PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 00, Number 0, Pages 000–000
S 0002-9939(XX)0000-0

POSITIVE DEFINITE MATRICES AND THE S-DIVERGENCE

SUVRIT SRA

(Communicated by)

Abstract. Hermitian positive definite (hpd) matrices form a self-dual convex cone whose interior is a Riemannian manifold of nonpositive curvature. The manifold view is endowed with a geodesically convex distance function but the convex view is not. Drawing motivation from convex optimization, we introduce the *S-Divergence*, a "distance-like" function on the convex cone of hpd matrices. We study basic properties of the S-divergence that connect it intimately with the Riemannian distance. In particular, we show that (i) its square-root is a distance; and that (ii) it exhibits several (nonpositive-curvature-like) properties akin to the Riemannian distance.

1. Introduction

Hermitian positive definite (hpd) matrices form a manifold of nonpositive curvature [11, Ch.10], [6, Ch.6], whose closure is a self-dual convex cone. The manifold is important in various applications [21], while the conic view is fundamental to convex optimization [20, 10] and nonlinear Perron-Frobenius theory [18], for instance.

The manifold view comes with a natural distance function, while the conic view does not. Drawing motivation from convex optimization we introduce on the hpd cone a distance-like function: the *S-Divergence*. We prove several results connecting the S-Divergence to the Riemannian distance. Our main result is that the square-root of this divergence is actually a distance; our remaining results explore geometric and analytic properties common to the S-Divergence and the Riemannian distance.

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

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

The set of positive definite (henceforth *positive*) matrices in $\mathbb{H}_n$ is denoted by $\mathbb{P}_n$. We say A is *positive semidefinite* if $\langle x,\, Ax\rangle \geq 0$ for all x and write $A \geq 0$. The inequality $A \geq B$ means $A - B \geq 0$. The *Frobenius norm* of a matrix X is $\|X\|_{\mathrm{F}} := \sqrt{\mathrm{tr}(X^*X)}$, while $\|X\|$ denotes the operator norm. Let f be an analytic function on $\mathbb{C}$, and let A have the eigendecomposition $A = U\Lambda U^*$ with unitary U, then $f(A) = Uf(\Lambda)U^*$ with $f(\Lambda)$ equal to the diagonal matrix $\mathrm{Diag}[f(\lambda_1), \ldots, f(\lambda_n)]$.

The set $\mathbb{P}_n$ is a well-studied differentiable Riemannian manifold, with the Riemannian metric given by the differential form $ds = \|A^{-1/2}dAA^{-1/2}\|_{\mathrm{F}}$. This metric

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

This work was done when 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*—see [23].

induces the *Riemannian distance* (see e.g., [6, Ch. 6]):

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

The focus of this paper is on a complement to (1.2), namely, the *S-Divergence*:[1]

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

This divergence was proposed as a numerical alternative to the Riemannian distance δ_R in [14]—there, it was used in an application to image-search, primarily for its empirical benefits. This initial empirical success of S-Divergence motivated us to investigate δ_S more closely in this paper.

Contributions. The present paper goes substantially beyond our initial conference version [23]. The main differences are: (i) Theorems 4.1, 4.5, 4.6, 4.7, and 4.9, which establish several new geometric and analytic similarities between δ_S and δ_R; (ii) the joint geodesic convexity of δ_S^2 (Prop. 4.3, Theorem 4.4); and (iii) new "conic" contraction results for δ_S and δ_R that uncover properties akin to those exhibited Hilbert's projective metric and Thompson's part metric [18] (Prop. 4.11, Corollary 4.12, Theorem 4.14, Theorem 4.15, and Corollary 4.16).

Related work. While our paper was under preparation (in 2011), we became aware of a concurrent paper of Chebbi and Moakher (CM) [12] who considered a single parameter family of divergences that generalize (1.3). Our work differs from CM in several key aspects. (1) CM prove δ_S to be a distance for commuting matrices only. We note in passing that the commuting case is essentially equivalent to the scalar case. The noncommuting case is much harder, and was also conjectured by [12]. We prove the noncommuting case too, and note that our consideration of it was agnostic of CM [12]. (2) We establish several theorems that uncover geometric and analytic similarities between δ_S^2 and δ_R. (3) A question closely related to δ_S being a distance is whether the matrix $[\det(X_i+X_j)^{-\beta}]_{i,j=1}^m$ is positive semidefinite for arbitrary matrices $X_1,\ldots,X_m \in \mathbb{P}_n$, every integer $m \geq 1$, and every scalar $\beta \geq 0$. CM considered special cases of this question. We provide a complete characterization of β necessary and sufficient for the above matrix to be semidefinite.

2. The S-Divergence

Consider a differentiable strictly convex function $f : \mathbb{R} \to \mathbb{R}$; 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:*[2]

$$D_f(x,y) := f(x) - f(y) - f'(y)(x-y). \tag{2.1}$$

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

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

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 the squared *Frobenius norm* $\frac{1}{2}\|X-Y\|_{\mathrm{F}}^2$. If

[1] It is a divergence because although nonnegative, definite, and symmetric, it is *not* a metric.

[2] 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].

$f(x) = x \log x - x$ on $(0, \infty)$, then $\operatorname{tr} f(X) = \operatorname{tr}(X \log X - X)$ and (2.2) yields the (unnormalized) *von Neumann Divergence* of quantum information theory.

The asymmetry of Bregman divergences can be sometimes undesirable. This has led researchers to consider symmetric divergences, among which the most popular is the "*Jensen-Shannon / Bregman*" divergence

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

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

$$S_f(X, Y) \quad = \quad \tfrac{1}{2}\big(\operatorname{tr} f(X) + \operatorname{tr} f(Y)\big) - \operatorname{tr} f\big(\tfrac{X+Y}{2}\big). \tag{2.4}$$

The S-Divergence (1.3) can be obtained from (2.4) by setting $f(x) = -\log x$, so that $f(X) = -\log\det(X)$, the barrier function for the positive definite cone [20]. The S-Divergence may be 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 may be easily verified.

Proposition 2.1. *Let $\lambda(X)$ be the vector of eigenvalues of X, and $\operatorname{Eig}(X)$ the diagonal matrix with $\lambda(X)$ on its diagonal. 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 δ_S distance

In this section we present our main result: *the square-root δ_S of the S-Divergence is actually a distance*. Previous authors [12, 14] conjectured this result; both appealed to classical ideas from harmonic analysis [3, Ch. 3] to establish the commutative case. But the noncommutative case requires a different approach, as we show below.

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

The proof of Theorem 3.1 depends on several results, some of which we prove below.

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

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

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])**.** *Let $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be negative definite. There is a Hilbert space $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ and a mapping $x \mapsto \varphi(x)$ from $\mathcal{X} \to \mathcal{H}$ such that*

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

Moreover, negative definiteness of ψ is necessary for such a mapping to exist.

Theorem 3.3 helps prove the triangle inequality for the scalar case.[3]

Lemma 3.4. *Let δ_s be the scalar version of δ_S, i.e.,*

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

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

[3]The idea of invoking Schoenberg's theorem (Theorem 3.3) for for establishing the commutative case (which actually is essentially just the scalar case of Lemma 3.4) was also used by [12]; we arrived at it independently following the classic route of harmonic analysis [3, Ch. 3].

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 $\beta > 0$ and $x, y > 0$. This in turn follows from the inner-product representation

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

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

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

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

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

$$\delta_S(X, Y) \leq \delta_S(X, Z) + \delta_S(Y, Z) \tag{3.4}$$

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

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

Theorem 3.7 ([5, VI.7])**.** *Let $A, B > 0$. Let $\lambda^{\downarrow}(X)$ denote the vector of eigenvalues of X arranged in decreasing order; define $\lambda^{\uparrow}(X)$ likewise. Then,*

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

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

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

Proof. In (3.5), dividing by $2^n \sqrt{\det(A)\det(B)}$ we obtain

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

Since determinants are invariant to permutation of eigenvalues, we can rearrange the leftmost and rightmost terms so that upon taking logarithms, (3.6) follows. □

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*

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

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

Proof. (Theorem 3.1). We need to show that δ_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 Prop. 2.1), proving the triangle inequality reduces to showing that

$$\delta_S(I, D) \leq \delta_S(I, Z) + \delta_S(D, Z). \tag{3.8}$$

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

$$\delta_S(I, D^{\downarrow}) \le \delta_S(I, \mathrm{Eig}^{\downarrow}(Z)) + \delta_S(D^{\downarrow}, \mathrm{Eig}^{\downarrow}(Z)). \tag{3.9}$$

Prop. 2.1(i) implies that $\delta_S(I, D) = \delta_S(I, D^{\downarrow})$ and $\delta_S(I, Z) = \delta_S(I, \mathrm{Eig}^{\downarrow}(Z))$, while Corollary 3.8 shows that $\delta_S(D^{\downarrow}, \mathrm{Eig}^{\downarrow}(Z)) \le \delta_S(D, Z)$. Combining these inequalities, we immediately obtain (3.8). □

We now turn our attention to a related connection that enjoys importance in some applications: kernel functions arising from δ_S.

3.1. **Hilbert space embedding.** Since δ_S is a metric, which for scalars embeds isometrically into Hilbert space (Lemma 3.4), it is natural to ask whether $\delta_S(X, Y)$ also admits such an embedding. But as we already noted, such an embedding does not exist. Theorem 3.3 implies that a Hilbert space embedding exists if and only if $\delta_S^2(X, Y)$ is a negative definite kernel; equivalently, iff 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 $\beta > 0$. It suffices to check whether the matrix

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

is positive definite for every $m \ge 1$ and arbitrary positive matrices $X_1, \ldots, X_m \in \mathbb{P}_n$. Unfortunately, a quick numerical experiment reveals that H_β can be indefinite.

This leads us to the weaker question: *for what choices of β is $H_\beta \ge 0$?*

Theorem 3.10 answers this question for H_β formed from symmetric real positive definite matrices.

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

$$\beta \in \left\{\tfrac{j}{2} : j \in \mathbb{N},\ \text{and } 1 \le j \le (n-1)\right\} \cup \left\{\gamma : \gamma \in \mathbb{R}, \text{and } \gamma > \tfrac{1}{2}(n-1)\right\}. \tag{3.11}$$

Proof. Please refer to the longer version of this paper [22]. □

Theorem 3.10 shows that $e^{-\beta\delta_S^2}$ is not always a kernel, while for commuting matrices $e^{-\beta\delta_S^2}$ is always a positive definite kernel. This raises the following:

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 δ_R

4.1. **Geometric mean.** We begin by studying an object that connects δ_R and δ_S^2 most intimately: the matrix geometric mean. For positive matrices A and B, the *matrix geometric mean* (MGM) is denoted by $A\sharp B$, and is given by the formula

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

The MGM (4.1) has a host of attractive properties—see for instance the classic paper [1]. The following variational characterization is important [7]:

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

Surprisingly, the MGM enjoys a similar characterization even under δ_S^2.

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

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

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

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

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

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

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 as $\nabla h(X) = 0$ has a unique positive solution and h goes to $+\infty$ at the boundary. Equidistance follows easily from $A\sharp B = B\sharp A$ and Prop. 2.1. □

4.2. **Geodesic convexity.** The above derivation concludes optimality of the MGM from first principles. In this section, we show that δ_S^2 *is* actually jointly geodesically convex, hereafter 'g-convex', (see (4.5)), a property also satisfied by δ_R.

Before proving Theorem 4.4, we recall two results; the first implies the second[4].

Theorem 4.2 ([16])**.** *The GM of $A, B \in \mathbb{P}_n$ is given by the variational formula*

$$A\sharp B = \max\left\{X \in \mathbb{H}_n \mid \left[\begin{smallmatrix} A & X \\ X & B \end{smallmatrix}\right] \geq 0\right\}.$$

Proposition 4.3 (Joint-concavity (see e.g. [16]))**.** *Let $A, B, C, D > 0$. Then,*

$$(A\sharp B) + (C\sharp D) \leq (A+C)\sharp(B+D). \tag{4.4}$$

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

Proof. Since δ_S^2 is continuous, it suffices to show that for $X_1, X_2, Y_1, Y_2 > 0$,

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

From Prop. 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

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

which when combined with the identity

$$-\tfrac{1}{2}\log\det\big((X_1\sharp X_2)(Y_1\sharp Y_2)\big) = -\tfrac{1}{4}\log\det(X_1Y_1) - \tfrac{1}{4}\log\det(X_2Y_2)$$

yields inequality (4.5), establishing joint g-convexity. □

4.3. **Basic contraction results.** In this section we show that δ_S and δ_R share several contraction properties. We state our results either in terms of δ_S^2 or of δ_S, depending on whichever appears more elegant.

[4]It is a minor curiosity to note that the mixed-mean inequality for matrix geometric and arithmetic means proved in [19, Thm. 2] is a special case of Prop. 4.3.

4.3.1. *Power-contraction.* The metric δ_R satisfies (e.g., [6, Exercise 6.5.4])

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

Theorem 4.5 shows that S-Divergence satisfies the same relation.

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

$$\delta_S^2(A^t, B^t) \leq t\delta_S^2(A, B). \tag{4.7}$$

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

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

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

4.3.2. *Contraction on geodesics.* The curve

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

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 δ_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 $\gamma(t)$ be defined by* (4.8)*. Then,*

$$\delta_S^2(A, \gamma(t)) \leq t\delta_S^2(A, B), \qquad 0 \leq t \leq 1. \tag{4.9}$$

Proof. The proof follows upon observing that

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

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

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

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

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

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

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

$$\det{}^{1/t}\Big(\tfrac{A^t+B^t}{2}\Big) \quad \leq \quad \det{}^{1/u}\Big(\tfrac{A^u+B^u}{2}\Big). \tag{4.12}$$

Proof (Theorem 4.7). (i): Note that $\delta_R(X, Y) = \|\log E^{\downarrow}(XY^{-1})\|_{\mathrm{F}}$. We must show

$$\tfrac{1}{t}\|\log E^{\downarrow}(A^tB^{-t})\|_{\mathrm{F}} \leq \tfrac{1}{u}\|\log E^{\downarrow}(A^uB^{-u})\|_{\mathrm{F}}.$$

Equivalently, for vectors of eigenvalues we may prove

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

The log-majorization relation stated in [5, Theorem IX.2.9] says

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

to which we apply the map $x \mapsto \|x\|_2$ immediately obtaining (4.13). Notice, that we have in fact proved the more general result

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

where Φ is a symmetric gauge function (a permutation invariant absolute norm).

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

$$\tfrac{1}{t}\log\det\big((A^t+B^t)/2\big) - \tfrac{t}{2}\log\det(A^tB^t) \leq \tfrac{1}{u}\log\det\big((A^u+B^u)/2\big) - \tfrac{u}{2}\log\det(A^uB^u).$$

But this inequality is immediate from Prop. 4.8 and the monotonicity of log. □

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]:

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

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 [17].

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

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

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

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

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 $P = (A+X)^{-1}$, $Q = (B+X)^{-1}$, so

$$\nabla^2 g(A) = \tfrac{1}{2}(P\otimes P + Q\otimes Q) - \Big(\tfrac{P^{-1}+Q^{-1}}{2}\Big)^{-1} \otimes \Big(\tfrac{P^{-1}+Q^{-1}}{2}\Big)^{-1}.$$

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

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

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

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

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

4.4. **Conic contraction.** We establish below (Theorem 4.14) a compression property for the S-Divergence, which it shares with the well-known Hilbert and Thompson metrics on convex cones [18, Ch.2]. We start with some preparatory results.

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

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

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

Inequality (4.17) follows once we note the factorization

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

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

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

Proof. Since $A \geq G$, it follows from Prop. 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 $\Pi : \mathbb{P}_n \to \mathbb{P}_k$ be a positive linear map. Then,*

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

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,*

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

Proof. We may equivalently show that

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

But Prop. 4.13 shows $P^*(A\sharp B)P \leq (P^*AP)\sharp(P^*BP)$, which implies that

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

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

Theorem 4.14 relates δ_S to the classical Hilbert and Thompson metrics on convex cones, which also satisfy similar results (for the wider class of order-preserving subhomogenous maps [18]); this explains the name "conic contraction." Theorem 4.14 also extends to δ_R, as noted in Corollary 4.16, which itself follows from Theorem 4.15 (we believe that this theorem must exist in the literature—see [22] for our proof).

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

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

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

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

where Φ is any symmetric gauge function.

We conclude by noting a bi-Lipschitz-like inequality between δ_S and δ_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 [18]. Then, we have the following bounds*

$$8\delta_S^2(A, B) \leq \delta_R^2(A, B) \leq 2\delta_T(A, B)\big(\delta_S^2(A, B) + n\log 2\big). \tag{4.24}$$

References

1. T. Ando, *Concavity of certain maps on positive definite matrices and applications to Hadamard products*, Lin. Alg. Appl. **26** (1979), no. 0, 203–241.
2. A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, *Clustering with Bregman Divergences*, SIAM Int. Conf. on Data Mining (Florida), April 2004.
3. C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic analysis on semigroups: theory of positive definite and related functions*, GTM, vol. 100, Springer, 1984.
4. K. V. Bhagwat and R. Subramanian, *Inequalities between means of positive operators*, Math. Proc. Camb. Phil. Soc. **83** (1978), no. 3, 393–401.
5. R. Bhatia, *Matrix Analysis*, Springer, 1997.
6. ———, *Positive Definite Matrices*, Princeton University Press, 2007.
7. R. Bhatia and J. A. R. Holbrook, *Riemannian geometry and matrix geometric means*, Linear Algebra Appl. **413** (2006), 594–618.
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.
9. P. Bougerol, *Kalman Filtering with Random Coefficients and Contractions*, SIAM J. Control Optim. **31** (1993), no. 4, 942–959.
10. S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, March 2004.
11. M. R. Bridson and A. Haeflinger, *Metric Spaces of Non-Positive Curvature*, Springer, 1999.
12. Z. Chebbi and M. Moahker, *Means of hermitian positive-definite matrices based on the log-determinant α-divergence function*, Lin. Alg. Appl. **436** (2012), 1872–1889.
13. P. Chen, Y. Chen, and M. Rao, *Metrics defined by Bregman divergences: Part I*, Communications on Mathematical Sciences **6** (2008), 9915–926.
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. T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley-Interscience, 1991.
16. F. Kubo and T. Ando, *Means of positive linear operators*, Mathematische Annalen **246** (1980), 205–224.
17. H. Lee and Y. Lim, *Invariant metrics, contractions and nonlinear matrix equations*, Nonlinearity **21** (2008), 857–878.
18. B. Lemmens and R. Nussbaum, *Nonlinear Perron-Frobenius Theory*, Cambridge Univ. Press, 2012.
19. B. Mond and J.E. Pečarić., *A mixed arithmetic-mean-harmonic-mean matrix inequality*, Lin. Alg. Appl. **237** (1996), 449–454.
20. Yu. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, 1987.
21. F. Nielsen and R. Bhatia (eds.), *Matrix Information Geometry*, Springer, 2013.
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, MA
E-mail address: suvrit@mit.edu