Tyler's M-estimator Through the Lens of Convex-Concave Programming

Daniel Cederberg
Stanford University

Abstract

Tyler's M-estimator is a popular method for robust covariance estimation. Although Tyler proposed a widely used fixed-point iteration to compute the estimator nearly four decades ago, the theoretical properties of the fixed-point iteration have remained elusive. In particular, no deterministic global convergence rate has been established, despite decades of research. In this paper, we resolve this longstanding question by interpreting the fixed-point iteration as an instance of the convex-concave procedure, and analyzing it through a novel combination of relative smoothness and Riemannian optimization. Our analysis requires us to navigate two distinct geometric frameworks: one induced by the Kullback–Leibler divergence, and the other one arising from the Riemannian geometry on the manifold of positive definite matrices. This interplay between the two geometries also allows us to prove that a popular regularized variant of the fixed-point iteration converges linearly. Beyond contributing to the theoretical understanding of Tyler's M-estimator, our analysis demonstrates how the integration of Euclidean and Riemannian perspectives can lead to new insights into the design and analysis of statistical algorithms.

1 INTRODUCTION

Elliptical probability distributions are widely used in robust statistics and probabilistic machine learning beacause of their ability to model heavy-tailed and non-Gaussian data (Delmas et al., 2024). This rich family of distributions includes the Gaussian distribution, the multivariate t-distribution, and other heavy-tailed distributions sometimes used in practice. Elliptical distributions are characterized by their ellipsoidal level sets, with a so-called *shape matrix* capturing the geometry of the distribution, such as the principal directions. When the covariance matrix exists, it is proportional to the shape matrix; however, the shape matrix remains well-defined even when second moments are infinite. Estimating the shape matrix of an elliptical distribution is a fundamental problem, underpinning many methods in high-dimensional statistics such as principal component analysis.

Tyler's M-estimator (Tyler, 1987) is a popular estimator for the shape matrix of an elliptical distribution. Given m samples $x_1, \dots, x_m \in \mathbf{R}^n$ drawn from an n-dimensional distribution, it is defined as a positive definite matrix $\Sigma^\star \in \mathbf{S}^n$ satisfying the fixed-point equation

$$\Sigma^\star = \frac{n}{m} \sum_{i=1}^{m} \frac{1}{x_i^T (\Sigma^\star)^{-1} x_i} x_i x_i^T. \tag{1}$$

In addition to establishing powerful statistical properties of the estimator, such as distribution-free strong consistency, asymptotic normality, and minimax optimality with respect to the asymptotic variance, Tyler (1987) also proposed a widely used fixed-point iteration for computing it. (We describe it in §2.) His work has inspired decades of research on the estimator focusing on analyzing its properties and incorporating prior structure. However, no deterministic global convergence rate for his fixed-point iteration has been known so far. In this paper, we propose a simple yet powerful perspective on Tyler's M-estimator based on the convex-concave procedure (CCP) (Yuille and Rangarajan, 2001). By combining this perspective with tools from Riemannian optimization on the manifold of positive definite matrices (Boumal, 2023) and the concept of relative smoothness (Lu et al., 2018) with respect to the Kullback–Leibler divergence, we establish the first deterministic global convergence rate for Tyler's fixed-point iteration — a result that has long

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

been sought in the literature. This perspective also allows us to derive the first linear convergence rate for a popular regularized variant of the fixed-point iteration (Ollila and Tyler, 2014) that (unlike a previous result) holds across all levels of regularization. Finally, we show that this perspective unifies and extends existing methods for computing Tyler's M-estimator under structural constraints, including sparsity of the shape matrix (or its inverse) and factor model structure.

1.1 Related work and outline

This work contributes to two main areas of research: the study of Tyler's M-estimator, and, more broadly, the integration of Euclidean and Riemannian perspectives for analyzing optimization algorithms. We discuss prior work in both areas below.

Tyler's M-estimator. Tyler's M-estimator has been extensively studied in both the statistics and optimization communities (Tyler, 1987; Kent and Tyler, 1988; Wiesel and Zhang, 2015; Zhang and Wiesel, 2016; Danon and Garber, 2022). A popular line of work seeks to improve Tyler's M-estimator when the number of samples is few by incorporating regularization (Chen et al., 2011; Ollila and Tyler, 2014; Sun et al., 2014; Ollila et al., 2021), or by imposing structural constraints such as approximate sparsity (Goes et al., 2020), sparsity of the inverse shape matrix (Zhang et al., 2013), low-rank structure (Bouchard et al., 2021; Cederberg, 2026), or more general convex constraints such as Toeplitz structure (Soloveychik and Wiesel, 2014; Sun et al., 2016). Despite the rich literature on the topic, including specific efforts to analyze the convergence behavior of Tyler's fixed-point iteration and its regularized variant (Franks and Moitra, 2020; Goes et al., 2020), no deterministic global convergence rate has been obtained so far. (We discuss existing convergence results in §2.)

Riemannian optimization and CCP. Riemannian optimization provides a framework for extending classical optimization algorithms to settings where the domain is a Riemannian manifold rather than a Euclidean space. This perspective has led to natural generalizations of several Euclidean algorithms. In particular, the classical Euclidean CCP (Yuille and Rangarajan, 2001), originally designed for minimizing the difference of two convex functions, has been extended to handle differences of geodesically convex functions (Souza and Oliveira, 2015; Almeida et al., 2020; Bergmann et al., 2023). Closely related work has analyzed the application of Euclidean CCP to geodesically convex objectives expressed as differences of convex functions (Weber and Sra, 2023; Cheng and Weber, 2024). However, while Tyler's M-estimator can be computed by minimizing a geodesically convex function that is as a difference of convex functions, those results do not yield any convergence rate for Tyler's fixed-point iteration, for the following three reasons:

1. In this paper, we show that Tyler's fixed-point iteration is an instance of CCP applied to a maximum likelihood (ML) estimation problem *after a change of variables*. While Weber and Sra (2023); Cheng and Weber (2024) identify that the objective function of the ML estimation problem is a difference of convex functions and can be optimized using CCP, doing so without the variable transformation leads to a CCP subproblem that does not have an analytical solution. In contrast, after the change of variables, the CCP subproblem has an analytical solution and explicitly yields Tyler's fixed-point iteration.

2. The convergence analysis in Weber and Sra (2023) relies on the assumption that the component of the objective function that is linearized in each iteration of CCP is smooth in the classical Euclidean sense. This assumption is violated by the objective function of Tyler's ML estimation problem.

3. The analysis in Weber and Sra (2023) assumes that the initial sublevel set of the objective function is bounded. This assumption fails in the case of Tyler's ML estimation problem. Specifically, the radius $R > 0$ in Weber and Sra (2023, Theorem 3.5) is infinite for Tyler's ML estimation problem.

To overcome the second limitation, we rely on the notion of relative smoothness (Lu et al., 2018), rather than classical Euclidean smoothness. As in our analysis of Tyler's celebrated fixed-point iteration, relative smoothness has also recently played a key role in the analysis of other statistical algorithms, such as the Sinkhorn iteration and the expectation-maximization (EM) algorithm (Mishchenko, 2019; Kunstner et al., 2021; Aubin-Frankowski et al., 2022).

Outline. The rest of the paper is organized as follows. In §2, we present background on Tyler's M-estimator, the convex-concave procedure, relative smoothness, and Riemannian optimization. In §3, we show how Tyler's fixed-point iteration can be interpreted as an instance of CCP applied to an ML estimation problem. In §4, we present our main contribution of deriving global convergence rates for Tyler's fixed-point iteration and its regularized variant. We present our conclusions in §5. The supplementary material contains omitted proofs and examples of how the

CCP perspective can be used to derive algorithms for computing Tyler's M-estimator under various structural constraints.

2 BACKGROUND AND PRELIMINARIES

This paper uses tools and perspectives from several areas, including Tyler's M-estimator, convex-concave programming, relative smoothness, and Riemannian optimization. In this section, we review the key concepts from each of these domains that are relevant to our analysis.

2.1 Tyler's M-estimator

Computation via fixed-point iterations. Tyler's M-estimator, as defined in (1), can be interpreted as a solution to the optimization problem

$$\text{minimize} \ \log\det\Sigma + \frac{n}{m}\sum_{i=1}^{m}\log(x_i^T\Sigma^{-1}x_i) \tag{2}$$

with variable $\Sigma \in \mathbf{S}^n$. This equivalence follows from setting the gradient of the objective function of (2) to zero, which yields an equation that is equivalent to (1). Problem (2) is the maximum likelihood (ML) estimation problem for the shape matrix of a so-called *angular central Gaussian distribution*, which is the distribution obtained by normalizing an elliptically distributed random vector to have unit length (see Wiesel and Zhang (2015) for more details). In his seminal paper, Tyler (1987) proposed to solve the fixed-point equation (1), or equivalently the ML estimation problem (2), using the fixed-point iteration

$$\Sigma_{k+1} = \frac{n}{m}\sum_{i=1}^{m}\frac{1}{x_i^T\Sigma_k^{-1}x_i}x_ix_i^T. \tag{3}$$

Tyler originally formulated the iteration in terms of the inverse matrix Σ_k^{-1} and also included a normalization step in each iteration. However, in work appearing shortly thereafter, Kent and Tyler (1988) reformulated the iteration in terms of Σ_k and showed that the normalization step is unnecessary. As a result, the iteration (3) is commonly referred to as *Tyler's fixed-point iteration* in the modern literature (see, for example, Wiesel and Zhang (2015, §3)). While Tyler (1987) originally derived the iterative scheme (3) as a fixed-point iteration, it can also be interpreted as a majorization-minimization algorithm (Wiesel and Zhang, 2015) and as an EM algorithm (Cederberg, 2026). In contrast to our CCP-based interpretation, none of these alternative perspectives offer convergence rate results.

A widely used extension of Tyler's M-estimator, particularly effective when the number of samples is small, is to incorporate regularization (Chen et al., 2011; Sun et al., 2014; Ollila et al., 2021). Ollila and Tyler (2014) introduced a regularized version of the objective function of (2) by scaling the second term and adding a trace-inverse penalty, leading to the optimization problem

$$\begin{aligned}\text{minimize} \ \log\det\Sigma + \frac{(1-\alpha)n}{m}\sum_{i=1}^{m}\log(x_i^T\Sigma^{-1}x_i) \\ + \alpha\,\mathbf{Tr}(\Sigma^{-1})\end{aligned} \tag{4}$$

with variable $\Sigma \in \mathbf{S}^n$, where $\alpha \in (0,1)$ is a regularization parameter. Setting the gradient of the objective function of (4) to zero yields a fixed-point equation, which the authors propose solving via the iterative scheme

$$\Sigma_{k+1} = \frac{(1-\alpha)n}{m}\sum_{i=1}^{m}\frac{1}{x_i^T\Sigma_k^{-1}x_i}x_ix_i^T + \alpha I, \tag{5}$$

where $I \in \mathbf{R}^{n\times n}$ is the identity matrix. We will refer to this iterative scheme as *Tyler's regularized fixed-point iteration*.

Existing convergence results. Despite the popularity of Tyler's fixed-point iteration, not much is known about its convergence properties. One of the first *asymptotic* convergence results was established by Kent and Tyler (1988, Theorem 2), who proved that a solution to (1) exists and is unique (up to a positive scaling) under a geometric condition that avoids concentration of observed samples on lower-dimensional subspaces. They also proved that the fixed-point iteration (3) converges to a solution of (1) whenever the geometric condition is satisfied, but no convergence rate was established.

In more recent work, Goes et al. (2020) proved a *non-asymptotic* convergence rate for Tyler's *regularized* fixed-point iteration (5), assuming that the regularization parameter α is sufficiently large. To state their result, let $R \in (0,1)$ and assume $\lambda \in \mathbf{R}$ satisfies

$$\lambda > \max\left\{\frac{n}{m}\left(3+\frac{1}{R}\right)\left\|\sum_{i=1}^{m}\frac{x_ix_i^T}{\|x_i\|_2^2}\right\|_2 - 1, 0\right\} \tag{6}$$

where $\|\cdot\|_2$ with a matrix argument denotes the spectral norm. Then for $\alpha = \lambda/(1+\lambda) \in (0,1)$, the kth iterate of Tyler's regularized fixed-point iteration (5) satisfies

$$\|\Sigma_k - \Sigma_\alpha^\star\|_2 \le R^k\|\Sigma_0 - \Sigma_\alpha^\star\|_2,$$

where $\Sigma_\alpha^\star$ is the solution of the regularized ML estimation problem (4). It is important to note that this

result relies on very restrictive assumptions regarding the level of regularization. For example, in our simulations (see Appendix A.1 for details), the condition (6) yields a value of $\alpha \approx 0.985$, which effectively renders the influence of the observed samples on the estimation negligible. As a consequence, the solution to the regularized ML estimation problem (4) is close to the identity matrix, independent of the observed samples.

In other recent work, Franks and Moitra (2020) established a non-deterministic convergence guarantee for Tyler's fixed-point iteration under the assumption that the number of samples is sufficiently large. Their main result states that there exist constants $C > 0$ and $c > 0$ such that if the number of samples satisfies $m \geq Cn\log^2(n)$, then with probability at least $1-\exp(-mc/(n\log^2(n)))$, the kth iterate Σ_k of Tyler's fixed-point iteration (3) satisfies

$$\|I - (\Sigma^\star)^{1/2}\Sigma_k^{-1}(\Sigma^\star)^{1/2}\|_F \leq \epsilon$$

after

$$k = \mathcal{O}(\log\det\Sigma + n + \log(1/\epsilon))$$

iterations, where $\Sigma^\star$ is a solution of (1), Σ is the true shape matrix of the underlying elliptical distribution normalized to have trace one, and $\|\cdot\|_F$ denotes the Frobenius norm. The proof is based on advanced mathematical concepts such as operator scaling and quantum expanders, making it highly non-trivial and non-transparent. For example, it is difficult to gain any intuition of what affects how large the two constants C and c are.

2.2 Convex-concave procedure

The convex-concave procedure (Yuille and Rangarajan, 2001) is a general framework for solving optimization problems involving functions that are expressed as differences of convex functions. For a problem of the form

$$\text{minimize } f(Y) - g(Y) \tag{7}$$

where $f : \mathbf{S}^n \to \mathbf{R}$ is convex and $g : \mathbf{S}^n \to \mathbf{R}$ is convex and differentiable, CCP linearizes the second term around the current iterate $Y_k \in \mathbf{int}\,\mathbf{dom}\,g$ and computes the next iterate by solving the convex subproblem

$$Y_{k+1} \in \operatorname*{argmin}_Y\{f(Y) - (g(Y_k) + \langle\nabla g(Y_k), Y - Y_k\rangle)\}.$$

The function $Y \mapsto f(Y) - (g(Y_k) + \langle\nabla g(Y_k), Y - Y_k\rangle)$ that is minimized in each iteration is bounded from below as long as (7) has an optimal solution $Y^\star$, since for any $Y \in \mathbf{S}^n$ we have

$$\begin{aligned} f(Y) - (g(Y_k) + \langle\nabla g(Y_k), Y - Y_k\rangle) &\geq f(Y) - g(Y) \\ &\geq f(Y^\star) - g(Y^\star). \end{aligned}$$

Furthermore, CCP is a descent algorithm in the sense that $F(Y_{k+1}) \leq F(Y_k)$. Recently, it was shown in Yurtsever and Sra (2022); Faust et al. (2023) that CCP for solving (7) converges to a stationary point at a sublinear rate in the sense that for any $K \geq 1$, there exists $0 \leq k < K$ such that

$$f(Y_k) - f(Y) - \langle\nabla g(Y_k), Y_k - Y\rangle \leq \mathcal{O}(1/K)$$

for all $Y \in \mathbf{S}^n$. (In particular, this convergence guarantee applies to all the methods we present in this paper.) Note that while problem (7) appears to be unconstrained, any convex constraints can be incorporated into the domain of f, as f is not assumed to be differentiable. For more details and variations of CCP, we refer the reader to Lipp and Boyd (2016).

2.3 Relative smoothness

Relative smoothness is a generalization of the classical (Euclidean) smoothness assumption in optimization. Instead of requiring the gradient of a function $f : \mathbf{S}^n \to \mathbf{R}$ to be Lipschitz continuous with respect to the Frobenius norm, relative smoothness is defined with respect to a convex and differentiable *kernel function* $h : \mathbf{S}^n \to \mathbf{R}$. We say that a convex function f is L-smooth relative to h if there exists a constant $L > 0$ such that for all $Y, Z \in \mathbf{int}\,\mathbf{dom}\,f$,

$$f(Y) \leq f(Z) + \langle\nabla f(Z), Y - Z\rangle + LD_h(Y, Z), \tag{8}$$

where $D(Y, Z) = h(Y) - (h(Z) + \langle\nabla h(Z), Y - Z\rangle)$ is the *Bregman divergence* induced by h. In this paper, we will use the kernel $h(Y) = -(1/2)\log\det Y$, whose Bregman divergence is the Kullback–Leibler (KL) divergence

$$D_{\mathrm{KL}}(Y\|Z) = \frac{1}{2}\left(\mathbf{Tr}(Z^{-1}Y) + \log\det(ZY^{-1}) - n\right). \tag{9}$$

For more background on relative smoothness, we refer the reader to Lu et al. (2018).

2.4 Riemannian optimization

Riemannian optimization generalizes classical optimization in Euclidean spaces to curved spaces known as *Riemannian manifolds*. A particularly important example in practice is the manifold of positive definite matrices, denoted by $\mathbf{S}^n_{++}$. The key idea is to replace the standard Euclidean geometry on $\mathbf{S}^n_{++}$ — such as the trace inner product and the Frobenius norm — with a Riemannian geometry that better captures the structure and invariances of positive definite matrices. In this paper, we make use of several concepts from Riemannian optimization on $\mathbf{S}^n_{++}$, which we introduce below. For a more comprehensive treatment, we refer the reader to Boumal (2023).

The set of positive definite matrices forms a smooth Riemannian manifold when equipped with the so-called *affine-invariant metric*. Under this metric, the shortest curve between two matrices $X \in \mathbf{S}^n_{++}$ and $Y \in \mathbf{S}^n_{++}$ (also called the *geodesic*) is given by (Bhatia, 2007, eq. (6.11))

$$\gamma(t; X, Y) = X^{1/2}(X^{-1/2}YX^{-1/2})^t X^{1/2}, \quad t \in [0, 1],$$

and it remains entirely in $\mathbf{S}^n_{++}$. The corresponding *geodesic distance* between X and Y is (Bhatia, 2007, eq. (6.13))

$$D_{\text{geo}}(X, Y) = \|\log(X^{-1/2}YX^{-1/2})\|_F. \tag{10}$$

This metric is linear along the geodesic path in the sense that (Bhatia, 2007, eq. (6.12))

$$D_{\text{geo}}(\gamma(t; X, Y), X) = tD_{\text{geo}}(X, Y), \quad t \in [0, 1]. \tag{11}$$

Two elementary properties of the geodesic distance, following from (10) and the fact that $\log(A^{-1}) = -\log(A)$ for $A \in \mathbf{S}^n_{++}$, are

$$D_{\text{geo}}(X, Y) = D_{\text{geo}}(X^{-1}, Y^{-1})$$
$$D_{\text{geo}}(X, Y) = D_{\text{geo}}(Y, X).$$

The concept of a geodesic path allows us to generalize the notion of standard Euclidean convexity, which is defined along straight lines, to geodesic convexity, which is defined along geodesics. A function $f : \mathbf{S}^n_{++} \to \mathbf{R}$ is said to be *geodesically convex* if for all $X, Y \in \mathbf{S}^n_{++}$,

$$f(\gamma(t; X, Y)) \leq (1 - t)f(X) + tf(Y), \qquad t \in [0, 1].$$

Furthermore, f is *geodesically μ-strongly convex* on a geodesically convex set $\mathcal{A} \subseteq \mathbf{S}^n_{++}$ if there exists $\mu > 0$ such that for all $X, Y \in \mathcal{A}$ and $t \in [0, 1]$,

$$f(\gamma(t; X, Y)) \leq (1 - t)f(X) + tf(Y) - \frac{t(1 - t)\mu}{2} D_{\text{geo}}(X, Y).$$

In this paper, we make use of the fact that Tyler's cost function, *i.e.*, the objective function in (2), is geodesically convex (Wiesel and Zhang, 2015, Theorem 5.1). We also use the result that if $f(X)$ is geodesically convex, then the function $X \mapsto f(X^{-1})$ is also geodesically convex (Wiesel and Zhang, 2015, Lemma 1.17).

3 TYLER'S FIXED-POINT ITERATION AS CCP

In this section, we show that both Tyler fixed-point iteration and its regularized variant can be interpreted as instances of CCP. This interpretation forms the basis of our convergence analysis in §4. Furthermore, in Appendix A, we show how the CCP framework can be used to derive algorithms for computing Tyler's M-estimator under various structural constraints such as sparsity of the shape matrix or its inverse, or factor model structure.

To show that Tyler's fixed-point iteration (3) can be interpreted as an instance of CCP, we apply the change of variables $Y = \Sigma^{-1}$ in (2), resulting in the equivalent formulation

$$\text{minimize} \ -\log\det Y - (-n/m)\sum_{i=1}^m \log(x_i^T Y x_i) \tag{12}$$

with variable $Y \in \mathbf{S}^n$. Here the first term $Y \mapsto -\log\det Y$ is well-known to be convex, and the second term $Y \mapsto -(n/m)\sum_{i=1}^m \log(x_i^T Y x_i)$ is also convex since the ith term of the sum is the composition of the concave function $u \mapsto \log u$ with the linear function $Y \mapsto x_i^T Y x_i$, and is thus concave (Boyd and Vandenberghe, 2004, §3). Hence, the objective function of (12) is a difference of convex functions and can be minimized using CCP. Linearizing the second term around $Y_k \in \mathbf{S}^n_{++}$ yields the subproblem

$$\text{minimize} \ -\log\det Y + \mathbf{Tr}(S_k Y)$$

where

$$S_k = \frac{n}{m}\sum_{i=1}^m \frac{1}{x_i^T Y_k x_i} x_i x_i^T. \tag{13}$$

Setting the gradient of the objective function of this subproblem to zero yields $-Y^{-1} + S_k = 0$, so the subproblem has the analytical solution $Y_{k+1} = S_k^{-1}$, which is equivalent to Tyler's fixed-point iteration (3). This shows that Tyler's classical fixed-point iteration is simply *CCP applied to the underlying ML estimation problem* (2) *after a change of variables*. In §4, we use this simple observation to establish a long-sought deterministic global convergence rate for Tyler's fixed-point iteration. A similar argument (see Appendix A.2 for details) also shows that the regularized Tyler's fixed-point iteration (5) can be interpreted as CCP applied to the regularized ML estimation problem (4) after a change of variables.

4 CONVERGENCE ANALYSIS

In this section, we present our main technical contributions. First, we establish the first known deterministic global convergence rate for Tyler's fixed-point iteration (3). Second, we derive a new linear convergence rate for Tyler's regularized fixed-point iteration (5) that, unlike the previous result (6), holds for *any* level of regularization.

Our analysis combines the CCP perspective and relative smoothness with tools from Riemannian optimization. Combining relative smoothness with Riemannian

geometry requires us to navigate two distinct geometric frameworks: the relative smoothness component of our analysis operates within the geometry induced by the KL divergence (9), while the Riemannian geometry is based on the geodesic distance (10). In general, the KL divergence is neither upper nor lower bounded by the geodesic distance, which makes the analysis challenging. The following technical lemma, however, allows us to relate the KL divergence to the geodesic distance along geodesics between the iterates and optimal solutions. This relationship will prove sufficient for our purposes.

Lemma 1. *For any $Y, Z \in \mathbf{S}^n_{++}$ and any $t \in [0,1]$,*

$$D_{KL}(\gamma(t;Y,Z)\|Y) \leq C_{YZ} D_{geo}(\gamma(t;Y,Z),Y)^2,$$

where

$$C_{YZ} = \frac{a_{YZ} - \log a_{YZ} - 1}{2\log^2(a_{YZ})}$$

for any $a_{YZ} \in \mathbf{R}$ satisfying $a_{YZ} \geq \lambda_{\max}(Y^{-1}Z)$. (For $a_{YZ} = 1$, we define $C_{YZ} = \lim_{a\to 1}(a - \log a - 1)/(2\log^2(a)) = 1/4$.)

While we defer the proof of Lemma 1 to Appendix B, we emphasize that this lemma is a purely geometric result, independent of Tyler's M-estimator. The constant $C_{YZ} > 0$ in Lemma 1 grows slower than linearly with $\lambda_{\max}(Y^{-1}Z)$ (see Figure 1). In our analysis, Y will correspond to an iterate of Tyler's fixed-point iteration, and Z to an optimal solution of the corresponding ML estimation problem. The explicit form of the fixed-point iteration will allow us to control $\lambda_{\max}(Y^{-1}Z)$, ensuring that it does not exceed its initial value. Thus, if we interpret C_{YZ} as a measure of the discrepancy between the geometry induced by the KL divergence and the Riemannian geometry, Tyler's fixed-point iteration (very roughly speaking) keeps this discrepancy under control, which is crucial for our analysis. We believe that Lemma 1 might have applications beyond the scope of this paper, for analyzing other statistical algorithms acting on positive definite matrices that control the quantity $\lambda_{\max}(Y^{-1}Z)$.

Our next result hints at how relative smoothness enters our analysis and will be helpful to quantify the progress made in each CCP iteration.

Lemma 2. *The function $w(Y) = -\log(x^TYx)$ where $x \in \mathbf{R}^n$, $x \neq 0$ is fixed is 2-smooth relative to the logdet kernel $h(Y) = -(1/2)\log\det Y$.*

The proof of Lemma 2 is based on matrix calculus and the Cauchy-Schwarz inequality, and is given in Appendix B.

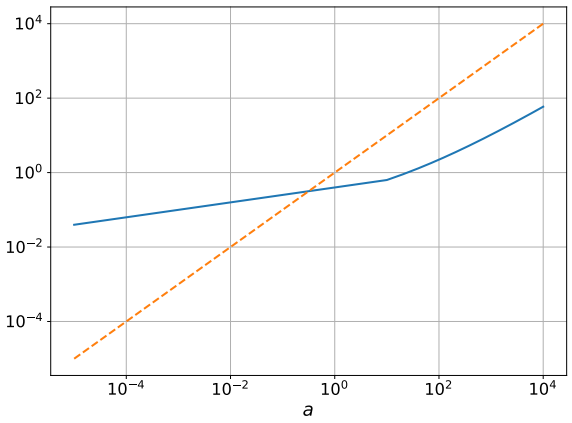

Figure 1: The function $a \mapsto (a - \log a - 1)/(2\log^2(a))$ (in blue) versus the linear function $a \mapsto a$ (in orange).

4.1 Analysis of Tyler's fixed-point iteration

Throughout our analysis we assume that Tyler's M-estimator exists, *i.e.*, that the ML estimation problem (2) has a solution. This is the case if the observed samples $x_1, \ldots, x_m$ are not concentrated on a lower-dimensional subspace in the sense that (Kent and Tyler, 1988)

$$m > \frac{N(L)}{\dim L} n,$$

for any linear subspace $L \subseteq \mathbf{R}^n$, where $N(L)$ is the number of points of $x_1, \ldots, x_m$ lying in L. If the samples are drawn from a continuous distribution, this condition is satisfied with probability one when $m > n$.

Preliminaries. We let $G(\Sigma)$ and $F(Y)$ be the objective functions of (2) before and after the variable change $Y = \Sigma^{-1}$,

$$G(\Sigma) = \log\det\Sigma + \frac{n}{m}\sum_{i=1}^m \log(x_i^T\Sigma^{-1}x_i)$$
$$F(Y) = -\log\det Y + \frac{n}{m}\sum_{i=1}^m \log(x_i^TYx_i).$$

We can write $F(Y)$ as a difference of convex functions as $F(Y) = f(Y) - g(Y)$, where

$$f(Y) = -\log\det Y, \qquad g(Y) = -\frac{n}{m}\sum_{i=1}^m \log(x_i^TYx_i).$$

From Lemma 2 it follows that $g(Y)$ is L-smooth relative to $h(Y) = -(1/2)\log\det Y$ with $L = 2n$.

Tyler's M-estimator is scale-invariant in the sense that if $\Sigma^\star$ is a solution to (2), then $t\Sigma^\star$ is also a solution to (2) for any $t > 0$. (The underlying reason is that

$G(t\Sigma) = G(\Sigma)$ for any $t > 0$.) To obtain a unique estimator, it is common to normalize it to have trace n (Wiesel and Zhang, 2015). In the rest of the paper, we let $\Sigma^\star \in \mathbf{S}^n_{++}$ be the unique solution to (2) with trace n, and we let $Y^\star = (\Sigma^\star)^{-1}$. For $k \geq 0$, we let $Y_k = \Sigma_k^{-1}$ be the inverse of the kth iterate of Tyler's fixed-point iteration (3). While we assume that we initialize the fixed-point iteration at $\Sigma_0 = I$, our analysis easily extends to any initialization. Finally, we introduce the following variant of the initial sublevel set:

$$\mathcal{L}_0^G = \{\Sigma \in \mathbf{S}^n_{++} \mid b_0\Sigma^\star \preceq \Sigma \preceq a_0\Sigma^\star,\ G(\Sigma) \leq G(\Sigma_0)\}$$

where $a_0 = \lambda_{\max}((\Sigma^\star)^{-1})$, $b_0 = \lambda_{\min}((\Sigma^\star)^{-1})$, and $\preceq$ denotes the Loewner order on symmetric matrices. The corresponding set in the variable $Y = \Sigma^{-1}$ is

$$\mathcal{L}_0^F = \{Y \in \mathbf{S}^n_{++} \mid (1/a_0)Y^\star \preceq Y \preceq (1/b_0)Y^\star, \\ F(Y) \leq F(Y_0)\}.$$

Convergence result. We begin with a lemma that establishes the monotonicity of the largest and smallest eigenvalue of the matrix $Y_k^{-1}Y^\star$ along the iterates (Y_k) of Tyler's fixed-point iteration. This lemma (with proof given in Appendix C) is crucial for our analysis, as it allows us to prove that the iterates remain in $\mathcal{L}_0^F$.

Lemma 3. *For any $k \geq 0$, we have $\lambda_{\max}(Y_{k+1}^{-1}Y^\star) \leq \lambda_{\max}(Y_k^{-1}Y^\star)$ and $\lambda_{\min}(Y_{k+1}^{-1}Y^\star) \geq \lambda_{\min}(Y_k^{-1}Y^\star)$. Furthermore, $Y_k \in \mathcal{L}_0^F$.*

Next, we specialize Lemma 1 to upper bound the KL divergence along the geodesic between Y_k and $Y^\star$ in terms of the geodesic distance. This provides the essential bridge connecting the KL geometry of relative smoothness to the Riemannian geometry.

Lemma 4. *For any $k \geq 0$ and $t \in [0,1]$, we have*

$$D_{KL}(\gamma(t; Y_k, Y^\star)\|Y_k) \leq C D_{geo}(\gamma(t; Y_k, Y^\star), Y_k)^2,$$

where

$$C = \frac{a_0 - \log a_0 - 1}{2\log^2(a_0)}. \tag{14}$$

Proof. The result follows from combining Lemma 3, which implies $\lambda_{\max}(Y_k^{-1}Y^\star) \leq a_0$, with Lemma 1 using $Y = Y_k$ and $Z = Y^\star$. □

We point out that the case $a_0 = 1$ only happens if $\Sigma^\star = I$, since $a_0 = 1/\lambda_{\min}(\Sigma^\star)$ and $\mathbf{Tr}(\Sigma^\star) = n$. In this case Tyler's fixed-point iteration terminates immediately, since $\Sigma_0 = I$ is already optimal and satisfies (1).

Now we are ready to derive the first known deterministic convergence rate for Tyler's fixed-point iteration. The proof borrows ideas from an argument by Nesterov (2013, Theorem 4) that was also used by Mairal (2013, Proposition 2.2).

Theorem 1. *Let $R > 0$ be such that $D_{geo}(\Sigma, \Sigma^\star) \leq R$ for all $\Sigma \in \mathcal{L}_0^G$. Tyler's fixed point-iteration (3) for solving (2) satisfies*

$$G(\Sigma_k) - G(\Sigma^\star) \leq \frac{4LCR^2}{k+2}, \quad k \geq 1. \tag{15}$$

Here $L = 2n$ is the smoothness constant of $g(Y)$ relative to the logdet kernel, and C is defined in (14).

Proof. For any $Y \in \mathbf{S}^n_{++}$ we have

$$\begin{aligned} F(Y_{k+1}) &= f(Y_{k+1}) - g(Y_{k+1}) \\ &\leq f(Y_{k+1}) - (g(Y_k) + \langle \nabla g(Y_k), Y_{k+1} - Y_k\rangle) \\ &\leq f(Y) - (g(Y_k) + \langle \nabla g(Y_k), Y - Y_k\rangle) \\ &= F(Y) + g(Y) - (g(Y_k) + \langle \nabla g(Y_k), Y - Y_k\rangle) \\ &\leq F(Y) + L D_{\mathrm{KL}}(Y\|Y_k). \end{aligned}$$

Here, the first inequality is a consequence of the convexity of g, the second follows from the definition of Y_{k+1} as a minimizer of the CCP subproblem, and the last inequality uses the relative smoothness of g. Since this inequality holds for any $Y \in \mathbf{S}^n_{++}$, it particularly holds on the geodesic $\gamma(t; Y_k, Y^\star)$. Writing $\gamma_t = \gamma(t; Y_k, Y^\star)$ for brevity, we have

$$\begin{aligned} F(Y_{k+1}) &\leq \min_{t\in[0,1]} \{F(\gamma_t) + L D_{\mathrm{KL}}(\gamma_t\|Y_k)\} \\ &\leq \min_{t\in[0,1]} \left\{F(\gamma_t) + LC D_{\mathrm{geo}}(\gamma_t, Y_k)^2\right\} \\ &= \min_{t\in[0,1]} \left\{F(\gamma_t) + LCt^2 D_{\mathrm{geo}}(Y^\star, Y_k)^2\right\} \\ &\leq \min_{t\in[0,1]} \big\{(1-t)F(Y_k) + tF(Y^\star) \\ &\qquad + LCt^2 D_{\mathrm{geo}}(Y^\star, Y_k)^2\big\}. \end{aligned}$$

Here, the second inequality is Lemma 4, the equality follows from (11), and the last inequality uses geodesic convexity of F. Note that

$$\begin{aligned} D_{\mathrm{geo}}(Y^\star, Y_k) &= D_{\mathrm{geo}}(Y_k, Y^\star) = D_{\mathrm{geo}}(Y_k^{-1}, (Y^\star)^{-1}) \\ &= D_{\mathrm{geo}}(Y_k^{-1}, \Sigma^\star). \end{aligned}$$

Since $Y_k \in \mathcal{L}_0^F$ it holds that $Y_k^{-1} \in \mathcal{L}_0^G$, implying that $D_{\mathrm{geo}}(Y_k^{-1}, \Sigma^\star) \leq R$. Thus,

$$\begin{aligned} F(Y_{k+1}) \leq \min_{t\in[0,1]} \big\{&(1-t)F(Y_k) + tF(Y^\star) \\ &+ LCR^2t^2\big\}. \end{aligned}$$

Subtracting $F(Y^\star)$ from both sides shows that

$$\Delta_{k+1} \leq \min_{t\in[0,1]} \left\{(1-t)\Delta_k + LCR^2t^2\right\},$$

where $\Delta_k = F(Y_k) - F(Y^\star) = G(\Sigma_k) - G(\Sigma^\star)$. Turning this inequality into a convergence rate requires some algebra.

If $\Delta_k \geq 2LCR^2$, then the minimum is attained at $t = 1$ and we have $\Delta_{k+1} \leq LCR^2$. If $\Delta_k \leq 2LCR^2$, then the minimum is attained at $t = \Delta_k/(2LCR^2)$ and we have

$$\Delta_{k+1} \leq \Delta_k \left(1 - \frac{\Delta_k}{4LCR^2}\right).$$

Hence, $\Delta_{k+1}^{-1} \geq \Delta_k^{-1}(1 - \Delta_k/(4LCR^2))^{-1} \geq \Delta_k^{-1} + 1/(4LCR^2)$, where we have used the fact that $(1 - x)^{-1} \geq 1 + x$ for $x \in (0, 1)$.

To summarize, if $\Delta_0 \geq 2LCR^2$, then $\Delta_1 \leq LCR^2$, and thus $\Delta_{k+1}^{-1} \geq \Delta_1^{-1} + k/(4LCR^2) \geq (k + 4)/(4LCR^2)$. Otherwise, we have $\Delta_{k+1}^{-1} \geq \Delta_0^{-1} + (k+1)/(4LCR^2) \geq (k + 3)/(4LCR^2)$. The desired result now follows. □

4.2 Analysis of Tyler's regularized fixed-point iteration

We now turn to the regularized version (5) of Tyler's fixed-point iteration. While the overall analysis parallels that of the unregularized case, there is one important distinction: in the unregularized case, we showed that the iterates remain within the *bounded* set $\mathcal{L}_0^F$. In contrast, for the regularized case, we are not able to establish a similar result. However, we are able to prove that the iterates remain within a geodesically convex set on which the regularized objective function is strongly geodesically convex. This enables us to derive a linear convergence rate.

Throughout our analysis we assume that Tyler's regularized M-estimator exists, *i.e.*, that the regularized ML estimation problem (4) has a solution. This is the case if the observed samples $x_1, \ldots, x_m$ satisfy (Ollila and Tyler, 2014)

$$m > (1 - \alpha)\frac{N(L)}{\dim L}n,$$

for any linear subspace $L \subseteq \mathbf{R}^n$. When this condition is satisfied, the solution of (4) is unique.

Preliminaries. We let $G_\alpha(\Sigma)$ and $F_\alpha(Y)$ be the objective functions of (4) before and after the variable change $Y = \Sigma^{-1}$,

$$\begin{aligned} G_\alpha(\Sigma) &= \log\det\Sigma + \frac{(1-\alpha)n}{m}\sum_{i=1}^m \log(x_i^T\Sigma^{-1}x_i) \\ &\quad + \alpha\,\mathbf{Tr}(\Sigma^{-1}) \\ F_\alpha(Y) &= -\log\det Y + \frac{(1-\alpha)n}{m}\sum_{i=1}^m \log(x_i^T Y x_i) \\ &\quad + \alpha\,\mathbf{Tr}(Y). \end{aligned}$$

We can write $F_\alpha(Y)$ as a difference of convex functions as $F_\alpha(Y) = f_\alpha(Y) - g_\alpha(Y)$, where

$$\begin{aligned} f_\alpha(Y) &= -\log\det Y + \alpha\,\mathbf{Tr}(Y) \\ g_\alpha(Y) &= -\frac{(1-\alpha)n}{m}\sum_{i=1}^m \log(x_i^T Y x_i). \end{aligned}$$

From Lemma 2 it follows that $g_\alpha(Y)$ is L_α-smooth relative to $h(Y) = -(1/2)\log\det Y$ with $L_\alpha = 2n(1 - \alpha)$.

Convergence result. Our first result is similar to Lemma 3 and establishes that the largest eigenvalue of the matrix $Y_k^{-1}Y_\alpha^\star$ is bounded from above. However, the proof is different from that of Lemma 3 and is given in Appendix D. In the rest of §4.2, we let Σ_k be the kth iterate of Tyler's regularized fixed-point iteration (5), where the dependence on a given $\alpha \in (0, 1)$ is omitted. We use the initialization $\Sigma_0 = I$ and we let $\Sigma_\alpha^\star$ be the unique solution to the regularized ML estimation problem (4), $Y_\alpha^\star = (\Sigma_\alpha^\star)^{-1}$, and $Y_k = \Sigma_k^{-1}$ for $k \geq 0$.

Lemma 5. *For any $k \geq 0$, we have*

$$\lambda_{\max}(Y_k^{-1}Y_\alpha^\star) \leq \lambda_{\max}((\Sigma_\alpha^\star)^{-1}).$$

Our next result is similar to Lemma 4 and has an identical proof.

Lemma 6. *For $k \geq 0$ and $t \in [0, 1]$, we have*

$$D_{KL}(\gamma(t; Y_k, Y_\alpha^\star)\|Y_k) \leq C_\alpha D_{geo}(\gamma(t; Y_k, Y_\alpha^\star), Y_k)^2,$$

where

$$C_\alpha = \frac{\lambda_{\max}((\Sigma_\alpha^\star)^{-1}) - \log\lambda_{\max}((\Sigma_\alpha^\star)^{-1}) - 1}{2\log^2(\lambda_{\max}((\Sigma_\alpha^\star)^{-1}))}. \quad (16)$$

The next results (with proofs given in Appendix D) show that the regularized objective function is geodesically strongly convex on a geodesically convex set that contains the iterates. Let $\kappa^\star = \lambda_{\max}(\Sigma_\alpha^\star)/\lambda_{\min}(\Sigma_\alpha^\star)$ be the condition number of the optimal solution and define the set

$$\mathcal{A} = \left\{Y \in \mathbf{S}_{++}^n \mid \lambda_{\min}(Y) \geq 1/\kappa^\star\right\}.$$

Lemma 7. *It holds that $Y_\alpha^\star \in \mathcal{A}$, and $Y_k \in \mathcal{A}$ for $k \geq 0$.*

Lemma 8. *The set $\mathcal{A}$ is geodesically convex.*

Lemma 9. *The function $r(Y) = \mathbf{Tr}(Y)$, $\mathbf{dom}\, r = \mathbf{S}_{++}^n$, is geodesically μ-strongly convex on $\mathcal{A}$ with $\mu = 1/\kappa^\star$.*

We are now ready to prove that Tyler's regularized fixed-point iteration converges linearly.

Theorem 2. *Tyler's regularized fixed-point iteration* (5) *for solving* (4) *satisfies that for any* $k \geq 1$*,*

$$G_\alpha(\Sigma_k) - G_\alpha(\Sigma_\alpha^\star) \leq \beta^k (G_\alpha(\Sigma_0) - G_\alpha(\Sigma_\alpha^\star)),$$

where

$$\beta = \begin{cases} 1/2 & \text{if } \alpha \geq 4\kappa^\star L_\alpha C_\alpha, \\ 1 - \alpha/(8\kappa^\star L_\alpha C_\alpha) & \text{otherwise.} \end{cases}$$

Here, $L_\alpha = 2n(1-\alpha)$ *is the smoothness constant of* $g_\alpha(Y)$ *relative to the logdet kernel, and* C_α *is defined in* (16)*.*

Proof. By mimicking the start of the proof of Theorem 1, we obtain

$$\begin{aligned} F_\alpha(Y_{k+1}) \leq \min_{t\in[0,1]} \{&(1-t)F_\alpha(Y_k) + tF_\alpha(Y_\alpha^\star) \\ &+ L_\alpha C_\alpha t^2 D_{\text{geo}}(Y_\alpha^\star, Y_k)^2\}. \end{aligned} \tag{17}$$

From Lemma 9 it follows that F_α is geodesically $(\alpha\mu)$-strongly convex on the geodesically convex set $\mathcal{A}$ that contains both Y_k and $Y_\alpha^\star$. Hence, (Boumal, 2023, Theorem 11.21)

$$D_{\text{geo}}(Y_\alpha^\star, Y_k)^2 \leq \frac{2}{\alpha\mu}(F_\alpha(Y_k) - F_\alpha(Y_\alpha^\star)).$$

By combining this inequality with (17) and subtracting $F_\alpha(Y_\alpha^\star)$ from both sides, we obtain

$$\Delta_{k+1} \leq \min_{t\in[0,1]} \left\{1 - t + \frac{2L_\alpha C_\alpha}{\alpha\mu} t^2\right\} \Delta_k$$

where $\Delta_k = F_\alpha(Y_k) - F_\alpha(Y_\alpha^\star) = G_\alpha(\Sigma_k) - G_\alpha(\Sigma_\alpha^\star)$. If $\alpha\mu \geq 4L_\alpha C_\alpha$, then the minimum is attained at $t = 1$ and we have

$$\Delta_{k+1} \leq \frac{2L_\alpha C_\alpha}{\alpha\mu}\Delta_k \leq \frac{1}{2}\Delta_k.$$

Otherwise, the minimum is attained for the value $t = \alpha\mu/(4L_\alpha C_\alpha) \in (0, 1]$, and we have

$$\Delta_{k+1} \leq \left(1 - \frac{\alpha\mu}{8L_\alpha C_\alpha}\right)\Delta_k.$$

This concludes the proof. □

5 CONCLUSIONS

We have established the first deterministic global convergence rate for Tyler's renowned fixed-point iteration, thereby closing a significant gap in the literature. We have also derived the first linear convergence rate for its regularized variant that holds uniformly across all levels of regularization (assuming Tyler's regularized M-estimator exists). Our analysis leverages an interpretation of these algorithms as instances of the convex-concave procedure, combined with tools from relative smoothness and Riemannian optimization. To the best of our knowledge, this work is the first to analyze a Euclidean algorithm, the convex-concave procedure, using concepts based on relative smoothness and Riemannian optimization. We expect that this approach will find broader applicability, such as in the analysis of algorithms for computing Brascamp–Lieb constants (Garg et al., 2017) through the lens of geodesic convexity (Sra et al., 2018).

Acknowledgements

We thank Kasper Johansson for helpful discussions on Riemannian optimization. We also thank four anonymous reviewers for their feedback.

References

Almeida, Y., da Cruz Neto, J., Oliveira, P., and Souza, J. (2020). A Modified Proximal Point Method for DC Functions on Hadamard Manifolds. *Computational Optimization and Applications*, 76:649–673.

ApS, M. (2025). MOSEK modeling cookbook.

Aubin-Frankowski, P.-C., Korba, A., and Léger, F. (2022). Mirror Descent with Relative Smoothness in Measure Spaces, with Application to Sinkhorn and EM. *Advances in Neural Information Processing Systems*, 35:17263–17275.

Bartholomew, D., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. John Wiley & Sons.

Bergmann, R., Ferreira, O., Santos, E., and Souza, J. (2023). The Difference of Convex Algorithm on Hadamard Manifolds.

Bhatia, R. (2007). *Positive Definite Matrices*. Princeton University Press, Princeton.

Bien, J. and Tibshirani, R. (2011). Sparse Estimation of a Covariance Matrix. *Biometrika*, 98(4):807–820.

Bouchard, F., Breloy, A., Ginolhac, G., Renaux, A., and Pascal, F. (2021). A Riemannian Framework for Low-Rank Structured Elliptical Models. *IEEE Transactions on Signal Processing*, 69:1185–1199.

Boumal, N. (2023). *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Cederberg, D. (2026). T-Rex: Fitting a Robust Factor Model via Expectation-Maximization. *IEEE Transactions on Signal Processing*.

Chaudhuri, S., Drton, M., and Richardson, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika*, 94(1):199–216.

Chen, Y., Wiesel, A., and Hero, A. (2011). Robust Shrinkage Estimation of High-Dimensional Covariance Matrices. *IEEE Transactions on Signal Processing*, 59(9):4097–4107.

Cheng, A. and Weber, M. (2024). Structured Regularization for Constrained Optimization on the SPD Manifold. *arXiv preprint arXiv:2410.09660*.

Dahl, J., Vandenberghe, L., and Roychowdhury, V. (2008). Covariance Selection for Nonchordal Graphs via Chordal Embedding. *Optimization Methods & Software*, 23(4):501–520.

Danon, L. and Garber, D. (2022). Frank-Wolfe-based Algorithms for Approximating Tyler's M-Estimator. *Advances in Neural Information Processing Systems*, 35:3637–3648.

Delmas, J., El Korso, M., Fortunati, S., and Pascal, F. (2024). *Elliptically Symmetric Distributions in Signal Processing and Machine Learning*, volume 1. Springer.

Dempster, A. (1972). Covariance Selection. *Biometrics*, pages 157–175.

Diamond, S. and Boyd, S. (2016). CVXPY: A Python-embedded Modeling Language for Convex Optimization. *Journal of Machine Learning Research*, 17(83):1–5.

Fatima, G., Babu, P., and Stoica, P. (2024). Two New Algorithms for Maximum Likelihood Estimation of Sparse Covariance Matrices with Applications to Graphical Modeling. *IEEE Transactions on Signal Processing*, 72:958–971.

Faust, O., Fawzi, H., and Saunderson, J. (2023). A Bregman Divergence View on the Difference-of-Convex Algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 3427–3439. PMLR.

Finegold, M. and Drton, M. (2011). Robust Graphical Modeling of Gene Networks Using Classical and Alternative t-Distributions. *The Annals of Applied Statistics*, 5(2A):1057–1080.

Franks, W. and Moitra, A. (2020). Rigorous Guarantees for Tyler's M-Estimator via Quantum Expansion. In *Conference on Learning Theory*, pages 1601–1632. PMLR.

Garg, A., Gurvits, L., Oliveira, R., and Wigderson, A. (2017). Algorithmic and Optimization Aspects of Brascamp-Lieb Inequalities, via Operator Scaling. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 397–409.

Goes, J., Lerman, G., and Nadler, B. (2020). Robust Sparse Covariance Estimation by Thresholding Tyler's M-Estimator. *The Annals of Statistics*, 48(1):86–110.

Kauermann, G. (1996). On a Dualization of Graphical Gaussian Models. *Scandinavian Journal of Statistics*, pages 105–116.

Kent, J. and Tyler, D. (1988). Maximum Likelihood Estimation for the Wrapped Cauchy Distribution. *Journal of Applied Statistics*, 15(2):247–254.

Kunstner, F., Kumar, R., and Schmidt, M. (2021). Homeomorphic-Invariance of EM: Non-Asymptotic Convergence in KL Divergence for Exponential Families via Mirror Descent. In *International Conference on Artificial Intelligence and Statistics*, pages 3295–3303. PMLR.

Lauritzen, S. (1996). *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. Oxford University Press.

Lipp, T. and Boyd, S. (2016). Variations and Extension of the Convex–Concave Procedure. *Optimization and Engineering*, 17:263–287.

Lu, H., Freund, R., and Nesterov, Y. (2018). Relatively Smooth Convex Optimization by First-Order Methods, and Applications. *SIAM Journal on Optimization*, 28(1):333–354.

Mairal, J. (2013). Optimization with First-Order Surrogate Functions. In *International Conference on Machine Learning*, pages 783–791. PMLR.

Mishchenko, K. (2019). Sinkhorn Algorithm as a Special Case of Stochastic Mirror Descent. *arXiv preprint arXiv:1909.06918*.

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

Ollila, E., Palomar, D., and Pascal, F. (2021). Shrinking the Eigenvalues of M-Estimators of Covariance Matrix. *IEEE Transactions on Signal Processing*, 69:256–269.

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

Pascal, F., Chitour, Y., and Quek, Y. (2014). Generalized Robust Shrinkage Estimator and Its Application to STAP Detection Problem. *IEEE Transactions on Signal Processing*, 62(21):5640–5651.

Soloveychik, I. and Wiesel, A. (2014). Tyler's Covariance Matrix Estimator in Elliptical Models with Convex Structure. *IEEE Transactions on Signal Processing*, 62(20):5251–5259.

Souza, J. and Oliveira, P. (2015). A Proximal Point Algorithm for DC Functions on Hadamard Manifolds. *Journal of Global Optimization*, 63:797–810.

Sra, S., Vishnoi, N., and Yildiz, O. (2018). On Geodesically Convex Formulations for the Brascamp-Lieb Constant. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*, pages 25–1. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Sun, Y., Babu, P., and Palomar, D. (2016). Robust Estimation of Structured Covariance Matrix for Heavy-Tailed Elliptical Distributions. *IEEE Transactions on Signal Processing*, 64(14):3576–3590.

Sun, Y., Babu, P., and Palomar, D. P. (2014). Regularized Tyler's Scatter Estimator: Existence, Uniqueness, and Algorithms. *IEEE Transactions on Signal Processing*, 62(19):5143–5156.

Tyler, D. (1987). A Distribution-Free M-Estimator of Multivariate Scatter. *The Annals of Statistics*, 15(1):234–251.

Weber, M. and Sra, S. (2023). Global Optimality for Euclidean CCCP under Riemannian Convexity. In *International Conference on Machine Learning*, pages 36790–36803. PMLR.

Wiesel, A. and Zhang, T. (2015). Structured Robust Covariance Estimation. *Foundations and Trends® in Signal Processing*, 8(3):127–216.

Xu, J. and Lange, K. (2022). A Proximal Distance Algorithm for Likelihood-Based Sparse Covariance Estimation. *Biometrika*, 109(4):1047–1066.

Yuille, A. and Rangarajan, A. (2001). The Concave-Convex Procedure (CCCP). *Advances in Neural Information Processing Systems*, 14.

Yurtsever, A. and Sra, S. (2022). CCCP is Frank-Wolfe in disguise. *Advances in Neural Information Processing Systems*, 35:35352–35364.

Zhang, T. and Wiesel, A. (2016). Automatic Diagonal Loading for Tyler's Robust Covariance Estimator. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pages 1–5.

Zhang, T., Wiesel, A., and Greco, M. S. (2013). Multivariate Generalized Gaussian Distribution: Convexity and Graphical Models. *IEEE Transactions on Signal Processing*, 61(16):4141–4148.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Not Applicable.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes.
   (b) Complete proofs of all theoretical results. Yes.
   (c) Clear explanations of any assumptions. Yes.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results. Not Applicable.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Not Applicable.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Not Applicable.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Not Applicable.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable.
   (b) The license information of the assets, if applicable. Not Applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.
   (d) Information about consent from data providers/curators. Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

ORGANIZATION OF THE SUPPLEMENTARY MATERIAL

- Section A: Structured Tyler's M-estimator
- Section B: Omitted proofs - General
- Section C: Omitted proofs - Tyler's fixed-point iteration
- Section D: Omitted proofs - Tyler's regularized fixed-point iteration

A STRUCTURED TYLER'S M-ESTIMATOR

A.1 Regularization parameter simulations

To better understand how restrictive condition (6) is for the choice of regularization parameter in Tyler's regularized fixed-point iteration, we conduct an experiment where we fix $R = 0.99$ and $n = 50$, and then vary the number of samples $m \in \{25, 50, 75, \dots, 500\}$. As ground truth, we use the sample covariance matrix computed from the returns of n assets in the S&P 500 index over a two-year period. Using this covariance matrix, we generate Gaussian data with zero mean and compute the regularization parameter as $\alpha = \lambda/(1+\lambda)$ where we pick λ according to (6). For each value of m, we repeat the experiment 50 times and report the average value of α. The result, shown in Figure 2, indicates that condition (6) leads to the value of the regularization parameter being very close to one, effectively making the shape matrix estimate almost entirely independent of the observed data. The code for the simulation is available at `https://github.com/dance858/Tyler-CCP-analysis`.

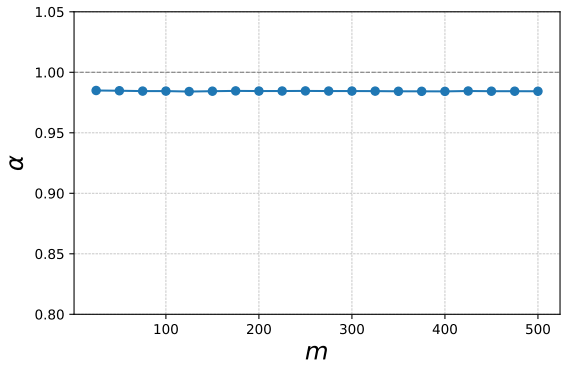

Figure 2: The average value of the regularization parameter α as a function of the number of samples m.

A.2 Regularized Tyler's fixed-point iteration as CCP

To show that Tyler's regularized fixed-point iteration (5) can be interpreted as an instance of CCP, we make the variable change $Y = \Sigma^{-1}$ in (4). The objective function after the variable change is of the form $f_\alpha(Y) - g_\alpha(Y)$, where

$$f_\alpha(Y) = -\log\det Y + \alpha\,\mathbf{Tr}(Y), \qquad g_\alpha(Y) = -\frac{(1-\alpha)n}{m}\sum_{i=1}^{m}\log(x_i^T Y x_i).$$

Both $f_\alpha(Y)$ and $g_\alpha(Y)$ are convex, so the objective function is a difference of convex functions and CCP-compatible. Linearizing the concave term $g_\alpha(Y)$ around $Y_k \in \mathbf{S}^n_{++}$ yields the subproblem

$$\text{minimize} \;\; -\log\det Y + \alpha\,\mathbf{Tr}(((1-\alpha)S_k + \alpha I)Y)$$

with variable $Y \in \mathbf{S}^n_{++}$, where $S_k = (n/m)\sum_{i=1}^m (1/x_i^T Y_k x_i) x_i x_i^T$. This subproblem has the analytical solution $Y_{k+1} = (1-\alpha)S_k + \alpha I$, which is the update step in Tyler's regularized fixed-point iteration (5).

A.3 Tyler's M-estimator with known inverse sparsity pattern

The sparsity pattern of the inverse covariance matrix of a Gaussian random vector determines the conditional independence structure between its components, with a zero in position (i, j) indicating that components i and j are conditionally independent given the rest (Dempster, 1972). This property forms the foundation of *Gaussian graphical models* (Lauritzen, 1996, Chapter 7), and the Gaussian ML estimate of the covariance matrix conforming to a known conditional independence structure can be obtained by solving

$$\begin{array}{ll} \text{minimize} & \log\det\Sigma + \mathbf{Tr}(S\Sigma^{-1}) \\ \text{subject to} & (\Sigma^{-1})_{ij} = 0, \qquad (i,j) \notin V, \end{array} \tag{18}$$

where $\Sigma \in \mathbf{S}^n$ is the variable, and the problem data are the sample covariance matrix $S \in \mathbf{S}^n_+$ and the set V of components that are allowed to be conditionally dependent given the rest. This problem is convex in Σ^{-1} and specialized algorithms for solving it have been proposed (Dahl et al., 2008).

When the observed data is either more heavy-tailed than the Gaussian distribution or contains outliers, we can obtain a more robust estimator by incorporating the known inverse sparsity pattern into Tyler's M-estimator. If the data is not Gaussian, a zero in position (i, j) of the inverse covariance matrix does not mean that components i and j are conditionally independent given the rest. Nevertheless, it remains popular to impose sparsity in the inverse covariance matrix as a proxy for conditional independence (Finegold and Drton, 2011). Furthermore, incorporating sparsity in the inverse covariance matrix reduces the effective number of parameters to be estimated.

To incorporate a known inverse sparsity pattern into Tyler's M-estimator, we solve

$$\begin{array}{ll} \text{minimize} & \log\det\Sigma + (n/m)\sum_{i=1}^m \log(x_i^T\Sigma^{-1}x_i) \\ \text{subject to} & \Sigma^{-1}_{ij} = 0, \qquad (i,j) \notin V. \end{array}$$

After the variable change $Y = \Sigma^{-1}$ we obtain the equivalent formulation

$$\begin{array}{ll} \text{minimize} & -\log\det Y - (-n/m)\sum_{i=1}^m \log(x_i^T Y x_i) \\ \text{subject to} & Y_{ij} = 0, \qquad (i,j) \notin V. \end{array}$$

In this formulation the constraints are linear, and the objective function is a difference of convex functions and hence CCP-compatible. Linearizing the second term of the objective around $Y_k \in \mathbf{S}^n_{++}$ yields the subproblem

$$\begin{array}{ll} \text{minimize} & -\log\det Y + \mathbf{Tr}(S_k Y) \\ \text{subject to} & Y_{ij} = 0, \qquad (i,j) \notin V, \end{array} \tag{19}$$

where S_k is defined in (13). This is the same as the Gaussian covariance selection problem (18), but with the sample covariance matrix S replaced by the weighted sample covariance matrix S_k.

A.4 Tyler's M-estimator with known sparsity pattern

The sparsity pattern of the covariance matrix itself of a Gaussian random vector determines the marginal independence structure between its components, with a zero in position (i, j) indicating that components i and j are marginally independent. This structure naturally gives rise to a *covariance graph* (Kauermann, 1996), a type of graphical model in which nodes represent variables and edges represent marginal dependencies. Imposing sparsity in the covariance matrix not only facilitates interpretability through graph structure but also reduces the number of parameters to estimate. Early contributions to sparse covariance estimation include Chaudhuri et al. (2007); Bien and Tibshirani (2011), while more recent references are Xu and Lange (2022); Fatima et al. (2024).

To incorporate a known sparsity pattern into Tyler's M-estimator, we solve

$$\begin{array}{ll} \text{minimize} & \log\det\Sigma + (n/m)\sum_{i=1}^m \log(x_i^T\Sigma^{-1}x_i) \\ \text{subject to} & \Sigma_{ij} = 0, \qquad (i,j) \notin V, \end{array}$$

with variable $\Sigma \in \mathbf{S}^n$, and where V is the set of indices that are allowed to be nonzero. The first term is well-known to be concave (Boyd and Vandenberghe, 2004, Chapter 3) and the second term can be shown to be

convex (see Lemma 10 below). Hence, the objective function is a difference of convex functions and the problem is CCP-compatible. By linearizing the concave term we get the subproblem

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}(\Sigma_k^{-1}\Sigma) + (n/m)\sum_{i=1}^m \log(x_i^T\Sigma^{-1}x_i) \\ \text{subject to} & \Sigma_{ij} = 0, \qquad (i,j) \notin V, \end{array} \tag{20}$$

with variable $\Sigma \in \mathbf{S}^n$.

An important practical feature of CCP is that the subproblem is convex and can typically be specified easily and solved using domain-specific languages (DSLs) for convex optimization, such as CVXPY (Diamond and Boyd, 2016). DSLs parse the high-level description of a convex optimization problem and transform it into a standard form suitable for numerical solvers that handle generic problem classes, such as second-order cone programs or semidefinite programs (ApS, 2025). To specify a problem in a DSL, the objective and constraints must be constructed by combining a set of *atomic functions* according to a set of *composition rules*. Although the second term in the objective function of (20) is convex, it is not immediately clear how to express it in a way that complies with the disciplined convex programming (DCP) rules required by such DSLs. To formulate it in a DCP-compliant manner, we use the fact that for fixed $\Sigma \in \mathbf{S}^n_{++}$ and fixed $x \in \mathbf{R}^n$, the value of $\log(x^T\Sigma^{-1}x)$ is equal to the optimal value of the convex problem

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \begin{bmatrix} 1 & vx^T \\ vx & \Sigma \end{bmatrix} \succeq 0 \\ & v \geq e^{-t/2}, \end{array}$$

with variables $t \in \mathbf{R}$ and $v \in \mathbf{R}$ (see Lemma 11 below.) This allows us to write the subproblem (20) as

$$\begin{array}{lll} \text{minimize} & \mathbf{Tr}(\Sigma_k^{-1}\Sigma) + (n/m)\sum_{i=1}^m t_i & \\ \text{subject to} & \Sigma_{ij} = 0, & (i,j) \notin V, \\ & \begin{bmatrix} 1 & v_i x_i^T \\ v_i x_i & \Sigma \end{bmatrix} \succeq 0, & i = 1, \ldots, m, \\ & v_i \geq e^{-t_i/2}, & i = 1, \ldots, m, \end{array}$$

with variables $\Sigma \in \mathbf{S}^n$, $t \in \mathbf{R}^m$, and $v \in \mathbf{R}^m$. This problem is convex and DCP-compliant, so it can be easily specified and solved using CVXPY.

Lemma 10. *Let $x \neq 0$ be fixed. The function $q(\Sigma) = \log(x^T\Sigma^{-1}x)$ with domain $\mathbf{dom}\, q = \mathbf{S}^n_{++}$ is convex.*

Proof. We prove convexity by restricting the function $q(\Sigma) = \log(x^T\Sigma^{-1}x)$ to an arbitrary line, given by $\Sigma = Z + tV$, where $Z, V \in \mathbf{S}^n$. We define $f(t) = q(Z + tV)$ and restrict f to the interval of values of t for which $Z + tV \succ 0$. Without loss of generality we assume that $Z \succ 0$. We have

$$f(t) = \log(x^T(Z+tV)^{-1}x) = \log(x^TZ^{-1/2}(I + tZ^{-1/2}VZ^{-1/2})^{-1}Z^{-1/2}x).$$

Consider the eigenvalue decomposition $Q\Lambda Q^T = Z^{-1/2}VZ^{-1/2}$ where $\Lambda = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$, and let $z = Q^TZ^{-1/2}x$. Then

$$f(t) = \log(z^T(I + t\Lambda)^{-1}z) = \log\left(\sum_{i=1}^n \frac{z_i^2}{1+t\lambda_i}\right).$$

Hence, $f(t) = \log g(t)$ where $g(t) = \sum_{i=1}^n z_i^2/(1+t\lambda_i)$. The second derivative is

$$f''(t) = \frac{g(t)g''(t) - g'(t)^2}{g(t)^2}.$$

As we show below, it follows from the Cauchy-Schwarz inequality that the nominator is nonnegative, thus implying that f is convex. To see that the nominator is nonnegative, let $a_i = z_i^2 \geq 0$ and write it as

$$g(t)g''(t) - g'(t)^2 = \left(\sum_{i=1}^n \frac{a_i}{1+t\lambda_i}\right) \cdot \left(2\sum_{i=1}^n \frac{a_i\lambda_i^2}{(1+t\lambda_i)^3}\right) - \left(\sum_{i=1}^n \frac{a_i\lambda_i}{(1+t\lambda_i)^2}\right).$$

Applying the Cauchy-Schwarz inequality

$$\left(\sum_{i=1}^n b_i^2\right)\left(\sum_{i=1}^n c_i^2\right) \geq \left(\sum_{i=1}^n b_i c_i\right)^2$$

with

$$b_i = \left(\frac{a_i}{1+t\lambda_i}\right)^{1/2}, \qquad c_i = \frac{\sqrt{a_i}\lambda_i}{(1+t\lambda_i)^{3/2}}$$

concludes the proof. □

Lemma 11. *For fixed $\Sigma \in \mathbf{S}^n_{++}$ and fixed non-zero $x \in \mathbf{R}^n$, the value of $\log(x^T\Sigma^{-1}x)$ is equal to the optimal value of the convex problem*

$$\begin{array}{ll} \textit{minimize} & t \\ \textit{subject to} & \begin{bmatrix} 1 & vx^T \\ vx & \Sigma \end{bmatrix} \succeq 0 \\ & v \geq e^{-t/2}, \end{array}$$

with variables $t \in \mathbf{R}$ and $v \in \mathbf{R}$.

Proof. The value of $\log(x^T\Sigma^{-1}x)$ is equal to the optimal value of

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \log(x^T\Sigma^{-1}x) \leq t, \end{array}$$

with variable $t \in \mathbf{R}$. By applying the exponential function to both sides of the constraint, we conclude that this is equivalent to the problem

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & 1 - e^{-t}x^T\Sigma^{-1}x \geq 0, \end{array}$$

with variable $t \in \mathbf{R}$. This is in turn equivalent to

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & 1 - v^2x^T\Sigma^{-1}x \geq 0 \\ & v = e^{-t/2}, \end{array}$$

with variables $t \in \mathbf{R}$ and $v \in \mathbf{R}$. By relaxing the equality constraint to an inequality constraint, we obtain the problem

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & 1 - v^2x^T\Sigma^{-1}x \geq 0 \\ & v \geq e^{-t/2}. \end{array}$$

There always exists an optimal solution $(v^\star, t^\star)$ to this problem such that $v^\star = e^{-t^\star/2}$. Hence, the relaxation is tight. Furthermore, using the Schur complement (Boyd and Vandenberghe, 2004, §A.5), the constraint $1 - v^2x^T\Sigma^{-1}x \geq 0$ is equivalent to

$$\begin{bmatrix} 1 & vx^T \\ vx & \Sigma \end{bmatrix} \succeq 0,$$

which concludes the proof. □

A.5 Tyler's M-estimator with factor model

Another popular structural assumption in covariance estimation is the *statistical factor model*, which assumes that the covariance matrix can be decomposed as the sum of a low-rank positive semidefinite matrix and a diagonal matrix (Bartholomew et al., 2011). Specifically, we assume that the covariance matrix $\Sigma \in \mathbf{S}^n$ is of the form

$$\Sigma = FF^T + D, \tag{21}$$

where $D \in \mathbf{R}^{n\times n}$ is diagonal with positive diagonal entries, and $F \in \mathbf{R}^{n\times r}$ with $r \leq n$. Note that the covariance structure (21) only has $n(r+1)$ parameters to estimate, which typically is much fewer than the $n(n+1)/2$ parameters that must be estimated for an unstructured covariance matrix.

To incorporate the covariance structure (21) into Tyler's M-estimator, we solve

$$\text{minimize } \log\det(FF^T + D) + \frac{n}{m}\sum_{i=1}^{m}\log(x_i^T(FF^T + D)^{-1}x_i), \tag{22}$$

where the variables are the diagonal matrix $D \in \mathbf{R}^{n\times n}$ and $F \in \mathbf{R}^{n\times n}$.

An equivalent problem formulation. From the matrix inversion lemma (Boyd and Vandenberghe, 2004, §C.4) we know that the inverse of a low-rank plus diagonal matrix has a similar low-rank structure. More precisely,

$$(FF^T + D)^{-1} = D^{-1} - D^{-1}F(I + F^TD^{-1}F)^{-1}F^TD^{-1} = E - GG^T$$

where $E \in \mathbf{R}^{n\times n}$ and $G \in \mathbf{R}^{n\times r}$ are given by

$$E = D^{-1}, \qquad G = D^{-1}F(I + F^TD^{-1}F)^{-1/2}. \tag{23}$$

This suggests that problem (22) is equivalent to

$$\text{minimize } h(G, E) = -\log\det(E - GG^T) + \frac{n}{m}\sum_{i=1}^{m}\log(x_i^T(E - GG^T)x_i), \tag{24}$$

where the variables are $G \in \mathbf{R}^{n\times r}$ and the diagonal matrix $E \in \mathbf{R}^{n\times n}$. This intuition is formally verified by the following theorem.

Theorem 3. *Let $(G^\star, e^\star)$ be a solution of (24), and set $E^\star = \mathbf{diag}(e^\star)$. Then the matrix $I - (G^\star)^T(E^\star)^{-1}G^\star$ is positive definite and a solution $(F^\star, D^\star)$ of (22) is given by*

$$\begin{aligned} d^\star &= 1/e^\star \qquad \textit{(element-wise division)} \\ F^\star &= (E^\star)^{-1}G^\star(I - (G^\star)^T(E^\star)^{-1}G^\star)^{-1/2}. \end{aligned} \tag{25}$$

Proof. It holds that $E^\star - G^\star(G^\star)^T \succ 0$, since if a solution exists it must belong to the domain of the objective function. This further implies that $E^\star \succ 0$. From results on Schur complements (see, for example, (Boyd and Vandenberghe, 2004, Appendix A.5.5)), it follows that $E^\star - G^\star(G^\star)^T \succ 0$ if and only if $I - (G^\star)^T(E^\star)^{-1}G^\star \succ 0$.

For any (F, d) that belongs to the domain of the objective function $f(F, d)$ of (22), we can use (23) to construct (G, e) satisfying $h(G, e) = f(F, d)$. Hence, the optimal value of (24) is a lower bound on the optimal value of (22). Furthermore, $(F^\star, D^\star)$ given by (25) satisfies $f(F^\star, d^\star) = h(G^\star, e^\star)$, which is the lower bound on the optimal value of (22), and $(F^\star, D^\star)$ given by (25) must therefore be optimal. □

The practical consequence of this result is that we can solve (22) by solving (24), and then recover the solution to (22) using (25). The objective function of (24) is as a difference of convex functions, and we can therefore solve (24) using CCP. Specifically, we can write the objective function of (24) as

$$h(G, e) = -\log\det\begin{bmatrix} E & G \\ G^T & I \end{bmatrix} + \frac{n}{m}\sum_{i=1}^{m}\log(x_i^TEx_i - \|G^Tx_i\|_2^2).$$

The first term is convex in (G, e), since $A \mapsto -\log\det A$ is convex and

$$(G, e) \mapsto \begin{bmatrix} E & G \\ G^T & I \end{bmatrix}$$

is linear in (G, e). The second term is concave in (G, e). To see this, note that $x_i^TEx_i$ is linear in e, and $G \mapsto \|G^Tx_i\|_2^2$ is convex, since $\|\cdot\|_2^2$ is convex and $G \mapsto G^Tx_i$ is linear. Hence, $(G, e) \mapsto x_i^TEx_i - \|G^Tx_i\|_2^2$ is concave, and since the logarithm preserves concavity see (Boyd and Vandenberghe, 2004, §3) it follows that the second term of $h(G, e)$ is concave. The objective function $h(G, e)$ is therefore the sum of a convex and a concave term, and hence CCP-compatible. The subproblem obtained by linearizing the concave term is easy to specify and solve using CVXPY.

B OMITTED PROOFS - GENERAL

Lemma 12. *The function $w(Y) = -\log(a^T Y a)$, $\mathbf{dom}\, w = \mathbf{S}^n_{++}$, where $a \in \mathbf{R}^n$ is fixed is 2-smooth relative to $h(Y) = -(1/2)\log\det Y$.*

Proof. The Euclidean second directional derivatives at $Y \in \mathbf{S}^n_{++}$ in the direction $U \in \mathbf{S}^n$ are given by

$$\nabla^2 h(Y)[U,U] = \frac{1}{2}\,\mathbf{Tr}(Y^{-1}UY^{-1}U), \qquad \nabla^2 w(Y)[U,U] = \frac{(a^TUa)^2}{(a^TYa)^2}.$$

To show that $w(Y)$ is 2-smooth relative to $h(Y)$, we need to show that (Lu et al., 2018, Proposition 1.1)

$$\nabla^2 w(Y)[U,U] \le 2\nabla^2 h(Y)[U,U].$$

With $Z = Y^{-1/2}UY^{-1/2}$ and $x = Y^{1/2}a$, we have

$$a^TUa = a^TY^{1/2}Y^{-1/2}UY^{-1/2}Y^{1/2}a = x^TZx.$$

Furthermore,

$$0 \le x^TZx = \mathbf{Tr}(Zxx^T) \le \|Z\|_F\|xx^T\|_F = \|Z\|_F\|x\|_2^2$$

where we have used the Cauchy-Schwarz inequality. Thus, we have

$$(a^TUa)^2 \le \|Z\|_F^2\|x\|_2^4$$

where

$$\begin{aligned}\|x\|_2^4 &= \|Y^{1/2}a\|_2^4 = (a^TYa)^2\\ \|Z\|_F^2 &= \|Y^{-1/2}UY^{-1/2}\|_F^2 = \mathbf{Tr}(Y^{-1/2}UY^{-1/2}Y^{-1/2}UY^{-1/2}) = \mathbf{Tr}(Y^{-1}UY^{-1}U).\end{aligned}$$

The desired result follows. □

Lemma 13. *For any $Y, Z \in \mathbf{S}^n_{++}$ and any $t \in [0,1]$,*

$$D_{KL}(\gamma(t;Y,Z)\|Y) \le C_{YZ}D_{geo}(\gamma(t;Y,Z),Y)^2,$$

where

$$C_{YZ} = \frac{a_{YZ} - \log a_{YZ} - 1}{2\log^2(a_{YZ})}$$

for any $a_{YZ} \in \mathbf{R}$ satisfying $a_{YZ} \ge \lambda_{\max}(Y^{-1}Z)$. (For $a_{YZ} = 1$, we define $C_{YZ} = \lim_{a\to 1}(a - \log a - 1)/(2\log^2(a)) = 1/4$.)

Proof. Fix $Y, Z \in \mathbf{S}^n_{++}$ and let $A = \log(Y^{-1/2}ZY^{-1/2})$. Using (11), we have

$$D_{\mathrm{geo}}(\gamma(t;Y,Z),Y) = tD_{\mathrm{geo}}(Z,Y) = t\|A\|_F.$$

Furthermore, $\gamma(t;Y,Z) = Y^{1/2}e^{tA}Y^{1/2}$, so

$$\begin{aligned}\mathbf{Tr}(Y^{-1}\gamma(t;Y,Z)) &= \mathbf{Tr}(e^{tA})\\ \log\det(Y^{-1}\gamma(t;Y,Z)) &= \log\det(e^{tA}) = t\,\mathbf{Tr}(A),\end{aligned}$$

implying that

$$\begin{aligned}D_{\mathrm{KL}}(\gamma(t;Y,Z)\|Y) &= \frac{1}{2}\left(\mathbf{Tr}(Y^{-1}\gamma(t;Y,Z)) + \log\det(Y\gamma(t;Y,Z)^{-1}) - n\right)\\ &= \frac{1}{2}\left(\mathbf{Tr}(Y^{-1}\gamma(t;Y,Z)) - \log\det(Y^{-1}\gamma(t;Y,Z)) - n\right)\\ &= \frac{1}{2}\left(\mathbf{Tr}(e^{tA}) - t\,\mathbf{Tr}(A) - n\right).\end{aligned}$$

Let λ_i, $i = 1, \dots, n$, be the eigenvalues of A. Then

$$D_{\mathrm{KL}}(\gamma(t; Y, Z) \| Y) = \frac{1}{2} \sum_{i=1}^{n} (e^{t\lambda_i} - t\lambda_i - 1)$$
$$D_{\mathrm{geo}}(\gamma(t; Y, Z), Y)^2 = t^2 \sum_{i=1}^{n} \lambda_i^2.$$

Hence, we want find a constant $C_{YZ} > 0$ such that

$$\frac{1}{2} \sum_{i=1}^{n} (e^{t\lambda_i} - t\lambda_i - 1) \leq C_{YZ} t^2 \sum_{i=1}^{n} \lambda_i^2, \qquad t \in (0, 1],$$

whenever λ_i, $i = 1, \dots, n$, are the eigenvalues of $A = \log(Y^{-1/2} Z Y^{-1/2})$. (The inequality clearly holds for $t = 0$.) If $\lambda_k = 0$ for some $k \in \{1, \dots, n\}$, then the corresponding terms in the left and right hand sides are zero, and we can simply remove them. Thus, assume that $\lambda_i \neq 0$ for all $i = 1, \dots, n$.

For fixed $t \in (0, 1]$, consider the function $f_t(\lambda) = (e^{t\lambda} - t\lambda - 1)/(t^2\lambda^2)$ for $\lambda \neq 0$ and extended by continuity to $f_t(0) = 1/2$. Then

$$\begin{aligned} \frac{\sum_{i=1}^{n} (e^{t\lambda_i} - t\lambda_i - 1)}{t^2 \sum_{j=1}^{n} \lambda_j^2} &= \sum_{i=1}^{n} \frac{e^{t\lambda_i} - t\lambda_i - 1}{t^2 \lambda_i^2} \cdot \frac{\lambda_i^2}{\sum_{j=1}^{n} \lambda_j^2} \\ &= \sum_{i=1}^{n} f_t(\lambda_i) \cdot \frac{\lambda_i^2}{\sum_{j=1}^{n} \lambda_j^2} \\ &\leq \max_{i=1,\dots,n} f_t(\lambda_i). \end{aligned}$$

The function $\lambda \mapsto f_t(\lambda)$ is increasing for fixed $t \in (0, 1]$, (see Lemma 14 below), and $t \mapsto f_t(\lambda)$ is also increasing (see Lemma 15 below), which together with the above inequality implies that for any $t \in (0, 1]$,

$$\frac{\sum_{i=1}^{n} (e^{t\lambda_i} - t\lambda_i - 1)}{t^2 \sum_{j=1}^{n} \lambda_j^2} \leq f_t(\lambda_{\max}(A)) \leq f_1(\lambda_{\max}(A)).$$

Note that

$$f_1(\lambda_{\max}(\log(Y^{-1/2} Z Y^{-1/2}))) = f_1(\lambda_{\max}(\log(Y^{-1} Z))) = f_1(\log(\lambda_{\max}(Y^{-1} Z))) \leq f_1(\log(a_{YZ})),$$

where $a_{YZ} \geq \lambda_{\max}(Y^{-1} Z)$. Hence, it suffices to choose

$$C_{YZ} = \frac{f_1(\log(a_{YZ}))}{2} = \frac{a_{YZ} - \log(a_{YZ}) - 1}{2 \log(a_{YZ})^2},$$

which concludes the proof. □

Lemma 14. *For fixed $t \in (0, 1]$, the function $f_t(\lambda) = (e^{t\lambda} - t\lambda - 1)/(t^2\lambda^2)$ for $\lambda \neq 0$ and extended by continuity to $f_t(0) = 1/2$, is increasing.*

Proof. The function f_t is continuous, so it suffices to show that the derivative is nonnegative for all $\lambda \neq 0$. The first derivative is

$$\frac{\partial f_t(\lambda)}{\partial \lambda} = \frac{t^2 \lambda \left(t\lambda(e^{t\lambda} - 1) - 2(e^{t\lambda} - t\lambda - 1) \right)}{(t^2\lambda^2)^2}.$$

The sign of the derivative is determined by the sign of the numerator. With $z = t\lambda$, we can write the numerator (excluding a positive factor t) as

$$z \left(z(e^z - 1) - 2(e^z - z - 1) \right) = z \left(e^z(z - 2) + z + 2 \right).$$

This expression is positive for all $z \neq 0$. □

Lemma 15. *For fixed $\lambda \neq 0$, the function $t \mapsto f_t(\lambda) = (e^{t\lambda} - t\lambda - 1)/(t^2\lambda^2)$ with domain $\mathbf{dom}\, f = (0, 1]$ is increasing.*

Proof. The proof is identical to the proof of Lemma 14, but with t and λ swapped. □

C OMITTED PROOFS - TYLER'S FIXED-POINT ITERATION

For the definition of Y_k, $Y^\star$, and $\mathcal{L}_0^F$ in the following result, see §4.1 of the main text.

Lemma 16. *For any $k \geq 0$, we have $\lambda_{\max}(Y_{k+1}^{-1}Y^\star) \leq \lambda_{\max}(Y_k^{-1}Y^\star)$ and $\lambda_{\min}(Y_{k+1}^{-1}Y^\star) \geq \lambda_{\min}(Y_k^{-1}Y^\star)$. Furthermore, $Y_k \in \mathcal{L}_0^F$.*

Proof. Note that $Y_k^{-1}Y^\star = \Sigma_k(\Sigma^\star)^{-1}$. We first prove that $\lambda_{\max}(\Sigma_{k+1}(\Sigma^\star)^{-1}) \leq \lambda_{\max}(\Sigma_k(\Sigma^\star)^{-1})$. With $y_i = (\Sigma^\star)^{-1/2}x_i$, $i = 1, \ldots, m$, we have

$$\begin{aligned}(\Sigma^\star)^{-1/2}\Sigma_{k+1}(\Sigma^\star)^{-1/2} &= (\Sigma^\star)^{-1/2}\left(\frac{n}{m}\sum_{i=1}^m \frac{x_i x_i^T}{x_i^T\Sigma_k^{-1}x_i}\right)(\Sigma^\star)^{-1/2} \\ &= (\Sigma^\star)^{-1/2}\left(\frac{n}{m}\sum_{i=1}^m \frac{x_i x_i^T}{y_i^T(\Sigma^\star)^{1/2}\Sigma_k^{-1}(\Sigma^\star)^{1/2}y_i}\right)(\Sigma^\star)^{-1/2}.\end{aligned}$$

For any $i = 1, \ldots, m$,

$$y_i^T(\Sigma^\star)^{1/2}\Sigma_k^{-1}(\Sigma^\star)^{1/2}y_i \geq \lambda_{\min}((\Sigma^\star)^{1/2}\Sigma_k^{-1}(\Sigma^\star)^{1/2})y_i^T y_i = \lambda_{\min}(\Sigma_k^{-1}\Sigma^\star)y_i^T y_i = \frac{1}{\lambda_{\max}(\Sigma_k(\Sigma^\star)^{-1})}y_i^T y_i.$$

Hence,

$$\begin{aligned}(\Sigma^\star)^{-1/2}\Sigma_{k+1}(\Sigma^\star)^{-1/2} &\preceq \lambda_{\max}(\Sigma_k(\Sigma^\star)^{-1})(\Sigma^\star)^{-1/2}\left(\frac{n}{m}\sum_{i=1}^m \frac{x_i x_i^T}{y_i^T y_i}\right)(\Sigma^\star)^{-1/2} \\ &= \lambda_{\max}(\Sigma_k(\Sigma^\star)^{-1})(\Sigma^\star)^{-1/2}\left(\frac{n}{m}\sum_{i=1}^m \frac{x_i x_i^T}{x_i^T(\Sigma^\star)^{-1}x_i}\right)(\Sigma^\star)^{-1/2} \\ &= \lambda_{\max}(\Sigma_k(\Sigma^\star)^{-1})I,\end{aligned}$$

where the last equality follows from the fact that $\Sigma^\star$ satisfies (1). We therefore have

$$\lambda_{\max}((\Sigma^\star)^{-1/2}\Sigma_{k+1}(\Sigma^\star)^{-1/2}) \leq \lambda_{\max}(\Sigma_k(\Sigma^\star)^{-1}).$$

Note that

$$\lambda_{\max}((\Sigma^\star)^{-1/2}\Sigma_{k+1}(\Sigma^\star)^{-1/2}) = \lambda_{\max}(\Sigma_{k+1}(\Sigma^\star)^{-1})$$

due to a similarity transformation. Hence,

$$\lambda_{\max}(\Sigma_{k+1}(\Sigma^\star)^{-1}) \leq \lambda_{\max}(\Sigma_k(\Sigma^\star)^{-1}).$$

Next, we prove that $\lambda_{\min}(Y_{k+1}^{-1}Y^\star) \geq \lambda_{\min}(Y_k^{-1}Y^\star)$. As above, we have

$$(\Sigma^\star)^{-1/2}\Sigma_{k+1}(\Sigma^\star)^{-1/2} = (\Sigma^\star)^{-1/2}\left(\frac{n}{m}\sum_{i=1}^m \frac{x_i x_i^T}{y_i^T(\Sigma^\star)^{1/2}\Sigma_k^{-1}(\Sigma^\star)^{1/2}y_i}\right)(\Sigma^\star)^{-1/2}.$$

For any $i = 1, \ldots, m$,

$$y_i^T(\Sigma^\star)^{1/2}\Sigma_k^{-1}(\Sigma^\star)^{1/2}y_i \leq \lambda_{\max}((\Sigma^\star)^{1/2}\Sigma_k^{-1}(\Sigma^\star)^{1/2})y_i^T y_i = \lambda_{\max}(\Sigma_k^{-1}\Sigma^\star)y_i^T y_i = \frac{1}{\lambda_{\min}(\Sigma_k(\Sigma^\star)^{-1})}y_i^T y_i.$$

Hence,

$$\begin{aligned}(\Sigma^\star)^{-1/2}\Sigma_{k+1}(\Sigma^\star)^{-1/2} &\succeq \lambda_{\min}(\Sigma_k(\Sigma^\star)^{-1})(\Sigma^\star)^{-1/2}\left(\frac{n}{m}\sum_{i=1}^m \frac{x_i x_i^T}{y_i^T y_i}\right)(\Sigma^\star)^{-1/2} \\ &= \lambda_{\min}(\Sigma_k(\Sigma^\star)^{-1})(\Sigma^\star)^{-1/2}\left(\frac{n}{m}\sum_{i=1}^m \frac{x_i x_i^T}{x_i^T(\Sigma^\star)^{-1}x_i}\right)(\Sigma^\star)^{-1/2} \\ &= \lambda_{\min}(\Sigma_k(\Sigma^\star)^{-1})I.\end{aligned}$$

From eigenvalue similarity it follows that $\lambda_{\min}(\Sigma_{k+1}(\Sigma^\star)^{-1}) \geq \lambda_{\min}(\Sigma_k(\Sigma^\star)^{-1})$. This concludes the proof of the first part of the lemma.

To see that $Y_k \in \mathcal{L}_0^F$, note that above we have shown $(\Sigma^\star)^{-1/2}\Sigma_{k+1}(\Sigma^\star)^{-1/2} \succeq \lambda_{\min}(\Sigma_k(\Sigma^\star)^{-1})I$, or equivalently, $\Sigma_{k+1} \succeq \lambda_{\min}(\Sigma_k(\Sigma^\star)^{-1})\Sigma^\star$. Using that $\lambda_{\min}(\Sigma_k(\Sigma^\star)^{-1})$ is increasing in k, we conclude that $\Sigma_{k+1} \succeq \lambda_{\min}((\Sigma^\star)^{-1})\Sigma^\star$. A similar argument, based on the inequality $(\Sigma^\star)^{-1/2}\Sigma_{k+1}(\Sigma^\star)^{-1/2} \preceq \lambda_{\max}(\Sigma_k(\Sigma^\star)^{-1})I$ that we showed above, shows that $\Sigma_{k+1} \preceq \lambda_{\max}((\Sigma^\star)^{-1})\Sigma^\star$. Hence, for any $k \geq 0$,

$$\lambda_{\min}((\Sigma^\star)^{-1})\Sigma^\star \preceq \Sigma_k \preceq \lambda_{\max}((\Sigma^\star)^{-1})\Sigma^\star,$$

or equivalently

$$(1/\lambda_{\min}((\Sigma^\star)^{-1}))Y^\star \succeq Y_k \succeq (1/\lambda_{\max}((\Sigma^\star)^{-1}))Y^\star.$$

The statement that $Y_k \in \mathcal{L}_0^F$ now follows from combining this inequality with the fact that CCP is a descent algorithm (see §2.2 of the main text) so $F(Y_k) \leq F(Y_0)$ for $k \geq 0$. □

D OMITTED PROOFS - TYLER'S REGULARIZED FIXED-POINT ITERATION

For the definition of Y_k and $Y_\alpha^\star$ used in the following results, see §4.2 of the main text.

Lemma 17. *For any $k \geq 0$, we have*

$$\lambda_{\max}(Y_k^{-1}Y_\alpha^\star) \leq \lambda_{\max}((\Sigma_\alpha^\star)^{-1}).$$

Proof. The result clearly holds for $k = 0$ since we initialize at $\Sigma_0 = Y_0 = I$. Assume the result holds for some $k \geq 0$. We prove that $\lambda_{\max}(Y_{k+1}^{-1}Y_\alpha^\star) = \lambda_{\max}(\Sigma_{k+1}(\Sigma_\alpha^\star)^{-1}) \leq \lambda_{\max}((\Sigma_\alpha^\star)^{-1})$. With $y_i = (\Sigma_\alpha^\star)^{-1/2}x_i$, we have

$$\begin{aligned}
(\Sigma_\alpha^\star)^{-1/2}\Sigma_{k+1}(\Sigma_\alpha^\star)^{-1/2} &= (\Sigma_\alpha^\star)^{-1/2}\left(\frac{(1-\alpha)n}{m}\sum_{i=1}^m \frac{x_i x_i^T}{y_i^T(\Sigma_\alpha^\star)^{1/2}\Sigma_k^{-1}(\Sigma_\alpha^\star)^{1/2}y_i} + \alpha I\right)(\Sigma_\alpha^\star)^{-1/2} \\
&\preceq (\Sigma_\alpha^\star)^{-1/2}\left(\frac{(1-\alpha)n\lambda_{\max}(\Sigma_k(\Sigma_\alpha^\star)^{-1})}{m}\sum_{i=1}^m \frac{x_i x_i^T}{y_i^T y_i} + \alpha I\right)(\Sigma_\alpha^\star)^{-1/2} \\
&= (\Sigma_\alpha^\star)^{-1/2}\left(\frac{(1-\alpha)n\lambda_{\max}(\Sigma_k(\Sigma_\alpha^\star)^{-1})}{m}\sum_{i=1}^m \frac{x_i x_i^T}{x_i^T(\Sigma_\alpha^\star)^{-1}x_i} + \alpha I\right)(\Sigma_\alpha^\star)^{-1/2} \\
&= \lambda_{\max}(\Sigma_k(\Sigma_\alpha^\star)^{-1})(\Sigma_\alpha^\star)^{-1/2}\left(\frac{(1-\alpha)n}{m}\sum_{i=1}^m \frac{x_i x_i^T}{x_i^T(\Sigma_\alpha^\star)^{-1}x_i} + \alpha I\right)(\Sigma_\alpha^\star)^{-1/2} \\
&\quad + \alpha(1-\lambda_{\max}(\Sigma_k(\Sigma_\alpha^\star)^{-1}))(\Sigma_\alpha^\star)^{-1} \\
&= \lambda_{\max}(\Sigma_k(\Sigma_\alpha^\star)^{-1})I + \alpha(1-\lambda_{\max}(\Sigma_k(\Sigma_\alpha^\star)^{-1}))(\Sigma_\alpha^\star)^{-1},
\end{aligned}$$

where we have used the fact that $\Sigma_\alpha^\star$ satisfies

$$\Sigma_\alpha^\star = \frac{(1-\alpha)n}{m}\sum_{i=1}^m \frac{x_i x_i^T}{x_i^T(\Sigma_\alpha^\star)^{-1}x_i} + \alpha I.$$

This can be shown by setting the gradient of the objective function of Tyler's regularized maximum likelihood estimation problem to zero. Next, we consider two cases.

- If $\lambda_{\max}(\Sigma_k(\Sigma_\alpha^\star)^{-1}) \geq 1$, then we can drop the last term to conclude that

  $$(\Sigma_\alpha^\star)^{-1/2}\Sigma_{k+1}(\Sigma_\alpha^\star)^{-1/2} \preceq \lambda_{\max}(\Sigma_k(\Sigma_\alpha^\star)^{-1})I,$$

  implying that the maximum eigenvalue of the left-hand side is upper bounded by the maximum eigenvalue of the right-hand side. Note that

  $$\lambda_{\max}((\Sigma_\alpha^\star)^{-1/2}\Sigma_{k+1}(\Sigma_\alpha^\star)^{-1/2}) = \lambda_{\max}(\Sigma_{k+1}(\Sigma_\alpha^\star)^{-1}).$$

  Thus, it follows that

  $$\lambda_{\max}(\Sigma_{k+1}(\Sigma_\alpha^\star)^{-1}) \leq \lambda_{\max}(\Sigma_k(\Sigma_\alpha^\star)^{-1}) \leq \lambda_{\max}((\Sigma_\alpha^\star)^{-1}).$$

- Suppose $a_k = \lambda_{\max}(\Sigma_k(\Sigma_\alpha^\star)^{-1}) < 1$. Then we have

$$a_{k+1} \leq a_k + \alpha(1-a_k)\lambda_{\max}((\Sigma_\alpha^\star)^{-1}).$$

Note that

$$\begin{aligned}\sup_{a_k\in[0,1]} \left\{a_k + \alpha(1-a_k)\lambda_{\max}((\Sigma_\alpha^\star)^{-1})\right\} &= \sup_{a_k\in[0,1]} \left\{\alpha\lambda_{\max}((\Sigma_\alpha^\star)^{-1}) + (1-\alpha\lambda_{\max}((\Sigma_\alpha^\star)^{-1}))a_k\right\} \\ &= \max\left\{\alpha\lambda_{\max}((\Sigma_\alpha^\star)^{-1}), 1\right\} \\ &\leq \max\left\{\lambda_{\max}((\Sigma_\alpha^\star)^{-1}), 1\right\}.\end{aligned}$$

Now we use the fact that $\mathbf{Tr}((\Sigma_\alpha^\star)^{-1}) = n$ (Pascal et al., 2014, Propostion 3.1) to conclude that $\lambda_{\max}((\Sigma_\alpha^\star)^{-1}) \geq 1$. Hence, it follows that

$$a_{k+1} \leq \lambda_{\max}((\Sigma_\alpha^\star)^{-1}).$$

This concludes the proof. □

For the next results we recall that $\mathcal{A} = \{Y \in \mathbf{S}_{++}^n \mid \lambda_{\min}(Y) \geq 1/\kappa^\star\}$ where $\kappa^\star = \lambda_{\max}(\Sigma_\alpha^\star)/\lambda_{\min}(\Sigma_\alpha^\star)$.

Lemma 18. *It holds that $Y_\alpha^\star \in \mathcal{A}$, and $Y_k \in \mathcal{A}$ for $k \geq 0$.*

Proof. Verifying that $Y_\alpha^\star \in \mathcal{A}$ is straightforward since

$$\frac{1}{\kappa^\star} = \frac{\lambda_{\min}(\Sigma_\alpha^\star)}{\lambda_{\max}(\Sigma_\alpha^\star)} = \frac{\lambda_{\min}(Y_\alpha^\star)}{\lambda_{\max}(Y_\alpha^\star)},$$

and $\lambda_{\max}(Y_\alpha^\star) \geq 1$ since $\mathbf{Tr}(Y_\alpha^\star) = n$ (Pascal et al., 2014, Proposition 3.1).

From Lemma 17 we have $\lambda_{\max}(Y_k^{-1}Y_\alpha^\star) \leq \lambda_{\max}(Y_\alpha^\star)$. Combining this together with the fact that $\lambda_{\max}(Y_k^{-1}Y_\alpha^\star) \geq \lambda_{\max}(Y_k^{-1})\lambda_{\min}(Y_\alpha^\star)$, we have

$$\lambda_{\max}(Y_k^{-1}) \leq \frac{\lambda_{\max}(Y_\alpha^\star)}{\lambda_{\min}(Y_\alpha^\star)}.$$

Hence,

$$\lambda_{\min}(Y_k) \geq \frac{\lambda_{\min}(Y_\alpha^\star)}{\lambda_{\max}(Y_\alpha^\star)}.$$

□

Lemma 19. *The set $\mathcal{A}$ is geodesically convex.*

Proof. We need to show that for any $Y_1, Y_2 \in \mathcal{A}$, the geodesic $\gamma(t; Y_1, Y_2)$, $t \in [0,1]$, connecting Y_1 and Y_2 is contained in $\mathcal{A}$.

Let $Y_1 \in \mathcal{A}$ and $Y_2 \in \mathcal{A}$. A property of the geodesic $\gamma(t;\cdot,\cdot)$ is that for any $A, B \in \mathbf{S}_{++}^n$, it holds that (Wiesel and Zhang, 2015, Theorem 1.6)

$$\gamma(t; A, B) \preceq (1-t)A + tB, \qquad t \in [0,1].$$

Hence, with $A = Y_1^{-1}$ and $B = Y_2^{-1}$, we have

$$\gamma(t; Y_1^{-1}, Y_2^{-1}) \preceq (1-t)Y_1^{-1} + tY_2^{-1}, \qquad t \in [0,1].$$

Hence, for any $t \in [0,1]$,

$$\begin{aligned}\lambda_{\max}(\gamma(t; Y_1^{-1}, Y_2^{-1})) &\leq \lambda_{\max}((1-t)Y_1^{-1} + tY_2^{-1}) \\ &\leq (1-t)\lambda_{\max}(Y_1^{-1}) + t\lambda_{\max}(Y_2^{-1}) \\ &= \frac{1-t}{\lambda_{\min}(Y_1)} + \frac{t}{\lambda_{\min}(Y_2)} \\ &\leq (1-t)\kappa^\star + t\kappa^\star = \kappa^\star.\end{aligned}$$

We now use that $\gamma(t; Y_1^{-1}, Y_2^{-1}) = \gamma(t; Y_1, Y_2)^{-1}$ (Wiesel and Zhang, 2015, Lemma 1.8), so

$$\frac{1}{\lambda_{\min}(\gamma(t; Y_1, Y_2))} \leq \kappa^\star,$$

which implies the desired result. □

Lemma 20. *The function $r(Y) = \mathbf{Tr}(Y)$, $\mathbf{dom}\, r = \mathbf{S}^n_{++}$, is geodesically μ-strongly convex on $\mathcal{A}$ with $\mu = 1/\kappa^\star$.*

Proof. The function $r(Y)$ is geodesically μ-strongly convex on $\mathcal{A}$ if for all $Y \in \mathcal{A}$ (Boumal, 2023, Theorem 11.23),

$$\nabla^2 r(Y)[U, U] \succeq \mu \|U\|_Y^2, \qquad U \in \mathbf{S}^n,$$

where $\|U\|_Y^2 = \mathbf{Tr}(Y^{-1}U^T Y^{-1}U)$ is the Riemannian metric on $\mathbf{S}^n_{++}$ and $\nabla^2 r(Y)[U, U]$ is the Riemannian second directional derivative of r at Y in the direction U. The Euclidean gradient and Hessian of $r(Y)$ is I and 0, respectively. Thus, using (Boumal, 2023, Chapter 11.7), we have

$$\nabla^2 r(Y)[U] = \frac{1}{2}(YU + UY), \qquad U \in \mathbf{S}^n,$$

where $\nabla^2 r(Y)[U]$ is the Riemannian directional derivative of the Riemannian gradient at Y in the direction U. Furthermore, if we denote the Riemannian inner product at Y by $\langle A, B\rangle_Y = \mathbf{Tr}(Y^{-1}AY^{-1}B)$ for $A, B \in \mathbf{S}^n_{++}$, then the Riemannian second directional derivative of r at Y in the direction $U \in \mathbf{S}^n$ is given by

$$\begin{aligned}\nabla^2 r(Y)[U, U] &= \frac{1}{2}\langle YU + UY, U\rangle_Y = \frac{1}{2}\,\mathbf{Tr}(Y^{-1}(YU + UY)Y^{-1}U) \\ &= \mathbf{Tr}(UY^{-1}U).\end{aligned}$$

To lower bound $\nabla^2 r(Y)[U, U]$ in terms of the Riemannian metric $\|U\|_Y^2 = \mathbf{Tr}(Y^{-1}UY^{-1}U)$, we write

$$\mathbf{Tr}(UY^{-1}U) = \mathbf{Tr}(YY^{-1}UY^{-1}U) \geq \lambda_{\min}(Y)\,\mathbf{Tr}(Y^{-1}UY^{-1}U) = \lambda_{\min}(Y)\|U\|_Y^2.$$

Hence, for any $Y \in \mathcal{A}$, we have

$$\nabla^2 r(Y)[U, U] \geq (1/\kappa^\star)\|U\|_Y^2.$$

Thus, $r(Y)$ is geodesically μ-strongly convex on $\mathcal{A}$ with $\mu = 1/\kappa^\star$. □