SIAM J. OPTIM.
Vol. 25, No. 1, pp. 713–739

CONIC GEOMETRIC OPTIMIZATION ON THE MANIFOLD OF POSITIVE DEFINITE MATRICES*

SUVRIT SRA† AND RESHAD HOSSEINI‡

Abstract. We develop *geometric optimization* on the manifold of Hermitian positive definite (HPD) matrices. In particular, we consider optimizing two types of cost functions: (i) geodesically convex (g-convex) and (ii) log-nonexpansive (LN). G-convex functions are nonconvex in the usual Euclidean sense but convex along the manifold and thus allow global optimization. LN functions may fail to be even g-convex but still remain globally optimizable due to their special structure. We develop theoretical tools to recognize and generate g-convex functions as well as cone theoretic fixed-point optimization algorithms. We illustrate our techniques by applying them to maximum-likelihood parameter estimation for elliptically contoured distributions (a rich class that substantially generalizes the multivariate normal distribution). We compare our fixed-point algorithms with sophisticated manifold optimization methods and obtain notable speedups.

Key words. manifold optimization, geometric optimization, geodesic convexity, log-nonexpansive, conic fixed-point theory, Thompson metric, vector transport, Riemannian BFGS

AMS subject classifications. 65K10, 49Q99, 26A51, 47H10

DOI. 10.1137/140978168

1. Introduction.

Hermitian positive definite (HPD) matrices possess a remarkably rich geometry that is a cornerstone of modern convex optimization [38] and convex geometry [9]. In particular, HPD matrices form a convex cone, the strict interior of which is a differentiable Riemannian manifold which is also a prototypical CAT(0) space (i.e., a metric space of nonpositive curvature [12]). This rich structure enables "geometric optimization" on the set of HPD matrices—enabling us to solve certain problems that may be nonconvex in the Euclidean sense but are convex in the manifold sense (see section 2 or [49]) or, failing that, still have enough geometry (see section 4) to admit efficient optimization.

This paper formally develops *conic geometric optimization*[1] for HPD matrices. We present key results that help us recognize geodesic convexity (g-convexity); we also present sufficient conditions that place even several non–geodesically convex functions within the grasp of geometric optimization.

Motivation. We begin by noting that the widely studied class of *geometric programs* ultimately reduces to conic geometric optimization on 1×1 HPD matrices (i.e., positive scalars; see Remark 2.10). Geometric programming has enjoyed great success across a spectrum of applications—see, e.g., the survey of [11]; we hope this paper

*Received by the editors July 18, 2014; accepted for publication (in revised form) January 22, 2015; published electronically March 31, 2015.
http://www.siam.org/journals/siopt/25-1/97816.html
†Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA (suvrit@gmail.com). A part of this work was done while the first author was at Carnegie Mellon University, Pittsburgh (Machine Learning Department) on leave from the MPI for Intelligent Systems, Tübingen, Germany. A preliminary fraction of this paper was presented at the 2013 *Neural Information Processing Systems* conference.
‡School of ECE, College of Engineering, University of Tehran, Tehran, Iran (reshadh@gmail.com).
[1] To the best of our knowledge the name "geometric optimization" has not been previously attached to g-convex and cone theoretic HPD matrix optimization, though several scattered examples do exist. Our theorems offer a formal starting point for recognizing HPD geometric optimization problems.

helps conic geometric optimization gain wider exposure.

Perhaps the best known conic geometric optimization problem is computation of the Karcher (Fréchet) mean of a set of HPD matrices, a topic that has attracted great attention within matrix theory [7, 48, 8, 25], computer vision [17], radar imaging [41, Part II], and medical imaging [52, 16]; we refer the reader to the recent book [41] for additional applications and references. Another basic geometric optimization problem arises as a subroutine in image search and matrix clustering [18].

Conic geometric optimization problems also occur in several other areas such as statistics (covariance shrinkage) [15], nonlinear matrix equations [31], Markov decision processes, and more broadly in the fascinating areas of nonlinear Perron–Frobenius theory [32].

As a concrete illustration of our ideas, we discuss the task of *maximum-likelihood estimate* (mle) for *elliptically contoured distributions* (ECDs) [13, 21, 37]; see section 5. We use ECDs to illustrate our theory not only because of their instructive value but also because of their importance in a variety of applications [42].

Outline. The main focus of this paper is on recognizing and constructing certain structured nonconvex functions of HPD matrices. In particular, section 2 studies the class of geodesically convex functions, while section 4 introduces "log-nonexpansive" (LN) functions. We present a limited-memory BFGS algorithm in section 3, where we also present a derivation for the *parallel transport* which we could not find elsewhere in the literature. Even though manifold optimization algorithms apply to both classes of functions, for LN functions we advance fixed-point theory and algorithms separately in section 4. We present an application of geometric optimization in section 5, where we consider statistical inference with ECDs. Numerical results are the subject of section 6.

2. Geodesic convexity for HPD matrices.

Geodesic convexity (g-convexity) is a classical concept in geometry and analysis; it is used extensively in the study of Hadamard manifolds and metric spaces of *nonpositive curvature* [12, 43], i.e., metric spaces having a g-convex distance function. The concept of g-convexity has been previously explored in nonlinear optimization [45], but its importance and applicability in statistical applications and optimization has only recently gained more attention [49, 51]. It is worth noting that geometric programming [11] ultimately relies on "geometric-mean" convexity [40], i.e., $f(x^\alpha y^{1-\alpha}) \leq [f(x)]^\alpha [f(y)]^{1-\alpha}$, which is nothing but logarithmic g-convexity on 1×1 HPD matrices (positive scalars).

To introduce g-convexity on $n \times n$ HPD matrices we begin by recalling some key definitions; see [12, 43] for extensive details.

DEFINITION 2.1 (g-convex sets). *Let $\mathcal{M}$ be a d-dimensional connected C^2 Riemannian manifold. A set $\mathcal{X} \subset \mathcal{M}$ is called* geodesically convex *if any two points of $\mathcal{X}$ are joined by a geodesic lying in $\mathcal{X}$. That is, if $x, y \in \mathcal{X}$, then there exists a shortest path $\gamma : [0,1] \to \mathcal{X}$ such that $\gamma(0) = x$ and $\gamma(1) = y$.*

DEFINITION 2.2 (g-convex functions). *Let $\mathcal{X} \subset \mathcal{M}$ be a g-convex set. A function $\phi : \mathcal{X} \to \mathbb{R}$ is called* geodesically convex *if for any $x, y \in \mathcal{X}$, we have the inequality*

$$\phi(\gamma(t)) \leq (1-t)\phi(\gamma(0)) + t\phi(\gamma(1)) = (1-t)\phi(x) + t\phi(y), \tag{2.1}$$

where $\gamma(\cdot)$ is the geodesic $\gamma : [0,1] \to \mathcal{X}$ with $\gamma(0) = x$ and $\gamma(1) = y$.

2.1. Recognizing g-convexity.

Unlike scalar g-convexity, for matrices, recognizing g-convexity is not so easy. Indeed, for scalars, a function $f : \mathbb{R}_{++} \to \mathbb{R}$ is log-g-convex (and hence g-convex) if and only if $\log \circ f \circ \exp$ is convex. A similar characterization does not seem to exist for HPD matrices, primarily due to the noncommutativity of matrix multiplication. We develop some theory below for helping to recognize and construct g-convex functions.

To define g-convex functions on HPD matrices recall that $\mathbb{P}_d$ is a differentiable Riemannian manifold where geodesics between points are available in closed form. Indeed, the tangent space to $\mathbb{P}_d$ at any point can be identified with the set of Hermitian matrices, and the inner product on this space leads to a Riemannian metric on $\mathbb{P}_d$. At any point $A \in \mathbb{P}_d$, this metric is given by the differential form $ds = \|A^{-1/2} dA A^{-1/2}\|_{\mathrm{F}}$; for $A, B \in \mathbb{P}_d$ there is a unique geodesic path [6, Thm. 6.1.6]

$$\gamma(t) = A \#_t B := A^{1/2}(A^{-1/2} B A^{-1/2})^t A^{1/2}, \quad t \in [0,1]. \tag{2.2}$$

The midpoint of this path, namely, $A\#_{1/2}B$, is called the *matrix geometric mean*, which is an object of great interest [6, 7, 25, 8]; we drop the 1/2 and denote it simply by $A\#B$. Starting from the geodesic (2.2), many g-convex functions can be constructed by extending monotonic convex functions to matrices. To that end, first recall the fundamental operator inequality [2] (where $\preceq$ denotes the Löwner partial order):

$$A \#_t B \preceq (1-t)A + tB. \tag{2.3}$$

Theorem 2.3 uses the operator inequality (2.3) to construct "tracial" g-convex functions.

THEOREM 2.3. *Let $h : \mathbb{R}_+ \to \mathbb{R}$ be monotonically increasing and convex; let $\lambda : \mathbb{P}_n \to \mathbb{R}^n_+$ denote the eigenvalue map and $\lambda^\downarrow(\cdot)$ its decreasingly sorted version. Then, $\sum_{j=1}^k h(\lambda_j^\downarrow(\cdot))$ is g-convex for each $1 \le k \le n$.*

Proof. It suffices to establish midpoint convexity. Inequality (2.3) implies that

$$\lambda_j(A\#B) \le \lambda_j\left(\tfrac{A+B}{2}\right) \quad \text{for} \quad 1 \le j \le n.$$

Since h is monotonic, for $1 \le k \le n$ it follows that

$$\sum\nolimits_{j=1}^k h(\lambda_j^\downarrow(A\#B)) \le \sum\nolimits_{j=1}^k h(\lambda_j^\downarrow\left(\tfrac{A+B}{2}\right)). \tag{2.4}$$

Lidskii's theorem [5, Thm. III.4.1] yields the majorization $\lambda^\downarrow\left(\frac{A+B}{2}\right) \prec \frac{\lambda^\downarrow(A)+\lambda^\downarrow(B)}{2}$, which combined with a celebrated result of [23][2] and convexity of h yields

$$\sum_{j=1}^k h(\lambda_j^\downarrow\left(\tfrac{A+B}{2}\right)) \le \sum_{j=1}^k h\left(\tfrac{\lambda_j^\downarrow(A)+\lambda_j^\downarrow(B)}{2}\right) \le \tfrac{1}{2}\sum_{j=1}^k h(\lambda_j^\downarrow(A)) + \tfrac{1}{2}\sum_{j=1}^k h(\lambda_j^\downarrow(B)).$$

Now invoke inequality (2.4) to conclude that $\sum_{j=1}^k h(\lambda_j^\downarrow(\cdot))$ is g-convex. ∎

Example 2.4. Theorem 2.3 shows that the following functions are g-convex: (i) $\phi(A) = \operatorname{tr}(e^A)$; (ii) $\phi(A) = \operatorname{tr}(A^\alpha)$ for $\alpha \ge 1$; (iii) $\lambda_1^\downarrow(e^A)$; (iv) $\lambda_1^\downarrow(A^\alpha)$ for $\alpha \ge 1$.

We now construct examples of g-convex functions different from those obtained via Theorem 2.3. Let us start with a motivating example.

Example 2.5. Let $z \in \mathbb{C}^d$. The function $\phi(A) := z^* A^{-1} z$ is g-convex. To prove this claim it suffices to verify midpoint convexity: $\phi(A\#B) \le \frac{1}{2}\phi(A) + \frac{1}{2}\phi(B)$ for

[2] For a more recent textbook exposition, see, e.g., [40, Thm. 1.5.4].

$A, B \in \mathbb{P}_d$. Since $(A\#B)^{-1} = A^{-1}\#B^{-1}$ and $A^{-1}\#B^{-1} \preceq \frac{A^{-1}+B^{-1}}{2}$ [6, 4.16], it follows that $\phi(A\#B) = z^*(A\#B)^{-1}z \leq \frac{1}{2}(z^*A^{-1}z + z^*B^{-1}z) = \frac{1}{2}(\phi(A) + \phi(B))$.

Below we substantially generalize this example, but first we give some background.

Definition 2.6 (positive linear map). *A linear map Φ from Hilbert space $\mathcal{H}_1$ to a Hilbert space $\mathcal{H}_2$ is called* positive *if for $0 \preceq A \in \mathcal{H}_1$, $\Phi(A) \succeq 0$. It is called* strictly positive *if $\Phi(A) \succ 0$ for $A \succ 0$; finally, it is called* unital *if $\Phi(I) = I$.*

Lemma 2.7 (see [6, Ex. 4.1.5]). *Define the* parallel sum *of HPD matrices A, B as*

$$A : B := [A^{-1} + B^{-1}]^{-1}.$$

Then, for any positive linear map $\Pi : \mathbb{P}_d \to \mathbb{P}_k$, we have

$$\Phi(A : B) \preceq \Phi(A) : \Phi(B).$$

Building on Lemma 2.7, we are ready to state a key theorem that helps us recognize and construct g-convex functions (see Theorem 2.14, for instance). This result is by itself not new (e.g., it follows from the classic paper [30]); due to its key importance we provide our own proof below for completeness.

Theorem 2.8. *Let $\Phi : \mathbb{P}_d \to \mathbb{P}_k$ be a strictly positive linear map. Then,*

$$\Phi(A\#_t B) \preceq \Phi(A)\#_t\Phi(B), \qquad t \in [0, 1] \quad \text{for } A, B \in \mathbb{P}_d. \tag{2.5}$$

Proof. The key insight of the proof is to use the integral identity [3]:

$$\int_0^1 \frac{\lambda^{\alpha-1}(1-\lambda)^{\beta-1}}{[\lambda a^{-1} + (1-\lambda)b^{-1}]^{\alpha+\beta}} d\lambda = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} a^\alpha b^\beta.$$

Using $\alpha = 1 - t$ and $\beta = t > 0$, for $C \succeq 0$ this yields the integral representation

$$C^t = \frac{\Gamma(1)}{\Gamma(t)\Gamma(1-t)} \int_0^1 \frac{\left[\lambda C^{-1} + (1-\lambda)I\right]^{-1}}{\lambda^t (1-\lambda)^{1-t}} d\lambda, \tag{2.6}$$

where Γ is the usual Gamma function. Since $A\#_t B = A^{1/2}(A^{-1/2}BA^{-1/2})^t A^{1/2}$, using (2.6), we may write it as

$$A\#_t B = \textstyle\int_0^1 \left[(1-\lambda)A^{-1} + \lambda B^{-1}\right]^{-1} d\mu(\lambda), \tag{2.7}$$

for a suitable measure $d\mu(\lambda)$. Applying the map Φ to both sides of (2.7) we obtain

$$\begin{aligned} \Phi(A\#_t B) &= \textstyle\int_0^1 \Phi\big(\left[(1-\lambda)A^{-1} + \lambda B^{-1}\right]^{-1}\big) d\mu(\lambda) \\ &= \textstyle\int_0^1 \Phi(\bar{A} : \bar{B}) d\mu(\lambda), \end{aligned}$$

where $\bar{A} = (1-\lambda)^{-1}A$ and $\bar{B} = \lambda^{-1}B$. Using Lemma 2.7 and the linearity of Φ, we see

$$\begin{aligned} \textstyle\int_0^1 \Phi(\bar{A} : \bar{B}) d\mu(\lambda) &\preceq \textstyle\int_0^1 \big(\Phi(\bar{A}) : \Phi(\bar{B})\big)\, d\mu(\lambda) \\ &= \textstyle\int_0^1 \left[(1-\lambda)\Phi(A)^{-1} + \lambda\Phi(B)^{-1}\right]^{-1} d\mu(\lambda) \\ &\overset{(2.7)}{=} \Phi(A)\#_t\Phi(B), \end{aligned}$$

which completes the proof. $\square$

A corollary of Theorem 2.8 (that subsumes Example 2.5) follows.

COROLLARY 2.9. *Let* $A, B \in \mathbb{P}_d$, *and let* $X \in \mathbb{C}^{d\times k}$ *have full column rank; then*

$$\operatorname{tr} X^*(A\#_t B)X \le [\operatorname{tr} X^*AX]^{1-t}[\operatorname{tr} X^*BX]^t, \qquad t \in [0,1]. \tag{2.8}$$

Proof. Use the positive linear map $A \mapsto \operatorname{tr} X^*AX$ in Theorem 2.8. ∎

Remark 2.10. Corollary 2.9 actually proves a result stronger than g-convexity: it shows *log-g-convexity*, i.e., $\phi(X\#Y) \le \sqrt{\phi(X)\phi(Y)}$, so that $\log\phi$ is g-convex. It is easy to verify that if ϕ_1, ϕ_2 are log-g-convex, then both $\phi_1\phi_2$ and $\phi_1 + \phi_2$ are log-g-convex.

We mention now another corollary to Theorem 2.8; we note in passing that it subsumes a more complicated result of Gurvits and Samorodnitsky [22, Lem. 3.2].

COROLLARY 2.11. *Let* $A_i \in \mathbb{C}^{d\times k}$ *with* $k \le d$ *such that* $\operatorname{rank}([A_i]_{i=1}^m) = k$; *also let* $B \succeq 0$. *Then* $\phi(X) := \log\det(B + \sum_i A_i^* X A_i)$ *is g-convex on* $\mathbb{P}_d$.

Proof. By our assumption on A_i and B, the map $\Phi = S \mapsto B + \sum_i A_i^* X A_i$ is strictly positive. Theorem 2.8 implies that $\Phi(X\#Y) = B + \sum_i A_i^*(X\#Y)A_i \preceq \Phi(X)\#\Phi(Y)$. This operator inequality is stronger than what we require. Indeed, since log det is monotonic and determinants are multiplicative, from this inequality it follows that

$$\begin{aligned}\phi(S\#R) &= \log\det\Phi(S\#R) \le \log\det(\Phi(S)\#\Phi(R))\\ &\le \tfrac{1}{2}\log\det\Phi(S) + \tfrac{1}{2}\log\det\Phi(R) = \tfrac{1}{2}\phi(S) + \tfrac{1}{2}\phi(R).\end{aligned}$$

Observe that the above result extends to $\phi(X) = \log\det\big(B + \int_0^\infty A_\lambda^* X A_\lambda d\mu(\lambda)\big)$, where μ is some positive measure on $(0,\infty)$. ∎

Remark 2.12. Corollary 2.11 may come as a surprise to some readers because $\log\det(X)$ is well known to be concave (in the Euclidean sense), and yet $\log\det(B + A^*XA)$ turns out to be g-convex; moreover, $\log\det(X)$ is g-linear, i.e., both g-convex and g-concave.

Example 2.13. In [48] (see also [18, 14]) a dissimilarity function to compare a pair of HPD matrices is studied. Specifically, for $X, Y \succ 0$, this function is called the S-Divergence and is defined as

$$S(X,Y) := \log\det\left(\tfrac{X+Y}{2}\right) - \tfrac{1}{2}\log\det(X) - \tfrac{1}{2}\log\det(Y). \tag{2.9}$$

Divergence (2.9) proves useful in several applications [48, 18, 14], and very recently its joint g-convexity (in both variables) was discovered [48]. Corollary 2.11 along with Remark 2.12 yields g-convexity of $S(X,Y)$ in either X or Y separately.

We are now ready to state our next key g-convexity result. A similar result was obtained in [51]; our result is somewhat more general as it allows incorporation of positive linear maps. Moreover, our proof technique is completely different.

THEOREM 2.14. *Let* $h : \mathbb{P}_k \to \mathbb{R}$ *be nondecreasing (in Löwner order) and g-convex. Let* $r \in \{\pm 1\}$, *and let* Φ *be a positive linear map. Then,* $\phi(S) = h(\Phi(S^r))$ *is g-convex.*

Proof. It suffices to prove midpoint g-convexity. Since $r \in \{\pm 1\}$, $(X\#Y)^r = X^r\#Y^r$. Thus, applying Theorem 2.8 to Φ and noting that h is nondecreasing it follows that

$$h(\Phi(X\#Y)^r) = h(\Phi(X^r\#Y^r)) \le h(\Phi(X^r)\#\Phi(Y^r)). \tag{2.10}$$

By assumption h is g-convex, so the last inequality in (2.10) yields

$$h(\Phi(X^r)\#\Phi(Y^r)) \le \tfrac{1}{2}h(\Phi(X^r)) + \tfrac{1}{2}h(\Phi(Y^r)) = \tfrac{1}{2}\phi(X) + \tfrac{1}{2}\phi(Y). \tag{2.11}$$

Notice that if h is strictly g-convex, then $\phi(S)$ is also strictly g-convex. $\square$

Example 2.15. Let $h = \log\det(X)$ and $\Phi(X) = B + \sum_i A_i^* X A_i$. Then, $\phi(X) = \log\det(B + \sum_i A_i^* X^r A_i)$ is g-convex. With $h(X) = \operatorname{tr}(X^\alpha)$ for $\alpha \geq 1$, $\operatorname{tr}(B + \sum_i A_i^* X^r A_i)^\alpha$ is g-convex.

Next, Theorem 2.16 presents a method for creating essentially logarithmic versions of our "tracial" g-convexity result Theorem 2.3.

THEOREM 2.16. *If $f : \mathbb{R} \to \mathbb{R}$ is convex, then $\phi(\cdot) := \sum_{i=1}^k f(\log \lambda_i^{\downarrow}(\cdot))$ is g-convex for each $1 \leq k \leq n$. If $h : \mathbb{R} \to \mathbb{R}$ is nondecreasing and convex, $\phi(\cdot) = \sum_{i=1}^k h(|\log \lambda(\cdot)|)$ is g-convex for each $1 \leq k \leq n$.*

To prove Theorem 2.16 we will need the following majorization.

LEMMA 2.17. *Let $\prec_{\log}$ denote the* log-majorization *order; i.e., for $x, y \in \mathbb{R}_{++}^n$ ordered nonincreasingly, we say $x \prec_{\log} y$ if $\prod_{i=1}^{n-1} x_i \leq \prod_{i=1}^{n-1} y_i$ and $\prod_{i=1}^n x_i = \prod_{i=1}^n y_i$. Then, for $A, B \in \mathbb{P}_n$ and $t \in [0,1]$, we have the log-majorization between the eigenvalues:*

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

Proof. The first majorization follows from a recent result of [35]. The second follows easily from $\lambda_1(AB) \leq \sigma_1(AB) \leq \sigma_1(A)\sigma_1(B) = \lambda_1(A)\lambda_1(B)$ (the final equality holds since $A, B \in \mathbb{P}_n$). Apply this inequality to the antisymmetric (Grassmann) exterior product $\wedge^k(AB)$, since $\sigma_1(\wedge^k(AB)) = \prod_{j=1}^k \sigma_j(AB)$ (see, e.g., [5, I; IV.2]); then we obtain $\lambda_1(\wedge^k(AB)) \leq \sigma_1(\wedge^k(AB))$. Now set $A \leftarrow A^{1-t}$, $B \leftarrow B^t$, and use the multiplicativity $\wedge^k(AB) = \wedge^k A \wedge^k B$ to complete the proof. $\square$

Proof. (See Theorem 2.16.) From Lemma 2.17 we have the majorization

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

on taking logarithms, this majorization may be written equivalently as

$$\log \lambda(A\#_t B) \prec (1-t)\log\lambda(A) + t\log\lambda(B). \tag{2.12}$$

Applying a classical result of [23] on majorization under convex functions, from (2.12) we obtain the inequality

$$\begin{aligned}
\phi(A\#_t B) = \sum\nolimits_{i=1}^k f(\log \lambda_i(A\#_t B)) &\leq \sum\nolimits_{i=1}^k f\left((1-t)\log\lambda_i(A) + t\log\lambda_i(B)\right) \\
&\leq (1-t)\sum\nolimits_{i=1}^k f(\log\lambda_i(A)) + t\sum\nolimits_{i=1}^k f(\log\lambda_i(B)) \\
&= (1-t)\phi(A) + t\phi(B).
\end{aligned}$$

Applying the Ky–Fan norm $\sum_{i=1}^k \sigma_i(\cdot)$, that is, the sum of top-k singular values, to (2.12), we obtain the weak majorization (see, e.g., [5, II] for more on majorization)

$$\sigma(\log A\#_t B) \prec_w \sigma\left[(1-t)\log\lambda(A) + t\log\lambda(B)\right] \prec_w (1-t)\sigma(\log A) + t\sigma(\log B). \tag{2.13}$$

Since h is monotone and convex, (2.13) yields g-convexity of $\sum_{i=1}^k h(|\log\lambda_i(\cdot)|)$. $\square$

COROLLARY 2.18. *Let $\Phi : \mathbb{R}^n \to \mathbb{R}_+$ be a symmetric gauge function (i.e., Φ is a norm invariant to permutation and sign changes). Also, let $X \in GL_n(\mathbb{C})$. Then, $\Phi(\sigma(\log(X^*AX))$ is g-convex.*

Proof. Note that $X^*(A\#B)X = (X^*AX)\#(X^*BX)$; now apply Theorem 2.16. $\square$

Example 2.19. Consider $\delta_R(A, X) := \|\log(X^{-1/2}AX^{-1/2})\|_{\mathrm{F}}$ the Riemannian distance between $A, X \in \mathbb{P}_d$ (see [6, Ch. 6]). Since $\|\log\lambda(X^{-1/2}AX^{-1/2})\|_2 =$

$\|\sigma(\log X^{-1/2}AX^{-1/2})\|_2$, it follows from Corollary 2.18 that $A \mapsto \delta_R(A,X)$ is g-convex (see also [6, Cor. 6.1.11]).

This immediately shows that the computations of the Fréchet (Karcher) mean and median of HPD matrices (also known as geometric mean and median of HPD matrices, respectively) are g-convex optimization problems; formally, these problems are given by

$$\min_{X\succ 0} \quad \sum\nolimits_{i=1}^m w_i\delta_R(X,A_i) \qquad \text{(geometric median)},$$
$$\min_{X\succ 0} \quad \sum\nolimits_{i=1}^m w_i\delta_R^2(X,A_i) \qquad \text{(geometric mean)},$$

where $\sum_i w_i = 1$, $w_i \geq 0$, and $A_i \succ 0$ for $1 \leq i \leq m$. The latter problem has received great interest in the literature [36, 6, 7, 48, 8, 25, 41], and its optimal solution is unique owing to the (strict) g-convexity of its objective. The former problem is less well known but in some cases proves more favorable [4, 41]; again, despite the nonconvexity of the objective, its g-convexity ensures that every local solution is global.

We conclude this section by using Lemma 2.17 to prove the following log-convexity analogue to Theorem 2.16 (cf. the scalar case studied in [39, Prop. 2.4]).

THEOREM 2.20. *Let $f(x) = \sum_{j\geq 0} a_jx^j$ be real analytic with $a_j \geq 0$ for $j \geq 0$ and radius of convergence R. Then, $\phi(\cdot) = \prod_{i=1}^k f(\lambda_i(\cdot))$ is log-g-convex on matrices with spectrum in $(0,R)$.*

Proof. It suffices to verify that $\log\phi(A\#B) \leq \frac{1}{2}\log\phi(A) + \frac{1}{2}\log\phi(B)$. Since $f' \geq 0$, we have

$$\begin{aligned}
\phi(A\#B) = \prod\nolimits_{i=1}^k f(\lambda_i(A\#B)) &\leq \prod\nolimits_{i=1}^k f(\lambda_i^{1/2}(A)\lambda_i^{1/2}(B)) && \text{(using Lemma 2.17)}\\
&\leq \prod\nolimits_{i=1}^k \sqrt{f(\lambda_i(A))}\sqrt{f(\lambda_i(B))} && \text{(Cauchy–Schwarz on power-series of } f)\\
&= \sqrt{\phi(A)}\sqrt{\phi(B)}.
\end{aligned}$$

Taking logarithms, we see that $\phi(\cdot)$ is log-g-convex (and hence also g-convex). ∎

Example 2.21. Some examples of f that satisfy the conditions of Theorem 2.20 are exp, sinh on $(0,\infty)$, $-\log(1-x)$ and $(1+x)/(1-x)$ on $(0,1)$; see [39] for more examples.

2.2. Multivariable g-convexity. We now describe an extension of g-convexity to multiple matrices; a two-variable version was also partially explored in [49, 51], though under a different name. We begin our multivariable extension by recalling a few basic properties of the Kronecker product [34].

LEMMA 2.22. *Let $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p\times q}$. Then, $A\otimes B := [a_{ij}B] \in \mathbb{R}^{mp\times nq}$ satisfies the following:*

(i) $(A\otimes B)^* = A^*\otimes B^*$.

(ii) $(A\otimes B)^{-1} = A^{-1}\otimes B^{-1}$.

(iii) *Assuming that the respective products exist,*

$$(AC\otimes BD) = (A\otimes B)(C\otimes D). \tag{2.14}$$

(iv) $A\otimes(B\otimes C) = (A\otimes B)\otimes C$.

(v) *If $A = UD_1U^*$ and $B = VD_2V^*$, then $(A\otimes B) = (U\otimes V)(D_1\otimes D_2)(U\otimes V)^*$.*

(vi) *Let $A, B \succeq 0$ and $t \in \mathbb{R}$; then*

$$(A\otimes B)^t = A^t\otimes B^t. \tag{2.15}$$

(vii) *If* $A \succeq B$ *and* $C \succeq D$*, then* $(A \otimes C) \succeq (B \otimes D)$*.*

Proof. Identities (i)–(iii) are classic; (v) follows easily from (i) and (iv), while (vi) and (vii) follow from (v); and (vii) is an easy exercise. ∎

We will need the following easy but key result on tensor products of geometric means.

LEMMA 2.23. *Let* $A, B \in \mathbb{P}_{d_1}$ *and* $C, D \in \mathbb{P}_{d_2}$*. Then,*

$$(A\#B) \otimes (C\#D) = (A \otimes C)\#(B \otimes D). \tag{2.16}$$

Proof. Denote $\gamma(X, Y) := (X^{-1/2}YX^{-1/2})^{1/2}$. Observe that

$$\begin{aligned}\gamma(A, B) \otimes \gamma(C, D) &= (A^{-1/2}BA^{-1/2})^{1/2} \otimes (C^{-1/2}DC^{-1/2})^{1/2} \\ &= [(A^{-1/2}BA^{-1/2}) \otimes (C^{-1/2}DC^{-1/2})]^{1/2} \\ &= [(A \otimes C)^{-1/2}(B \otimes D)(A \otimes C)^{-1/2}]^{1/2} \\ &= \gamma(A \otimes C, B \otimes D),\end{aligned}$$

where the second equality follows from Lemma 2.22(iii), while the third follows from Lemma 2.22(ii), (iii), and (vi). A similar manipulation then shows that

$$\begin{aligned}(A\#B) \otimes (C\#D) &= (A^{1/2}\gamma(A, B)A^{1/2}) \otimes (C^{1/2}\gamma(C, D)C^{1/2}) \\ &= (A^{1/2} \otimes C^{1/2})(\gamma(A, B) \otimes \gamma(C, D))(A^{1/2} \otimes C^{1/2}) \\ &= (A \otimes C)^{1/2}(\gamma(A, B) \otimes \gamma(C, D))(A \otimes C)^{1/2} \\ &= (A \otimes C)^{1/2}\gamma(A \otimes C, B \otimes D)(A \otimes C)^{1/2} \\ &= (A \otimes C)\#(B \otimes D),\end{aligned}$$

which concludes the proof. ∎

Lemma 2.23 inductively extends to the multivariable case, so that

$$\textstyle\bigotimes_{i=1}^m (A_i\#B_i) = \left(\bigotimes_{i=1}^m A_i\right) \# \left(\bigotimes_{i=1}^m B_i\right). \tag{2.17}$$

Using identity (2.17), we thus obtain the following multivariate analogue to Theorem 2.16.

THEOREM 2.24. *Let* h *be an increasing convex function on* $\mathbb{R}_+ \to \mathbb{R}$*. Then, the map* $\prod_{i=1}^m \operatorname{tr} h(X_i)$ *is jointly g-convex; i.e.,* $\operatorname{tr} h(\bigotimes_{i=1}^m X_i)$ *is g-convex in its variables.*

Proof. Let $(A_1, B_1), \ldots, (A_m, B_m)$ be pairs of HPD matrices of arbitrary sizes (such that, for each i, A_i and B_i are of the same size). Let j index the eigenvalues of the tensor product $\bigotimes_{i=1}^m (A_i\#B_i)$. Then, starting with identity (2.17), we obtain

$$\begin{aligned}\lambda_j\left[\textstyle\bigotimes_{i=1}^m (A_i\#B_i)\right] &= \lambda_j\left[\left(\textstyle\bigotimes_{i=1}^m A_i\right) \# \left(\textstyle\bigotimes_{i=1}^m B_i\right)\right] \le \tfrac{1}{2}\lambda_j\left[\textstyle\bigotimes_{i=1}^m A_i + \textstyle\bigotimes_{i=1}^m B_i\right], \\ \operatorname{tr} h\left(\textstyle\bigotimes_{i=1}^m (A_i\#B_i)\right) &= \sum_j h\left(\lambda_j\left[\textstyle\bigotimes_{i=1}^m (A_i\#B_i)\right]\right) \le \sum_j h\left(\tfrac{1}{2}\lambda_j\left[\textstyle\bigotimes_{i=1}^m A_i + \textstyle\bigotimes_{i=1}^m B_i\right]\right) \\ &\le \tfrac{1}{2}\sum_j h(\lambda_j(\textstyle\bigotimes_{i=1}^m A_i)) + \tfrac{1}{2}\sum_j h(\lambda_j(\textstyle\bigotimes_{i=1}^m B_i)) \\ &= \tfrac{1}{2}\operatorname{tr} h(\textstyle\bigotimes_{i=1}^m A_i) + \tfrac{1}{2}\operatorname{tr} h(\textstyle\bigotimes_{i=1}^m B_i) \\ &= \tfrac{1}{2}\prod_{i=1}^m \operatorname{tr} h(A_i) + \tfrac{1}{2}\prod_{i=1}^m \operatorname{tr} h(B_i),\end{aligned}$$

which shows the desired multivariable g-convexity of the map $\operatorname{tr} h(\bigotimes_{i=1}^m X_i)$. ∎

Again, using (2.17) we obtain the following multivariate analogue to Theorem 2.8.

THEOREM 2.25. *Let $(X_1, Y_1), \ldots, (X_m, Y_m)$ be pairs of HPD matrices of arbitrary sizes (such that, for each i, X_i and Y_i are of the same size). Let $\Phi_i : \mathcal{H}_i \to \mathcal{H}'_i$ be a positive linear map for each i, and let Φ be the* positive multilinear map *defined by $\Phi \equiv \otimes_{i=1}^m A_i \mapsto \otimes_{i=1}^m \Phi_i(A_i)$. Then,*

$$\Phi(\otimes_{i=1}^m (X_i \# Y_i)) \preceq \Phi(\otimes_i X_i) \# \Phi(\otimes_i Y_i). \tag{2.18}$$

Proof. Expanding the definition of Φ, we have

$$\begin{aligned}\Phi(\textstyle\bigotimes_i (X_i \# Y_i)) &= \textstyle\bigotimes_i \Phi_i(X_i \# Y_i) \preceq \bigotimes_i [\Phi_i(X_i) \# \Phi_i(Y_i)] \\ &= \big[\textstyle\bigotimes_i \Phi_i(X_i)\big] \# \big[\bigotimes_i \Phi_i(Y_i)\big] = \Phi(\bigotimes_i X_i) \# \Phi(\bigotimes_i Y_i).\end{aligned}$$

The operator inequality (2.18) then follows upon invoking Theorem 2.8 and Lemma 2.22(viii). ◻

Building on Theorem 2.25, we also derive a generalization to Theorem 2.14.

THEOREM 2.26. *Let $h : \otimes_i \mathcal{H}'_i \to \mathbb{R}$ be nondecreasing (in Löwner order) and g-convex. Let $r_i \in \{\pm 1\}$, and let $\Phi : \otimes_i \mathcal{H}_i \to \otimes_i \mathcal{H}'_i$ be a strictly positive multilinear map. Then, $\phi(X_1, \ldots, X_m) = (h \circ \Phi)(\bigotimes_i X_i^{r_i})$ is jointly g-convex (i.e., g-convex in $X_1, \ldots, X_m$).*

Proof. Since ϕ is continuous, it suffices to establish midpoint g-convexity.

$$\begin{aligned}(h \circ \Phi)(\textstyle\bigotimes_i (X_i \# Y_i)^{r_i}) &= (h \circ \Phi)(\textstyle\bigotimes_i (X_i^{r_i} \# Y_i^{r_i})) \\ &\preceq h\big(\Phi(\textstyle\bigotimes_i X_i^{r_i}) \# \Phi(\bigotimes_i Y_i^{r_i})\big) \\ &\preceq \tfrac{1}{2}\left((h \circ \Phi)(\textstyle\bigotimes_i X_i^{r_i}) + (h \circ \Phi)(\bigotimes_i Y_i^{r_i})\right) \\ &= \tfrac{1}{2}\left(\phi(X_1, \ldots, X_m) + \phi(Y_1, \ldots, Y_m)\right).\end{aligned}$$

Since h is nondecreasing, using Theorem 2.25 the first inequality follows. The second one follows as h is g-convex, which completes the proof. ◻

Using identities (2.15) and (2.17) with Lemma 2.17, we obtain the following log-majorizations.

PROPOSITION 2.27. *Let $(A_i, B_i)_{i=1}^m$ be pairs of HPD matrices of compatible sizes. Then,*

$$\begin{aligned}\lambda(\textstyle\bigotimes_{i=1}^m A_i \#_t B_i) &\prec_{\log} \lambda([\textstyle\bigotimes_{i=1}^m A_i]^{1-t} [\bigotimes_{i=1}^m B_i]^t), \qquad t \in [0, 1], \\ \lambda([\textstyle\bigotimes_{i=1}^m A_i]^{1-t} [\bigotimes_{i=1}^m B_i]^t) &\prec_{\log} \lambda[\textstyle\bigotimes_{i=1}^m A_i^{1-t}] \lambda[\bigotimes_{i=1}^m B_i^t].\end{aligned}$$

Proposition 2.27 grants us the following multivariate analogue to Theorem 2.16.

THEOREM 2.28. *If $f : \mathbb{R} \to \mathbb{R}$ is convex, then $\phi(\cdot) := \sum_{j=1}^k f(\log \lambda_j(\bigotimes_{i=1}^m X_i))$ is g-convex on $\{X_i \in \mathbb{P}_n\}_{i=1}^m$ for each $1 \leq k \leq n$. If $h : \mathbb{R} \to \mathbb{R}$ is nondecreasing and convex, then $\phi(\cdot) = \sum_{j=1}^k h(|\log \lambda_j(\bigotimes_{i=1}^m X_i)|)$ is g-convex for $1 \leq k \leq n$.*

Theorem 2.28 brings us to the end of our theoretical results on recognizing and constructing g-convex functions. We are now ready to devote our attention to optimization algorithms. In particular, we first discuss manifold optimization [1] techniques in section 3. Then, in section 4 we introduce a special class of functions that overlaps with g-convex functions, but not entirely, and admits simpler "conic fixed-point" algorithms.

3. Manifold optimization for g-convex functions. Since $\mathbb{P}_d$ is a smooth manifold, we can use optimization techniques based on exploiting smooth manifold structure. In addition to common concepts such as tangent vectors and derivatives along manifolds, different optimization methods need a subset of new definitions and explicit expressions for inner products, gradients, retractions, vector transport, and Hessians [1, 24].

Since $\mathbb{P}_d$ can be viewed as a submanifold of the Euclidean space $\mathbb{R}^{2d^2}$, most of the concepts of importance to our study can be defined by using the embedding structure of Euclidean space. The tangent space at any point is the space $\mathbb{H}_d$ of $d \times d$ Hermitian matrices. The derivative of a function on the manifold in any direction in the tangent space is simply the embedded Euclidean derivative in that direction.

For several optimization algorithms, two different inner product formulations were tested in [25] for $\mathbb{P}_d$. The authors observed that the intrinsic inner product leads to the best convergence speed for the tested algorithms. We too observed that the intrinsic inner product yields more than a hundred times faster convergence for our algorithms compared to the induced inner product of Euclidean space. The *intrinsic inner product* of two tangent vectors at point X on the manifold is given by

$$g_X(\eta, \xi) = \operatorname{tr}(\eta X^{-1} \xi X^{-1}), \qquad \eta, \xi \in \mathbb{H}_d. \tag{3.1}$$

This intrinsic inner product leads to geodesics of the form (2.2). Now that we have set up an inner product tensor, we can define the gradient direction as the direction of the maximum change. The inner product between the gradient vector and a vector in the tangent space is equal to the gradient of the function in that direction. If $\operatorname{grad}^{\mathrm{H}} f(X) = \frac{1}{2}(\operatorname{grad} f(X) + (\operatorname{grad} f(X))^*)$ is the Hermitian part of Euclidean gradient, then the gradient in intrinsic metric is given by

$$\operatorname{grad}^{\mathrm{HPD}} f(X) = X \operatorname{grad}^{\mathrm{H}} f(X) X.$$

The simplest gradient descent approach, namely, steepest descent, also needs the notion of projection of a vector in the tangent space onto a point on the manifold. Such a projection is called *retraction*. If the manifold is Riemannian, a particular retraction is the exponential map, i.e., moving along a geodesic. If the inner product is the induced inner product of the manifold, then the retraction is normal retraction on the Euclidean space which is obtained by summing the point on the manifold and the vector on the tangent space. The intrinsic inner product of (3.1) of the Riemannian manifold leads to the following exponential map:

$$R_X^{\mathrm{HPD}}(\xi) = X^{1/2} \exp(X^{-1/2} \xi X^{-1/2}) X^{1/2}, \qquad \xi \in \mathbb{H}_d. \tag{3.2}$$

From a numerical perspective, our experiments revealed that the following equivalent representation of the retraction (3.2) gives the best computational speed:

$$R_X^{\mathrm{HPD}}(\xi) = X \exp(X^{-1} \xi), \qquad \xi \in \mathbb{H}_d. \tag{3.3}$$

Definitions of the gradient and retraction suffice for implementing steepest descent on $\mathbb{P}_d$. For approaches such as conjugate gradients or quasi-Newton methods, we need to relate the tangent vector at one point to the tangent vector at another point, i.e., we need to define *vector transport*. A special case of vector transport on a Riemannian manifold is parallel transport: for the induced Euclidean metric, parallel transport is simply the identity map. In order to compute the parallel transport one first needs to compute the *Levi-Civita connection*. This connection is a way to compute directional derivatives of vector fields. It is a map from the Cartesian product of tangent bundles to the tangent bundle:

$$\nabla : T\mathcal{M} \times T\mathcal{M} \to T\mathcal{M},$$

where $T\mathcal{M}$ is the tangent bundle of manifold $\mathcal{M}$ (i.e., the space of smooth vector fields on $\mathcal{M}$). It can be verified that for the intrinsic metric (3.1) the following connection satisfies all the needed properties (see, e.g., [25]):

$$\nabla^{\text{HPD}}_{\zeta_X}\xi_X = D\xi(X)[\zeta_X] - \tfrac{1}{2}(\zeta_X X^{-1}\xi_X + \xi_X X^{-1}\zeta_X),$$

where $D\xi(X)$ denotes the classical Fréchet derivative of $\xi(X)$. ξ_X and ζ_X are vector fields on the manifold $\mathbb{H}_d$. Subindex X is used to discriminate a vector field from a tangent vector.

Consider $P(t)$, a vector field along the geodesic curve $\gamma(t)$. Parallel transport along a curve is given by the differential equation

$$D_t P(t) = \nabla_{\dot\gamma(t)} P(t) = 0 \qquad \text{s.t. } P(0) = \eta.$$

For the intrinsic metric, the above equation becomes

$$\dot P(t) - \tfrac{1}{2}(\dot\gamma(t) X_t^{-1} P(t) + P(t) X_t^{-1}\dot\gamma(t)) = 0.$$

The geodesic passing through $\gamma(0) = X$ with $\dot\gamma = \xi$ is given by

$$\gamma(t) = X^{1/2}\exp(tX^{-1/2}\xi X^{-1/2})X^{1/2}.$$

For $t = 1$ we get the retraction (3.2). It can be shown that along the geodesic curve the following equation gives the parallel transport:

$$P(t) = X^{1/2}\exp(t\tfrac{1}{2}X^{-1/2}\xi X^{-1/2})X^{-1/2}\eta X^{-1/2}\exp(t\tfrac{1}{2}X^{-1/2}\xi X^{-1/2})X^{1/2}.$$

Thus, parallel transport for the intrinsic inner product is given by

$$\mathcal{T}^{\text{HPD}}_{X,Y}(\eta) = X^{1/2}(X^{-1/2}YX^{-1/2})^{1/2}X^{-1/2}\eta X^{-1/2}(X^{-1/2}YX^{-1/2})^{1/2}X^{1/2}.$$

It is important to note that this parallel transport can be written in a compact form that is also computationally more advantageous, namely,

$$\mathcal{T}^{\text{HPD}}_{X,Y}(\eta) = E\eta E^*, \quad \text{where} \quad E = (YX^{-1})^{1/2}. \tag{3.4}$$

We are now ready to describe a quasi-Newton method on $\mathbb{P}_d$. Different algorithms such as conjugate-gradient, BFGS, and trust-region methods for the Riemannian manifold $\mathbb{P}_d$ are explained in [25]. Here we only provide details for a limited memory version of Riemannian BFGS (RBFGS). The RBFGS algorithm for general retraction and vector transport was originally explained in [44], and the proof of convergence appeared in [46], although for a slightly different version. It was proved that for g-convex functions and with line-search that satisfies Wolfe conditions, the RBFGS algorithm has a (local) superlinear convergence rate. The RBFGS algorithm can be transformed into a limited-memory L-RBFGS algorithm by unrolling the update step of the approximate Hessian computation as shown in Algorithm 1. As may be apparent from the algorithm, parallel transport and its inverse can be the computational bottlenecks. One possible speedup is to store the matrix E and its inverse in (3.4).

4. Geometric optimization for log-nonexpansive functions. Though manifold optimization is powerful and widely applicable (see, e.g., the excellent toolbox [10]), for a special class of geometric optimization problems we may be able to circumvent its heavy machinery in favor of potentially much simpler algorithms.

Algorithm 1 L-RBFGS.

Given: Riemannian manifold $\mathcal{M}$ with Riemannian metric g; vector transport $\mathcal{T}$ on $\mathcal{M}$ with associated retraction R; initial value X_0; a smooth function f
Set initial $H_{\text{diag}} = 1/\sqrt{g_{X_0}(\text{grad}f(X_0), \text{grad}f(X_0))}$
for $k = 0, 1, \ldots$ **do**
  Obtain descent direction ξ_k by unrolling the RBFGS method
    $\xi_k \leftarrow \text{HessMul}(-\text{grad}f(X_k), k)$
  Use line-search to find α s.t. $f(R_{X_k}(\alpha\xi_k))$ is sufficiently smaller than $f(X_k)$
  Calculate $X_{k+1} = R_{X_k}(\alpha\xi_k)$
  Define $S_k = \mathcal{T}_{X_k,X_{k+1}}(\alpha\xi_k)$
  Define $Y_k = \text{grad}f(X_{k+1}) - \mathcal{T}_{X_k,X_{k+1}}(\text{grad}f(X_k))$
  Update $H_{\text{diag}} = g_{X_{k+1}}(S_k, Y_k)/g_{X_{k+1}}(Y_k, Y_k)$
  Store Y_k; S_k; $g_{X_{k+1}}(S_k, Y_k)$; $g_{X_{k+1}}(S_k, S_k)/g_{X_{k+1}}(S_k, Y_k)$; H_{diag}
end for
return X_k

function HessMul(P, k)
if $k > 0$ **then**
  $P_k = P - \frac{g_{X_{k+1}}(S_k, P_{k+1})}{g_{X_{k+1}}(Y_k, S_k)} Y_k$
  $\hat{P} = \mathcal{T}^{-1}_{X_k,X_{k+1}} \text{HessMul}(\mathcal{T}_{X_k,X_{k+1}} P_k, k-1)$
  return $\hat{P} - \frac{g_{X_{k+1}}(Y_k, \hat{P})}{g_{X_{k+1}}(Y_k, S_k)} S_k + \frac{g_{X_{k+1}}(S_k, S_k)}{g_{X_{k+1}}(Y_k, S_k)} P$
else
  return $H_{\text{diag}} P$
end if
end function

This motivation underlies the material developed in this section, where ultimately our goal is to obtain fixed-point iterations by viewing $\mathbb{P}_d$ as a convex cone instead of a Riemannian manifold. This viewpoint is grounded in nonlinear Perron–Frobenius theory [32], and it proves to be of practical value for our application in section 5. Notably, for certain problems we can obtain globally optimal solutions even without g-convexity. We believe the general conic optimization theory developed in this section may be of wider interest.

Consider thus the minimization problem

$$\min_{S \succ 0} \quad \Phi(S), \tag{4.1}$$

where Φ is a continuously differentiable real-valued function on $\mathbb{P}_d$. Since the constraint set $\{S \succ 0\}$ is an open subset of a Euclidean space, the first-order optimality condition for (4.1) is similar to that of unconstrained optimization. A point S^* is a candidate local minimum of Φ only if its gradient at this point is zero, that is,

$$\nabla\Phi(S^*) = 0. \tag{4.2}$$

The nonlinear (matrix) equation (4.2) could be solved using numerical techniques such as Newton's method. However, such approaches can be computationally more demanding than the original optimization problem, especially because they involve the (inverse of) the second derivative $\nabla^2\Phi$ at each iteration. We propose exploiting a fixed-point iteration that offers a simpler method for solving (4.2). *More importantly,* the fixed-point technique allows one to show that under certain conditions the solution to (4.2) is unique and therefore potentially a global minimum (essentially, if the global minimum is attained, then it must be this unique stationary point).

Assume therefore that (4.2) is rewritten as the fixed-point equation

$$S^* = \mathcal{G}(S^*). \tag{4.3}$$

Then, a fixed point of the map $\mathcal{G} : \mathbb{P}_d \to \mathbb{P}_d$ is a potential solution (since it is a stationary point) to the minimization problem (4.1). The natural question is how to find such a fixed point and, starting with a feasible $S_0 \succ 0$, whether it suffices to perform the Picard iteration

$$S_{k+1} \leftarrow \mathcal{G}(S_k), \quad k = 0, 1, \ldots. \tag{4.4}$$

Iteration (4.4) is (usually) *not* a fixed-point iteration when cast in a normed vector space; the conic geometry of $\mathbb{P}_d$ alluded to previously suggests that it might be better to analyze the iteration using a non-vectorial metric.

We provide below a class of sufficient conditions ensuring convergence of (4.4). Therein, the correct metric space in which to study convergence is neither the Euclidean (or Banach) space $\mathbb{R}^n$ nor the Riemannian manifold $\mathbb{P}_d$ with distance (5.5). Instead, a conic metric proves more suitable, namely, the Thompson part metric, an object of great interest in nonlinear Perron–Frobenius theory [32, 31].

Our sufficient conditions stem from the following key definition.

DEFINITION 4.1 (log-nonexpansive). *Let $f : (0, \infty) \to (0, \infty)$. We say f is* log-nonexpansive *(LN) on a compact interval $I \subset (0, \infty)$ if there exists a constant $0 \le q \le 1$ such that*

$$|\log f(t) - \log f(s)| \le q|\log t - \log s| \quad \forall s, t \in I. \tag{4.5}$$

*If $q < 1$, we say f is q-*log-contractive*. If for every $s \neq t$ it holds that*

$$|\log f(t) - \log f(s)| < |\log t - \log s| \quad \forall s, t \quad s \neq t,$$

we say f is log-contractive.

We use LN functions in a concrete optimization task in section 4.2. The proofs therein rely on core properties of the Thompson metric and contraction maps in the associated metric space; we cover requisite background in section 4.1. The content of section 4.1 is of independent interest as the theorems therein provide techniques for establishing contractivity (or nonexpansivity) of nonlinear maps from $\mathbb{P}_d$ to $\mathbb{P}_k$.

4.1. Thompson metric and contractive maps. On $\mathbb{P}_d$, the *Thompson metric* is defined as (cf. δ_R which uses $\|\cdot\|_F$)

$$\delta_T(X, Y) := \|\log(Y^{-1/2} X Y^{-1/2})\|, \tag{4.6}$$

where $\|\cdot\|$ is the usual operator norm (largest singular value), and log is the matrix logarithm. Let us recall some core (known) properties of (4.6); for details please see [31, 32, 33].

PROPOSITION 4.2. *Unless noted otherwise, all matrices are assumed to be HPD.*

$$\begin{aligned}
&\text{(4.7a)} & \delta_T(X^{-1}, Y^{-1}) &= \delta_T(X, Y), \\
&\text{(4.7b)} & \delta_T(B^*XB, B^*YB) &= \delta_T(X, Y), \qquad B \in GL_n(\mathbb{C}), \\
&\text{(4.7c)} & \delta_T(X^t, Y^t) &\leq |t|\delta_T(X, Y) \qquad \textit{for } t \in [-1, 1], \\
&\text{(4.7d)} & \delta_T\Big(\sum\nolimits_i w_i X_i, \sum\nolimits_i w_i Y_i\Big) &\leq \max_{1 \leq i \leq m} \delta_T(X_i, Y_i), \qquad w_i \geq 0, w \neq 0, \\
&\text{(4.7e)} & \delta_T(X + A, Y + A) &\leq \frac{\alpha}{\alpha + \beta}\delta_T(X, Y), \qquad A \succeq 0,
\end{aligned}$$

where $\alpha = \max\{\|X\|, \|Y\|\}$ *and* $\beta = \lambda_{\min}(A)$.

We now prove a powerful refinement to (4.7b), which shows contraction under "compression."

THEOREM 4.3. *Let* $X \in \mathbb{C}^{d \times p}$*, where* $p \leq d$ *have full column rank. Let* A, $B \in \mathbb{P}_d$*. Then,*

$$\delta_T(X^*AX, X^*BX) \leq \delta_T(A, B). \tag{4.8}$$

Proof. Let $A_C = X^*AX$ and $B_C = X^*BX$ denote the "compressions" of A and B, respectively; these compressions are invertible since X is assumed to have full column rank. The largest generalized eigenvalue of the pencil (A, B) is given by

$$\lambda_1(A, B) := \lambda_1(A^{-1}B) = \max_{x \neq 0} \frac{x^*Bx}{x^*Ax}. \tag{4.9}$$

Starting with (4.9) we have the following relations:

$$\begin{aligned}
\lambda_1(A^{-1}B) = \lambda_1(A^{-1/2}BA^{-1/2}) \quad &= \quad \max_{x \neq 0} \frac{x^*A^{-1/2}BA^{-1/2}x}{x^*x} \\
= \max_{w \neq 0} \frac{w^*Bw}{(A^{1/2}w)^*(A^{1/2}w)} \quad &= \quad \max_{w \neq 0} \frac{w^*Bw}{w^*Aw} \\
\geq \max_{w = Xp, p \neq 0} \frac{w^*Bw}{w^*Aw} \quad &= \quad \max_{p \neq 0} \frac{p^*X^*BXp}{p^*X^*AXp} \\
= \max_{p \neq 0} \frac{p^*B_Cp}{p^*A_Cp} \quad &= \quad \lambda_1(A_C^{-1}B_C) = \lambda_1(A_C^{-1/2}BA_C^{-1/2}).
\end{aligned}$$

Similarly, we can show that $\lambda_1(B^{-1}A) = \lambda_1(B^{-1/2}AB^{-1/2}) \geq \lambda_1(B_C^{-1/2}A_CB_C^{-1/2})$. Since A, B and the matrices A_C, B_C are all positive, we may conclude

$$\max\left\{\log \lambda_1(A_U^{-1}B_U), \log \lambda_1(B_U^{-1}A_U)\right\} \leq \max\left\{\lambda_1(A^{-1}B), \log \lambda_1(B^{-1}A)\right\}, \tag{4.10}$$

which is nothing but the desired claim $\delta_T(X^*AX, X^*BX) \leq \delta_T(A, B)$. ∎

Theorem 4.3 can be extended to encompass more general "compression" maps, namely, to those defined by operator monotone functions, a class that enjoys great importance in matrix theory; see, e.g., [5, Ch. V] and [6].

THEOREM 4.4. *Let* f *be an operator monotone (i.e., if* $X \preceq Y$*, then* $f(X) \preceq f(Y)$*) function on* $(0, \infty)$ *such that* $f(0) \geq 0$*. Then,*

$$\delta_T(f(X), f(Y)) \leq \delta_T(X, Y), \qquad X, Y \in \mathbb{P}_d. \tag{4.11}$$

Proof. If f is operator monotone with $f(0) \geq 0$, then it admits the integral representation [5, (V.53)]

$$f(t) = \gamma + \beta t + \int_0^\infty \frac{\lambda t}{\lambda + t} d\mu(\lambda), \tag{4.12}$$

where $\gamma = f(0)$, $\beta \geq 0$, and $d\mu(t)$ is a nonnegative measure. Using (4.12) we get

$$f(A) = \gamma I + \beta A + \int_0^\infty (\lambda A)(\lambda I + A)^{-1} d\mu(\lambda) =: \gamma I + \beta A + M(A).$$

Similarly, we obtain $f(B) = \gamma I + \beta B + M(B)$. Now, consider at first

$$\begin{aligned}
\delta_T(M(A), M(B)) &= \delta_T(\textstyle\int \lambda A(\lambda I + A)^{-1} d\mu(t), \int \lambda A(\lambda I + A)^{-1} d\mu(t)) \\
&\leq \max_\lambda \delta_T(\lambda A(\lambda I + A)^{-1}, \lambda B(\lambda I + B)^{-1}) \\
&\leq \max_\lambda \delta_T((\lambda A^{-1} + I)^{-1}, (\lambda B^{-1} + I)^{-1}) \\
&= \max_\lambda \delta_T(I + \lambda A^{-1}, I + \lambda B^{-1}) \\
&\leq \max_\lambda \frac{\bar{\alpha}}{\bar{\alpha} + 1} \delta_T(\lambda A^{-1}, \lambda B^{-1}), \qquad \bar{\alpha} := \max\{\|A^{-1}\|, \|B^{-1}\|\}, \\
&= \frac{\bar{\alpha}}{\bar{\alpha} + 1} \delta_T(A, B) < \delta_T(A, B).
\end{aligned}$$

Next, defining $\alpha := \max\{\|\beta A + M(A)\|, \|\beta B + M(B)\|\}$, we can use the above contraction to help prove contraction for the map f as follows:

$$\begin{aligned}
\delta_T(f(A), f(B)) &= \delta_T(\gamma I + \beta A + M(A), \gamma I + \beta B + M(B)) \\
&\leq \frac{\alpha}{\alpha + \gamma} \delta_T(\beta A + M(A), \beta B + M(B)) \\
&\leq \frac{\alpha}{\alpha + \gamma} \max\{\delta_T(\beta A, \beta B), \delta_T(M(A), M(B))\} \\
&\leq \frac{\alpha}{\alpha + \gamma} \delta_T(A, B).
\end{aligned}$$

Moreover, for $A \neq B$ the inequality is strict if $f(0) > 0$. ∎

Example 4.5. Let $X \in \mathbb{C}^{d \times k}$, and let $f = t^r$ for $t \in (0, \infty)$ and $r \in (0, 1)$. Then,

$$\begin{aligned}
\delta_T((X^*AX)^r, (X^*BX)^r) &\leq \delta_T(A, B) \qquad \forall A, B \in \mathbb{P}_d, \\
\delta_T(X^*A^rX, X^*B^rX) &\leq \delta_T(A, B) \qquad \forall A, B \in \mathbb{P}_d.
\end{aligned}$$

Theorem 4.3 and Theorem 4.4 together yield the following general result.

COROLLARY 4.6. *Let $\Phi : \mathbb{P}_d \to \mathbb{P}_k$ ($k \leq d$) and $\Psi : \mathbb{P}_k \to \mathbb{P}_r$ ($r \leq k$) be completely positive (see, e.g.,* [6, *Ch.* 3]*) maps. Then,*

$$\delta_T(f(\Phi(X)), f(\Phi(Y))) \leq \delta_T(X, Y), \qquad X, Y \in \mathbb{P}_d, \tag{4.13}$$

$$\delta_T(\Psi(f(X)), \Psi(f(Y))) \leq \delta_T(X, Y), \qquad X, Y \in \mathbb{P}_k. \tag{4.14}$$

Proof. We prove (4.13); the proof of (4.14) is similar and hence is omitted. From Theorem 4.4 it follows that $\delta_T(f(\Phi(X)), f(\Phi(Y))) \leq \delta_T(\Phi(X), \Phi(Y))$. Since Φ is completely positive, it follows from a result of [19] and [29] that there exist matrices $V_j \in \mathbb{C}^{d\times k}$, $1 \le j \le dk$, such that

$$\Phi(X) = \sum\nolimits_{i=1}^{nk} V_j^* X V_j, \qquad X \in \mathbb{P}_d.$$

Theorem 4.3 and property (4.7d) imply that $\delta_T(\Phi(X), \Phi(Y)) \le \delta_T(X, Y)$, which proves (4.13). ∎

4.1.1. Thompson log-nonexpansive maps. Let $\mathcal{G}$ be a map from $\mathcal{X} \subseteq \mathbb{P}_d \to \mathcal{X}$. Analogous to (4.5), we say $\mathcal{G}$ is (Thompson) *log-nonexpansive* if

$$\delta_T(\mathcal{G}(X), \mathcal{G}(Y)) \le \delta_T(X, Y) \qquad \forall X, Y \in \mathcal{X};$$

the map is called *log-contractive* if the inequality is strict. We present now a key result that justifies our nomenclature and the analogy to (4.5): it shows that the sum of a log-contractive map and an LN map is log-contractive. This behavior is a striking feature of the nonpositive curvature of $\mathbb{P}_d$; such a result does *not* hold in normed vector spaces.

THEOREM 4.7. *Let $\mathcal{G}$ be an LN map and $\mathcal{F}$ be a log-contractive map. Then, their sum $\mathcal{G} + \mathcal{F}$ is log-contractive.*

Proof. We start by writing the Thompson metric in an alternative form [32]:

$$\delta_T(A, B) = \max\{\log W(A/B), \log W(B/A)\}, \tag{4.15}$$

where $W(A/B) := \inf\{\lambda > 0, A \preceq \lambda B\}$. Let $\lambda = \exp(\delta_T(X, Y))$; then it follows that $X \preceq \lambda Y$. Since $\mathcal{G}$ is nonexpansive in δ_T, using (4.15) it further follows that

$$\mathcal{G}(X) \quad \preceq \quad \lambda \mathcal{G}(Y),$$

and $\mathcal{F}$ is log-contractive map; we obtain the inequality

$$\mathcal{F}(X) \quad \prec \quad \lambda^t \mathcal{F}(Y), \qquad \text{where } t \le 1.$$

Write $\mathcal{H} := \mathcal{G} + \mathcal{F}$; then, we have the following inequalities:

$$\begin{aligned}
\mathcal{H}(X) &\prec \lambda \mathcal{H}(Y) + (\lambda^t - \lambda)\mathcal{F}(Y),\\
\mathcal{H}(Y)^{-1/2}\mathcal{H}(X)\mathcal{H}(Y)^{-1/2} &\prec \lambda I + (\lambda^t - \lambda)\mathcal{H}(Y)^{-1/2}\mathcal{F}(Y)\mathcal{H}(Y)^{-1/2},\\
\mathcal{H}(Y)^{-1/2}\mathcal{H}(X)\mathcal{H}(Y)^{-1/2} &\prec \lambda I + (\lambda^t - \lambda)\lambda_{\min}(\mathcal{H}(Y)^{-1/2}\mathcal{F}(Y)\mathcal{H}(Y)^{-1/2})I.
\end{aligned}$$

As $\lambda_{\max}(\mathcal{H}(Y)^{-1/2}\mathcal{H}(X)\mathcal{H}(Y)^{-1/2}) > \lambda_{\max}(\mathcal{H}(X)^{-1/2}\mathcal{H}(Y)\mathcal{H}(X)^{-1/2})$, using (4.15) we obtain

$$\delta_T(\mathcal{H}(X), \mathcal{H}(Y)) < \delta_T(X, Y) + \log\big(1 + \lambda_{\min}(\mathcal{H}(Y)^{-1/2}\mathcal{F}(Y)\mathcal{H}(Y)^{-1/2})\left[\lambda^{t-1} - 1\right]\big). \tag{4.16}$$

We also have the following eigenvalue inequality:

$$\lambda_{\min}(\mathcal{H}(Y)^{-1/2}\mathcal{F}(Y)\mathcal{H}(Y)^{-1/2}) \le \frac{\lambda_{\min}(\mathcal{F}(Y))}{\lambda_{\max}(\mathcal{G}(Y)) + \lambda_{\min}(\mathcal{F}(Y))}. \tag{4.17}$$

Combining inequalities (4.16) and (4.17), we see that

$$\delta_T(\mathcal{H}(X), \mathcal{H}(Y)) < \delta_T(X, Y) + \log\big(1 + \tfrac{\lambda_{\min}(\mathcal{F}(Y))}{\lambda_{\max}(\mathcal{G}(Y)) + \lambda_{\min}(\mathcal{F}(Y))}\left[\exp(\delta_T(X, Y))^{t-1} - 1\right]\big). \tag{4.18}$$

Similarly, since $\lambda_{\max}\big(\mathcal{H}(Y)^{-1/2}\mathcal{H}(X)\mathcal{H}(Y)^{-1/2}\big) < \lambda_{\max}\big(\mathcal{H}(X)^{-1/2}\mathcal{H}(Y)\mathcal{H}(X)^{-1/2}\big)$, we also obtain the bound (notice that we now have $\mathcal{F}(X)$ instead of $\mathcal{F}(Y)$)

(4.19)
$$\delta_T(\mathcal{H}(X),\mathcal{H}(Y)) < \delta_T(X,Y) + \log\big(1 + \tfrac{\lambda_{\min}(\mathcal{F}(X))}{\lambda_{\max}(\mathcal{G}(X))+\lambda_{\min}(\mathcal{F}(X))}\big[\exp(\delta_T(X,Y))^{t-1}-1\big]\big).$$

Combining (4.18) and (4.19) into a single inequality, we get

$$\delta_T(\mathcal{H}(X),\mathcal{H}(Y)) < \delta_T(X,Y) + \log\big(1 + \tfrac{\lambda_{\min}(\mathcal{F}(X),\mathcal{F}(Y))}{\lambda_{\max}(\mathcal{G}(X),\mathcal{G}(Y))+\lambda_{\min}(\mathcal{F}(X),\mathcal{F}(Y))}\big[\exp(\delta_T(X,Y))^{t-1}-1\big]\big).$$

As the second term is ≤ 0, the inequality is strict, proving log-contractivity of $\mathcal{H}$. ∎

Using log-contractivity, we can finally state our main result for this section.

THEOREM 4.8. *If $\mathcal{G}$ is log-contractive and* (4.3) *has a solution, then this solution is unique and iteration* (4.4) *converges to it.*

Proof. If (4.22) has a solution, then from a theorem of [20] it follows that the log-contractive map $\mathcal{G}$ yields iterates that stay within a compact set and converge to a unique fixed point of $\mathcal{G}$. This fixed point is positive definite by construction (starting from a positive definite matrix, none of the operations in (4.22) violates positivity). Thus, the unique solution is positive definite. ∎

4.2. Example of LN optimization. To illustrate how to exploit LN functions for optimization, let us consider the following minimization problem:

$$\min_{S\succ 0} \quad \Phi(S) := \tfrac{1}{2}n\log\det(S) - \sum\nolimits_i \log\varphi(x_i^T S^{-1}x_i), \tag{4.20}$$

which arises in maximum-likelihood estimation of ECDs (see section 5 for further examples and details) and also M-estimation of the scatter matrix [27].

The first-order necessary optimality condition for (4.20) stipulates that a candidate solution $S \succ 0$ must satisfy

$$\frac{\partial\Phi(S)}{\partial S} = 0 \quad \Longleftrightarrow \quad \tfrac{1}{2}nS^{-1} + \sum_{i=1}^{n}\frac{\varphi'(x_i^T S^{-1}x_i)}{\varphi(x_i^T S^{-1}x_i)}S^{-1}x_i x_i^T S^{-1} = 0. \tag{4.21}$$

Defining $h \equiv -\varphi'/\varphi$, (4.21) may be rewritten more compactly in matrix notation as the equation

$$S = \tfrac{2}{n}\sum\nolimits_{i=1}^{n} x_i h(x_i^T S^{-1}x_i)x_i^T = \tfrac{2}{n}Xh(D_S)X^T, \tag{4.22}$$

where $h(D_S) := \mathrm{Diag}(h(x_i^T S^{-1}x_i))$, and $X = [x_1,\ldots,x_m]$. We then solve the nonlinear equation (4.22) via a fixed-point iteration. Introducing the nonlinear map $\mathcal{G}: \mathbb{P}_d \to \mathbb{P}_d$ that maps S to the right-hand side of (4.22), we use fixed-point iteration (4.4) to find the solution. In order to show that the Picard iteration converges (to the unique fixed point), it is enough to show that $\mathcal{G}$ is log-contractive (see Theorem 4.8). The following proposition gives a sufficient condition on h, under which the map is log-contractive.

PROPOSITION 4.9. *Let h be LN. Then, the map $\mathcal{G}$ in* (4.4) *is LN. Moreover, if h is log-contractive, then $\mathcal{G}$ is log-contractive.*

Proof. Let $S, R \succ 0$ be arbitrary. Then, we have the following chain of inequalities:

$$\begin{aligned}\delta_T(\mathcal{G}(S),\mathcal{G}(R)) &= \delta_T\big(\tfrac{2}{n}Xh(D_S)X^T,\ \tfrac{2}{n}Xh(D_R)X^T\big)\\ &\leq \delta_T\big(h(D_S),h(D_R)\big) \quad\leq\quad \max_{1\leq i\leq n}\delta_T\big(h(x_i^T S^{-1}x_i),h(x_i^T R^{-1}x_i)\big)\\ &\leq \max_{1\leq i\leq n}\delta_T\big(x_i^T S^{-1}x_i, x_i^T R^{-1}x_i\big) \quad\leq\quad \delta_T\big(S^{-1},R^{-1}\big) = \delta_T(S,R).\end{aligned}$$

The first inequality follows from (4.7b) and Theorem 4.3; the second inequality follows since $h(D_S)$ and $h(D_R)$ are diagonal; the third follows from (4.7d); and the fourth follows from another application of Theorem 4.3, while the final equality is via (4.7a). This proves log-nonexpansivity (i.e., nonexpansivity in δ_T). If in addition h is log-contractive and $S \neq R$, then the second inequality above is strict; that is,

$$\delta_T(\mathcal{G}(S), \mathcal{G}(R)) < \delta_T(S, R) \quad \forall S, R \quad \text{and} \quad S \neq R. \qquad \square$$

If h is merely LN (not log-contractive), it is still possible to show uniqueness of (4.22) up to a constant. Our proof depends on the compression property of δ_T proved in Theorem 4.3.

THEOREM 4.10. *Let the data $\mathcal{X} = \{x_1, \ldots, x_n\}$ span the whole space. If h is LN, and $S_1 \neq S_2$ are solutions to* (4.22)*, then iteration* (4.4) *converges to a solution, and $S_1 \propto S_2$.*

Proof. Without loss of generality, assume that $S_1 = I$. Let $S_2 \neq cI$. Theorem 4.3 implies that

$$\begin{aligned}
&\delta_T\big(x_i h(x_i^T S_2^{-1} x_i) x_i^T, x_i h(x_i^T S_1^{-1} x_i) x_i\big) \\
&\le \delta_T\big(h(x_i^T S_2^{-1} x_i), h(x_i^T x_i)\big) \le \delta_T\big(x_i^T S_2^{-1} x_i, x_i^T x_i\big) = \left|\log \tfrac{x_i^T S_2^{-1} x_i}{x_i^T x_i}\right|.
\end{aligned}$$

As per assumption, the data span the whole space. Since $S_2 \neq cI$, we can find x_1 such that

$$\left|\log \tfrac{x_1^T S_2^{-1} x_1}{x_1^T x_1}\right| < \delta_T(S_2, I).$$

Therefore, we obtain the following inequality for point x_1:

$$\delta_T\big(x_1 h(x_i^T S_2^{-1} x_1) x_1^T, x_1 h(x_1^T S_1^{-1} x_1) x_1\big) < \delta_T(S_2, S_1). \tag{4.23}$$

Using Proposition 4.9 and invoking Theorem 4.7, it then follows that

$$\delta_T(\mathcal{G}(S_2), \mathcal{G}(S_1)) < \delta_T(S_2, S_1).$$

But this means that S_2 cannot be a solution to (4.22), which is a contradiction. Therefore, $S_2 \propto S_1$. $\square$

4.2.1. Computational efficiency. So far we have not addressed computational efficacy of the fixed-point algorithm. The rate of convergence depends heavily on the contraction factor, and, as we will see in the experiments, without further care one obtains poor contraction factors that can lead to a very slow convergence. We briefly discuss below a useful speedup technique that seems to have a dramatic impact on the empirical convergence speed (see Figure 2).

At the fixed point S^* we have $\mathcal{G}(S^*) = S^*$, or equivalently for a new map $\mathcal{M}$ we have

$$\mathcal{M}(S^*) := {S^*}^{-1/2} \mathcal{G}(S^*) {S^*}^{-1/2} = I.$$

Therefore, one way to analyze the convergence behavior is to assess how fast $\mathcal{M}(S_k)$ converges to identity. Using the theory developed beforehand, it is easy to show that

$$\delta_T(M(S_{k+1}), I) \le \eta \delta_T(M(S_k), I),$$

where η is the contraction factor between S_k and S_{k+1}, so that

$$\delta_T(\mathcal{G}(S_{k+1}), \mathcal{G}(S_k)) < \eta \delta_T(S_{k+1}, S_k).$$

To increase the convergence speed we may replace S_{k+1} by its scaled version $\alpha_k S_{k+1}$ such that

$$\delta_T(\mathcal{M}(\alpha_k S_{k+1}), I) \leq \delta_T(\mathcal{M}(S_{k+1}), I).$$

One can do a search to find a good α_k. Clearly, the sequence $S_{k+1} = \alpha_k \mathcal{G}(S_k)$ converges at a faster pace. We will see in the numerical results section that scaling with α_k has a remarkable effect on the convergence speed. An intuitive reason why this happens is that the additional scaling factor can resolve the problematic cases where the contraction factor becomes small. These problematic cases are those where both the smallest and the largest eigenvalues of $\mathcal{M}(S_k)$ become smaller (or larger) than one, whereby the contraction factor (for $\mathcal{G}$) becomes small, which may lead to a very slow convergence. The scaling factor, however, makes the smallest eigenvalues of $\mathcal{M}(S_k)$ always smaller and its largest eigenvalue larger than one. One way to avoid the search is to choose α_k such that $\operatorname{trace}(\mathcal{M}(S_{k+1})) = d$—though with a small caveat: empirically this simple choice of α_k works very well, but our convergence proof no longer holds. Extending our convergence theory to incorporate this specific choice of scaling α_k is a part of our future work. In all simulations in the result section, α_k is selected by ensuring $\operatorname{trace}(\mathcal{M}(S_{k+1})) = d$.

5. Application to elliptically contoured distributions. In this section we present details for a concrete application of conic geometric optimization: mle for ECDs [13, 21, 37]. We use ECDs as a platform for illustrating geometric optimization because ECDs are widely important (see, e.g., the survey [42]) and are instructive in illustrating our theory.

First, we give some basics. If an ECD has density on $\mathbb{R}^d$, it assumes the form[3]

$$\forall\, x \in \mathbb{R}^d, \qquad \mathscr{E}_\varphi(x; S) \propto \det(S)^{-1/2} \varphi(x^T S^{-1} x), \tag{5.1}$$

where $S \in \mathbb{P}_d$ is the scatter matrix and $\varphi : \mathbb{R} \to \mathbb{R}_{++}$ is the *density generating function* (dgf). If the ECD has finite covariance, then the scatter matrix is proportional to the covariance matrix [13].

Example 5.1. Let $\varphi(t) = e^{-t/2}$; then (5.1) reduces to the multivariate Gaussian density. For

$$\varphi(t) = t^{\alpha - d/2} \exp\bigl(-(t/b)^\beta\bigr), \tag{5.2}$$

where α, b, $\beta > 0$ are fixed, density (5.1) yields the rich class called *Kotz-type distributions* that have powerful modeling abilities [26, section 3.2]; they include as special cases multivariate power exponentials, elliptical gamma, and multivariate W-distributions, for instance. Other examples include multivariate student-t, multivariate logistic, and Weibull dgfs (see section 5.2).

5.1. Maximum likelihood parameter estimation. Let $(x_1, \ldots, x_n)$ be i.i.d. samples from an ECD $\mathscr{E}_\varphi(S)$. Ignoring constants, the log-likelihood is

$$\mathcal{L}(x_1, \ldots, x_n; S) = -\tfrac{1}{2} n \log\det S + \sum\nolimits_{i=1}^{n} \log \varphi(x_i^T S^{-1} x_i). \tag{5.3}$$

[3]For simplicity we describe only mean zero families; the extension to the general case is easy.

To compute an mle we equivalently consider the minimization problem (4.20), which we restate here for convenience:

$$\min_{S \succ 0} \quad \Phi(S) := \tfrac{1}{2} n \log\det(S) - \sum\nolimits_i \log \varphi(x_i^T S^{-1} x_i). \tag{5.4}$$

Unfortunately, (5.4) is in general very difficult: Φ may be nonconvex and may have multiple local minima (observe that $\log\det(S)$ is concave in S and we are minimizing). Since statistical estimation relies on having access to globally optimal estimates, it is important to be able to solve (5.4) globally. These difficulties notwithstanding, using our theory we identify a rich class of ECDs for which we can solve (5.4) globally. Some examples are already known [42, 27, 51], but our techniques yield results that are strictly more general: they subsume previous examples while advancing the broader idea of geometric optimization over HPD matrices.

Building on sections 2 and 4, we divide our study into the following three classes of dgfs:

(i) Geodesically convex (g-convex): This class contains functions for which the negative log-likelihood $\Phi(S)$ is g-convex. Some members of this class have been previously studied (though sometimes without recognizing or directly exploiting g-convexity).

(ii) Log-nonexpansive (LN): This is a new class introduced in this paper. It exploits the "nonpositive curvature" property of the HPD manifold. To the best of our knowledge, this class of ECDs was beyond the grasp of previous methods [51, 27, 49]. The iterative algorithm for finding the global minimum of the objective is similar to that of the class LC.

(iii) Log-convex (LC): We cover this class for completeness; it covers the case of log-convex φ but leads to nonconvex Φ (due to the $-\log\varphi$ term). However, the structure of the problem is such that one can derive an efficient algorithm for finding a local minumum of the objective function.

As illustrated in Figure 1, these classes can overlap. When a function is in the overlap between LC and GC, one can be sure that the iterative algorithm derived for class LN will converge to a unique minimum. Table 1 summarizes the applicability of fixed-point or manifold optimization methods on different classes of dgfs.

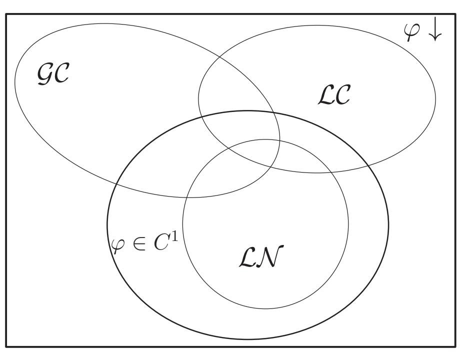

Fig. 1. *Overview of dgf classes for nonincreasing φ.*

5.2. mle for distributions in class g-convex. If the log-likelihood is strictly g-convex, then (5.4) cannot have multiple solutions. Moreover, for any local optimization method that ensures a local solution to (5.4), g-convexity ensures that this solution is globally optimal.

TABLE 1

Applicability of the different algorithms: **Yes** *means a preferred algorithm;* Can* *denotes applicability on a case-by-case basis; and* Can *signifies possible applicability of method.*

Problem class	Manifold opt.	Fixed-point
$\mathcal{GC}$	**Yes**	Can*
$\mathcal{LN}$	Can	**Yes**
$\mathcal{LC}$	Can	**Yes**

First we state a corollary of Theorem 2.14 that helps us recognize g-convexity of ECDs. We remark that a result equivalent to Corollary 5.2 was also recently discovered in [51]. Theorem 2.14 is more general and uses a completely different argument founded on matrix-theoretic results.

COROLLARY 5.2. *Let $h : \mathbb{R}_{++} \to \mathbb{R}$ be g-convex (i.e., $h(x^{1-\lambda}y^{\lambda}) \le (1-\lambda)h(x) + \lambda h(y)$). If h is nondecreasing, then for $r \in \{\pm 1\}$, $\phi : \mathbb{P}_d \to \mathbb{R} : S \mapsto \sum_i h(x_i^T S^r x_i) \pm \log\det(S)$ is g-convex. Furthermore, if h is strictly g-convex, then ϕ is also strictly g-convex.*

Proof. The proof is immediate from Theorem 2.14 since $x_i^T S^r x_i$ is a positive linear map. ∎

For reference, we summarize several examples of strictly g-convex ECDs in Corollary 5.3.

COROLLARY 5.3. *The negative log-likelihood* (5.4) *is strictly g-convex for the following distributions:* (i) *Kotz with $\alpha \le \frac{d}{2}$ (its special cases include Gaussian, multivariate power exponential, multivariate W-distribution with shape parameter smaller than one, and elliptical gamma with shape parameter $\nu \le \frac{d}{2}$);* (ii) *multivariate-t;* (iii) *multivariate Pearson type II with positive shape parameter; and* (iv) *elliptical multivariate logistic distribution.*[4]

Even though g-convexity ensures that every local solution will be globally optimal, we must first ensure that there exists a solution at all; that is, *does* (5.4) *have a solution?* Answering this question is nontrivial even in special cases [27, 51]. We provide below a fairly general result that helps establish existence.

THEOREM 5.4. *Let $\Phi(S)$ satisfy the following:* (i) *$-\log\varphi(t)$ is lower semicontinuous (lsc) for $t > 0$, and* (ii) *$\Phi(S) \to \infty$ as $\|S\| \to \infty$ or $\|S^{-1}\| \to \infty$; then $\Phi(S)$ attains its minimum.*

Proof. Consider the metric space $(\mathbb{P}_d, d_R)$, where d_R is the Riemannian distance,

$$d_R(A, B) = \|\log(A^{-1/2}BA^{-1/2})\|_{\mathrm{F}}, \qquad A, B \in \mathbb{P}_d. \tag{5.5}$$

If $\Phi(S) \to \infty$ as $\|S\| \to \infty$ or as $\|S^{-1}\| \to \infty$, then $\Phi(S)$ has bounded lower-level sets in $(\mathbb{P}_d, d_R)$. It is a well-known result in variational analysis that an lsc function which has bounded lower-level sets in a metric space attains its minimum [47]. Since $-\log\varphi(t)$ is lsc and $\log\det(S^{-1})$ is continuous, $\Phi(S)$ is lsc on $(\mathbb{P}_d, d_R)$. Therefore, it attains its minimum. ∎

A key consequence of this theorem is its utility is in showing existence of solutions to (5.4) for a variety of different ECDs. We show an example application to Kotz-type distributions [26, 28] below. For these distributions, the function $\Phi(S)$ assumes the

[4]The dgfs of different distributions are brought here for the reader's convenience. Multivariate power exponential: $\phi(t) = \exp(-t^{\nu}/b)$, $\nu > 0$. Multivariate W-distribution: $\phi(t) = t^{\nu-1}\exp(-t^{\nu}/b)$, $\nu > 0$. Elliptical gamma: $\phi(t) = t^{\nu-d/2}\exp(-t/b)$, $\nu > 0$. Multivariate t: $\phi(t) = (1 + t/\nu)^{-(\nu+d)/2}$, $\nu > 0$. Multivariate Pearson type II: $\phi(t) = (1-t)^{\nu}$, $\nu > -1, 0 \le t \le 1$. Elliptical multivariate logistic: $\phi(t) = \exp(-\sqrt{t})/(1+\exp(-\sqrt{t}))^2$.

form

$$K(S) = \tfrac{n}{2}\log\det(S) + (\tfrac{d}{2}-\alpha)\sum_{i=1}^{n}\log x_i^T S^{-1}x_i + \sum_{i=1}^{n}\left(\tfrac{x_i^T S^{-1} x_i}{b}\right)^{\beta}. \tag{5.6}$$

Lemma 5.5 shows that $K(S) \to \infty$ whenever $\|S^{-1}\| \to \infty$ or $\|S\| \to \infty$.

LEMMA 5.5. *Let the data $\mathcal{X} = \{x_1, \ldots, x_n\}$ span the whole space and for $\alpha < \frac{d}{2}$ satisfy*

$$\frac{|\mathcal{X} \cap L|}{|\mathcal{X}|} < \frac{d_L}{d - 2\alpha}, \tag{5.7}$$

where L is an arbitrary subspace with dimension $d_L < d$ and $|\mathcal{X} \cap L|$ is the number of datapoints that lie in the subspace L. If $\|S^{-1}\| \to \infty$ or $\|S\| \to \infty$, then $K(S) \to \infty$.

Proof. If $\|S^{-1}\| \to \infty$ and since the data span the whole space, it is possible to find a datum x_1 such that $t_1 = x_1^T S^{-1} x_1 \to \infty$. Since

$$\lim_{t\to\infty} c_1 \log(t) + t^{c_2} + c_3 \to \infty$$

for constants c_1, c_3, and $c_2 > 0$, it follows that $K(S) \to \infty$ whenever $\|S^{-1}\| \to \infty$.

If $\|S\| \to \infty$ and $\|S^{-1}\|$ is bounded, then the third term in expression of $K(S)$ is bounded. Assume that d_L is the number of eigenvalues of S that go to ∞ and $|\mathcal{X} \cap L|$ is the number of data that lie in the subspace spanned by these eigenvalues. Then in the limit when eigenvalues of S go to ∞, $K(S)$ converges to the following limit:

$$\lim_{\lambda\to\infty} \tfrac{n}{2} d_L \log\lambda + (\tfrac{d}{2} - \alpha)|\mathcal{X} \cap L| \log\lambda^{-1} + c.$$

Apparently if $\frac{n}{2}d_L + (\frac{d}{2} - \alpha)|\mathcal{X} \cap L| > 0$, then $K(S) \to \infty$, and the proof is complete. ∎

It is important to note that overlap condition (5.7) can be fulfilled easily by assuming that the number of datapoints is larger than their dimensionality and that they are noisy. Using Lemma 5.5 with Theorem 5.4, we obtain the following key result for Kotz-type distributions.

THEOREM 5.6 (mle existence). *If the data samples satisfy condition* (5.7)*, then the log-likelihood of Kotz-type distribution has a maximizer (i.e., there exists an mle).*

5.2.1. Optimization algorithm. Once existence is ensured, one may use any local optimization method to minimize (5.4) to obtain the desired mle. For members of the class g-convex that do not lie in class LN or class LC, we recommend invoking the manifold optimization techniques summarized in section 3.

5.3. mle for distributions in class LN. For negative log-likelihoods (5.4) in class LN, we can circumvent the heavy machinery of manifold optimization and obtain simple fixed-point algorithms by appealing to the contraction results developed in section 4. We note that some members of class g-convex may also turn out to lie in class LN, so the discussion below also applies to them.

As an illustrative example of these results, consider the problem of finding the minimum of negative log-likelihood solution of Kotz-type distribution (5.6). If the corresponding nonlinear equation (4.22) with corresponding $h(.) = (\frac{d}{2} - \alpha)(.)^{-1} + \frac{\beta}{b^\beta}(.)^{\beta-1}$ has a positive definite solution, then it is a candidate mle; if it is unique, then it is the desired solution to (5.6).

But how should we solve (4.22)? This is where the theory developed in section 4 comes into play. Convergence of the iteration (4.4) as applied to (4.22) can be obtained from Theorem 4.10. But in the Kotz case we can actually show a stronger result that helps ensure better geometric convergence rates for the fixed-point iteration.

LEMMA 5.7. *If* $c \geq 0$ *and* $-1 < \tau < 1$, *then* $g(x) = cx + x^\tau$ *is log-contractive.*

Proof. Without loss of generality, assume $t = ks$ with $k \geq 1$. Assume that $g(t) \geq g(s)$:

$$\begin{aligned}
\log g(t) &= \log(ct + t^\tau) \\
&= \log(kcs + k^\tau s^\tau) \\
&= \quad \log(k(cs + s^\tau) + k^\tau s^\tau - ks^\tau) \\
&= \log k(cs + s^\tau)\Big(1 + \frac{k^\tau s^\tau - ks^\tau}{k(cs + s^\tau)}\Big) \\
&= \log k + \log g(s) + \log\Big(1 + \frac{s^\tau(k^{\tau-1} - 1)}{(cs + s^\tau)}\Big), \\
|\log g(t) - \log g(s)| &= |\log t - \log s| + \log\Big(1 + \frac{s^\tau(k^{\tau-1} - 1)}{(cs + s^\tau)}\Big).
\end{aligned}$$

Since the second term is negative, g is log-contractive. Consider the other case, $g(t) \geq g(s)$, that could happen only when $\tau \leq 0$:

$$\begin{aligned}
\log g(s) &= \log(cs + s^\tau) \\
&= \log(ct/k + k^{|\tau|}t^\tau) \\
&= \log(k(ct + t^\tau) + k^{|\tau|}t^\tau + ct/k - ckt - kt^\tau) \\
&= \quad \log k(ct + t^\tau)\Big(1 + \frac{k^{|\tau|}t^\tau + ct/k - ckt - kt^\tau}{k(ct + t^\tau)}\Big) \\
&= \log k + \log g(t) + \log\Big(1 + \frac{ct\left(k^{-2} - 1\right) + t^\tau(k^{|\tau|-1} - 1)}{(ct + t^\tau)}\Big), \\
|\log g(t) - \log g(s)| &= |\log t - \log s| + \log\Big(1 + \frac{ct\left(\frac{1}{k^2} - 1\right) + t^\tau(k^{|\tau|-1} - 1)}{(ct + t^\tau)}\Big).
\end{aligned}$$

In this case, the second term is also negative. Therefore, h is log-contractive. ∎

We assume that $\tau = \beta - 1$ and $c = \frac{b^\beta(d/2-\alpha)}{\beta}$; knowing that $h(.) = g(\beta b^{-\beta}(.))$ has the same contraction factor as $g(.)$, we infer from Lemma 5.7 that h in the iteration (4.22) for the Kotz-type distributions with $0 < \beta < 2$ and $\alpha \leq \frac{d}{2}$ is log-contractive. Based on Theorem 5.6, $K(S)$ has at least one minimum. Thus, using Theorem 4.8, we have the following main convergence result.

THEOREM 5.8. *If the data samples satisfy* (5.7), *then iteration* (4.22) *for Kotz-type distributions with* $0 < \beta < 2$ *and* $\alpha \leq \frac{d}{2}$ *converges to a unique fixed point.*

5.4. mle for distributions in class LC. For completeness, we briefly mention class LC here, which is perhaps one of the most studied classes of ECDs, at least from an algorithmic point of view [27]. Therefore, we only discuss it summarily and present our new results.

For the class LC, we assume that the dgf φ is log-convex. Without the assumptions that are typically made in the literature, it can be that neither the GC nor the LN analysis applies to class LC. However, the optimization problem still has structure that allows simple and efficient algorithms. Specifically, here the objective function $\Phi(S)$ can be written as a difference of two convex functions by introducing the variable $P = S^{-1}$, wherewith we have $\Phi(P) = -an \log\det(P) - \sum_i \log\varphi(x_i^T P x_i)$.

To this representation of Φ we may now apply the convex-concave procedure (CCCP) [50] to search for a locally optimal point. The method operates as follows:

$$P^{k+1} \leftarrow \underset{P \succ 0}{\operatorname{argmin}} \quad -\tfrac{n}{2} \log\det(P) + \operatorname{tr}\Big(P \sum\nolimits_i h(x_i^T P^k x_i) x_i x_i^T\Big),$$

which yields the update

$$P^{k+1} = \Big(\tfrac{2}{n} \sum\nolimits_i h(x_i^T P^k x_i) x_i x_i^T\Big)^{-1}. \tag{5.8}$$

Because P^{k+1} is constructed using the CCCP procedure, it can be shown that the sequence $\{\Phi(P^k)\}$ is monotonically decreasing. Furthermore, since we assumed h to be nonnegative, the iteration stays within the positive semidefinite cone. If the cost function goes to infinity whenever the covariance matrix is singular, then using Theorem 5.4 we can conclude that iteration converges to a positive definite matrix. Thus, we can state the following key result for class LC.

THEOREM 5.9 (convergence). *Assume that $\Phi(P)$ goes to infinity whenever P reaches the boundary of $\mathbb{P}^d$, i.e., $\|P\| \to \infty \vee \|P^{-1}\| \to \infty \implies \Phi(P) \to \infty$. Furthermore if $-\log\varphi$ is concave and h is nonnegative, then each step of the iterative algorithm given in* (5.8) *decreases the cost function, and furthermore it converges to a positive definite solution.*

A similar theorem, but under stricter conditions, was established in [27]. Knowing that the iterative algorithm in (5.8) is the same as (4.22) and using Theorem 5.9 with the existence result of Theorem 5.6 and the uniqueness result of Corollary 5.3, we can state the following theorem for Kotz-type distributions (cf. Theorem 5.8).

THEOREM 5.10. *If the data samples satisfy condition* (5.7)*, then iteration* (4.22) *for Kotz-type distributions with $\beta \geq 1$ and $\alpha \leq \frac{d}{2}$ converges to a unique fixed point.*

Theorems 5.10 and 5.8 together show that the iteration (4.22) for Kotz-type distributions with $\alpha \leq \frac{d}{2}$ and regardless of the value of β always converges to the unique mle whenever it exists.

6. Numerical results. We briefly illustrate the numerical performance of our fixed-point iteration. The key message here is that our fixed-point iterations solve nonconvex problems that are further complicated by a positive definiteness constraint. But by construction the fixed-point iterations satisfy the constraint, so no extra eigenvalue computation is needed, which can provide substantial computational savings. In contrast, a general nonlinear solver must perform constrained optimization, which may be unduly expensive.

We show two experiments (Figures 2 and 3) to demonstrate the scalability of the fixed-point iteration with increasing dimensionality of the input matrix and for varying β parameter of the Kotz distribution which influences convergence rate of our fixed-point iteration. For all simulations, we sampled 10,000 datapoints from the Kotz-type distribution with given α and β parameters and a random covariance matrix.

We note that the problems are nonconvex with an open set as a constraint—this precludes direct application of semidefinite programming or approaches such as gradient-projection (projection requires closed sets). We also tried interior-point methods, but we did not include them in the comparisons because of their extremely slow convergence speed on this problem. So we choose to show the result of (Riemannian) manifold optimization techniques [1].

We compare our fixed-point iteration against four different manifold optimization methods: (i) steepest descent (SD); (ii) conjugate gradients (CG); (iii) trust-region (TR); and (iv) limited-memory RBFGS (denoted as LBFGS below), which implements Algorithm 1. All methods are implemented in MATLAB (including the fixed-point iteration); for manifold optimization we extend the MANOPT toolbox [10] to support the HPD manifold[5] as well as Algorithm 1.

From Figure 2 we see that the basic fixed-point algorithm (FP) does not perform better than SD, the simplest manifold optimization method. Moreover, even when FP performs better than CG, TR, or LBFGS (Figure 3), it seems to closely follow SD. However, the scaling idea introduced in section 4.2 leads to a fixed-point method (FP2) that outperforms all other methods, both with increasing dimensionality and varying β. The scale is chosen by ensuring $\text{trace}(\mathcal{M}(S_{k+1})) = d$.

These results merely indicate that the fixed-point approach can be competitive. A more thorough experimental study to assess our algorithms remains to be undertaken.

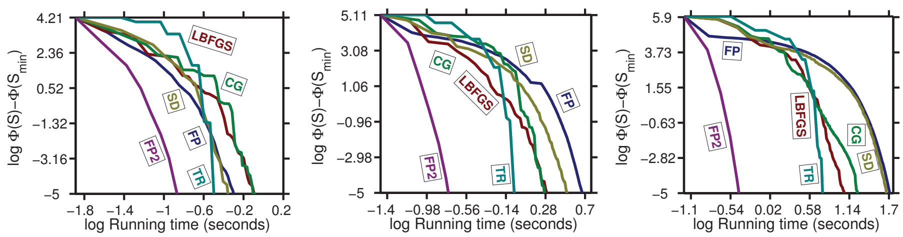

FIG. 2. *Comparison of the running times of the fixed-point iterations and four different manifold optimization techniques to maximize a Kotz-likelihood with $\beta = 0.5$ and $\alpha = 1$ (see text for details). FP denotes normal fixed-point iteration, and FP2 is the fixed-point iteration with scaling factor. Manifold optimization methods are steepest descent (SD), conjugate gradient (CG), limited-memory RBFGS (LBFGS), and trust-region (TR) . The plots show (from left to right) running times for estimating $S \in \mathbb{P}_d$ for $d \in \{4, 16, 64\}$.*

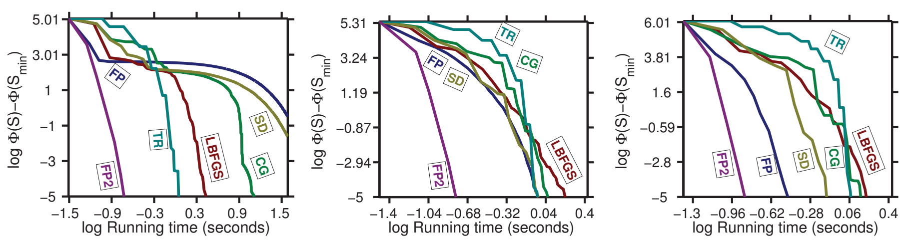

FIG. 3. *In the Kotz-type distribution, when β gets close to zero or 2 or when α gets close to zero, the contraction factor becomes smaller, which can impact the convergence rate. This figure shows running time variance for Kotz-type distributions with $d = 16$ and $\alpha = 2\beta$ for different values of $\beta \in \{0.1, 1, 1.7\}$.*

7. Conclusion. We studied geometric optimization for minimizing certain nonconvex functions over the set of positive definite matrices. We showed key results that help us recognize geodesic convexity; we also introduced a new class of log-nonexpansive functions which contains functions that need not be geodesically convex

[5]The newest version of the MANOPT toolbox ships with an implementation of the HPD manifold, but we use our own implementation as it includes some utilities specific to LBFGS.

but can still be optimized efficiently. Key to our ideas was a construction of fixed-point iterations in a suitable metric space on positive definite matrices.

Additionally, we developed and applied our results in the context of maximum likelihood estimation for elliptically contoured distributions, covering instances substantially beyond the state-of-the-art. We believe that the general geometric optimization techniques that we developed in this paper will prove to be of wider use and interest beyond our motivating examples and applications. Moreover, developing a more extensive geometric optimization numerical package is an ongoing project.

REFERENCES

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2009.

[2] T. Ando, *Concavity of certain maps of positive definite matrices and applications to Hadamard products*, Linear Algebra Appl., 26 (1979), pp. 203–221.

[3] T. Ando, C.-K. Li, and R. Mathias, *Geometric means*, Linear Algebra Appl., 385 (2004), pp. 305–334.

[4] M. Arnaudon, F. Barbaresco, and L. Yang, *Riemannian medians and means with applications to radar signal processing*, IEEE J. Selected Topics Signal Process., 7 (2013), pp. 595–604.

[5] R. Bhatia, *Matrix Analysis*, Springer, New York, 1997.

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

[7] R. Bhatia and R. L. Karandikar, *Monotonicity of the matrix geometric mean*, Math. Ann., 353 (2012), pp. 1453–1467.

[8] D. A. Bini and B. Iannazzo, *Computing the Karcher mean of symmetric positive definite matrices*, Linear Algebra Appl., 438 (2013), pp. 1700–1710.

[9] G. Blekherman, P. A. Parrilo, and R. R. Thomas, eds., *Semidefinite Optimization and Convex Algebraic Geometry*, SIAM, Philadelphia, 2013.

[10] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, *Manopt, a MATLAB toolbox for optimization on manifolds*, J. Machine Learning Res., 15 (2014), pp. 1455–1459.

[11] S. Boyd, S.-J. Kim, L. Vandenberghe, and A. Hassibi, *A tutorial on geometric programming*, Optimization Engrg., 8 (2007), pp. 67–127.

[12] M. R. Bridson and A. Haeflinger, *Metric Spaces of Non-Positive Curvature*, Springer, New York, 1999.

[13] S. Cambanis, S. Huang, and G. Simons, *On the theory of elliptically contoured distributions*, J. Multivariate Anal., 11 (1981), pp. 368–385.

[14] Z. Chebbi and M. Moahker, *Means of Hermitian positive-definite matrices based on the log-determinant α-divergence function*, Linear Algebra Appl., 436 (2012), pp. 1872–1889.

[15] Y. Chen, A. Wiesel, and A. O. Hero, *Robust shrinkage estimation of high-dimensional covariance matrices*, IEEE Trans. Signal Process., 59 (2011), pp. 4097–4107.

[16] G. Cheng, H. Salehian, and B. C. Vemuri, *Efficient recursive algorithms for computing the mean diffusion tensor and applications to DTI segmentation*, in Proceedings of the European Conference on Computer Vision (ECCV), Vol. 7, Springer, New York, 2012, pp. 390–401.

[17] G. Cheng and B. C. Vemuri, *A novel dynamic system in the space of SPD matrices with applications to appearance tracking*, SIAM J. Imaging Sci., 6 (2013), pp. 592–615.

[18] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos, *Jensen-Bregman LogDet divergence for efficient similarity computations on positive definite tensors*, IEEE Trans. Pattern Anal. Machine Intell., 35 (2013), pp. 2161–2174.

[19] M.-D. Choi, *Completely positive linear maps on complex matrices*, Linear Algebra Appl., 10 (1975), pp. 285–290.

[20] M. Edelstein, *On fixed and periodic points under contractive mappings*, J. London Math. Soc., s1-37 (1962), pp. 74–79.

[21] A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions*, Chapman and Hall/CRC, Boca Raton, FL, 1999.

[22] L. Gurvits and A. Samorodnitsky, *A deterministic algorithm for approximating mixed discriminant and mixed volume, and a combinatorial corollary*, Discrete Comput. Geom., 27 (2002), pp. 531–550.

[23] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Some simple inequalities satisfied by convex functions*, Messenger Math., 58 (1929), pp. 145–152.
[24] F. Hiai and D. Petz, *Riemannian metrics on positive definite matrices related to means.* II, Linear Algebra Appl., 436 (2012), pp. 2117–2136.
[25] B. Jeuris, R. Vandebril, and B. Vandereycken, *A survey and comparison of contemporary algorithms for computing the matrix geometric mean*, Electron. Trans. Numer. Anal., 39 (2012), pp. 379–402.
[26] S. Kotz, K.-T. Fang, and K. W. Ng, *Symmetric Multivariate and Related Distributions*, Chapman and Hall, London, 1990.
[27] J. T. Kent and D. E. Tyler, *Redescending M-estimates of multivariate location and scatter*, Ann. Statist., 19 (1991), pp. 2102–2119.
[28] S. Kotz, N. L. Johnson, and D. W. Boyd, *Series representations of distributions of quadratic forms in normal variables.* I. *Central case*, Ann. Math. Statist., 38 (1967), pp. 823–837.
[29] K. Kraus, *General state changes in quantum theory*, Ann. Phys., 64 (1971), pp. 311–335.
[30] F. Kubo and T. Ando, *Means of positive linear operators*, Math. Ann., 246 (1980), pp. 205–224.
[31] H. Lee and Y. Lim, *Invariant metrics, contractions and nonlinear matrix equations*, Nonlinearity, 21 (2008), pp. 857–878.
[32] B. Lemmens and R. Nussbaum, *Nonlinear Perron-Frobenius Theory*, Cambridge University Press, Cambridge, UK, 2012.
[33] Y. Lim and M. Pálfia, *Matrix power means and the Karcher mean*, J. Funct. Anal., 262 (2012), pp. 1498–1514.
[34] C. F. Van Loan, *The ubiquitous Kronecker product*, J. Comput. Appl. Math., 123 (2000), pp. 85–100.
[35] J. S. Matharu and J. S. Aujla, *Some inequalities for unitarily invariant norms*, Linear Algebra Appl., 436 (2012), pp. 1623–1631.
[36] M. Moakher, *A differential geometric approach to the geometric mean of symmetric positive-definite matrices*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 735–747.
[37] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, John Wiley, New York, 1982.
[38] Yu. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.
[39] C. P. Niculescu, *Convexity according to the geometric mean*, Math. Inequalities Appl., 3 (2000), pp. 155–167.
[40] C. Niculesu and L. E. Persson, *Convex Functions and Their Applications: A Contemporary Approach*, Science & Business 13, Springer, New York, 2006.
[41] F. Nielsen and R. Bhatia, eds., *Matrix Information Geometry*, Springer, New York, 2013.
[42] E. Ollila, D.E. Tyler, V. Koivunen, and H. V. Poor, *Complex elliptically symmetric distributions: Survey, new results and applications*, IEEE Trans. Signal Process., 60 (2011), pp. 5597–5625.
[43] A. Papadopoulos, *Metric Spaces, Convexity and Nonpositive Curvature*, European Mathematical Society, Zürich, Switzerland, 2005.
[44] C. Qi, K. A. Gallivan, and P.-A. Absil, *Riemannian BFGS algorithm with applications*, in Recent Advances in Optimization and Its Applications in Engineering, Springer, Berlin, Heidelberg, 2010, pp. 183–192.
[45] T. Rapcsák, *Geodesic convexity in nonlinear optimization*, J. Optim. Theory Appl., 69 (1991), pp. 169–183.
[46] W. Ring and B. Wirth, *Optimization methods on Riemannian manifolds and their application to shape space*, SIAM J. Optim., 22 (2012), pp. 596–627.
[47] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, New York, 1998.
[48] S. Sra, *Positive definite matrices and the S-divergence*, arXiv preprint, arXiv: 1110.1773v4, http://arxiv.org/abs/1110.1773v4, 2013.
[49] A. Wiesel, *Geodesic convexity and covariance estimation*, IEEE Trans. Signal Process., 60 (2012), pp. 6182–89.
[50] A. L. Yuille and A. Rangarajan, *The concave-convex procedure*, Neural Comput., 15 (2003), pp. 915–936.
[51] T. Zhang, A. Wiesel, and S. Greco, *Multivariate generalized Gaussian distribution: Convexity and graphical models*, IEEE Trans. Signal Process., 60 (2013), pp. 5597–5625.
[52] H. Zhu, H. Zhang, J.G. Ibrahim, and B.S. Peterson, *Statistical analysis of diffusion tensors in diffusion-weighted magnetic resonance imaging data*, J. Amer. Statist. Assoc., 102 (2007), pp. 1085–1102.