Global optimality for Euclidean CCCP under Riemannian convexity

Melanie Weber [1] **Suvrit Sra** [2]

Abstract

We study geodesically convex (g-convex) problems that can be written as a difference of Euclidean convex functions. This structure arises in key applications such as matrix scaling, M-estimators of scatter matrices, and Brascamp-Lieb inequalities. In particular, we exploit this structure to make use of the Convex-Concave Procedure (CCCP), which helps us bypass potentially expensive Riemannian operations and leads to very competitive solvers. Importantly, unlike existing theory for CCCP that ensures convergence to stationary points, we exploit the overall g-convexity structure and provide iteration complexity results for *global optimality*. We illustrate our results by specializing them to a few concrete optimization problems that have been previously studied in the machine learning literature. We hope our work spurs the study of mixed Euclidean-Riemannian optimization algorithms.

1. Introduction

We study optimization problems that assume the form

$$\min_{x\in\mathcal{M}} \quad \phi(x) = f(x) - h(x), \tag{1.1}$$

where $\phi : \mathcal{M} \to \mathbb{R}$ is a geodesically convex (*g-convex*) function on a Riemannian manifold $\mathcal{M}$ (within an ambient Euclidean space), while h is a smooth Euclidean convex function. Problems with such structure most commonly arise when the manifold $\mathcal{M}$ arises from the interior of a Euclidean convex cone, such as the nonegative orthant or the set of semi-definite matrices.

Some problems relevant to machine learning where the model (1.1) arises: Tyler's and related M-estimators (Tyler, 1987; Wiesel, 2012; Ollila and Tyler, 2014; Sra and Hosseini, 2015; Franks and Moitra, 2020); certain geometric optimization problems (Sra and Hosseini, 2015; Bacák, 2014); metric learning (Zadeh et al., 2016); robust subspace recovery (Zhang, 2016); matrix barycenters (Sra, 2016a); matrix-square roots (Sra, 2016b); computation of Brascamp-Lieb constants (Weber and Sra, 2022a; Sra et al., 2018; Allen-Zhu et al., 2018; Bennett et al., 2008); certain Wasserstein bounds on entropy (Courtade et al., 2017); learning Determinantal Point Processes (DPPs) (Mariet and Sra, 2015), optimistic likelihood estimation (Nguyen et al., 2019).

A powerful tool for solving (1.1) is Riemannian optimization (Absil et al., 2009; Boumal, 2020; Boumal et al., 2014). Under suitable regularity conditions on the objective and its gradients, both local (Udriste, 1994; da Cruz Neto et al., 1998; Absil et al., 2009) and global convergence rates for many Riemannian methods can be attained (Zhang and Sra, 2016; Bento et al., 2017). The corresponding algorithms typically need Riemannian objects such as exponential maps, geodesics, and parallel transports (or retractions and vector transports). But for g-convex problems that with the special difference of convex (DC) structure (1.1), one may wonder *if simpler, potentially more efficient methods exist?*

Our goal is to answer this question, where the idea is to exploit the DC structure of (1.1) via the well-known Convex-Concave Procedure (CCCP) of Yuille and Rangarajan (2001). This method assumes differentiability of $h(x)$ to construct a monotonically decreasing sequence $\{\phi(x_k)\}_{k\geq 0}$ of objectives by successively minimizing upper bounds on ϕ. In general, for nonconvex problems, known CCCP analyses can ensure only *asymptotic convergence* (Lanckriet and Sriperumbudur, 2009), and at most to stationary points (Le Thi and Pham Dinh, 2018; Yurtsever and Sra, 2022). But our nonconvex cost function is not arbitrary: it is geodesically convex. Our aim is therefore to understand the much stronger result on *non-asymptotic* convergence rates to ϵ-global optimality, i.e., to ensure $\phi(x) - \inf \phi \leq \epsilon$.

1.1. Main contributions

1. We identify the DC structure (1.1) across several Riemannian optimization problems. Subsequently, we develop global non-asymptotic convergence guarantees for CCCP (Alg. 1) applied to these problems. To our knowledge, this work presents the ***first*** general class of nonconvex DC optimization problems (beyond the Polyak-Łojasiewicz class) for which global iteration complexity of CCCP could be established.

[1]Harvard University [2]MIT. Correspondence to: Melanie Weber <mweber@seas.harvard.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

2. We illustrate the benefits of using Euclidean CCCP for several applications, including M-estimators of scatter matrices, barycenters of positive definite matrices, and computation of the Brascamp-Lieb constant (which also comes up in "operator scaling / Sinkhorn" procedures). Importantly, our theory provides *non-asymptotic* convergence guarantees, where previously *only asymptotic* convergence guarantees were available. Moreover, our theory offers a transparent analysis for several existing algorithms that were previously obtained either in an *ad hoc* manner, or had a much more involved analysis.

Notably, our theoretical analysis turns out be simple, in that it does not require any deep tools from Riemannian geometry and it is largely built on existing Euclidean results. Our focus in the paper is thus less on the analysis itself, but more on its far-reaching implications, as well toward drawing attention to the following key realization:

"*Many Riemannian optimization problems can be solved efficiently via a Euclidean lens.*"

Ultimately, we hope this realization paves the way for a broader study of mixed Riemannian-Euclidean optimization. On a more technical note, we remark that while the Euclidean view bypasses the usual Riemannian tools, and can thus potentially be computationally more efficient, the CCCP approach requires an oracle more powerful than a gradient oracle, which makes it a less general choice. Nevertheless, for several important applications, we show that such an oracle is actually available. We illustrate the practicality of the resulting methods in numerical experiments.

1.2. Related work

CCCP and DC programming. Riemannian DC problems have been studied recently, notably in (Almeida et al., 2020; Souza and Oliveira, 2015; Ferreira et al., 2021). This line of work studies the difference of *geodesically* convex functions as opposed to difference of *Euclidean* convex functions. We follow a different approach that generalizes (Mairal, 2015) to problems of the form 1.1. The methods of (Souza and Oliveira, 2015; Ferreira et al., 2021) involve solving nonconvex subproblems at each iteration, whereas the Euclidean CCCP approach requires solving convex ones.

Riemannian optimization. Riemannian optimization has recently seen a surge of interest in machine learning. Generalizations of classical Euclidean algorithms to the Riemannian setting have been studied for convex (Zhang and Sra, 2016), nonconvex (Boumal et al., 2019), stochastic (Bonnabel, 2013; Zhang and Sra, 2016; Zhang et al., 2016) and constrained problems (Weber and Sra, 2022b; 2021), among others. A textbook treatment can be found in (Absil et al., 2009; Boumal, 2020), whereas the works (Udriste, 1994; Bacák, 2014; Zhang and Sra, 2016) focus on the geodesically convex setting. Some recent work has considered schemes that bypass expensive Riemannian operations in specific applications (Gao et al., 2019; Ablin and Peyré, 2022). However, there has been little work on methods that combine insights from both Euclidean and Riemannian viewpoints.

Recent connection between CCCP and Frank-Wolfe. Very recently, Yurtsever and Sra (2022) highlighted a deep connection between CCCP and the Frank-Wolfe algorithm (Frank and Wolfe, 1956). However, in nonconvex settings (such as 1.1) their results only guarantee non-asymptotic convergence rates to stationary points, whereas our results ensure convergence to *global optimality* (by adding an L-smoothness assumption and exploiting geodesic convexity).

2. Background and Notation

2.1. Riemannian geometry

A *manifold* $\mathcal{M}$ is a locally Euclidean space that is equipped with a differential structure. Its *tangent spaces* $T_x\mathcal{M}$ consist of the tangent vectors at points $x \in \mathcal{M}$. We focus on *Riemannian manifolds*, i.e., smooth manifolds with an inner product $\langle u, v\rangle_x$ defined on $T_x\mathcal{M}$ for each $x \in \mathcal{M}$. To map between a manifold and its tangent space, we define *exponential maps* $\mathrm{Exp} : T_x\mathcal{M} \to \mathcal{M}$, $y = \mathrm{Exp}_x(g_x) \in \mathcal{M}$, given with respect to a geodesic $\gamma : [0,1] \times \mathcal{M} \times \mathcal{M} \to \mathcal{M}$, where $\gamma(0; x, y) = x$, $\gamma(1; x, y) = y$ and $\dot\gamma(0; x, y) = g_x$. The *inverse* exponential map $\mathrm{Exp}^{-1} : \mathcal{M} \to T_x\mathcal{M}$ defines a diffeomorphism from the neighborhood of $x \in \mathcal{M}$ onto the neighborhood of $0 \in T_x\mathcal{M}$ with $\mathrm{Exp}_x^{-1}(x) = 0$. The inner product structure on $T_x\mathcal{M}$ defines a norm $\|v\|_x := \sqrt{\langle v, v\rangle_x}$ for $v \in T_x\mathcal{M}$. We define the geodesic distance of $x, y \in \mathcal{M}$ as $d(x, y)$. Finally, we note that to ease exposition, we limit our attention to Hadamard manifolds (complete, connected Riemannian manifolds with globally nonpositive curvature) as they present the simplest setting for discussing geodesic convexity.

The goal of this paper is the optimization of functions $\phi : \mathcal{M} \to \mathbb{R}$. If ϕ is differentiable, then its gradient $\operatorname{grad}\phi(x)$ is defined as the vector $v \in T_x\mathcal{M}$ with $D\phi(x)[v] = \langle \operatorname{grad}\phi(x), v\rangle_x$. We say that ϕ is *geodesically convex* (short: *g-convex*), if for all $x, y \in \mathcal{M}$

$$\phi(y) \geq \phi(x) + \langle \operatorname{grad}\phi(y), \mathrm{Exp}_x^{-1}(y)\rangle_x. \tag{2.1}$$

In the applications considered in this paper, $\mathcal{M}$ will be the manifold of positive definite matrices, i.e.,

$$\mathcal{M} = \mathbb{P}_d := \{X \in \mathbb{R}^{d\times d} : X = X^T,\ X \succ 0\}, \tag{2.2}$$

i.e., the set of all real symmetric matrices with only positive eigenvalues. We can define a Riemannian structure on $\mathbb{P}_d$

with respect to the inner product

$$\langle A,\, B\rangle_X = \operatorname{tr}\left(X^{-1}AX^{-1}B\right) \quad X \in \mathbb{P}_d,\\ A, B \in T_X(\mathbb{P}_d) = \mathbb{S}_d\ ,$$

where $\mathbb{S}_d$ denotes the space of symmetric matrices.

Throughout the paper, $\|\cdot\|_2$ will denote the Euclidean norm.

Remark 2.1. Observe that we did not define the usual Lipschitz continuity of Riemannian gradients, as we will not be using it. Instead, we will blend the Riemannian view with the Euclidean, and will require Euclidean L-smoothness, which for a C^1 function $h(\cdot)$ is defined as:

$$\|\nabla h(x) - \nabla h(y)\|_2 \le L\|x - y\|_2.$$

Note that here, $\nabla h(\cdot)$ denotes the *Euclidean* gradient.

2.2. Difference of convex functions

Our goal is efficiently minimize g-convex functions that are *difference of (Euclidean) convex* (short: DC) functions. The original motivation for this paper arose from objectives $\phi(\cdot)$ that can be written as $f(\cdot) - h(\cdot)$ of the form

$$\underbrace{-\sum_{j=1}^{n}\log\det(X_j)}_{f(X_1,\ldots,X_n)} \underbrace{-\Big[-\log\det\Big(\sum_{j=1}^{n}A_j^*X_jA_j\Big)\Big]}_{h(X_1,\ldots,X_n)}, \quad (2.3)$$

$$\underbrace{-\log\det(X)}_{f(X)} \underbrace{-\Big[\sum\nolimits_j -\log\det(A_j^*XA_j)\Big]}_{h(X)}. \quad (2.4)$$

As shown in (Sra et al., 2018), objectives (2.3) and (2.4) are g-convex. Since $\log\det(X)$ is Euclidean concave, it is clear that (2.3) and (2.4) are DC programs of the form (1.1). We will revisit problem (2.4) later in Section 4.4.

3. CCCP with global iteration complexity via g-convexity

We propose a Euclidean CCCP method for solving g-convex DC problems. Our proposed method (Alg. 1) utilizes insights on the structure of problem 1.1 from both Euclidean and Riemannian viewpoints. Importantly, we exploit the g-convexity of the DC objective to obtain a non-asymptotic iteration complexity, i.e., a non-asymptotic convergence rate to the global optimum, while exploiting Euclidean Lipschitz-smoothness to control CCCP iterates. This approach is in contrast to the standard CCCP approach that typically only guarantees asymptotic convergence.

The analysis in this section relies on the following manifold-dependent assumption on the relation between Euclidean and geodesic distances:

Assumption 3.1. Let $x, y \in \mathcal{M}$. We have $\|x - y\|_2 \le \alpha_{\mathcal{M}}(d(x, y))$, where $\alpha_{\mathcal{M}}$ is a bounded and positive function that depends on the geometry of $\mathcal{M}$ only.

We note below a **crucial point**, namely that Assumption 3.1 is fulfilled for important instances of Problem 1.1. Indeed, if $\mathcal{M}$ is an embedded submanifold (e.g., the unit sphere or the hyperboloid), Assumption 3.1 holds with $\alpha_{\mathcal{M}}$ being the identity. If $\mathcal{M} = \mathbb{P}_d$ (which includes all applications discussed in §4), we can show the following useful bound:

Lemma 3.2. *Let $X, Y \in \mathbb{P}_d$. Then,*

$$\|X - Y\|_{\mathrm{F}}^2 \le \sqrt{2}\frac{e^{d(X,Y)} - 1}{e^{d(Y,Y)}}\max\{\|X\|_{\mathrm{F}}, \|Y\|_{\mathrm{F}}\}\ .$$

We defer the proof of Lemma 3.2 to the appendix.

3.1. Algorithm: Euclidean CCCP

Algorithm 1 Euclidean CCCP for Riemannian DC

Input: $x_0 \in \mathcal{M}$, K
for $k = 0, 1, \ldots, K-1$ **do**
  Let $Q(x, x_k) = f(x) - h(x_k) - \langle\nabla h(x_k),\, x - x_k\rangle$.
  $x_{k+1} \leftarrow \operatorname{argmin}_{x\in\mathcal{M}} Q(x, x_k)$.
end for
Output: x_K

Recall from (1.1) that we have $\phi(x) = f(x) - h(x)$. Since $-h(x)$ is concave, we can upper bound it as

$$-h(x) \le -h(y) - \langle\nabla h(y),\, x - y\rangle. \quad (3.1)$$

We build on the classical CCCP method and use gradients to linearize the concave part $-h(x)$ of the objective. Observe that only smoothness of h is required, and f can be non-smooth. CCCP utilizes the bound (3.1) in its update rule. In each iteration it seeks to minimize the upper bound

$$\phi(x) \le Q(x, y) := f(x) - h(y) - \langle\nabla h(y),\, x - y\rangle.$$

Since we must ensure that $x \in \mathcal{M}$, the CCCP update step (*CCCP oracle*) in our case is

$$x_{k+1} \leftarrow \operatorname*{argmin}_{x\in\mathcal{M}} Q(x, x_k), \qquad k = 0, 1, \ldots. \quad (3.2)$$

The resulting algorithm is described schematically in Alg. 1.

3.2. Convergence analysis

The convergence of Algorithm 1 can be established via an easy adaptation of the Euclidean MISO algorithm (convex case) (Mairal, 2015, Prop. 2.7) to g-convex Riemannian DC problems. Our non-asymptotic convergence analysis requires us to place some regularity assumptions on the gradient $\nabla h(x)$, which we discuss below.

We first recall the notion of first-order surrogates and recall some of their basic properties (Mairal, 2015).

Definition 3.3 (First-order surrogate functions)**.** Let $\psi : \mathcal{M} \to \mathbb{R}$. We say that ψ is a *first-order surrogate* of ϕ near $x \in \mathcal{M}$, if

1. $\psi(z) \geq \phi(z)$ for all minimizers z of ψ;
2. the approximation error $\theta(z) := \psi(z) - \phi(z)$ is L-smooth, $\theta(x) = 0$ and $\nabla\theta(x) = 0$.

Lemma 3.4. *Let ψ be a first-order surrogate of ϕ near $x \in \mathcal{M}$. Let further $\theta(z) := \psi(z) - \phi(z)$ be L-smooth and $z' \in \mathcal{M}$ a minimizer of ψ. Then:*

1. $|\theta(z)| \leq \frac{L}{2}\|x - z\|^2$;
2. $\phi(z') \leq \phi(z) + \frac{L}{2}\|x - z\|^2$.

The proof of this lemma is straightforward. For completeness, we provide a proof in the appendix. We can now state our main convergence results:

Theorem 3.5. *Let $d(x_0, x^*) \leq R$ for some $x_0 \in \mathcal{M}$ with $\phi(x) \leq \phi(x_0)$. If the functions $Q(x, x_k)$ in Alg. 1 are first-order surrogate functions, then*

$$\phi(x_k) - \phi(x^*) \leq \frac{2L\alpha^2_{\mathcal{M}}(R)}{k+2} \qquad \forall k \geq 1. \tag{3.3}$$

Remark 3.6. Theorem 3.5 and the variants discussed below are subject to a bounded set assumptions. Such conditions also permit non-trivial (i.e., non-constant) g-convex functions on both non-compact and compact manifolds. Ensuring that iterates remain within this set is guaranteed, provided that the condition holds for the initial level set since CCCP is a descent method.

To prove this theorem, we first derive a condition under which the CCCP oracle is defined via first-order surrogates:

Lemma 3.7. *The function $Q(x, x_k)$ as defined in Alg. 1 is a first-order surrogate of ϕ near x_k, if h is L-smooth.*

Proof. We have to show that $Q(x, x_k)$ fulfills all conditions in Definition 3.3. Note that, by construction, condition (1) is fulfilled, i.e., $Q(x, x_k) \geq \phi(x)$ for all x and hence also for all minimizers. Let

$$\begin{aligned}\theta(x) &:= Q(x, x_k) - \phi(x) \\ &= h(x) - h(x_k) - \langle \nabla h(x_k),\, x - x_k \rangle \,.\end{aligned}$$

We see that this term is L-smooth whenever h is L-smooth. Moreover, $\theta(x_k) = 0$, and $\nabla\theta(x_k) = 0$. □

We can now prove Theorem 3.5:

Proof. Using the assumption that $Q(x, x_k)$ is a first-order surrogate of ϕ at x_k, Lemma 3.7 together with Lemma 3.4(2) implies

$$\begin{aligned}\phi(x_k) &\leq \min_{x \in \mathcal{M}} \left[\phi(x) + \tfrac{L}{2}\|x - x_{k-1}\|^2\right] \\ &\leq \min_{x \in \mathcal{M}} \left[\phi(x) + \tfrac{L}{2}\alpha^2_{\mathcal{M}}(d(x, x_{k-1}))\right] \,,\end{aligned}$$

where the second inequality follows from Assumption 3.1. We now follow Nesterov's classical proof technique (Nesterov, 2013) to see that

$$\begin{aligned}\phi(x_k) &\leq \min_{s \in [0,1]} \left[\phi(\gamma(s; x_{k-1}, x^*)) + \tfrac{Ls^2}{2}\alpha^2_{\mathcal{M}}(d(x^*, x_{k-1}))\right] \\ &\leq \min_{s \in [0,1]} \left[s\phi(x^*) + (1-s)\phi(x_{k-1}) + \tfrac{Ls^2}{2}\alpha^2_{\mathcal{M}}(R)\right] \,,\end{aligned}$$

where we have replaced the minimization over $x \in \mathcal{M}$ with minimization over the geodesic $\gamma(s; x_{k-1}, x^*)$ and inserted the bound $d(x^*, x_{k-1}) \leq R$. Since $\phi(x_k)$ is a monotonically decreasing sequence, we can invoke the bounded level-set assumption ($\phi(x) \leq \phi(x_0)\ \forall x \in \mathcal{M}$) to obtain

$$\begin{aligned}\phi(x_k) - \phi(x^*) \leq \min_{s \in [0,1]} \big[&(1-s)(\phi(x_{k-1}) - \phi(x^*)) \\ &+ \tfrac{1}{2}L\alpha^2_{\mathcal{M}}(R)s^2\big] \,.\end{aligned} \tag{3.4}$$

Let $\Delta_k := \phi(x_k) - \phi(x^*)$.

We introduce the shorthand $\Delta_k := \phi(x_k) - \phi(x^*)$, with which we have

$$\Delta_k \leq \min_{s \in [0,1]} \left[(1-s)\Delta_{k-1} + \tfrac{1}{2}L\alpha^2_{\mathcal{M}}(R)s^2\right] \,.$$

We want to find s, such that the right hand side is minimized, which is equivalent to minimizing over the geodesic $\gamma(s; x_{k-1}, x^*)$. We now distinguish between two cases regarding the value of Δ_{k-1} relative to the term $L\alpha^2_{\mathcal{M}}(R)$:

1. If $\Delta_{k-1} \geq L\alpha^2_{\mathcal{M}}(R)$, then the minimum is attained on the boundary, i.e., $s^* = 1$, whereby we immediately have $\Delta_k \leq \frac{1}{2}L\alpha^2_{\mathcal{M}}(R)$.
2. Otherwise, assume $\Delta_{k-1} \leq L\alpha^2_{\mathcal{M}}(R)$. Then Eq. 3.4 implies that the sequence $(\Delta_k)_k$ is monotonically decreasing; the minimum is attained at $s^* = \frac{\Delta_{k-1}}{L\alpha^2_{\mathcal{M}}(R)}$, which implies $\Delta_k \leq \Delta_{k-1}\left(1 - \frac{\Delta_{k-1}}{2L\alpha^2_{\mathcal{M}}(R)}\right)$ or equivalently

$$\begin{aligned}\Delta_k^{-1} &\geq \Delta_{k-1}^{-1}\left(1 - \frac{\Delta_{k-1}}{2L\alpha^2_{\mathcal{M}}(R)}\right) \\ &\geq \Delta_{k-1}^{-1} + \frac{1}{2L\alpha^2_{\mathcal{M}}(R)} \,.\end{aligned}$$

Here, the second inequality follows from $(1-x)^{-1} \geq 1 + x\ \forall x \in (0, 1)$.

Now iteratively apply the two inequalities to conclude. □

3.3. Implementing the CCCP oracle

The complexity of Alg. 1 relies crucially on the complexity of the CCCP oracle. In the following section, we discuss several instances where the CCCP oracle has a closed form solution, resulting in a competitive algorithm.

Remark 3.8. Note that if the CCCP oracle can be solved in closed form for a specific objective ϕ, then Alg. 1 is metric independent. In this context, it is interesting to consider under which metric, for which ϕ is g-convex, the constant $\alpha_{\mathcal{M}}$ in Thm. 3.5 is smallest.

However, in general, a closed-form solution may not always be available. In this section, we discuss several instances of this setting. First, we investigate an *inexact variant* of Alg. 1, where we solve the CCCP oracle only approximately. Secondly, we investigate a CCCP approach that exploits *finite-sum structure* (Alg. 2), which we encounter in many problems of the form 1.1.

We remark that while we require h to be smooth and Euclidean convex, we have imposed no such constraint on f. In particular, both Lem. 3.7 and Thm. 3.5 hold, if f is *non-differentiable*. This widens the range of problems that can be analyzed with our CCCP framework.

3.3.1. Inexact CCCP oracle

In general, we may only be able to solve the CCCP oracle approximately. Therefore, we complement our analysis of Alg. 1 with the study of an *inexact* variant. We assume that in iteration k, we perform an inexact CCCP update, i.e., we compute an ϵ-approximate minimum $\tilde{Q}_k$, such that

$$\tilde{Q}_k \leq \min_{x \in \mathcal{M}} Q(x, x_k) + \frac{1}{2} L \alpha_{\mathcal{M}}^2(R) s_k^2 \epsilon \,, \tag{3.5}$$

where s_k again denotes the step size defined in the proof of Thm. 3.5. A simple adaption of our convergence proof above gives the following non-asymptotic guarantee:

Theorem 3.9. *Let $d(x_0, x^*) \leq R$ for some $x_0 \in \mathcal{M}$ with $\phi(x) \leq \phi(x_0)$ and let $Q(x, x_k)$ be first-order surrogate functions. Let $\left(\tilde{Q}_k\right)_{k \geq 0}$ be a sequence of ϵ-approximate CCCP updates in the sense of Eq. 3.5. Then*

$$\phi(x_k) - \phi(x^*) \leq \frac{2 L \alpha_{\mathcal{M}}^2(R)(1+\epsilon)}{k+2} \qquad \forall\, k \geq 1. \tag{3.6}$$

The proof is a simple adaption of the proof of Thm. 3.5 and can be found in the appendix.

Remark 3.10. We note that the convergence analysis in this section provides an iteration complexity, i.e., a complexity guarantee in the sense of a bound on the number of iterations required to achieve an ϵ-approximate solution to the original problem. A more refined convergence analysis would also include the impact of the "degree of inexactness" of the CCCP oracle, because in general, one might not be able to implement the oracle exactly (although for the applications presented in this paper, the oracle is *indeed* exact).

3.3.2. Exploiting Finite-sum Structure

In applications, we frequently encounter instaces of (1.1), where h has a *finite-sum structure* $h(x) := \frac{1}{m}\sum_{i=1}^m h_i(x)$, where the h_i are L-smooth. Notice that in this case, computing the CCCP step requires m gradient evaluations, which may be expensive, if m is large. Instead of recomputing the full surrogate as in Alg. 1, we could make only incremental updates to the surrogate in each iteration. We outline an incremental update scheme in Alg. 2, which requires only *one*, instead of m gradient evaluations, significantly reducing the complexity of the CCCP oracle.

We note that several of the applications presented in section 4 have a finite-sum structure. However, in those cases, the CCCP oracle can actually be solved in closed-form, resulting in a very competitive implementation of Alg. 1.

For the convergence analysis we again follow closely the analysis of the MISO algorithm (Mairal, 2015, Prop. 3.1). We show the following result:

Theorem 3.11. *Let $d(x_0, x^*) \leq R$ for some $x_0 \in \mathcal{M}$ with $\phi(x) \leq \phi(x_0)$. Assume that $g_{i_k}^k$ as defined in Alg. 2 is a first-order surrogate of h_{i_k} near x_{k-1}. Then Alg. 2 converges almost surely.*

We defer the proof details to the appendix.

Remark 3.12. An additional speed up in Alg. 2 can be achieved by applying the SPIDER technique (Fang et al., 2018; Nguyen et al., 2017) to the variance-reduced approximation of the gradient. Here, the batch size decreases in later epochs, reducing the cost of the gradient oracle.

4. Applications

In this section we present several applications that possess the DC structure (1.1). All our examples are drawn from the manifold of positive definite matrices, since a large number of practical matrix estimation problems are known in this setting (Wiesel, 2012; Sra and Hosseini, 2015), and it serves to best illustrate the practical aspects.

Importantly, for the applications presented here, our framework provides for the CCCP approach (Alg. 1) either (1) the first *non-asymptotic* guarantees on suboptimality (for barycenters of SPD matrices via S-divergence, sec. 4.3) or (2) a simple and competitive algorithm and analysis (e.g., Brascamp-Lieb constants, sec. 4.4). We summarize previous convergence results in Tab. 1. Moreover,s ince the CCCP oracle (line 4, Alg. 1) is solvable in closed-form, it renders our approach into a practical method attractive for downstream applications.

Algorithm 2 Incremental CCCP for Riemannian DC with finite-sum structure

Input: $x_0 \in \mathcal{M}$, K
Set $g^0(x) := \frac{1}{m}\sum_{i=1}^m h_i(x_0) - \langle \nabla h_i(x_0),\, x - x_0\rangle$.
for $k = 1, \ldots, K$ **do**
  Choose $i_k \sim [m]$ randomly.
  Set $g_{i_k}^k(x) := h_{i_k}(x_k) - \langle \nabla h_{i_k}(x_k),\, x - x_k\rangle$ and $g_i^k \triangleq g_i^{k-1}$ for $i \neq i_k$.
  Set $Q(x, x_k) := f(x) - g^k(x)$.
  $x_{k+1} \leftarrow \operatorname{argmin}_{x\in\mathcal{M}} Q(x, x_k)$.
end for
Output: x_K

Problem	Asymp	Non-asymp
Matrix scaling	✓	✓∗
Tyler's M-estimator	✓	✓∗
PD Matrix square root	✓	✗
PD Barycenter	✓	✗
Brascamp-Lieb	✓	✓∗
DPP	✓	✗

Table 1. Summary of existing convergence guarantees for CCCP algorithms for Riemannian DC problems. ($*$ denotes a simplified analysis with our approach.)

4.1. Matrix scaling

This application is purely expository, but we include it as a likely more familiar problem. For diagonal positive definite matrices, g-convexity reduces to ordinary convexity after a global change of variables as in geometric programming (Boyd et al., 2007). A canonical problem here is matrix scaling, for which perhaps the best known method is the classical Sinkorn algorithm (Sinkhorn and Knopp, 1967), though the problem has witnessed considerable recent progress too (Allen-Zhu et al., 2017; Cohen et al., 2017; Altschuler et al., 2017). We comment only on the most basic version of the problem; see (Yuille and Rangarajan, 2003) for more details.

We are given an $n \times n$ positive matrix M for which we seek to compute diagonal scaling matrices D and E such that DME is doubly stochastic, i.e., its rows and columns sum to 1. Sinkhorn's algorithm is known to be obtained by applying CCCP to minimize the cost function

$$\min_{x>0}\ \ \phi(x) := -\sum\nolimits_j \log x_j + \sum\nolimits_i \log\Big(\sum\nolimits_j x_j M_{ij}\Big)\ . \tag{4.1}$$

Here, $\{x_j\}$ are the diagonal elements of E, while the diagonal elements of D are $1/\{\sum_j x_j M_{ij}\}$. Observe that $\phi(x)$ given by (4.1) is actually g-convex,[1] and thus this problem is indeed of the form (1.1). Further, it can be verified that the $h(x)$ part of (4.1) satisfies the L-smoothness assumption, since the logarithm is L-smooth on a domain with a positive uniform lower bound.

[1] This observation follows immediately from the g-convexity of the BL problem (2.3), of which problem (4.1) is known to be a special case.

4.2. M-estimators for scatter matrices

Estimating the shape of a covariance matrix for high-dimensional data is an important problem in statistics. One important class of covariance estimators, based on elliptically contoured distributions, is Tyler's M-estimator (Ollila and Tyler, 2014). There are several important asymptotic properties of this estimator, and it has been extensively studied; for additional details and discussion we refer the reader to the papers (Franks and Moitra, 2020; Sra and Hosseini, 2015; Wiesel, 2012; Wiesel et al., 2015; Ollila and Tyler, 2014; Zhang, 2016; Tyler, 1987). The best known algorithms for computing Tyler's M-estimator arise from carefully constructed fixed-point iterations. The convergence analysis of those fixed-point iterations utilize the Hilbert projective metrics, in a manner analogous to Birkhoff's use of the Hilbert projective metric for the convergence analysis of problems closely related to matrix scaling (Birkhoff, 1957). Following our discussion above, Algorithm 1 delivers a transparent method for obtaining Tyler's estimator by solving (4.2), at least in the cases where g-convexity applies; see also (Sra and Hosseini, 2015) for additional discussion.

The resulting optimization problem involves obtaining a scatter matrix by maximizing a likelihood of the form

$$\mathcal{L}(X, A) := -\frac{n}{2}\log\det(X) + \sum\nolimits_i \log f(a_i^T X^{-1} a_i), \tag{4.2}$$

where f is a so-called "distance generating function". The likelihood (4.2) generalizes the usual multivariate Gaussian to the much larger class of Elliptically contoured distributions. Assuming that $\log f$ is concave and monotonic, it is easily seen that (4.2) can be equivalently written as a g-convex minimization problem (after reversing signs) of the form (1.1). Empirically, the much faster run times obtained via CCCP (which yields a fixed-point iteration for solving (4.2)) has been explicitly highlighted in (Sra and Hosseini, 2015; Hosseini et al., 2016). Notably, (4.2) has a

finite-sum structure, which we can exploit by using Alg. 2 to compute the M-estimator with a faster gradient oracle.

4.3. Matrix square root and barycenter of PD matrices

The S-Divergence (Sra, 2016a) between two positive definite matrices X, Y is defined as

$$\delta_s(X,Y) := \log\det\left(\tfrac{(X+Y)}{2}\right) - \tfrac{1}{2}\log\det(XY). \quad (4.3)$$

Matrix square root. Suppose M is a positive definite matrix. In (Jain et al., 2017) the authors proposed a gradient descent based method to compute the square root of M. A faster algorithm was obtained in (Sra, 2016b) who proposed the following iteration

$$X \leftarrow (X+I)^{-1} + (X+M)^{-1},$$

which was obtained as a certain fixed-point iteration to compute the *barycenter*

$$\min_{X\succ 0} \quad \delta_s^2(X,I) + \delta_s^2(X,M). \quad (4.4)$$

More generally, in (Sra, 2016a) the *barycenter* version of (4.4) is studied. Here, given n positive definite matrices $A_1,\dots,A_n$, one seeks to solve

$$\min_{X\succ 0} \quad \sum_{i=1}^n w_i \delta_s^2(X,A_i). \quad (4.5)$$

Using the defintion (4.3) of δ_s, it is immediate that (4.5) is a difference of Euclidean convex functions; its g-convexity is more involved but follows from (Sra, 2016a). By applying our CCCP Algorithm, one immediately recovers a proof of convergence for the fixed-point iteration proposed in (Sra, 2016a) for solving (4.5).

4.4. Brascamp-Lieb Constant

Now we come to what is perhaps the most interesting application of the problem structure (1.1). Indeed, as previously, this application is the one that motivated us to develop the method studied in this paper. Specifically, we study Brascamp-Lieb (short: BL) inequalities that form a central class of inequalities in functional analysis and probability theory, offering a great generalization to the basic Hölder inequality, and being intimately related with entropy inequalities too. As a special instance of the Operator Scaling problem (Garg et al., 2017), they relate to a range of problems in various areas of mathematics and theoretical computer science (Bennett et al., 2008).

The computation of BL constants can be formulated as an optimization task on $\mathbb{P}_d$:

$$F(X) = -\log\det(X) + \sum\nolimits_i w_i \log\det(\Phi_i(X)), \quad (4.6)$$

where $\Phi_i(X) = A_i^* X A_i$, and $w_i \geq 0$ with $w^T 1 = 1$. Notably, the objective is g-convex (Sra et al., 2018), which allows for applying Algorithm 1 with global convergence guarantees. Since $\log\det(\Phi(X))$ is a concave function of X, it can be upper bounded as

$$\log\det(\Phi_i(X)) \leq \log\det(\Phi(Z)) + \operatorname{tr}(Z^{-1}\Phi_i(X) - Z). \quad (4.7)$$

Using (4.7) we thus have the following upper bound

$$\begin{aligned} F(X) &\leq -\log\det(X) + \log\det(Z) \\ &\quad + \sum\nolimits_i w_i \operatorname{tr}(Z^{-1}\Phi_i(X)) - d \\ &=: g(X,Z). \end{aligned}$$

The CCCP update step is

$$X_{k+1} \leftarrow \operatorname*{argmin}_{X>0} g(X, X_k), \quad (4.8)$$

which results in iteration of the map

$$X_{k+1} = \left[\sum\nolimits_i w_i A_i (A_i^* X_k A_i)^{-1} A_i^*\right]^{-1}. \quad (4.9)$$

Our analysis delivers non-asymptotic guarantees for computing BL constants, a result that was obtained by analyzing a more involved operator Sinkhorn iteration in (Garg et al., 2017), as well as, more recently in (Weber and Sra, 2022a) with more involved tools from Finslerian geometry.

4.5. Determinantal Point Processes

Discrete Determinantal Point Processes are often characterized by a positive semidefinite matrix, the *DPP kernel*. Consider the problem of fitting a DPP kernel L to observations $(Y_1,\dots,Y_n) \subseteq \mathcal{Y}$ (Gillenwater et al., 2014):

$$\max_{L\succeq 0} \sum_{i=1}^n \log\det(L_{Y_i}) - n\log\det(I+L), \quad (4.10)$$

where for each i, $L_{Y_i} = U_i^* L U_i$ for a suitable partial isometry U_i. Notice that (4.10) is g-convex; remarkably, exploiting this g-convexity was already stated as an open problem in (Mariet and Sra, 2015), who presented a fast fixed-point iteration toward solving (4.10). Assuming that $L \succ 0$, any critical point of (4.10) is a solution to the matrix equation

$$\sum_{i=1}^n U_i (U_i^* L U_i)^{-1} U_i^* - n(I+L)^{-1} = 0\,. \quad (4.11)$$

To solve (4.11) Mariet and Sra (2015) propose the following fixed-point iteration:

$$L_{k+1} \leftarrow \frac{1}{n}\sum_{i=1}^n U_i (U_i^* L_k U_i)^{-1} U_i^* - (I+L_k)^{-1} + L_k^{-1},$$

whose validity they justify by appealing to the CCCP framework. But the convergence analysis in (Mariet and Sra, 2015) is limited to local guarantees, whereas our CCCP framework provides *global, non-asymptotic guarantees*.

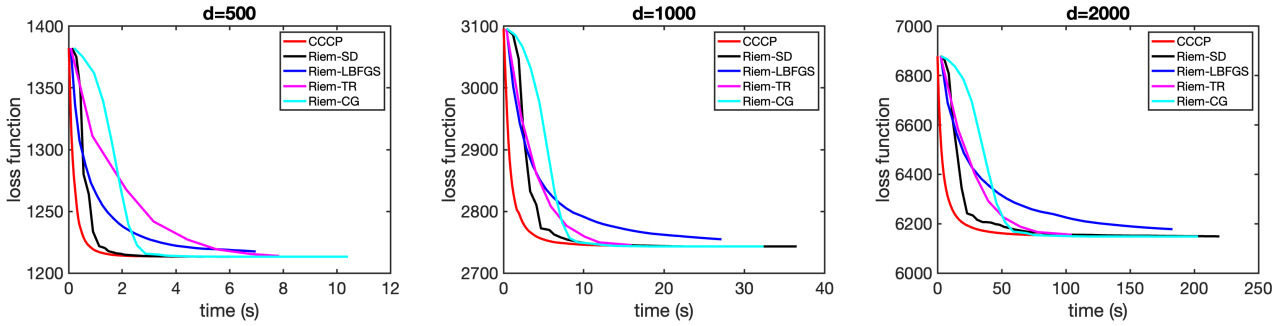

Figure 1. Performance of CCCP approach in comparison with Riemannian Steppest Descent (Riem-SD), Riemannian LBFGS (Riem-LBFGS), Riemannian Trustregions (Riem-TR) and Riemannian Conjugate Gradient (Riem-CG) for computing matrix square roots with inputs of dimension d.

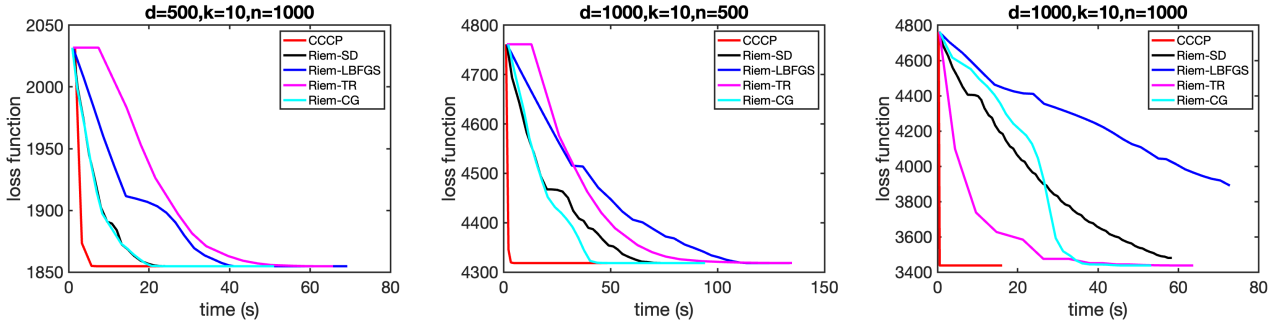

Figure 2. Performance of CCCP approach in comparison with Riemannian Steppest Descent (Riem-SD), Riemannian LBFGS (Riem-LBFGS), Riemannian Trustregions (Riem-TR) and Riemannian Conjugate Gradient (Riem-CG) for computing Brascamp-Lieb constants for n input matrices of size $d \times k$. Here, the loss function is the value of the objective 4.6, i.e., the value attained by the BL constant.

4.6. Experiments

To demonstrate the efficiency of our proposed approach, we complement our discussion with experimental results for two of the applications discussed above. We show that CCCP performs competitively against several popular Riemannian Optimization methods for the problem of computing matrix square roots (Fig. 1) and for computing Brascamp-Lieb constants (Fig. 2). In both experiments we compare against the Manopt (Boumal et al., 2014) implementation of the algorithms for inputs of different sizes. We do note, however, that this advantage in running time is more pronounced for larger problems, as expected.

5. Conclusion

We consider geodesically convex optimization problems that admit a Euclidean difference of convex (DC) decomposition. We analyzed the global iteration complexity of Euclidean CCCP applied to solving such problems, where geodesic convexity played an important role to help bound function suboptimality, while Euclidean smoothness (of one of the DC components) helped control the progress of CCCP. While simple, this work captures a sufficiently valuable class of nonconvex optimization problems for which CCCP can be shown to converge globally. We illustrate our ideas on several important applications where such a DC structure arises, and for which CCCP either delivers a new convergent algorithm, or helps us explain the convergence of an existing algorithm.

An important question in this context is whether there exist an efficiently computable DC representation for any geodesically convex cost function? Since $\mathcal{M}$ is a manifold, it is an open set. Hence, nonconvex nonsmooth functions that satisfy bounded-variation admit a DC representation; moreover, in case $\phi \in C^2$ (i.e., twice continuously differentiable), there always is a DC representation, regardless of g-convexity (Konno et al., 1997). The key challenge is whether one can efficiently find such a representation. This problem seems to be of considerable difficulty. In Appendix 1.1 we give an example of how the well-known Riemannian distance function $d_R(X,Y) = \|\log(X^{-1/2}YX^{-1/2})\|_{\mathrm{F}}$ on the positive definite matrices admits such a DC representation, albeit one that seems quite intricate as it involves integrating over infinitely many functions.

We hope that our work spurs not only an investigation of the fundamental question raised above, but of better algo-

rithms and complexity analysis for CCCP and other related procedures when applied to the class of g-convex functions studied in this paper. We believe that it should be possible to drop the dependence on the gradient Lipschitzness L in the CCCP method studied in this work, but expect that a completely different approach will be needed to analyze the method. Finally, in the same vein, it will be valuable to extend our study to non-differentiable g-convex functions that enjoy a Euclidean DC representation. We leave investigation of these important problems to the future.

Acknowledgments

SS acknowledges support from the NSF-CAREER grant (IIS-1846088). Part of this work was done while MW was at the University of Oxford, supported by a Hooke Research Fellowship and a Postdoctoral Enrichment Award by the Alan Turing Institute. The authors thank Pierre-Antoine Absil for helpful comments on the manuscript.

References

P. Ablin and G. Peyré. Fast and accurate optimization on the orthogonal manifold without retraction. In *International Conference on Artificial Intelligence and Statistics*, pages 5636–5657. PMLR, 2022.

P.-A. Absil, R. Mahony, and R. Sepulchre. Optimization algorithms on matrix manifolds. In *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

Z. Allen-Zhu, Y. Li, R. Oliveira, and A. Wigderson. Much faster algorithms for matrix scaling. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 890–901. IEEE, 2017.

Z. Allen-Zhu, A. Garg, Y. Li, R. Oliveira, and A. Wigderson. Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing. In *Proceedings of the 50th Annual ACM Symposium on Theory of Computing, Los Angeles, CA, USA, June 25-29, 2018*, 2018. To appear.

Y. T. Almeida, J. X. da Cruz Neto, P. R. Oliveira, and J. C. d. O. Souza. A modified proximal point method for DC functions on Hadamard manifolds. *Computational Optimization and Applications*, 76(3):649–673, 2020.

J. Altschuler, J. Niles-Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *Advances in neural information processing systems*, 30, 2017.

M. Bacák. Convex analysis and optimization in Hadamard spaces. In *Convex Analysis and Optimization in Hadamard Spaces*. de Gruyter, 2014.

J. Bennett, A. Carbery, M. Christ, and T. Tao. The Brascamp-Lieb inequalities: finiteness, structure and extremals. *Geometric and Functional Analysis*, 17(5):1343–1415, 2008.

G. C. Bento, O. P. Ferreira, and J. G. Melo. Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 173(2):548–562, 2017.

R. Bhatia. *Positive definite matrices*. Princeton University Press, 2009.

G. Birkhoff. Extensions of Jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227, 1957.

S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58 (9):2217–2229, 2013.

N. Boumal. An introduction to optimization on smooth manifolds. *Available online, May*, 3, 2020.

N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research*, 15(1):1455–1459, 2014.

N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

S. Boyd, S.-J. Kim, L. Vandenberghe, and A. Hassibi. A tutorial on geometric programming. *Optimization and engineering*, 8(1):67–127, 2007.

M. B. Cohen, A. Madry, D. Tsipras, and A. Vladu. Matrix scaling and balancing via box constrained Newton's method and interior point methods. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 902–913. IEEE, 2017.

T. A. Courtade, M. Fathi, and A. Pananjady. Wasserstein stability of the entropy power inequality for log-concave random vectors. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 659–663. IEEE, 2017.

J. da Cruz Neto, L. De Lima, and P. Oliveira. Geodesic algorithms in Riemannian geometry. *Balkan J. Geom. Appl*, 3(2):89–100, 1998.

C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

O. P. Ferreira, E. M. Santos, and J. C. O. Souza. The difference of convex algorithm on riemannian manifolds. *arXiv preprint arXiv:2112.05250*, 2021.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110, 1956.

W. C. Franks and A. Moitra. Rigorous guarantees for Tyler's M-estimator via quantum expansion. In *Conference on*

Learning Theory, pages 1601–1632. PMLR, 2020.

B. Gao, X. Liu, and Y.-x. Yuan. Parallelizable algorithms for optimization problems with orthogonality constraints. *SIAM Journal on Scientific Computing*, 41(3):A1949–A1983, 2019.

A. Garg, L. Gurvits, R. M. de Oliveira, and A. Wigderson. Algorithmic and optimization aspects of Brascamp-Lieb inequalities, via operator scaling. In *Proceedings of the 49th Annual ACM Symposium on Theory of Computing, Montreal, QC, Canada, June 19-23, 2017*, pages 397–409, 2017. doi: 10.1145/3055399.3055458. URL http://doi.acm.org/10.1145/3055399.3055458.

J. A. Gillenwater, A. Kulesza, E. Fox, and B. Taskar. Expectation-maximization for learning determinantal point processes. *Advances in Neural Information Processing Systems*, 27, 2014.

R. Hosseini, S. Sra, L. Theis, and M. Bethge. Inference and mixture modeling with the elliptical Gamma distribution. *Computational Statistics & Data Analysis*, 101:29–43, 2016.

P. Jain, C. Jin, S. Kakade, and P. Netrapalli. Global convergence of non-convex gradient descent for computing matrix squareroot. In *Artificial Intelligence and Statistics*, pages 479–488. PMLR, 2017.

H. Konno, P. T. Thach, and H. Tuy. DC functions and DC sets. In *Optimization on Low Rank Nonconvex Structures*, pages 47–76. Springer, 1997.

G. Lanckriet and B. K. Sriperumbudur. On the convergence of the concave-convex procedure. *Advances in neural information processing systems*, 22, 2009.

H. A. Le Thi and T. Pham Dinh. Dc programming and DCA: thirty years of developments. *Mathematical Programming*, 169(1):5–68, 2018.

J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

Z. Mariet and S. Sra. Fixed-point algorithms for learning determinantal point processes. In *International Conference on Machine Learning*, pages 2389–2397. PMLR, 2015.

Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017.

V. A. Nguyen, S. Shafieezadeh Abadeh, M.-C. Yue, D. Kuhn, and W. Wiesemann. Calculating optimistic likelihoods using (geodesically) convex optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

E. Ollila and D. E. Tyler. Regularized m-estimators of scatter matrix. *IEEE Transactions on Signal Processing*, 62(22):6059–6070, 2014.

M. Šilhavý. A functional inequality related to analytic continuation. 2015.

R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

D. A. Snyder. On the relation of Schatten norms and the Thompson metric. *arXiv:1608.03301*, 2016.

J. C. d. O. Souza and P. R. Oliveira. A proximal point algorithm for DC fuctions on Hadamard manifolds. *Journal of Global Optimization*, 63(4):797–810, 2015.

S. Sra. Positive definite matrices and the S-divergence. *Proceedings of the American Mathematical Society*, 144 (7):2787–2797, 2016a.

S. Sra. On the matrix square root via geometric optimization. *The Electronic Journal of Linear Algebra*, 31:433–443, 2016b.

S. Sra and R. Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, 25(1):713–739, 2015.

S. Sra, N. K. Vishnoi, and O. Yildiz. On geodesically convex formulations for the Brascamp-Lieb constant. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

D. E. Tyler. A distribution-free M-estimator of multivariate scatter. *The annals of Statistics*, pages 234–251, 1987.

C. Udriste. *Convex Functions and Optimization Methods on Riemannian Manifolds*, volume 297. Springer Science & Business Media, 1994.

M. Weber and S. Sra. Projection-free nonconvex stochastic optimization on Riemannian manifolds. *IMA Journal of Numerical Analysis*, 2021.

M. Weber and S. Sra. Computing Brascamp-Lieb constants through the lens of Thompson geometry. *arXiv preprint arXiv:2208.05013*, 2022a.

M. Weber and S. Sra. Riemannian optimization via Frank-Wolfe methods. *Mathematical Programming*, 2022b.

A. Wiesel. Geodesic convexity and covariance estimation. *IEEE transactions on signal processing*, 60(12):6182–6189, 2012.

A. Wiesel, T. Zhang, et al. Structured robust covariance estimation. *Foundations and Trends® in Signal Processing*, 8(3):127–216, 2015.

A. L. Yuille and A. Rangarajan. The concave-convex procedure (CCCP). *Advances in neural information processing systems*, 14, 2001.

A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural computation*, 15(4):915–936, 2003.

A. Yurtsever and S. Sra. Cccp is frank-wolfe in disguise. *arXiv preprint arXiv:2206.12014*, 2022.

P. Zadeh, R. Hosseini, and S. Sra. Geometric mean metric learning. In *International conference on machine learning*, pages 2464–2471. PMLR, 2016.

H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638, 2016.

H. Zhang, S. J Reddi, and S. Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. *Advances in Neural Information Processing Systems*, 29, 2016.

T. Zhang. Robust subspace recovery by Tyler's M-estimator. *Information and Inference: A Journal of the IMA*, 5(1): 1–21, 2016.

A. DC representation of Riemannian distance on $\mathbb{P}_d$

We will need the following useful integral representation of the squared logarithm:

Lemma A.1. *Let $x > 0$. Then, the following representation holds*

$$(\log x)^2 = \int_0^\infty [\log(1+tx) + \log(t+x) - \log x - 2\log(1+t)]\frac{dt}{t}. \tag{A.1}$$

Proof. This is easily verified using, e.g., Mathematica. However, for a more general theory of such integrals we refer the interested reader to (Šilhavý, 2015). □

We are now ready to state our result on the DC representation of Riemannian distance.

Theorem A.2. *Let $X, Y \in \mathbb{P}_d$ and let $d_R(X,Y) = \|\log(X^{-1/2}YX^{-1/2})\|_{\mathrm{F}}$ be the Riemannian distance between them. Then, $d_R^2(X,Y)$ is g-convex jointly in (X,Y) and it admits a DC representation*

$$d_R^2(X,Y) = \int_0^\infty [f_t(X,Y) - h_t(X,Y)]d\mu(t), \tag{A.2}$$

where f_t and h_t are convex, and μ is a suitable measure.

Proof. It is well-known that d_R is jointly g-convex—see e.g., (Bhatia, 2009, Ch.6) for a proof. Consequently, d_R^2 is also g-convex. In deriving our proof of the DC representation of d_R^2, we will also obtain an alternative (and to our knowledge, a new) proof of this joint g-convexity as a byproduct.

Begin with observing that $d_R^2(X,Y) = \|\log(X^{-1/2}YX^{-1/2})\|_{\mathrm{F}} = \sum_i (\log\lambda_i(X^{-1}Y))^2$. For brevity, we write $\lambda_i \equiv \lambda_i(X^{-1}Y)$ and $\ldet \equiv \log\det$; then, using the integral (A.1) we have

$$\begin{aligned}
d_R^2(X,Y) &= \sum_i \textstyle\int_0^\infty [\log(1+t\lambda_i) + \log(t+\lambda_i) - \log\lambda_i - 2\log(1+t)]\,\frac{dt}{t} \\
&= \textstyle\int_0^\infty \left[\ldet(I+tX^{-1}Y) + \ldet(tI+X^{-1}Y) - \ldet(X^{-1}Y) - 2n\log(1+t)\right]\frac{dt}{t} \\
&= \textstyle\int_0^\infty [\ldet(X+tY) + \ldet(tX+Y) - 2\ldet(X) - \ldet(Y) - 2n\log(1+t)]\,\frac{dt}{t} \\
&= \textstyle\int_0^\infty [-2\ldet(X) - \ldet(Y) - 2n\log(1+t) - (-\ldet(X+tY) - \ldet(tX+Y))]\,\frac{dt}{t} \\
&= \textstyle\int_0^\infty [f_t(X,Y) - h_t(X,Y)]\,\frac{dt}{t},
\end{aligned}$$

where $f_t(X,Y) = -2\ldet(X) - \ldet(Y) - 2n\log(1+t)$ and $h_t(X,Y) = -\ldet(X+tY) - \ldet(tX+Y)$. Convexity of both f_t and h_t is immediate from the well-known convexity of $-\log\det(X)$ on $X \succ 0$. □

B. Proof Details

B.1. Relation between Euclidean and Riemannian metrics on $\mathbb{P}_d$

Lemma B.1. *Let $x, y \in \mathbb{P}_d$. Then the Euclidean and Riemannian distance relate as*

$$\|x-y\|_2^2 \le \sqrt{2}\frac{e^{d(x,y)}-1}{e^{d(x,y)}}\max\{\|x\|_2, \|y\|_2\}\,.$$

Proof. Recall that the Thompson metric δ_T and the Riemannian distance d for positive definite matrices are given by

$$\begin{aligned}
\delta_T(x,y) &:= \|\log\left(x^{-\frac{1}{2}}yx^{-\frac{1}{2}}\right)\| \\
d(x,y) &:= \|\log\left(x^{-\frac{1}{2}}yx^{-\frac{1}{2}}\right)\|_F\,,
\end{aligned}$$

where $\|\cdot\|$ denotes the operator norm and $\|\cdot\|_F$ the Frobenius norm. It is well-known that $\|x\| \le \|x\|_F$ for $x \in \mathbb{P}_d$. This implies $\delta_T(x,y) \le d(x,y)$. The claim follows from a relation between the Euclidean distance and the Thompson metric, established by Snyder (2016):

$$\|x-y\|_2^2 \le \sqrt{2}\frac{e^{\delta_T(x,y)}-1}{e^{\delta_T(x,y)}}\max\{\|x\|_2, \|y\|_2\}\,.$$

□

B.2. Properties of surrogate functions

We sketch a proof for Lem. 3.4:

Lemma B.2. *Let ψ be a first-order surrogate of ϕ near $x \in \mathcal{M}$. Let further $\theta(z) := \psi(z) - \phi(z)$ be L-smooth and $z' \in \mathcal{M}$ a minimizer of ψ. Then:*

1. $|\theta(z)| \leq \frac{L}{2}\|x - z\|^2$;
2. $\phi(z') \leq \phi(z) + \frac{L}{2}\|x - z\|^2$.

Proof. For (1) recall a classical inequality, which follows from the L-smoothness of the surrogate function:

$$|\theta(z) - \theta(x) - \langle \nabla\theta(x),\, x - z\rangle| \leq \frac{L}{2}\|x - z\|_2^2 \,.$$

The claim follows from $\theta(x) = 0$ and $\nabla\theta(x) = 0$.

For (2), note that we have by construction

$$\phi(z') \leq \psi(z') \leq \psi(z) = \phi(z) - \theta(z) \,.$$

Inserting (1) directly gives the claim. □

B.3. Inexact CCCP oracle

For completeness, we give a proof of Thm. 3.7:

Theorem B.3. *Let $d(x, x^*) \leq R$ for all $x \in \mathcal{M}$, $\phi(x) \leq \phi(x_0)$ and let $Q(x, x_k)$ be first-order surrogate functions. Let $\big(\tilde{Q}_k\big)_{k\geq 0}$ be a sequence of ϵ-approximate CCCP updates in the sense of Eq. 3.6. Then*

$$\phi(x_k) - \phi(x^*) \leq \frac{2L\alpha_{\mathcal{M}}^2(R)(1+\epsilon)}{k+2} \qquad \forall\, k \geq 1. \tag{B.1}$$

Proof. Replacing exact with inexact CCCP updates, we have

$$\phi(x_k) \leq \min_{x\in\mathcal{M}} \left[\phi(x) + \frac{L}{2}\|x - x_{k-1}\|^2 + \tfrac{1}{2}L\alpha_{\mathcal{M}}^2(R)s^2\epsilon\right] \,.$$

Following the steps of the proof of Thm. 3.5 to Eq.(3.5), we get

$$\begin{aligned}\phi(x_k) - \phi(x^*) \leq \min_{s\in[0,1]} \big[&(1-s)(\phi(x_{k-1}) - \phi(x^*)) \\ &+ \tfrac{1}{2}L\alpha_{\mathcal{M}}^2(R)s^2(1+\epsilon)\big] \,.\end{aligned}$$

The claim follows from an analysis the step-sizes analogously to the proof of Thm. 3.5. □

B.4. Exploiting Finite-sum Structure

We give a proof of Thm. 3.8:

Theorem B.4. *Let again $d(x, x^*) \leq R$ for all $x \in \mathcal{M}$ and $\phi(x) \leq \phi(x_0)$. Assume that $g_{i_k}^k$ as defined in Alg. 2 is a first-order surrogate of h_{i_k} near x_{k-1}. Then Alg. 2 converges almost surely.*

Proof. As outlined in Alg. 2, we use the following majorization to construct the CCCP oracle:

$$\begin{aligned} g^k(x) &:= \frac{1}{m}\sum_{i=1}^m g_i^k(x) \\ g_i^k(x) &= \begin{cases} h_{i_k}(x_k) - \langle \nabla h_{i_k}(x_k),\, x - x_k\rangle, & \text{if } i = i_k \\ g_i^{k-1}, & \text{if } i \neq i_k \end{cases} \,. \end{aligned}$$

By construction, this gives

$$g^k(x) = g^{k-1}(x) + \frac{g^k_{i_k}(x) - g^{k-1}_{i_k}(x)}{m} \ . \tag{B.2}$$

Observe that

$$\begin{aligned} g^k(x_k) &\overset{(1)}{\leq} g^k(x_{k-1}) \\ &\overset{(2)}{=} g^{k-1}(x_{k-1}) + \frac{g^k_{i_k}(x_{k-1}) - g^{k-1}_{i_k}(x_{k-1})}{m} \\ &\overset{(3)}{=} g^{k-1}(x_{k-1}) + \frac{h_{i_k}(x_{k-1}) - g^k_{i_k}(x_{k-1})}{m} \\ &\overset{(4)}{\leq} g^{k-1}(x_{k-1}) \ . \end{aligned}$$

Here, (1) follows from x_k being the argmin determined in the CCCP step and (2) from Eq. B.2. By assumption, $g^k_{i_k}$ is a first-order surrogate of h_{i_k} near x_{k-1}, which implies by Def. 3.3(2) that $g^k_{i_k}(x_{k-1}) = h_{i_k}(x_{k-1})$ and therefore (3). (4) follows from $g^{k-1}_{i_k}$ being a majorization of h_{i_k}. With this, $\{\big(g^k(x_k)\big)\}_{k\geq 0}$ is monotonically decreasing. Due to the level-set assumption this ensures that the sequence converges almost surely.

Taking expectations in the chain of inequalities, we get monotone convergence of $\{\mathbb{E}\big[g^k(x_k)\big]\}_{k\geq 0}$. For the analysis of the approximation error $\big(g^k(x_k) - h(x_k)\big)$, note that

$$\begin{aligned} &\mathbb{E}\big[\sum_{k=0}^{\infty} g^k_{i_k}(x_k) - h_{i_{k+1}}(x_k)\big] \\ &\quad\overset{(5)}{=} \sum_{k=0}^{\infty} \mathbb{E}\big[g^k_{i_{k+1}}(x_k) - h_{i_{k+1}}(x_k)\big] \\ &\quad\overset{(6)}{=} \sum_{k=0}^{\infty} \mathbb{E}\Big[\mathbb{E}\big[g^k_{i_{k+1}}(x_k) - h_{i_{k+1}}(x_k)|\mathcal{F}_k\big]\Big] \\ &\quad= \sum_{k=0}^{\infty} \mathbb{E}\big[g^k(x_k) - h(x_k)\big] \\ &\quad\overset{(5)}{=} \mathbb{E}\big[\sum_{k=0}^{\infty} g^k(x_k) - h(x_k)\big] < \infty \ . \end{aligned}$$

Here, (5) follows from the Beppo Levi lemma; in (6), we have rewritten the previous equality with respect to the sigma-field $\mathcal{F}_k$ generated by the x_k. With this, we have that $\{\big(g^k(x_k) - h(x_k)\big)\} \to 0$ almost surely. Now, following the proof of Thm. 3.5, we conclude that the sequence of objective values generated by Alg. 2 convergences to the optimum almost surely. □