A)) = \det(\text{Eig}^\downarrow(A))$ , we can rewrite the leftmost and rightmost terms above, so that upon taking logarithms we obtain (3.6).  $\square$

The final result we need is a classic lemma from linear algebra.

**Lemma 3.9.** *If  $A > 0$  and  $B$  is Hermitian, then there is a matrix  $P$  such that*

$$(3.7) \quad P^*AP = I \quad \text{and} \quad P^*BP = D, \quad \text{where } D \text{ is diagonal.}$$

Equipped with the above results, we are ready to prove Theorem 3.1.

*Proof of Theorem 3.1.* We need to show that  $\delta_S$  is symmetric, nonnegative, definite, and that it satisfies the triangle inequality. Symmetry and nonnegativity are obvious, while definiteness follows from strict convexity of  $-\log \det(X)$ , the seed function that generates the  $S$ -divergence. The only difficulty is posed by the triangle inequality.

Let  $X, Y, Z > 0$  be arbitrary. From Lemma 3.9 we know that there is a matrix  $P$  such that  $P^*XP = I$  and  $P^*YP = D$ . Since  $Z > 0$  is arbitrary and congruence preserves positive definiteness, we may write just  $Z$  instead of  $P^*ZP$ . Also, since  $\delta_S(P^*XP, P^*YP) = \delta_S(X, Y)$  (see Proposition 2.1), proving the triangle inequality reduces to showing that

$$(3.8) \quad \delta_S(I, D) \leq \delta_S(I, Z) + \delta_S(D, Z).$$

Consider now the diagonal matrices  $D^\downarrow$  and  $\text{Eig}^\downarrow(Z)$ . Corollary 3.6 asserts

$$(3.9) \quad \delta_S(I, D^\downarrow) \leq \delta_S(I, \text{Eig}^\downarrow(Z)) + \delta_S(D^\downarrow, \text{Eig}^\downarrow(Z)).$$

Proposition 2.1(i) implies that  $\delta_S(I, D) = \delta_S(I, D^\downarrow)$  and  $\delta_S(I, Z) = \delta_S(I, \text{Eig}^\downarrow(Z))$ , while Corollary 3.8 shows that  $\delta_S(D^\downarrow, \text{Eig}^\downarrow(Z)) \leq \delta_S(D, Z)$ . Combining these inequalities, we immediately obtain (3.8).  $\square$

We now briefly consider a closely related topic of importance in some applications: kernel functions arising from  $\delta_S$ .

**3.1. Hilbert space embedding.** Since  $\delta_S$  is a metric that embeds isometrically into Hilbert space (Lemma 3.4) when restricted to scalars, it is natural to ask whether it also admits such an embedding for matrices. But as already noted, such an embedding does not exist. Specifically, Theorem 3.3 shows that a Hilbert space embedding exists if and only if  $\delta_S^2(X, Y)$  is a negative definite kernel; equivalently, if and only if the map (cf. Lemma 3.4)

$$e^{-\beta \delta_S^2(X, Y)} = \frac{\det(X)^\beta \det(Y)^\beta}{\det((X + Y)/2)^\beta}$$

is a positive definite kernel for  $\beta > 0$ . This in turn is equivalent to the matrix

$$(3.10) \quad H_\beta = [h_{ij}] = [\det(X_i + X_j)^{-\beta}], \quad 1 \leq i, j \leq m,$$

being positive definite for every  $m \geq 1$  and arbitrary positive matrices  $X_1, \dots, X_m \in \mathbb{P}_n$ . A quick numerical check shows, however, that  $H_\beta$  can be indefinite. Thus, we are led to the weaker question: *for what choices of  $\beta$  is  $H_\beta \geq 0$ ?*

Theorem 3.10 provides an answer for (real) symmetric positive definite matrices.

**Theorem 3.10.** *Let  $X_1, \dots, X_m$  be real symmetric matrices in  $\mathbb{P}_n$ . The  $m \times m$  matrix  $H_\beta$  defined by (3.10) is positive definite, if and only if  $\beta$  satisfies*

$$(3.11) \quad \beta \in \left\{ \frac{j}{2} : j \in \mathbb{N} \text{ and } 1 \leq j \leq (n-1) \right\} \cup \left\{ \gamma : \gamma \in \mathbb{R} \text{ and } \gamma > \frac{1}{2}(n-1) \right\}.$$

We refer the interested reader to the longer version of this paper [22] for details.

Theorem 3.10 says that  $e^{-\beta \delta_S^2}$  is not always a kernel, though for commuting matrices  $e^{-\beta \delta_S^2}$  is always a kernel. This discrepancy raises the following question:

**Open problem.** Determine necessary and sufficient conditions on a set  $\mathcal{X} \subset \mathbb{P}_n$ , so that  $e^{-\beta \delta_S^2(X, Y)}$  is a kernel function on  $\mathcal{X} \times \mathcal{X}$  for all  $\beta > 0$ .

#### 4. GEOMETRIC AND ANALYTIC SIMILARITIES WITH $\delta_R$

**4.1. Geometric mean.** We begin by studying an object that connects  $\delta_R$  and  $\delta_S^2$  most intimately: the matrix geometric mean. For positive  $A$  and  $B$ , the *matrix geometric mean* (MGM) is denoted by  $A \sharp B$ , and is given by the formula [6, Ch. 4]

$$(4.1) \quad A \sharp B := A^{1/2} (A^{-1/2} B A^{-1/2})^{1/2} A^{1/2}.$$

The MGM enjoys a host of attractive properties—see for instance the classic paper [1]; of these, the following variational characterization [7] is important:

$$(4.2) \quad \begin{aligned} A \sharp B &= \operatorname{argmin}_{X > 0} \delta_R^2(A, X) + \delta_R^2(B, X) \quad \text{and} \\ \delta_R(A, A \sharp B) &= \delta_R(B, A \sharp B). \end{aligned}$$

Surprisingly, the MGM enjoys a similar characterization even under  $\delta_S^2$ .

**Theorem 4.1.** *Let  $A, B > 0$ . Then,*

$$(4.3) \quad A\sharp B = \operatorname{argmin}_{X>0} [h(X) := \delta_S^2(X, A) + \delta_S^2(X, B)].$$

*Moreover,  $A\sharp B$  is equidistant from  $A$  and  $B$ , i.e.,  $\delta_S(A, A\sharp B) = \delta_S(B, A\sharp B)$ .*

*Proof.* If  $A = B$ , then clearly  $X = A$  minimizes  $h(X)$ . Assume, therefore, that  $A \neq B$ . Ignoring the constraint  $X > 0$  for the moment, we see that any stationary point of  $h(X)$  must satisfy  $\nabla h(X) = 0$ . This condition translates into

$$\nabla h(X) = \left(\frac{X+A}{2}\right)^{-1} \frac{1}{2} + \left(\frac{X+B}{2}\right)^{-1} \frac{1}{2} - X^{-1} = 0 \implies B = XA^{-1}X.$$

The latter equation is a Riccati equation whose *unique* positive solution is  $X = A\sharp B$  [6, Prop 1.2.13]. It remains to show that the stationary point  $A\sharp B$  is actually a local minimum. Consider the Hessian

$$2\nabla^2 h(X) = X^{-1} \otimes X^{-1} - [(X+A)^{-1} \otimes (X+A)^{-1} + (X+B)^{-1} \otimes (X+B)^{-1}].$$

Writing  $P = (X+A)^{-1}$ ,  $Q = (X+B)^{-1}$ , and using  $\nabla h(X) = 0$  we obtain

$$2\nabla^2 h(X) = (Q \otimes P) + (P \otimes Q) > 0.$$

Thus,  $X = A\sharp B$  is a *strict* local minimum of  $h(X)$ . This local minimum is the global minimum since  $\nabla h(X) = 0$  has a unique positive solution and  $h$  goes to  $+\infty$  at the boundary. Equidistance follows easily from  $A\sharp B = B\sharp A$  and Proposition 2.1.  $\square$

**4.2. Geodesic convexity.** In this section, we show that  $\delta_S^2$  is jointly geodesically convex (hereafter ‘g-convex’), an important property also satisfied by  $\delta_R$ . Before proving our g-convexity result (Theorem 4.4), we recall two useful facts.

**Theorem 4.2** ([18]). *The MGM of  $A, B \in \mathbb{P}_n$  is given by the variational formula*

$$A\sharp B = \max\{X \in \mathbb{H}_n \mid \begin{bmatrix} A & X \\ X & B \end{bmatrix} \geq 0\}.$$

**Proposition 4.3** (Joint-concavity (see, e.g., [18])). *Let  $A, B, C, D > 0$ . Then,*

$$(4.4) \quad (A\sharp B) + (C\sharp D) \leq (A+C)\sharp(B+D).$$

**Theorem 4.4.** *The function  $\delta_S^2(X, Y)$  is jointly g-convex for  $X, Y > 0$ .*

*Proof.* It suffices to show that for  $X_1, X_2, Y_1, Y_2 > 0$  we have

$$(4.5) \quad \delta_S^2(X_1\sharp X_2, Y_1\sharp Y_2) \leq \frac{1}{2}\delta_S^2(X_1, Y_1) + \frac{1}{2}\delta_S^2(X_2, Y_2).$$

From Proposition 4.3 it follows that  $X_1\sharp X_2 + Y_1\sharp Y_2 \leq (X_1 + Y_1)\sharp(X_2 + Y_2)$ . Since  $\log \det$  is monotonic and determinants are multiplicative, it then follows that

$$\log \det \left( \frac{X_1\sharp X_2 + Y_1\sharp Y_2}{2} \right) \leq \log \det \left( \frac{(X_1 + Y_1)\sharp(X_2 + Y_2)}{2} \right),$$

which when combined with the identity

$$-\frac{1}{2} \log \det((X_1\sharp X_2)(Y_1\sharp Y_2)) = -\frac{1}{4} \log \det(X_1 Y_1) - \frac{1}{4} \log \det(X_2 Y_2)$$

yields inequality (4.5), establishing joint g-convexity.  $\square$

**4.3. Basic contraction results.** We now show that  $\delta_S$  and  $\delta_R$  exhibit similar contraction properties. The results are stated in terms of  $\delta_S^2$  or  $\delta_R$ , depending on which appears more elegant.

4.3.1. *Power-contraction.* The metric  $\delta_R$  satisfies (e.g., [6, Ex. 6.5.4])

$$(4.6) \quad \delta_R(A^t, B^t) \leq t\delta_R(A, B), \quad \text{for } A, B > 0 \text{ and } t \in [0, 1].$$

The S-divergence satisfies the same relation.

**Theorem 4.5.** *Let  $A, B > 0$ , and let  $t \in [0, 1]$ . Then,*

$$(4.7) \quad \delta_S^2(A^t, B^t) \leq t\delta_S^2(A, B).$$

*Moreover, if  $t \geq 1$ , then the inequality gets reversed.*

*Proof.* Recall that for  $t \in [0, 1]$ , the map  $X \mapsto X^t$  is operator concave. Thus,  $\frac{1}{2}(A^t + B^t) \leq \left(\frac{A+B}{2}\right)^t$ ; by monotonicity of the determinant it then follows that

$$\delta_S^2(A^t, B^t) = \log \frac{\det\left(\frac{1}{2}(A^t + B^t)\right)}{\det(A^t B^t)^{1/2}} \leq \log \frac{\det\left(\frac{1}{2}(A + B)\right)^t}{\det(AB)^{t/2}} = t\delta_S^2(A, B).$$

The reverse inequality for  $t \geq 1$  follows by considering  $\delta_S^2(A^{1/t}, B^{1/t})$ .  $\square$

4.3.2. *Contraction on geodesics.* The curve

$$(4.8) \quad \gamma(t) := A^{1/2}(A^{-1/2}BA^{-1/2})^t A^{1/2}, \quad \text{for } t \in [0, 1],$$

parameterizes the *unique* geodesic between the positive matrices  $A$  and  $B$  on the manifold  $(\mathbb{P}_n, \delta_R)$  [6, Thm. 6.1.6]. On this curve  $\delta_R$  satisfies

$$\delta_R(A, \gamma(t)) = t\delta_R(A, B), \quad t \in [0, 1].$$

The S-divergence satisfies a similar, albeit slightly weaker, result.

**Theorem 4.6.** *Let  $A, B > 0$ , and let  $\gamma(t)$  be defined by (4.8). Then,*

$$(4.9) \quad \delta_S^2(A, \gamma(t)) \leq t\delta_S^2(A, B), \quad 0 \leq t \leq 1.$$

*Proof.* The proof follows upon observing that

$$\delta_S^2(A, \gamma(t)) = \delta_S^2(I, (A^{-1/2}BA^{-1/2})^t) \stackrel{(4.7)}{\leq} t\delta_S^2(I, A^{-1/2}BA^{-1/2}) = t\delta_S^2(A, B). \quad \square$$

4.3.3. *A power-monotonicity property.* We show below that on matrix powers,  $\delta_S^2$  and  $\delta_R$  exhibit a similar monotonicity property reminiscent of a power-means inequality.

**Theorem 4.7.** *Let  $A, B > 0$ , and let scalars  $t$  and  $u$  satisfy  $1 \leq t \leq u < \infty$ . Then,*

$$(4.10) \quad t^{-1}\delta_R(A^t, B^t) \leq u^{-1}\delta_R(A^u, B^u)$$

$$(4.11) \quad t^{-1}\delta_S^2(A^t, B^t) \leq u^{-1}\delta_S^2(A^u, B^u).$$

To our knowledge, inequality (4.10) is also new. Before proving Theorem 4.7, we first state a “power-means” inequality on determinants (which also follows from the monotonicity theorem of [4]; see [22] for an alternative proof).

**Proposition 4.8.** *Let  $A, B > 0$ , and let scalars  $t, u$  satisfy  $1 \leq t \leq u < \infty$ . Then,*

$$(4.12) \quad \det^{1/t}\left(\frac{A^t + B^t}{2}\right) \leq \det^{1/u}\left(\frac{A^u + B^u}{2}\right).$$

*Proof of Theorem 4.7.* (i) Since  $\delta_R(X, Y) = \|\log E^\downarrow(XY^{-1})\|_F$ , we must show that

$$\frac{1}{t} \|\log E^\downarrow(A^t B^{-t})\|_F \leq \frac{1}{u} \|\log E^\downarrow(A^u B^{-u})\|_F.$$

Writing this inequality in terms of vectors of eigenvalues, we need to show that

$$(4.13) \quad \|\log \lambda^{1/t}(A^t B^{-t})\|_2 \leq \|\log \lambda^{1/u}(A^u B^{-u})\|_2.$$

This inequality follows readily from the log-majorization [5, Theorem IX.2.9]

$$\log \lambda^{1/t}(A^t B^{-t}) \prec \log \lambda^{1/u}(A^u B^{-u}),$$

upon applying the map  $x \mapsto \|x\|_2$ , which yields (4.13). We have in fact proved the more general result

$$\frac{1}{t} \|\log E^\downarrow(A^t B^{-t})\|_\Phi \leq \frac{1}{u} \|\log E^\downarrow(A^u B^{-u})\|_\Phi,$$

where  $\Phi$  is a symmetric gauge function (i.e., a permutation invariant absolute norm).

(ii) To prove (4.11) we must show that

$$\begin{aligned} & \frac{1}{t} \log \det((A^t + B^t)/2) - \frac{t}{2} \log \det(A^t B^t) \\ & \leq \frac{1}{u} \log \det((A^u + B^u)/2) - \frac{u}{2} \log \det(A^u B^u). \end{aligned}$$

This inequality is immediate from Proposition 4.8 and the monotonicity of log.  $\square$

**4.3.4. Contraction under translation.** The last basic contraction result that we prove is an analogue of the following shrinkage property [9, Prop. 1.6]:

$$(4.14) \quad \delta_R(A + X, A + Y) \leq \frac{\alpha}{\alpha + \beta} \delta_R(X, Y), \quad \text{for } A \geq 0 \text{ and } X, Y > 0,$$

where  $\alpha = \max\{\|X\|, \|Y\|\}$  and  $\beta = \lambda_{\min}(A)$ . This result plays a crucial role in deriving contractive maps for certain nonlinear matrix equations [19].

**Theorem 4.9.** *Let  $X, Y > 0$  and  $A \geq 0$ . Then the function*

$$(4.15) \quad g(A) := \delta_S^2(A + X, A + Y)$$

*is monotonically decreasing and convex in  $A$ .*

*Proof.* We must show that if  $A \leq B$ , then  $g(A) \geq g(B)$ . Equivalently, we show that the gradient  $\nabla_A g(A) \leq 0$ , which follows easily since

$$\nabla_A g(A) = \left( \frac{(A+X)+(A+Y)}{2} \right)^{-1} - \frac{1}{2} (A+X)^{-1} - \frac{1}{2} (A+Y)^{-1} \leq 0,$$

as the map  $X \mapsto X^{-1}$  is operator convex. It remains to prove convexity of  $g$ . Consider therefore its Hessian  $\nabla^2 g(A)$ . Let  $P = (A+X)^{-1}$ ,  $Q = (A+Y)^{-1}$ , so

$$\nabla^2 g(A) = \frac{1}{2} (P \otimes P + Q \otimes Q) - \left( \frac{P^{-1}+Q^{-1}}{2} \right)^{-1} \otimes \left( \frac{P^{-1}+Q^{-1}}{2} \right)^{-1}.$$

Using matrix convexity of  $X \mapsto X^{-1}$  we obtain

$$\nabla^2 g(A) \geq \frac{1}{2} (P \otimes P + Q \otimes Q) - \frac{P+Q}{2} \otimes \frac{P+Q}{2} = \frac{1}{2} (P-Q) \otimes (P-Q),$$

which is positive definite since by assumption  $P \geq Q$ .  $\square$

**Corollary 4.10.** *Let  $X, Y > 0$ ,  $A \geq 0$ , and  $\beta = \lambda_{\min}(A)$ . Then,*

$$(4.16) \quad \delta_S^2(A + X, A + Y) \leq \delta_S^2(\beta I + X, \beta I + Y) \leq \delta_S^2(X, Y).$$



**4.4. Conic contraction.** This section proves a compression property for  $\delta_S$ , which it shares with the well-known Hilbert and Thompson metrics on cones [20, Ch.2].

**Proposition 4.11.** *Let  $P \in \mathbb{C}^{n \times k}$  ( $k \leq n$ ) have full column rank. The function  $f : \mathbb{P}_n \rightarrow \mathbb{R} \equiv X \mapsto \log \det(P^*XP) - \log \det(X)$  is operator decreasing.*

*Proof.* It suffices to show that  $\nabla f(X) \leq 0$ . This amounts to establishing that

$$(4.17) \quad P(P^*XP)^{-1}P^* \leq X^{-1} \quad \Leftrightarrow \quad \begin{bmatrix} X^{-1} & P \\ P^* & P^*XP \end{bmatrix} \geq 0.$$

Inequality (4.17) follows once we note the factorization

$$\begin{bmatrix} X^{-1} & P \\ P^* & P^*XP \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & P^* \end{bmatrix} \begin{bmatrix} X^{-1} & I \\ I & X \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & P \end{bmatrix}. \quad \square$$

**Corollary 4.12.** *Let  $X, Y > 0$ . Let  $A = \left(\frac{X+Y}{2}\right)$ ,  $G = X\sharp Y$ , and  $P \in \mathbb{C}^{n \times k}$  ( $k \leq n$ ) have full column rank. Then,*

$$(4.18) \quad \frac{\det(P^*AP)}{\det(P^*GP)} \leq \frac{\det(A)}{\det(G)}.$$

*Proof.* Since  $A \geq G$ , it follows from Proposition 4.11 that

$$\log \det(P^*AP) - \log \det(A) \leq \log \det(P^*GP) - \log \det(G).$$

Rearranging, and using the fact that  $P^*AP \geq P^*GP$ , we obtain (4.18).  $\square$

**Theorem 4.13** ([1, Thm. 3]). *Let  $\Pi : \mathbb{P}_n \rightarrow \mathbb{P}_k$  be a positive linear map. Then,*

$$(4.19) \quad \Pi(A\sharp B) \leq \Pi(A)\sharp\Pi(B), \quad \text{for } A, B \in \mathbb{P}_n.$$

We are now ready to prove the main theorem of this section.

**Theorem 4.14.** *Let  $P \in \mathbb{C}^{n \times k}$  ( $k \leq n$ ) have full column rank. Then,*

$$(4.20) \quad \delta_S^2(P^*AP, P^*BP) \leq \delta_S^2(A, B), \quad \text{for } A, B \in \mathbb{P}_n.$$

*Proof.* We may equivalently show that

$$(4.21) \quad \frac{\det\left(\frac{P^*(A+B)P}{2}\right)}{\sqrt{\det(P^*AP)\det(P^*BP)}} \leq \frac{\det\left(\frac{A+B}{2}\right)}{\sqrt{\det(AB)}}.$$

But Theorem 4.13 asserts that  $P^*(A\sharp B)P \leq (P^*AP)\sharp(P^*BP)$ , whereby

$$\frac{1}{\sqrt{\det(P^*AP)\det(P^*BP)}} = \frac{1}{\det[(P^*AP)\sharp(P^*BP)]} \leq \frac{1}{\det(P^*(A\sharp B)P)}.$$

Consequently, an invocation of Corollary 4.12 concludes the argument.  $\square$

Theorem 4.14 relates  $\delta_S$  to the classical Hilbert and Thompson metrics on convex cones, which satisfy similar inequalities (for a wider class of order-preserving maps [20]); hence the name “conic contraction”. Theorem 4.14 also extends to  $\delta_R$ , as noted in Corollary 4.16, which follows from Theorem 4.15. We believe that this theorem must exist in the literature—see [22, Thm. 4.17] for a proof.

**Theorem 4.15.** *Let  $A, B \in \mathbb{P}_n$  and  $P \in \mathbb{C}^{n \times k}$  ( $k \leq n$ ) have full column rank. Then,*

$$(4.22) \quad \lambda_j^\downarrow(P^*AP(P^*BP)^{-1}) \leq \lambda_j^\downarrow(AB^{-1}), \quad \text{for } 1 \leq j \leq k.$$

**Corollary 4.16.** *Let  $P \in \mathbb{C}^{n \times k}$  ( $k \leq n$ ) have full column rank. Then,*

$$(4.23) \quad \delta_{\Phi}(P^*AP, P^*BP) \leq \delta_{\Phi}(A, B) := \|\log(B^{-1/2}AB^{-1/2})\|_{\Phi},$$

where  $\Phi$  is any symmetric gauge function.

We conclude by noting a bi-Lipschitz-like inequality between  $\delta_S$  and  $\delta_R$ .

**Theorem 4.17** ([22]). *Let  $A, B \in \mathbb{P}_n$ . Let  $\delta_T(A, B) = \|\log(B^{-1/2}AB^{-1/2})\|$  denote the Thompson-part metric [20]. Then, we have the bounds*

$$(4.24) \quad 8\delta_S^2(A, B) \leq \delta_R^2(A, B) \leq 2\delta_T(A, B)(\delta_S^2(A, B) + n \log 2).$$

## REFERENCES

- [1] T. Ando, *Concavity of certain maps on positive definite matrices and applications to Hadamard products*, Linear Algebra Appl. **26** (1979), 203–241, DOI 10.1016/0024-3795(79)90179-4. MR535686 (80f:15023)
- [2] Arindam Banerjee, Srujana Merugu, Inderjit Dhillon, and Joydeep Ghosh, *Clustering with Bregman divergences*, Proceedings of the Fourth SIAM International Conference on Data Mining, SIAM, Philadelphia, PA, 2004, pp. 234–245. MR2388444
- [3] Christian Berg, Jens Peter Reus Christensen, and Paul Ressel, *Harmonic analysis on semi-groups*, Graduate Texts in Mathematics, vol. 100, Springer-Verlag, New York, 1984. Theory of positive definite and related functions. MR747302 (86b:43001)
- [4] K. V. Bhagwat and R. Subramanian, *Inequalities between means of positive operators*, Math. Proc. Cambridge Philos. Soc. **83** (1978), no. 3, 393–401. MR0467372 (57 #7231)
- [5] Rajendra Bhatia, *Matrix analysis*, Graduate Texts in Mathematics, vol. 169, Springer-Verlag, New York, 1997. MR1477662 (98i:15003)
- [6] Rajendra Bhatia, *Positive definite matrices*, Princeton Series in Applied Mathematics, Princeton University Press, Princeton, NJ, 2007. MR2284176 (2007k:15005)
- [7] Rajendra Bhatia and John Holbrook, *Riemannian geometry and matrix geometric means*, Linear Algebra Appl. **413** (2006), no. 2–3, 594–618, DOI 10.1016/j.laa.2005.08.025. MR2198952 (2007c:15030)
- [8] A. Bhattacharyya, *On a measure of divergence between two statistical populations defined by their probability distributions*, Bull. Calcutta Math. Soc. **35** (1943), 99–109. MR0010358 (6,7b)
- [9] Philippe Bougerol, *Kalman filtering with random coefficients and contractions*, SIAM J. Control Optim. **31** (1993), no. 4, 942–959, DOI 10.1137/0331041. MR1227540 (94h:93078)
- [10] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, Cambridge, 2004. MR2061575 (2005d:90002)
- [11] Martin R. Bridson and André Haefliger, *Metric spaces of non-positive curvature*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 319, Springer-Verlag, Berlin, 1999. MR1744486 (2000k:53038)
- [12] Zeineb Chebbi and Maher Moakher, *Means of Hermitian positive-definite matrices based on the log-determinant  $\alpha$ -divergence function*, Linear Algebra Appl. **436** (2012), no. 7, 1872–1889, DOI 10.1016/j.laa.2011.12.003. MR2889964
- [13] Pengwen Chen, Yunmei Chen, and Murali Rao, *Metrics defined by Bregman divergences*, Commun. Math. Sci. **6** (2008), no. 4, 915–926. MR2511699 (2010g:94038)
- [14] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos, *Efficient Similarity Search for Covariance Matrices via the Jensen-Bregman LogDet Divergence*, Int. Conf. Computer Vision (ICCV), Nov. 2011, pp. 2399–2406.
- [15] Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, Wiley Series in Telecommunications, John Wiley & Sons, Inc., New York, 1991. A Wiley-Interscience Publication. MR1122806 (92g:94001)
- [16] Miroslav Fiedler, *Bounds for the determinant of the sum of hermitian matrices*, Proc. Amer. Math. Soc. **30** (1971), 27–31. MR0286814 (44 #4021)
- [17] S. G. Gindikin, *Invariant generalized functions in homogeneous domains* (Russian), Funkcional. Anal. i Priložen. **9** (1975), no. 1, 56–58, English transl., Functional Anal. Appl. **9** (1975), no. 1, 50–52. MR0377423 (51 #13595)

- [18] Fumio Kubo and Tsuyoshi Ando, *Means of positive linear operators*, Math. Ann. **246** (1979/80), no. 3, 205–224, DOI 10.1007/BF01371042. MR563399 (84d:47028)
- [19] Hosoo Lee and Yongdo Lim, *Invariant metrics, contractions and nonlinear matrix equations*, Nonlinearity **21** (2008), no. 4, 857–878, DOI 10.1088/0951-7715/21/4/011. MR2399829 (2009j:37084)
- [20] Bas Lemmens and Roger Nussbaum, *Nonlinear Perron-Frobenius theory*, Cambridge Tracts in Mathematics, vol. 189, Cambridge University Press, Cambridge, 2012. MR2953648
- [21] Yurii Nesterov and Arkadii Nemirovskii, *Interior-point polynomial algorithms in convex programming*, SIAM Studies in Applied Mathematics, vol. 13, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994. MR1258086 (94m:90005)
- [22] S. Sra, *Positive definite matrices and the S-Divergence*, arXiv:1110.1773 (2011).
- [23] S. Sra, *A new metric on the manifold of kernel matrices with application to matrix geometric means*, Adv. Neural Inf. Proc. Syst., December 2012.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MASSACHUSETTS 02139  
E-mail address: `suvrit@mit.edu`