Demystifying the characterization of SDP matrices in mathematical programming

Daniel Porumbel

June 18, 2022

Argument

This manuscript was written because I found no other introduction to SDP programming that targets the same audience. This work is intended to be accessible to anybody who does not hate maths, who knows what a derivative is and accepts (or has a proof of) results like $\det(AB) = \det(A)\det(B)$. If you know this, I think you can understand most of this text without buying other books; my goal is not to remind/enumerate a list of results but to (try to) make the reader examine (the proofs of) these results so as to get full insight into them.

A first difference compared to other existing introductions to SDP is that this work comes out of a mind that was itself struggling to understand and not from a long established expert. *This may seem to be only a weakness, but, paradoxically, it is both a weakness and a strength.* I really did not try to overpower the reader or to transform the author–reader relationship into a formal professor–student relationship. My goal was to minimize the distance between the author and the reader as much as possible. Another strength comes from the fact that many long-acknowledged experts tend to forget the difficulties of beginners; I think this did not happened to me. Other experts try to make all proofs as short as possible and to dismiss as unimportant certain key results they have seen thousands of time in their career. I also avoided this, even if I did shorten a few proofs when I revised this manuscript two years after it was first written. However, I also kept certain proofs that seem longer than necessary because I feel they offer more insight; an important goal is to capture the "spirit" of each proven result instead of reducing it to a flow of formulae.

The very first key step towards mastering SDP programming is to get full insight into the eigen-decomposition of real symmetric matrices. Many other published introductions to SDP programming that merit our (research work) consideration address this eigen-decomposition in a way that shows their target audience is different from mine. They usually list the eigen-decomposition without proof, while I give two proofs to really familiarize the reader with this key concept.[1]

If you can say "Ok, I still vaguely remember the eigen-decomposition (and other key SDP properties) from my (under-)graduate studies some $n \geq 5$ years ago; I don't need a proof", then you do not belong to my target audience. I am highly skeptical that such approach can lead to anything but superficial learning. Anyhow, my brain functions in the most opposite manner. *I like to learn by checking all proofs by myself and I can't stand taking things for granted*. The only unproven facts from this manuscript are the fundamental theorem of algebra and two results from Section 5.3.2.3. But I do provide complete proofs, for instance, for *the Cholesky decomposition of SDP matrices, the strong duality theorem for linear conic programming (including SDP programming), six equivalent formulations of the Lovász theta number, the copositive formulation of the maximum stable, a few convexification results for quadratic programs and many others.* I tried to prove everything by myself, so that certain proofs are original although this introduction was not intended to be research work; of course I got help multiple times from the internet, articles and books, the references being indicated as footnote citations.

[1]In "Handbook of Semidefinite Programming Theory, Algorithms, and Applications" by H. Wolkowicz, R. Saigal and L. Vandenberghe, the eigen-decomposition (called spectral theorem) is listed with no proof in Chapter 2 "Convex Analysis on Symmetric Matrices". The introduction of the "Handbook on Semidefinite, Conic and Polynomial Optimization" by M. Anjost and J.B. Lasserre refers the reader to the (700 pages long) book "Matrix analysis" by Horn and Johnson. As a side remark, the introductions of both these handbooks are rather short (14 or resp. 22 pages) and they mainly remind/enumerate different key results pointing to other books for proofs. In "Semidefinite Programming for Combinatorial Optimization", by C. Helmberg, the eigen-decomposition is presented in an appendix and redirects the reader to the same "Matrix analysis" book. The slides of the course "Programmation linaire et optimisation combinatoire" of Frédéric Roupin for the "Master Parisien de Recherche Oprationnelle" (`lipn.univ-paris13.fr/~roupin/docs/MPROSDPRoupin2018-partie1.pdf`) provide many results from my manuscript but no proof is given. The MIT course "Introduction to Semidefinite Programming " by R. Freund does not even provide the SDP definition or the eigen-decomposition. The book "Convex Optimization" by S. Boyd and L. Vandenberghe starts using SDP matrices from the beginning (*e.g.*, to define ellipsoids in Section 2.2.2) without defining the concept of SDP matrix, not even in appendix. The argument could extend to other non-trivial concepts that are taken as pre-requisite in above works. For instance, the above "Convex Optimization" book introduces the square root of an SDP matrix (in five lines in Appendix A.5.2), without showing the uniqueness – the proof takes half a page in Appendix B.4 of this manuscript.

The most essential building blocks are presented in the first part. One should really master this first part before jumping to the second one; the essentials from the first part may even be generally useful for reading other SDP work. In fact, the goal of this manuscript is to give you all the tools needed to move to the next level and carry out research work.

Contents

APPENDIX

PART 1: THE ESSENTIAL BUILDING BLOCKS

1 Characterization of semidefinite positive (SDP) matrices

1.1 Real symmetric matrices, eigenvalues and the eigendecomposition

This light introduction aims at familiarizing the reader with the main concepts of real symmetric matrices, eigenvalues and the eigenvalue decomposition. Experts on this topic can skip to Section 1.2 or even to further sections. Absolute beginners should first consult Appendix A to get the notion of matrix rank, (sub-)space dimension, (principal) minor, or to recall how $\det(A) = 0 \iff \exists \mathbf{x}$ s. t. $A\mathbf{x} = \mathbf{0}$. To familiarize with such introductory concepts it is also useful to solve a few exercises, but the only exercise I propose is to ask the reader to prove by himself all theorems whose proof does not exceed half a page.

Given (real symmetric) matrix A, we say λ is a (real) *eigenvalue* of A if there exists *eigenvector* $\mathbf{v}$ such that $A\mathbf{v} = \lambda\mathbf{v}$. Notice that by multiplying the eigenvector with a constant we obtain another eigenvector. An eigenvector is called *unitary* if it has a norm of 1, *i.e.*, $|\mathbf{v}|^2 = \mathbf{v}^\top\mathbf{v} = \sum_{i=1}^n v_i^2 = 1$. We say $\mathbf{v}$ is an eigenvector if it satisfies $\lambda I_n\mathbf{v} - A\mathbf{v} = \mathbf{0}$, equivalent to $(\lambda I_n - A)\mathbf{v} = \mathbf{0}$, which also means $\det(\lambda I_n - A) = 0$.

We can also define the eigenvalues as the roots of the *characteristic polynomial* $\det(xI_n - A)$. Indeed, if $\det(\lambda_1 I_n - A) = 0$, there exists a (real) eigenvector $\mathbf{v}_1$ such that $(\lambda_1 I_n - A)\mathbf{v}_1 = 0$, equivalent to $A\mathbf{v}_1 = \lambda_1\mathbf{v}_1$. The fact that $\mathbf{v}_1$ needs to exist is formally proven in Prop A.2.2; this Prop A.2.2 proves the statement in the more general context of complex matrices because we should not yet take for granted that root λ_1 is not complex, *i.e.*, $\lambda_1 I_n - A$ could have complex entries in principle. However, it *is* possible to show that a real symmetric matrix has only real eigenvalues and real proper eigenvectors. The proof is not really completely obvious, it takes a third of a page and it is given in appendix (Prop. B.1.2). By developing the characteristic polynomial, one can also prove that the determinant is the product of the eigenvalues (Prop. A.2.4).

The characteristic polynomial has degree n, and so, it has n roots (eigenvalues), but some of them can have multiplicities greater than 1, *i.e.*, some eigenvalues can appear more than once. However, each eigenvalue is associated to at least one eigenvector. An eigenvalue with multiplicity greater than 1 can have more than one eigenvector. The following result is called the *eigendecomposition* of the real symmetric matrix A.

$$
\begin{aligned}
A &= \overset{\substack{\text{unitary eigenvectors of } A \\ \downarrow \quad \downarrow \quad \dots \quad \downarrow}}{\begin{bmatrix} v_{11} & v_{21} & \dots & v_{n,1} \\ v_{21} & v_{22} & \dots & v_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n,1} & v_{n,2} & \dots & v_{n,n} \end{bmatrix}}
\overset{\text{eigenvalues of } A}{\begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}}
\begin{bmatrix} v_{11} & v_{12} & \dots & v_{1,n} \\ v_{12} & v_{22} & \dots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1,n} & v_{2,n} & \dots & v_{n,n} \end{bmatrix}
\begin{array}{l} \leftarrow \\ \leftarrow \\ \vdots \\ \leftarrow \end{array}
\ \text{transposed eigenvectors} \\
&= \begin{bmatrix} \mathbf{v}_1 \ \mathbf{v}_2 \ \dots \mathbf{v}_n \end{bmatrix} \mathtt{diag}(\lambda_1,\ \lambda_2,\ \dots \lambda_n) \begin{bmatrix} \mathbf{v}_1^\top \\ \mathbf{v}_2^\top \\ \vdots \\ \mathbf{v}_n^\top \end{bmatrix} = U\Lambda U^\top \qquad (1.1.1) \\
&= \sum_{i=1}^n \lambda_i \mathbf{v}_i \mathbf{v}_i^\top,
\end{aligned}
$$

where $\mathbf{v}_1, \mathbf{v}_2, \dots \mathbf{v}_n$ are the unitary (column) eigenvectors of A associated to (some repeated) eigenvalues $\lambda_1,\ \lambda_2, \dots \lambda_n$ and $\Lambda = \mathtt{diag}(\lambda_1,\ \lambda_2\ \dots, \lambda_n)$ is the diagonal matrix with $\Lambda_{ii} = \lambda_i\ \forall i \in [1..n]$. The eigenvectors are unitary and orthogonal, meaning that $\mathbf{v}_i^\top\mathbf{v}_j = 0,\ \forall i, j \in [1..n],\ i \neq j$ and $\mathbf{v}_i^\top\mathbf{v}_i = 1\ \forall i \in [1..n]$. This directly leads to

$$
[\mathbf{v}_1\ \mathbf{v}_2\ \dots \mathbf{v}_n]^\top [\mathbf{v}_1\ \mathbf{v}_2\ \dots \mathbf{v}_n] = I_n, \text{ equivalent to } [\mathbf{v}_1\ \mathbf{v}_2\ \dots \mathbf{v}_n][\mathbf{v}_1\ \mathbf{v}_2\ \dots \mathbf{v}_n]^\top = I_n, \qquad (1.1.2)
$$

using the very well known property $XY = I \implies YX = I$ (Prop. A.2.5).

The simple case of distinct eigenvalues

It is important to familiarize with this decomposition. For this, let us first examine a proof that works for symmetric matrices A with distinct eigenvalues (all with multiplicity one). We will first show that the eigenvectors of symmetric A are orthogonal. Let $\mathbf{v}_1,\ \mathbf{v}_2, \ldots \mathbf{v}_n$ be the unitary eigenvectors of resp. $\lambda_1,\ \lambda_2, \ldots \lambda_n$. For any $i, j \in [1..n],\ i \neq j$, we can write $\mathbf{v}_i^\top A \mathbf{v}_j = \mathbf{v}_i^\top \lambda_j \mathbf{v}_j = \lambda_j \mathbf{v}_i^\top \mathbf{v}_j$ and also $\mathbf{v}_i^\top A \mathbf{v}_j = \lambda_i \mathbf{v}_i^\top \mathbf{v}_j$ based on $\mathbf{v}_i^\top A = (\mathbf{v}_i^\top A)^{\top\top} = (A^\top \mathbf{v}_i)^\top = (A\mathbf{v}_i)^\top = \lambda_i \mathbf{v}_i^\top$. This leads to $\lambda_j \mathbf{v}_i^\top \mathbf{v}_j = \lambda_i \mathbf{v}_i^\top \mathbf{v}_j$, and, using $\lambda_i \neq \lambda_j$, we obtain $\mathbf{v}_i^\top \mathbf{v}_j = 0\ \forall i, j \in [1..n],\ i \neq j$.

We now construct the eigen-decomposition. Using $\lambda_i \mathbf{v}_i = A\mathbf{v}_i\ \forall i \in [1..n]$, we obtain $A[\mathbf{v}_1\ \mathbf{v}_2\ \ldots \mathbf{v}_n] = [\lambda_1 \mathbf{v}_1\ \lambda_2 \mathbf{v}_2\ \ldots \lambda_n \mathbf{v}_n]$, equivalent to $A[\mathbf{v}_1\ \mathbf{v}_2\ \ldots \mathbf{v}_n] = [\mathbf{v}_1\ \mathbf{v}_2\ \ldots \mathbf{v}_n]\mathtt{diag}(\lambda_1,\ \lambda_2, \ldots \lambda_n)$. We now multiply both sides at right by $[\mathbf{v}_1\ \mathbf{v}_2\ \ldots \mathbf{v}_n]^\top$ which is equal to $[\mathbf{v}_1\ \mathbf{v}_2\ \ldots \mathbf{v}_n]^{-1}$ by virtue (1.1.2); we thus obtain $A = [\mathbf{v}_1\ \mathbf{v}_2\ \ldots \mathbf{v}_n]\mathtt{diag}(\lambda_1,\ \lambda_2, \ldots \lambda_n)[\mathbf{v}_1\ \mathbf{v}_2\ \ldots \mathbf{v}_n]^\top$, which is exactly (1.1.1).

Generalizing the proof for the case of eigenvalues with different multiplicities

We now prove the eigen-decomposition in the general case by extending the above construction. We need two steps

Step. 1 We first show that the above construction for eigenvalues with unitary multiplicities can be readily generalized to the case where there is a different eigenvector for each of the k_i repetitions of each root λ_i of the characteristic polynomial (*i.e.*, for any eigenvalue λ_i). In this case, there are k_i linearly independent eigenvectors associated to λ_i; we say that the eigenspace dimension k_i' of λ_i (the geometric multiplicity k_i') is equal to the algebraic multiplicity k_i. This step is relatively straightforward and it is proved in full detail (and slightly generalized) in Prop. B.2.2.

Step. 2 We need to show that the eigenspace of λ_i has dimension k_i. To show this, we use the notion of similar matrices: X and Y are similar if there exists U such that $X = UYU^{-1}$; we say that Y is the representation of X in the basis composed by the columns of U. Two simple properties (Prop. B.1.4 and B.1.5) show that: (i) two similar matrices have the same characteristic polynomial so that any common eigenvalue λ_i is repeated k_i times and (ii) similar matrices have the same eigenspace dimension k_i' for any common eigenvalue λ_i. Assuming $k_i' < k_i$ for symmetric $A \in \mathbb{R}^{n\times n}$, we can construct $U \in \mathbb{R}^{n\times n}$ by putting k_i' orthogonal unitary eigenvectors of λ_i on the first k_i' columns and by filling the other columns (introducing $n - k_i'$ unitary vectors perpendicular on the first k_i' columns) such that $U^\top U = I_n$. We obtain $U^\top AU = \begin{bmatrix} \lambda_i I_{k_i'} & \mathbf{0} \\ \mathbf{0} & D \end{bmatrix}$. Since similar matrices have the same characteristic polynomial and $k_i' < k$, λ_i needs to be a root of the characteristic polynomial of D, and so, λ_i needs to have an eigenvector in D. This means that the eigenspace of λ_i in $\begin{bmatrix} \lambda_i I_{k_i'} & \mathbf{0} \\ \mathbf{0} & D \end{bmatrix}$ has dimension at least $k_i' + 1$, contradicting the fact that similar matrices have the same eigenspace dimension for λ_i. As such, the assumption $k_i' < k_i$ was false and we obtain $k_i' = k_i$, *i.e.*, λ_i has the same algebraic and geometric multiplicity k_i. For the skeptical, this Step 2 is proven in full detail in Prop. B.2.3.

1.2 Equivalent SDP definitions

Definition 1.2.1. *A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive semidefinite (SDP) if the following holds for any vector $\mathbf{x} \in \mathbb{R}^n$*

$$\mathbf{x}^\top A\mathbf{x} = \mathbf{x} \bullet A\mathbf{x} = A \bullet \mathbf{x}\mathbf{x}^\top = \mathit{trace}(X\mathbf{x}\mathbf{x}^\top) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j \geq 0, \tag{1.2.1}$$

where $\bullet$ stands for the scalar product. If the above inequality is always strict, the matrix is positive definite. If the inequality is reversed, the matrix is negative semidefinite.

Proposition 1.2.2. *A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive semidefinite (resp. definite) if and only if all eigenvalues λ_i verify $\lambda_i \geq 0$ (resp. > 0).*

Proof.
$\Longrightarrow$
We consider A is positive semidefinite (resp. definite). Assume there exist an eigenvalue $\lambda < 0$ (resp $\lambda \leq 0$) associated to eigenvector $\mathbf{v}$. We have $\mathbf{v}^\top A\mathbf{v} = \mathbf{v}^\top \lambda \mathbf{v} = \lambda(v_1^2 + v_2^2 + \ldots v_n^2)$. If $\lambda < 0$ (resp $\lambda \leq 0$), then

$\mathbf{v}^\top A\mathbf{v} < 0$ (resp $\mathbf{v}^\top A\mathbf{v} \leq 0$). Thus, A is not positive semidefinite (resp. definite). This is a contradiction, and so, the assumption $\lambda < 0$ (resp $\lambda \leq 0$) was false. We need to have $\lambda \geq 0$ (resp $\lambda > 0$).
$\Longleftarrow$
Without loss of generality, we suppose

$$\lambda_1 \leq \lambda_2 \leq \ldots \lambda_n$$

We consider λ_1 satisfies $\lambda_1 \geq 0$ (resp $\lambda_1 > 0$). We consider the eigenvalue decomposition of symmetric matrix A, as constructed in (1.1.1) – see also (B.2.3) of Proposition B.2.1.

$$A = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top, \tag{1.2.2}$$

where $\mathbf{v}_1,\ \mathbf{v}_2, \ldots, \mathbf{v}_n$ are the unitary orthonormal eigenvectors of A. We consider the following minimization problem over all unitary vectors $\mathbf{x} \in \mathbb{R}^n$ and we will show it is non-negative (resp. strictly positive):

$$\min_{|\mathbf{x}|=1} \mathbf{x} \bullet A\mathbf{x} \tag{1.2.3}$$

The above formula $\mathbf{x} \bullet A\mathbf{x}$ with unitary $\mathbf{x}$ is actually a particular case of the Rayleigh ratio/quotient usually written under the form $R(A, \mathbf{x}) = \frac{\mathbf{x} \bullet A\mathbf{x}}{\mathbf{x} \bullet \mathbf{x}}$.

Lemma 1.2.2.1. *The minimum value of (1.2.3) above is the smallest eigenvalue λ_1. This minimum is attained by the unit eigenvector $\mathbf{v}_1$ of λ_1. This lemma actually holds for any real symmetric matrix A. This also implies that the maximum eigenvalue can be determined by calculating the minimum eigenvalue of $-A$, i.e., $\lambda_n = -\lambda_{min}(-A) = -\min_{|\mathbf{x}|=1} -\mathbf{x} \bullet A\mathbf{x} = \max_{|\mathbf{x}|=1} \mathbf{x} \bullet A\mathbf{x}$.*

Proof. We can write $\mathbf{x}$ in basis $\mathbf{v}_1,\ \mathbf{v}_2, \ldots \mathbf{v}_n$. This is always possible because the equation $[\mathbf{v}_1\ \mathbf{v}_2\ \ldots \mathbf{v}_n]\mathbf{a} = \mathbf{x}$ always has the solution $\mathbf{a} = [\mathbf{v}_1\ \mathbf{v}_2\ \ \ldots \mathbf{v}_n]^\top \mathbf{x}$ simply using (1.1.2). We can write

$$\mathbf{x} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \ldots a_n\mathbf{v}_n \tag{1.2.4}$$

Since $|\mathbf{x}| = 1$, we have $\sum_{i=1}^n a_i^2 \mathbf{v}_i \bullet \mathbf{v}_i + 2\sum_{i=1}^n \sum_{j=1, j\neq i}^n a_i a_j \mathbf{v}_i \bullet \mathbf{v}_j = 1$. Since $\mathbf{v}_i \bullet \mathbf{v}_j = 0\ \forall i \neq j$ and $\mathbf{v}_i \bullet \mathbf{v}_i = 1$, we obtain $1 = \sum_{i=1}^n a_i^2$. We now replace this in (1.2.3) and we obtain

$$\begin{aligned}
\min_{|\mathbf{x}|=1} \mathbf{x} \bullet A\mathbf{x} &= \min_{\sum_{i=1}^n a_i^2 = 1} \left(\sum_{i=1}^n a_i \mathbf{v}_i\right) \bullet A \left(\sum_{i=1}^n a_i \mathbf{v}_i\right) \\
&= \min_{\sum_{i=1}^n a_i^2 = 1} \left(\sum_{i=1}^n a_i \mathbf{v}_i\right) \bullet \left(\sum_{i=1}^n a_i \lambda_i \mathbf{v}_i\right) \\
&= \min_{\sum_{i=1}^n a_i^2 = 1} \sum_{i=1}^n \lambda_i a_i^2 \\
&\geq \min_{\sum_{i=1}^n a_i^2 = 1} \sum_{i=1}^n \lambda_1 a_i^2 \\
&= \lambda_1,
\end{aligned} \tag{1.2.5}$$

The inequality is not strict. Using $\mathbf{x} = \mathbf{v}_1$ we obtain $\mathbf{v}_1 \bullet A\mathbf{v}_1 = \lambda_1 \mathbf{v}_1 \bullet \mathbf{v}_1 = \lambda_1$. The proof has not used the fact that $\lambda_1 \geq 0$, and so, it can be applied for any symmetric matrix A. □

We recall that any $\mathbf{y} \in \mathbb{R}^n$ can be written as $\mathbf{y} = \alpha\mathbf{x}$ such that $|\mathbf{x}| = 1$ (technically $\alpha = |\mathbf{y}| = \sqrt{\sum_{i=1}^n y_i^2}$). We obtain $\mathbf{y}^\top A\mathbf{y} = \alpha^2\mathbf{x}^\top A\mathbf{x} = \alpha^2\mathbf{x} \bullet A\mathbf{x} \geq \alpha^2\lambda_1$. Observe this minimum $\alpha^2\lambda_1$ can always be attained by $\mathbf{y} = \alpha\mathbf{v}_1$. If $\lambda_1 \geq 0$ (resp. $\lambda_1 > 0$), the matrix A is positive semidefinite (resp. definite). □

Proposition 1.2.3. *We say that matrices S and T are* `congruent` *if there is a non-singular Q such that $T = Q^\top SQ$. Two congruent matrices have the same SDP status, i.e., S is semidefinite positive if and only if T is semidefinite positive and S is definite positive if and only if T is definite positive:*

$$S \succeq \mathbf{0} \iff T \succeq \mathbf{0} \text{ and } S \succ \mathbf{0} \iff T \succ \mathbf{0}.$$

Proof.

$\implies$

If S is SDP, then $\mathbf{x}^\top T\mathbf{x} = \mathbf{x}^\top Q^\top SQ\mathbf{x} = (Q\mathbf{x})^\top S(Q\mathbf{x}) \geq 0 \implies T \succeq \mathbf{0}$. Using the fact that Q is non-singular, we also obtain that $Q\mathbf{x}$ is zero only when $\mathbf{x}$ is zero, and so, $S \succ \mathbf{0} \implies (Q\mathbf{x})^\top S(Q\mathbf{x}) > 0 \forall \mathbf{x} \neq \mathbf{0} \implies T \succ \mathbf{0}$.

$\impliedby$

The converse proof is identical, because we can write $S = \left(Q^{-1}\right)^\top TQ^{-1}$ and apply the above two lines argument on swapped S and T. We simply used $(Q^\top)^{-1} = (Q^{-1})^\top$, which is equivalent to $Q^\top(Q^{-1})^\top = I_n$, which follows from transposing $Q^{-1}Q = I_n$. □

We now use the above result[2] to introduce a (very practical) remark on how certain well-known elementary row/column operations preserve the SDP status.

Proposition 1.2.4. *It is well known that the operations below preserve the determinant; we now prove that they also preserve the SDP status: (a) the initial matrix A is SDP if only if the transformed matrix A' is SDP and (b) A is positive definite if and only if the A' is positive definite. Finally, A and A' are also congruent.*

(i) add row i multiplied by $z_{ji} \in \mathbb{R}$ to row j and then column i multiplied by z_{ji} to column j

(ii) perform a sequence of row operations as above and then the corresponding (transposed) column operations

(iii) permute the rows of A and then permute in the same manner (apply the same permutation) on the columns of A

Proof. The row operation from (i) amounts to performing $(I_n + Z_{j\swarrow i})\, A$, where $Z_{j\swarrow i}$ is a matrix that contains only one non-zero element: put z_{ji} at row j and column i. The column operation from (i) amounts to multiplying at right by $(I_n + Z_{j\swarrow i})^\top = I_n + Z_{i\swarrow j}$. The final matrix resulting from operation (i) is:

$$A' = (I_n + Z_{j\swarrow i})\, A\, (I_n + Z_{j\swarrow i})^\top$$

As such, A and A' become congruent (notice $\det(I_n + Z_{j\swarrow i}) = 1$), finishing the proof by Prop 1.2.3 above. The operation (ii) simply consists of applying (i) several times, leading to the following congruent matrices:

$$A' = (I_n + Z_{j_1\swarrow i_1})\,(I_n + Z_{j_2\swarrow i_2}) \ldots \left(I_n + Z_{j_p\swarrow i_p}\right) A \left(I_n + Z_{j_p\swarrow i_p}\right)^\top \ldots (I_n + Z_{j_2\swarrow i_2})^\top (I_n + Z_{j_1\swarrow i_1})^\top$$

The operation (iii) does not change the SDP status because the permutation (reordering) that transforms A into A' can be applied to transform $\mathbf{x}$ into some $\mathbf{x}'$ so that $A \bullet \mathbf{x}\mathbf{x}^\top = A' \bullet \mathbf{x}'\mathbf{x}'^\top$, and this operation can also be reversed. □

Proposition 1.2.5. *A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive semidefinite if and only if all principal minors (recall Def A.2.1) are non-negative. This is equivalent to stating $A \succeq \mathbf{0} \iff det([A]_J) \geq 0\ \forall J \subseteq [1..n]$, where the operator $[\cdot]_J$ represents the principal minor obtained by selecting rows J and columns J.*

We will later see that A is positive ***`definite`*** *if and only if all leading principal minors are strictly positive (Sylvester criterion, see Prop. 1.5.2), which is also equivalent to $det([A]_J) > 0\ \forall J \subseteq [1..n]$ using Prop. 1.6.3.*

Proof.

$\implies$

Take any subset of indices $J \subseteq [1..n]$ and consider any vector $\overline{\mathbf{x}}_J \in \mathbb{R}^n$ that contains non-zero elements only on positions J. Let $\mathbf{x}_J$ be the $|J|$-dimensional vector obtained by extracting/keeping only the positions J of $\overline{\mathbf{x}}_J$. Using the SDP definition (1.2.1), the following needs to hold:

$$A \bullet \overline{\mathbf{x}}_J\overline{\mathbf{x}}_J^\top \geq 0 \tag{1.2.6}$$

Since $\overline{\mathbf{x}}_J\overline{\mathbf{x}}_J^\top$ contains non-zero elements only on lines J and columns J, we can re-write above formula as:

$$[A]_J \bullet \mathbf{x}_J\mathbf{x}_J^\top \geq 0 \tag{1.2.7}$$

[2] The same result appears in Section 6.9.1 (p. 91) of the lecture notes of Maurício de Oliveira, available on-line as of 2019 at http://maecourses.ucsd.edu/~mdeolive/mae280b/lecture/lecture6.pdf.

Since this holds for any $\mathbf{x}_J$, the principal minor $[A]_J$ is SDP. This means that the eigenvalues of $[A]_J$ are non-negative (Prop. 1.2.2), and so, the determinant of $[A]_J$ is non-negative because it is equal to the product of its eigenvalues (Prop. A.2.4).

$\Longleftarrow$

Let r be the rank of A. Based on Prop. A.1.2, A has at least a non-zero *principal* minor of order r. We can reorder the rows and columns of A to move this principal minor in the upper-left corner; we obtain a (permuted) matrix $A' = \left[\begin{smallmatrix} \overline{\mathcal{A}} & B^\top \\ B & C \end{smallmatrix}\right]$, where $\overline{\mathcal{A}}$ is non singular; A, A' and $\overline{\mathcal{A}}$ have rang r. Prop 1.2.4.(iii), certifies that A' has the same SDP status as A. Since A' has rang r, the bottom $n-r$ rows (*i.e.*$[\,B\ C\,]$), can be written as a linear combination of the first r rows (*i.e.*, $[\,\overline{\mathcal{A}}\ B^\top\,]$). We can subtract this linear combination of the first r rows from the last $n-r$ rows to cancel them (make them zero). After performing the transposed operation on the columns, we obtain a matrix $A'' = \left[\begin{smallmatrix} \overline{\mathcal{A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right]$ that has the same SDP status as A' and A, using Prop. 1.2.4.(ii). To prove $A, A', A'' \succeq \mathbf{0}$, it is enough to solve the following (sub-)problem:

$$\text{Any } \textit{non-singular} \text{ symmetric } \overline{\mathcal{A}} \in \mathbb{R}^{r\times r} \text{ that satisfies } \det([\overline{\mathcal{A}}]_J) \geq 0\ \forall J \subseteq [1..r] \text{ is SDP} \qquad (*)$$

First, notice $\overline{\mathcal{A}}_{r,r} > 0$ because $\overline{\mathcal{A}}_{r,r} = 0$ would lead to $\overline{\mathcal{A}}_{i,r} = 0\ \forall i \in [1..r-1]$ (otherwise the 2×2 minor of $\overline{\mathcal{A}}$ selecting rows/columns i and r would be negative) and if the last column of $\overline{\mathcal{A}}$ is zero then $\det(\overline{\mathcal{A}}) = 0$, contradiction. We can now subtract the last row r from each other row $i \in [1..r-1]$ in such a manner (*i.e.*, after multiplying it by $\frac{\overline{\mathcal{A}}_{i,r}}{\overline{\mathcal{A}}_{r,r}}$) to cancel all elements on the last column above $\overline{\mathcal{A}}_{r,r}$. We then apply the same row operations but transposed (generating column operations). This leads to a matrix $\left[\begin{smallmatrix} \overline{\mathcal{A}}_{r-1} & \mathbf{0} \\ \mathbf{0} & \overline{\mathcal{A}}_{r,r} \end{smallmatrix}\right]$ with the same SDP status as $\overline{\mathcal{A}}$ by virtue of Prop. 1.2.4.(ii). Also, any principal minor $\overline{\mathcal{A}}_J$ of $\overline{\mathcal{A}}_{r-1}$ is non-singular because it corresponds to a principal minor $\left[\begin{smallmatrix} \overline{\mathcal{A}}_J & \mathbf{0} \\ \mathbf{0} & \overline{\mathcal{A}}_{r,r} \end{smallmatrix}\right]$ whose determinant is $\det(\overline{\mathcal{A}}_J)\cdot \overline{\mathcal{A}}_{r,r}$ where $\overline{\mathcal{A}}_{r,r} > 0$. Thus, to prove (*), it is enough to prove $\overline{\mathcal{A}}_{r-1} \succeq \mathbf{0}$. But since $\overline{\mathcal{A}}_{r-1}$ satisfies all conditions of (*), we have actually obtained the same problem (*) on a dimension reduced by one. We can repeat this until we remain with a matrix of size 1 and this proves $\overline{\mathcal{A}} \succeq 0$, enough to certify $A \succeq 0$. ☐

1.3 Schur complements, the self-duality of the SDP cone and related properties

Proposition 1.3.1. *(Schur complements particular case) The* $(n+1)\times(n+1)$ *matrix* $\left[\begin{smallmatrix} 1 & \mathbf{b}^\top \\ \mathbf{b} & C \end{smallmatrix}\right]$ *is SDP if and only if* $C - \mathbf{b}\mathbf{b}^\top$ *is SDP.*

Proof. We will give two proofs. The first one produces a congruent matrix using row/column operations. The second one is actually a formalization of the first, but it uses a "magical" short decomposition.

Proof 1) Let us subtract the first row of $\left[\begin{smallmatrix} 1 & \mathbf{b}^\top \\ \mathbf{b} & C \end{smallmatrix}\right]$ from all other rows $i+1$ ($\forall i \in [1..n]$) premultiplying $[1\ \mathbf{b}^\top]$ with b_i. We then perform the transposed operation on the resulting matrix. The two operations lead to

$$\begin{bmatrix} 1 & \mathbf{b}^\top \\ \mathbf{b} & C \end{bmatrix} \to \begin{bmatrix} 1 & \mathbf{b}^\top \\ \mathbf{0}_n & C - \mathbf{b}\mathbf{b}^\top \end{bmatrix} \to \begin{bmatrix} 1 & \mathbf{0}_n^\top \\ \mathbf{0}_n & C - \mathbf{b}\mathbf{b}^\top \end{bmatrix} \tag{1.3.1}$$

Using Prop 1.2.4, the above two operations together lead to a congruent matrix (at right) with the same SDP status as the initial one (at left). The second matrix is SDP if and only if $C - \mathbf{b}\mathbf{b}^\top$ is SDP (the "$\Rightarrow$" implication follows from performing a scalar product with any $\mathbf{x}\mathbf{x}^\top$ with $x_0 = 0$ and the "$\Leftarrow$" implication follows from the fact that the sum of two SDP matrices is SDP).

Proof 2) We formalize (1.3.1) using matrix multiplications. The first transformation can be realized by:

$$\begin{bmatrix} 1 & \mathbf{b}^\top \\ \mathbf{0}_n & C - \mathbf{b}\mathbf{b}^\top \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}_n^\top \\ -\mathbf{b} & I_n \end{bmatrix} \begin{bmatrix} 1 & \mathbf{b}^\top \\ \mathbf{b} & C \end{bmatrix} \tag{1.3.2}$$

and the second one by:

$$\begin{bmatrix} 1 & \mathbf{0}_n^\top \\ \mathbf{0}_n & C - \mathbf{b}\mathbf{b}^\top \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{b}^\top \\ \mathbf{0}_n & C - \mathbf{b}\mathbf{b}^\top \end{bmatrix} \begin{bmatrix} 1 & -\mathbf{b}^\top \\ \mathbf{0}_n & I_n \end{bmatrix} \tag{1.3.3}$$

Combining (1.3.2)-(1.3.3), we obtain:

$$\begin{bmatrix} 1 & \mathbf{0}_n^\top \\ \mathbf{0}_n & C - \mathbf{b}\mathbf{b}^\top \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}_n^\top \\ -\mathbf{b} & I_n \end{bmatrix} \begin{bmatrix} 1 & \mathbf{b}^\top \\ \mathbf{b} & C \end{bmatrix} \begin{bmatrix} 1 & -\mathbf{b}^\top \\ \mathbf{0}_n & I_n \end{bmatrix}$$

We obtain again that $\left[\begin{smallmatrix}1 & \mathbf{0}_n^\top \\ \mathbf{0}_n & C-\mathbf{b}\mathbf{b}^\top\end{smallmatrix}\right]$ is congruent to $\left[\begin{smallmatrix}1 & \mathbf{b}^\top \\ \mathbf{b} & C\end{smallmatrix}\right]$, because $\det\left[\begin{smallmatrix}1 & \mathbf{0}_n^\top \\ -\mathbf{b} & I_n\end{smallmatrix}\right] = 1$. This finishes the proof by virtue of Prop 1.2.3. □

Proposition 1.3.2. *(Schur complements general case) Given positive definite $A \in \mathbb{R}^{m\times m}$, the $(n+m)\times(n+m)$ matrix $\left[\begin{smallmatrix}A & B^\top \\ B & C\end{smallmatrix}\right]$ is SDP if and only if $C - BA^{-1}B^\top$ is SDP.*

Proof. As in previous Prop. 1.3.1, we want to subtract from the (bottom) n rows that cover $B \in \mathbb{R}^{n\times m}$ a linear combination of the top m rows so as to cancel (make zero) all terms of B. We look for a matrix $X \in \mathbb{R}^{n\times m}$ such that $XA = -B$; incidentally, X and B have the same size because the multiplication with a square matrix conserves the size (of A). By this operation, each row i of X generates a linear combination of the rows of A that equals the negative of row i of B. The transpose of this operation is applied on columns to cancel $B^\top$. It is clear that $X = -BA^{-1}$. This explains the bottom-left term of matrix U below.

$$\begin{bmatrix} A & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & C - BA^{-1}B^\top \end{bmatrix} = \underbrace{\begin{bmatrix} I_m & \mathbf{0}_{m\times n} \\ -BA^{-1} & I_n \end{bmatrix}}_{U} \begin{bmatrix} A & B^\top \\ B & C \end{bmatrix} \underbrace{\begin{bmatrix} I_m & -A^{-1}B^\top \\ \mathbf{0}_{n\times m} & I_n \end{bmatrix}}_{U^\top} = (A^{-1})^\top B^\top$$

Notice that $U^\top$ is written after applying the simplification $(A^{-1})^\top = A^{-1}$ (which follows from $I = (A^{-1}A)^\top = A^\top\left(A^{-1}\right)^\top \implies \left(A^{-1}\right)^\top = \left(A^\top\right)^{-1} = A^{-1}$). Since $\det(U) = 1$, we obtain that $\left[\begin{smallmatrix}A & \mathbf{0}_{m\times n} \\ \mathbf{0}_{n\times m} & C-BA^{-1}B^\top\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}A & B^\top \\ B & C\end{smallmatrix}\right]$ are congruent. We finish with Prop 1.2.3 as for Prop 1.3.1 above. □

Proposition 1.3.3. *A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is SDP if and only if $A \bullet B \geq 0$ for any SDP matrix B. We say that the cone S_n^+ of SDP matrices is self-dual.*

Proof. We apply the eigen-decomposition (1.1.1) on A and B: :

$$A = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top, \quad B = \sum_{i=1}^{n} \lambda'_i \mathbf{u}_i \mathbf{u}_i^\top, \tag{1.3.4}$$

Lemma 1.3.3.1. *Given vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, we have*

$$(\mathbf{u}\mathbf{u}^\top) \bullet (\mathbf{v}\mathbf{v}^\top) = (\mathbf{u} \bullet \mathbf{v})^2$$

Proof.

$$(\mathbf{u}\mathbf{u}^\top) \bullet (\mathbf{v}\mathbf{v}^\top) = \mathbf{v}^\top(\mathbf{u}\mathbf{u}^\top)\mathbf{v} = (\mathbf{v}^\top\mathbf{u})(\mathbf{u}^\top\mathbf{v}) = (\mathbf{u}^\top\mathbf{v})^\top(\mathbf{u}^\top\mathbf{v}) = (\mathbf{u} \bullet \mathbf{v})^2$$

□

$\implies$
If A is SDP, then $\lambda_i \geq 0 \forall i \in [1..n]$. Using substitution (1.3.4), we obtain

$$A \bullet B = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda'_j (\mathbf{v}_i\mathbf{v}_i^\top) \bullet (\mathbf{u}_j\mathbf{u}_j^\top) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda'_j (\mathbf{v}_i \bullet \mathbf{u}_j)^2,$$

where we used Lemma 1.3.3.1 for the last equality. This shows that $A \bullet B \geq 0$.
$\impliedby$
Let us take some $i \in [1..n]$ and consider $B = \mathbf{v}_i^\top \mathbf{v}_i$, where recall that $\mathbf{v}_i$ is a unit eigenvector of A. Since $A \bullet B \geq 0$, we deduce $A \bullet \mathbf{v}_i^\top \mathbf{v}_i \geq 0$, or $\mathbf{v}_i^\top A \mathbf{v}_i \geq 0$, equivalent to $\mathbf{v}_i^\top \lambda_i \mathbf{v}_i \geq 0$. This means that $\lambda_i \geq 0$. All eigenvalues of A are non-negative, and so, A is SDP. □

Proposition 1.3.4. *If A, B are SDP, then $A \bullet B \geq 0$ and $A \bullet B = 0 \iff AB = \mathbf{0}_{n\times n}$.*

Proof. We apply the eigen-decomposition (1.1.1) listing the only terms with non-zero eigenvalues:

$$A = \sum_{i=1}^{r} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top, \quad B = \sum_{i=1}^{r'} \lambda'_i \mathbf{u}_i \mathbf{u}_i^\top \tag{1.3.5}$$

where r and r' are the ranks of A and resp. B (the number of non-zero eigenvalues, see Prop A.1.7). We now use Lemma 1.3.3.1 to calculate

$$A \bullet B = \sum_{i=1}^{n}\sum_{j=1}^{r} \lambda_i\lambda'_j(\mathbf{v}_i\mathbf{v}_i^\top) \bullet (\mathbf{u}_j\mathbf{u}_j^\top) = \sum_{i=1}^{n}\sum_{j=1}^{r'} \lambda_i\lambda'_j(\mathbf{v}_i \bullet \mathbf{u}_j)^2 \geq 0$$

If $A \bullet B = 0$, then all terms $\mathbf{v}_i \bullet \mathbf{u}_j$ with $i \in [1..r]$ and $j \in [1..r']$ need to be zero (recall we only use strictly positive eigenvalues). Now observe

$$AB = \sum_{i=1}^{r}\sum_{j=1}^{r'} \lambda_i\lambda_j(\mathbf{v}_i\mathbf{v}_i^\top)(\mathbf{u}_j\mathbf{u}_j^\top) = \sum_{i=1}^{r}\sum_{j=1}^{r'} \lambda_i\lambda_j\mathbf{v}_i(\mathbf{v}_i^\top\mathbf{u}_j)\mathbf{u}_j^\top = \sum_{i=1}^{r}\sum_{j=1}^{r'} \lambda_i\lambda_j\mathbf{v}_i \cdot 0 \cdot \mathbf{u}_j^\top = \mathbf{0}_{n\times n}$$

We still need to show the converse: $AB = \mathbf{0}_{n\times n} \implies A \bullet B = 0$. Taking any $k \in [1..n]$, the k^{th} diagonal element of AB is $0 = \sum_{\ell=1}^{n} a_{k\ell}b_{k\ell}$ (we used the symmetry of B). Summing up for all k we obtain $0 = \sum_{k=1}^{n}\sum_{\ell=1}^{n} a_{k\ell}b_{k\ell} = A \bullet B$. □

1.4 Three easy ways to generate (semi-)definite positive matrices

There are at least three easy ways of generating semidefinite (or definite) positive matrices.

1. Generate $A \in \mathbb{R}^{n\times n}$ by taking $A = V^\top V$ for any $V \in \mathbb{R}^{p\times n}$. It is easy to verify $\mathbf{x}^\top V^\top V\mathbf{x} = (V\mathbf{x})^\top(V\mathbf{x}) = |V\mathbf{x}|^2 \geq 0\ \forall \mathbf{x} \in \mathbb{R}^n$. If V has rank n, then $V\mathbf{x}$ is non-zero for any non-zero $\mathbf{x}$, and so, $|V\mathbf{x}|^2 > 0\ \forall \mathbf{x} \in \mathbb{R}^n - \{\mathbf{0}\}$, meaning that $A = V^\top V$ is positive definite. If $S \succ \mathbf{0}$ and V has rank n, we also have $V^\top SV \succ \mathbf{0}$. As a side remark, notice $rank(A) = rank(V)$ using Prop. A.1.8 (based on the rank–nullity theorem)

2. Take a diagonally dominant matrix such that $A_{ii} \geq r_i = \sum_{\substack{j=1\\ j\neq i}}^{n} |A_{ij}|\ \forall i \in [1..n]$. Any eigenvalue λ of such matrices verify $|\lambda - A_{ii}| \leq r_i$ for some $i \in [1..n]$ by virtue of the (relatively easy to prove) Gershgorin circle Theorem A.2.8.[3] If $A_{ii} \geq r_i\ \forall i \in [1..n]$, we need to have $\lambda \geq 0$. The matrix $A + \varepsilon I_n$ is *positive definite* for any $\varepsilon > 0$.

3. Take $A = |X|I_n + X$, where $|X| = \sqrt{X \bullet X}$. This follows from the fact that the minimum eigenvalue of X is greater than or equal to $-|X|$, by virtue of Prop. 1.4.1 below.

Proposition 1.4.1. *Given symmetric $X \in \mathbb{R}^{n\times n}$, the Frobenius norm $|X| = \sqrt{\sum_{i,j=1}^{n} X_{ij}^2} = \sqrt{X \bullet X}$ is equal to $\sqrt{\lambda_1^2 + \lambda_2^2 + \dots \lambda_n^2}$, where $\lambda_1,\ \lambda_2, \dots \lambda_n$ are the eigenvalues of X. This means that the maximum eigenvalue of X is at most $|X|$ and the minimum eigenvalue is at least $-|X|$.*

Proof. Standard calculations can confirm $|X| = \sqrt{X \bullet X} = \sqrt{\mathtt{trace}(XX)}$; and more generally we have $\mathtt{trace}(XY) = X \bullet Y$. We apply the eigendecomposition (1.1.1) to write symmetric X in the form $X = U\mathtt{diag}(\lambda_1,\ \lambda_2,\ \dots\lambda_n)U^\top$, where $U^\top = U^{-1}$. We obtain $XX = U\mathtt{diag}(\lambda_1^2,\ \lambda_2^2,\ \dots\lambda_n^2)U^\top$. This means that the eigenvalues of XX are $\lambda_1^2,\ \lambda_2^2,\ \dots\lambda_n^2$, see also Prop B.1.4. Since the trace is the sum of the eigenvalues (see Prop. A.2.4), we obtain

$$|X| = \sqrt{\mathtt{trace}(XX)} = \sqrt{\lambda_1^2 + \lambda_2^2 + \dots + \lambda_n^2}. \tag{1.4.1}$$

Is is clear X can have no eigenvalue strictly larger than $|X|$ or strictly lower than $-|X|$, because this would violate (1.4.1).[4] □

[3] I found this approach at page 4 of the Habilitation thesis (*Habilitationsschrift*) of Christoph Helmberg "Semidefinite Programming for Combinatorial Optimization", Technical University of Berlin (*Technische Universität Berlin*), The Zuse Institute Berlin (*Konrad-Zuse-Zentrum für Informationstechnik Berlin*), ZIB-report ZR-00-34, available on-line as of 2017 at http://opus4.kobv.de/opus4-zib/files/602/ZR-00-34.pdf.

[4] I first found this result in Section 10.1 of the lecture notes of Robert Freund "Introduction to Semidefinite Programming (SDP)", available on-line as of 2017 at https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-251j-introduction-to-mathematical-programming-fall-2009/readings/MIT6_251JF09_SDP.pdf.

1.5 Positive definite matrices: unique Cholesky factorization and Sylvester criterion

Proposition 1.5.1. *(Cholesky factorization of positive definite matrices) A real symmetric matrix A is positive definite if and only if it can be factorized as:*

$$A = RR^\top = \begin{bmatrix} r_{11} & 0 & 0 & \dots & 0 \\ r_{21} & r_{22} & 0 & \dots & 0 \\ r_{31} & r_{32} & r_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \dots & r_{nn} \end{bmatrix} \begin{bmatrix} r_{11} & r_{21} & r_{31} & \dots & r_{n1} \\ 0 & r_{22} & r_{32} & \dots & r_{n2} \\ 0 & 0 & r_{33} & \dots & r_{n3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & r_{nn} \end{bmatrix},$$

where the diagonal terms are strictly positive. The factorization is unique.
`Practical hint:` It could be useful to interpret $A = RR^\top$ in the sense that A_{ij} is the product of rows i and j of R; only the first i (resp. j) components of row i (resp. j) are non-zero.

Proof.
$\Longleftarrow$
Take any non-zero $\mathbf{x} \in \mathbb{R}^n - \{\mathbf{0}_n\}$ and observe that $\mathbf{x}^\top A\mathbf{x} = \mathbf{x}^\top RR^\top \mathbf{x} = (R^\top \mathbf{x})^\top (R^\top \mathbf{x})$. Writing $\mathbf{y} = R^\top \mathbf{x}$, this value is equal to $\sum_{i=1}^n y_i^2 \geq 0$. This inequality is strict because the only $\mathbf{y} \in \mathbb{R}^n$ such that $\sum_{i=1}^n y_i^2 = 0$ is $\mathbf{y} = \mathbf{0}_n$ and because $\mathbf{y} = R^\top \mathbf{x}$ can not be $\mathbf{0}_n$ for any non-zero $\mathbf{x}$ (since $det(R^\top) = r_{11}r_{22}\dots r_{nn} > 0$).
$\Longrightarrow$
We proceed by induction. The implication is obviously true for $n = 1$. We suppose that there exists a unique factorization:

$$[A]_{n-1} = [R]_{n-1}[R]_{n-1}^\top = \begin{bmatrix} r_{11} & 0 & \dots & 0 \\ r_{21} & r_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ r_{n-1,1} & r_{n-1,2} & \dots & r_{n-1,n-1} \end{bmatrix} \begin{bmatrix} r_{11} & r_{21} & \dots & r_{n-1,1} \\ 0 & r_{22} & \dots & r_{n-1,2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & r_{n-1,n-1} \end{bmatrix},$$

where $r_{11},\ r_{22},\ \dots r_{n-1,n-1} > 0$. We will prove that this decomposition can be extended to a $n \times n$ decomposition for matrix A. The values $r_{1n},\ r_{2n}, \dots r_{n-1,n}$ are set to zero by definition to preserve the decomposition of $[A]_{n-1}$. We can exactly determine $r_{n,1},\ r_{n,2}, \dots r_{n,n}$ using the following calculations:

(a) $r_{n1} = \dfrac{a_{n1}}{r_{11}}$, based on $a_{n1} = \mathbf{r}_{n,\times} \bullet \mathbf{r}_{1,\times}$, where $\mathbf{r}_{i,\times}$ is the i^{th} line of R.

(b) $r_{n2} = \dfrac{a_{n2} - r_{21}r_{n1}}{r_{22}}$, based on $a_{n2} = \mathbf{r}_{n,\times} \bullet \mathbf{r}_{2,\times}$.

(c) $r_{ni} = \dfrac{a_{ni} - r_{i1}r_{n1} - r_{i2}r_{n2} - \dots - r_{i,i-1}r_{n,i-1}}{r_{ii}}$ for any $i \leq n-1$, based on $a_{ni} = \mathbf{r}_{n,\times} \bullet \mathbf{r}_{i,\times}$.

(d) $r_{nn} = \sqrt{a_{nn} - \sum_i^{n-1} r_{ni}^2}$, so that the value of r_{nn} that makes the factorisation $A = RR^\top$ work can be potentially non-real, *e.g.*, we could have $r_{nn} = i$ so that $i^2 = -1$.

We only still need to show r_{nn} is real. Since $[R]_{n-1}$ is non-singular by the induction hypothesis, there exists $\mathbf{x} \in \mathbb{R}^{n-1}$ so that $\mathbf{x}^\top [R]_{n-1} = [r_{n,1}\ r_{n,2}, \dots r_{n,n-1}]$. By bordering $\mathbf{x}$ with a n^{th} component of value -1, we obtain $[\mathbf{x}^\top \quad -1]R = [\mathbf{0}_{n-1} \quad -r_{nn}]$. This means that $[\mathbf{x}^\top \quad -1]A\left[\begin{smallmatrix}\mathbf{x}\\-1\end{smallmatrix}\right] = [\mathbf{x}^\top \quad -1]RR^\top\left[\begin{smallmatrix}\mathbf{x}\\-1\end{smallmatrix}\right] = [\mathbf{0}_{n-1} \quad -r_{nn}][\mathbf{0}_{n-1} \quad -r_{nn}]^\top = r_{nn}^2$. We thus have $r_{nn}^2 > 0$ because $A \succ \mathbf{0}$; recalling how r_{nn} was determined at point (d) above, it is clear that r_{nn} can not be an imaginary number, *i.e.*, we can only have $r_{nn} \in \mathbb{R}_+$. □

Proposition 1.5.2. *(Sylvester criterion) A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive definite if and only all leading principal minors are strictly positive. By symmetrically permuting the rows and columns, this is equivalent to the fact that any nested sequence of principal minors contains only strictly positive minors. Principal minors A'' and A' corresponding to rows/columns J'' and resp. J' are nested if and only if $J'' \subsetneq J'$ and $|J''| + 1 = |J'|$.*

This theorem can be tackled from many different angles. The interested reader may be able to find a proof by himself (in less than an hour) if all material presented until here (including Appendix A) has been acquired. We present below three proofs, so as to gain an insight from every possible angle.

Proof 1

We proceed by induction. Both implications are obviously true for $n = 1$. We now show how to move from $n-1$ to n. We can use the fact that the $(n-1)\times(n-1)$ leading principal minor $[A]_{n-1}$ has a non-null determinant. Since $\det[A]_{n-1} \neq 0$, there exists $\mathbf{x} \in \mathbb{R}^{n-1}$ such that $\mathbf{x}^\top [A]_{n-1}$ is equal to the first $n-1$ positions of the last row of A. After subtracting this linear combination of the first $n-1$ rows of A from the last row followed by the transposed operation on columns, we obtain a matrix of the following form:

$$A' = \begin{bmatrix} & & & 0 \\ & [A]_{n-1} & & 0 \\ & & & \vdots \\ 0 & 0 & \dots & a'_{n,n} \end{bmatrix}.$$

We finish by applying Prop 1.2.4: the above subtraction does not change the SDP status or the determinant, and so, it is enough the prove the Sylvester criterion for a matrix of the form of A' above. And this is obvious. For the direct implication, simply check that $A' \succ \mathbf{0} \implies a'_{n,n} > 0 \implies \det(A') = \det([A]_{n-1}) \cdot a'_{n,n} > 0$. For the converse, we use that $\det(A') > 0 \implies a'_{n,n} > 0$. This proves $A' \succ \mathbf{0}$ because $A' \cdot \mathbf{x}\mathbf{x}^\top = [x_1\ x_2 \dots x_{n-1}][A]_{n-1}[x_1\ x_2 \dots x_{n-1}]^\top + a'_{n,n}x_n^2$, which is strictly positive for any $\mathbf{x} \neq \mathbf{0}$ because $[A]_{n-1} \succ \mathbf{0}$ and $a'_{n,n} > 0$. □

Proof 2

$\Longrightarrow$

Using the above Cholesky factorization of positive definite matrices (Prop. 1.5.1), A can be written $A = RR^\top$, where R is a lower triangular matrix with strictly positive diagonal elements $r_{11},\ r_{22}, \dots r_{kk} > 0$. One can simply verify that $[A]_k = [R]_k[R]_k^\top$, where the operator $[\cdot]_k$ represents the leading principal minor of size k. We obtain that $det([A]_k) = det([R]_k)det([R]_k^\top) = (r_{11}r_{22}\dots r_{kk})^2 > 0$, by virtue of $r_{11}, r_{22}, \dots r_{kk} > 0$.

$\Longleftarrow$

We proceed by induction. The implication is obviously true for $n = 1$. Suppose it is true for $n-1$, so that the $(n-1)\times(n-1)$ leading principal minor $[A]_{n-1}$ is positive definite.

We need to show that A is positive definite as well. Assume the contrary: A has an eigenvalue $\lambda_u \leq 0$ with unit eigenvector u. Since $det(A) > 0$ is the product of the eigenvalues of A, λ_u can not be zero, and so, $\lambda_u < 0$. Using again the fact that $det(A)$ is the product of the eigenvectors, we obtain that A needs to have (at least) another negative eigenvalue $\lambda_v < 0$ with unit eigenvector $\mathbf{v}$.

There exist $a_u, a_v \in \mathbb{R}$ (not both 0) such that $a_u u_n + a_v v_n = 0$, *i.e.*, if $u_n = 0$ take $a_u = 1,\ a_v = 0$ and if $u_n \neq 0$ take $a_u = \frac{-v_n}{u_n},\ a_v = 1$. We define $\mathbf{x} = a_u\mathbf{u} + a_v\mathbf{v}$, so that $x_n = 0$. We compute

$$\begin{aligned}\mathbf{x}^\top A\mathbf{x} = \mathbf{x} \bullet A\mathbf{x} &= (a_u\mathbf{u} + a_v\mathbf{v}) \bullet A(a_u\mathbf{u} + a_v\mathbf{v}) \\ &= (a_u\mathbf{u} + a_v\mathbf{v}) \bullet (\lambda_u a_u\mathbf{u} + \lambda_v a_v\mathbf{v}) \\ &= \lambda_u a_u^2\mathbf{u} \bullet \mathbf{u} + \lambda_v a_v^2\mathbf{v} \bullet \mathbf{v} \\ &= \lambda_u a_u^2 + \lambda_v a_v^2 \\ &< 0\end{aligned}$$

The last inequality follows from $\lambda_u, \lambda_v < 0$ and from the fact that a_u and a_v are not both zero. We now develop $\mathbf{x}^\top A\mathbf{x} = A \bullet \mathbf{x}\mathbf{x}^\top$ and notice that matrix $\mathbf{x}\mathbf{x}^\top$ has non-zero elements only on the first $n-1$ rows and columns. As such, $A \bullet \mathbf{x}\mathbf{x}^\top < 0$ simplifies to $[A]_{n-1} \bullet [x_1\ x_2 \dots x_{n-1}][x_1\ x_2 \dots x_{n-1}]^\top < 0$, which violates the induction hypothesis that $[A]_{n-1}$ is positive definite. We obtained a contradiction on the existence of an eigenvalue $\lambda_u \leq 0$. This means that all eigenvalues of A are strictly positive. By simply applying Prop. 1.2.2, we obtain that A is positive definite. □

Proof 3

$\Longrightarrow$

We can use the proof of the "$\Longrightarrow$" implication of Prop. 1.2.5, but the inequalities (1.2.6)-(1.2.7) become strict. This means that every principal minor of A is positive definite, and, using Prop. 1.2.2, the principal minor has only strictly positive eigenvalues. The determinant of the minor is strictly positive, as it is the product of the eigenvalues (Prop. A.2.4).

$\Longleftarrow$

We proceed by induction. The statement is true for $n = 1$. Suppose it is true for $n-1$. The $(n-1)\times(n-1)$ leading principal minor is positive definite, and so, its minimum eigenvalue $\lambda_{\min}([A]_{n-1})$ is strictly positive. We then apply lemma below to show that A has at least $n-1$ strictly positive eigenvalues, *i.e.*, $0 < \lambda_{\min}([A]_{n-1}) \le \lambda_2 \le \lambda_3 \le \dots \lambda_n$. This ensures that $\lambda_1\lambda_2\dots\lambda_n = \det(A) > 0$ (recall Prop. A.2.4) can only hold because $\lambda_1 > 0$, so that actually all eigenvalues need to be positive which proves $A \succ \mathbf{0}$ (via Prop 1.2.2).

Lemma 1.5.2.1. *Consider symmetric matrix A with eigenvalues $\lambda_1 \le \lambda_2 \le \dots \le \lambda_n$. Any principal minor A' of order $n-1$ with eigenvalues $\lambda'_1 \le \lambda'_2 \le \dots \le \lambda'_{n-1}$ verifies $\lambda'_1 \le \lambda_2$.*

Proof. We consider A' is obtained from A by removing row i and column i. Denote by $\mathbf{u}_1$ and $\mathbf{u}_2$ the unit orthogonal eigenvectors of A corresponding to λ_1 and resp. λ_2, recall (1.1.1). One can surely find $\alpha_1,\ \alpha_2 \in \mathbb{R}$ (not both zero) such that vector $\mathbf{u} = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2$ satisfies $u_i = 0$. Furthermore, using an appropriate scaling, the values of α_1 and α_2 can be chosen such that $\alpha_1^2 + \alpha_2^2 = 1$. Notice that $|\mathbf{u}|^2 = (\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2)\bullet(\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2) = \alpha_1^2\mathbf{u}_1\bullet\mathbf{u}_1 + \alpha_2^2\mathbf{u}_2\bullet\mathbf{u}_2 + 2\alpha_1\alpha_2\mathbf{u}_1\bullet\mathbf{u}_2 = \alpha_1^2|\mathbf{u}_1|^2 + \alpha_2^2|\mathbf{u}_2|^2 = \alpha_1^2 + \alpha_2^2 = 1$, where we used that $\mathbf{u}_1$ and $\mathbf{u}_2$ are unitary orthogonal eigenvectors.

Let us calculate $\mathbf{u}\bullet A\mathbf{u}$ (a particular form of the Rayleigh ratio $R(A,\mathbf{u}) = \frac{\mathbf{u}\bullet A\mathbf{u}}{\mathbf{u}\bullet\mathbf{u}}$). We have $\mathbf{u}\bullet A\mathbf{u} = \mathbf{u}\bullet(\alpha_1\lambda_1\mathbf{u}_1 + \alpha_2\lambda_2\mathbf{u}_2) = (\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2)\bullet(\alpha_1\lambda_1\mathbf{u}_1 + \alpha_2\lambda_2\mathbf{u}_2) = \alpha_1^2\lambda_1 + \alpha_2^2\lambda_2 \le \lambda_2$, where we used again that $\mathbf{u}_1$ and $\mathbf{u}_2$ are unitary orthogonal eigenvectors, followed by $\lambda_1 \le \lambda_2$. Let $\mathbf{u}'$ be $\mathbf{u}$ without component i and we obtain $\mathbf{u}\bullet A\mathbf{u} = A\bullet\mathbf{u}\mathbf{u}^\top = A'\bullet\mathbf{u}'\mathbf{u}'^\top = \mathbf{u}'\bullet A'\mathbf{u}'$, where we used the fact that row and column i of $\mathbf{u}\mathbf{u}^\top$ are zero based on $u_i = 0$. We obtained that the unitary $\mathbf{u}'$ yields $\mathbf{u}'\bullet A'\mathbf{u}' \le \lambda_2$. Using Lemma 1.2.2.1, the smallest eigenvalue of A' is less than or equal to $\mathbf{u}'\bullet A'\mathbf{u}' \le \lambda_2$, *i.e.*, $\lambda'_1 \le \lambda_2$. $\square$

The following lemma could be generally useful, although it is not used for other proofs in this document.

Lemma 1.5.2.2. *Consider symmetric matrix A with eigenvalues $\lambda_1 \le \lambda_2 \le \dots \le \lambda_n$. Any principal minor A' of order $n-1$ with eigenvalues $\lambda'_1 \le \lambda'_2 \le \dots \le \lambda'_{n-1}$ satisfies:*[5]

$$\lambda_1 \le \lambda'_1 \le \lambda_2 \tag{1.5.1a}$$

$$\lambda_{n-1} \le \lambda'_{n-1} \le \lambda_n \tag{1.5.1b}$$

Proof. We have already proved in Lemma 1.5.2.1 that $\lambda'_1 \le \lambda_2$. To show that $\lambda_1 \le \lambda'_1$, let us consider the unitary eigenvector $\mathbf{v}'$ of λ'_1 in A' and notice $\mathbf{v}'\bullet A'\mathbf{v}' = \lambda'_1$. If A' is obtained from A by removing row i and column i, we can say $\mathbf{v}'$ is obtained by removing position i from an $\mathbf{v}\in\mathbb{R}^n$ such that $v_i = 0$. We can write $\lambda'_1 = \mathbf{v}'\bullet A'\mathbf{v}' = \mathbf{v}\bullet A\mathbf{v}$. We recall Lemma 1.2.2.1 that states that λ_1 is the minimizer of the Rayleigh ratio, and so, we obtain: $\lambda_1 = \min\limits_{|\mathbf{x}|=1}\mathbf{x}\bullet A\mathbf{x} \le \mathbf{v}\bullet A\mathbf{v} = \lambda'_1$.

To prove (1.5.1b) it is enough to apply (1.5.1a) on matrices $-A$ and $-A'$. $\square$

1.6 Cholesky decomposition of *semidefinite* positive matrices

We provide two proofs. The first one is much shorter, but the factorization arises somehow out of the blue. The second one is longer and a bit more general, discussing along the way several properties that are generally useful (Prop. 1.6.3 and 1.6.4), leading the reader to a deeper insight into the factorization.

---

[5]It is possible to prove that

$$\lambda_1 \le \lambda'_1 \le \lambda_2 \le \lambda'_2 \le \dots \lambda'_{n-2} \le \lambda_{n-1} \le \lambda'_{n-1} \le \lambda_n$$

The proof lies outside the scope of this document, but it could use the following argument by contradiction. Assume $\lambda'_1,\ \lambda'_2, \dots \lambda'_k < \lambda_k$ and we can derive a contradiction. All vectors of the subspace S' generated by the first k eigenvectors of A' have a Rayleigh ratio $\le \lambda'_k$. All vectors of the subspace S generated by the last $n-k+1$ eigenvectors of A have a Rayleigh ratio $\ge \lambda_k$. But the subspaces S and S' need to have an intersection of at least dimension 1, because $dim(S) + dim(S') = n-k+1+k = n+1$. We obtained a contradiction: the Rayleigh ratio over this intersection needs to be both $\ge \lambda_k$ and $\le \lambda'_k$. An analogous reversed argument could be used to show that $\lambda'_k \le \lambda_{k+1}$. Different proofs can be found by searching key words "Interlacing eigenvalues" on the internet, the one from the David Williamson's course (`http://people.orie.cornell.edu/dpw/orie6334/lecture4.pdf`) is the most related to the above ideas. Another proof can be found at least in [Ikramov H., Recueil de problèmes d'algébre lineaire, publisher MIR - Moscou, 1977], exercise 7.4.35.

1.6.1 A short proof using the square root and the QR decompositions

1.6.1.1 Square root factorisations of SDP matrices

Given matrix $A \succeq \mathbf{0}$, we are looking for matrices K such that $KK = A$. We apply the eigendecomposition (1.1.1) and write $A = U\Lambda U^\top$, where $\Lambda = \mathtt{diag}(\lambda_1, \lambda_2, \ldots \lambda_n)$ contains the non-negative eigenvalues of A. Take $D = \mathtt{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots \sqrt{\lambda_n}\right)$ and consider $K = UDU^\top$. One can easily check that this K is a square root of A: $KK = UDU^\top UDU^\top = UDDU^\top = U\Lambda U^\top = A$, where we used $U^\top U = I_n$ from (1.1.2). This K is called the principal square root of A and it is SDP, because it is similar and congruent to $\mathtt{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots \sqrt{\lambda_n}\right)$. It is the only SDP square root of A (Appendix B.4).

Remark 1.6.1. *There are* `multiple symmetric` *matrices $K \in \mathbb{R}^{n\times n}$ such that $A = KK$ that can be found by taking $K = UDU^\top$ for any* `symmetric` *D such that $DD = \Lambda$. Examples of such symmetric matrices D can be found by taking $D = \mathtt{diag}\left(\pm\sqrt{\lambda_1}, \pm\sqrt{\lambda_2}, \ldots \pm\sqrt{\lambda_n}\right)$.*

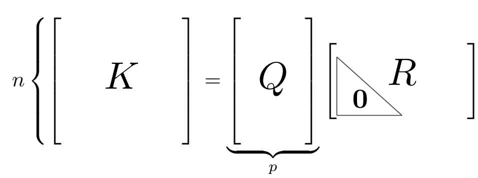
Non-symmetric matrices K such that $A = KK$ do exist; in fact, they can be infinitely many, because there can be infinitely many matrices $D \in \mathbb{R}^{n\times n}$ such that $DD = \mathtt{diag}(\lambda_1, \lambda_2, \ldots \lambda_n)$. To see this, notice that $I_2 = \begin{bmatrix} \frac{s}{t} & \frac{r}{t} \\ \frac{r}{t} & -\frac{s}{t} \end{bmatrix} \begin{bmatrix} \frac{s}{t} & \frac{r}{t} \\ \frac{r}{t} & -\frac{s}{t} \end{bmatrix}$ for infinitely many Pythagorean triples (r, s, t) such that $r^2 + s^2 = t^2$. If we have $\lambda_1 = \lambda_2$, the above 2×2 construction can be extended to a $n \times n$ matrix in which the leading principal minor of size 2×2 is given by this 2×2 construction.

As a side remark, we can prove that $rank(A) = rank(K)$ if K is symmetric by applying Prop. A.1.8 on $A = K^\top K$. This is no longer true with non-symmetric matrices, *e.g.*, $\mathbf{0} = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}$ has rank zero while the square root factor has rank 1.

1.6.1.2 The Cholesky factorization proof

Proposition 1.6.2. *(Cholesky factorization of positive semidefinite matrices) Any real SDP matrix A can be factorized as $A = LL^\top$ where L is a lower triangular matrix with non-negative diagonal elements. This proposition is slightly weaker than Prop 1.6.7 where we will show (using a longer proof) that there always exists a factorization in which the value of L_{nn} depends on the ranks of A and of the leading principal minor of size $n - 1$ of A.*

Proof.
We know there exists a symmetric $K \in \mathbb{R}^{n\times n}$ such that $A = KK$ using Remark 1.6.1 above. We now apply the QR decomposition on K and write $K = QR$ where $Q \in \mathbb{R}^{n\times p}$ satisfies $Q^\top Q = I_p$ and $R \in \mathbb{R}^{p\times n}$ is upper triangular for some $p \leq n$—see the proof in Prop. B.3.1. We can develop:



$$A = KK = K^\top K = (QR)^\top QR = R^\top Q^\top QR = R^\top I_p R = R^\top R,$$

which is very close to a Cholesky decomposition, because $R^\top$ is lower triangular. However, $R^\top$ has only $p \leq n$ columns, but we can extended to a $n \times n$ matrix by adding $n - p$ zero columns. This way, $R^\top$ is transformed into a lower triangular *square* matrix L such that $L\,L^\top = R^\top R = A$. The last point to address is the fact that the QR decomposition from Prop. B.3.1 does not state that the diagonal elements of R are non-negative. There might exist multiple $i \in [1..p]$ such that $L_{ii} = R_{ii} < 0$. We can overcome this with a simple trick. The product $LL^\top$ does not change if we negate all columns i of L satisfying $L_{ii} < 0$, because $A_{jj'}$ is the dot product of rows j and j' of L which does not change if both rows j and j' negate some column(s) i. This leads to a factorisation in which the factors have a non-negative diagonal. [6] ☐

1.6.2 A longer Cholesky proof providing more insight into the properties of SDP matrices

Proposition 1.6.3. *If SDP matrix $A \in \mathbb{R}^{n\times n}$ has some null principal minor (i.e., $\exists J \subset [1..n]$ such that $det([A]_J) = 0$), then $A \not\succ \mathbf{0}$ and $det(A) = 0$.*

[6] I first found this proof in an answer of user *loup blanc* on the on-line forum `https://math.stackexchange.com/questions/1331451/how-to-prove-cholesky-decomposition-for-positive-semidefinite-matrices`. It also appear in Corollary 7.2.9 of the book “Matrix Analysis” by Roger Horn and Charles Johnson, second edition, Cambridge University Press, 2013.

Proof 1. One of the eigenvalues of $[A]_J$ need to be zero so that $[A]_J \not\succ \mathbf{0}$ and $[A]_J \bullet \mathbf{x}_J\mathbf{x}_J^\top = 0$ for some vector $\mathbf{x}_J$ with $|J|$ components. Construct $\mathbf{x} \in \mathbb{R}^n$ by keeping the values of $\mathbf{x}_J$ on the positions J of $\mathbf{x}$ and by filling the rest with zeros. It is not hard to check that $A \bullet \mathbf{x}\mathbf{x}^\top = [A]_J \bullet \mathbf{x}_J\mathbf{x}_J^\top = 0$, so that then $A \not\succ \mathbf{0}$ and $det(A)$ is zero as the product of the eigenvalues. □

Proof 2. Re-order the rows and columns of A so that the minor $[A]_J$ becomes a leading principal minor; the Sylvester criterion is thus violated, which shows $A \not\succ \mathbf{0}$ □.

Proposition 1.6.4. *If we are given an SDP matrix $A \in \mathbb{R}^{n\times n}$ written under the form*

$$A = \begin{bmatrix} & & & b_1 \\ & [A]_{n-1} & & b_2 \\ & & & \vdots \\ b_1 & b_2 & \dots & b_n \end{bmatrix},$$

then $[b_1\ b_2\ \dots b_{n-1}]$ can be written as a linear combination of the rows of $[A]_{n-1}$. This combination uses only rows J where $J \subset [1..n-1]$ such that $[A]_J$ (matrix obtained by selecting rows J and columns J) is a non-null principal minor of maximum order (the rank of $[A]_{n-1}$).

We provide two proofs. The first one is much shorter, but the second one provides more insight into the arrangement of the matrices.

Proof 1. Assume $[b_1\ b_2\ \dots b_{n-1}]$ does not belong to the row image (set of linear combinations of the rows) of $[A]_{n-1}$. Using the rank-nullity Theorem A.1.3, we have $rank([A]_{n-1}) + nullity([A]_{n-1}) = n-1$. The dimension of the image $\mathtt{img}([A]_{n-1})$ plus the dimension of $\mathtt{null}([A]_{n-1})$ is equal to $n-1$. Since the two spaces are perpendicular (any $\mathbf{x}_0 \in \mathtt{null}([A]_{n-1})$ satisfies $[A]_{n-1}^i\mathbf{x}_0 = 0$ for any row i of $[A]_{n-1}$), the sums between elements of $\mathtt{null}([A]_{n-1})$ and elements of $\mathtt{img}([A]_{n-1})$ cover the whole (transposed) $\mathbb{R}^{n-1}$. As such, we can write $[b_1\ b_2\ \dots b_{n-1}] = \mathbf{b}_{\mathtt{img}}^\top + \mathbf{b}_0^\top$, where $\mathbf{b}_{\mathtt{img}}^\top \in \mathtt{img}([A]_{n-1})$ and $\mathbf{b}_0 \in \mathtt{null}([A]_{n-1})$ with $\mathbf{b}_0 \neq \mathbf{0}$. We can now calculate:

$$\begin{aligned} [-t\mathbf{b}_0^\top\ 1]A \begin{bmatrix} -t\mathbf{b}_0 \\ 1 \end{bmatrix} &= t^2\mathbf{b}_0^\top[A]_{n-1}\mathbf{b}_0 - 2t(\mathbf{b}_{\mathtt{img}}^\top + \mathbf{b}_0^\top)\mathbf{b}_0 + b_n \\ &= 0 - 2t\mathbf{b}_0^\top\mathbf{b}_0 + b_n = b_n - 2t|\mathbf{b}_0|^2 \to -\infty \end{aligned}$$

We obtained a contradiction, the assumption $[b_1\ b_2\ \dots b_{n-1}] \notin \mathtt{img}([A]_{n-1})$ was false. □

Proof 2. If $[A]_{n-1}$ is non-singular, the conclusion is obvious: $\mathbf{x}^\top[A]_{n-1} = [b_1\ b_2\ \dots b_{n-1}]$ has solution $\mathbf{x}^\top = [b_1\ b_2\ \dots b_{n-1}][A]_{n-1}^{-1}$.

We hereafter consider $[A]_{n-1}$ has rank $r < n-1$. Based on Prop. A.1.2, A has at least a non-zero *principal* minor of order r. Without loss of generality, we permute the rows and columns of A until this non-zero minor is positioned in the upper-left corner – this does not change the determinant or the SDP status (Prop 1.2.4). Let $[A]_r$ be this leading principal minor. Consider the solution $\mathbf{x}$ of the system $\mathbf{x}^\top[A]_r = \mathbf{b}_r^\top$, where $\mathbf{b}_r$ is $\mathbf{b}$ reduced to positions $[1..r]$. This solution exists and it has value $\mathbf{x}^\top = \mathbf{b}_r^\top[A]_r^{-1}$.

We will show that $[b_1\ b_2\ \dots b_{n-1}]$ can be written as a linear combination of the rows $[1..r]$ of $[A]_{n-1}$, the coefficients of this combination being $\mathbf{x}^\top$. Take any $i \in [r+1..n-1]$. Let us consider the minor obtained by selecting rows and columns $[1..r] \cup \{i, n\}$ of A, see left matrix below. We subtract from the last row the linear combination of the first r rows defined by $\mathbf{x}$, so as to cancel the first r positions of the last row, followed by the transposed operation on columns. We obtain the right matrix below.

$$\begin{bmatrix} & & & & a_{i,1} & b_1 \\ & & & & a_{i,2} & b_2 \\ & [A]_r & & & \vdots & \vdots \\ & & & & a_{i,r} & b_r \\ a_{i,1} & a_{i,2} & \dots & a_{i,r} & a_{i,i} & b_i \\ b_1 & b_2 & \dots & b_r & b_i & b_n \end{bmatrix} \xrightarrow[\text{that do not change the determinant}]{\text{row and column operations}} \begin{bmatrix} & & & & a_{i,1} & 0 \\ & & & & a_{i,2} & 0 \\ & [A]_r & & & \vdots & \vdots \\ & & & & a_{i,r} & 0 \\ a_{i,} & a_{i,2} & \dots & a_{i,r} & a_{i,i} & \widehat{b_i} \\ 0 & 0 & \dots & 0 & \widehat{b_i} & \widehat{b_n} \end{bmatrix},$$

where $\widehat{b_i} = b_i - \sum_{j\in[1..r]} x_j a_{ij}$. The determinant of this right matrix can be calculated (see the Leibniz formula or the Laplace formula for determinants) as follows:

$$\widehat{b_n} \cdot \det\left([A]_{[1..r]\cup\{i\}}\right) - \widehat{b_i}^2 \cdot \det\left([A]_r\right),$$

where $[A]_{[1..r]\cup\{i\}}$ is the $(r+1)\times(r+1)$ upper left minor of above matrix. This $(r+1)\times(r+1)$ minor needs to be null because $[A]_{n-1}$ has rang r and $[1..r]\cup\{i\}\subseteq[1..n-1]$. On the other hand we have $det([A]_r) > 0$, so that the above determinant simplifies to $-\widehat{b_i}^2 \cdot \det([A]_r)$. But this is the determinant of a matrix obtained from a minor of A after performing linear operations with rows and columns of A. Since A is SDP, this determinant needs to be non-negative, and so, we need to have $\widehat{b_i} = 0$, *i.e.*, $b_i - \sum_{j\in[1..r]} x_j a_{ij} = 0$, meaning that b_i is a linear combination (defined by $\mathbf{x}$) of the first r rows of column i (recall $a_{ij} = a_{ji}$). Recall i was chosen at random from $[r+1..n-1]$, so that actually all elements b_i of $[b_1\ b_2\ \dots b_{n-1}]$ can be written as a linear combination (defined by $\mathbf{x}$) of the first r rows. □

Corollary 1.6.5. *If $A \succeq \mathbf{0}$ and $A_{jj} = 0$ for some $j \in [1..n]$, then the row and column j contain only zeros.*

Corollary 1.6.6. *If $A \succeq \mathbf{0}$ and $a_{11} = a_{12} = a_{21} = a_{22}$, then $a_{i1} = a_{i2}\ \forall i \in [1..n]$.*

Proposition 1.6.7. *(Cholesky factorization of positive semidefinite matrices) A real symmetric matrix A is positive semidefinite if and only if it can be factorized as:*

$$A = RR^\top = \begin{bmatrix} r_{11} & 0 & 0 & \dots & 0 \\ r_{21} & r_{22} & 0 & \dots & 0 \\ r_{31} & r_{32} & r_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \dots & r_{nn} \end{bmatrix} \begin{bmatrix} r_{11} & r_{21} & r_{31} & \dots & r_{n1} \\ 0 & r_{22} & r_{32} & \dots & r_{n2} \\ 0 & 0 & r_{33} & \dots & r_{n3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & r_{nn} \end{bmatrix}, \tag{1.6.1}$$

where the diagonal terms are non-negative.

The factorization is not always unique. There exists a factorization that has $r_{nn} = 0$ only if $rank(A) = rank([A]_{n-1})$, where $[A]_{n-1}$ is the leading principal minor of size $(n-1)\times(n-1)$.

Proof.
$\Longleftarrow$
Take any non-zero $\mathbf{x} \in \mathbb{R}^n - \{\mathbf{0}_n\}$ and observe that $\mathbf{x}^\top A\mathbf{x} = \mathbf{x}^\top RR^\top\mathbf{x} = (R^\top\mathbf{x})^\top(R^\top\mathbf{x}) \geq 0$. This is enough to prove that A is SDP.
$\Longrightarrow$
We proceed by induction. The implication is obviously true for $n = 1$. We suppose that there exists a factorization:

$$[A]_{n-1} = [R]_{n-1}[R]_{n-1}^\top = \begin{bmatrix} r_{11} & 0 & \dots & 0 \\ r_{21} & r_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ r_{n-1,1} & r_{n-1,2} & \dots & r_{n-1,n-1} \end{bmatrix} \begin{bmatrix} r_{11} & r_{21} & \dots & r_{n-1,1} \\ 0 & r_{22} & \dots & r_{n-1,2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & r_{n-1,n-1} \end{bmatrix},$$

where $r_{11},\ r_{22},\ \dots r_{n-1,n-1} \geq 0$. We will prove that this decomposition can be extended to a $n\times n$ decomposition for matrix A. The values $r_{1n},\ r_{2n}, \dots r_{n-1,n}$ (the elements above the diagonal on the last column of R) are set to zero by definition.

The main difficulty is to determine the last row $\mathbf{r}^\top = [\,\mathbf{r}_{n-1}^\top\ r_n\,]$ of R. We write

$$A = \begin{bmatrix} [A]_{n-1} & \mathbf{b}_{n-1} \\ \mathbf{b}_{n-1}^\top & b_n \end{bmatrix}$$

Using Prop. 1.6.4, there exists $\mathbf{x} \in \mathbb{R}^{n-1}$ such that $\mathbf{b}_{n-1}^\top = \mathbf{x}^\top[A]_{n-1}$. This leads to

$$\begin{bmatrix} [A]_{n-1} \\ \mathbf{b}_{n-1}^\top \end{bmatrix} = \begin{bmatrix} I_{n-1} \\ \mathbf{x}^\top \end{bmatrix}[A]_{n-1} = \begin{bmatrix} I_{n-1} \\ \mathbf{x}^\top \end{bmatrix}[R]_{n-1}[R]_{n-1}^\top = \begin{bmatrix} [R]_{n-1} \\ \mathbf{x}^\top[R]_{n-1} \end{bmatrix}[R]_{n-1}^\top.$$

Using the notational shorthand $\mathbf{r}_{n-1}^\top = \mathbf{x}^\top [R]_{n-1}$, we can write:

$$\begin{bmatrix} [A]_{n-1} \\ \mathbf{b}_{n-1}^\top \end{bmatrix} = \begin{bmatrix} [R]_{n-1} \\ \mathbf{r}_{n-1}^\top \end{bmatrix} [R]_{n-1}^\top.$$

In fact, we can already fix the first $n-1$ positions of the last row (of R) to $\mathbf{r}^\top = \left[\begin{smallmatrix} \mathbf{r}_{n-1}^\top & r_n \end{smallmatrix} \right]$. Then, let us extend the above equality to work with $n \times n$ matrices, by adding: (i) a column with zeros to the left factor of the product and (ii) a row with zeros and a column $\mathbf{r}_{n-1}$ to the right factor. We obtain

$$\begin{bmatrix} [R]_{n-1} & \mathbf{0} \\ \mathbf{r}_{n-1}^\top & 0 \end{bmatrix} \begin{bmatrix} [R]_{n-1} & \mathbf{0} \\ \mathbf{r}_{n-1}^\top & 0 \end{bmatrix}^\top = \begin{bmatrix} [A]_{n-1} & \mathbf{b}_{n-1} \\ \mathbf{b}_{n-1}^\top & z \end{bmatrix} = A_z,$$

where $z = \mathbf{r}_{n-1}^\top \mathbf{r}_{n-1}$. The matrix at right has the same rank as the matrices at left (use Prop. A.1.8), that is precisely the rank of $[R]_{n-1}$ which is equal to the rank of $[A]_{n-1}$ (by applying again Prop. A.1.8 on $[A]_{n-1} = [R]_{n-1} R_{n-1}^\top$). In short, A_z has the same rank as $[A]_{n-1}$.

Let us now study the minor $\left[\begin{smallmatrix} \overline{A} & \overline{\mathbf{b}} \\ \overline{\mathbf{b}}^\top & z \end{smallmatrix} \right]$ of A_z obtained by bordering the largest order non-zero minor $\overline{A}$ of $[A]_{n-1}$ with the last column and row of A_z (keeping only the positions that constitute the minor). The determinant of this new minor of A_z has to be zero because A_z has the same rank as $[A]_{n-1}$. We now compare A to A_z; let us write the bottom-right term b_n of A in the form $b_n = z + \alpha$. By replacing z with $z + \alpha$, the above minor evolves to $\left[\begin{smallmatrix} \overline{A} & \overline{\mathbf{b}} \\ \overline{\mathbf{b}}^\top & z+\alpha \end{smallmatrix} \right]$ which is a matrix of determinant $\alpha \det\left(\overline{A}\right)$. Since A is SDP and $\overline{A}$ is a non-zero minor of A, this determinant has to be non-negative, and so, we have $\alpha \geq 0$. This enables us to finish the construction by setting

$$\begin{bmatrix} [R]_{n-1} & \mathbf{0} \\ \mathbf{r}_{n-1}^\top & \sqrt{\alpha} \end{bmatrix} \begin{bmatrix} [R]_{n-1} & \mathbf{0} \\ \mathbf{r}_{n-1}^\top & \sqrt{\alpha} \end{bmatrix}^\top = \begin{bmatrix} [A]_{n-1} & \mathbf{b}_{n-1} \\ \mathbf{b}_{n-1}^\top & z+\alpha \end{bmatrix} = A,$$

Finally, if $\texttt{rank}(A) = \texttt{rank}([A]_{n-1})$ then we can say $\texttt{rank}(A) = \texttt{rank}([A]_{n-1}) = \texttt{rank}([R]_{n-1}) = \texttt{rank}(R)$, and so, we need to have $r_{nn} = 0$ because any $r_{nn} > 0$ would make $\alpha = r_{nn}^2 > 0$, leading to $\texttt{rank}(R) > \texttt{rank}([R]_{n-1})$. On the other hand, if $\texttt{rank}(A) > \texttt{rank}([A]_{n-1})$, then we have $r_{nn} = \sqrt{\alpha} > 0$. This justifies the last statement of the proposition. $\square$

Corollary 1.6.8. *If a SDP matrix A is not positive definite, the Cholesky factorization may or may no be unique.*

Proof. The following factorization is unique:

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$$

because it requires $1 = r_1^2$ and $1 = r_1 r_2 + r_2^2$, imposing $r_1 = 1$ and $r_2 = 0$.

The following factorization is not unique:

$$\begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 2 \\ 2 & 2 & 13 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 2 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

This is the decomposition that the proof of the above theorem would construct by taking $\mathbf{x}^\top = \left[\begin{smallmatrix} 2 & 0 \end{smallmatrix} \right]$ on the induction basis of the previous 2×2 decomposition of $[A]_{n-1} = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix} \right]$. Precisely, this proof calculates $\left[\begin{smallmatrix} r_{31} & r_{32} \end{smallmatrix} \right] = \mathbf{x}^\top [R]_{n-1} = \left[\begin{smallmatrix} 2 & 0 \end{smallmatrix} \right] \left[\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix} \right] = \left[\begin{smallmatrix} 2 & 0 \end{smallmatrix} \right]$. However, more generally, the coefficients r_{32} and r_{33} are active only in the equation of a_{33}, *i.e.*, $13 = 2^2 + r_{32}^2 + r_{33}^2$. We could take $r_{32} = 1$ and $r_{33} = \sqrt{8}$. Also, we could take $r_{32} = 3$ and $r_{33} = 0$, leading to a different factorization:

$$\begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 2 \\ 2 & 2 & 13 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 2 & 3 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix}$$

Unlike the previous factorization constructed by the proof of the above theorem, this decomposition has a null bottom-right diagonal term, although $rank(A) > rank([A]_2)$. However, notice that $rank(R) = rank([R]_{n-1}) + 1$, where $[R]_{n-1}$ is the leading principal minor of size 2×2. Since $rank(RR^\top) = rank(R)$ (use Prop. A.1.8), the rank of A is equal to the rank of R and the rank of $[A]_{n-1}$ equals the rank of $[R]_{n-1}$. ☐

1.7 Any $A \succeq 0$ has infinitely many factorizations $A = VV^\top$ related by rotations and reflections

Corollary 1.7.1. *Any SDP matrix $A \in \mathbb{R}^{n\times n}$ can be factorized in a form $A = VV^\top$. Generating such V is simply equivalent to taking n vectors $\mathbf{v}_1^\top$, $\mathbf{v}_2^\top$, $\dots$ $\mathbf{v}_n^\top$ (the rows of V) such that $A_{ij} = \mathbf{v}_i^\top \mathbf{v}_j = \mathbf{v}_i \bullet \mathbf{v}_j$ $\forall i, j \in [1..n]$. There are infinitely many such matrices V (or vectors $\mathbf{v}_1$, $\mathbf{v}_2$, $\dots$ $\mathbf{v}_n$) for a fixed A. We can say that any SDP matrix A can be constructed by choosing a set of n vectors of $\mathbb{R}^n$.*

Proof. For a given $A \succeq 0$, we have actually already presented three ways of computing a factor V such that $A = VV^\top$. We recall these three ways at points (1)-(3) below. At point (4), we show that from any such factor V we can further generate infinitely many other factors.

(1) Use the above Cholesky decomposition of SDP matrices to write $A = RR^\top$ and take $V = R$ as needed.

(2) Use the eigendecomposition (1.1.1) to write $A = U\Lambda U^\top$, where $\Lambda = \texttt{diag}(\lambda_1, \lambda_2, \dots \lambda_n)$. Since $\lambda_i \geq 0$ $\forall i \in [1..n]$, we can define real matrix $\sqrt{\Lambda} = \texttt{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots \sqrt{\lambda_n})$ and write $A = U\sqrt{\Lambda}\sqrt{\Lambda}U^\top = (U\sqrt{\Lambda})(U\sqrt{\Lambda})^\top = VV^\top$ with $V = U\sqrt{\Lambda}$.

(3) Use one of the multiple square root decompositions $A = KK$ with symmetric K from Remark 1.6.1. This gives $V = K$ and $V^\top = K$.

(4) We can generate infinitely many more decomposition from any V determined as above. For this, let us consider any unitary orthogonal matrix $\mathcal{R}$, *i.e.*, a matrix such that $\mathcal{R}^\top\mathcal{R} = I_n$. There are infinitely-many such matrices $\mathcal{R}$, each one of them representing a composition of rotation and reflection operators.[7] Now check that $V_{\text{new}}V_{\text{new}}^\top = (V\mathcal{R})(V\mathcal{R})^\top = V\mathcal{R}\mathcal{R}^\top V^\top = VV^\top = A$.

☐

As a side remark, the factorizations mentioned at above points (3) and (4) are related. In fact, Remark 1.6.1 used at (3) constructs the factorisation by applying a particular case of the technique used at (4) on $A = VV^\top = U\sqrt{\Lambda}\sqrt{\Lambda}U^\top$. More exactly, recall that symmetric K from Remark 1.6.1 was generated by setting $K = UDU^\top$ for any `symmetric` D such that $DD = \Lambda$. Such a D could be found by taking $D = \sqrt{\Lambda}\mathcal{R}$, where $\mathcal{R} = \texttt{diag}(e_1,\ e_2, \dots, e_n)$ with $e_i = \pm 1$ $\forall i \in [1..n]$; this $\mathcal{R}$ can actually be seen as a composition of reflexion operators like matrix $\mathcal{R}$ at point (4) above. We can write $A = VV^\top = U\sqrt{\Lambda}\sqrt{\Lambda}U^\top = U\sqrt{\Lambda}\mathcal{R}\mathcal{R}^\top\sqrt{\Lambda}U^\top = UDD^\top U^\top = UDDU^\top = UDU^\top UDU^\top = KK$.

Remark 1.7.2. *Given two factorizations $A = VV^\top = UU^\top$, it is always possible to write $V = U\mathcal{R}$, where $\mathcal{R}$ satisfies $\mathcal{R}\mathcal{R}^\top = I_n$, i.e., $\mathcal{R}$ represents a composition of rotation and reflection operators. The rows of V can thus be obtained from the rows of U by applying a composition of rotations and reflections.*

Proof. Let r be the rank of V, U and A (recall Prop A.1.8) and J a set of rows such that $\lfloor V \rfloor_J$ has rank $|J| = r$, where the operator $\lfloor\cdot\rfloor_J$ represents the given matrix reduced to rows J. Any row $\mathbf{v}_i$ of V outside J (*i.e.*, such that $i \in [1..n] \setminus J$) can be written as linear combination of rows J using coefficients $\mathbf{x} \in \mathbb{R}^r$: $\mathbf{v}_i = \mathbf{x}^\top \lfloor V \rfloor_J$

Let's examine row i of the product $A = VV^\top = UU^\top$ for any fixed $i \in [1..n] \setminus J$. Replacing above $\mathbf{v}_i = \mathbf{x}^\top \lfloor V \rfloor_J$ in $\mathbf{a}_i = \mathbf{v}_i V^\top$, we obtain $\mathbf{a}_i = \mathbf{x}^\top \lfloor V \rfloor_J V^\top$. But now notice that $\lfloor V \rfloor_J V^\top$ actually represents the rows J of matrix A; we can thus replace $\lfloor V \rfloor_J V^\top = \lfloor A \rfloor_J$ and obtain $\mathbf{a}_i = \mathbf{x}^\top \lfloor A \rfloor_J$.

We will now prove that $\mathbf{u}_i = \mathbf{x}^\top \lfloor U \rfloor_J$; we introduce notational shortcuts $\mathbf{u}_i = \overline{\mathbf{u}_i} + \mathbf{z} = \mathbf{x}^\top \lfloor U \rfloor_J + \mathbf{z}$ and we will show $\mathbf{z} = \mathbf{0}$. Let us first calculate $\overline{\mathbf{u}_i} U^\top = \mathbf{x}^\top \lfloor U \rfloor_J U^\top = \mathbf{x} \lfloor A \rfloor_J = \mathbf{a}_i$. Using $A = UU^\top$, we also have $\mathbf{u}_i U^\top = \mathbf{a}_i$ which means that $(\mathbf{u}_i - \overline{\mathbf{u}_i})\, U^\top = \mathbf{z}U^\top = \mathbf{0}^\top$. Taking row i, we obtain $\mathbf{z}\mathbf{u}_i = 0$, or $\mathbf{z}\left(\overline{\mathbf{u}_i} + \mathbf{z}\right)^\top = 0$. We now simply replace $\overline{\mathbf{u}_i} = \mathbf{x}^\top \lfloor U \rfloor_J$ and obtain $\mathbf{z}\left(\mathbf{x}^\top \lfloor U \rfloor_J + \mathbf{z}\right)^\top = 0 \implies \mathbf{z}\lfloor U \rfloor_J^\top \mathbf{x} + \mathbf{z}\mathbf{z}^\top = 0$. But $\mathbf{z}\lfloor U \rfloor_J^\top$

[7] Using $\mathcal{R}\,\mathcal{R}^\top = I_n$, notice $\mathbf{x}^\top\mathbf{y} = \mathbf{x}^\top\mathcal{R}\,\mathcal{R}^\top\mathbf{y} = (\mathbf{x}^\top\mathcal{R})(\mathbf{y}^\top\mathcal{R})^\top$, *i.e.*, the operator that maps $\mathbf{x}^\top \to \mathbf{x}^\top\mathcal{R}$ preserves the angles, and so, it needs to be a composition of rotations and reflections.

represents the columns J of $\mathbf{z}U^\top = \mathbf{0}^\top$, *i.e.*, $\mathbf{z}\lfloor U\rfloor_J^\top = \mathbf{0}^\top$ is a row vector (of length $|J| = r$) that only contains zeros. This leads to $\mathbf{z}\mathbf{z}^\top = 0$ and this shows that $\mathbf{z} = 0$, leading to $\mathbf{u}_i = \mathbf{x}^\top \lfloor U\rfloor_J$. We actually obtained that all rows $[1..n] - J$ of the equality $A = VV^\top = UU^\top$ represent merely linear combinations of the rows J of this equality.

This equality is thus a (linear combination) consequence of $\lfloor A\rfloor_J = \lfloor V\rfloor_J V^\top = \lfloor U\rfloor_J U^\top$, where $\lfloor V\rfloor_J$ and $\lfloor U\rfloor_J$ have full rank $|J| = r$. We now replace all rows $[1..n] - J$ of V with $n - r$ unit orthogonal vectors rows that are also perpendicular to $\lfloor V\rfloor_J^\top$ and obtain matrix $\overline{V}$; this is possible because these unit orthogonal vectors are simply a basis for the null space of $\lfloor V\rfloor_J^\top$ which has dimension $n-r$ by virtue of the rank-nullity theorem. One can check that the product $\overline{V}\,\overline{V}^\top$ is a matrix $\overline{A}$ such that $\overline{a}_{j,j'} = a_{j,j'}$ $\forall j, j' \in J$, $\overline{a}_{i,i} = 1$ $\forall i \in [1..n] - J$ and $\overline{a}_{i,j} = \overline{a}_{j,i} = 0$ $\forall i \in [1..n] - J, j \in J$. We perform a similar operation on U and obtain matrix $\overline{U}$ (filling rows $[1..n] - J$ with a different basis than for V) and one can check that similarly we have

$$\overline{U}\,\overline{U}^\top = \overline{V}\,\overline{V}^\top = \overline{A}$$

We can now use that that $\overline{U}, \overline{V}$ and $\overline{A}$ are non-singular and invertible. This ensures that there exists $\mathcal{R}$ such that $\overline{V} = \overline{U}\mathcal{R}$. We obtain $\overline{U}\,\overline{U}^\top = \overline{U}\mathcal{R}\mathcal{R}^\top\overline{U}^\top$. Multiplying at left with $\overline{U}^{-1}$ and at right with $\overline{U}^{\top^{-1}}$, we obtain that $\mathcal{R}\mathcal{R}^\top = I_n$. Now recall that $\overline{V} = \overline{U}\mathcal{R}$ contains rows J inherited from V and resp. U, so that we also have $\lfloor V\rfloor_J = \lfloor U\rfloor_J\mathcal{R}$. This can easily be extended to $V = U\mathcal{R}$ because each missing row (*i.e.*, each $i \in [1..n] - J$) is a linear combination of rows J (*i.e.*, $\mathbf{v}_i = \mathbf{x}\lfloor V\rfloor_J = \mathbf{x}\lfloor U\rfloor_J\mathcal{R} = \mathbf{u}_i\mathcal{R}$). ☐

1.8 Convex functions have an SDP Hessian assuming the Hessian is symmetric

Notice proposition below requires the Hessian matrix to be symmetric. This condition was omitted from certain texts (see Section 3.1.4 of the book “Convex Optimization” by Stephen Boyd and Lieven Vandenberghe, p. 71, Cambridge University Press, 2004) but we address it in our work. Convex functions with asymmetric non-SDP Hessians do exist, see Example C.3.1 in Appendix C. For such cases, the convexity condition should actually evolve from $\nabla^2 f(\mathbf{y}) \succeq \mathbf{0}$ $\forall \mathbf{y} \in \mathbb{R}^n$ (*i.e.*, SDP Hessian) to $\nabla^2 f(\mathbf{y}) + \nabla^2 f(\mathbf{y})^\top \succeq \mathbf{0}$ $\forall \mathbf{y} \in \mathbb{R}^n$.

Proposition 1.8.1. *A twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ with a symmetric Hessian matrix $\nabla^2 f(\mathbf{y}) = H_\mathbf{y}$ defined by terms $h_{ji}^\mathbf{y} = \dfrac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{y})$ for all $\mathbf{y} \in \mathbb{R}^n$ and $i, j \in [1..n]$ is convex if and only if $H_\mathbf{y} \succeq \mathbf{0}$ $\forall \mathbf{y} \in \mathbb{R}^n$, i.e., if and only if the Hessian is SDP in all points.*

Proof. Let us consider any $\mathbf{y} \in \mathbb{R}^n$. We take any direction $\mathbf{v} \in \mathbb{R}^n$ and define $g : \mathbb{R} \to \mathbb{R}$ via

$$g(t) = f(\mathbf{y} + t\mathbf{v})$$

Using the chain rule to the gradient,[8] we obtain

$$g'(t) = \underbrace{\nabla f\,(\mathbf{y} + t\mathbf{v})}_{\text{row vector}}\,\mathbf{v} = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(\mathbf{y} + t\mathbf{v})v_i \tag{1.8.1}$$

We derivate again in t to obtain $g''(t)$. For this, we observe that the derivative in t of any term $\frac{\partial f}{\partial x_i}(\mathbf{y}+t\mathbf{v})$ can be calculated as in (1.8.1) using the chain rule, *i.e.*, $\left(\frac{\partial f}{\partial x_i}(\mathbf{y} + t\mathbf{v})\right)' = \nabla \frac{\partial f}{\partial x_i}(\mathbf{y} + t\mathbf{v}) \bullet \mathbf{v} = \sum_{j=1}^n \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{y} + t\mathbf{v})v_j$. Summing up over all $i \in [1..n]$, we obtain:

$$g''(t) = \sum_{i=1}^n \sum_{j=1}^n \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{y} + t\mathbf{v})v_j v_i = H_{\mathbf{y}+t\mathbf{v}} \bullet \left(\mathbf{v}\mathbf{v}^\top\right) \tag{1.8.2}$$

$$\implies g''(0) = H_\mathbf{y} \bullet \left(\mathbf{v}\mathbf{v}^\top\right) = \mathbf{v}^\top H_\mathbf{y}\mathbf{v}$$

[8] If you are unfamiliar with gradients, consider that $\nabla f(\mathbf{y})$ is the hyperplane tangent to the function graph at $\mathbf{y}$. Using notational shortcut $\nabla f(\mathbf{y}) = [\,\nabla_1\ \nabla_2\ \ldots \nabla_n\,]$, the function value (of this hyperplane) increases by $\epsilon\nabla_i$ when one performs a step of length ϵ along direction x_i. Let us further study this hyperplane: its value increases by ∇_i when one performs a unit step (of length 1) along direction x_i from any starting point. What happens if one moves along some other direction $\mathbf{v}$? Answer: this is equivalent to advancing a step of v_1 along $\mathbf{x}_1$, followed by a step of v_2 along $\mathbf{x}_2$, etc, leading to a total increase of $\nabla_1 v_1 + \nabla_2 v_2 + \ldots \nabla_n v_n = [\,\nabla_1\ \nabla_2\ \ldots \nabla_n\,]\,\mathbf{v}$. As a side remark, we can also see $[\,\nabla_1\ \nabla_2\ \ldots \nabla_n\,]$ as a gradient direction (vector). The increase (of the hyperplane) when one moves along some direction $\mathbf{v}$ is given by the scalar product between $\mathbf{v}$ and this gradient vector, equivalent to the projection of $\mathbf{v}$ along this gradient. Among all vectors of the unit sphere, the vector/direction that makes the hyperplane value change the most (in absolute value) is the one that is collinear to the gradient direction.

If f is convex, then g is convex for any direction $\mathbf{v} \in \mathbb{R}^n$. As such, $g(0) = \mathbf{v}^\top H_{\mathbf{y}} \mathbf{v} \geq 0 \ \forall \mathbf{v} \in \mathbb{R}^n$, which means that $H_{\mathbf{y}}$ is SDP for any chosen $\mathbf{y} \in \mathbb{R}^n$.

Conversely, if $H_{\mathbf{y}}$ is SDP for any $\mathbf{y} \in \mathbb{R}^n$, then g is convex for any $\mathbf{v} \in \mathbb{R}^n$ because $g''(t) \geq 0 \ \forall t \in \mathbb{R}$ by virtue of (1.8.2). The convexity definition ensures that $\alpha f(\mathbf{y}_1) + \beta f(\mathbf{y}_2) \geq f(\alpha \mathbf{y}_1 + \beta \mathbf{y}_2)$ for all $\alpha, \beta \geq 0$ such that $\alpha + \beta = 1$, *i.e.*, the line joining $f(\mathbf{y}_1)$ and $f(\mathbf{y}_2)$ is above or equal to the function value evaluated at any point $\alpha \mathbf{y}_1 + \beta \mathbf{y}_2$ on the segment $[\mathbf{y}_1, \mathbf{y}_2]$. This can be instantiated as follows:

$$\beta f(\mathbf{y} - \alpha \mathbf{v}) + \alpha f(\mathbf{y} + \beta \mathbf{v}) \geq f(\mathbf{y}) \quad \forall \mathbf{y}, \mathbf{v} \in \mathbb{R}^n, \alpha, \beta \geq 0 \text{ such that } \alpha + \beta = 1 \tag{1.8.3}$$

To prove f is convex, we need to show that for any $\mathbf{y}_1, \ \mathbf{y}_2 \in \mathbb{R}^n$, $\alpha \geq 0$ and $\beta = 1 - \alpha$ we have $\beta f(\mathbf{y}_1) + \alpha f(\mathbf{y}_2) \geq f(\beta \mathbf{y}_1 + \alpha \mathbf{y}_2)$. But this reduces to (1.8.3): if we fix $\mathbf{v} = \mathbf{y}_2 - \mathbf{y}_1$ and replace $\beta \mathbf{y}_1 + \alpha \mathbf{y}_2 = \mathbf{y}$, one can check $\mathbf{y}_1 = \mathbf{y} - \alpha \mathbf{v}$ and $\mathbf{y}_2 = \mathbf{y} + \beta \mathbf{v}$, *i.e.*, we obtain (1.8.3) for $\mathbf{y}$ and $\mathbf{v}$ defined above. ☐

2 Primal-Dual SDP programs and optimization considerations

2.1 Primal and dual SDP programs

2.1.1 Main duality

Proposition 2.1.1. *The dual of a primal SDP program is an SDP program and the following degeneracy-related properties hold:*

(a) If the primal is unbounded, the dual is infeasible.

(b) If the primal is infeasible, the dual can be unbounded, infeasible or non-degenerate.

We will later see that that if the primal is feasible and bounded (non-degenerate), there might be a duality gap with regards to the optimal value of the dual, and, even more, the dual can even be infeasible.

Proof. Let us introduce the first SDP program in variables $\mathbf{x} \in \mathbb{R}^n$, using matrices of an arbitrary order.

$$(SDP) \begin{cases} \min \quad \sum_{i=1}^{n} c_i x_i & \text{(2.1.1a)} \\ \quad s.t \quad \sum_{i=1}^{n} A_i x_i \succeq B & \text{(2.1.1b)} \\ \qquad \mathbf{x} \in \mathbb{R}^n & \text{(2.1.1c)} \end{cases}$$

The inequalities (2.1.1b) are often called *linear matrix inequalities*. Let us now relax them (or penalize their potential violation) using Lagrangian multipliers $Y \succeq \mathbf{0}$ to obtain the Lagrangian function:

$$\mathscr{L}(\mathbf{x}, Y) = \sum_{i=1}^{n} c_i x_i - Y \bullet (\sum A_i x_i - B)$$

Observe that if $\mathbf{x}$ satisfies (2.1.1b), we get awarded in the above Lagrangian because the Lagrangian term subtracts the product of two SDP matrices (non-negative by virtue of Prop. 1.3.4). As in linear programming, we use the convention that $OPT(SDP) = \infty$ if (2.1.1b) has no feasible solution (*i.e.*, we have an infeasible program) and $OPT(SDP) = -\infty$ if (2.1.1a)-(2.1.1c) can be indefinitely small (*i.e.*, we have an unbounded program). However, in all cases, $\min_{\mathbf{x} \in \mathbb{R}^n} \mathscr{L}(\mathbf{x}, Y)$ is a relaxation of (SDP) and we can write:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \mathscr{L}(\mathbf{x}, Y) \leq OPT(SDP), \ \forall Y \succeq \mathbf{0} \tag{2.1.2}$$

We now develop the expression of the Lagrangian:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \mathscr{L}(\mathbf{x}, Y) = \min_{\mathbf{x} \in \mathbb{R}^n} Y \bullet B + \sum_{i=1}^{n} (c_i - Y \bullet A_i) x_i$$

If there is a single $i \in [1..n]$ such that $c_i - Y \bullet A_i \neq 0$, the above minimum is $-\infty$ (unbounded), by using an appropriate value of x_i. To have a bounded $\min_{\mathbf{x}\in\mathbb{R}^n} \mathscr{L}(\mathbf{x}, Y)$, the matrix Y needs to satisfy $c_i - Y \bullet A_i = 0$ for all $i \in [1..n]$. Notice that if we actually consider a non-negative variable $x_i \geq 0$ for some $i \in [1..n]$, the condition becomes $c_i - Y \bullet A_i \geq 0$ for such i. We are interested in finding:

$$\max_{Y \succeq \mathbf{0}} \min_{\mathbf{x}\in\mathbb{R}^n} \mathscr{L}(\mathbf{x}, Y)$$

Or $A_i \bullet Y \leq c_i$ if $x_i \geq 0$ is added to (SDP)

that can be written:

$$(DSDP) \begin{cases} \max\ B \bullet Y & (2.1.3a) \\ s.t\ A_i \bullet Y = c_i\ \forall i \in [1..n] & (2.1.3b) \\ Y \succeq \mathbf{0} & (2.1.3c) \end{cases}$$

Based on (2.1.2), we obtain:

$$OPT(DSDP) \leq OPT(SDP). \tag{2.1.4}$$

The case of degenerate programs is addressed below, proving points (a) and (b) of the conclusion.

(a) In above (2.1.4), we can say that that $OPT(DSDP) = -\infty$ if $(DSDP)$ is not feasible, because $\min\limits_{\mathbf{x}\in\mathbb{R}^n} \mathscr{L}(\mathbf{x}, Y) = -\infty\ \forall Y \succeq \mathbf{0}$ in this case. It is clear that if $(DSDP)$ is feasible, then (SDP) can not be unbounded from below. Thus, if $OPT(SDP) = -\infty$ (unbounded from below), then $(DSDP)$ needs to be infeasible.

(b) If (SDP) is infeasible, we can infer nothing about the dual, *i.e.*, $(DSDP)$ can be unbounded, infeasible or non-degenerate. If you are familiar with linear programming, then it is well-known that the dual of an infeasible LP is infeasible or unbounded. By writing such LPs in an SDP form, one can obtain the examples (i) and (i) below. The most difficult is to find a primal SDP whose dual is non-degenerate. Such programs can be found by exploiting a phenomenon of "clenching" in the development of $\sum_{i=1}^n A_i x_i - B \succeq \mathbf{0}$ or $Y \succeq \mathbf{0}$, *e.g.*, we can use Corollary 1.6.5 to force certain rows or columns of an SDP matrix to be zero, pushing it to a certain form (or to infeasibility). For instance, the SDP matrix in example (iii) below need to have $y = 0$: the zero at position (2,2) forces the second row and the second column to contain only zeros by virtue of Corollary 1.6.5. This makes this program infeasible. On the dual side, we have $Y_{11} = 0$ because x has a coefficient of zero in the primal objective function and $Y_{33} = 1$ because $Y_{12} + Y_{21} + Y_{33} = 1$, which leads to a unique dual feasible solution. $\square$

$$\begin{aligned} &\min\ x \\ &s.t. \begin{bmatrix} x & 0 \\ 0 & -x \end{bmatrix} \succeq \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned}$$

(i) unbounded dual

$$\begin{aligned} &\min\ x \\ &s.t. \begin{bmatrix} x-y & 0 \\ 0 & -x+y \end{bmatrix} \succeq \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned} \qquad \begin{aligned} &\min\ -y \\ &s.t. \begin{bmatrix} x & 0 & 0 \\ 0 & -x & 0 \\ 0 & 0 & y \end{bmatrix} \succeq \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

(ii) infeasible duals

$$\begin{aligned} &\min\ y \\ &s.t. \begin{bmatrix} x & y & 0 \\ y & 0 & 0 \\ 0 & 0 & y \end{bmatrix} \succeq \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned}$$

(iii) non-degenerate dual

Proposition 2.1.2. *Program (DSDP) from (2.1.3a)-(2.1.3c) can be written in the primal form of (SDP) from (2.1.1a)-(2.1.1c).*

Proof. We first solve the system $A_i \bullet Y = c_i\ \forall i \in [1..n]$. If this system has no solution, then the given (DSDP) program is infeasible; in this case, any infeasible (SDP) can be considered (by convention) equivalent to the given infeasible (DSDP) – assuming that both programs have the same optimization (min/max) direction.

The non-degenerate case is the essential one: we consider from now on that the system $A_i \bullet Y = c_i\ \forall i \in [1..n]$ has at least a feasible solution $-B'$. The set of all solutions is given by

$$Y = -B' + \sum_{j=1}^{k} A'_j x'_j, \tag{2.1.5}$$

where $A'_1,\ A'_2,\ldots A'_k$ are a basis (maximum set of independent vectors) of the null space of $\{A_i:\ i\in[1..n]\}$ (see the null space definition in (A.1.2)).[9] The matrices A'_j with $j\in[1..k]$ and A_i with $i\in[1..n]$ satisfy:

$$A_i \bullet A'_j = 0,\ \forall i\in[1..n],\ j\in[1..k] \text{ and } -A_i \bullet B' = c_i,\ \forall i\in[1..n]. \tag{2.1.6}$$

The space spanned by (the linear combinations of) A_i ($\forall i\in[1..n]$) and A'_j ($\forall j\in[1..k]$) need to cover the whole space of symmetric matrices of the size of Y, by virtue of the rank-nullity Theorem A.1.3 that could be applied on the vectorized versions of these matrices. One can confirm that any feasible Y can be expressed in the form (2.1.5): just notice that $Y+B'$ belongs to the null space of $\{A_i:\ i\in[1..n]\}$, and so, it can be expressed as a linear combination of the basis $A'_1,\ A'_2,\ldots A'_k$.

Replacing (2.1.5) in (2.1.3a)-(2.1.3c), we obtain:

$$\begin{aligned} &\max -B\bullet B' + \sum_{j=1}^{k}(B\bullet A'_j)x'_j \\ &s.t.\ \sum_{j=1}^{k} A'_j x'_j \succeq B' \\ &\mathbf{x}'\in\mathbb{R}^k, \end{aligned} \tag{2.1.7}$$

which is an SDP program in the primal form (2.1.1a)-(2.1.1c).

□

Proposition 2.1.3. *If we have $x_j\ge 0$ for certain variables $J\subseteq[1..n]$ of (2.1.1a)-(2.1.1c), we can still write an equivalent program without explicit non-negative variables by incorporating the non-negativities in new rows and columns of (2.1.1b). The dual can be written in the form (2.1.3a)-(2.1.3c), but there is an equivalent dual in which equalities (2.1.3b) become $A_j\bullet Y\le c_j$ for all $j\in J$.*

Proof. We define matrices A'_i ($\forall i\in[1..n]$) and B' by bordering A_i and resp. B with $|J|$ rows and columns that contain only zeros except at the following new positions: A'_j contains 1 at a position (m_j,m_j) that correspond to the bordering row and column associated to $j\in J$. We can drop $x_j\ge 0$ but write an equivalent program (2.1.1a)-(2.1.1c) with matrices A'_i ($\forall i\in[1..n]$) and B'.

The dual of this program has the form (2.1.3a)-(2.1.3c) and it contains constraints of the form $A'_i\bullet Y'=c_i\ \forall i\in[1..n]$. For $i\notin J$, $A'_i\bullet Y'=c_i$ is equivalent to $A_i\bullet Y=c_i$. For $j\in J$, $A'_j\bullet Y'=c_j$ becomes $Y'_{m_j,m_j}+A_j\bullet Y=c_j$, which is equivalent to $A_i\bullet Y\le c_i$, because $Y'_{m_j,m_j}\ge 0$ does not play a role elsewhere in the program. The objective function $B'\bullet Y'$ is equivalent to $B\bullet Y$. The initial dual with matrices A'_i ($\forall i\in[1..n]$) and B' can be equivalently written with matrices A_i ($\forall i\in[1..n]$) and B by using inequality constraints $A_j\bullet Y\le c_j\ \forall j\in J$. Finally, notice that $Y'\succeq\mathbf{0}\implies Y\succeq\mathbf{0}$ because Y is a principal minor of Y (use the definition from Prop. 1.2.5). □

Proposition 2.1.4. *(the case of multiple constraints) Suppose one needs to impose multiple constraints in (2.1.1a)-(2.1.1c):*

$$\sum_{i=1}^{n} A_i^j x_i^j \succeq B^j \tag{2.1.8}$$

for $j\in[1..n']$. Notices that the involved matrices can have order 1 for some $j\in[1..n']$, i.e., (2.1.8) for such j is a linear constraint. However, all these constraints can be incorporated in a unique constraint of the form (2.1.1b) expressed with aggregated block-diagonal matrices with n' blocks. An aggregated dual can be expressed in the canonical form (2.1.3a)-(2.1.3b) using aggregated block-diagonal matrices. This aggregated dual is equivalent to a dual in which we have n' dual matrix variables.

[9] Assuming $A_1,\ A_2,\ldots A_n\in\mathbb{R}^{m\times m}$ are linearly independent, we have $k=\frac{m(m+1)}{2}-n$. We used the rank-nullity Theorem A.1.3 with the fact the set of symmetric matrices of size $m\times m$ has dimension $\frac{m(m+1)}{2}$, using the symmetry constraints.

Proof. We define aggregated block-diagonal matrices A'_i ($\forall i \in [1..n]$) and B' with n' blocks defined by A_i^j and resp. B^j for all $j \in [1..n']$:

$$B' = \begin{bmatrix} B^1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & B^2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & B^{n'} \end{bmatrix} \text{ and } A'_i = \begin{bmatrix} A_i^1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & A_i^2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & A_i^{n'} \end{bmatrix} \forall i \in [1..n]$$

Constraints (2.1.8) are equivalent to a unique constraint in aggregated block-diagonal matrices

$$\sum_{i=1}^{n} A'_i \succeq B'.$$

We obtain an aggregated program (2.1.1a)-(2.1.1c) expressed with block-diagonal matrices A'_i ($\forall i \in [1..n]$) and B'.

We can now construct the dual of this aggregated primal program. We obtain an aggregated dual program of the form (2.1.3a)-(2.1.3c) expressed in aggregated variables $Y' \succeq 0$. The dual objective function is:

$$B' \bullet Y' = \sum_{j=1}^{n'} B^j \bullet Y^j, \tag{2.1.9}$$

where Y^j is obtained from Y' by keeping only the rows and columns that correspond to block B^j inside B' ($\forall j \in [1..n']$). The dual constraints are $A'_i \bullet Y' = c_i$ ($\forall i \in [1..n]$) and they can also be written as:

$$\sum_{j=1}^{n'} A_i^j \bullet Y^j = c_i \ \forall i \in [1..n]. \tag{2.1.10}$$

Notice that the variables $y'_{i',j'}$ outside the n' diagonal blocks of Y' play no role in the constraints or in the objective function of the dual. Also, $Y' \succeq \mathbf{0}$ implies $Y^j \succeq \mathbf{0}$ $\forall j \in [1..n']$ because all Y^j are principal minors of Y (use the definition from Prop. 1.2.5). The aggregated dual of the form (2.1.3a)-(2.1.3b) in variables Y' is equivalent to a dual in variables Y^j (with $j \in [1..n]$) using objective (2.1.9) and constraints (2.1.10). □

2.1.2 The dual of the dual is the initial program

The remaining of Section 2.1 is devoted to a few properties that may seem a bit boring and easy to trust, because they only ask to verify certain equivalences between the dual and the primal forms. However, the exercise of verifying these properties may offer a good insight into the different ways of expressing the same SDP program and into the different ways of understanding its space of feasible solutions.

We now provide two results on the dual of the dual. The first one uses a new type of Lagrangian duality, while the second one only uses the first duality from Prop. 2.1.1.

Proposition 2.1.5. *By dualizing twice the* (SDP) *from* (2.1.1a)-(2.1.1c) *we obtain the initial* (SDP) *and the following properties hold*

(a) If the $(DSDP)$ *from* (2.1.3a)-(2.1.3c) *is unbounded, its dual* (SDP) *is infeasible.*

(b) If $(DSDP)$ *is infeasible,* (SDP) *can be unbounded, infeasible or non-degenerate.*

Proof. Let us calculate the Lagrangian dual of (DSDP) from (2.1.3a)-(2.1.3c) and verify that we obtain the (SDP) from (2.1.1a)-(2.1.1c). We relax constraints (2.1.3b) using coefficients $\mathbf{x}' \in \mathbb{R}^n$:

$$\mathscr{L}'(Y, \mathbf{x}') = B \bullet Y + \sum_{i=1}^{n} (c_i - A_i \bullet Y) x'_i. \tag{2.1.11}$$

For any Y that satisfies (2.1.3b), the value of above Lagrangian is $B \bullet Y$, *i.e.*, the objective value (2.1.3a) of Y in $(DSDP)$. If *no* $Y \succeq \mathbf{0}$ satisfies (2.1.3b), we say $OPT(DSDP) = -\infty$, adopting a similar convention as in linear programming or in Prop. 2.1.1. We also say and $OPT(DSDP) = \infty$ if (2.1.3a)-(2.1.3c) can be indefinitely large (*i.e.*, we have an unbounded program). However, in all cases, we can state:

$$\max_{Y \succeq 0} \mathscr{L}'(Y, \mathbf{x}') \geq OPT(DSDP),\ \forall \mathbf{x}' \in \mathbb{R}^n \tag{2.1.12}$$

We now re-write the above Lagrangian:

$$\max_{Y \succeq 0} \mathscr{L}'(Y, \mathbf{x}') = \max_{Y \succeq 0} \left(B - \sum_{i=1}^{n} A_i x'_i \right) \bullet Y + \sum_{i=1}^{n} c_i x'_i$$

We will show that this expression can only be bounded if $B - \sum\limits_{i=1}^{n} A_i x'_i \preceq \mathbf{0}$. Notice that $Y = \mathbf{0}$ leads to $\left(B - \sum\limits_{i=1}^{n} A_i x'_i \right) \bullet Y = 0$. We need values $\mathbf{x}'$ such that $\left(\sum\limits_{i=1}^{n} A_i x'_i - B \right) \bullet Y \geq 0\ \forall Y \succeq \mathbf{0}$. Using Prop 1.3.3, this can only hold if $\sum\limits_{i=1}^{n} A_i x'_i - B \succeq \mathbf{0}$. For such $\mathbf{x}'$, we have $\max\limits_{Y \succeq 0} \mathscr{L}'(Y, \mathbf{x}') = \sum_{i=1}^{n} c_i x'_i$. We can write:

$$\min_{\mathbf{x}' \in \mathbb{R}^n} \max_{Y \succeq 0} \mathscr{L}'(Y, \mathbf{x}') = \begin{cases} \min & \sum\limits_{i=1}^{n} c_i x'_i \\ s.t & \sum A_i x'_i \succeq B \\ & \mathbf{x}' \in \mathbb{R}^n, \end{cases}$$

which is exactly the (SDP) from (2.1.1a)-(2.1.1c). Based on (2.1.12), we discover (2.1.4) again:

$$OPT(DSDP) \leq OPT(SDP). \tag{2.1.14}$$

We now address points (a) and (b) of the conclusion.

(a) It is clear that if (SDP) is feasible, then $(DSDP)$ can not be unbounded. This means that if $(DSDP)$ unbounded, then (SDP) is infeasible.

(b) If $(DSDP)$ is infeasible, we can infer nothing about (SDP), *i.e.*, (SDP) could be unbounded, infeasible, or non-degenerate. One can find examples of unbounded or infeasible duals by generalizing the linear programming examples. For instance, if if $(DSDP) = \max\{y :\ y = -1,\ y \geq 0\}$, then $(SDP) = \min\{-x :\ x \geq 1\}$ is unbounded. A pair of infeasible primal-dual programs can simply be taken from example (ii) provided at the end of the proof of Prop 2.1.1. To find an example of an infeasible (DSDP) with a feasible (SDP), it is enough to take example (iii) at the end of the proof of Prop. 2.1.1 and to change the right-hand side and the objective function of the (SDP).

If $(DSDP) = \max\left\{ \begin{bmatrix} 1&0&0\\0&0&0\\0&0&0 \end{bmatrix} \bullet Y :\ \begin{bmatrix} 1&0&0\\0&0&0\\0&0&0 \end{bmatrix} \bullet Y = 0,\ \begin{bmatrix} 0&1&0\\1&0&0\\0&0&1 \end{bmatrix} \bullet Y = -1, \right\}$, then we obtain $(SDP) = \min\ \left\{ -x_2 :\ \begin{bmatrix} x_1&x_2&0\\x_2&0&0\\0&0&x_2 \end{bmatrix} \succeq \begin{bmatrix} 1&0&0\\0&0&0\\0&0&0 \end{bmatrix} \right\}$ that has solution $x_2 = 0$ (apply Corollary 1.6.5 on the fact that the middle element is zero) and $x_1 \geq 1$ of objective value zero. □

Proposition 2.1.6. *Assuming the dual $(DSDP)$ from (2.1.3a)-(2.1.3c) is feasible, we can write it into the primal form (2.1.7) as described by Prop. 2.1.2. If we apply the first duality from Prop. 2.1.1 on this primal form, we obtain a dual that is equivalent to the primal (SDP) from (2.1.1a)-(2.1.1c).*

Proof. The proof relies on a few arguments from the proof of Prop. 2.1.2. First, recall that the feasible dual program (2.1.3a)-(2.1.3c) can be written in the primal form (2.1.7). We re-write (2.1.7) as follows:

$$-B \bullet B' - \begin{pmatrix} \min \sum\limits_{i=1}^{k} (-B \bullet A'_i) x'_i \\ s.t \sum\limits_{i=1}^{k} A'_i x'_i \succeq B' \\ \mathbf{x}' \in \mathbb{R}^k. \end{pmatrix}.$$

Recall that $A'_1,\ A'_2, \ldots A'_k$ and B' arise from solving $A_i \bullet Y = c_i$, *i.e.*, any solution Y can be written in the form $Y = -B' + \sum_{i=1}^{k} A'_i x'_i$, where

$$A'_1, \ldots, A'_k \text{ are a basis of the null space of } A_1, \ldots, A_n \text{ and } -B' \bullet A_i = c_i \ \forall i \in [1..n] \qquad (*)$$

We now apply the first duality from Prop. 2.1.1 on the above program in (SDP) form and obtain:

$$-B \bullet B' - \begin{pmatrix} \max\ B' \bullet Y \\ s.t\ A'_i \bullet Y = -B \bullet A'_i\ \forall i \in [1..k] \\ Y \succeq \mathbf{0}. \end{pmatrix}$$

This dual could be infeasible even if the corresponding primal is feasible (we have already presented such examples, see also Prop. 2.2.3). However, the system of linear equations $A'_i \bullet Y = -B \bullet A'_i\ \forall i \in [1..k]$ has at least the feasible solution $Y = -B$. The existence of this solution is sufficient to place us in the non-degenerate case of the proof of Prop. 2.1.2, which leads to re-writing the above program in the primal form. For this, we consider the origin $-B$ and the basis $A_1,\ A_2, \ldots, A_n$ that is orthogonal to all $A'_i,\ \forall i \in [1..k]$. Recall (*) that $A'_1,\ A'_2, \ldots, A'_k$ are a basis for the null space of $A_1,\ A_2, \ldots A_n$. As such, the space generated by (the linear combinations of) $A_1,\ A_2, \ldots A_n$ and $A'_1,\ A'_2, \ldots A'_k$ cover the whole space of symmetric matrices. Any $Y \succeq \mathbf{0}$ can be written in the form $\sum\limits_{i=1}^{n} x_i A_i - B$. The above program can thus be re-written as:

$$\begin{aligned} -B \bullet B' - \max\ & \left(-B \bullet B' + \sum_{i=1}^{n} (B' \bullet A_i) x_i\right) \\ s.t.\ & \sum_{i=1}^{n} A_i x_i \succeq B \\ & \mathbf{x} \in \mathbb{R}^n. \end{aligned}$$

Recall now $B' \bullet A_i = -c_i,\ \forall i \in [1..n]$ from (*). By replacing this in above program and simplifying $-B \bullet B'$, we obtain:

$$\begin{aligned} -\max\ & \left(\sum_{i=1}^{n} -c_i x_i\right) \\ s.t.\ & \sum_{i=1}^{n} A_i x_i \succeq B \\ & \mathbf{x} \in \mathbb{R}^n, \end{aligned}$$

which is exactly (2.1.1a)-(2.1.1c). □

2.1.3 From the primal form to the dual form

The transformation from the primal form to the dual form is more difficult and it is not always possible.

Proposition 2.1.7. *The (SDP) program in the primal form (2.1.1a)-(2.1.1c) can be written in the dual form (2.1.3a)-(2.1.3c), provided that the matrices $A_1,\ A_2, \ldots A_n$ from (2.1.1b) are linearly independent.*

Proof. We write $Y = \sum_{i=1}^{n} A_i x_i - B$ and notice that (2.1.1b) stipulates $Y \succeq \mathbf{0}$. We need to show that the set of such Y can be expressed as the solution set of a system of equations $A'_i \bullet Y = c'_i\ \forall i \in [1..k]$, *i.e.*, of the form (2.1.3b). However, the most difficult task is to write objective function (2.1.1a) in the form $B' \bullet Y$ as in (2.1.3a); for this, we need to express $x_1,\ x_2, \ldots x_n$ as linear combinations of variables Y. We will see that $x_1,\ x_2, \ldots x_n$ can be exactly determined from Y when $A_1,\ A_2, \ldots A_n$ are linearly independent.

Consider the following program in the form (2.1.1a)-(2.1.1c) in which A_1 and A_2 are *not* linearly independent, since $A_1 = A_2 = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$.

$$\begin{aligned} \min\ & 2x_1 + 3x_2 \\ s.t.\ & \begin{bmatrix} 1 & x1 + x2 \\ x_1 + x_2 & 1 \end{bmatrix} \succeq \mathbf{0} \\ & x_1,\ x_2 \in \mathbb{R} \end{aligned}$$

If we try to write $Y = \left[\begin{smallmatrix} 1 & x1+x2 \\ x_1+x_2 & 1 \end{smallmatrix}\right]$, then the set of feasible symmetric matrices Y is the set of solutions of the system $Y \bullet \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right] = 1$ and $Y \bullet \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right] = 1$. However, it is impossible to express the objective function value in terms of variables Y. What is the objective value of $Y = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$? It can actually be any real number or even $-\infty$, depending on the choice of x_1 and x_2 such that $0 = x_1 + x_2$.

We hereafter assume that $A_1,\ A_2, \dots A_n$ are linearly independent and we provide an algorithm for finding a system of equations whose solutions Y can all be written in the form $Y = \sum_{i=1}^n A_i x_i - B$. Let us work with "vectorized" versions of the matrices: the notation $\overline{X}$ represents a column vector containing the diagonal and the upper triangular elements of symmetric matrix X. The size of this vector is $m = \frac{q(q+1)}{2}$, where q is the order of X. We have $Y = \sum_{i=1}^n A_i x_i - B \Longleftrightarrow \overline{Y} = \sum_{i=1}^n \overline{A}_i x_i - \overline{B}$. Using aggregate matrix $\overline{A} \in \mathbb{R}^{m \times n}$ such that $\overline{A} = [\overline{A}_1\ \overline{A}_2 \dots \overline{A}_n]$, we can write:

$$\overline{Y} + \overline{B} = \overline{A}\mathbf{x} \tag{2.1.15}$$

Since $A_1,\ A_2, \dots A_n$ are linearly independent, $\overline{A}$ has rank n; this also ensures $m \geq n$. Without loss of generality, we can consider that the first n rows of $\overline{A}$ are linearly independent and form a non-null minor $[\overline{A}]_n$. The equations corresponding to the first n rows of (2.1.15) can be written $[\overline{Y}]_n + [\overline{B}]_n = [\overline{A}]_n \mathbf{x}$. As such, we can deduce $\mathbf{x} = [\overline{A}]_n^{-1}([\overline{Y}]_n + [\overline{B}]_n)$. This ensures that the objective value $\sum_{i=1}^n c_i x_i$ from (2.1.1a) can be written as a linear combination of the $[\overline{Y}]_n$ values, *i.e.*, in the form $B' \bullet Y$ as in (2.1.3a). We now replace $\mathbf{x}$ in (2.1.15) and obtain:

$$\overline{Y} + \overline{B} = \overline{A}[\overline{A}]_n^{-1}([\overline{Y}]_n + [\overline{B}]_n). \tag{2.1.16}$$

The first n rows of above formula are redundant. The remaining $k = m - n$ rows actually represent a system of linear equations: notice that each element of the left-hand $\overline{Y}$ is expressed as a linear combination of the free variables $[\overline{Y}]_n$ (plus a fixed coefficient). Now check than any such linear equation in variables Y can be rewritten in the form $A'_i \bullet Y = c'_i$ (for any $i \in [1..k]$).

We have just shown that any solution $\overline{Y}$ of (2.1.15) satisfies a system of equations of the form $A'_i \bullet Y = c'_i$ (with $i \in [1..k]$). We now prove the converse: any solution of this system can be written in the form (2.1.15) for a certain $\mathbf{x} \in \mathbb{R}^n$; more exactly, this $\mathbf{x}$ is the unique solution of $[\overline{Y}]_n = [\overline{A}]_n \mathbf{x} - [\overline{B}]_n$. To show this, it is enough to see that any solution of the above system satisfies (2.1.16) by construction and that, after replacing $[\overline{Y}]_n = [\overline{A}]_n \mathbf{x} - [\overline{B}]_n$, (2.1.16) becomes

$$\begin{aligned}
\overline{Y} + \overline{B} &= \overline{A}[\overline{A}]_n^{-1}([\overline{Y}]_n + [\overline{B}]_n) \\
&= \overline{A}[\overline{A}]_n^{-1}([\overline{A}]_n \mathbf{x} - [\overline{B}]_n + [\overline{B}]_n) \\
&= \overline{A}[\overline{A}]_n^{-1}([\overline{A}]_n \mathbf{x}) \\
&= \overline{A}\mathbf{x}
\end{aligned}$$

By "devectorizing" $\overline{Y},\ \overline{B},\ \overline{A}_1,\ \overline{A}_2, \dots \overline{A}_n$ into symmetric matrices $Y,\ B,\ A_1,\ A_2, \dots A_n$ (put the corresponding elements on the diagonal and resp. on the symmetric matrix positions), we obtain $Y = \sum_{i=1}^n A_i x_i - B$. $\square$

Example 2.1.8. *We apply the above proof of Prop. 2.1.7 on an example showing how to rewrite a program (2.1.1a)-(2.1.1c) in the form of (2.1.3a)-(2.1.3c). Consider*

$$\begin{aligned}
&\min\ x_1 + 2x_2 + x_3 + 2x_4 \\
&s.t. \begin{bmatrix} 0 & x_1 + x_3 & x_4 \\ x_1 + x_3 & x_2 + x_3 + 1 & x_4 + 2x_2 \\ x_4 & x_4 + 2x_2 & x_1 + x_2 + 2 \end{bmatrix} \succeq \mathbf{0} \\
&\quad x_1,\ x_2,\ x_3,\ x_4 \in \mathbb{R}
\end{aligned} \tag{2.1.17}$$

We write a formula corresponding to (2.1.15) but without "vectorization":

$$\begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{12} & y_{22} - 1 & y_{23} \\ y_{13} & y_{23} & y_{33} - 2 \end{bmatrix} = \begin{bmatrix} 0 & x_1 + x_3 & x_4 \\ x_1 + x_3 & x_2 + x_3 & x_4 + 2x_2 \\ x_4 & x_4 + 2x_2 & x_1 + x_2 \end{bmatrix} \tag{2.1.18}$$

As in paragraph below (2.1.15), we will express the four variables $x_1,\ x_2,\ x_3$ and x_4 in terms of four variables Y. We can not choose Y_{11}, because the corresponding row of (2.1.15) has only zeros in the matrix $\overline{A}$. Let us *choose:* $y_{12} = x_1 + x_3$, $y_{22} - 1 = x_2 + x_3$, $y_{33} - 2 = x_1 + x_2$ *and* $y_{13} = x_4$. *We obtain*

$$x_1 = \frac{y_{12} + y_{33} - y_{22} - 1}{2},\ x_2 = \frac{y_{22} + y_{33} - y_{12} - 3}{2},\ x_3 = \frac{y_{12} + y_{22} - y_{33} + 1}{2} \text{ and } x_4 = y_{13}. \tag{2.1.19}$$

Our objective function can be written:

$$\frac{y_{12} + y_{33} - y_{22} - 1}{2} + 2\frac{y_{22} + y_{33} - y_{12} - 3}{2} + \frac{y_{12} + y_{22} - y_{33} + 1}{2} + 2y_{13} = y_{22} + y_{33} + 2y_{13} - \frac{3}{2}. \tag{2.1.20}$$

We now replace the values of $x_1,\ x_2,\ x_3,\ x_4$ *in the remaining equations of* (2.1.18). *We have* $y_{11} = 0$ *and* $y_{23} = y_{13} + y_{22} + y_{33} - y_{12} - 3$, *or* $3 = y_{13} + y_{22} + y_{33} - y_{12} - y_{23}$. *Combining these two equations with* (2.1.20), *we obtain the program:*

$$\begin{aligned}
\min \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \bullet Y - \frac{3}{2} \\
\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \bullet Y = 0 \\
\begin{bmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & 1 \end{bmatrix} \bullet Y = 3 \\
Y \succeq \mathbf{0}
\end{aligned}$$

One can check that any feasible solution Y *of above program can be written as in* (2.1.17). *For this, it is enough to determine variables* $x_1,\ x_2,\ x_3,\ x_4$ *from* Y *using* (2.1.19) *and check that all positions of* Y *and the objective function from above program can be written using variables* $\mathbf{x}$ *which leads to* (2.1.17). *For instance, to check that* y_{23} *can be written as* $x_4 + 2x_2$ *with* x_2 *and* x_4 *determined by* (2.1.19), *we write:* $x_4 + 2x_2 = y_{13} + y_{22} + y_{33} - y_{12} - 3 = y_{23}$, *where we used the second constraint on* Y *for the last equality.*

We used several times in this section the transformation from the dual form into the primal form. All these transformations rely on solving the system $A_i \bullet Y = c_i$, with $i \in [1..n]$. You might have noticed there are several ways to solve this system. Just to give an example, consider that $y_1 + y_2 = 10$ can be solved in two manners: (a) take $y_1 = x_1$ as a free variable and all solutions have the form $[y_1\ y_2] = [x_1\ \ x_1 - 10]$ or (b) take solution $[5\ 5]$ and all solutions take the form $[5\ 5] + x_1[-1\ 1]$. We can say that the approach (a) was used in the Example 2.1.8 above, while the approach (b) was used in Prop 2.1.2. They are both related, because switching from (a) to (b) is equivalent to performing a change of variable in the resulting primal, *e.g.*, take (2.1.1a)-(2.1.1c) and use variables $x'_i = x_i + 7$ instead of x_i $\forall i \in [1..n]$. Such a change leads to a different primal form, in which the objective $\sum_{i=1}^n c_i x_i$ evolves to $\sum_{i=1}^n c_i x'_i - 7n$.

2.2 Relations between the primal optimum and the dual optimum

Proposition 2.2.1. *(Complementary Slackness) If* $\overline{\mathbf{x}}$ *and* $\overline{Y}$ *are the optimum primal and resp. dual solutions of SDP and resp. DSDP (in* (2.1.1a)-(2.1.1c) *and* (2.1.3a)-(2.1.3c) *resp.), the duality gap can be written:*

$$OPT(SDP) - OPT(DSDP) = \overline{Y} \bullet \left(\sum_{i=1}^n A_i \overline{x_i} - B \right) \tag{2.2.1}$$

There is no strict complementarity as in linear programming, i.e., the matrices of the above product might share an eigenvector whose eigenvalue is zero in both matrices for any optimal $\overline{\mathbf{x}}$ *and* $\overline{Y}$.

Proof. It is enough to develop

$$\begin{aligned}
OPT(SDP) - OPT(DSDP) &= \sum_{i=1}^n c_i \overline{x_i} - B \bullet \overline{Y} \\
&= \sum_{i=1}^n \left(A_i \bullet \overline{Y} \right) \overline{x_i} - B \bullet \overline{Y} \\
&= \overline{Y} \bullet \left(\sum_{i=1}^n A_i \overline{x_i} - B \right)
\end{aligned}$$

When strong duality holds (*i.e.*, when $OPT(SDP) - OPT(DSDP) = 0$), we can observe the following using the eigen-decomposition (1.1.1):

(a) the eigenvectors of $(\sum_{i=1}^n A_i\overline{x_i} - B)$ with non-zero eigenvalues belong to the space generated by the eigenvectors of $\overline{Y}$ with eigenvalue 0.

(b) the eigenvectors of $\overline{Y}$ with non-zero eigenvalues belong to the space generated by the eigenvectors of $(\sum_{i=1}^n A_i\overline{x_i} - B)$ with eigenvalue 0.

In intuitive terms, we can say that any eigenvector of $(\sum_{i=1}^n A_i\overline{x_i} - B)$ with a non-zero eigenvalue can be seen as an eigenvector of $\overline{Y}$ with eigenvalue 0 and vice-versa. Notice we did *not* claim that $rank\,(\sum_{i=1}^n A_i\overline{x_i} - B) = nullity(\overline{Y})$. Take for instance the matrices $\begin{bmatrix} 1&0&0\\0&0&0\\0&0&0 \end{bmatrix}$ and $\begin{bmatrix} 0&0&0\\0&0&0\\0&0&1 \end{bmatrix}$. They satisfy above conditions (a) and (b), but the first one has rank 1 and the second one has nullity 2. However, if there exist multiple primal and dual optimal solutions, do some of them satisfy $rank\,(\sum_{i=1}^n A_i\overline{x_i} - B) = nullity(\overline{Y})$?

This is called the strict complementarity property: any eigenvector of $(\sum_{i=1}^n A_i\overline{x_i} - B)$ with a zero eigenvalue is an eigenvector of $\overline{Y}$ with non-zero eigenvalue. In linear programming, a theorem of Goldman and Tucker[10] states that there always exist primal-dual solutions $\mathbf{x}$ and $\mathbf{y}$ that are *strictly* complementary, *i.e.*, every zero of $A\mathbf{x} - \mathbf{b}$ corresponds to non-zero of $\mathbf{y}$. In SDP programming, this strict complementarity property no longer holds. Matrices $(\sum_{i=1}^n A_i\overline{x_i})$ and $\overline{Y}$ can share an eigenvector whose eigenvalue is zero in both matrices. Consider the following primal-dual programs, both expressed in the form of (2.1.1a)-(2.1.1c); the dual was transformed as described in Prop. 2.1.2 (see below).

$$\begin{aligned} &\min\ x_3 \\ &s.t. \begin{bmatrix} x_1 & x_2 & 0 & 0 \\ x_2 & 0 & 0 & 0 \\ 0 & 0 & x_2 & 0 \\ 0 & 0 & 0 & x_3-2 \end{bmatrix} \succeq \mathbf{0} \\ &\quad x_1,\ x_2,\ x_3 \in \mathbb{R} \end{aligned} \qquad\qquad \begin{aligned} &\max\ 2y_{44} \\ &s.t.\ Y = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & a^2+\Delta^2 & 0 & a \\ 0 & 0 & 0 & 0 \\ 0 & a & 0 & 1 \end{bmatrix} \\ &\quad Y \succeq 0 \end{aligned}$$

The optimum primal solution $\overline{\mathbf{x}}$ satisfies $\overline{x}_2 = 0$ (row 2 and column 2 need to be zero, because position (2,2) is zero) and $\overline{x}_3 = 2$. The matrix $\sum\limits_{i=1}^n A_i\overline{x_i} - B$ is $\begin{bmatrix} \overline{x}_{11} & 0&0&0\\ 0&0&0&0\\0&0&0&0\\0&0&0&0 \end{bmatrix}$ which has at maximum rank 1.

Any feasible Z satisfies $z_{11} = 0$ because the coefficient of x_1 is zero in the primal objective function. This forces row 1 and column 1 of Z to have only zeros. The dual constraint corresponding to x_2 is $2z_{12} + z_{33} = 0$; since $z_{12} = 0$, we have $z_{13} = 0$. The dual constraint corresponding to x_3 imposes $z_4 = 1$. There is no constraint on $z_{24} = a$; z_{22} needs to be greater than or equal to a^2 so as to have a non-negative principal minor corresponding to rows/columns 2 and 4; we can write $z_{22} = a^2 + \Delta^2$.

On the dual side, any feasible Y satisfies $y_{11} = 0$ because the coefficient of x_1 is zero in the primal objective function. This forces row 1 and column 1 of Y to have only zeros. The dual constraint corresponding to x_2 is $2y_{12} + y_{33} = 0$; since $y_{12} = 0$, we have $y_{13} = 0$. The dual constraint corresponding to x_3 imposes $y_4 = 1$. There is no constraint on $y_{24} = a$; y_{22} needs to be greater than or equal to a^2 so as to have a non-negative principal minor corresponding to rows/columns 2 and 4; we can write $y_{22} = a^2 + \Delta^2$. The nullity of any feasible Y is at least 2. Both matrices have eigenvector $\begin{bmatrix} 0&0&1&0 \end{bmatrix}^\top$ with eigenvalue 0. □

Proposition 2.2.2. *(Non-zero duality gap) If $\overline{\mathbf{x}}$ and $\overline{Y}$ are the optimum primal and resp. dual solutions of SDP and resp. DSDP (in (2.1.1a)-(2.1.1c) and (2.1.3a)-(2.1.3c) resp.), the duality gap $OPT(SDP) - OPT(DSDP)$ is not necessarily 0.*

Proof. Consider the following primal-dual programs, where the dual is actually written in the primal form, as obtained after applying the transformation from Prop. 2.1.2.

[10] I first learned of this theorem from the article "Semidefinite programming" by M. Overton and H. Wolkowicz, a foreword for a special issue on SDP of Mathematical Programming (Volume 77, Issue 1, 1997).

$$\begin{aligned}
&\min\ x_3\\
&s.t.\ \begin{bmatrix} x_1 & x_2 & 0 & 0\\ x_2 & 0 & 0 & 0\\ 0 & 0 & x_2 & 1-x_3\\ 0 & 0 & 1-x_3 & x_3 \end{bmatrix} \succeq \mathbf{0}\\
&\quad x_1,\ x_2,\ x_3 \in \mathbb{R}
\end{aligned}
\qquad\qquad
\begin{aligned}
&\max\ \ -2y_{34}\\
&s.t.\ Y = \begin{bmatrix} 0 & 0 & 0 & 0\\ 0 & a^2+\Delta^2 & 0 & a\\ 0 & 0 & 0 & 0\\ 0 & a & 0 & 1 \end{bmatrix}\\
&\quad Y \succeq 0
\end{aligned}$$

The optimum primal solution satisfies $\overline{x}_2 = 0$ (row 2 and column 2 need to be zero using Corollary 1.6.5, because position (2,2) is zero) and $\overline{x}_3 = 1$ (row 3 and column 3 need to be zero because position (3,3) is $x_2 = 0$). The primal optimum solution is 1.
On the dual side, y_{11} needs to be zero because the objective coefficient of x_1 is zero. As such, all elements on row 1 and column 1 of Y need to be zero. The constraint corresponding to x_2 stipulates $2y_{12} + y_{33} = 0$. Since $y_{12} = 0$, we need to have $y_{33} = 0$. This means that the row 3 and column 3 contain only zeros, and so, $y_{34} = 0$, *i.e.*, the *dual optimum objective value is 0*. Finally, let us fill the remaining elements of the optimal Y and check its feasibility. The dual constraint corresponding to x_3 imposes $y_4 = 1$. There is no constraint on $y_{24} = a$; y_{22} needs to be greater than or equal to a^2 so as to have a non-negative principal minor corresponding to rows/columns 2 and 4. □

Proposition 2.2.3. *The dual DSDP (2.1.3a)-(2.1.3c) of a feasible SDP (2.1.1a)-(2.1.1c) is not necessarily feasible.*

Proof. We have actually already shown an example that shows this at point (b) in the proof of Prop 2.1.5. Let us modify a bit this example for the sake of diversity, by adding a variable x_3. Consider the following primal program that has at least the feasible solution $x_1 = x_2 = x_3 = 0$.

$$\begin{aligned}
&\min\ x_3 - x_2\\
&s.t.\ \begin{bmatrix} x_1 & x_2 & x_3\\ x_2 & 0 & 0\\ x_3 & 0 & x_2 \end{bmatrix} \succeq \mathbf{0}\\
&\quad x_1,\ x_2,\ x_3 \in \mathbb{R}
\end{aligned}$$

The dual of this program is infeasible. Since x_1 has a null objective function coefficient, y_{11} needs to be zero. This means (Corollary 1.6.5) that all elements on row 1 and column 1 of Y need to be zero. We thus obtain $y_{13} = 0$. On the other hand, the dual constraint corresponding to x_3 stipulates that $y_{13} = \frac{1}{2}$, which is a contradiction. Finally, notice $-x_2$ is not necessary in the primal objective value. □

2.3 Strong duality

Several results from this section (including the final proof of the strong duality) are taken from a course of Anupam Gupta, also using arguments from the lecture notes of László Lovász.[11]

2.3.1 Basic facts on the cone of SDP matrices and the cone of definite positive matrices

Proposition 2.3.1. *The SDP matrices of size $n \times n$ form a closed convex cone S_n^+ (see also Def. 7.2.1). The set of positive* ***definite*** *matrices form an open cone* $\texttt{interior}(S_n^+)$*, see also the interior Definition 7.2.3. The closure of* $\texttt{interior}(S_n^+)$ *is* S_n^+.

Proof. Most statements follow by applying the definitions. First, it is easy to prove that any $X \succeq \mathbf{0}$ for which there is a non-zero $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{v}^\top X \mathbf{v} = 0$ does not belong to the interior of S_n^+. We only need to show there is *no* open ball centered at X which is completely contained in S_n^+ (see also the interior Definition 7.2.3). This follows from the fact that S_n^+ contains no element from the set $\{X - \varepsilon I :\ \varepsilon > 0\}$ because $\mathbf{v}^\top (X - \varepsilon I)\mathbf{v} = 0 - \varepsilon |\mathbf{v}|^2 < 0$, where $|\mathbf{v}|$ is the 2-norm $\sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$ of $\mathbf{v}$.

On the other hand, a matrix $X \succ \mathbf{0}$ does belong to the interior $\texttt{interior}(S_n^+)$ as it does contain an open ball centered at X. It is not hard to show there exists a sufficiently small $\varepsilon > 0$ such that $X + \varepsilon Y$ remains positive definite for any symmetric matrix of bounded 2-norm. Take any $\mathbf{v} \in R^n$ such that $|\mathbf{v}| = 1$.

[11] As of 2017, they are available, respectively at http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15859-f11/www/notes/lecture12.pdf and http://www.ime.usp.br/~fmario/sdp/lovasz.pdf.

Let $\lambda_1 > 0$ be the minimum eigenvalue of X so that $\mathbf{v}^\top X\mathbf{v} \geq \lambda_1$ using Lemma 1.2.2.1. We now develop $\mathbf{v}^\top (X + \varepsilon Y)\mathbf{v} = \mathbf{v}^\top X\mathbf{v} + \varepsilon\mathbf{v}^\top Y\mathbf{v} \geq \lambda_1 + \varepsilon\mathbf{v}^\top Y\mathbf{v}$. Since both $\mathbf{v}$ and Y have bounded norm, $\mathbf{v}^\top Y\mathbf{v}$ is bounded; this way, there exists a sufficiently small $\varepsilon > 0$ that makes $\varepsilon\mathbf{v}^\top Y\mathbf{v}$ ridiculously small compared to λ_1. This shows that $\mathbf{v}^\top (X + \varepsilon Y)\mathbf{v} > 0$ for any $\mathbf{v}$ of norm 1, equivalent to $X + \varepsilon Y \succ \mathbf{0}$.

It is easy to verify that both S_n^+ and $\texttt{interior}(S_n^+)$ are convex cones. If $X \succeq \mathbf{0}$ (resp. $X \succ \mathbf{0}$), for any $\alpha > 0$ we have $\alpha X \succeq \mathbf{0}$ (resp. $\alpha X \succ \mathbf{0}$) because $\mathbf{v}^\top(\alpha X)\mathbf{v} = \alpha\mathbf{v}^\top X\mathbf{v} \geq 0$ (resp. > 0) for any $\mathbf{v} \in \mathbb{R}^n - \{\mathbf{0}\}$. One can also confirm the convexity: for any $\alpha \in (0, 1)$ and $X, Y \succeq \mathbf{0}$ (resp. $X, Y \succ \mathbf{0}$), we have that $Z_\alpha = \alpha X + (1-\alpha)Y$ verifies $Z_\alpha \succeq \mathbf{0}$ (resp. $Z_\alpha \succ \mathbf{0}$), because $\mathbf{v}^\top Z_\alpha\mathbf{v} = \alpha\mathbf{v}^\top X\mathbf{v} + (1-\alpha)\mathbf{v}^\top Y\mathbf{v} \geq 0$ (resp. > 0) for any $\mathbf{v} \in \mathbb{R}^n - \{\mathbf{0}\}$.

To prove that S_n^+ is closed, we need to show S_n^+ contains all its limit points. Assume the contrary: there is a sequence $\{X_i\}$ with $X_i \succeq \mathbf{0}\ \forall i \in \mathbb{N}^*$ such that $\lim\limits_{i\to\infty} X_i = Z \not\succeq \mathbf{0}$. This means there exists a non-zero rank 1 matrix $V = \mathbf{v}\mathbf{v}^\top$ such that $Z \bullet V = -a < 0$. As such, $\lim\limits_{i\to\infty} X_i \bullet V = Z \bullet V = -a < 0$. The convergence definition states that for any $\varepsilon > 0$ there exists an $m \in \mathbb{N}$ such that $X_i \bullet V \in [-a-\varepsilon, -a+\varepsilon]$ for any $i \geq m$. Taking an $\varepsilon < a$, we obtain that $X_m \bullet V < 0$ which means $X_m \not\succeq \mathbf{0}$, contradiction. All limit points $\lim\limits_{i\to\infty} X_i$ have to belong to S_n^+.

We still need to prove that the closure of $\texttt{interior}(S_n^+)$ is S_n^+. Since all limit points of S_n^+ belong to S_n^+ (above paragraph), all limit points of $\texttt{interior}(S_n^+) \subset S_n^+$ have to belong to S_n^+ as well. We only need to prove that any $X \succeq \mathbf{0}$ is a limit point of a sequence $\{X_i\}$ such that $X_i \succ \mathbf{0}\ \forall i \in \mathbb{N}^*$. It is enough to take $X_i = X + \frac{1}{i}I_n$ and one can check that $X_i = X + \frac{1}{i}I_n \succ \mathbf{0}$ because $\mathbf{v}^\top(X + \frac{1}{i}I_n)\mathbf{v} = \mathbf{v}^\top X\mathbf{v} + \frac{1}{i}\mathbf{v}^\top I\mathbf{v} \geq \frac{1}{i}|\mathbf{v}|^2 > 0$ for all $\mathbf{v} \in \mathbb{R}^n - \{\mathbf{0}\}$. The convergence to X is easy to prove using $\lim\limits_{i\to\infty} \dfrac{1}{i}I_n = \mathbf{0}$.

□

2.3.2 The proof of the strong duality

We need the following proposition.

Proposition 2.3.2. *Let $F(\mathbf{x}) = \sum_{i=1}^{n} x_i A_i - B$ for any $\mathbf{x} \in \mathbb{R}^n$, where all matrices B and A_i (with $i \in [1..n]$) are symmetric.*

$$F(\mathbf{x}) \not\succ \mathbf{0}\ \forall \mathbf{x} \in \mathbb{R}^n \iff$$

$$\exists Y \succeq \mathbf{0},\ Y \neq \mathbf{0} \text{ such that } A_i \bullet Y = 0\ \forall i \in [1..n] \text{ and } F(\mathbf{x}) \bullet Y = -B \bullet Y \leq 0,\ \forall \mathbf{x} \in \mathbb{R}^n. \tag{2.3.1}$$

In other words, if the sub-space generated by $A_1, A_2, \dots A_n$ with basis $-B$ does not touch $\texttt{interior}(S_n^+)$, then this sub-space belongs a hyperplane $\{X :\ X \bullet Y = -B \bullet Y,\ X \text{ symmetric}\}$.

Proof.
$\Longleftarrow$
If $F(\mathbf{x}) \bullet Y \leq 0$, $F(\mathbf{x})$ can not be positive definite because any $Z \succ \mathbf{0}$ verifies $Z \bullet Y > 0$ for non-zero $Y \succeq \mathbf{0}$. To check this, use the eigenvalue decomposition (*i.e.*, (B.2.3) of Proposition B.2.1) to write Y as a sum of non-zero rank 1 matrices of the form $\mathbf{v}\mathbf{v}^\top$. We have $Z \bullet \mathbf{v}\mathbf{v}^\top > 0$ by Definition 1.2.1.
$\Longrightarrow$
The interior $\texttt{interior}(S_m^+)$ of the SDP cone does not intersect the image of F. We can apply the hyperplane separation Theorem C.4.1: there exists a non-zero symmetric $Y \in \mathbb{R}^{m\times m}$ and a $c \in \mathbb{R}$ such that

$$F(\mathbf{x}) \bullet Y \leq c \leq X \bullet Y\ \forall \mathbf{x} \in \mathbb{R}^n,\ \forall X \in \texttt{interior}(S_m^+) \tag{2.3.2}$$

It is clear that we can not have $c > 0$ because $X \bullet Y$ can be arbitrarily close to 0 by choosing $X = \varepsilon I_m$ for an arbitrarily small $\varepsilon > 0$. We now prove $X \bullet Y \geq 0\ \forall X \in \texttt{interior}(S_m^+)$. Let us assume the contrary: $\exists X \in \texttt{interior}(S_m^+)$ such that $X \bullet Y = c' < 0$. By the cone property of $\texttt{interior}(S_m^+)$, we have $tX \in \texttt{interior}(S_m^+)\ \forall t > 0$. The value $(tX) \bullet Y = tc'$ can be arbitrarily low by choosing an arbitrarily large t, and so, $(tX) \bullet Y$ can be easily less than c, contradiction. This means that $X \bullet Y \geq 0$ for all $X \in \texttt{interior}(S_m^+)$.

Based on (2.3.2) and on the fact that $c \leq 0$, we obtain:

$$F(\mathbf{x}) \bullet Y \leq 0 \leq X \bullet Y\ \forall \mathbf{x} \in \mathbb{R}^n,\ \forall X \in \texttt{interior}(S_m^+)$$

We prove that $Y \succeq \mathbf{0}$. For this, we show that $\overline{X} \bullet Y \geq 0\ \forall \overline{X} \in S_m^+$. Assume the contrary: there is some $\overline{X} \in S_m^+$ such that $\overline{X} \bullet Y < 0$. For any $\varepsilon > 0$, we have $\overline{X} + \varepsilon I_m \in \texttt{interior}(S_m^+)$. For a small enough ε, $(\overline{X} + \varepsilon I_m) \bullet Y$ remains strictly negative, which contradicts $\overline{X} + \varepsilon I_m \in \texttt{interior}(S_m^+)$. We obtain $\overline{X} \bullet Y \geq 0\ \forall \overline{X} \succeq \mathbf{0}$. Using the fact that the SDP cone is self-dual (Prop 1.3.3), we obtain $Y \succeq \mathbf{0}$. We have just found an SDP matrix Y such that

$$F(\mathbf{x}) \bullet Y \leq 0\ \forall \mathbf{x} \in \mathbb{R}^n. \tag{2.3.3}$$

We still need to show that $F(\mathbf{x})$ is constant for every $\mathbf{x} \in \mathbb{R}$, which is equivalent to $A_i \bullet Y = 0\ \forall i \in [1..n]$. This is not difficult. If there is a single $\mathbf{x} \in \mathbb{R}^n$ such that $F(\mathbf{x}) = F(\mathbf{0}) + \Delta$ with $\Delta \neq 0$, notice $F(t\mathbf{x})$ can become arbitrarily large using an appropriate value of t (*i.e.*, use $t \to \infty$ for $\Delta > 0$ or $t \to -\infty$ otherwise), which is impossible. We thus need to have:

$$F(\mathbf{x}) \bullet Y = F(\mathbf{0}) \bullet Y = -B \bullet Y \leq 0\ \forall \mathbf{x} \in \mathbb{R}^n,$$

where we used (2.3.3) for the last inequality. □

For the reader's convenience, we repeat the definitions of the SDP program (2.1.1a)-(2.1.1c) and resp. of its dual (2.1.3a)-(2.1.3c):

$$(SDP) \quad \min\left\{\sum_{i=1}^{n} c_i x_i : \ \sum_{i=1}^{n} A_i x_i \succeq B,\ \mathbf{x} \in \mathbb{R}^n\right\}, \tag{2.3.4}$$

and

$$(DSDP) \quad \max\left\{B \bullet Y : \ A_i \bullet Y = c_i\ \forall i \in [1..n],\ Y \succeq \mathbf{0}\right\}. \tag{2.3.5}$$

Theorem 2.3.3. *If the primal (SDP) from (2.3.4) is bounded and has a strictly feasible solution (Slater's interiority condition), then the primal and the dual optimal values are the same and the dual $(DSDP)$ from (2.3.5) reaches this optimum value. Recall (Prop. 2.1.1) that if (SDP) is unbounded, then $(DSDP)$ is infeasible.*

Proof. Let p be the optimal primal value. The system $\sum\limits_{i=1}^{n} c_i x_i < p$ and $\sum_{i=1}^{n} A_i x_i \succeq B$ has no solution. We define

$$A_i' = \begin{bmatrix} -c_i & \mathbf{0}_n^\top \\ \mathbf{0}_n & A_i \end{bmatrix}\ \forall i \in [1..n] \text{ and } B' = \begin{bmatrix} -p & \mathbf{0}_n^\top \\ \mathbf{0}_n & B \end{bmatrix}$$

and observe that $\sum\limits_{i=1}^{n} A_i' x_i - B' \not\succ \mathbf{0}\ \forall \mathbf{x} \in \mathbb{R}^n$ (we can *not* say $\sum\limits_{i=1}^{n} A_i' x_i - B' \not\succeq \mathbf{0}$, as the optimal solution $\mathbf{x}$ can cancel the top-left term of the expression). We can thus apply Prop. 2.3.2 (implication "$\Longrightarrow$") and conclude there is some non-zero $Y' \succeq \mathbf{0}$ such that $A_i' \bullet Y' = 0\ \forall i \in [1..n]$ and $-B' \bullet Y' \leq 0$. Writing $Y' = \begin{bmatrix} t & \cdots \\ \vdots & Y \end{bmatrix}$, we obtain:

$$tc_i = A_i \bullet Y,\ \forall i \in [1..n] \tag{2.3.6}$$

and

$$-B \bullet Y \leq -tp. \tag{2.3.7}$$

We now prove $t > 0$ by contradiction. Supposing $t = 0$, we obtain $A_i \bullet Y = 0\ \forall i \in [1..n]$ and $-B \bullet Y \leq 0$. Applying again Prop. 2.3.2 (implication "$\Longleftarrow$"), we conclude $\sum\limits_{i=1}^{n} A_i x_i - B \not\succ \mathbf{0}\ \forall \mathbf{x} \in \mathbb{R}^n$, which contradicts the fact the primal (2.3.4) is strictly feasible. We need to have $t > 0$.

Taking SDP matrix $\overline{Y} = \frac{1}{t} Y$, (2.3.6)-(2.3.7) become: $c_i = A_i \bullet \overline{Y}\ \forall i \in [1..n]$ and $B \bullet \overline{Y} \geq p$. In other words, $\overline{Y}$ is a feasible solution in the dual (2.3.5) and it has an objective value $B \bullet \overline{Y} \geq p$. Using the weak duality (2.1.4), $B \bullet \overline{Y} \leq p$, and so, $B \bullet \overline{Y} = p$, *i.e.*, the dual achieves the optimum primal value. □

Theorem 2.3.4. *If the dual $(DSDP)$ from (2.3.5) is bounded and has a strictly feasible solution, then the primal and the dual optimal values are the same and the primal (SDP) from (2.3.4) reaches this optimum value. Recall (Prop. 2.1.5) that if $(DSDP)$ is unbounded, then (SDP) is infeasible.*

Proof. Apply Theorem 2.3.3 and Prop. 2.1.6. The main idea is to write the dual $(DSDP)$ from (2.3.5) in the primal form (this is possible using Prop. 2.1.2). Theorem 2.3.3 states that the dual of this primal form reaches the optimum solution. But the dual of this primal form is exactly equivalent to the primal (SDP) from (2.3.4) by virtue of Prop. 2.1.6. □

Theorem 2.3.5. *If both (SDP) and $(DSDP)$ are strictly feasible, then $OPT(SDP) = OPT(DSDP)$ and this value is reached by both programs.*

Proof. Using Prop. 2.1.1, if $(DSDP)$ has a feasible solution, then (SDP) is bounded. Using Prop. 2.1.5, if (SDP) has a feasible solution, $(DSDP)$ is bounded. We can now apply Theorems 2.3.3 and 2.3.4 to obtain the desired result. □

2.3.3 Further properties on the intersection of the SDP cone with a sub-space

Proposition 2.3.6. *Let $F(\mathbf{x}) = \sum_{i=1}^{n} x_i A_i - B$. If the image of F (i.e., the space spanned by A_i with $i \in [1..n]$ and basis $-B$) does not intersect the SDP cone S_m^+, then there exists $Y \succeq \mathbf{0}$ such that $F(\mathbf{x}) \bullet Y = -B \bullet Y < 0 \ \forall \mathbf{x} \in \mathbb{R}^n$.*

Proof.
$\Longleftarrow$
We know that $Y \succeq \mathbf{0}$ satisfies $F(\mathbf{x}) \bullet Y < 0$ for any $\mathbf{x} \in \mathbb{R}^n$. Since all $X \succeq \mathbf{0}$ verify $X \bullet Y \geq 0$, $F(\mathbf{x})$ can not belong to the SDP cone.
$\Longrightarrow$
There exists a sufficiently small ε such that (the image of) $F_\varepsilon(\mathbf{x}) = \sum_{i=1}^{n} x_i A_i - B + \varepsilon I_m$ still does not intersect the SDP cone. This means that this image does not intersect $\texttt{interior}(S_m^+)$ either and we can apply Prop 2.3.2 on F_ε. There exists a non-zero $Y \succeq \mathbf{0}$ such that $F_\varepsilon(\mathbf{x}) \bullet Y = (-B + \varepsilon I_m) \bullet Y \leq 0 \ \forall \mathbf{x} \in \mathbb{R}^n$. It is now enough to check that for any $\mathbf{x} \in \mathbb{R}^n$ the following value is constant: $F(\mathbf{x}) \bullet Y = F_\varepsilon(\mathbf{x}) \bullet Y - \varepsilon I_m \bullet Y < 0$; we used $I_m \bullet Y > 0$ which follows from $\texttt{trace}(Y) > 0$ (or apply Prop. 1.3.4). □

Proposition 2.3.7. *Let $F(\mathbf{x}) = \sum_{i=1}^{n} x_i A_i - B$ such that $\exists \mathbf{x}_0 \in \mathbb{R}^n$ such that $F(\mathbf{x}_0) = \mathbf{0}$. If the image of F intersects the SDP cone S_m^+ only in the origin $\mathbf{0}$, then there exists $Y \succ \mathbf{0}$ such that $F(\mathbf{x}) \bullet Y = 0 \ \forall \mathbf{x} \in \mathbb{R}^n$.*

Proof.
$\Longleftarrow$
We know that $Y \succ \mathbf{0}$ satisfies $F(\mathbf{x}) \bullet Y = 0$ for any $\mathbf{x} \in \mathbb{R}^n$. It is not hard to prove that all non-zero $X \in S_m^+$ satisfy $X \bullet Y > 0$, because Prop. 1.3.4 states that $X \bullet Y = 0 \implies XY = \mathbf{0}$, which leads to $X = \mathbf{0}Y^{-1} = \mathbf{0}$ (because $Y \succ \mathbf{0}$ is non-singular). Thus $F(\mathbf{x})$ can not cover any non-zero SDP matrix.
$\Longrightarrow$
Let $S_m^* = \{X \in S_m^+ : \ \texttt{trace}(X) = 1\}$. It is not hard to check that S_m^* is convex, closed and bounded. The image $\texttt{img}(F)$ of F is a closed convex set.

We show that $S_m^* - \texttt{img}(F) = \{X^a - X^b : \ X^a \in S_m^*, \ X^b \in \texttt{img}(F)\}$ is closed. We take any convergent sequence $\{X_i^a - X_i^b\}$ and we will show the limit point belongs to $S_m^* - \texttt{img}(F)$. It is not hard to see that any $X^a \in S_m^*$ is bounded in the sense that it satisfies $X_{ij}^a \leq 1 \ \forall i, j \in [1..m]$, because the 2×2 minor corresponding to rows/columns i and j has to be non-negative. The sequence $\{X_i^b\}$ (with $i \to \infty$) needs to be bounded, because otherwise $\{X_i^a - X_i^b\}$ would be unbounded, and so, non-convergent. Using the Bolzano-Weierstrass theorem (Theorem C.4.9), the bounded sequence $\{X_i^a\}$ has a convergent sub-sequence $\{X_{n_i}^a\}$. Using the Bolzano-Weierstrass theorem again, the sub-sequence $\{X_{n_i}^b\}$ contains a convergent sub-sub-sequence $\{X_{m_i}^b\}$, with $\{m_i\} \subseteq \{n_i\}$. Since S_m^* and $\texttt{img}(F)$ are closed, $\lim_{i \to \infty} X_{m_i}^a = X^a \in S_m^*$ and $\lim_{i \to \infty} X_{m_i}^b = X^b \in \texttt{img}(F)$, and so, $\lim_{i \to \infty} X_{m_i}^a - X_{m_i}^b = X^a - X^b \in S_m^* - \texttt{img}(F)$, *i.e.*, $S_m^* - \texttt{img}(F)$ contains all its limit points.

Since $\mathbf{0} \notin S_m^* - \texttt{img}(F)$, the simple separation Theorem C.4.5 states there is an Y such that $(X^a - X^b) \bullet Y > 0, \ \forall X^a \in S_m^*, \ X^b \in \texttt{img}(F)$. This is equivalent to

$$X^a \bullet Y > X^b \bullet Y, \ \forall X^a \in S_m^*, \ X^b \in \texttt{img}(F). \tag{2.3.8}$$

Since $\mathbf{0} \in \texttt{img}(F)$, we have $X^a \bullet Y > 0 \ \forall X^a \in S_m^*$. This is equivalent to $X^a \bullet Y > 0 \ \forall X \in S_m^+ \setminus \{\mathbf{0}\}$, and so, $Y \bullet (\mathbf{v}\mathbf{v}^\top) > 0 \ \forall \mathbf{v} \in \mathbb{R}^n - \{\mathbf{0}\}$, *i.e.*, $Y \succ \mathbf{0}$.

We still need to show that $F(\mathbf{x}) \bullet Y = 0\ \forall \mathbf{x} \in \mathbb{R}^n$. Assume there exists $\mathbf{x} \in \mathbb{R}^n$ such that $F(\mathbf{x}) \bullet Y = \Delta \neq 0$. Recall from hypothesis that there exists $\mathbf{x}_0 \in \mathbb{R}^n$ such that $F(\mathbf{x}_0) = \mathbf{0}$. We can write $\mathbf{x} = \mathbf{x}_0 + \mathbf{x}_1$ and we obtain $F(\mathbf{x}_0 + \mathbf{x}_1) = d$. For any $\alpha \in \mathbb{R}$, we have $F(\mathbf{x}_0 + \alpha\mathbf{x}_1) \bullet Y = \alpha\Delta$, and so, $F(\mathbf{x}_0 + \alpha\mathbf{x}_1) \bullet Y$ can be arbitrarily large, violating (2.3.8). $\square$

2.4 The difficulty of exactly solving (SDP) and algorithmic comments

We first notice that $\max\left\{y_{12} : \begin{bmatrix} 1 & y_{12} \\ y_{21} & 2 \end{bmatrix} \succeq \mathbf{0}\right\}$ is $\sqrt{2}$. This means it is rather unlikely that a purely numerical algorithm based on binary encodings can exactly optimize any SDP program. It is however possible to find the optimum value of any SDP program up to any specified additive error (precision) using ellipsoid, interior point, spectral or conic bundle methods.

We now show that exactly solving SDP is at least as hard as the square-root sum problem whose exact complexity is still an open problem. Consider the following feasibility program in variables $x_2,\ x_3, \dots x_n$.[12]

$$\begin{bmatrix} 1 & k-(x_2+x_3+\cdots+x_n) & & & & & & & \\ k-(x_2+x_3+\cdots+x_n) & a_1 & & & & & & & \\ & & 1 & x_2 & & & & & \\ & & x_2 & a_2 & & & & & \\ & & & & 1 & x_3 & & & \\ & & & & x_3 & a_3 & & & \\ & & & & & & \ddots & & \\ & & & & & & & 1 & x_n \\ & & & & & & & x_n & a_n \end{bmatrix} \succeq \mathbf{0}.$$

This program is feasible if and only if $k \leq \sum_{i=1}^n \sqrt{a_i}$. This is the square-root sum problem: given k and $a_1,\ a_2, \dots a_n$, decide if $k \leq \sum_{i=1}^n \sqrt{a_i}$. It is still an open question if this problem is polynomial or not, using at least the on-line forum http://www.openproblemgarden.org/op/complexity_of_square_root_sum. However, the same link reads that the SDP optimum can be determined (approximated) with any specified additive accuracy (error) ε using the interior point method or the ellipsoid algorithm in time polynomial in the size of the instance and $\log 1/\varepsilon$.

To show the difficulty of exactly solving SDP programs, we consider the following minimization problem.

$$\begin{aligned} &\min x \\ &\begin{bmatrix} x & 1 & 3 \\ 1 & x+2 & 0 \\ 3 & 0 & x+1 \end{bmatrix} \succeq \mathbf{0} \end{aligned}$$

The determinant of the whole matrix is $f(x) = x^3 + 3x^2 - 8x - 19$. We first present an intuitive figure (graph) of f on the right[13] and then we will formally determine the values of $x \geq 0$ such that $f(x) \geq 0$. We are not interested in values $x < 0$ because the leading principal minor of size 1 would be $x < 0$. The optimum of above program is clearly greater than 0.

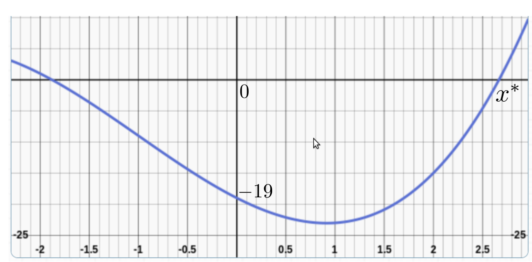


Figure 1: The graph of $f(x) = x^3 + 3x^2 - 8x - 19$

By calculating the second derivative $f''(x) = 6x + 6$, it is clear that f is strictly convex over $[0, \infty)$. Since $f'(0) = -8$, the function first decreases when going from 0 to ∞ and, after a point, it increases. This means f has a unique root $x^* \in [0, \infty)$ and the optimum of above program is at least x^*. Before determining this root, notice the figure intuitively shows that $x^* > 2.6$. We can also formally check that $f(2.6) = -1.944$, and, by convexity, we surely have $x^* > 2.6$. Other principal minors do not pose any problem because they are surely non-negative for any $x^* > 2.6$. The most problematic one is the one associated to rows and columns 1 and 3. Since $2.6 * 3.6 = 9.36 > 9$, this principal minor is positive for $x = 2.6$ and it remains so by increasing x.

[12] See the article "Semidefinite programming and combinatorial optimization" by Michel Goemans, in the *International Congress of Mathematicians*, Volume III, Documenta Mathematica, Extra Vol. ICM III, 1998, 657-666, page 2. The article is available on-line as of 2017 at https://www.math.uni-bielefeld.de/documenta/xvol-icm/17/Goemans.MAN.ps.gz.

[13] Obtained using http://derivative-calculator.net.

This means that the optimum value of above program is equal to the largest root x^* of the cubic equation $f(x)=0$. To determine this root, let us write $f(x)=(x+1)^3-11(x+1)-9$. Using Cardano's formula,[14] we obtain:

$$x^* = \sqrt[3]{\frac{9}{2}+\sqrt{\frac{3137}{108}}i}+\sqrt[3]{\frac{9}{2}-\sqrt{\frac{3137}{108}}i}-1 \;\approx\; 2.6679286.$$

It is highly unlikely that an SDP algorithm could exactly determine such values in polynomial time, solving cubic equations like $f(x)=0$ or possibly other higher degree equations. On the other hand, the numerical methods mentioned above can determine the optimum value up to any specified additive error.

3 Interesting SDP programs

3.1 An SDP program does not always reach its min (inf) or max (sup) value

The value

$$\min\left\{t:\ \begin{bmatrix} t & 1\\ 1 & t'\end{bmatrix}\succeq \mathbf{0}\right\}$$

is zero but *no* feasible solution reaches this optimum objective value of zero. However, we could use "min" to actually mean "inf". Since the program is bounded and strictly feasible (*e.g.*, for $t=t'=2$, we have $\left[\begin{smallmatrix}2&1\\1&2\end{smallmatrix}\right]\succ\mathbf{0}$) we can use the strong duality (Theorem 2.3.3) to conclude that the dual does achieve the optimum 0. One can easily check that the dual has only one feasible point $\left[\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right]$ with objective value 0.

An example of a program in the dual form (2.1.3a)-(2.1.3c) can be found by re-formulating the above primal program using the method from Section 2.1.3; aiming at a maximization form, one may find:

$$\max\left\{-Y_{11}:\ \begin{bmatrix} 0 & 1\\ 1 & 0\end{bmatrix}\bullet Y=2,\ Y\succeq\mathbf{0}\right\}.$$

3.2 The lowest and greatest eigenvalue using the SDP duality

Consider symmetric matrix $X\in\mathbb{R}^{m\times m}$ with eigenvalues $\lambda_1\leq\lambda_2\leq\cdots\leq\lambda_n$. The following program reports the smallest eigenvalue.

$$\lambda_1=\begin{cases}\max & t\\ \quad s.t & X-tI_m\succeq\mathbf{0}\\ & t\in\mathbb{R}\end{cases}$$

Observe this program is bounded and strictly feasible (take $t\to-\infty$), and so, the strong duality Theorem 2.3.3 states that the dual reaches the optimum value λ_1. By dualizing after transforming "$\max t$" into "$-\min -t$", we obtain an objective "$-\max -X\bullet Y$", that is equivalent to "$\min X\bullet Y$".

$$\lambda_1=\begin{cases}\min & X\bullet Y\\ \quad s.t & \mathtt{trace}(Y)=1\\ & Y\succeq\mathbf{0}\end{cases}$$

We have obtained a more general version of Lemma 1.2.2.1 in which the rank of Y is not necessarily 1 (as in Lemma 1.2.2.1). Furthermore, the lowest eigenvalue of $-X$ is $-\lambda_n$. The first above program for $-X$ can be written, after replacing t with $-t$, as $-\lambda_n=\max\{-t:\ tI_m-X\succeq\mathbf{0},\ t\in\mathbb{R}\}$. As such, λ_n can be obtained by the following primal-dual programs.

$$\lambda_n=\begin{cases}\min & t\\ s.t & tI_m-X\succeq\mathbf{0}\\ & t\in\mathbb{R}\end{cases}\qquad(3.2.1a)\qquad\qquad \lambda_n=\begin{cases}\max & X\bullet Y\\ s.t & \mathtt{trace}(Y)=1\\ & Y\succeq\mathbf{0}\end{cases}\qquad(3.2.1b)$$

[14] See the wikipedia article https://en.wikipedia.org/wiki/Cubic_function.

3.3 Change of variable in SDP programs

Let us focus on the dual SDP ($DSDP$) form from (2.1.3a)-(2.1.3c); recall its constraints have the form $A_i \bullet Y = c_i \ \forall i \in [1..n]$, where $Y \succeq \mathbf{0}$ are the variables. We will now change variables Y into variables $Z = QYQ^\top$ for a non-singular matrix Q. Recalling Prop. 1.2.3, such Z and Y are *congruent* and they satisfy $Z \succeq \mathbf{0} \iff Y \succeq \mathbf{0}$. Notice that if $Y = \mathbf{y}\mathbf{y}^\top$, then the variable change maps $\mathbf{y}\mathbf{y}^\top$ to $Q\mathbf{y}\mathbf{y}^\top Q^\top = (Q\mathbf{y})(Q\mathbf{y})^\top$, *i.e.*, we can also say vector $\mathbf{y}$ is mapped to $\mathbf{z} = Q\mathbf{y}$.

One could try to express the dual program ($DSDP$) in variables Z by simply replacing Y with $Q^{-1}ZQ^{-1^\top}$. But developing $A_i \bullet Y = A_i \bullet (Q^{-1}ZQ^{-1^\top})$ by brute force could be quite painful and lead to a mess. We can obtain a more elegant reformulation using the lemma below.

Lemma 3.3.0.1. *Given symmetric matrices $A, B \in \mathbb{R}^{m\times m}$ and non-singular matrix R, the following holds:*

$$A \bullet B = \left(R^{-1^\top} A R^{-1}\right) \bullet \left(RBR^\top\right)$$

Proof. Using the eigendecomposition (1.1.1) can write $A = \sum_{i=1}^m \lambda_i^a \mathbf{a}_i\mathbf{a}_i^\top$ and $B = \sum_{i=1}^m \lambda_i^b \mathbf{b}_i\mathbf{b}_i^\top$. Given that the scalar product is distributive, to prove the lemma, it is enough to show $\left(\mathbf{a}_i\mathbf{a}_i^\top\right) \bullet \left(\mathbf{b}_j\mathbf{b}_j^\top\right) = \left(R^{-1^\top}\mathbf{a}_i\mathbf{a}_i^\top R^{-1}\right) \bullet \left(R\mathbf{b}_j\mathbf{b}_j^\top R^\top\right) \ \forall i, j \in [1..n]$. We can calculate

$$\begin{aligned}
\left(R^{-1^\top}\mathbf{a}\mathbf{a}^\top R^{-1}\right) \bullet \left(R\mathbf{b}\mathbf{b}^\top R^\top\right) &= \left(\left(R^{-1^\top}\mathbf{a}\right)\left(R^{-1^\top}\mathbf{a}\right)^\top\right) \bullet \left((R\mathbf{b})(R\mathbf{b})^\top\right) \\
&= \left(\left(R^{-1^\top}\mathbf{a}\right) \bullet (R\mathbf{b})\right)^2 && \text{(we applied Lemma 1.3.3.1)} \\
&= \left(\mathbf{a}^\top R^{-1}R\mathbf{b}\right)^2 \\
&= \left(\mathbf{a}^\top\mathbf{b}\right)^2 \\
&= \left(\mathbf{a}\mathbf{a}^\top\right) \bullet \left(\mathbf{b}\mathbf{b}^\top\right). && \text{(we applied Lemma 1.3.3.1)}
\end{aligned}$$

which finishes the proof. □

Based on this lemma, we have $B \bullet Y = \left(Q^{-1^\top}BQ^{-1}\right) \bullet \left(QYQ^\top\right) = \left(Q^{-1^\top}YQ^{-1}\right) \bullet Z$. By applying the same calculations on $A_i \bullet X \ \forall i \in [1..n]$, we obtain that the dual ($DSDP$) program is equivalent to:

$$(DSDP_Z)\begin{cases}\max \ \left(Q^{-1^\top}BQ^{-1}\right) \bullet Z \\ \quad s.t \ \left(Q^{-1^\top}A_iQ^{-1}\right) \bullet Z = c_i \ \forall i \in [1..n] \\ \qquad Z \succeq \mathbf{0} \qquad\qquad \text{(recall } Z \succeq \mathbf{0} \iff Y \succeq \mathbf{0} \text{ since } Y \text{ and } Z \text{ are congruent)}\end{cases}$$

3.4 Convex quadratic programming is a particular case of SDP programming

We consider a convex quadratic program with n variables $x_1, x_2, \ldots x_n \in \mathbb{R}$:

$$\min \ \mathbf{x}^\top A_0\mathbf{x} + \mathbf{b}_0^\top\mathbf{x} \tag{3.4.1a}$$

$$s.t. \ \mathbf{x}^\top A_i\mathbf{x} + \mathbf{b}_i^\top\mathbf{x} \le c_i \ \forall i \in [1..p] \tag{3.4.1b}$$

Particularizing Prop. 1.8.1, a quadratic function $\mathbf{x}^\top A_i\mathbf{x}$ ($\forall i \in [0..p]$) is convex if and only if the Hessian $2A_i$ is SDP, *i.e.*, we need to have $A_i \succeq \mathbf{0} \ \forall i \in [0..p]$.

3.4.1 Reformulation using the (Cholesky) factorization of SDP matrices

Each A_i (with $i \in [0..p]$) can be factorized as $A_i = R_iR_i^\top$ using any of the presented decompositions of SDP matrices (*e.g.*, Cholesky, eigenvalue or square root, see Corollary 1.7.1). Let us move the objective function into the constraint set by introducing a real variable c_0 that has to be minimized such that $\mathbf{x}^\top A_0\mathbf{x}+\mathbf{b}_0^\top\mathbf{x} \le c_0$. After simple algebraic manipulations, the above program (3.4.1a)-(3.4.1b) can be written as:

$$\begin{aligned} \min \quad & c_0 \\ s.t. \quad & c_i - \mathbf{b}_i^\top \mathbf{x} - \mathbf{x}^\top R_i R_i^\top \mathbf{x} \ge 0 \ \forall i \in [0..p] \end{aligned}$$

Using the Schur complements Property 1.3.2, the above program is further equivalent to program below. Indeed, by applying the Schur complement, one can easily check that a feasible solution of above program is feasible in program below and vice-versa.

$$\begin{aligned} \min \quad & c_0 \\ s.t. \quad & \begin{bmatrix} I_n & R_i^\top \mathbf{x} \\ \mathbf{x}^\top R_i & c_i - \mathbf{b}_i^\top \mathbf{x} \end{bmatrix} \succeq \mathbf{0} \quad \forall i \in [0..p], \end{aligned}$$

which is a program in variables $x_1,\ x_2, \dots x_n$ and c_0. This is an SDP program with $p+1$ constraints (that could be expressed in an aggregated form (2.1.1a)-(2.1.1c) with a unique constraint, see also Prop. 2.1.4).

3.4.2 Reformulation by relaxing $\mathbf{x}\mathbf{x}^\top$ into $X \succeq \mathbf{x}\mathbf{x}^\top$

We will show that (3.4.1a)-(3.4.1b) is equivalent to the following SDP program:

$$\min \quad A_0 \bullet X + \mathbf{b}_0^\top \mathbf{x} \tag{3.4.2a}$$

$$s.t. \quad A_i \bullet X + \mathbf{b}_i^\top \mathbf{x} \le c_i \ \forall i \in [1..p] \tag{3.4.2b}$$

$$\begin{bmatrix} 1 & \mathbf{x}^\top \\ \mathbf{x} & X \end{bmatrix} \succeq \mathbf{0} \tag{3.4.2c}$$

First, it is clear that any feasible solution $\mathbf{x}$ of (3.4.1a)-(3.4.1b) can be associated to a feasible solution $X = \mathbf{x}\mathbf{x}^\top$ of (3.4.2a)-(3.4.2c) that has the same objective value. Conversely, it is possible to show that a feasible solution (3.4.2a)-(3.4.2c) corresponds to feasible solution of (3.4.1a)-(3.4.1b), but we still need a few lines. Any X and $\mathbf{x}$ that satisfy (3.4.2c) also satisfy $X \succeq \mathbf{x}\mathbf{x}^\top$ by virtue of the Schur complements Prop. 1.3.1. Thus, X can be written in the form $X = \mathbf{x}\mathbf{x}^\top + S$ for some $S \succeq \mathbf{0}$. Since $A_i \succeq \mathbf{0}\ \forall i \in [0..p]$, each product $A_i \bullet S$ is non-negative. This ensures the fact that $\mathbf{x}\mathbf{x}^\top$ satisfy all constraints (3.4.1b) and, similarly, the objective value (3.4.2a) satisfies $A_0 \bullet \left(\mathbf{x}\mathbf{x}^\top + S\right) + \mathbf{b}_0^\top \mathbf{x} \ge A_0 \bullet \mathbf{x}\mathbf{x}^\top + \mathbf{b}_0^\top \mathbf{x}$. In other words, $\mathbf{x}$ is a feasible solution of (3.4.1a)-(3.4.1b) and its objective value is no worse than that of X and $\mathbf{x}$ in (3.4.2a).

The above (3.4.2a)-(3.4.2c) is actually an SDP program that can be easily written in the standard form (2.1.1a)-(2.1.1c). Both (3.4.2b) and (3.4.2c) are linear matrix inequalities like (2.1.1b) in variables $x_1, x_2, \dots, x_n$ and $X_{11}, X_{12}, \dots X_{nn}$; (3.4.2b) uses "$1 \times 1$ matrices" and (3.4.2c) uses $(n+1) \times (n+1)$ matrices. We thus have two inequalities of the form (2.1.1b) that can be aggregated into a unique constraint (2.1.1b) that uses block-diagonal matrices (Prop. 2.1.4). As a final side remark, similar programs may even have constraints like $\begin{bmatrix} -y & \frac{1}{2}\mathbf{b}_0^\top \\ \frac{1}{2}\mathbf{b}_0 & A_0 \end{bmatrix} \bullet \begin{bmatrix} 1 & \mathbf{x}^\top \\ \mathbf{x} & X \end{bmatrix} \le 0$ that contain variables in both factors, because of variable y in the left-hand term. However, after developing the scalar product, the variable y simply arises as a term in a final sum as in (2.1.1b).

3.4.3 Unconstrained quadratic programming reduces to SDP programming

We consider an *unconstrained* quadratic program :

$$\min \quad \mathbf{x}^\top A\mathbf{x} + \mathbf{b}^\top \mathbf{x} \tag{3.4.3}$$

If A is not SDP, we can show that this program is unbounded from below. For this, let us take any eigenvector $\mathbf{v}$ such that $A\mathbf{v} = -\alpha\mathbf{v}$ with $\alpha > 0$. Consider the function $f(t) = (t\mathbf{v}^\top)A(t\mathbf{v}) + \mathbf{b}^\top\mathbf{v} = -\alpha t^2 \mathbf{v}^\top\mathbf{v} + t\mathbf{b}^\top\mathbf{v}$. This is clearly a strictly concave function in t and it goes to $-\infty$ when t goes to ∞.

To solve (3.4.3) by SDP programming, one first has to check if A is SDP or not. The necessity of this step makes the approach from Section 3.4.1 unusable here, because we here need a method that handles at the same time both the case $A \not\succeq \mathbf{0}$ and $A \succeq \mathbf{0}$.

Let us now show how to address the case $A \succeq \mathbf{0}$ by SDP programming. We first need the following result which should not be taken from granted (although it is omitted from certain lecture notes).

Proposition 3.4.1. *If an unconstrained quadratic program $\mathbf{x}^\top A\mathbf{x} + \mathbf{b}^\top\mathbf{x}$ (i.e., a polynomial of degree 2) is bounded from below, there is a solution $\overline{\mathbf{x}}$ that does reach the minimum value. The polynomial is convex ($A \succeq \mathbf{0}$) and the gradient in $\overline{\mathbf{x}}$ is $\mathbf{0}$.*

This does ***not hold for polynomials of any degree or for convex functions****. Indeed, $p(x_1, x_2) = (1 - x_1x_2)^2 + x_1^2$ does not reach its minimum (infimum) $0 = \lim_{m\to\infty} p(\frac{1}{m}, m)$. Function $h(x) = e^x$ is infinitely differentiable, convex and bounded from below, but there is no x with $h'(x) = 0$ that does reach the minimum value. On the other hand, it is actually possible to prove than a convex polynomial (of any degree) always reaches its minimum value, but the proof of such result is outside the scope of this document.*[15]

Proof. We showed above that the polynomial is bounded only if $A \succeq \mathbf{0}$.

Let us show that if there is *no* $\overline{\mathbf{x}} \in \mathbb{R}^n$ that cancels the gradient, the program is *unbounded*. The gradient can be written $\nabla\left(\mathbf{x}^\top A\mathbf{x} + \mathbf{b}^\top\mathbf{x}\right) = 2A\mathbf{x} + \mathbf{b}$, and so, there is *no* $\overline{\mathbf{x}} \in \mathbb{R}^n$ such that $2A\overline{\mathbf{x}} = -\mathbf{b}$. This means that $-\mathbf{b}$ (or $\mathbf{b}$) does not belong to the image (set of linear combinations of the columns) of $2A$. Using a similar (transposed) argument as in the first paragraph of the first proof of Prop. 1.6.4, $-\mathbf{b}$ needs to have the form $-\mathbf{b} = -\mathbf{b}_{\mathtt{img}} - \mathbf{b}_0$, where $-\mathbf{b}_{\mathtt{img}} \in \mathtt{img}(2A)$ and $-\mathbf{b}_0 \in \mathtt{null}(2A)$ with $\mathbf{b}_0 \neq \mathbf{0}$, where $\mathtt{img}(2A)$ is the image of $2A$ and $\mathtt{null}(2A)$ is the null space of A (see the null space definition in (A.1.2)).

Taking $\mathbf{x} = t\mathbf{b}_0$, we can now define $h(t) = (t\mathbf{b}_0)^\top A(t\mathbf{b}_0) + (\mathbf{b}_{\mathtt{img}} + \mathbf{b}_0)^\top(t\mathbf{b}_0) = 0 + t\mathbf{b}_0^\top\mathbf{b}_0 = t|\mathbf{b}_0|^2$, where we used $\mathbf{b}_0 \in \mathtt{null}(A) \implies A\mathbf{b}_0 = \mathbf{0}$. Since $\mathbf{b}_0 \neq \mathbf{0}$, we have $|\mathbf{b}_0| \neq 0$, and so, $h(t)$ is a non-zero linear function that is clearly unbounded from below. Since $\mathbf{x} = t\mathbf{b}_0$ is a perfectly feasible solution of our unconstrained program, this program is unbounded from below. We proved that if there is *no* stationary point $\overline{\mathbf{x}} \in \mathbb{R}^n$ that cancels the gradient, the program is *unbounded*. This means that a bounded convex quadratic program needs to have some stationary point $\overline{\mathbf{x}} \in \mathbb{R}^n$ such that $\nabla\left(\mathbf{x}^\top A\mathbf{x} + \mathbf{b}^\top\mathbf{x}\right)_{\overline{\mathbf{x}}} = 2A\overline{\mathbf{x}} + \mathbf{b} = \mathbf{0}$. For a convex function, the stationary point needs to be the global minimum, and so, $\overline{\mathbf{x}}$ reaches the minimum value of $\mathbf{x}^\top A\mathbf{x} + \mathbf{b}^\top\mathbf{x}$. □

We hereafter consider (3.4.3) is convex ($A \succeq \mathbf{0}$).

Proposition 3.4.2. *The optimum value of unconstrained convex quadratic program (3.4.3) is equal to:*

$$t^* = \max \quad t \tag{3.4.4a}$$

$$s.t. \begin{bmatrix} -t & \frac{1}{2}\mathbf{b}^\top \\ \frac{1}{2}\mathbf{b} & A \end{bmatrix} \succeq \mathbf{0}, \tag{3.4.4b}$$

where we use the convention $t^* = -\infty$ if (3.4.4b) is infeasible for any $t \in \mathbb{R}$, this case being equivalent to the fact that (3.4.3) is unbounded from below.

Proof. It is enough to prove that: (i) if (3.4.3) is unbounded, then (3.4.4a)-(3.4.4b) is infeasible and (ii) if (3.4.3) is bounded then its optimum value is equal to t^* in (3.4.4a)-(3.4.4b).

Based on Prop 3.4.1, the above case (i) can only arise if either A is not SDP so that (3.4.4b) is clearly infeasible, or if (3.4.3) has no stationary point. If there is no stationary point, then there is no $\overline{\mathbf{x}}$ such that $2A\overline{\mathbf{x}} + \mathbf{b} = \mathbf{0}$. This means that $\frac{1}{2}\mathbf{b}^\top$ can not be written as a linear combination of (the rows of) A. Using Prop. 1.6.4, we obtain that (3.4.4b) is infeasible

In the non-degenerate case (ii), Prop. 3.4.1 guarantees that the quadratic function has a stationary point $\overline{\mathbf{x}}$ so that $2A\overline{\mathbf{x}} = -\mathbf{b}$. This stationary point minimizes (3.4.3) so that the optimum value is $\overline{\mathbf{x}}^\top A\overline{\mathbf{x}} + \mathbf{b}^\top\overline{\mathbf{x}} = -\overline{\mathbf{x}}^\top\frac{1}{2}\mathbf{b} + \mathbf{b}^\top\overline{\mathbf{x}} = \frac{1}{2}\overline{\mathbf{x}}^\top\mathbf{b}$. We will prove that $t^* = \frac{1}{2}\overline{\mathbf{x}}^\top\mathbf{b}$.

Using Prop. 1.2.4, the SDP status of $\begin{bmatrix} -t & \frac{1}{2}\mathbf{b}^\top \\ \frac{1}{2}\mathbf{b} & A \end{bmatrix}$ does not change if we add to the first row a linear combination of the other rows. If we add $\overline{\mathbf{x}}^\top\left[\frac{1}{2}\mathbf{b}\ A\right]$ to the first row followed by the same (transposed) operation on the columns, we obtain the matrix $\begin{bmatrix} -t+\frac{1}{2}\overline{\mathbf{x}}^\top\mathbf{b} & \mathbf{0} \\ \mathbf{0} & A \end{bmatrix}$. This latter matrix is SDP only if $-t+\frac{1}{2}\overline{\mathbf{x}}^\top\mathbf{b} \geq 0$, which means that the optimum t^* value is $t^* = \frac{1}{2}\overline{\mathbf{x}}^\top\mathbf{b}$, which is also the optimum value of the unconstrained quadratic function. □

[15] For a proof, see the response of R. Israel on my question asked on the on-line forum https://math.stackexchange.com/questions/2341163/does-a-convex-polynomial-always-reaches-its-minimum-value/. The example $(1 - x_1x_2)^2 + x_1^2$ is taken from a response of J.P. McCarthy on https://math.stackexchange.com/questions/279497/polynomial-px-y-with-inf-mathbbr2-p-0-but-without-any-point-where.

Finally, it is interesting that the linear matrix inequality (3.4.4b) is equivalent to the following system of 2^n linear inequalities: $\det\begin{pmatrix} -t & \frac{1}{2}\mathbf{b}_J^\top \\ \frac{1}{2}\mathbf{b}_J & A_J \end{pmatrix} \geq 0 \ \forall J \subseteq [1..n]$, where $\mathbf{b}_J$ and resp. A_J are obtained from $\mathbf{b}$ and resp. A by selecting rows or columns J. When $A \succeq \mathbf{0}$, (3.4.4a)-(3.4.4b) can actually be written as a linear program.

3.5 An LP with equality constraints as an SDP program in the dual form

Consider the Linear Program (LP):

$$\begin{aligned} &\max\ b_1x_1 + b_2x_2 + \ldots b_nx_n \\ &\ s.t.\ c_0^j + c_1^jx_1 + c_2^jx_2 + \ldots c_n^jx_n = 0\ \forall j \in [1..m] \end{aligned}$$

There are two ways of converting this LP into an SDP. The direct method consists of replacing each equality with two inequalities and of applying Prop. 2.1.4 on the resulting system of inequalities. We obtain an aggregated SDP programs expressed with aggregated diagonal matrices.

The second method uses the property $\left(\mathbf{c}^j{\mathbf{c}^j}^\top\right) \bullet \left(\begin{bmatrix}1\\ \mathbf{x}\end{bmatrix}[1\ \mathbf{x}^\top]\right) = \left(\mathbf{c}^j \bullet \begin{bmatrix}1\\ \mathbf{x}\end{bmatrix}\right)^2$ proved in Lemma 1.3.3.1, where $\mathbf{c}^j = [c_0^j\ c_1^j\ c_2^j \ldots c_n^j]^\top\ \forall j \in [1..m]$. We can thus write:

$$\mathbf{c}^j \bullet \begin{bmatrix}1\\ \mathbf{x}\end{bmatrix} = 0 \iff \left(\mathbf{c}^j{\mathbf{c}^j}^\top\right) \bullet \begin{bmatrix}1 & \mathbf{x}^\top\\ \mathbf{x} & \mathbf{x}\mathbf{x}^\top\end{bmatrix} = 0. \tag{3.5.1}$$

Thus, any feasible solution $\mathbf{x}$ of the above LP can be directly converted to a feasible solution of the SDP program below by simply setting $X = \mathbf{x}\mathbf{x}^\top$ and performing a sum over all $j \in [1..m]$.

$$\begin{aligned} &\max\ b_1x_1 + b_2x_2 + \ldots b_nx_n \\ &\ s.t.\ \left(\sum_{j=1}^m \mathbf{c}^j{\mathbf{c}^j}^\top\right) \bullet \begin{bmatrix}1 & \mathbf{x}^\top\\ \mathbf{x} & X\end{bmatrix} = 0 \\ &\qquad \begin{bmatrix}1 & \mathbf{x}^\top\\ \mathbf{x} & X\end{bmatrix} \succeq \mathbf{0} \end{aligned}$$

The last point to prove is that any feasible solution or $\left[\begin{smallmatrix}1 & \mathbf{x}^\top\\ \mathbf{x} & X\end{smallmatrix}\right]$ of the above SDP program can be associated to a feasible solution $\mathbf{x}$ of the LP. First, we notice that

$$\left(\sum_{j=1}^m \mathbf{c}^j{\mathbf{c}^j}^\top\right) \bullet \begin{bmatrix}1 & \mathbf{x}^\top\\ \mathbf{x} & X\end{bmatrix} = 0 \implies \left(\mathbf{c}^j{\mathbf{c}^j}^\top\right) \bullet \begin{bmatrix}1 & \mathbf{x}^\top\\ \mathbf{x} & X\end{bmatrix} = 0\ \forall j \in [1..m],$$

because for each $j \in [1..m]$ we have a non-negative product of two SDP matrices in the left-hand side sum, *i.e.*, $\left(\mathbf{c}^j{\mathbf{c}^j}^\top\right) \bullet \left[\begin{smallmatrix}1 & \mathbf{x}^\top\\ \mathbf{x} & X\end{smallmatrix}\right] \geq 0\ \forall j \in [1..m]$. We can now use that $\left[\begin{smallmatrix}1 & \mathbf{x}^\top\\ \mathbf{x} & X\end{smallmatrix}\right] \succeq \mathbf{0}$ is equivalent to $X \succeq \mathbf{x}\mathbf{x}^\top$ so as to obtain

$$\begin{aligned} \left(\mathbf{c}^j{\mathbf{c}^j}^\top\right) \bullet \begin{bmatrix}1 & \mathbf{x}^\top\\ \mathbf{x} & X\end{bmatrix} = 0 &\implies \left(\mathbf{c}^j{\mathbf{c}^j}^\top\right) \bullet \begin{bmatrix}0 & \mathbf{0}\\ \mathbf{0} & X - \mathbf{x}\mathbf{x}^\top\end{bmatrix} + \left(\mathbf{c}^j{\mathbf{c}^j}^\top\right) \bullet \begin{bmatrix}1 & \mathbf{x}^\top\\ \mathbf{x} & \mathbf{x}\mathbf{x}^\top\end{bmatrix} = 0 \\ &\implies \left(\mathbf{c}^j{\mathbf{c}^j}^\top\right) \bullet \begin{bmatrix}1 & \mathbf{x}^\top\\ \mathbf{x} & \mathbf{x}\mathbf{x}^\top\end{bmatrix} = 0, \end{aligned}$$

where we used the fact that all involved matrices are SDP and that their scalar products are non-negative. We can now apply (3.5.1) to obtain the implication below, which confirms $\mathbf{x}$ is feasible in the initial LP.

$$\left(\mathbf{c}^j{\mathbf{c}^j}^\top\right) \bullet \begin{bmatrix}1 & \mathbf{x}^\top\\ \mathbf{x} & \mathbf{x}\mathbf{x}^\top\end{bmatrix} = 0 \implies \mathbf{c}^j \bullet \begin{bmatrix}1\\ \mathbf{x}\end{bmatrix} = 0\ \forall j \in [1..m].$$

PART 2: MORE ADVANCED SDP PROGRAMMING

4 Six equivalent formulations of the Lovász theta number $\vartheta(G)$

4.1 A first SDP formulation of the theta number

4.1.1 The primal form (ϑ_G)

We consider a graph $G = ([1..n], E)$. We introduce the Lovász theta number using the following program based on SDP matrix $\overline{Z} \in \mathbb{R}^{n\times n}$.

$$(\vartheta_G)\begin{cases} \min\ \ t & (4.1.1a)\\ \ \ s.t.\ \ \overline{z}_{ii} = t-1 \ \ \forall i \in [1..n] & (4.1.1b)\\ \qquad\ \ \overline{z}_{ij} = -1 \qquad \forall \{i,j\} \notin E & (4.1.1c)\\ \qquad\ \ \overline{Z} \succeq \mathbf{0}. & (4.1.1d)\end{cases}$$

We use the notational convention $\vartheta(G) = OPT(\vartheta_G)$. We consider above (ϑ_G) as a primal SDP program of the form (2.1.1a)-(2.1.1c) in which the variables are $t \in \mathbb{R}^n$ and $\overline{z}_{ij} \in \mathbb{R}^n\ \forall\{i,j\} \in E$. The matrix B from (2.1.1a)-(2.1.1c) contains ones on the diagonal and on all positions (i,j) corresponding to $\{i,j\} \notin E$. As a side remark, using (3.2.1a), we observe that (ϑ_G) returns the maximum eigenvalue of the matrix B described above.

Theorem 4.1.1. *($\vartheta(G) \geq \alpha(G)$) The Lovász theta number $\vartheta(G)$ is greater than or equal to the maximum stable $\alpha(G)$ of G.*

Proof. Let us consider the maximum stable $J \in [1..n]$ with $|J| = \alpha(G)$. If we restrict $\overline{Z}$ from (4.1.1a)-(4.1.1d) to its minor corresponding to rows J and columns J, we obtain $tI_{\alpha(G)} - \mathbb{1} \succeq \mathbf{0}$, where $\mathbb{1}$ is a matrix in which all elements are equal to 1. The fact that $t \geq \alpha(G)$ follows from the next lemma (that needs to hold for $n = \alpha(G)$). $\square$

Lemma 4.1.1.1. *The lowest $z \in \mathbb{R}$ for which the $n \times n$ matrix below is SDP is $z = n - 1$.*

$$A_n(z) = \underbrace{\begin{bmatrix} z & -1 & -1 & \dots & -1\\ -1 & z & -1 & \dots & -1\\ -1 & -1 & z & \dots & -1\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ -1 & -1 & -1 & \dots & z\end{bmatrix}}_{n} \succeq \mathbf{0}$$

We provide four proofs so as to master multiple proof techniques and explore a variety of SDP tools presented in this manuscript.

Proof 1. We can write $A_n(z) = (z+1) - \mathbb{1}$, where $\mathbb{1}$ is a matrix filled with $n \times n$ ones. Using (3.2.1a), the minimum value $(z+1)$ such that $(z+1) - \mathbb{1} \succeq \mathbf{0}$ is the maximum eigenvalue of $\mathbb{1}$. But what are the eigenvalues and eigenvectors of $\mathbb{1}$? Since all positions of $\mathbb{1}\mathbf{v}$ must be equal for any eigenvector $\mathbf{v} \in \mathbb{R}^n$, we obtain that either (i) the eigenvalue λ_v of v is zero or (ii) all elements of $\mathbf{v}$ are equal so that $\lambda_v = n$. The sought value $z + 1$ is thus n, leading to $z = n - 1$. $\square$.

Proof 2. It is easy to see that the Frobenius norm of $A_n(0)$ is $|A_n(0)| = \sqrt{\sum_{i,j=1}^{n} A_n(0)_{ij}} = \sqrt{(n-1)^2} = n-1$.

Using Prop. 1.4.1, the minimum eigenvalue of $A_n(0)$ is at least $-|A_n(0)| = -(n-1)$. This is enough to guarantee that $A_n(n-1) = (n-1)I_n + A_n(0) \succeq \mathbf{0}$. We show $A_n(z) \not\succeq \mathbf{0}$ for any $z < n-1$ by noticing $[1\ 1\ 1 \dots 1]A_n(z)[1\ 1\ 1 \dots 1]^\top < 0$ for any $z < n-1$. This proves that $z = n-1$ is the lowest z such that $A_n(z) \succeq \mathbf{0}$. $\square$

Proof 3. We apply the Gershgorin circle Theorem A.2.8 to show $A_n(n-1) \succeq \mathbf{0}$. The theorem states that any eigenvalue λ of $A_n(n-1)$ needs to satisfy $|\lambda - A_n(n-1)_{ii}| \leq \sum_{\substack{j=1\\ j\neq i}}^{n} |A_n(n-1)_{ij}|$ for some $i \in [1..n]$. However, for any $i \in [1..n]$ this reduces to $|\lambda - (n-1)| \leq n-1$, which means $\lambda \geq 0$, leading to $A_n(n-1) \succeq \mathbf{0}$. To show $A_n(z) \not\succeq \mathbf{0}$ for any $z < n-1$, use the method from the last two sentences of above *Proof 2.* $\square$

Proof 4. We proceed by induction. The lemma is clearly true for $n = 1$. For any $n \geq 2$, we need to have $z > 0$ because $z = 0$ would make negative any 2×2 minor (use Prop 1.2.5). We apply the Schur complement from Prop 1.3.2. Using the notations A, B and C from Prop 1.3.2, we write $A_n(z) = \left[\begin{smallmatrix} A & B^\top \\ B & C \end{smallmatrix}\right]$ with $A = z$ (which satisfies condition $A \succ \mathbf{0}$), $B^\top = -\mathbb{1}_{n-1}^\top = \underbrace{[-1\ \ -1\ \ \cdots -1]}_{n-1 \text{ positions}}$, and $C = A_{n-1}(z)$, *i.e.*, $A_{n-1}(z)$ is the $(n-1)\times(n-1)$ bottom-right minor. Using the Schur complement, we obtain $A_n(z) \succeq \mathbf{0} \iff C - BA^{-1}B^\top \succeq \mathbf{0} \iff A_{n-1}(z) - \frac{1}{z}\mathbb{1}_{n-1}\mathbb{1}_{n-1}^\top \succeq \mathbf{0}$. This last matrix inequality boils down to:

$$\underbrace{\begin{bmatrix} z-\frac{1}{z} & -1-\frac{1}{z} & \dots & -1-\frac{1}{z} \\ -1-\frac{1}{z} & z-\frac{1}{z} & \dots & -1-\frac{1}{z} \\ \vdots & \vdots & \ddots & \vdots \\ -1-\frac{1}{z} & -1-\frac{1}{z} & \dots & z-\frac{1}{z} \end{bmatrix}}_{n-1} \succeq \mathbf{0} \iff \underbrace{\begin{bmatrix} \frac{(z+1)(z-1)}{z} & -\frac{z+1}{z} & \dots & -\frac{z+1}{z} \\ -\frac{z+1}{z} & \frac{(z+1)(z-1)}{z} & \dots & -\frac{z+1}{z} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{z+1}{z} & -\frac{z+1}{z} & \dots & \frac{(z+1)(z-1)}{z} \end{bmatrix}}_{n-1} \succeq \mathbf{0}$$

$$\iff \underbrace{\begin{bmatrix} z-1 & -1 & \dots & -1 \\ -1 & z-1 & \dots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \dots & z-1 \end{bmatrix}}_{n-1} \succeq \mathbf{0}$$

We have obtained that $A_n(z) \succeq \mathbf{0} \iff A_{n-1}(z-1) \succeq \mathbf{0}$. We can use the induction hypothesis: the lowest $z-1$ such that $A_{n-1}(z-1) \succeq \mathbf{0}$ is $z - 1 = n-2$, and so, $z = n-1$ is the lowest z such that $A_n(z) \succeq \mathbf{0}$. $\square$

Theorem 4.1.2. *($\vartheta(G) \leq \chi(\overline{G})$) The Lovász theta number $\vartheta(G)$ is less than or equal to the clique cover number of G. This clique cover number is the chromatic number of the complementary graph $\overline{G}$.*

Proof. Consider a partition $(C_1, C_2, \ldots C_k)$ of the vertex set $[1..n]$ such that each C_i (with $i \in [1..k]$) is a clique. Applying Lemma 4.1.1.1 above, we have

$$\underbrace{\begin{bmatrix} k-1 & -1 & -1 & \dots & -1 \\ -1 & k-1 & -1 & \dots & -1 \\ -1 & -1 & k-1 & \dots & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & -1 & \dots & k-1 \end{bmatrix}}_{k} \succeq \mathbf{0}$$

This means (use the Cholesky factorisation from Prop 1.6.7 or other decompositions from Corollary 1.7.1) there exist k-dimensional vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_k$ such that $\mathbf{v}_i \cdot \mathbf{v}_i = k-1$ and $\mathbf{v}_i \cdot \mathbf{v}_j = -1\ \forall i, j \in [1..k], i \neq j$.

Since $\{C_1, C_2, \ldots C_k\}$ cover the vertex set $[1..n]$, all vertices $u \in [1..n]$ can be associated to vector $\mathbf{v}^u = \mathbf{v}_{\ell(u)}$, where $\ell(u)$ is the clique that contains u (we have $u \in C_{\ell(u)}$). We actually associate the same vector to all vertices of a clique.

We now define the matrix $\overline{Z}$ by setting $\overline{z}_{ij} = \mathbf{v}^i \cdot \mathbf{v}^j$. This is a Gram matrix that is clearly SDP (use Prop. A.1.8). Notice that: (i) $\overline{z}_{ii} = k-1\ \forall i \in [1..n]$ and (ii) if $i, j \in [1..n]$ do not belong to the same clique, then $\overline{z}_{ij} = -1$, and so, $\overline{z}_{ij} = -1\ \forall \{i, j\} \notin E$. Properties (i) and (ii) are enough to ensure that $\overline{Z}$ can be written in the form (4.1.1b)-(4.1.1c) with $t = k$. We have constructed a feasible solution of (ϑ_G) from (4.1.1a)-(4.1.1d) with objective value k. Taking $k = \chi(\overline{G})$, we have $OPT(\vartheta_G) \leq \chi(\overline{G})$.

This proof is inspired from a related result of a lecture note of Anupam Gupta.[16] $\square$

[16] See (11.3) of the document available, as of 2017, at http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15859-f11/www/notes/lecture11.pdf.

Using $\alpha(G) = \omega(\overline{G})$ (where ω denotes the maximum clique size), we also obtain the "sandwich" property as a direct consequence of Theorem 4.1.1 and Theorem 4.1.2.

$$\omega(\overline{G}) \leq \vartheta(G) \leq \chi(\overline{G}) \tag{4.1.2}$$

The following corollary of Lemma 4.1.1.1 can be generally useful, but we do not use it in this document.

Corollary 4.1.3. *When $n \geq 2$ unitary vectors are as spread out as much as possible, the dot product of any pair of them is $\frac{-1}{n-1}$.*

Proof. By "spread out as much as possible", we want $t = \max\{\mathbf{v}_i \bullet \mathbf{v}_j : \; i,j \in [1..n], \; i \neq j\}$ to be as low as possible. Lemma 4.1.1.1 states that $nI_n - \mathbb{1} \succeq \mathbf{0}$, where $\mathbb{1}$ is a matrix with all elements equal to 1. By dividing all terms by $\frac{1}{n-1}$, we obtain a matrix $T \succeq \mathbf{0}$ with 1 on the diagonal and with $\frac{-1}{n-1}$ on all non-diagonal positions. This means that $t \leq \frac{-1}{n-1}$. This t value is optimal. If we decrease any non-diagonal term(s) of T, we obtain a matrix $T' \not\succeq \mathbf{0}$, since $[1\ 1 \dots 1]T'[1\ 1 \dots 1]^\top < 0$. $\square$

4.1.2 The dual form of (ϑ_G)

We now introduce the dual program $(D\vartheta_G)$ of (ϑ_G) from (4.1.1a)-(4.1.1d). Let $D\vartheta(G) = OPT(D\vartheta_G)$.

$$(D\vartheta_G) \begin{cases} \max \quad \displaystyle\sum_{i=1}^{n} Y_{ii} + \sum_{\{i,j\} \notin E} Y_{ij} = \mathbb{1} \bullet Y & (4.1.3a) \\ s.t. \quad \mathtt{trace}(Y) = 1 & (4.1.3b) \\ \qquad Y_{ij} = 0 \quad \forall \{i,j\} \in E & (4.1.3c) \\ \qquad Y \succeq \mathbf{0}, & (4.1.3d) \end{cases}$$

where (4.1.3a) simplifies to $\mathbb{1} \bullet Y$ because $Y_{ij} = 0 \ \forall \{i,j\} \in E$ by virtue of (4.1.3c). Notice that both (ϑ_G) and $(D\vartheta_G)$ are strictly feasible (take a sufficiently large t in (ϑ_G) and $Y = \frac{1}{n} I_n$ in $(D\vartheta_G)$), and so, we can apply the strong duality Theorem 2.3.5 to state that the optimum value is reached by both programs.

Proposition 4.1.4. *The "sandwich" property $\alpha(G) \leq \vartheta(G) \leq \chi(\overline{G})$ obtained in (4.1.2) in the previous Section 4.1.1 can also be proved using the dual $(D\vartheta_G)$.*

Proof of $\alpha(G) \leq \vartheta(G)$
Take the largest stable S of G. Construct matrix Y^S such that $Y^S_{ij} = \frac{1}{|S|} \ \forall i,j \in S$ and $Y^S_{ij} = 0$ otherwise. It is not hard to check that Y^S is a feasible solution of $(D\vartheta_G)$ in (4.1.3a)-(4.1.3d) with objective value $|S|$. This is enough to state $\alpha(G) \leq D\vartheta(G) = \vartheta(G)$.
Proof of $\vartheta(G) \leq \chi(\overline{G})$
Consider a clique cover $\{C_1, \ C_2, \dots C_k\}$ of G. Without loss of generality, we can re-order $[1..n]$ such that all C_i (with $i \in [1..k]$) represent segments of $[1..n]$, *i.e.*, C_1 contains the first $|C_1|$ elements $\big[1..|C_1|\big]$, $C_2 = \big[|C_1| + 1..|C_1| + |C_2|\big]$, etc. We can decompose Y of $(D\vartheta(G))$ from (4.1.3a)-(4.1.3d) into k^2 blocks:

$$Y = \begin{bmatrix} \mathcal{Y}_{11} & \mathcal{Y}_{12} & \dots & \mathcal{Y}_{1k} \\ \mathcal{Y}_{21} & \mathcal{Y}_{22} & \dots & \mathcal{Y}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{Y}_{k1} & \mathcal{Y}_{k2} & \dots & \mathcal{Y}_{kk} \end{bmatrix},$$

where block $\mathcal{Y}_{ij}$ has size $|C_i| \times |C_j|$ for any $i,j \in [1..k]$. Notice that $\mathcal{Y}_{ii}$ is diagonal because C_i is a clique ($\forall i \in [1..k]$). This means $\sum_{i=1}^{k} \mathbb{1} \bullet \mathcal{Y}_{ii} = 1$, based on $\mathtt{trace}(Y) = 1$. We will show that

$$\mathbb{1} \bullet Y = \sum_{i=1}^{k} \sum_{j=1}^{k} \mathbb{1} \bullet \mathcal{Y}_{ij} \leq k. \tag{4.1.4}$$

The principal minor of Y associated to rows and columns C_i and C_j needs to be SDP. We thus obtain $\begin{bmatrix} \mathcal{Y}_{ii} & \mathcal{Y}_{ij} \\ \mathcal{Y}_{ji} & \mathcal{Y}_{jj} \end{bmatrix} \succeq \mathbf{0}$. Take $\mathbf{x}^\top = [\underbrace{1\ 1 \dots 1}_{|C_i|} \ \underbrace{-1\ -1\ \cdots -1}_{|C_j|}]$. Using $\begin{bmatrix} \mathcal{Y}_{ii} & \mathcal{Y}_{ij} \\ \mathcal{Y}_{ji} & \mathcal{Y}_{jj} \end{bmatrix} \bullet \mathbf{x}\mathbf{x}^\top \geq 0$, we obtain:

$$\mathbb{1} \bullet \mathcal{Y}_{ij} \leq \frac{\mathbb{1} \bullet \mathcal{Y}_i + \mathbb{1} \bullet \mathcal{Y}_j}{2}, \ \forall i,j \in [1..k]$$

Notice this holds (with equality) for terms $i = j \in [1..k]$. We can simply now obtain (4.1.4) by applying

$$\sum_{i=1}^{k}\sum_{j=1}^{k} \mathbb{1} \bullet \mathcal{Y}_{ij} \le \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{\mathbb{1} \bullet \mathcal{Y}_i + \mathbb{1} \bullet \mathcal{Y}_j}{2} = \frac{2k\sum_{i=1}^{k} \mathbb{1} \bullet \mathcal{Y}_i}{2} = k.$$

4.2 A second SDP formulation (ϑ'_G) of the theta number

We start from the following well-known linear program for the maximum stable $\alpha(G)$.

$$(\alpha_G)\begin{cases} \max & \sum_{i=1}^{n} y_i \\ s.t. & y_i + y_j \le 1 \ \forall \{i,j\} \in E \\ & \mathbf{y} \in \{0,1\}^n \end{cases}$$

We now introduce a first SDP relaxation of above (α_G).

$$(\vartheta'_G)\begin{cases} \max & \sum_{i=1}^{n} y_{0i} & \text{(4.2.2a)} \\ s.t. & y_{00} = 1 & \text{(4.2.2b)} \\ & y_{ij} = 0 \quad \forall \{i,j\} \in E & \text{(4.2.2c)} \\ & y_{ii} = y_{0i} \ \forall i \in [1..n] & \text{(4.2.2d)} \\ & Y \succeq \mathbf{0} & \text{(4.2.2e)} \end{cases}$$

It is not hard to check that (ϑ'_G) is a relaxation of (α_G). Take an optimal solution $\mathbf{y}$ of (α_G) and construct $Y = \left[\begin{smallmatrix} 1 \\ \mathbf{y} \end{smallmatrix}\right] [1 \ \mathbf{y}^\top]$ that is feasible for (ϑ'_G) and has the same objective value as in (α_G).

Theorem 4.2.1. *The optimum value $\vartheta'(G)$ of (ϑ'_G) from (4.2.2a)-(4.2.2e) is equal to the Lovász theta number i.e., to the optimum value $\vartheta(G)$ of (ϑ_G) from (4.1.1a)-(4.1.1d).*

Proof. We first show that (ϑ'_G) is equivalent to the following relaxation of (α_G).

$$(\widetilde{\vartheta'}_G)\begin{cases} \max & \sum_{i=1}^{n} 2y_{0i} - y_{ii} & \text{(4.2.3a)} \\ s.t. & y_{00} = 1 & \text{(4.2.3b)} \\ & y_{ij} = 0 \ \forall \{i,j\} \in E & \text{(4.2.3c)} \\ & Y \succeq \mathbf{0} & \text{(4.2.3d)} \end{cases}$$

We take an optimal solution Y of $(\widetilde{\vartheta'}_G)$ and we will show it satisfies $y_{ii} = y_{0i} \ \forall i \in [1..n]$. Consider any $i \in [1..n]$. If $y_{ii} = 0$, then y_{0i} needs to be zero as well. If $y_{ii} \neq 0$, we can multiply row i and column i of Y by $\frac{y_{0i}}{y_{ii}}$ to obtain a matrix that is still feasible for $(\widetilde{\vartheta'}_G)$ and whose objective value is greater than or equal to that of Y. This simply follows from $2y_{0i}\frac{y_{0i}}{y_{ii}} - y_{ii}\frac{y_{0i}^2}{y_{ii}^2} \ge 2y_{0i} - y_{ii} \overset{y_{ii}>0}{\iff} y_{0i}^2 \ge 2y_{0i}y_{ii} - y_{ii}^2 \iff (y_{0i} - y_{ii})^2 \ge 0$. This proves $OPT(\widetilde{\vartheta'}_G) = OPT(\vartheta'_G)$. We will show that $OPT(\widetilde{\vartheta'}_G) = OPT(\vartheta_G) = \vartheta(G)$.

Let us write the dual of $(\widetilde{\vartheta'}_G)$.

$$(D\widetilde{\vartheta'}_G)\begin{cases} \min \ t & \text{(4.2.4a)} \\ s.t. \ \begin{bmatrix} t & -\mathbb{1}^\top \\ -\mathbb{1} & Z \end{bmatrix} = \begin{bmatrix} t & -1 & -1 & -1 & \dots & -1 \\ -1 & 1 & z_{12} & z_{13} & \dots & z_{n1} \\ -1 & z_{21} & 1 & z_{23} & \dots & z_{n2} \\ -1 & z_{31} & z_{32} & 1 & \dots & z_{n3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & z_{n1} & z_{n2} & z_{n3} & \dots & z_{nn} \end{bmatrix} \succeq \mathbf{0} & \text{(4.2.4b)} \\ \quad z_{ij} = 0 \ \forall \{i,j\} \notin E & \text{(4.2.4c)} \end{cases}$$

Programs $(\widetilde{\vartheta'}_G)$ in (4.2.3a)-(4.2.3d) and $(D\widetilde{\vartheta'}_G)$ in (4.2.4a)-(4.2.4c) are strictly feasible. For $(\widetilde{\vartheta'}_G)$, it is enough to take $Y = I$. For $(D\widetilde{\vartheta'}_G)$, we have $\left[\begin{smallmatrix} n+1 & -\mathbb{1}^\top \\ -\mathbb{1} & I_n \end{smallmatrix}\right] \succ \mathbf{0}$, as certified by applying the Sylvester criterion (Prop. 1.5.2) in reversed order (from bottom-right to top-left). The determinant of the whole matrix is 1: if we add the last n rows to the first one, we obtain a first row with only one non-zero element of value 1 in the upper-left corner. As such, using the strong duality Theorem 2.3.5, we have $OPT(\widetilde{\vartheta'}_G) = OPT(D\widetilde{\vartheta'}_G)$ and this value is effectively reached by both programs.

We now reformulate $(D\widetilde{\vartheta'}_G)$ and let us focus on (4.2.4b). Using the Schur complement Prop. 1.3.2, we have

$$\begin{bmatrix} t & -\mathbb{1}^\top \\ -\mathbb{1} & Z \end{bmatrix} \succeq \mathbf{0} \iff Z - \frac{1}{t}\mathbb{1} \succeq \mathbf{0} \iff \frac{tZ - \mathbb{1}}{t} \succeq \mathbf{0} \iff tZ - \mathbb{1} \succeq \mathbf{0},$$

where we used several times $t > 0$ (notice $t = 0$ would render $(D\widetilde{\vartheta'}_G)$ infeasible because of the upper-left 2×2 minor of the matrix in (4.2.4b)). Writing $\overline{Z} = tZ - \mathbb{1}$, we notice $\overline{z}_{ii} = t - 1\ \forall i \in [1..n]$ (because $z_{ii} = 1$ in (4.2.4b)) and $\overline{z}_{ij} = -1\ \forall \{i,j\} \notin E$ (because $z_{ij} = 0$ in (4.2.4c)). If $\{i,j\} \in E$, we have $\overline{z}_{ij} = tz_{ij} - 1$, *i.e.*, $\overline{z}_{ij}$ can actually take any real value, independent from the other terms of the matrix. Replacing this in $(D\widetilde{\vartheta'}_G)$, we obtain that $(D\widetilde{\vartheta'}_G)$ is equivalent to following program which is exactly (ϑ_G) from (4.1.1a)-(4.1.1d).

$$(\vartheta_G)\begin{cases} \min & t \\ s.t. & \overline{z}_{ii} = t - 1 \quad \forall i \in [1..n] \\ & \overline{z}_{ij} = -1 \qquad \forall \{i,j\} \notin E \\ & \overline{Z} \succeq \mathbf{0}. \end{cases}$$

All programs presented during the proof have the same optimum, *i.e.*, we proved that $OPT(\vartheta'_G) = OPT(\widetilde{\vartheta'}_G) = OPT(D\widetilde{\vartheta'}_G) = OPT(\vartheta_G) = \vartheta(G)$. The arguments from this proof are based on a reversed version of the proof from Section 3.1. of the PhD thesis of of Nebojša Gvozdenović.[17] □

4.3 A formulation $\vartheta''(G)$ of the theta number without SDP matrices

We associate an unit vector $\mathbf{u}^i$ for each vertex $i \in [1..n]$. We say that the unit vectors $\{\mathbf{u}^1, \mathbf{u}^2, \ldots \mathbf{u}^n\}$ constitute an *orthonormal representation* of $G = ([1..n], E)$ if and only if $\mathbf{u}^i \cdot \mathbf{u}^j = 0\ \forall \{i,j\} \notin E,\ i \neq j$. We introduce the following function:

$$\vartheta''(G) = \min\left\{ \max_{i\in[1..n]} \frac{1}{(\mathbf{c}\cdot\mathbf{u}^i)^2} :\ |\mathbf{c}| = 1,\ \{\mathbf{u}^1, \mathbf{u}^2, \ldots \mathbf{u}^n\} \text{ is an orthonormal representation of } G\right\} \quad (4.3.1)$$

Given any $\mathbf{c}$ and $\{\mathbf{u}^1, \mathbf{u}^2, \ldots \mathbf{u}^n\}$ that yield an optimum in above (4.3.1), we can apply a rotation R that maps $\mathbf{c}$ into unitary vector $R\mathbf{c} = [1\ 0\ 0\ 0 \ldots 0]^\top$ and $\{\mathbf{u}^1, \mathbf{u}^2, \ldots \mathbf{u}^n\}$ into $\{\mathbf{u}'^1, \mathbf{u}'^2, \ldots \mathbf{u}'^n\}$. A rotation is always represented by a matrix R that is unitary and orthonormal,[18] and so, does not change scalar products or norms. Indeed, if $R^\top R = I$, for any two vectors $\mathbf{v}$ and $\mathbf{v}'$ we have $\mathbf{v}^\top\mathbf{v}' = \mathbf{v}^\top R^\top R\mathbf{v}' = (R\mathbf{v})^\top(R\mathbf{v}')$, and so, the mapping $\mathbf{v} \to R\mathbf{v}$ does not change scalar products or norms. This means that $[1\ 0\ 0\ 0 \ldots 0]^\top$ and $\{\mathbf{u}'^1, \mathbf{u}'^2, \ldots \mathbf{u}'^n\}$ is also an optimal solution for (4.3.1). As such, we can simplify (4.3.1) to:

$$\vartheta''(G) = \min\left\{ \max_{i\in[1..n]} \left(\frac{1}{u_1^i}\right)^2 :\ \{\mathbf{u}^1, \mathbf{u}^2, \ldots \mathbf{u}^n\} \text{ is an orthonormal representation of } G\right\} \quad (4.3.2)$$

[17] As of 2019, it is available on-line at http://pure.uva.nl/ws/files/4245957/54393_thesis.pdf.

[18] We actually still need to prove there exists a rotation of the space that maps the unitary vector $\mathbf{c}$ to $\overline{\mathbf{c}} = [1\,0\,0\,0\ldots 0]^\top \in \mathbb{R}^k$. This task would be very simple if $\mathbf{c}$ had the form $\mathbf{c} = [c_1\ c_2\ 0\ 0\ 0 \ldots 0]^\top$, *i.e.*, it would be enough to apply a rotation on the (space of) the first two coordinates and keep all remaining $k-2$ coordinates fixed. We will construct a basis so as to make the representation of $\mathbf{c}$ in this basis always have the above form. Let us take an orthonormal basis $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^k$ (with $\mathbf{v}_1 = \overline{\mathbf{c}}$) of the sub-space spanned by $\mathbf{c}$ and $\overline{\mathbf{c}}$ and complete it to a full orthonormal basis by adding $\mathbf{v}_3,\ \mathbf{v}_4 \ldots \mathbf{v}_k$. Writing $V = [\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3 \ldots \mathbf{v}_k]$, any vector $\mathbf{v} \in \mathbb{R}^k$ can be represented in the new basis as $\mathbf{v} = V\mathbf{v}'$, *i.e.*, $\mathbf{v}' = V^{-1}\mathbf{v}$ is the expression of $\mathbf{v}$ in the new basis. Notice $\mathbf{c}$ and $\overline{\mathbf{c}}$ can be written as linear combinations of $\mathbf{v}_1$ and $\mathbf{v}_2$, and so, their representations $\mathbf{c}'$ and $\overline{\mathbf{c}}'$ in the new basis only use the first two coordinates. After checking that $\mathbf{c}'$ and $\overline{\mathbf{c}}'$ are unitary, we can define $\alpha = \arccos \mathbf{c}' \cdot \overline{\mathbf{c}}'$. The matrix $R_2 = \left[\begin{smallmatrix} \cos\alpha & sin\alpha \\ -\sin\alpha & \cos\alpha \end{smallmatrix}\right]$ rotates by α any 2-dimensional point, and so, it maps $\overline{\mathbf{c}}$ to $\mathbf{c}'$ in the space of the first two coordinates (or $\mathbf{c}'$ to $\overline{\mathbf{c}}$, case in which we replace α by $-\alpha$). We extend R_2 to $R_k \in \mathbb{R}^{k\times k}$ by putting an one on each new diagonal position (i,i) with $i > 2$ and zeros on all other positions (i,j) with $i > 2$ or $j > 2$. R_k rotates the first two coordinates of any vector $\mathbf{v}' \in \mathbb{R}^k$ and leaves untouched the rest of values $v'_3,\ v'_3 \ldots v'_k$. Let us calculate $VR_kV^\top\mathbf{c} = VR_kV^\top V\mathbf{c}' = VR_k\mathbf{c}' = V\overline{\mathbf{c}}' = \overline{\mathbf{c}}$. Since $VR_kV^\top$ is orthonormal, it is a rotation matrix that performs the desired rotation.

Theorem 4.3.1. $\vartheta''(G) = \vartheta(G)$

Proof. We will first show $\vartheta(G) \leq \vartheta''(G)$ and then $\vartheta''(G) \leq \vartheta(G)$.
$\vartheta(G) \leq \vartheta''(G)$
We will start from an optimal orthonormal representation of G and we will construct a feasible solution of $(D\widetilde{\vartheta'}_G)$ from (4.2.4a)-(4.2.4c) that we proved (Theorem 4.2.1) to reach the optimum $\vartheta(G)$. Without loss of generality, we can consider $|u_1^1| \leq |u_1^i|\ \forall i \in [1..n]$, and so, (4.3.2) states that $\vartheta''(G) = \left(\frac{1}{u_1^1}\right)^2$. Let us construct vectors $\mathbf{v}^0,\ \mathbf{v}^1,\ \mathbf{v}^2, \ldots \mathbf{v}^n$ such that $\mathbf{v}^0 = [-\frac{1}{u_1^1}\ 0\ 0\ 0 \ldots 0]^\top$ and for all $i \in [1..n]$ we set $\mathbf{v}^i = \frac{u_1^1}{u_1^i}\mathbf{u}^i$. We construct the Gram matrix Z' of these vectors such that $Z'_{ij} = \mathbf{v}^i \bullet \mathbf{v}^j\ \forall i,j \in [0..n]$. We have $Z' \succeq \mathbf{0}$ using Prop. A.1.8, $Z'_{00} = \left(\frac{1}{u_1^1}\right)^2 = \vartheta''(G)$, and $Z'_{ij} = 0\ \forall\{i,j\} \notin E$, using $\mathbf{v}^i \bullet \mathbf{v}^j = \frac{u_1^1}{u_1^i}\frac{u_1^1}{u_1^j}\mathbf{u}^i \bullet \mathbf{u}^j$ and the fact that $\mathbf{u}^1,\ \mathbf{u}^2, \ldots \mathbf{u}^n$ are a orthonormal representation of G. We also have $Z'_{0i} = -1$, because $v_1^i = u_1^1\ \forall i \in [1..n]$.

The diagonal elements are $Z'_{ii} = \left(\frac{u_1^1}{u_1^i}\right)^2 \mathbf{u}^i \bullet \mathbf{u}^i \leq 1\ \forall i \in [1..n]$, where we used $|u_1^1| \leq |u_1^i|$ and the fact that $\mathbf{u}_1,\ \mathbf{u}_2,\ \mathbf{u}_3, \ldots \mathbf{u}_n$ are unitary. We transform Z' into a matrix Z by only increasing the diagonal elements up to $Z_{ii} = 1\ \forall i \in [1..n]$; Z remains SDP as the sum of Z and a positive diagonal matrix. This matrix Z is a feasible solution in $(D\widetilde{\vartheta'}_G)$ in (4.2.4a)-(4.2.4c) with objective value $Z'_{00} = \left(\frac{1}{u_1^1}\right)^2 = \vartheta''(G)$. Since $OPT(D\widetilde{\vartheta'}_G) = \vartheta(G)$ as shown in the proof of Theorem 4.2.1, we have $\vartheta(G) \leq \vartheta''(G)$.

$\vartheta''(G) \leq \vartheta(G)$
We start from (ϑ_G) in (4.1.1a)-(4.1.1d) and we will construct an orthogonal representation of value $\vartheta(G)$. Notice the optimal solution $\overline{Z}$ of (4.1.1a)-(4.1.1d) verifies $\overline{Z}_{ii} = \vartheta(G) - 1$ and $\overline{Z}_{ij} = -1\ \forall\{i,j\} \notin E$. Using the Cholesky factorisation from Prop. 1.6.7 (or the eigenvalue or the square root factorization as in Corollary 1.7.1), there exist $\mathbf{v}_1,\ \mathbf{v}_2, \ldots \mathbf{v}_n$ such that $\overline{Z}_{ij} = \mathbf{v}_i \bullet \mathbf{v}_j$. We construct the representation $\mathbf{u}_i = \frac{1}{\sqrt{\vartheta(G)}}\begin{bmatrix}1\\ \mathbf{v}_i\end{bmatrix}\ \forall i \in [1..n]$. We verify the following:

- $|\mathbf{u}_i|^2 = \frac{1}{\vartheta(G)}(1 + \mathbf{v}_i \bullet \mathbf{v}_i) = \frac{\vartheta(G)}{\vartheta(G)} = 1\ \forall i \in [1..n]$, *i.e.*, the representation is unitary;
- for $\{i,j\} \notin E$, we have $\mathbf{u}_i \bullet \mathbf{u}_j = \frac{1}{\vartheta(G)}(1 + \mathbf{v}_i \bullet \mathbf{v}_j) = 0$, *i.e.*, the representation is orthonormal;
- the value of the representation in (4.3.2) is $\left(\frac{1}{\frac{1}{\sqrt{\vartheta(G)}}}\right)^2 = \vartheta(G)$ because $u_1^i = \frac{1}{\sqrt{\vartheta(G)}}\ \forall i \in [1..n]$.

This shows $\vartheta''(G) \leq \vartheta(G)$, which finishes the proof.

The proof of $\vartheta(G) \leq \vartheta''(G)$ is personal. The proof of $\vartheta''(G) \leq \vartheta(G)$ is adapted from the proof of "$\vartheta_1(G,w) \leq \vartheta_2(G,w)$" from Section 6 of the survey article "The sandwich theorem" of Donald Knuth.[19] □

4.4 A fourth formulation $\vartheta^\ell(G)$ of the theta number

We consider an orthonormal representation $\mathbf{u}^1,\ \mathbf{u}^2, \ldots \mathbf{u}^n$ of $\overline{G}$, *i.e.*, we have $\mathbf{u}^i \bullet \mathbf{u}^j = 0\ \forall\{i,j\} \in E$. The *leaning* of this representation is $\sum_{i=1}^{n}(\mathbf{e}_1 \bullet \mathbf{u}^i)^2 = \sum_{i=1}^{n}(\mathbf{u}_1^i)^2$, where $\mathbf{e}_1 = [1\ \underbrace{0\ 0 \ldots 0}_{n-1}]^\top$. One can think of this *leaning* term in the sense that any vector is leaning somehow on (casts its shadow on) the first dimension. We introduce:

$$\vartheta^\ell(G) = \max\left\{\sum_{i=1}^{n}(\mathbf{e}_1 \bullet \mathbf{u}^i)^2 :\ \{\mathbf{u}^1, \mathbf{u}^2, \ldots \mathbf{u}^n\} \text{ is an orthonormal representation of } \overline{G}\right\} \tag{4.4.1}$$

Proposition 4.4.1. $\vartheta^\ell(G) = \vartheta(G)$

[19] Published in the *Electronic Journal of Combinatorics* in 1994 (1), available at `http://www.combinatorics.org/ojs/index.php/eljc/article/view/v1i1a1/pdf`.

Proof. We will show that generating an orthonormal representation of $\overline{G}$ in (4.4.1) is equivalent to generating a feasible solution of (ϑ'_G) in (4.2.2a)-(4.2.2e) with the same objective value. We start by noticing that generating a feasible solution $Y \in \mathbb{R}^{(n+1)\times(n+1)}$ of (ϑ'_G) is equivalent to generating $n+1$ vectors $\mathbf{c}, \overline{\mathbf{v}}^1, \overline{\mathbf{v}}^2, \dots \overline{\mathbf{v}}^n$ such that $Y_{ij} = \overline{\mathbf{v}}^i \bullet \overline{\mathbf{v}}^j \ \forall i, j \in [1..n]$ and $Y_{0i} = Y_{i0} = \mathbf{c} \bullet \overline{\mathbf{v}}^i \ \forall i \in [1..n]$. Given constraints (4.2.2b)-(4.2.2d), these vectors need to satisfy: $\mathbf{c} \bullet \mathbf{c} = Y_{00} = 1$, $\overline{\mathbf{v}}^i \bullet \overline{\mathbf{v}}^j = 0 \ \forall \{i, j\} \in E$, and $\mathbf{c} \bullet \overline{\mathbf{v}}^i = \overline{\mathbf{v}}^i \bullet \overline{\mathbf{v}}^i = d_i^2, \ \forall i \in [1..n]$. The objective value of Y in (ϑ'_G) is $\sum_{i=1}^{n} d_i^2$.

Since there exists an orthonormal rotation matrix R that maps $\mathbf{c}$ to $R\mathbf{c} = \mathbf{e}_1$ and leaves unchanged scalar products (see Footnote 18, p. 43), generating vectors $\mathbf{c}, \overline{\mathbf{v}}^1, \overline{\mathbf{v}}^2, \dots \overline{\mathbf{v}}^n$ with above properties is equivalent to generating vectors $R\mathbf{c} = \mathbf{e}_1, R\overline{\mathbf{v}}^1 = \mathbf{v}^1, R\overline{\mathbf{v}}^2 = \mathbf{v}^2, \dots R\overline{\mathbf{v}}^n = \mathbf{v}^n$ with the same pairwise scalar products, *i.e.*, such that $\mathbf{v}^i \bullet \mathbf{v}^j = 0 \ \forall \{i, j\} \in E$ and $\mathbf{e} \bullet \mathbf{v}^i = \mathbf{v}^i \bullet \mathbf{v}^i = d_i^2, \ \forall i \in [1..n]$ because $RR^\top = I$. Remark $v_1^i = d_i^2$.

Generating vectors $\mathbf{e}_1, \mathbf{v}^1, \mathbf{v}^2, \dots \mathbf{v}^n$ with above properties is equivalent to generating vectors $\mathbf{e}_1, \mathbf{u}^1, \mathbf{u}^2, \dots \mathbf{u}^n$, where $\mathbf{u}^i = \frac{1}{d_i}\mathbf{v}^i \ \forall i \in [1..n]$ (if $d_i = 0$, we have $\mathbf{v}^i = \mathbf{0}$ and we set $\mathbf{u}^i = \left[\begin{smallmatrix}\mathbf{v}^i\\1\end{smallmatrix}\right] = \left[\begin{smallmatrix}\mathbf{0}\\1\end{smallmatrix}\right]$ and extend all other $\mathbf{u}^j \ \forall j \in [1..n] - \{i\}$ with a 0 element). One can check that $|\mathbf{u}^i|^2 = \mathbf{u}^i \bullet \mathbf{u}^i = \frac{1}{d_i^2}\mathbf{v}^i \bullet \mathbf{v}^i = 1 \ \forall i \in [1..n]$ and $\mathbf{u}^i \bullet \mathbf{u}^j = \frac{1}{d_i}\frac{1}{d_j}\mathbf{v}^i \bullet \mathbf{v}^j = 0 \ \forall \{i, j\} \in E$. This means that $\mathbf{u}^1, \ \mathbf{u}^2, \dots \mathbf{u}^n$ is an orthonormal representation of $\overline{G}$ such that $u_1^i = d_i \ \forall i \in [1..n]$. The leaning of $\mathbf{u}^1, \ \mathbf{u}^2, \dots \mathbf{u}^n$ in (4.4.1) is $\sum_{i=1}^{n} d_i^2$, *i.e.*, the same as the objective value of feasible matrix Y in (ϑ'_G) in (4.2.2a)-(4.2.2e). This confirms $\vartheta^\ell(G) = OPT(\vartheta'_G) = \vartheta(G)$. □

4.5 Two formulations of the theta number using maximum eigenvalues

4.5.1 A formulation only using the maximum eigenvalue

Let us introduce:

$$\vartheta^{\lambda_{\max}}(G) = \max\left\{\lambda_{\max}(Z) : \ Z \succeq \mathbf{0}, \ Z_{ii} = 1 \ \forall i \in [1..n], \ z_{ij} = 0 \ \forall \{i, j\} \in E\right\}, \tag{4.5.1}$$

where $\lambda_{\max}(Z)$ is the maximum eigenvalue of Z.

Theorem 4.5.1. $\vartheta^{\lambda_{\max}}(G) = \vartheta(G)$

Proof. Using Lemma 1.2.2.1, the maximum eigenvalue of Z can be computed as $\lambda_{\max}(Z) = \max\limits_{|\mathbf{x}|=1} Z \bullet (\mathbf{x}\mathbf{x}^\top)$. As such, (4.5.1) is equivalent to:

$$\vartheta^{\lambda_{\max}}(G) = \max\left\{Z \bullet (\mathbf{x}\mathbf{x}^\top) : \ Z \succeq \mathbf{0}, \ Z_{ii} = 1 \ \forall i \in [1..n], \ Z_{ij} = 0 \ \forall \{i, j\} \in E, \ |\mathbf{x}| = 1\right\} \tag{4.5.2}$$

We show that any feasible solution $(Z, \mathbf{x})$ of above program can be mapped to a feasible solution Y of $(D\vartheta_G)$ from (4.1.3a)-(4.1.3d). We showed in Section 4.1.2 that $\vartheta(G) = OPT(D\vartheta_G)$. For the reader's convenience, we recall below the definition of this program.

$$D\vartheta(G) = \max\left\{Y \bullet \mathbb{1} : \ Y \succeq \mathbf{0}, \ Y_{ij} = 0 \ \forall \{i, j\} \in E, \ \mathtt{trace}(Y) = 1\right\} \tag{4.5.3}$$

Consider matrix Y obtained by multiplying each row and column i of Z with x_i. In other words, $Y_{ij} = Z_{ij}x_ix_j$. It is not hard to check that Y reaches in (4.5.3) the same objective value $Y \bullet \mathbb{1} = \sum\limits_{i,j\in[1..n]} Z_{ij}x_ix_j = Z \bullet (\mathbf{x}\mathbf{x}^\top)$. Furthermore, Y satisfies all constraints of (4.5.3), *i.e.*, $\mathtt{trace}(Y) = \sum_{i=1}^{n} x_i^2 = |\mathbf{x}|^2 = 1$.

By reversing the above transformation, it is quite straightforward to prove the converse: any feasible solution Y of (4.5.3) can be mapped to a feasible solution $(Z, \mathbf{x})$ of (4.5.2). For this, we take $x_i = \sqrt{Y_{ii}} \ \forall i \in [1..n]$. We obtain Z by dividing each row and column i of Y with x_i for all i such that $x_i > 0$. More exactly, we obtain $Z_{ij} = \frac{Y_{ij}}{x_ix_j} \ \forall i, j \in [1..n], \ x_i > 0, x_j > 0$. If $x_i = 0$, we set $Z_{ii} = 1$ and $Z_{ij} = 0 \ \forall j \in [1..n] - \{i\}$. It is not hard to check that $(Z, \mathbf{x})$ satisfies the constraints of (4.5.2). For instance, we have $|\mathbf{x}|^2 = \mathtt{trace}(Y) = 1$ and $Z_{ii} = \frac{Y_{ii}}{\sqrt{Y_{ii}}^2} = 1$. One can also check $Y \bullet \mathbb{1} = \sum\limits_{i,j\in[1..n]} Y_{ij} = \sum\limits_{i,j\in[1..n]} Z_{ij}x_ix_j = Z \bullet (\mathbf{x}\mathbf{x}^\top)$, where we used $Y_{ij} = Z_{ij}x_ix_j$ that holds even if $x_i = 0$ or $x_j = 0$.

This proves (4.5.2) and (4.5.3) are equivalent, which means $\vartheta^{\lambda_{\max}}(G) = D\vartheta(G) = \vartheta(G)$. □

4.5.2 A formulation using the maximum and the minimum eigenvalue

Let us note $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ the minimum and resp. the maximum eigenvalue of X. We introduce

$$\vartheta^{\lambda}(G) = \max\left\{1 - \frac{\lambda_{\max}(X)}{\lambda_{\min}(X)} :\ X_{ii} = 0\ \forall i \in [1..n],\ X_{ij} = 0\ \forall \{i,j\} \in E\right\}, \tag{4.5.4}$$

where we use the convention $\frac{\lambda_{\max}(X)}{\lambda_{\min}(X)} = 0$ if $\lambda_{\max}(X) = \lambda_{\min}(X) = 0$. Since the only SDP matrix with zeros on the diagonal is $\mathbf{0}$, any feasible $X \neq \mathbf{0}$ in above (4.5.4) verifies $X \not\succeq \mathbf{0}$ and $X \not\preceq \mathbf{0}$, and so, $\lambda_{\max}(X) > 0 > \lambda_{\min}(X)$. Any $X \neq \mathbf{0}$ yields a value $1 - \frac{\lambda_{\max}(X)}{\lambda_{\min}(X)} > 1$ in (4.5.4). If $X = \mathbf{0}$ is the optimal solution of (4.5.4), then G must be a clique, so that constraints $X_{ij} = 0\ \forall\{i,j\} \in E$ force every non-diagonal element of X to be zero. In this case, we have $\alpha(G) = \vartheta(G) = \chi(\overline{G}) = \vartheta^{\lambda}(G) = 1$.

Theorem 4.5.2. $\vartheta^{\lambda}(G) = \vartheta(G)$

Proof. We take the optimal X in (4.5.4) and we construct a feasible solution Z of (4.5.1) with the same objective value. As stated above, $X = \mathbf{0}$ leads to $\vartheta^{\lambda}(G) = \vartheta(G) = 1$ and any $X \neq \mathbf{0}$ satisfies $\lambda_{\min}(X) < 0$. Considering $\lambda_{\min}(X) < 0$, we can construct $X' = \frac{X}{|\lambda_{\min}(X)|}$ and notice $\frac{\lambda_{\max}(X)}{\lambda_{\min}(X)} = \frac{\lambda_{\max}(X')}{\lambda_{\min}(X')} = -\lambda_{\max}(X')$. This means X' is also optimal in (4.5.4) and has $\lambda_{\min}(X) = -1$, more exactly X' lead (4.5.4) to objective value $1 + \lambda_{\max}(X')$. Consider now $Z = X' + I$ and notice $\lambda_{\min}(Z) = 0$, and so, $Z \succeq \mathbf{0}$. One can easily check Z is feasible in (4.5.1) and has objective value $\lambda_{\max}(X' + I) = \lambda_{\max}(X') + 1$, *i.e.*, the same as the objective value of X or X' in (4.5.4). This shows $\vartheta^{\lambda}(G) \leq \vartheta^{\lambda_{\max}}(G) = \vartheta(G)$.

We still have to prove $\vartheta^{\lambda_{\max}}(G) \leq \vartheta^{\lambda}(G)$. We now attempt to reverse the above process. Let Z be the optimal solution of (4.5.1) and take $X = Z - I$. Notice X is feasible in (4.5.4) and that

$$\vartheta^{\lambda_{\max}}(G) = \lambda_{\max}(Z - I + I) = 1 + \lambda_{\max}(X) = 1 + \frac{\lambda_{\max}(X)}{1} \leq 1 - \frac{\lambda_{\max}(X)}{\lambda_{\min}(X)}, \tag{4.5.5}$$

where we used $0 > \lambda_{\min}(X) \geq -1$ that we prove now. In fact, we already left aside the particular case $\lambda_{\min}(X) = 0$ because that would imply $X \succeq 0$ which would further lead (by applying $\mathtt{diag}(X) = \mathbf{0}$) to $X = 0$, which means the optimal Z is $Z = X + I = I$ and the graph is a clique. We can not have $\lambda_{\min}(X) > 0$ because that would imply $X \succ \mathbf{0}$, impossible when $\mathtt{diag}(X) = \mathbf{0}$. We now notice $Z = X + I \succeq \mathbf{0} \implies \lambda_{\min}(X + I) \geq 0 \implies \lambda_{\min}(X) \geq -1$. This shows $0 > \lambda_{\min}(X) \geq -1$, confirming the last inequality of (4.5.5). This finishes the proof because (4.5.5) simplifies to $\vartheta^{\lambda_{\max}}(G) \leq 1 - \frac{\lambda_{\max}(X)}{\lambda_{\min}(X)} \leq \vartheta^{\lambda}(G)$. $\square$

4.6 The theta number $\vartheta(G)$ is bounded by the fractional chromatic number $\chi^*(\overline{G})$ of $\overline{G}$

4.6.1 The fractional chromatic number

Let $\mathscr{C}$ be the set of cliques of G. In the primal-dual programs below we introduce the most standard formulation of the fractional chromatic number of $\overline{G}$. It is not hard to see that the standard chromatic number $\chi(\overline{G})$ is an upper bound for these programs, by taking $\lambda_{C_i} = 1\ \forall i \in [1..\chi(\overline{G})]$, where $\{C_1,\ C_2, \dots C_{\chi(\overline{G})}\}$ is the optimal coloring of $\overline{G}$ (clique covering of G).

$$(\hat{\chi}^*_{\overline{G}}) \left\{ \begin{array}{lll} \min & \sum\limits_{C \in \mathscr{C}} \lambda_C & \\ s.t. & \sum\limits_{\substack{C \in \mathscr{C} \\ C \ni i}} \lambda_C \geq 1 & \forall i \in [1..n] \\ & \lambda_C \geq 0 & \forall C \in \mathscr{C} \end{array} \right. \qquad (D\hat{\chi}^*_{\overline{G}}) \left\{ \begin{array}{lll} \max & \sum\limits_{i=1}^{n} x_i & \\ s.t. & \sum\limits_{i \in C} x_i \leq 1 & \forall C \in \mathscr{C} \\ & x_i \geq 0 & \forall i \in [1..n] \end{array} \right.$$

These primal-dual programs can be modified as follows. Let us drop the non-negativity constraint $\mathbf{x} \geq \mathbf{0}$ in the dual $(D\hat{\chi}^*_{\overline{G}})$. We prove by contradiction that the resulting program has only non-negative optimal solutions. Assume there is an optimal solution $\mathbf{x}$ such that $x_j < 0$ for some $j \in [1..n]$. We show that by increasing x_j to 0, $\mathbf{x}$ remains feasible. For this, take any clique $C \ni j$; since $C - \{j\}$ is naturally a clique and $\mathbf{x}$ is feasible, we have $\sum_{i \in C - \{j\}} x_i \leq 1$. If we now increase x_j to 0, the new solution satisfies

$\sum_{i\in C} x_i = \sum_{i\in C-\{j\}} x_i \le 1$. This new solution is still feasible and has a higher objective value, and so, we obtained a contradiction. The assumption $x_j < 0$ was false. The above programs are thus equivalent to the next ones, *i.e.*, $OPT(\hat{\chi}^*_{\overline{G}}) = OPT(D\hat{\chi}^*_{\overline{G}}) = OPT(\chi^*_{\overline{G}}) = OPT(D\chi^*_{\overline{G}}) = \chi^*(\overline{G})$

$$(\chi^*_{\overline{G}}) \begin{cases} \min & \sum\limits_{C\in\mathscr{C}} \lambda_C \\ s.t. & \sum\limits_{\substack{C\in\mathscr{C}\\ C\ni i}} \lambda_C = 1 \quad \forall i \in [1..n] \\ & \lambda_C \ge 0 \qquad\quad \forall C \in \mathscr{C} \end{cases} \tag{4.6.1a}$$

$$(D\chi^*_{\overline{G}}) \begin{cases} \max & \sum\limits_{i=1}^{n} x_i \\ s.t. & \sum\limits_{i\in C} x_i \le 1 \quad \forall C \in \mathscr{C} \\ & x_i \in \mathbb{R} \qquad \forall i \in [1..n] \end{cases} \tag{4.6.1b}$$

4.6.2 A hierarchy of SDP programs sandwiched between $\vartheta(G)$ and $\chi^*(\overline{G})$

We will use the $(D\widetilde{\vartheta}'_G)$ formulation of $\vartheta(G)$ from (4.2.4a)-(4.2.4c). For the reader's convenience, we repeat the definition of this program, multiplying by -1 the first row and column of the SDP matrix (this does not change its SDP status).

$$\left(D\widetilde{\vartheta}'_G\right) \begin{cases} \min & t & (4.6.2a) \\ s.t. & \begin{bmatrix} t & \mathbb{1}^\top \\ \mathbb{1} & Z \end{bmatrix} \succeq \mathbf{0} & (4.6.2b) \\ & z_{ii} = 1 \qquad \forall i \in [1..n] & (4.6.2c) \\ & z_{ij} = 0 \qquad \forall \{i,j\} \notin E & (4.6.2d) \end{cases}$$

Notice that $(\chi^*_{\overline{G}})$ in (4.6.1a) is defined using variables indexed by a subset of the power set $\mathbb{P}([1..n])$ of the vertex set $[1..n]$. This gives us some intuitions that we might need programs with variables indexed by certain subsets of $[1..n]$. Let us introduce $\mathbb{P}_r([1..n]) = \{S \subseteq [1..n] : \ |S| \le r\} \ \forall r \in [1..n]$. Given a vector $\mathbf{y}$ indexed by all $I \in \mathbb{P}_{2r}([1..n])$, we can write $\mathbf{y} = (y_I)_{I\in\mathbb{P}_{2r}([1..n])}$. Let us now introduce a matrix $M_r(\mathbf{y})$ with rows and columns indexed by $\mathbb{P}_r([1..n])$ such that $M_r(\mathbf{y})_{I,J} = \mathbf{y}_{I\cup J}$. We can also compactly write:

$$M_r(\mathbf{y}) = (y_{I\cup J})_{I,J\in\mathbb{P}_r([1..n])}. \tag{4.6.3}$$

We now introduce the following program:

$$\psi^r(G) = \left\{ \min y_\emptyset : \ M_r(\mathbf{y}) \succeq \mathbf{0}, \ y_{\{i\}} = 1 \ \forall i \in [1..n], \ y_{\{i,j\}} = 0 \ \forall \{i,j\} \notin E \right\}, \tag{4.6.4}$$

where one could notice that $M_r(\mathbf{y})_{I,I} = M_r(\mathbf{y})_{\emptyset,I} \ \forall I \in \mathbb{P}_r([1..n])$, in particular $M_r(\mathbf{y})_{\{i\},\{i\}} = M_r(\mathbf{y})_{\emptyset,\{i\}} = y_{\{i\}} = 1 \ \forall i \in [1..n]$.

Theorem 4.6.1. $\vartheta(G) = \psi^1(G) \le \psi^2(G) \le \cdots \le \psi^{\omega(G)}(G) \le \chi^*(\overline{G})$.

Proof. We first show that (4.6.4) with $r = 1$ is equivalent to $\left(D\widetilde{\vartheta}'_G\right)$ in (4.6.2a)-(4.6.2d). It is actually enough to carefully "decode" all notations to see that the two programs are simply identical up to a notational translation. With this goal, we can replace $M_1(\mathbf{y})_{\emptyset,\emptyset} = y_\emptyset$ with t, $M_1(\mathbf{y})_{\{i\},\{j\}} = y_{\{i,j\}} = 0$ with $z_{ij} = 0 \ \forall \{i,j\} \notin E$, and $M_1(\mathbf{y})_{\{i\},\{j\}} = y_{\{i,j\}}$ with $z_{ij} \ \forall \{i,j\} \in E$; $M_{\emptyset,\{i\}} = y_{\{i\}} = 1$ for $i \in [1..n]$ is simply translated into the vectors $\mathbb{1}$ and $\mathbb{1}^\top$ that border Z in (4.6.2b). One can check that we have just mapped one program into another, which guarantees that $\vartheta(G) = OPT\left(D\widetilde{\vartheta}'_G\right) = \psi^1(G)$.

We now show that $\psi^r(G) \le \psi^{r+1}(G) \ \forall r \in [1..n]$. Notice it is not possible to border a solution $M_r(\mathbf{y})$ of $\psi^r(G)$ with zeros to obtain a solution $M_{r+1}(\mathbf{y}')$ of $\psi^{r+1}(G)$, because certain elements of $M_{r+1}(\mathbf{y}')$ are inherited from $M_r(\mathbf{y}')$, *e.g.*, a proper bordering imposes $M_{r+1}(\mathbf{y}')_{\emptyset,[1..r+1]} = M_r(\mathbf{y})_{\{1\},[1..r]}$. Take the solution $\mathbf{y}'$ that achieves the optimum value $\psi^{r+1}(G)$. The key is to notice $M_r(\mathbf{y}')$ is a principal minor of $M_{r+1}(\mathbf{y}')$, and so, $M_r(\mathbf{y}') \succeq \mathbf{0}$. By removing from $\mathbf{y}'$ all indices $S \in \mathbb{P}_{2r+2}[1..n] - \mathbb{P}_{2r}[1..n]$, we obtain a vector $\mathbf{y}$ indexed by $\mathbb{P}_{2r}[1..n]$ that generates a feasible solution $M_r(\mathbf{y})$ of $\psi^r(G)$. Indeed, $M_r(\mathbf{y}) \succeq \mathbf{0}$ because $M_r(\mathbf{y}) = M_r(\mathbf{y}')$ is a principal minor of $M_{r+1}(\mathbf{y}')$; all other constraints in (4.6.4) concern the values of $\mathbf{y}$ on sets of $\mathbb{P}_2([1..n])$ that are inherited from $\mathbf{y}'$. The objective value is the same, *i.e.*, $y_\emptyset = y'_\emptyset$. This is enough to state $\psi^r(G) \le \psi^{r+1}(G)$, because (4.6.4) has a minimizing objective. The value $\psi^r(G)$ could be even strictly less that $y_\emptyset$ because $\psi^r(G)$ is obtained with a program (4.6.4) with fewer constraints than $\psi^{r+1}(G)$, *i.e.*, it involves smaller SDP matrices $M_r(\mathbf{y})$. In other words, there might exist some $\overline{\mathbf{y}}$ indexed by $\mathbb{P}_{2r}([1..n])$ that achieves a lower value $\psi^r(G)$ and that can not be extended to some feasible $\overline{\mathbf{y}}'$ indexed by $\mathbb{P}_{2i+2}([1..n])$.

We finally show $\psi^{\omega(G)}(G) \leq \chi^*(\overline{G})$. We take the optimum solution λ of $(\chi^*_{\overline{G}})$ from (4.6.1a) and construct a feasible solution with the same objective value in (4.6.4) with $r = \omega(G) = \mathbb{P}_{\omega(G)}([1..n])$. Take any clique $C \in \mathscr{C} \subseteq \mathbb{P}_{\omega(G)}([1..n])$ and construct $\mathbf{y}^C$ indexed by $\mathbb{P}_{2\omega(G)}([1..n])$ such that

$$y_S^C = \begin{cases} 1 & \text{if } S \subseteq C \\ 0 & \text{if } S \nsubseteq C \end{cases} \tag{4.6.5}$$

One only needs to decode notations to see that $M_{\omega(G)}(\mathbf{y}^C) = \overline{\mathbf{y}^C}\,\overline{\mathbf{y}^C}^\top \in \{0,1\}^{P_{\omega(G)}([1..n]) \times P_{\omega(G)}([1..n])}$, where $\overline{\mathbf{y}^C}$ is a reduced version of $\mathbf{y}^C$ that contains only $\mathbb{P}_{\omega(G)}([1..n])$ elements $\overline{y_S^C} = y_S^C$ with $S \in \mathbb{P}_{\omega(G)}([1..n])$. Indeed, if $M_{\omega(G)}(\mathbf{y}^C)_{I,J} = 0$, then $I \cup J \nsubseteq C$ using (4.6.5), and so, we have $I \nsubseteq C$ or $J \nsubseteq C$, which means that $y_I^C = 0$ or $y_J^C = 0$. Also, if $M_{\omega(G)}(\mathbf{y}^C)_{I,J} = 1$, then we have $I \cup J \subseteq C$ by virtue of (4.6.5), which means that $I, J \subseteq C$, and so, $y_I^C = y_J^C = 1$. This confirms that $M_{\omega(G)}(\mathbf{y}^C) = \overline{\mathbf{y}^C}\,\overline{\mathbf{y}^C} \succeq \mathbf{0}$.

By applying (4.6.3) on $\mathbf{y} = \sum\limits_{C \in \mathscr{C}} \lambda_C \mathbf{y}^C$ we obtain $M_{\omega(G)}(\mathbf{y}) = M_{\omega(G)}\left(\sum\limits_{C \in \mathscr{C}} \lambda_C \mathbf{y}^C\right) = \sum\limits_{C \in \mathscr{C}} \lambda_C M_{\omega(G)}\left(\mathbf{y}^C\right)$. As a sum of PSD matrices $M_{\omega(G)}(\mathbf{y}^C)$ resp. multiplied by positive scalars λ_C, the matrix $M_{\omega(G)}(\mathbf{y})$ is SDP. We now check the two non-SDP constraints of (4.6.4). First, we have $y_{\{i\}} = \sum\limits_{C \in \mathscr{C}} \lambda_C \mathbf{y}^C_{\{i\}} = \sum\limits_{C \in \mathscr{C}, C \ni i} \lambda_C = 1\ \forall i \in [1..n]$, using (4.6.1a). Secondly, $y_{\{i,j\}} = 0\ \forall \{i,j\} \notin E$ also holds because the non-edge $\{i,j\}$ belongs to no clique. We have just constructed a feasible solution $\mathbf{y}$ in (4.6.4) for $r = \omega(G)$ with objective value $y_\emptyset = \sum\limits_{C \in \mathscr{C}} \lambda_C \mathbf{y}^C_\emptyset = \sum\limits_{C \in \mathscr{C}} \lambda_C = \chi^*(\overline{G})$. This is enough to conclude $\psi^{\omega(G)}(G) \leq \chi^*(\overline{G})$.

Parts of this proof are a simplification of the proof of Theorem 3.1. from the article "The operator Ψ for the chromatic number of a graph" of Nebojša Gvozdenović and Monique Laurent.[20] □

5 A taste of copositive optimization and sum of squares hierarchies

5.1 Introducing the completely positive and the copositive cones

Let us try to produce better relaxations and reformulations by replacing the cone S_n^+ of SDP matrices with a smaller cone. For this, we introduce the cone of *completely positive matrices*:

$$C^{n*} = \left\{ X \in S_n :\ X = \sum_{i=1}^{k} \mathbf{y}_i \mathbf{y}_i^\top \text{ with } \mathbf{y}_i \geq \mathbf{0}\ \forall i \in [1..k] \right\} \tag{5.1.1}$$

$$= \texttt{conv}\left\{ \mathbf{y}\mathbf{y}^\top : \mathbf{y} \geq \mathbf{0} \right\}, \tag{5.1.2}$$

where $S_n \subsetneq \mathbb{R}^{n \times n}$ is the set of real symmetric matrices and the operator $\texttt{conv}(...)$ produces all convex combinations of the elements from the set given as argument. It is clear that any X from (5.1.1) can be written as a convex combination of rank-1 matrices as in (5.1.2). For this, it is enough to write $X = \sum_{i=1}^k \mathbf{y}_i \mathbf{y}_i^\top = \sum_{i=1}^k \frac{1}{k}\left(\sqrt{k}\mathbf{y}_i\right)\left(\sqrt{k}\mathbf{y}_i\right)^\top$. Any $X \in C^{n*}$ can be written $X = YY^\top$ with $Y = [\mathbf{y}_1\ \mathbf{y}_2\ \ldots \mathbf{y}_k]$, and so, it is clear that X is also SDP using Prop. A.1.8. The difference between a completely positive matrix X and an SDP one is that above Y needs to be non-negative. Any SDP matrix $\widetilde{X} \succeq \mathbf{0}$ can be written as $\widetilde{X} = \widetilde{Y}\widetilde{Y}^\top$ using various decompositions (*e.g.*, Cholesky, eigenvalue or square root, see Corollary 1.7.1), but $\widetilde{Y}$ is not necessarily non-negative. For any $n \geq 2$, we have $C^{n*} \subsetneq S_n^+$.

We can define the dual of a cone C (with regards to the scalar product) using the formula $C^* = \{Y :\ X \bullet Y \geq 0\ \forall X \in C\}$. We already proved in Prop. 1.3.3 that the SDP cone is self-dual, *i.e.*, $(S_n^+)^* = S_n^+$. Since C^{n*} is smaller than S_n^*, its dual cone might be larger than S_n^+. Indeed, we introduce the *cone of copositive matrices*

$$C^n = \left\{ X \in S_n :\ X \bullet \mathbf{y}\mathbf{y}^\top \geq 0\ \forall \mathbf{y} \geq \mathbf{0} \right\} \tag{5.1.3}$$

[20] Published in *SIAM Journal on Optimization* in 2008, vol 19(2), pp. 572–591, available on-line as of 2017 at http://homepages.cwi.nl/~monique/files/SIOPTGL1.pdf.

such that $C^n = (C^{n*})^*$.

Let $\mathcal{N}^n \subsetneq S_n$ be the set of non-negative symmetric matrices. The following hierarchy of inclusions

$$C^{n*} \subset S_n^+ \cap \mathcal{N}^n \subset S_n^+ \subset S_n^+ + \mathcal{N}^n \subset C^n \tag{5.1.4}$$

follows from two short arguments. First, any $X \in C^{n^*}$ satisfies $X \succeq \mathbf{0}$ (see above) and $X \geq \mathbf{0}$ (see the definition (5.1.1)). Secondly, any $S \succeq \mathbf{0}$ and $N \geq \mathbf{0}$ verify $S \bullet \mathbf{y}\mathbf{y}^\top \geq 0$ and $N \bullet \mathbf{y}\mathbf{y}^\top \geq 0$ for any $\mathbf{y} \geq \mathbf{0}$.

5.2 Reformulating a homogeneous quadratic program as a copositive problem

We consider the following problem with a *homogeneous objective* function and *non-negative variables*:

$$(QP_+) \begin{cases} \min & Q \bullet \mathbf{x}\mathbf{x}^\top & (5.2.1a) \\ s.t & \mathbf{a}^\top \mathbf{x} = b & (5.2.1b) \\ & \mathbf{x} \in \mathbb{R}^n_+, & (5.2.1c) \end{cases}$$

where $\mathbf{a} > 0$ is a *strictly positive parameter*. One should keep in mind that $\mathbf{x}$ is non-negative.

5.2.1 Solving (QP_+) is NP-hard

Not surprisingly, solving this program is NP hard. This follows from the fact that it contains the maximum stable problem as a particular case,[21] as a consequence of the following result.

Proposition 5.2.1. *Consider a graph G with adjacency matrix A^G. The maximum stable $\alpha(G)$ can be determined by solving the following program, which is a particular case of (QP_+) from above (5.2.1a)-(5.2.1c).*

$$\frac{1}{\alpha(G)} = \min \left\{ \left(A^G + I_n\right) \bullet \mathbf{x}\mathbf{x}^\top : \; \mathbf{e}_n^\top \mathbf{x} = 1 \right\}, \tag{5.2.2}$$

where $\mathbf{e}_n^\top = \underbrace{[1\ 1\ 1 \dots 1]}_{n \text{ positions}}$.

Proof. We use the technique described next. Consider any feasible solution $\mathbf{x}$ of above program (5.2.2). We say that an edge $\{i,j\}$ is *supported* by $\mathbf{x}$ if $x_i, x_j > 0$. We will show that if $\mathbf{x}$ has more than zero supported edges, then it be transformed into a no-worse solution $\mathbf{x}'$ that has strictly fewer supported edges. For this, let us start by evaluating the contribution of x_i and x_j to the objective value of edge $\{i,j\}$:

$$\begin{aligned} \mathtt{val}(x_i, x_j) &= x_i^2 + 2x_i x_j + x_j^2 + 2 \sum_{k \in [1..n] - \{i,j\}} A^G_{ki} x_k x_i + 2 \sum_{k \in [1..n] - \{i,j\}} A^G_{kj} x_k x_j & (5.2.3) \\ &= (x_i + x_j)^2 + z_i x_i + z_j x_j, & (5.2.4) \end{aligned}$$

where the values z_i and z_j are determined by the above sums and do not depend on x_i or x_j. Consider now the function $f : [0, x_i + x_j] \to \mathbb{R}$ defined by $f(t) = \mathtt{val}(t, x_i + x_j - t)$, *i.e.*, we replace x_i with t and x_j with $x_i + x_j - t$ in above (5.2.3)-(5.2.4). This way $f(t)$ can be written

$$f(t) = (x_i + x_j)^2 + z_i t + z_j (x_i + x_j - t) = (z_i - z_j)t + (x_i + x_j)^2 + z_j (x_i + x_j)$$

The only-non constant term is $(z_i - z_j)t$. We obtain that f is a linear function that reaches its minimum value at one of the two bounds of $[0, x_i + x_j]$, *i.e.*, a value $\bar{t}$ that minimizes f is either $\bar{t} = 0$ or $\bar{t} = x_i + x_j$. We can say $f\left(\bar{t}\right) = \mathtt{val}\left(\bar{t}, x_i + x_j - \bar{t}\right)$ is at least as good as $f(x_i) = \mathtt{val}(x_i, x_j)$.

If we define $x'_i = \bar{t}$, $x'_j = x_i + x_j - \bar{t}$ and $x'_k = x_k \ \forall k \in [1..n] - \{i,j\}$, we obtain that $\mathbf{x}'$ is at least as good as $\mathbf{x}$. Notice that the edge $\{i,j\}$ is no longer supported in $\mathbf{x}'$ because x'_i or x'_j is equal to zero. Any other supported edge in $\mathbf{x}'$ is also supported in $\mathbf{x}$, because $x'_\ell > 0 \implies x_\ell > 0 \ \forall \ell \in [1..n]$. This means that

[21] If we allow the objective to be non-homogeneous, we obtain a particular case of unconstrained quadratic programming *in non-negative variables*. This problem is NP-hard because it is at least as hard as the bi-partition problem, *i.e.*, one can solve the (bi-)partition problem for elements $a_1, a_2, \dots a_n$ by solving $\min_{x_i, x'_i \geq 0} \sum_i (x_i + x'_i - 1)^2 + x_i x'_i + \left(\sum_i x_i a_i - \frac{1}{2} \sum_i a_i\right)^2$. More generally, unconstrained quadratic programming is not NP-hard because it reduces to SDP programming (Section 3.4.3).

$\mathbf{x}'$ has at least one supported edge less than $\mathbf{x}$. By repeating above operation iteratively for all supported edges, we will eventually find some $\overline{\mathbf{x}}'$ that is at least as good as $\mathbf{x}$ and has no supported edge.

This means that the non-zero elements of $\overline{\mathbf{x}}'$ generate a stable S of G. The objective value of $\overline{\mathbf{x}}'$ can be written $I_{|S|} \bullet \overline{\mathbf{x}}\, \overline{\mathbf{x}}^\top$, where $\overline{\mathbf{x}}$ is a vector containing only the $|S|$ non-zero values of $\overline{\mathbf{x}}'$. We can still say $\mathbf{e}_{|S|}^\top \overline{\mathbf{x}} = 1$, *i.e.*, the sum of the elements of $\overline{\mathbf{x}}$ (or $\overline{\mathbf{x}}'$) is one. We now show that all elements of $\overline{\mathbf{x}}$ are equal if $\overline{\mathbf{x}}'$ is optimal. Assume the contrary: there is $i, j \in [1..|S|]$ such that $\overline{x}_i \neq \overline{x}_j$. The contribution of $\overline{x}_i$ and $\overline{x}_j$ to the objective function is $\overline{\mathtt{val}}(\overline{x}_i, \overline{x}_j) = \overline{x}_i^2 + \overline{x}_j^2 = \frac{1}{2}\left(2\overline{x}_i^2 + 2\overline{x}_j^2\right) > \frac{1}{2}\left(\overline{x}_i^2 + \overline{x}_j^2 + 2\overline{x}_i\overline{x}_j\right) = 2\left(\frac{\overline{x}_i+\overline{x}_j}{2}\right)^2$. By replacing $\overline{x}_i$ and $\overline{x}_j$ with $\left(\frac{\overline{x}_i+\overline{x}_j}{2}\right)$ we obtain a better objective value, while still respecting the constraint (the sum of the elements remains the same). This means that $\overline{\mathbf{x}}'$ is not optimal, which is a contradiction. As such, the optimal $\overline{\mathbf{x}}'$ has the same value $\frac{1}{|S|}$ at all positions $i \in S$ and its objective value is $\sum_{i=1}^{|S|} \frac{1}{|S|^2} = \frac{|S|}{|S|^2} = \frac{1}{|S|}$.

We started from an arbitrary solution $\mathbf{x}$ and we constructed a solution $\overline{\mathbf{x}}'$ that is no worse than $\mathbf{x}$ and that has no supported edge; this means there is always an optimal solution with no supported edges. We then showed that if $\overline{\mathbf{x}}'$ is optimal, it has to contain the value $\frac{1}{|S|}$ on all positions of some stable S of G; the objective value of this solution is $\frac{1}{|S|}$. But this value can only be optimal if S is a maximum stable of G with $|S| = \alpha(G)$. We obtained that the optimum of the program from (5.2.2) is indeed $\frac{1}{\alpha(G)}$. It is achieved by taking $x_i = \frac{1}{|S|}$ for all $i \in S$ and $x_i = 0$ if $i \notin S$ for a maximum stable S. $\square$

According to the article "Copositive Optimization" by Immanuel Bomze, Mirjam Dür and Chung-Piaw Teo,[22] the above result dates back to the 1960s. However, the proof is personal.

5.2.2 The reformulation of (QP_+) as a copositive program

Let us consider the following program associated to (QP_+) from (5.2.1a)-(5.2.1c).

$$C(QP_+) \begin{cases} \min \quad Q \bullet X & (5.2.5a) \\ s.t \quad A \bullet X = (\mathbf{a}\mathbf{a}^\top) \bullet X = b^2 & (5.2.5b) \\ \qquad X \in C^{n*}, & (5.2.5c) \end{cases}$$

where recall $A = \mathbf{a}\mathbf{a}^\top$ respects $A \geq \mathbf{0}$ and $A_{ii} > 0\ \forall i \in [1..n]$ because (5.2.1a)-(5.2.1c) imposes $a_i > 0 \forall i \in [1..n]$, *i.e.*, $\mathbf{a}$ is a strictly positive parameter. Since the above program minimizes a linear function, its optimum might only be achieved by an extreme point (or along an extreme ray) of the feasible area defined by (5.2.5b)-(5.2.5c).

Proposition 5.2.2. *We consider a symmetric matrix $A \geq \mathbf{0}$ with a strictly positive diagonal ($A_{ii} > 0\ \forall i \in [1..n]$) and some $b \in \mathbb{R}$. The extreme solutions (vertices) of the set below are the rank-1 matrices of the form $X = \mathbf{u}\mathbf{u}^\top$ with $\mathbf{u} \geq \mathbf{0}$. This set has no extreme rays.*

$$[C_A^{n*}] = \left\{ X \in C^{n*} :\ A \bullet X = b^2 \right\} \tag{5.2.6}$$

Proof. We will prove three facts:

(i) A rank-1 completely positive matrix X such that $AX = b^2$ is an extreme solution of $[C_A^{n*}]$.

(ii) A completely positive matrix X of rank higher than one can not be an extreme solution of $[C_A^{n*}]$.

(iii) The set $[C_A^{n*}]$ has *no* extreme ray.

(i)
We first prove that a rank-1 matrix $X = \mathbf{y}\mathbf{y}^\top$ (with $\mathbf{y} \geq \mathbf{0}$) such that $AX = b^2$ is an extreme solution of this $[C_A^{n*}]$ set. Assume the contrary: there is symmetric non-zero $M \in \mathbb{R}^{n \times n}$ such that $X - M, X + M \in [C_A^{n*}]$. Based on (5.1.1), we can write $X - M = Y_a Y_a^\top$ and $X + M = Y_b Y_b^\top$; combining the two, we obtain $X = \frac{1}{2}[Y_a Y_b][Y_a Y_b]^\top$. But since $rank(X) = rank\left(\mathbf{y}\mathbf{y}^\top\right) = 1$, we can apply Prop. A.1.8 to conclude

[22] Published in the *Optima* 89 newsletter in august 2012, pp. 2-8, available on line at http://www.mathopt.org/Optima-Issues/optima89.pdf.

$rank(X) = rank\left([Y_aY_b][Y_aY_b]^\top\right) = rank([Y_aY_b])$, and so, we have $rank([Y_aY_b]) = 1$. The columns of Y_a and Y_b need to be multiples of $\mathbf{y}$. The matrices X and $X - M$ are multiples of $\mathbf{y}\mathbf{y}^\top$, and so, M is a multiple of X, *i.e.*, $M = tX$ with $t \neq 0$. This contradicts $X - M \in [C_A^{n*}]$ because $A\,(X - M) = b^2 - tb^2 \neq b^2$. The assumption $X - M, X + M \in [C_A^{n*}]$ for some $M \neq \mathbf{0}$ was false, and so, X is an extreme solution.
(ii)
We now show that any matrix $X \in C^{n*}$ of rank higher than 1 can not be an extreme solution of $[C_A^{n*}]$. Assume there exists k non-zero vectors $\mathbf{y}_1,\ \mathbf{y}_2, \dots \mathbf{y}_k \geq \mathbf{0}$ such that $rank[\mathbf{y}_1\ \mathbf{y}_2\ \dots \mathbf{y}_k] > 1$ (*i.e.*, they are not all multiples of the same vector) and $X = \sum_{i=1}^k \mathbf{y}_i\mathbf{y}_i^\top$, recall definition (5.1.1). Without loss of generality, we suppose $\mathbf{y}_1$ and $\mathbf{y}_2$ are linearly independent. We can write:

$$A \bullet X = \underbrace{A \bullet \mathbf{y}_1\mathbf{y}_1^\top}_{t_1} + \underbrace{A \bullet \mathbf{y}_2\mathbf{y}_2^\top}_{t_2} + A \bullet \sum_{i=3}^{k} \mathbf{y}_i\mathbf{y}_i^\top. \tag{5.2.7}$$

Notice we have $t_1,\ t_2 > 0$, because $A \bullet \mathbf{y}_1\mathbf{y}_1^\top \geq \mathtt{diag}(A) \bullet \mathtt{diag}(\mathbf{y}_1\mathbf{y}_1^\top) > 0$ follows from the fact that A and $\mathbf{y}_1$ are non-negative, the diagonal of A is strictly positive and $\mathbf{y}_1$ is non-null. An analogous argument proves $t_2 > 0$. We now introduce the following family of matrices depending on a parameter α.

$$X_\alpha = X + \alpha \underbrace{\left(\frac{1}{t_1}\mathbf{y}_1\mathbf{y}_1^\top - \frac{1}{t_2}\mathbf{y}_2\mathbf{y}_2^\top\right)}_{\substack{\neq 0 \text{ because } \mathbf{y}_1 \text{ and} \\ \mathbf{y}_2 \text{ are independent}}} \tag{5.2.8}$$

Notice X_α remains completely positive for a sufficiently small positive or negative α — more exactly the limits of α are $\alpha \in [-t_1, t_2]$. Let us check if $X_\alpha \in [C_A^{n*}]$ by calculating the scalar product with A:

$$AX_\alpha = AX + \alpha\left(\frac{1}{t_1}A \bullet \mathbf{y}_1\mathbf{y}_i^\top - \frac{1}{t_2}A \bullet \mathbf{y}_2\mathbf{y}_2^\top\right) = AX + \alpha\left(\frac{t_1}{t_1} - \frac{t_2}{t_2}\right) = AX$$

We can move from X back and forward along non-zero $\left(\frac{1}{t_1}\mathbf{y}_1\mathbf{y}_i^\top - \frac{1}{t_2}\mathbf{y}_2\mathbf{y}_2^\top\right)$ until we reach (the above) limits of α. Thus, such X can not be an extreme solution of $[C_A^{n*}]$.
(iii)
We finally show (by contradiction) that $[C_A^{n*}]$ can not contain an extreme ray of the form $X + tZ$ with $t > 0$ and $Z \neq \mathbf{0}_{n\times n}$. Assume such Z exists. We can easily notice $Z \geq \mathbf{0}$ because $C^{n*} \subset \mathbb{R}^n_+$. Since $A \bullet Z$ has to be zero and the diagonal of A is strictly positive, we also obtain $\mathtt{diag}(Z) = \mathbf{0}$, which means Z is *not* SDP by applying Corollary 1.6.5. There exists a vector $\mathbf{u}$ such that $Z \bullet \mathbf{u}\mathbf{u}^\top = -z < 0$. But now notice that $(X + tZ) \bullet \mathbf{u}\mathbf{u}^\top = X \bullet \mathbf{u}\mathbf{u}^\top - tz$ can become negative for a sufficiently large t, which means $X + tZ$ is not SDP. This is a contradiction because $C^{n*} \subset S_n^+$ in (5.1.4). □

The above Prop. 5.2.2 leads to the fact that the optimal solution of $(C(QP_+))$ from (5.2.5a)-(5.2.5c) has the form $X = \mathbf{y}\mathbf{y}^\top$ with $\mathbf{y} \geq \mathbf{0}$. This means $\mathbf{y}$ is an optimal solution of (QP_+) from (5.2.1a)-(5.2.1c). Indeed, if (QP_+) would have a solution $\overline{\mathbf{y}}$ of better quality, $\overline{X} = \overline{\mathbf{y}}\,\overline{\mathbf{y}}^\top$ would also be a solution of better quality than X in $(C(QP_+))$. This means $(C(QP_+))$ is an exact reformulation of (QP_+).

The difficulty of $(C(QP_+))$ is hidden in the cone constraint. Indeed, checking membership in C^{n*} is NP-hard. In particular, if we try to factorise $X \in C^{n*}$ into $X = YY^\top$ using any of the decompositions presented for SDP matrices (*e.g.*, Cholesky, eigenvalue or square root, see Corollary 1.7.1), the factor Y is *not necessarily* non-negative. It is still an open question whether checking C^{n*} membership is in NP (making the problem NP-complete, because it is NP-Hard) or not. Checking membership in the dual cone C^n is even co-NP complete.[23] Under a legitimate well-accepted (but still open) assumption co-NP$\neq$ NP, a co-NP complete problem can not belong to NP – if that were the case, all co-NP problems would belong to NP, which is unlikely. This way, it is very likely that checking C^n membership is *not* even in NP. For more details on such aspects, we refer the reader to the article "On the computational complexity of membership problems for the completely positive cone and its dual" by Peter Dickinson and Luuk Gijben.[24]

[23] Recall a decision problem is NP (reps. co-NP) if and only if there is a polynomial-time algorithm that can verify "yes" (resp. "no") instances.

[24] Published in *Computational optimization and applications* in 2014, vol 57(2), pp. 403–415, available on-line at http://www.optimization-online.org/DB_FILE/2011/05/3041.pdf.

5.2.3 Comparing with the SDP relaxation of (QP_+)

We now investigate the reasons why replacing C^{n*} with S_n^+ does not lead to such a strong result (reformulation). In other words, we replace the completely positive (A-)constrained set $[C_A^{n*}]$ from (5.2.6) with an SDP set $[SDP_A]$ using the same constraint defined by A. More exactly, we investigate why the characterization of extreme solutions from Prop. 5.2.2 of $[C_A^{n*}]$ does not hold in the same way for

$$[SDP_A] = \left\{ X \succeq \mathbf{0} : \; A \bullet X = b^2 \right\}. \tag{5.2.9}$$

First, we can still say that the rank-1 matrices $X \succeq \mathbf{0}$ such that $A \bullet X = b^2$ are extreme solutions of $[SDP_A]$. It is enough to check that the arguments for point (i) from the proof of Prop. 5.2.2 still hold for $[SDP_A]$. However, one should bear in mind that a rank-1 SDP matrix $X = \mathbf{y}\mathbf{y}^\top$ might *not* verify $\mathbf{y} \geq 0$, and so, $\mathbf{y}$ might *not* be a feasible solution of the initial program $[QP_+]$ from (5.2.5a)-(5.2.5c).

Secondly, the SDP set might have extreme rays and the proof of (iii) from Prop. 5.2.2 fails in the SDP case. This follows from the fact that even if $A \geq \mathbf{0}$, we can still find SDP matrices Z such that $A \bullet Z = 0$. This means there might well be matrices $X + tZ \in [SDP_A], \forall t > 0$, based on $A \bullet (A + tZ) = A \bullet X = b^2$. If $\exists Z \succeq \mathbf{0}$ such that $A \bullet Z = 0$ and $Q \bullet Z < 0$, the SDP relaxation of $[QP_+]$ from (5.2.5a)-(5.2.5c) is unbounded. We will assume that this relaxation is not unbounded, *i.e.*, $A \bullet Z = 0 \implies Q \bullet Z \geq 0$.

Finally, we tackle the point (iii) of the proof of Prop. 5.2.2. We can still prove there is no extreme solution of $[SDP_A]$ with rank higher than 1. For this, we can still write (5.2.7) with independent $\mathbf{y}_1$ and $\mathbf{y}_2$. However, we can no longer state $t_1, \; t_2 > 0$. If we have $t_1 = 0$, then $\mathbf{y}_1\mathbf{y}_1^\top$ is a ray of $[SDP_A]$ and it is clear X is not an extreme point because we can add or subtract multiples of $\mathbf{y}_1\mathbf{y}_1^\top$ from the description $X = \sum_{i=1}^k \mathbf{y}_i\mathbf{y}_i^\top$ and remain in $[SDP_A]$. The same happens if $t_2 = 0$. We can hereafter assume $t_1 \neq 0$ and $t_2 \neq 0$. This way, we can still construct the family X_α of matrices from (5.2.8). One can check that X_α remains in the $[SDP_A]$ for sufficiently small values of α, *i.e.*, check that if $\alpha \in [-\varepsilon, +\varepsilon]$ with $\varepsilon < \min(|t_1|, |t_2|)$, then the coefficients of $\mathbf{y}_1\mathbf{y}_1^\top$ and $\mathbf{y}_2\mathbf{y}_2^\top$ in the sum composing X_α remain strictly positive. This is enough to guarantee that X is not an extreme solution.

To summarize, we have found two differences between the completely positive (re-)*formulation* and the SDP *relaxation*. First, the feasible area in the SDP case can have extreme rays and the objective is unbounded if there is $Z \succeq \mathbf{0}$ such that $A \bullet Z = 0$ and $Q \bullet Z < 0$. Secondly, it the objective is not unbounded, the optimal solution has rank 1 like in the completely positive case, but it is not necessarily non-negative.

5.3 Relaxations of the copositive formulation of the maximum stable

5.3.1 A second completely positive formulation of the maximum stable

We have already introduced the $\alpha(G)$ formulation (5.2.2) and proven it in Prop. 5.2.1. We will show that

$$\alpha(G) = \max \left\{ \mathbf{e}_n\mathbf{e}_n^\top \bullet X : \; \left(A^G + I_n\right) \bullet X = 1, \; X \in C^{n*} \right\}, \tag{5.3.1}$$

where A^G is the adjacency matrix of graph G and recall $\mathbf{e}_n^\top = \underbrace{[1\ 1\ 1 \ldots 1]}_{\text{n positions}}$.

First, notice this program is very similar to $C(QP_+)$ from (5.2.5a)- (5.2.5c). In particular, the matrix $A^G + I_n$ satisfies exactly all conditions imposed on A in Prop. 5.2.2. As such, an optimal solution of (5.3.1) can be achieved by an extreme solution of the feasible area that has the form $X = \mathbf{y}\mathbf{y}^\top$ with $\mathbf{y} \geq \mathbf{0}$ by virtue of Prop. 5.2.2. It is enough to prove the following:

$$\alpha(G) = \max \left\{ \mathbf{e}_n\mathbf{e}_n^\top \bullet \mathbf{y}\mathbf{y}^\top : \; \left(A^G + I_n\right) \bullet \mathbf{y}\mathbf{y}^\top = 1 \right\}, \tag{5.3.2}$$

Consider a feasible solution $\mathbf{x}$ of (5.2.2) with objective value $\left(A^G + I_n\right) \bullet \mathbf{x}\mathbf{x}^\top = \frac{1}{t}$. Let us define $\mathbf{y} = \sqrt{t}\mathbf{x}$ and one can calculate $\left(A^G + I_n\right) \bullet \mathbf{y}\mathbf{y}^\top = \sqrt{t}^2 \left(A^G + I_n\right) \bullet \mathbf{x}\mathbf{x}^\top = t\frac{1}{t} = 1$, *i.e.*, $\mathbf{y}$ is feasible in (5.3.2). Based on $\mathbf{e}_n^\top\mathbf{x} = 1$, we obtain $\mathbf{e}_n^\top\mathbf{y} = \sqrt{t}$, and so, $\mathbf{e}_n\mathbf{e}_n^\top \bullet \mathbf{y}\mathbf{y}^\top = \sqrt{t}^2 = t$. From a feasible solution $\mathbf{x}$ of (5.2.2) with value $\frac{1}{t}$, we constructed a feasible solution in (5.3.2) with objective value t. The converse is also possible. Consider any feasible solution $\mathbf{y}$ of (5.3.2) of objective value $(\mathbf{e}_n^\top\mathbf{y})^2 = t$ (notice any feasible $\mathbf{y}$ is non-zero, and so, $t > 0$). Let us define $\mathbf{x} = \frac{\mathbf{y}}{\sqrt{t}}$. From $(\mathbf{e}_n^\top\mathbf{y})^2 = t$, we have $(\mathbf{e}_n^\top\mathbf{x})^2 = \left(\frac{1}{\sqrt{t}}\mathbf{e}_n^\top\mathbf{y}\right)^2 = \frac{1}{\sqrt{t}^2}t = 1$, *i.e.*, $\mathbf{x}$ is feasible in (5.2.2). The objective value of $\mathbf{x}$ is $\left(A^G + I_n\right) \bullet \mathbf{x}\mathbf{x}^\top = \frac{1}{\sqrt{t}^2}\left(A^G + I_n\right) \bullet \mathbf{y}\mathbf{y}^\top = \frac{1}{t}$.

From a feasible solution of (5.2.2) with objective value $\frac{1}{t}$ we can construct a feasible solution of (5.3.2) with objective value t. Conversely, from a feasible solution of (5.3.2) with objective value t we can construct a feasible solution of (5.2.2) with objective value $\frac{1}{t}$. This is enough to guarantee that the optimum of (5.3.2) is 1 divided by the optimum of (5.2.2), *i.e.*, it is $\frac{1}{\frac{1}{\alpha(G)}} = \alpha(G)$.

5.3.2 Sum-of-squares relaxations of the copositive formulation of $\alpha(G)$

5.3.2.1 The copositive formulation of $\alpha(G)$

The dual of (5.3.1) can be calculated as in the SDP case, see also the description of primal-dual conic programs from Section 7.2.2. We can apply the technique used in the proof of Prop. 2.1.5, but we need the dual cone of C^{n*}, *i.e.*, C^n. The dual of (5.3.1) can thus be written as follows:

$$\alpha(G) = \min\left\{t:\ t\left(A^G + I_n\right) - \mathbf{e}_n\mathbf{e}_n^\top \in C^n\right\}. \tag{5.3.3}$$

It is possible to show that both (5.3.1) and (5.3.3) are strictly feasible. Let us start with (5.3.1) and notice that if $Y \in \texttt{interior}\left(C^{n*}\right)$, then $\texttt{diag}(Y) \neq 0$ and $\left(A^G + I_n\right) \bullet Y \geq \texttt{diag}(Y) \bullet \texttt{diag}(I_n) > 0$. As such, we can define $X = \frac{Y}{(A^G+I_n)\bullet Y}$ that also belongs to $\texttt{interior}\left(C^{n*}\right)$ and is feasible in (5.3.1). We only need to show that C^{n*} has a non-empty interior, *i.e.*, we have to generate some strictly feasible matrix of C^{n*}. For this, consider the set $\mathcal{A}^{(0,1)} = \{nI_n + A \in S_n:\ 0 < A_{ij} < 1\}$. For any $nI_n + A \in \mathcal{A}^{(0,1)}$ and $M \in S_n$, there is a sufficiently small $\varepsilon > 0$ such that $nI_n + A - \varepsilon M,\ nI_n + A + \varepsilon M \in \mathcal{A}^{(0,1)}$. Because of this last property, to prove $\mathcal{A}^{(0,1)} \subsetneq \texttt{interior}\left(C^{n*}\right)$ it is now enough to show $\mathcal{A}^{(0,1)} \subsetneq C^{n*}$. We can write

$$\mathcal{A}^{(0,1)} \ni nI_n + A = \sum_{i<j} A_{ij}E^{ij} + \sum_{i\in[1..n]}\left(n + A_{ii} - \sum_{j\neq i} A_{ij}\right)E^{ii}, \tag{5.3.4}$$

where E^{ij} is a matrix full of zeros except at positions (i,i), (i,j), (j,i) and (j,j) where it has ones. All terms in above (5.3.4) can be written under the form $\mathbf{y}\mathbf{y}^\top$ for a non-negative $\mathbf{y}$. For $A_{ij}E^{ij}$, it is enough to take a $\mathbf{y}$ full of zeros except for $y_i = y_j = \sqrt{A_{ij}}$. For $\left(n + A_{ii} - \sum_{j\neq i} A_{ij}\right)E^{ii}$, we take a $\mathbf{y}$ full of zeros, except for $y_i = \sqrt{n + A_{ii} - \sum_{j\neq i} A_{ij}} > \sqrt{n + A_{ii} - (n-1)} > 0$. This allows one to write $nI_n + A$ in the form required by the C^{n*} definition (5.1.1).

We now show (5.3.3) has strictly feasible solutions. It is enough to show there is a sufficiently large t such that $\left(t\left(A^G + I_n\right) - \mathbf{e}_n\mathbf{e}_n^\top\right) \bullet \mathbf{y}\mathbf{y}^\top > 0$ for any non-negative $\mathbf{y} \geq \mathbf{0}$. Using $\left(t\left(A^G + I_n\right) - \mathbf{e}_n\mathbf{e}_n^\top\right) \bullet \mathbf{y}\mathbf{y}^\top \geq \left(tI_n - \mathbf{e}_n\mathbf{e}_n^\top\right) \bullet \mathbf{y}\mathbf{y}^\top$, it suffices to show that this last scalar product is strictly positive if t is large enough. But this simply follows from the fact that the matrix $tI_n - \mathbf{e}_n\mathbf{e}_n^\top$ becomes diagonally dominant and can have arbitrarily large eigenvalues when $t \to \infty$, *i.e.*, we can easily have $tI_n - \mathbf{e}_n\mathbf{e}_n^\top \succ \mathbf{0}$.

Since both (5.3.1) and (5.3.3) are strictly feasible, we can apply the strong duality Theorem 7.2.8 for linear conic programming that states that both programs have the same optimum value and they effectively reach this value.

5.3.2.2 A "natural" strengthening bounded by the Lovász number

There are several cone hierarchies and relaxations used to approximate the copositive cone C^n. Let us first recall hierarchy (5.1.4) and define a "natural" strengthening (more constrained restricted version) that replaces C^n with $S_n^+ + \mathcal{N}^n$, where $\mathcal{N}^n \subsetneq S_n$ is the set of non-negative symmetric matrices. The strengthening can simply be written by modifying (5.3.3):

$$\alpha^0(G) = \min\left\{t:\ t\left(A^G + I_n\right) - \mathbf{e}_n\mathbf{e}_n^\top \in S_n^+ + \mathcal{N}^n\right\} \quad [\geq \alpha(G)]\,. \tag{5.3.5}$$

Since the dual cone of $S_n^+ + \mathcal{N}^n$ is $S_n^+ \cap \mathcal{N}^n$,[25] the dual can be written as follows:

$$\alpha^0(G) = \max\left\{\mathbf{e}_n\mathbf{e}_n^\top \bullet X:\ \left(A^G + I_n\right) \bullet X = 1,\ X \in S_n^+ \cap \mathcal{N}^n\right\}. \tag{5.3.6}$$

[25]To see this, simply notice $S_n^+ \subsetneq S_n^+ + \mathcal{N}^n$, and so, $\left(S_n^+ + \mathcal{N}^n\right)^* \subsetneq (S_n^+)^* = S_n^+$. Similarly, $\left(S_n^+ + \mathcal{N}^n\right)^* \subsetneq (\mathcal{N}^n)^* = \mathcal{N}^n$. As such, $\left(S_n^+ + \mathcal{N}^n\right)^* \subset S_n^+ \cap \mathcal{N}^n$. One can check that $X \in S_n^+ \cap \mathcal{N}^n$ and $Y = Y_1 + Y_2$ with $Y_1 \in S_n^+$ and $Y_2 \in \mathcal{N}^n$ yield $X \bullet (Y_1 + Y_2) = X \bullet Y_1 + X \bullet Y_2 \geq 0$, using the fact that X is both SDP and non-negative.

The first program (5.3.5) is strictly feasible by taking a sufficiently large t, so that $t\left(A^G+I_n\right)-\mathbf{e}_n\mathbf{e}_n^\top = \underbrace{tI_n-\mathbf{e}_n\mathbf{e}_n^\top}_{\succ\mathbf{0}}+\underbrace{tA^G}_{\geq\mathbf{0}}$ and $t\left(A^G+I_n\right)-\mathbf{e}_n\mathbf{e}_n^\top\pm\varepsilon M=\underbrace{tI_n-\mathbf{e}_n\mathbf{e}_n^\top\pm\varepsilon M}_{\succ\mathbf{0}}+\underbrace{tA^G}_{\geq\mathbf{0}}$ using a sufficiently small $\varepsilon>0$ for any $M\in S_n$. The dual (5.3.6) is strictly feasible taking $X=\frac{tI_n+\mathbf{e}_n\mathbf{e}_n^\top}{(A^G+I_n)\bullet(tI_n+\mathbf{e}_n\mathbf{e}_n^\top)}$ for a sufficiently large t. We can apply the strong duality Theorem 7.2.8 for linear conic programming to state that both programs have the same optimum value $\alpha^0(G)$ and they effectively reach this value.

We now prove the following:

$$\alpha^0(G)=\min\Big\{t:\ tI_n+\sum_{\{i,j\}\in E:i<j}t_{ij}E_{ij}-\mathbf{e}_n\mathbf{e}_n^\top\in S_n^+ +\mathcal{N}^n\Big\},\tag{5.3.7}$$

where t_{ij} are decision variables and E_{ij} is a matrix that contains a value of one at positions (i,j) and (j,i) and zeros everywhere else. This is a relaxation of (5.3.5) because it lifts the constraints $t_{ij}=t\ \forall\{i,j\}\in E$. However, we can show that any feasible solution of (5.3.7) with objective value t can be transformed into a feasible solution of (5.3.5) with the same objective value t.

Take any edge $\{i,j\}\in E$ such that $t_{ij}\neq t$. If $t_{ij}<t$, one can simply increase t_{ij} to t and remain feasible because the resulting matrix is the old matrix plus a non-negative increase that belongs to $\mathcal{N}^n$. If $t_{ij}>t$, let us decrease t_{ij} to the minimum value $\bar{t}$ that keeps the resulting matrix in $S_n^+ +\mathcal{N}^n$. Let us focus on the 2×2 minor corresponding to i and j. If $\bar{t}>t$, this minor is not SDP, and so, a decrease from $\bar{t}$ down towards t would only represent a decrease of the $\mathcal{N}^n$ component the matrix in $S_n^+ +\mathcal{N}^n$. The only possible case that can forbid any decrease below $\bar{t}$ is $\bar{t}=t$.

As a relaxation of (5.3.5), the new program (5.3.7) remains strictly feasible. This means that the following dual does achieve the optimum value $\alpha^0(G)$.

$$\alpha^0(G)=\max\Big\{\mathbf{e}_n\mathbf{e}_n^\top\bullet X:\ \left(A^G+I_n\right)\bullet X=1,\ X_{ij}=0\ \forall\{i,j\}\in E,\ X\in S_n^+\cap\mathcal{N}^n\Big\}\tag{5.3.8}$$

$$=\max\Big\{\mathbf{e}_n\mathbf{e}_n^\top\bullet X:\ I_n\bullet X=1,\ X_{ij}=0\ \forall\{i,j\}\in E,\ X\succeq\mathbf{0},X\geq\mathbf{0}\Big\}.\tag{5.3.9}$$

It is clear that above (5.3.9) *without the non-negativity* constraint $X\geq 0$ is equivalent to the $(D\vartheta_G)$ formulation from (4.1.3a)-(4.1.3d). As such, $\alpha^0(G)\leq OPT(D\vartheta_G)=\vartheta(G)$. This "natural" strengthening of the copositive formulation of the maximum stable is bounded by the Lovász theta number. Since we have $\alpha(G)\leq\alpha^0(G)$ in (5.3.5), we can write $\alpha(G)\leq\alpha^0(G)\leq\vartheta(G)$.

5.3.2.3 The sum of squares (SOS) hierarchy

This is the only subsection where I will use a few results without an explicit proof. In all cases, I will leave a comment on the margin of the document.

This subsection is devoted to the SOS approach due to Parilo, De Klerk and Pasechnik as cited in the "Copositive Optimization" article of the *Optima 89* newsletter (see Footnote 22, p. 50) or in the article "Semidefinite Bounds for the Stability Number of a Graph via Sums of Squares of Polynomials" by Nebojša Gvozdenović and Monique Laurent.[26]

The general SOS setting

First, notice that $M\in C^n$ (recall Def. (5.1.3)) $\iff$ $p_M(\mathbf{x})=M\bullet[x_1^2\ x_2^2\ \dots x_n^2]^\top[x_1^2\ x_2^2\ \dots x_n^2]=\sum_{i,j=1}^n M_{ij}x_i^2x_j^2\geq 0\ \forall\mathbf{x}\in\mathbb{R}^n$. This means that the polynomial $p_M(\mathbf{x})$ takes only non-negative values over the reals. We can start using results related to non-negative polynomials. There are at least two conditions that guarantee that a polynomial is always non-negative: (i) all coefficients are positive and all powers of the variables are even, (ii) it can be written as a sum of squares (SOS). One of the earliest work on these aspects dates back to Hilbert who studied the classes of non-negative polynomials that can always be written as an SOS. He proved, for instance, that non-negative univariate polynomials ($n=1$) and non-negative polynomials of degree 2 can always be decomposed into a SOS, see also Hilbert's 17[th] problem. However, even if a non-negative polynomial does not respect any of above conditions (i) or (ii), we could multiply it by another non-negative polynomial and obtain a product that does respect (i) or (ii).

[26] Published in *Mathematical Programming* in 2007, vol 110(1), pp 145–173, available on-line as of 2017 at http://oai.cwi.nl/oai/asset/11672/11672D.pdf.

Applying this last argument, we notice that $p_M(\mathbf{x})$ is non-negative if and only if

$$p_M^{(r)} = \left(\sum_{i=1}^{n} x_i^2\right)^r p_M(\mathbf{x}) \tag{5.3.10}$$

is non-negative. Pólya proved in the 1970s the following theorem.

Theorem 5.3.1. *If $p_M(\mathbf{x})$ is a homogeneous polynomial (with all terms of the same degree) that is strictly positive over $\mathbb{R}^n - \{0\}$, then $p_M^{(r)}$ has only non-negative coefficients for a sufficiently large r.*

Considering our p_M above of the form $p_M(\mathbf{x}) = M \bullet [x_1^2\ x_2^2\ \dots x_n^2]^\top [x_1^2\ x_2^2\ \dots x_n^2]$, we notice that $p_M^{(r)}$ accepts an obvious SOS decomposition for a sufficiently large r, *i.e.*, all terms of $p_M^{(r)}$ are squares multiplied by a positive value. We now can define the cone:

Missing: a proof for the result of Pólya (Theorem 5.3.1).

$$\mathcal{K}_n^{(r)} = \{M \in S_n :\ p_M^{(r)} \text{ can be written as an SOS}\}. \tag{5.3.11}$$

Using Theorem 5.3.1, we have $\texttt{interior}(C^n) = \bigcup\limits_{r\geq 1} \mathcal{K}_n^{(r)}$. We also have $\mathcal{K}_n^{(i)} \subseteq K_n^{(i+1)}$ because switching from i to $i+1$ is equivalent to multiplying an SOS form with $\sum_{i=1}^n x_i^2$, which develops into a sum of squared (and thus SOS) terms. Replacing C^n with $\mathcal{K}_n^{(r)}$ in (5.3.3), we obtain the following strengthening (restriction) of (5.3.3):

$$\alpha^{(i)}(G) = \min\left\{t :\ t\left(A^G + I_n\right) - \mathbf{e}_n\mathbf{e}_n^\top \in \mathcal{K}_n^{(i)}\right\} \tag{5.3.12}$$

and we naturally have

$$\alpha^{(0)}(G) \geq \alpha^{(1)}(G) \geq \alpha^{(2)}(G) \cdots \geq \alpha^{(r)}(G) \to \alpha(G),$$

using $\mathcal{K}_n^{(i)} \subseteq K_n^{(i+1)}$, with the convergence following from $\texttt{interior}(C^n) = \lim\limits_{r\to\infty} \mathcal{K}_n^{(r)}$. As such, we easily have $\alpha(G) = \left\lfloor \alpha^{(r)}(G) \right\rfloor$ for a sufficiently large r. A sufficient value of r is $r = \alpha(G)^2$, as proved in Theorem 4.1. of the article "Approximation of the stability number of a graph via copositive programming" by Etienne de Klerk and Dmitri Pasechnik.[27] It is conjectured that $\alpha^{(r)}(G) = \alpha(G)$ for $r \geq \alpha(G) - 1$, see also Conjecture 1 in the article indicated at Footnote 26 (p. 54).

Missing: the proof for the sufficiency of $r = \alpha(G)^2$.

We will show that optimizing over some $\mathcal{K}_n^{(i)}$ might be easier than optimizing over C^n, particularly for bounded values of i. While checking membership in C^n is NP-hard and probably not NP (assuming NP≠co-NP, see last paragraph of Section 5.2.2), checking membership in $K_n^{(i)}$ requires solving an SDP with $\binom{n+i+1}{i+2} \times \binom{n+i+1}{i+2}$ variables, as we will see later in Remark 5.3.3. More generally, we will show one can determine an SOS decomposition of any polynomial of degree $2d$ by solving and SDP with at most $\binom{n+d}{d} \times \binom{n+d}{d}$ variables, using results from the Phd thesis "Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization" by Pablo Parrilo.[28]

Characterizing $K_n^{(0)}$

Let us start with $K_n^{(0)}$ and we will show that $\alpha^{(0)}(G)$ is actually equal to the natural strengthening $\alpha^0(G)$ from (5.3.5) which is bounded by the Lovász theta number $\vartheta(G)$ as described in Subsection 5.3.2.2. However, let us first present a small example.

Example 5.3.2. *Consider $p_M^{(0)}(\mathbf{x}) = p_M(\mathbf{x}) = x_1^4 + x_2^4 + 3(x_1x_2)^2$ that accepts multiple SOS decompositions, e.g., $(x_1^2 + x_2^2)^2 + (x_1x_2)^2$ or $(x_1^2)^2 + (x_2^2)^2 + (x_1x_2)^2$. Such decompositions can be determined by solving an SDP with a null objective function (SDP feasibility problem).*

Proof. We will later show (see Remark 5.3.3) that for such polynomial, the only terms that can appear in the squared expressions are x_1^2, x_2^2 and x_1x_2. Let us define $\overline{\mathbf{x}}^\top = [x_1^2,\ x_2^2,\ x_1x_2]$ and write $p_M(\mathbf{x}) = \overline{M} \bullet \overline{\mathbf{x}}\ \overline{\mathbf{x}}^\top$, with $\overline{M} \in S_3$.

[27] Published in *SIAM journal on optimization* in 2002, vol 12(4), pp. 875–892, available on-line as of 2017 at https://dr.ntu.edu.sg/handle/10220/6790.

[28] Defended in 2000 at California Institute of Technology, available on-line as of 2017 at http://www.mit.edu/~parrilo/pubs/files/thesis.pdf.

Step 1 This last equality $p_M(\mathbf{x}) = \overline{M} \bullet \overline{\mathbf{x}}\, \overline{\mathbf{x}}^\top$ does not allow one to uniquely identify $\overline{M}$. This comes from the fact that $(x_1x_2)^2$ can be written both as a square of x_1x_2 (involving a non-zero $\overline{M}_{33}$ in this last equality) and as a product of x_1^2 and x_2^2 (involving $\overline{M}_{12}$). We can write $p_M(\mathbf{x}) = p_M(\mathbf{x}) + \lambda_{12}\left(-x_1^2x_2^2 + (x_1x_2)^2\right)$. This way, $\overline{M}$ can take the form

$$\overline{M}_{\lambda_{12}} = \begin{bmatrix} 1 & \frac{3-\lambda_{12}}{2} & 0 \\ \frac{3-\lambda_{12}}{2} & 1 & 0 \\ 0 & 0 & \lambda_{12} \end{bmatrix}.$$

Step 2 Assume there is an SOS decomposition $p_M(\mathbf{x}) = \sum_{i=1}^k (\mathbf{v}_i \bullet \overline{\mathbf{x}})^2 = \sum_{i=1}^k \mathbf{v}_i\mathbf{v}_i^\top \bullet \overline{\mathbf{x}}\, \overline{\mathbf{x}}^\top = VV^\top \bullet \overline{\mathbf{x}}\, \overline{\mathbf{x}}^\top$. This can only happen if $\overline{M} = VV^\top$ is SDP (see also Prop. A.1.8). The above decomposition $(x_1^2 + x_2^2)^2 + (x_1x_2)^2$ corresponds to $\lambda_{12} = 1$. We can find a decomposition by solving an SDP program $\overline{M}_{\lambda_{12}} \succeq \mathbf{0}$ with null objective function, *i.e.*, we only want to know if there exists feasible $\overline{M}_{\lambda_{12}} \succeq \mathbf{0}$.

Both steps can generate infinitely-many SOS decompositions. At Step 1, there is a feasible matrix $\overline{M}_{\lambda_{12}}$ for each $\lambda_{12} \in R$. At Step 2, any SDP matrix $\overline{M}_{\lambda_{12}}$ accepts infinitely many factorizations $\overline{M} = VV^\top$, as stated in Corollary 1.7.1. □

Let us return to general case of the polynomial $p_M(\mathbf{x})$ of (homogeneous) degree $2d = 4$, *i.e.*, all terms have degree $d = 4$. The squared expressions contain no free terms because $p_M(\mathbf{x})$ has no free term. They can neither contain terms of degree $d - 1 = 1$ like x_i because there is no way other squared expression cancel some x_i^2. We define $\overline{\mathbf{x}}$ to contain all monomials of degree $d = 2$ and we attempt to write $p_M(\mathbf{x}) = \overline{M} \bullet \overline{\mathbf{x}}\, \overline{\mathbf{x}}^\top$. As already stated in the proof of above example, an SOS decomposition $p_M(\mathbf{x}) = \sum_{i=1}^k (\mathbf{v}_i \bullet \overline{\mathbf{x}})^2 = \sum_{i=1}^k \mathbf{v}_i\mathbf{v}_i^\top \bullet \overline{\mathbf{x}}\, \overline{\mathbf{x}}^\top = VV^\top \bullet \overline{\mathbf{x}}\, \overline{\mathbf{x}}^\top$ exists if and only there is some SDP $\overline{M} = VV^\top$ such that $p_M(\mathbf{x}) = \overline{M} \bullet \overline{\mathbf{x}}\, \overline{\mathbf{x}}^\top$. All expressions of the form $p_M(\mathbf{x}) = \overline{M} \bullet \overline{\mathbf{x}}\, \overline{\mathbf{x}}^\top$ can be found by writing

$$\begin{aligned} p_M(\mathbf{x}) = p_M(\mathbf{x}) &+ \sum_{i<j} \lambda_{ij}\left(-(x_i^2)(x_j^2) + (x_ix_j)^2\right) \\ &+ \sum_{\substack{j<j' \\ i\neq j, i\neq j'}} \mu_{ijj'}\left((x_ix_j)(x_ix_{j'}) - (x_i^2)(x_jx_{j'})\right) \\ &+ \sum_{\substack{i'\neq j', j\neq i', j\neq j' \\ i<j, i<i', i<j'}} \mu_{ii'jj'}\left((x_ix_j)(x_{i'}x_{j'}) - (x_ix_{i'})(x_jx_{j'})\right), \end{aligned}$$

where "$i < j, i < i', i < j'$" in the last sum comes from the fact that (i) permuting i and j is equivalent to permuting i' and j', (ii) permuting i and i' is equivalent to permuting j and j', and (iii) permuting i and j' is equivalent to permuting i' and j. A feasible $\overline{M}$ can take the following form:

$$\overline{M}_{\boldsymbol{\lambda},\boldsymbol{\mu}} = \begin{bmatrix} M_{11} & M_{12}-\lambda_{12} & \dots & M_{1n}-\lambda_{1n} & * & * & \dots & * \\ M_{12}-\lambda_{12} & M_{22} & \dots & M_{2n}-\lambda_{2n} & * & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ M_{n1}-\lambda_{n1} & M_{n2}-\lambda_{n2} & \dots & M_{nn} & * & * & \dots & * \\ * & * & \dots & * & 2\lambda_{12} & * & \dots & * \\ * & * & \dots & * & * & 2\lambda_{13} & \dots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ * & * & \dots & * & * & * & \dots & 2\lambda_{(n-1),n} \end{bmatrix}, \tag{5.3.13}$$

where the asterisks stand for null terms or linear combinations of the $\boldsymbol{\mu}$ variables. However, if $\overline{M}_{\boldsymbol{\lambda},\boldsymbol{\mu}} \succeq \mathbf{0}$ for some $\boldsymbol{\mu} \neq \mathbf{0}$, then $\overline{M}_{\boldsymbol{\lambda},\mathbf{0}} \succeq \mathbf{0}$, because any minor of $\overline{M}_{\boldsymbol{\lambda},\mathbf{0}}$ (replace above asterisks with zeros) can be seen as a diagonal of blocks that also appear in the minors of $\overline{M}_{\boldsymbol{\lambda},\boldsymbol{\mu}}$ if $\boldsymbol{\mu} \neq \mathbf{0}$. All these blocks need to have a non-negative determinant in both matrices. As such, we can find an SOS decomposition if and only if $\overline{M}_{\boldsymbol{\lambda},\mathbf{0}} \succeq \mathbf{0}$.

We now investigate the relation between the above $\overline{M}_{\boldsymbol{\lambda},\mathbf{0}}$ and the initial matrix M that defines $p_M(\mathbf{x}) = M \bullet [x_1^2\ x_2^2\ \dots x_n^2]^\top [x_1^2\ x_2^2\ \dots x_n^2]$. We obtain relatively easily that if $\overline{M}_{\boldsymbol{\lambda},\mathbf{0}} \succeq \mathbf{0}$, then M can be written as an SDP matrix (the leading principal minor of size $n \times n$ of $\overline{M}_{\boldsymbol{\lambda},\mathbf{0}} \succeq \mathbf{0}$) plus a non-negative matrix filled with all the λ_{ij} values. In other words, $M \in S_n^+ + \mathcal{N}^n$, and so, $K_n^{(0)} \subseteq S_n^+ + \mathcal{N}^n$.

We can also show $S_n^+ + \mathcal{N}^n \subseteq K_n^{(0)}$. This follows from the fact that if $S_n^+ + \mathcal{N}^n \ni M = [M]' + [\boldsymbol{\lambda}]'$ with $[M]' \succeq \mathbf{0}$ and $[\boldsymbol{\lambda}]' \geq \mathbf{0}$, we can also write M as the sum of an SDP matrix and a non-negative matrix $[\boldsymbol{\lambda}]$ such that $\texttt{diag}([\boldsymbol{\lambda}]) = \mathbf{0}$. For this, let us (re-)write $M = \underbrace{[M]' + \texttt{diag}([\boldsymbol{\lambda}]')}_{[M]\succeq \mathbf{0}} + \underbrace{[\boldsymbol{\lambda}]' - \texttt{diag}([\boldsymbol{\lambda}]')}_{[\boldsymbol{\lambda}]\geq \mathbf{0}} = [M] + [\boldsymbol{\lambda}]$ such that $[M] \succeq \mathbf{0}$, $[\boldsymbol{\lambda}] \geq \mathbf{0}$ and $[\boldsymbol{\lambda}]$ has zeros on the diagonal. We can thus construct a matrix $\overline{M}_{\boldsymbol{\lambda},\mathbf{0}} \succeq \mathbf{0}$ as above from these two matrices $[M]$ and $[\boldsymbol{\lambda}]$. This means $p_M(\mathbf{x})$ is an SOS, so that $M \in K_n^{(0)}$. We have just proved that $S_n^+ + \mathcal{N}^n \subseteq K_n^{(0)}$ so that $K_n^{(0)} = S_n^+ + \mathcal{N}^n$.

We obtained that $\alpha^{(0)}(G)$ and $\alpha^0(G)$ have the same feasible area. This means that $\alpha^{(0)}(G)$ is the natural strengthening $\alpha^0(G)$ from (5.3.5) bounded by the theta number $\vartheta(G)$ as described in Subsection 5.3.2.2.

Studying $K_n^{(i)}$ and finding the SOS decomposition of $p_M^{(i)}$

Recall the definitions (5.3.10) and (5.3.11) of $p_M^{(i)}$ and respectively $K_n^{(i)}$. We simply obtain that $p_M^{(i)}$ is a polynomial of degree $2d = 2(i+2)$. We can find an SOS decomposition of a polynomial of degree $2d$ using the following approach already exemplified above. For most general polynomials of degree $2d$, we define a column vector $\overline{\mathbf{x}}$ containing all monomials of degree less than or equal to d. We have that $p_M^{(i)}$ is an SOS if and only if we can write it $p_M^{(i)}(\mathbf{x}) = \overline{M} \bullet \overline{\mathbf{x}}\,\overline{\mathbf{x}}^\top$ with $\overline{M} \succeq \mathbf{0}$. We can write $\sum_{j=1}^k (\mathbf{v}_j \bullet \overline{\mathbf{x}})^2 = \sum_{j=1}^k \mathbf{v}_j \mathbf{v}_j^\top \bullet \overline{\mathbf{x}}\,\overline{\mathbf{x}}^\top = VV^\top \bullet \overline{\mathbf{x}}\,\overline{\mathbf{x}}^\top$ if and only if there exists $\overline{M} = VV^\top \succeq \mathbf{0}$ such that $p_M^{(i)} = \overline{M} \bullet \overline{\mathbf{x}}\,\overline{\mathbf{x}}^\top$. The number of (independent) squared expressions is equal to the rank of V which is equal to the rank of $\overline{M}$ (see Prop. A.1.8).

Checking membership in $K_n^{(i)}$ reduces to checking SDP membership for a matrix $\overline{M}$ of the form (5.3.13) but of much larger size and with more variables like $\boldsymbol{\lambda}$ or $\boldsymbol{\mu}$. For instance, a term like $x_i^2 \cdot M_{ij}(x_i)^2(x_j)^2 = M_{ij}x_i^4 x_j$ of $p_M^{(i)}(\mathbf{x})$ leads to a constraint $M_{ij} = \sum_{ab=x_i^4 x_j^2} \overline{M}_{a,b}$, where $\overline{M}_{a,b}$ indicates the $\overline{M}$ term corresponding to monomials a and b in $\overline{M}$. An optimization problem over $K_n^{(i)}$ can be transformed into an SDP problem by replacing an initial constraint of the form $M \in K_n^{(i)}$ into some $\overline{M} \succeq \mathbf{0}$, but where the size of $\overline{M}$ can be much larger than $n \times n$, recall $\overline{\mathbf{x}}$ contains all monomials of degree less than or equal to d.

5.3.2.4 The size of the SDP programs used for computing SOS decompositions

In the most general setting, the length of $\overline{\mathbf{x}}$ is given by the number of monomials of degree less than or equal to d. A classical combinatorics method known as "stars and bars" can determine this number as $\binom{n+d}{d}$; it relies on encoding any monomial using a string of stars and bars that arises quite out of the blue (*e.g.*, $x_1^2 x^3 x_4$ is $\star\star||\star|\star$). However, I prefer to show it using a different argument that I personally find more natural. Consider the following set $\{x_1,\ x_2, \dots x_n,\ c_1,\ c_2, \dots c_d\}$ of cardinal $n+d$. Choosing some $x_{i_1},\ x_{i_2}, \dots x_{i_k}$ amounts to building an initial expression $e = x_{i_1}x_{i_2}\dots x_{i_k}$ (or $e=1$ if no x_i term is chosen). Then, choosing some c_j amounts to the following: (i) if the j^{th} factor of e exists, copy (insert) it at position $j+1$ or (ii) if the j^{th} factor of e does not exist (e is shorter), do not modify e. Under this interpretation, any monomial of degree less than or equal to d corresponds to choosing d elements from $\{x_1,\ x_2, \dots x_n,\ c_1,\ c_2, \dots c_d\}$.

- x_1 corresponds to choosing $\{x_1,\ c_2,\ c_3, \dots c_d\}$. We start with expression $e = x_1$ and the copying elements $c_2,\ c_3, \dots c_d$ do not modify e because e has no factors at positions $2,\ 3, \dots d$.
- $x_1^2 x_2$ corresponds to choosing $\{x_1,\ x_2,\ c_1,\ c_4,\ c_5, \dots c_d\}$. Indeed, we first have the expression $e = x_1x_2$; then, c_1 will insert a copy of x_1 at position 2 to obtain $e = x_1x_1x_2$. The c_4 element does not change e, because e has no factor at position 4. Same applies to $c_5,\ c_6 \dots c_d$, and so, the final e is $e = x_1x_1x_2$.
- $x_5x_7^3x_9$ corresponds to choosing $\{x_5,\ x_7,\ x_9,\ c_2,\ c_3,\ c_6,\ c_7, \dots c_d\}$. We start with $e = x_5x_7x_9$ and c_2 inserts at position 3 a copy of the second element x_7, leading to $e = x_5x_7x_7x_9$. Then, c_3 duplicates the third element x_7, generating $e = x_5x_7x_7x_7x_9$. Elements $c_6,\ c_7, \dots c_d$ will perform no modification on e because e has no factor at positions $6,\ 7, \dots d$.
- x_7^d corresponds to choosing $\{x_7,\ c_1,\ c_2, \dots c_{d-1}\}$. Indeed, we start with $e = x_7$ and c_1 duplicates x_7 leading to x_7x_7. Then c_2 duplicates the second term, leading to $x_7x_7x_7$. Applying this for all $c_1,\ c_2, \dots c_{d-1}$, we obtain that x_7 is duplicated $d-1$ times, and so, the final expression is $\underbrace{x_7x_7\dots x_7}_{d \text{ times}}$.

- $x_7^2 x_9^{d-2}$ corresponds to choosing $\{x_7,\ x_9,\ c_1,\ c_3,\ c_4,\ \dots c_{d-1}\}$. We start with $e = x_7x_9$ and c_1 duplicates x_7 leading to $e = x_7x_7x_9$. Since c_2 is not chosen, x_7 is not duplicated again. On the other hand, x_9 is duplicated in cascade $d-3$ times and we obtain $e = x_7x_7\underbrace{x_9x_9x_9\dots x_9}_{d-2 \text{ times}}$.

- 1 corresponds to choosing $c_1,\ c_2, \dots c_d$.

Remark 5.3.3. *The polynomial $p_M^{(i)}$ of degree $2d = 2(i+2)$ also has the property that it contains only monomials of degree $2d$. In this case, we do not need to construct $\overline{\mathbf{x}}$ using all monomials of degree less than or equal to d, but we only need monomials of degree exactly d. The number of such monomials is*

$$\binom{n+d-1}{d} = \binom{n+i+1}{i+2}.$$

Proof. First, since $p_M^{(i)}$ has no free member, we can *not* use a squared expression with a free term, because *no* other squared expression can cancel a free term. We can repeat this argument by induction. Let us take any $d' < d$; the induction basis is that there is *no* term of degree strictly smaller than d' in any squared expression. Now notice we can not have a squared expression of the form $(e+x)^2$ with `degree(x)`$=d'$, because we can not cancel the monomial x^2. Indeed, such monomial could only be canceled by some $(e + x' - x'')^2$ with $x^2 = x'x''$ and the induction hypothesis states there are *no* terms such as x' or x'' of degree strictly smaller than d' in any squared expression. We proved by induction that there are *no* terms of degree $d' < d$ in our squared expressions, *i.e.*, $\overline{\mathbf{x}}$ is constructed using only monomials of degree d.

Investigating in detail the combinatorial argument above this proposition, we notice that expressions of degree exactly d are only constructed when c_d is not chosen (hint: let c_ℓ be the first chosen among the c_i's and notice there are at least ℓ elements chosen among the x_i's because none of $c_1,\ c_2, \dots c_{\ell-1}, c_d$ is chosen). The number of monomials of degree exactly d is thus $\binom{n+d-1}{d}$. $\square$

A final remark sometimes partially overlooked is that we took somehow for granted that we can *not* have an SOS decomposition with some terms of degree $\overline{d} > d$. However, this follows from the following argument. Take the maximum degree of a term in a squared expression and assume this value is $\overline{d} > d$. Assume there is an SOS decomposition $\sum (e_i + f_i)^2$, where the polynomials f_i's contain all terms of degree $\overline{d}$. The monomials of degree $2\overline{d}$ are those resulting from (summing up) the terms f_i^2. We need to have $\sum f_i^2 = 0$, which means *any* f_i yields $f_i^2 = 0$, and so, $f_i = 0$.

5.4 Further characterization of the completely positive and the copositive cones

We first recall below the following basic hierarchy from (5.1.4) of Section 5.1.

$$C^{n*} \subset S_n^+ \cap \mathcal{N}^n \subset S_n^+ \subset S_n^+ + \mathcal{N}^n \subset C^n, \tag{5.4.1}$$

We showed in Footnote 25 (p. 53) that $(S_n^+ + \mathcal{N}^n)^* = S_n^+ \cap \mathcal{N}^n$. We will now prove the converse: $(S_n^+ \cap \mathcal{N}^n)^* = S_n^+ + \mathcal{N}^n$. I could not easily find a direct proof using the properties of these matrices. However, it is direct consequence of the following general proposition combined with the fact that $S_n^+ + \mathcal{N}^n$ is closed (which follows from Prop. 2.3.1 and the fact that $\mathcal{N}^n$ is closed). More exactly, property below enables us to derive the following implication: $(S_n^+ + \mathcal{N}^n)^{**} = (S_n^+ \cap \mathcal{N}^n)^* \implies S_n^+ + \mathcal{N}^n = (S_n^+ \cap \mathcal{N}^n)^*$.

Proposition 5.4.1. *Given any convex cone C, we have `closure`$(C) = C^{**}$.*

Proof. We first show `closure`$(C) \subseteq C^{**}$. Assume the opposite: there is a sequence $(\mathbf{c}_i)$ with all $\mathbf{c}_i \in C$ and $\lim_{i\to\infty} \mathbf{c}_i = \mathbf{c} \notin C^{**}$. This means there is $\mathbf{c}^* \in C^*$ such that $\mathbf{c} \cdot \mathbf{c}^* = z < 0$. But since $\lim_{i\to\infty} \mathbf{c}_i = \mathbf{c}$, there exists some $\mathbf{c}_i$ sufficiently close to $\mathbf{c}$ such that $\mathbf{c}_i \cdot \mathbf{c}^* \in [z-\varepsilon, z+\varepsilon]$, for any $\varepsilon > 0$. For a sufficiently small ε, this means $\mathbf{c}_i \cdot \mathbf{c}^* < 0$ which is a contradiction of $\mathbf{c}^* \in C^*$ and $\mathbf{c}_i \in C$. As such, $\lim_{i\to\infty} \mathbf{c}_i = \mathbf{c} \notin C^{**}$ is false $\implies$ `closure`$(C) \subseteq C^{**}$.

We now show $\mathbf{b} \notin$ `closure`$(C) \implies \mathbf{b} \notin C^{**}$. We apply the simple separation Theorem C.4.5 on `closure`(C) and $\{\mathbf{b}\}$ to obtain there is a strictly separating inequality $\mathbf{c} \cdot \mathbf{c}^* > \mathbf{b} \cdot \mathbf{c}^*\ \forall \mathbf{c} \in$ `closure`(C). Since $\mathbf{0} \in C$, we have $\mathbf{0} \cdot \mathbf{c}^* > \mathbf{b} \cdot \mathbf{c}^* \implies \mathbf{b} \cdot \mathbf{c}^* < 0$. We now show $\mathbf{c}^* \in C^*$, *i.e.*, $\mathbf{c} \cdot \mathbf{c}^* \geq 0\ \forall \mathbf{c} \in C$. Assume $\exists \overline{\mathbf{c}} \in C$ such that $\overline{\mathbf{c}} \cdot \mathbf{c}^* < 0$. Using the cone property, $t\overline{\mathbf{c}} \in C$ for any arbitrarily large t. This means

$t\bar{\mathbf{c}} \bullet \mathbf{c}^*$ can be arbitrarily low, easily less than $\mathbf{b} \bullet \mathbf{c}^*$, which is a contradiction. The assumption $\exists \bar{\mathbf{c}} \in C$ such that $\bar{\mathbf{c}} \bullet \mathbf{c}^* < 0$ was false. All $\mathbf{c} \in C$ satisfy $\mathbf{c} \bullet \mathbf{c}^* \geq 0$. We produced $\mathbf{c}^* \in C^*$ such that $\mathbf{b} \bullet \mathbf{c}^* < 0$ for any $\mathbf{b}$ outside $\texttt{closure}(C)$. □

Using the results from Section 5.3.2.3, the above hierarchy (5.4.1) is refined to the hierarchy below. Recalling Pólya's Theorem 5.3.1, we have $\texttt{interior}(C^n) = \mathcal{K}_n^{(\infty)} = \lim_{i\to\infty} \mathcal{K}_n^{(i)}$.

$$C^{n*} \subset S_n^+ \cap \mathcal{N}^n \subset S_n^+ \subset S_n^+ + \mathcal{N}^n = \mathcal{K}_n^{(0)} \subset \mathcal{K}_n^{(1)} \subset \mathcal{K}_n^{(2)} \subset \dots \mathcal{K}_n^{(\infty)} \subset C^n, \tag{5.4.2}$$

We can also build an outer approximation hierarchy for C^n. Consider $\mathbb{N}_r^n = \left\{ \mathbf{z} \in \mathbb{N}^n : \sum_{j=1}^n z_i \leq r \right\}$ and define $\mathscr{P}_n^{(r)} = \left\{ X \in S_n : X \bullet \mathbf{z}\mathbf{z}^\top \geq 0 \ \forall \mathbf{z} \in \mathbb{N}_r^n \right\}$.[29] Since $\mathscr{P}_n^{(r)}$ contains less constraints than C^n, we simply obtain $C^n \subset \mathscr{P}_n^{(r)} \ \forall r \in \mathbb{N}$. Notice also $\mathbb{N}_r^n \subsetneq \mathbb{N}_{r+1}^n$, and so $C^n \subset \mathscr{P}_n^{(r+1)} \subset \mathscr{P}_n^{(r)} \forall r \in \mathbb{N}$. Writing $\mathscr{P}_n^{(\infty)} = \lim_{r\to\infty} \mathscr{P}_n^{(r)}$, we will prove that

$$C^n = \mathscr{P}_n^{(\infty)} \subset \cdots \subset \mathscr{P}_n^{(3)} \subset \mathscr{P}_n^{(2)} \subset \mathscr{P}_n^{(1)}$$

We still have to show the first equality; the other inclusions were proved in above paragraph. We know $C^n \subseteq \mathscr{P}_n^{(\infty)}$, because C^n contains more constraints than $\mathscr{P}_n^{(\infty)}$. We only need to show that $X \in \mathscr{P}_n^{(\infty)} \implies X \in C^n$. Assume the opposite: $\exists X \in \mathscr{P}_n^{(\infty)}$ such that $X \notin C^n$. This means there exists $\mathbf{y} \in \mathbb{R}_+^n$ such that $X \bullet \mathbf{y}\mathbf{y}^\top < 0$. Now we can construct a rational $\mathbf{y}_q$ such that $X \bullet \mathbf{y}_q\mathbf{y}_q^\top < 0$. Such a $\mathbf{y}_q$ can simply be constructed by taking sufficiently many decimals of of $\mathbf{y}$, enough to make $X \bullet \mathbf{y}_q\mathbf{y}_q^\top$ as close as necessary to $X \bullet \mathbf{y}\mathbf{y}^\top$. By multiplying rational $\mathbf{z}_q$ with the least common denominator, we obtain an *integer* $\mathbf{z}_i$ such that $X \bullet \mathbf{z}_i\mathbf{z}_i^\top < 0$, which contradicts $X \in \mathscr{P}_n^{(\infty)}$. The assumption $X \notin C^n$ was false, and so, $C^n = \mathscr{P}_n^{(\infty)}$.

The final remark is that all inclusions in the basic hierarchy (5.4.1) are strict. Indeed, $S_n^+ \cap \mathcal{N}^n \subsetneq S_n^+$ simply because $\left[\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right] \in S_2^+ - S_2^+ \cap \mathcal{N}^2$. It is also very easy to check $\left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ belongs to $S_2^+ + \mathcal{N}^2$, but not to S_2^+. We can derive $C^{n*} \subsetneq S_n^+ \cap \mathcal{N}^n$ from the fact that C^{n*} and $S_n^+ \cap \mathcal{N}^n$ have disjoint dual cones C^n and resp. $S_n^+ + \mathcal{N}^n$. Indeed, we can show $S_n^+ + \mathcal{N}^n \subsetneq C^n$ using the so-called Horn-matrix.[30]

$$H = \begin{bmatrix} 1 & -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 \\ -1 & 1 & 1 & -1 & 1 \end{bmatrix}$$

5.4.1 The Horn matrix is copositive

We will show this without decomposing $H \bullet \mathbf{x}\mathbf{x}^\top$ and calculating. We start from some $\mathbf{x} \geq 0$ and we investigate the evolution of function $H \bullet \mathbf{x}\mathbf{x}^\top$ as we change variables $\mathbf{x}$ in a particular way described next. The key is to notice that if we can decrease x_i and x_j by any $\varepsilon \in (0, min(x_i, x_j)]$, the value $H \bullet \mathbf{x}\mathbf{x}^\top$ decreases, for any $i, j \in [1..5]$ with $i \neq j$. Indeed, take any two lines i and j and notice they contain four times $\left[\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right]$ or $\left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$ and one times $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$. The contribution to $H \bullet \mathbf{x}\mathbf{x}^\top$ of a sub-matrix $\left[\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right]$ or $\left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$ is a multiple of $x_i - x_j$. Since we decrease *both variables* x_i and x_j by ε, the value $x_i - x_j$ does not change. On the other hand, the contribution of a sub-matrix $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$ is $x_\ell(x_i + x_j)$ for some $\ell \notin \{i, j\}$ that has to verify $x_\ell \geq 0$. Decreasing both x_i and x_j can only decrease $x_\ell(x_i + x_j)$. An analogous (transposed) phenomenon happens on the columns i and j. This way, decreasing x_i and x_j by the same $\varepsilon > 0$ can only decrease $H \bullet \mathbf{x}\mathbf{x}^\top$.

We can sequentially decrease pairs of non-zero variables x_i and x_j until we end up with a final $\mathbf{x}$ that contains only one non-zero variable and whose value $H \bullet \mathbf{x}\mathbf{x}^\top$ is less than or equal to the initial one. This finishes the proof, because this final value is a multiple of a diagonal element of H that is surely non-negative.

[29] Some papers define $\mathbb{N}_r^n$ using $\sum_{j=1}^n z_i = r$. However, as far as I can see, this does no longer ensures the property $\mathscr{P}_n^{(r+1)} \subset \mathscr{P}_n^{(r)}$ that we need afterwords. We would have $X = \left[\begin{smallmatrix} 7 & -8 & 0 \\ -8 & 7 & 0 \\ 0 & 0 & 2 \end{smallmatrix}\right] \in \mathscr{P}_3^{(3)}$ but $X \notin \mathscr{P}_3^{(2)}$.

[30] For an exact reference, check the comments on equation (7) from the "Copositive Optimization" article of the *Optima 89* newsletter, see the `http` link from Footnote 22, p. 50.

5.4.2 The Horn matrix does not belong to $S_5^+ + \mathcal{N}^5$

The easiest way to see H is not SDP is by noticing it does not respect the structure from Prop. 1.6.4. Indeed, the first four elements of the last row can not be written as a linear combination of the rows of $[H]_{[1..4]}$, because of columns 2 and 3, where $[H]_I$ is the principal minor of A that selects rows I and columns I. We still need to show $H - N$ is not SDP for any $N \geq \mathbf{0}$. Assume the contrary: there is $N \geq \mathbf{0}$ such that $H - N \succeq \mathbf{0}$ and we sequentially conclude:

(i) The operation $H - N$ can not decrease any diagonal element H_{ii}, because that would lead to $\det\left([H-N]_{\{i,j\}}\right) < 0$ for $j = i + 1 \texttt{ mod } 5$. Indeed, the product of the elements on the main diagonal of such $[H-N]_{\{i,j\}}$ would become less than 1 if $N_{ii} > 0$ or $N_{jj} > 0$, which is strictly less that the product of the elements of the second diagonal (remark $(H-N)_{ij} \leq H_{ij} = -1$). We thus need to have $\texttt{diag}(N) = \mathbf{0}$.

(ii) The operation $H - N$ can not decrease any H_{ij} for $j = i + 1 \texttt{ mod } 5$, because this would lead to $\det\left([H-N]_{\{i,j\}}\right) < 0$. Indeed, this would increase the product of the elements of the second diagonal of $[H-N]_{\{i,j\}}$ while the main diagonal does not change by virtue of $\texttt{diag}(N) = \mathbf{0}$ from above point (i). We obtain that all negative elements of H can not be decreased by the operation $H - N$.

(iii) Notice we have $H^3 = [H]_{[1..3]} = [H]_{[2..4]} = [H]_{[3..5]}$. Since we need such matrices H^3 to remain SDP after subtracting N, the extremal elements of the second diagonal (*i.e.*, H^3_{13} and H^3_{31}) need to remain unchanged. This follows from applying Prop. 1.6.4 and the fact that the other elements of any H^3 are not changed by the $H - N$ operation by virtue of (i) and (ii) above. We obtain $N_{13} = N_{24} = N_{25} = 0$ and the same applies to the transposed elements.

(iv) The only elements of N that can still be non-zero are N_{14}, N_{25} and their transposed, so that $|H - N|_{\{1,3,5\}} = H_{\{1,3,5\}}$. Since the determinant of this 3×3 matrix is -4, $H - N$ can not be SDP.

We obtained there is no $N \geq \mathbf{0}$ that can lead to $H - N \succeq \mathbf{0}$, and thus, $H \notin S_5^+ + \mathcal{N}^5$.

5.5 A final short property: the Schur complement does not apply in C^{n*}

I finish with a final remark not directly related to other property from this chapter. In certain proofs, I was tempted to use a "Schur complement property" with completely positive matrices but this is not possible.

Proposition 5.5.1. *The Schur complement property from Prop. 1.3.2 does not hold for completely positive matrices. In particular, if* $\begin{bmatrix} 1 & \mathbf{y}^\top \\ \mathbf{y} & Y \end{bmatrix} \in C^{(n+1)*}$, *we do not necessarily have* $Y - \mathbf{y}\mathbf{y}^\top \in C^{n*}$.

Proof. It is enough to give an example. We will exhibit a 2×2 matrix Y with some zeros and a vector $\mathbf{y}$ with no zero. It is enough to take the sum $[1\ 0\ a]^\top [1\ 0\ a] + [1\ a\ 0]^\top [1\ a\ 0]$ and divide it by 2 to obtain the completely positive matrix below.

$$\begin{bmatrix} 1 & a/2 & a/2 \\ a/2 & a^2/2 & 0 \\ a/2 & 0 & a^2/2 \end{bmatrix}$$

The non-diagonal terms of Y are zero. If we try to apply the Schur complement, we obtain negative entries on the non-diagonal terms of $Y - \mathbf{y}\mathbf{y}^\top$, so that this matrix can not be completely positive. $\square$

6 SDP relaxations and convexifications of quadratic programs

Let us give a short warning: an important difficulty in this section comes from parsing a number of (Lagrangian) notations. It may be useful to print this section (twice) to more easily jump from one formula or notation to another.

6.1 The most general quadratic program: SDP relaxation and total Lagrangian

We presented SDP reformulations and relaxations of *convex* quadratic programs in Section 3.4. We now introduce a quadratic program in its most general form, not necessarily convex. Notice this form could include $0-1$ quadratic programs by adding $x_i^2 = x_i \ \forall i \in [1..n]$ to the set of quadratic constraints (6.1.1b).

$$(QP)\begin{cases} \min \quad Q \bullet \mathbf{x}\mathbf{x}^\top + \mathbf{c}^\top \mathbf{x} & (6.1.1a) \\ s.t \quad B_i \bullet \mathbf{x}\mathbf{x}^\top + \mathbf{d}_i^\top \mathbf{x} = (\leq) e_i \quad \forall i \in [1..p] & (6.1.1b) \\ \quad \mathbf{x} \in \mathbb{R}^n & (6.1.1c) \end{cases}$$

The basic SDP relaxation of above (QP) program is the following:

$$SDP(QP)\begin{cases} \min \quad Q \bullet X + \mathbf{c}^\top \mathbf{x} & (6.1.2a) \\ s.t \quad B_i \bullet X + \mathbf{d}_i^\top \mathbf{x} = (\leq) e_i \quad \forall i \in [1..p] & (6.1.2b) \\ \quad \begin{bmatrix} 1 & \mathbf{x}^\top \\ \mathbf{x} & X \end{bmatrix} \succeq \mathbf{0} & (6.1.2c) \end{cases}$$

Let us start with a rather negative result: the SDP relaxation of $\min\{-x^2 : x = 0\} = 0$ is $\min\{-X_{11} : x = 0,\ X \succeq xx^\top = 0\} = -\infty$. This means that the above basic SDP relaxation can be arbitrarily bad in the worst case. However, we will overcome this with different strengthening methods in the next subsections.

Let us introduce the Total Lagrangian (TL) in variables $\mathbf{x}$ of (QP) from (6.1.1a)-(6.1.1c). We can write:

$$\left(TL^{\mathbf{X}}(QP)\right) = \max_{\boldsymbol{\mu}} \mathscr{L}_{TL^{\mathbf{X}}(QP)}(\boldsymbol{\mu}) \tag{6.1.3a}$$

$$\mathscr{L}_{TL^{\mathbf{X}}(QP)}(\boldsymbol{\mu}) = \min_{\mathbf{x}\in\mathbb{R}^n} \left(Q + \sum_{i=1}^{p} \mu_i B_i\right) \bullet \mathbf{x}\mathbf{x}^\top + \left(\mathbf{c}^\top + \sum_{i=1}^{p} \mu_i \mathbf{d}_i^\top\right)\mathbf{x} - \boldsymbol{\mu}^\top \mathbf{e} \tag{6.1.3b}$$

Notice we have $\mu_i \geq 0$ if we have an inequality constraint $B_i \bullet \mathbf{x}\mathbf{x}^\top + d_i^\top x \leq e_i$ for some $i \in [1..p]$.

Theorem 6.1.1. *The optimum total Lagrangian $\left(TL^{\mathbf{X}}(QP)\right)$ of a quadratic program (QP) is equal to the optimum value of the dual of the basic SDP relaxation $(SDP(QP))$ of (QP) from (6.1.2a)-(6.1.2c).*

Proof. To compute $\mathscr{L}_{TL^{\mathbf{X}}(QP)}(\boldsymbol{\mu})$ for a fixed $\boldsymbol{\mu}$, we can use the results for unconstrained quadratic programs from Section 3.4.3, more exactly Proposition 3.4.2 that can be applied on $\mathscr{L}_{TL^{\mathbf{X}}(QP)}(\boldsymbol{\mu})$ as follows:

$$\mathscr{L}_{TL^{\mathbf{X}}(QP)}(\boldsymbol{\mu}) = \max_t \quad t - \boldsymbol{\mu}\mathbf{e}$$
$$s.t. \quad \begin{bmatrix} -t & \frac{1}{2}\left(\mathbf{c}^\top + \sum_{i=1}^p \mu_i \mathbf{d}_i^\top\right) \\ \frac{1}{2}\left(\mathbf{c} + \sum_{i=1}^p \mu_i \mathbf{d}_i\right) & Q + \sum_{i=1}^p \mu_i B_i \end{bmatrix} \succeq \mathbf{0}.$$

Replacing this in (6.1.3a), we obtain:

$$\left(TL^{\mathbf{X}}(QP)\right) = \max_{t,\boldsymbol{\mu}} \quad t - \sum_{i=1}^{p} \mu_i \mathbf{e}_i \tag{6.1.4a}$$

$$s.t. \quad \begin{bmatrix} -t & \frac{1}{2}\left(\mathbf{c}^\top + \sum_{i=1}^p \mu_i \mathbf{d}_i^\top\right) \\ \frac{1}{2}\left(\mathbf{c} + \sum_{i=1}^p \mu_i \mathbf{d}_i\right) & Q + \sum_{i=1}^p \mu_i B_i \end{bmatrix} \succeq \mathbf{0}, \tag{6.1.4b}$$

which is exactly the dual of $SDP(QP)$ from (6.1.2a)-(6.1.2c). Recall that an inequality constraint $B_i \bullet \mathbf{x}\mathbf{x}^\top + d_i^\top \mathbf{x} \leq e_i$ leads to non-negative $\mu_i \geq 0$ in the total Lagrangian. When dualizing such μ_i of above (6.1.4a)-(6.1.4b), we obtain a inequality constraint in $SDP(QP)$, see also Prop. 2.1.3. However, we can certainly write $\left(TL^{\mathbf{X}}(QP)\right) = OPT(DUAL(SDP(QP)))$. One should bear in mind that there might be a duality gap between $(DUAL(SDP(QP)))$ and $(SDP(QP))$. □

The fact that $\left(TL^{\mathbf{X}}(QP)\right) \leq OPT(SDP(QP))$ can also be shown by noticing

$$\left(TL^{\mathbf{X}}(QP)\right) = \left(TL^{X}(QP)\right), \tag{6.1.5}$$

where $\left(TL^{X}(QP)\right)$ is the total Lagrangian of $(SDP(QP))$, *i.e.*, it is defined using the same formulas (6.1.3a)-(6.1.3b) but we replace the term $\mathbf{x}\mathbf{x}^\top$ from (6.1.3b) with X and add constraint $X \succeq \mathbf{x}\mathbf{x}^\top$ to (6.1.3b). To show $\mathscr{L}_{TL^{X}(QP)}(\boldsymbol{\mu}) = \mathscr{L}_{TL^{\mathbf{x}}(QP)}(\boldsymbol{\mu})$, it is enough to show there is an optimal solution $(X, \mathbf{x})$ of $\mathscr{L}_{TL^{X}(QP)}$ such that $X = \mathbf{x}\mathbf{x}^\top$. We consider two cases: (a) $Q + \sum_{i=1}^{p} \mu_i B_i \not\succeq \mathbf{0}$ and (b) $Q + \sum_{i=1}^{p} \mu_i B_i \succeq \mathbf{0}$. In the first case (a), we take an $\mathbf{x} \in \mathbb{R}^n$ such that $(Q + \sum_{i=1}^{p} \mu_i B_i) \bullet \mathbf{x}\mathbf{x}^\top < 0$. By replacing $\mathbf{x}$ with $t\mathbf{x}$ we obtain a quadratic concave function in t that converges to $-\infty$; by taking $X = \mathbf{x}\mathbf{x}^\top$ $\mathscr{L}_{TL^{X}(QP)}(\boldsymbol{\mu})$ converges to $-\infty$ as well. In the case (b), there is no use to consider any $X = Y + \mathbf{x}\mathbf{x}^\top$ with some non-zero $Y \succeq \mathbf{0}$: the existence of such Y can only increase the value of $\mathscr{L}_{TL^{X}(QP)}(\boldsymbol{\mu})$ by $(Q + \sum_{i=1}^{p} \mu_i B_i) \bullet Y \geq 0$. This confirms that $\left(TL^{\mathbf{x}}(QP)\right) = \left(TL^{X}(QP)\right) \leq OPT(SDP(QP))$.

The dual of $(SDP(QP))$ is actually constructed by relaxing all constraints from (6.1.2a)-(6.1.2c), *i.e.*, the proof of the general dualization property (Proposition 2.1.5) starts by computing in (2.1.11) the total SDP Lagrangian using coefficients $\mathbf{x}'$. These coefficients $\mathbf{x}'$ may include $\boldsymbol{\mu}$ and also the dual variable (t) of the constraint the puts a value of 1 in the upper right corner of the matrix from (6.1.2c). This explains the very close relationship between $DUAL(SDP(QP))$ and $\left(TL^{X}(QP)\right)$. The above argument could actually extend to a second proof of Theorem 6.1.1. Many convexification ideas presented next have certain roots in the fact that $\left(TL^{X}(QP)\right)$ is very related to $DUAL(SDP(QP))$.

6.2 Partial and total Lagrangians for quadratic programs with linear equality constraints

We consider a version of the general quadratic program (QP) from (6.1.1a)-(6.1.1c) in which we explicitly separate linear equality constraints $A\mathbf{x} = \mathbf{b}$. Furthermore, we consider adding $\overline{p}$ redundant constraints $\overline{B}_j \bullet \mathbf{x}\mathbf{x}^\top + \overline{\mathbf{d}}_j^\top \mathbf{x} = \overline{e}_j \ \forall j \in [1..\overline{p}]$ that are always satisfied for all $\mathbf{x} \in \mathbb{R}^n$ such that $A\mathbf{x} = \mathbf{b}$. An example of such redundant constraint is $A_j^\top A_j \bullet \mathbf{x}\mathbf{x}^\top = b_j^2$, where A_j is row j of A. By summing up over all $j \in [1..\overline{p}]$ we can also obtain $A^\top A \bullet \mathbf{x}\mathbf{x}^\top = \sum_{j=1}^{\overline{p}} b_j^2$. Many other examples can be found, *e.g.*, we will later see the redundant constraint sets from Example 6.3.4 and Example 6.3.5. However, for now, it is enough to say that such constraints have no impact on the initial quadratic program, but they can be useful to convexify it (*i.e.*, so that the factor of the quadratic term becomes SDP) or to strengthen its SDP relaxation.

We formulate

$$
(QP_=)\begin{cases}
\min\ \ Q \bullet \mathbf{x}\mathbf{x}^\top + \mathbf{c}^\top \mathbf{x} & (6.2.1a)\\
\ \ s.t\ \ A\mathbf{x} = \mathbf{b} & (6.2.1b)\\
\qquad \overline{B}_j \bullet \mathbf{x}\mathbf{x}^\top + \overline{\mathbf{d}}_j^\top \mathbf{x} = \overline{e}_j \qquad \forall j \in [1..\overline{p}] & (6.2.1c)\\
\qquad B_i \bullet \mathbf{x}\mathbf{x}^\top + \mathbf{d}_i^\top \mathbf{x} = (\leq) e_i \ \ \forall i \in [1..p] & (6.2.1d)\\
\qquad \mathbf{x} \in \mathbb{R}^n & (6.2.1e)
\end{cases}
$$

The formulation of the SDP relaxation leads to a program $SDP(QP_=)$ defined by writting the SDP form (6.1.2a)-(6.1.2c) corresponding to above $(QP_=)$. More exactly, we obtain the following SDP relaxation of $(QP_=)$.

$$
SDP\,(QP_=)\begin{cases}
\min\ \ Q \bullet X + \mathbf{c}^\top \mathbf{x} & (6.2.2a)\\
\ \ s.t\ \ A\mathbf{x} = \mathbf{b} & (6.2.2b)\\
\qquad \overline{B}_j \bullet X + \overline{\mathbf{d}}_j^\top \mathbf{x} = \overline{e}_j \qquad \forall j \in [1..\overline{p}] & (6.2.2c)\\
\qquad B_i \bullet X + \mathbf{d}_i^\top \mathbf{x} = (\leq) e_i \ \ \forall i \in [1..p] & (6.2.2d)\\
\qquad \begin{bmatrix} 1 & \mathbf{x}^\top \\ \mathbf{x} & X \end{bmatrix} \succeq \mathbf{0} & (6.2.2e)
\end{cases}
$$

6.2.1 The partial Lagrangians of $(QP_=)$ and $SDP(QP_=)$

We introduce below the Partial Lagrangian (PL) in variables $\mathbf{x}$ of $(QP_=)$.

$$(PL_{\mathbf{X}}(QP_=)) = \max_{\boldsymbol{\mu}} \mathscr{L}_{PL\mathbf{X}(QP_=)}(\boldsymbol{\mu}) \tag{6.2.3a}$$

$$\mathscr{L}_{PL\mathbf{X}(QP_=)}(\boldsymbol{\mu}) = \begin{cases} \min & (Q + \sum_{i=1}^{p} \mu_i B_i) \bullet \mathbf{x}\mathbf{x}^\top + \left(\mathbf{c}^\top + \sum_{i=1}^{p} \mu_i \mathbf{d}_i^\top\right) \mathbf{x} - \boldsymbol{\mu}^\top \mathbf{e} \\ s.t. & A\mathbf{x} = b \\ & \mathbf{x} \in \mathbb{R}^n \end{cases} \tag{6.2.3b}$$

This partial Lagrangian does not dualize the linear equality constraints $A\mathbf{x} = \mathbf{b}$ from (6.2.1b) and ignores the redundant constraints (6.2.1c). In other words, we obtain the partial Lagrangian of the initial program without redundant constraints. Notice that adding redundant constraints to above $\mathscr{L}_{PL\mathbf{X}(QP_=)}(\boldsymbol{\mu})$ would not change its value, because $\mathscr{L}_{PL\mathbf{X}(QP_=)}(\boldsymbol{\mu})$ already imposes $A\mathbf{x} = b$, and so, the redundant constraints are satisfied.

The above partial Lagrangian is equal to the following augmented partial Lagrangian that also dualizes the redundant constraints (6.2.1c).

$$(PL_{\mathbf{X}}(QP_=)) = \max_{\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}} \mathscr{L}_{PL\mathbf{X}(QP_=)}(\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}) \tag{6.2.4a}$$

$$\mathscr{L}_{PL\mathbf{X}(QP_=)}(\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}) = \begin{cases} \min & \left(Q + \sum_{i=1}^{p} \mu_i B_i + \sum_{j=1}^{\overline{p}} \overline{\mu}_j \overline{B}_j\right) \bullet \mathbf{x}\mathbf{x}^\top + \\ & \left(\mathbf{c}^\top + \sum_{i=1}^{p} \mu_i \mathbf{d}_i^\top + \sum_{j=1}^{\overline{p}} \overline{\mu}_j \overline{\mathbf{d}}_j^\top\right) \mathbf{x} \\ & -\boldsymbol{\mu}^\top \mathbf{e} - \overline{\boldsymbol{\mu}}^\top \overline{\mathbf{e}} \\ s.t. & A\mathbf{x} = b \\ & \mathbf{x} \in \mathbb{R}^n \end{cases} \tag{6.2.4b}$$

Notice we slightly abused notations, because we used $\mathscr{L}_{PL\mathbf{X}(QP_=)}$ both as a function with one argument $\boldsymbol{\mu}$ in (6.2.3b) or as an augmented function with two arguments $\boldsymbol{\mu}$ and $\overline{\boldsymbol{\mu}}$ in (6.2.4b) just above. However, it is not hard to check that the value of this partial Lagrangian only depends on the first argument, *i.e.*, since the $\overline{\boldsymbol{\mu}}$ terms are multiplied with (redundant) expressions that are equal to zero, we can write:

$$\mathscr{L}_{PL\mathbf{X}(QP_=)}(\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}) = \mathscr{L}_{PL\mathbf{X}(QP_=)}(\boldsymbol{\mu}) \ \forall \boldsymbol{\mu}, \overline{\boldsymbol{\mu}},$$

Let us now introduce the augmented partial Lagrangian in variables X (such that $X \succeq \mathbf{x}\mathbf{x}^\top$) of the SDP relaxation $(SDP\,(QP_=))$ from (6.2.2a)-(6.2.2e).

$$\left(PL^{X}(QP_=)\right) = \max_{\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}} \mathscr{L}_{PL^{X}(QP_=)}(\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}) \tag{6.2.5a}$$

$$\mathscr{L}_{PL^{X}(QP_=)}(\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}) = \begin{cases} \min & \left(Q + \sum_{i=1}^{p} \mu_i B_i + \sum_{j=1}^{\overline{p}} \overline{\mu}_j \overline{B}_j\right) \bullet X + \\ & \left(\mathbf{c}^\top + \sum_{i=1}^{p} \mu_i \mathbf{d}_i^\top + \sum_{j=1}^{\overline{p}} \overline{\mu}_j \overline{\mathbf{d}}_j^\top\right) \mathbf{x} \\ & -\boldsymbol{\mu}^\top \mathbf{e} - \overline{\boldsymbol{\mu}}^\top \overline{\mathbf{e}} \\ s.t. & A\mathbf{x} = b \\ & \begin{bmatrix} 1 & \mathbf{x}^\top \\ \mathbf{x} & X \end{bmatrix} \succeq \mathbf{0} \end{cases} \tag{6.2.5b}$$

The redundant constraints are dualized in objective function terms that are no longer redundant. In this relaxation, we have *no* guarantee that the term of $\overline{\mu}_j$ is zero, *i.e.*, we no longer have $\overline{B}_j \bullet X + \overline{\mathbf{d}}_j^\top \mathbf{x} - \overline{e}_j = 0 \ \forall j \in [1..\overline{p}]$, as the only constraint on X is $X \succeq \mathbf{x}\mathbf{x}^\top$. It is not hard to see that the above program is: (i) a Lagrangian relaxation of $(SDP\,(QP_=))$ and (ii) an SDP relaxation of $\mathscr{L}_{PL\mathbf{X}(QP_=)}(\boldsymbol{\mu}, \overline{\boldsymbol{\mu}})$. Combining (i) and (ii), we directly obtain:

$$\left(PL^{X}(QP_=)\right) \leq OPT(SDP\,(QP_=)), \tag{6.2.6a}$$

$$\left(PL^{X}(QP_=)\right) \leq \left(PL^{\mathbf{X}}(QP_=)\right). \tag{6.2.6b}$$

6.2.2 The total Lagrangian using equality constraints

The total Lagrangian can be expressed as in (6.1.3a)-(6.1.3b). However, technically, we now separate the dual variables $\boldsymbol{\beta}$ associated to the linear equality constraints. We also separate the dual variables $\overline{\boldsymbol{\mu}}$ associated with the redundant constraints. Thus, the total Lagrangian from (6.1.3a)-(6.1.3b) evolves to:

$$\left(TL^{\mathbf{X}}(QP_{=})\right) = \max_{\boldsymbol{\mu},\overline{\boldsymbol{\mu}},\boldsymbol{\beta}} \mathscr{L}_{TL^{\mathbf{X}}(QP_{=})}(\boldsymbol{\mu},\overline{\boldsymbol{\mu}},\boldsymbol{\beta}) \tag{6.2.7a}$$

$$\mathscr{L}_{TL^{\mathbf{X}}(QP_{=})}(\boldsymbol{\mu},\overline{\boldsymbol{\mu}},\boldsymbol{\beta}) = \begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^n} & \left(Q + \sum_{i=1}^{p}\mu_i B_i + \sum_{j=1}^{\overline{p}}\overline{\mu}_j\overline{B}_j\right)\bullet\mathbf{x}\mathbf{x}^\top \\ & +\left(\mathbf{c}^\top + \sum_{i=1}^{p}\mu_i\mathbf{d}_i^\top + \sum_{j=1}^{\overline{p}}\overline{\mu}_j\overline{\mathbf{d}}_j^\top + \boldsymbol{\beta}^\top A\right)\mathbf{x} \\ & -\boldsymbol{\mu}^\top\mathbf{e} - \overline{\boldsymbol{\mu}}^\top\overline{\mathbf{e}} - \boldsymbol{\beta}^\top\mathbf{b} \end{cases} \tag{6.2.7b}$$

We can define the SDP version of the above total Lagrangian by replacing $\mathbf{x}\mathbf{x}^\top$ with X in (6.2.7b) and by adding constraint $X \succeq \mathbf{x}\mathbf{x}^\top$. We thus construct the total Lagrangian $\mathscr{L}_{TL^{X}(QP)}(\boldsymbol{\mu},\overline{\boldsymbol{\mu}},\boldsymbol{\beta})$ in variables X (and $\mathbf{x}$). For any fixed $\boldsymbol{\mu}$ and $\overline{\boldsymbol{\mu}}$, the value $\max\limits_{\boldsymbol{\beta}}\mathscr{L}_{TL^{X}(QP)}(\boldsymbol{\mu},\overline{\boldsymbol{\mu}},\boldsymbol{\beta})$ can be seen as a Lagrangian of $\mathscr{L}_{PL^{X}(QP_{=})}(\boldsymbol{\mu},\overline{\boldsymbol{\mu}},)$. As such, the optimum SDP total Lagrangian $\left(TL^{X}(QP_{=})\right) = \max\limits_{\boldsymbol{\mu},\overline{\boldsymbol{\mu}},\boldsymbol{\beta}}\mathscr{L}_{TL^{X}(QP_{=})}(\boldsymbol{\mu},\overline{\boldsymbol{\mu}},\boldsymbol{\beta})$, satisfies:

$$\left(TL^{X}(QP_{=})\right) \leq \left(PL^{X}(QP_{=})\right). \tag{6.2.8}$$

Proposition 6.2.1. *The total Lagrangian* $\left(TL^{\mathbf{X}}_{\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*}(QP_{=})\right)$ *for* <u>*fixed* $\boldsymbol{\mu}^*$ *and* $\overline{\boldsymbol{\mu}}^*$</u> *(formally defined as the maximum of (6.2.11.a) below) is equal to OPT* $\left(\mathscr{L}_{PL^{X}(QP_{=})}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*)\right)$ *from* (6.2.5b). *This is enough to ensure:*

$$\left(TL^{\mathbf{X}}(QP_{=})\right) = \left(TL^{X}(QP_{=})\right) = \left(PL^{X}(QP_{=})\right). \tag{6.2.9}$$

Proof. Let us introduce notational shortcuts $Q_{\boldsymbol{\mu},\overline{\boldsymbol{\mu}}} = Q + \sum_{i=1}^{p}\mu_i B_i + \sum_{j=1}^{\overline{p}}\overline{\mu}_j\overline{B}_j$ and $\mathbf{c}^\top_{\boldsymbol{\mu},\overline{\boldsymbol{\mu}}} = \mathbf{c}^\top + \sum_{i=1}^{p}\mu_i\mathbf{d}_i^\top + \sum_{j=1}^{\overline{p}}\overline{\mu}_j\overline{\mathbf{d}}_j^\top$. Starting from the right term of (6.2.9), notice that $\left(\mathscr{L}_{PL^{X}(QP_{=})}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*)\right)$ from (6.2.5b) can be compactly written:

$$\mathscr{L}_{PL^{X}(QP_{=})}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*) = \begin{cases} \min & Q_{\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*}\bullet X + \mathbf{c}^\top_{\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*}\mathbf{x} - {\boldsymbol{\mu}^*}^\top\mathbf{e} - {\overline{\boldsymbol{\mu}}^*}^\top\overline{\mathbf{e}} \\ s.t. & A\mathbf{x} = b \\ & \begin{bmatrix} 1 & \mathbf{x}^\top \\ \mathbf{x} & X \end{bmatrix} \succeq \mathbf{0} \end{cases} \tag{6.2.10}$$

The total Lagrangian (6.2.7a)-(6.2.7b) with fixed $\boldsymbol{\mu}^*$ and $\overline{\boldsymbol{\mu}}^*$, can be written in compact notations using *only variables* $\boldsymbol{\beta}$ associated to the linear equality constraint $A\mathbf{x} = \mathbf{b}$.

$$\left(TL^{\mathbf{X}}_{\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*}(QP_{=})\right) =$$

$$= \begin{cases} \max\limits_{\boldsymbol{\beta}} & \mathscr{L}_{TL^{\mathbf{X}}(QP_{=})}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*,\boldsymbol{\beta}) \\ s.t. & \mathscr{L}_{TL^{\mathbf{X}}(QP_{=})}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*,\boldsymbol{\beta}) = \begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^n} & Q_{\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*}\bullet\mathbf{x}\mathbf{x}^\top + \left(\mathbf{c}^\top_{\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*} + \boldsymbol{\beta}^\top A\right)\mathbf{x} \\ & -{\boldsymbol{\mu}^*}^\top\mathbf{e} - {\overline{\boldsymbol{\mu}}^*}^\top\overline{\mathbf{e}} - \boldsymbol{\beta}^\top\mathbf{b} \end{cases} \end{cases} \tag{6.2.11a-b}$$

$$= \begin{cases} \max\limits_{t,\boldsymbol{\beta}} & t - {\boldsymbol{\mu}^*}^\top\mathbf{e} - {\overline{\boldsymbol{\mu}}^*}^\top\overline{\mathbf{e}} - \boldsymbol{\beta}^\top\mathbf{b} \\ s.t. & \begin{bmatrix} -t & \frac{\mathbf{c}^\top_{\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*} + \boldsymbol{\beta}^\top A}{2} \\ \frac{\mathbf{c}_{\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*} + A^\top\boldsymbol{\beta}}{2} & Q_{\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*} \end{bmatrix} \succeq \mathbf{0} \end{cases} \tag{6.2.11c-d}$$

where we applied Proposition 3.4.2 from Section 3.4.3, similarly to what we did for the total Lagrangian formulation (6.1.4a)-(6.1.4b). We obtained in (6.2.11c)-(6.2.11d) the dual of $\mathscr{L}_{PL^{X}(QP_{=})}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$ from (6.2.10). Since this last program is strictly feasible (it has no constraint on X other than $X \succeq \mathbf{x}\mathbf{x}^\top$), we can apply the strong duality Theorem 2.3.4 to conclude:

$$OPT\left(DUAL\left(\mathscr{L}_{PL^{X}(QP_{=})}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)\right)\right) = \left(TL^{\mathbf{X}}_{\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*}(QP_{=})\right) = OPT\left(\mathscr{L}_{PL^{X}(QP_{=})}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)\right)$$

We can also notice that $\left(\mathscr{L}_{PL^{X}(QP_{=})}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)\right)$ could be either bounded or unbounded ($-\infty$); recall these two cases are also separated in the strong duality Theorem 2.3.4. If it is unbounded, then (6.2.11d) is surely infeasible meaning that the total Lagrangian (6.2.11a)–(6.2.11b) has to be unbounded ($-\infty$).

Till here, we worked with an arbitrary fixed solution $(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$. From now on, let us consider that $(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$ is the optimal solution of the SDP partial Lagrangian $\left(PL^{X}(QP_{=})\right)$ from (6.2.5a); we proved above that $\left(PL^{X}(QP_{=})\right) = OPT(\mathscr{L}_{PL^{X}(QP_{=})}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*))$ is effectively reached by the total Lagrangian $(TL^{\mathbf{X}}(QP_{=}))$ in a point defined by above $\boldsymbol{\mu}^*$ and $\overline{\boldsymbol{\mu}}^*$ and some $\boldsymbol{\beta}^*$. Combining this with $\left(TL^{\mathbf{X}}(QP_{=})\right) = \left(TL^{X}(QP_{=})\right) \leq \left(PL^{X}(QP_{=})\right)$ from (6.1.5) and (6.2.8), we obtain the sought inequality (6.2.9) that we recall below for the reader's convenience. □

$$\left(TL^{\mathbf{X}}(QP_{=})\right) = \left(TL^{X}(QP_{=})\right) = \left(PL^{X}(QP_{=})\right) \tag{6.2.12}$$

Now recall Theorem 6.1.1 on total Lagrangians that states $OPT(DUAL(SDP(P_{=}))) = \left(TL^{\mathbf{X}}(QP_{=})\right)$. Combining this with above (6.2.12) and (6.2.6a)-(6.2.6b), we obtain the following fundamental hierarchies (inequalities):

$$OPT(DUAL(SDP(P_{=}))) = \left(TL^{\mathbf{X}}(QP_{=})\right) = \left(TL^{X}(QP_{=})\right) = \left(PL^{X}(QP_{=})\right) \leq \left(PL^{\mathbf{X}}(QP_{=})\right) \tag{6.2.13a}$$

$$OPT(DUAL(SDP(P_{=}))) = \left(TL^{\mathbf{X}}(QP_{=})\right) = \left(TL^{X}(QP_{=})\right) = \left(PL^{X}(QP_{=})\right) \leq OPT(SDP(P_{=})) \tag{6.2.13b}$$

The values of $\left(PL^{\mathbf{X}}(QP_{=})\right)$ and $OPT(SDP(P_{=}))$ can not be ordered in the most general setting. We will provide Example 6.2.3 in which $\left(PL^{\mathbf{X}}(QP_{=})\right) < OPT(SDP(P_{=}))$ and Example 6.3.3 in which $\left(PL^{\mathbf{X}}(QP_{=})\right) > OPT(SDP(P_{=}))$. As a side remark, recall from equations (6.2.6a)-(6.2.6b) that $(PL^{X}(QP_{=}))$ is a relaxation of both $(PL^{\mathbf{X}}(QP_{=}))$ and $OPT(SDP(P_{=}))$.

6.2.3 Using convexifications to obtain $\left(PL^{X}(QP_{=})\right) = \left(PL^{\mathbf{X}}(QP_{=})\right)$

We here show how $(PL^{\mathbf{X}}(QP_{=}))$ becomes equal to $(PL^{X}(QP_{=}))$ by exploiting the relaxation of the redundant constraints. The following proposition is a (quite deeply) modified and adapted version of a theorem from a lecture note of Frédéric Roupin,[31] but the proof is personal.

Proposition 6.2.2. *If $\left(PL^{\mathbf{X}}(QP_{=})\right) = -\infty$ it is clear that **all** programs from the hierarchy (6.2.13a) are unbounded. If $\left(PL^{\mathbf{X}}(QP_{=})\right) \neq -\infty$, we can take an optimal solution $\boldsymbol{\mu}^*$ of $\left(PL^{\mathbf{X}}(QP_{=})\right)$ in (6.2.3a)-(6.2.3b). If there exists $\overline{\boldsymbol{\mu}}^*$ that convexifies the augmented formulation of $\left(PL^{\mathbf{X}}(QP_{=})\right)$ from (6.2.4a)-(6.2.4b), then $\left(PL^{X}(QP_{=})\right) = \left(PL^{\mathbf{X}}(QP_{=})\right)$. This means that the first fundamental hierarchy (6.2.13a) collapses.*

Proof. We have $\left(PL^{X}(QP_{=})\right) \leq \left(PL^{\mathbf{X}}(QP_{=})\right)$ from (6.2.13a). Since we say (slightly abusing notations) that $\left(PL^{\mathbf{X}}(QP_{=})\right) = \mathscr{L}_{PL^{\mathbf{X}}(QP_{=})}(\boldsymbol{\mu}^*) = \mathscr{L}_{PL^{\mathbf{X}}(QP_{=})}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$, we can write

[31] See slide 57 of http://lipn.univ-paris13.fr/~roupin/docs/MPROSDPRoupin2017-partie2.pdf.

$\mathscr{L}_{PL^X(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*) \le \left(PL^X(QP_=)\right) \le (PL^{\mathbf{X}}(QP_=)) = \mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$. It is thus enough to show that $\mathscr{L}_{PL^X(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*) = \mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$.

Since the Hessian of $\mathscr{L}_{PL^X(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$ is equal to the Hessian of $\mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$, both programs are convex and have an SDP Hessian. Writing the optimal solution of $\mathscr{L}_{PL^X(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$ as $X = \mathbf{x}\mathbf{x}^\top + Y$, we notice $Y \succeq \mathbf{0}$ can only increase the objective value of $\mathscr{L}_{PL^X(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$, and so, an optimal Y can be $Y = \mathbf{0}$. An optimal solution of $\mathscr{L}_{PL^X(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$ satisfies $X = \mathbf{x}\mathbf{x}^\top$, *i.e.*, $\mathscr{L}_{PL^X(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*) = \mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*) = \left(PL^{\mathbf{X}}(QP_=)\right)$. $\square$

We now provide an example in which the hierarchy (6.2.13a) collapses, but the last inequality in (6.2.13b) is strict, *i.e.*, there is a duality gap between $(DUAL(SDP(P_=)))$ and $(SDP(P_=))$. The example relies on the fact that $\min\left\{x_{12} : \begin{bmatrix} 0 & x_{12} & 0 \\ x_{12} & x_{22} & 0 \\ 0 & 0 & 1+x_{12} \end{bmatrix} \succeq \mathbf{0}\right\} = 0$, while the dual of this program has optimum value -1.

Example 6.2.3. *We present the $(QP_=)$ example on the left and its partial Lagrangian $\left(PL^{\mathbf{X}}(QP_=)\right)$ on the right. After that, we will compare $(SDP(QP_=))$ and $(DUAL(SDP(QP_=))$.*

$$
(QP_=)\begin{cases}
\min & \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \bullet \mathbf{x}\mathbf{x}^\top \\
& s.t.\ x_4 = 0 \\
\overline{\mu}_{44} : & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \bullet \mathbf{x}\mathbf{x}^\top = 0 \\
\mu_{23} : & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \bullet \mathbf{x}\mathbf{x}^\top = 0 \\
\mu_{13} : & \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \bullet \mathbf{x}\mathbf{x}^\top = 0 \\
\mu_{11} : & \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \bullet \mathbf{x}\mathbf{x}^\top = 0 \\
\mu_{12} : & \begin{bmatrix} 0 & -\frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \bullet \mathbf{x}\mathbf{x}^\top = 1 \\
& \mathbf{x} \in \mathbb{R}^4
\end{cases}
$$

The first above constraint is a linear equality constraint. The next one is a redundant constraint, i.e., we have $\overline{p} = 1$ with regards to the canonical formulation (6.2.1a)-(6.2.1e) of $(QP_=)$. The other constraints are classical quadratic constraints.

The left program has only two feasible solutions $\mathbf{x}^\top = [0\ 0\ 1\ 0]$ and $\mathbf{x}^\top = [0\ 0\ -1\ 0]$ both of value 0.[32] *We now formulate the (augmented) partial Lagrangian.*

$$
\begin{aligned}
\left(PL^{\mathbf{X}}(QP_=)\right) &= \max_{\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}} \min_{\substack{\mathbf{x} \in \mathbb{R}^4 \\ x_4 = 0}} \\
& \begin{bmatrix} \mu_{11} & \frac{1}{2} - \frac{\mu_{12}}{2} & \mu_{13}+\mu_{11} & 0 \\ \frac{1}{2} - \frac{\mu_{12}}{2} & 0 & \mu_{23}+\mu_{11} & 0 \\ \mu_{13}+\mu_{11} & \mu_{23}+\mu_{11} & \mu_{12} & 0 \\ 0 & 0 & 0 & 1+\overline{\mu}_{44} \end{bmatrix} \bullet \mathbf{x}\mathbf{x}^\top - \mu_{12} \\
&= [\boldsymbol{\mu}] \bullet \mathbf{x}\mathbf{x}^\top - \mu_{12}
\end{aligned}
$$

The above (inner) minimization problem can be unbounded only if $[\boldsymbol{\mu}]_{3\times 3} \succeq \mathbf{0}$, where $[\boldsymbol{\mu}]_{3\times 3}$ is the leading principal minor of size 3×3 of $[\boldsymbol{\mu}]$. The variable $\overline{\mu}_{44}$ is redundant and does not appear in the non-augmented formulation (6.2.3a)-(6.2.3b) of $\left(PL^{\mathbf{X}}(QP_=)\right)$. However, notice that Prop 6.2.2 does hold, i.e., the augmented partial $\mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*, \overline{\mu}^*_{44})$ from the inner minimization problem is convex for any optimal solution $(\boldsymbol{\mu}^*, \overline{\mu}^*_{44})$ of $\left(PL^{\mathbf{X}}(QP_=)\right)$. Since $[\boldsymbol{\mu}^*]_{3\times 3} \succeq \mathbf{0}$ and $[\boldsymbol{\mu}^*]_{22} = 0$, we need to have $[\boldsymbol{\mu}^*]_{12} = 0$, i.e., $\mu^*_{12} = 1$. The rest of the elements can be, for instance, $\mu^*_{11} = \mu^*_{13} = \mu^*_{23} = \overline{\mu}^*_{44} = 0$. However, for any choice of these latter elements, the objective value of the minimization problem will be -1 (notice $\mathbf{x} = \mathbf{0}$ leads to an objective value of $-\mu^*_{12} = -1$), and so, $\left(PL^{\mathbf{X}}(QP_=)\right) = -1$.

The SDP relaxation $SDP(QP_=)$ can be written: $\min\left\{X_{12} : X = \begin{bmatrix} 0 & X_{12} & 0 & 0 \\ X_{12} & X_{22} & 0 & 0 \\ 0 & 0 & 1+X_{12} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \succeq \mathbf{x}\mathbf{x}^\top \succeq \mathbf{0}\right\} = 0$. *We have $OPT(SDP(QP_=)) = 0$ and this solution is achieved by $X_{12} = 0$, $X_{22} = 0$, $X_{33} = 1 + X_{12} = 1$ and $\mathbf{x} = [0\ 0\ 0\ 0]^\top$ (or $\mathbf{x} = [0\ 0\ 1\ 0]^\top$); notice X_{12} can not be strictly less than 0 because $X_{11} = 0$. One could calculate that $OPT(DUAL(SDP(QP_=))) = -1$. All values in the hierarchies (6.2.13a)-(6.2.13b) are equal to -1 except $OPT(SDP(QP_=)) = 0$.*

[32] Constraints μ_{23} and μ_{13} enforce $x_2x_3 = x_1x_3 = 0$; combining this with constraint μ_{11} that states $x_1^2 + 2x_2x_3 + 2x_1x_3 = 0$, we obtain $x_1 = 0$. Constraint μ_{12} enforces $-x_1x_2 + x_3^2 = 1 \implies x_3^2 = 1$. Finally, $x_2 = 0$ follows from constraint μ_{23}.

6.3 The case of $0-1$ quadratic programs: partial and total Lagrangians

A part of this section aims at proving results related to the QCR method of Alain Billionnet, Sourour Elloumi and Marie-Christine Plateau (see link in Footnote 38, p. 96) and to the article of Alain Faye and Frédéric Roupin indicated in Footnote 33, p. 71. However, I think the presentation style (using longer and more detailed arguments) and the order of the theorems from this manuscript is completely different.

6.3.1 Main characterization

We interpret a $0-1$ quadratic program exactly as a particular case of $(QP_=)$ from (6.2.1a)-(6.2.1e) in which the first n constraints (6.2.1d) are $x_i^2 = x_i$. We can thus have $p \geq n$ non-redundant quadratic constraints. All results from the previous Subsection 6.2 do hold in this new $0-1$ context. In fact, the $0-1$ programs are in some sense simpler because we can prove $OPT(DUAL(SDP(QP_=))) = OPT(SDP(QP_=))$, *i.e.*, the hierarchy (6.2.13b) collapses and we can no longer have a duality gap $\left(PL^{\mathbf{X}}(QP_=)\right) < OPT(SDP(QP_=))$ as in Example 6.2.3.

Theorem 6.3.1. *The following fundamental hierarchies hold for $0-1$ quadratic programs.*

$$OPT(DUAL(SDP(P_=))) = \left(TL^{\mathbf{X}}(QP_=)\right) = \left(TL^{X}(QP_=)\right) = \left(PL^{X}(QP_=)\right) = OPT(SDP(P_=)) \quad (6.3.1a)$$
$$\leq \left(PL^{\mathbf{X}}(QP_=)\right) \quad (6.3.1b)$$

Proof. The first equality (6.3.1a) follows from the fact that we can show that the hierarchy (6.2.13b) collapses, *i.e.*, all inequalities in (6.2.13b) are equalities in a $0-1$ context. To show this, it is enough to prove there is no duality gap between $(DUAL(SDP(P_=)))$ and $(SDP(P_=))$. This is a consequence of the strong duality Theorem 2.3.3, considering that $(DUAL(SDP(P_=)))$ is strictly feasible and bounded – the unbounded case $OPT(DUAL(SDP(P_=))) = \infty$ would lead to to $(OPT(SDP(P_=))) = \infty$ by virtue of $OPT(DUAL(SDP(P_=))) \leq OPT(SDP(P_=))$. However, the dual is always strictly feasible because the SDP constraint (6.1.4b) can be written under the form:

$$Y + Y_\mu = Y + \begin{bmatrix} -t & -\frac{1}{2}\mu_1 & -\frac{1}{2}\mu_2 & \dots & -\frac{1}{2}\mu_n \\ -\frac{1}{2}\mu_1 & \mu_1 & 0 & \dots & 0 \\ -\frac{1}{2}\mu_2 & 0 & \mu_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ -\frac{1}{2}\mu_n & 0 & 0 & \dots & \mu_n \end{bmatrix} \succeq \mathbf{0}.$$

The above matrix can be strictly feasible (positive definite) by taking $\mu_1 = \mu_2 \cdots = \mu_n = M$ for a sufficiently large M, so that the bottom-right $n \times n$ minor $[Y + \mathtt{diag}(\boldsymbol{\mu})]_{n\times n}$ become positive definite. A sufficient M value can be 1 minus the lowest eigenvalue of Y. Using the Sylvester criterion (Prop. 1.5.2) in reversed order, we can prove that the above matrix $Y + Y_\mu$ is positive definite by showing that the whole determinant is positive. This can always be the case by taking a sufficiently large value of $-t$. By developing the Leibniz formula for the determinant of the matrix $Y + Y_\mu$, there will be a term $-t\det\left([Y + \mathtt{diag}(\boldsymbol{\mu})]_{n\times n}\right)$ that can be arbitrarily large, so as to make the determinant as high as possible. Using the strong duality as mentioned above, we obtain $OPT(DUAL(SDP(P_=))) = OPT(SDP(P_=))$, and so, the hierarchy (6.2.13b) collapses into (6.3.1a). Finally, (6.3.1b) follows from (6.2.13a). □

The above proof does show that $OPT(DUAL(SDP(P_=))) = OPT(SDP(P_=))$, but the strong duality theorem only guarantees that $(SDP(P_=))$ does reach the optimum value. The program $(DUAL(SDP(P_=)))$ does not necessarily effectively reach its optimum value. This will become clear in the following example.

6.3.2 Two examples: $OPT(DUAL(SDP(P_=)))$ may not reach its own optimum value and $OPT(SDP(P_=))$ may be strictly lower than $\left(PL^{\mathbf{X}}(QP_=)\right)$

Example 6.3.2. *We modify Example 6.2.3 as follows:*

– *The constraint associated to dual variable μ_{12} becomes*

$$\begin{bmatrix} 0 & -\frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \bullet \mathbf{x}\mathbf{x}^\top = 0$$

– *We add four constraints $x_i^2 = x_i$ with $i \in [1..4]$ whose dual values are μ_1, μ_2, μ_3 and μ_4. These constraints imply that all variables are binary.*

The $SDP(P_{=})$ relaxation of $(QP_{=})$ has the solution $X = \mathbf{0}_{4\times 4}$ of objective value 0. We will show that $(DUAL(SDP(P_{=})))$ converges to 0, even if there is no feasible solution with value 0 in this dual. For this, we first write $(DUAL(SDP(P_{=})))$ as follows:

$$(DUAL(SDP(P_{=}))) \begin{cases} \max \quad t \\ \quad s.t. \begin{bmatrix} -t & \frac{-\mu_1}{2} & \frac{-\mu_2}{2} & \frac{-\mu_3}{2} & \frac{-\mu_4}{2} \\ \frac{-\mu_1}{2} & \mu_{11}+\mu_1 & \frac{1}{2}-\frac{\mu_{12}}{2} & \mu_{13}+\mu_{11} & 0 \\ \frac{-\mu_2}{2} & \frac{1}{2}-\frac{\mu_{12}}{2} & \mu_2 & \mu_{23}+\mu_{11} & 0 \\ \frac{-\mu_3}{2} & \mu_{13}+\mu_{11} & \mu_{23}+\mu_{11} & -\mu_{12}+\mu_3 & 0 \\ \frac{-\mu_4}{2} & 0 & 0 & 0 & 1+\overline{\mu}_{44}+\mu_4 \end{bmatrix} \succeq \mathbf{0}, \end{cases}$$

A solution of value zero of the above program would clearly set $t = 0$. This would imply that Row 1 contains only zeros, so that $\mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$. Furthermore, because μ_2 stands alone on the diagonal at position (3,3), row 3 has to contain only zeros as well. This means $\mu_{12} = 1$ which leads to a negative value at position (4,4) on the diagonal. The resulting matrix can not be SDP.

We can prove the optimal solution converges to zero. Let us take $\mu_1 = 0$, $\mu_2 = \varepsilon$, $\mu_3 = 0$, $\mu_4 = 0$, $\mu_{11} = M$, $\mu_{12} = 0$, $\mu_{13} = -\mu_{11} = -M$, $\mu_{23} = -\mu_{11} = -M$, $\overline{\mu}_{44} = 0$. The variables ε and M stand, resp. , for a very small and a very large positive value. The above program simplifies to:

$$\max \quad t$$

$$s.t.\ [\boldsymbol{\mu}_{M,\varepsilon}] = \begin{bmatrix} -t & 0 & -\varepsilon/2 & 0 & 0 \\ 0 & M & 1/2 & 0 & 0 \\ -\varepsilon/2 & 1/2 & \varepsilon & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \succeq \mathbf{0},$$

We take a sufficiently large value of M $(> \frac{1}{4\varepsilon})$, so as to be able to have $\det([\boldsymbol{\mu}_{M,\varepsilon}]_{3\times 3}) > 0$ for small values of $-t$, where $[\boldsymbol{\mu}_{M,\varepsilon}]_{3\times 3}$ is the leading principal minor of size 3×3. However, the optimum value of above program is $-\frac{\varepsilon}{4}$, associated to $-t = \frac{\varepsilon}{4}$. Recall M can be as large as necessary to ensure $\det([\boldsymbol{\mu}_{M,\varepsilon}]_{3\times 3}) > 0$. The objective value of this solution is $-\frac{\varepsilon}{4}$ and its limit is $\lim_{\varepsilon\to 0} -\frac{\varepsilon}{4} = 0$.

We continue with an example in which the fundamental inequality (6.3.1b) is strict, *i.e.*, $OPT(SDP(P_{=})) < \big(PL^{\mathbf{X}}(QP_{=})\big)$.

Example 6.3.3. *We introduce a $0-1$ program on the left, the partial Lagrangian on the right and then we will analyse the SDP relaxation.*

$$(QP_{=}) \begin{cases} \min \quad \left[\begin{smallmatrix} 0 & -1 \\ -1 & 0 \end{smallmatrix}\right] \bullet \mathbf{x}\mathbf{x}^\top \\ \quad s.t.\ x_1 + x_2 = 1 \\ \quad \mu_1 : \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right] \bullet \mathbf{x}\mathbf{x}^\top - x_1 = 0 \\ \quad \mu_2 : \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right] \bullet \mathbf{x}\mathbf{x}^\top - x_2 = 0 \\ \qquad \mathbf{x} \in \mathbb{R}^2 \end{cases}$$

Notice that the first constraint in the above program is a linear equality constraint. We have no redundant constraints, i.e., we have $\overline{p} = 0$ with regards to the canonical formulation (6.2.1a)-(6.2.1e) of $(QP_{=})$.

The only feasible solutions of $(QP_{=})$ are $[x_1\ x_2] = [1\ 0]$ and $[x_1\ x_2] = [0\ 1]$, both of value $OPT(QP_{=}) = 0$. We now formulate the partial Lagrangian.

$$\Big(PL^{\mathbf{X}}(QP_{=})\Big) = \max_{\mu_1,\mu_2} \min_{\substack{\mathbf{x}\in\mathbb{R}^2 \\ x_1+x_2=1}} \left[\begin{smallmatrix} \mu_1 & -1 \\ -1 & \mu_2 \end{smallmatrix}\right] \bullet \mathbf{x}\mathbf{x}^\top - \mu_1 x_1 - \mu_2 x_2$$

We know by the Lagrangian definition that $\big(PL^{\mathbf{X}}(QP_{=})\big) \leq OPT(QP_{=})$. We will show that for $[\mu_1\ \mu_2] = [-1\ -1]$, the Lagrangian reaches $OPT(QP_{=}) = 0$. Indeed, replacing these $[\mu_1\ \mu_2]$ values in above formula, we obtain

$$\min_{\substack{\mathbf{x}\in\mathbb{R}^2 \\ x_1+x_2=1}} \left[\begin{smallmatrix} -1 & -1 \\ -1 & -1 \end{smallmatrix}\right] \bullet \mathbf{x}\mathbf{x}^\top + x_1 + x_2 = -(x_1+x_2)^2 + (x_1+x_2) = 0$$

For now, we have $\left(PL^{\mathbf{X}}(QP_=)\right) = OPT(QP_=) = 0$ *and let us turn towards* $(SDP(QP_=))$*. We notice that* $[x_1\ x_2] = [0.5\ 0.5]$ *combined with* $X = \left[\begin{smallmatrix}0.5 & 0.25\\ 0.25 & 0.5\end{smallmatrix}\right] = \mathbf{x}\mathbf{x}^\top + 0.25I_2 = \left[\begin{smallmatrix}0.25 & 0.25\\ 0.25 & 0.25\end{smallmatrix}\right] + \left[\begin{smallmatrix}0.25 & 0\\ 0 & 0.25\end{smallmatrix}\right]$ *is a feasible solution of* $(SDP(QP_=))$ *with objective value* -0.5*. This leads to*

$$OPT(SDP(QP_=)) \leq -0.5 < 0 = \left(PL^{\mathbf{X}}(QP_=)\right),$$

and so, the inequality (6.3.1b) can be strict.

6.3.3 The limit of the strongest convexification is $\left(PL^{\mathbf{X}}(QP_=)\right)$

6.3.3.1 The strongest convexifications can lead to $\left(PL^{X}(QP_=)\right) = \left(PL^{\mathbf{X}}(QP_=)\right)$

We recall Proposition 6.2.2. It states that if $\boldsymbol{\mu}^*$ is an optimal solution of the non-augmented (6.2.3a)-(6.2.3b) formulation of $\left(PL^{\mathbf{X}}(QP_=)\right)$ and there there exists $\overline{\boldsymbol{\mu}}^*$ that convexifies the augmented (6.2.4a)-(6.2.4b) formulation of $\left(PL^{\mathbf{X}}(QP_=)\right)$, then $\left(PL^{X}(QP_=)\right) = \left(PL^{\mathbf{X}}(QP_=)\right)$. In other words, if the convexification is successful (for the optimal $\boldsymbol{\mu}^*$ chosen above), both hierarchies (6.3.1a)-(6.3.1b) collapse (in the binary case).

Let $Q_{\boldsymbol{\mu}^*}$ be the factor of the quadratic term $\mathbf{x}\mathbf{x}^\top$ associated to the optimal solution $\boldsymbol{\mu}^*$ of the non-augmented formulation (6.2.3a)-(6.2.3b) of $\left(PL^{\mathbf{X}}(QP_=)\right)$. The factor of $\mathbf{x}\mathbf{x}^\top$ in the augmented formulation (6.2.4a)-(6.2.4b) of $\left(PL^{\mathbf{X}}(QP_=)\right)$ has the form $Q_{\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}} = Q_{\boldsymbol{\mu}^*} + \sum_{j=1}^{\overline{p}} \overline{\mu}_j \overline{B}_j$ where the $\overline{B}_j's$ with $j \in [1..\overline{p}]$ are the quadratic factors of $\mathbf{x}\mathbf{x}^\top$ in the redundant constraints (6.2.1c) of the $(QP_=)$ definition.

Let us first investigate the case in which if Q_{μ^*} is not positive over $\{\mathbf{y} \in \mathbb{R}^n : A\mathbf{y} = \mathbf{0}\}$, *i.e.*, over the null space $\texttt{null}(A)$ of A (see also (A.1.2)). In this case, we can show that $\left(PL^{\mathbf{X}}(QP_=)\right) = -\infty$. Indeed, if there is some $\mathbf{y} \in \mathbb{R}^n$ such that $A\mathbf{y} = \mathbf{0}$ and $Q_{\mu^*} \bullet \mathbf{y}\mathbf{y}^\top = -z < 0$, we can take any feasible solution $\mathbf{x} \in \mathbb{R}^n$ of $\mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*)$ from the non-augmented $\left(PL^{\mathbf{X}}(QP_=)\right)$ formulation (6.2.3a)-(6.2.3b). This $\mathbf{x}$ satisfies $A\mathbf{x} = \mathbf{b}$ and notice that $\mathbf{x} + t\mathbf{y}$ satisfies $A(\mathbf{x} + t\mathbf{y}) = \mathbf{b}$ as well $\forall t > 0$. If we write the objective value of $\mathbf{x} + t\mathbf{y}$ as a function of t, we obtain a polynomial of degree 2 with the leading term $-zt^2$. This is a concave polynomial that is unbounded from below regardless of its non-quadratic terms. In this case it is not difficult to achieve $\left(PL^{X}(QP_=)\right) = \left(PL^{\mathbf{X}}(QP_=)\right) = -\infty$.

We hereafter focus on the contrary case: Q_{μ^*} is positive over $\texttt{null(A)}$. We are looking for redundant constraints that can convexify any quadratic factor Q_μ that is non-negative (positive) over $\texttt{null}(A)$. In fact, it may be enough to convexify the matrix Q_{μ^*} associated to the optimal solution μ^* of $\left(PL^{\mathbf{X}}(QP_=)\right)$, but let us keep in mind a broader objective. However, if it is possible to convexify Q_{μ^*} into $Q_{\mu^*,\overline{\boldsymbol{\mu}}^*} \succeq \mathbf{0}$, we can apply Proposition 6.2.2, and so, both the fundamental hierarchies (6.3.1a)-(6.3.1b) collapse – also recall Theorem 6.3.1. In such a case, we obtain that $OPT(SDP(Q_=)) = \left(PL^{X}(QP_=)\right)$ reaches the limit $\left(PL^{\mathbf{X}}(QP_=)\right)$.

However, if the redundant constraints are not strong enough to convexify $Q_{\boldsymbol{\mu}^*}$ for the above chosen optimal $\boldsymbol{\mu}^*$, they might be able to convexify some Q_μ for some other $\boldsymbol{\mu}$. If this happens, we are certain that $\left(PL^{X}(QP_=)\right)$ is bounded, although its value may be $\mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}) < \mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*) = \left(PL^{\mathbf{X}}(QP_=)\right)$. Finally, keep in mind that only $\left(PL^{\mathbf{X}}(QP_=)\right)$ does not depend on the chosen redundant constraints in the fundamental hierarchy (6.3.1a)-(6.3.1b), *i.e.*, one can consider that all other terms in this hierarchy are actually indexed by a set of chosen redundant constraints.

6.3.3.2 Examples of redundant constraints of different strengths

Example 6.3.4. *(A unique redundant constraint) A redundant constraint can be constructed by observing that* $(A_i\mathbf{x} - b_i)^2 = 0$ *for any* $i \in [1..p]$*, where* A_i *is the row* i *of* A*. We obtain* $(A_i\mathbf{x})^2 - 2b_iA_i\mathbf{x} + b_i^2 = 0$*, equivalent to*

$$A_i^\top A_i \bullet \mathbf{x}\mathbf{x}^\top - 2b_iA_i\mathbf{x} + b_i^2 = 0, \tag{6.3.2}$$

by virtue of Lemma 1.3.3.1. By summing over all $i \in [1..p]$*, we obtain*

$$A^\top A \bullet \mathbf{x}\mathbf{x}^\top - 2\mathbf{b}^\top A\mathbf{x} + \mathbf{b}^\top\mathbf{b} = 0 \tag{6.3.3}$$

As long as we consider a partial and not a total Lagrangian, we have $2b_iA_i\mathbf{x} = 2b_i^2$ because $A\mathbf{x} = \mathbf{b}$ is active; thus, the constraint (6.3.2) reduces to $A_i^\top A_i \bullet \mathbf{x}\mathbf{x}^\top = b_i^2$ and (6.3.3) reduces to

$$A^\top A \bullet \mathbf{x}\mathbf{x}^\top = \mathbf{b}^\top\mathbf{b} \tag{6.3.4}$$

Example 6.3.5. *(A set of redundant constraints) We first use constraints $A\mathbf{x} = b$. For each $i \in [1..p]$, we generate a set of constraints $x_jA_i\mathbf{x} = x_j\mathbf{b}_i$ for all $j \in [1..n]$. There exists a linear combination of these constraints that generate the constraint from previous Example 6.3.4.*

Proof. For each $i \in [1..p]$, we perform the following. We multiply by A_{ij} the constraint $x_jA_i\mathbf{x} = x_jb_i$, $\forall j \in [1..n]$. By summing up over all j, we obtain $\left(\sum_{j=1}^{n} A_{ij}x_j\right) A_i\mathbf{x} = \left(\sum_{j=1}^{n} A_{ij}x_j\right) b_i$ that can be written $A_i\mathbf{x}A_i\mathbf{x} = A_i\mathbf{x}b_i$, or furthermore $A_i^\top A_i \bullet \mathbf{x}\mathbf{x}^\top - b_iA_i\mathbf{x} = 0$. We now multiply $-A_i\mathbf{x} + b_i = 0$ by b_i to obtain $-b_iA_i\mathbf{x} + b_i^2 = 0$ and we add this to the previous equality to obtain $A_i^\top A_i \bullet \mathbf{x}\mathbf{x}^\top - 2b_iA_i\mathbf{x} + b_i^2 = 0$, which is exactly (6.3.2). ☐

The above constraint sets are not equivalent in the general non-binary case

Remark 6.3.6. *In the general non-binary case, the first redundant constraint set (Example 6.3.4) might be weaker. For instance, consider $Q_{\mu^*} = \left[\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right]$ and a unique linear constraint $A\mathbf{x} = \mathbf{b}$ with $A = [0\ 1]$ and $\mathbf{b} = 0$. The constraint generated by the first Example 6.3.4 is (6.3.3), i.e., $A^\top A \bullet \mathbf{x}\mathbf{x}^\top = 0$, equivalent to $\left[\begin{smallmatrix}0&0\\0&1\end{smallmatrix}\right] \bullet \mathbf{x}\mathbf{x}^\top = 0$. This $A^\top A = \left[\begin{smallmatrix}0&0\\0&1\end{smallmatrix}\right]$ matrix can not convexify Q_{μ^*}. On the other hand, the second redundant constraint set (Example 6.3.5) generates $x_1A\mathbf{x} = x_1b = 0$ (or $x_1x_2 = 0$, i.e., $A' \bullet \mathbf{x}\mathbf{x}^\top = \left[\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right] \bullet \mathbf{x}\mathbf{x}^\top = 0$) and $x_2A\mathbf{x} = x_2b = 0$ (or $x_2x_2 = 0$, i.e., $A'' \bullet \mathbf{x}\mathbf{x}^\top = \left[\begin{smallmatrix}0&0\\0&1\end{smallmatrix}\right] \bullet \mathbf{x}\mathbf{x}^\top = 0$). The matrices $A' = \left[\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right]$ and $A'' = \left[\begin{smallmatrix}0&0\\0&1\end{smallmatrix}\right]$ can easily convexify Q_{μ^*}.*

Remark 6.3.7. *The second redundant constraint set (Example 6.3.5) can convexify any matrix Q_{μ^*} that is non-negative over $\texttt{null}(A)$. This way, we can always apply Prop. 6.2.2 and collapse the hierarchy (6.2.13a), i.e., the convexification is optimal and it reaches the limit value $\left(PL^{\mathbf{X}}(QP_{=})\right)$. The proof of this convexification is given in Appendix C.2.1, see more exactly Prop C.2.2.*

If Q_{μ^*} is strictly positive over $\texttt{null(A)}$ (i.e., $\mathbf{u}^\top Q\mathbf{u} > 0\ \forall \mathbf{u} \in \texttt{null(A)} - \{\mathbf{0}\}$), Q_{μ^*} can be convexified more easily. We can use any unique redundant constraint of the form $\mathbf{x}^\top A^\top SA\mathbf{x} = \mathbf{b}^\top S\mathbf{b}$ for some $S \succ \mathbf{0}$, e.g., for instance (6.3.4) corresponds to choosing $S = I_n$. In other words, we can always construct $Q_{\mu^*} + \lambda A^\top SA \succeq \mathbf{0}$ for a sufficiently large λ, see the proof in Prop. C.2.1.

The above constraint sets are equivalent in the binary case

For the binary case, we will prove below (Prop. 6.3.8) that the SDP programs integrating the SDP versions of above redundant constraint sets (Examples 6.3.4 and 6.3.5) are equivalent. Recall that in the 0-1 case the convexified total and partial Lagrangians reach $OPT(SDP(P_{=}))$ as stated in (6.3.1a). Using above two statements, the two redundant constraint sets make the convexified total and partial Lagrangians reach the same value $OPT(SDP(P_{=}))$, *i.e.*, they are equivalent.

Proposition 6.3.8. *The SDP constraints associated to the redundant constraints from Example 6.3.4 and Example 6.3.5 are equivalent. This means that the best convexifications (of the total or partial Lagrangians) achieved by the two redundant constraint sets have <u>the same value in the $0-1$ case</u>, i.e., that of the SDP bound $OPT(SDP(P_{=}))$ expressed using either set of redundant constraints.*

Proof. Notice using Examples 6.3.4 and 6.3.5 that the two SDP constraint sets are respectively:

$$A^\top A \bullet X - 2\mathbf{b}^\top A\mathbf{x} + \mathbf{b}^\top\mathbf{b} = 0 \tag{6.3.5a}$$

$$\sum_{k=1}^{n} A_{ik}X_{jk} = x_jb_i,\ \forall j \in [1..n],\ i \in [1..p] \tag{6.3.5b}$$

The implication (6.3.5b) $\Longrightarrow$ (6.3.5a) can be constructed by applying on constraints (6.3.5b) the linear combination presented in the proof of Example 6.3.5. This linear combination leads to (6.3.5a).

It is more difficult to show the converse (6.3.5a) $\implies$ (6.3.5b). We write (6.3.5a) as $0 = A^\top A \bullet (X - \mathbf{x}\mathbf{x}^\top) + A^\top A \bullet \mathbf{x}\mathbf{x}^\top - 2\mathbf{b}^\top A\mathbf{x} + \mathbf{b}^\top\mathbf{b} = A^\top A \bullet (X - \mathbf{x}\mathbf{x}^\top) + 0$, where we used (6.3.3) which is a consequence of $A\mathbf{x} = \mathbf{b}$ (this constraint does appear in both SDP formulations even if it is not necessary as it can be inferred from (6.3.5a) or (6.3.5b) respectively). We thus obtain $A^\top A \bullet (X - \mathbf{x}\mathbf{x}^\top) = 0$ and Prop. 1.3.4 implies that $A^\top A(X - \mathbf{x}\mathbf{x}^\top) = \mathbf{0}$. Taking row $r \in [1..n]$ and column $c \in [1..n]$ of this product, we obtain:

$$\begin{aligned}
0 &= \sum_{k=1}^{n} (A^\top A)_{rk} (X - \mathbf{x}\mathbf{x}^\top)_{kc} \\
&= \sum_{k=1}^{n} \left(\sum_{i=1}^{p} A_{ir} A_{ik} \right) (X_{kc} - x_k x_c) \\
&= \sum_{k=1}^{n} \sum_{i=1}^{p} A_{ir} \left(A_{ik} X_{kc} - A_{ik} x_k x_c \right) \\
&= \sum_{k=1}^{n} \sum_{i=1}^{p} A_{ir} A_{ik} X_{kc} - \sum_{i=1}^{p} A_{ir} b_i x_c && \text{(we used } A_i\mathbf{x} = b_i\text{)} \\
&= \sum_{i=1}^{p} A_{ir} \left(\sum_{k=1}^{n} A_{ik} X_{ck} - x_c b_i \right)
\end{aligned}$$

Since the last formula holds for all $r \in [1..n]$, we can reformulate it in terms of the rows A_i of A, obtaining $\sum_{i=1}^{p} A_i \left(\sum_{k=1}^{n} A_{ik} X_{ck} - x_c b_i \right) = \mathbf{0}$. This is a linear combination of the rows A_i of A. Assuming a legitimate condition $rank(A) = p$ (*i.e.*, the constraints $A\mathbf{x} = \mathbf{b}$ are linearly independent), this linear combination can lead to $\mathbf{0}$ only if $\sum_{k=1}^{n} A_{ik} X_{ck} - x_c b_i = 0 \ \forall i \in [1..p]$. Since this holds for any $c \in [1..n]$, we have obtained (6.3.5b). This proof is taken from Prop. 5 of the paper "Partial Lagrangian relaxation for General Quadratic Programming" by Alain Faye and Frédéric Roupin.[33] □

6.3.4 Collapsing both hierarchies by convexification and an associated `Branch-and-bound`

We here focus on solving *binary* equality-constrained quadratic programming. The main idea is that we can use the best convexification constructed in Section 6.3.4.1 to determine fast lower bounds for a `Branch-and-bound` (Section 6.3.4.2) that solves the initial binary equality-constrained quadratic problem. As a side remark, certain convexification ideas below can well apply to the non-binary problem as well.

6.3.4.1 Determining the best convexification coefficients $\boldsymbol{\mu}^*$ and $\overline{\boldsymbol{\mu}}^*$ by solving $DUAL(SDP(P_=))$

According to (6.3.1a), in the binary case, the value of a convexified total Lagrangian reaches $OPT(SDP(P_=))$. However, the quality of $OPT(SDP(P_=))$ is dependent on the redundant constraints it integrates. Let us write the total Lagrangian from (6.2.7a)-(6.2.7b) in a more compact form:

$$\left(TL^{\mathbf{X}}(QP_=)\right) = \max_{\boldsymbol{\mu},\overline{\boldsymbol{\mu}},\boldsymbol{\beta}} \mathscr{L}_{TL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}, \boldsymbol{\beta}) \tag{6.3.6a}$$

$$\mathscr{L}_{TL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}, \boldsymbol{\beta}) = \left\{ \begin{array}{ll} \min\limits_{\mathbf{x}\in\mathbb{R}^n} & Q_{\boldsymbol{\mu},\overline{\boldsymbol{\mu}}} \bullet \mathbf{x}\mathbf{x}^\top + \left(\mathbf{c}_{\boldsymbol{\mu},\overline{\boldsymbol{\mu}}}^\top + \boldsymbol{\beta}^\top A\right)\mathbf{x} \\ & -\boldsymbol{\mu}^\top \mathbf{e} - \overline{\boldsymbol{\mu}}^\top \overline{\mathbf{e}} - \boldsymbol{\beta}^\top \mathbf{b}, \end{array} \right. \tag{6.3.6b}$$

where $Q_{\boldsymbol{\mu},\overline{\boldsymbol{\mu}}} = Q + \sum_{i=1}^{p} \mu_i B_i + \sum_{j=1}^{\overline{p}} \overline{\mu}_j \overline{B}_j$ and $\mathbf{c}_{\boldsymbol{\mu},\overline{\boldsymbol{\mu}}}^\top = \mathbf{c}^\top + \sum_{i=1}^{p} \mu_i \mathbf{d}_i^\top + \sum_{j=1}^{\overline{p}} \overline{\mu}_j \overline{\mathbf{d}}_j^\top$.

Let us first show that $\left(TL^{\mathbf{X}}(QP_=)\right)$ is bounded. For this, we can use the fact that in the $0-1$ case the objective function contains n terms of the form $\mu_i x_i^2 - \mu_i x_i$ (for all $i \in [1..n]$) that could always strictly convexify any matrix. In other words, there always exist $\boldsymbol{\mu}^o$ and $\overline{\boldsymbol{\mu}}^o$ such that $Q_{\boldsymbol{\mu}^o,\overline{\boldsymbol{\mu}}^o} \succ \mathbf{0}$, *i.e.*, take a sufficiently large μ_i^o in the terms $\mu_i^o x_i^2 - \mu_i^o x_i$ (for all $i \in [1..n]$). We have $Q_{\boldsymbol{\mu}^o,\overline{\boldsymbol{\mu}}^o}$ invertible by virtue of $\det\left(Q_{\boldsymbol{\mu}^o,\overline{\boldsymbol{\mu}}^o}\right) > 0$. The (transposed) gradient of the objective function of $\mathscr{L}_{TL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^o, \overline{\boldsymbol{\mu}}^o, \boldsymbol{\beta})$ from (6.3.6b) is $\nabla_{\boldsymbol{\mu}^o,\overline{\boldsymbol{\mu}}^o}^\top(\mathbf{x}) = 2\mathbf{x}^\top Q_{\boldsymbol{\mu}^o,\overline{\boldsymbol{\mu}}^o} + \mathbf{c}_{\boldsymbol{\mu},\overline{\boldsymbol{\mu}}}^\top + \boldsymbol{\beta}^\top A$. Since $Q_{\boldsymbol{\mu}^o,\overline{\boldsymbol{\mu}}^o}$ is invertible, there exists a stationary point

[33] Published in *A Quarterly Journal of Operations* in 2007, vol 5(1), pp. 75-88.

$\mathbf{x}^o$ in which $\nabla^\top_{\boldsymbol{\mu}^o,\overline{\boldsymbol{\mu}}^o}(\mathbf{x}^o) = \mathbf{0}$. The stationary point of a (strictly) convex function is its minimizer, and so, we are sure that $OPT(DUAL(SDP(P_=))) = \left(TL^{\mathbf{X}}(QP_=)\right) = OPT(SDP(P_=))$ is bounded.

Let us study what happens if we determine $\boldsymbol{\mu}^*$ and $\overline{\boldsymbol{\mu}}^*$ from the optimal values of $DUAL(SDP(P_=))$ from (6.1.4a)-(6.1.4b), *i.e.*, solve (6.1.4a)-(6.1.4b) and retrieve the optimal value of $\boldsymbol{\mu}$ (think of $\boldsymbol{\mu}$ and $\overline{\boldsymbol{\mu}}$ as merged into a unique $\boldsymbol{\mu}$). He hereafter consider that the coefficients $\boldsymbol{\mu}^*$ and $\overline{\boldsymbol{\mu}}^*$ are fixed. We will show that using these coefficients in the augmented partial Lagrangian allows it to reach the same value as the total Lagrangian $\left(TL^{\mathbf{X}}(QP_=)\right)$. We recall the definition of the augmented partial Lagrangian from (6.2.4b) using more compact notations for fixed $\boldsymbol{\mu}^*$ and $\overline{\boldsymbol{\mu}}^*$:

$$\mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*) = \begin{cases} \min & Q_{\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*} \bullet \mathbf{x}\mathbf{x}^\top + \mathbf{c}^\top_{\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*}\mathbf{x} - \boldsymbol{\mu}^{*\top}\mathbf{e} - \overline{\boldsymbol{\mu}}^{*\top}\overline{\mathbf{e}} \\ s.t. & A\mathbf{x} = b \\ & \mathbf{x} \in \mathbb{R}^n \end{cases} \tag{6.3.7}$$

For the fixed $\boldsymbol{\mu}^*$ and $\overline{\boldsymbol{\mu}}^*$ determined above, we have $Q_{\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*} \succeq \mathbf{0}$ because of the bottom-right term of (6.1.4b). As such, the above program has the same optimum value as its SDP version obtained by replacing $\mathbf{x}\mathbf{x}^\top$ with X in (6.3.7) and adding $X \succeq \mathbf{x}\mathbf{x}^\top$. This SDP partial Lagrangian $\mathscr{L}_{PL^{X}(QP_=)}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*)$ is formally defined by (6.2.5b).

One can easily check that $\left(TL^{\mathbf{X}}(QP_=)\right) = \mathscr{L}_{TL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*,\boldsymbol{\beta}^*) \leq \mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*) = \mathscr{L}_{PL^{X}(QP_=)}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*) \leq \left(PL^{X}(QP_=)\right)$. Both inequalities follow from Lagrangian relaxation relations. But since we know $OPT(DUAL(SDP(P_=))) = \left(TL^{\mathbf{X}}(QP_=)\right) = \left(PL^{X}(QP_=)\right) = OPT(SDP(P_=))$ from the fundamental hierarchy (6.3.1a), we can (re-)write:

$$OPT\left(\mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*)\right) = \left(TL^{\mathbf{X}}(QP_=)\right) = OPT(DUAL(SDP(QP_=))) = OPT(SDP(QP_=)). \tag{6.3.8}$$

Remark 6.3.9. *In the* 0–1 *case, the coefficients* $\boldsymbol{\mu}^*$ *and* $\overline{\boldsymbol{\mu}}^*$ *determined by solving* $DUAL(SDP(P_=))$ *as explained above can be used in the partial Lagrangian program* $\left(\mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*)\right)$ *and make this program reach the value* $OPT(SDP(QP_=))$. *If* $(QP_=)$ *integrates the redundant constraints from Example 6.3.5, the hierarchy (6.2.13a) collapses as stated in Remark 6.3.7, i.e., we can write* $\left(\mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*)\right) = \left(TL^{\mathbf{X}}(QP_=)\right) = \left(PL^{\mathbf{X}}(QP_=)\right)$. *Applying (6.3.8), we obtain* $\left(\mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*)\right) = OPT(SDP(QP_=)) = \left(PL^{\mathbf{X}}(QP_=)\right)$, *i.e., both (6.3.1a)-(6.3.1b) collapse; the convexification is optimal, reaching its limit value* $\left(PL^{\mathbf{X}}(QP_=)\right)$ *from (6.3.1b). In fact, since we are in the* $0-1$ *case, we can also use the redundant constraint from Example 6.3.4 and obtain the same value* $OPT(SDP(Q_=)) = \left(PL^{\mathbf{X}}(QP_=)\right)$ *using Prop. 6.3.8.*

Finally, we showed in Example 6.3.2 that the optimal solution of $DUAL(SDP(QP_=))$ might only converge towards $OPT(SDP(QP_=))$ without effectivelly reaching this value. In this case, we can have a sequence of solutions $(\boldsymbol{\mu}^*_i,\overline{\boldsymbol{\mu}}^*_i)$ whose objective values converge towards the optimum. We can apply the same calculations from any $\boldsymbol{\mu}^*_i,\overline{\boldsymbol{\mu}}^*_i$ that reach an objective value arbitrarily close to the optimum $OPT(SDP(QP_=))$.

6.3.4.2 Using the optimal convexification coefficients in a convex quadratic `Branch-and-bound`

Let us here consider that the only quadratic constraints of $(QP_=)$ are of the form $x_i^2 = x_i\ \forall i \in [1..n]$ (*i.e.*, integrality constraints). However, $(QP_=)$ can integrate various redundant constraint sets, *e.g.*, like those from Example 6.3.4 or Example 6.3.5. In fact, these two constraint sets produce the same SDP relaxation value $OPT(SDP(QP_=))$ in the $0-1$ case, by virtue of Prop. 6.3.8. If we use any of these constraint sets, we obtain $OPT(SDP(QP_=)) = \left(PL^{\mathbf{X}}(QP_=)\right)$ as stated in Remark 6.3.9. This means that the convexification is optimal, because it reaches its limit $\left(PL^{\mathbf{X}}(QP_=)\right)$.

Once the optimal coefficients $\boldsymbol{\mu}^*$ and $\overline{\boldsymbol{\mu}}^*$ have been found, we focus on the fact that the augmented partial Lagrangian $\mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*)$ from (6.3.7) is convex with linear constraints and has the optimum value $OPT(SDP(QP_=))$ by virtue of (6.3.8). From now on, we can apply a convex quadratic solver to optimize by `Branch-and-bound` the binary version of $\mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*,\overline{\boldsymbol{\mu}}^*)$ from (6.3.7). One can verify that the objective value of $\mathbf{x}$ in this program is equal to the value of $\mathbf{x}$ in $(QP_=)$ whenever we have $A\mathbf{x} = \mathbf{b}$ and $\mathbf{x}$ is binary (*i.e.*, all quadratic constraints are satisfied). By solving the binary version of $\mathscr{L}_{PL\mathbf{x}_{(QP_=)}}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$ we actually solve $(QP_=)$. Optimizing a convex quadratic program takes usually less time than solving an SDP. We solved a single SDP program (*i.e.*, $(SDP(QP_=))$ and its dual equal to the total Lagrangian) to obtain the best convexification coefficients $\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*$; we use the quality of the SDP bound at the root of the `Branch-and-bound` tree. More importantly, at each `Brand-and-bound` node some of the $\mathbf{x}$ variables are fixed to binary values and we can still solve a reduced partial Lagrangian only using the remaining variables. The value of this reduced partial Lagrangian can be used to prune the node.

Appendix C.2.2 briefly discusses a more refined approach that further restricts the feasible area of $\mathscr{L}_{PL\mathbf{x}_{(QP_=)}}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$ by imposing the additional constraint $x_i \in [0,1]$, equivalent to $x_i^2 \leq x_i\ \forall i \in [1..n]$.

7 Basic elements of several other research topics: under construction

7.1 Approximation algorithms using SDP programming

We here only present a (famous) SDP 0.8785–approximation algorithm for Max–Cut, but one should keep in mind there are also other SDP approximation algorithms that exploit a similar approach.

Proposition 7.1.1. *(Goemans-Williamson SDP approximation algorithm) Consider a weighted graph $G = ([1..n], E)$ with weights $w_{ij} \geq 0\ \{i,j\} \in E$ and $w_{ij} = 0\ \forall \{i,j\} \notin E$. The Max-Cut problem requires splitting $[1..n]$ in two sub-sets so as to maximize the total weighted sum of the edges with end vertices in different subsets. The optimum value $OPT(MC_w)$ satisfies*

$$0.8785 \cdot OPT(SDP_w) < OPT(MC_w) \leq OPT(SDP_w),$$

where (SDP_w) is the SDP program:

$$(SDP_w) \begin{cases} \max & \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{1}{2} w_{ij}(1 - X_{ij}) & (7.1.1a) \\ s.t & \boldsymbol{diag}(X) = I_n & (7.1.1b) \\ & X \succeq \mathbf{0} & (7.1.1c) \end{cases}$$

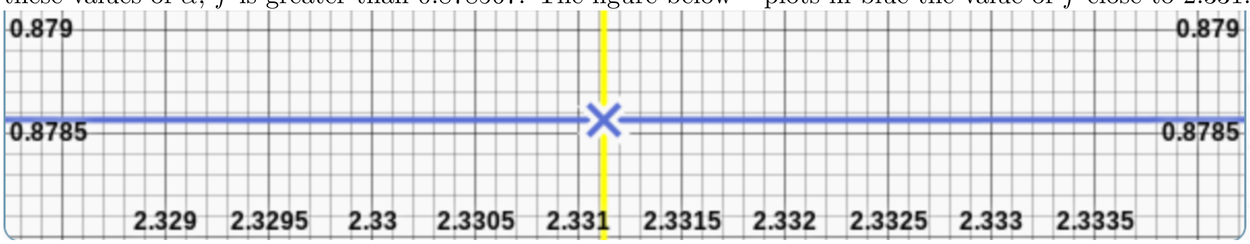
Proof. One can formulate the Max-Cut problem using variables $z_i \in \{-1,\ 1\}$ such that $z_i \neq z_j$ implies that vertices i and j belong to different subsets, and so, edge $\{i,j\}$ has a contribution w_{ij} to the objective function. We can formulate Max-Cut as $\max \left\{ \sum\limits_{i=1}^{n} \sum\limits_{i=j+1}^{n} \frac{1}{2} w_{ij} (1 - z_i z_j) :\ z_i \in \{-1, 1\}\ \forall i \in [1..n] \right\}$. We now apply the following relaxation: we transform z_i into a vector $\mathbf{y}_i$ (with $i \in [1..n]$) of the unit sphere. The product $z_i z_j$ is generalized to $X_{ij} = \mathbf{y}_i \cdot \mathbf{y}_j$, and so, X is a Gram matrix that needs to be SDP (use Prop. A.1.8) and that satisfies $X_{ii} = \mathbf{y}_i \cdot \mathbf{y}_i = 1$ (because $\mathbf{y}_i$ belongs to the unit sphere for all $i \in [1..n]$). We thus obtain that the above (SDP_w) from (7.1.1a)-(7.1.1c) is a relaxation of the Max-Cut problem, and so, $OPT(MC_w) \leq OPT(SDP_w)$.

We now prove $0.8785 OPT(SDP_w) < OPT(MC_w)$. From a feasible solution of (SDP_w) from (7.1.1a)-(7.1.1c) we can generate a feasible Max-Cut solution as follows. Take any vector $\mathbf{v} \in \mathbb{R}^n$ and set $z_i = -1$ if $\mathbf{y}_i \cdot \mathbf{v} \leq 0$ and $z_i = 1$ otherwise. For different vectors $\mathbf{v}$ we obtain different Max-Cut solutions. But the probability of separating $\mathbf{y}_i$ and $\mathbf{y}_j$ (with $i,j \in [1..n]$) so as to give rise to an objective function contribution of w_{ij} is exactly $\dfrac{\arccos(\mathbf{y}_i \cdot \mathbf{y}_j)}{\pi}$. The expected value resulting from taking different $\mathbf{v}$ vectors is $\sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{ij} \dfrac{\arccos(\mathbf{y}_i \cdot \mathbf{y}_j)}{\pi}$. This expected value needs to be less than or equal to $OPT(MC_w)$.

We will show that the expected value is greater than $0.8785 \cdot OPT(SDP_w)$. For this, notice each term $\dfrac{1}{2} w_{ij}(1 - \mathbf{y}_i \cdot \mathbf{y}_j)$ from (7.1.1a) corresponds to $w_{ij} \dfrac{\arccos(\mathbf{y}_i \cdot \mathbf{y}_j)}{\pi}$ in the above sum (representing the expected value). We calculate the minimum of

$$\frac{\frac{\arccos(\mathbf{y}_i \cdot \mathbf{y}_j)}{\pi}}{\frac{1}{2}(1 - \mathbf{y}_i \cdot \mathbf{y}_j)} = \frac{2}{\pi} \frac{\arccos(\mathbf{y}_i \cdot \mathbf{y}_j)}{1 - \mathbf{y}_i \cdot \mathbf{y}_j} = \frac{2}{\pi} \frac{\alpha}{1 - \cos(\alpha)} = f(\alpha).$$

For $\alpha = 0$ we have $\mathbf{y}_i \bullet \mathbf{y}_j = 1$ and the correspondence $\frac{1}{2}w_{ij}(1 - \mathbf{y}_i \bullet \mathbf{y}_j) \to w_{ij}\frac{\arccos(\mathbf{y}_i \bullet \mathbf{y}_j)}{\pi}$ is equivalent to $0 \to 0$, so we can ignore such terms. It is enough to prove $f(\alpha) = \frac{2}{\pi}\frac{\alpha}{1-\cos(\alpha)} > 0.8785$ for any $\alpha \in (0, \pi]$. The derivative in α is $f'(\alpha) = \frac{2\,(1 - \alpha\sin(\alpha) - \cos(\alpha))}{\pi\,(\cos(\alpha) - 1)^2}$. The denominator of $f'(\alpha)$ is thus positive over the whole interval of interest $\alpha \in (0, \pi]$. The derivative of the numerator is $-2\alpha\cos(\alpha)$, which is negative over $\alpha \in \left[0, \frac{\pi}{2}\right)$ and non-negative over $\alpha \in \left[\frac{\pi}{2}, \pi\right]$. As such, the numerator is 0 in $\alpha = 0$ and it decreases as α increases up to $\frac{\pi}{2}$ and starts increasing after $\frac{\pi}{2}$. This numerator continue to increase even after the point $\overline{\alpha}$ where $f'(\overline{\alpha}) = 0$. Consequently, the numerator is negative for $\alpha \leq \overline{\alpha}$ (and so is $f'(\alpha)$) and positive for $\alpha \geq \overline{\alpha}$ (and so is $f'(\alpha)$). Thus, the stationary point $\overline{\alpha}$ reaches the minimum of f. We obtain that for $\alpha = 2.33112237$ the derivative is slightly negative and for $\alpha = 2.33112238$ it is slightly positive. For both these values of α, f is greater than 0.878567. The figure below[34] plots in blue the value of f close to 2.331.



Using numerical arguments (see figure above), the value of f is always greater than 0.8785. $\square$

As a side remark, it was proved that the above ratio of the Goemans-Williamson algorithm is essentially optimal if the Unique Games Conjecture holds.[35]

7.2 Strong duality in the more general context of linear conic programming

All SDP and linear programs presented in this work are actually particular cases of more general linear conic programs. We here only present how the SDP strong duality actually holds in linear conic programming. The line of reasoning is a relatively direct generalization of analogous results from the SDP case. As mentioned in the first paragraph of Section 2.3, the initial ideas are taken from a course of Anupam Gupta, also using arguments from the lecture notes of László Lovász (see Footnote 11, p. 29).

7.2.1 A preliminary conic separation result

We need Prop. 7.2.4 below that generalizes Prop. 2.3.2. We first recall three basic definitions to establish a rigorous framework.

Definition 7.2.1. *A convex cone $\mathcal{C} \subset \mathbb{R}^m$ is a set closed under linear combinations with positive coefficients. In particular, if $X, Y \in \mathcal{C}$, then $tX \in \mathcal{C}$ $\forall t > 0$ and $X + Y \in \mathcal{C}$.*
If the cone is not convex, we only have $X \in \mathcal{C} \implies tX \in \mathcal{C}$ $\forall t > 0$.

Definition 7.2.2. *The dual (convex) cone $\mathcal{C}^*$ of cone $\mathcal{C}$ is defined by*

$$\mathcal{C}^* = \left\{ Y \in \mathbb{R}^m : \ \overline{X} \bullet Y \geq 0\ \forall \overline{X} \in \mathcal{C} \right\}.$$

Definition 7.2.3. *The interior of cone $\mathcal{C}$ is the cone*

$$\texttt{interior}(\mathcal{C}) = \{ Z \in \mathcal{C} :\ \exists \varepsilon > 0 \text{ s.t. } |Z' - Z| < \varepsilon \implies Z' \in \mathcal{C} \}.$$

In other words, the set $\texttt{interior}(\mathcal{C})$ contains an open ball around each of its elements.
The interior is a cone because if $\mathcal{C}$ contains a ball around Z, it also contains a ball scaled by a factor of t (shrinked or enlarged by t) around tZ $\forall t > 0$.

---

[34] The figure and the derivatives are obtained using http://derivative-calculator.net.

[35] In the article "Optimal inapproximability results for MAX-CUT and other 2-variable CSPs?" by Subhash Khot, Guy Kindler, Elchanan Mossel and Ryan O'Donnell, published in SIAM Journal on Computing in 2007, vol 37(1), pp 319–357, a preliminary version is available at https://www.cs.cmu.edu/~odonnell/papers/maxcut.pdf.

Proposition 7.2.4. *Let* $F(\mathbf{x}) = \sum_{i=1}^{n} x_i A_i - B$ *for any* $\mathbf{x} \in \mathbb{R}^n$*, where* $B \in \mathbb{R}^m$ *and* $A_i \in \mathbb{R}^m$ $\forall i \in [1..n]$*. We consider a closed convex cone* $\mathcal{C} \in \mathbb{R}^m$ *and its dual* $\mathcal{C}^*$*. The following needs to hold.*

$$F(\mathbf{x}) \notin \mathtt{interior}(\mathcal{C}) \ \forall \mathbf{x} \in \mathbb{R}^n \iff$$

$$\exists Y \in \mathcal{C}^*, \ Y \neq \mathbf{0} \text{ such that } F(\mathbf{x}) \bullet Y = -B \bullet Y \leq 0, \ \forall \mathbf{x} \in \mathbb{R}^n. \tag{7.2.1}$$

We say that $F(\mathbf{x})$ *belongs to the hyperplane* $\{X : \ X \bullet Y = -B \bullet Y\}$*. Above equation (7.2.1) implies* $A_i \bullet Y = 0 \ \forall i \in [1..n]$.

Proof.
$\Longleftarrow$
If $F(\mathbf{x}) \bullet Y \leq 0$ for some non-null $Y \in \mathcal{C}^*$, $F(\mathbf{x})$ can not belong to the interior of $\mathcal{C}$ because any $Z \in \mathtt{interior}(\mathcal{C})$ satisfies $Z \bullet Y > 0$. Indeed, if $Z \in \mathtt{interior}(\mathcal{C})$, then $\mathcal{C}$ contains a ball around Z, and so, there exists $\varepsilon > 0$ such that $Z - \varepsilon Y \in \mathcal{C}$. If we assume $Z \bullet Y = 0$, then $(Z - \varepsilon Y) \bullet Y = -\varepsilon |Y|^2 < 0$, since $Y \neq \mathbf{0}$. This is a contradiction on $Y \in \mathcal{C}^*$ and $Z - \varepsilon Y \in \mathcal{C}$. As such, any $Z \in \mathtt{interior}(\mathcal{C})$ satisfies $Z \bullet Y > 0$, and so, $F(\mathbf{x}) \notin \mathtt{interior}(\mathcal{C})$.
$\Longrightarrow$
We know that $\mathtt{interior}(\mathcal{C})$ does not intersect the image of F. We can apply the hyperplane separation Theorem C.4.1: there exists a non-zero $Y \in \mathbb{R}^m$ and a $c \in \mathbb{R}$ such that

$$F(\mathbf{x}) \bullet Y \leq c \leq X \bullet Y \ \forall \mathbf{x} \in \mathbb{R}^n, \ \forall X \in \mathtt{interior}(\mathcal{C}) \tag{7.2.2}$$

It is clear that we can not have $c > 0$ because $X \bullet Y$ can be arbitrarily close to 0 by choosing $X = \varepsilon Z \in \mathtt{interior}(\mathcal{C})$ for an arbitrarily small $\varepsilon > 0$ and some $Z \in \mathtt{interior}(\mathcal{C})$—recall $\varepsilon Z \in \mathtt{interior}(\mathcal{C})$, because $\mathtt{interior}(\mathcal{C})$ is a cone (Def. 7.2.3). We now prove $X \bullet Y \geq 0 \ \forall X \in \mathtt{interior}(\mathcal{C})$. Let us assume the contrary: $\exists X \in \mathtt{interior}(\mathcal{C})$ such that $X \bullet Y = c' < 0$. Since $\mathtt{interior}(\mathcal{C})$ is a cone (Def 7.2.3), we have $tX \in \mathtt{interior}(\mathcal{C}) \ \forall t > 0$. The value $(tX) \bullet Y = tc'$ can be arbitrarily low by choosing an arbitrarily large t, and so, $(tX) \bullet Y$ can be easily less than c, contradiction. This means that $X \bullet Y \geq 0 \ \forall X \in \mathtt{interior}(\mathcal{C})$. Using $c \leq 0$, (7.2.2) simplifies to

$$F(\mathbf{x}) \bullet Y \leq 0 \leq X \bullet Y \ \forall \mathbf{x} \in \mathbb{R}^n, \ \forall X \in \mathtt{interior}(\mathcal{C})$$

It is not hard to prove that $Y \in \mathcal{C}^*$. For this, we will show that $\overline{X} \bullet Y \geq 0 \ \forall \overline{X} \in \mathcal{C}$, relying on the above proved fact $\overline{X} \bullet Y \geq 0 \ \forall \overline{X} \in \mathtt{interior}(\mathcal{C})$. Assume the contrary: there is some $\overline{X} \in \mathcal{C}$ such that $\overline{X} \bullet Y < 0$. For any $\varepsilon > 0$, we have $\overline{X} + \varepsilon Z \in \mathtt{interior}(S_m^+)$ for any $Z \in \mathtt{interior}(\mathcal{C})$—because of the cone property and of the fact that if $\mathcal{C}$ contains a ball centered at Z, then $\mathcal{C}$ contains this ball shrinked by ε centered at $\overline{X} + \varepsilon Z$. For a small enough ε, $(\overline{X} + \varepsilon Z) \bullet Y$ remains strictly negative, which contradicts $\overline{X} + \varepsilon Z \in \mathtt{interior}(\mathcal{C})$. We obtain $\overline{X} \bullet Y \geq 0 \ \forall \overline{X} \in \mathcal{C}$. This means that $Y \in \mathcal{C}^*$ and Y satisfies

$$F(\mathbf{x}) \bullet Y \leq 0 \ \forall \mathbf{x} \in \mathbb{R}^n. \tag{7.2.3}$$

We write $d = F(\mathbf{0}) \bullet Y = -B \bullet Y$. Assume for the sake of contradiction there is an $\mathbf{x} \in \mathbb{R}^n$ such that $F(\mathbf{x}) = d + d'$ with $d' \neq 0$. Let us calculate $F\left(\frac{-d+1}{d'}\mathbf{x}\right) \bullet Y = d + \frac{-d+1}{d'} d' = d - d + 1 = 1$, which contradicts (7.2.3). The assumption $d' \neq 0$ was false, and so, we obtain

$$F(\mathbf{x}) \bullet Y = F(\mathbf{0}) \bullet Y = -B \bullet Y \leq 0, \ \forall \mathbf{x} \in \mathbb{R}^n$$

where we used (7.2.3) for the last inequality. $\square$

7.2.2 Recalling the weak duality for conic programming

Let us introduce the first conic linear program in variables $\mathbf{x} \in \mathbb{R}^n$ based on closed convex cone C, generalizing (SDP) from (2.1.1a)-(2.1.1c).

$$(\texttt{ConicLP})\begin{cases} \min \quad \sum_{i=1}^{n} c_i x_i & (7.2.4a) \\ \quad s.t \quad \sum_{i=1}^{n} A_i x_i - B \in C & (7.2.4b) \\ \quad\quad \mathbf{x} \in \mathbb{R}^n & (7.2.4c) \end{cases}$$

By relaxing the conic constraint (7.2.4b) with multipliers $Y \in C^*$, we obtain the following Lagrangian:

$$\mathscr{L}(\mathbf{x}, Y) = \sum_{i=1}^{n} c_i x_i - Y \bullet (\sum A_i x_i - B)$$

Notice that if $\mathbf{x}$ satisfies (7.2.4b), we get awarded in the above Lagrangian because $Y \bullet (\sum A_i x_i - B) \geq 0$ when $Y \in C^*$ and $\sum A_i x_i - B \in C$. Thus, $\min\limits_{\mathbf{x} \in \mathbb{R}^n} \mathscr{L}(\mathbf{x}, Y)$ is a relaxation of $(\texttt{ConicLP})$:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \mathscr{L}(\mathbf{x}, Y) \leq OPT(\texttt{ConicLP}), \tag{7.2.5}$$

where $OPT(\texttt{ConicLP})$ can also be unbounded $(-\infty)$.

Let us write:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \mathscr{L}(\mathbf{x}, Y) = \min_{\mathbf{x} \in \mathbb{R}^n} Y \bullet B + \sum_{i=1}^{n} (c_i - Y \bullet A_i) x_i$$

If there is a single $i \in [1..n]$ such that $c_i - Y \bullet A_i \neq 0$, the above minimum is $-\infty$ (unbounded), by using an appropriate value of x_i. To have a bounded $\min_{\mathbf{x} \in \mathbb{R}^n} \mathscr{L}(\mathbf{x}, Y)$, Y needs to verify $c_i - Y \bullet A_i = 0$ for all $i \in [1..n]$. We are interested in finding:

$$\max_{Y \in C^*} \min_{\mathbf{x} \in \mathbb{R}^n} \mathscr{L}(\mathbf{x}, Y)$$

that can be written:

$$(\texttt{DConicLP})\begin{cases} \max \ B \bullet Y & (7.2.6a) \\ \quad s.t \ A_i \bullet Y = c_i \ \forall i \in [1..n] & (7.2.6b) \\ \quad\quad Y \in C^* & (7.2.6c) \end{cases}$$

Based on (7.2.5), we obtain the weak duality:

$$OPT(\texttt{DConicLP}) \leq OPT(\texttt{ConicLP}). \tag{7.2.7}$$

In particular, if $OPT(\texttt{ConicLP}) = -\infty$, then $(\texttt{DConicLP})$ needs to be infeasible, because otherwise the above (7.2.7) would limit $OPT(\texttt{ConicLP})$ from going to $-\infty$.

7.2.3 Strong duality for the primal form

The following theorem generalizes Theorem 2.3.3 of the SDP case.

Theorem 7.2.5. *If the primal $(\texttt{ConicLP})$ from (7.2.4a)-(7.2.4c) is bounded and has a strictly feasible solution (Slater's interiority condition), then the primal and the dual optimal values are the same and the dual $(\texttt{DConicLP})$ from (7.2.6a)-(7.2.6c) reaches this optimum value. Recall (last phrase of Section 7.2.2) that if $(\texttt{ConicLP})$ is unbounded, then $(\texttt{DConicLP})$ is infeasible.*

Proof. Let p be the optimal primal value. The system $\sum\limits_{i=1}^{n} c_i x_i < p$ and $\sum_{i=1}^{n} A_i x_i - B \in C$ has no solution. We define

$$\overline{A}_i = \begin{bmatrix} -c_i \\ A_i \end{bmatrix} \ \forall i \in [1..n] \text{ and } \overline{B} = \begin{bmatrix} -p \\ B \end{bmatrix},$$

and the closed convex cone

$$\overline{C} = \left\{ \begin{bmatrix} x \\ X \end{bmatrix} : \ x \geq 0, \ X \in C \right\}$$

Notice that $\sum_{i=1}^{n} \overline{A}_i x_i - \overline{B} \notin \texttt{interior}(\overline{C}) \ \forall \mathbf{x} \in \mathbb{R}^n$ (we can *not* say $\sum_{i=1}^{n} \overline{A}_i x_i - \overline{B} \notin \overline{C}$, as the optimal solution $\mathbf{x}$ can cancel the value $-p$ on the first position of the expression). We can thus apply Prop. 7.2.4 (implication "$\Longrightarrow$") and conclude there is some non-zero $\overline{Y} \in \overline{C}^*$ such that $\overline{A}_i \bullet \overline{Y} = 0$ and $-\overline{B} \bullet \overline{Y} \leq 0$. Writing $\overline{Y} = \left[\begin{smallmatrix} t \\ Y \end{smallmatrix} \right]$ with $t \geq 0$ and $Y \in C^*$ (necessary and sufficient conditions to have $\overline{Y} \in \overline{C}^*$), we obtain:

$$tc_i = A_i \bullet Y, \ \forall i \in [1..n] \tag{7.2.8}$$

and

$$-B \bullet Y \leq -tp. \tag{7.2.9}$$

Assume $t = 0$. We obtain $A_i \bullet Y = 0 \ \forall i \in [1..n]$ and $-B \bullet Y \leq 0$ with $Y \in C^*$ as discussed above. Applying again Prop. 7.2.4 ("$\Longleftarrow$" implication), we conclude $\sum_{i=1}^{n} A_i x_i - B \notin \texttt{interior}(C) \ \forall \mathbf{x} \in \mathbb{R}^n$, which contradicts the fact the primal (`ConicLP`) from (7.2.4a)-(7.2.4c) is strictly feasible. We thus need to have $t > 0$.

Taking $\widehat{Y} = \frac{1}{t} Y \in C^*$ (recall Def. 7.2.1 and Def. 7.2.2), the above (7.2.8)-(7.2.9) become: $c_i = A_i \bullet \widehat{Y} \ \forall i \in [1..n]$ and $B \bullet \widehat{Y} \geq p$. In other words, $\widehat{Y}$ is a feasible solution in the dual (`DConicLP`) from (7.2.6a)-(7.2.6c) and it has an objective value $B \bullet \widehat{Y} \geq p$. Using the weak duality (7.2.7), the feasible solution $\widehat{Y}$ of (`DConicLP`) needs to satisfy $B \bullet \widehat{Y} \leq p$, and so, $B \bullet \widehat{Y} = p$, *i.e.*, the dual achieves the optimum primal value in $\widehat{Y}$. □

7.2.4 Strong duality for dual forms

We start from the conic dual. For the reader's convenience, we repeat the definitions of (`DConicLP`) from (7.2.6a)-(7.2.6c):

$$(\texttt{DConicLP}) \begin{cases} \max \ B \bullet Y & (7.2.10a) \\ \quad s.t. \ A_i \bullet Y = c_i \ \forall i \in [1..n] & (7.2.10b) \\ \quad\quad Y \in C^* & (7.2.10c) \end{cases}$$

The following is a generalization of Prop. 2.1.2.

Proposition 7.2.6. *Program (`DConicLP`) from (7.2.10a)-(7.2.10c) is equivalent to (`ConicLP'`) from (7.2.13a)-(7.2.13c) below which is a program in the primal form (7.2.4a)-(7.2.4c), provided that the system of linear equations $A_i \bullet Y = c_i \ \forall i \in [1..n]$ from (7.2.10b) has at least a feasible a solution.*

Proof. We first solve the system $A_i \bullet Y = c_i \ \forall i \in [1..n]$. This system has at least a feasible solution $-B'$. The set of all solutions is given by

$$Y = -B' + \sum_{j=1}^{k} A'_j x'_j \text{ with } x'_j \in \mathbb{R} \, \forall j \in [1..k], \tag{7.2.11}$$

where $A'_1, \ A'_2, \ldots A'_k$ are a basis of the null space of $\{A_i : \ i \in [1..n]\}$ satisfying:

$$A_i \bullet A'_j = 0, \ \forall i \in [1..n], \ j \in [1..k]. \tag{7.2.12}$$

The space spanned by (the linear combinations of) A_i ($\forall i \in [1..n]$) and A'_j ($\forall j \in [1..k]$) need to cover the whole space $\mathbb{R}^m$, by virtue of the rank-nullity Theorem A.1.3. Replacing (7.2.11) in the objective function of (`DConicLP`) from (7.2.10a), we obtain objective function $\max \left(-B' + \sum_{j=1}^{k} A'_j x'_j \right) \bullet B = \max -B \bullet B' +$

$\sum_{j=1}^k (B \bullet A'_j)x'_j = -\min B \bullet B' + \sum_{j=1}^k(-B \bullet A'_j)x'_j$. The dual (DConicLP) can be written:

$$(\texttt{ConicLP}')\begin{cases} -\min\ B \bullet B' + \sum_{j=1}^{k}(-B \bullet A'_j)x'_j & (7.2.13a)\\ \quad s.t.\ \sum_{j=1}^{k} A'_j x'_j - B' \in C^* & (7.2.13b)\\ \quad\quad \mathbf{x}' \in \mathbb{R}^k, & (7.2.13c)\end{cases}$$

which is a linear conic program in the primal form (7.2.4a)-(7.2.4c). □

Applying the duality (ConicLP) → (DConicLP) from Section 7.2.2 on above (ConicLP′), we obtain (DConicLP′) below. Notice we proved in Prop. 5.4.1 that closure$(C) = C^{**}$. Since we said from the beginning (first paragraph of Section 7.2.2) that C is closed, we can use $C = C^{**}$.

$$(\texttt{DConicLP}')\begin{cases} -\max\ B \bullet B' + B' \bullet Y' & (7.2.14a)\\ \quad s.t.\ A'_j \bullet Y' = -B \bullet A'_j\ \forall j \in [1..k] & (7.2.14b)\\ \quad\quad Y' \in C^{**} = C & (7.2.14c)\end{cases}$$

The system of equations (7.2.14b) has at least the solution $Y' = -B$ and this allows us to express (DConicLP′) in the primal form using (the approach from) Prop 7.2.6. Any solution of this system could be written as $-B$ plus a linear combination of vectors from the null space of $\{A'_j :\ j \in [1..k]\}$. But we said in Prop 7.2.6 (see (7.2.12) and above) that $\{A'_j :\ j \in [1..k]\}$ are a basis of the null space of $\{A_i :\ i \in [1..n]\}$. Vectors A'_j (with $j \in [1..k]$) and A_i (with $i \in [1..n]$) span the whole space $\mathbb{R}^m$, by virtue of the rank-nullity Theorem A.1.3. The null space of $\{A'_j :\ j \in [1..k]\}$ can thus be generated by the linear combinations of (a maximum set of independent vectors of) $\{A_i :\ i \in [1..n]\}$. This is enough to allow us to say that all solutions Y' of (7.2.14b) verify:

$$Y' = -B + \sum_{i=1}^{n} A_i x_i \text{ with } x_i \in \mathbb{R}^n\ \forall i \in [1..n].$$

Replacing this in the objective function (7.2.14a) of (DConicLP′) above, we obtain the objective function in variables $x_1,\ x_2, \ldots x_i$:

$$-\max B \bullet B' + \left(-B + \sum_{i=1}^{n} A_i x_i\right) \bullet B' = -\max \sum_{i=1}^{n}(A_i \bullet B')x_i = -\max \sum_{i=1}^{n} -c_i x_i = \min \sum_{i=1}^{n} c_i x_i,$$

where we used $A_i \bullet -B' = c_i$ from the first line of the proof of Prop. 7.2.6. This way, replacing the value of Y', program (DConicLP′) from (7.2.14a)-(7.2.14c) is completely equivalent to:

$$(\texttt{ConicLP})\begin{cases} \min \quad \sum_{i=1}^{n} c_i x_i & (7.2.15a)\\ \quad s.t \quad \sum_{i=1}^{n} A_i x_i - B \in C & (7.2.15b)\\ \quad\quad \mathbf{x} \in \mathbb{R}^n, & (7.2.15c)\end{cases}$$

which is exactly the initial program from (7.2.4a)-(7.2.4c).

Theorem 7.2.7. *If the dual (DConicLP) from (7.2.10a)-(7.2.10c) (or (7.2.6a)-(7.2.6c)) is bounded and has a strictly feasible solution, then the primal and the dual optimal values are the same and the primal (ConicLP) from (7.2.15a)-(7.2.15c) (or (7.2.4a)-(7.2.4c)) reaches this optimum value.*

Proof. Since (DConicLP) is feasible, we can apply Prop. 7.2.6 and obtain that (DConicLP) can be reformulated as (ConicLP′) from (7.2.13a)-(7.2.13c). This (ConicLP′) needs to be bounded and to have a strictly feasible solution as well, and so, we can apply Theorem 7.2.5 that states that the dual of (ConicLP′) reaches the optimum value of (ConicLP′). But the dual of (ConicLP′) is (DConicLP′) from (7.2.14a)-(7.2.14c), which is completely equivalent to (ConicLP) from (7.2.15a)-(7.2.15c). We obtained that (ConicLP) reaches the optimum value of (DConicLP). □

Theorem 7.2.8. *If both the primal* (`ConicLP`) *from* (7.2.15a)-(7.2.15c) *(or* (7.2.4a)-(7.2.4c)*) and the dual* (`DConicLP`) *from* (7.2.10a)-(7.2.10c) *(or* (7.2.6a)-(7.2.6c)*) are strictly feasible, then they have the same optimum value and they both reach it.*

Proof. Using weak duality (7.2.7), both programs need to be bounded. The conclusion then simply follows from combining Theorem 7.2.5 and Theorem 7.2.7. □

7.3 Polynomial Optimization

7.4 Algorithms for SDP optimization

A On ranks, determinants and space dimensions

This is the most elementary section from this document. If you are an absolute beginner, you could read first App. A.2 before App. A.1.

A.1 The rank-nullity theorem and other interesting rank properties

Definition A.1.1. *Given matrix $A \in \mathbb{R}^{n\times m}$, the rank* $\textbf{\textit{rank}}(A)$ *has two equivalent definitions:*

(a) the largest order of any non-zero minor of A (see Def A.2.1)

(b) the maximum number of independent rows (or columns) of A, i.e., the dimension of the space generated by the rows (or columns) of A using linear combinations (see also Def. A.2.7 for the notion of dimension).

Proof.
$(a) \implies (b)$
We show that the existence of a non-zero minor of order r. Without loss of generality, we permute rows and columns until the non-zero minor is positioned in the upper-left corner, and, so we can consider that $det([A]_r) \neq 0$, where $[A]_r$ is the leading principal minor of size $r \times r$, where $r = rank(A)$. Since the rows and columns of $[A]_r$ are independent, the corresponding rows and columns in the full matrix are linearly independent (in A). We have at least r independent rows and columns.

We will show that any row $\mathbf{a}_i$ with $i > r$ is dependent of rows $\mathbf{a}_1,\ \mathbf{a}_2, \dots \mathbf{a}_r$. Since $det([A]_r) \neq 0$, there exist $x_1, x_2, \dots x_r \in \mathbb{R}$ such that

$$[a_{i1}\ a_{i2}\ \dots a_{ir}] = [x_1\ x_2\ \dots x_r][A]_r \tag{A.1.1}$$

Take any column $j \in [r+1..m]$ and consider the matrix obtained by bordering $[A]_r$ with row i and column j.

$$\begin{bmatrix} & & & & a_{1j} \\ & & & & a_{2j} \\ & & \Large [A]_r & & \vdots \\ & & & & a_{rj} \\ a_{i1} & a_{i2} & \dots & a_{ir} & a_{ij} \end{bmatrix}$$

We subtract from the last row the first r rows respectively multiplied by $x_1,\ x_2, \dots x_r$. By virtue of (A.1.1), the above matrix becomes:

$$\begin{bmatrix} & & & & a_{1j} \\ & & & & a_{2j} \\ & \Large [A]_r & & & \vdots \\ & & & & a_{rj} \\ 0 & 0 & \dots & 0 & a_{ij} - \sum_{k=1}^{r} x_k a_{kj} \end{bmatrix}$$

Since $[A]_r$ is the largest non zero minor, both above bordered matrices need to have a zero determinant. But the determinant of the last above matrix is $(a_{ij} - \sum_{k=1}^{r} x_k a_{kj})\, det([A]_r)$. Since $det([A]_r) \neq 0$, we obtain that $a_{ij} = \sum_{k=1}^{r} x_k a_{kj}$. Since this needs to hold for all $j \in [r+1..m]$, we obtain $\mathbf{a}_i = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \dots x_r\mathbf{a}_r$, and so, row i is dependent on the first r rows. The number of independent rows is thus exactly r. An analogous

argument can be used to obtain that the number of independent columns is r. This is the dimension of the space generated by the rows (resp. columns).
$(b) \implies (a)$
The determinant of any minor that contains rows $r+1$ rows (resp. columns) is zero, because one of these rows (resp. columns) can be written as a linear combination of the other r. The order of the largest non-zero minor is thus at most r. It is easy to see that this largest order can not be $r-1$. If that were the case, we could apply the above $(a) \implies (b)$ proof to show that all rows are dependent on some $r-1$ rows. $\square$

Proposition A.1.2. *Given symmetric matrix $A \in \mathbb{R}^{n \times n}$ of rank r, A has at least one non-zero principal minor of order r.*

Proof. The rank definition ensures the existence of a set of rows J (with $|J| = r$) such that any other row $i \in [1..n] - J$ can be written as a linear combination of rows J. This means that the matrix A reduced to rows J has full rank r. By symmetry, each column of this latter matrix with r rows can be written as a linear combination of the columns J. This means that the matrix A reduced to rows J and (then) to columns J has full rank r, *i.e.*, the principal minor corresponding to rows J and columns J is non-zero. $\square$

The above proposition is sometimes referred to as the "principal minor theorem" and it also holds if A is skew-symmetric, *i.e.*, if $A^\top = -A$.

Theorem A.1.3. *(Rank-nullity theorem) Given $A \in \mathbb{R}^{n \times m}$, the null space (kernel) of A is given by*

$$\mathtt{null}(A) = \{\mathbf{x} \in \mathbb{R}^m : \ A\mathbf{x} = \mathbf{0}_n\} \tag{A.1.2}$$

We denote by $nullity(A)$ the dimension (maximum number of linearly independent vectors—see also Def. A.2.7) of $\mathtt{null}(A)$. The following equation holds:

$$m = rank(A) + nullity(A) \tag{A.1.3}$$

Proof. Without loss of generality, we permute rows and columns until a non-zero principal minor of size $r = rank(A)$ is positioned in the upper-left corner, so that the leading principal minor $[A]_r$ has a non-null determinant. We want to study the set of solutions of

$$A\mathbf{x} = \mathbf{0}_n \tag{A.1.4}$$

Using Def A.1.1, the last $n-r$ rows of A can be written as a linear combination of the first r rows. It is enough to investigate only the first r equations in above system (A.1.4). We can say that $\mathbf{x}$ is a solution of above system if and only if it is a solution of the following:

$$\left[\begin{array}{c|cccc} \multirow{4}{*}{$\Large [A]_r$} & a_{1,r+1} & a_{1,r+2} & \dots & a_{1,m} \\ & a_{2,r+1} & a_{2,r+2} & \dots & a_{2,m} \\ & \vdots & \vdots & \ddots & \vdots \\ & a_{r,r+1} & a_{r,r+2} & \dots & a_{r,m} \end{array}\right] \mathbf{x} = \mathbf{0}_r$$

Using simple notational shortcuts, we write the above as follows:

$$\begin{bmatrix} [A]_r & [A]_{m-r} \end{bmatrix} \begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_{m-r} \end{bmatrix} = \mathbf{0}_r, \tag{A.1.5}$$

where $[A]_{m-r}$ is the A minor obtained by selecting the first r rows and last $m-r$ columns; $\mathbf{x}_r$ selects the first r components of $\mathbf{x}$ and $\mathbf{x}_{m-r}$ selects the last $m-r$. We can further re-write (A.1.5) above as:

$$[A]_r\mathbf{x}_r + [A]_{m-r}\mathbf{x}_{m-r} = \mathbf{0}_r$$

We can now write $\mathbf{x}_r$ as a function of $\mathbf{x}_{m-r}$, more exactly:

$$\mathbf{x}_r = -[A]_r^{-1}[A]_{m-r}\mathbf{x}_{m-r}$$

We can say that each of the first r components of $\mathbf{x}$ (*i.e.*, $\mathbf{x}_r$) can be written as a linear combination of the last $m-r$ components (*i.e.*, $\mathbf{x}_{m-r}$). The dimension of the space generated by all above solutions $\mathbf{x}$ of (A.1.4) reduces to the dimension of the space generated by the last $m-r$ components $\mathbf{x}_{m-r}$, *i.e.*, this dimension is $m-r$, which confirms (A.1.3). $\square$

Proposition A.1.4. *Given $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times p}$, we have*

$$rank(AB) \leq min\big(rank(A), rank(B)\big) \tag{A.1.6}$$

Proof. Take any $\mathbf{x} \in \mathbb{R}^n$ in the null space of $A^\top$, *i.e.*, $A^\top\mathbf{x} = \mathbf{0}_m$ or $\mathbf{x}^\top A = \mathbf{0}_m^\top$. We observe that $\mathbf{x}^\top AB = \mathbf{0}_m^\top B = \mathbf{0}_p$. This means that $\mathbf{x}$ also belongs to the null space of $(AB)^\top$. This means that the null space of $(AB)^\top$ is greater than or equal to the null space of $A^\top$. We can write $nullity\big((AB)^\top\big) \geq nullity(A)$. Using the rank-nullity theorem (Th A.1.3), we have $rank(A^\top) + nullity(A^\top) = n$ and $rank\big((AB)^\top\big) + nullity\big((AB)^\top\big) = n$. This means that $rank\big((AB)^\top\big) \leq rank(A^\top)$, equivalent to $rank(AB) \leq rank(A)$. Analogously, we can obtain $rank(AB) \leq rank(B)$, which leads to (A.1.6). $\square$

Proposition A.1.5. *Given any $A \in \mathbb{R}^{n\times n}$ and invertible U, the following holds:*

$$rank(AU) = rank(A)$$

Proof. Using (A.1.6), we have $rank(A) \geq rank(AU) \geq rank(AUU^{-1}) = rank(A)$, which means that all inequalities are actually equalities. $\square$

Proposition A.1.6. *Any similar matrices $A, B \in \mathbb{R}^{n\times n}$ (i.e., such that $B = U^{-1}AU$ for some $U \in R^{n\times n}$) have the same rank.*

Proof. This simply follows from applying Proposition A.1.5 twice, once for U and once for U^{-1}. $\square$

Proposition A.1.7. *The rank of symmetric $A \in \mathbb{R}^{n\times n}$ is equal to n minus the multiplicity of the eigenvalue 0. We can write:*

$$n = rank(A) + eigenmult(0)$$

Proof. We use the eigendecomposition (as proved in Prop B.2.1):

$$A = U\Lambda U^\top,$$

where U is invertible ($U^{-1} = U^\top$) and Λ is a diagonal matrix with the eigenvalues of A on the diagonal. Using Prop A.1.6, A and Λ have the same rank. Since Λ is diagonal, its rank is equal to the number of non-zero elements on the diagonal, *i.e.*, n minus the multiplicity of eigenvalue 0. Additionally, remark that we proved in Prop. B.2.3 and Prop. B.2.5 that in symmetric matrices the algebraic multiplicity of an eigenvalue is equal to its geometric multiplicity, hence we use the term multiplicity to refer to both. $\square$

Proposition A.1.8. *Given $V \in \mathbb{R}^{m\times n}$, the matrix $A = V^\top V \in \mathbb{R}^{n\times n}$ is SDP. We say A is the Gram matrix of the column vectors of V. Furthermore, $rank(V) = rank(V^\top V) = rank(V^\top SV)$ for any positive definite $S \in \mathbb{R}^{m\times m}$.*

Proof.
1) To see $V^\top V$ is SDP, notice that for any $\mathbf{x} \in \mathbb{R}^n$, we have $\mathbf{x}^\top(V^\top V)\mathbf{x} = (\mathbf{x}^\top V^\top)(V\mathbf{x}) = (V\mathbf{x})^\top(V\mathbf{x}) = |V\mathbf{x}|^2$. Also, by writing $\mathbf{y} = V\mathbf{x}$, this value simply becomes $\sum_{i=1}^n y_i^2 \geq 0$.
2) $rank(V^\top V) = rank(V)$ follows from $rank(V^\top SV) = rank(V)$ with $S = I_m$. We now prove the equality for an arbitrary $S \succ \mathbf{0}$. We will show $V^\top SV$ and V have the same null space so as to apply the rank-nullity Theorem A.1.3. It is clear that any $\mathbf{x} \in \mathbb{R}^n$ in the null space of V belongs to the null space of $V^\top SV$, because $V\mathbf{x} = \mathbf{0}_m \Longrightarrow V^\top SV\mathbf{x} = \mathbf{0}_n$. We now prove the converse: $V^\top SV\mathbf{x} = \mathbf{0}_n \Longrightarrow V\mathbf{x} = \mathbf{0}_m$. We multiply both sides of $V^\top SV\mathbf{x} = \mathbf{0}_n$ by $\mathbf{x}^\top$ and we obtain $\mathbf{x}^\top V^\top SV\mathbf{x} = 0$, equivalent to $(V\mathbf{x})^\top S(V\mathbf{x}) = 0$. Using $S \succ \mathbf{0}$, this value can only be zero if $V\mathbf{x} = \mathbf{0}_m$. We obtained that the null space of V is equal to the null space of $V^\top SV$. Using the rank-nullity Theorem A.1.3, the two matrices need to have the same rank. $\square$

A.2 Results on determinants and space dimensions

Definition A.2.1. *Given matrix A of size $n \times m$, a minor of A is a sub-matrix obtained by selecting only some rows $J_1 \subseteq [1..n]$ and some columns $J_2 \subseteq [1..m]$ of A. A principal minor $[A]_J$ is a minor obtained by selecting the same rows and columns $J = J_1 = J_2$. A principal minor $[A]_J$ is a leading principal minor if $J = [1..p]$ for some $p \leq n, m$. We say $[A]_J$ is null (or zero) if $\det([A]_J) = 0$. The order of $[A_J]$ is $|J|$.*

A.2.1 Very elementary results on matrices and determinants

Proposition A.2.2. *Given complex matrix $A \in \mathbb{C}^{n\times n}$, $det(A) = 0 \iff \exists \mathbf{u} \in \mathbb{C}^n - \{\mathbf{0}\}$ such that $A\mathbf{u} = \mathbf{0}$.*

Proof.
$\Longleftarrow$
Take $\mathbf{u} \neq 0$ such that $A\mathbf{u} = \mathbf{0}$. Without loss of generality, we assume $u_1 \neq 0$. We can write:

$$\begin{bmatrix} a_{11} \\ a_{12} \\ \vdots \\ a_{1n} \end{bmatrix} + \frac{u_2}{u_1} \begin{bmatrix} a_{21} \\ a_{22} \\ \vdots \\ a_{2n} \end{bmatrix} + \frac{u_3}{u_1} \begin{bmatrix} a_{31} \\ a_{32} \\ \vdots \\ a_{3n} \end{bmatrix} + \ldots\ldots + \frac{u_n}{u_1} \begin{bmatrix} a_{n1} \\ a_{n2} \\ \vdots \\ a_{nn} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

We use the fact that adding a multiple of a column to another column does not change the determinant. By performing all additions of column multiples from above formula, we obtain only zeros in the first column. The determinant of the resulting matrix can only be zero.
$\Longrightarrow$
We proceed by induction. For $n = 1$, the implication is obviously true. Assume that it holds for $n - 1$. The implication is also obvious if $A = \mathbf{0}_{n\times n}$. We assume that A has some non-zero elements and without loss of generality we can use $a_{11} \neq 0$ (it is enough to permute lines/columns to obtain this).

We want to find $\mathbf{u} \in \mathbb{C}^n - \{0\}$ such that $A\mathbf{u} = \mathbf{0}$. A can be changed to a form in which a_{11} becomes 1: it is enough to divide first line by the initial a_{11} to obtain an equivalent system of equations. To simplify notations, we can continue assuming $a_{11} = 1$. We perform Gaussian elimination on first column using pivot $a_{11} = 1$. We want to solve:

$$\begin{bmatrix} 1 & a_{12} & a_{13} & \ldots & a_{1n} \\ 0 & a_{22} - a_{12} & a_{23} - a_{13} & \ldots & a_{2n} - a_{1n} \\ 0 & a_{32} - a_{12} & a_{33} - a_{13} & \ldots & a_{3n} - a_{1n} \\ \vdots & & & & \\ 0 & a_{n2} - a_{12} & a_{n3} - a_{13} & \ldots & a_{nn} - a_{1n} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Using compact notations, this can also be written:

$$\begin{bmatrix} 1 & \mathbf{b}^\top \\ \mathbf{0}_{n-1} & B \end{bmatrix} \begin{bmatrix} u_1 \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{0}_{n-1} \end{bmatrix} \tag{A.2.1}$$

Since above row multiplications and addition did not change the determinant of the matrix, we obtain $det\left(\begin{smallmatrix} 1 & \mathbf{b}^\top \\ \mathbf{0}_{n-1} & B \end{smallmatrix}\right) = 0$, which means $det(B) = 0$. We can use the induction hypothesis: there exists non-zero $\mathbf{v} \in \mathbb{C}^{n-1}$ such that $B\mathbf{v} = \mathbf{0}_{n-1}$. The last n-1 equations of above system are satisfied. We still need to find u_1 such that $u_1 + \mathbf{b}^\top\mathbf{v} = 0 \implies u_1 = -\mathbf{b}^\top\mathbf{v}$. We have just found a solution $\mathbf{u} = \left[\begin{smallmatrix} u_1 \\ \mathbf{v} \end{smallmatrix}\right] = 0$ for (A.2.1). It is easy to check that $\mathbf{u}$ verifies $A\mathbf{u} = 0$, *i.e.*, it is enough to reverse the Gaussian elimination since the first equation/line is verified by $\mathbf{u}$. $\square$

Corollary A.2.3. *If we replace $\mathbb{C}$ by $\mathbb{R}$ in the proof of above Proposition A.2.2, everything remains correct. The solution of the system will be real, because $u_1 = -\mathbf{b}^\top\mathbf{v}$ is real using the induction hypothesis that $\mathbf{v} \in \mathbb{R}^{n-1}$.*

Proposition A.2.4. *The determinant of a matrix A is the product of its eigenvalues. The trace of A is equal to the sum of the eigenvalues.*

Proof. Consider the characteristic polynomial $det(xI - A)$. The eigenvalues of A are the roots of $det(xI - A)$, and so,

$$det(xI - A) = (x - \lambda_1)(x - \lambda_2)\ldots(x - \lambda_n). \tag{A.2.2}$$

1) It is enough to evaluate this equation in $x = 0$ and we obtain $det(-A) = (-1)^n\lambda_1\lambda_2\ldots\lambda_n$. We also have $det(-A) = (-1)^n det(A)$, because any term in the Leibniz formula for determinants is a product of n elements of the matrix. This simply leads to $(-1)^n det(-A) = (-1)^n\lambda_1\lambda_2\ldots\lambda_n$, and so, $det(A) = \lambda_1\lambda_2\ldots\lambda_n$

2) We evaluate the term corresponding to x^{n-1} of both sides of (A.2.2). In the right side we obtain the term $-\sum_{i=1}^{n} \lambda_i x^{n-1}$. We will show that in the lefthand side we obtain $-\sum_{i=1}^{n} \mathbf{a}_{ii} x^{n-1}$. Using the Leibniz formula for determinants, we observe that the only determinant terms that make x arise $n-1$ times in $det(xI - A)$ are those that use $n-1$ diagonal terms of the form $x - a_{ii}$ with $i \in [1..n]$. Such a determinant term needs to also use the n^{th} diagonal value as well. To find the term of x^{n-1} in $det(xI - A)$ we need to develop $(x - a_{11})(x - a_{22}) \ldots (x - a_{nn})$. The x^{n-1} term is $-\sum_{i=1}^{n} \mathbf{a}_{ii} x^{n-1}$. $\square$

Proposition A.2.5. *Given $A, B \in \mathbb{C}^n$, if $AB = I_n$ then $BA = I_n$. We want a proof for the lazy, without calculating some left A^{-1} or some right B^{-1}.*

Proof. We first show that the columns of B are linearly independent. This follows from $\boldsymbol{\alpha} \in \mathbb{C}^n - \{\mathbf{0}\} \implies (AB)\boldsymbol{\alpha} \neq \mathbf{0}$. It is clear we can not have $B\boldsymbol{\alpha} = \mathbf{0}$ because this would make $AB\boldsymbol{\alpha} = \mathbf{0}$. The columns of B need to be linearly independent, and so, their linear combinations cover a space of dimension n, *i.e.*, $\mathbb{C}^n$ (see also Prop. A.2.7). As such, for any column $\mathbf{x}_i$ of $I_n = [\mathbf{x}_1\ \mathbf{x}_2 \ldots \mathbf{x}_n]$, there exist a linear combination of the columns of B that is equal to $\mathbf{x}_i$, *i.e.*, $\exists \mathbf{y}_i$ s. t. $\mathbf{x}_i = B\mathbf{y}_i$. Joining toghether all columns $\mathbf{x}_i$ of I_n, there exists $Y = [\mathbf{y}_1\ \mathbf{y}_2 \ldots \mathbf{y}_n]$ s. t. $I_n = BY$. We finish by $Y = (AB)Y = A(BY) = A$ which proves $BA = I_n$. $\square$

A.2.2 The dimension of a (sub-)space

Definition A.2.6. *The vectors $\mathbf{x}_1,\ \mathbf{x}_2, \ldots \mathbf{x}_k \in \mathbb{R}^n$ are affinely indepedent if there is no $\lambda_1,\ \lambda_2, \ldots \lambda_k$ with $\sum_{i=1}^{k} \lambda_i = 0$ such that $\sum_{i=1}^{k} \lambda_i \mathbf{x}_i = \mathbf{0}_n$. This is equivalent to the fact that $\mathbf{x}_1 - \mathbf{x}_k,\ \mathbf{x}_2 - \mathbf{x}_k, \ldots \mathbf{x}_{k-1} - \mathbf{x}_k$ are linearly independent.*

Proof. It is enough to show that

$$\mathbf{x}_1,\ \mathbf{x}_2, \ldots \mathbf{x}_k \text{ affinely dependent} \iff \mathbf{x}_1 - \mathbf{x}_k,\ \mathbf{x}_2 - \mathbf{x}_k, \ldots \mathbf{x}_{k-1} - \mathbf{x}_k \text{ linearly dependent}$$

$\implies$
Given $\lambda_1,\ \lambda_2, \ldots \lambda_k$ with $\sum_{i=1}^{k} \lambda_i = 0$ such that $\sum_{i=1}^{k} \lambda_i \mathbf{x}_i = \mathbf{0}_n$, we obtain

$$\sum_{i=1}^{k-1} \lambda_i (\mathbf{x}_i - \mathbf{x}_k) = \underbrace{\sum_{i=1}^{k} \lambda_i (\mathbf{x}_i - \mathbf{x}_k)}_{\text{we added } +\lambda_k(\mathbf{x}_k - \mathbf{x}_k)} = \sum_{i=1}^{k} \lambda_i \mathbf{x}_i - \sum_{i=1}^{k} \lambda_i \mathbf{x}_k = \mathbf{0}_n - 0\mathbf{x}_k = \mathbf{0}_n$$

$\impliedby$
We consider there is $\lambda_1,\ \lambda_2, \ldots \lambda_{k-1}$ such that $\sum_{i=1}^{k-1} \lambda_i (\mathbf{x}_i - \mathbf{x}_k) = 0$, equivalent to $\sum_{i=1}^{k-1} \lambda_i \mathbf{x}_i - \sum_{i=1}^{k-1} \lambda_i \mathbf{x}_k = 0$. By taking $\lambda_k = -\sum_{i=1}^{k-1} \lambda_i$, this reduces to $\sum_{i=1}^{k-1} \lambda_i \mathbf{x}_i + \lambda_k \mathbf{x}_k = 0$. $\square$

Definition A.2.7. *A subspace $S \subseteq \mathbb{R}^n$ has dimension k if $k+1$ is the maximum number of affinely independent vectors $\mathbf{x}_0, \mathbf{x}_1,\ \mathbf{x}_2 \ldots x_k \in S$. This is equivalent to the existence of maximum k linearly independent vectors $\mathbf{x}_1 - \mathbf{x}_0,\ \mathbf{x}_2 - \mathbf{x}_0, \ldots \mathbf{x}_k - \mathbf{x}_0$ by virtue of Def A.2.6. We say that $\mathbf{x}_0$ is an orgin and $\mathbf{x}_1 - \mathbf{x}_0,\ \mathbf{x}_2 - \mathbf{x}_0, \ldots \mathbf{x}_k - \mathbf{x}_0$ are a basis of S.*

If S contains $\mathbf{0}_n$, we can take $\mathbf{x}_0 = \mathbf{0}_n$ and the definition is equivalent to the existence of maximum k linearly independent vectors $\mathbf{x}_1,\ \mathbf{x}_2, \ldots \mathbf{x}_k$.

A.2.3 Every eigenvalue belongs to a Gershgorin disc

Theorem A.2.8. *(Gershgorin circle theorem) Given complex matrix $A \in \mathbb{C}^{n \times n}$, every eigenvalue λ belongs to a Gershgorin disk of the following form for some $i \in [1..n]$.*

$$\left\{ z \in \mathbb{C} :\ |z - A_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |A_{ij}| \right\}$$

Proof. Consider eigenvalue λ with an eigenvector $\mathbf{v} \in \mathbb{C}^n - \{\mathbf{0}\}$. Take an index $i \in [1..n]$ such that $|v_i| \geq |v_j|\ \forall j \in [1..n]$. This means that $\left|\dfrac{v_j}{v_i}\right| \leq 1\ \forall j \in [1..n]$. We now develop position i of $A\mathbf{v} = \lambda\mathbf{v}$ and we obtain

$\lambda v_i = \sum_{j=1}^{n} A_{ij} v_j$. Dividing this by v_i leads to $\lambda = A_{ii} + \sum_{\substack{j=1\\j\neq i}}^{n} A_{ij}\frac{v_j}{v_i}$. We can further develop:

$$\begin{aligned}
|\lambda - A_{ii}| &= \left|\sum_{\substack{j=1\\j\neq i}} A_{ij}\frac{v_j}{v_i}\right| \\
&\leq \sum_{\substack{j=1\\j\neq i}} \left|A_{ij}\frac{v_j}{v_i}\right| && \text{(we used the triangle inequality)} \\
&\leq \sum_{\substack{j=1\\j\neq i}}^{n} |A_{ij}|, && \text{(we used } \left|\frac{v_j}{v_i}\right| \leq 1\text{)}
\end{aligned}$$

which proves that λ belongs to the Gershgorin disk associated to i. □

B Three decompositions: eigenvalue, QR and square root

B.1 Preliminaries on eigen-values/vectors and similar matrices

Proposition B.1.1. *Given matrix $A \in \mathbb{R}^{n\times n}$, one can always find some $\lambda \in \mathbb{C}$ and $\mathbf{u} \in \mathbb{C}^n$ such that $A\mathbf{u} = \lambda\mathbf{u}$, i.e., there is at least one eigenvalue associated to an eigenvector.*

Proof. We consider the characteristic polynomial $det(xI - A) = 0$. Using the fundamental theorem of algebra, this polynomial has at least one complex root λ. We have $det(A - \lambda I) = 0$. Using Proposition A.2.2, there exists some $\mathbf{u} \in \mathbb{C}^n$ such that $(A - \lambda I)\mathbf{u} = \mathbf{0} \implies A\mathbf{u} = \lambda\mathbf{u}$. □

Proposition B.1.2. *All eigenvalues and proper eigenvectors of a real symmetric matrix are real (non complex).*

Proof. Take symmetric matrix $A \in \mathbb{R}^{n\times n}$, as well as λ and $\mathbf{u}$ such that $\lambda\mathbf{u} = A\mathbf{u}$. Let $\mathbf{u}^*$ be the conjugate transpose of $\mathbf{u}$, *i.e.*, transpose $\mathbf{u}$ and negate all terms that contain i. We have:

$$\mathbf{u}^*\lambda\mathbf{u} = \mathbf{u}^*A\mathbf{u}$$

We take the conjugate transpose of both sides (transpose and then negate imaginary terms) and obtain:

$$\mathbf{u}^*\lambda^*\mathbf{u} = \mathbf{u}^*A\mathbf{u}.$$

We used the fact that the conjugate of an expression can be obtained by conjugating each of the expression's terms,[36] as well as $(\mathbf{u}^*)^* = \mathbf{u}$ and $(A^*)^* = A$. The two above expressions have the same right-hand side and need to be equal: $\mathbf{u}^*\lambda\mathbf{u} = \mathbf{u}^*\lambda^*\mathbf{u} \implies \lambda(\mathbf{u}^*\mathbf{u}) = \lambda^*(\mathbf{u}^*\mathbf{u})$. Since $(a+bi)(a-bi) = a^2+b^2$, it is easy to check that $\sum_i^n u_i^* u_i$ is a strictly positive real (unless $\mathbf{u} = \mathbf{0}$). As such, $\lambda = \lambda^*$, *i.e.*, λ is real.

We now show that a proper eigenvector $\mathbf{u}$ is real. Suppose $\mathbf{u} = \mathbf{u}_a + \mathbf{u}_b i$, with $\mathbf{u}_a, \mathbf{u}_b \in \mathbb{R}^n$. Since $(A - \lambda I)(\mathbf{u}_a + \mathbf{u}_b i) = \mathbf{0}$, we obtain: $(A - \lambda I)\mathbf{u}_a = \mathbf{0}$ and $(A - \lambda I)\mathbf{u}_b = \mathbf{0}$. This means that $\mathbf{u}$ is merely a combination of other eigenvectors $\mathbf{u}_a$ and $\mathbf{u}_b$ of λ. We consider that the proper eigenvectors are $\mathbf{u}_a$ and $\mathbf{u}_b$. One can always multiply them by complex numbers and combine them to obtain eigenvectors like $\mathbf{u}$. □

Definition B.1.3. *(algebraic and geometric multiplicity) The algebraic multiplicity of an eigenvalue λ of matrix $A \in \mathbb{C}^{n\times n}$ is the multiplicity of root λ in the characteristic polynomial $det(xI - A)$. The geometric multiplicity of λ is the dimension of the eigen space $\{\mathbf{u} \in \mathbb{C}^n : A\mathbf{u} = \lambda\mathbf{u}\}$, i.e., the maximum number of linearly independent eigenvectors of λ (this is how we calculate the dimension of any space that includes $\mathbf{0}$, recall Def. A.2.7). The two multiplicities are not necessarily equal.*

[36] If you are unfamiliar with complex numbers, take the product of two complex numbers: $a_1a_2 - b_1b_2 + (a_1b_1 + a_2b_2)i = (a_1 + b_1i)(a_2 + b_2i)$. By conjugating each term in the right-hand side, we obtain the conjugate of the left-hand side, *i.e.*, $a_1a_2 - b_1b_2 - (a_1b_1 + a_2b_2)i$. For additions, the property is even easier to verify.

Proof. We give an example in which the two multiplicities are not equal. Take $A = \left[\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right]$. The characteristic polynomial is $(x-1)^2 = 0$, and so, the multiplicity of eigenvalue $\lambda = 1$ is 2. To compute the geometric multiplicity, we determine the solutions of $u_1 = u_1 + 2u_2$ and $u_2 = 0u_1 + u_2$. We obtain $u_2 = 0$ and u_1 can take any value. The eigen space of λ contains all vectors $\mathbf{u}$ with $u_2 = 0$; the dimension of this space is 1. ☐

Proposition B.1.4. *Any similar matrices $A, B \in \mathbb{C}^{n\times n}$ (i.e., such that $B = U^{-1}AU$ for some non-singular $U \in C^{n\times n}$) have the same characteristic polynomial. Consequently, A and B have the same eigenvalues with the same algebraic multiplicities. We also say that B is the representation of A in the basis determined by the columns of U; the change of basis matrix from this basis to the canonical basis is exactly U.*

Proof.

$$\begin{aligned} det(xI - B) &= det(xI - U^{-1}AU) \\ &= det(xU^{-1}U - U^{-1}AU) \\ &= det(U^{-1}(xI - A)U) \\ &= det(U^{-1}) \cdot det(xI - A) \cdot det(U) \\ &= det(xI - A) \end{aligned}$$

☐

Proposition B.1.5. *Any two similar matrices $A, B \in \mathbb{C}^{n\times n}$ have the same geometric multiplicity for any common eigenvalue λ.*

Proof. Take any fixed λ and $\mathbf{u}_B \in \mathbb{C}^n$ such that $B\mathbf{u}_B = \lambda\mathbf{u}_B$. We can write $\lambda\mathbf{u}_B = U^{-1}AU\mathbf{u}_B$, which is equivalent to $\lambda U\mathbf{u}_B = AU\mathbf{u}_B$. As such, A has eigenvector $\mathbf{u}_A = U\mathbf{u}_B$ with eigenvalue λ. There is a bijection between the eigenvectors of B and the eigenvectors of A, given by above transformation $\mathbf{u}_B \to \mathbf{u}_A = U\mathbf{u}_b$. To check the bijectivity, notice the injectivity follows from $U\mathbf{u}_B = U\mathbf{u}'_B \implies U(\mathbf{u}_B - \mathbf{u}'_B) = \mathbf{0} \implies \mathbf{u}_B = \mathbf{u}'_B$ based on the non-singularity of U. The surjectivity follows from $\forall \mathbf{u}_A \in \mathbb{C}^n$, $\exists \mathbf{u}_B = U^{-1}\mathbf{u}_A \in \mathbb{C}^n$ such that $\mathbf{u}_A = U\mathbf{u}_B$. This bijection shows that the geometric multiplicities of λ are the same. ☐

B.2 The eigenvalue decomposition

Proposition B.2.1. *(Eigendecomposition) Any symmetric matrix $A \in \mathbb{R}^{n\times n}$ can be decomposed as follows:*

$$A = U\Lambda U^\top \tag{B.2.1}$$

$$= \sum_{i=1}^{n} \lambda_i \mathbf{u}_{\times,i}\mathbf{u}_{\times,i}^\top, \tag{B.2.2}$$

where $\Lambda = \boldsymbol{diag}(\lambda_1, \lambda_2, \ldots \lambda_n)$ contains the eigenvalues of A and U contains columns $\mathbf{u}_{\times,1}$, $\mathbf{u}_{\times,2}, \ldots \mathbf{u}_{\times,n}$ that represent the orthonormal unit eigenvectors of A.

Proof. We provide two proofs (Appendix B.2.1 and Appendix B.2.2) for showing the key fact (B.2.1):

- The first proof was actually given in Section 1.1 and we here only repeat it in greater detail and a bit generalized. It relies on the equality between the geometric and algebraic multiplicities of each eigenvalue.
- Use the Schur decomposition of complex matrices and above (B.2.1) becomes a simple re-writing of (B.2.9) in Prop. B.2.5. Reading this second proof is useful to develop a general culture (on complex or asymmetric matrices).

Both above proofs also show that U contains n orthonormal unit eigenvectors of A and that this way we have $U^\top = U^{-1}$. After showing (B.2.1) with either proof, (B.2.2) simply follows from developing $A = U\Lambda U^\top$:

$$A = [\mathbf{u}_{\times,1}, \mathbf{u}_{\times,2}, \ldots, \mathbf{u}_{\times,n}]\Lambda \begin{bmatrix} \mathbf{u}_{\times,1}^\top \\ \mathbf{u}_{\times,2}^\top \\ \vdots \\ \mathbf{u}_{\times,n}^\top \end{bmatrix} = [\mathbf{u}_{\times,1}, \mathbf{u}_{\times,2}, \ldots, \mathbf{u}_{\times,n}] \begin{bmatrix} \lambda_1\mathbf{u}_{\times,1}^\top \\ \lambda_2\mathbf{u}_{\times,2}^\top \\ \vdots \\ \lambda_n\mathbf{u}_{\times,n}^\top \end{bmatrix} = \sum_{i=1}^{n} \lambda_i \mathbf{u}_{\times,i}\mathbf{u}_{\times,i}^\top$$

Writing eigenvector $\mathbf{v}_i = \mathbf{u}_{\times,i} \in \mathbb{R}^n$, the above formula can be expressed in a very compact form:

$$A = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top, \tag{B.2.3}$$

where $\mathbf{v}_1,\ \mathbf{v}_2, \ldots \mathbf{v}_n$ are orthonormal eigenvectors (pairwise orthogonal and of unitary norm). □

B.2.1 Proof using the equality of the geometric and algebraic multiplicities

Proposition B.2.2. *Consider (possibly non-symmetric) matrix $A \in \mathbb{R}^{n\times n}$ such that each eigenvalue λ_i has the same geometric and algebraic multiplicity k_i. This means root λ_i arises k_i times in the characteristic polynomial and there are k_i linearly independent eigenvectors of λ_i. Matrix A has the following eigen-decomposition (diagonalization):*

$$A = U \boldsymbol{diag}(\lambda_1,\ \lambda_2, \ldots \lambda_n) U^{-1} \tag{B.2.4}$$

where U contains n orthonormal eigenvectors of A and $\boldsymbol{diag}(\lambda_1,\ \lambda_2, \ldots \lambda_n)$ is a diagonal matrix with the (possibly complex) eigenvalues on the diagonal.
If A is symmetric, we have $\lambda_1,\ \lambda_2, \ldots \lambda_n \in \mathbb{R}$ (Prop. B.1.2). We also prove $U^{-1} = U^\top$, and so, (B.2.4) becomes:

$$A = U \boldsymbol{diag}(\lambda_1,\ \lambda_2, \ldots \lambda_n) U^\top \tag{B.2.5}$$

Proof. Since each eigenvalue λ_i (with $i \in [1..n]$) has the same geometric and algebraic multiplicity k_i, we can say each repetition of λ_i as root of the characteristic polynomial can be associated to a different eigenvector. The eigenspace of λ_i has dimension k_i and we can surely find an orthonormal basis of this space to represent the k_i eigenvectors associated to the k_i repetitions of root λ_i. The sums of the algebraic multiplicities is n because the characteristic polynomial has degree n, and so, we can fill an $n \times n$ matrix U with the eigenvectors of $\lambda_1,\ \lambda_2, \ldots \lambda_n$.

Since by multiplying A with any eigenvector (column) of U we obtain the eigenvector multiplied by its eigenvalue, we can write $U\mathtt{diag}(\lambda_1,\ \lambda_2, \ldots \lambda_n) = AU$. By multiplying at right with U^{-1}, we obtain (B.2.4).

If A is symmetric, we can show $U^{-1} = U^\top$ using the fact that the eigenvectors are orthonormal. We already said above that the eigenvectors corresponding to the same eigenvalue can be chosen to be orthonormal. Any eigenvectors $\mathbf{v}_i$ and $\mathbf{v}_j$ corresponding to distinct eigenvalues $\lambda_i \neq \lambda_j$ need to be orthogonal. To see this, notice $\mathbf{v}_i^\top A \mathbf{v}_j = \mathbf{v}_i^\top \lambda_j \mathbf{v}_j = \lambda_j \mathbf{v}_i^\top \mathbf{v}_j$ and also $\mathbf{v}_i^\top A \mathbf{v}_j = \lambda_i \mathbf{v}_i^\top \mathbf{v}_j$ based on $\mathbf{v}_i^\top A = (\mathbf{v}_i^\top A)^{\top\top} = (A^\top \mathbf{v}_i)^\top = (A\mathbf{v}_i)^\top = \lambda_i \mathbf{v}_i^\top$. This leads to $\lambda_j \mathbf{v}_i^\top \mathbf{v}_j = \lambda_i \mathbf{v}_i^\top \mathbf{v}_j$, and, using $\lambda_i \neq \lambda_j$, we obtain $\mathbf{v}_i^\top \mathbf{v}_j = 0\ \forall i, j \in [1..n],\ i \neq j$. This shows $U^\top U = I_n$, and so, $U^{-1} = U^\top$. By simply replacing U^{-1} with $U^\top$ in (B.2.4), we obtain (B.2.5) as needed. □

Proposition B.2.3. *Any eigenvalue λ of real symmetric matrix A has the same algebraic and geometric multiplicity.*

Proof. We assume the characteristic polynomial of A has a factor $(x - \lambda)^s$, *i.e.*, the algebraic multiplicity of λ is s. Let t be the dimension of the eigenspace of λ, *i.e.*, the geometric multiplicity is t. We will show $s = t$. Consider t orthonormal eigenvectors of λ, generated by taking an orthonormal basis of the eigenspace of λ. We construct an unitary matrix V by putting these eigenvectors on the first t columns and by filling the remaining $n - t$ columns with other vectors that generate an orthonormal basis of $\mathbb{R}^n$ together with $\mathbf{v}_1,\ \mathbf{v}_2, \ldots \mathbf{v}_t$. We thus have $V^{-1} = V^\top$ and we can compute:

$$V^\top A V = V^\top [\lambda \mathbf{v}_1\ \lambda \mathbf{v}_2\ \ldots \lambda \mathbf{v}_t\ B] = \begin{bmatrix} \lambda I_t & C \\ \mathbf{0}_{n-t,t} & D \end{bmatrix},$$

where the zeros on the first t columns are due to the fact that each $\mathbf{v}_i$ with $i \in [1..t]$ is orthogonal to all the other column vectors of V. By transposing above formula, we obtain the same matrix because $\left(V^\top A V\right)^\top = V^\top A^\top V = V^\top A V$. This means that $\begin{bmatrix} \lambda I_t & C \\ \mathbf{0}_{n-t,t} & D \end{bmatrix} = \begin{bmatrix} \lambda I_t & C \\ \mathbf{0}_{n-t,t} & D \end{bmatrix}^\top$, and so, C must be zero and D must be diagonal. We can write:

$$V^{-1} A V = V^\top A V = \begin{bmatrix} \lambda I_t & \mathbf{0}_{t,n-t} \\ \mathbf{0}_{n-t,t} & D \end{bmatrix}.$$

This means that matrices A and $\begin{bmatrix} \lambda I_t & \mathbf{0}_{t,n-t} \\ \mathbf{0}_{n-t,t} & D \end{bmatrix}$ are similar, and so, they have the same characteristic polynomial by virtue of Prop. B.1.4. The characteristic polynomial of the second matrix is $(x-\lambda)^t \det(xI-D)$. This directly shows that we can *not* have $t > s$. This would be equivalent to the existence of a term $(x-\lambda)^t$ with $t > s$ in the characteristic polynomial of A, which is impossible because the algebraic multiplicity of λ is s.

We now prove by contradiction that $t = s$. Assuming the contrary, the only remaining case is $t < s$. This means that $\det(xI - D)$ has to contain a term $(x-\lambda)^{s-t}$ because $\det(xI - A)$ contains a term $(x-\lambda)^s$. This way, λ is an eigenvalue of D for which there exists at least an eigenvector $\mathbf{d} \in \mathbb{R}^{n-t}$. We will show that λ has a geometric multiplicity higher than t in $\begin{bmatrix} \lambda I_t & \mathbf{0}_{t,n-t} \\ \mathbf{0}_{n-t,t} & D \end{bmatrix}$, which contradicts Prop. B.1.5, *i.e.*, similar matrices must have the same eigenvalue multiplicities for a common eigenvalue λ. It is not hard to check that λ has at least the following $t+1$ eigenvectors in $\begin{bmatrix} \lambda I_t & \mathbf{0}_{t,n-t} \\ \mathbf{0}_{n-t,t} & D \end{bmatrix}$:

$$\begin{bmatrix} 1 \\ \left.\begin{matrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{matrix}\right\} \text{n-1} \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ \left.\begin{matrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{matrix}\right\} \text{n-2} \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 1 \\ \left.\begin{matrix} 0 \\ 0 \\ \vdots \\ 0 \end{matrix}\right\} \text{n-3} \end{bmatrix} \quad \dots \quad \begin{bmatrix} \left.\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}\right\} \text{t-1} \\ 1 \\ \left.\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}\right\} \text{n-t} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \left.\begin{matrix} 0 \\ \vdots \\ 0 \\ 0 \end{matrix}\right\} \text{t} \\ \\ \mathbf{d} \\ \\ \end{bmatrix}.$$

The last vector is an eigenvector, because it is enough to check that the top first t positions of the product with $\begin{bmatrix} \lambda I_t & \mathbf{0}_{t,n-t} \\ \mathbf{0}_{n-t,t} & D \end{bmatrix}$ are $\lambda I_t \mathbf{0}_t + \mathbf{0}_{t,n-t}\mathbf{d} = \mathbf{0}_t$ and the bottom $n-t$ positions are $\mathbf{0}_{n-t,t}\mathbf{0}_t + D\mathbf{d} = \lambda\mathbf{d}$ since $\mathbf{d}$ is an eigenvector of D. We obtained that λ has geometric multiplicity at least $t+1$ in $\begin{bmatrix} I_t & \mathbf{0}_{t,n-t} \\ \mathbf{0}_{n-t,t} & D \end{bmatrix}$ and t in A, contradicting Prop. B.1.5 as stated above. This ensures the only possible case is $t = s$. $\square$

B.2.2 Proof using the Schur triangulation of general complex square matrices

Theorem B.2.4. *(Schur decomposition) Given any $A \in \mathbb{C}^{n\times n}$, there exists unitary matrix $U \in \mathbb{C}^{n\times n}$ (i.e., such that the conjugate transpose satisfies $U^*U = I$) for which:*

$$U^*AU = T, \tag{B.2.6}$$

where T is an upper triangular matrix. The diagonal elements of T are the eigenvalues of A. The number of times eigenvalue λ appears on the diagonal of T is the algebraic multiplicity of λ in A.

Proof. We proceed by induction. For $n = 1$, the theorem is clearly true. Assume it holds for $n-1$ and we prove it also holds for n.

Consider an eigenvalue λ and an eigenvector $\mathbf{u}$ (they exists as proved by Proposition B.1.1). Without loss of generality we assume $\mathbf{u}$ is unitary, *i.e.*, $\mathbf{u}^*\mathbf{u} = 1$ (from a non-unitary eigenvector we can easily obtain an unitary one by dividing each of its term by a non-negative real number).

We construct an unitary matrix $\overline{U} \in \mathbb{C}^{n\times n}$ with $\mathbf{u}$ on the first column. We write $\overline{U} = [\mathbf{u}, V]$ with $V \in C^{n\times(n-1)}$. The construction of V can be done column by column as follows. The first unitary column $\mathbf{v}_{\times,1}$ of V is chosen by solving $\mathbf{u}^*\mathbf{v}_{\times,1} = 0$ in variables $v_{11}, v_{21}, \dots v_{n1}$. The existence of a solution for this equation can follow from the more general Proposition A.2.2; and once a solution is found, we easily make $\mathbf{v}_{\times,1}$ unitary by dividing all terms by the initial norm $|\mathbf{v}_{\times,1}|$. The $i^{\textbf{th}}$ unitary column $\mathbf{v}_{\times,i}$ (for $i \leq n-1$) is chosen by solving the following system in variables $v_{1i}, v_{2i}, \dots v_{ni}$: (a) $\mathbf{u}^*\mathbf{v}_{\times,i} = 0$ and (b) $\mathbf{v}^*_{\times,j}\mathbf{v}_{\times,i} = 0$ for all $j \leq i-1$. There are at most $n-1$ equations for n variables, and so, a solution has to exist (for the skeptical, the coefficients of the $n-1$ equations can be put in a $n\times n$ matrix filled with zeros on the last row, so as to obtain a null determinant and apply Proposition A.2.2). We obtain:

$$\overline{U}^*A\overline{U} = \begin{bmatrix} \mathbf{u}^* \\ V^* \end{bmatrix} A[\mathbf{u}, V] = \begin{bmatrix} \mathbf{u}^* \\ V^* \end{bmatrix} [\lambda\mathbf{u}, AV] = \begin{bmatrix} \lambda & \mathbf{u}^*AV \\ \lambda V^*\mathbf{u} & V^*AV \end{bmatrix} = \begin{bmatrix} \lambda & \mathbf{u}^*AV \\ \mathbf{0}_{n-1} & V^*AV \end{bmatrix}$$

As a side remark, remark that if A is hermitian ($A = A^*$) or real symmetric, the top-right term is also zero: $\mathbf{u}^*AV = (A\mathbf{u})^*V = \lambda\mathbf{u}^*V = \mathbf{0}^\top_{n-1}$. In fact, it is possible to particularize the above line of proof to directly prove that real symmetric matrices are diagonalizable (and produce an eigen-decomposition), but it may be useful now to stay a bit more on the general case.

Let us use more compact notations for above equation:

$$\overline{U}^* A \overline{U} = \begin{bmatrix} \lambda & \mathbf{b}^\top \\ \mathbf{0}_{n-1} & B \end{bmatrix} \tag{B.2.7}$$

Using the induction hypothesis, there is some unitary matrix $U_B \in \mathbb{C}^{(n-1)\times(n-1)}$ such that $U_B^* B U_B = T_B$ is upper triangular. We construct unitary matrix $\widehat{U} = \left[\begin{smallmatrix} 1 & \mathbf{0}_{n-1}^\top \\ \mathbf{0}_{n-1} & U_B \end{smallmatrix}\right]$ and we obtain:

$$\begin{aligned} \widehat{U}^* \overline{U}^* A \overline{U} \widehat{U} &= \begin{bmatrix} 1 & \mathbf{0}_{n-1}^\top \\ \mathbf{0}_{n-1} & U_B^* \end{bmatrix} \begin{bmatrix} \lambda & \mathbf{b}^\top \\ \mathbf{0}_{n-1} & B \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}_{n-1}^\top \\ \mathbf{0}_{n-1} & U_B \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}_{n-1}^\top \\ \mathbf{0}_{n-1} & U_B^* \end{bmatrix} \begin{bmatrix} \lambda & \mathbf{b}^\top U_B \\ \mathbf{0}_{n-1} & B U_B \end{bmatrix} \\ &= \begin{bmatrix} \lambda & \mathbf{b}^\top U_B \\ \mathbf{0}_{n-1} & U_B^* B U_B \end{bmatrix} = \begin{bmatrix} \lambda & \mathbf{b}^\top U_B \\ \mathbf{0}_{n-1} & T_B \end{bmatrix} = T \end{aligned} \tag{B.2.8}$$

Since T_B is upper triangular by the induction hypothesis, T is also upper triangular. Noting $U = \overline{U}\widehat{U}$, we obtain (B.2.6). It is not hard to check that U is unitary: $U^* U = \widehat{U}^* \overline{U}^* \overline{U} \widehat{U} = \widehat{U}^* \widehat{U} = I$.

We still need to prove that the diagonal elements of T are the eigenvalues of A. Applying Proposition B.1.4, similar matrices A and T have the same characteristic polynomial. But the characteristic polynomial of upper triangular matrix T is $(x - t_{11})(x - t_{22}) \dots (x - t_{nn})$. The diagonal elements of T coincide thus with the roots of the characteristic polynomial of A and T. □

U does not necessarily contain the eigenvectors of A as columns, even if the construction starts from an eigenvector of A. Consider, for instance, the matrix $\left[\begin{smallmatrix} 1 & -1 \\ 1 & -1 \end{smallmatrix}\right]$. The characteristic polynomial is x^2 so that root $\lambda = 0$ has algebraic multiplicity 2. However, the matrix has only one eigenvector $\mathbf{u} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$, because it has rank 1; the geometric multiplicity of eigenvalue $\lambda = 0$ is 1. The Schur decomposition constructs U (equal to $\overline{U}$ because $U_B = 1$ in the proof) by putting $\mathbf{u}$ on the first column and by filling the other columns of U so as to make U orthonormal. We obtain $U = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}$ and $U^* = U^\top = U$. The decomposition is $\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}$. U does not contain an eigenvector on the second column.

Proposition B.2.5. *If $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix, the Schur decomposition computes a diagonalization of A. There exists an unitary (orthonormal) matrix $U \in \mathbb{R}^{n\times n}$ such that*

$$T = U^* A U = U^\top A U \tag{B.2.9}$$

is a diagonal matrix with the eigenvalues of A on the diagonal. The columns of U are the eigenvectors of A which needs to be real (Prop. B.1.2), and $U \in \mathbb{R}^{n\times n}$ so that $U^* = U^\top$.

Proof. Theorem B.2.4 shows that there is a decomposition (B.2.9) that generates an upper triangular matrix T. We apply the conjugate transpose on both sides of (B.2.9):

$$T^* = (U^* A U)^* = U^* A^* (U^*)^* = U^* A U = T$$

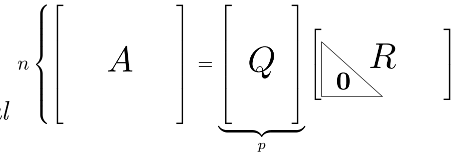
The equality $T^* = T$ can only hold if T is a diagonal matrix with real elements on the diagonal. Recalling that Theorem B.2.4 shows that the diagonal elements of T are the eigenvalues of A (each taken with its algebraic multiplicity), we can write $T = \mathtt{diag}(\lambda_1, \lambda_2, \dots \lambda_n)$. Multiplying (B.2.9) by U at left (and using that U is unitary, *i.e.*, $UU^* = U^* U = I_n$, as stated by Theorem B.2.4), we obtain: $U\mathtt{diag}(\lambda_1, \lambda_2, \dots \lambda_n) = AU$. The column i (with $i \in [1..n]$) on both sides can be written: $\lambda_i u_{\times,i} = A u_{\times,i}$. This shows that column i of U is an eigenvector associated to λ_i. Using Proposition B.1.2, this column $u_{\times,i}$ contains only real elements, and so, U is real. □

B.3 The QR decomposition of real matrices

Proposition B.3.1. *Any matrix $A \in \mathbb{R}^{n\times n}$ can be decomposed as*

$$A = QR,$$

where $Q \in \mathbb{R}^{n\times p}$ is an orthonormal matrix (its columns are orthogonal unit vectors meaning $Q^\top Q = I_p$) and $R \in \mathbb{R}^{p\times n}$ is upper triangular.



Proof. In preamble, let us first (try to) capture the "spirit" of the factorization we want to prove. Q has to contain a number of p unit vectors that actually represent an orthonormal basis of the space spanned by them. Then, $A = QR$ means that each column $\mathbf{a}_k$ of A can be written as a linear combination of the first k elements of this basis (the first k columns of Q). It is easy to obtain this linear combination when $\mathbf{a}_k$ belongs to the space spanned by these first k elements of the basis. We need to construct an increasingly larger orthonormal basis that first covers only $\mathbf{a}_1$, then $\mathbf{a}_1$ and $\mathbf{a}_2$, then $\mathbf{a}_1,\ \mathbf{a}_2$ and $\mathbf{a}_3$, etc. Geometrically, we can first simply take the unit vector $\mathbf{e}_1$ collinear with $\mathbf{a}_1$; at step 2, we take a vector $\mathbf{e}_2$ perpendicular to $\mathbf{a}_1$ or $\mathbf{e}_1$ such that $\mathbf{e}_1$ and $\mathbf{e}_2$ determine the same 2D hyperplane as $\mathbf{a}_1$ and $\mathbf{a}_2$; at the third step we take a vector $\mathbf{e}_3$ perpendicular on this 2D hyperplane so as to determine the same 3D (sub-)space as $\mathbf{a}_1,\ \mathbf{a}_2,\ \mathbf{a}_3$. This is the goal of the Gram-Schmidt process presented next.

We now formally present the Gram-Schmidt process on the columns of $A = [\mathbf{a}_1,\ \mathbf{a}_2, \dots \mathbf{a}_n]$. This process takes the column vectors $\mathbf{a}_1,\ \mathbf{a}_2, \dots \mathbf{a}_n$ and generates an orthogonal set of vectors that spans the same subspace as $\mathbf{a}_1,\ \mathbf{a}_2, \dots \mathbf{a}_n$. Let us define the *normalized vector* $\mathbf{e} \in \mathbb{R}^n$ of $\mathbf{u} \in \mathbb{R}_n$ as $\mathbf{e} = \frac{\mathbf{u}}{|\mathbf{u}|}$ if $\mathbf{u} \neq \mathbf{0}$, or $\mathbf{e} = \mathbf{0}$ if $\mathbf{u} = \mathbf{0}$ (degenerate normalized vector). Notice a non-degenerate normalized vector $\mathbf{e}$ has norm 1 because $\mathbf{e} \bullet \mathbf{e} = \frac{\mathbf{u}\bullet\mathbf{u}}{|\mathbf{u}|^2} = 1$. We define the projection operator `proj` by setting:

$$\mathtt{proj}_{\mathbf{e}}\mathbf{a} = (\mathbf{e} \bullet \mathbf{a})\mathbf{e} = \mathtt{proj}_{\mathbf{u}}\mathbf{a} = \left(\frac{\mathbf{u}}{|\mathbf{u}|} \bullet \mathbf{a}\right)\frac{\mathbf{u}}{|\mathbf{u}|} = \frac{\mathbf{u}\bullet\mathbf{a}}{\mathbf{u}\bullet\mathbf{u}}\mathbf{u}, \tag{B.3.1}$$

where $\mathbf{e}$ is the normalized vector of $\mathbf{u}$. If $\mathbf{u} = \mathbf{e} = 0$, we define $\mathtt{proj}_{\mathbf{0}}\mathbf{a} = \mathbf{0}$. This represents the projection of $\mathbf{a}$ on $\mathbf{e}$ or $\mathbf{u}$. The Gram-Schmidt process constructs the following sequence:

$$\begin{aligned}
\mathbf{u}_1 &= \mathbf{a}_1\\
\mathbf{u}_2 &= \mathbf{a}_2 - \mathtt{proj}_{\mathbf{u}_1}\mathbf{a}_2\\
\mathbf{u}_3 &= \mathbf{a}_3 - \mathtt{proj}_{\mathbf{u}_1}\mathbf{a}_3 - \mathtt{proj}_{\mathbf{u}_2}\mathbf{a}_3\\
&\ \,\vdots\\
\mathbf{u}_n &= \mathbf{a}_n - \sum_{i=1}^{n-1}\mathtt{proj}_{\mathbf{u}_i}\mathbf{a}_n
\end{aligned}$$

Geometrically, this construction work as follows: to compute $\mathbf{u}_k$, it projects $\mathbf{a}_k$ orthogonally onto the subspace U generated by $\mathbf{u}_1,\ \mathbf{u}_2, \dots \mathbf{u}_{k-1}$, which is the same as the subspace generated by $\mathbf{a}_1,\ \mathbf{a}_2, \dots \mathbf{a}_{k-1}$. The vector $\mathbf{u}_k$ is then defined to be the difference between $\mathbf{a}_k$ and this projection, guaranteed to be orthogonal to all of the vectors in the subspace U.

We now show formally that $\mathbf{u}_k$ is orthogonal to all $\mathbf{u}_1,\ \mathbf{u}_2, \dots \mathbf{u}_{k-1}$. Assume by induction $\mathbf{u}_{k'} \bullet \mathbf{u}_j = 0\ \ \forall j, k' \in [1..n],\ j < k' \leq k-1$. To show $\mathbf{u}_k \bullet \mathbf{u}_j = 0\ \forall j < k$, we first observe $\mathbf{u}_k \bullet \mathbf{u}_j = 0$ if $\mathbf{u}_j = \mathbf{0}$. If $\mathbf{u}_j \neq \mathbf{0}$, we calculate:

$$\begin{aligned}
\mathbf{u}_k \bullet \mathbf{u}_j &= \left(\mathbf{a}_k - \sum_{i=1}^{k-1}\mathtt{proj}_{\mathbf{u}_i}\mathbf{a}_k\right) \bullet \mathbf{u}_j\\
&= \mathbf{a}_k \bullet \mathbf{u}_j - \sum_{\substack{i\in[1..k-1]\\ \mathbf{u}_i\neq\mathbf{0}}}\frac{\mathbf{u}_i\bullet\mathbf{a}_k}{\mathbf{u}_i\bullet\mathbf{u}_i}\mathbf{u}_i \bullet \mathbf{u}_j && \text{(we developped } \mathtt{proj}_{\mathbf{u}_i}\mathbf{a}_k \text{ using (B.3.1) for } \mathbf{u}_i \neq \mathbf{0})\\
&= \mathbf{a}_k \bullet \mathbf{u}_j - \frac{\mathbf{u}_j\bullet\mathbf{a}_k}{\mathbf{u}_j\bullet\mathbf{u}_j}\mathbf{u}_j \bullet \mathbf{u}_j && \text{(we used } \mathbf{u}_i \bullet \mathbf{u}_j = 0 \text{ for } i \neq j \text{ and } i, j \in [1..k-1])\\
&= \mathbf{a}_k \bullet \mathbf{u}_j - \mathbf{u}_j \bullet \mathbf{a}_k = 0
\end{aligned}$$

Notice we can have $\mathbf{u}_k = \mathbf{0}$ for certain $k \in [1..n]$. However, we can also calculate

$$\mathbf{u}_k \bullet \mathbf{u}_k = \left(\mathbf{a}_k - \sum_{i=1}^{k-1} \mathtt{proj}_{\mathbf{u}_i}\mathbf{a}_k\right) \bullet \mathbf{u}_k = \mathbf{a}_k \bullet \mathbf{u}_k - \sum_{\substack{i\in[1..k-1]\\ \mathbf{u}_i\neq\mathbf{0}}} \frac{\mathbf{u}_i \bullet \mathbf{a}_k}{\mathbf{u}_i \bullet \mathbf{u}_i}\mathbf{u}_i \bullet \mathbf{u}_k = \mathbf{a}_k \bullet \mathbf{u}_k,$$

meaning that $\mathbf{u}_k = \mathtt{proj}_{\mathbf{u}_k}\mathbf{u}_k = \mathtt{proj}_{\mathbf{u}_k}\mathbf{a}_k$. This allows us to write the equation at step k of above Gram-Schmidt process as:

$$\begin{aligned}\mathbf{a}_k = \mathbf{u}_k + \sum_{i=1}^{k-1}\mathtt{proj}_{\mathbf{u}_i}\mathbf{a}_k &= \mathtt{proj}_{\mathbf{u}_k} a_k + \sum_{i=1}^{k-1}\mathtt{proj}_{\mathbf{u}_i}\mathbf{a}_k \\ &= \sum_{i=1}^{k}\mathtt{proj}_{\mathbf{u}_i}\mathbf{a}_k = \sum_{i=1}^{k}\mathtt{proj}_{\mathbf{e}_i}\mathbf{a}_k && \text{(see (B.3.1))} \\ &= \sum_{i=1}^{k}(\mathbf{e}_i \bullet \mathbf{a}_k)\mathbf{e}_i,\end{aligned}$$

where $\mathbf{e}_i$ is the normalized vector of $\mathbf{u}_i$. Since this holds for any $k \in [1..n]$, we can write it in matrix form:

$$A = [\mathbf{a}_1\ \mathbf{a}_2 \dots \mathbf{a}_n] = [\mathbf{e}_1\ \mathbf{e}_2 \dots \mathbf{e}_n]\begin{bmatrix}\mathbf{e}_1 \bullet \mathbf{a}_1 & \mathbf{e}_1 \bullet \mathbf{a}_2 & \mathbf{e}_1 \bullet \mathbf{a}_3 & \dots & \mathbf{e}_1 \bullet \mathbf{a}_n \\ 0 & \mathbf{e}_2 \bullet \mathbf{a}_2 & \mathbf{e}_2 \bullet \mathbf{a}_3 & \dots & \mathbf{e}_2 \bullet \mathbf{a}_n \\ 0 & 0 & \mathbf{e}_3 \bullet \mathbf{a}_3 & \dots & \mathbf{e}_3 \bullet \mathbf{a}_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \mathbf{e}_n \bullet \mathbf{a}_n\end{bmatrix} = \overline{Q}\ \overline{R}$$

Notice that some columns of $\overline{Q} = [\mathbf{e}_1\ \mathbf{e}_2 \dots \mathbf{e}_n]$ can be zero. We can transform the $n \times n$ matrix $\overline{Q}$ into a $n \times p$ matrix Q by removing $n-p$ zero columns. At the same time, we need to remove the corresponding rows of $\overline{R}$ and we obtain matrix $R \in \mathbb{R}^{p\times n}$ that remains upper triangular, *i.e.*, the elements below the diagonal are zero, leading to $A = QR$ as needed. □

B.4 An SDP matrix has a unique SDP square root factor

We introduced the square root decompositions in Section 1.6.1.1. Let's examine the square root matrices $K \in \mathbb{R}^{n\times n}$ such that $KK = A \succeq \mathbf{0}$. Using the eigendecomposition (1.1.1), we have $KK = A = U\Lambda U^\top$ where Λ is diagonal and U satisfies $UU^\top = I_n$. We can thus write $KK = KUU^\top K = U\Lambda U^\top \implies U^\top KUU^\top KU = \Lambda$. Using notation $D = U^\top KU$, we obtain $DD = \Lambda$. In other words, K need to have the form $K = UDU^\top$ for some $D \in \mathbb{R}^{n\times n}$ such that $DD = \Lambda$.

Proposition B.4.1. *Given SDP matrix $A \in \mathbb{R}^{n\times n}$, there exists a unique matrix $K \in \mathbb{R}^{n\times n}$ such that $KK = A$ and $K \succeq \mathbf{0}$. This K is called the principal square root of A.*

Proof. As described in above paragraph, any K such that $KK = A$ satisfies $K = UDU^\top$; the columns of U are the unitary orthonormal eigenvectors of A, so that $UU^\top = I$. We notice $\mathbf{x}^\top D\mathbf{x} = \mathbf{x}^\top U^\top UDU^\top U\mathbf{x} = (U\mathbf{x})^\top K(U\mathbf{x}) \geq 0\ \forall \mathbf{x} \in \mathbb{R}^n$, and so, D is also SDP. To prove the unicity of SDP matrix K, it is enough to show there is a unique SDP matrix $D \in \mathbb{R}^{n\times n}$ such that $DD = \mathtt{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$, where $\lambda_1, \lambda_2, \dots, \lambda_n$ are the eigenvalues of A. We apply the eigendecomposition on SDP matrix D: we obtain $D = VEV^\top$, where $E = \mathtt{diag}(e_1, e_2, \dots e_n) \geq \mathbf{0}$ and $VV^\top = V^\top V = I_n$.

Using $DD = \Lambda$, we have $VEV^\top VEV^\top = \Lambda$, and so, $VE^2V^\top = \Lambda$, or $VE^2 = \Lambda V$. Taking any $i, j \in [1..n]$, we have $(VE^2)_{ij} = (\Lambda V)_{ij}$, equivalent to $V_{ij}e_j^2 = \lambda_i V_{ij}$. We observe the following property:

$$e_j^2 \neq \lambda_i \implies V_{ij} = 0 \tag{*}$$

To prove that $K = UDU^\top$ is unique, it is enough to show that D is unique. We will exactly determine the value of D by showing $D = VEV^\top = \sqrt{\Lambda}$, where $\sqrt{\Lambda} = \mathtt{diag}\left(\sqrt{\lambda_1},\ \sqrt{\lambda_2}, \dots, \sqrt{\lambda_n}\right)$. For this, it is enough to prove $VE = \sqrt{\Lambda}V$, which is somehow a consequence of above $VE^2 = \Lambda V$. More exactly, we need to show that we find the same value at position (i, j) of both sides of $VE = \sqrt{\Lambda}V$ for any $i, j \in [1..n]$. For this, we have to show $V_{ij}e_j = \sqrt{\lambda_i}V_{ij}$. If $e_j = \sqrt{\lambda_i}$, this is clearly true. If $e_j \neq \sqrt{\lambda_i}$, we have $e_j^2 \neq \lambda_i$ (recall both e_j and λ_i are non-negative eigenvalues of SDP matrices), and so, $(*)$ states that $V_{ij} = 0$, showing $V_{ij}e_j = \sqrt{\lambda_i}V_{ij} = 0$. □

C Useful related facts

We provide two classical results, a proposition related to the completely positive cone, finishing with an example of a convex function with an asymmetric non-SDP Hessian.

C.1 Optimality conditions for linearly-constrained quadratic programs

Proposition C.1.1. *Consider the following linearly-constrained quadratic optimization problem, based on (not necessarily SDP) symmetric matrix $Q \in \mathbb{R}^{n\times n}$, full-rank matrix $A \in \mathbb{R}^{p\times n}$ with $p \leq n$ and $\mathbf{b} \in \mathbb{R}^p$.*

$$(QP_=)\begin{cases}\min \quad p(\mathbf{x}) = \mathbf{x}^\top Q\mathbf{x} + \mathbf{c}^\top\mathbf{x} & \text{(C.1.1a)}\\ \quad s.t.\ A\mathbf{x} = \mathbf{b} & \text{(C.1.1b)}\\ \qquad\ \mathbf{x}\in\mathbb{R}^n & \text{(C.1.1c)}\end{cases}$$

The solution $\mathbf{x}^*$ is optimal if and only if the following conditions are satisfied:

$$2Q\mathbf{x}^* + \mathbf{c} = A^\top\boldsymbol{\mu} \text{ for some } \boldsymbol{\mu}\in\mathbb{R}^p \tag{C.1.2a}$$

$$\mathbf{z}^\top Q\mathbf{z} \geq 0\ \forall \mathbf{z} \in \mathtt{null}(A) = \{\mathbf{z}\in\mathbb{R}^n :\ A\mathbf{z} = \mathbf{0}\} \tag{C.1.2b}$$

It can be a useful exercise to give three proofs using different techniques.

Proof 1. We solve by force the system $A\mathbf{x} = \mathbf{b}$. Since A has full rank p, the null space $\mathtt{null}(A)$ of A has dimension $n-p$ by virtue of the rank-nullity Theorem A.1.3. Let $\mathbf{z}_1,\ \mathbf{z}_2,\ldots\mathbf{z}_{n-p}$ be a basis of the null space. Any solution $\mathbf{x}$ of the system considered above has the form $\mathbf{x} = \mathbf{v}+\sum_{i=1}^{n-p}\mathbf{z}_i y_i$ where $A\mathbf{v} = \mathbf{b}$. Constructing matrix $Z = [\mathbf{z}_1\ \mathbf{z}_2\ \ldots\mathbf{z}_{n-p}]$, we can write the solutions of the system under the form $\mathbf{v} + Z\mathbf{y}$ with $\mathbf{y}\in\mathbb{R}^{n-p}$.

We can write $(QP_=)$ from (C.1.1a)-(C.1.1c) under the form:

$$(QP_=)\begin{cases}\min \quad (\mathbf{v}+Z\mathbf{y})^\top Q(\mathbf{v}+Z\mathbf{y}) + \mathbf{c}^\top(\mathbf{v}+Z\mathbf{y}) & \text{(C.1.3a)}\\ \quad s.t.\ \mathbf{y}\in\mathbb{R}^{n-p} & \text{(C.1.3b)}\end{cases}$$

We re-write above (C.1.3a) as:

$$\mathbf{y}^\top Z^\top QZ\mathbf{y} + 2\mathbf{v}^\top QZ\mathbf{y} + \mathbf{v}^\top Q\mathbf{v} + \mathbf{c}^\top\mathbf{v} + \mathbf{c}^\top Z\mathbf{y},$$

where we used $\mathbf{v}^\top QZ\mathbf{y} = (\mathbf{v}^\top QZ\mathbf{y})^\top = \mathbf{y}Z^\top Q\mathbf{v}$. Using Prop. 3.4.1, this unconstrained quadratic program is bounded from below if and only if it is convex and the gradient vanishes at the optimal solution $\mathbf{y}^*$. The necessary and sufficient conditions for the optimality of $\mathbf{x}^* = \mathbf{v} + \mathbf{y}^*$ are:

(a) $Z^\top QZ \succeq \mathbf{0}$, which is equivalent to the fact that Q is positive over $\mathtt{null}(A)$, *i.e.*, we obtain the second order condition (C.1.2b).

(b) The optimal solution $\mathbf{y}^*$ needs to cancel the (column vector) gradient $2Z^\top QZ\mathbf{y}^* + 2Z^\top Q\mathbf{v} + Z^\top\mathbf{c} = Z^\top(2Q(\mathbf{v}+Z\mathbf{y}^*)+\mathbf{c}) = Z^\top(2Q\mathbf{x}^*+\mathbf{c})$. Since the rows of $Z^\top$ are a transposed basis for $\mathtt{null}(A)$, the above gradient can only cancel if $2Q\mathbf{x}^*+\mathbf{c}$ belongs to the transposed row image of A, *i.e.*, there is some $\boldsymbol{\mu}\in\mathbb{R}^p$ such that $2Q\mathbf{x}^*+\mathbf{c} = A^\top\boldsymbol{\mu}$, which is exactly the first order condition (C.1.2a).

Finally, there is a degenerate case $p = n$ in which above Z and $\mathbf{y}$ have dimension 0 and they disappear. In this case, (C.1.2a) holds because $A^\top\boldsymbol{\mu}$ with $\boldsymbol{\mu}\in\mathbb{R}^n$ can cover the whole space $\mathbb{R}^n$ since A is a square full rank matrix. The second order condition (C.1.2b) holds because it reduces to nothing using $\mathtt{null}(A) = \{\mathbf{0}\}$. The system has thus only one feasible solution $\mathbf{x}^* = A^{-1}\mathbf{b}$ that satisfies both conditions above. □

Proof 2.
$\Longrightarrow$ The first order condition (C.1.2a) follows by applying the method of Lagrange multipliers, obtaining a particular case of the KKT conditions.[37] However, in our case the argument reduces to the following. The

[37] See also my document "Trying to demystify the Karush-Kuhn-Tucker conditions", available on-line at http://cedric.cnam.fr/~porumbed/papers/kkt.pdf.

gradient $\nabla p(\mathbf{x}^*)$ is perpendicular in $\mathbf{x}^*$ to a surface (level set) on which p takes the constant value $p(\mathbf{x}^*)$. Why does not p increase or decrease by moving backward or forward from $\mathbf{x}^*$ along some $\mathbf{z} \in \texttt{null}(A)$? Because the gradient in $\mathbf{x}^*$ is perpendicular to $\mathbf{z}$. Indeed, using the chain rule, the derivative in $t = 0$ of $f(t) = p(\mathbf{x}^* + t\mathbf{z})$ is equal to $f'(0) = \nabla p(\mathbf{x}^*) \bullet \mathbf{z} = 0$. If this were not zero, one could move backward or forward from $\mathbf{x}^*$ along $\mathbf{z}$ by some $\varepsilon > 0$ to decrease p. We obtain that the gradient $\nabla p(\mathbf{x}^*)$ is perpendicular to $\texttt{null}(A)$, and so, it belongs to the transposed row image of A, *i.e.*, $\nabla p(\mathbf{x}^*) = A^\top \boldsymbol{\mu}$ for some $\boldsymbol{\mu} \in \mathbb{R}^p$, which is exactly (C.1.2a).

We prove $\mathbf{z}^\top Q\mathbf{z} \geq 0 \forall \mathbf{z} \in \texttt{null}(A)$ by assuming the opposite: $\exists \mathbf{z} \in \texttt{null}(A)$ such that $\mathbf{z}^\top Q\mathbf{z} = -\varepsilon < 0$. All $\mathbf{x} + t\mathbf{z}$ are feasible since they satisfy $A(\mathbf{x} + t\mathbf{z}) = A\mathbf{x} = \mathbf{b}$. The function $f(t) = p(\mathbf{x} + t\mathbf{z})$ has degree 2 and the quadratic factor is $\mathbf{z}^\top Q\mathbf{z}t^2 = -\varepsilon t^2$. This is a concave function that goes to $-\infty$ in both directions. This means $\mathbf{x}^*$ is not minimal. We obtained a contradiction from $\mathbf{z}^\top Q\mathbf{z} = -\varepsilon < 0$. The second order condition (C.1.2b) needs to hold.

$\Longleftarrow$ We suppose both conditions (C.1.2a)-(C.1.2b) are satisfied by some $\mathbf{x}^*$ such that $A\mathbf{x}^* = \mathbf{b}$. We will prove that $p(\mathbf{x}) \geq p(\mathbf{x}^*)$ for any feasible $\mathbf{x} = \mathbf{x}^* + \mathbf{z}$, with $\mathbf{z} \in \texttt{null}(A)$. Consider the function $f(t) = p(\mathbf{x}^* + t\mathbf{z})$. Using the chain rule and (C.1.2a), we obtain $f'(0) = \nabla p(\mathbf{x}^*) \bullet \mathbf{z} = \nabla p(\mathbf{x}^*)^\top \mathbf{z}$. Replacing the gradient with the right hand side of (C.1.2a), this is further equal to $\boldsymbol{\mu}^\top A\mathbf{z}$ which is equal to 0 because $\mathbf{z} \in \texttt{null}(A)$. We thus obtained $f'(0) = 0$. We will show f is convex. The only quadratic factor in t of $p(\mathbf{x}^* + t\mathbf{z})$ is $\mathbf{z}^\top Q\mathbf{z}t^2$ and its second derivative is $2\mathbf{z}^\top Q\mathbf{z}$ which is non-negative by virtue of the second order condition (C.1.2b). This means that f is convex and reaches its minimum at $t = 0$, *i.e.*, $p(\mathbf{x}^*) \leq p(\mathbf{x})$ for all feasible $\mathbf{x}$. □

Proof 3.

$\Longrightarrow$ We show the second order condition (C.1.2b) using the second proof above. We can prove the first order condition (C.1.2a) using convexification results. We first apply Prop. C.2.2 that states there exists $\boldsymbol{\lambda} \in \mathbb{R}^{n\times p}$ such that $p_{\boldsymbol{\lambda}}(\mathbf{x}) = p(\mathbf{x}) + \sum \lambda_{ji}(x_j A_i \mathbf{x} - x_j \mathbf{b}_i)$ is convex, where A_i is the row i of A. Notice $p_{\boldsymbol{\lambda}}(\mathbf{x}) = p(\mathbf{x})$ whenever $A\mathbf{x} = \mathbf{b}$. We can write the following program

$$p(\mathbf{x}^*) = \min_{\mathbf{x}} \{p_{\boldsymbol{\lambda}}(\mathbf{x}) : \; A\mathbf{x} = \mathbf{b}\} \tag{C.1.4}$$

as an SDP program because it is convex. Using the results from Section 6.2.2, this SDP program takes the form of $\left(PL^{X}(QP_{=})\right)$ from (6.2.5a)-(6.2.5b) with $\boldsymbol{\mu} = \mathbf{0}$ and $\overline{\boldsymbol{\mu}} = \mathbf{0}$. We can write $p(\mathbf{x}^*) = OPT(\text{C.1.4}) = \left(PL^{\mathbf{x}}(QP_{=})\right) = \left(PL^{X}(QP_{=})\right)$ because the Hessian of all these programs except the first one is SDP; this means that the first essential hierarchy (6.2.13a) collapses (like in Remark 6.3.7). As such, the total Lagrangian of (C.1.4) reaches $p(\mathbf{x}^*)$. The total Lagrangian for optimal Lagrangian multipliers $\boldsymbol{\beta} \in \mathbb{R}^p$ is

$$\min_{\mathbf{x}\in\mathbb{R}^n} p_{\boldsymbol{\lambda}}(\mathbf{x}) + \boldsymbol{\beta}^\top A\mathbf{x} - \boldsymbol{\beta}^\top \mathbf{b}. \tag{C.1.5}$$

Evaluating this total Lagrangian in $\mathbf{x}^*$, we obtain the value $p(\mathbf{x}^*)$. Since the total Lagrangian reaches $p(\mathbf{x}^*)$, the minimum of (C.1.5) needs to be $p(\mathbf{x}^*)$. Using Prop. 3.4.1, this can only be the case if the gradient in $\mathbf{x}^*$ of (C.1.5) is zero.

Let us calculate the gradient of $p(\mathbf{x}) + \sum \lambda_{ji}(x_j A_i \mathbf{x} - x_j \mathbf{b}_i) + \boldsymbol{\beta}^\top A\mathbf{x}$ in $\mathbf{x}^*$. The gradient of each term $x_j A_i \mathbf{x} - x_j \mathbf{b}_i$ is computed as follows. For any $k \neq j$, the partial derivative on x_k is $x_j A_{ik}$. The partial derivative on x_j is $2A_{ij}x_j + \sum_{k\neq j} A_{ik}x_k - b_i = x_j A_{ij} + A_i\mathbf{x} - b_i$. The second term vanishes in feasible $\mathbf{x}^*$. We obtain that the gradient of $x_j A_i \mathbf{x} - x_j \mathbf{b}_i$ in $\mathbf{x}^*$ is $x_j^* A_i^\top$. We obtain $\nabla \left(p(\mathbf{x}) + \sum \lambda_{ji}(x_j A_i \mathbf{x} - x_j \mathbf{b}_i) + \boldsymbol{\beta}^\top A\mathbf{x}\right)_{\mathbf{x}^*} = \nabla p(\mathbf{x}^*) + \sum \lambda_{ji} x_j^* A_i^\top + A^\top \boldsymbol{\beta}$. The last two terms belong to the transposed row image of A. Since the gradient needs to be zero, the first term $\nabla p(\mathbf{x}^*)$ has to belong to the transposed row image of A as well, *i.e.*, $\nabla p(\mathbf{x}^*) = A^\top \boldsymbol{\mu}$ for some $\boldsymbol{\mu} \in \mathbb{R}^p$, which is exactly (C.1.2a).

$\Longleftarrow$ Using the fact that Q is non-negative over $\texttt{null}(A)$, we can use a relatively similar (reversed) argument as in the above "$\Longrightarrow$" proof. First, as above, we apply Prop. C.2.2 that states there exists $\boldsymbol{\lambda} \in \mathbb{R}^{n\times p}$ such that $p_{\boldsymbol{\lambda}}(\mathbf{x}) = p(\mathbf{x}) + \sum \lambda_{ji}(x_j A_i \mathbf{x} - x_j \mathbf{b}_i)$ is convex. Notice $p_{\boldsymbol{\lambda}}(\mathbf{x}) = p(\mathbf{x})$ whenever $A\mathbf{x} = \mathbf{b}$. The total Lagrangian of $\min \{p_{\boldsymbol{\lambda}}(\mathbf{x}) : \; A\mathbf{x} = \mathbf{b}\}$ is $p_{\boldsymbol{\lambda}}(\mathbf{x}) + \boldsymbol{\beta}^\top A\mathbf{x} - \boldsymbol{\beta}^\top \mathbf{b}$ which is a convex polynomial for any value of $\boldsymbol{\beta} \in \mathbb{R}^p$. The value of this total Lagrangian in $\mathbf{x}^*$ is $p(\mathbf{x}^*)$. It is enough to show that the gradient of the total Lagrangian in $\mathbf{x}^*$ is zero for an appropriate $\boldsymbol{\beta} \in \mathbb{R}^p$. Using the calculations from the above "$\Longrightarrow$" proof, the gradient in $\mathbf{x}^*$ is $\nabla p(\mathbf{x}^*) + \sum \lambda_{ji} x_j^* A_i^\top + A^\top \boldsymbol{\beta} = A^\top \boldsymbol{\mu} + \sum \lambda_{ji} x_j^* A_i^\top + A^\top \boldsymbol{\beta}$. The first two terms belong to the transposed row image of A, and so, they can be canceled by an appropriate $\boldsymbol{\beta} \in \mathbb{R}^p$. We obtain that the total Lagrangian reaches the value $p(\mathbf{x}^*)$, and so, $\mathbf{x}^*$ needs to be a minimizer of p (the Lagrangian is always less than or equal to the constrained optimum of p or $p_{\boldsymbol{\lambda}}$). $\square$

C.2 More insight and detail into the convexifications from Section 6

C.2.1 Constraints that can be used to convexify any matrix non-negative over $\texttt{null}(A)$

We consider a full-rank matrix $A \in \mathbb{R}^{p\times n}$ associated to linear constraint $A\mathbf{x} = \mathbf{b}$. Based on these constraints, one can generate various redundant quadratic constraints that are surely satisfied when $A\mathbf{x} = \mathbf{b}$, see Section 6.3.3.2 for examples. We showed in Section 6.3.3.1 that a partial Lagrangian subject to $A\mathbf{x} = \mathbf{b}$ can reach the optimum value only when the Lagrangian multipliers construct a matrix Q that is non-negative over the null space $\texttt{null}(A)$ of A (see (A.1.2) for the null space definition); if this does not happen, the partial Lagrangian converges to $-\infty$. We here discuss convexifications that can make Q non-negative over the whole $\mathbb{R}^n$ (*i.e.*, SDP) using the Lagrangian multipliers associated to the redundant quadratic constraints.

The first paragraph of Section 2.3. from "Partial Lagrangian relaxation for General Quadratic Programming" (see Footnote 33, p. 71) states the following result as already known in the literature. If Q is strictly positive over $\texttt{null}(A)$, there exists $V \in \mathbb{R}^{n\times n}$ such that $Q + A^\top V A \succeq \mathbf{0}$. We prove below a generalization of this result. This proof is not taken from existing work and we think it is original; it uses the Bolzano-Weierstrass theorem C.4.9 (any bounded sequence contains a convergent sub-sequence).

Proposition C.2.1. *Consider any $Q \in \mathbb{R}^{n\times n}$ strictly positive over $\texttt{null}(A)$, i.e., $\mathbf{u}^\top Q\mathbf{u} > 0\ \forall \mathbf{u} \in \texttt{null}(A) - \{\mathbf{0}\}$. If $B \succeq \mathbf{0}$ satisfies*

$$B \bullet \mathbf{u}\mathbf{u}^\top = 0 \iff \mathbf{u} \in \texttt{null}(A), \tag{C.2.1}$$

then there exists $\lambda > 0$ such that $Q + \lambda B \succeq \mathbf{0}$. In other words, we can convexify Q using the Lagrangian multiplier of a (redundant) quadratic constraint with quadratic factor B. The matrix B can be for instance $B = A^\top S A$ for any $S \succ \mathbf{0}$, generalizing the redundant constraint from Example 6.3.4.

Proof. We define a set of particular interest: $\widetilde{X} = \{\mathbf{x} \in \mathbb{R}^n :\ |\mathbf{x}| = 1,\ Q \bullet \mathbf{x}\mathbf{x}^\top < 0\}$. We need to make the elements $\mathbf{x}$ of this set verify $(Q + \lambda B) \bullet \mathbf{x}\mathbf{x}^\top \geq 0$. Consider the function $f : \widetilde{X} \to \mathbb{R}$ defined by

$$f(\mathbf{x}) = \frac{B \bullet \mathbf{x}\mathbf{x}^\top}{|Q \bullet \mathbf{x}\mathbf{x}^\top|}$$

Based on (C.2.1), $B \succeq \mathbf{0}$ is strictly positive over all $\mathbf{x}$ outside the null space of A. Since $\widetilde{X} \subseteq \mathbb{R}^n - \texttt{null}(A)$, we easily obtain $f(\mathbf{x}) > 0$ $\mathbf{x} \in \widetilde{X}$.

We will prove by contradiction that $\inf f(\mathbf{x}) > 0$. Assuming the contrary (*i.e.*, $\inf f(\mathbf{x}) = 0$), there exists a sequence $(\mathbf{x}_i)$ with $\mathbf{x}_i \in \widetilde{X}$ such that $\lim\limits_{i\to\infty} f(\mathbf{x}_i) = 0$. Using the Bolzano-Weierstrass Theorem C.4.9, there exists a sub-sequence $\{\mathbf{x}_{n_i}\}$ such that $\lim\limits_{i\to\infty} \mathbf{x}_{n_i} = \widetilde{\mathbf{x}}$. We will show we can obtain a contradiction for each of the following three cases that cover $\widetilde{\mathbf{x}} \in \mathbb{R}^n$:

(i) $\widetilde{\mathbf{x}} \in \widetilde{X}$

(ii) $\widetilde{\mathbf{x}} \in \texttt{null}(A)$

(iii) $\widetilde{\mathbf{x}} \in \mathbb{R}^n - \texttt{null}(A) - \widetilde{X}$

For case (i), it is enough to notice that the sequence $f(\mathbf{x}_{n_i})$ can be arbitrarily close to $f(\widetilde{\mathbf{x}}) > 0$, which contradicts that $f(\mathbf{x}_i)$ converges to zero.

For case (ii), we obtain by hypothesis that $Q \bullet \widetilde{\mathbf{x}}\,\widetilde{\mathbf{x}}^\top > 0$. This leads to $Q \bullet \widetilde{\mathbf{x}_\varepsilon}\,\widetilde{\mathbf{x}_\varepsilon}^\top > 0$ for any $\widetilde{\mathbf{x}_\varepsilon}$ such that $|\widetilde{\mathbf{x}_\varepsilon} - \widetilde{\mathbf{x}}| < \varepsilon$ for a sufficiently small ε. But this would mean that f is undefined in a sufficiently small ball around $\widetilde{\mathbf{x}}$, and so, we can *not* have $\lim\limits_{i\to\infty} \mathbf{x}_{n_i} = \widetilde{\mathbf{x}}$ with $\mathbf{x}_{n_i} \in \widetilde{X}\ \forall i \in \mathbb{N}$, contradiction.

For case (iii), we can use (C.2.1) and $\widetilde{\mathbf{x}} \notin \texttt{null}(A)$ to obtain $B \bullet \widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^\top = z > 0$. When $i \to \infty$, the value $B \bullet \mathbf{x}_{n_i}\mathbf{x}_{n_i}^\top$ can become arbitrarily close to z and $Q \bullet \mathbf{x}_{n_i}\mathbf{x}_{n_i}^\top$ can become arbitrarily close to 0. This latter fact ($\lim_{i\to\infty} Q \bullet \mathbf{x}_{n_i}\mathbf{x}_{n_i}^\top = 0$) follows from $Q \bullet \mathbf{x}_{n_i}\mathbf{x}_{n_i} < 0\ \forall i \in \mathbb{N}$ (because $\mathbf{x}_{n_i} \in \widetilde{X}$) and $Q \bullet \widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^\top \geq 0$ (because $\widetilde{\mathbf{x}} \notin \widetilde{X}$). Combining the above convergence properties of $B \bullet \mathbf{x}_{n_i}\mathbf{x}_{n_i}^\top$ and $Q \bullet \mathbf{x}_{n_i}\mathbf{x}_{n_i}^\top$, we obtain that $f(\mathbf{x}_{n_i}) = \frac{B \bullet \mathbf{x}_{n_i}\mathbf{x}_{n_i}^\top}{-Q \bullet \mathbf{x}_{n_i}\mathbf{x}_{n_i}^\top}$ can become arbitrarily large, contradicting that $f(\mathbf{x}_i)$ converges to zero.

This means there exists a (possibly large) $\lambda > 0$ such that $f(\mathbf{x}) = \frac{B \bullet \mathbf{x}\mathbf{x}^\top}{|Q \bullet \mathbf{x}\mathbf{x}^\top|} > \frac{1}{\lambda}\ \forall \mathbf{x} \in \widetilde{X}$. We obtain $\lambda B \bullet \mathbf{x}\mathbf{x}^\top + Q \bullet \mathbf{x}\mathbf{x}^\top > 0$ for all $\mathbf{x}$ such that $Q \bullet \mathbf{x}\mathbf{x}^\top < 0$ (we developed the $\widetilde{X}$ definition, forgetting that all its elements are unitary). For the remaining cases, *i.e.*, for all $\mathbf{x} \in \mathbb{R}^n$ such that $Q \bullet \mathbf{x}\mathbf{x}^\top \geq 0$, we also obtain $(Q + \lambda B) \bullet \mathbf{x}\mathbf{x}^\top \geq 0$ simply because $B \succeq \mathbf{0}$ and $\lambda > 0$. Combining both cases above, we get $(Q + \lambda B) \bullet \mathbf{x}\mathbf{x}^\top \geq 0\ \forall \mathbf{x} \in \mathbb{R}^n$. □

The next proofs are modified versions of the proofs from (Section 2.3 of) "Partial Lagrangian relaxation for General Quadratic Programming" by Alain Faye and Frédéric Roupin (see Footnote 33, p. 71).

Proposition C.2.2. *We are given a full rank matrix $A \in \mathbb{R}^{p\times n}$ associated to constraints $A\mathbf{x} = \mathbf{b}$. Consider any $Q \in \mathbb{R}^{n\times n}$ non-negative over `null(A)`, i.e., $\mathbf{u}^\top Q\mathbf{u} \geq 0\ \forall \mathbf{u} \in$ `null(A)` $- \{\mathbf{0}\}$. There exists a linear combination of the redundant constraints $x_j A_i \mathbf{x} - x_j \mathbf{b}_i = 0$ (where A_i is the row i of A, with $j \in [1..n]$ and $i \in [1..p]$) from Example 6.3.5 that can be added to $Q \bullet \mathbf{x}\mathbf{x}^\top$ to transform Q into an SDP matrix. Equivalently, if $\overline{A}_{j\swarrow i}$ is the $n \times n$ matrix with only a non-zero row j that contains A_i, then there always exist $\lambda_{ji} \in \mathbb{R}$ (for all $j \in [1..n]$ and $i \in [1..p]$) such that $Q + \sum_{j,i} \lambda_{ji} \left(\overline{A}^\top_{j\swarrow i} + \overline{A}_{j\swarrow i}\right) \succeq \mathbf{0}$.*

Proof. We will show in Prop. C.2.3 there exists $\overline{W} \in \mathbb{R}^{n\times p}$ such that $Q + A^\top \overline{W}^\top + \overline{W}A \succeq \mathbf{0}$. We can write $\overline{W} = \sum W_{ji} E_{ji}$ where the E_{ji}'s represent the canonical base indexed by $j \in [1..n]$ and $i \in [1..p]$, *i.e.*, $E_{ji} \in \mathbb{R}^{n\times p}$ has a value of one at position (j, i) and only zeros at all other positions. We have

$$\begin{aligned} A^\top \overline{W}^\top + \overline{W}A &= \sum_{j,i} W_{ji} \left(A^\top E^\top_{ji} + E_{ji} A\right) \\ &= \sum_{j,i} W_{ji} \left(\overline{A}^\top_{j\swarrow i} + \overline{A}_{j\swarrow i}\right), \end{aligned}$$

which concludes the proof, with the values $\lambda_{ji} = W_{ji}$.

□

Proposition C.2.3. *We are given a full rank matrix $A \in \mathbb{R}^{p\times n}$ associated to constraints $A\mathbf{x} = \mathbf{b}$. Consider any $Q \in \mathbb{R}^{n\times n}$ non-negative over `null`(A), i.e., $\mathbf{u}^\top Q\mathbf{u} \geq 0\ \forall \mathbf{u} \in$ `null(A)` $- \{\mathbf{0}\}$. There exists $\overline{W} \in \mathbb{R}^{n\times p}$ such that $Q + A^\top \overline{W}^\top + \overline{W}A \succeq \mathbf{0}$.*

Proof. We recall the QR decomposition from Prop. B.3.1 and the Gram-Schmidt orthogonalization process described in the proof of this Prop. B.3.1. We apply this process up to the last column p of $A^\top$, so as to factorize $A^\top = U^\top R$, where U has the size of A, $UU^\top = I_p$ and $R \in \mathbb{R}^{p\times p}$ is upper triangular. One can see this as a QR decomposition restricted to the first p columns, *i.e.*, it can be extended to a full QR decomposition by adding $n - p$ zero columns to $A^\top$ and R. However, we can equivalently factorize $A = R^\top U$. Notice that the (first $i \leq p$) rows of U span the same subspace as the (first $i \leq p$) rows of A. Using Prop. A.1.4, we have $p = rank(A) \leq rank(R^\top), rank(U)$, and so, U and $R^\top$ are full rank and $R^\top$ is invertible. If we find $W \in \mathbb{R}^{n\times p}$ such that $S + U^\top W^\top + WU \succeq \mathbf{0}$, we can use equation below and finish the proof:

$$WU = W\left(R^\top\right)^{-1} R^\top U = \underbrace{W\left(R^\top\right)^{-1}}_{\overline{W}} A \text{ and } U^\top W^\top = A^\top \overline{W}^\top. \tag{C.2.2}$$

It is enough to show there exists $W \in \mathbb{R}^{n\times p}$ such that $Q + U^\top W^\top + WU \succeq \mathbf{0}$, where the (first $i \leq p$) rows of U are an orthonormal basis spanning the same subspace as the (first $i \leq p$) rows of A. Let B be a matrix whose columns are an orthonormal basis of `null`(A) = `null`(U). Using the rank-nullity Theorem A.1.3, we have $B \in \mathbb{R}^{n,n-p}$. The fact that Q is non-negative over `null(A)` is equivalent to $B^\top QB \succeq \mathbf{0}$; this follows from

$$\mathbf{x}^\top B^\top QB\mathbf{x} \geq 0\ \forall \mathbf{x} \in \mathbb{R}^{n-p} \iff (B\mathbf{x})^\top Q(B\mathbf{x}) \geq 0\ \forall \mathbf{x} \in \mathbb{R}^{n-p} \iff \mathbf{y}^\top Q\mathbf{y} \geq 0\ \forall \mathbf{y} \in \texttt{null}(A) \tag{C.2.3}$$

Let L_i be the sub-space spanned by $\mathbf{u}_{i+1},\ \mathbf{u}_{i+2}, \ldots \mathbf{u}_p$ and (the columns of) B, where $\mathbf{u}_i$ is the row i of U written as a column vector. This sub-space has dimension $n - i$ and is perpendicular on $\mathbf{u}_1,\ \mathbf{u}_2, \ldots \mathbf{u}_i$ (recall $UB = 0$). We can also write:

$$L_i = \texttt{img}\left(\mathbf{u}_{i+1},\ \mathbf{u}_{i+2}, \ldots \mathbf{u}_p, B\right) = \left\{\mathbf{y} \in \mathbb{R}^n :\ \mathbf{u}_j \mathbf{y} = 0\ \forall j \in [1..i]\right\}, \tag{C.2.4}$$

where $\mathtt{img}(...)$ is the sub-space generated by the column vectors of the matrices given as arguments. In particular, we have $L_0 = \mathbb{R}^n$ and $L_p = \mathtt{img}(B) = \mathtt{null}(A)$. We will construct a matrix $Q_i \in \mathbb{R}^{n\times n}$ that is non-negative on L_i, by induction on i from $i = p$ (with $Q_p = Q$) down to $i = 0$ (when $i = 0$, Q_0 non-negative over $\mathbb{R}^n$ is equivalent to $Q_0 \succeq \mathbf{0}$). Using an argument as the one from (C.2.3), a matrix $Q_i \in \mathbb{R}^{n\times n}$ is non-negative over L_i if and only if $M^i = [\mathbf{u}_{i+1},\ \mathbf{u}_{i+2}, \dots \mathbf{u}_p, B]^\top Q_i [\mathbf{u}_{i+1},\ \mathbf{u}_{i+2}, \dots \mathbf{u}_p, B] \succeq \mathbf{0}$. We will prove we can use Q_i associated to $M^i \succeq \mathbf{0}$ to construct Q_{i-1} such that $M^{i-1} \succeq \mathbf{0}$. We will thus iteratively construct $Q_i = Q + \sum\limits_{j=i+1}^{p} \mathbf{u}_j\mathbf{w}_j^\top + \mathbf{w}_j\mathbf{u}_j^\top$ with i from p down to 0. At the last iteration we will obtain:

$$Q_0 = Q + \sum_{j=1}^{p} \mathbf{u}_j\mathbf{w}_j^\top + \mathbf{w}_j\mathbf{u}_j^\top = Q + U^\top W^\top + WU \tag{C.2.5}$$

as needed, where $W = [\mathbf{w}_1\ \mathbf{w}_2\ \mathbf{w}_3 \dots \mathbf{w}_p]$. We will use several times the following property:

$$\mathbf{u}_a^\top \left(\mathbf{u}_j\mathbf{w}_j^\top + \mathbf{w}_j\mathbf{u}_j^\top\right)\mathbf{u}_b = 0\ \forall \mathbf{u}_a, \mathbf{u}_b \in \left\{\mathbf{u}_1,\ \mathbf{u}_2, \dots, \mathbf{u}_{j-1}, \mathbf{u}_{j+1},\ \mathbf{u}_{j+2}, \dots, \mathbf{u}_p,\ B\right\}, \tag{C.2.6}$$

where one can read B as an enumeration of column vectors (slightly abusing notations). This simply follows from $\mathbf{u}_a^\top\mathbf{u}_j = \mathbf{u}_j^\top\mathbf{u}_b = 0$ with $\mathbf{u}_a$ and $\mathbf{u}_b$ from the above set.

Now we present the induction step. We can assume Q_i is already constructed and we need to determine $Q_{i-1} = Q_i + \mathbf{u}_i\mathbf{w}_i^\top + \mathbf{w}_i\mathbf{u}_i^\top$. More exactly, the goal is to find some $\mathbf{w}_i \in \mathbb{R}^n$ such that $M^{i-1} = [\mathbf{u}_i,\ \mathbf{u}_{i+1}, \dots \mathbf{u}_p, B]^\top Q_{i-1} [\mathbf{u}_i,\ \mathbf{u}_{i+1}, \dots \mathbf{u}_p, B] \succeq \mathbf{0}$, as argued above. At each transition $i \to i-1$, we can use the induction hypothesis that Q_i is associated to $M^i \succeq \mathbf{0}$ (recall this is surely true for $Q_p = Q$ and $M^p = B^\top QB$ by hypothesis). Let us develop the formula of M^{i-1} that we will construct to be SDP.

$$\begin{aligned}
M^{i-1} &= \begin{bmatrix} \mathbf{u}_i^\top \\ \mathbf{u}_{i+1}^\top \\ \vdots \\ \mathbf{u}_p^\top \\ B^\top \end{bmatrix} \left(Q + \sum_{j=i}^{p} \mathbf{u}_j\mathbf{w}_j^\top + \mathbf{w}_j\mathbf{u}_j^\top\right)[\mathbf{u}_i,\ \mathbf{u}_{i+1}, \dots \mathbf{u}_p, B] = \left[\begin{array}{c|c} M_{1,1}^{i-1} & M_{1,[2..n-i+1]}^{i-1} \\ \hline M_{[2..n-i+1],1}^{i-1} & M_{[2..n-i+1],[2..n-i+1]}^{i-1} \end{array}\right] \\
&= \left[\begin{array}{c|c} \mathbf{u}_i^\top \left(Q + \sum\limits_{j=i}^{p} \mathbf{u}_j\mathbf{w}_j^\top + \mathbf{w}_j\mathbf{u}_j^\top\right)\mathbf{u}_i & \mathbf{u}_i^\top \left(Q + \sum\limits_{j=i}^{p} \mathbf{u}_j\mathbf{w}_j^\top + \mathbf{w}_j\mathbf{u}_j^\top\right)[\mathbf{u}_{i+1}, \dots \mathbf{u}_p, B] \\ \hline \begin{bmatrix} \mathbf{u}_{i+1}^\top \\ \vdots \\ \mathbf{u}_p^\top \\ B^\top \end{bmatrix} \left(Q + \sum\limits_{j=i}^{p} \mathbf{u}_j\mathbf{w}_j^\top + \mathbf{w}_j\mathbf{u}_j^\top\right)\mathbf{u}_i & M^i \end{array}\right],
\end{aligned} \tag{C.2.7}$$

where we used (C.2.6) with $j = i$ and $\mathbf{u}_a, \mathbf{u}_b \in \{\mathbf{u}_{i+1},\ \mathbf{u}_{i+2}, \dots, \mathbf{u}_p,\ B\}$ to remain with the M^i term in the bottom-right cell. Let us develop the *first line* of M^{i-1}:

- the first position is $M_{1,1}^{i-1} = \mathbf{u}_i^\top \left(Q + \sum\limits_{j=i}^{p} \mathbf{u}_j\mathbf{w}_j^\top + \mathbf{w}_j\mathbf{u}_j^\top\right)\mathbf{u}_i = \mathbf{u}_i^\top \left(Q + \mathbf{u}_i\mathbf{w}_i^\top + \mathbf{w}_i\mathbf{u}_i^\top\right)\mathbf{u}_i$ by virtue of (C.2.6). We can further develop this into $M_{1,1}^{i-1} = \mathbf{u}_i^\top Q\mathbf{u}_i + \mathbf{w}_i^\top\mathbf{u}_i + \mathbf{u}_i^\top\mathbf{w}_i$.
- on the remaining $n - i$ positions, we have:

$$\begin{aligned}
M_{1,[2..n-i+1]}^{i-1} &= \mathbf{u}_i^\top \left(Q + \sum_{j=i}^{p} \mathbf{u}_j\mathbf{w}_j^\top + \mathbf{w}_j\mathbf{u}_j^\top\right)[\mathbf{u}_{i+1}, \dots \mathbf{u}_p, B] \\
&= \mathbf{u}_i^\top Q [\mathbf{u}_{i+1}, \dots \mathbf{u}_p, B] + \mathbf{u}_i^\top\mathbf{u}_i\mathbf{w}_i^\top [\mathbf{u}_{i+1}, \dots \mathbf{u}_p, B] && \text{(we used (C.2.6))} \\
&\quad + \mathbf{u}_i^\top \left(\sum_{j=i}^{p} \mathbf{w}_j\mathbf{u}_j^\top\right)[\mathbf{u}_{i+1}, \dots \mathbf{u}_p, B] \\
&= \mathbf{u}_i^\top Q [\mathbf{u}_{i+1}, \dots \mathbf{u}_p, B] + \mathbf{w}_i^\top [\mathbf{u}_{i+1}, \dots \mathbf{u}_p, B] \\
&\quad + \mathbf{u}_i^\top [\mathbf{w}_{i+1}, \dots \mathbf{w}_p, \mathbf{0}] && \text{(we used (C.2.6))}
\end{aligned}$$

Using the orthonormality properties of $\mathbf{u}_{i+1}, \dots \mathbf{u}_p$ and B, the last term can be written

$$\begin{aligned}\mathbf{u}_i^\top [\mathbf{w}_{i+1}, \dots \mathbf{w}_p, \mathbf{0}] &= \mathbf{u}_i^\top \underbrace{[\mathbf{w}_{i+1}, \dots \mathbf{w}_p, \mathbf{0}] [\mathbf{u}_{i+1}, \dots \mathbf{u}_p, B]^\top}_{P_i} [\mathbf{u}_{i+1}, \dots \mathbf{u}_p, B] \\ &= \mathbf{u}_i^\top P_i [\mathbf{u}_{i+1}, \dots \mathbf{u}_p, B],\end{aligned}$$

where $P_i \in \mathbb{R}^{n \times n}$ does not depend on the vector $\mathbf{w}_i$ we need to determine. We will need the following:

$$P_i \mathbf{u}_i = \mathbf{0}_{n\times 1} \text{ and } \mathbf{u}_i^\top P_i^\top = \mathbf{0}_{1\times n} \tag{C.2.8}$$

Finally, by simplifying above formulas, the first row of M^{i-1} can be written:

$$M_1^{i-1} = \left[\mathbf{u}_i^\top Q \mathbf{u}_i + 2\mathbf{u}_i^\top \mathbf{w}_i,\ \left(\mathbf{u}_i^\top Q + \mathbf{w}_i^\top + \mathbf{u}_i^\top P_i\right) [\mathbf{u}_{i+1}, \dots \mathbf{u}_p, B]\right]$$

We need to determine $\mathbf{w}_i$ such that $M^{i-1} \succeq \mathbf{0}$, given that we can rely on the induction hypothesis $M^i \succeq \mathbf{0}$. The simplest way to construct an SDP matrix M^{i-1} is to generate only zeros on the first row M_1^{i-1}. First, we would like $\mathbf{w}_i^\top$ to cancel the terms $\mathbf{u}_i^\top Q + \mathbf{u}_i^\top P_i$ in above M_1^{i-1} formula. As such, $\mathbf{w}_i$ integrates a first term $-\left(\mathbf{u}_i^\top Q + \mathbf{u}_i^\top P_i\right)^\top$. A second term of $\mathbf{w}_i$ is $z\mathbf{u}_i$; this second term does not change the canceled positions of the first row (see point (b) below), but it can make $M_{1,1}^{i-1} \geq 0$ for a sufficiently large z. Thus, we set $\mathbf{w}_i = -\left(Q + P_i^\top\right)\mathbf{u}_i + z\mathbf{u}_i$, where the value of z will be determined at point (a) below. However, this $\mathbf{w}_i$ vector leads to the following values on the first row of M^{i-1}.

(a) $M_{1,1}^{i-1} = \mathbf{u}_i^\top Q \mathbf{u}_i + 2\mathbf{u}_i^\top \left(-\left(Q + P_i^\top\right)\mathbf{u}_i + z\mathbf{u}_i\right) = \mathbf{u}_i^\top Q \mathbf{u}_i - 2\mathbf{u}_i^\top Q \mathbf{u}_i - 2\mathbf{u}_i^\top P_i^\top \mathbf{u}_i + 2\mathbf{u}_i^\top z \mathbf{u}_i = -\mathbf{u}_i^\top Q \mathbf{u}_i + 2z\mathbf{u}_i^\top \mathbf{u}_i = -\mathbf{u}_i^\top Q \mathbf{u}_i + 2z$, where we used (C.2.8) to cancel the $P_i^\top$ term. We set $z = \frac{1}{2}\left(\mathbf{u}_i^\top Q \mathbf{u}_i\right)$ to make $M_{1,1}^{i-1} = 0$, but larger values can also be chosen.

(b) $M_{1,[2..n-i+1]}^{i-1} = \left(\mathbf{u}_i^\top Q + \left(-\left(Q + P_i^\top\right)\mathbf{u}_i + z\mathbf{u}_i\right)^\top + \mathbf{u}_i^\top P_i\right) [\mathbf{u}_{i+1}, \dots \mathbf{u}_p, B] = z\mathbf{u}_i^\top [\mathbf{u}_{i+1}, \dots \mathbf{u}_p, B] = \mathbf{0}_{1\times(n-i)}$.

The resulting first row of M^{i-1} is filled with zeros, and so needs to be the first column by symmetry. Recalling (C.2.7), we have $M^{i-1} = \left[\begin{smallmatrix} 0 & \mathbf{0} \\ \mathbf{0} & M^i \end{smallmatrix}\right]$. Using the induction hypothesis $M^i \succeq \mathbf{0}$, we obtain $M^{i-1} \succeq \mathbf{0}$, which finishes the induction step. At the last iteration, we obtain $M^0 \succeq \mathbf{0}$, which is enough to guarantee $Q_0 \succeq \mathbf{0}$ using arguments discussed above. Recalling (C.2.5) and (C.2.2), this finishes the proof. □

C.2.2 Refining the `Branch-and-bound` for equality-constrained binary quadratic programming from Section 6.3.4.2

An approach like in Section 6.3.4.2 can be found in the article "Improving the performance of standard solvers for quadratic 0-1 programs by a tight convex reformulation: The QCR method" by Alain Billionnet, Sourour Elloumi and Marie-Christine Plateau.[38] They use the redundant constraints from Example 6.3.5 which yield the optimum value $OPT(SDP(QP_=)) = \left(PL^{\mathbf{X}}(QP_=)\right)$ as stated in Remark 6.3.9. However, instead of using $\mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$, they work with a restricted version $\mathscr{L}^{[0-1]}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$ of this program in which they also impose $x_i \in [0, 1]$, equivalent to $x_i^2 \leq x_i\ \forall i \in [1..n]$. However, they show that this program has the same objective value $OPT(SDP(QP_=))$ using an argument based on the Slater's interiority condition, stating that the total dual Lagrangian of $\mathscr{L}^{[0-1]}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$ has no duality gap—see the implication $(D1) \to (D2)$. However, we can here use a different argument.

If we construct $\mathscr{L}^{[0-1]}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$, $\mathscr{L}^{[0-1]}_{PL^{X}(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$ and $SDP^{[0-1]}(P_=)$ from resp. $\mathscr{L}_{PL^{\mathbf{X}}(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$, $\mathscr{L}_{PL^{X}(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)$ and $SDP(P_=)$ by adding constraints $x_i^2 \leq x_i$ (or

[38] Published in *Discrete Applied Mathematics* in 2009, vol 157 (6), pp. 1185-1197, a draft is available at `http://cedric.cnam.fr/fichiers/RC1120.pdf`.

resp. $X_{ii} \le x_i$) $\forall i \in [1..n]$ in resp. (6.3.7), (6.2.5b) and (6.2.2a)-(6.2.2e), then the following hierarchy holds and it naturally collapses:

$$\begin{aligned}
OPT\left(SDP(P_=)\right) &= OPT\left(\mathscr{L}_{PL\mathbf{X}(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)\right) && \text{(C.2.9a)}\\
&\le OPT\left(\mathscr{L}^{[0-1]}_{PL\mathbf{X}(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)\right) && \text{(C.2.9b)}\\
&= OPT\left(\mathscr{L}^{[0-1]}_{PL}X_{(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)\right) && \text{(C.2.9c)}\\
&\le OPT\left(SDP^{[0-1]}(P_=)\right) && \text{(C.2.9d)}\\
&= OPT\left(SDP(P_=)\right) && \text{(C.2.9e)}
\end{aligned}$$

Proof. The first equality (C.2.9a) is taken from (6.3.8). The inequality (C.2.9b) follows from the fact that the "$[0-1]$" version of the partial Lagrangian (a minimization program) is more constrained. The equality (C.2.9c) is due to the fact that the quadratic factor $Q_{\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*}$ of both programs (in variables X or $\mathbf{x}$) is SDP and that $X = \mathbf{x}\mathbf{x}^\top$ respects all constraints of the SDP program in variables X. The inequality (C.2.9d) follows from the fact that $\left(\mathscr{L}^{[0-1]}_{PL}X_{(QP_=)}(\boldsymbol{\mu}^*, \overline{\boldsymbol{\mu}}^*)\right)$ is a Lagrangian of $(SDP^{[0-1]}(P_=))$. Finally, (C.2.9e) holds because $(SDP(P_=))$ already integrates binary constraints $X_{ii} = x_i \ \forall i \in [1..n]$. ☐

More advanced convexifications can be found in the work of A. Billionnet, S. Elloumi and A. Lambert, *e.g.*, see papers "Extending the QCR method to general mixed integer programs." and "Exact quadratic convex reformulations of mixed-integer quadratically constrained problems".[39] However, for the moment, such methods lie outside the scope of this non-research document; further progress is unessential for now.

C.3 A convex function with an asymmetric Hessian

We next provide an example of a convex function with an asymmetric Hessian. This shows that a statement like "A twice differentiable function is convex if and only if its Hessian is SDP" is technically not complete, because a convex function can have an asymmetric non-SDP Hessian. This case is omitted from certain textbooks (see a reference in the first paragraph of Section 1.8) but we addressed it in our work by requiring the Hessian to be symmetric in Prop. 1.8.1.

Example C.3.1. *The following function f is convex for any $\mu \ge 9$ and has a non-symmetric Hessian in $\mathbf{0}$, i.e., $\nabla^2 f(\mathbf{0})$ is not symmetric.*

$$f(x,y) = \begin{cases} \dfrac{x^3 y}{x^2+y^2} + \mu x^2 + \mu y^2 & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } (x,y) = 0 \end{cases} \tag{C.3.1}$$

Proof. The gradient of f at $(0,0)$ can not be computed algebraically. We can however obtain $\dfrac{\partial f}{\partial x}(0,0) = \lim_{\varepsilon \to 0} \dfrac{f(\varepsilon,0) - f(0)}{\varepsilon} = \lim_{\varepsilon \to 0} \dfrac{\mu \varepsilon^2}{\varepsilon} = 0$ and similarly $\dfrac{\partial f}{\partial y}(0,0) = 0$. The gradient of f is thus:

$$\nabla f(x,y) = \begin{cases} \left(\dfrac{x^2 y \left(x^2+3y^2\right)}{\left(x^2+y^2\right)^2} + 2\mu x, \dfrac{x^3\left(x^2-y^2\right)}{\left(x^2+y^2\right)^2} + 2\mu y \right) & \text{if } (x,y) \neq (0,0) \\ (0,0) & \text{if } (x,y) = 0 \end{cases}$$

Let us now calculate the Hessian $\nabla^2 f(\mathbf{x})$. As above, we apply the derivative formula to calculate

[39] Both published in *Mathematical Programming*, resp .in 2012 (vol. 131(1), pp. 381-401) and 2016 (vol 158(1), pp 235-266).

$\nabla^2 f(0,0)$:

$$\frac{\partial^2 f}{\partial x \partial x}\Big(0,0\Big) = \lim_{\varepsilon\to 0} \frac{\frac{\partial f}{\partial x}\Big(\varepsilon,0\Big) - \frac{\partial f}{\partial x}\Big(0,0\Big)}{\varepsilon} = \lim_{\varepsilon\to 0} \frac{2\mu\varepsilon}{\varepsilon} = 2\mu$$

$$\frac{\partial^2 f}{\partial y \partial x}\Big(0,0\Big) = \lim_{\varepsilon\to 0} \frac{\frac{\partial f}{\partial x}\Big(0,\varepsilon\Big) - \frac{\partial f}{\partial x}\Big(0,0\Big)}{\varepsilon} = \lim_{\varepsilon\to 0} \frac{0}{\varepsilon} = 0$$

$$\frac{\partial^2 f}{\partial x \partial y}\Big(0,0\Big) = \lim_{\varepsilon\to 0} \frac{\frac{\partial f}{\partial y}\Big(\varepsilon,0\Big) - \frac{\partial f}{\partial y}\Big(0,0\Big)}{\varepsilon} = \lim_{\varepsilon\to 0} \frac{\frac{\varepsilon^3(\varepsilon^2-0^2)}{(\varepsilon^2+0^2)^2}}{\varepsilon} = \lim_{\varepsilon\to 0} \frac{\varepsilon}{\varepsilon} = 1$$

$$\frac{\partial^2 f}{\partial y \partial y}\Big(0,0\Big) = \lim_{\varepsilon\to 0} \frac{\frac{\partial f}{\partial y}\Big(0,\varepsilon\Big) - \frac{\partial f}{\partial y}\Big(0,0\Big)}{\varepsilon} = \lim_{\varepsilon\to 0} \frac{2\mu\varepsilon}{\varepsilon} = 2\mu$$

The Hessian of f is thus:

$$\nabla^2 f(x,y) = \begin{cases} \begin{bmatrix} \frac{2xy^3\left(3y^2-x^2\right)}{\left(x^2+y^2\right)^3} + 2\mu & \frac{x^2\left(x^4+6y^2x^2-3y^4\right)}{\left(x^2+y^2\right)^3} \\ \frac{x^2\left(x^4+6y^2x^2-3y^4\right)}{\left(x^2+y^2\right)^3} & \frac{2x^3y\left(y^2-3x^2\right)}{\left(x^2+y^2\right)^3} + 2\mu \end{bmatrix} & \text{if } (x,y) \neq (0,0) \\ \begin{bmatrix} 2\mu & 1 \\ 0 & 2\mu \end{bmatrix} & \text{if } (x,y) = 0 \end{cases}$$

All fractions in the Hessian have the form $\frac{p_1(x,y)}{p(x,y)}$, where p_1 and p are both homogeneous polynomials of degree 6. We easily obtain $\frac{p_1(tx,ty)}{p(tx,ty)} = \frac{t^6 p_1(x,y)}{t^6 p(x,y)} = \frac{p_1(x,y)}{p(x,y)}$. The graph of such fraction can be seen as a surface that has the same height (value) on each ray starting from (but not touching) the origin. The image of $\frac{p_1(x,y)}{p(x,y)}$ over $\mathbb{R}^2 - \{0\}$ is equal to the image of $\frac{p_1(x,y)}{p(x,y)}$ over unit circle $x^2+y^2=1$, which needs to be bounded.

We now calculate the bounds of the fractions. First, notice $\frac{p_1(x,y)}{p(x,y)} = p_1(x,y)$ over the unit circle, since $p(x,y) = (x^2+y^2)^6$. Each monomial of degree 6 of $p_1(x,y)$ belongs to the interval $[-1,1]$. Using this, we obtain, for instance, that $\frac{\partial^2 f}{\partial x \partial x} - 2\mu \in \left[\min\left(2xy^3\left(3y^2-x^2\right)\right), \max\left(2xy^3\left(3y^2-x^2\right)\right)\right] \subset [-8,8]$ when $x^2+y^2=1$. By applying the same approach on all fractions, we obtain:

$$-\begin{bmatrix} 8 & 10 \\ 10 & 8 \end{bmatrix} \leq \nabla^2 f(x,y) - \begin{bmatrix} 2\mu & 0 \\ 0 & 2\mu \end{bmatrix} \leq \begin{bmatrix} 8 & 10 \\ 10 & 8 \end{bmatrix}$$

By taking any $\mu \geq 9$, we obtain $\nabla^2 f(x,y) \succeq \mathbf{0}$ for $(x,y) \neq (0,0)$. On the other hand, we can *never* state $\nabla^2(0,0) \succeq \mathbf{0}$ because $\nabla^2(0,0)$ is not symmetric. We thus need to use $\nabla^2(0,0) + \nabla^2(0,0)^\top \succeq \mathbf{0}$.

The idea is taken from the last pages of the article "On second derivatives of convex functions" by Richard Dudley,[40] but we are the first to calculate an explicit minimum value of μ. The derivatives were calculated at `http://www.derivative-calculator.net`. $\square$

[40]Published in *Mathematica Scandinavica* in 1978, vol 41, pp 159–174, available on-line as of 2017 at `http://www.mscand.dk/article/download/11710/9726`, see also the discussion on the on-line math forum `https://math.stackexchange.com/questions/1181713/convex-function-with-non-symmetric-hessian`.

C.4 The separating hyperplane theorem

C.4.1 General theorems and their reduction to a particular case

Theorem C.4.1. *(Hyperplane separation theorem) Given two disjoint convex sets $X, Y \subset \mathbb{R}^n$, there exist a non-zero $\mathbf{v} \in \mathbb{R}^n$ and a real number c such that*

$$\mathbf{v} \bullet \mathbf{x} \geq c \geq \mathbf{v} \bullet \mathbf{y}, \tag{C.4.1}$$

for any $\mathbf{x} \in X$ and $\mathbf{y} \in Y$. The hyperplane $\{\mathbf{u} \in \mathbb{R}^n : \ \mathbf{v} \bullet \mathbf{u} = c\}$ separates X and Y.

Proof. We will show that the general theorem reduces to a simpler theorem version in which $Y = \{\mathbf{0}\}$, *i.e.*, the hyperplane **0**-separation Theorem C.4.2.

Consider the set $Z = X - Y = \{\mathbf{x} - \mathbf{y} : \ \mathbf{x} \in X, \ \mathbf{y} \in Y\}$. The set Z is convex: take $\mathbf{z}_a = \mathbf{x}_a - \mathbf{y}_a$, $\mathbf{z}_b = \mathbf{x}_b - \mathbf{y}_b$ and any $\alpha \in [0,1]$ and observe $\alpha \mathbf{z}_a + (1-\alpha)\mathbf{z}_b = \alpha(\mathbf{x}_a - \mathbf{y}_a) + (1-\alpha)(\mathbf{x}_b - \mathbf{y}_b) = \alpha \mathbf{x}_a + (1-\alpha)\mathbf{x}_b - (\alpha \mathbf{y}_a + (1-\alpha)\mathbf{y}_b)$. Since X and Y are convex, $\mathbf{x}_\alpha = \alpha \mathbf{x}_a + (1-\alpha)\mathbf{x}_b \in X$ and $\mathbf{y}_\alpha = \alpha \mathbf{y}_a + (1-\alpha)\mathbf{y}_b \in Y$, and so, $\mathbf{z}_a + (1-\alpha)\mathbf{z}_b = \mathbf{x}_\alpha - \mathbf{y}_\alpha \in Z$, *i.e.*, Z is convex.

We apply the hyperplane **0**-separation Theorem C.4.2 on Z and $\mathbf{0}$ (observe $\mathbf{0} \notin Z$ because X and Y are disjoint) and obtain there is a non-zero $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{v} \bullet \mathbf{z} \geq 0 \ \forall \mathbf{z} \in Z$. This means that

$$\mathbf{v} \bullet \mathbf{x} \geq \mathbf{v} \bullet \mathbf{y}, \ \forall \mathbf{x} \in X \text{ and } \forall \mathbf{y} \in Y. \tag{C.4.2}$$

We obtain $\inf_{\mathbf{x} \in X} \mathbf{v} \bullet \mathbf{x} \geq \sup_{\mathbf{y} \in Y} \mathbf{v} \bullet \mathbf{y}$, because otherwise (C.4.2) would be violated by some $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{v} \bullet \mathbf{x}$ is close enough to $\inf_{\mathbf{x} \in X} \mathbf{v} \bullet \mathbf{x}$ and $\mathbf{v} \bullet \mathbf{y}$ is close enough to $\sup_{\mathbf{y} \in Y} \mathbf{v} \bullet \mathbf{y}$. Taking $c = \frac{\inf_{\mathbf{x} \in X} \mathbf{v} \bullet \mathbf{x} + \sup_{\mathbf{y} \in Y} \mathbf{v} \bullet \mathbf{y}}{2}$, (C.4.2) can be written in the form (C.4.1). ☐

Theorem C.4.2. *(Hyperplane **0**-separation theorem) Given convex set $X \subset \mathbb{R}^n$ that does not contain **0**, there exist a non-zero $\mathbf{v} \in \mathbb{R}^n$ such that*

$$\mathbf{v} \bullet \mathbf{x} \geq 0,$$

for any $\mathbf{x} \in X$.

We give two proofs. The first one is based on induction and it takes a bit more than one page. It is a personal proof; I doubt it can also be found in classical textbooks. The second proof takes 2.5 pages and it follows well-established arguments (some of them using convergent sequences) that I could found on the Internet.[41] The first proof essentially relies on Theorem C.4.4 from Appendix C.4.2; the second one relies on Theorem C.4.5 and Theorem C.4.6 from Appendix C.4.3.

Proof 1 For any $\mathbf{u} \in \mathbb{R}^n$, let $f(\mathbf{u})$ be the largest t such that $t\mathbf{u} \in X$, or $-\infty$ if no $t\mathbf{u}$ belongs to X. If $f(\mathbf{u}) > 0$ for all non-zero $\mathbf{u} \in \mathbb{R}^n$, then $\mathbf{0} \in X$: it is enough to take $\mathbf{u}$ and $-\mathbf{u}$ and observe that the segment joining $f(\mathbf{u})\mathbf{u}$ and $f(-\mathbf{u}) - \mathbf{u}$ contains $\mathbf{0}$. Since X does not contain $\mathbf{0}$, there exist some non-zero $\mathbf{u} \in \mathbb{R}^n$ such that $f(\mathbf{u}) \leq 0$. The theorem then follows from applying Theorem C.4.4. ☐

Proof 2 We first prove that the closure $\overline{X}$ of X (*i.e.*, the set X along with all its limit points) is convex. Take any $\mathbf{x}, \ \mathbf{y} \in \overline{X}$ and consider two sequences $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$ in X that converge to $\mathbf{x}$ and resp. $\mathbf{y}$. Such sequences always exist because, by definition, $\overline{X}$ is the set of the limit points of all sequences of X. Take any $\alpha \in [0,1]$; it is enough to show $\alpha\mathbf{x} + (1-\alpha)\mathbf{y} \in \overline{X}$ to prove that $\overline{X}$ is convex. We have $\mathbf{z}_i = \alpha\mathbf{x}_i + (1-\alpha)\mathbf{y}_i \in X$, because X is convex. We next observe that $z = \lim\limits_{i\to\infty} z_i \in \overline{X}$, because $\overline{X}$ contains all limit points of X. But $\alpha\mathbf{x} + (1-\alpha)\mathbf{y} = \alpha \lim\limits_{i\to\infty} \mathbf{x}_i + (1-\alpha) \lim\limits_{i\to\infty} \mathbf{y}_i = \lim\limits_{i\to\infty} \mathbf{z}_i = \mathbf{z} \in \overline{X}$, and so, $\overline{X}$ is convex.

If $\mathbf{0}$ does not belong to the closure $\overline{X}$, then the conclusion follows from the Simple Separation Theorem C.4.5 applied on $\mathbf{0}$ and $\overline{X}$. If $\mathbf{0}$ belongs to the closure of X, the conclusion follows from the Simple Supporting Hyperplane Theorem C.4.6 applied on $\overline{X}$ and $\mathbf{0}$ as a boundary point of $\overline{X}$ ($\mathbf{0}$ does not belong to the interior, because it does not belong to X). ☐

The following variant can be generally useful, but we do not need it in this document.

[41] I used the Wikipedia article en.wikipedia.org/wiki/Hyperplane_separation_theorem and the course of Peter Norman www.unc.edu/~normanp/890part4.pdf.

Theorem C.4.3. *(Separation theorem for open set X) Given two disjoint convex sets $X, Y \subset \mathbb{R}^n$ such that X is open, there exist a non-zero $\mathbf{v} \in \mathbb{R}^n$ and a real number c such that*

$$\mathbf{v} \bullet \mathbf{x} > c \geq \mathbf{v} \bullet \mathbf{y}, \forall \mathbf{x} \in X,\ \mathbf{y} \in Y \tag{C.4.3}$$

The closure $\overline{X}$ of X satisfies

$$\mathbf{v} \bullet \overline{\mathbf{x}} \geq c \geq \mathbf{v} \bullet \mathbf{y}, \forall \overline{\mathbf{x}} \in \overline{X},\ \mathbf{y} \in Y \tag{C.4.4}$$

Proof. Using the standard hyperplane separation Theorem C.4.1, there is a non-zero $\mathbf{v} \in \mathbb{R}^n$ and some $c \in \mathbb{R}$ such that:

$$\mathbf{v} \bullet \mathbf{x} \geq c \geq \mathbf{v} \bullet \mathbf{y}, \forall \mathbf{x} \in X,\ \mathbf{y} \in Y \tag{C.4.5}$$

For the sake of contradiction, assume there exists some $\mathbf{x} \in X$ such that $\mathbf{v} \cdot \mathbf{x} = c$. Since X is open, X contains an open ball around $\mathbf{x}$, and so, for a sufficiently small $\epsilon > 0$, we have $\mathbf{x} - \epsilon\mathbf{v} \in X$. But $\mathbf{v} \cdot (\mathbf{x} - \epsilon\mathbf{v}) = c - \epsilon|\mathbf{v}|^2 < c$ which contradicts (C.4.5). The assumption $\mathbf{v} \cdot \mathbf{x} = c$ was false, and so, (C.4.5) becomes (C.4.3).

We still have to prove (C.4.4). Assume there is some $\overline{\mathbf{x}}$ in $\overline{X}$ such that $\mathbf{v} \cdot \overline{\mathbf{x}} < c$. Since $\overline{\mathbf{x}}$ has to be the limit point of some sequence $\{\mathbf{x}_i\}$ with elements $\mathbf{x}_i \in X\ \forall i \in \mathbb{N}^*$, we deduce that $\lim\limits_{i\to\infty} \mathbf{v} \cdot \mathbf{x}_i = \mathbf{v} \cdot \overline{\mathbf{x}}$. For any $\epsilon > 0$ there exists some $m \in \mathbb{N}^*$ such that $|\mathbf{v} \cdot \mathbf{x}_i - \mathbf{v} \cdot \overline{\mathbf{x}}| < \epsilon\ \forall i \geq m$. Taking any $\epsilon < c - \mathbf{v} \cdot \overline{\mathbf{x}}$, we have $\mathbf{v} \cdot \mathbf{x}_m < c$, which contradicts (C.4.3). The assumption $\mathbf{v} \cdot \overline{\mathbf{x}} < c$ was false, which proves (C.4.4). ☐

C.4.2 Proving the theorem using personal arguments

Theorem C.4.4. *(Hyperplane $\mathbf{0}$-separation theorem in presence of open rays) Consider convex set $X \in \mathbb{R}^n$ (that may contain $\mathbf{0}$ or not) and let $f(\mathbf{u})$ be the largest t such that $t\mathbf{u} \in X$ for any $\mathbf{u} \in \mathbb{R}^n$. If $f(\mathbf{u}) \leq 0$, we say that ray $\mathbf{u}$ is open, because there is no $\epsilon > 0$ such that $\epsilon\mathbf{u} \in X$. If there is at least an open ray, then there exist some non-zero $\mathbf{v} \in \mathbb{R}^n$ such that*

$$\mathbf{v} \bullet \mathbf{x} \geq 0,\ \forall \mathbf{x} \in X.$$

Proof. We proceed by induction. We first prove it for $n = 2$ using the notion of angle.

Lemma C.4.4.1. *The theorem holds for $n = 2$.*

Proof. Without loss of generality, we consider the open ray $\mathbf{u} \in \mathbb{R}^2$ is unitary. We associate $\mathbf{u}$ with an angle $\theta = 0$. Using a slight notation abuse, let $f(\theta) = f(\mathbf{u}_\theta)$, where $\mathbf{u}_\theta$ is an unitary vector of $\mathbb{R}^2$ that makes an angle of θ with $\mathbf{u}$ measured clockwise. Technically, $\mathbf{u}_\theta$ satisfies $\mathbf{u} \cdot \mathbf{u}_\theta = \cos(\theta)$ and it is the first unitary vector with this property found by moving clockwise from $\mathbf{u}$. Let Θ be the set of angles θ for which $f(\theta) > 0$. Θ belongs to segment $(0, 2\pi)$ because it does not contain $\theta = 0$ and it might be open. However, Θ needs to have an infimum $\inf(\Theta)$ and a supremum $\sup(\Theta)$, see also Prop. C.4.7.

Assume for the sake of contradiction that $\sup(\Theta) - \inf(\Theta) > \pi$. This means there are two angles $\theta_M, \theta_m \in \Theta$ close enough to $\sup(\Theta)$ and resp. $\inf(\Theta)$ so that $\theta_M - \theta_m > \pi$. By convexity, the segment that joins $f(\theta_M)\mathbf{u}_{\theta_M}$ and $f(\theta_m)\mathbf{u}_{\theta_m}$ is included in X and it also intersects the segment $[\mathbf{0}, \mathbf{u}]$ in some point $\epsilon\mathbf{u}$ with $\epsilon > 0$. This contradicts the fact that $\mathbf{u}$ is an open ray.

We can now consider $\sup(\Theta) - \inf(\Theta) \leq \pi$. The line $\left\{t\mathbf{u}^* :\ t \in \mathbb{R}, \mathbf{u}^* \in \mathbb{R}^2 - \{\mathbf{0}\}\right\}$ that goes through $\mathbf{0}$ and makes an angle of $\inf(\Theta)$ with $\mathbf{u}$ (clockwise) has the whole X on one side; all points in X make an angle in $[\inf(\Theta), \sup(\Theta)]$ with $\mathbf{u}$ (clockwise). We can take $\mathbf{v}$ one of the two vectors perpendicular to $\mathbf{u}^*$ in $\mathbf{0}$ and obtain $\mathbf{v} \cdot \mathbf{x} \geq 0$ for all $\mathbf{x} \in X$. ☐

Now consider $n > 2$. Take a 2-dimensional sub-space S_2 that contains the unitary vector $\mathbf{u}$, *i.e.*, $S_2 = \{t\mathbf{u} + t'\mathbf{u}' :\ t, t' \in \mathbb{R}\}$ for some unitary $\mathbf{u}'$ such that $\mathbf{u} \cdot \mathbf{u}' = 0$. The intersection of two convex sets is convex, and so, $X_2 = S_2 \cap X$ is convex. Using above lemma for $n = 2$, there exists some unitary $\mathbf{v}_2 \in S_2$ such that $\mathbf{v}_2 \cdot \mathbf{x}_2 \geq 0,\ \forall \mathbf{x}_2 \in X_2$. Notice that $f(-\mathbf{v}_2) \leq 0$ because all $-t\mathbf{v}_2$ with $t > 0$ do not belong to X or X_2, since $\mathbf{v}_2 \cdot (-t\mathbf{v}_2) < 0$.

Take any *unitary* $\mathbf{u}_2 \in S_2$ such that

$$\mathbf{v}_2 \bullet \mathbf{u}_2 = 0. \tag{C.4.6}$$

Consider the $(n-1)$-dimensional sub-space $S_{n-1} \subsetneq \mathbb{R}^n$ perpendicular on $\mathbf{u}_2$, *i.e.*, $S_{n-1} = \{\mathbf{x} \in X : \ \mathbf{u}_2 \bullet \mathbf{x} = 0\}$. Observe $\mathbf{v}_2$ and $-\mathbf{v}_2$ belong to S_{n-1}, using (C.4.6). We now project the whole space X on S_{n-1}, *i.e.*, we obtain the set $X_{n-1} = \{\mathbf{x} - \mathbf{u}_2(\mathbf{u}_2 \bullet \mathbf{x}) : \ \mathbf{x} \in X\}$. One can easily check that all elements of X_{n-1} satisfy $\mathbf{u}_2 \bullet (\mathbf{x} - \mathbf{u}_2(\mathbf{u}_2 \bullet \mathbf{x})) = \mathbf{u}_2 \bullet \mathbf{x} - \mathbf{u}_2 \bullet \mathbf{u}_2(\mathbf{u}_2 \bullet \mathbf{x}) = \mathbf{u}_2 \bullet \mathbf{x} - \mathbf{u}_2 \bullet \mathbf{x} = 0$.

We now define function $f_{n-1} : S_{n-1} \to \mathbb{R} \cup \{-\infty\}$ in the same style as f, *i.e.*, $f_{n-1}(\mathbf{s}_{n-1})$ is the smallest t for which $t\mathbf{s}_{n-1} \in X_{n-1}$, for any $\mathbf{s}_{n-1} \in S_{n-1}$. We showed above that $f(-\mathbf{v}_2) \leq 0$. We can also prove $f_{n-1}(-\mathbf{v}_2) \leq 0$. Recall we have $\mathbf{v}_2 \bullet \mathbf{x}_2 \geq 0, \ \forall \mathbf{x}_2 \in X_2$. Consider now the projection $\mathbf{x}_{n-1} = \mathbf{x}_2 - \mathbf{u}_2(\mathbf{u}_2 \bullet \mathbf{x}_2)$ and notice that $\mathbf{v}_2 \bullet \mathbf{x}_{n-1} = \mathbf{v}_2 \bullet (\mathbf{x}_2 - \mathbf{u}_2(\mathbf{u}_2 \bullet \mathbf{x}_2)) = \mathbf{v}_2 \bullet \mathbf{x}_2 - \mathbf{v}_2 \bullet \mathbf{u}_2(\mathbf{u}_2 \bullet \mathbf{x}_2) = \mathbf{v}_2 \bullet \mathbf{x}_2 \geq 0$ (we used (C.4.6) for the last equality). The elements $-t\mathbf{v}_2$ with $t > 0$ can not belong to the projection of X_2, because $\mathbf{v}_2 \bullet (-t\mathbf{v}_2) < 0$ and the elements $\mathbf{x}_{n-1}$ of the projection verify $\mathbf{v}_2 \bullet \mathbf{x}_{n-1} \geq 0$. Finally, remark we do not lose generality by restricting the argument to the projections of X_2: all elements of X that could project on $-t\mathbf{v}_2$ could only belong to X_2, *i.e.*, the space generated by $\mathbf{v}_2$ and $\mathbf{u}_2$.

We can easily check that X_{n-1} is convex. Consider $\mathbf{x}_{n-1} \in X_{n-1}$ as the projection of $\mathbf{x}_{n-1} + a\mathbf{u}_2 \in X$ and $\mathbf{y}_{n-1} \in X_{n-1}$ as the projection of $\mathbf{y}_{n-1} + b\mathbf{u}_2 \in X$. Using the convexity of X, the following holds for any $\alpha \in [0,1]$: $\alpha(\mathbf{x}_{n-1} + a\mathbf{u}_2) + (1-\alpha)(\mathbf{y}_{n-1} + b\mathbf{u}_2) \in X$. We can re-write this as: $\alpha\mathbf{x}_{n-1} + (1-\alpha)\mathbf{y}_{n-1} + (\alpha a + (1-\alpha)b)\mathbf{u}_2 \in X$, and so, $\alpha\mathbf{x}_{n-1} + (1-\alpha)\mathbf{y}_{n-1} \in X_{n-1}$ (one can easily verify that the scalar product of this with $\mathbf{u}_2$ is zero).

We now apply the induction hypothesis on set X_{n-1} with $f_{n-1}(-\mathbf{v}_2) \leq 0$ in the sub-space S_{n-1} (notice this is a full $(n-1)$-dimensional space where all elements can be written as a linear combination of a canonical basis perpendicular to $\mathbf{u}_2$). We obtain there is non-zero $\mathbf{v}_{n-1} \in S_{n-1}$ such that $\mathbf{v}_{n-1} \bullet \mathbf{x}_{n-1} \geq 0$ for all $\mathbf{x}_{n-1} \in X_{n-1}$. This means that $\mathbf{v}_{n-1} \bullet (\mathbf{x}_{n-1} + a\mathbf{u}_2) \geq 0$ for any $\mathbf{x}_{n-1} \in X_{n-1}$ and any $a \in \mathbb{R}$, because $\mathbf{v}_{n-1}$ is perpendicular on $\mathbf{u}_2$. It is easy to check that $X \subseteq \{\mathbf{x}_{n-1} + a\mathbf{u}_2 : \ \mathbf{x}_{n-1} \in X_{n-1}, \ a \in \mathbb{R}\}$. This is enough to conclude that $\mathbf{v}_{n-1}\mathbf{x} \geq 0$ for any $\mathbf{x} \in X$. ☐

C.4.3 Proving the theorem using well-established textbook arguments

Theorem C.4.5. *(Simple separation theorem) Given convex closed set $X \subset \mathbb{R}^n$ and some $\mathbf{y} \in \mathbb{R}^n$ such that $\mathbf{y} \notin X$, there exist a non-zero $\mathbf{v} \in \mathbb{R}^n$ such that*

$$\mathbf{v} \bullet \mathbf{x} > \mathbf{v} \bullet \mathbf{y}, \ \forall \mathbf{x} \in X. \tag{C.4.7}$$

Proof. We need the following lemma.

Lemma C.4.5.1. *Given closed convex set X, there exist a unique $\mathbf{x} \in X$ such that $|\mathbf{x}| = \inf\{|\mathbf{x}'| : \ \mathbf{x}' \in X\}$, where $|\cdot|$ is the norm (length), e.g., $|\mathbf{x}| = \sqrt{\mathbf{x} \bullet \mathbf{x}}$.*

Proof. The Wikipedia proof is kind of magical for my taste, with a few tricks arising rather out of the blue. I provide a more natural and even simpler (without Cauchy sequences) proof.

Let $\delta = \inf\{|\mathbf{x}'| : \ \mathbf{x}' \in X\}$. We first show X contains a sequence $\{\mathbf{x}_i\}$ such that $\lim\limits_{i\to\infty} |\mathbf{x}_i| = \delta$. For instance, we can consider $\epsilon_1 = 1$, $\epsilon_2 = \frac{1}{2}$, $\epsilon_3 = \frac{1}{3}$, ... For each ϵ_i ($i \in \mathbb{N}^*$), X needs to contain some $\mathbf{x}_i$ such that $|\mathbf{x}_i| < \delta + \epsilon_i$, because otherwise we would have $\delta + \epsilon_i \leq \inf\{|\mathbf{x}'| : \ \mathbf{x}' \in X\}$, impossible since $\delta = \inf\{|\mathbf{x}'| : \ \mathbf{x}' \in X\}$. The sequence $\{\mathbf{x}_i\}$ constructed this way satisfies $\lim\limits_{i\to\infty} |\mathbf{x}_i| = \delta$.

This sequence $\{\mathbf{x}_i\}$ of elements of $\mathbb{R}^n$ needs to contain a convergent subsequence using the Bolzano-Weierstrass Theorem C.4.9, *i.e.*, there exists a sub-sequence $\{\mathbf{x}_{n_i}\}$ such that $\lim\limits_{i\to\infty} \mathbf{x}_{n_i} = \mathbf{x}$. Since X is closed, it contains all limit points, and so, $\mathbf{x} \in X$. It is not hard now to check that the sub-sequence $\{|\mathbf{x}_{n_i}|\}$ converges to δ. Since for any ϵ there exists $m \in \mathbb{N}^*$ such that $|\mathbf{x}_j| < \delta + \epsilon \ \forall j \geq m$, there must be some $n_m \geq m$ (the sub-sequence is infinite) such that $|\mathbf{x}_{n_j}| < \delta + \epsilon \ \forall n_j \geq n_m$. This confirms $\delta = \lim\limits_{i\to\infty} |\mathbf{x}_{n_i}| = |\mathbf{x}|$.

We still need to show that $\mathbf{x}$ is the unique element of minimum norm. Suppose there exists $\mathbf{y} \in X - \{\mathbf{x}\}$ such that $|\mathbf{y}| = \delta$. By convexity, $\frac{\mathbf{x}+\mathbf{y}}{2} \in X$. We can calculate $|\frac{\mathbf{x}+\mathbf{y}}{2}|^2 = \frac{\mathbf{x} \bullet \mathbf{x} + \mathbf{y} \bullet \mathbf{y} + 2\mathbf{x} \bullet \mathbf{y}}{4} = \frac{\delta^2 + \mathbf{x} \bullet \mathbf{y}}{2}$. We will show $\mathbf{x} \bullet \mathbf{y} < \delta^2$. For this, it is enough to observe that $0 < |\mathbf{x} - \mathbf{y}|^2 = \mathbf{x} \bullet \mathbf{x} + \mathbf{y} \bullet \mathbf{y} - 2\mathbf{x} \bullet \mathbf{y} = 2\delta^2 - 2\mathbf{x} \bullet \mathbf{y}$, *i.e.*, $\mathbf{x} \bullet \mathbf{y} < \delta^2$. We obtained that $|\frac{\mathbf{x}+\mathbf{y}}{2}|^2 < \delta^2$, contradiction. There is no $\mathbf{y} \neq \mathbf{x}$ in X such that $|\mathbf{y}| = \delta$. ☐

We will first prove the theorem for $\mathbf{y} = \mathbf{0}$ and then we will use a simple translation argument to extend it for an arbitrary $\mathbf{y}$. Using the lemma, let us take $\mathbf{v} \in X$ of minimum norm $\delta > 0$ (because $\mathbf{y} = \mathbf{0} \notin X$). We will prove the following:

$$\mathbf{v} \bullet \mathbf{x} > 0, \ \forall \mathbf{x} \in X \tag{C.4.8}$$

Consider any $\mathbf{x} \in X$ and write $\Delta = \mathbf{x} - \mathbf{v}$. Keeping in mind the goal of showing $\mathbf{v} \cdot \Delta \geq 0$, consider a function $f : [0,1] \to \mathbb{R}$ defined by $f(t) = (\mathbf{v} + t\Delta) \cdot (\mathbf{v} + t\Delta)$. By convexity, we simply have $\mathbf{v} + t\Delta \in X\ \forall t \in [0,1]$, and, using the above lemma, we also obtain $f(t) > f(0) = \gamma\ \forall t > 0$. This means the derivative in 0 can not be negative; we need to have $f'(t) \geq 0$. Since $f'(t) = 2\mathbf{v} \cdot \Delta + t^2 \Delta \cdot \Delta$, this means $\mathbf{v} \cdot \Delta \geq 0$, enough to show $\mathbf{v} \cdot \mathbf{x} = \mathbf{v} \cdot (\mathbf{v} + \Delta) = \gamma + \mathbf{v} \cdot \Delta \geq \gamma > 0$.

Finally, if $\mathbf{y} \neq \mathbf{0}$, it is enough to consider set $X' = \{\mathbf{x} - \mathbf{y} :\ \mathbf{x} \in X\}$ which is convex and does not contain $\mathbf{0}$. We can apply the theorem for $\mathbf{0}$ and X' and we obtain there is $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{v} \cdot (\mathbf{x} - \mathbf{y}) > 0$, $\forall \mathbf{x} \in X$, which is equivalent to (C.4.7) as needed. ☐

Theorem C.4.6. *(Simple supporting hyperplane theorem) Given convex closed set $X \subset \mathbb{R}^n$ such that $\mathbf{0}$ is a boundary point of X, there exist a non-zero $\mathbf{v} \in \mathbb{R}^n$ such that*

$$\mathbf{v} \bullet \mathbf{x} \geq 0,\ \forall \mathbf{x} \in X. \tag{C.4.9}$$

Proof. Since $\mathbf{0}$ is a boundary point, there exists a sequence $\{\mathbf{x}_i\}$ of exterior points (*i.e.*, $\mathbf{x}_i \notin X$, $\forall i \in \mathbb{N}^*$) that converges to $\mathbf{0}$. Using the simple separation Theorem C.4.5, for each $\mathbf{x}_i$ there exists a non-zero unitary $\mathbf{v}_i \in \mathbb{R}^n$ such that

$$\mathbf{v}_i \bullet \mathbf{x} > \mathbf{v}_i \bullet \mathbf{x}_i\ \forall \mathbf{x} \in X. \tag{C.4.10}$$

Without loss of generality, we can consider all $\mathbf{v}_i$ are unitary, *i.e.*, $|\mathbf{v}_i| = 1\ \forall i \in \mathbb{N}^*$. As such, the sequence $\{\mathbf{v}_i\}$ is bounded, and so, we can apply the Bolzano-Weierstrass Theorem C.4.9 to conclude that $\{\mathbf{v}_i\}$ contains a convergent sub-sequence $\{\mathbf{v}_{n_i}\}$ such that $\lim\limits_{i\to\infty} \mathbf{v}_{n_i} = \mathbf{v}$.

Assume there is some $\mathbf{x} \in X$ such that $\mathbf{v} \cdot \mathbf{x} = -a < 0$. We derive a contradiction using a limiting argument. We take an $\epsilon < |a|$; since $\lim\limits_{i\to\infty} \mathbf{v}_{n_i} \cdot \mathbf{x} = \mathbf{v} \cdot \mathbf{x} = -a$, there exits some $m \in \mathbb{N}^*$ such that $\mathbf{v}_{n_i} \cdot \mathbf{x} \in [-a - \epsilon, -a + \epsilon]$ for all $i \geq m$. Applying (C.4.10), we obtain that all these $i \geq m$ satisfy $\mathbf{v}_{n_i} \cdot \mathbf{x}_{n_i} < -a + \epsilon < 0$. This contradicts the fact that $\lim\limits_{i\to\infty} \mathbf{x}_{n_i} = 0$. Indeed, if $\mathbf{v}_{n_i}$ can become arbitrarily close to $\mathbf{v}$ while at the same time $\mathbf{x}_{n_i}$ becomes arbitrarily close to $\mathbf{0}$, then the product $\mathbf{v}_{n_i} \cdot \mathbf{x}_{n_i}$ can also become arbitrarily close to 0. ☐

C.4.3.1 Convergence theorems on sequences

We need several convergence results on sequences for the classical proof of the hyperplane separation theorem, *i.e.*, in particular for Theorems C.4.5 and C.4.6.

Proposition C.4.7. *Any bounded set $S \subsetneq \mathbb{R}$ has a unique finite least upper bound (or supremum) $\sup(S)$. Equivalently, a unique $\inf(S)$ also exists and is finite.*

Proof. To avoid unessential complication, we will assume that S contains at least a positive number. Any set S' can be transformed to this form by applying a simple translation $S = \{s' - s'_0 :\ s \in S'\}$ for some fixed $s'_0 \in S'$. If $\sup(S)$ exists, then $\sup(S') = \sup(S) + s'_0$.

We will determine $\sup(S)$ as a real number a written in the decimal expansion as $a = a_0.a_1a_2a_3\ldots$ with potentially infinite number of digits. We can suppose that $a_0.a_1a_2\ldots a_n99\ldots9$ with $a_n \neq 9$ is equal to $a_0.a_1a_2\ldots a_{n-1}(a_n+1)$. We choose the decimals as follows.

– a_0 is the greatest integer that is not a strict upper bound for S (*i.e.*, that is not strictly greater than all elements of S). A finite a_0 value must exists because S is bounded.

– a_1 is the greatest digit such that $a_0.a_1$ is not a strict upper bound of S

– a_2 is the greatest digit such that $a_0.a_1a_2$ is not a strict upper bound of S

$\vdots$

– a_n is the greatest digit such that $a_0.a_1a_2\ldots a_n$ is not a strict upper bound of S

$\vdots$

We now prove that $a = a_0.a_1a_2a_3\ldots$ is an upper bound of S. Assume there exists an $s \in S$ such that $s > a$. This means there exists an index n such that s can be written $a_0.a_1a_2\ldots a_{n-1}s_ns_{n+1}s_{n+2}\ldots$ where $s_n > a_n$. Since a_n is the greatest digit such that $a_0.a_1a_2\ldots a_n$ is not a strict upper bound for S, we obtain that $a_0.a_1a_2\ldots a_{n-1}s_n$ is a strict upper bound for S. This implies that $s \geq a_0.a_1a_2\ldots a_{n-1}s_n$ is also a strict upper bound of S, which contradicts $s \in S$. There can be no $s \in S$ such that $s > a$.

To prove the uniqueness, we still have to show that a is the minimum upper bound, *i.e.*, there is no other upper bound $a' < a$. For the sake of contradiction, assume there exists such an upper bound $a' < a$; it can be written $a' = a_0.a_1a_2\ldots a_{n-1}a'_na'_{n+1}a'_{n+2}\ldots$ such that $a'_n < a_n$ for some n and $a'_{n+1}, a'_{n+2}\ldots$ are not all 9—such a number would reduce to $a_0.a_1a_2\ldots a_{n-1}(a'_n+1)$. We thus obtain that $a' < a_0.a_1a_2\ldots a_{n-1}a_n$. Since a' is an upper bound, $a_0.a_1a_2\ldots a_{n-1}a_n$ needs to be a strict upper bound. This contradicts the choice of a_n as the greatest digit such that $a_0.a_1a_2\ldots a_{n-1}a_n$ is *not* a strict upper bound of S.

By combining the two above paragraphs, we obtain that $a = a_0.a_1a_2a_3\ldots$ is the least upper bound $\sup(S)$. Recall we can suppose that $a_0.a_1a_2\ldots a_n99\ldots 9$ with $a_n \neq 9$ is equal to $a_0.a_1a_2\ldots a_{n-1}(a_n+1)$. ☐

Proposition C.4.8. *Any bounded monotone sequence $\{a_i\}$ of real numbers is convergent.*

Proof. Since the sequence is monotone, we can consider it is non-decreasing, the non-increasing case being completely analogous. Let $A = \{a_i : \ i \in \{1, 2, \ldots \infty\}\}$. Since A is bounded, the least upper bound property (Prop. C.4.7) states that $a = \sup(A)$ exists and is finite. For any $\epsilon > 0$, there needs to exist some positive integer n such that $a_n > a - \epsilon$, because otherwise $a - \epsilon$ would be an upper bound lower than $\sup(A)$, impossible. Since $\{a_i\}$ is non-decreasing, all integers $n' > n$ verify $a_{n'} \geq a_n > a - \epsilon$. This is exactly the definition of the fact that $\lim\limits_{i\to\infty} a_i = a$. ☐

Theorem C.4.9. *(Bolzano–Weierstrass theorem) Any bounded sequence $\{\mathbf{x}_i\}$ of $\mathbb{R}^n$ contains a convergent subsequence.*

Proof. We first show the theorem for $n = 1$.

Lemma C.4.9.1. *Any bounded sequence $\{x_i\}$ of $\mathbb{R}$ contains a convergent subsequence.*

Proof. We consider the set of maxima $N = \{i \in \{1, 2, \ldots\} : \ x_j < x_i, \forall j > i\}$. We distinguish three cases depending on the cardinal of N:

1. If $N = \emptyset$, then for any index n_i, there exists a position n_{i+1} such that $x_{n_{i+1}} \geq x_{n_i}$. The subsequence $x_{n_1},\ x_{n_2},\ x_{n_3}, \ldots$ is monotone non-decreasing, and so, convergent using Prop. C.4.8.

2. If N is not finite, then N contains an infinite sequence of indices $n_1 < n_2 < n_3 < \ldots$ such that $x_{n_1} > x_{n_2} > x_{n_3} \ldots$. The subsequence $x_{n_1},\ x_{n_2},\ x_{n_3}\ldots$ is monotone decreasing, and using Prop. C.4.8, it is convergent.

3. If $|N| = t$ with $t \in \mathbb{N} - \{0\}$, then N contains a finite sequence of indices $n_1 < n_2 < n_3 < \ldots n_t$ such that $x_{n_1} > x_{n_2} > x_{n_3} \cdots > x_{n_t}$. The set $N_{>n_t} = \{i \in \{n_t+1, n_t+2, n_t+3, \ldots\} : \ x_j < x_i, \forall j > i\}$ is empty. We can thus apply the argument of case 1 and obtain that the infinite sequence x_{n_t+1}, x_{n_t+2}, x_{n_t+3}, $\ldots$ contains a convergent sub-sequence. ☐

We now generalize the result for any $n > 1$. Considering the first position of the sequence $\{\mathbf{x}_i\}$, the above lemma shows that $\{\mathbf{x}_i\}$ contains a sub-sequence $\{\mathbf{x}_i^1\}$ whose first position converges to some y_1. We now consider the second position of $\{\mathbf{x}_i^1\}$. Using the lemma again, we obtain that $\{\mathbf{x}_i^1\}$ contains a sub-sequence $\{\mathbf{x}_i^2\}$ whose second position converges to some y_2. We observe that the first position of $\{\mathbf{x}_i^2\}$ converges to y_1 and the second to y_2. The argument can be repeated to find sub-sequences $\{\mathbf{x}_i^1\}$, $\{\mathbf{x}_i^2\}$, $\ldots$, $\{\mathbf{x}_i^n\}$ such that $\{\mathbf{x}_i^n\}$ converges to $[y_1\ y_2\ \ldots y_n]^\top$. ☐

References

References are provided throughout the document as footnote citations. This is because I wanted to make each reference readily available to the reader. I acknowledge again that I mentioned throughout the document the work of the following people (lecture notes and papers only, excluding web-sites and responses on mathematical on-line forums), in the order of apparition: Maurício de Oliveira, Christoph Helmberg, Robert Freund, David Williamson, Roger Horn, Charles Johnson H. Ikramov, Stephen Boyd, Lieven Vandenberghe, Anupam Gupta, László Lovász, Michael Overton, Henry Wolkowicz, Michel Goemans, Nebojša Gvozdenović, Donald Knuth, Monique Laurent, Immanuel Bomze, Mirjam Dür, Chung-Piaw Teo, Peter Dickinson, Luuk Gijben, Etienne de Klerk, Dmitri Pasechnik, Pablo Parrilo, Frédéric Roupin, Alain Billionnet, Sourour Elloumi, Marie-Christine Plateau, Alain Faye, Amélie Lambert, Peter Norman, Subhash Khot, Guy Kindler, Elchanan Mossel, Ryan O'Donnell and Richard Dudley.