OPT2024: 16th Annual Workshop on Optimization for Machine Learning

Structured Regularization on the SPD Manifold

Andrew Cheng ANDREWCHENG@G.HARVARD.EDU
Melanie Weber MWEBER@SEAS.HARVARD.EDU
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 side information on the solution. 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 propose a structured regularization framework based on symmetric gauge functions and *disciplined geodesically convex programming*. We show that the regularizer preserves crucial structure in the objective, including geodesic convexity. This allows for solving the regularized problem with a fast unconstrained method with a global optimality certificate. We demonstrate the effectiveness of our approach in numerical experiments on two examples, the computation of the Karcher mean of SPD matrices and Optimistic Gaussian Likelihood estimation.

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 [15, 19, 25], robust subspace recovery [27], the computation of Brascamp-Lieb constants [23], and learning determinantal point processes [12], among others. We are particularly interested in constraints that encode *side information*, such as a coarse estimate of the solution, which can be enforced as a ball constraint $\mathcal{X} = \mathcal{B}_R(\hat{X}) \overset{\text{def}}{=} \{X \in \mathbb{P}_d : \delta(X, \hat{X}) \leq \rho\}$ with respect to some metric δ on $\mathbb{P}_d$. A notable example of this problem class is *optimistic likelihood estimation* [14].

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, this implies a global optimality certificate 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) [11] and projection-free Riemannian Frank-Wolfe (R-FW) methods [21, 22]. 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 propose a regularization approach based on *symmetric gauge functions*, which allows for preserving desirable properties, such as geodesic convexity and difference of convex (DC) structure in the objective. 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 [18, 25]. In settings where the DC objectives is geodesically convex, we can leverage a Riemannian analysis to obtain global optimality certificates [24]. 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 enables the design of new regularizers for a variety of programs. We demonstrate the utility of our approach in numerical experiments.

2. Background

2.1. Geometry of the SPD manifold

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*. In particular, if $\mathbb{P}_d$ is 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 $\gamma(t) = A^{1/2}\left(A^{-1/2}BA^{-1/2}\right)^t A^{1/2}$ $(0 \leq t \leq 1)$. The *Riemannian metric* corresponding to this geometry is given by $\delta_R(A, B) = \left\|\log A^{-1/2}BA^{-1/2}\right\|_2$.

The Euclidean geometry of $\mathbb{P}_d$ is induced by endowing the symmetric positive definite matrices with the smooth inner product $\langle A, B\rangle = \operatorname{tr}(A^\top B)$ for all $A, B \in \mathbb{P}_d$. 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 [13]. We further use the following convexity notions.

Definition 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 $\gamma : [0,1] \to \mathbb{P}_d$ between them lies entirely in S, i.e., the image satisfies $\gamma([0,1]) \subseteq S$.*

Definition 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 $\gamma(0) \neq \gamma(1)$.*

2.2. Difference of Convex (DC) Optimization

Optimization tasks on the SPD manifold 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 idea of convex-concave procedures (short: CCCP) is to iteratively minimize a majorization surrogate function instead of the original, non-convex objective (see 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 [9], but due to non-convexity, a 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 3 ([24]) *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 $\phi(x_k) - \phi(x^*) \leq \frac{2L\alpha_{\mathcal{M}}^2(R)}{k+2}$ $(\forall k \geq 1)$, where $\alpha_{\mathcal{M}}$ depends on the geometry of the manifold and L characterizes the smoothness of $h(\cdot)$.*

3. Structured Regularization

Regularization approach The properties of symmetric gauge functions [1, 10] will form the basis for the design of our structured regularization approach.

Definition 4 (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 $\sigma_n : \mathbb{R}^d \to \mathbb{R}^d$ (known as* 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 $\alpha_k \in \{\pm 1\}$ (known as* gauge invariant *or* absolute property*).*

With an abuse of notation, we denote $\Phi : \mathbb{P}_d \to \mathbb{R}$ as $\Phi(A) \overset{\text{def}}{=} \Phi(\lambda(A))$, i.e., $\Phi(A)$ acts on the eigenspectrum of A. Notably Φ induces metrics d_Φ and norms $\|\cdot\|_\Phi$ that are particularly well-suited for our mixed Euclidean-Riemannian perspective. In particular, $\|A\|_\Phi \overset{\text{def}}{=} \Phi(\lambda(A))$ is g-convex; d_Φ is a complete metric on the convex cone of $\mathbb{P}_d$ and g-convex. We defer a more detailed discussion to section C.3.

These observations imply that regularizing problem 1 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 introduce a hyperparameter $\beta > 0$

to control the strength of the regularizer. For the specific example of the ball constraint discussed above, the 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

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

is g-convex and DC, too.

Designing regularizers via disciplined programming. To design regularizers for a wide range of constrained tasks, we take a *disciplined programming* approach [4, 6]: We refer to a repository of known g-convex functions (for e.g., symmetric gauge functions) and apply g-convexity preserving operations to construct g-convex functions that act as regularizers. Below, we give two examples to illustrate this idea. More details on basic g-convex functions and g-convexity preserving operations can be found in [4].

Proposition 1 *The following are g-convex functions [4].*

1. *All symmetric gauge functions functions $\Phi : \mathbb{P}_d \to \mathbb{R}$ defined by $\Phi(A) \overset{def}{=} \Phi(\lambda(A))$ are g-convex. This includes the ℓ_p-Schatten norms for $p \geq 1$ and the Ky-fan norms [1].*
2. *The log-determinant $\log\det : \mathbb{P}_d \to \mathbb{R}_{++}$ is both g-convex and g-concave. Moreover, $\log\det(\cdot)$ is Euclidean concave.*

Proposition 2 *The following operations preserve g-convexity [4].*

1. *If $f_i : \mathbb{P}_d \to \mathbb{R}$ are g-convex then $f(X) = \sum_{i=1}^n f_i(X)$ is g-convex for $\alpha_i \geq 0$.*
2. *If $f : \mathbb{P}_d \to \mathbb{R}$ is a strictly positive linear map, i.e. $f(X)$ is linear and is positive definite whenever X is positive definite then the function $g(X) = \log\det f(X)$ is g-convex.*

Example 1 (Diagonal Loading [26]) *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 $R_\Phi(X) \overset{def}{=} trX^{-1} + \log\det X = \|X^{-1}\|_\Phi + d_\Phi(X, I_d)$, where Φ is the Schatten 1-norm. Then $R_\Phi(X)$ is g-convex and DC.*

Example 2 (S-divergence [18]) *For fixed $Y \in \mathbb{P}_d$, the s-divergence $\delta_S^2 : \mathbb{P}_d \to \mathbb{R}_{++}$ defined by $\delta_S^2(X, Y) \overset{def}{=} \log\det\left(\frac{X+Y}{2}\right) - \frac{1}{2}\log\det(XY)$ is g-convex and DC.*

4. Applications

Karcher Mean The Karcher mean [7, 8] corresponds to the centroid of SPD matrices. 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 $\min_{X \in \mathbb{P}_d} \overset{\text{def}}{=} \sum_{i=1}^m w_i \delta_R^2(X, A_i)$. Remarkably, [18] showed that the Karcher mean problem can be reformulated in terms of the s-divergence, a symmetric gauge function:

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

Importantly, this formulation is g-convex and DC, which allows for deriving an effective CCCP approach [18]:

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

We demonstrate the competitive performance of this approach in comparison with first-order Riemannian methods in Fig. 1.

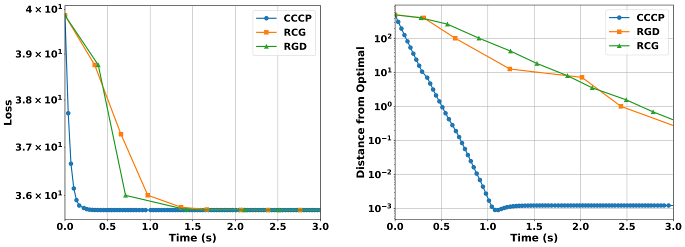

Figure 1: **Karcher Mean** for inputs with $m = 100$ and $d = 100$. CCCP outperforms Riemannian first-order methods.

Gaussian optimistic likelihood Consider a set of i.i.d data points $x = (x_1, \ldots, x_n) \in \mathbb{R}^d$ generated from 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 $|C| < \infty$. The true Gaussian distribution can be determined by solving

$$c^\star \in \operatorname*{arg\,min}_{c \in \mathcal{C}} \left\{ \phi\left(\Sigma_c; x\right) \overset{\text{def}}{=} \frac{1}{n} \sum_{k=1}^{n} x_k^\top \Sigma_c^{-1} x_k + \log\det \Sigma_c \right\} \ . \tag{3}$$

Problem (3) has applications in machine learning (e.g. quadratic discriminant analysis [17]) and in statistics (e.g. Bayesian inference [16]). In general, Σ_c is unknown, but we can obtain an estimator $\hat{\Sigma}_c$ from the data. Problem (3) is highly sensitive to misspecification of the candidate distributions $\mathcal{N}(0, \Sigma_c)$, due to which the following constrained formulation can be more effective [14]:

$$\min_{\Sigma \in \mathcal{B}_R(\hat{\Sigma}_c; \rho_c)} \phi(\Sigma; x) \qquad \text{where} \qquad \mathcal{B}_R(\hat{\Sigma}_c; \rho_c) \overset{\text{def}}{=} \left\{ \Sigma \in \mathbb{P}_d : \delta_R\left(\Sigma, \hat{\Sigma}_c\right) \leq \rho_c \right\} \ . \tag{4}$$

Problem (4) is an instance of problem 1 with side information. The shared properties of δ_S^2 and δ_R (see Apx. C.3) allows for a structured regularization of the form

$$\operatorname*{arg\,min}_{\Sigma \in \mathbb{P}_d} \left\{ \hat{\phi}(\Sigma) \overset{\text{def}}{=} \operatorname{tr}\left(S\Sigma^{-1}\right) + \log\det\Sigma + \beta\delta_S^2\left(\Sigma, \hat{\Sigma}\right) \right\} . \tag{5}$$

The regularization preserves the g-convexity and DC structure, which allows for applying a CCCP approach (see Algorithm 2 in Apx. B.1). Numerical results for this approach can be found in Figure 2.

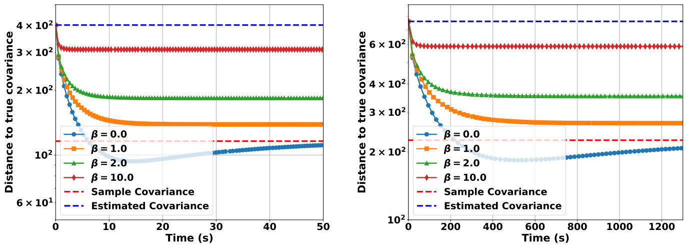

Figure 2: **Optimistic Gaussian Likelihood** for inputs of size $n = 100$, $d = 30$ (left) and $n = 1500, d = 100$ (right). As we increase β, Algorithm 2 converges to a solution $\hat{\Sigma}_\beta$ closer to $\hat{\Sigma}$. Increasing β also results in faster convergence. At $\beta = 0$, the algorithm converges to the sample covariance, i.e., $\hat{\Sigma}_\beta = S$.

5. Discussion

In this paper we introduced a structured regularization approach for constrained optimization on the SPD manifold. Our regularizers rely 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. An extended version of this paper [3] constructs structured regularizers for a wider range of problems and discusses additional applications.

We believe that our proposed approach opens up new directions for constrained optimization on Riemannian manifolds that circumvents the costly subroutines of previous constrained Riemannian optimization approaches. While this paper only discusses constrained optimization on the SPD manifold, we believe that many of the ideas could be extended to more general Cartan-Hadamard manifolds.

Acknowledgements

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] Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer, 1997.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] Andrew Cheng and Melanie Weber. Structured regularization for constrained optimization on the spd manifold. *arXiv preprint arXiv:2410.09660*, 2024.

[4] Andrew Cheng, Vaibhav Dixit, and Melanie Weber. Disciplined geodesically convex programming. *arXiv:2407.05261*, 2024.

[5] Anoop Cherian, Suvrit Sra, Arindam Banerjee, and Nikolaos Papanikolopoulos. Efficient similarity search for covariance matrices via the Jensen-Bregman LogDet divergence. *International Conference on Computer Vision*, pages 2399–2406, 2011.

[6] Michael Grant, Stephen Boyd, and Yinyu Ye. *Disciplined Convex Programming*, pages 155–210. Springer US, Boston, MA, 2006.

[7] Ben Jeuris, Raf Vandebril, and Bart Vandereycken. A survey and comparison of contemporary algorithms for computing the matrix geometric mean. *Electronic Transactions on Numerical Analysis*, 39:379–402, 2012.

[8] Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30:509–541, 1977.

[9] Gert Lanckriet and Bharath K Sriperumbudur. On the convergence of the concave-convex procedure. *Advances in Neural Information Processing Systems*, 22, 2009.

[10] Yongdo Lim. Convex geometric means. *Journal of Mathematical Analysis and Applications*, 404(1):115–128, 2013.

[11] Changshuo Liu and Nicolas Boumal. Simple algorithms for optimization on Riemannian manifolds with constraints. *arXiv:1901.10000*, 2019.

[12] Zelda Mariet and Suvrit Sra. Fixed-point algorithms for learning determinantal point processes. In *International Conference on Machine Learning*, pages 2389–2397. PMLR, 2015.

[13] Yurii Nesterov and Arkadi Nemirovski. Interior-point polynomial algorithms in convex programming. In *SIAM Studies in Applied Mathematics*, 1994.

[14] Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh, Man-Chung Yue, Daniel Kuhn, and Wolfram Wiesemann. Calculating optimistic likelihoods using (geodesically) convex optimization. In *Neural Information Processing Systems*, 2019.

[15] Esa Ollila and David E. Tyler. Regularized M-Estimators of Scatter Matrix. *IEEE Transactions on Signal Processing*, 62(22):6059–6070, 2014. doi: 10.1109/TSP.2014.2360826.

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

[17] Theofanis Sapatinas. Discriminant Analysis and Statistical Pattern Recognition. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 168(3):635–636, 06 2005.

[18] Suvrit Sra. Positive definite matrices and the s-divergence. *arXiv:1110.1773*, 2013.

[19] David E Tyler. A distribution-free M-estimator of multivariate scatter. *The Annals of Statistics*, pages 234–251, 1987.

[20] Nisheeth K. Vishnoi. Geodesic convex optimization: Differentiation on manifolds, geodesics, and convexity. *arXiv:1806.06373*, 2018.

[21] Melanie Weber and Suvrit Sra. Projection-free nonconvex stochastic optimization on Riemannian manifolds. *IMA Journal of Numerical Analysis*, 42(4):3241–3271, 2021.

[22] Melanie Weber and Suvrit Sra. Riemannian optimization via Frank-Wolfe methods. *Mathematical Programming*, 2022.

[23] Melanie Weber and Suvrit Sra. Computing Brascamp-Lieb Constants through the lens of Thompson Geometry. *arXiv:2208.05013*, 2022.

[24] Melanie Weber and Suvrit Sra. Global optimality for Euclidean CCCP under Riemannian convexity. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 36790–36803. PMLR, 23–29 Jul 2023.

[25] Ami Wiesel. Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing*, 60(12):6182–6189, 2012.

[26] Ami Wiesel and Teng Zhang. Structured robust covariance estimation. *Foundations and Trends in Signal Processing*, 8(3):127–216, 2015.

[27] Teng Zhang. Robust subspace recovery by Tyler's M-estimator. *Information and Inference: A Journal of the IMA*, 5(1):1–21, 2016.

Appendix A. CCCP Algorithm

Algorithm 1: Convex-Concave Procedure (CCCP)

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

Appendix B. Details on Applications

B.1. Optimistic Gaussian Likelihood

Figure 2 was created by following an experimental setup similar to that of [14]. 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 $\{1, 2, \dots, 50\}$.

The CCCP algorithm applied to the Gaussian optimistic likelihood takes the form of Algorithm 2.

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 $k = 0, \dots, K-1$ do
    Precompute $\Sigma_k^{-1} + \beta\left(\Sigma_k + \hat{\Sigma}\right)^{-1}$
    for $\ell = 0, \dots, L-1$ 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 $\Sigma_{k+1} \leftarrow \Sigma_L$
    end
end
Output: $\Sigma_K$
```

B.2. Experimental Setup

Karcher mean. We sample $G_1, \dots, G_m$ 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. The gap in performance only widens as m and d increases.

Optimistic Gaussian 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., $\hat{\Sigma}_\beta = S$.

Appendix C. Deferred Proofs

C.1. Disciplined Programming with symmetric gauge functions

The following proof shows that symmetric gauge functions are g-convex on $\mathbb{P}_d$.

Proof [Proposition 1] Symmetric gauge functions are g-convex as was proven in the previous proposition. To show that $f : \mathbb{P}_d \to \mathbb{R}_{++}$ is indeed g-concave with respect to the Euclidean metric we refer the reader to Section 3.1.5 [2]. Let $X, Y \in \mathbb{P}_d$ and $\gamma : [0, 1] \to \mathbb{P}_d$ be the geodesic segment connecting $\gamma(0) = A$ to $\gamma(1) = B$. For $t \in [0, 1]$

$$\begin{aligned}
\log\det\left(\gamma(t)\right) &= \log\det\left(X^{1/2}(X^{-1/2}YX^{-1/2})^t X^{1/2}\right) \\
&= \log\left(\det(X)\det(X^{-1})^t\det(Y)^t\right) \\
&= \log\det(X) - t\log\det(X) + t\log\det(Y) \\
&= (1-t)\log\det(X) + t\log\det(Y).
\end{aligned}$$

■

Proof [Proposition 2] For the proofs of (1) and (2), we refer the reader to Proposition 1 [4] and Proposition 5.8 [20], respectively. ■

Lemma 5 (Proposition 5 [4]) *Let $f : \mathbb{P}_d \to \mathbb{R}$ be g-convex. Then $g(X) = f(X^{-1})$ is also g-convex.*

Proof [Example 1] We can express $\mathrm{tr}X^{-1}$ as $f(X) = \mathrm{tr}(X^{-1}) = \|X^{-1}\|_\Phi$. By Proposition 3 and Lemma 5 we have $f(X)$ is g-convex. By Proposition 1, $\log\det(\cdot)$ is g-convex and the result follows from that fact that the sum of the two g-convex functions is g-convex. ■

Proof [Example 2] See the proof of Proposition 10 in [4]. ■

C.2. Properties of symmetric gauge functions

Symmetric gauge functions Φ induces unitarily invariant norms that are g-convex (Proposition 3) on the $\mathbb{P}_d$ manifold and complete metrics d_Φ on the convex cone $\mathbb{P}_d$. Hence they play particularly well with our mixed Riemmanian-Euclidean optimization perspective.

Proposition 3 *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 geodesically convex.*

Proof [Proposition 3] 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 [1], 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

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

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

$$A\#B \preceq \frac{A+B}{2} \implies \lambda^{\downarrow}(A\#B) \preceq \lambda^{\downarrow}\left(\frac{A+B}{2}\right). \tag{6}$$

Thus using (6) 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 [1] implies

$$\lambda^{\downarrow}\left(\frac{A+B}{2}\right) \prec_w \lambda^{\downarrow}\left(\frac{A}{2}\right) + \lambda^{\downarrow}\left(\frac{B}{2}\right) \ . \tag{7}$$

Also we know that Φ satisfies the *strongly isotone* property (see Page 45 [1]), 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 (7) gives

$$\Phi\left(\lambda\left(\frac{A+B}{2}\right)\right) = \Phi\left(\lambda^{\downarrow}\left(\frac{A+B}{2}\right)\right) \leq \Phi\left(\lambda^{\downarrow}\left(\frac{A}{2}\right) + \lambda^{\downarrow}\left(\frac{B}{2}\right)\right). \tag{8}$$

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

$$\begin{aligned}
\|A\#B\|_\Phi &= \Phi(\lambda(A\#B)) \\
&\leq \Phi\left(\lambda\left(\frac{A+B}{2}\right)\right) && \text{(Applying monotonicity of } \Phi \text{ to (6))} \\
&\leq \Phi\left(\lambda^{\downarrow}\left(\frac{A}{2}\right) + \lambda^{\downarrow}\left(\frac{B}{2}\right)\right) && \text{(Apply (8))} \\
&= \Phi\left(\frac{1}{2}\lambda^{\downarrow}(A) + \frac{1}{2}\lambda^{\downarrow}(B)\right) && \text{(Property of eigenvalues: } \lambda(A/2) = \tfrac{1}{2}\lambda(A)) \\
&\leq \frac{\Phi(\lambda^{\downarrow}(A)) + \Phi(\lambda^{\downarrow}(B))}{2} && (\Phi \text{ is a norm; triangle inequ., pos. homogeneity}) \\
&= \frac{\Phi(\lambda(A)) + \Phi(\lambda(B))}{2} && \text{(Remove } \downarrow \text{ by permutation invariance of } \Phi) \\
&\overset{\text{def}}{=} \frac{\|A\|_\Phi + \|B\|_\Phi}{2}
\end{aligned}$$

which proves the midpoint criterion for g-convexity. ■

Definition 6 *For a continuous segment $\gamma : [0,1] \to \mathbb{P}_d$ we define its length w.r.t a symmetric gauge function $\Phi : \mathbb{R}^n \to \mathbb{R}_+$ as*

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

We define the distance between $A, B \in \mathbb{P}_d$ with respect to Φ as

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

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 7 (Theorem 2.2 [10]) *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. *$d_\Phi\left(A\#_tB, A\#_sB\right) = |s-t|d_\Phi(A,B)$ for all $t, s \in [0,1]$;*
4. *$d_\Phi\left(A\#_tB, C\#_tD\right) \leq (1-t)d_\Phi(A,C) + td_\Phi(B,D)$ for all $t \in [0,1]$.*

Theorem 8 (Theorem 2.2 + Proposition 3.5 [10]) *We can explicitly express $d_\Phi(A,B) = \left\|\log\left(A^{-1/2}BA^{-1/2}\right)\right\|_\Phi$ and d_Φ is a complete metric distance on the convex cone of $\mathbb{P}_d$ and satisfies*

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 $\alpha \geq 1$.*

C.3. S-divergence and the Riemannian metric

The computation of the Riemannian metric requires computing the generalized eigenvalues of A and B, which introduces a computational bottleneck. To address this problem, [5] introduced a *symmetrized log-det based matrix divergence*, also known as the *S-divergence*. Sra et al. [18] discuss 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 (Table 4.1 [18]) *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.*** $\delta_S^2(A + X, B + X) \le \delta_S^2(A, B)$.
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 is also satisfied by the Riemannian metric δ_R. See Table 4 [18] 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 5 (Theorem 4.19 [18]) *Let $A, B \in \mathbb{P}_d$. Then, we have the following bound*

$$8\delta_S^2(A, B) \le \delta_R^2(A, B).$$

Proposition 6 *Fix $\hat{\Sigma} \in \mathbb{P}_d$ and fix $\alpha > 0$. Define the sets*

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

and

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

Then the subset-inequality

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

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

Proof By Proposition 5, we have the inequality

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

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

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

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 implementation 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 can be seen as a relaxation of the original problem. 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 [5].