Deterministic guarantees for Burer–Monteiro factorizations of smooth semidefinite programs

NICOLAS BOUMAL
Mathematics Department and Program in Applied and Computational Mathematics, Princeton University

VLADISLAV VORONINSKI
Helm.ai

AND

AFONSO S. BANDEIRA
Department of Mathematics and Center for Data Science, Courant Institute of Mathematical Sciences, New York University

Abstract

We consider semidefinite programs (SDPs) with equality constraints. The variable to be optimized is a positive semidefinite matrix X of size n. Following the Burer–Monteiro approach, we optimize a factor Y of size $n \times p$ instead, such that $X = YY^\top$. This ensures positive semidefiniteness at no cost and can reduce the dimension of the problem if p is small, but results in a non-convex optimization problem with a quadratic cost function and quadratic equality constraints in Y. In this paper, we show that if the set of constraints on Y regularly defines a smooth manifold, then, despite non-convexity, first- and second-order necessary optimality conditions are also sufficient, provided p is large enough. For smaller values of p, we show a similar result holds for almost all (linear) cost functions. Under those conditions, a global optimum Y maps to a global optimum $X = YY^\top$ of the SDP. We deduce old and new consequences for SDP relaxations of the generalized eigenvector problem, the trust-region subproblem and quadratic optimization over several spheres, as well as for the Max-Cut and Orthogonal-Cut SDPs which are common relaxations in stochastic block modeling and synchronization of rotations.

https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21830

1 Introduction

We consider semidefinite programs (SDPs) of the form

$$\text{(SDP)} \qquad f^\star = \min_{X \in \mathbb{S}^{n\times n}} \langle C, X\rangle \quad \text{subject to} \quad \mathscr{A}(X) = b,\ X \succeq 0,$$

where $\mathbb{S}^{n\times n}$ is the set of real symmetric matrices of size n, $C \in \mathbb{S}^{n\times n}$ is the cost matrix, $\langle C, X\rangle = \mathrm{Tr}(C^\top X)$, $\mathscr{A} : \mathbb{S}^{n\times n} \to \mathbb{R}^m$ is a linear operator capturing m equality constraints with right-hand side $b \in \mathbb{R}^m$, and the variable X is symmetric, positive

semidefinite. Let $A_1, \ldots, A_m \in \mathbb{S}^{n\times n}$ be the constraint matrices such that $\mathscr{A}(X)_i = \langle A_i, X\rangle$, and let

$$\mathscr{C} = \left\{X \in \mathbb{S}^{n\times n} : \mathscr{A}(X) = b \text{ and } X \succeq 0\right\} \tag{1.1}$$

be the search space of (SDP), assumed non empty.

Interior point methods solve (SDP) in polynomial time [23]. In practice however, for n beyond a few thousands, such algorithms run out of memory (and time), prompting research for alternative solvers. Crucially, if $\mathscr{C}$ is compact, then (SDP) admits a global optimum of rank at most r, where $\frac{r(r+1)}{2} \leq m$ [24, 7]—we review this fact in Section 2.2. Thus, if one restricts $\mathscr{C}$ to matrices of rank at most p with $\frac{p(p+1)}{2} \geq m$, the optimal value remains unchanged. This restriction is easily enforced by factorizing $X = YY^\top$ where Y has size $n \times p$, yielding a quadratically constrained quadratic program:

$$\min_{Y\in\mathbb{R}^{n\times p}} \langle CY, Y\rangle \quad \text{subject to} \quad \mathscr{A}(YY^\top) = b. \tag{P}$$

In general, (P) is non-convex because its search space

$$\mathscr{M}_p = \left\{Y \in \mathbb{R}^{n\times p} : \mathscr{A}(YY^\top) = b\right\} \tag{1.2}$$

is non-convex. (When p is clear from context or unimportant, we just write $\mathscr{M}$.)

Non-convexity makes it a priori unclear how to solve (P). Still, the benefits are that $\mathscr{M}$ requires no conic constraint and can be lower dimensional than $\mathscr{C}$. This has motivated Burer and Monteiro [12, 13] to try to solve (P) using local optimization methods, with surprisingly good results. They developed theory in support of this observation (details below). About their results, Burer and Monteiro write:

> "*How large must we take p so that the local minima of* (P) *are guaranteed to map to global minima of* (SDP)*? Our theorem asserts that we need only*[1] $\frac{p(p+1)}{2} > m$ *(with the important caveat that positive-dimensional faces of* (SDP) *which are 'flat' with respect to the objective function can harbor non-global local minima).*"
>
> — End of Section 3 in [13], mutatis mutandis.

The caveat—the existence or non-existence of non-global local optima, or their potentially adverse effect for local optimization algorithms—was not further discussed. How mild this caveat really is (as stated) is hard to gauge, considering $\mathscr{C}$ can have a continuum of faces.

Contributions

In this paper, we identify settings where the non-convexity of (P) is benign, in the sense that second-order necessary optimality conditions are sufficient for global optimality—an unusual property for a non-convex problem. This paper extends a

[1] The condition on p and m is slightly, but inconsequentially, different in [13].

previous conference paper by the same authors [11]. Our core assumption is as follows.

Assumption 1.1. For a given p such that $\mathscr{M}$ (1.2) is non-empty, constraints on (SDP) defined by $A_1, \ldots, A_m \in \mathbb{S}^{n\times n}$ and $b \in \mathbb{R}^m$ satisfy at least one of the following:

a. $\{A_1Y, \ldots, A_mY\}$ are linearly independent in $\mathbb{R}^{n\times p}$ for all $Y \in \mathscr{M}$; or
b. $\{A_1Y, \ldots, A_mY\}$ span a subspace of constant dimension in $\mathbb{R}^{n\times p}$ for all Y in an open neighborhood of $\mathscr{M}$ in $\mathbb{R}^{n\times p}$.

In either case, let m' denote the dimension of the space spanned by $\{A_1Y, \ldots, A_mY\}$. (By assumption, m' is independent of the choice of $Y \in \mathscr{M}$.)

Under Assumption 1.1, $\mathscr{M}$ is a smooth manifold, which is why we say such an (SDP) is *smooth*. Furthermore, if the assumption holds for several values of p, then m' is the same for all. Formal statements follow; proofs are in Appendix A.

Proposition 1.2. *Under Assumption 1.1, $\mathscr{M}$ is an embedded submanifold of $\mathbb{R}^{n\times p}$ of dimension $np - m'$.*

Proposition 1.3. *If Assumption 1.1 holds for some p, it holds for all $p' \leq p$ such that $\mathscr{M}_{p'}$ is non-empty. Furthermore, if Assumption 1.1a holds for $p = n$, then it holds for all p' such that $\mathscr{M}_{p'}$ is non-empty. In both cases, m' is independent of p.*

Examples of SDPs satisfying Assumption 1.1 are detailed in Section 5 (they all satisfy Assumption 1.1a for $p = n$). The assumption itself is further discussed in Section 6. Our first main result is as follows, where $\operatorname{rank}\mathscr{A}$ can be replaced by m if preferred. Optimality conditions are derived in Section 2.

Theorem 1.4. *Let p be such that $\frac{p(p+1)}{2} > \operatorname{rank}\mathscr{A}$ and such that Assumption 1.1 holds. For almost any cost matrix $C \in \mathbb{S}^{n\times n}$, if $Y \in \mathscr{M}$ satisfies first- and second-order necessary optimality conditions for (P), then Y is globally optimal and $X = YY^\top$ is globally optimal for (SDP).*

The proof combines two intermediate results (Proposition 3.1 and Lemma 3.3 below):

(1) If Y is *column-rank deficient* and satisfies first- and second-order necessary optimality conditions for (P), then it is globally optimal and $X = YY^\top$ is optimal for (SDP); and
(2) If $\frac{p(p+1)}{2} > \operatorname{rank}\mathscr{A}$, then, for almost all C, every Y which satisfies first-order necessary optimality conditions is column-rank deficient.

The first step is a variant of well-known results [12, 13, 17]. The second step is new and crucial, as it allows to formally exclude the existence of spurious local optima, thus resolving the caveat raised by Burer and Monteiro generically in C.

Theorem 1.4 is a statement about the optimization problem itself, not about specific algorithms. If $\mathscr{C}$ is compact, then so is $\mathscr{M}$ and known algorithms for

optimization on manifolds converge to *second-order critical points*,[2] regardless of initialization [10]. Thus, provided p is large enough, for almost any cost matrix C, such algorithms generate sequences which converge to global optima of (P). Each iteration requires a polynomial number of arithmetic operations.

In practice, the algorithm is stopped after a finite number of iterations, at which point one can only guarantee approximate satisfaction of first- and second-order necessary optimality conditions. Ideally, this should lead to a statement of approximate optimality. We are only able to make that statement for large values of p. We state this result informally here, and give a precise statement in Corollary 4.5 below.

Theorem 1.5 (Informal)**.** *Assume $\mathscr{C}$ is compact and Assumption 1.1 holds for $p = n+1$. Then, for any cost matrix $C \in \mathbb{S}^{n\times n}$, if $Y \in \mathscr{M}_{n+1}$ approximately satisfies first- and second-order necessary optimality conditions for* (P)*, then it is approximately globally optimal and $X = YY^\top$ is approximately globally optimal for* (SDP)*, in terms of attained cost value.*

Theorem 1.4 does not exclude the possibility that a zero-measure subset of cost matrices C may pose difficulties. Theorem 1.5 does apply for all cost matrices, but requires a large value of p. A complementary result in this paper, which comes with a more geometric proof, constitutes a refinement of the caveat raised by Burer and Monteiro [13] in the excerpt quoted above. It states that a suboptimal second-order critical point Y must map to a face $\mathscr{F}_{YY^\top}$ of the convex search space $\mathscr{C}$ whose dimension is large (rather than just positive) when p itself is large. The facial structure of $\mathscr{C}$ is discussed in Section 2.2. The following is a consequence of Corollary 2.9 and Theorem 3.4 below.

Theorem 1.6. *Let Assumption 1.1 hold for some p. Let $Y \in \mathscr{M}$ be a second-order critical point of* (P)*. If* $\operatorname{rank}(Y) < p$*, or if* $\operatorname{rank}(Y) = p$ *and* $\dim \mathscr{F}_{YY^\top} < \frac{p(p+1)}{2} - m' + p$*, then Y is globally optimal for* (P) *and $X = YY^\top$ is globally optimal for* (SDP)*.*

Combining this theorem with bounds on the dimension of faces of $\mathscr{C}$ allows us to conclude the optimality of second-order critical points for *all* cost matrices C, with bounds on p that are smaller than n. Implications of these theorems for examples of SDPs are treated in Section 5, including the trust-region subproblem, Max-Cut and Orthogonal-Cut.

Notation

$\mathbb{S}^{n\times n}$ is the set of real, symmetric matrices of size n. A symmetric matrix X is positive semidefinite ($X \succeq 0$) if and only if $u^\top X u \geq 0$ for all $u \in \mathbb{R}^n$. For matrices A, B, the standard Euclidean inner product is $\langle A, B\rangle = \operatorname{Tr}(A^\top B)$. The associated

[2] Points which satisfy first- and second-order necessary optimality conditions. Compactness of $\mathscr{C}$ ensures a minimum is attained in (P), hence also that second-order critical points exist.

(Frobenius) norm is $\|A\| = \sqrt{\langle A, A\rangle}$. Id is the identity operator and I_n is the identity matrix of size n. The variable $m' \leq m$ is defined in Assumption 1.1. The adjoint of $\mathscr{A}$ is $\mathscr{A}^*$, such that $\mathscr{A}^*(y) = y_1 A_1 + \cdots + y_m A_m$.

2 Geometry and optimality conditions

We first discuss the smooth geometry of (P) and the convex geometry of (SDP), as well as optimality conditions for both.

2.1 For the non-convex problem (P)

Endow $\mathbb{R}^{n\times p}$ with the classical Euclidean metric $\langle U_1, U_2\rangle = \mathrm{Tr}(U_1^\top U_2)$, corresponding to the Frobenius norm: $\|U\|^2 = \langle U, U\rangle$. As stated in Proposition 1.2, under Assumption 1.1 for a given p, the search space $\mathscr{M}$ of (P) defined in (1.2) is a submanifold of $\mathbb{R}^{n\times p}$ of dimension $\dim \mathscr{M} = np - m'$. Furthermore, the tangent space to $\mathscr{M}$ at Y is a subspace of $\mathbb{R}^{n\times p}$ obtained by linearizing the equality constraints.

Lemma 2.1. *Under Assumption 1.1, the tangent space at Y to $\mathscr{M}$, $\mathrm{T}_Y\mathscr{M}$, obeys*

$$\begin{aligned} \mathrm{T}_Y\mathscr{M} &= \left\{\dot{Y} \in \mathbb{R}^{n\times p} : \mathscr{A}(\dot{Y}Y^\top + Y\dot{Y}^\top) = 0\right\} \\ &= \left\{\dot{Y} \in \mathbb{R}^{n\times p} : \langle A_i Y, \dot{Y}\rangle = 0 \text{ for } i = 1, \ldots, m\right\}. \end{aligned} \tag{2.1}$$

Proof. By definition, $\dot{Y} \in \mathbb{R}^{n\times p}$ is a tangent vector to $\mathscr{M}$ at Y if and only if there exists a curve $\gamma\colon \mathbb{R} \to \mathscr{M}$ such that $\gamma(0) = Y$ and $\dot{\gamma}(0) = \dot{Y}$, where $\dot{\gamma}$ is the derivative of γ. Then, $\mathscr{A}(\gamma(t)\gamma(t)^\top) = b$ for all t. Differentiating on both sides yields $\mathscr{A}(\dot{\gamma}(t)\gamma(t)^\top + \gamma(t)\dot{\gamma}(t)^\top) = 0$. Evaluating at $t = 0$ confirms $\mathrm{T}_Y\mathscr{M}$ is included in the subspace (2.1). To conclude, use the fact that both subspaces have the same dimension under Assumption 1.1, by Proposition 1.2. □

Each tangent space is equipped with a restriction of the metric $\langle\cdot,\cdot\rangle$, thus making $\mathscr{M}$ a *Riemannian* submanifold of $\mathbb{R}^{n\times p}$. From (2.1), it is clear that the $A_i Y$ span the normal space at Y:

$$\mathrm{N}_Y\mathscr{M} = \mathrm{span}\{A_1 Y, \ldots, A_m Y\}. \tag{2.2}$$

An important tool is the orthogonal projector $\mathrm{Proj}_Y\colon \mathbb{R}^{n\times p} \to \mathrm{T}_Y\mathscr{M}$:

$$\mathrm{Proj}_Y Z = \operatorname*{argmin}_{\dot{Y}\in \mathrm{T}_Y\mathscr{M}} \|\dot{Y} - Z\|. \tag{2.3}$$

We have the following lemma to characterize it.

Lemma 2.2. *Under Assumption 1.1, the orthogonal projector is given by:*

$$\mathrm{Proj}_Y Z = Z - \mathscr{A}^*\left(G^\dagger \mathscr{A}(ZY^\top)\right) Y,$$

where $\mathscr{A}^*\colon \mathbb{R}^m \to \mathbb{S}^{n\times n}$ is the adjoint of $\mathscr{A}$, $G = G(Y)$ is a Gram matrix defined by $G_{ij} = \langle A_i Y, A_j Y\rangle$, and $G^\dagger$ denotes the Moore–Penrose pseudo-inverse of G.

Furthermore, if $Y \mapsto Z(Y)$ is differentiable in an open neighborhood of $\mathcal{M}$ in $\mathbb{R}^{n\times p}$, then $Y \mapsto \mathrm{Proj}_Y Z(Y)$ is differentiable at all Y in $\mathcal{M}$.

Proof. Orthogonal projection is along the normal space, so that $\mathrm{Proj}_Y Z \in \mathrm{T}_Y\mathcal{M}$ and $Z - \mathrm{Proj}_Y Z \in \mathrm{N}_Y\mathcal{M}$ (2.2). From the latter we infer there exists $\mu \in \mathbb{R}^m$ such that

$$Z - \mathrm{Proj}_Y Z = \sum_{i=1}^{m} \mu_i A_i Y = \mathscr{A}^*(\mu)Y,$$

since the adjoint of $\mathscr{A}$ is $\mathscr{A}^*(\mu) = \mu_1 A_1 + \cdots + \mu_m A_m$ by definition. Multiply on the right by $Y^\top$ and apply $\mathscr{A}$ to obtain

$$\mathscr{A}(ZY^\top) = \mathscr{A}(\mathscr{A}^*(\mu)YY^\top),$$

where we used $\mathscr{A}(\mathrm{Proj}_Y(Z)Y^\top) = 0$ since $\mathrm{Proj}_Y(Z) \in \mathrm{T}_Y\mathcal{M}$. The right-hand side expands into

$$\mathscr{A}(\mathscr{A}^*(\mu)YY^\top)_i = \left\langle A_i, \sum_{j=1}^{m} \mu_j A_j YY^\top \right\rangle = \sum_{j=1}^{m} \left\langle A_i Y, A_j Y \right\rangle \mu_j = (G\mu)_i.$$

Thus, any μ satisfying $G\mu = \mathscr{A}(ZY^\top)$ will do. Without loss of generality, we pick the smallest norm solution: $\mu = G^\dagger \mathscr{A}(ZY^\top)$. The function $Y \mapsto G^\dagger$ is continuous and differentiable at $Y \in \mathcal{M}$ provided G has constant rank in an open neighborhood of Y in $\mathbb{R}^{n\times p}$ [16, Thm. 4.3], which is the case under Assumption 1.1. □

Problem (P) minimizes

$$g(Y) = \langle CY, Y \rangle \tag{2.4}$$

over $\mathcal{M}$, where g is defined over $\mathbb{R}^{n\times p}$. Its classical (Euclidean) gradient at Y is $\nabla g(Y) = 2CY$. The Riemannian gradient of g at Y, $\mathrm{grad}\, g(Y)$, is defined as the unique tangent vector at Y such that, for all tangent $\dot{Y}$, $\langle \mathrm{grad}\, g(Y), \dot{Y} \rangle = \langle \nabla g(Y), \dot{Y} \rangle$. This is given by the projection of the classical gradient onto the tangent space [3, eq. (3.37)]:

$$\mathrm{grad}\, g(Y) = \mathrm{Proj}_Y\left(\nabla g(Y)\right) = 2\,\mathrm{Proj}_Y\left(CY\right) = 2\left(C - \mathscr{A}^*\left(G^\dagger \mathscr{A}(CYY^\top)\right)\right)Y.$$

This motivates the definition of S as follows, with $G_{ij} = \left\langle A_i Y, A_j Y \right\rangle$:

$$S = S(Y) = S(YY^\top) = C - \mathscr{A}^*(\mu), \qquad \text{with} \qquad \mu = G^\dagger \mathscr{A}(CYY^\top). \tag{2.5}$$

This is indeed well defined since G_{ij} is a function of $YY^\top$. We get a convenient formula for the gradient:

$$\mathrm{grad}\, g(Y) = 2SY. \tag{2.6}$$

In the sequel, S will play a major role.

Turning toward second-order derivatives, the Riemannian Hessian of g at Y is a symmetric operator on the tangent space at Y obtained as the projection of the derivative of the Riemannian gradient vector field [3, eq. (5.15)]. The latter

is indeed differentiable owing to Lemma 2.2. With D denoting classical Fréchet differentiation, writing $S = S(Y)$ and $\dot{S} = \mathrm{D}(Y \mapsto S(Y))(Y)[\dot{Y}]$,

$$\mathrm{Hess}\, g(Y)[\dot{Y}] = \mathrm{Proj}_Y\big(\mathrm{Dgrad}\, g(Y)[\dot{Y}]\big) = 2\mathrm{Proj}_Y\big(\dot{S}Y + S\dot{Y}\big) = 2\mathrm{Proj}_Y\big(S\dot{Y}\big). \tag{2.7}$$

The projection of $\dot{S}Y$ vanishes because $\dot{S} = \mathscr{A}^*(\nu)$ for some $\nu \in \mathbb{R}^m$ so that $\dot{S}Y = \sum_{i=1}^m \nu_i A_i Y$ is in the normal space at Y (2.2).

These differentials are relevant for their role in necessary optimality conditions of (P).

Definition 2.3. $Y \in \mathscr{M}$ is a *(first-order) critical point* for (P) if

$$\frac{1}{2}\mathrm{grad}\, g(Y) = SY = 0, \tag{2.8}$$

where S is a function of Y (2.5). If furthermore $\mathrm{Hess}\, g(Y) \succeq 0$, that is (using the fact that Proj_Y is self-adjoint),

$$\forall \dot{Y} \in \mathrm{T}_Y\mathscr{M}, \quad \frac{1}{2}\langle \dot{Y}, \mathrm{Hess}\, g(Y)[\dot{Y}]\rangle = \langle \dot{Y}, S\dot{Y}\rangle \geq 0, \tag{2.9}$$

then Y is a *second-order critical point* for (P).

Proposition 2.4. *Under Assumption 1.1, all local (and global) minima of* (P) *are second-order critical points.*

Proof. These are standard necessary optimality conditions on manifolds, see [31, Rem. 4.2 and Cor. 4.2]. □

Thus, the central role of S in necessary optimality conditions for the non-convex problem is clear. Its role for the convex problem is elucidated next.

2.2 For the convex problem (SDP)

The search space of (SDP) is the convex set $\mathscr{C}$ defined in (1.1), assumed non-empty. Geometry-wise, we are primarily interested in the facial structure of $\mathscr{C}$ [27, §18].

Definition 2.5. A *face* of $\mathscr{C}$ is a convex subset $\mathscr{F}$ of $\mathscr{C}$ such that every (closed) line segment in $\mathscr{C}$ with a relative interior point in $\mathscr{F}$ has both endpoints in $\mathscr{F}$. The empty set and $\mathscr{C}$ itself are faces of $\mathscr{C}$.

For example, the non-empty faces of a cube are its vertices, edges, facets and the cube itself. By [27, Thm. 18.2], the collection of relative interiors of the non-empty faces forms a partition of $\mathscr{C}$ (the relative interior of a singleton is the singleton). That is, each $X \in \mathscr{C}$ is in the relative interior of exactly one face of $\mathscr{C}$, called $\mathscr{F}_X$. The dimension of a face is the dimension of the lowest dimensional affine subspace which contains that face. Of particular interest are the zero-dimensional faces of $\mathscr{C}$ (singletons).

Definition 2.6. $X \in \mathscr{C}$ is an *extreme point* of $\mathscr{C}$ if $\dim \mathscr{F}_X = 0$.

In other words, X is extreme if it does not lie on an open line segment included in $\mathscr{C}$. If $\mathscr{C}$ is compact, it is the convex hull of its extreme points [27, Cor. 18.5.1]. Of importance to us, if $\mathscr{C}$ is compact, (SDP) always attains its minimum at one of its extreme points since the linear cost function of (SDP) is (a fortiori) concave [27, Cor. 32.3.2]. The faces of $\mathscr{C}$ can be described explicitly as follows. The proof is in Appendix B.

Proposition 2.7. *Let $X \in \mathscr{C}$ have rank p and let $\mathscr{F}_X$ be its associated face (that is, X is in the relative interior of $\mathscr{F}_X$.) Then, with $Y \in \mathscr{M}_p$ such that $X = YY^\top$,*

$$\mathscr{F}_X = \left\{ X' = Y(I_p + A)Y^\top : A \in \ker \mathscr{L}_X \text{ and } I_p + A \succeq 0 \right\}, \tag{2.10}$$

where $\mathscr{L}_X : \mathbb{S}^{p\times p} \to \mathbb{R}^m$ is defined by:

$$\mathscr{L}_X(A) = \mathscr{A}(YAY^\top) = \left(\left\langle Y^\top A_1 Y, A \right\rangle, \ldots, \left\langle Y^\top A_m Y, A \right\rangle \right)^\top. \tag{2.11}$$

Thus, the dimension of $\mathscr{F}_X$ is the dimension of the kernel of $\mathscr{L}_X$. Since the dimension of $\mathbb{S}^{p\times p}$ is $\frac{p(p+1)}{2}$ and $\operatorname{rank}(\mathscr{L}_X) \leq m'$, the rank-nullity theorem gives a lower bound:

$$\dim \mathscr{F}_X = \frac{p(p+1)}{2} - \operatorname{rank} \mathscr{L}_X \geq \frac{p(p+1)}{2} - m'. \tag{2.12}$$

For extreme points, $\dim \mathscr{F}_X = 0$; then, $\frac{p(p+1)}{2} = \operatorname{rank} \mathscr{L}_X \leq m'$. Solving for p (the rank of X) shows extreme points have small rank, namely,

$$\dim \mathscr{F}_X = 0 \implies \operatorname{rank}(X) \leq p^* \triangleq \frac{\sqrt{8m'+1}-1}{2}. \tag{2.13}$$

Since (SDP) attains its minimum at an extreme point for compact $\mathscr{C}$, we recover the known fact that one of the optima has rank at most p^*. This approach to proving that statement is well known [24, Thm. 2.1].

Optimality conditions for (SDP) are easily stated once S (2.5) is introduced—it acts as a dual certificate, known in closed form owing to the underlying smooth geometry of $\mathscr{M}$. We need a first general fact about SDPs (Assumption 1.1 is not required.)

Proposition 2.8. *Let $X \in \mathscr{C}$ and let $S = C - \mathscr{A}^*(\nu)$ for some $\nu \in \mathbb{R}^m$ (as is the case in (2.5) for example). If $S \succeq 0$ and $\langle S, X \rangle = 0$, then X is optimal for (SDP).*

Proof. First, use $S \succeq 0$: for any $X' \in \mathscr{C}$, since $X' \succeq 0$ and $\mathscr{A}(X) = \mathscr{A}(X')$,

$$0 \leq \langle S, X' \rangle = \langle C, X' \rangle - \langle \mathscr{A}^*(\nu), X' \rangle = \langle C, X' \rangle - \langle \nu, \mathscr{A}(X) \rangle.$$

Concentrating on the last term, use $\langle S, X \rangle = 0$:

$$\langle \nu, \mathscr{A}(X) \rangle = \langle \mathscr{A}^*(\nu), X \rangle = \langle C, X \rangle - \langle S, X \rangle = \langle C, X \rangle.$$

Hence, $\langle C, X \rangle \leq \langle C, X' \rangle$, which shows X is optimal. □

Since (SDP) is a relaxation of (P), this leads to a corollary of prime importance.

Corollary 2.9. *Let Assumption 1.1 hold for some p. If Y is a critical point for (P) as defined by (2.8) and S (2.5) is positive semidefinite, then $X = YY^\top$ is globally optimal for (SDP) and Y is globally optimal for (P).*

Proof. Since Y is a critical point, $SY = 0$; thus, $\langle S, X\rangle = 0$ and Proposition 2.8 applies. □

A converse of Proposition 2.8 holds under additional conditions which are satisfied by all examples in Section 5. Thus, for those cases, for a critical point Y, $YY^\top$ is optimal if and only if S is positive semidefinite. We state it here for completeness (this result is not needed in the sequel.)

Proposition 2.10. *Let $X \in \mathscr{C}$ be a global optimum of (SDP) and assume strong duality holds. Let Assumption 1.1a hold with $p = \operatorname{rank}(X)$. Then, $S \succeq 0$ and $\langle S, X\rangle = 0$, where $S = S(X)$ is as in (2.5).*

Proof. Consider the dual of (SDP):

$$\text{(DSDP)} \qquad \max_{\nu \in \mathbb{R}^m} \langle b, \nu\rangle \text{ subject to } C - \mathscr{A}^*(\nu) \succeq 0.$$

Since we assume strong duality and X is optimal, there exists ν optimal for the dual such that $\langle C, X\rangle = \langle b, \nu\rangle$. Using $\langle b, \nu\rangle = \langle \mathscr{A}(X), \nu\rangle = \langle X, \mathscr{A}^*(\nu)\rangle$, this implies

$$0 = \langle C, X\rangle - \langle b, \nu\rangle = \langle C - \mathscr{A}^*(\nu), X\rangle.$$

Since both $C - \mathscr{A}^*(\nu)$ and X are positive semidefinite, $(C - \mathscr{A}^*(\nu))X = 0$. As a result, by definition of μ and G (2.5),

$$\mu = G^\dagger \mathscr{A}(CX) = G^\dagger \mathscr{A}(\mathscr{A}^*(\nu)X) = G^\dagger G\nu = \nu,$$

where we used $G^\dagger = G^{-1}$ under Assumption 1.1a and

$$(G\nu)_i = \sum_j G_{ij}\nu_j = \sum_j \langle A_i, A_j X\rangle \nu_j = \langle A_i, \mathscr{A}^*(\nu)X\rangle = \mathscr{A}(\mathscr{A}^*(\nu)X)_i.$$

Thus, $S = C - \mathscr{A}^*(\mu) = C - \mathscr{A}^*(\nu)$ has the desired properties. This concludes the proof, and shows uniqueness of the dual certificate. □

3 Optimality of second-order critical points

We aim to show that second-order critical points of (P) are global optima, provided p is sufficiently large. To this end, we first recall a known result about *rank-deficient* second-order critical points.[3]

Proposition 3.1. *Let Assumption 1.1 hold for some p and let $Y \in \mathscr{M}$ be a second-order critical point for (P). If $\operatorname{rank}(Y) < p$, then $S(Y) \succeq 0$ so that Y is globally optimal for (P) and so is $X = YY^\top$ for (SDP).*

[3] Optimality of rank deficient *local optima* is shown (under different assumptions) in [13, 17], with the proof in [17] actually only requiring second-order criticality.

Proof. The proof parallels the one in [17]. By Corollary 2.9, it is sufficient to show that $S = S(Y)$ (2.5) is positive semidefinite. Since $\operatorname{rank}(Y) < p$, there exists $z \in \mathbb{R}^p$ such that $z \neq 0$ and $Yz = 0$. Furthermore, for all $x \in \mathbb{R}^n$, the matrix $\dot{Y} = xz^\top$ is such that $Y\dot{Y}^\top = 0$. In particular, $\dot{Y}$ is a tangent vector at Y (2.1). Since Y is second-order critical, inequality (2.9) holds, and here simplifies to:

$$0 \leq \left\langle \dot{Y}, S\dot{Y} \right\rangle = \left\langle xz^\top, Sxz^\top \right\rangle = \|z\|^2 \cdot x^\top S x.$$

This holds for all $x \in \mathbb{R}^n$. Thus, S is positive semidefinite. □

Corollary 3.2. *Let Assumption 1.1 hold for some $p \geq n$. Then, any second-order critical point $Y \in \mathscr{M}$ of (P) is globally optimal, and $X = YY^\top$ is globally optimal for (SDP).*

Proof. For $p > n$ (with $p = n+1$ being the most interesting case), points in $\mathscr{M}$ are necessarily column-rank deficient, so that the corollary follows from Proposition 3.1. For $p = n$, if Y is rank deficient, use the same proposition. Otherwise, Y is invertible and $SY = 0$ (2.8) implies $S = 0$, which is a fortiori positive semidefinite. By (2.5), this only happens if $C = \mathscr{A}^*(\mu)$ for some μ, in which case the cost function $\langle C, X \rangle = \langle \mathscr{A}^*(\mu), X \rangle = \langle \mu, b \rangle$ is constant over $\mathscr{C}$. □

In this paper, we aim to secure optimality of second-order critical points for p less than n. As indicated by Proposition 3.1, the sole concern in that respect is the possible existence of *full-rank* second-order critical points. We first give a result which excludes the existence of full-rank first-order critical points (thus, a fortiori of second-order critical points) for *almost all* cost matrices C, provided p is sufficiently large. The argument is by dimensionality counting.

Lemma 3.3. *Let p be such that $\frac{p(p+1)}{2} > \operatorname{rank} \mathscr{A}$ and such that Assumption 1.1 holds. Then, for almost all C, all critical points of (P) are column-rank deficient.*

Proof. Let $Y \in \mathscr{M}$ be a critical point for (P). By the definition of $S(Y) = C - \mathscr{A}^*(\mu(Y))$ (2.5) and the first-order condition $S(Y)Y = 0$ (2.8), we have

$$\operatorname{rank} Y \leq \operatorname{null}(C - \mathscr{A}^*(\mu(Y))) \leq \max_{\nu \in \mathbb{R}^m} \operatorname{null}(C - \mathscr{A}^*(\nu)), \tag{3.1}$$

where null denotes the nullity (dimension of the kernel). This first step in the proof is inspired by [30, Thm. 3]. If the right-hand side evaluates to ℓ, then there exists ν and $M = C - \mathscr{A}^*(\nu)$ such that $\operatorname{null}(M) = \ell$. Writing $C = M + \mathscr{A}^*(\nu)$, we find that

$$C \in \mathscr{N}_\ell + \operatorname{im}(\mathscr{A}^*), \tag{3.2}$$

where $\mathscr{N}_\ell$ denotes the set of symmetric matrices of size n with nullity ℓ and the $+$ is a set-sum. The set $\mathscr{N}_\ell$ has dimension

$$\dim \mathscr{N}_\ell = \frac{n(n+1)}{2} - \frac{\ell(\ell+1)}{2}. \tag{3.3}$$

Assume the right-hand side of (3.1) evaluates to p or more. Then, a fortiori,

$$C \in \bigcup_{\ell=p,\ldots,n} \mathscr{N}_\ell + \operatorname{im}(\mathscr{A}^*). \tag{3.4}$$

The set on the right-hand side contains all "bad" matrices C, that is, those for which (3.1) offers no information about the rank of Y. The dimension of that set is bounded as follows, using the fact that the dimension of a finite union is at most the maximal dimension, and the dimension of a finite sum of sets is at most the sum of the set dimensions:

$$\begin{aligned}\dim\left(\bigcup_{\ell=p,\ldots,n} \mathscr{N}_\ell + \operatorname{im}(\mathscr{A}^*)\right) &\leq \dim\left(\mathscr{N}_p + \operatorname{im}(\mathscr{A}^*)\right) \\ &\leq \frac{n(n+1)}{2} - \frac{p(p+1)}{2} + \operatorname{rank}\mathscr{A}.\end{aligned}$$

Since $C \in \mathbb{S}^{n\times n}$ lives in a space of dimension $\frac{n(n+1)}{2}$, almost no C verifies (3.4) if

$$\frac{n(n+1)}{2} - \frac{p(p+1)}{2} + \operatorname{rank}\mathscr{A} < \frac{n(n+1)}{2}.$$

Hence, if $\frac{p(p+1)}{2} > \operatorname{rank}\mathscr{A}$, for almost all C, critical points have $\operatorname{rank}(Y) < p$. □

Theorem 1.4 follows as an easy corollary of Proposition 3.1 and Lemma 3.3.

In order to make a statement valid for *all* C, we further explore the implications of second-order criticality on the definiteness of S. For large p (though still smaller than n), we expect full-rank second-order critical points should indeed be optimal. The intuition is as follows. If $Y \in \mathscr{M}$ is a second-order critical point of rank p, then, by (2.8), $SY = 0$ which implies S has a kernel of dimension at least p. Furthermore, by (2.9), S has "positive curvature" along directions in $\mathrm{T}_Y\mathscr{M}$, whose dimension grows with p. Overall, the larger p, the more conditions force S to have nonnegative eigenvalues. The main concern is to avoid double counting, as the two conditions are redundant along certain directions: this is where the facial structure of $\mathscr{C}$ comes into play.

The following theorem refines this intuition. We use $\otimes$ for Kronecker products and vec to vectorize a matrix by stacking its columns on top of each other, so that $\operatorname{vec}(AXB) = (B^\top \otimes A)\operatorname{vec}(X)$. A real number a is rounded down as $\lfloor a \rfloor$.

Theorem 3.4. *Let p be such that Assumption 1.1 holds. Let $Y \in \mathscr{M}$ be a second-order critical point for (P). The matrix $X = YY^\top$ belongs to the relative interior of the face $\mathscr{F}_X$ (2.10). If $\operatorname{rank}(Y) = p$, then $S = S(X)$ (2.5) has at most*

$$\left\lfloor \frac{\dim \mathscr{F}_X - \Delta}{p} \right\rfloor \tag{3.5}$$

negative eigenvalues, where

$$\Delta = \frac{p(p+1)}{2} - m'. \tag{3.6}$$

In particular, if $\dim \mathscr{F}_X < \Delta + p$*, then* S *is positive semidefinite and both* X *and* Y *are globally optimal.*

Proof. Consider the subspace $\mathrm{vec}(\mathrm{T}_Y\mathscr{M})$ of vectorized tangent vectors at Y: it has dimension $k \triangleq \dim \mathscr{M}$. Pick $U \in \mathbb{R}^{np\times k}$ with columns forming an orthonormal basis for that subspace: $U^\top U = I_k$. Then, $U^\top(I_p \otimes S)U$ has the same spectrum as $\frac{1}{2}\mathrm{Hess}\, g(Y)$. Indeed, for all $\dot{Y} \in \mathrm{T}_Y\mathscr{M}$ there exists $x \in \mathbb{R}^k$ such that $\mathrm{vec}(\dot{Y}) = Ux$, and, by (2.9),

$$\frac{1}{2}\langle \dot{Y}, \mathrm{Hess}\, g(Y)[\dot{Y}]\rangle = \langle \dot{Y}, S\dot{Y}\rangle = \langle Ux, (I_p \otimes S)Ux\rangle = \langle x, U^\top(I_p \otimes S)Ux\rangle.$$

In particular, $U^\top(I_p \otimes S)U$ is positive semidefinite since Y is second-order critical.

Let $V \in \mathbb{R}^{np\times p^2}, V^\top V = I_{p^2}$, have columns forming an orthonormal basis of the space spanned by the vectors $\mathrm{vec}(YR)$ for $R \in \mathbb{R}^{p\times p}$: such V exists because $\mathrm{rank}(Y) = p$. Indeed, $\mathrm{vec}(YR) = (I_p \otimes Y)\mathrm{vec}(R)$ and $I_p \otimes Y \in \mathbb{R}^{np\times p^2}$ then has full rank p^2. Since Y is a critical point, $SY = 0$ by (2.8), which implies $(I_p \otimes S)V = 0$.

Let k' denote the dimension of the space spanned by the columns of both U and V, and let $W \in \mathbb{R}^{np\times k'}, W^\top W = I_{k'}$, be an orthonormal basis for this space. It follows that $M = W^\top(I_p \otimes S)W$ is positive semidefinite. Indeed, for any z, there exist x, y such that $Wz = Ux + Vy$. Hence, $z^\top Mz = x^\top U^\top(I_p \otimes S)Ux \geq 0$.

Let $\lambda_0 \leq \cdots \leq \lambda_{n-1}$ denote the eigenvalues of S, and let $\tilde{\lambda}_0 \leq \cdots \leq \tilde{\lambda}_{np-1}$ denote the eigenvalues of $I_p \otimes S$. The latter are simply the eigenvalues of S repeated p times, thus: $\tilde{\lambda}_i = \lambda_{\lfloor i/p\rfloor}$. Let $\mu_0 \leq \cdots \leq \mu_{k'-1}$ denote the eigenvalues of M. The Cauchy interlacing theorem states that, for all i,

$$\tilde{\lambda}_i \leq \mu_i \leq \tilde{\lambda}_{i+np-k'}. \tag{3.7}$$

In particular, since $M \succeq 0$, we have $0 \leq \mu_0 \leq \lambda_{\lfloor (np-k')/p\rfloor}$. It remains to determine k'.

From Proposition 1.2, recall that $k = \dim \mathscr{M} = np - m'$. We now investigate how many new dimensions V adds to U. All matrices $R \in \mathbb{R}^{p\times p}$ admit a unique decomposition as

$$R = R_{\mathrm{skew}} + R_{\ker \mathscr{L}} + R_{(\ker \mathscr{L})^\perp},$$

where R_{skew} is skew-symmetric, $R_{\ker \mathscr{L}}$ is in the kernel of $\mathscr{L}_X$ (2.11) and $R_{(\ker \mathscr{L})^\perp}$ is in the orthogonal complement of the latter in $\mathbb{S}^{p\times p}$. Recalling the definition of tangent vectors (2.1), it is clear that $\dot{Y} = YR_{\mathrm{skew}}$ is tangent. Similarly, $\dot{Y} = YR_{\ker \mathscr{L}}$ is tangent because of the definition of $\mathscr{L}_X$ (2.11). Thus, vectorized versions of these are already in the span of U. On the other hand, by definition, $YR_{(\ker \mathscr{L})^\perp}$ is not tangent at Y (if it is nonzero). This raises k' (the rank of W) by $\dim (\ker \mathscr{L}_X)^\perp = \frac{p(p+1)}{2} - \dim \ker \mathscr{L}_X$. Since $\dim \ker \mathscr{L}_X = \dim \mathscr{F}_X$, we have:

$$k' = np - m' + \frac{p(p+1)}{2} - \dim \mathscr{F}_X = np + \Delta - \dim \mathscr{F}_X. \tag{3.8}$$

Thus, $np - k' = \dim \mathscr{F}_X - \Delta$. Combine with $\lambda_{\lfloor (np-k')/p \rfloor} \geq 0$ to conclude. □

Theorem 1.6 follows easily from Corollary 2.9 and Theorem 3.4.

Remark 3.5. What does it take for a second-order critical point $Y \in \mathscr{M}$ to be suboptimal? For local optima, the quote from Burer and Monteiro [13, §3] in the introduction readily states that Y must have rank p, and the face $\mathscr{F}_X$ (with $X = YY^\top$) must be positive dimensional and such that the cost function $\langle C, X \rangle$ is constant over $\mathscr{F}_X$. Here, under Assumption 1.1 for p, Theorem 3.4 states that if Y is second-order critical and is suboptimal, then $\mathscr{F}_X$ must have dimension $\Delta + p$ or higher. Since (2.12) suggests generic faces at rank p have dimension Δ, this further shows thats suboptimal second-order critical points, if they exist, can only occur if the cost function is constant over a high-dimensional face of $\mathscr{C}$.

To use Theorem 3.4 in a particular application, one needs to obtain upper bounds on the dimensions of faces of $\mathscr{C}$. We follow this path for a number of examples in Section 5.

4 Near optimality of near second-order critical points

Under Assumption 1.1, problem (P) is an example of smooth optimization over a smooth manifold. This suggests using *Riemannian optimization* to solve it [3], as already proposed by Journée et al. [17] in a similar context. Importantly, known algorithms—in particular, the *Riemannian trust-region method* (RTR)—converge to second-order critical points regardless of initialization [2]. We state here a recent computational result to that effect [10].

Proposition 4.1. *Under Assumption 1.1, if $\mathscr{C}$ is compact, RTR initialized with any $Y_0 \in \mathscr{M}$ produces in $\mathscr{O}(1/\varepsilon_g^2 \varepsilon_H + 1/\varepsilon_H^3)$ iterations a point $Y \in \mathscr{M}$ such that*

$$g(Y) \leq g(Y_0), \qquad \|\operatorname{grad} g(Y)\| \leq \varepsilon_g, \qquad \text{and} \qquad \operatorname{Hess} g(Y) \succeq -\varepsilon_H \operatorname{Id},$$

where g (2.4) is the cost function of (P).

Proof. Apply the main results of [10] using the fact that g has locally Lipschitz continuous gradient and Hessian in $\mathbb{R}^{n \times p}$ and $\mathscr{M}$ is a compact submanifold of $\mathbb{R}^{n \times p}$. □

Importantly, only a finite number of iterations of any algorithm can be run in practice, so that only approximate second-order critical points can be computed. Thus, it is of interest to establish whether approximate second-order critical points are also approximately optimal. As a first step, we give a soft version of Corollary 2.9. We remark that the condition $I_n \in \operatorname{im} \mathscr{A}^*$ is satisfied in all examples of Section 5.

Lemma 4.2. *Let Assumption 1.1 hold for some p and assume $\mathscr{C}$ (1.1) is compact. For any Y on the manifold $\mathscr{M}$, if $\|\operatorname{grad} g(Y)\| \leq \varepsilon_g$ and $S(Y) \succeq -\frac{\varepsilon_H}{2} I_n$, then the optimality gap at Y with respect to (SDP) is bounded as*

$$0 \leq 2(g(Y) - f^\star) \leq \varepsilon_H R + \varepsilon_g \sqrt{R}, \tag{4.1}$$

where $f^\star$ is the optimal value of (SDP) and $R = \max_{X \in \mathscr{C}} \operatorname{Tr}(X) < \infty$ measures the size of $\mathscr{C}$.

If $I_n \in \operatorname{im}(\mathscr{A}^*)$, the right-hand side of (4.1) can be replaced by $\varepsilon_H R$. This holds in particular if all $X \in \mathscr{C}$ have same trace and $\mathscr{C}$ has a relative interior point (Slater condition).

Proof. By assumption on $S(Y) = C - \mathscr{A}^*(\mu(Y))$ (2.5) with $\mu(Y) = G^\dagger \mathscr{A}(CYY^\top)$,

$$\begin{aligned} \forall X' \in \mathscr{C}, \quad -\frac{\varepsilon_H}{2}\operatorname{Tr}(X') \leq \langle S(Y), X' \rangle &= \langle C, X' \rangle - \langle \mathscr{A}^*(\mu(Y)), X' \rangle \\ &= \langle C, X' \rangle - \langle \mu(Y), b \rangle. \end{aligned}$$

This holds in particular for X' optimal for (SDP). Thus, we may set $\langle C, X' \rangle = f^\star$; and certainly, $\operatorname{Tr}(X') \leq R$. Furthermore,

$$\langle \mu(Y), b \rangle = \langle \mu(Y), \mathscr{A}(YY^\top) \rangle = \langle C - S(Y), YY^\top \rangle = g(Y) - \langle S(Y)Y, Y \rangle.$$

Combining the displayed equations and using $\operatorname{grad} g(Y) = 2S(Y)Y$ (2.8), we find

$$0 \leq 2(g(Y) - f^\star) \leq \varepsilon_H R + \langle \operatorname{grad} g(Y), Y \rangle. \tag{4.2}$$

In general, we do not assume $I_n \in \operatorname{im}(\mathscr{A}^*)$ and we get the result by Cauchy–Schwarz on (4.2) and $\|Y\| = \sqrt{\operatorname{Tr}(YY^\top)} \leq \sqrt{R}$:

$$0 \leq 2(g(Y) - f^\star) \leq \varepsilon_H R + \varepsilon_g \sqrt{R}.$$

But if $I_n \in \operatorname{im}(\mathscr{A}^*)$, then we show that Y is a normal vector at Y, so that it is orthogonal to $\operatorname{grad} g(Y)$. Formally: there exists $\nu \in \mathbb{R}^m$ such that $I_n = \mathscr{A}^*(\nu)$, and

$$\langle \operatorname{grad} g(Y), Y \rangle = \langle \operatorname{grad} g(Y) Y^\top, I_n \rangle = \langle \mathscr{A}(\operatorname{grad} g(Y) Y^\top), \nu \rangle = 0,$$

since $\operatorname{grad} g(Y) \in \mathrm{T}_Y \mathscr{M}$ (2.1). This indeed allows us to simplify (4.2).

To conclude, we show that if $\mathscr{C}$ has a relative interior point X' (that is, $\mathscr{A}(X') = b$ and $X' \succ 0$) and if $\operatorname{Tr}(X)$ is constant for X in $\mathscr{C}$, then $I_n \in \operatorname{im}(\mathscr{A}^*)$. Indeed, $\mathbb{S}^{n \times n} = \operatorname{im}(\mathscr{A}^*) \oplus \ker \mathscr{A}$, so there exist $\nu \in \mathbb{R}^m$ and $M \in \ker \mathscr{A}$ such that $I_n = \mathscr{A}^*(\nu) + M$. Thus, for all X in $\mathscr{C}$,

$$0 = \operatorname{Tr}(X - X') = \left\langle \mathscr{A}^*(\nu) + M, X - X' \right\rangle = \left\langle M, X - X' \right\rangle.$$

This implies M is orthogonal to all $X - X'$. These span $\ker \mathscr{A}$ since X' is interior. Indeed, for any $H \in \ker \mathscr{A}$, since $X' \succ 0$, there exists $t > 0$ such that $X \triangleq X' + tH \succeq 0$ and $\mathscr{A}(X) = b$, so that $X \in \mathscr{C}$. Hence, $M \in \ker \mathscr{A}$ is orthogonal to $\ker \mathscr{A}$. Consequently, $M = 0$ and $I_n = \mathscr{A}^*(\nu)$. □

The lemma above involves a condition on the spectrum of S. Next, we show this condition is satisfied under an assumption on the spectrum of $\operatorname{Hess} g$ and rank deficiency.

Lemma 4.3. *Let Assumption 1.1 hold for some p. If $Y \in \mathscr{M}$ is column-rank deficient and $\operatorname{Hess} g(Y) \succeq -\varepsilon_H \operatorname{Id}$, then $S(Y) \succeq -\frac{\varepsilon_H}{2} I_n$.*

Proof. By assumption, there exists $z \in \mathbb{R}^p$, $\|z\| = 1$ such that $Yz = 0$. Thus, for any $x \in \mathbb{R}^n$, we can form $\dot{Y} = xz^\top$: it is a tangent vector since $Y\dot{Y}^\top = 0$ (2.1), and $\|\dot{Y}\|^2 = \|x\|^2$. Then, condition (2.9) combined with the assumption on $\operatorname{Hess} g(Y)$ tells us

$$-\varepsilon_H \|x\|^2 \leq \langle \dot{Y}, \operatorname{Hess} g(Y)[\dot{Y}] \rangle = 2\langle \dot{Y}, S\dot{Y} \rangle = 2\langle xz^\top zx^\top, S \rangle = 2x^\top Sx.$$

This holds for all $x \in \mathbb{R}^n$, hence $S \succeq -\frac{\varepsilon_H}{2} I_n$ as required. □

We now combine the two previous lemmas to form a soft optimality statement.

Theorem 4.4. *Assume $\mathscr{C}$ is compact and let $R < \infty$ be the maximal trace of any X feasible for (SDP). For some p, let Assumption 1.1 hold for both p and $p+1$. For any $Y \in \mathscr{M}_p$, form $\tilde{Y} = [Y | 0_{n\times 1}]$ in $\mathscr{M}_{p+1}$. The optimality gap at Y is bounded as*

$$0 \leq 2(g(Y) - f^\star) \leq \sqrt{R}\|\operatorname{grad} g(Y)\| - R\lambda_{\min}(\operatorname{Hess} g(\tilde{Y})). \tag{4.3}$$

If all $X \in \mathscr{C}$ have the same trace R and there exists a positive definite feasible X, then the bound

$$0 \leq 2(g(Y) - f^\star) \leq -R\lambda_{\min}(\operatorname{Hess} g(\tilde{Y})) \tag{4.4}$$

holds. If $p > n$, the bounds hold with $\tilde{Y} = Y$ (and Assumption 1.1 only needs to hold for p.)

Proof. Since $\tilde{Y}\tilde{Y}^\top = YY^\top$, $S(\tilde{Y}) = S(Y)$; in particular, we have $g(\tilde{Y}) = g(Y)$ and $\|\operatorname{grad} g(\tilde{Y})\| = \|\operatorname{grad} g(Y)\|$. Since $\tilde{Y}$ has deficient column rank, apply Lemmas 4.2 and 4.3. For $p > n$, there is no need to form $\tilde{Y}$ as Y itself necessarily has deficient column rank. □

This works well with Proposition 4.1. Indeed, equation (4.3) also implies the following:

$$\lambda_{\min}(\operatorname{Hess} g(\tilde{Y})) \leq -\frac{2(g(Y) - f^\star) - \sqrt{R}\|\operatorname{grad} g(Y)\|}{R}.$$

That is, an approximate critical point Y in $\mathscr{M}_p$ which is far from optimal (for (SDP)) maps to a comfortably-escapable approximate saddle point $\tilde{Y}$ in $\mathscr{M}_{p+1}$. This can be helpful for the development of optimization algorithms.

For $p = n+1$, the bound in Theorem 4.4 can be controlled a priori: approximate second-order critical points are approximately optimal, for any C.[4]

[4] With $p = n+1$, problem (P) is no longer lower dimensional than (SDP), but retains the advantage of not involving a positive semidefiniteness constraint.

Corollary 4.5. *Assume $\mathscr{C}$ is compact. Let Assumption 1.1 hold for $p = n+1$. If $Y \in \mathscr{M}_{n+1}$ satisfies both $\|\operatorname{grad} g(Y)\| \leq \varepsilon_g$ and $\operatorname{Hess} g(Y) \succeq -\varepsilon_H \operatorname{Id}$, then Y is approximately optimal in the sense that (with $R = \max_{X\in\mathscr{C}} \operatorname{Tr}(X)$):*

$$0 \leq 2(g(Y) - f^\star) \leq \varepsilon_g \sqrt{R} + \varepsilon_H R.$$

Under the same condition as in Theorem 4.4, the bound holds with right-hand side $\varepsilon_H R$ instead.

Theorem 1.5 is an informal statement of this corollary.

5 Applications

In all applications below, Assumption 1.1a holds for all p such that the search space is non-empty. For each one, we deduce the consequences of Theorems 1.4 and 1.6. For the latter, the key part is to investigate the facial structure of the SDP. As everywhere else in the paper, $\|x\|$ denotes the 2-norm of vector x and $\|X\|$ denotes the Frobenius norm of matrix X.

5.1 Generalized eigenvalue SDP

The generalized symmetric eigenvalue problem admits a well-known extremal formulation:

$$\text{(EIG)} \qquad \min_{x\in\mathbb{R}^n} x^\top C x \quad \text{subject to} \quad x^\top B x = 1,$$

where C, B are symmetric of size $n \geq 2$. The usual relaxation by lifting introduces $X = xx^\top$ and discards the constraint $\operatorname{rank}(X) = 1$ to obtain this SDP (which is also the Lagrangian dual of the dual of (EIG)):

$$\text{(EIG-SDP)} \qquad \min_{X\in\mathbb{S}^{n\times n}} \langle C, X\rangle \quad \text{subject to} \quad \langle B, X\rangle = 1,\ X \succeq 0.$$

Let $\mathscr{C}$ denote the search space of (EIG-SDP). It is non-empty and compact if and only if $B \succ 0$, which we now assume. A direct application of (2.13) guarantees all extreme points of $\mathscr{C}$ have rank 1, so that it always admits a solution of rank 1: the SDP relaxation is always tight, which is well known. Under our assumption, B admits a Cholesky factorization as $B = R^\top R$ with $R \in \mathbb{R}^{n\times n}$ invertible. The corresponding Burer–Monteiro formulation at rank p reads:

$$\text{(EIG-BM)} \qquad \min_{Y\in\mathbb{R}^{n\times p}} \langle CY, Y\rangle \quad \text{subject to} \quad \|RY\|^2 = 1.$$

Let $\mathscr{M}$ denote its search space. Assumption 1.1a holds for any $p \geq 1$ with $m' = 1$. Indeed, for all $Y \in \mathscr{M}$, $\{BY\}$ spans a subspace of dimension 1, since $BY = R^\top RY$, $RY \neq 0$ and $R^\top$ is invertible. Thus, Theorem 1.4 readily states that for $p \geq 2$, for almost all C, all second-order critical points of (EIG-BM) are optimal.

We can do better. The facial structure of $\mathscr{C}$ is easily described. Recalling (2.12), for all $X = YY^\top \in \mathscr{C}$ we have $\dim \mathscr{F}_X = \frac{p(p+1)}{2} - 1$, since $Y^\top BY \neq 0$. Hence, by Theorem 1.6, for any value of $p \geq 1$, all second-order critical points of (EIG-BM)

are optimal (for any C). In particular, for $p = 1$ (EIG) and (EIG-BM) coincide and we get:

Corollary 5.1. *All second-order critical points of* (EIG) *are optimal.*

This is a well-known fact, though usually proven by direct inspection of necessary optimality conditions.

5.2 Trust-region subproblem SDP

The trust-region subproblem consists in minimizing a quadratic on a sphere, with $n \geq 2$:

$$\text{(TRS)} \qquad \min_{x \in \mathbb{R}^n} x^\top A x + 2b^\top x + c \quad \text{subject to} \quad \|x\|^2 = 1.$$

It is not difficult to produce (A, b, c) such that (TRS) admits suboptimal second-order critical points. The usual lifting here introduces

$$X = \begin{pmatrix} x \\ 1 \end{pmatrix} \begin{pmatrix} x^\top & 1 \end{pmatrix} = \begin{pmatrix} xx^\top & x \\ x^\top & 1 \end{pmatrix}, \qquad \text{and} \qquad C = \begin{pmatrix} A & b \\ b^\top & c \end{pmatrix}.$$

The quadratic cost and constraint are linear in X, yielding this SDP relaxation:

$$\text{(TRS-SDP)} \quad \min_{X \in \mathbb{S}^{n \times n}} \langle C, X \rangle \quad \text{subject to} \quad \mathrm{Tr}(X_{1:n,1:n}) = 1,\ X_{n+1,n+1} = 1,\ X \succeq 0.$$

Let $\mathscr{C}$ denote the search space of (TRS-SDP). It is non-empty and compact. Here too, a direct application of (2.13) guarantees the SDP relaxation is always tight (it always admits a solution of rank 1), which is a well-known fact related to the S-lemma [25]. The Burer–Monteiro relaxation at rank p reads:

$$\text{(TRS-BM)} \quad \min_{Y_1 \in \mathbb{R}^{n \times p}, y_2 \in \mathbb{R}^p} \langle CY, Y \rangle \quad \text{subject to} \quad \|Y_1\|^2 = 1,\ \|y_2\|^2 = 1, \quad \text{with } Y = \begin{pmatrix} Y_1 \\ y_2^\top \end{pmatrix}.$$

Let $\mathscr{M}$ denote its search space. After verifying Assumption 1.1 holds (see below), application of Theorem 1.4 guarantees that for $p \geq 2$ and for *almost* all (A, b, c), second-order critical points of (TRS-BM) are optimal. We can further strengthen this result by looking at the faces of $\mathscr{C}$, as we do now.

Lemma 5.2. *Assumption 1.1a holds for any $p \geq 1$ with $m' = 2$. Furthermore, for $X \in \mathscr{C}$ of rank p,*

$$\dim \mathscr{F}_X = \begin{cases} 0 & \text{if } p = 1, \\ \frac{p(p+1)}{2} - 2 & \text{if } p \geq 2. \end{cases}$$

Proof. The constraints of (SDP) are defined by

$$A_1 = \begin{pmatrix} I_n & 0_{n \times 1} \\ 0_{1 \times n} & 0 \end{pmatrix}, \qquad b_1 = 1, \qquad A_2 = \begin{pmatrix} 0_{n \times n} & 0_{n \times 1} \\ 0_{1 \times n} & 1 \end{pmatrix}, \qquad b_2 = 1.$$

For $Y \in \mathscr{M}$, we have

$$A_1 Y = \begin{pmatrix} Y_1 \\ 0_{1\times p} \end{pmatrix}, \qquad A_2 Y = \begin{pmatrix} 0_{n\times p} \\ y_2^\top \end{pmatrix}.$$

These are nonzero and always linearly independent, so that $\dim \operatorname{span}\{A_1 Y, A_2 Y\} = 2$ for all $Y \in \mathscr{M}$, which confirms Assumption 1.1a holds with $m' = 2$.

The facial structure of $\mathscr{C}$ is simple as well. Let $X \in \mathscr{C}$ have rank p and consider $Y \in \mathscr{M}$ such that $X = YY^\top$. To use (2.12), note that:

$$Y^\top A_1 Y = Y_1^\top Y_1, \qquad Y^\top A_2 Y = y_2 y_2^\top.$$

These are nonzero. For $p = 1$, they are scalars: they span a subspace of dimension 1. Then, $\dim \mathscr{F}_X = 1 - 1 = 0$. For $p > 1$, we argue they are linearly independent. Indeed, if they are not, there exists $\alpha \neq 0$ such that $Y_1^\top Y_1 = \alpha \cdot y_2 y_2^\top$. If so, Y_1 must have rank 1 with row space spanned by y_2, so that $Y_1 = z y_2^\top$ for some $z \in \mathbb{R}^n$, and $\|z\| = 1$. As a result, Y itself has rank 1, which is a contradiction. Thus, $\dim \mathscr{F}_X = \frac{p(p+1)}{2} - 2$, as announced. □

Combining the latter with Theorem 1.6 yields the following new result, which holds for *all* (A, b, c). Notice that for $p = 1$, the theorem correctly allows second-order critical points to be suboptimal in general.

Corollary 5.3. *For $p \geq 2$, all second-order critical points of* (TRS-BM) *are globally optimal.*

A second-order critical point Y of (TRS-BM) with $p = 2$ is thus always optimal. If Y has rank 1, it is straightforward to extract a solution of (TRS) from it. If Y has rank 2,[5] it maps to a face of dimension 1. The endpoints of that face have rank 1 and are also optimal. The following lemma shows these can be computed easily from Y by solving two scalar equations.

Lemma 5.4. *Let $Y \in \mathscr{M}$ be a second-order critical point of* (TRS-BM) *with $p = 2$, and let $z \in \mathbb{R}^2$ satisfy $\|Y_1 z\|^2 = 1$ and $y_2^\top z = 1$. Then, $Y_1 z$ is a global optimum of* (TRS).

Proof. If $\operatorname{rank}(Y) = 1$, then $Y_1 = x y_2^T$ for some $x \in \mathbb{R}^n$, and $\|Y_1\| = 1, \|y_2\| = 1$ ensure $\|x\| = 1$. Solutions to $y_2^T z = 0$ are of the form $z = y_2 + u$, where $y_2^T u = 0$. For any such z, $Y_1 z = x$, which is indeed optimal for (TRS) since Y is globally optimal for (TRS-BM) and x attains the same cost for the restricted problem (TRS).

Now assume $\operatorname{rank}(Y) = 2$. By (2.10), the one-dimensional face $\mathscr{F}_{YY^\top}$ contains all matrices of the form $Y(I_2 - M)Y^\top$ such that $I_2 - M \succeq 0$ and $\langle I_2 - M, Y_1^\top Y_1 \rangle = 0$, $\langle I_2 - M, y_2 y_2^\top \rangle = 0$. This face has two extreme points of rank 1, for which $I_2 - M$ is a positive semidefinite matrix of rank 1, so that $I_2 - M = zz^\top$ for some $z \in \mathbb{R}^2$. Given that Y is feasible, the conditions on z are $\|Y_1 z\|^2 = 1$ and $y_2^\top z = \pm 1$. These

[5] This can happen, notably if (A, b, c) forms a so-called *hard case* TRS (details omitted.) This observation shows that it is indeed necessary to exclude some non-trivial matrices C in Lemma 3.3.

equations define an ellipse in $\mathbb{R}^2$ and two parallel lines, totaling four intersections $\pm z, \pm z'$ which can be computed explicitly. Fixing $y_2^\top z = +1$ allows to identify the two extreme points of the face. Since the cost function is constant along that face, either extreme point yields a global optimum in the same way as above. □

5.3 Optimization over several spheres

The trust-region subproblem generalizes to optimization of a quadratic function over k spheres, possibly in different dimensions $n_1, \ldots, n_k \geq 2$:

$$\text{(Spheres)} \qquad \min_{x_i \in \mathbb{R}^{n_i}, i=1\ldots k} x^\top C x \quad \text{subject to} \quad \|x_1\| = \cdots = \|x_k\| = 1,$$
$$\text{with } x^\top = \begin{pmatrix} x_1^\top & \cdots & x_k^\top & 1 \end{pmatrix}.$$

The variable x is in $\mathbb{R}^{n+1}$, with $n = n_1 + \cdots + n_k$. Since the last entry of x is 1, this indeed covers all possible quadratic functions of $x_1, \ldots, x_k$. The SDP relaxation by lifting reads:

$$\min_{X \in \mathbb{R}^{(n+1)\times(n+1)}} \langle C, X \rangle \quad \text{subject to} \quad \operatorname{Tr}(X_{11}) = \cdots = \operatorname{Tr}(X_{kk}) = 1,$$
$$\text{(Spheres-SDP)} \qquad X_{n+1,n+1} = 1, X \succeq 0,$$

where X_{ij} denotes the block of size $n_i \times n_j$ of matrix X, in the obvious way. This SDP has a non-empty compact search space and $k+1$ independent constraints, so that by (2.13) it always admits a solution of rank at most $p^* = \frac{\sqrt{8k+9}-1}{2}$. The Burer–Monteiro relaxation at rank p reads:

(Spheres-BM)

$$\min_{Y \in \mathbb{R}^{(n+1)\times p}} \langle CY, Y \rangle \quad \text{subject to} \quad \|Y_1\| = \cdots = \|Y_k\| = 1, \|y\| = 1,$$
$$\text{with } Y^\top = \begin{pmatrix} Y_1^\top & \cdots & Y_k^\top & y \end{pmatrix},$$

where $Y_i \in \mathbb{R}^{n_i \times p}$ and $y \in \mathbb{R}^p$. It is easily checked that Assumption 1.1a holds for all $p \geq 1$. Thus, Theorem 1.4 gives this result:

Corollary 5.5. *For $p > \frac{\sqrt{8k+9}-1}{2}$ and for almost all C, all second-order critical points of* (Spheres-BM) *are optimal and map to optima of* (Spheres-SDP).

To apply Theorem 1.6, we first investigate the facial structure of the SDP.

Lemma 5.6. *Let Y be feasible for* (Spheres-BM) *and have full rank p. The dimension of the face of the search space of* (Spheres-SDP) *at $YY^\top$ obeys:*

$$\dim \mathscr{F}_{YY^\top} \leq \frac{p(p+1)}{2} - 2$$

if $p \geq 2$, and $\dim \mathscr{F}_{YY^\top} = 0$ if $p = 1$.

Proof. Following (2.12),

$$\dim \mathscr{F}_{YY^\top} = \frac{p(p+1)}{2} - \dim \operatorname{span}\left(Y_1^\top Y_1, \ldots, Y_k^\top Y_k, yy^\top\right).$$

Since Y is feasible, each defining element of the span is nonzero, so that the dimension is at least 1. If $p = 1$, these elements are scalars: they span $\mathbb{R}$. Now consider $p \geq 2$ and assume for contradiction that the span has dimension one. Then, all defining elements are equal up to scaling. In other words: $Y_i^\top Y_i = \alpha_i \cdot yy^\top$ for some nonzero α_i. If so, Y_i has rank 1 and there exists $z_i \in \mathbb{R}^{n_i}$ such that $Y_i = z_i y^\top$. In turn, this implies Y has rank 1, which is a contradiction. Thus, the span has dimension at least two. □

Corollary 5.7. *For $p \geq \max(2,k)$, all second-order critical points of* (Spheres-BM) *are optimal and map to optima of* (Spheres-SDP) *(for any C).*

For $k = 1$, this recovers the main result about the trust-region subproblem. If the cost function in (Spheres) is a homogeneous quadratic, then it can be written as

$$\text{(SpheresH)} \qquad \min_{x_i \in \mathbb{R}^{n_i}, i=1\ldots k} x^\top C x \quad \text{subject to} \quad \|x_1\| = \cdots = \|x_k\| = 1,$$
$$\text{with } x^\top = \begin{pmatrix} x_1^\top & \cdots & x_k^\top \end{pmatrix}.$$

The corresponding relaxation and Burer–Monteiro formulations read:

$$\text{(SpheresH-SDP)} \qquad \min_{X \in \mathbb{R}^{n\times n}} \langle C, X\rangle \quad \text{subject to} \quad \mathrm{Tr}(X_{11}) = \cdots = \mathrm{Tr}(X_{kk}) = 1, X \succeq 0,$$

and:

$$\text{(SpheresH-BM)} \qquad \min_{Y \in \mathbb{R}^{n\times p}} \langle CY, Y\rangle \quad \text{subject to} \quad \|Y_1\| = \cdots = \|Y_k\| = 1,$$
$$\text{with } Y^\top = \begin{pmatrix} Y_1^\top & \cdots & Y_k^\top \end{pmatrix}.$$

Assumption 1.1a holds for all $p \geq 1$ with $m' = k$. A similar analysis of the facial structure yields the following corollary of Theorem 1.6.

Corollary 5.8. *For almost all C, provided $p > \frac{\sqrt{8k+1}-1}{2}$, all second-order critical points of* (SpheresH-BM) *are optimal and map to optima of* (SpheresH-SDP)*. If $p \geq k$, the result holds for all C.*

For $k = 1$, this recovers the results of (EIG) with $B = I_n$.

5.4 Max-Cut and Orthogonal-Cut SDP

Let $n = qd$ for some integers q, d. Consider the semidefinite program

$$\text{(OrthoCut)} \qquad \min_{X \in \mathbb{S}^{n\times n}} \langle C, X\rangle \quad \text{subject to} \quad \mathrm{sbd}(X) = I_n,\ X \succeq 0,$$

where $\mathrm{sbd}\colon \mathbb{S}^{n\times n} \to \mathbb{S}^{n\times n}$ preserves the diagonal blocks of size $d \times d$ and zeros out all other blocks. Specifically, with X_{ij} denoting the (i,j)th block of size $d \times d$ in matrix X,

$$\mathrm{sbd}(X)_{ij} = \begin{cases} X_{ii} & \text{if } i = j, \\ 0_{d\times d} & \text{otherwise.} \end{cases}$$

For example, with $d = 1$, the constraint $\mathrm{sbd}(X) = I_n$ is equivalent to $\mathrm{diag}(X) = \mathbf{1}$ and this SDP is the Max-Cut SDP [15]. For general d, diagonal blocks of X of size $d \times d$ are constrained to be identity matrices: this SDP is known as Orthogonal-Cut [6, 9]. Among other uses, it appears as a relaxation of synchronization on $\mathbb{Z}_2 = \{\pm 1\}$ [5, 21, 1] and synchronization of rotations [28, 14], with applications in stochastic block modeling (community detection) and SLAM (simultaneous localization and mapping for robotics).

The Stiefel manifold $\mathrm{St}(p,d)$ is the set of matrices of size $p \times d$ with orthonormal columns. The Burer–Monteiro formulation of (OrthoCut) is an optimization problem over q copies of $\mathrm{St}(p,d)$:

(OrthoCut-BM)

$$\min_{Y_1,\ldots,Y_q \in \mathbb{R}^{p\times d}} \langle CY, Y\rangle \quad \text{subject to} \quad Y_k^\top Y_k = I_d \ \forall k,\ Y^\top = \begin{bmatrix} Y_1 & \cdots & Y_q \end{bmatrix}.$$

For $d = 1$, this problem captures one side of the Grothendieck inequality [18, eq. (1.1)]. Assumption 1.1a holds for all $p \geq d$ with $m' = q\frac{d(d+1)}{2}$ (which is the number of constraints). Theorem 1.4 applies as follows.

Corollary 5.9. *If $p > \frac{\sqrt{1+4n(d+1)}-1}{2}$, for almost all C, any second-order critical point Y of* (OrthoCut-BM) *is a global optimum, and $X = YY^\top$ is globally optimal for* (OrthoCut).

In order to apply Theorem 1.6, we must investigate the facial structure of

$$\mathscr{C} = \{X \in \mathbb{S}^{n\times n} : \mathrm{sbd}(X) = I_n, X \succeq 0\}.$$

The following result generalizes a result in [19, Thm. 3.1(i)] to $d \geq 1$.

Theorem 5.10. *If $X \in \mathscr{C}$ has rank p, then the face $\mathscr{F}_X$ (2.10) has dimension bounded as:*

$$\frac{p(p+1)}{2} - n\frac{d+1}{2} \ \leq\ \dim \mathscr{F}_X \ \leq\ \frac{p(p+1)}{2} - p\frac{d+1}{2}. \tag{5.1}$$

If p is an integer multiple of d, the upper bound is attained for some X.

The proof is in Appendix C. Combining this with Theorem 1.6 yields the following result.

Corollary 5.11. *If $p > \frac{d+1}{d+3}n$, any second-order critical point Y for* (OrthoCut-BM) *is globally optimal, and $X = YY^\top$ is globally optimal for* (OrthoCut). *In particular, for Max-Cut SDP ($d = 1$), the requirement is $p > \frac{n}{2}$.*

Proof. If Y is rank deficient, use Proposition 3.1. Otherwise, since $\mathrm{rank}(X) = p$, Theorem 5.10 gives $\dim \mathscr{F}_X \leq \frac{p(p+1)}{2} - p\frac{d+1}{2}$ and Theorem 1.6 gives optimality if

$$\dim \mathscr{F}_X < \frac{p(p+1)}{2} - n\frac{d+1}{2} + p.$$

This is the case provided $(n-p)(d+1) < 2p$, that is, if $p > \frac{d+1}{d+3}n$. □

6 Discussion of the assumptions

We now discuss the assumptions that appear in the main theorems.

The starting point of this investigation is the hope to solve (SDP) by solving (P) instead. For smooth, non-convex optimization problems, even verifying local optimality is usually hard [22]. Thus, we wish to restrict our attention to efficiently computable points, such as points which satisfy first- and second-order Karush–Kuhn–Tucker (KKT) conditions for (P)—see [12, §2.2] and [29, §3]. This only helps if global optima satisfy the latter, that is, if KKT conditions are necessary for optimality.

A global optimum Y necessarily satisfies KKT conditions if *constraint qualifications* (CQs) hold at Y [29]. The standard CQs for equality constrained programs are Robinson's conditions or metric regularity (they are here equivalent). They read as follows:

(CQ) CQs hold at $Y \in \mathscr{M}$ if $A_1Y, \ldots, A_mY$ are linearly independent in $\mathbb{R}^{n\times p}$.

Considering all cost matrices C, global optima could, a priori, be anywhere in $\mathscr{M}$. Thus, we require CQs to hold at all Y in $\mathscr{M}$ rather than only at the (unknown) global optima. This leads to Assumption 1.1a. Adding redundant constraints (for example, duplicating $\langle A_1, X\rangle = b_1$) would break the CQs, but does not change the optimization problem. This is allowed by Assumption 1.1b.

In general, (SDP) may not have an optimal solution. One convenient way to guarantee that it does is to require $\mathscr{C}$ to be compact, which is why this assumption appears in Theorem 1.5 to bound optimality gaps for approximate second-order critical points. When $\mathscr{C}$ is compact, one furthermore gets the guarantee that at least one of the global optima is an extreme point of $\mathscr{C}$, which leads to the guarantee that at least one of the global optima has rank p bounded as $\frac{p(p+1)}{2} \leq m'$ (2.13). The other way around, it is possible to pick the cost matrix C such that the unique solution to (SDP) is an extreme point of maximal rank, which can be as large as allowed by (2.13). This justifies why, in Theorem 1.4, the bound on p is essentially optimal. The compactness assumption could conceivably be relaxed, provided candidate global optima remain bounded. This could plausibly come about by restricting attention to positive definite cost matrices C.

One restriction in particular in Theorem 1.4 merits further investigation: the exclusion of a zero-measure set of cost matrices ("bad C"). From the trust-region subproblem example in Section 5.2, we know that it is necessary (in general) to allow the exclusion of a zero-measure set of cost matrices in Lemma 3.3. Yet, in that same example, the excluded cost matrices do not give rise to suboptimal second-order critical points (as we proved through a different argument involving Theorem 1.6.) Thus, it remains unclear whether or not a zero-measure set of cost matrices must be excluded in Theorem 1.4. Resolving this question is key to gain deeper understanding of the relationship between (SDP) and (P).

Finally, we connect the notion of smooth SDP used in this paper to the more standard notion of non-degeneracy in SDPs as defined in [4, Def. 5]. Informally: for linearly independent A_i, non-degeneracy at all points is equivalent to smoothness. The proof is in Appendix D.

Definition 6.1. X is *primal non-degenerate* for (SDP) if it is feasible and $\mathrm{T}_X + \ker \mathscr{A} = \mathbb{S}^{n\times n}$, where T_X is the tangent space at X to the manifold of symmetric matrices of rank r embedded in $\mathbb{S}^{n\times n}$, where $r = \mathrm{rank}(X)$.

Proposition 6.2. *Let $A_1,\ldots,A_m$ defining $\mathscr{A}$ be linearly independent. Then, Assumption 1.1a holds for all p such that $\mathscr{M}_p$ is non-empty if and only if all $X \in \mathscr{C}$ are primal non-degenerate.*

7 Conclusions and perspectives

We have shown how, under Assumption 1.1 and extra conditions (on p, compactness, and the cost matrix), the Burer–Monteiro factorization approach to solving (SDP) is "safe", despite non-convexity. For future research, it is of interest to determine if the proposed assumptions can be relaxed. Furthermore, it is important for practical purposes to determine whether approximate second-order critical points are approximately optimal for values of p well below n (an example of this for a specific context is given in [5]). One possible way forward is a smoothed analysis of the type developed recently in [8, 26], though these early works leave plenty of room for improvement.

Appendix A: Consequences and properties of Assumption 1.1

Proof of Proposition 1.2. The set $\mathscr{M}$ is defined as the zero level set of $\Phi\colon \mathbb{R}^{n\times p} \to \mathbb{R}^m$ where $\Phi(Y) = \mathscr{A}(YY^\top) - b$. The differential of Φ at Y, $\mathrm{D}\Phi(Y)$, has rank equal to the dimension of the space spanned by $\{A_1Y,\ldots,A_mY\}$. Under Assumption 1.1a, $\mathrm{D}\Phi(Y)$ has full rank m on $\mathscr{M}$ and the result follows from [20, Corollary 5.14]. Under Assumption 1.1b, $\mathrm{D}\Phi(Y)$ has constant rank m' in a neighborhood of $\mathscr{M}$ and the result follows from [20, Theorem 5.12]. □

Proof of Proposition 1.3. First, let Assumption 1.1a hold for some p, and consider $p' < p$ such that $\mathscr{M}_{p'}$ is non-empty. For any $Y' \in \mathscr{M}_{p'}$, form $Y = \left[Y' | 0_{n\times(p-p')}\right] \in \mathbb{R}^{n\times p}$. Clearly, Y is in $\mathscr{M}_p$, so that

$$m = \dim \mathrm{span}\{A_1Y,\ldots,A_mY\} = \dim \mathrm{span}\{A_1Y',\ldots,A_mY'\},$$

as desired. For $p = n$, we now consider the case $p' > n$. Let $Y' \in \mathscr{M}_{p'}$ and consider its full SVD, $Y' = U\Sigma V^\top$, with $\Sigma \in \mathbb{R}^{n\times p'}$. Then, $Y'V$ is in $\mathscr{M}_{p'}$ as well. Since the last $p'-n$ columns of Σ are zero, we have $Y'V = U\Sigma = \left[Y | 0_{n\times(p'-n)}\right]$ with $Y \in \mathscr{M}_n$.

Thus, as desired,

$$\begin{aligned}\dim \operatorname{span}\{A_1Y', \ldots, A_mY'\} &= \dim \operatorname{span}\{A_1Y'V, \ldots, A_mY'V\} \\ &= \dim \operatorname{span}\{A_1Y, \ldots, A_mY\} \\ &= m.\end{aligned}$$

Second, let Assumption 1.1b hold for some p, and consider $p' < p$ such that $\mathscr{M}_{p'}$ is non-empty. For any $Y' \in \mathscr{M}_{p'}$, form $Y = \left[Y' | 0_{n\times(p-p')}\right] \in \mathscr{M}_p$. By assumption, there exists an open ball B_Y in $\mathbb{R}^{n\times p}$ of radius $\varepsilon = \varepsilon(Y) > 0$ centered at Y such that

$$\dim \operatorname{span}\{A_1\tilde{Y}, \ldots, A_m\tilde{Y}\} = m'$$

for all $\tilde{Y} \in B_Y$. Let $B_{Y'}$ be the open ball in $\mathbb{R}^{n\times p'}$ of radius $\varepsilon(Y)$ and center Y'. For any $\tilde{Y}' \in B_{Y'}$, form $\tilde{Y} = \left[\tilde{Y}' | 0_{n\times(p-p')}\right]$. Since $\|\tilde{Y} - Y\| = \|\tilde{Y}' - Y'\| \leq \varepsilon$, we have $\tilde{Y} \in B_Y$, so that

$$m' = \dim \operatorname{span}\{A_1\tilde{Y}, \ldots, A_m\tilde{Y}\} = \dim \operatorname{span}\{A_1\tilde{Y}', \ldots, A_m\tilde{Y}'\}.$$

Thus, Assumption 1.1b holds with the open neighborhood of $\mathscr{M}_{p'}$ consisting of the union of all balls $B_{Y'}$ for $Y' \in \mathscr{M}_{p'}$ as described above. □

Appendix B: The facial structure of $\mathscr{C}$

Proof of Proposition 2.7. The construction follows [24] and applies for any linear equality constraints. We first show that if X' is of the form in (2.10), then it must be in $\mathscr{F}_X$. This is clear if $X' = X$. Otherwise, pick $t > 0$ such that $I_p - tA \succeq 0$. Then, X' and $X'' = Y(I_p - tA)Y^\top$ define a closed line segment in $\mathscr{C}$ whose relative interior contains X. By Definition 2.5, this implies X' (and X'') are in $\mathscr{F}_X$.

The other way around, we now show that any point in $\mathscr{F}_X$ must be of the form of X' in (2.10). Let $W \in \mathbb{S}^{n\times n}$ be such that $X' = X + W$. Since X is in the relative interior of $\mathscr{F}_X$ which is convex, there exists $t > 0$ such that $X - tW \in \mathscr{F}_X$. Let $Y_\perp \in \mathbb{R}^{n\times(n-p)}$ be such that $M = \begin{bmatrix} Y & Y_\perp \end{bmatrix}$ is invertible. We can express $X = YY^\top$ and W as

$$X = M\begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix}M^\top \qquad \text{and} \qquad W = M\begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}M^\top.$$

Then, explicitly, these two matrices must belong to $\mathscr{C}$:

$$X + W = M\begin{bmatrix} I_p + A & B \\ B^\top & C \end{bmatrix}M^\top, \quad \text{and} \quad X - tW = M\begin{bmatrix} I_p - tA & -tB \\ -tB^\top & -tC \end{bmatrix}M^\top.$$

In particular, they must both be positive semidefinite, which implies $C \succeq 0$ and $-tC \succeq 0$, so that $C = 0$. By Schur's complement, it follows that $B = 0$. Thus, $W = YAY^\top$ for some $A \in \mathbb{S}^{p\times p}$ such that $I_p + A \succeq 0$. Furthermore, $\mathscr{A}(X') = \mathscr{A}(X + W) = b$, so that $\mathscr{A}(W) = 0$. The latter is equivalent to $\mathscr{L}_X(A) = 0$ using (2.11). □

Appendix C: Faces of the Ortho-Cut SDP

Proof of Theorem 5.10. Consider the definition of $\mathscr{L}_X$ (2.11) and inequality (2.12): the latter covers the lower bound and shows we need rank $\mathscr{L}_X \geq p(d+1)/2$ for the upper bound, that is, we need to show the condition $\mathscr{L}_X(A) = 0$ imposes at least $p(d+1)/2$ linearly independent constraints on $A \in \mathbb{S}^{p\times p}$.

Let $Y \in \mathscr{M}_p$ be such that $X = YY^\top$, and let $y_1, \ldots, y_n \in \mathbb{R}^p$ denote the rows of Y, transposed. Greedily select p linearly independent rows of Y, in order, such that row i is picked iff it is linearly independent from rows y_1 to y_{i-1}. This is always possible since Y has rank p. Write $t = \{t_1 < \cdots < t_p\}$ to denote the indices of selected rows. Write $s_k = \{((k-1)d+1), \ldots, kd\}$ to denote the indices of rows in slice $Y_k^\top$, and let $c_k = s_k \cap t$ be the indices of selected rows in that slice.

We make use of the following fact [19, Lem. 2.1]: for $x_1, \ldots, x_p \in \mathbb{R}^p$ linearly independent, the $p(p+1)/2$ symmetric matrices $x_i x_j^\top + x_j x_i^\top$ form a basis of $\mathbb{S}^{p\times p}$. Defining $E_{ij} = y_i y_j^\top + y_j y_i^\top = E_{ji}$, this means $\mathscr{E} = \{E_{t_\ell, t_{\ell'}} : \ell, \ell' = 1 \ldots p\}$ forms a basis of $\mathbb{S}^{p\times p}$ ($\mathscr{E}$ is a set, so that E_{ij} and E_{ji} contribute only one element). Similarly, since each slice $Y_k^\top$ has orthonormal rows, matrices in $\{E_{ij} : i, j \in s_k\}$ are linearly independent.

The constraint $\mathscr{L}_X(A) = 0$ means $\langle A, E_{ij}\rangle = 0$ for each k and for each $i, j \in s_k$. To establish the theorem, we need to extract a subset $\mathscr{T}$ of at least $p(d+1)/2$ of these $qd(d+1)/2$ constraint matrices, and guarantee their linear independence. To this end, let

$$\mathscr{T} = \{E_{ij} : k \in \{1, \ldots, q\} \text{ and } i \in c_k \subseteq s_k, j \in s_k\}. \tag{C.1}$$

That is, for each slice k, $\mathscr{T}$ includes all constraints of that slice which involve at least one of the selected rows. For each slice k, there are $|c_k|d - \frac{|c_k|(|c_k|-1)}{2}$ such constraints—note the correction for double-counting the E_{ij}'s where both i and j are in c_k. Thus, using $|c_1| + \cdots + |c_q| = p$, the cardinality of $\mathscr{T}$ is:

$$|\mathscr{T}| = \sum_{k=1}^{q} \left[|c_k|d - \frac{|c_k|(|c_k|-1)}{2} \right] = p(d+1/2) - \frac{1}{2}\sum_{k=1}^{q} |c_k|^2. \tag{C.2}$$

We first show matrices in $\mathscr{T}$ are linearly independent. Then, we show $|\mathscr{T}|$ is large enough.

Consider one $E_{ij} \in \mathscr{T}$: $i, j \in s_k$ for some k and $i = t_\ell$ for some ℓ (otherwise, permute i and j). By construction of t, we can expand y_j in terms of the rows selected in slices 1 to k, i.e., $y_j = \sum_{\ell'=1}^{\ell_k} \alpha_{j,\ell'} y_{t_{\ell'}}$, where $\ell_k = |c_1| + \cdots + |c_k|$. As a result, E_{ij} expands in the basis $\mathscr{E}$ as follows: $E_{ij} = \sum_{\ell'=1}^{\ell_k} \alpha_{j,\ell'} E_{t_\ell, t_{\ell'}}$. As noted before, E_{ij}'s in $\mathscr{T}$ contributed by a same slice k are linearly independent. Furthermore, they expand in only a subset of the basis $\mathscr{E}$, namely, $\mathscr{E}^{(k)} = \{E_{t_\ell, t_{\ell'}} : \ell_{k-1} < \ell \leq \ell_k, \ell' \leq \ell_k\}$: t_ℓ is a selected row of slice k and $t_{\ell'}$ is a selected row of some slice between 1 and k. For $k \neq k'$, $\mathscr{E}^{(k)}$ and $\mathscr{E}^{(k')}$ are disjoint; in fact, they form a partition of $\mathscr{E}$. Hence, elements of $\mathscr{T}$ are linearly independent.

It remains to lower bound (C.2). To this end, use $|c_k| \leq d$ and $|c_1| + \cdots + |c_q| = p$ to get:

$$\sum_{k=1}^{q} |c_k|^2 \leq \max_{x \in \mathbb{R}^q : \|x\|_\infty \leq d, \|x\|_1 = p} \|x\|^2 = \left\lfloor \frac{p}{d} \right\rfloor d^2 + \left(p - \left\lfloor \frac{p}{d} \right\rfloor d\right)^2 \leq pd.$$

Indeed, the maximum in x is attained by making as many of the entries of x as large as possible, that is, by setting $\lfloor p/d \rfloor$ entries to d and setting one other entry to $p - \lfloor p/d \rfloor d$ if the latter is nonzero. This many entries are available since $p \leq qd = n$. That this is optimal can be verified using KKT conditions. In combination with (C.2), this confirms at least $p(d+1/2) - pd/2 = p(d+1)/2$ linearly independent constraints act on A, thus upper bounding $\dim \mathscr{F}_X$.

To conclude, we argue that the proposed upper bound is essentially tight. Indeed, build $Y \in \mathscr{M}_p$ by repeating q times the d first rows of I_p, then by replacing its p first rows with I_p (to ensure Y has full rank). If p/d is an integer, then exactly the p/d first slices each contribute $d(d+1)/2$ independent constraints, i.e., $\dim \mathscr{F}_{YY^\top} = p(p+1)/2 - p(d+1)/2$. □

Appendix D: Equivalence of global non-degeneracy and smoothness

Proof of Proposition 6.2. By Proposition 1.3, it is sufficient to consider the case $p = n$. Consider $X \in \mathscr{C}$ of rank r and a diagonalization $X = QDQ^\top$, where $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)$ and $Q = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}$ is orthogonal of size n with $Q_1 \in \mathbb{R}^{n \times r}$. By [4, Thm. 6], since $A_1, \ldots, A_m$ are linearly independent, X is primal non-degenerate if and only if the matrices

$$B_k = \begin{bmatrix} Q_1^\top A_k Q_1 & Q_1^\top A_k Q_2 \\ Q_2^\top A_k Q_1 & 0 \end{bmatrix}, \quad k = 1 \ldots, m$$

are linearly independent. The B_k are linearly dependent if and only if there exist $\alpha_1, \ldots, \alpha_m$ not all zero such that $\alpha_1 B_1 + \cdots + \alpha_m B_m = 0$. Considering the first r columns of the B_k, the latter holds if and only if $\sum_k \alpha_k Q^\top A_k Q_1 = 0$, which holds if and only if $\sum_k \alpha_k A_k Q_1 = 0$. For any $Y \in \mathbb{R}^{n \times p}$ such that $X = YY^\top$, since $\operatorname{span}(Y) = \operatorname{span}(Q_1)$, we have $\sum_k \alpha_k A_k Q_1 = 0$ if and only if $\sum_k \alpha_k A_k Y = 0$. This shows the B_k are linearly dependent if and only if the $A_k Y$ are linearly dependent. Thus, X is primal non-degenerate if and only if $\{A_1 Y, \ldots, A_m Y\}$ are linearly independent. Overall, primal non-degeneracy holds at all $X \in \mathscr{C}$ if and only if Assumption 1.1a holds. □

Acknowledgment.

NB was partially supported by NSF grant DMS-1719558. Part of this work was done while NB was with the D.I. at Ecole normale supérieure de Paris and INRIA's SIERRA team. ASB was partially supported by NSF grants DMS-1712730 and DMS-1719545. Part of this work was done while ASB was with the Mathematics Department at MIT and partially supported by NSF grant DMS-1317308

Bibliography

[1] Abbé, E.; Bandeira, A.; Hall, G. Exact recovery in the stochastic block model. *Information Theory, IEEE Transactions on* **62** (2016), no. 1, 471–487.

[2] Absil, P.-A.; Baker, C. G.; Gallivan, K. A. Trust-region methods on Riemannian manifolds. *Foundations of Computational Mathematics* **7** (2007), no. 3, 303–330.

[3] Absil, P.-A.; Mahony, R.; Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2008.

[4] Alizadeh, F.; Haeberly, J.-P.; Overton, M. Complementarity and nondegeneracy in semidefinite programming. *Mathematical Programming* **77** (1997), no. 1, 111–128.

[5] Bandeira, A.; Boumal, N.; Voroninski, V.: On the low-rank approach for semidefinite programs arising in synchronization and community detection, in *Proceedings of The 29th Conference on Learning Theory, COLT 2016, New York, NY, June 23–26*, 2016 .

[6] Bandeira, A.; Kennedy, C.; Singer, A. Approximating the little Grothendieck problem over the orthogonal and unitary groups. *Mathematical Programming* (2016), 1–43.

[7] Barvinok, A. Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry* **13** (1995), no. 1, 189–202.

[8] Bhojanapalli, S.; Boumal, N.; Jain, P.; Netrapalli, P.: Smoothed analysis for low-rank solutions to semidefinite programs in quadratic penalty form, in *Proceedings of the 31st Conference On Learning Theory*, *Proceedings of Machine Learning Research*, vol. 75, edited by S. Bubeck; V. Perchet; P. Rigollet, PMLR, 2018 pp. 3243–3270. Available at: http://proceedings.mlr.press/v75/bhojanapalli18a.html

[9] Boumal, N. A Riemannian low-rank method for optimization over semidefinite matrices with block-diagonal constraints. *arXiv preprint arXiv:1506.00575* (2015).

[10] Boumal, N.; Absil, P.-A.; Cartis, C. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis* (2018).

[11] Boumal, N.; Voroninski, V.; Bandeira, A. The non-convex Burer–Monteiro approach works on smooth semidefinite programs. in *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee; M. Sugiyama; U. V. Luxburg; I. Guyon; R. Garnett, pp. 2757–2765, Curran Associates, Inc., 2016.

[12] Burer, S.; Monteiro, R. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming* **95** (2003), no. 2, 329–357.

[13] Burer, S.; Monteiro, R. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming* **103** (2005), no. 3, 427–444.

[14] Eriksson, A.; Olsson, C.; Kahl, F.; Chin, T.-J.: Rotation averaging and strong duality, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018 pp. 127–135.

[15] Goemans, M.; Williamson, D. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)* **42** (1995), no. 6, 1115–1145.

[16] Golub, G. H.; Pereyra, V. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis* **10** (1973), no. 2, 413–432.

[17] Journée, M.; Bach, F.; Absil, P.-A.; Sepulchre, R. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization* **20** (2010), no. 5, 2327–2351.

[18] Khot, S.; Naor, A. Grothendieck-type inequalities in combinatorial optimization. *Communications on Pure and Applied Mathematics* **65** (2012), no. 7, 992–1035.

[19] Laurent, M.; Poljak, S. On the facial structure of the set of correlation matrices. *SIAM Journal on Matrix Analysis and Applications* **17** (1996), no. 3, 530–547.

[20] Lee, J. *Introduction to Smooth Manifolds*, *Graduate Texts in Mathematics*, vol. 218, Springer-Verlag New York, 2012, 2nd ed.

[21] Mei, S.; Misiakiewicz, T.; Montanari, A.; Oliveira, R.: Solving SDPs for synchronization and MaxCut problems via the Grothendieck inequality, in *Proceedings of the 2017 Conference on Learning Theory*, *Proceedings of Machine Learning Research*, vol. 65, edited by S. Kale; O. Shamir, PMLR, Amsterdam, Netherlands, 2017 pp. 1476–1515. Available at: http://proceedings.mlr.press/v65/mei17a.html

[22] Murty, K.; Kabadi, S. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming* **39** (1987), no. 2, 117–129.

[23] Nesterov, Y.; Nemirovskii, A. *Interior-point polynomial algorithms in convex programming*, SIAM, 1994.

[24] Pataki, G. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of operations research* **23** (1998), no. 2, 339–358.

[25] Polik, I.; Terlaky, T. A survey of the S-lemma. *SIAM Review* **49** (2007), no. 3, 371–418.

[26] Pumir, T.; Jelassi, S.; Boumal, N. Smoothed analysis of the low-rank approach for smooth semidefinite programs. in *Advances in Neural Information Processing Systems 31*, edited by S. Bengio; H. Wallach; H. Larochelle; K. Grauman; N. Cesa-Bianchi; R. Garnett, pp. 2283–2292, Curran Associates, Inc., 2018.

[27] Rockafellar, R. *Convex analysis*, Princeton University Press, Princeton, NJ, 1970.

[28] Rosen, D. M.; Carlone, L.; Bandeira, A. S.; Leonard, J. J. A certifiably correct algorithm for synchronization over the special euclidean group. *arXiv preprint arXiv:1611.00128* (2016).

[29] Ruszczyński, A. *Nonlinear optimization*, Princeton University Press, Princeton, NJ, 2006.

[30] Wen, Z.; Yin, W. A feasible method for optimization with orthogonality constraints. *Mathematical Programming* **142** (2013), no. 1–2, 397–434.

[31] Yang, W.; Zhang, L.-H.; Song, R. Optimality conditions for the nonlinear programming problems on Riemannian manifolds. *Pacific Journal of Optimization* **10** (2014), no. 2, 415–434.

Received April 2018.