Structured Regularization for Constrained Optimization on the SPD Manifold

Andrew Cheng[1] and Melanie Weber[1]

[1]School of Engineering and Applied Sciences, Harvard University.

Abstract

Matrix-valued optimization tasks, including those involving symmetric positive definite (SPD) matrices, arise in a wide range of applications in machine learning, data science and statistics. Classically, such problems are solved via constrained Euclidean optimization, where the domain is viewed as a Euclidean space and the structure of the matrices (e.g., positive definiteness) enters as constraints. More recently, geometric approaches that leverage parametrizations of the problem as unconstrained tasks on the corresponding matrix manifold have been proposed. While they exhibit algorithmic benefits in many settings, they cannot directly handle additional constraints, such as inequality or sparsity constraints. A remedy comes in the form of constrained Riemannian optimization methods, notably, Riemannian Frank-Wolfe and Projected Gradient Descent. However, both algorithms require potentially expensive subroutines that can introduce computational bottlenecks in practise. To mitigate these shortcomings, we introduce a class of structured regularizers, based on symmetric gauge functions, which allow for solving constrained optimization on the SPD manifold with faster unconstrained methods. We show that our structured regularizers can be chosen to preserve or induce desirable structure, in particular convexity and "difference of convex" structure. We demonstrate the effectiveness of our approach in numerical experiments.

Keywords: Riemannian optimization, constrained optimization, positive definite matrices, SPD manifold, difference of convex optimization

1 Introduction

We study constrained optimization problems of the form

$$\min_{x \in \mathcal{X} \subset \mathbb{P}_d} \phi(x) \,, \tag{1}$$

where $\phi : \mathbb{P}_d \to \mathbb{R}$ is a smooth function defined on the symmetric, positive definite matrices $\mathbb{P}_d$ and $\mathcal{X} \subset \mathbb{P}_d$ a subset defined by geometric constraints. Problems of this form arise in many settings, including the computation of Tyler's M-estimators [1–3], robust subspace recovery [4], barycenters problems [5], matrix-square roots [6], the computation of Brascamp-Lieb constants [7], and learning determinantal point processes [8], among others.

Classical approaches for this class of problems include constrained Euclidean optimization, where the domain in problem 1 is viewed as a Euclidean space and the geometric structure of the problem enters as constraints. However, it is often beneficial to encode the positive definiteness constraint explicitly in the parametrization of the domain by solving problem 1 as a constrained problem on the manifold of symmetric positive definite matrices (SPD manifold). For instance, if the objective ϕ is geodesically convex with respect to the Riemannian metric, then one can provide global optimality certificates for first-order methods in the Riemannian setting. Consequently, several constrained Riemannian optimization methods have been proposed, including variants of Riemannian Projected Gradient Descent (R-PGD) [9] and projection-free Riemannian Frank-Wolfe (R-FW) methods [10, 11]. However, several shortcomings arise, which limit the applicability of those methods in practise. First, both R-PGD and R-FW rely on subroutines for implicitly imposing constraints, which can be costly in the geometric setting. Second, the geometric tools needed to implement Riemannian optimization methods, including Riemannian gradients, exponential maps, and parallel transport operators, often introduce significant computational overhead compared to their Euclidean counterparts. To mitigate both limitations, we will apply a *mixed Euclidean-Riemannian perspective*, which leverages the computational efficiency of Euclidean methods while admitting global optimality certificates thanks to geodesic convexity.

We propose a class of regularizers based on symmetric gauge functions, which allows for relaxing several types of constraints that frequently arise in geometric optimization. We show that this *structured regularization* of constrained optimization tasks can preserve desirable properties, such as geodesic convexity and difference of convex (DC) structure. Moreover, in some settings, DC structure may be induced by a suitable regularization. Optimization tasks with DC objectives can be solved using *Convex-Concave Procedures* (short: *CCCP*), a class of Euclidean solvers that can often numerically outperform classical first-order methods in practise [3, 5]. In settings where the DC objectives is geodesically convex, we can leverage a Riemannian analysis to obtain global optimality certificates [12]. We will show that this lens applies readily to our regularized objectives, allowing us to leverage CCCP with global optimality guarantees in the constrained setting. To the best of our knowledge, this represents

the first application of CCCP to constrained, geodesically convex programs. Importantly, our structured regularizers are highly modular, which simplifies the design of new regularizers for a variety of programs.

We present a convergence analysis of CCCP applied to our regularized objectives and discuss the computational complexity of CCCP, specifically with regards to the complexity of a crucial subroutine, the *CCCP oracle*, which lies at the heart of the algorithm. We propose effective solvers for this subroutine for several classical SPD optimization tasks. Our results allow for recovering and reinterpreting previously known specialized fixed point approaches for related problems in our structured regularization framework. We corroborate our theoretical finding with numerical experiments, which illustrate the competitive performance of the proposed methods.

1.1 Related Works

Riemannian optimization

Optimization on geometric domains has received significant attention in the machine learning and statistics communities motivated by a wide range of applications that involve structured data and models. This has led to the generalizations of many classical Euclidean algorithms to the Riemannian setting, including for geodesically convex [13–15] and nonconvex [16] problems, as well as constrained [9, 10] and stochastic [11, 15, 17, 18] settings. Constrained Riemannian optimization has largely focused on projection-based [9] and projection-free [10] first-order methods, which ensure the feasibility of the iterates via subroutines that can be expensive in practise. To the best of our knowledge, handling constraints via regularizers that are explicitly designed to preserve desirable structure in the objective (geodesic convexity, DC structure) has not been studied systematically in the prior Riemannian optimization literature.

Regularization in SPD optimization

Regularization techniques have been studied for many problems on the manifold of positive definite matrices. Notable examples include covariance estimation [19, 20], where the regularizer enforces sparsity constraints and optimistic likelihood estimation, where side information is leveraged for regularization [21]. Usually, the regularizer is designed with a specific task in mind. In contrast, here we aim to design a general class of regularizers and investigate its properties for preserving desirable structure in the objective.

DC optimization

The optimization of DC function has been extensively studied in the Euclidean optimization literature [22]. The CCCP algorithm has emerged as a popular solver for problems of this structure. More recently, geometric optimization problems with DC structure have received interest, including differences of g-convex functions [23–25] and differences of Euclidean convex functions that are g-convex [12]. While the former rely on Riemannian tools, such as exponential maps and Riemannian gradients, the latter can be implemented using purely Euclidean tools [12]. To the best of our knowledge

no extensions of CCCP to constrained geometric problems have been considered in the prior literature.

1.2 Summary of Contributions

We briefly summarize the main contributions of our work.

1. We introduce a class of structured regularizers for constrained optimization on the manifold of symmetric positive definite matrices. Our regularizers are based on symmetric gauge functions, whose inherent algebraic properties provide the regularizers with a modular structure. This allows for the design of custom regularizers for a range of constrained problems.
2. We show that structured regularizers can be chosen to preserve or induce desirable structure in the objective, in particular difference of convex structure and geodesic convexity. The former allows for leveraging a simple CCCP approach to find solutions, and the latter to guarantee their global optimality.
3. To the best of our knowledge, we provide the first analysis of Euclidean CCCP applied to constrained geodesically convex programs. This approach has notable computational benefits compared to existing constrained Riemannian optimization approaches.
4. We illustrate the utility of our approach in several sets of numerical experiments, highlighting the computational efficiency and numerical stability of the proposed CCCP methods in applications.

2 Background

2.1 Riemannian Geometry of $\mathbb{P}_d$

Throughout this paper we consider the set of real symmetric square matrices with strictly positive eigenvalues, denoted by

$$\mathbb{P}_d \stackrel{\text{def}}{=} \{X \in \mathbb{R}^{d\times d} : X \succ 0\}.$$

A *manifold* $\mathcal{M}$ is a topological space that is locally Euclidean with a tangent space $\mathcal{T}_x\mathcal{M}$ associated to each point $x \in \mathcal{X}$. If $\mathcal{M}$ is *smooth* and has a smoothly varying inner product $\langle u, v\rangle_x$ defined on $\mathcal{T}_x\mathcal{M}$ for $x \in \mathcal{M}$ then it is a *Riemannian manifold.*

Here, we focus on the algorithmic benefits of viewing $\mathbb{P}_d$ under both the *affine-invariant Riemannian* geometry and the *Euclidean* geometry. When endowed with the *affine-invariant* inner product

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

the positive definite matrices form a Riemannian manifold. Here, the tangent space $\mathbb{H}_d$ is the space of $d \times d$ real symmetric matrices. Under this geometry, given two points $A, B \in \mathbb{P}_d$ there is an explicit parametrization for the *unique geodesic* that interpolates

A to B given by

$$\gamma(t) = A^{1/2}\left(A^{-1/2}BA^{-1/2}\right)^t A^{1/2}, \quad 0 \le t \le 1. \tag{2}$$

The geodesic given in (2) is referred to as the *weighted geometric mean* of A and B. The midpoint of the geodesic denoted by $A\sharp B \overset{\text{def}}{=} \gamma(1/2)$ is known as the *geometric mean* of A and B. Furthermore, the *Riemannian metric* corresponding to this geometry is given by

$$\delta_R(A, B) = \left\|\log A^{-1/2}BA^{-1/2}\right\|_F. \tag{3}$$

It is important to remark that the resulting Riemannian manifold $\mathbb{P}_d$ is a *Cartan-Hadamard manifold*, i.e., it is *complete*, *simply connected*, and has *non-positive curvature*. The Cartan-Hadamard manifold setting is particularly suitable for geodesically convex analysis and optimization [14].

The Euclidean geometry of $\mathbb{P}_d$ is induced by endowing the symmetric positive definite matrices with the smooth inner product

$$\langle A, B\rangle = \operatorname{tr}(A^\top B) \qquad \forall A, B \in \mathbb{P}_d\,,$$

in which case the corresponding *Euclidean metric* is the Frobenius norm

$$d(A, B) = \|A - B\|_F.$$

In this case, we can view the set $\mathbb{P}_d$ as a *convex cone*, i.e., a set closed under conic combinations. This conic perspective lends itself to convex analysis and optimization [26].

We further provide definitions for several convexity and smoothness notions that will be used throughout the paper.

Definition 2.1 (Geodesic convexity of sets)**.** *We say that a set $S \subseteq \mathbb{P}_d$ is geodesically convex (short: g-convex) if for any two points $A, B \in \mathbb{P}_d$, the unique geodesic $\gamma : [0, 1] \to \mathbb{P}_d$ given by (2) lies entirely in S, i.e., the image satisfies $\gamma([0, 1]) \subseteq S$.*

Definition 2.2 (Geodesic convexity of functions)**.** *We say that $\phi : S \to \mathbb{R}$ is a geodesically convex function if $S \subseteq \mathbb{P}_d$ is geodesically convex and $f \circ \gamma : [0, 1] \to \mathbb{R}$ is (Euclidean) convex for each geodesic segment $\gamma : [0, 1] \to \mathbb{P}_d$ whose image is in S with $\gamma(0) \neq \gamma(1)$.*

Recall that the Riemannian gradient on $\mathbb{P}_d$ can be computed as follows. Let $\bar{\phi} : \mathbb{H}_d \to \mathbb{R}$ be a function over symmetric matrices with the Euclidean metric from $\mathbb{R}^{n\times n}$. Denote ϕ as the restriction $\phi \overset{\text{def}}{=} \bar{\phi}|_{\mathbb{P}_d}$ to the SPD manifold equipped with the invariant metric. Then the Riemannian gradient of ϕ and the Euclidean gradient of $\bar{\phi}$ are related as $\operatorname{grad}\phi(X) = X \operatorname{grad}\bar{\phi}(X)X$ for all $(X, V) \in \mathcal{T}_X(\mathbb{P}_d)$.

Definition 2.3. *If $\phi : \mathbb{P}_d \to \mathbb{R}$ is differentiable then its gradient $\operatorname{grad}\phi(X)$ is defined as the unique vector $V \in \mathcal{T}_X(\mathbb{P}_d)$ with $D\phi(X)[V] = \langle \operatorname{grad}\phi(X), V\rangle_X$.*

Although the Euclidean perspective provides fast algorithms with provable convergence guarantees, many classical optimization tasks are nonconvex under the Euclidean metric. However, a large subset of these tasks admit a geodesically convex formulation with respect to the Riemannian metric allowing provable convergence to global optima. Unfortunately, Riemannian optimization techniques can often introduce nontrivial computational overhead stemming from the cost of computing geometric tools, such as geodesics and Riemannian gradients. By adopting a *mixed Euclidean-Riemannian perspective* one can simultaneously reap the computational benefits of the Euclidean perspective and the theoretical guarantees obtained via a Riemannian analysis.

2.2 Difference of Convex (DC) Optimization

Optimization tasks on the positive definite matrices frequently exhibit a special structure, where the objective function can be written as a difference of two convex functions. Formally, we consider instances of problem 1, where $\phi(x) = f(x) - h(x)$ with $f(\cdot), h(\cdot)$ Euclidean convex and $h(\cdot)$ smooth. The convexity of $h(\cdot)$ directly implies that

$$-h(x) \leq -h(y) - \langle \nabla h(y), x - y \rangle \ .$$

We can use this inequality to upper bound the objective, which defines the following surrogate function:

$$\phi(x) \leq Q(x, y) \stackrel{\text{def}}{=} f(x) - h(y) - \langle \nabla h(y), x - y \rangle \ .$$

The idea of convex-concave procedures (short: CCCP) is to iteratively minimize this surrogate function instead of the original, non-convex objective (Algorithm 1). Notably, this algorithm is purely Euclidean and does not require the computation of Riemannian tools, such as exponential maps or parallel transport operators.

With a purely Euclidean analysis one can show that this algorithm converges asymptotically to a stationary point of the underlying objective [27], but due to non-convexity as non-asymptotic convergence analysis is challenging in the general case. However, if $\phi(\cdot)$ is in addition geodesically convex, then sublinear, global convergence guarantees can be obtained for the (purely Euclidean) CCCP algorithm:

Theorem 2.4 ([12])**.** *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*

$$\phi(x_k) - \phi(x^*) \leq \frac{2L\alpha_{\mathcal{M}}^2(R)}{k+2} \qquad \forall k \geq 1 \ , \tag{4}$$

where $\alpha_{\mathcal{M}}$ depends on the geometry of the manifold and L characterizes the smoothness of $h(\cdot)$.

We note that the CCCP algorithm cannot explicitly handle constraints in its standard form. The structured regularization techniques introduced in this work will allow for applying CCCP approaches to constrained tasks.

Algorithm 1 Convex-Concave Procedure (CCCP)

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

2.3 Symmetric Gauge Functions and Unitarily Invariant Norms on $\mathbb{P}_d$

In this section we recall classical results on symmetric gauge functions and their algebraic structure. The properties of this class of functions will form the basis for the design of our structured regularizers, introduced in the next section. A more comprehensive overview of symmetric gauge functions can be found in [28] and [29, Ch. IV].

We begin with a formal definition of *symmetric gauge functions*:

Definition 1 (Symmetric Gauge Functions.)**.** *A function $\Phi : \mathbb{R}^d \to \mathbb{R}_+$ is called a* symmetric gauge function *if*

1. *Φ is a norm.*
2. *$\Phi(\sigma_d(x)) = \Phi(x)$ for all $x \in \mathbb{R}^d$ and all permutation maps $\sigma_n : \mathbb{R}^d \to \mathbb{R}^d$. This is known as the* symmetric property*.*
3. *$\Phi(\alpha_1 x_1, \ldots, \alpha_d x_d) = \Phi(x_1, \ldots, x_d)$ for all $x \in \mathbb{R}^d$ and $\alpha_k \in \{\pm 1\}$. This is known as the* gauge invariant *or* absolute property*.*

It is easy to see that symmetric gauge functions are closed under positive scaling, which directly follows from the fact that it is a norm.

Proposition 2.5 (Closure under positive scaling)**.** *If $\Phi : \mathbb{R}^d \to \mathbb{R}$ is a symmetric gauge function then $\alpha\Phi(x)$ is a symmetric gauge function for all $\alpha > 0$.*

Observe that the class of symmetric gauge functions is closed under addition.

Proposition 2.6 (Closure under addition)**.** *If $\Phi_1, \ldots \Phi_n$ are symmetric gauge functions, then so is $\Phi = \Phi_1 + \cdots + \Phi_n$.*

Proof. We prove this for the case $n = 2$. We know that if Φ_1 and Φ_2 are norms on $\mathbb{R}^n$ and then so is $\Phi = \Phi_1 + \Phi_2$. Let $P : \mathbb{R}^n \to \mathbb{R}^n$ be a permutation matrix and fix some $x \in \mathbb{R}^n$. Then

$$\Phi(Px) \stackrel{\text{def}}{=} \Phi_1(Px) + \Phi_2(Px) = \Phi_1(x) + \Phi_2(x) \stackrel{\text{def}}{=} \Phi(x).$$

This establishes the symmetric property of Φ. To show gauge invariance, let $\{\alpha_j\}_{j=1}^n \subseteq \{\pm 1\}$. Then

$$\begin{aligned}\Phi(\alpha_1 x_1, \ldots, \alpha_n x_n) &= \Phi_1(\alpha_1 x_1, \ldots, \alpha_n x_n) + \Phi_2(\alpha_1 x_1, \ldots, \alpha_n x_n) \\ &= \Phi_1(x_1, \ldots, x_n) + \Phi_2(x_1, \ldots, x_n) \\ &= \Phi(x).\end{aligned}$$

Thus Φ satisfies all the properties of a symmetric gauge invariant function. We can induct to prove the general case $n > 2$. □

Recall the following definition of the dual of a norm on $\mathbb{R}^d$.

Definition 2. *If Φ is a norm on $\mathbb{R}^d$ the dual of Φ is denoted by $\Phi^* : \mathbb{R}^d \to \mathbb{R}_+$ and defined as*

$$\Phi^*(x) \stackrel{def}{=} \sup_{\{y \in \mathbb{R}^d : \Phi(y) \leq 1\}} |\langle x, y \rangle|.$$

where $\langle \cdot, \cdot \rangle$ is the cannonical inner product of $\mathbb{R}^d$.

The class of symmetric gauge functions are closed under taking the dual.

Proposition 2.7 (Closure under the dual)**.** *If $\Phi : \mathbb{R}^d \to \mathbb{R}_+$ is a symmetric gauge function then so is Φ^*.*

Proof. We simultaneously prove the symmetry and gauge invariance property of Φ^*. To this end, let $P_d \circ \mathbb{Z}_2$ be the group of permutations and sign changes of the coordinates. Since Φ is a gauge function we have

$$\{y \in \mathbb{R}^d : \Phi(y) \leq 1\} = \{y \in \mathbb{R}^d : \Phi(Py) \leq 1\} \qquad P \in P_d \circ \mathbb{Z}_2.$$

Moreover, observe that $|\langle Px, Py \rangle| = |\langle x, y \rangle|$ for all $x, y \in \mathbb{R}^n$ and all $P \in P_d \circ \mathbb{Z}_2$. Hence for all $P \in P_d \circ \mathbb{Z}_2$, we have

$$\begin{aligned}
\Phi^*(Px) &= \sup \{|\langle Px, y \rangle| : \Phi(y) \leq 1\} \\
&= \sup \left\{|\langle Px, y \rangle| : \Phi(P^{-1}y) \leq 1\right\} \\
&= \sup \{|\langle Px, P\tilde{y} \rangle| : \Phi(\tilde{y}) \leq 1\} \qquad \text{(Change-of-variables } \tilde{y} \leftarrow P^{-1}y\text{)} \\
&= \sup \{|\langle x, \tilde{y} \rangle| : \Phi(\tilde{y}) \leq 1\} \\
&= \Phi^*(x).
\end{aligned}$$

It is a standard fact that the dual of a norm is a norm. Hence Φ^* is indeed a symmetric gauge function. □

The following corollary illustrates that the family of ℓ_p norms are closed under the ℓ_p transformations.

Corollary 2.8 (Exercise IV.1.9 [29])**.** *Let Φ be a symmetric gauge function and let $p \geq 1$. Let*

$$\Phi^{(p)}(x) = \left[\Phi\left(|x|^p\right)\right]^{1/p}$$

then $\Phi^{(p)}$ is a symmetric gauge function. Suppose Φ_p denotes the family of ℓ_p norms, then

$$\Phi_{p_1}^{(p_2)} = \Phi_{p_1 p_2} \qquad \forall p_1, p_2 \geq 1.$$

3 Structured Regularization

We consider regularizations of problem 1 of the form

$$\min_{X \in \mathbb{P}_d} \hat{\phi}(X) \stackrel{\text{def}}{=} \phi(X) + R(X)\,, \tag{5}$$

where the regularizer $R : \mathbb{P}_d \to \mathbb{R}$ is determined by the structure of the constraints $\mathcal{X}$ and the objective ϕ. In this section, we will show a construction of R via symmetric gauge functions for two types of constraints and examine the properties of the regularized objective $\hat{\phi}$. We will further discuss the construction of a wider range of regularizers following similar ideas.

Sparsity regularization

In the first instance of problem 1 that we consider $\mathcal{X}$ induces sparsity in the eigenspectrum of the solution, i.e., it encourages low-rank solutions. The corresponding *sparsity regularization problem* is of the form

$$\hat{\phi}(X) \stackrel{\text{def}}{=} \phi(X) + \beta S(X) \,, \tag{6}$$

where $S : \mathbb{P}_d \to \mathbb{R}$ and $\beta \geq 0$ a tunable hyperparameter. Problems of this form arise for instance in the computation of statistical estimator, such as covariance estimation [19], Gaussian graphical models [30], which we discuss later in the paper, as well as in recommender systems [31].

Ball constraint regularization

In the second class of problems that we consider, side information on the solution is given by a coarse empirical estimate and we want to confine our optimization procedure within a neighborhood of this estimator. Let $d(X, \hat{X}) : \mathbb{P}_d \times \mathbb{P}_d \to \mathbb{R}_+$ denote a metric on the of $\mathbb{P}_d$, $\hat{X} \in \mathbb{P}_d$ some fixed nominal estimate of the optimum $X^* \stackrel{\text{def}}{=} \arg\min_{X \in \mathbb{P}_d} \phi(X)$, and $\beta \geq 0$ a tunable hyperparameter. Then the *ball constraint regularization* problem is given by

$$\hat{\phi}(X) \stackrel{\text{def}}{=} \phi(X) + \beta d(X, \hat{X}) \,. \tag{7}$$

Problems of this form arise in the computation of statistical estimators, e.g., the Karcher mean [32] and optimistic likelihood [21], both of which we discuss later in the paper.

Outline

This section is structured as follows. We will motivate symmetric gauge functions as a natural starting point from which we can generate the sparsity and the ball constraint regularizers. For each class of regularizers we will first show that they arise from symmetric gauge functions. Second, we justify their value by showcasing their desirable algorithmic properties for g-convex optimization. Third, we identify examples in which they arise in the prior optimization literature and show that rewriting these well-known regularizers in terms of our two classes of regularizers provides insight into their algorithmic properties. Fourth, we show that we can generate novel sparsity or ball constraint regularizers in a principled manner from these symmetric gauge functions. Moreover, leveraging the algebraic structure of symmetric gauge function, a wide range of new regularizers for other types of objectives and constraints can

be constructed. Finally, we present a principled way of tuning the hyperparameters introduced by these regularizers.

3.1 Symmetric Gauge Functions as Sparsity and Ball Constraint Regularizers

Let $\Phi \in \mathcal{SG}(\mathbb{P}_d)$ be a *symmetric gauge function* (short: *SG*) and let $\hat{X} \in \mathbb{P}_d$ be fixed. In this section, we will show how to find the corresponding sparsity and ball constraint regularizers, i.e., $S_\Phi(X)$ and $d_\Phi(X, \hat{X})$, respectively.

3.1.1 SG as Sparsity Regularizers

There is a deep connection between symmetric gauge functions and unitarily invariant norms by a theorem of von Neumann. We leverage this connection in order to find $S_\Phi(X)$.

Definition 3 (Unitarily Invariant Norms)**.** *We say that a norm $\|\cdot\|$ on $\mathbb{R}^{d\times d}$ is unitarily (or orthogonally) invariant if $\|UAV\| = \|A\|$ for all orthogonal operators U, V on $\mathbb{R}^{d\times d}$ and all $A \in \mathbb{R}^{d\times d}$.*

The next theorem is von Neumann's [33] result on the correspondence between symmetric gauge functions and unitarily invariant norms. We apply the result of Theorem IV.2.1 [29] to the context of $\mathbb{P}_d$.

Theorem 3.1 (von Neumann)**.** *Given a symmetric gauge function Φ on $\mathbb{R}^d$, define a function on $\mathbb{P}_d$ as*

$$\|A\|_\Phi = \Phi(\lambda(A))$$

where $\lambda(A)$ denotes the singular values of A. Then this defines a unitarily invariant norm on $\mathbb{P}_d$. Conversely, given any unitarily invariant norm $\|\cdot\|$ on $\mathbb{P}_d$, define a function on $\mathbb{R}^d$ by

$$\Phi_{\|\cdot\|}(x) = ||\operatorname{diag}(x)||,$$

where $\operatorname{diag}(x)$ *is the diagonal matrix with entries $x_1, \ldots, x_d$ on its diagonal. Then this defines a symmetric gauge function on $\mathbb{R}^d$.*

Remark 3.2. Given a symmetric gauge function $\Phi \in \mathcal{SG}(\mathbb{P}_d)$, we propose using the corresponding unitarily invariant norm as our sparsity regularizer, that is, $S_\Phi(X) = \|X\|_\Phi$.

Example 3.3. *We provide two well-known examples of symmetric gauge function and their associated unitarily invariant norms. Let $A \in \mathbb{P}_d$ and $\lambda(A) = \{\lambda_1, \ldots, \lambda_d\} \in \mathbb{R}^d$ denote the ordered eigenvalues, i.e., $\lambda_1 \geq \cdots \geq \lambda_d > 0$. First, consider the class of Schatten p-norms defined as*

$$\|A\|_p = \Phi_p(\lambda(A)) = \left(\sum_{j=1}^{n} |\lambda_j(A)|^p\right)^{\frac{1}{p}}, \qquad 1 \leq p < \infty,$$

$$\|A\|_\infty = \Phi_\infty(\lambda(A)) = \lambda_1(A) = \|A\|.$$

Second, consider the class of Ky Fan k-norms defined as

$$\|A\|_{(k)} = \sum_{j=1}^{k} |\lambda_j^{\downarrow}(A)|, \quad 1 \leq k \leq n$$

which is the sum of the first k largest eigenvalues of A.

3.1.2 SG as Ball Constraints

In addition, for every symmetric gauge function $\Phi \in \mathcal{SG}(\mathbb{P}_d)$ there exists a complete metric d_Φ on the convex cone of $\mathbb{P}_d$. These metrics can provide a way to encourage our iterates to stay within a neighborhood of an estimated solution.

We can define the length of a path $\gamma : [0,1] \to \mathbb{P}_d$ with respect to Φ as follows.

Definition 4. *For a path $\gamma : [0,1] \to \mathbb{P}_d$ we define its length w.r.t a symmetric gauge function $\Phi : \mathbb{R}^n \to \mathbb{R}_+$ as*

$$L_\Phi(\gamma) \stackrel{def}{=} \int_0^1 \left\| \gamma^{-1/2}(t)\gamma'(t)\gamma^{-1/2}(t) \right\|_\Phi dt.$$

This gives rise to the following metric d_Φ associated with Φ.

Definition 5. *For a given symmetric gauge function $\Phi : \mathbb{R}^d \to \mathbb{R}_+$ we define the distance between $A, B \in \mathbb{P}_d$ with respect to Φ as*

$$d_\Phi(A,B) \stackrel{def}{=} \inf \{ L_\Phi(\gamma) : \gamma \text{ is a path from } A \text{ to } B \}.$$

It turns out that d_Φ can be expressed in terms of the unitarily invariant norm $\|\cdot\|_\Phi$. Moreover d_Φ is a complete metric on the convex cone of $\mathbb{P}_d$ with several nice properties illustrated by the following theorem.

Theorem 3.4 (Theorem 2.2 [28])**.** *We have $d_\Phi(A,B) = \left\| \log\left(A^{-1/2}BA^{-1/2}\right) \right\|_\Phi$ and d_Φ is a complete metric distance on the convex cone of $\mathbb{P}_d$ such that for $A, B \in \mathbb{P}_d$ and for invertible matrix M,*

1. $d_\Phi(A,B) = d_\Phi\left(A^{-1}, B^{-1}\right) = d_\Phi\left(MAM^*, MBM^*\right)$;
2. $d_\Phi(A\#B, A) = d_\Phi(A\#B, B) = \frac{1}{2} d_\Phi(A,B)$, *where* $A\#B = A\#_{\frac{1}{2}}B$;
3. $d_\Phi\left(A\#_t B, A\#_s B\right) = |s-t| d_\Phi(A,B)$ *for all* $t, s \in [0,1]$;
4. $d_\Phi\left(A\#_t B, C\#_t D\right) \leq (1-t) d_\Phi(A,C) + t d_\Phi(B,D)$ *for all* $t \in [0,1]$.

The following examples show that d_Φ recovers well-known metrics for specific symmetric gauge functions Φ.

Example 3.5. *If we specify Φ to be the Schatten 2-norm then the corresponding metric corresponds to the Riemannian affine-invariant metric on $\mathbb{P}_d$. Moreover, if we choose Φ to be the ∞-Schatten norm then the associated metric is the Thompson metric [34] on the positive definite cone. Namely,*

$$d_\infty(A,B) = \max\{\log M(B/A), \log M(A/B)\}$$

where $M(B/A) \stackrel{def}{=} \inf\{\alpha > 0 : B \leq \alpha A\} = \lambda_1\left(A^{-1/2}BA^{-1/2}\right) = \lambda_1\left(A^{-1}B\right)$.

3.2 Properties of Symmetric Gauge Function Regularizers

In this section we discuss useful properties of general symmetric gauge functions, and our regularizers $S_\Phi(X)$, $d_\Phi(X, \hat{X})$ in particular. We will discuss the algorithmic implications of these results below in sec. 4.1.

3.2.1 Preserving g-convexity

The following proposition shows that unitarily invariant norms are g-convex on $\mathbb{P}_d$ which implies that our regularized objective (Eq. 6) remains g-convex. We will need the following definition.

Definition 6. *Let $x, y \in \mathbb{R}^d$. We say x* is weakly-submajorized by *y if*

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

We denote this by $x \prec_w y$.

Proposition 3.6 (Unitarily Invariant Norms are g-convex)**.** *Let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be a symmetric gauge function. Then the unitarily invariant norm $\|\cdot\|_\Phi : \mathbb{P}_d \to \mathbb{R}_+$ defined by $\|\cdot\|_\Phi = \Phi(\lambda(A))$ is g-convex.*

Proof. To show $\|\cdot\|_\Phi$ is g-convex it suffices to verify midpoint g-convexity. We use the notation $\lambda^{\downarrow}(A) \preceq \lambda^{\downarrow}(B)$ to denote $\lambda_j(A) \leq \lambda_j(B)$ for $j = 1, \ldots, d$ for the spectrum ordered in decreasing order, i.e.,

$$\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_d(A) \qquad \text{and} \qquad \lambda_1(B) \geq \lambda_2(B) \geq \cdots \geq \lambda_d(B)\ .$$

It is known that symmetric gauge functions are monotone [29], that is, if $\lambda^{\downarrow}(A) \preceq \lambda^{\downarrow}(B)$ then

$$\|A\|_\Phi = \Phi(\lambda^{\downarrow}(A)) \leq \Phi(\lambda^{\downarrow}(B)) = \|B\|_\Phi\ ,$$

where the equalities follow from the permutation invariance property of Φ. For $A, B \in \mathbb{P}_d$ the weighted geometric mean satisfies [35, Exercise 6.5.6]:

$$A \#_t B \preceq (1-t)A + tB \qquad \text{for } t \in [0, 1].$$

Recall that if $A \succeq B$ then $\lambda^{\downarrow}(A) \succeq \lambda^{\downarrow}(B)$ which follows from the min-max theorem:

$$\lambda_k(A) = \min_{\substack{U \subset \mathbb{C}^n \\ \dim(U)=k}} \max_{x \in U \setminus \{0\}} \frac{x^\top A x}{x^\top x} \geq \min_{\substack{U \subset \mathbb{C}^n \\ \dim(U)=k}} \max_{x \in U \setminus \{0\}} \frac{x^\top B x}{x^\top x} = \lambda_k(B),$$

for $k \in [d]$ where the inequality follows from the fact that $A \succeq B \implies A - B \succeq 0$, that is $x^\top (A - B)x \geq 0$ for all vectors $x \neq 0$. Hence setting $t = 1/2$ in the geometric mean we have

$$A \# B \preceq \frac{A+B}{2} \implies \lambda^{\downarrow}(A \# B) \preceq \lambda^{\downarrow}\left(\frac{A+B}{2}\right). \tag{8}$$

Thus using (8) and applying the monotonicity and permutation invariance of Φ we get

$$\Phi(\lambda(A\#B)) = \Phi(\lambda^{\downarrow}(A\#B)) \leq \Phi\left(\lambda^{\downarrow}\left(\frac{A+B}{2}\right)\right) = \Phi\left(\lambda\left(\frac{A+B}{2}\right)\right).$$

Moreover, Exercise II.1.14 [29] implies

$$\lambda^{\downarrow}\left(\frac{A+B}{2}\right) \prec_w \lambda^{\downarrow}\left(\frac{A}{2}\right) + \lambda^{\downarrow}\left(\frac{B}{2}\right). \tag{9}$$

We further know that Φ satisfies the *strongly isotone* property (see [29, page 45]), i.e.,

$$x \prec_w y \implies \Phi(x) \leq \Phi(y) \qquad \forall x, y \in \mathbb{R}^n_+.$$

Applying permutation invariance of Φ with (9) gives

$$\Phi\left(\lambda\left(\frac{A+B}{2}\right)\right) = \Phi\left(\lambda^{\downarrow}\left(\frac{A+B}{2}\right)\right) \leq \Phi\left(\lambda^{\downarrow}\left(\frac{A}{2}\right) + \lambda^{\downarrow}\left(\frac{B}{2}\right)\right). \tag{10}$$

Finally, we have for all $A, B \in \mathbb{P}_d$,

$$\begin{aligned}
\|A\#B\|_\Phi &= \Phi(\lambda(A\#B)) \\
&\leq \Phi\left(\lambda\left(\frac{A+B}{2}\right)\right) && \text{(Applying monotonicity of } \Phi \text{ to (8))} \\
&\leq \Phi\left(\lambda^{\downarrow}\left(\frac{A}{2}\right) + \lambda^{\downarrow}\left(\frac{B}{2}\right)\right) && \text{(Apply (10))} \\
&= \Phi\left(\frac{1}{2}\lambda^{\downarrow}(A) + \frac{1}{2}\lambda^{\downarrow}(B)\right) && \text{(Property of eigenvalues: } \lambda(A/2) = \frac{1}{2}\lambda(A)) \\
&\leq \frac{\Phi(\lambda^{\downarrow}(A)) + \Phi(\lambda^{\downarrow}(B))}{2} && (\Phi \text{ is a norm; triangle inequ., pos. homogeneity}) \\
&= \frac{\Phi(\lambda(A)) + \Phi(\lambda(B))}{2} && \text{(Remove } \downarrow \text{ by permutation invariance of } \Phi) \\
&\overset{\text{def}}{=} \frac{\|A\|_\Phi + \|B\|_\Phi}{2},
\end{aligned}$$

which proves the midpoint criterion for g-convexity. □

This implies that regularizing a g-convex objective with a symmetric gauge function (or its corresponding unitarily invariant norm) will maintain g-convexity. Moreover, since symmetric gauge functions are closed under positive scaling, we can always select a hyperparameter $\beta > 0$ to control the regularity of Φ.

Remark 3.7. *We remark that functions that are convex with respect to the Euclidean metric may not be g-convex with respect to the Riemannian metric and vice-versa. For*

example, let $y_1, \dots, y_n \in \mathbb{R}^d$ be non-zero vectors and define the functions $f, g : \mathbb{P}_d \to \mathbb{R}$ by

$$f(X) \stackrel{def}{=} \sum_{ij=1}^{d} |X_{ij}| \qquad \text{and} \qquad g(X) \stackrel{def}{=} \sum_{i=1}^{n} \log\left(y_i^\top X y_i\right).$$

$f(X)$ is Euclidean convex since it is a norm but it is not g-convex with respect to the Riemannian metric (see [36, Proposition 11]). On the other hand, $g(X)$ is g-convex with respect to the Riemannian metric but not Euclidean convex (see [36, Proposition 12]).

3.2.2 Properties of Sparsity Regularizers

The following result shows that symmetric gauge functions are closed under a ℓ_p-*transformation*, which provides us with additional flexibility to tune its regularity.

Proposition 3.8. *Let Φ be a symmetric gauge function and $p \geq 1$. Define the ℓ_p-transformation of Φ to be*

$$\Phi^{(p)}(x) \stackrel{def}{=} \left[\Phi(|x|^p)\right]^{1/p}.$$

Then $\Phi^{(p)}$ is also a symmetric gauge function.

Proof. Gauge invariance and symmetry follows directly from the definition of $\Phi^{(p)}$. It remains to show that $\Phi^{(p)}$ is a norm. The absolute homogeneity and positive definiteness property of $\Phi^{(p)}$ follows directly from definition. The triangle inequality directly follows by applying Minkowski's inequality for symmetric gauge functions [29, Theorem IV.1.8]. □

Thus we can write Problem 6 as the following problem with more fine-grained regularity on the sparsity. For any $p \geq 1$ and $\beta > 0$, we can rewrite Problem 6 as

$$\hat{\phi}(X) = \operatorname*{arg\,min}_{X \in \mathbb{P}_d} \phi(X) + \beta \Phi^{(p)}(\lambda(X)), \tag{11}$$

which is g-convex. We note that if ϕ is g-convex and DC, then so is $\hat{\phi}$ in (11). However, if ϕ is not DC then $\hat{\phi}$ is not necessarily DC. In that case, we may construct a symmetric gauge function regularizer to induce DC structure and sparse solutions (see Proposition 3.24 below for an example).

3.2.3 Properties of Ball Constraint Regularizers

Similarly, we have the following desirable algorithmic properties of d_Φ.

Proposition 3.9 (Proposition 3.5 [28])**.** *Let $\Phi \in \mathcal{SG}(\mathbb{P}_d)$ be a symmetric gauge function and $d_\Phi : \mathbb{P}_d \times \mathbb{P}_d \to \mathbb{R}$ be the corresponding metric. Then*

1. *Every d_Φ-ball is geodesically convex in $\mathbb{P}_d$.*
2. *The map $d_\Phi^\alpha(\cdot, Z) : \mathbb{P}_d \to \mathbb{R}$ is geodesically convex for any $\alpha \geq 1$.*

Proposition 3.9 implies that our ball constraint regularizer preserves desirable properties: For a g-convex and DC objective $\phi : \mathbb{P}_d \to \mathbb{R}$ and an appropriate choice of $\beta > 0$

and $\alpha \geq 1$, the regularized problem

$$\arg\min_{X \in \mathbb{P}_d} f(X) + \beta d_\Phi^\alpha(X, Z)$$

is g-convex and DC, too. If DC structure is not present, one can use a regularizer that encodes the S-divergence, a function of symmetric gauge functions, to induce DC structure while retaining g-convexity (see Example 3.23 for more details).

3.3 Designing new regularizers from symmetric gauge functions

While the sparsity and ball constraints discussed above allow for regularizing many common constrained problem, the idea of leveraging symmetric gauge functions as regularizers applies more broadly. In this section we discuss how to build a rich class of regularizers for a wide range of constrained problems on the SPD manifold, using symmetric gauge functions and their convexity-preserving properties.

3.3.1 Disciplined Geodesically Convex Programming with symmetric gauge functions

Disciplined Geodesically Convex Programming (short: DGCP) [36] is a framework for testing and verifying the geodesic convexity of nonlinear programs on Cartan-Hadamard manifolds. DGCP presents a set of g-convex functions (*atoms*) and g-convex preserving operations (*rules*) for the $\mathbb{P}_d$ manifold. To test and verify convexity, a function is expressed in terms of atoms using only convexity preserving operations. Below, we discuss g-convex preserving operations on the symmetric gauge functions, which can be used to design new g-convex regularizers on $\mathbb{P}_d$.

Rules

Proposition 3.10. *Let $S \subseteq \mathbb{P}_d$ be a g-convex subset. Suppose the functions $f_i : S \to \mathbb{R}$ are g-convex for $i = 1, \ldots, n$. Then the following functions are g-convex*

1. $f(X) = \max_{i \in \{1,\ldots,n\}} f_i(X)$
2. $g(X) = \sum_{i=1}^n \alpha_i f_i(X)$ *for* $\alpha_1 \ldots, \alpha_n \geq 0$.

The classes of positive linear and affine maps will be useful in identifying and constructing g-convex functions.

Definition 3.11 (Positive Linear Map [35])**.** *A linear map $\Phi : \mathbb{P}_d \to \mathbb{P}_m$ is positive when $A \succeq 0$ implies $\Phi(A) \succeq 0$ for all $A \in \mathbb{P}_m$. We say that Φ is strictly positive when $A \succ 0$ implies that $\Phi(A) \succ 0$.*

Definition 3.12 (Positive Affine Map [36])**.** *Let $B \succeq 0$ be a fixed symmetric positive semi-definite matrix and $\Phi : \mathbb{P}_d \to \mathbb{P}_d$ a positive linear operator. Then the function $\phi : \mathbb{P}_d \to \mathbb{P}_d$ defined by $\phi(X) \stackrel{def}{=} \Phi(X) + B$ is a positive affine operator.*

Notably, strictly positive linear maps are g-convex.

Proposition 3.13 (Proposition 5.8 [37])**.** *Let $\Phi(X)$ be a strictly positive linear operator from $\mathbb{P}_d$ to $\mathbb{P}_m$. Then $\Phi(X)$ is g-convex with respect to the Löwner order on $\mathbb{P}_m$ over $\mathbb{P}_d$ with respect to the canonical Riemannian inner product $g_X(U, V) \stackrel{def}{=}$*

$\operatorname{tr}\left[X^{-1}UX^{-1}V\right]$. *In other words, for any geodesic* $\gamma : [0,1] \to \mathbb{P}_d$ *we have that*

$$\Phi(\gamma(t)) \preceq (1-t)\Phi(\gamma(0)) + t\Phi(\gamma(1)) \quad \forall t \in [0,1]\,.$$

We can further leverage the following g-convexity preserving compositions.

Proposition 3.14 (Proposition 7 [36])**.** *Let* $\phi(X) \stackrel{def}{=} \Phi(X) + B$ *where* $\Phi(X)$ *is a positive linear map and* $B \succeq 0$*. Let* $f : \mathbb{P}_d \to \mathbb{P}_m$ *be g-convex and monotonically increasing, i.e.,* $f(X) \preceq f(Y)$ *whenever* $X \preceq Y$*. Then the function* $g(X) \stackrel{def}{=} f(\phi(X))$ *is g-convex.*

Proposition 3.15 (Proposition 2 [36])**.** *Suppose* $f : \mathbb{P}_d \to \mathbb{R}$ *is g-convex. If* $h : \mathbb{R} \to \mathbb{R}$ *is nondecreasing and Euclidean convex, then* $h \circ f$ *is g-convex on* $\mathbb{P}_d$*.*

Finally, the variable substitution $X = X^{-1}$ preserves g-convexity.

Proposition 3.16 ([36])**.** *Let* $f : \mathbb{P}_d \to \mathbb{R}$ *be g-convex. Then* $g(X) = f(X^{-1})$ *is also g-convex.*

Atoms

DGCP provides a fundamental set of atoms, which, when combined with its *rules*, can be used to design novel regularizers for optimization on the $\mathbb{P}_d$ manifold.

Example 3.17 (G-convex Atoms)**.** *The following functions* $\phi(X) : \mathbb{P}_d \to \mathbb{P}_m$ *are g-convex with respect to the Riemannian affine invariant metric.*

1. $\phi(X) = \log\det X$
2. $\phi(X) = tr(X)$
3. *If* $S \succeq 0$ *and* $\phi(X) = tr(SX)$
4. *Let* $M \succeq 0$ *and* M *has no zero rows and* $\phi(X) = M \odot X$ *(ϕ is a strictly positive linear map)*
5. $\phi(X) = \Phi(\lambda(X))$ *for any symmetric gauge function* $\Phi(X) : \mathbb{R}^d \to \mathbb{R}$
6. *Let* $y \in \mathbb{R}^d$ *be a nonzero vector and* $r \in \{-1, 1\}$*. Then the function* $\phi(X) = y^\top X^r y$ *is strictly g-convex*
7. *Let* $h_i \in \mathbb{R}^d$ *be nonzero vectors for* $i = 1, \ldots, n$ *and* $r \in \{-1, 1\}$*. Then* $\phi(X) = \log\left(\sum_{i=1}^n h_i^\top X^r h_i\right)$ *is strictly g-convex.*

3.3.2 Designing new regularizers

The utility of the DGCP framework, in conjunction with the class of symmetric functions, is now evident. By using the class of symmetric gauge functions $\mathcal{SG}(\mathbb{P}_d)$ along with the set of g-convex *atoms*, and applying g-convex preserving *rules*, we can design novel g-convex (and possibly DC) regularizers for optimization problems on $\mathbb{P}_d$. Conversely, DGCP allows us to principally decompose known regularizers in terms of symmetric gauge functions and atoms to gain intuition of the properties that they induce. In short, DGCP and symmetric gauge functions presents a principled way of reasoning about regularization on $\mathbb{P}_d$.

Below, we discuss a range of examples. Some of them recover known regularizers, reinterpreted via symmetric gauge functions. In particular, we will see that many classical regularizers can be written as compositions and transformations of the Schatten

1-norm, which induces sparsity. We also introduce novel regularizers that have not been studied previously.

Known regularizers as symmetric gauge functions

Lemma 3.18 (Regularizing Top-k Eigenvalues)**.** *Let $\Phi : \mathbb{P}_d \to \mathbb{R}$ be the k-*Ky-Fan norm *restricted to the set of positive definite matrices. For $k \in \{1, \ldots, d\}$, and $h(x) = |x|^p$ for $p \geq 1$. Then $R_\Phi(X) = \left|\sum_{j=1}^k \lambda_j^\downarrow(X)\right|^p$ is g-convex.*

This follows directly from Proposition 3.15.

Proposition 3.19 (Low-rank Regularizer)**.** *Let $\Phi : \mathbb{P}_d \to \mathbb{R}$ be the n-*Ky-Fan norm *or the* Schatten 1-norm*. Then the low-rank inducing trace regularizer can be written as $R_\Phi(X) = \Phi(\lambda(X)) = tr(X)$.*

Note that we can apply Proposition 3.15 with $h(x) = \exp(x)$ to get another g-convex regularizer $R_\Phi(X) = \exp(\mathrm{tr}(X))$.

The following proposition illustrates two structure-inducing properties of the log-determinant barrier function $f(X) \stackrel{\text{def}}{=} \log\det X$, which is a g-linear (i.e. g-convex and g-concave) function. First, it encourages the iterates to be low-rank by penalizing the sum of the log of their eigenvalues. Second, it encourages the iterates to stay close to the identity matrix I_d via the metric $d_\Phi(X, I_d)$.

Proposition 3.20 (Log-Det Barrier function)**.** *Let Φ be the Schatten 1-norm and let $Z = I_d$ fixed. Then we have*

$$\log\det X = d_\Phi(X, I_d) = \sum_{j=1}^d \log \lambda_j(X).$$

Next we introduce another class of sparsity-inducing regularizers, defined via Schatten p-norms.

Proposition 3.21 (Smooth Schatten p-Functions)**.** *Define R_Φ^p as*

$$R_\Phi^p(X) \stackrel{def}{=} \mathrm{Tr}\,(X + \gamma I)^{p/2} = \sum_{i=1}^d (\lambda_i(X) + \gamma)^{p/2} = \|X + \gamma I\|_\Phi^{p/2}$$

where Φ is the Schatten 1-norm. This is known as the smooth Schatten p-function [38].

The Schatten p-function is differentiable for $p > 0$ and Euclidean convex (short: e-convex) and g-convex for $p \geq 1$. For $p \in [0, 1)$, $R_\Phi^{(p)}$ is used in the iteratively reweighted least squares (IRLS-p) algorithm to obtain low-rank solutions under affine constraints, i.e., the affine rank minimization problem [38]. The level of sparsity is regulated by p where stronger sparsity is induced for smaller values of p.

Remark 3.22. Interestingly, if we take $p \to 0$, we obtain a familiar g-convex (in fact, g-linear) regularizer:

$$\lim_{p\to 0}\frac{R_\Phi^{(p)}(X)-d}{p} = \frac{1}{2}\sum_{i=1}^{d}\frac{\left[(\lambda_i(X)+\gamma)^{p/2}-1\right]}{\frac{p}{2}}$$
$$= \frac{1}{2}\sum_{i=1}^{d}\log\left(\lambda_i(X)+\gamma\right) \qquad \left(\lim_{p\to 0}\frac{x^p-1}{p} = \log x \text{ for } x\in\mathbb{R}_{++}\right)$$
$$= \frac{1}{2}\log\det\left(X+\gamma I\right) \ ,$$

which is used in the IRLS-0 algorithm in [38].

The following lemma gives a useful connection between the S-divergence and the Schatten 1-norm.

Lemma 3.23 (S-Divergence)**.** *Choose Φ to be the Schatten 1-norm. The S-divergence $\delta_S^2(X,Y):\mathbb{P}_d\times\mathbb{P}_d\to\mathbb{R}_+$ is defined by*

$$\delta_S^2(X,Y) \stackrel{def}{=} \log\det\left(\frac{X+Y}{2}\right) - \frac{1}{2}\log\det(XY)$$

and can be expressed as

$$\delta_S^2(X,Y) = d_\Phi(X,X+Y) - \frac{1}{2}d_\Phi(X,Y) - \log 2\|I_d\|_\Phi.$$

Proof. To see this, we perform simple algebra on the definition of $\delta_S^2(\cdot,\cdot)$:

$$\delta_S^2(X,Y) = \log\det X + \log\det\left(I_d + X^{-1}Y\right) - n\log 2 - \frac{1}{2}\left[\log\det X + \log\det Y\right]$$
$$= \log\det\left(X^{-1}Y + I_d\right) - \frac{1}{2}\log\det X^{-1}Y - n\log 2$$
$$= \sum_{i=1}^{d}\left[\log(1+\lambda_i(X^{-1}Y)\right] - \frac{1}{2}\sum_{i=1}^{d}\lambda_i(X^{-1}Y) - n\log 2$$
$$= d_\Phi(X,X+Y) - \frac{1}{2}d_\Phi(X,Y) - \log 2\|I_d\|_\Phi \ ,$$

where Φ is the Schatten 1-norm. □

$\delta_S^2(\cdot,Y)$ has many desirable algorithmic properties. It is g-convex and DC, a metric, and its gradient can be computed particularly efficiently. It can be seen as a symmetrized version of the *Log-Determinant Divergence.* The S-divergence is extensively studied in [5].

Proposition 3.24 (Diagonal Loading)**.** *We can sum the log-det barrier and the trace-inverse regularizer to get the diagonal loading regularizer $R_\Phi(X):\mathbb{P}_d\to\mathbb{R}$ defined by*

$$R_\Phi(X) \stackrel{def}{=} trX^{-1} + \log\det X = \|X^{-1}\|_\Phi + d_\Phi(X,I_d) \ ,$$

where Φ is the Schatten 1-norm. Then $R_\Phi(X)$ is g-convex and DC in X.

Proof. Observe that $R_\Phi(X)$ is g-convex since it is a sum of a unitarily invariant norm $\|\cdot\|_\Phi$ and distance metric $d_\Phi(\cdot,\cdot)$ corresponding to a symmetric gauge function Φ. The DC structure follows from the fact that $f(X) = \log\det(X)$ and $g(X) = \mathrm{tr}X^{-1}$ are concave and convex, respectively. □

This regularizer is also known as the *shrinkage to identity* regularizer for covariance estimation [20]. It is used when the number of samples is small relative to the number of features d. It encourages the solution to be close to the identity I_d which is illustrated by the $d_\Phi(X, I_d)$ term. Another way to check this is to show that $R_\Phi(X)$ is minimized at I by setting its gradient to zero.

Symmetric gauge functions in applications

In many classical problems, the objective itself can be written in terms of symmetric gauge functions without explicit regularizers. The techniques discussed above, including the DGCP framework, can be applied to these problems, too. We illustrate this on three examples.

Example 3.25 (Square Root)**.** *Sra [6] introduced a formulation of the problem of computing the square root of $A \in \mathbb{P}_d$ using the S-divergence (see Example 3.23):*

$$\min_{X\in\mathbb{P}_d} \left\{ \phi(X) \stackrel{def}{=} \delta_S^2(X, A) + \delta_S^2(X, I) \right\}.$$

In fact, this formulation provides a parametrization of the problem in terms of transformations and compositions of symmetric gauge functions. It has several desirable algorithmic properties, including DC structure and g-convexity. We will discuss this example in more detail below.

Example 3.26 (Karcher Mean)**.** *The Karcher mean problem [12] is the solution to the following problem. Given data $\{A_1, \ldots, A_m\} \in \mathbb{P}_d$ and $w \in \mathbb{R}^m_+$ such that $\sum_{i=1}^m w_i = 1$ we solve*

$$\min_{X\in\mathbb{P}_d} \left\{ \phi(X) \stackrel{def}{=} \sum_{i=1}^m w_i \delta_R^2(X, A_i) \right\},$$

where $\delta_R(X, A_i) \stackrel{def}{=} \|X^{-1/2} A_i X^{-1/2}\|_F$ is the Riemannian metric. Sra [5] showed that the Karcher mean problem can be reformulated as

$$\min_{X\in\mathbb{P}_d} \left\{ \phi(X) \stackrel{def}{=} \sum_{i=1}^m w_i \delta_S^2(X, A_i) \right\},$$

where δ_S^2 is again the S-divergence. The problem is g-convex and DC.

Example 3.27 (Tyler's Estimator with Diagonal Loading)**.** *Tyler's estimator [1] is a well-known robust covariance estimator. It is defined as the solution to the g-convex*

and DC objective $\phi(\Sigma) : \mathbb{P}_d \to \mathbb{R}$ *defined by*

$$\arg\min_{\Sigma \in \mathbb{P}_d} \phi(\Sigma) \stackrel{def}{=} \frac{d}{n} \sum_{i=1}^{n} \log\left(x_i^T \mathbf{\Sigma}^{-1} x_i\right) + \log\det(\Sigma) \ ,$$

where $\{x_i\}_{i=1}^n \subseteq \mathbb{R}^d$ *are observed data samples. Wiesel and Zhang [20] introduce the penalty* $R_\Phi(\Sigma) \stackrel{def}{=} tr\left(\Sigma^{-1}\right) + \log\det\Sigma$ *which encourages a solution towards the identity matrix (see Proposition 3.24). By proposition 3.24 the regularizer is g-convex and DC. Hence*

$$\hat{\phi}(\Sigma) = \frac{d}{n} \sum_{i=1}^{n} \log\left(x_i^T \mathbf{\Sigma}^{-1} x_i\right) + \log\det(\Sigma) + \beta\left(tr\left(\Sigma^{-1}\right) + \log\det\Sigma\right)$$

is g-convex and DC.

Example 3.28 (Normalized Regularized Tyler Estimator)**.** *Suppose we are given* n *independent realizations* $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ *drawn from an unknown zero mean distribution* $f(x)$ *with covariance* Σ*. To estimate* Σ*, one can use the normalized regularized Tyler estimator (see [20, Section 4.2.3]) which is (up to scaling) the minimizer of*

$$R_\Phi(Y) = \frac{d}{n} \sum_{i=1}^{n} \log\left[(1-\alpha) x_i^\top Y^{-1} x_i + \alpha \frac{\|x_i\|^2}{p} \operatorname{Tr}\left\{Y^{-1}\right\}\right] + \log\det Y \ ,$$

where $\alpha \in [0, 1]$ *is a hyperparameter. One can rewrite* $R(Y)$ *as*

$$R_\Phi(Y) = d_\Phi(D, I_d) + d_\Phi(Y, I_d) \ ,$$

where Φ *is the Schatten* 1*-norm and* D *is the diagonal matrix with entries*

$$D_{ii} = tr\left(Y^{-1}\left((1-\alpha) x_i x_i^\top + \alpha \|x_i\|^2 d^{-1} I_d\right)\right) \qquad i = 1, \ldots, d.$$

4 Solving Structured Regularized Problems via CCCP

4.1 Exploiting the structural properties of R_Φ

Recall that in Section 3.2 we discussed that for any $\Phi \in \mathcal{SG}(\mathbb{P}_d)$ we can combine transformations of S_Φ and d_Φ to obtain a g-convex (and possibly DC) regularizer R_Φ. Table 4.1 illustrates the "g-convex+DC" property of the regularized problem $\hat{\phi}$ for varying g-convex and DC properties of ϕ and R_Φ.

For completeness, we will provide a proof of these results.

Proposition 4.1. *We prove the assertions in Table 4.1. Let* $\phi : \mathbb{P}_d \to \mathbb{R}$ *be the objective function,* $R_\Phi(X) : \mathbb{P}_d \to \mathbb{R}_+$ *our generated regularizer for some* $\Phi \in \mathcal{SG}(\mathbb{P}_d)$*, and* $\hat{\phi} : \mathbb{P}_d \to \mathbb{R}$ *to be the resulting regularized objective. The following holds*

1. *If* $\phi(X)$ *is g-convex and* $R_\Phi(X)$ *is g-convex then* $\hat{\phi}(X)$ *is g-convex.*

2. *If $\phi(X)$ is g-convex and e-convex and $R_\Phi(X)$ is g-convex and DC then $\hat{\phi}(X)$ is g-convex and DC.*
3. *If $\phi(X)$ is g-convex and DC and $R_\Phi(X)$ is g-convex and e-convex then $\hat{\phi}(X)$ is g-convex and DC.*
4. *If $\phi(X)$ is g-convex and DC and $R_\Phi(X)$ is g-convex and DC then $\hat{\phi}(X)$ is g-convex and DC.*

Proof. Recall the structure of the regularized objective

$$\hat{\phi}(X) = \phi(X) + \beta R_\Phi(X) \qquad \text{for some } \beta > 0.$$

1. The g-convexity of $\hat{\phi}(X)$ follows from the fact that the sum of two g-convex functions is g-convex.
2. The g-convexity of $\hat{\phi}(X)$ again follows from the fact that the sum of two g-convex functions is g-convex. Write $R_\Phi(X) = f(X) - h(X)$ where $f, h : \mathbb{P}_d \to \mathbb{R}$ are Euclidean convex functions. This implies DC. Moreover, we have

$$\hat{\phi}(X) = (\phi(X) + \beta f(X)) - \beta h(X).$$

   Observe that $\phi(X) + \beta f(X)$ is e-convex and thus $\phi(X)$ is g-convex.
3. Write $\phi(X) = f(X) - h(X)$ where $f, h : \mathbb{P}_d \to \mathbb{R}$ are e-convex. Then

$$\hat{\phi}(X) = (f(X) + \beta R_\Phi(X)) - h(X)$$

   which is g-convex and DC.
4. Write $\phi(X) = f_1(X) - h_1(X)$ and $R_\Phi(X) = f_2(X) - h_2(X)$ where $f_k, h_k : \mathbb{P}_d \to \mathbb{R}$ are e-convex for $k = 1, 2$. Write

$$\hat{\phi}(X) = (f_1(X) + \beta f_2(X)) - (h_1(X) + \beta h_2(X))$$

   which is clearly g-convex and DC.

□

Remark 4.2. Table 4.1 illustrates the interaction between the convexity and DC structure of $\phi(X)$ and $R_\Phi(X)$ and the resulting regularized problem $\hat{\phi}(X)$. In particular, due to the desirable properties of g-convex and DC, Table 4.1 highlights the desirable properties of $R_\Phi(X)$, ranked from greatest to least: g-convex and DC, g-convex and e-convex, and g-convex.

4.2 Complexity analysis

In the previous section we discussed desirable properties that our class of regularizers induces in the reparametrized objective with respect to convexity and difference of convex structure (Tab. 4.1). We will now discuss the algorithmic implications of this induced structure, in particular, how it allows for applying CCCP algorithms in the constrained setting, as well as implications on the iteration and oracle complexities of the resulting optimization routine.

Table 1 Properties of $\hat{\phi}(X) = \phi(X) + \beta R_\Phi(X)$ where R_Φ is a regularizer consisting of (transformations of) S_Φ and $d_\Phi(X, \hat{X})$ for a specified $\Phi \in \mathcal{SG}(\mathbb{P}_d)$. All regularized problems are g-convex but *not* Euclidean convex.

Original		**Regularized**		**Example**
$\phi(X)$	$\mathcal{X}$	$R_\Phi(X)$	$\hat{\phi}(X)$	
g-cvx + DC, ¬ e-cvx	$\mathbb{P}_d$	g-cvx + DC	g-cvx + DC	Tyler's Estimator w/ Diagonal Loading (Ex. 3.27, Sec. 4.2.2 [20]) & Karcher Mean w/ S-Divergence (Ex. 3.26)
g-cvx, ¬ DC, ¬ e-cvx	$\mathbb{P}_d$	g-cvx	g-cvx	Karcher Mean w/ Riemannian metric (Ex. 3.26)
g-cvx+DC, ¬ e-cvx	B_R	g-cvx + DC	g-cvx + DC	Optimistic Likelihood Problem (Sect. 5.2)
g-cvx, ¬ DC, e-cvx	B_R	g-cvx + DC	g-cvx + DC	Linear Regression on $\mathbb{P}_d$ with S-divergence (Sect. 5.2)

4.2.1 Iteration complexity

The *iteration complexity* of an algorithm refers to the number of iterations required to reach an ϵ-accurate solution and hence provides an important lens for non-asymptotic convergence analysis. Classical Riemannian first-order methods [9, 10] for constrained optimization on manifolds converge to the global optimum of problem 1 at a sublinear rate, whereas for Euclidean constrained methods, only sublinear convergence to a stationary point can be guaranteed. Our regularization approach, which reparametrizes problem 1 as an *unconstrained* problem, further allows for leveraging CCCP (Alg. 1). The regularized objective fulfills the conditions of Theorem 2.4, which guarantees an at most sublinear iteration complexity:

Corollary 4.3. *Let $\hat{\phi}$ denote a regularization of problem 1 that is g-convex and DC. Then Alg. 1 converges to a global optimum at a sublinear rate.*

Note that this result provides only an upper bound on the iteration complexity. We will see in the following section that, in some instances, faster, linear convergence can be obtained by leveraging special structure in the regularized objective to solve the CCCP oracle in closed form.

4.2.2 Oracle complexity

However, the iteration complexity alone does not provide a full characterization of the convergence rate. In order to understand the efficiency of an iterative algorithm, we further need to analyze the *cost per iteration*. In the case of constrained Riemannian optimization, this cost is dominated by subroutines that enforce the constraints: In R-PGD [9], a projection onto the feasible region is computed, which can be costly. While R-FW [10] is projection-free, it requires a call to a linear oracle. This subroutine has closed form solutions or reductions to efficient oracles for specific domains and sets of constraints [10, 39], but is in general non-convex, which can present a challenge in practise. Since both approaches are *Riemannian* algorithms, the respective

subroutines, as well as the subsequent computation of the next iterate, require the implementation of Riemannian tools, such as exponential maps, Riemannian gradients and parallel transport operators. In contrast, CCCP requires only the implementation of Euclidean tools, which is often much faster.

Like the Riemannian constrained optimization approaches, each CCCP iteration requires a call to a subroutine, the *CCCP oracle* (see line 4 in Alg. 1). In some cases, it is possible to solve the minimization of the linear surrogate function therein in closed form, which renders the CCCP approach into a simple fixed-point algorithm. We will discuss examples of this form in the next section. Even if such a fixed-point algorithm cannot be derived, the oracle complexity is lower than that of the Riemannian approaches. First, the CCCP oracle is convex and no Riemannian tools are required to solve this subproblem numerically. Second, the structure of our class of regularizers, specifically that of our two canonical examples (ball and sparsity regularizers) provides additional algorithmic benefits, as we discuss below.

4.2.3 Complexity of regularization

Suppose our objective $\phi : \mathbb{P}_d \to \mathbb{R}$ is g-convex and DC, that is, ϕ is g-convex and can be written as $\phi(X) = f_1(X) + h_1(X)$ for convex functions $f_1, h_1 : \mathbb{P}_d \to \mathbb{R}$. Moreover, suppose that $R : \mathbb{P}_d \to \mathbb{R}$ is a g-convex and DC regularizer that can be written as $R(X) = f_2(X) + h_2(X)$. In the case that $R(X)$ is convex then $h_2(X) = 0$. Applying Algorithm 1 to the structured regularized problem

$$\arg\min_{X \in \mathbb{P}_d} \{\phi(X) + \beta R(X) = (f_1(X) + \beta f_2(X)) - (h_1(X) + \beta h_2(X))\}$$

requires solving the convex optimization problem

$$\arg\min_{X \in \mathbb{P}_d} \{Q(X, Y) = (f_1(X) + \beta f_2(X)) - (h_1(Y) + h_2(Y)) - \langle \nabla(h_1 + h_2)(Y), X - Y \rangle\}. \tag{12}$$

In general, one can directly solve (12) with gradient descent, which amounts to computing the updates

$$\begin{aligned} X_{\ell+1} &\leftarrow X_\ell - \eta \nabla_X Q(X_\ell, Y) \\ \text{where} \qquad Q(X_\ell, Y) &= \nabla(f_1 + f_2)(X_\ell) - \nabla(h_1 + h_2)(Y). \end{aligned} \tag{13}$$

This requires computing the gradients ∇f_i and ∇h_i for $i = 1, 2$ in each iteration. The complexity of the subroutine depends crucially on the number of calls to the corresponding gradient oracles. Choosing a regularizer that has efficiently computable gradients could allow for mitigating possible bottlenecks. We will discuss an example in the next section, where the ball constraint regularizer is parametrized via the S-divergence, which comes with numerical benefits for the gradient computation.

We contrast these considerations with the subroutines that handle constraints in R-PGD and R-FW. Both require calls to *Riemannian* gradient oracles. This requires an additional projection, which introduces computational overhead (see sec. 2.1).

Moreover, as discussed above, the subroutines generally require solving a nonconvex problem, which can be much more challenging than the CCCP oracle. An analysis of oracle complexities for certain sets of constraints in projection-free and projection-based methods can be found in [10, 11, 40].

4.2.4 Ball constraint via S-divergence

The computation of the Riemannian metric (Eq. 3) requires computing the generalized eigenvalues of A and B, which introduces a computational bottleneck. To address this problem, [41] introduced a *symmetrized log-det based matrix divergence*, also known as the *S-divergence* (Lem. 3.23). Sra [5] discusses the relationship of the Riemannian metric δ_R and the S-divergence δ_S^2 and its algorithmic implications. We present relevant properties of the S-divergence and its relation to the Riemannian metric δ_R.

Proposition 4.4 (Table 4.1 [5])**.** *Let $A, B, X \in \mathbb{P}_d$. The S-divergence δ_S^2 satisfies the following properties*

1. ***Invariant Under Inversions.*** $\delta_S\left(A^{-1}, B^{-1}\right) = \delta_S(A, B)$
2. ***Invariant Under Conjugation.*** $\delta_S(X^*AX, X^*BX) = \delta_S(A, B)$
3. ***Bi-G-convex.*** $\delta_S^2(X, Y)$ *is g-convex in* X, Y
4. ***Lower Bounded By Shifts.*** $\delta_S^2(A + X, B + X) \leq \delta_S^2(A, B)$.
5. ***Geodesic As S-divergence.*** $A \sharp B = \arg\min_{X \in \mathbb{P}_d} \delta_S^2(X, A) + \delta_S^2(X, B)$

Every property listed in Proposition 4.4 is also satisfied by the Riemannian metric δ_R, see [5, Table 4] for more shared properties of δ_S^2 and δ_R. Moreover, we can relate the size of the metric balls induced by the δ_R and δ_S^2 via the following proposition.

Proposition 4.5 (Theorem 4.19 [5])**.** *Let $A, B \in \mathbb{P}_d$. Then we have $8\delta_S^2(A, B) \leq \delta_R^2(A, B)$.*

Proposition 4.6. *Fix $\hat{\Sigma} \in \mathbb{P}_d$ and fix $\alpha > 0$. Define the sets*

$$\mathcal{B}_R(\hat{\Sigma}; \alpha) \stackrel{def}{=} \{A \in \mathbb{P}_d : \delta_R(A, \hat{\Sigma}) \leq \alpha\}$$

and

$$\mathcal{B}_S(\hat{\Sigma}; \alpha) \stackrel{def}{=} \{A \in \mathbb{P}_d : \delta_S^2(A, \hat{\Sigma}) \leq \alpha\}.$$

Then the subset-inequality

$$\mathcal{B}_R(\hat{\Sigma}; \alpha) \subseteq \mathcal{B}_S(\hat{\Sigma}; C\alpha)$$

holds for $C \geq \frac{\alpha}{8}$.

Proof. By Proposition 4.5, we have the inequality

$$2\sqrt{2}\delta_S(A, B) \leq \delta_R(A, B) \qquad \forall A, B \in \mathbb{P}_d.$$

Let $\alpha > 0$ and suppose $A \in \mathcal{B}_R(\hat{\Sigma}; \alpha)$. By definition and applying the inequality above, we have

$$\begin{aligned} \delta_R(A, \hat{\Sigma}) \leq \alpha &\implies 2\sqrt{2}\delta_S(A, \hat{\Sigma}) \leq \alpha \\ &\implies \delta_S^2(A, \hat{\Sigma}) \leq \frac{1}{8}\alpha^2 \,. \end{aligned}$$

Hence $A \in \mathcal{B}_S(\hat{\Sigma}; C\alpha)$ for any $C \geq \frac{\alpha}{8}$. Since $A \in \mathcal{B}_R(\hat{\Sigma}; \alpha)$ was arbitrarily selected we have

$$\mathcal{B}_R(\hat{\Sigma}; \alpha) \subseteq \mathcal{B}_S(\hat{\Sigma}; C\alpha) \qquad \forall C \geq \frac{\alpha}{8}.$$

□

This suggests that the S-divergence can be leveraged for an efficient relaxation of the ball constraint regularizer: Suppose we have an optimization problem constrained to lie within a Riemannian distance ball $\mathcal{B}_R(\cdot; \alpha)$ of radius $\alpha > 0$. Since the S-divergence ball $\mathcal{B}_S(\cdot; C\alpha)$ is a superset of the Riemannian distance ball, we can replace the Riemannian distance ball with the S-divergence ball with radius $C\alpha$ for some $C \geq \alpha/8$. This alludes to a more general relaxation technique which we discuss now.

Computational considerations

Computing the S-divergence $\delta_S^2(A, B)$ requires 3 Cholesky factorizations for $A + B, A$ and B, whereas computing the Riemannian metric requires computing generalized eigenvalues at a cost of $4d^3$ flops for positive definite matrices. The cost gap between δ_S^2 and δ_R only widens when considering their gradients

$$\begin{aligned}\nabla_A \delta_R^2(A, B) &= A^{-1} \log\left(AB^{-1}\right) \\ \nabla_A \delta_S^2(A, B) &= (A + B)^{-1} - \frac{1}{2}A^{-1}.\end{aligned}$$

This is particularly well-illustrated in Table 2 [41]. Hence, in many of our applications we will replace δ_R with the computationally advantageous δ_S^2.

Relaxing the Riemannian Ball Constraint Problem

Let $\phi : \mathbb{P}_d \to \mathbb{R}$ be g-convex and DC. We would like to solve the following problem

$$\begin{aligned} X^* &\stackrel{\text{def}}{=} \operatorname*{arg\,min}_{X \in \mathbb{P}_d} \phi(X) \qquad \text{subject to} \qquad \mathcal{B}_R(\hat{X}, r) \\ \text{where} \qquad \mathcal{B}_R(\hat{X}, r) &\stackrel{\text{def}}{=} \{X \in \mathbb{P}_d : \delta_R(X, \hat{X}) \leq r\}. \end{aligned} \tag{14}$$

We use the notation $\hat{X} \in \mathbb{P}_d$ to denote additional information as to where the true solution X^* may lie and the radius $r > 0$ corresponds to the confidence of such information. In light of Proposition 4.6 we can relax Problem 15 by replacing the Riemannian ball $\mathcal{B}_R(\hat{X}, r)$ with the S-divergence ball $\mathcal{B}_S(\hat{X}, r/8)$. Then we can design the corresponding relaxed structured regularization problem as

$$X^* \stackrel{\text{def}}{=} \operatorname*{arg\,min}_{X \in \mathbb{P}_d} \phi(X) + \beta \delta_S^2\left(X, \hat{X}\right) \tag{15}$$

for $\beta > 0$. We remark that Problem 15 can be naturally applied to maximum likelihood problems [21]. In the experiments section, we will illustrate this relaxation procedure on the example of optimistic likelihood estimation.

5 Applications

In this section we illustrate our structured regularization framework on the square root problem, the Karcher mean problem, and the optimistic likelihood problem. In particular, we present Euclidean algorithms to solve for *global* solutions of Euclidean nonconvex (but g-convex) optimization problems that have traditionally been viewed through a Riemannian lens.

5.1 Square Root and Karcher Mean: Fixed-Point Solvers

We have previously discussed structured relaxations of the Karcher mean (Example 3.26) and the square root (Example 3.25) problems using the S-divergence. Since both objectives are g-convex and DC, applying the CCCP algorithm will provably converge to the optimal solution (Theorem 2.4): G-convexity ensures that every local optimum is a global one and the CCCP algorithm will always converge to one such global optimum (see section 2.2 for details).

Since the CCCP algorithm is a Euclidean algorithm, it circumvents the computational overhead introduced by Riemannian optimization tools. Moreover, the CCCP approach gives rise to fast fixed-point algorithms, whenever the CCCP oracle can be solved in closed form. The following propositions establish such fixed-point approaches for the square root problem and the Karcher mean problem. Recall that the square root of a SPD matrix is given by the solution to the optimization problem

$$\min_{X \in \mathbb{P}_d} \phi(X) \stackrel{\text{def}}{=} \delta_S^2(X, A) + \delta_S^2(X, I) \;, \tag{16}$$

where the objective is composed of symmetric gauge functions. This formulation allows for the following effective CCCP approach:

Proposition 5.1 (Fixed-point approach for Square Root [6])**.** *Applying the CCCP algorithm to the problem of computing the square root of an SPD matrix is equivalent to iterating the fixed point map*

$$X \leftarrow \left[(X + A)^{-1} + (X + I)^{-1}\right]^{-1} \;. \tag{17}$$

Recall that the Karcher mean problem [32, 42] is the solution to the following problem: Given data $\{A_1, \ldots, A_m\} \in \mathbb{P}_d$ and $w \in \mathbb{R}_+^m$ such that $\sum_{i=1}^m w_i = 1$ we solve

$$\min_{X \in \mathbb{P}_d} \stackrel{\text{def}}{=} \sum_{i=1}^m w_i \delta_R^2(X, A_i) \;,$$

which can be rewritten as [5]

$$\min_{X \in \mathbb{P}_d} \phi(X) \stackrel{\text{def}}{=} \sum_{i=1}^m w_i \delta_S^2(X, A_i) \;. \tag{18}$$

Then the following effective fixed-point approach can be derived:

Proposition 5.2 (Fixed-point approach for Karcher Mean [12])**.** *Applying the CCCP algorithm to the Karcher mean problem is equivalent to the fixed-point algorithm*

$$X \leftarrow \left[\sum_{i=1}^{m} w_i \left(\frac{X + A_i}{2}\right)^{-1}\right]^{-1} \qquad k = 0, 1, \ldots \ . \tag{19}$$

5.2 The Optimistic Likelihood Problem

Next we consider the problem of computing optimistic likelihoods. Let $\phi(x)$ denote the negative log-likelihood and suppose we want to incorporate additional side information on the solution X^*, such as an estimator $\hat{X} \in \mathbb{P}_d$. One natural formulation is the following constrained optimization problem:

$$\begin{aligned} X^* &\overset{\text{def}}{=} \operatorname*{arg\,min}_{X \in \mathbb{P}_d} \phi(X) \qquad \text{subject to} \qquad \mathcal{B}_R(\hat{X}, r) \\ \text{where} \qquad \mathcal{B}_R(\hat{X}, r) &\overset{\text{def}}{=} \{X \in \mathbb{P}_d : \delta_R(X, \hat{X}) \leq r\} \ . \end{aligned} \tag{20}$$

Note that $\phi(X)$ is g-convex and DC. The radius $r > 0$ can be interpreted a confidence interval for $\hat{X}$ with smaller values corresponding to higher confidence. Incorporating such information via ball constraints can be advantageous in statistical problems, as we illustrate below.

The optimistic likelihood problem was first introduced by [21] for the special case where ϕ is the *Gaussian* negative log-likelihood. In this section, we will develop an unconstrained formulation of (20) via structured regularization that can be solved efficiently using CCCP. In the subsequent sections we will show that our framework can extend beyond the Gaussian case to the class of Kotz distributions [43].

5.2.1 Optimistic Gaussian Likelihood

The *Optimistic Gaussian Likelihood problem* [21] concerns the following setting: Consider a set of i.i.d data points $x = (x_1, \ldots, x_n) \in \mathbb{R}^d$ generated from precisely one of several Gaussian distributions $\{\mathcal{N}(0, \Sigma_c)\}_{c \in \mathcal{C}}$ with zero mean and covariance Σ_c indexed by $c \in \mathcal{C}$ where $|C| < \infty$. Since we can parametrize the family of zero mean Gaussian distributions with $\mathbb{P}_d$, i.e., $\mathcal{N}(0, \Sigma) \simeq \Sigma$, we denote the set of candidate distributions $\{\mathcal{N}(0, \Sigma_c)\}_{c \in \mathcal{C}}$ simply as $\{\Sigma_c\}_{c \in \mathcal{C}}$. The goal is to determine the true Gaussian distribution by solving the following maximum likelihood problem over the class $\mathcal{C}$:

$$c^\star \in \operatorname*{arg\,min}_{c \in \mathcal{C}} \left\{ \phi\left(\Sigma_c; x\right) \overset{\text{def}}{=} \frac{1}{n} \sum_{k=1}^{n} x_k^\top \Sigma_c^{-1} x_k + \log\det \Sigma_c \right\} \ . \tag{21}$$

In general, Σ_c is unknown, but we can obtain an estimator $\hat{\Sigma}_c$ from the data. Moreover, Problem (21) is highly sensitive to misspecification of the candidate distributions $\mathcal{N}(0, \Sigma_c)$. [21] addresses this by solving the following *constrained g-convex* optimization

problem for $c \in \mathcal{C}$:

$$\min_{\Sigma \in \mathcal{B}_R(\hat{\Sigma}_c; \rho_c)} \phi(\Sigma; x) \qquad \text{where} \qquad \mathcal{B}_R(\hat{\Sigma}_c; \rho_c) \stackrel{\text{def}}{=} \left\{ \Sigma \in \mathbb{P}_d : \delta_R\left(\Sigma, \hat{\Sigma}_c\right) \leq \rho_c \right\}. \tag{22}$$

Problem (21) has applications in machine learning (e.g. quadratic discriminant analysis [44]) and in statistics (e.g. Bayesian inference [45]). Problem (22) is itself a natural formulation of maximum likelihood estimation with side information. That is, we want to compute a Σ^* that maximizes the likelihood of observing the data $x_1, \ldots, x_n$ while staying close to $\hat{\Sigma}$ in order to leverage prior information.

The Riemannian metric on $\mathbb{P}_d$ presents a natural notion of distance on densities that are parametrized by their covariance matrices. Among other desirable properties, it satisfies desirable statistical properties such as invariances under congruences and affine transformations (see Proposition 4.4 or Table 4.1 in [5])

$$\begin{aligned} \delta_R\left(\Sigma^{-1}, \hat{\Sigma}^{-1}\right) &= \delta_R\left(\Sigma, \hat{\Sigma}\right) \\ \delta_R\left(A\Sigma A^\top, A\hat{\Sigma}A^\top\right) &= \delta_R\left(\Sigma, \hat{\Sigma}\right), \end{aligned} \tag{23}$$

for all invertible $A \in \mathbb{R}^{d \times d}$ and $\hat{\Sigma}, \Sigma \in \mathbb{P}_d$ [46].

In [21], the authors propose to solve Problem (22) via R-PGD. However, their approach requires expensive computational operations such as inverse fractional powers of matrices arising from the projection onto $\mathcal{B}_R \subseteq \mathbb{P}_d$. We present a relaxation of Problem (22) that can be efficiently solved using the CCCP algorithm, circumventing the need for such expensive computational machinery.

Per the discussion of the computational advantages of the S-divergence δ_S^2 (see Section 4.2.4) and the fact that it satisfies the desirable statistical properties of (23) (see Proposition 4.4) we can leverage Proposition 4.6 to relax Problem (22) as follows:

$$\arg\min_{\Sigma \in \mathbb{P}_d} \left\{ \phi(\Sigma) \stackrel{\text{def}}{=} \operatorname{tr}\left(S\Sigma^{-1}\right) + \log\det\Sigma \right\} \qquad \text{subject to} \quad \Sigma \in \mathcal{B}_S\left(\hat{\Sigma}; r\right), \tag{24}$$

where $S = \frac{1}{N}\sum_{m=1}^{N}\left(x_m - \hat{\mu}\right)\left(x_m - \hat{\mu}\right)^\top$ denotes the sample covariance matrix and $\mathcal{B}_S\left(\hat{\Sigma}; \rho\right) = \{X \in \mathbb{P}_d : \delta_S^2(X, \hat{\Sigma}) \leq \rho\}$ is the *S-divergence ball* with radius $\rho > 0$ centered at $\hat{\Sigma}$. Finally, performing structured regularization, we can rewrite the problem as

$$\arg\min_{\Sigma \in \mathbb{P}_d} \left\{ \hat{\phi}(\Sigma) \stackrel{\text{def}}{=} \operatorname{tr}\left(S\Sigma^{-1}\right) + \log\det\Sigma + \beta\delta_S^2\left(\Sigma, \hat{\Sigma}\right) \right\} \tag{25}$$

for some hyperparameter $\beta > 0$. The objective $\hat{\phi}(\Sigma)$ is (strictly) g-convex and hence has a unique global minimum (if it exists). Solving this problem via CCCP yields Algorithm 2. We argue that this approach can be used to find the global optimum.

Corollary 5.3. *Initialize $\Sigma_0 \in \mathbb{P}_d$ and denote $\{X_{\ell+1}\}$ as the iterates generated by Algorithm 2. For a suitable choice of step sizes $\{\eta_\ell\}$, the sequence $X_{\ell+1}$ converges to the unique solution Σ^* of Problem (25).*

Proof. Observe that $\hat{\phi}(0) = \phi(X)$ is undefined and $\phi(X)$ is continuous on $\mathbb{P}_d$. This implies the optimum is attained in $\mathbb{P}_d$. Since $\hat{\phi}(\Sigma)$ is of g-convex and DC structure, the CCCP algorithm converges to the global minima [12]. □

Algorithm 2 CCCP for Optimistic Gaussian Likelihood

Input: $\Sigma_0, \hat{\Sigma} \in \mathbb{P}_d$, $K, L \in \mathbb{N}$, $\beta > 0$ and $\{\eta_\ell\} \subseteq \mathbb{R}_{++}$
for $k = 0, \ldots, K-1$ **do**
  Precompute $\Sigma_k^{-1} + \beta \left(\Sigma_k + \hat{\Sigma}\right)^{-1}$
  for $\ell = 0, \ldots, L-1$ **do**
    $\Sigma_{\ell+1} \leftarrow \Sigma_\ell - \eta_\ell \left(-\Sigma_\ell^{-1} S \Sigma_\ell^{-1} - \frac{\beta}{2}\Sigma_\ell^{-1} + \Sigma_k^{-1} + \beta \left(\Sigma_k + \hat{\Sigma}\right)^{-1}\right)$
    Update $\Sigma_{k+1} \leftarrow \Sigma_L$
Output: Σ_K

5.2.2 Generalizing beyond the Gaussian setting

In the previous section, we formulated Problem (25) to find the maximum likelihood estimator of a *Gaussian* density in the presence of side information. This formulation can be readily generalized to Kotz-type distributions [43, 47] and the multivariate log-normal distribution.

5.2.3 Kotz-type distributions

We generalize the optimistic likelihood problem to a special class of *elliptically contoured distributions*, also known as *Kotz-type distributions* [43, 47]. For simplicity, we consider the case of zero-mean Kotz-type distributions.

If the density of a (zero-mean) Kotz-type distribution exists on $\mathbb{R}^d$ then it is of the form

$$f(x; \Sigma) \propto \det(\Sigma)^{-1/2} g\big(x^\top \Sigma^{-1} x\big) \,,$$

where $\Sigma \in \mathbb{P}_d$ and $g : \mathbb{R} \to \mathbb{R}_{++}$ is known as the *density generator*.

Example 5.4. *The density generator $g(t) = \exp(-t/2)$ recovers the multivariate Gaussian distribution.*

In this section, we focus on the Kotz-type distribution that arise from specifying the density generator to be of the form

$$g(t) = t^{\alpha - d/2} \exp\left(- \left(\frac{t}{b}\right)^\beta \right)$$

for distribution parameters $\alpha \in (0, d/2], \beta, b > 0$. The Kotz-type distributions with these parameters encompass many well-known instances, including Gaussian and multivariate power exponential distributions, the multivariate-W distribution with shape

parameter less than one, as well as the elliptical gamma distribution with shape parameter less than $d/2$, among others (see, e.g., sec. 5 in [47] or [43] for more examples).

Given $x_1, \ldots, x_n$ i.i.d observations sampled from a Kotz-type distribution, its negative log-likelihood assumes the form[1]

$$K(\Sigma; x_1, \ldots, x_n) = \frac{n}{2} \log \det(\Sigma) + \left(\frac{d}{2} - \alpha\right) \sum_{i=1}^{n} \log\left(x_i^T \Sigma^{-1} x_i\right) + \sum_{i=1}^{n} \left(\frac{x_i^T \Sigma^{-1} x_i}{b}\right)^{\beta}. \tag{26}$$

In this section, we show that $K(\Sigma)$ given in (26) satisfies the g-convex and DC property for $\alpha \in (0, d/2]$ and for $b, \beta > 0$. Moreover, we can find a global optimum (if it exists) using CCCP. The next two results provide sufficient conditions such that the maximum likelihood estimator (MLE) of (26) exists.

Proposition 5.5 (Lemma 44 [47]). *Let the data $\mathcal{X} = \{x_1, \ldots, x_n\}$ span the whole space and for $\alpha < \frac{d}{2}$ satisfy*

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

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

The sufficient condition given in Proposition 5.5 can be satisfied in the under-parametrized noisy setting. That is, the number of data points satisfies $n \leq d$ and they are perturbed by noise.

Theorem 5.6 (Existence of MLE, Theorem 45 [47]). *If the data samples satisfy Proposition 5.5, then the negative log-likelihood of a Kotz-type distribution (26) has a minimizer (i.e., there exists an MLE).*

The main result in the section is the following proposition.

Proposition 5.7. *Let $\alpha \leq d/2$ and $b, \beta > 0$ be fixed distribution parameters for a Kotz-type distribution. Then its negative log-likelihood given by (26) is g-convex and DC.*

We remark that Proposition 5.7 was first established in [47]. However, our proof of Proposition 5.7 places an emphasis on establishing the g-convex and DC structure of the Kotz distribution via a DGCP analysis. We believe that such an approach will be fruitful in establishing g-convexity and DC structure for other objective functions.

The following lemmas will help us establish Proposition 5.7.

Lemma 5.8. *Let $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d \setminus \{0\}$ be fixed nonzero vectors. For $0 < \alpha \leq d/2$ the function $f : \mathbb{P}_d \to \mathbb{R}_{++}$ defined by*

$$f(\Sigma) \stackrel{def}{=} \left(\frac{d}{2} - \alpha\right) \sum_{i=1}^{n} \log\left(x_i^T \Sigma^{-1} x_i\right)$$

is Euclidean convex and (strictly) g-convex.

[1] We ignore any constants in the log-likelihood.

Proof. To establish g-convexity, we observe that the function $g_i(\Sigma) = \log(x_i^T \Sigma^{-1} x_i)$ is a strictly g-convex atom (see item 7 in Example 3.17). Hence

$$f(\Sigma) = \Big(\frac{d}{2} - \alpha\Big) \sum_{i=1}^{n} g_i(\Sigma)$$

is strictly g-convex since conic combinations preserve (strict) g-convexity.

To establish Euclidean convexity of the component functions we express $\sum_{i=1}^{n} \log(x_i^T A^{-1} x_i)$ as a log determinant of a product of matrices. Define a block diagonal matrix B such that

$$B = \operatorname{diag}(A^{-1}, A^{-1}, \ldots, A^{-1}) \,,$$

where each of the n blocks consists of the matrix A^{-1}. Thus, B is a $dn \times dn$ matrix. Now define a $dn \times n$ matrix Y by stacking the vectors x_i in a specific block form:

$$Y = \begin{bmatrix} x_1 & 0 & 0 & \cdots & 0 \\ 0 & x_2 & 0 & \cdots & 0 \\ 0 & 0 & x_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & x_n \end{bmatrix}$$

Here, each x_i is a d-dimensional column vector, and there are n such vectors. Consider the matrix product $Y^T B Y$. This is an $n \times n$ matrix where the ith diagonal element of $Y^T B Y$ is given by:

$$(Y^T B Y)_{ii} = x_i^T A^{-1} x_i$$

and the off-diagonal elements are zero because of the block structure of Y and B. Therefore, $Y^T B Y = \mathbf{diag}(x_1^T A^{-1} x_1, x_2^T A^{-1} x_2, \ldots, x_n^T A^{-1} x_n)$. Taking the logarithm of both sides, we get:

$$\log(\det(Y^T B Y)) = \log\Big(\prod_{i=1}^{n} (x_i^T A^{-1} x_i)\Big) \,.$$

Standard logarithm laws imply $\log(\det(Y^T BY)) = \sum_{i=1}^n \log(x_i^T A^{-1} x_i)$. We can prove the Euclidean convexity of the right hand side as follows:

$$\begin{aligned}
\sum_{i=1}^n \log(x_i^T A^{-1} x_i) &= \log\left(\det(Y^T BY)\right) \\
&= \log\left(\det(B)\det(YY^\top\right) \\
&= \log\det(B) + \log\det Y^\top Y \\
&= \log\det\left(\mathbf{diag}(A^{-1},\ldots,A^{-1})\right) + \log\det\left(\mathbf{diag}(\|x_1\|^2,\ldots,\|x_n\|^2)\right) \\
&= -\log\det\left(\mathbf{diag}(A,\ldots,A)\right) + C(x_1,\ldots,x_n) \\
&= -\log\left(\det(A)\right)^n + C(x_1,\ldots,x_n) \\
&= -n\log\det(A) + C(x_1,\ldots,x_n).
\end{aligned}$$

Since $A \mapsto -\log\det(A)$ is Euclidean convex on $\mathbb{P}_d$ then it follows $f(A)$ is convex. □

Lemma 5.9. *Let $\{x_1,\ldots x_n\} \subseteq \mathbb{R}^d \setminus \{0\}$ be nonzero vectors and fix $\beta > 0$. The function $g : \mathbb{P}_d \to \mathbb{R}_{++}$ defined by*

$$g(\Sigma) = \sum_{i=1}^n \left(\frac{x_i^\top \Sigma^{-1} x_i}{b}\right)^\beta$$

is Euclidean convex and g-convex.

Proof. Without loss of generality we assume $b = 1$ (since $b > 0$ it can be absorbed by the vectors x_i). Define the functions $h_i(\Sigma) := x_i^\top \Sigma x_i$ and $g(t) := t^\beta$. Observe that $h_i(\Sigma)$ is a g-convex atom (see Example 3.17(6)) for $1 \le i \le n$, and in particular it is a positive linear map. Also $g(t)$ is g-convex and non-decreasing on the positive reals since for all $a, b > 0$ and we have

$$g\left(a \sharp_\theta b\right) = g\left(a^{1-\theta} b^\theta\right) = \left(a^\beta\right)^{1-\theta}\left(b^\beta\right)^\theta \le (1-\theta)a^\beta + \theta b^\beta \quad \forall \theta \in [0,1]\,,$$

which follows from the (weighted) AM-GM inequality. Hence by Proposition 3.14 the function

$$g \circ h_i(\Sigma) = \left(x_i^\top \Sigma x_i\right)^\beta$$

is g-convex. By applying Proposition 3.16 the function

$$\psi_i(\Sigma) = g \circ h_i(\Sigma^{-1}) = \left(x_i^\top \Sigma^{-1} x_i\right)^\beta$$

is g-convex. To establish Euclidean convexity, observe that

$$\begin{aligned}
g_i(\Sigma) = \left(x_i^\top \Sigma^{-1} x_i\right)^\beta &= \exp\left(\log\left(x_i^\top \Sigma^{-1} x_i\right)^\beta\right) \\
&= \exp\left(\beta \log\left(x_i^\top \Sigma^{-1} x_i\right)\right) = \exp(\beta \ell_i(\Sigma))\,,
\end{aligned}$$

where $\ell_i(\Sigma) = \log\left(x_i^\top \Sigma^{-1} x_i\right)$ is Euclidean convex by Lemma 5.8. Since the exponential function $t \mapsto \exp(\beta t)$ is monotonically increasing and Euclidean convex on $\mathbb{R}_{++}$ we have that $g_i(\Sigma) = \exp(\beta \ell_i(\Sigma))$ is convex for each $1 \leq 1 \leq n$ which follows by standard convex compositional rules. Thus we conclude that

$$g(\Sigma) = \sum_{i=1}^{n} \exp(\beta \ell_i(\Sigma)) = \sum_{i=1}^{n} \left(\frac{x_i^T \Sigma^{-1} x_i}{b}\right)^{\beta}$$

is Euclidean convex. □

We can now show Proposition 5.7.

Proof. The g-convexity of $K(\Sigma)$ follows from the fact that $\Sigma \mapsto \log\det\Sigma$ is g-linear and applying Lemmas 5.8,5.9. The DC structure of $K(\Sigma)$ follows from the fact that $\Sigma \mapsto \log\det(\Sigma)$ is concave. The second summand is convex by Lemma 5.8. The third summand is convex by Lemma 5.9. Hence the Kotz-type likelihood is a difference of convex functions. □

We remark that Section 5.4 of [47] provides a fixed-point algorithm to efficiently solve the problem $\arg\min_{\Sigma \in \mathbb{P}_d} K(\Sigma)$. However, we are interested in solving the corresponding optimistic likelihood problem

$$\arg\min_{\Sigma \in \mathbb{P}_d} K(\Sigma) + \gamma \delta_S^2(\Sigma, \hat{\Sigma}) \tag{27}$$

for some fixed $\hat{\Sigma} \in \mathbb{P}_d$ and $\gamma > 0$. We apply the CCCP algorithm, where we solve the surrogate minimization problem (CCCP oracle) via gradient descent (see Algorithm 3). Note that Algorithm 3 is structurally similar to Algorithm 2 where we replaced the sample covariance $S = \frac{1}{n} X X^\top$ with the matrix $X D_\ell X^\top$. Since the objective (27) is g-convex and DC, Algorithm 3 converges to the global optimum whenever it exists.

Algorithm 3 CCCP on Optimistic Kotz Likelihood

Input: $\Sigma_0, \hat{\Sigma} \in \mathbb{P}_d$, $K, L \in \mathbb{N}$, $\alpha \in (0, d/2)$, $\beta, \gamma > 0$ and $\{\eta_\ell\} \subseteq \mathbb{R}_{++}$

Set $D_\ell \overset{\text{def}}{=} \mathbf{diag}\left(\frac{\alpha - \frac{d}{2}}{x_i^\top \Sigma_\ell^{-1} x_i} - \frac{\beta}{b^\beta \left(x_i^\top \Sigma_\ell^{-1} x_i\right)^{\beta - 1}} : 1 \leq i \leq d\right)$.

for $k = 0, \ldots, K-1$ **do**

  Precompute $\frac{n}{2}\Sigma_k^{-1} + \gamma\left(\Sigma_k + \hat{\Sigma}_k\right)^{-1}$

  for $\ell = 0, \ldots, L-1$ **do**

    $\Sigma_{\ell+1} \leftarrow \Sigma_\ell - \eta_\ell \left(\Sigma_\ell^{-1}\left(X D_\ell X^\top\right)\Sigma_\ell^{-1} - \frac{\gamma}{2}\Sigma_\ell^{-1} + \frac{n}{2}\Sigma_k^{-1} + \gamma\left(\Sigma_k + \hat{\Sigma_k}\right)^{-1}\right)$

    Update $\Sigma_{k+1} \leftarrow \Sigma_L$.

Output: Σ_K

Algorithm 4 CCCP on Optimistic Multivariate T-Likelihood

Input: $\Sigma_0, \hat{\Sigma} \in \mathbb{P}_d$, $K, L \in \mathbb{N}$, $\gamma, \nu > 0$ and $\{\eta_\ell\} \subseteq \mathbb{R}_{++}$
for $k = 0, \ldots, K-1$ **do**
  Precompute $\gamma \left(\Sigma_k + \hat{\Sigma}_k\right)^{-1} + \frac{n}{2}\Sigma_k^{-1}$.
  for $\ell = 0, \ldots, L-1$ **do**
    $\Sigma_{\ell+1} \leftarrow \Sigma_\ell - \eta_\ell \Big(-\frac{\nu+d}{2}\Sigma_\ell^{-1} \big(\sum_{i=1}^n \frac{x_i x_i^\top}{\nu + x_i^\top \Sigma_\ell^{-1} x_i} \big) \Sigma_\ell^{-1} - \frac{\gamma}{2}\Sigma_\ell^{-1} + \gamma \big(\Sigma_k + \hat{\Sigma}_k\big)^{-1} + \frac{n}{2}\Sigma_k^{-1} \Big)$
    Update $\Sigma_{k+1} \leftarrow \Sigma_L$
Output: Σ_K

5.2.4 Multivariate T-Distribution

In this section, we illustrate how one can apply Algorithm 3 to a specific Kotz-type distribution. We focus on the d-dimensional mean zero multivariate t-distribution with $\nu > 0$ degrees of freedom, parameterized by the scatter matrix $\Sigma \in \mathbb{P}_d$. We denote this distribution as $\mathrm{MVT}(\Sigma; \nu)$. It has the density

$$f_X(x; \nu) = \frac{\Gamma[(\nu+d)/2]}{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2}|\Sigma|^{1/2}} \left[1 + \frac{1}{\nu}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]^{-(\nu+d)/2}.$$

Its negative log-likelihood is proportional to $T_\nu(\Sigma)$ given by

$$T_\nu(\Sigma) = \frac{n}{2}\log\det\Sigma + \frac{\nu+d}{2}\sum_{i=1}^n \log\left(1 + \frac{1}{\nu}x_i^\top \Sigma^{-1} x_i\right) + \gamma \log\det\left(\frac{\Sigma + \hat{\Sigma}}{2}\right) - \frac{\gamma}{2}\log\det\left(\Sigma\hat{\Sigma}\right). \tag{28}$$

The multivariate t-distribution is a generalization of several well-known distributions: We recover the Student's t-distribution by taking $d = 1$ and $\Sigma = 1$. Taking $\nu \to +\infty$ recovers the multivariate normal density with mean 0 and covariance Σ. With $\nu = 1$ we recover the d-variate Cauchy distribution. Taking $\frac{\nu+d}{2}$ to be an integer recovers the d-variate Pearson type VII distribution [48].

Due to its ability to capture tail events (i.e., data involving errors with heavier than Gaussian tails), the multivariate t-distribution is often used for robust statistical modelling, especially when the assumptions of normality is violated [49, 50].

The corresponding optimistic likelihood problem with side information $\hat{\Sigma} \in \mathbb{P}_d$ can be written as

$$\Sigma^* = \operatorname*{arg\,min}_{\Sigma \in \mathbb{P}_d} \left\{ \hat{\phi}(\Sigma) \stackrel{\text{def}}{=} T_\nu(\Sigma) + \gamma\delta_S^2\left(\Sigma, \hat{\Sigma}\right) \right\}, \tag{29}$$

where $\gamma > 0$ is a hyperparameter that denotes our confidence in $\hat{\Sigma}$. We solve problem (29) using again a CCCP approach. To adapt Algorithm 3 to this new setting, we can group the convex and concave terms of $\hat{\phi}(\Sigma)$ into the functions $f(\Sigma)$ and $g(\Sigma)$,

respectively:

$$f(\Sigma) = \frac{\nu + d}{2} \sum_{i=1}^{n} \log\left(1 + \frac{1}{\nu} x_i^\top \Sigma^{-1} x_i\right) - \frac{\gamma}{2} \log\det\left(\Sigma\hat{\Sigma}\right)$$
$$g(\Sigma) = \frac{n}{2} \log\det\Sigma + \gamma \log\det\left(\frac{\Sigma + \hat{\Sigma}}{2}\right)$$

Their gradients are computed as

$$\nabla f(\Sigma) = -\frac{\nu + d}{2}\Sigma^{-1}\left(\sum_{i=1}^{n} \frac{x_i x_i^\top}{\nu + x_i^\top \Sigma^{-1} x_i}\right)\Sigma^{-1} - \frac{\gamma}{2}\Sigma^{-1}$$
$$\nabla g(\Sigma) = \gamma\left(\Sigma + \hat{\Sigma}\right)^{-1} + \frac{n}{2}\Sigma^{-1}$$

The convex CCCP surrogate function can be written as

$$Q(\Sigma, \Sigma_k) = f(\Sigma) + \langle g(\Sigma_k), \Sigma - \Sigma_k \rangle$$

for convex f and concave g. Thus to solve for $\arg\min_{\Sigma \in \mathbb{P}_d} Q(\Sigma, \Sigma_k)$ we can apply the gradient descent steps

$$\Sigma_{\ell+1} \leftarrow \Sigma_\ell - \eta_\ell \left(\nabla f(\Sigma_\ell) + \nabla g\left(\Sigma_k\right)\right) \ ,$$

for a sequence of step-sizes $\{\eta_\ell\}$. The resulting approach is shown schematically in Algorithm 4.

5.2.5 Complexity Analysis

We analyze the complexity of the discussed CCCP approaches on the example of Algorithm 4. Each outer loop requires computing the term $\gamma\left(\Sigma_k + \hat{\Sigma}_k\right)^{-1} + \frac{n}{2}\Sigma_k^{-1}$ once. This requires two matrix inversions and one matrix addition. Each inner loop requires computing $\nabla f(\Sigma_\ell) + \nabla g(\Sigma_k)$, which can be computed as follows. First, construct a matrix $X = (x_1, \ldots, x_n) \in \mathbb{R}^{d \times n}$. Second, compute Σ_ℓ^{-1}. Third, compute the vector $(\nu + x_i^\top \Sigma^{-1} x_i : i \in [n])$ by first computing the quantity

$$X \odot \left(\Sigma_\ell^{-1} X\right)$$

and summing along its last dimension. Call the resulting vector y. This requires one matrix multiplication between $d \times d$ and $d \times n$ matrices, one Hadamard product, and n^2 scalar additions. For simplicity we can construct the diagonal matrix D_ℓ from the following vector

$$D_\ell \stackrel{\text{def}}{=} \mathbf{diag}\left(\nu + x_i^\top \Sigma^{-1} x_i : i \in [n]\right) = \mathbf{diag}\left(\left(\nu \mathbf{1} + y\right)^{\odot -1}\right) \ ,$$

where $\odot - 1$ denotes element-wise inversion. Finally we obtain

$$\sum_{i=1}^{n} \frac{x_i x_i^\top}{\nu + x_i^\top \Sigma_\ell^{-1} x_i} = X D_\ell X^\top \ ,$$

which requires one matrix multiplication between two $n \times d$ matrices. Then the gradient step requires two additional $d \times d$ matrix products and three matrix-matrix additions. Hence, each inner loop requires at most one matrix inversion and four matrix-matrix multiplications.

5.2.6 Log Elliptically Contoured Distributions.

A *log elliptically contoured distributions* is a probability distribution whose logarithm lies in the class of elliptically contoured distributions. It turns out some log elliptically contoured distributions also satisfy the g-convex and DC structure. One example is the mean zero multivariate log-normal distribution $\mathrm{LogMVN}(0, \Sigma)$. Given i.i.d observations $x_1, \ldots, x_n \in \mathbb{R}^d_{++}$ from $\mathrm{LogMVN}(0, \Sigma)$ its negative log-likelihood $\psi(\Sigma)$ is proportional to

$$\psi(\Sigma) \propto \frac{n}{2} \log\det \Sigma + \frac{1}{2} \sum_{j=1}^{m} \log(x_i)^\top \Sigma^{-1} \log(x_i) \tag{30}$$

where we define the operation $\log x_i \in \mathbb{R}^d$ to denote elementwise logarithm.

Observe that the g-convex and DC structure of (30) follows from the fact that $f(\Sigma) = \log\det \Sigma$ is Euclidean concave and g-linear and that the matrix fractional function $g_i(\Sigma) = \log(x_i)^\top \Sigma^{-1} \log(x_i)$ is Euclidean convex for each $1 \leq i \leq d$. The g-convexity of $g_i(\Sigma)$ follows from it being a g-convex atom (see Example 3.17(6)).

Hence one can also derive an algorithm like Algorithm 2 and Algorithm 3 to solve the *log-normal optimistic likelihood* optimization problem

$$\begin{aligned} &\arg\min_{\Sigma \in \mathbb{P}_d} \Psi(\Sigma) + \beta \delta_S^2(\Sigma, \hat{\Sigma}) \\ &\text{where} \quad \Psi(\Sigma) \stackrel{\text{def}}{=} \frac{n}{2} \log\det \Sigma + \frac{1}{2} \sum_{j=1}^{m} \log(x_i)^\top \Sigma^{-1} \log(x_i) \end{aligned}$$

where $\beta > 0$ is our regularization hyperparameter.

5.3 Linear Regression on the $\mathbb{P}_d$ manifold.

Our framework naturally encompasses linear regression on the manifold of positive definite matrices endowed with the affine invariant metric [51]. Let $X \in \mathbb{R}^{d \times d}$ be our data and $y \in \mathbb{R}$ be our observed target. We aim to minimize the quadratic loss function $f(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$. For simplicity, we consider the model

$$\hat{y} = f(W) = \mathrm{tr}(WX) \ , \tag{31}$$

wherein the least square problem becomes:

$$\min_{W \in \mathbb{P}_d} \frac{1}{2} \left(\operatorname{tr}(W \operatorname{Sym}(X)) - y\right)^2 \tag{32}$$

with $\operatorname{Sym}(X) = \frac{X+X^\top}{2}$ (see [51, sec. 5] for more details). The flexibility of our structured regularization framework allows us to regularize (32) to incorporate side information. For instance, if we are given an estimator $\hat{W} \in \mathbb{P}_d$ we can reformulate the problem as

$$\min_{W \in \mathbb{P}_d} \frac{1}{2} \left(\operatorname{tr}\left(W \operatorname{Sym}(X)\right) - y\right)^2 + \beta d_\Phi(W, \hat{W}) \tag{33}$$

for a specified symmetric gauge function Φ (see Section 3.1.2). Based on our discussions about its computational advantages and similarity to the Riemannian distance (see Section 4.2.4), one can replace d_Φ with the S-divergence, δ_S^2, i.e., a regularizer based on symmetric gauge functions (see Example 3.23).

Moreover, we can induce sparsity by adding a sparsity inducing regularizer $R_\Phi(W)$, obtaining

$$\min_{W \in \mathbb{P}_d} \frac{1}{2} \left(\operatorname{tr}\left(W \operatorname{Sym}(X)\right) - y\right)^2 + \beta R_\Phi(W) \,. \tag{34}$$

We note that since d_Φ and R_Φ are g-convex regularizers, the resulting optimization problems (33) and (34) are both g-convex and hence can be solved using standard first-order Riemannian iterative methods. However, if one specifies d_Φ in (33) to be the S-divergence regularizer or replace $R_\Phi(W)$ with the diagonal loading regularizer (see Example 3.24) then the objective becomes g-convex *and* DC.

One can adapt (33) or (34) to the kernel learning or Mahalanobis distance learning problem (see sections 9.1 and 9.2 [51]). For example, one can use (33) to *anchor* the solution to $\hat{W} \in \mathbb{P}_d$ where $\hat{W}$ is an estimated Mahalanobis matrix from an auxiliary dataset that is representative of the data of interest. Alternatively, one can use (34) to encourage convergence to low-rank Mahalanbois matrices.

6 Experiments

6.1 Square Root Problem

We consider the problem of computing the square root of a matrix $A \in \mathbb{P}_d$. In this section, we demonstrate the competitive performance of CCCP against standard first-order Riemannian approaches for this problem. Recall that in this case, the CCCP oracle can be solved in closed form, rendering CCCP into a simple fixed point approach (see Eq. 17).

Data Generation

We generate both medium-conditioned and ill-conditioned data. In the *medium-conditioned* case, we construct

$$A = GG^\top \qquad \text{where} \qquad G_{ij} \overset{\text{i.i.d.}}{\sim} N(0,1).$$

We further consider the Hilbert matrix

$$H_{ij} = \frac{1}{i+j-1},$$

a notable example of a very ill-conditioned matrix.

Results

We showcase the performance of computing the square roots $A^{\frac{1}{2}}$ and $H^{\frac{1}{2}}$ with CCCP and benchmark against two Riemannian Conjugate Gradient (RCG) and Riemannian Gradient Descent (RGD).

Figures 1 and 2 give results for all three algorithms in both cases. The stepsizes for RGD and RCG are determined by backtracking line search. In contrast, the fixed-point algorithm does not require a stepsize. As a reference point, the true optimum X^* is computed using NumPy's square root function. The CCCP algorithm exhibits superior performance in terms of runtime. In the ill-conditioned case, we observe that RGD and RCG are unable to converge to the optimum. In contrast, CCCP attains the optimum even for ill-conditioned data. We believe this is due to the fact that the fixed-point algorithm computes the inverse of $X_k + A$ (medium-conditioned) rather than of A (ill-conditioned), which is required by the gradient-based methods. This is illustrated in Figure 3 and particularly evident, if we cross-compare the accuracy achieved by the three approaches:

$$\begin{aligned} \|H - \hat{H}_{\mathrm{CCCP}}\|_F &= 8.9 \times 10^{-5} \\ \|H - \hat{H}_{\mathrm{RGD}}\|_F &= 129.791 \\ \|H - \hat{H}_{\mathrm{RCG}}\|_F &= 126.335 \end{aligned} \quad .$$

We discuss this observation in more detail in the next section.

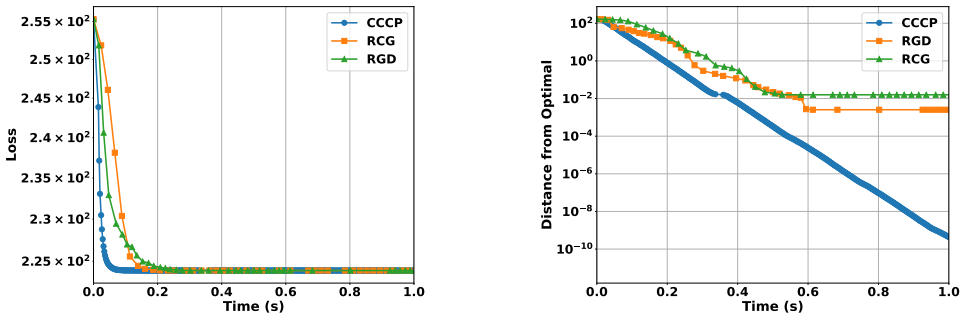

Fig. 1 We apply the fixed-point algorithm (see Proposition 5.1) to the medium conditioned $H \in \mathbb{R}^{200\times 200}$. We initialized all algorithms at $X_0 = 3I_d$. Although all three methods are of the same order in terms of per-iteration-complexity, the fixed-point method exhibits superior runtime performance. The stepsizes for RGD and RCG are chosen using backtracking line search. In contrast, the fixed-point algorithm does not need a stepsize. Distance is measured in terms of the Frobenius norm.

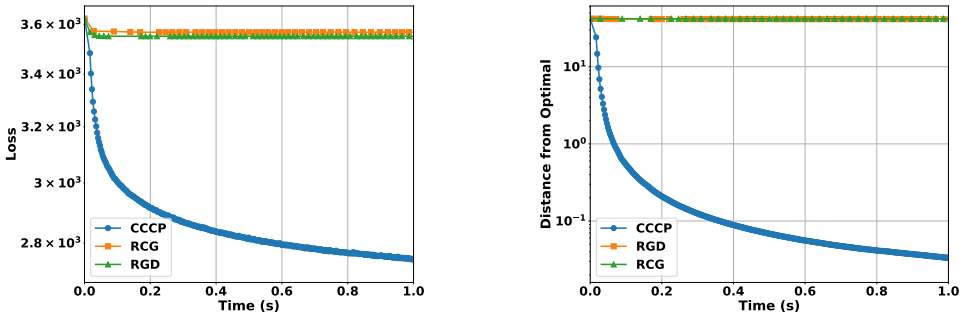

Fig. 2 We apply the fixed-point algorithm (see Proposition 5.1) to the ill-conditioned Hilbert matrix where we took dimension $d = 200$. We initialized all algorithms at $X_0 = 3I_d$. The benchmarks fail to converge whereas CCCP exhibits robustness to ill conditioning.

Gradient Steps

The poor convergence of the Riemannian first-order methods holds across different very ill-conditioned matrices beyond the Hilbert matrix. For example, one can take the ill-conditioned linear low k-rank projections of $GG^\top$ with small perturbation δI for $\delta \approx 0$. The first-order methods fail to converge for this ill-conditioned matrix as well. In the following, we discuss this observation and a possible explanation on the example of the Hilbert matrix. Recall that RGD preforms updates (for some stepsize $\eta > 0$)

$$X \leftarrow \operatorname{Exp}(-\eta \operatorname{grad} \phi(X)) \,,$$

where the exponential map is defined as

$$\operatorname{Exp}_X(tV) = X^{1/2} \exp\left(tX^{-1/2} V X^{-1/2}\right) X^{1/2}.$$

Inserting this and the Riemannian gradient $\operatorname{grad} \phi(X)$ (see sec. 2.1) above gives

$$\operatorname{Exp}(-\eta \operatorname{grad} \phi(X)) = X^{1/2} \exp\left(-\eta X^{1/2} \nabla_X \bar{\phi}(X) X^{1/2}\right) X^{1/2}$$

with

$$\nabla \bar{\phi}(X) = \frac{1}{2}\left(\frac{X+A}{2}\right)^{-1} + \frac{1}{2}\left(\frac{X+I}{2}\right)^{-1} - X^{-1} \,.$$

We suspect that computing the inverse of X^{-1}, i.e., inverting the Hilbert matrix at each iteration of the first-order methods results in numerical instability leading to the exhibited poor convergence behaviour. In contrast, the CCCP approach (Eq. 17) is robust to ill-conditioned matrices. Adding the identity $X + I$ improves the condition number of X before taking inverses. Although $X + A$ does not have better conditioning than A in general, we observe that in cases where A is ill-conditioned, the matrix $X + A$ actually becomes well-conditioned in practice. Heuristically, this follows from

Weyl's inequality which implies

$$\kappa(A+B) \leq \frac{\lambda_{\max}(A)+\lambda_{\max}(B)}{\lambda_{\min}(A)+\lambda_{\min}(B)} \qquad \text{for} \qquad A, B \in \mathbb{P}_d \,,$$

where $\kappa(X)$ is the condition number of X. This is demonstrated in Figure 3.

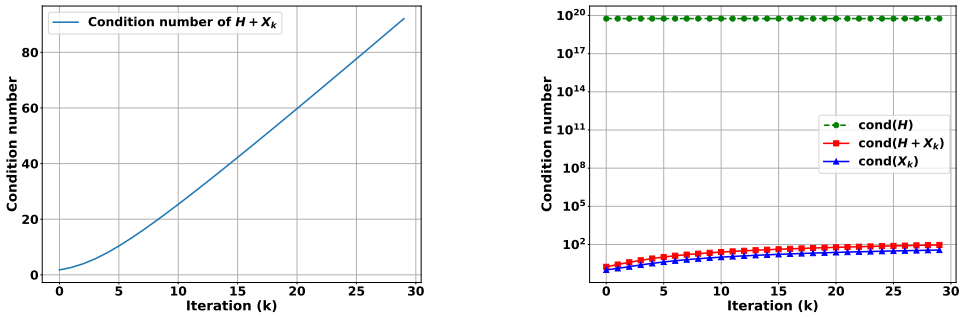

Fig. 3 We generate the Hilbert matrix $H \in \mathbb{R}^{200\times 200}$. We plot the condition number of $H + X_k$ where X_k is the k-th iterate of the fixed-point algorithm 17 and compare this to the condition number of H. Clearly, $H + X_k$ is much better conditioned than H. This trend also holds for other very ill-conditioned matrices.

6.2 Karcher Mean

In this section, we compare the performance of the CCCP approach, i.e., the fixed point approach given by Eq. 19, to RGD and RCG for the Karcher mean problem (Eq. 18).

Data generation

We focus on the medium-conditioned case, where we sample $G_1, \ldots, G_m$ random matrices, each with i.i.d standard Gaussian entries, and construct the data points $A_k \overset{\text{def}}{=} G_k G_k^\top$. A proxy for the true optimum is obtained by averaging the last iterate of the three algorithms upon convergence.

Results

Figures 4 and 5 show the convergence of all three algorithms. We see that the CCCP algorithm exhibits superior runtime compared to the two Riemannian first-order methods. Notably, the gap between CCCP and the two gradient-based approaches only widens as we increase the dimensionality and number of data samples.

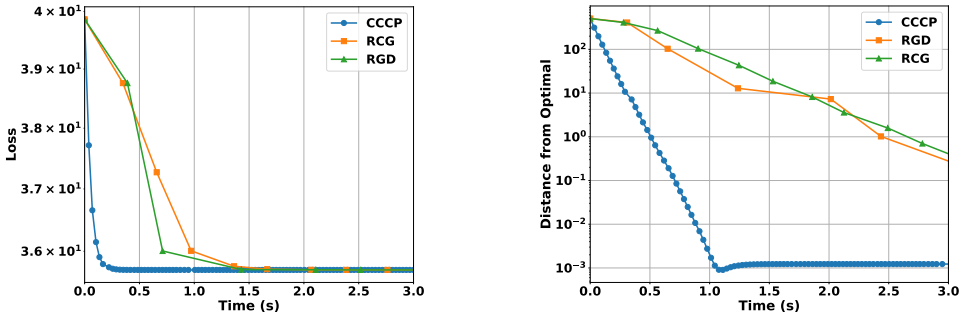

Fig. 4 Karcher Mean. $m = 100$ and $d = 100$. . CCCP demonstrates superior runtime complexity.

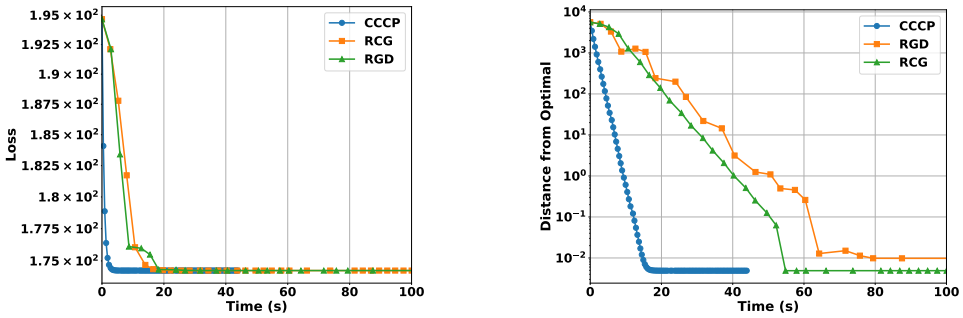

Fig. 5 Karcher Mean. $m = 100$ and $d = 500$. We observe that the gap between the runtime performance between CCCP and the benchmarks widens as we take the dimension d to be larger.

6.3 Optimistic Gaussian Likelihood

In this section, we test the performance of CCCP on the problem of computing the optimistic likelihood, introduced in sec. 5.2.1:

$$\underset{\Sigma \in \mathbb{P}_d}{\arg\min} \left\{ \hat{\phi}(\Sigma) \overset{\text{def}}{=} \operatorname{tr}\left(S\Sigma^{-1}\right) + \log\det\Sigma + \beta\delta_S^2\left(\Sigma, \hat{\Sigma}\right) \right\}. \tag{35}$$

Data Generation

We follow an experimental setup similar to that of [21]. In particular, we generate the true covariance Σ and its estimate $\hat{\Sigma} \in \mathbb{P}_d$ as follows. First we draw a Gaussian random matrix A with i.i.d. entries $A_{ij} \sim \mathcal{N}(0,1)$. Then we symmetrize and ensure it is positive definite via $\Sigma = \frac{1}{2}\left(A + A^\top\right) + \delta I$. To construct $\hat{\Sigma}$ we conduct the eigenvalue decomposition $\Sigma = Q\Lambda Q^\top$ and replace the eigenvalues in Λ with a random diagonal matrix $\hat{D}$ whose diagonal elements are sampled independently and uniformly from $\{1, 2, \ldots, 50\}$.

Our experiments in Figure 7 illustrates the output of Algorithm 2 and its distance from the true covariance Σ. The trend of the curves matches the intuition that increasing $\beta > 0$ and therefore increasing the confidence in $\hat{\Sigma}$ results in a solution closer to $\hat{\Sigma}$. We also note that higher regularization $\beta > 0$ results in faster convergence as

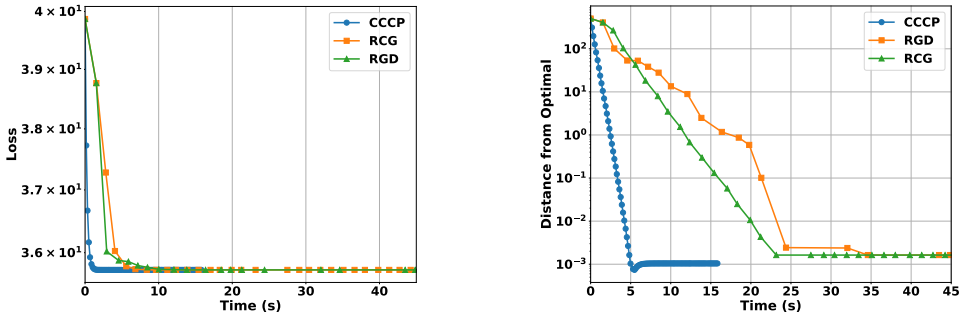

Fig. 6 Karcher Mean. $m = 500$ and $d = 100$. We observe that the gap between the runtime performance between CCCP and the benchmarks widens as we take the number of samples m to be larger.

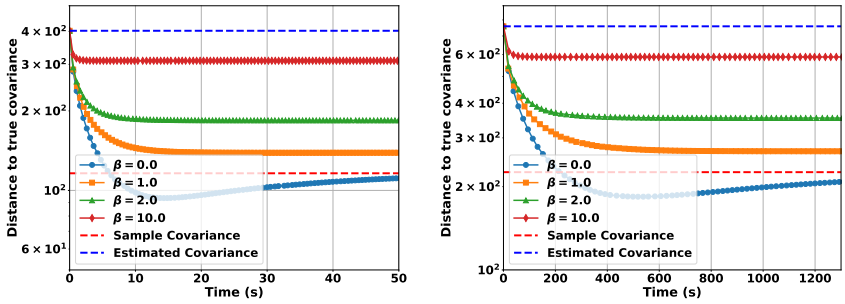

Fig. 7 Algorithm 2 on Gaussian Optimistic Likelihood. We sampled $n = 100$ independent Gaussian vectors of dimension $d = 30$ for the left plot. Meanwhile, the right plot was generated with $n = 1500$ and $d = 100$. We initialized our iterate at our estimate $\hat{\Sigma}$. As we increase β, Algorithm 2 converges to a solution $\hat{\Sigma}_\beta$ closer to $\hat{\Sigma}$. At $\beta = 0$, the algorithm converges to the sample covariance, i.e., $\hat{\Sigma}_\beta = S$. Refer to table 2 for the distances between $\|\hat{\Sigma}_\beta - \Sigma^*\|_F$ and $\|\hat{\Sigma}_\beta - \hat{\Sigma}\|_F$ for varying β.

exhibited in Figure 7. Moreover, Table 2 illustrates the interpolation behaviour of $\hat{\Sigma}_\beta$ between S and $\hat{\Sigma}$ as a function of β.

β	$\|\hat{\Sigma}_\beta - S\|_F$	$\|\hat{\Sigma}_\beta - \hat{\Sigma}\|_F$
0	0.149	412.009
1	116.545	296.022
2	178.140	234.583
10	316.485	96.094

β	$\|\hat{\Sigma}_\beta - S\|_F$	$\|\hat{\Sigma}_\beta - \hat{\Sigma}\|_F$
0	21.741	780.233
1	231.662	566.262
2	351.602	446.663
10	617.255	180.798

Table 2 $\hat{\Sigma}_\beta$ denotes the output of Algorithm 2 with different β values. The left table corresponds to the case $n = 100$ and $d = 30$. The right table corresponds to $n = 1500$ and $d = 100$.

6.4 Optimistic Multivariate T-Likelihood

We perform an experiment similar to that in the previous section for Algorithm 4. We generate d-dimensional random vectors $x_1, \ldots, x_n \overset{\text{iid}}{\sim} \text{MVT}(\Sigma; \nu)$ and observe the output of Algorithm 4 for different values of γ. We observe that incorporating $\hat{\Sigma}$ improves the distance from optimality. However, we observed that this method requires a high number of data points (See Figure 9).

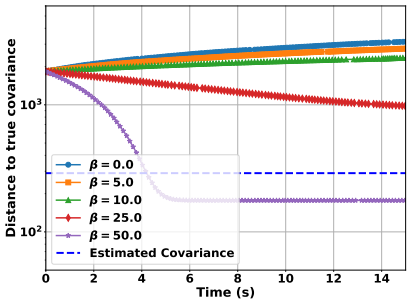

Fig. 8 Multivariate T-Distribution Optimistic Likelihood. We sampled $n = 1200$ i.i.d. multivariate t vectors of dimension $d = 15$. The plot indicates similar behaviour of that to the Gaussian optimistic likelihood problem: higher regularization encourages solution to be nearer $\hat{\Sigma}$ and also exhibits faster convergence.

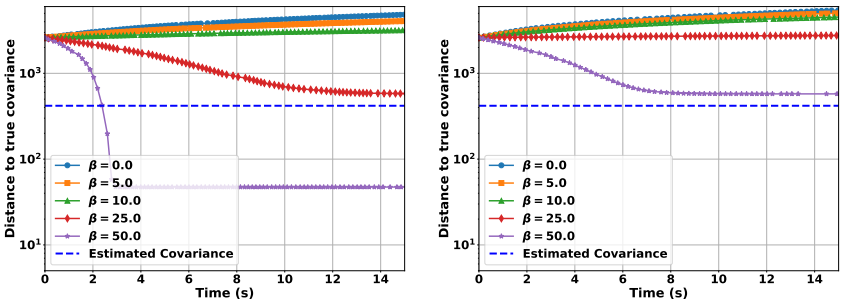

Fig. 9 Multivariate T-Distribution Optimistic Likelihood. Both plots are $d = 30$ and $\nu = 5$. The left plot is generated with $n = 1500$ and the right with $n = 3000$. This method requires a high number of samples for it to correspond with the theory. In particular, the right-side plot agreees with the theory: higher regularization corresponds to closer solutions to $\hat{\Sigma}$ in a monotonic fashion. The left-hand side with insufficient data points violates this.

7 Discussion

In this paper we introduced structured regularization approaches for constrained optimization on the SPD manifold. We considered different classes of constraints with

a particular focus on sparsity and ball constraints. Our structured regularization approach relies on symmetric gauge functions, whose algebraic properties give rise to a modular framework that allows for designing regularizers that preserve desirable properties of the orginial objective, specifically geodesic convexity and difference of convex structure. We illustrate the utility of our approach on a range of data science and machine learning applications.

We believe that our proposed approach opens up new directions for constrained optimization on Riemannian manifolds that circumvents the potentially costly subroutines of standard constrained Riemannian optimization approaches, such as R-PGD and R-FW. While we focus on two specific classes of constraints for most of the paper, we believe that our approach could be applied much more broadly. Our discussion on disciplined programming with symmetric gauge functions may serve as a starting point for future work in this direction. The introduction of new regularizers for structured constraints could significantly widen the range of applications in machine learning and data science. Furthermore, while this paper only discusses constrained optimization on the SPD manifold, we believe that many of the ideas could be extended to other Cartan-Hadamard manifolds.

Acknowledgements

We thank Bobak Kiani and Vaibhav Dixit for helpful discussions and comments.

This work was supported by the Harvard Dean's Competitive Fund for Promising Scholarship and NSF award 2112085. AC is partially supported by a NSERC Postgraduate Fellowship.

References

[1] Tyler, D. E. A distribution-free M-estimator of multivariate scatter. *The Annals of Statistics* 234–251 (1987).

[2] Ollila, E. & Tyler, D. E. Regularized M-Estimators of Scatter Matrix. *IEEE Transactions on Signal Processing* **62**, 6059–6070 (2014).

[3] Wiesel, A. Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing* **60**, 6182–6189 (2012).

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

[5] Sra, S. Positive definite matrices and the s-divergence. *arXiv:1110.1773* (2013).

[6] Sra, S. On the matrix square root via geometric optimization. *arXiv:1507.08366* (2015).

[7] Weber, M. & Sra, S. Computing Brascamp-Lieb Constants through the lens of Thompson Geometry. *arXiv:2208.05013* (2022).

[8] Mariet, Z. & Sra, S. *Fixed-point algorithms for learning determinantal point processes*, 2389–2397 (PMLR, 2015).

[9] Liu, C. & Boumal, N. Simple algorithms for optimization on Riemannian manifolds with constraints. *arXiv:1901.10000* (2019).

[10] Weber, M. & Sra, S. Riemannian optimization via Frank-Wolfe methods. *Mathematical Programming* (2022).

[11] Weber, M. & Sra, S. Projection-free nonconvex stochastic optimization on Riemannian manifolds. *IMA Journal of Numerical Analysis* **42**, 3241–3271 (2021).

[12] Weber, M. & Sra, S. *Global optimality for Euclidean CCCP under Riemannian convexity*, Vol. 202, 36790–36803 (PMLR, 2023).

[13] Udriste, C. *Convex Functions and Optimization Methods on Riemannian Manifolds* Vol. 297 (Springer Science & Business Media, 1994).

[14] Bacak, M. *Convex Analysis and Optimization in Hadamard Spaces* (De Gruyter, Berlin, München, Boston, 2014).

[15] Zhang, H. & Sra, S. *First-order methods for geodesically convex optimization*, 1617–1638 (2016).

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

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

[18] Zhang, H., J. Reddi, S. & Sra, S. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. *Advances in Neural Information Processing Systems* **29** (2016).

[19] Bien, J. & Tibshirani, R. Sparse estimation of a covariance matrix. *Biometrika* **98**, 807–820 (2011).

[20] Wiesel, A. & Zhang, T. Structured robust covariance estimation. *Foundations and Trends in Signal Processing* **8**, 127–216 (2015).

[21] Nguyen, V. A., Shafieezadeh-Abadeh, S., Yue, M.-C., Kuhn, D. & Wiesemann, W. *Calculating optimistic likelihoods using (geodesically) convex optimization* (2019).

[22] Tuy, H. *D.C. Optimization: Theory, Methods and Algorithms*, 149–216 (Springer US, Boston, MA, 1995).

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

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

[25] Ferreira, O. P., Santos, E. M. & Souza, J. C. O. The difference of convex algorithm on riemannian manifolds. *arXiv preprint arXiv:2112.05250* (2021).

[26] Nesterov, Y. & Nemirovski, A. *Interior-point polynomial algorithms in convex programming* (1994).

[27] Lanckriet, G. & Sriperumbudur, B. K. On the convergence of the concave-convex procedure. *Advances in Neural Information Processing Systems* **22** (2009).

[28] Lim, Y. Convex geometric means. *Journal of Mathematical Analysis and Applications* **404**, 115–128 (2013).

[29] Bhatia, R. *Matrix Analysis* Vol. 169 (Springer, 1997).

[30] Uhler, C. Gaussian graphical models: An algebraic and geometric perspective. *arXiv:1707.04345* (2017).

[31] Vandereycken, B. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization* **23**, 1214–1236 (2013).

[32] Karcher, H. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics* **30**, 509–541 (1977).

[33] Von Neumann, J. *Some matrix-inequalities and metrization of matric space* (1937).

[34] Thompson, A. C. On certain contraction mappings in a partially ordered vector space. *Proceedings of the American Mathematical Society* **14**, 438–443 (1963).

[35] Bhatia, R. *Positive Definite Matrices* Princeton Series in Applied Mathematics (Princeton University Press, Princeton, NJ, USA, 2007).

[36] Cheng, A., Dixit, V. & Weber, M. Disciplined geodesically convex programming. *arXiv:2407.05261* (2024).

[37] Vishnoi, N. K. Geodesic convex optimization: Differentiation on manifolds, geodesics, and convexity. *arXiv:1806.06373* (2018).

[38] Mohan, K. & Fazel, M. Iterative reweighted algorithms for matrix rank minimization. *Journal of Machine Learning Research* **13**, 3441–3473 (2012).

[39] Scieur, D., Kerdreux, T., d'Aspremont, A. & Pokutta, S. Strong convexity of sets in Riemannian manifolds. *arXiv:2312.03583* (2023).

[40] Complexity of linear minimization and projection on some sets. *Operations Research Letters* **49**, 565–571 (2021).

[41] Cherian, A., Sra, S., Banerjee, A. & Papanikolopoulos, N. Efficient similarity search for covariance matrices via the Jensen-Bregman LogDet divergence. *International Conference on Computer Vision* 2399–2406 (2011).

[42] Jeuris, B., Vandebril, R. & Vandereycken, B. A survey and comparison of contemporary algorithms for computing the matrix geometric mean. *Electronic Transactions on Numerical Analysis* **39**, 379–402 (2012).

[43] Fang, K.-T., Kotz, S. & Ng, K. W. *Symmetric Multivariate and Related Distributions* (Chapman and Hall/CRC, 2018).

[44] Sapatinas, T. Discriminant Analysis and Statistical Pattern Recognition. *Journal of the Royal Statistical Society Series A: Statistics in Society* **168**, 635–636 (2005).

[45] Price, L. F., Drovandi, C. C., Lee, A. & Nott, D. J. Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics* **27**, 1–11 (2017).

[46] Said, S., Bombrun, L., Berthoumieu, Y. & Manton, J. H. Riemannian Gaussian distributions on the space of symmetric positive definite matrices. *IEEE Transactions on Information Theory* **63**, 2153–2170 (2017).

[47] Sra, S. & Hosseini, R. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization* **25**, 713–739 (2015).

[48] Kotz, S. & Nadarajah, S. *Multivariate t-Distributions and Their Applications* (Cambridge University Press, 2004).

[49] Lange, K. L., Little, R. J. A. & Taylor, J. M. G. Robust statistical modeling using the t-distribution. *Journal of the American Statistical Association* **84**, 881–896 (1989).

[50] Nadarajah, S. & Kotz, S. Mathematical properties of the multivariate t-distribution. *Acta Applicandae Mathematicae* **89**, 53–84 (2005).

[51] Meyer, G., Bonnabel, S. & Sepulchre, R. Regression on fixed-rank positive semidefinite matrices: A Riemannian approach. *Journal of Machine Learning Research* **12**, 593–625 (2011).