Rank Reduction with Convex Constraints

Christian Grussler

LUND
UNIVERSITY

Department of Automatic Control

PhD Thesis TFRT-1117
ISBN 978-91-7753-080-0 (print)
ISBN 978-91-7753-081-7 (web)
ISSN 0280–5316

Department of Automatic Control
Lund University
Box 118
SE-221 00 LUND
Sweden

Printed in Sweden by MediaTryck.
Lund 2017

Abstract

This thesis addresses problems which require low-rank solutions under convex constraints. In particular, the focus lies on model reduction of positive systems, as well as finite dimensional optimization problems that are convex, apart from a low-rank constraint.

Traditional model reduction techniques try to minimize the error between the original and the reduced system. Typically, the resulting reduced models, however, no longer fulfill physically meaningful constraints. This thesis considers the problem of model reduction with internal and external positivity constraints. Both problems are solved by means of balanced truncation. While internal positivity is shown to be preserved by a symmetry property; external positivity preservation is accomplished by deriving a modified balancing approach based on ellipsoidal cone invariance.

In essence, positivity preserving model reduction attempts to find an infinite dimensional low-rank approximation that preserves nonnegativity, as well as Hankel structure. Due to the non-convexity of the low-rank constraint, this problem is even challenging in a finite dimensional setting. In addition to model reduction, the present work also considers such finite dimensional low-rank optimization problems with convex constraints. These problems frequently appear in applications such as image compression, multivariate linear regression, matrix completion and many more.

The main idea of this thesis is to derive the largest convex minorizers of rank-constrained unitarily invariant norms. These minorizers can be used to construct optimal convex relaxations for the original non-convex problem. Unlike other methods such as nuclear norm regularization, this approach benefits from having verifiable a posterior conditions for which a solution to the convex relaxation and the corresponding non-convex problem coincide. It is shown that this applies to various numerical examples of well-known low-rank optimization problems. In particular, the proposed convex relaxation performs significantly better than nuclear norm regularization. Moreover, it can be observed that a careful choice among the proposed convex relaxations may have a tremendous positive impact on matrix completion.

Computational tractability of the proposed approach is accomplished in two ways. First, the considered relaxations are shown to be representable by semi-definite programs. Second, it is shown how to compute the proximal mappings, for both, the convex relaxations, as well as the non-convex problem. This makes it possible to apply first order method such as so-called Douglas-Rachford splitting. In addition to the convex case, where global convergence of this algorithm is guaranteed, conditions for local convergence in the non-convex setting are presented.

Finally, it is shown that the findings of this thesis also extend to the general class of so-called atomic norms that allow us to cover other non-convex constraints.

Acknowledgments

First and foremost, I am most grateful to my supervisor Anders Rantzer, who supported me in pursuing a PhD at this wonderful department. He never seems to be out of ideas, and in our many discussions I had the privilege to learn a lot about research, and probably even about myself. He created opportunities for me to visit other universities, as well as to meet inspiring people. But the greatest gift that I received from him was his unconditional and constant confidence. With the same amount of gratefulness, I would like to thank my co-supervisor Pontus Giselsson. He has only supervised me for about a year, but his role in this thesis cannot be overstated. Working with him was pure fun, motivation, and inspiration. Together we had the unique opportunity to start exploring a new research area, side by side. I am also thankful to have worked with my second co-supervisor Andrey Ghulchak, who was another great source of new viewpoints and ideas.

Among the many inspiring people that I have met abroad, I would like to give a special mention to Armin Zare and Mihailo Jovanović. Thank you Armin for the countless Skype calls, and the many pleasant discussions. The same applies to my Master's thesis supervisor and collaborator Tobias Damm. He has always had a sympathetic ear for mathematic problems, and provided me with his help whenever possible.

The department of Automatic Control was always more than just a working place. I consider myself very lucky to have been a member of it. I would like to thank all the current and former staff for their interest, support, and time. Especially, I would like to express my gratitude to the heart(s) of this department, Eva Schildt, Eva Westin, Ingrid Nilsson, Mika Nishimura, and Monika Rasmusson. You make this department run, and, because of you, this department has a unique pleasant atmosphere. I would also like to thank my office neighbor Fredrik Magnusson for never being too busy to listen to my mathematics problems. This helped me a lot to sort out my thoughts.

The thesis would be far worse to read without the helpful comments from Giulia Giordano, Luc Muhirwa, Michelle Chong and Carolina Lidström. I am especially grateful for the help of Jochen Kall. He was not only a proofreader, but has also helped me with all my LaTeX-problems. At the same time I would like to thank (King) Richard Pates for taming the "beast paper" and Leif Andersson for all his support.

Even though it is already 10 years since I left, I am constantly thankful for having been a student of Helmut Blauth. In all stages of my university career, I have benefited from many great people. However, none of that would have happened without the fun and encouragement that I received from you in our mathematics classes together.

Finally, I thank my family for supporting me during my undergraduate years, which eventually made it possible to come to Sweden. I also appreciate that you have never asked a lot about what I am actually working with.

Contents

Preface

Contributions of the Thesis

The thesis consists of two chapters and four papers. The chapters provide mathematical background that is needed to understand the papers. This section describes the content of each chapter and the contribution of each paper.

Chapter 1 – Positivity Preserving Balanced Truncation

The first chapter consists of material relevant for the publications covered in Papers I and II. Basic control concepts, nonnegative matrices, positive systems and balanced truncation are covered, as far as they are needed in the context of positivity preserving balanced truncation. Based on that, the main challenges that arise from positivity preserving model reduction are discussed and linked to the contributions in Papers I and II.

Chapter 2 – Optimization

The second chapter covers some basic concepts in convex optimization, which form the basis for Papers III and IV. Further, Berge's maximum theorem as well as the Douglas-Rachford splitting algorithm are introduced.

Paper I

Grussler, C. and T. Damm (2012). "A symmetry approach for balanced truncation of positive linear systems". In: *51st IEEE Conference on Decision and Control (CDC)*, pp. 4308–4313.

This paper considers model order reduction of stable internally positive linear systems. It is shown how a symmetry characterization can be used in order to preserve positivity in balanced truncation. As a result, the reduced model has the additional feature of being symmetric. In contrast to other

methods, this approach works even in the absence of a positive state-space realization. Specifically, it is proven that balanced truncation to order one always preserves internal positivity. This also supplies a tractable necessary condition for the existence of an internally positive realization.

This paper is part of the first author's Master's Thesis. The topic has been suggested by the co-supervisor Anders Rantzer. All ideas are contributions of the first author, and have been derived in part under the supervision of the second author.

Paper II

Grussler, C. and A. Rantzer (2014). "Modified balanced truncation preserving ellipsoidal cone-invariance". In: *53rd IEEE Conference on Decision and Control (CDC)*, pp. 2365–2370.

The paper addresses model order reduction of stable linear systems that leave ellipsoidal (second-order) cones invariant. It is shown how balanced truncation can be modified to preserve ellipsoidal cone-invariance. Further, a numerically tractable method for verifying external positivity on a large class of systems is provided. As a consequence of these results, external positivity preserving model reduction can be performed within this class. The paper is the first of its kind in the sense that none of the problems has been addressed in the literature without involving internal positivity.

As part of a discussion of unpublished work, the second author has pointed out to the first author that invariance with respect to ellipsoidal cones is numerically verifiable. The rest of the paper is entirely the first author's contribution.

Paper III

Grussler, C., A. Rantzer, and P. Giselsson (2016). "Low-rank optimization with convex constraints". arXiv: 1606.01793.

The problem of low-rank approximation with convex constraints is considered. Given a data matrix, the objective of this paper is to find a low-rank approximation that meets rank and convex constraints while minimizing the distance to the data matrix in the Frobenius norm. It is proposed to use the largest convex minorizer (under-approximation) of the squared Frobenius norm and the rank constraint as a convex proxy, which can be combined with other convex constraints to form an optimal convex minorizer of the original non-convex problem. Unlike other approaches such as nuclear norm regularization or alternating minimization methods, this approach has the

advantage of having easily verifiable a posteriori conditions under which the solutions to the convex relaxation and the original non-convex problem coincide. The paper demonstrates that this is the case for several numerical examples of well-known low-rank optimization problems. In particular, the proposed convex relaxation consistently performs better than the nuclear norm heuristic and indicates tremendous benefits for the problem of matrix completion. Furthermore, the paper discusses why the proposed approach can also be considered as a regularization method.

The expressibility and computational tractability is of great importance for a convex relaxation. A closed-form expression for the proposed convex relaxation is provided, in addition to its representation as a semi-definite program. In order to deal with problems of large size, the so-called Douglas-Rachford splitting algorithm is applied to the convex relaxation as well as to the original non-convex problem. While in the convex case the algorithm is known to converge, there is no such guarantee in a non-convex setting. Nevertheless, if the proposed convex relaxation has a unique solution, it is shown that the convex and non-convex Douglas-Rachford iterations locally coincide. In particular, it is shown by an analytically tractable example that scaling the cost function in the non-convex Douglas-Rachford helps in finding solutions where the convex relaxation fails.

The work on this paper started with the observation of the first author, that good nonnegativity preserving low-rank approximation can be found using alternating minimization methods. These methods outperform nonnegative matrix factorization, as well as nuclear norm regularization. This motivated the second author to apply Lagrange duality, yielding the $r*$-norm relaxation. The rest of the paper has been derived by the first author as a result of discussions with the third author. The write up of the paper, as well as the applications and numerical examples, are entirely the work of the first author.

Paper IV

Grussler, C. and P. Giselsson (2016). “Low-rank inducing norms with optimality interpretations”. Preprint.

The paper considers rank constrained problems and introduces a family of low-rank inducing norms and regularizers of which the celebrated nuclear norm is a special case. A posteriori guarantees on solving an underlying rank constrained optimization problem with these convex relaxations are provided. In order to demonstrate the benefits that come with these new regularizers, three matrix completion problems are evaluated. In all examples, the nuclear norm heuristic is outperformed by convex relaxations

based on other low-rank inducing norms. In particular, for two of the problems, it can be proven that there exist low-rank inducing norms that succeed in recovering the partially unknown matrix, while the nuclear norm fails. Both of these low-rank inducing norms are shown to be representable as semi-definite programs and have cheaply computable proximal mappings. The latter makes it possible to solve also problems of large size with the help of scalable first-order methods. Finally, it is proven that these findings extend to the more general class of atomic norm problems. In particular, this allows us to solve corresponding vector-valued problems as well as problems with other non-convex constraints.

This paper has been jointly derived by both authors. The numerical examples were proposed by the first author, who has also carried out the write up for most of the parts.

Additional Publications

Grussler, C. and A. Rantzer (2015). “On optimal low-rank approximation of non-negative matrices”. In: *54th IEEE Conference on Decision and Control (CDC)*, pp. 5278–5283.

Grussler, C., A. Zare, M. R. Jovanovic, and A. Rantzer (2016). “The use of the $r*$ heuristic in covariance completion problems”. In: *55th IEEE Conference on Decision and Control (CDC)*.

1

Positivity Preserving Balanced Truncation

This chapter provides background material on positive systems and balanced truncation. In the context of positivity preserving model reduction, the main challenges are discussed, and linked to Papers I and II.

1.1 Basic Control Theory

In this section, basic concepts of linear control theory are recalled, which can be found in standard text books, such as [Zhou et al., 1996; Datta, 2004; Antoulas, 2005].

Linear Systems

Let $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{k\times n}$ and $D \in \mathbb{R}^{k\times m}$ define a linear time-invariant continuous-time system

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{S}$$

where x is the *state*, u the *input*, y the *output* and n the *order* of the system. The solution to system (S) can be explicitly computed as

$$x(t) = e^{At}x(0) + \int_0^t e^{A(t-\tau)}Bu(\tau)d\tau, \tag{1.1}$$

where $x(0)$ is the *initial state* and $t \geq 0$. Since x depends on A and B only, x is also referred as the state to (A, B).

Transfer Function and System Realization

The *transfer function* of the system (S) is given by $G(s) = C(sI-A)^{-1}B+D$ and (A, B, C, D) is referred to as a *realization* or *state-space representation* of G. There is no unique realization, because for any invertible $T \in \mathbb{R}^{n\times n}$ it holds that

$$G(s) = C(sI - A)^{-1}B + D = CT(sI - T^{-1}AT)^{-1}T^{-1}B + D.$$

Hence, also $(T^{-1}AT, T^{-1}B, CT, D)$ is a realization of G. In particular, if two realizations of G are driven by the same input with zero initial states, then their outputs coincide.

The realization (A, B, C, D) is called a *minimal* realization if G does not have a realization (A_l, B_l, C_l, D_l) such that $A_l \in \mathbb{R}^{l\times l}$ with $l < n$. Otherwise, (A, B, C, D) is said to be *non-minimal*. If (A, B, C, D) is said to be a realization of the system (S), it refers to the corresponding transfer function.

Controllability and Observability

The system (S) or (A, B) is said to be *controllable* if for any $x(0)$, $\bar{x} \in \mathbb{R}^n$ and $\bar{t} > 0$, there exists a (piecewise continuous) input u such that $x(\bar{t}) = \bar{x}$. Otherwise, it is said to be *uncontrollable*.

Proposition 1.1—[Zhou et al., 1996]
The following are equivalent:

i. (A, B) is controllable.

ii. The *controllability matrix*

$$\mathscr{C} := \begin{pmatrix} B & AB & \cdots & A^{n-1}B \end{pmatrix} \in \mathbb{R}^{n\times nm}$$

has rank n.

iii. For all $t > 0$

$$W_c(t) := \int_0^t e^{A\tau}BB^Te^{A^T\tau}d\tau \in \mathbb{R}^{n\times n}$$

is non-singular. □

In particular, $\mathrm{im}(\mathscr{C})$ is called the *controllable subspace* of (A, B), where $\mathrm{im}(K)$ and denotes the *image* of a matrix K. Further, it holds for all $t > 0$ that

$$\mathrm{im}(\mathscr{C}) = \mathrm{im}(W_c(t)).$$

Hence, if (A, B) is controllable, then $\mathrm{im}(\mathscr{C}) = \mathbb{R}^n$. The *orthogonal complement* to $\mathrm{im}(\mathscr{C})$ is called the *uncontrollable subspace*.

The system (S) or (A, C) is said to be *observable* if for an arbitrary $\bar{t} > 0$ the initial state can be reconstructed from knowing $u(t)$ and $y(t)$ on the interval $[0, \bar{t}]$. Otherwise, it is said be *unobservable*.

PROPOSITION 1.2—[ZHOU ET AL., 1996]
The following are equivalent:

i. (A, C) is observable

ii. The *observability matrix*

$$\mathscr{O} := \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} \in \mathbb{R}^{kn \times n}$$

has rank n.

iii. For all $t > 0$

$$W_o(t) := \int_0^t e^{A^T \tau} C^T C e^{A\tau} d\tau \in \mathbb{R}^{n \times n}$$

is non-singular. □

The subspace defined by $\ker(\mathscr{O})$ is called the *unobservable subspace* of (A, C), where $\ker(K)$ is the kernel of a matrix K. Further, it holds for all $t > 0$ that

$$\ker(\mathscr{O}) = \ker(W_o(t)).$$

Thus, if (A, C) is observable, then $\ker(\mathscr{O}) = \{0\}$. The orthogonal complement to $\mathrm{im}(\mathscr{O})$ is called the *observable subspace*.

Kalman decompositions

Let $r_c := \mathrm{rank}(\mathscr{C})$. Then there exists $T \in \mathbb{R}^{n \times n}$ such that

$$\tilde{A} := T^{-1}AT = \begin{pmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \end{pmatrix}, \ \tilde{B} := T^{-1}B = \begin{pmatrix} \tilde{B}_1 \\ 0 \end{pmatrix}, \ \tilde{C} := CT = \begin{pmatrix} \tilde{C}_1 & \tilde{C}_2 \end{pmatrix},$$

where $\tilde{A}_{11} \in \mathbb{R}^{r_c \times r_c}$, $\tilde{B}_1 \in \mathbb{R}^{r_c \times m}$ is a controllable pair (see e.g. [Zhou et al., 1996]). A realization of this type is said to be in *Kalman controllability form*. If $\tilde{x} = \begin{pmatrix} \tilde{x}_1^T & \tilde{x}_2^T \end{pmatrix}^T$ is the state to $(\tilde{A}, \tilde{B})$ with $\tilde{x}_1 \in \mathbb{R}^{r_c}$, then $\tilde{x}_2$ cannot be influenced by u and the controllable subspace of $(\tilde{A}, \tilde{B})$ is given by

$$\left\{ \begin{pmatrix} \tilde{x}_1 \\ 0 \end{pmatrix} \in \mathbb{R}^n : \tilde{x}_1 \in \mathbb{R}^{r_c} \right\}.$$

In particular, if the initial state of $(\tilde{A}, \tilde{B}, \tilde{C}, D)$ equals zero, then $\tilde{x}_2$ is zero and so the term $\tilde{C}_2\tilde{x}_2$ gives no contribution to the output. Hence, $\tilde{x}_2$ is not needed for a realization of (S). Indeed, if $r_c < n$, then (A, B, C, D) is non-minimal and $(\tilde{A}_{11}, \tilde{B}_1, \tilde{C}_1, D)$ is a realization of (S). Note that $(\tilde{A}_{11}, \tilde{B}_1, \tilde{C}_1, D)$ is obtained by *projecting onto the controllable subspace* in the sense that $\tilde{x}_1 = P\tilde{x}$ is the state to $(\tilde{A}_{11}, \tilde{B}_1)$ and

$$(\tilde{A}_{11}, \tilde{B}_1, \tilde{C}_1, D) = (P\tilde{A}P^T, P\tilde{B}, \tilde{C}P^T, D),$$

where $P := \begin{pmatrix} I & 0 \end{pmatrix} \in \mathbb{R}^{r_c \times n}$ with *identity matrix* $I \in \mathbb{R}^{r_c \times r_c}$.

Similarly, let $r_o := \operatorname{rank}(\mathscr{O})$. Then there exists $T \in \mathbb{R}^{n \times n}$ such that

$$\tilde{A} := T^{-1}AT = \begin{pmatrix} \tilde{A}_{11} & 0 \\ \tilde{A}_{21} & \tilde{A}_{22} \end{pmatrix}, \quad \tilde{B} := T^{-1}B = \begin{pmatrix} \tilde{B}_1 \\ \tilde{B}_2 \end{pmatrix}, \quad \tilde{C} := CT = \begin{pmatrix} \tilde{C}_1 & 0 \end{pmatrix},$$

where $\tilde{A}_1 \in \mathbb{R}^{r_o \times r_o}$, $\tilde{C}_1 \in \mathbb{R}^{k \times r_c}$ is an observable pair. A realization of this type is said to be in *Kalman observability form*. If $\tilde{x} = \begin{pmatrix} \tilde{x}_1^T & \tilde{x}_2^T \end{pmatrix}^T$ is the state to $(\tilde{A}, \tilde{B})$ with $\tilde{x}_1 \in \mathbb{R}^{r_o}$, then $\tilde{x}_2$ does not influence the output $\tilde{y} = \tilde{C}\tilde{x}$ and the observable subspace of $(\tilde{A}, \tilde{C})$ is given by

$$\left\{ \begin{pmatrix} \tilde{x}_1 \\ 0 \end{pmatrix} \in \mathbb{R}^n : \tilde{x}_1 \in \mathbb{R}^{r_o} \right\}.$$

This implies that $(\tilde{A}_{11}, \tilde{B}_1, \tilde{C}_1, D)$ is a realization of (S) and (A, B, C, D) is non-minimal if $r_o < n$. The observable realization $(\tilde{A}_{11}, \tilde{B}_1, \tilde{C}_1, D)$ is obtained by *projecting onto the observable subspace* in the sense that $\tilde{x}_1 = P\tilde{x}$ is the state to $(\tilde{A}_{11}, \tilde{B}_1)$ and

$$(\tilde{A}_{11}, \tilde{B}_1, \tilde{C}_1, D) = (P\tilde{A}P^T, P\tilde{B}, \tilde{C}P^T, D)$$

with $P := \begin{pmatrix} I & 0 \end{pmatrix} \in \mathbb{R}^{r_o \times n}$ and $I \in \mathbb{R}^{r_o \times r_o}$.

Since these decompositions can be applied successively, it follows that (A, B, C, D) is a minimal realization if and only if it is both controllable and observable.

Arnoldi Iteration

Assuming that (S) has a single input, the so-called *Arnoldi iteration* (see below) can be applied to $\operatorname{im}(\mathscr{C})$. The resulting unitary transformation V puts (V^TAV, V^TB, CV, D) into Kalman controllability form. Similarly, if the system has a single output, the Arnoldi iteration to $\operatorname{im}(\mathscr{O}^T)$ gives a unitary matrix that transforms the system into Kalman observability form.

The *Arnoldi iteration* is a variant of the Gram-Schmidt iteration, tailored to compute a particular orthogonal basis of the linear span of vectors

$$\{b, Ab, \ldots, A^n b\} \subset \mathbb{R}^n.$$

If this basis is collected in the columns of $V \in \mathbb{R}^{n\times n}$, then it holds that V^TAV is an *(upper) Hessenberg matrix*, i.e. it is zero below the first subdiagonal. Further, $V^Tb = \|b\|e_1$ with e_1 and $\|\cdot\|$ denoting the first canonical vector in $\mathbb{R}^n$ and the *Euclidean norm*, respectively. If A is symmetric, then Arnoldi iteration is equivalent to the so-called *Lanczos algorithm* (see [Trefethen and Bau III, 1997] for both methods).

Lyapunov Equations

Given $A, Q \in \mathbb{R}^{n\times n}$,

$$A^TX + XA + Q = 0 \tag{1.2}$$

is called a *Lyapunov equation* with the solution $X \in \mathbb{R}^{n\times n}$. Let the *spectrum* of $A \in \mathbb{R}^{n\times n}$ be denoted by $\sigma(A)$.

Lemma 1.1—[Zhou et al., 1996]
There exists a unique solution $X \in \mathbb{R}^{n\times n}$ to (1.2) if and only if $\lambda_i + \lambda_j \neq 0$ for all $\lambda_i, \lambda_j \in \sigma(A)$. □

For a (symmetric) *positive definite* $P \in \mathbb{R}^{n\times n}$, we write $P \succ 0$ and $P \succeq 0$ if it is *positive semi-definite*. Then

$$A^TX + XA + Q \preceq 0$$

is said to be a *Lyapunov inequality* with solution X.

Inertia

The *inertia* of a matrix $A \in \mathbb{R}^{n\times n}$ is the triplet $\mathrm{In}(A) = (p(A), z(A), n(A))$ consisting of the number of eigenvalues with positive, zero and negative real part, respectively, counting multiplicities. *Sylvester's law of inertia* (see [Horn and Johnson, 2012]) states that if $P \in \mathbb{R}^{n\times n}$ is non-singular, then

$$\mathrm{In}(A) = \mathrm{In}(P^TAP).$$

Proposition 1.3—[Antoulas, 2005]
Let $A, Q \in \mathbb{R}^{n\times n}$ with $Q \succeq 0$ and (A, Q) observable. If there exists a symmetric $X \in \mathbb{R}^{n\times n}$ such that

$$A^TX + XA + Q = 0,$$

then $\mathrm{In}(A) = \mathrm{In}(-X)$ and $z(A) = 0$. □

If $\mathrm{In}(A) = (0, 0, n)$, then $A \in \mathbb{R}^{n\times n}$ is called *Hurwitz* or *stable*. If A in (S) is Hurwitz, then (S) is also said to be stable. The eigenvalues of A are

also called the *poles* of a system. In particular, if the real part $\Re(\cdot)$ of an eigenvalue coincides with the *spectral abscissa*

$$\mu(A) := \max_{\lambda \in \sigma(A)} \Re(\lambda),$$

it is referred to as a *dominant pole*.

Controllability and Observability Gramian

LEMMA 1.2—[ZHOU ET AL., 1996]
Given a stable realization (A, B, C, D), the following holds:

i. The *controllability Gramian* $L_c := \lim_{t\to\infty} W_c(t)$ fulfills

$$AL_c + L_c A^T + BB^T = 0$$

and (A, B) is controllable if and only if $L_c \succ 0$.

ii. The *observability Gramian* $L_o := \lim_{t\to\infty} W_o(t)$ fulfills

$$A^T L_o + L_o A + C^T C = 0$$

and (A, C) is observable if and only if $L_o \succ 0$. □

Notice that for invertible $T \in \mathbb{R}^{n\times n}$, the Gramians to $(T^{-1}AT, T^{-1}B, CT, D)$ are obtained by

$$T^{-T} L_c T^{-T} \quad \text{and} \quad T^T L_o T,$$

where T^{-T} denotes the *inverse of the transpose* of T.

1.2 Nonnegative Matrices

A matrix $N = (n_{ij}) \in \mathbb{R}^{n\times m}$ is called *nonnegative* if all its elements are nonnegative, i.e. $n_{ij} \geq 0$ for all i and j. We use the abbreviation $N \in \mathbb{R}^{n\times m}_{\geq 0}$. By the *Perron-Frobenius Theorem* (see e.g. [Meyer, 2000]) it holds that the *spectral radius*

$$\rho(N) := \max_{\lambda \in \sigma(N)} |\lambda|$$

of $N \in \mathbb{R}^{n\times n}_{\geq 0}$ is an eigenvalue of N with corresponding nonnegative eigenvector.

Metzler matrices

A matrix $M \in \mathbb{R}^{n\times n}$ is said to be *Metzler* if there exists an $\alpha \in \mathbb{R}$ such that

$$M + \alpha I \in \mathbb{R}^{n\times n}_{\geq 0}.$$

Note that M is Metzler if and only if e^M is nonnegative [Haddad et al., 2010]. Moreover, it holds that $\mu(M) \in \sigma(M)$ with corresponding nonnegative eigenvector.

Proposition 1.4—[Berman and Plemmons, 1994]
Let $M \in \mathbb{R}^{n\times n}$ be Metzler, then the following are equivalent:

i. M is Hurwitz.

ii. There exists a diagonal $D \succ 0$ such that $MD + DM^T \prec 0$.

iii. There exists $\xi \in \mathbb{R}^n_{>0}$ such that $-M\xi \in \mathbb{R}^n_{>0}$.

iv. M is non-singular and $-M^{-1} \in \mathbb{R}^{n\times n}_{\geq 0}$. □

1.3 Positive Systems

In the following the classes of externally and internally positive systems are introduced. The state (input, output) of the system (S) is said to be nonnegative if it is nonnegative for all $t \geq 0$.

Internal Positivity

A system (S) is called *internally positive* if x and y are nonnegative, whenever the initial state $x(0)$ and u are nonnegative. This is equivalent to the matrix A being Metzler and B, C, D being nonnegative (see [Luenberger, 1979]). Thus internal positivity requires a certain realization which is not necessarily minimal. Since A is Metzler, it follows that internally positive systems have a dominant real pole.

Internally positive systems frequently appear in areas such as biomedicine, economics, data networks and many more, where y is the partial observation of a state x that represents nonnegative quantities of drugs, goods or bytes (see [Luenberger, 1979; Brown, 1980; Shorten et al., 2006; Haddad et al., 2010; Farina and Rinaldi, 2011]). Even though internally positive systems have been studied over the past decades [Luenberger, 1979; Ohta et al., 1984; Anderson et al., 1996; Son and Hinrichsen, 1996; Benvenuti and Farina, 2002], this class has only recently received significantly more attention (see e.g. [Tanaka and Langbort, 2011; Ebihara et al., 2012; Briat, 2011; Rantzer, 2015]). One of the main reasons for that lies in Proposition 1.4, which allows us for instance

- Stability verification in a *scalable* way, i.e. $A \in \mathbb{R}^{n\times n}$ being Metzler requires to determine n variables to verify stability, instead of n^2 (see [Rantzer, 2015]).
- H_∞-control with structured feedback controllers (see [Tanaka and Langbort, 2011; Lidström and Rantzer, 2015]).

External Positivity

A system (S) is called *externally positive* if y is nonnegative whenever the initial state $x(0) = 0$ and u is nonnegative. This is equivalent to the impulse response $Ce^{At}B+D$ being nonnegative for all $t \geq 0$ (see [Farina and Rinaldi, 2011]). Apart from the requirement that $D \in \mathbb{R}^{k\times m}_{\geq 0}$, external positivity is only an input-output property, which does not depend on a particular realization. Consequently, internal positivity implies external positivity, but not vice versa. The property that the system has a dominant real pole also transfers to externally positive systems. (see [Farina and Rinaldi, 2011]).

In general, verification of external positivity is known to be NP-hard (see [Blondel and Portier, 2002]). A computationally tractable (i.e. in polynomial time) sufficient condition for external positivity is derived in Paper II. Recently, the notions of internal and external positivity have been generalized in [Rantzer, 2012; Sootla and Mauroy, 2015; Altafini, 2016] to systems that only partially capture the properties of positive systems.

Positive Realization

Due to the significant benefits of internal positivity (see [Tanaka and Langbort, 2011; Ebihara et al., 2012; Briat, 2011; Rantzer, 2015]), it is natural to ask whether a given transfer function admits an internally positive realization. The following is a brief discussion on the computational difficulties that arise from this question. We restrict ourselves to the single-input-single-output (SISO) case, i.e. $k = m = 1$ in (S).

Given any minimal realization (A, B, C, D) of the SISO system (S), the *reachable cone* K_r is defined as the smallest closed convex cone containing

$$\{e^{At}B : t \geq 0\}.$$

The *observable cone* is given by

$$K_o := \{x : Ce^{At}x \geq 0 \text{ for all } t \geq 0\}.$$

Given $\Omega \subset \mathbb{R}^n$, a matrix $T \in \mathbb{R}^{n\times n}$ is said to *leave S invariant* if $T\omega \in \Omega$ for all $\omega \in \Omega$. Thus, K_r and K_o are left invariant by e^{At} for all $t \geq 0$. It can be shown that (S) is externally positive if and only if

$$K_r \subset K_o \quad \text{and} \quad D \geq 0,$$

see [Ohta et al., 1984]. Internal positivity can be characterized similarly. To this end, let a convex cone $K_p \subset \mathbb{R}^n$ be called *polyhedral* if there exists a matrix $P \in \mathbb{R}^{n \times l}$ such that $K_p = \{Py : y \in \mathbb{R}^l_{\geq 0}\}$.

Proposition 1.5—[Ohta et al., 1984]
Let (A, B, C, D) be a minimal realization to the SISO-system (S) with $D \geq 0$. Then the system admits an internally positive realization if and only if there exists a polyhedral convex cone K_p such that

i. K_p is left invariant by e^{At} for all $t \geq 0$.

ii. $K_r \subset K_p \subset K_o$. □

The first requirement is equivalent to $A + \gamma I$ leaving K_p invariant for some $\gamma \geq 0$ (see [Ohta et al., 1984]). Thus testing whether a given K_p fulfills the first requirement of Proposition 1.5 is possible via linear programming. However, finding $K_p = \{Py : y \in \mathbb{R}^l_{\geq 0}\}$ remains a numerically intractable problem, because neither l nor γ are known. In particular, only if the state dimension is lower than three, it is possible to show that $l = 2$ and every externally positive system admits a minimal internally positive realization (see [Farina and Rinaldi, 2011; Grussler, 2012]). Otherwise, l may be arbitrarily large even if a minimal realization of an externally positive system has a state of dimension three only (see [Farina and Rinaldi, 2011]).

This implies that verifying external via internal positivity can be arbitrarily difficult even for simple systems. In Paper II, a sufficient test for external positivity is provided by replacing K_p with a so-called *ellipsoidal* or *second-order cone*. The advantage is that finding an ellipsoidal cone K_p, which fulfills the same two conditions as in Proposition 1.5, usually only requires the solutions to a few semi-definite programs.

1.4 Balanced Truncation

(Cyber-)physical systems are described by mathematical models in terms of differential equations such as (S). Under various assumptions, e.g. different initial conditions or control strategies, these models can be used to simulate the behavior of the system. These simulations can help get a better understanding of the system as well as save resources where otherwise experimental trials need to be performed.

Due to the complexity of a system, the corresponding models often involve large number of differential equations, which leads to computational demanding simulations. This is especially evident in a growingly interconnected world with models that describe power systems, Internet traffic, transportation networks and many others.

A systematic way of dealing with this *curse of dimensionality* is to apply so-called *model reduction* techniques. Given the system (S), the goal of model reduction is to find a linear system with a similar input-output behavior, but a lower dimensional state. For this purpose, many methods have been developed (see [Moore, 1981; Glover, 1984; Gugercin and Antoulas, 2004; Antoulas, 2005; Gugercin et al., 2008]) of which the most popular are based on linear subspace projection.

Among the most widely used projection methods is the so-called *balanced truncation* method (see [Moore, 1981; Gugercin and Antoulas, 2004]). The advantages of this method are the existence of a simple error bound, the need for few computational steps and its intuitive interpretation in terms of energy functionals. The latter two advantages are some of the reasons why balanced truncation is part of many undergraduate textbooks (see e.g [Johansson, 1993; Glad and Ljung, 2000; Dullerud and Paganini, 2013]). Unfortunately, given a system (S), the computational effort of balanced truncation is determined by the complexity of solving a Lyapunov equation, which is typically of magnitude $\mathcal{O}(n^3)$. If n is very large, it is possible to apply large-scale methods such as [Gugercin et al., 2008] to reduce the system to a size where balanced truncation is applicable. The main ideas and computational steps of balanced truncation are outlined in the following.

Minimal realization via Gramians

In Section 1.1 it is shown that the Kalman decompositions can be used to compute a controllable realization by projecting on the controllable subspace. The same can be done by using the controllability Gramian.

Assume (A, B, C, D) is a stable realization of (S) with *block diagonal* controllability Gramian $L_c = \operatorname{blkdiag}(\Sigma_1, 0)$, where $\Sigma_1 \in \mathbb{R}^{r_c \times r_c}$ is positive definite. Let

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \quad C = \begin{pmatrix} C_1 & C_2 \end{pmatrix}$$

be partitioned such that $A_{11} \in \mathbb{R}^{r \times r}$. Then the controllable subspace is given by

$$\operatorname{im}(L_c) = \left\{ \begin{pmatrix} x_1 \\ 0 \end{pmatrix} \in \mathbb{R}^n : x_1 \in \mathbb{R}^{r_c} \right\}$$

and the projection onto the controllable subspace (A_{11}, B_1, C_1, D) is a controllable realization of (S). Note that there always exists a *unitary matrix* U, i.e. $U^T U = U U^T = I$, such that $L_c = U^T \Lambda U$, where $\Lambda \succeq 0$ is diagonal. Thus, Λ is the controllability Gramian to the system

$$(U^T A U, U^T B, CU, D)$$

and the assumption that L_c is blockdiagonal can always be met. Analogous considerations can be made for the observability Gramian, which allows us to obtain a minimal realization.

Energy functionals

Assuming that (A, B, C, D) is a stable minimal realization to (S), the following discussion gives an intuitive measure on what other state equations can be removed without changing the output behavior too much.

For a (piecewise continuous) function $u : \mathbb{R}_{\geq 0} \to \mathbb{R}^m$ the *L_2-norm* is defined by

$$\|u\|_2 := \lim_{t\to\infty} \sqrt{\int_0^t \|u(t)\|^2 \, dt}.$$

The L_2 norm can be considered as a measure for the energy content of u.

Let the controllability Gramian L_c be non-singular. If the initial state is zero and we wish to steer to $\bar{x}$ over the time interval $[0, \infty)$, then a control input u with minimal L_2-norm is given by $u(t) = B^T e^{A^T t} L_c^{-1} \bar{x}$ and

$$\|u\|_2^2 = \bar{x}^T L_c^{-1} \bar{x},$$

see [Zhou et al., 1996]. The interpretation of this result is that states $\bar{x}$ which need a small amount of energy are *easy to reach* from 0. In particular, if L_c is diagonal with decreasingly sorted diagonal entries, then $\bar{x} = e_n$ is hardest to reach. Therefore, the influence of inputs u on x_n can be considered to be small. However, truncating x_n, as with the projection onto the controllable subspace, may still give a bad approximation, because x_n could have great effect on the output.

If the observability Gramian L_o is non-singular, $x(0) = x_0$ and $u \equiv 0$, then $y(t) = Ce^{At}x_0$ and

$$\|y\|_2^2 = x_0^T L_o x_0.$$

The states x_0 that provide a small contribution to the energy of y can be interpreted as being *hard to observe*. For diagonal L_o with decreasingly sorted diagonal entries, $x_0 = e_n$ is the hardest to observe and its influence on y may be neglected. Again, truncating x_n may still give a bad approximation, because x_n could be significantly influenced by u.

Unlike projections onto observable and controllable subspaces, only states that are both hard to reach and hard to observe can safely be neglected.

Balanced Realization

Given a stable realization (A, B, C, D) with Gramians $L_c, L_o \succ 0$, the goal of *balancing* is to find a non-singular T such that the equivalent system $(T^{-1}AT, T^{-1}B, CT, D)$ has diagonal Gramians that coincide. This transformation can be found as follows.

Let an eigenvalue decomposition of L_c be given by $L_c = U\Lambda_c U^T$ with diagonal $\Lambda \succ 0$ and unitary U. Then there exists a unitary matrix V such that

$$U\Lambda^{\frac{1}{2}} L_o \Lambda^{\frac{1}{2}} U^T = V\Sigma_{co} V^T,$$

where $\Sigma_{co} \succ 0$ is diagonal and $\Lambda^{\frac{1}{2}}$ is the *matrix square root* of Λ, i.e.

$$\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}} = \Lambda.$$

Letting $\Sigma_{co}^{-\frac{1}{2}}$ denote the inverse of $\Sigma_{co}^{\frac{1}{2}}$, it follows that $T := U\Lambda^{\frac{1}{2}}\Sigma_{co}^{-\frac{1}{2}}$ is non-singular and

$$T^{-1}L_c T^{-T} = \Sigma_{co} = T^T L_o T.$$

Hence, $(T^{-1}AT, T^{-1}B, CT, D)$ has diagonal Gramians Σ_{co} and is called a *balanced realization*. With such a realization, states are equally difficult to reach and to observe. The eigenvalues of Σ_{co} are called the *Hankel singular values*. Note that the squared Hankel singular values coincide with $\sigma(L_c L_o)$, because $\Sigma_{co}^2 = T^{-1}L_c L_o T$.

Truncation and Error Bound

If (S) is a stable system with transfer function G, then the so-called H_∞*-norm* of G is defined as

$$\|G\|_\infty := \sup_{\omega\in\mathbb{R}} \|G(i\omega)\|,$$

where $\|\cdot\|$ denotes the spectral norm. Assuming that (S) has zero initial state, the H_∞-norm can be expressed as

$$\|G\|_\infty = \sup_{\|u\|_2\leq 1} \|y\|_2,$$

where y is the output of (S) with input u. This norm can be used in order to measure the error between the original system and its approximation.

PROPOSITION 1.6—[ZHOU ET AL., 1996]
Assume (A, B, C, D) is a balanced realization of a stable transfer function G with Gramians $L_c = L_o = \Sigma_{co} = \text{blkdiag}(\Sigma_1, \Sigma_2) \succ 0$ such that

$$\Sigma_1 = \text{blkdiag}(\sigma_1 I_{k_1}, \ldots, \sigma_r I_{k_r}) \quad \text{and} \quad \Sigma_2 = \text{blkdiag}(\sigma_{r+1} I_{k_{r+1}}, \ldots, \sigma_N I_{k_N}),$$

where $\sigma_1 > \cdots > \sigma_N > 0$ and $I_{k_i} \in \mathbb{R}^{k_i}$ for all i. Let

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \quad C = \begin{pmatrix} C_1 & C_2 \end{pmatrix}$$

be partitioned such that $A_{11} \in \mathbb{R}^{k \times k}$, where $k := \sum_i^r k_i$. Then (A_{11}, B_1, C_1, D) is a balanced, minimal and stable realization of G_r that satisfies

$$\|G - G_r\|_\infty \leq 2 \sum_{i=1}^{r} \sigma_i.$$

□

In [Beck et al., 1996; Sandberg and Rantzer, 2004] it has been shown that the error bound in Proposition 1.6 also holds if the system is balanced with respect to the so-called *generalized Gramians* $\tilde{L}_c$ and $\tilde{L}_o$ satisfying

$$A\tilde{L}_c + \tilde{L}_c A^T + BB^T \preceq 0,$$
$$A^T \tilde{L}_o + \tilde{L}_o A^T + C^T C \preceq 0.$$

The reduced system G_r remains stable but (A_{11}, B_1, C_1, D) may not be minimal. Furthermore, the corresponding *generalized Hankel singular values*, i.e. $\sigma(\tilde{L}_c \tilde{L}_o)$, cannot be smaller than the Hankel singular values (see Paper I).

Singular Perturbation

Assume (A, B, C, D), G and G_r are as in Proposition 1.6, then

$$\lim_{s \to \infty} G(s) = \lim_{s \to \infty} G_r(s) = D.$$

It can be shown (see [Liu and Anderson, 1989]) that also the reduced system

$$(A_{11} - A_{12}A_{22}^{-1}A_{21}, B_1 - A_{12}A_{22}^{-1}B_2, C_1 - C_2 A_{22}^{-1} A_{21}, D - C_2 A_{22}^{-1} B_2)$$

with transfer function $\hat{G}_r(s)$ fulfills the error bound in Proposition 1.6 and

$$\lim_{s \to 0} G(s) = \lim_{s \to 0} \hat{G}_r(s).$$

This variant of balanced truncation is referred to as *singular perturbation balanced truncation*.

Positivity Preservation

Approximating a positive system with a lower dimensional positive system is a valid desire, since negative values in the output of the reduced system may lack interpretability in terms of the original system dynamics. Furthermore, internal positivity is required, if the intention is to apply tools, such as in [Tanaka and Langbort, 2011; Briat, 2011; Rantzer, 2015], to the reduced system.

Commonly used model reduction methods such as balanced truncation, however, do generally neither preserve internal nor external positivity. The reason for this is threefold:

i. Necessary transformations, such as balancing a system, usually destroy internal positivity.

ii. A priori, it is unknown whether the truncated system is externally positive.

iii. Internally positive realizations can be high-dimensional compared to the corresponding minimal realization.

In particular, a transformation matrix $T \in \mathbb{R}^{n\times n}$ leaves $\mathbb{R}^{n\times n}_{\geq 0}$ invariant if and only if $T = PD$, where $P \in \mathbb{R}^{n\times n}$ is a permutation matrix and $D \succ 0$ is diagonal. If T is a balancing transformation with respect to the generalized Gramians $\tilde{L}_c$ and $\tilde{L}_o$, then

$$T^{-1}\tilde{L}_c\tilde{L}_oT = D^{-1}P^T\tilde{L}_c\tilde{L}_oPD$$

being diagonal implies that $\tilde{L}_c\tilde{L}_o$ is diagonal. In [Reis and Virnik, 2009] this observation is utilized, because Proposition 1.4 indeed guarantees the existence of diagonal $\tilde{L}_c$ and $\tilde{L}_o$. Thus reduced internally positivity preserving approximations can be obtained while having an a priori error bound. Similar ideas based on Proposition 1.4 are used in [Sootla and Rantzer, 2012]. The problem that arises from such approaches is that T is essentially just a permutation, which often yields conservative errors in the H_∞-norm.

In Paper I an alternative method is suggested which is based on a symmetry property of balanced realizations, allowing to preserve internal positivity. Further, the work in Paper II is the first of its kind that relaxes this requirement by preserving external positivity, only. As a consequence, it is possible to perform generalized balanced truncation steps on externally positive systems without requiring an internally positive realization to begin with.

Finally, note that there exist non-linear optimization methods such as [Feng et al., 2010; Li et al., 2011; Li et al., 2014] that intend to preserve internal positivity. Unfortunately, they are computationally intractable and are not guaranteed to converge (see [Grussler, 2012]).

2

Convex Optimization

This chapter provides background material on convex optimization which can be found in standard text books, such as [Luenberger, 1968; Rockafellar, 1970; Hiriart-Urruty and Lemaréchal, 1996; Bauschke and Combettes, 2011].

2.1 Convex Sets

In the remainder of this chapter, it is assumed that all Hilbert spaces are real, finite-dimensional and equipped with an *inner product* $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$. In particular, it is assumed that all sets are subsets of the Hilbert space $\mathcal{H}$ if not otherwise stated.

Convex Set

Given $x_1, \ldots, x_k \in \mathcal{H}$, a point of the form

$$\sum_{i=1}^{k} \alpha_i x_i \quad \text{with} \quad \sum_{i=1}^{k} \alpha_i = 1 \quad \text{and} \quad \alpha_i \geq 0 \text{ for all } i,$$

is said to be a *convex combination* of $x_1, \ldots, x_k$. Geometrically, the set of all convex combinations of x_1 and x_2 is the line segment between x_1 and x_2.

A set $\mathcal{C}$ is called *convex* if

$$\alpha x_1 + (1 - \alpha) x_2 \in \mathcal{C}$$

for all $x_1, x_2 \in \mathcal{C}$ and all α such that $0 \leq \alpha \leq 1$.

Convex Hull and Extreme Points

Given a set $\mathcal{S}$, the *convex hull* $\operatorname{conv}(\mathcal{S})$ is defined as the smallest convex set containing $\mathcal{S}$. In particular,

$$\operatorname{conv}(\mathcal{S}) = \left\{ \sum_{i=1}^{k} \alpha_i x_i : k > 0,\ \sum_{i=1}^{k} \alpha_i = 1,\ x_i \in \mathcal{S},\ \alpha_i > 0 \text{ for all } i \right\}.$$

Let $\text{cl}(S)$ denote the *topological closure* of S, then

$$\text{cl}(\text{conv}(\mathcal{S}))$$

is called the *closed convex hull of* $\mathcal{S}$. Notice, in general

$$\text{cl}(\text{conv}(\mathcal{S})) \neq \text{conv}(\text{cl}(\mathcal{S})).$$

Lemma 2.1—[Hiriart-Urruty and Lemaréchal, 1996]
If $\mathcal{S}$ is bounded (compact), then $\text{conv}(\mathcal{S})$ is bounded (compact). □

Let $\mathcal{C}$ be convex, then $e \in \mathcal{C}$ is said to be an *extreme point of* $\mathcal{C}$, if there does not exist $x_1, x_2 \in \mathcal{C} \setminus \{e\}$ such that e is a convex combination of x_1 and x_2. The *set of all extreme points* of the convex set $\mathcal{C}$ is denoted by $\text{ext}(\mathcal{C})$.

Lemma 2.2—[Hiriart-Urruty and Lemaréchal, 1996]
If $\mathcal{C}$ is a compact convex set, then

$$\mathcal{C} = \text{conv}(\text{ext}(\mathcal{C})).$$ □

Lemma 2.3—[Rockafellar, 1970]
Two closed convex sets $\mathcal{C}_1$ and $\mathcal{C}_2$ coincide if and only if

$$\sup_{x \in \mathcal{C}_1} \langle x, y \rangle = \sup_{x \in \mathcal{C}_2} \langle x, y \rangle.$$

for all $y \in \mathcal{H}$ □

Cone

A set $\mathcal{K}$ is said to be a *cone* if

$$\alpha x \in \mathcal{K}$$

for all $x \in \mathcal{K}$ and all $\alpha \geq 0$. Thus, $\mathcal{K}$ is a *convex cone* if

$$\alpha_1 x_1 + \alpha_2 x_2 \in \mathcal{K}$$

for all $x_1, x_2 \in \mathcal{K}$ and all $\alpha_1, \alpha_2 \geq 0$.

Convex Conic Hull

The *convex conic hull* of a set $\mathcal{S}$ is defined as the smallest convex cone containing $\mathcal{S}$ and therefore

$$\text{cone}(\mathcal{S}) = \left\{ \sum_{i=1}^{k} \alpha_i x_i : k > 0,\ x_i \in \mathcal{S},\ \alpha_i \geq 0 \text{ for all } i \right\}.$$

A cone is said to be *solid* if it contains an interior point.

Affine Subspace

Given $x_1, \ldots, x_k \in \mathcal{H}$, a point of the form

$$\sum_{i=1}^{k} \alpha_i x_i \quad \text{with} \quad \alpha_i \in \mathbb{R} \text{ for all } i$$

is said to be a *linear combination* of $x_1, \ldots, x_k$. Geometrically, the set of all linear combinations of x_1 and x_2 forms the line through x_1 and x_2.

A set $\mathcal{L}$ is said to be a *linear subspace* if

$$\alpha_1 x_1 + \alpha_2 x_2 \in \mathcal{L}$$

for all $x_1, x_2 \in \mathcal{L}$ and all $\alpha_1, \alpha_2 \in \mathbb{R}$. If $x_1, \ldots, x_k \in \mathcal{L}$ is the smallest number of non-zero elements such that

$$\mathcal{L} = \left\{ \sum_{i=1}^{k} \alpha_i x_i : \alpha_i \in \mathbb{R} \text{ for all } i \right\},$$

then k is called the dimension $\dim(\mathcal{L})$ of the linear subspace $\mathcal{L}$. In particular, if $\mathcal{L} = \{0\}$ then $\dim(\mathcal{L}) = 0$.

A shifted linear subspace $\mathcal{A}$ is called an *affine subspace* or *linear variety*, i.e. for all $x_0 \in \mathcal{A}$ it holds that

$$\mathcal{A} = \{x_0 + x : x \in \mathcal{L}(\mathcal{A})\},$$

where

$$\mathcal{L}(\mathcal{A}) := \{\alpha_1(x_1 - x_0) + \alpha_2(x_2 - x_0) : x_1, x_2 \in \mathcal{A},\ \alpha_1, \alpha_2 \in \mathbb{R}\}$$

is a linear subspace that is parallel to $\mathcal{A}$. The *dimension* of $\mathcal{A}$ is defined as

$$\dim(\mathcal{A}) := \dim(\mathcal{L}(\mathcal{A})).$$

The *affine hull* $\operatorname{aff}(\mathcal{S})$ of a set $\mathcal{S}$ is defined as the smallest affine subspace containing $\mathcal{S}$ and therefore

$$\operatorname{aff}(\mathcal{S}) = \left\{ \sum_{i=1}^{k} \alpha_i x_i : k > 0,\ \sum_{i=1}^{k} \alpha_i = 1,\ x_i \in \mathcal{S},\ \alpha_i \in \mathbb{R} \text{ for all } i \right\}.$$

Relative Interior

The *relative interior* of a convex set $\mathcal{C}$ is defined as

$$\operatorname{ri}(\mathcal{C}) := \{x \in \operatorname{aff}(\mathcal{C}) : \exists \delta > 0 \text{ such that } \operatorname{aff}(\mathcal{C}) \cap \mathcal{B}_\delta(x) \subset \mathcal{C}\},$$

where $\mathcal{B}_\delta(x) := \{x \in \mathcal{H} : \|x\| \leq \delta\}$.

Carathéodory's Theorem on the Convex Hull

Let $\mathcal{C}$ be a convex set such that $\dim(\text{aff}(\mathcal{C})) = k$, then *Carathéodory's Theorem on the Convex Hull* (see [Hiriart-Urruty and Lemaréchal, 1996]) says that

$$\mathcal{C} = \left\{ \sum_{i=1}^{k+1} \alpha_i x_i : \sum_{i=1}^{k+1} \alpha_i = 1,\ x_i \in \mathcal{C},\ \alpha_i \geq 0 \text{ for all } i \right\}.$$

In particular, if $\mathcal{C}$ is compact, then

$$\mathcal{C} = \left\{ \sum_{i=1}^{k+1} \alpha_i x_i : \sum_{i=1}^{k+1} \alpha_i = 1,\ x_i \in \text{ext}(\mathcal{C}),\ \alpha_i \geq 0 \text{ for all } i \right\}.$$

2.2 Convex Functions

In the following, let the *effective domain* of a function $f : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ be defined as

$$\text{dom}(f) := \{x \in \mathcal{H} : f(x) < \infty\}.$$

If $\text{dom}(f) \neq \emptyset$, then f is called *proper*. In the remainder it is assumed that all functions are proper.

Convex Function

A function $f : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ is called *convex* if

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2)$$

for all $x_1, x_2 \in \text{dom}(f)$ and all α with $0 \leq \alpha \leq 1$. If $-f$ is convex, then f is called *concave*. The effective domain of a convex function f is convex and f is continuous on $\text{ri}(dom(f))$.

Strictly Convex Function

A convex function $f : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ is called *strictly convex* if

$$f(\alpha x_1 + (1-\alpha)x_2) < \alpha f(x_1) + (1-\alpha) f(x_2)$$

for all $x_1, x_2 \in \text{dom}(f)$ with $x_1 \neq x_2$ and for all α with $0 \leq \alpha \leq 1$.

Strongly Convex Function

A function $f : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ is said to be *strongly convex* with parameter $m > 0$ if

$$x \mapsto f(x) - \frac{m}{2}\|x\|^2$$

is convex.

Closed Function

A convex function $f : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ is said to be *closed* if the *epi-graph* of f

$$\operatorname{epi}(f) := \{(t, x) \in \mathbb{R} \times \mathcal{H} : f(x) \leq t\} \tag{2.1}$$

is a closed set.

Conjugate Function

The *conjugate function* to $f : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ at $x^* \in \mathcal{H}$ is defined as

$$f^*(x^*) := \sup_x [\langle x, x^* \rangle - f(x)].$$

Subdifferentials

Let $f : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ be a convex function, then

$$\partial f(x_0) := \{x_0^* \in \mathcal{H} : f(x) \geq f(x_0) + \langle x - x_0, x_0^* \rangle \text{ for all } x \in \mathcal{H}\}$$

is called the *subdifferential* of f at x_0.

Note that if $g : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ is another convex functional such that $\operatorname{ri}(\operatorname{dom}(g)) \cap \operatorname{ri}(\operatorname{dom}(f)) \neq \emptyset$, then

$$\partial[f(x_0) + g(x_0)] = \partial f(x_0) + \partial g(x_0),$$

where the addition is taken with respect to the *Minkowski sum*, i.e.

$$\partial f(x_0) + \partial g(x_0) := \{x + y : x \in \partial f(x_0),\ y \in \partial g(x_0)\}.$$

2.3 Optimality

An *optimization problem* is said to be *convex* if it can be written in the form

$$\underset{x}{\text{minimize}} \quad f(x) \tag{2.2}$$

where $f : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ is convex.

Minimum and Maximum

Let $f : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ and $x_0 \in \mathcal{H}$ be such that there exists $\delta > 0$ with

$$f(x_0) \leq f(x) \quad \text{for all} \quad x \in \mathcal{B}_\delta(x_0) \cap \operatorname{dom}(f). \tag{2.3}$$

Then $f(x_0)$ is called a *local minimum* with *local minimizer* x_0. If (2.3) is fulfilled for all $\delta > 0$, then $f(x_0)$ is a *global minimum* of f with *global minimizer* x_0. Analogously, one can define local (global) *maximum and maximizer* if the reversed inequality in (2.3) holds.

Global Optimality

The following is a summary of some well-known properties of the solutions to (2.2).

Proposition 2.1—see [Hiriart-Urruty and Lemaréchal, 1996]
Let $f : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ be convex and $\mu := \inf_x f(x)$, then

i. The solution set $\Sigma := \{x : f(x) = \mu\}$ is convex.

ii. $x^\star$ is a local minimizer of f if and only if $x^\star$ is a global minimizer.

iii. $f(x^\star) = \mu$ if and only if $0 \in \partial f(x^\star)$.

iv. If f is strictly convex, then Σ is either a singleton or the empty set.

v. If f is strongly convex, then Σ is a singleton.

vi. If $\mathrm{dom}(f)$ is compact, then f has a global minimum. □

Berge's Maximum Theorem

In the following, let $P(\mathcal{H})$ denote the *power set* of $\mathcal{H}$, i.e. the set of all subsets. For $\Theta \subset \mathbb{R}$, a multivalued function (correspondence) $F : \Theta \to P(\mathcal{H})$ is said to be *upper hemicontinuous* at $\theta \in \Theta$, if for all open sets $\mathcal{V} \subset \mathcal{H}$ with $F(\theta) \subset \mathcal{V}$, there exists $\delta > 0$ such that

$$F(y) \subset \mathcal{V} \text{ for all } \; y \in (\theta - \delta, \theta + \delta) \cap \Theta.$$

Further, F is called *lower hemicontinuous* at $\theta \in \Theta$ if for all open sets $\mathcal{V} \subset \mathcal{H}$ with $\mathcal{V} \cap F(x_0) \neq \emptyset$, there exists $\delta > 0$ such that

$$F(y) \cap \mathcal{V} \neq \emptyset \text{ for all } \; y \in (\theta - \delta, \theta + \delta) \cap \Theta.$$

If F is both upper and lower hemicontinuous at θ, then F is simply called *continuous* at θ. Finally, F is said to be *compact-valued (convex-valued)* at $\theta \in \Theta$ if $F(\theta)$ is a compact (convex) set.

The so-called *Berge's Maximum Theorem* (see [Berge, 1963; Sundaram, 1996]) provides conditions for the continuity of a parameter-dependent concave function and the so-called upper hemicontinuity of its set of maximizers.

Proposition 2.2—Berge's Maximum Theorem under Convexity
Let $\mathcal{S} \subset \mathcal{H}$ and $\Theta \subset \mathbb{R}$. Assume $F : \Theta \to P(\mathcal{H})$ is continuous, compact- and convex-valued on Θ and $f : \mathcal{S} \times \Theta \to \mathcal{H}$ is (jointly) continuous. Let

$$f^\star(\theta) := \max\{f(x, \theta) : x \in \mathcal{D}(\theta)\},$$
$$F^\star(\theta) := \{x \in F(\theta) : f(x, \theta) = f^\star(\theta)\}.$$

Then $f^\star$ is continuous and $F^\star$ is upper hemicontinuous on Θ. Furthermore, $F^\star$ is convex-valued on Θ if $f(\cdot, \theta)$ is concave on Θ. □

2.4 Duality Theory

Optimization problems are often not formulated as minimizing just a single function. If the cost function f in the *primal problem* (2.2) is split into several parts, it may be beneficial to consider the so-called corresponding *dual problem*, which can be easier to solve than the primal problem. There are several notions of duality such as *Fenchel, Lagrange* or *Gauge* duality (see [Rockafellar, 1970; Freund, 1987]). Here, the former two concepts are discussed.

Lagrange Duality

The following notational conventions are needed to introduce Lagrange duality. Let $\mathcal{Y}$ be a Hilbert space containing a solid cone $\mathcal{K} \subset \mathcal{Y}$. The cone $\mathcal{K}$ induces an *ordering* in the sense that if $x, y \in \mathcal{K}$ then

$$x \preceq_{\mathcal{K}} y \quad \text{if and only if} \quad x - y \in \mathcal{K}.$$

An operator $G : \Omega \to \mathcal{Y}$ is said to be *convex*, if $\Omega \subset \mathcal{H}$ is convex and

$$G(\alpha x_1 + (1-\alpha)x_2) \preceq_{\mathcal{K}} \alpha G(x_1) + (1-\alpha)G(x_2)$$

for all $x_1, x_2 \in \Omega$ and for all α with $0 \leq \alpha \leq 1$. Let $\mathcal{Z}$ be a Hilbert space, then $A : \mathcal{H} \to \mathcal{Z}$ is called a *linear operator* if

$$A(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 A(x_1) + \alpha_2 A(x_2)$$

for all $x_1, x_2 \in \mathcal{H}$ and for all $\alpha_1, \alpha_2 \in \mathbb{R}$.

A *primal problem in Lagrange duality* is commonly stated as

$$\begin{aligned} p_l \;:=\; \inf_{x \in \mathcal{D}} \quad & f_0(x) \\ \text{s.t.} \quad & G(x) \preceq_{\mathcal{K}} 0 \\ & A(x) = y_0 \end{aligned}$$

where $f_0 : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ and $G : \Omega \to \mathcal{Y}$ are convex, $A : \mathcal{H} \to \mathcal{Z}$ is linear with $y_0 \in \mathcal{Z}$ and $\mathcal{D} := \mathrm{dom}(f) \cap \Omega$. The associated *Lagrange dual problem* is

$$d_l := \max_{\substack{\Lambda \succeq_{\mathcal{K}} 0 \\ \Phi \in \mathcal{Z}}} \inf_{x \in \mathcal{D}} [f_0(x) + \langle \Lambda, G(x) \rangle + \langle \Phi, A(x) - y_0 \rangle].$$

It can be shown that $p_l \geq d_f$ and $p_l - d_l$ is referred to as the *duality gap*.

Proposition 2.3—Slater's condition
Assume that $p_l > -\infty$ and that there exists $x_0 \in \mathrm{ri}(\mathcal{D})$ with

$$G(x_0) \preceq_{\mathcal{K}} 0 \;\text{ and }\; A(x_0) = y_0,$$

then $p_l = d_l$. □

Fenchel Duality

A primal problem in Fenchel duality is usually stated as

$$p_f := \inf_x [f(x) + g(x)], \tag{2.4}$$

and its corresponding *Fenchel dual problem* is given by

$$d_f := -\min_{x^*} [f^*(x^*) + g^*(-x^*)],$$

where $f, g : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ are closed and convex. It holds that $p_f \geq d_f$ and $p_f - d_f$ is called the *duality gap*.

Proposition 2.4—Fenchel's duality theorem
Assume that $p_f > -\infty$ and

$$\text{ri}(\text{dom}(f)) \cap \text{ri}(\text{dom}(g)) \neq \emptyset, \qquad \square$$

then $p_f = d_f$.

In [Magnanti, 1974], it is shown that Fenchel duality is equivalent to Lagrange duality, i.e. Proposition 2.3 can be proven by Proposition 2.4 and vice versa.

2.5 Douglas-Rachford Splitting Algorithm

In numerical convex optimization, many solvers are so-called *interior point methods* (see e.g. [Boyd and Vandenberghe, 2004; Nocedal and Wright, 2006]). An advantage of these methods is that they usually exhibit fast convergence. However, this comes at the cost of having computationally demanding iterates (see .e.g [Toh et al., 1999; Peaucelle et al., 2002]), which grow unfavorably with the number of decision variables.

In order to maintain computability for large optimization problems, one can apply so-called *proximal splitting algorithms* (see [Combettes and Pesquet, 2011; Parikh and Boyd, 2014]). Among these algorithms is the celebrated *Douglas-Rachford Splitting Algorithm* (see [Douglas and Rachford, 1956; Lions and Mercier, 1979; Eckstein and Bertsekas, 1992; Bauschke and Combettes, 2011; Combettes and Pesquet, 2011]), which is discussed here.

Proximal Mapping

The proximal mapping of a closed convex function $f : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ is defined as

$$\text{prox}_{\gamma f}(z) := \underset{x}{\text{argmin}} \left(f(x) + \frac{1}{2\gamma}\|x - z\|^2 \right),$$

where $\gamma > 0$. Note that by Proposition 2.1 it holds that

$$z - x^\star \in \partial \gamma f(x^\star),$$

where $x^\star := \text{prox}_{\gamma f}(z)$.

Douglas-Rachford Iteration

Assume that the optimization problem is split as

$$\min_x \; [f(x) + g(x)], \tag{2.5}$$

where $f, g : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ are close and convex, and that there exist a solution to (2.5). Then the *Douglas-Rachford iteration* is given by

$$x^k = \text{prox}_{\gamma f}(z^{k-1}), \tag{2.6a}$$
$$y^k = \text{prox}_{\gamma g}(2x^k - z^{k-1}), \tag{2.6b}$$
$$z^k = z^{k-1} + \rho(y^k - x^k), \tag{2.6c}$$

where $\gamma > 0$ and $\rho \in (0, 2)$. A special instance of the Douglas-Rachford splitting algorithm is the so-called *alternating direction method of multipliers* (see [Glowinski and Marroco, 1975; Gabay and Mercier, 1976; Boyd et al., 2011]).

Fixed Points

In the following, it is explained why a limit point to (2.6a) is also a minimizer to (2.5). Rewriting (2.6a)–(2.6c) as

$$z^k = F(z^{k-1}) \tag{2.7}$$

with

$$F(z) := z + \rho \text{prox}_{\gamma g}(2\text{prox}_{\gamma f}(z) - z) - \rho \text{prox}_{\gamma f}(z) \tag{2.8}$$

shows that the Douglas-Rachford method is a *fixed-point iteration*. For any of these fixed-points $z^\star \in \mathcal{H}$ with

$$x^\star := \text{prox}_{\gamma f}(z^\star),$$
$$y^\star := \text{prox}_{\gamma g}(2\text{prox}_{\gamma f}(z^\star) - z^\star),$$

it holds that

$$z^\star - x^\star \in \partial\gamma f(x^\star), \quad 2x^\star - z^\star - y^\star \in \partial\gamma g(y^\star),$$

and $x^\star = y^\star$. This is implies that

$$0 \in \partial\gamma(f(x^\star) + g(x^\star)),$$

thus $x^\star$ is a solution to (2.5). Conversely, if a solution $x^\star$ to (2.5) fulfills

$$-w \in \partial\gamma f(x^\star) \text{ and } w \in \partial\gamma g(x^\star),$$

then $z^\star = x^\star - w^\star$ is a fixed point to (2.7).

Under the given assumptions on f and g the operator F is known to be *firmly nonexpansive*, i.e. for all $x, y \in \mathcal{H}$ it holds that

$$\langle F(x) - F(y), x - y\rangle \geq \|F(x) - F(y)\|^2.$$

This is the key property for showing that Douglas-Rachford iteration always converges (see [Douglas and Rachford, 1956; Lions and Mercier, 1979; Eckstein and Bertsekas, 1992]).

Bibliography

Altafini, C. (2016). "Minimal eventually positive realizations of externally positive systems". *Automatica* **68**, pp. 140–147.

Anderson, B. D. O., M. Deistler, L. Farina, and L. Benvenuti (1996). "Nonnegative realization of a linear system with nonnegative impulse response". *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* **43**:2, pp. 134–142.

Antoulas, A. (2005). *Approximation of Large-Scale Dynamical Systems*. SIAM.

Bauschke, H. H. and P. L. Combettes (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer New York.

Beck, C. L., J. Doyle, and K. Glover (1996). "Model reduction of multidimensional and uncertain systems". *IEEE Transactions on Automatic Control* **41**:10, pp. 1466–1477.

Benvenuti, L. and L. Farina (2002). "Positive and compartmental systems". *IEEE Transactions on Automatic Control* **47**:2, pp. 370–373.

Berge, C. (1963). *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity*. Courier Corporation.

Berman, A. and R. Plemmons (1994). *Nonnegative Matrices in the Mathematical Sciences*. SIAM.

Blondel, V. D. and N. Portier (2002). "The presence of a zero in an integer linear recurrent sequence is NP-hard to decide". *Linear Algebra and its Applications* **351**, pp. 91–98.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011). "Distributed optimization and statistical learning via the alternating direction method of multipliers". *Foundations and Trends in Machine Learning* **3**:1, pp. 1–122.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Briat, C. (2011). "Robust stability and stabilization of uncertain linear positive systems via integral linear constraints: L_1-gain and L_∞-gain characterization", pp. 6337–6342.

Brown, R. F. (1980). "Compartmental system analysis: state of the art". *IEEE Transactions on Biomedical Engineering* **BME-27**:1, pp. 1–11.

Combettes, P. L. and J.-C. Pesquet (2011). "Proximal splitting methods in signal processing". In: *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer New York, pp. 185–212.

Datta, B. N. (2004). *Numerical Methods for Linear Control Systems*. Vol. 1. Academic Press.

Douglas, J. and H. H. Rachford (1956). "On the numerical solution of heat conduction problems in two and three space variables". *Transactions of the American Mathematical Society* **82**:2, pp. 421–439.

Dullerud, G. E. and F. Paganini (2013). *A Course in Robust Control Theory: A Convex Approach*. Vol. 36. Texts in Applied Mathematics. Springer-Verlag New York.

Ebihara, Y., D. Peaucelle, and D. Arzelier (2012). "Optimal L1-controller synthesis for positive systems and its robustness properties". In: *2012 American Control Conference (ACC)*, pp. 5992–5997.

Eckstein, J. and D. P. Bertsekas (1992). "On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators". *Mathematical Programming* **55**:1, pp. 293–318.

Farina, L. and S. Rinaldi (2011). *Positive Linear Systems: Theory and Applications*. John Wiley & Sons.

Feng, J., J. Lam, Z. Shu, and Q. Wang (2010). "Internal positivity preserved model reduction". *Int. Journal of Control* **83**:3, pp. 575–584.

Freund, R. M. (1987). "Dual gauge programs, with applications to quadratic programming and the minimum-norm problem". *Mathematical Programming* **38**:1, pp. 47–67.

Gabay, D. and B. Mercier (1976). "A dual algorithm for the solution of nonlinear variational problems via finite element approximation". *Computers and Mathematics with Applications* **2**:1, pp. 17–40.

Glad, T. and L. Ljung (2000). *Control Theory*. CRC press.

Glover, K. (1984). "All optimal hankel-norm approximations of linear multivariable systems and their L_∞-error bounds". *International Journal of Control* **39**:6, pp. 1115–1193.

Glowinski, R. and A. Marroco (1975). "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problémes de dirichlet non linéaires". *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique* **9**, pp. 41–76.

Grussler, C. (2012). *Model reduction of positive systems*. M.Sc.Thesis. Lund University, Department of Automatic Control.

Gugercin, S., A. C. Antoulas, and C. Beattie (2008). "$\mathcal{H}_2$ model reduction for large-scale linear dynamical systems". *SIAM Journal on Matrix Analysis and Applications* **30**:2, pp. 609–638.

Gugercin, S. and A. C. Antoulas (2004). "A survey of model reduction by balanced truncation and some new results". *International Journal of Control* **77**:8, pp. 748–766.

Haddad, W. M., V. Chellaboina, and Q. Hui (2010). *Nonnegative and Compartmental Dynamical Systems*. Princeton University Press.

Hiriart-Urruty, J.-B. and C. Lemaréchal (1996). *Convex analysis and minimization algorithms II*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Horn, R. A. and C. R. Johnson (2012). *Matrix Analysis*. 2nd ed.

Johansson, R. (1993). *System Modeling and Identification*. Prentice hall.

Li, P., J. Lam, Z. Wang, and P. Date (2011). "Positivity-preserving H_∞ model reduction for positive system". *Automatica* **47**:7, pp. 1504–1511.

Li, X., C. Yu, H. Gao, and L. Zhang (2014). "A new approach to H_∞ model reduction for positive systems". *IFAC Proceedings Volumes* **47**:3, pp. 3809–3814.

Lidström, C. and A. Rantzer (2015). "Optimal distributed H-infinity state feedback for systems with symmetric and hurwitz state matrix". arXiv: 1510.00070.

Lions, P.-L. and B. Mercier (1979). "Splitting algorithms for the sum of two nonlinear operators". *SIAM Journal on Numerical Analysis* **16**:6, pp. 964–979.

Liu, Y. and B. D. O. Anderson (1989). "Singular perturbation approximation of balanced systems". *International Journal of Control* **50**:4, pp. 1379–1405.

Luenberger, D. (1979). *Introduction to Dynamic Systems: Theory, Models & Applications*. John Wiley & Sons.

Luenberger, D. G. (1968). *Optimization by Vector Space Methods*. John Wiley & Sons.

Magnanti, T. L. (1974). "Fenchel and lagrange duality are equivalent". *Mathematical Programming* **7**:1, pp. 253–258.

Meyer, C. D. (2000). *Matrix Analysis and Applied Linear Algebra*. Vol. 2. SIAM.

Moore, B. (1981). "Principal component analysis in linear systems: controllability, observability, and model reduction". *IEEE Transactions on Automatic Control* **26**:1, pp. 17–32.

Nocedal, J. and S. Wright (2006). *Numerical Optimization*. Springer Berlin Heidelberg.

Ohta, Y., H. Maeda, and S. Kodama (1984). "Reachability, observability, and realizability of continuous-time positive systems". *SIAM Journal on Control and Optimization* **22**:2, pp. 171–180.

Parikh, N. and S. Boyd (2014). "Proximal algorithms". *Foundations and Trends in Optimization* **1**:3, pp. 127–239.

Peaucelle, D., D. Henrion, Y. Labit, and K. Taitz (2002). "User's guide for SEDUMI INTERFACE 1.04". LAAS-CNRS, Toulouse.

Rantzer, A. (2012). "Optimizing positively dominated systems". In: *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 272–277.

Rantzer, A. (2015). "Scalable control of positive systems". *European Journal of Control* **24**, pp. 72–80.

Reis, T. and E. Virnik (2009). "Positivity preserving balanced truncation for descriptor systems". *SIAM Journal on Control and Optimization* **48**:4, pp. 2600–2619.

Rockafellar, R. T. (1970). *Convex Analysis*. 28. Princeton University Press.

Sandberg, H. and A. Rantzer (2004). "Balanced truncation of linear time-varying systems". *IEEE Transactions on Automatic Control* **49**:2, pp. 217–229.

Shorten, R., F. Wirth, and D. Leith (2006). "A positive systems model of TCP-like congestion control: asymptotic results". *IEEE/ACM Transactions on Networking* **14**:3, pp. 616–629.

Son, N. K. and D. Hinrichsen (1996). "Robust stability of positive continuous time systems". *Numerical Functional Analysis and Optimization* **17**:5-6, pp. 649–659.

Sootla, A. and A. Rantzer (2012). "Scalable positivity preserving model reduction using linear energy functions". In: *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 4285–4290.

Sootla, A. and A. Mauroy (2015). "Properties of eventually positive linear input-output systems". arXiv: 1509.08392.

Sundaram, R. K. (1996). *A First Course in Optimization Theory*. Cambridge University Press.

Tanaka, T. and C. Langbort (2011). “The bounded real lemma for internally positive systems and h-infinity structured static state feedback”. *IEEE Transactions on Automatic Control* **56**:9, pp. 2218–2223.

Toh, K.-C., M. J. Todd, and R. H. Tütüncü (1999). “SDPT3 – a MATLAB software package for semidefinite programming, version 1.3”. *Optimization Methods and Software* **11**:1-4, pp. 545–581.

Trefethen, L. N. and D. Bau III (1997). *Numerical Linear Algebra*. SIAM.

Zhou, K., J. C. Doyle, K. Glover, et al. (1996). *Robust and Optimal Control*. Vol. 40. Prentice Hall.

Paper I

A Symmetry Approach for Balanced Truncation of Positive Linear Systems

Christian Grussler Tobias Damm

Abstract

We consider model order reduction of positive linear systems and show how a symmetry characterization can be used in order to preserve positivity in balanced truncation. The reduced model has the additional feature of being symmetric.

1. Introduction

Mathematical modeling of biological, chemical and physical systems often leads to complex high-dimensional models, which are hard to analyze and simulate. Approximating high-order models by ones of reduced order is the central goal of model order reduction in control and has received considerable attention e.g. in [Moore, 1981; Fernando and Nicholson, 1983; Safonov and Chiang, 1989; Sandberg and Rantzer, 2004; Gugercin et al., 2008]. Here we consider linear time-invariant systems

$$G: \begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t), \end{cases} \tag{1}$$

with state vector $x \in \mathbb{R}^n$, input $u \in \mathbb{R}^m$ and output $y \in \mathbb{R}^p$, for small m, p and large dimension n. Our goal is to approximate (1) by a system of the same structure with the same m and p, but with smaller order $r < n$. For this purpose different methods have been developed, the most popular of which are based on linear subspace projection, such as balanced truncation [Moore, 1981] or Krylov subspace methods [Antoulas, 2005], [Gugercin et al., 2008].

In practice, one often deals with so called *(internally) positive* systems (see [Farina and Rinaldi, 2011]) whose output and state variables are nonnegative, whenever the input and initial states are confined to be nonnegative. Such systems occur, e.g. within the discretization of partial differential equations [Saad, 2003], transport models or compartmental systems [Luenberger, 1979]. It is desirable that the reduced system also is positive. Unfortunately, positive systems are defined on cones instead of linear subspaces [Ohta et al., 1984; Berman et al., 1989; Farina and Rinaldi, 2011], and therefore methods based on linear subspace projection typically do not preserve positivity. As a consequence, new methods have been developed in [Reis and Virnik, 2009; Feng et al., 2010; Li et al., 2011]. However, with rather conservative results regarding the H_∞-error and the computational effort.

In this paper we present several new results related to positivity preserving model order reduction. First, we show that balanced truncation to order one always gives a positive approximation. Then, for single-input single-output (SISO) systems, a symmetry condition for computing positive realizations is derived. Since any balanced realization of a SISO-system can be shown to be sign-symmetric with respect to the entries in A, B and C (see [Moore, 1981; Fernando and Nicholson, 1983]), we can describe a procedure to compute a positive reduced order model of a SISO-system, by just comparing signs in the sign-symmetric realization. In the worst case, this procedure only allows for the scalar approximation mentioned above, but

in practical examples, it also yields positive approximations of higher order with acceptable errors. These approximations have the additional property of being symmetric, which is desirable for instance, in case of linear networks with reaction-diffusion structure [Ishizaki et al., 2010].

2. Preliminaries

Throughout this paper we use the following notation for real matrices and vectors $X = (x_{ij})$. We say that X is *positive*, $X \gg 0$, if all its entries are positive ($x_{ij} > 0$ for all i, j). It is called *nonnegative*, $X \geqq 0$, if all entries are nonnegative ($x_{ij} \geq 0$ for all i, j). By $|X| = (|x_{ij}|) \geqq 0$ we denote the entrywise absolute value of X.

A square matrix X is *reducible*, if there exists a permutation matrix $P = [P_1, P_2]$ so that $P_2^T X P_1 = 0$. Otherwise, it is *irreducible* (see [Berman and Plemmons, 1994]). By $\sigma(X)$ we denote the spectrum of X.

If X is square and symmetric, then we write $X > 0$, or $X \geq 0$ if X is positive definite, or nonnegative definite, i.e. $\sigma(X) \subset [0, \infty[$. We also use these notations to describe the relation between two arbitrary elements, e.g. $A \geq B$ is defined by $A - B \geq 0$. A real vector valued function $u(t) \in \mathbb{R}^n$ is called *nonnegative* if and only if $u(t) \geqq 0$ for all t.

Proposition 1—Perron-Frobenius [Luenberger, 1979; Meyer, 2000]
If $A \geqq 0$ is irreducible, then there exist a real $\lambda_0 > 0$ and a vector $x_0 \gg 0$ such that

1. $Ax_0 = \lambda_0 x_0$.
2. $\lambda_0 \geq |\lambda|, \ \forall \lambda \in \sigma(A)$.
3. The algebraic multiplicity of λ_0 is one.

If $A \geqq 0$ is reducible, then there exists a real $\lambda_0 \geq 0$ and a vector $x_0 \geq 0$ such that

1. $Ax_0 = \lambda_0 x_0$.
2. $\lambda_0 \geq |\lambda|, \ \forall \lambda \in \sigma(A)$.

Moreover, there exists a permutation matrix π, such that

$$\pi^T A \pi = \begin{pmatrix} B_1 & * & * \\ & \ddots & * \\ & & B_k \end{pmatrix},$$

where each B_i is irreducible or equal to zero. In particular, if A is diagonalizable and λ_0 has multiplicity m_0, then A has m_0 linearly independent nonnegative eigenvectors. □

Next, let us define positive systems (see [Farina and Rinaldi, 2011]).

Definition 1—External Positivity
A linear system (A, B, C, D) as in (1) is called *externally positive* if and only if its output, corresponding to a zero initial state, is nonnegative for every nonnegative input. □

Definition 2—Internal Positivity
A linear system (1) is called (internally) positive if and only if its state and output are nonnegative for every nonnegative input and every nonnegative initial state. □

To characterize a continuous positive system, one needs the notion of a Metzler matrix (or Z-matrix) [Berman and Plemmons, 1994]. A matrix $A \in \mathbb{R}^{n\times n}$ is Metzler if there exists an $\alpha \in \mathbb{R}$ such that $A + \alpha I_n \geqq 0$, where I_n is the $n \times n$ identity matrix. If A is Metzler then $e^{At} \geqq 0$ for all $t \geq 0$ [Luenberger, 1979].

Proposition 2—[Farina and Rinaldi, 2011]
A continuous linear system (1) is positive if and only if A is Metzler and $B, C, D \geqq 0$. □

3. Balanced Truncation to Order One

In the following we consider asymptotically stable positive systems (A, B, C, D) as in (1). We assume the reader to be familiar with the concept of *standard balanced truncation* (see e.g. [Antoulas, 2005; Reis and Virnik, 2009]). In general, balanced truncation does not return a positive system – unless the system is reduced to the order $r = 1$.

Theorem 1—Positive Order-1 Balanced Truncation
If (A_1, B_1, C_1, D_1) is the reduced system of order 1, obtained by standard balanced truncation of (A, B, C, D), then it has a positive, asymptotically stable realization $(A_1, |B_1|, |C_1|, D_1)$ of order 1. □

Proof Let P and Q be the Gramians of (A, B, C, D), implicitly given by

$$AP + PA^T = -BB^T, \quad A^TQ + QA = -C^TC, \tag{2}$$

or in their explicit form by

$$P = \int_0^\infty e^{At}BB^Te^{A^Tt}dt, \quad Q = \int_0^\infty e^{A^Tt}C^TCe^{At}dt. \tag{3}$$

Obviously, P and Q are nonnegative and thus PQ, too. Balancing the system via a state-space transformation $x = T\xi$ yields

$$T^{-1}PQT = \text{blkdiag}\left(\sigma_1^2 I_{k_1}, \ldots, \sigma_N^2 I_{k_N}, 0\right)$$

with Hankel singular values $\sigma_1 > \cdots > \sigma_N$ and corresponding multiplicities $k_1, \ldots, k_N$ (see [Zhou et al., 1996]). Hence, the columns of T are eigenvectors of PQ and by Theorem 1 there exists a nonnegative right-eigenvector v_1 to the largest eigenvalue σ_1, i.e.

$$PQv_1 = \sigma_1 v_1 \text{ with } T = \left(v_1, \ldots, v_n\right).$$

Analogously, there exists a nonnegative left-eigenvector w_1 such that $T^{-1} = \left(w_1, \ldots, w_n\right)^T$. If $k_1 = 1$, the asymptotic stability of the reduced system of order one leads to

$$A_1 = w_1^T A v_1 < 0, \quad B_1 = w_1^T B \geqq 0, \quad C_1 = C v_1 \geqq 0, \quad D_1 = D \geqq 0.$$

If $k_1 > 1$, it could happen that $A_1 = 0$. But since the reduced system of order k_1 (belonging to all σ_1) is asymptotically stable, there must exist at least one asymptotically stable first order approximation. By Theorem 1 we conclude the reducibility of PQ and the positivity of each first order approximation. In both cases Theorem 2 concludes the proof.

Balanced truncation can also be performed by using $-v_1$ and $-w_1$. In this case we substitute B_1 and C_1 by their elementwise absolute values. □

In general, Theorem 1 does not transfer to *singular perturbation balanced truncation* (see [Reis and Virnik, 2009]). Further, Theorem 1 gives a necessary condition on the positivity, independent of its realization. By numerical experiments, we can observe that this is a strong condition. Many of the non-positive systems fail at this point.

4. The Positive Realization Problem

From the proof of Theorem 1 we can deduce that even in case of an approximation to order one, balanced truncation does not necessarily return a positive realization. However, it is straightforward to see, that every first order externally positive system has a positive realization of the same dimension. The same is true for second order SISO-systems (see [Ohta et al., 1984]). But higher-order externally positive systems do not necessarily admit an internally positive realization of the same dimension – even if they possess only real poles (see [Ohta et al., 1984] again). Knowing that balanced truncation always results in a minimal system, the positive realization problem and its connection to balanced realizations becomes the major obstacle beside the actual positivity preservation.

We call a linear system *quasi-symmetric* if $A = A^T$ and $C = kB^T$ for some $k > 0$. If $k = 1$ the system is said to be *symmetric* (see [Liu et al., 1998]).

THEOREM 2—POSITIVITY OF SYMMETRIC SYSTEMS
Every quasi-symmetric SISO system possesses a symmetric positive minimal realization, which can be computed by Arnoldi's (or Lanczos') algorithm□

Proof Let (A, B, C, D) be a quasi-symmetric system with Gramians P and Q. Then from (3) it follows that $Q = k^2P$. Diagonalization of kP gives

$$\begin{aligned} kP &= T^T \Sigma T \\ PQ &= k^2P^2 = T^T \Sigma^2 T = \tilde{T}^{-1} PQ \tilde{T} \end{aligned}$$

with $\tilde{T} = \frac{1}{\sqrt{k}} T$. Obviously, $\tilde{T}$ is a balancing transformation matrix and the balanced system is given by

$$(T^{-1}A\tilde{T}, \tilde{T}^{-1}B, C\tilde{T}) = (T^T AT, \sqrt{k}(BT)^T, \sqrt{k}BT).$$

Thus, it is always possible to find a symmetric minimal realization of a quasi-symmetric system. Arnoldi's algorithm (see [Trefethen and Bau III, 1997; Antoulas, 2005]) yields a unitary transformation matrix

$$V = \begin{pmatrix} \dfrac{B}{\|B\|_2} & * \end{pmatrix},$$

such that $V^T AV$ is upper Hessenberg with positive elements on its lower diagonal. If $A = A^T$ and $C = B^T$, this means that $V^T AV$ is Metzler and

$$CV = (B^T V) = \begin{pmatrix} \|B\|_2 & 0 & \cdots & 0 \end{pmatrix}.$$

Positivity now follows from Proposition 2. □

5. Symmetric Balanced Truncation

If balanced truncation of a SISO system results in a symmetric reduced model, then (by Theorem 2) we are able to compute its positive realization. To this end we recall the following important result of balanced SISO-systems (see also [Moore, 1981; Fernando and Nicholson, 1983]).

PROPOSITION 3
Let $G(s)$ be the transfer function of an arbitrary SISO-system. Then there exists a balanced realization (A, B, C, D) of $G(s)$, such that (A, B, C, D) is *sign symmetric*, i.e. $|A| = |A^T|$ and $|B| = |C^T|$. □

Proof Let (A, B, C, D) have simple Hankel singular values $\{\sigma_1, \ldots, \sigma_n\}$. By definition of a balanced system, its Lyapunov equations can be written as

$$\begin{aligned} A\Sigma + \Sigma A^T = -BB^T &\Leftrightarrow a_{ij}\sigma_j + \sigma_i a_{ji} = -b_i b_j, \\ A^T\Sigma + \Sigma A = -C^T C &\Leftrightarrow a_{ij}\sigma_i + \sigma_j a_{ji} = -c_i c_j, \end{aligned} \tag{4}$$

for $i, j = 1, \ldots, n$ and $\Sigma := \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$. In particular it holds for $i = j$:

$$2a_{ii}\sigma_i = -b_i^2 = -c_i^2 \;\Rightarrow\; b_i = \pm c_i. \tag{5}$$

If $i \neq j$ we can deduce from (4) and (5)

$$\begin{pmatrix} \sigma_j & \sigma_i \\ \sigma_i & \sigma_j \end{pmatrix} \begin{pmatrix} a_{ij} \\ a_{ji} \end{pmatrix} = \begin{pmatrix} b_i b_j \\ c_i c_j \end{pmatrix} = \begin{pmatrix} b_i b_j \\ \pm b_i b_j \end{pmatrix}.$$

Solving for $\begin{pmatrix} a_{ij} & a_{ji} \end{pmatrix}^T$ yields

$$\begin{pmatrix} a_{ij} \\ a_{ji} \end{pmatrix} = \frac{b_i b_j}{\sigma_j^2 - \sigma_i^2} \begin{pmatrix} \sigma_j \mp \sigma_i \\ \pm(\sigma_j \mp \sigma_i) \end{pmatrix}$$

and hence $a_{ij} = \pm a_{ji}$.

In case of multiple Hankel singular values we can assume w.l.o.g. $\Sigma = \operatorname{diag}(\sigma_1 I_{k_1}, \sigma_2, \ldots, \sigma_n)$ for $k_1 > 1$. By partitioning $A = \begin{pmatrix} A_1 & * \\ * & * \end{pmatrix}$ and $B = \begin{pmatrix} B_1 \\ * \end{pmatrix}$ correspondingly to $\sigma_1 I_{k_1}$, we can write $\sigma_1(A_1 + A_1^T) = B_1 B_1^T$. Diagonalizing

$$B_1 B_1^T = U^T \operatorname{diag}(\lambda, 0)\, U \quad \text{with } \lambda > 0$$

gives

$$\sigma_1(UA_1U^T + UA_1^T U^T) = UB_1B_1^T U^T = \operatorname{diag}(\lambda, 0)$$

and it follows for $\tilde{A} := UA_1U^T$, that $\tilde{a}_{ij} = -\tilde{a}_{ji}$. By $T := \operatorname{diag}(U, I)$ we define a balanced sign symmetric realization

$$(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) := (TAT^T, TB, CT^T, D).$$

□

Note that $b_i b_j = -c_i c_j$ if and only if $a_{ij} = -a_{ji}$. Hence, balanced truncation returns a k-th order symmetric approximation as long as $c_i = b_i$ for all $i = 1, \ldots, k$. From Theorem 1 we know that $k \geq 1$.

Theorems 1–3 provide the basis of the following *Symmetric Balanced Truncation Algorithm* (SBT).

ALGORITHM 1—SYMMETRIC BALANCED TRUNCATION ALGORITHM
Let G be a given linear system as in (1), then:

1. Compute a balanced realization (A_b, B_b, C_b, D_b).
2. Compare the entries of B_b and C_b in order to identify the smallest k, where $c_k \neq b_k$.
3. Perform the truncation of (A_b, B_b, C_b) to obtain a reduced symmetric system G_r of the order $r < k$.
4. Obtain a positive realization of G_r with the help of Lanczos Algorithm. □

Due to the symmetry constraint the reduced models possess only real eigenvalues. Thus, we can expect to approximate a system well, only if its dominating poles are real. Such systems often occur in the context of sparse large-scale systems, i.e. $n \gg 1000$. For such high dimensions balanced truncation may not be applicable and therefore a pre-approximation is required. In [Gugercin et al., 2008] it is shown empirically, that the Iterative Rational Krylov Algorithm (IRKA) gives comparable good results as balanced truncation. The same can be said about the size of the symmetric part after balancing a reduced model, which is obtained by IRKA. This makes IRKA an advisable pre-approximator for our method.

The applicability to large-scale systems and the general independence of a specific state-space representation can be considered the main advantages of SBT. This method is often preferable to those presented in [Reis and Virnik, 2009; Feng et al., 2010; Li et al., 2011], for the following reasons. The methods in [Feng et al., 2010] and [Li et al., 2011] have the common goal of satisfying the Bounded Real Lemma [Zhou et al., 1996] for the error-system, i.e. between the original and the reduce model. Both are using an iterative linearization approach and consequently do not have a convergence guarantee. The linear matrix inequalities (LMIs), which need to be solved, are usually very expensive to solve (see [Peaucelle et al., 2002]).

The method in [Reis and Virnik, 2009] is based on LMIs, consisting only of $2n$ variables. In the following we refer to this method as Generalized Balanced Truncation (GBT). It generalizes the idea of balanced truncation by using diagonal solutions $\tilde{P} \geq 0$ and $\tilde{Q} \geq 0$ of the LMIs

$$A\tilde{P} + \tilde{P}A^T \leq -BB^T, \quad A^T\tilde{Q} + \tilde{Q}A \leq -C^TC\,. \tag{6}$$

Such solutions exist, since A is Metzler (see [Berman and Plemmons, 1994]). Balanced truncation based on the generalized Gramians $\tilde{P}$ and $\tilde{Q}$ preserves the error formula [Beck et al., 1996], but the bound is more conservative, as the following proposition shows.

Proposition 4
Let (A, B, C, D) be a minimal system, $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of PQ given by (2), and $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_n$ be the eigenvalues of $\tilde{P}\tilde{Q}$ as defined in (2). Then $\lambda_i \leq \tilde{\lambda}_i$ for all $i = 1, \ldots, n$. □

Proof By subtracting equations (2) from the inequalities (6) it follows by the stability of the system [Zhou et al., 1996], that $\tilde{P} - P \geq 0$, or equivalently that $\tilde{P} \geq P > 0$. In the same way we obtain $\tilde{Q} \geq Q > 0$. It holds, that

$$\sigma(PQ) = \sigma(P^{-\frac{1}{2}}(PQ)P^{\frac{1}{2}}) = \sigma(RQR)$$

where $R = P^{\frac{1}{2}}$. Analogously, $\sigma(\tilde{P}\tilde{Q}) = \sigma(\tilde{R}\tilde{Q}\tilde{R})$ with $\tilde{R} \geq R > 0$. Since

$$\begin{aligned}\tilde{R}\tilde{Q}\tilde{R} - RQR &= \tilde{R}\tilde{Q}\tilde{R} - \tilde{R}Q\tilde{R} + \tilde{R}Q\tilde{R} - RQR = \\ &= R(\tilde{Q} - Q)R + Q^{-\frac{1}{2}}\left((Q^{\frac{1}{2}}\tilde{R}Q^{\frac{1}{2}})^2 - (Q^{\frac{1}{2}}RQ^{\frac{1}{2}})^2\right)Q^{-\frac{1}{2}},\end{aligned}$$

it follows by $\tilde{Q} \geq Q$, as well as $Q^{\frac{1}{2}}\tilde{R}Q^{\frac{1}{2}} \geq Q^{\frac{1}{2}}RQ^{\frac{1}{2}}$, that $\tilde{R}\tilde{Q}\tilde{R} \geq RQR$. The inequalities for the eigenvalues now follow from the Courant-Fischer theorem [Lancaster and Tismenetsky, 1985]. □

From a geometric point of view this is clear, since balancing with respect to the generalized Gramians does not project the system onto the controllable and observable subspace. In particular, standard balanced truncation with diagonal Gramians is essentially a permutation of the states followed by truncation.

In contrast, SBT inherits the good H_∞-error behavior of balanced truncation. For that reason even a small symmetric part often yields good results. In Section 6 we compare SBT and GBT numerically. Since GBT can also be used for singular perturbation balanced truncation, we always present the error of the better one.

6. Examples

We discuss some properties of SBT and compare it with the method in [Reis and Virnik, 2009].

6.1 Water Reservoirs

We start with the same water reservoir example as in [Reis and Virnik, 2009]. As schematically shown in Fig. 1, we consider a system of n connected water reservoirs. All reservoirs $R_1, \ldots, R_n$ are assumed to be located on the same level. Base area and fill level of reservoir R_i are denoted by a_i and h_i, respectively. Further, R_i and R_j are connected by a pipe of diameter

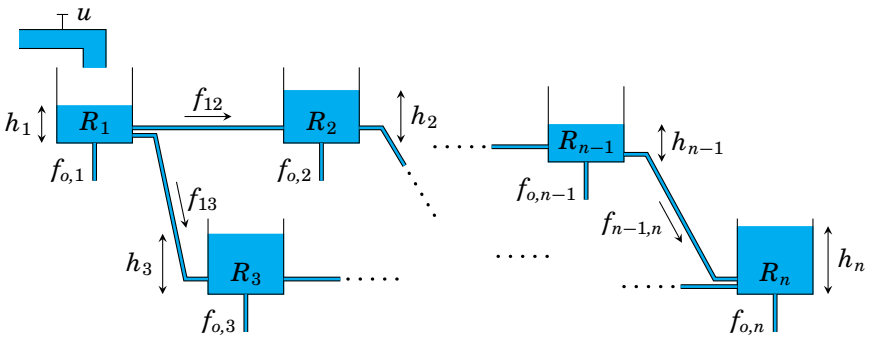


Figure 1. System of n water reservoirs.

$d_{ij} = d_{ji} \geq 0$, resulting in a flow f_{ij} from R_i to R_j, where f_{ij} is assumed to be linearly dependent on the pressure difference at both ends. The external inflow to reservoir R_1 serves as the single input of the system. The output is the sum of all outflows $f_{o,i}$ of R_i through a pipe with diameter $d_{o,i}$. According to Pascal's law the system flows are described by

$$f_{ij}(t) = d_{ij}^2 \cdot k \cdot (h_i(t) - h_j(t)),$$
$$f_{o,i}(t) = d_{o,i}^2 \cdot k \cdot (h_i(t) - h_j(t)),$$

where k is a constant representing gravity as well as viscosity and density of the medium. Thus, the fill level h_i of R_i is subject to the differential equation

$$\dot{h}_i = \frac{k}{a_i}\Big(- d_{o,i}^2 h_i(t) + \sum_{j=1}^{n} d_{ij}^2 (h_j(t) - h_i(t))\Big) + \frac{1}{a_i}\delta_{1i} u(t),$$

where $\delta_{1i} = 1$ if $i = 1$ and zero otherwise. Writing these equations as a linear state-space system results in a SISO-system (A, B, C, D) given by $B = \left(\frac{1}{a_1}, 0, \ldots, 0\right)^T$, $C = k\left(d_{o,1}^2 \quad \cdots \quad d_{o,n}^2\right)$ and a symmetric A with entries

$$a_{ij} := \frac{k}{a_i}\begin{cases} -d_{o,i}^2 - \sum_{m=1}^{n} d_{im}^2, & i = j \\ d_{ij}^2 & i \neq j, \end{cases} \quad \text{with} \quad d_{ii} := 0.$$

In [Reis and Virnik, 2009] the system is supposed to consist of two substructures of five reservoirs each. In both substructures each reservoir is connected to every other by a pipe of diameter 1, i.e $d_{ij} = 1$ for $i \neq j$ and $i, j = 1, \ldots, 5$ and $i, j = 6, \ldots, 10$, respectively. The connection of the substructures is established by a pipe of diameter $d_{1,10} = d_{10,1} = 0.2$, between

reservoir 1 and 10. For simplicity, $a_i = 1$ and $k = 1$. One can show that the transfer function is just $G(s) = \frac{1}{s+1}$.

Applying SBT to this system yields an exact realization of G. In contrast, since GBT does not return a minimal realization, we get $\tilde{G}(s) = \frac{3.039}{s+3.039}$, with a relative H_∞-error of 0.5014.

Now we modify the system to get a minimal example with unsymmetric A. First, we set $d_{o,i} = 0.01 \cdot i$ to get minimality. Further assume that the first substructure admits a flow from R_1 to R_j, but not vice versa, i.e. $d_{j1} = 0$ for $j = 2, \ldots, \frac{n}{2}$. For 50 water tanks per substructure, SBT gives a symmetric model of order 2

$$A_2 = \begin{pmatrix} -0.1305 & 0.0914 \\ 0.0914 & -0.2676 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0.0457 \\ 0 \end{pmatrix},$$
$$C_2 = \begin{pmatrix} 0.0457 & 0 \end{pmatrix}, \quad D_2 = 0,$$

with error 0.0032. About the same error is achieved by GBT only for reduction order 91. We conclude that SBT performs fairly well even for systems with non-symmetric A-matrix.

6.2 Heat Equation

Consider the two-dimensional heat equation

$$\dot{T} = \triangle T = \frac{\partial^2}{\partial x^2} T + \frac{\partial^2}{\partial y^2} T \tag{7}$$

on the unit square. The Dirichlet boundary conditions on the four edges are interpreted as inputs. Using a finite difference discretization on a uniform grid of step size $h = \frac{1}{N+1}$ sketched in Fig. 2 we get the relations

$$\triangle T_{ij} \approx -\frac{1}{h^2}(4T_{ij} - T_{i+1,j} - T_{i,j+1} - T_{j-1,j} - T_{i,j-1}),$$

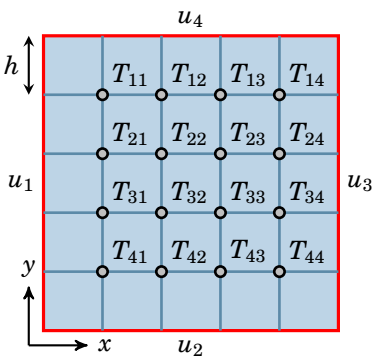


Figure 2. Discretized heat equation on the unit square.

for the temperatures at the inner grid points. Let A denote the $N^2 \times N^2$ Poisson-matrix and $B := [b_{ij}] \in \mathbb{R}^{N^2 \times 4}$, where $b_{ij} = 0$ except for the following cases:

$$\begin{aligned} b_{i1} &:= 1, && \text{for } i = 1, 2, \ldots, N \\ b_{i2} &:= 1, && \text{for } i = N, 2N, \ldots, N^2 \\ b_{i3} &:= 1, && \text{for } i = N(N-1)+1, N(N-1)+2, \ldots, N^2 \\ b_{i4} &:= 1, && \text{for } i = 1, N+1, \ldots, N(N-1)+1 \end{aligned}$$

This gives the discretized system

$$\dot{x} = \frac{1}{h^2} Ax + \frac{1}{h^2} Bu \quad \text{with } u \in \mathbb{R}^4 \text{ and } x \in \mathbb{R}^{N^2}. \tag{8}$$

As the output we take the average temperature, i.e.

$$y = \frac{1}{N^2} Cx, \text{ with } C := \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{1 \times N^2}.$$

For small h the system will be very large.

Starting the comparison between SBT and GBT with a SISO-system, i.e. $u_2 = u_3 = u_4 = 0$ and $N = 10$, yields for SBT a realization of order 15 without any error. In contrast, GBT gives a relative H_∞-error of $3.9087 \cdot 10^{-5}$ by just reducing one state. Moreover, if GBT halves the order it has nearly the same error as balanced truncation to order 1.

For $N = 50$, we get a system of order 2500, for which it takes GBT hours to calculate a reduced model, due to the high complexity of conventional

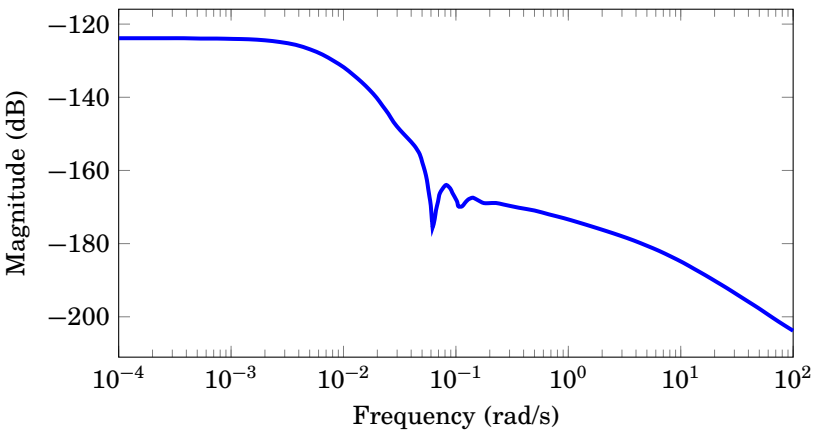


Figure 3. Bode plot: Error system of the heat equation with 2500 states (N = 50) and SBT of order 15.

LMI-solvers [Peaucelle et al., 2002]. In case of a large-scale system we apply IRKA to decrease the system to a order lower than 1000, followed by the usual symmetry argument. These computations consume less than half an hour and return a 15-th order model. The Bode diagram of the error system, as shown in Fig. 3, indicates that the reduction error is zero up to machine precision.

Applying balanced truncation to the full MISO-system results in a reduced system (A_r, B_r, C_r), with

$$A_r = A_r^T \quad \text{and} \quad C_r = B_r^1 = \cdots = B_r^4,$$

where $B_r^1, \ldots, B_r^4$ denote the columns of B_r. In case of $N = 10$ SBT returns as in the SISO-case a reduced system of order 15 with zero error. However, the error of reducing just one state by GBT increases to 0.0070.

7. Conclusion

We have presented a positivity preserving model reduction method for SISO-systems based on the sign-symmetry of balanced SISO-systems. It always yields at least some positive approximation since the reduced model of order 1 is guaranteed to be positive. Application of this idea to MIMO systems provides a necessary condition for positivity, which is preferable over a consideration of the impulse response [Farina and Rinaldi, 2011]. Furthermore, the reduction method works independently of a positive state-space realization. Hence, large-scale systems can be treated by pre-approximations with methods such as the Iterative Rational Krylov algorithm [Gugercin et al., 2008]. Besides, the method preserves and provides symmetry in the A-matrix.

References

Antoulas, A. (2005). *Approximation of Large-Scale Dynamical Systems*. SIAM.

Beck, C. L., J. Doyle, and K. Glover (1996). "Model reduction of multidimensional and uncertain systems". *IEEE Transactions on Automatic Control* **41**:10, pp. 1466–1477.

Berman, A. and R. Plemmons (1994). *Nonnegative Matrices in the Mathematical Sciences*. SIAM.

Berman, A., M. Neumann, and R. J. Stern (1989). *Nonnegative Matrices in Dynamic Systems*. Vol. 3. Wiley & Sons.

Farina, L. and S. Rinaldi (2011). *Positive Linear Systems: Theory and Applications*. John Wiley & Sons.

Feng, J., J. Lam, Z. Shu, and Q. Wang (2010). "Internal positivity preserved model reduction". *Int. Journal of Control* **83**:3, pp. 575–584.

Fernando, K. and H. Nicholson (1983). "On the structure of balanced and other principal representations of siso systems". *IEEE Transactions on Automatic Control* **28**:2, pp. 228–231.

Gugercin, S., A. C. Antoulas, and C. Beattie (2008). "$\mathcal{H}_2$ model reduction for large-scale linear dynamical systems". *SIAM Journal on Matrix Analysis and Applications* **30**:2, pp. 609–638.

Ishizaki, T., K. Kashima, and J. i. Imura (2010). "Extraction of 1-dimensional reaction-diffusion structure in siso linear dynamical networks". In: *49th IEEE Conference on Decision and Control (CDC)*, pp. 5350–5355.

Lancaster, P. and M. Tismenetsky (1985). *The Theory of Matrices: With Applications*. Elsevier.

Li, P., J. Lam, Z. Wang, and P. Date (2011). "Positivity-preserving H_∞ model reduction for positive system". *Automatica* **47**:7, pp. 1504–1511.

Liu, W., V. Sreeram, and K. Teo (1998). "Model reduction for state-space symmetric systems". *Systems & Control Letters* **34**:4, pp. 209–215.

Luenberger, D. (1979). *Introduction to Dynamic Systems: Theory, Models & Applications*. John Wiley & Sons.

Meyer, C. D. (2000). *Matrix Analysis and Applied Linear Algebra*. Vol. 2. SIAM.

Moore, B. (1981). "Principal component analysis in linear systems: controllability, observability, and model reduction". *IEEE Transactions on Automatic Control* **26**:1, pp. 17–32.

Ohta, Y., H. Maeda, and S. Kodama (1984). "Reachability, observability, and realizability of continuous-time positive systems". *SIAM Journal on Control and Optimization* **22**:2, pp. 171–180.

Peaucelle, D., D. Henrion, Y. Labit, and K. Taitz (2002). "User's guide for SEDUMI INTERFACE 1.04". LAAS-CNRS, Toulouse.

Reis, T. and E. Virnik (2009). "Positivity preserving model reduction". In: *Positive Systems: Proceedings of the third Multidisciplinary International Symposium on Positive Systems: Theory and Applications (POSTA)*. Springer Berlin Heidelberg, pp. 131–139.

Saad, Y. (2003). *Iterative Methods for Sparse Linear Systems*. Second. SIAM.

Safonov, M. G. and R. Y. Chiang (1989). "A schur method for balanced-truncation model reduction". *IEEE Transactions on Automatic Control* **34**:7, pp. 729–733.

Sandberg, H. and A. Rantzer (2004). "Balanced truncation of linear time-varying systems". *IEEE Transactions on Automatic Control* **49**:2, pp. 217–229.

Trefethen, L. N. and D. Bau III (1997). *Numerical Linear Algebra*. SIAM.

Zhou, K., J. C. Doyle, K. Glover, et al. (1996). *Robust and Optimal Control*. Vol. 40. Prentice Hall.

Acknowledgments

This work was part of the author's Diploma Thesis at TU Kaiserslautern and Lund University. It was supported by the Swedish Research Council through the LCCC Linnaeus Center.

Paper II

Modified Balanced Truncation Preserving Ellipsoidal Cone-Invariance

Christian Grussler Anders Rantzer

Abstract

We consider model order reduction of stable linear systems which leave ellipsoidal cones invariant. We show how balanced truncation can be modified to preserve cone-invariance. Additionally, this implies a method to perform external positivity preserving model reduction for a large class of systems.

1. Introduction

Cone-invariance of linear time-invariant systems is a common feature, which is appearing nowadays very frequently in the literature. This is due to an increased interest in systems with compartmental structure as they can be found in bio-medicine, economics, data networks and many more application areas (see [Luenberger, 1979; Brown, 1980; Shorten et al., 2006; Farina and Rinaldi, 2011]). For example, consider the linear time-invariant system

$$G: \begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t), \end{cases} \tag{1}$$

with state vector $x \in \mathbb{R}^n$, input $u \in \mathbb{R}^m$ and output $y \in \mathbb{R}^k$. Here x could stand for the temperature in n rooms within a building, influenced by the temperature of m radiators u. The temperature in k sensor locations, e.g. floors, is then represented by y. Consequently, x and y are confined to be nonnegative, whenever u is nonnegative. In the literature, such systems are referred to as being internally positive and as being externally positive if the system is positive from input to output (see Section 2).

Naturally, these systems often tend to be of large dimension n and need to be approximated with the help of model order reduction. Unfortunately, conventional model reduction methods (see e.g. [Moore, 1981; Glover, 1984; Gugercin and Antoulas, 2004; Antoulas, 2005; Gugercin et al., 2008]) do not preserve external positivity.

However, working with an approximation that is violating basic physical constraints by allowing for instance negative concentrations of chemical substance always leaves the question of how conclusive results on this basis are. Recently developed methods have tackled this problem by preserving internal positivity (see [Reis and Virnik, 2009; Feng et al., 2010; Li et al., 2011; Grussler and Damm, 2012; Sootla and Rantzer, 2012]), i.e. the invariance with respect to (w.r.t.) the nonnegative orthant.

The main goal of this work is to present a variant of balanced truncation, which guarantees to preserve invariance w.r.t. an ellipsoidal cone (see Section 2 and 3). An immediate consequence of this result is the preservation of external positivity under the assumption of ellipsoidal cone-invariance (see Section 4). Unlike internal positivity, our definition has the advantage of being computationally tractable and independent of a particular state-space realization (see Sections 2 and 4). In Section 6 we will see that ellipsoidal cone-invariance is often implied by internal positivity. Moreover, numerical experiments indicate that the error-difference between balanced truncation and our method appears to be fairly small (see Section 6).

2. Preliminaries

The following notations for real matrices and vectors $X = (x_{ij})$ are used throughout this paper. We say that $X \in \mathbb{R}^{m\times n}_{\geq 0}$ is *nonnegative*, if all entries are nonnegative ($x_{ij} \geq 0$ for all i, j). By $|X| = (|x_{ij}|)$ we denote the entry-wise absolute value of X and by x_i its i-th column, if not further specified.

If $X = X^T$, then we write $X \succ 0$, or $X \succeq 0$ if X is positive definite, or semi-definite, i.e. the set of eigenvalues of X, $\sigma(X) \subset [0, \infty[$. We also use these notations to describe the relation between two matrices, e.g. $A \succeq B$ defines $A - B \succeq 0$. A real vector valued function $u(t) \in \mathbb{R}^m$ is called *nonnegative* if and only if $u(t) \in \mathbb{R}^m_{\geq 0}$ for all $t \geq 0$. The inertia (p, z, n) of X is defined by the number of eigenvalues of X with positive, zero and negative real-parts, respectively counting multiplicities. Next, let us define cone-invariance.

Definition 1—Invariant cone
Let $\mathcal{K} \subset \mathbb{R}^n$ be a cone and $A \in \mathbb{R}^{n\times n}$. $\mathcal{K}$ is called A-invariant if and only if $A\mathcal{K} \subset \mathcal{K}$. $\mathcal{K}$ is called exponentially A-invariant if and only if $\forall t \geq 0 : e^{At}\mathcal{K} \subset \mathcal{K}$. □

Definition 2—Cone invariance
(A, B) is called cone-invariant w.r.t. a cone $\mathcal{K}$ if and only if $b_i \in \mathcal{K}$, for all i and $\mathcal{K}$ is exponentially invariant w.r.t. A. □

Similar to the introductory example, cone-invariance says: if the state-vector starts within a cone $\mathcal{K}$ then it will remain there for all nonnegative inputs u. Two important classes of cone-invariant systems are the so-called externally and internally positive systems, which will be discussed in Section 4.

In the following we define ellipsoidal cones, the essential ingredient for our main result. This class has been investigated in [Stern and Wolkowicz, 1991b; Stern and Wolkowicz, 1991a], which is why we adapt the notations.

Definition 3—Ellipsoidal cones
Let $Q = Q^T \in \mathbb{R}^{n\times n}$ with inertia $(n-1, 0, 1)$, then

$$\mathcal{K}_Q := \{x : x^T Q x \leq 0\}$$

is called an ellipsoidal double-cone. If $p \in \mathbb{R}^n$ is such that

$$\{p\}^\perp \cap \mathcal{K}_Q = \{0\}$$

where $\{p\}^\perp$ denotes the orthogonal complement of linear span $\{p\}$ of p, then we call $\mathcal{K}_{Q,p} := \{x : x^T Q x \leq 0, p^T x \geq 0\}$ an ellipsoidal cone. □

It is obvious that $\mathcal{K}_{Q,p}$ and $-\mathcal{K}_{Q,p} := \mathcal{K}_{Q,-p}$ are proper convex cones. In the following, we make the convention that $Q_n := \text{blkdiag}(I_{n-1}, -1)$ and $\mathcal{K}_{Q_n,e_n}$ is called the ice-cream cone, where e_n is the n-th canonical unit vector.

LEMMA 1
Let $\mathcal{K}_Q$ be an ellipsoidal double-cone and $\mathcal{K}_{Q,u_n}$ the corresponding ellipsoidal cone, where u_n is an eigenvector belonging to the negative eigenvalue of Q. Their dual sets can be parametrized as

$$\mathcal{K}_Q^* = \mathcal{K}_{Q^{-1}} \quad \text{and} \quad \mathcal{K}_{Q,u_n}^* = \mathcal{K}_{Q^{-1},u_n}.$$

□

Proof Let $\mathcal{K}_{Q,u_n}$ be an ellipsoidal cone with $\sigma(Q) = \{\lambda_1, \ldots, \lambda_n\}$, where

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n-1} > 0 > \lambda_n.$$

Diagonalizing Q with

$$U^T Q U = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) =: \Delta$$

defines a transformation matrix $T = |\Delta|^{\frac{1}{2}} U^T$, which gives

$$T\mathcal{K}_{Q,u_n} = \mathcal{K}_{T^{-T}QT^{-1},T^{-T}u_n} = \mathcal{K}_{Q_n,e_n}.$$

Since $\mathcal{K}_{Q_n,e_n}^* = \mathcal{K}_{Q_n,e_n}$ it follows that $(T\mathcal{K}_{Q,u_n})^* = \mathcal{K}_{Q_n,e_n}$ and therefore

$$\mathcal{K}_{Q,u_n}^* = T^T \mathcal{K}_{Q_n,e_n} = \mathcal{K}_{T^{-1}Q_nT^{-T},T^{-1}e_n} = \mathcal{K}_{U\Delta^{-1}U^T,Ue_n} = \mathcal{K}_{Q^{-1},u_n}.$$

By $\mathcal{K}_Q$ being independent of u_n we conclude the proof. □

The boundary of $\mathcal{K}_{Q,p}$ is given by

$$\partial\mathcal{K}_{Q,p} = \{x : x^T Q x = 0, p^T x > 0\} \cup \{0\}.$$

From this it follows that $p \in \operatorname{int}(\mathcal{K}_{Q,p}^*)$, which by the preceding Lemma is equivalent with $p^T Q^{-1} p < 0$. Thus, for given $Q = Q^T$ with inertia $(n-1, 0, 1)$ we conclude that $\mathcal{K}_{Q,p}$ is a proper convex cone if and only if

$$p^T Q^{-1} p < 0.$$

Moreover, $p \in \operatorname{int}(\mathcal{K}_{Q,p}^*)$ if and only if there exists $\tau > 0$ such that

$$\forall x \in \mathcal{K}_{Q,p} : x^T Q x + \tau x^T p p^T x > 0,$$

which is equivalent to

$$Q + \tau p p^T \succ 0.$$

Together with the main result in [Stern and Wolkowicz, 1991a], this leads to the following theorem.

THEOREM 1
Let $Q = Q^T$ with inertia $(n-1, 0, 1)$. Then $\mathcal{K}_{Q,p} := \{x : x^T Q x \leq 0, p^T x \geq 0\}$ is exponentially invariant w.r.t. A if and only if

$$\exists \gamma, \tau \in \mathbb{R} : A^T Q + QA + 2\gamma Q \preceq 0, \;\; Q + \tau p p^T \succ 0.$$

□

Notice that if $\sigma(A + \gamma I) \cap i\mathbb{R} = \emptyset$, then Q exists if and only if the inertia of $A+\gamma I$ and $-Q$ are equal (see e.g. [Datta, 2004]). Equivalently, A has a single dominant real eigenvalue $\lambda_{\max} \in \sigma(A)$ and $\sigma(A + \gamma I) \cap \mathbb{R}_{\geq 0} = \{\lambda_{\max} + \gamma\}$.

3. Central Theory

In the following we consider asymptotically stable systems as in (1), where (A, B) is invariant w.r.t. to an ellipsoidal double-cone. We assume the reader to be familiar with the concept of *standard balanced truncation* (BT) (see e.g. [Moore, 1981; Beck et al., 1996; Antoulas, 2005]).

In general, balanced truncation does not preserve the invariance with respect to an ellipsoidal cone – unless the system is reduced to order $r = 1$. To this end, we will modify the concept of balanced truncation to what we call *cone-balanced truncation*. For notational simplicity we start by deriving the main results for the case of a controllable system. Nonetheless, the reader should check that the results are still true in the uncontrollable case. Let us start with the first of two following modifications of balancing a system.

PROPOSITION 1
Given (A, B) and $\gamma > 0$, let $Q = Q^T$ with inertia $(n-1, 0, 1)$ and $P \succ 0$ fulfil

i. $A^T Q + QA + 2\gamma Q \preceq 0$,

ii. $b_j^T Q b_j < 0$ for all j,

iii. $AP + P^T = -BB^T$.

Then there exists $T \in \mathbb{R}^{n \times n}$ such that

$$T^{-1} P T^{-T} = \text{blkdiag}(\sigma_1, \sigma_2 I_{k_2}, \ldots, \sigma_s I_{k_s})$$
$$T^T Q T = \text{blkdiag}(-\sigma_1, \sigma_2 I_{k_2}, \ldots, \sigma_s I_{k_s})$$

where $\sigma_1 > \cdots > \sigma_s > 0$, $k_2 + \cdots + k_s = n - 1$ and

$$\sigma_1 \geq \sqrt{\sum_{i>1} \sigma_i^2}. \tag{2}$$

□

Proof Assume that P and Q are as in the claim and fulfil i. - iii.. We perform a singular value decomposition $P = U\Sigma_P U^T$ and define $L := U\Sigma_P^{\frac{1}{2}}$. By another singular value decomposition of L^TQL into

$$L^TQL = V\Sigma^2V^T$$

we define $T := LV\Sigma^{-\frac{1}{2}}$. Then we can verify that

$$\bar{P} := T^{-1}PT^{-T} \quad \text{and} \quad \bar{Q} := T^TQT$$

fulfill

$$\bar{P} = \Sigma^{\frac{1}{2}}V^TL^{-1}LL^TL^{-T}V\Sigma^{\frac{1}{2}} = \Sigma,$$
$$|\bar{Q}| = |\Sigma^{-\frac{1}{2}}V^TL^TQLV\Sigma^{-\frac{1}{2}}| = \Sigma,$$

with $\Sigma = \text{blkdiag}(\sigma_1 I_{k_1}, \ldots, \sigma_s I_{k_s})$, $\sigma_1 > \cdots > \sigma_s > 0$ and $k_1 + \cdots + k_s = n$. By Sylvester's law of inertia it follows that the inertia of T^TQT remains invariant, which is why $\bar{P}$ and $\bar{Q}$ are equal up to a sign-change on one of the diagonal entries.

We will show now that $\text{trace}(\bar{Q}) < 0$ implies that the sign-change occurs at σ_1 and $k_1 = 1$. To this end, assume without loss of generality that $P = I$ and $|Q| = \Sigma^2$, i.e.

$$A^TQ + QA + 2\gamma Q \preceq 0, \tag{3}$$
$$b_j^TQb_j < 0 \text{ for all } j, \tag{4}$$
$$A + A^T = -BB^T. \tag{5}$$

By substitution of $A = -BB^T - A^T$ in (3) we get

$$-(BB^T + A)Q - Q(BB^T + A^T) - 2\gamma Q \preceq -4\gamma Q. \tag{6}$$

Taking the trace over (6) and using

- $\sum_{j>0} b_j^TQb_j = \text{trace}(BB^TQ) = \text{trace}(QBB^T) < 0$
- $\text{trace}(AQ + QA^T + 2\gamma Q) = \text{trace}(A^TQ + QA + 2\gamma Q) \leq 0$

gives the following inequalities

$$-4\gamma\text{trace}(Q) \geq -2\text{trace}(BB^TQ) > 0$$
$$\Leftrightarrow \ \text{trace}(Q) \leq \frac{1}{2\gamma}\left(\text{trace}(BB^TQ)\right) < 0.$$

Therefore, by the inertia of Q and the assumption that $\sigma_1 > \cdots > \sigma_s > 0$, we conclude that the largest magnitude in Q is negative. □

If (A, B, C, D) is a system with (A, B), P, Q, γ and T as in Proposition 1, then truncating any of the last $n-1$ states of

$$(\bar{A}, \bar{B}, \bar{C}, \bar{D}) := (T^{-1}AT, T^{-1}B, CT, D)$$

preserves controllability as well as ellipsoidal cone-invariance. But, since T^TQT is indefinite we cannot apply the error-bound known from balanced truncation. Instead, we perform another balancing of $(\bar{A}, \bar{B}, \bar{C}, \bar{D})$ which will provide us with such.

Proposition 2
Let $(\bar{A}, \bar{B}, \bar{C}, \bar{D})$ be such that $(\bar{A}, \bar{B})$ is invariant w.r.t to $\mathcal{K}_{\bar{Q}}$ and

$$\bar{A}\bar{P} + \bar{P}\bar{A}^T = -\bar{B}\bar{B}^T$$

for diagonal $\bar{P} \succ 0$ with $\bar{P} = |\bar{Q}|$. Then

$$\exists \Delta \succ 0 : \ \bar{A}^T\Delta + \Delta\bar{A} \preceq -\bar{C}^T\bar{C}$$

with Δ being diagonal. □

Proof If $(\bar{A}, \bar{B}, \bar{C}, \bar{D})$ is as in the assumptions, then by Lemma 1 and Theorem 1 we conclude that

$$\bar{A}^T\bar{Q} + \bar{Q}\bar{A} + 2\gamma\bar{Q} \preceq 0, \tag{7}$$

$$\bar{A}\bar{P} + \bar{P}\bar{A}^T = -\bar{B}\bar{B}^T, \tag{8}$$

$$\bar{Q}^{-1} + \varepsilon\bar{b}_j\bar{b}_j^T \succ 0, \text{ for all } j, \tag{9}$$

for sufficiently large $\varepsilon > 0$. Multiplying (7) with $\bar{Q}^{-1}$ from the right and the left yields

$$\bar{A}\bar{Q}^{-1} + \bar{Q}^{-1}\bar{A}^T + 2\gamma\bar{Q}^{-1} \preceq 0 \tag{10}$$

and multiplying (8) by $2\gamma\varepsilon$ gives

$$2\gamma\varepsilon\bar{A}\bar{P} + 2\gamma\varepsilon\bar{P}\bar{A}^T + 2\gamma\varepsilon\bar{B}\bar{B}^T = 0, \tag{11}$$

where $2\gamma\varepsilon\sigma_1 - \sigma_1^{-1} > 0$. Adding up (10) and (11) results in

$$\bar{A}\Delta^{-1} + \Delta^{-1}\bar{A}^T + 2\gamma\left(\bar{Q}^{-1} + \varepsilon\bar{B}\bar{B}^T\right) \preceq 0$$

with $\Delta := (2\gamma\varepsilon\bar{P} + \bar{Q}^{-1})^{-1} \succ 0$. Finally, a proper scaling of Δ gives a diagonal solution to

$$\bar{A}^T\Delta + \Delta\bar{A} \preceq -\bar{C}^T\bar{C}. \tag{12}$$

□

Now, if $\bar{P}$ and Δ are as in Proposition 2 we can define a second balancing transformation

$$\bar{T} := \text{blkdiag}\left(1, \frac{\bar{p}_{22}}{\delta_{22}}, \ldots, \frac{\bar{p}_{nn}}{\delta_{nn}}\right)^{\frac{1}{4}}$$

such that

$$(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) := (\bar{T}^{-1}\bar{A}\bar{T}, \bar{T}^{-1}\bar{B}, \bar{C}\bar{T}, \bar{D})$$

fulfils

$$\begin{aligned}\tilde{A}\tilde{P} + \tilde{P}\tilde{A}^T &= -\tilde{B}\tilde{B}^T,\\ \tilde{A}^T\tilde{Q} + \tilde{Q}\tilde{A} &\preceq -\tilde{C}^T\tilde{C}\end{aligned}$$

where $\tilde{P}$ and $\tilde{Q}$ are diagonal and equal except for the first diagonal entry.

Definition 4—Cone-balanced
A linear system $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ is called cone-balanced if $\exists \gamma > 0$, $\tilde{P}, \tilde{Q} \succ 0$ and $\tilde{K} = \tilde{K}^T$ with inertia $(n-1, 0, 1)$ such that

$$\begin{aligned}\tilde{A}^T\tilde{K} + \tilde{K}\tilde{A} + 2\gamma\tilde{K} &\preceq 0,\\ \tilde{A}^T\tilde{Q} + \tilde{Q}\tilde{A} &\preceq -\tilde{C}^T\tilde{C},\\ \tilde{A}^T\tilde{P} + \tilde{P}\tilde{A} &= -\tilde{B}\tilde{B}^T,\end{aligned}$$

where $\tilde{P}$, $\tilde{Q}$ and $\tilde{K}$ are diagonal with

$$k_{11} < 0 \quad \text{and} \quad \tilde{p}_{22} = \tilde{q}_{22} \geq \cdots \geq \tilde{p}_{nn} = \tilde{q}_{nn}. \qquad \square$$

Again, truncating a cone-balanced system preserves ellipsoidal cone-invariance and it is well known (see e.g. [Beck et al., 1996]) that the error-bound result from standard balanced truncation carries over to the diagonal elements of $\tilde{P}$.

Theorem 2
Suppose $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ is a cone-balanced realization of a stable, minimal cone-invariant system with transfer function $\tilde{G}(s)$ and controllability Gramian $\tilde{P} = \text{blkdiag}\left(\tilde{\Sigma}_1, \tilde{\Sigma}_2\right)$,

$$\begin{aligned}\tilde{\Sigma}_1 &= \text{blkdiag}\left(\tilde{\sigma}_1, \tilde{\sigma}_2 I_{k_2} \ldots, \tilde{\sigma}_r I_{k_r}\right),\\ \tilde{\Sigma}_2 &= \text{blkdiag}\left(\tilde{\sigma}_{r+1} I_{k_{r+1}}, \ldots, \tilde{\sigma}_p I_{k_p}\right),\end{aligned}$$

where $\tilde{\sigma}_2 > \cdots > \tilde{\sigma}_r > \tilde{\sigma}_{r+1} > \cdots > \tilde{\sigma}_p > 0$.
Truncating the states corresponding to $\tilde{\Sigma}_2$ results in an approximation (A_r, B_r, C_r, D_r) of order $1 + \sum_{i=2}^{r} k_i$ with transfer function $G_r(s)$, which is cone-balanced, controllable and stable. Moreover, it holds for the H_∞-error

$$\|\tilde{G}(s) - G_r(s)\|_\infty \leq 2 \sum_{i=r+1}^{p} \tilde{\sigma}_i. \tag{13}$$

$\square$

It is known (see [Grussler and Damm, 2012]) that the $\tilde{\sigma}_i$ in (13) are always larger than the Hankel singular values. Nevertheless, we will see in Section 6 that we can get fairly close to them. The whole algorithm for *cone-balanced truncation* (CBT) is summarized in Algorithm 1.

Algorithm 1 Cone balanced truncation (CBT)

1: Let (A, B, C, D) be a minimal system.

2: **IF** (A, B) fulfils Proposition 1.
3: Find $T \in \mathbb{R}^{n\times n}$ such that $(\bar{A}, \bar{B}, \bar{C}, \bar{D}) := (T^{-1}AT, T^{-1}B, CT, D)$ has diagonal controllability Gramian $\bar{P}$ and $(\bar{A}, \bar{B})$ is invariant w.r.t. $\mathcal{K}_{\bar{Q}}$ with $\bar{P} = |\bar{Q}|$.
4: Minimize $\sum_{i>1} \delta_{ii}$ subject to

$$\bar{A}^T\Delta + \Delta\bar{A} \preceq -\bar{C}^T\bar{C}$$
$$\Delta := \operatorname{blkdiag}(\delta_{11}, \dots, \delta_{nn}) \succ 0.$$

5: Find a cone-balanced realization $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ with generalized singular values $\tilde{\sigma}_i := \sqrt{\bar{p}_{ii}\delta_{ii}},\ i > 1$.
6: Choose a reduced order according to (13) and truncate $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$.
7: **END**

4. Positive Systems

In the following we formally define externally and internally positive systems and compare them with ellipsoidal cone-invariant systems. After that, it will be evident why our result naturally extends to the class of externally positive systems.

Definition 5—External Positivity
A linear system (1) is called *externally positive* if and only if its output corresponding to a zero initial state is nonnegative for every nonnegative input. □

Proposition 3—[Farina and Rinaldi, 2011]
A linear system (A, B, C, D) is externally positive if and only if $\forall t \geq 0$: $Ce^{At}B \in \mathbb{R}^{k\times m}_{\geq 0}$ and $D \in \mathbb{R}^{k\times m}_{\geq 0}$. □

It is readily seen that every single-input-single-output (SISO) externally positive system (A, B, C) is invariant with respect to its so-called reachable and observable cone

$$R(A, B) := \operatorname{clcone}\{e^{At}B : t \geq 0\} \quad \text{and} \quad O(A, C) := \{x : \forall t \geq 0 : Ce^{At}x \geq 0\},$$

where cone$(\cdot)$ denotes the convex conic hull and cl$(\cdot)$ the topological closure (see [Ohta et al., 1984]).

Definition 6—Internal Positivity
A linear system (1) is called internally positive if and only if its state and output are nonnegative for every nonnegative input and every nonnegative initial state. □

Internal positivity of (1) requires that the nonnegative orthant $\mathbb{R}^n_{\geq 0}$ is exponentially invariant w.r.t. A. In [Berman and Plemmons, 1994] it is shown that this is the case if and only if A is Metzler, i.e.

$$\exists \alpha \geq 0 : A + \alpha I \in \mathbb{R}^{n\times n}_{\geq 0}.$$

Theorem 3—[Farina and Rinaldi, 2011]
A continuous linear system (A, B, C, D) is internally positive if and only if A is Metzler and B, C, D are nonnegative. □

Verification of external positivity is known to be NP-hard (see e.g. [Blondel and Portier, 2002]). Restricting oneself to internally positive systems is a convenient way to deal with this problem. But, as indicated in Theorem 3, internal positivity depends on very specific state-space realizations. Finding such a realization is known to be computationally difficult (see [Ohta et al., 1984; Anderson et al., 1996]). We believe, if one is only interested in external positivity, it is beneficial to look at externally positive systems, which are ellipsoidal cone-invariant. In this case, verification of external positivity can be performed with the help of convex optimization.

Theorem 4
Given (A, B, C, D) with $D \in \mathbb{R}^{k\times m}_{\geq 0}$, assume that there exists $Q = Q^T$ with inertia $(n-1, 0, 1)$ and $\gamma, \tau \in \mathbb{R}$ such that

1. $A^T Q + QA + 2\gamma Q \preceq 0$,
2. $b_j^T Q b_j < 0$ for all j,
3. $Q + \tau_i c_i^T c_i \succ 0$ for all i,
4. $CB \in \mathbb{R}^{k\times m}_{>0}$,

where c_i is the i-th row of C. Then (A, B, C, D) is externally positive. □

Proof The result follows directly from Lemma 1 and Theorem 1, which imply that (A, B, C, D) is invariant w.r.t. $\mathcal{K}_{Q,c_i^T}$, for all i.

Corollary 1
Assume $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ is a stable and externally positive system fulfilling Theorem 4. Then cone-balanced truncation preserves external positivity. □

Proof Assume w.l.o.g. that $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ is externally positive and cone-balanced w.r.t. $\mathcal{K}_{Q,c_i} = \mathcal{K}_{Q,e_1}$ for all i. Moreover, we assume that we reduce our system by one order with CBT to $(\bar{A}, \bar{B}, \bar{C}, \bar{D})$. Clearly, 1) – 3) in Theorem 4 are preserved after cone-balanced truncation. To see 4), observe that

$$b_j \in \mathcal{K}_{Q,e_1} \Rightarrow \begin{pmatrix} b_{1,j} & \cdots & b_{n-1,j} & 0 \end{pmatrix}^T \in \mathcal{K}_{Q,e_1} \Rightarrow \bar{c}_i \bar{b}_j \geq 0.$$

□

We will refer to this method as *positive cone-balanced truncation* (PCBT).

5. Discussion

In the previous two sections we have formally derived a solution to the following problems

I. *Ellipsoidal cone preserving model reduction*.

II. *External positivity preserving model reduction under the constraint of ellipsoidal cone-invariance*.

Now, we want to get some further insights into these problems and into the numerical computations involved.

First, notice it is straightforward to extend all results to discrete-time linear systems – for ellipsoidal cones see e.g. [Stern and Wolkowicz, 1991b]. Furthermore, Lemma 1, Proposition 1 and the duality between observability and controllability imply that if a non-minimal system is ellipsoidal cone-invariant then this is also true for its minimal realization.

A major draw-back of our method is the need to solve linear matrix inequalities (LMIs) in order to preserve external positivity – LMI-solvers are usually computational demanding (see e.g. [Peaucelle et al., 2002]). Given the NP-hardness of the verification problem, this is a small price to pay. Moreover, standard balanced truncation usually requires pre-reduction methods such as [Gugercin et al., 2008] to be able to handle large-scale systems. Hence, it is valid to assume a reduced system whose LMIs are sufficiently fast solvable.

Further observe, if one only wants to verify/preserve cone-invariance, it is often enough to consider Lyapunov equations. To see this, assume (A, B) is controllable with $\sigma(A+\gamma I) \cap i\mathbb{R} = \emptyset$ and no b_j is in the span of the eigenvectors belonging to the non-dominant eigenvalues. If $\mathcal{K}_Q$ is exponentially A-invariant then so is $e^{-At}K_Q,\ t \geq 0$, which by our assumptions implies

that $\exists t \geq 0 : b_j \in e^{-At}K_Q$, for all j. More explicitly, if $B \in \mathbb{R}^n$ one could solve

$$A^TQ + QA + 2\gamma Q = -R \preceq 0,$$
$$AP + PA^T + 2\gamma P = -BB^T.$$

Assuming without loss of generality that P is diagonal with $p_{11} < 0$, then

$$B^TQB = \text{trace}(BB^TQ) = \text{trace}(P(A^TQ + QA + 2\gamma Q)) = \text{trace}(PR)$$

yields that $B^TQB < 0$ for any $R \succeq 0$ with $r_{11} > 0$ and $r_{ii} = 0,\ i > 1$.

Corollary 2
Assume (A, B, C, D) is a minimal symmetric SISO-system, i.e. $A = A^T$ and $B = C^T$ and assume that A has single dominant eigenvalue. Then there exists $Q = Q^T$ fulfilling Theorem 4. □

Proof By the previous discussion it follows that there exists $Q = Q^T$ and u_n, such that $\mathcal{K}_{Q,u_n}$ is exponentially invariant w.r.t. A and $B \in \mathcal{K}_{Q,u_n}$. Hence, $C^T \in \mathcal{K}_{Q,u_n}$ and

$$\forall t \geq 0 : C^T e^{A\frac{t}{2}} e^{A\frac{t}{2}} B = \|e^{A\frac{t}{2}} B\|^2 > 0.$$

W.l.o.g. we can assume that

$$\mathcal{K}_{Q,u_n} = \mathcal{K}_{Q_n,e_n} \quad \text{and} \quad B^TQ_nB = CQ_nC^T < 0.$$

That implies that $C^T \in \text{int}(\mathcal{K}^*_{Q_n,e_n})$ which is why $\mathcal{K}_{Q,e_n} = \mathcal{K}_{Q,C^T}$. □

It is straightforward to show that all symmetric SISO-systems have an internally positive realization of the same dimension.

A practical procedure to deal with large-scale externally positive systems could be the following:

1. Reduce the system with help of a Krylov-subspace method (see [Antoulas, 2005; Gugercin et al., 2008]) to an order where Lyapunov equations can be solved efficiently.

2. Apply CBT to reduce the system to an order where the LMIs in Theorem 4 can be solved efficiently.

3. Use PCBT to verify external positivity and to reduce the system even further.

Our derivations deal with externally positive systems (A, B, C) where

$$CB \in \mathbb{R}^{k\times m}_{>0}.$$

To treat cases with zero entries in CB one could pre-approximate the original system with $(A, e^{A\epsilon}B, C)$ for $\epsilon > 0$. Then

$$\forall t > 0 : \; Ce^{A\epsilon+t}B > 0$$

by the assumption of a single dominant real pole. The error between those systems can be made arbitrarily small by the choice of ϵ.

Finally, note that if an externally positive system has a strictly dominant real pole of multiplicity one, then the system possesses a positive realization [Anderson et al., 1996]. Thus our method also preserves internally positive realizability. Unfortunately, internal positivity is not sufficient to ensure the requirements of Theorem 4.

6. Examples & Comparison

By considering some numerical examples, we discuss the quality of (positive) cone-balanced truncation. The results are compared to symmetric balanced truncation (SBT) in [Grussler and Damm, 2012] and standard balanced truncation (BT). Moreover, by the comparisons in [Grussler and Damm, 2012; Grussler, 2012] and [Sootla and Rantzer, 2012] it follows, that even a reduced model of order one often outperforms the methods in [Reis and Virnik, 2009; Feng et al., 2010; Li et al., 2011; Sootla and Rantzer, 2012].

Our comparison will always start from a minimal realization, which can be considered a pre-reduction. In order to make the solutions unique, we will add to minimize

$$\operatorname{trace}(Q + \tau C^T C)$$

in case of PCBT, which turned out to give good results. For CBT we use the same γ-shift as determined by PCBT and Q is given by

$$A^T Q + QA + 2\gamma Q = -C^T C.$$

6.1 Heat Equation

We begin with one of the examples given in [Grussler and Damm, 2012], the two-dimensional heat equation on a square

$$\dot{T} = \triangle T = \frac{\partial^2}{\partial x^2} T + \frac{\partial^2}{\partial y^2} T \tag{14}$$

with control of the Dirichlet boundary conditions of the four edges. Discretisation on a uniform grid leads to the following linear internally positive system:

$$\dot{T} = AT + Bu \;\text{ with } u \in \mathbb{R}^4 \text{ and } T \in \mathbb{R}^{N^2} \tag{15}$$

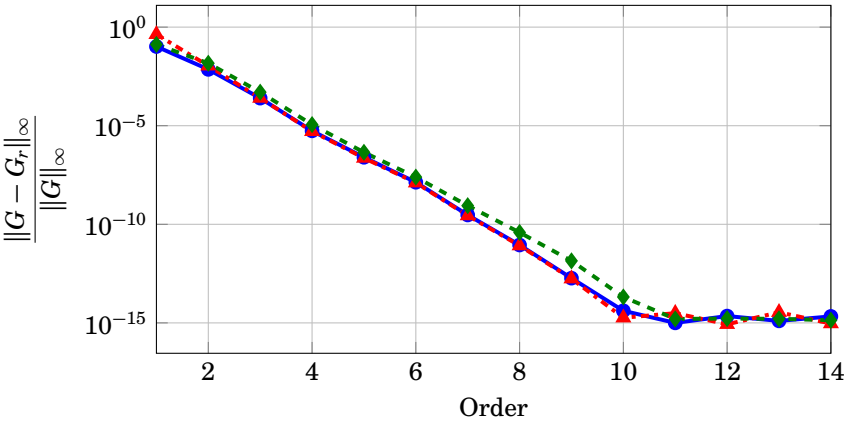


Figure 1. Normalized H_∞-error in heat experiment 1:
(S)BT: (symmetric) standard balanced truncation
CBT: cone preserving balanced truncation
PCBT: positivity preserving CBT

where A is the $N^2 \times N^2$-Poisson-matrix and $B := [b_{ij}] \in \mathbb{R}^{N^2 \times 4}$, where $b_{ij} = 0$ except for the following cases:

$$\begin{aligned} b_{i1} &:= 1, && \text{for } i = 1, 2, \dots, N \\ b_{i2} &:= 1, && \text{for } i = N, 2N, \dots, N^2 \\ b_{i3} &:= 1, && \text{for } i = N(N-1)+1, N(N-1)+2, \dots, N^2 \\ b_{i4} &:= 1, && \text{for } i = 1, N+1, \dots, N(N-1)+1 \end{aligned}$$

One may think of this example in the same way as in the one given in the introduction. In our first experiment the output is equal to the global average temperature, i.e.

$$y = \frac{1}{N^2} CT, \text{ with } C := \mathbf{1}_{N^2}^T := \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{1 \times N^2}.$$

In this case it was shown that SBT performs very well, because the minimal balanced realization is a symmetric system. Then by Corollary 2 it must be possible to apply PCBT. Repeating this experiment for (P)CBT with $N = 10$ gives the H_∞-error as shown in Figure 1. We observe, (P)CBT performs closely to (S)BT, the error-difference is due to numerical issues and a different sorting of the singular values. Moreover, PCBT does not suffer from disregarding the advantage of symmetry, as exploited by SBT. In fact, (P)CBT preserves it and an internally positive realization can be

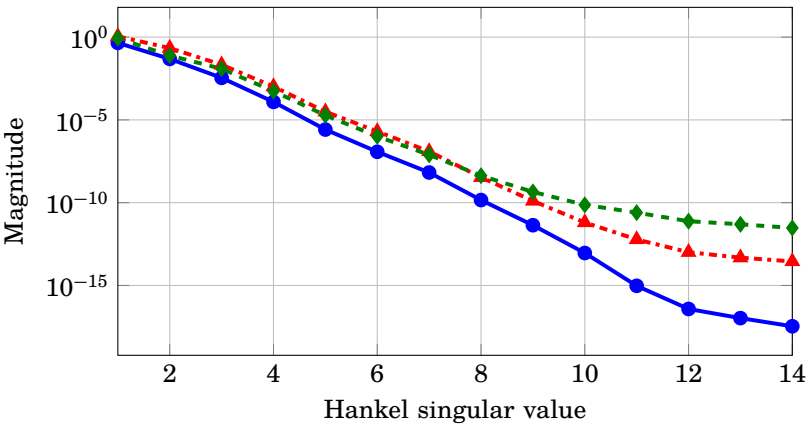


Figure 2. (Generalized) Hankel singular values in heat experiment 1:
(S)BT: (symmetric) standard balanced truncation
CBT: cone preserving balanced truncation
PCBT: positivity preserving CBT

found here as well. Also the error-bounds for (P)CBT lie within a good range as indicated in Figure 2, where the generalized Hankel singular values of P(CBT) result from Theorem 2.

Now, we modify this example by using the second and the fourth input only. Furthermore, we split the unit-square into 5 equally spaced vertical stripes and let y represent the average temperature in each of these zones, i.e.

$$C = \operatorname{blkdiag}\left(\mathbf{1}^T_{\frac{N^2}{5}}, \mathbf{1}^T_{\frac{N^2}{5}}, \mathbf{1}^T_{\frac{N^2}{5}}, \mathbf{1}^T_{\frac{N^2}{5}}, \mathbf{1}^T_{\frac{N^2}{5}}\right)$$

In this case, the minimal balanced system is no longer symmetric. Therefore, SBT will arrive with an approximation of order 1 and the same error as BT. Again, the normalized errors are shown in Figure 3.

6.2 Balanced truncation destroying positivity

It is readily verified that

$$G(s) = \frac{(s+1)^{10}}{(s+1)\prod_{k=2}^{3}(s+2-e^{\pm\sqrt{k}\pi})\prod_{k=4}^{9}(s+2-\frac{1}{k})}$$

defines an externally positive systems, which has an ellipsoidal cone-invariant realization. BT does not preserve these properties for the reduced models of order two and four. A comparison of the normalized errors is

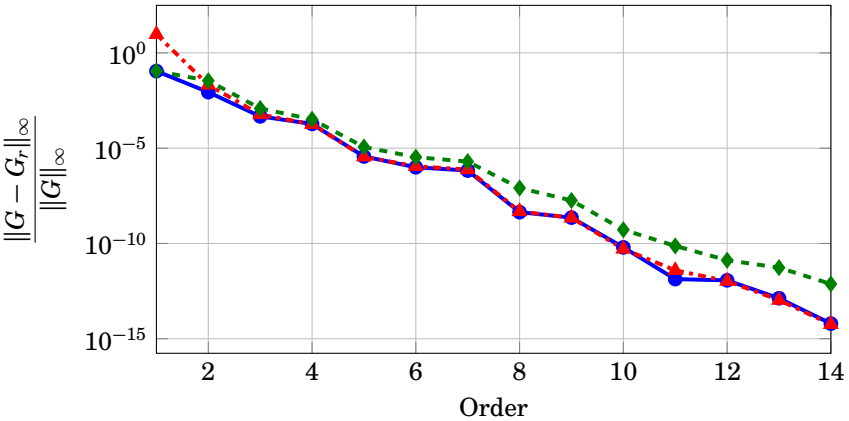


Figure 3. Normalized H_∞-error in heat experiment 2:
BT: standard balanced truncation
CBT: cone preserving balanced truncation
PCBT: positivity preserving CBT

presented in Figure 4. Although, both methods perform well, observe the comparably large error of PCBT for order two. Interestingly, other well established model reduction methods, such as [Glover, 1984] and [Gugercin et al., 2008] also destroy positivity for a better error performance.

7. Conclusion

We have not only presented a model reduction method which guarantees to preserve ellipsoidal cone-invariance but also defined a class of systems, which gives a broad intersection of some well studied cone-invariant systems. In fact, it seems that ellipsoidal cone-invariance is often implied by internal positivity. By that a numerical test for external/internal positivity has been established as well as a method for external positivity preserving model order reduction.

References

Anderson, B. D. O., M. Deistler, L. Farina, and L. Benvenuti (1996). "Nonnegative realization of a linear system with nonnegative impulse response". *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* **43**:2, pp. 134–142.

Antoulas, A. (2005). *Approximation of Large-Scale Dynamical Systems*. SIAM.

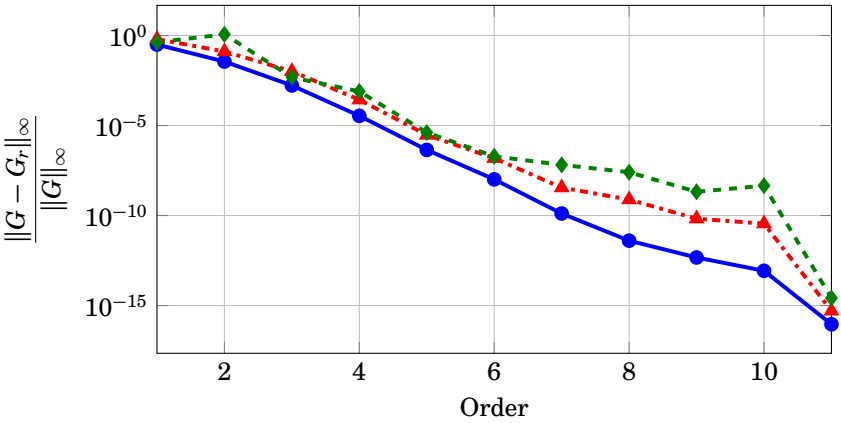


Figure 4. Normalized H_∞-error for the Academic Example:
BT: standard balanced truncation
CBT: cone preserving balanced truncation
PCBT: positivity preserving CBT

Beck, C. L., J. Doyle, and K. Glover (1996). "Model reduction of multidimensional and uncertain systems". *IEEE Transactions on Automatic Control* **41**:10, pp. 1466–1477.

Berman, A. and R. Plemmons (1994). *Nonnegative Matrices in the Mathematical Sciences*. SIAM.

Blondel, V. D. and N. Portier (2002). "The presence of a zero in an integer linear recurrent sequence is NP-hard to decide". *Linear Algebra and its Applications* **351**, pp. 91–98.

Brown, R. F. (1980). "Compartmental system analysis: state of the art". *IEEE Transactions on Biomedical Engineering* **BME-27**:1, pp. 1–11.

Datta, B. N. (2004). *Numerical Methods for Linear Control Systems*. Vol. 1. Academic Press.

Farina, L. and S. Rinaldi (2011). *Positive Linear Systems: Theory and Applications*. John Wiley & Sons.

Feng, J., J. Lam, Z. Shu, and Q. Wang (2010). "Internal positivity preserved model reduction". *Int. Journal of Control* **83**:3, pp. 575–584.

Glover, K. (1984). "All optimal hankel-norm approximations of linear multivariable systems and their L_∞-error bounds". *International Journal of Control* **39**:6, pp. 1115–1193.

Grussler, C. and T. Damm (2012). "A symmetry approach for balanced truncation of positive linear systems". In: *51st IEEE Conference on Decision and Control (CDC)*, pp. 4308–4313.

Grussler, C. (2012). *Model reduction of positive systems*. M.Sc.Thesis. Lund University, Department of Automatic Control.

Gugercin, S., A. C. Antoulas, and C. Beattie (2008). "$\mathcal{H}_2$ model reduction for large-scale linear dynamical systems". *SIAM Journal on Matrix Analysis and Applications* **30**:2, pp. 609–638.

Gugercin, S. and A. C. Antoulas (2004). "A survey of model reduction by balanced truncation and some new results". *International Journal of Control* **77**:8, pp. 748–766.

Li, P., J. Lam, Z. Wang, and P. Date (2011). "Positivity-preserving H_∞ model reduction for positive system". *Automatica* **47**:7, pp. 1504–1511.

Luenberger, D. (1979). *Introduction to Dynamic Systems: Theory, Models & Applications*. John Wiley & Sons.

Moore, B. (1981). "Principal component analysis in linear systems: controllability, observability, and model reduction". *IEEE Transactions on Automatic Control* **26**:1, pp. 17–32.

Ohta, Y., H. Maeda, and S. Kodama (1984). "Reachability, observability, and realizability of continuous-time positive systems". *SIAM Journal on Control and Optimization* **22**:2, pp. 171–180.

Peaucelle, D., D. Henrion, Y. Labit, and K. Taitz (2002). "User's guide for SEDUMI INTERFACE 1.04". LAAS-CNRS, Toulouse.

Reis, T. and E. Virnik (2009). "Positivity preserving balanced truncation for descriptor systems". *SIAM Journal on Control and Optimization* **48**:4, pp. 2600–2619.

Shorten, R., F. Wirth, and D. Leith (2006). "A positive systems model of TCP-like congestion control: asymptotic results". *IEEE/ACM Transactions on Networking* **14**:3, pp. 616–629.

Sootla, A. and A. Rantzer (2012). "Scalable positivity preserving model reduction using linear energy functions". In: *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 4285–4290.

Stern, R. J. and H. Wolkowicz (1991a). "Exponential nonnegativity on the ice cream cone". *SIAM Journal on Matrix Analysis and Applications* **12**:1, pp. 160–165.

Stern, R. J. and H. Wolkowicz (1991b). "Invariant ellipsoidal cones". *Linear Algebra and its Applications* **150**, pp. 81–106.

Acknowledgments

The authors are members of the LCCC Linnaeus Center and the eLLIIT Excellence Center at Lund University.

Paper III

Low-Rank Optimization with Convex Constraints

Christian Grussler Anders Rantzer Pontus Giselsson

Abstract

The problem of low-rank approximation with convex constraints, which often appears in data analysis, image compression, and model order reduction, is considered. Given a data matrix, the objective is to find a low-rank approximation that meets rank and convex constraints, while minimizing the distance to the data matrix in the Frobenius norm. The problem of matrix completion can be seen as a special case of this. Today, one of the most widely used techniques is to approximate this non-convex problem using convex nuclear norm regularization. In many situations, this technique does not give solutions with desirable properties. In this paper, we propose an alternative to the nuclear norm heuristic that promotes low-rank solutions. It is based on using the largest convex minorizer (under-approximation) of the squared Frobenius norm and the rank constraint as a convex proxy. This optimal convex proxy can be combined with other convex constraints to form an optimal convex minorizer of the original non-convex problem. With this approach, easily verifiable conditions are obtained, under which the solutions to the convex relaxation and the original non-convex problem coincide. Several numerical examples are provided for which that is the case. It is shown that the proposed convex relaxation consistently performs better than the nuclear norm heuristic, especially in the matrix completion case. The expressibility and computational tractability are of great importance for a convex relaxation so that they can be applied using standard software. A closed-form expression for the proposed convex relaxation is provided in addition to its representation as a semi-definite program. Furthermore, it is shown how to compute the proximal operator of the convex approximation. This allows the use of scalable first-order methods to solve convex approximation problems of large size.

Preprint.

1. Introduction

The rank captures many of the essential components of an otherwise complex operator. For instance, the rank of a matrix $N \in \mathbb{R}^{n\times m}$ equals the dimension of its column space. In other words, if a matrix has low rank then only a small number of basis vectors are needed to span its range and a possibly high dimensional subspace in $\mathbb{R}^m$ can be disregarded when studying $y = Nx$. Hence, if N is sufficiently close to a lower rank matrix, it may be sufficient to study the approximation $y \approx \hat{N}x$ where $\operatorname{rank}(\hat{N}) < \operatorname{rank}(N)$.

Due to this simplifying concept, many areas such as image analysis, model order reduction, multivariate linear regression, etc. desire a low-rank approximation (see [Izenman, 1975; Antoulas, 2005; Markovsky, 2008; Candès and Plan, 2010; Recht et al., 2010; Chandrasekaran et al., 2012; Reinsel and Velu, 1998; Hastie et al., 2015; Larsson and Olsson, 2016; Vidal et al., 2016]). In Sections 6 to 9 some of these applications are explained in greater depth.

For unitarily invariant norms an optimal low-rank approximation can be found by performing a singular value decomposition (SVD) (see Section 2). Unfortunately, these approximations usually do not fulfill structural constraints such as element-wise nonnegativity, Hankel-structure, prescribed entries, etc. (see [Higham, 2002; Chu et al., 2003; Berry et al., 2007; Markovsky, 2008; Candès and Recht, 2009; Olsson and Oskarsson, 2009; Recht et al., 2010; Reinsel and Velu, 1998]). Only in a few cases, an explicit solution to the constrained low-rank approximation problem is known (see [Antoulas, 2005; Markovsky, 2008; Reinsel and Velu, 1998]). For this reason, other concepts based on convex optimization have been developed (see [Fazel et al., 2001; Recht et al., 2010; Chandrasekaran et al., 2011; Larsson et al., 2014]). Many of them rely on nuclear norm regularization, which allows the incorporation of any convex constraint (see Section 5.1). Nevertheless, the question if this yields solutions to the non-convex problem is not addressed, unless one aims for a minimum rank solution (see [Candès and Recht, 2009; Recht et al., 2010]). Besides the nuclear norm heuristic, other commonly used heuristics, e.g. for element-wise nonnegativity are briefly considered in Section 6.

In this work, we study the optimal low-rank approximation problem with a prescribed target rank and convex constraints (see Problem 1). This is a continuation of the authors work [Grussler and Rantzer, 2015]. It is shown that a globally optimal solution to our non-convex problem can often be determined by convex optimization (see Section 3). In particular, if the SVD-approximation of a matrix is unique, then it is a solution to a semi-definite program (SDP). Even though the approach presented can be linked to the regularization method in [Larsson et al., 2014; Larsson and Olsson, 2016], we will see that the proposed method does not require a costly search

for a regularization parameter.

In Section 4 some computational aspects of the convexified problem are discussed. First, an SDP-representation of the convex proxy is presented, which allows the computation of solutions for small scale examples with SDP-representable constraints. Subsequently, we derive the so-called Douglas-Rachford iteration in order to deal with examples of larger size and sufficiently simple constraints (see Section 4.2). As a consequence, we will be able to prove local convergence of the Douglas-Rachford iterations of the original non-convex problem.

The paper is organized as follows. In Section 2 we recap the unconstrained low-rank approximation problem and define our main problem. The main approach is derived and discussed in Section 3 with some computational aspects examined in Section 4. Other known approaches, including the nuclear norm heuristic are discussed in Section 5. In Sections 6 to 9 some applications are presented that show the usefulness of this approach. Moreover, the examples are chosen to illustrate some properties and drawbacks of this method. Finally, we draw a conclusion and discuss future research in Section 10.

2. Background

The following notation for real matrices $X = (x_{ij}) \in \mathbb{R}^{n\times m}$ is used throughout this paper. If $X = X^T$, i.e. X is symmetric, then we write $X \in \mathbb{S}$. Moreover, if X is positive definite (semi-definite) we use the notation $X \succ 0$ ($X \succeq 0$). We also use these notations to describe the relation between two matrices, e.g. $A \succeq B$ means $A - B \succeq 0$.

The non-increasingly ordered singular values of $X \in \mathbb{R}^{n\times m}$, counted with multiplicity, are denoted by $\sigma_1(X) \geq \cdots \geq \sigma_{\min\{m,n\}}(X)$. Further, $\langle X, Y\rangle := \sum_{i=1}^{m}\sum_{j=n}^{n} x_{ij}y_{ij} = \operatorname{trace}(X^TY)$ defines the Frobenius inner-product for $X, Y \in \mathbb{R}^{n\times m}$. Correspondingly, the Frobenius norm is defined as

$$\|X\|_F := \sqrt{\sum_{i=1}^{m}\sum_{j=n}^{m} x_{ij}^2} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(X)}.$$

The Frobenius norm is unitarily invariant, i.e. $\|UXV\|_F = \|X\|_F$ for all unitary matrices U and V. A complete characterization of all unitarily invariant norms can be found in [Horn and Johnson, 2012]. This work mainly considers the unitarily invariant norms that are found in the following Lemma. A proof to this lemma is provided in Section A.3.

LEMMA 1
Let $M \in \mathbb{R}^{n\times m}$, and $r \in \mathbb{N}$ be such that $1 \le r \le q := \min\{m, n\}$. Then,

$$\|M\|_r := \sqrt{\sum_{i=1}^{r} \sigma_i^2(M)} = \sup_{\substack{\|X\|_F=1 \\ \operatorname{rank}(X)\le r}} \langle M, X\rangle \tag{1}$$

is a unitarily invariant norm with dual norm

$$\|M\|_{r*} := \max_{\|X\|_r\le 1} \langle M, X\rangle = \max_{\sum_i^r s_i^2 \le 1} \left[\sum_{i=1}^{r} \sigma_i(M)s_i + s_r \sum_{i=r+1}^{q} \sigma_i(M)\right].$$

Moreover,

$$\|M\|_1 \le \cdots \le \|M\|_q = \|M\|_F = \|M\|_{q*} \le \cdots \le \|M\|_{1*}. \tag{2}$$

$$\operatorname{rank}(M) \le r \text{ if and only if } \|M\|_r = \|M\|_F = \|M\|_{*r}. \tag{3}$$

□

Notice that $\|M\|_1 = \sigma_1(M)$ is equal to the spectral norm and its dual norm $\|M\|_{1*} = \sum_{i=1}^{\min\{m,n\}} \sigma_i(M)$ is equal to the nuclear (trace norm). These norms can be formulated using convex linear matrix inequalities (see [Fazel et al., 2001; Recht et al., 2010]). In Section 3 is shown that the same holds true for $\|\cdot\|_r^2$ and $\|\cdot\|_{r*}^2$. Unfortunately, there is no closed form expression for $\|\cdot\|_{r*}$. However, as discussed in [Freimer and Mudholkar, 1984], the necessary computations for evaluating dual norms of this form can be reduced to a one-dimensional parameter search.

Recently, the vector version of the $r*$-norm has appeared as under the names *k-support norm* (see [Argyriou et al., 2012]) or *overlapping norm* (see [Bach et al., 2012]). As a result, some authors have adopted that name for the matrix case (see [Lai et al., 2014; Eriksson et al., 2015; McDonald et al., 2015]). However, as for other vector/matrix norm pairings e.g. the ℓ_1 norm of the singular values is called the nuclear norm, we have chosen the $r*$-norm notation to distinguish between the matrix and vector case. This also avoids confusion if the k-support-norm is applied to the entries of a matrix and not its singular values.

2.1 Statements

Let us turn to the underlying problem of this work. We start with the traditional optimal low-rank approximation problem in $\mathbb{R}^{n\times m}$, which is formulated as follows. Given $N \in \mathbb{R}^{n\times m}$ and $r \in \mathbb{N}$ such that $1 \le r \le \min\{m, n\}$, find a solution $M^\star \in \mathbb{R}^{n\times m}$ to

$$\begin{aligned} &\text{minimize} && \|N - M\|_F^2 \\ &\text{subject to} && \operatorname{rank}(M) \le r \end{aligned} \tag{4}$$

In case of the Hilbert-Schmidt norm, the natural operator generalization of the Frobenius-norm, this problem has been solved by Schmidt and generalized by Mirsky to unitarily invariant norms (see [Antoulas, 2005]). The result is stated next.

Proposition 1
Let $N \in \mathbb{R}^{n\times m}$ and $r \in \mathbb{N}$ such that $1 \leq r \leq \min\{m, n\}$, then

$$\min_{\substack{M\in\mathbb{R}^{n\times m}\\ \operatorname{rank}(M)\leq r}} \|N - M\| = \|\operatorname{diag}(\sigma_{r+1}(N), \ldots, \sigma_{\min\{m,n\}}(N))\|,$$

holds for any unitarily invariant norm $\|\cdot\|$.

If an SVD of N is given by $N = \sum_{i=1}^{\min\{m,n\}} \sigma_i u_i v_i^T$, a solution to (4) can be derived as $M^\star = \operatorname{svd}_r(M) := \sum_{i=1}^{r} \sigma_i u_i v_i^T$, which we refer to as a *standard SVD-approximation*. This solution may not be unique if the norm does not depend on all singular values or if $\sigma_r(N) = \sigma_{r+1}(N)$. Nevertheless, with the Frobenius norm and $\sigma_r(N) \neq \sigma_{r+1}(N)$ the uniqueness of $M^\star$ is guaranteed.

However, this solution does not account for additional constraints. In this work, we look at the following extension of (4).

Problem 1
Given $N \in \mathbb{R}^{n\times m}$, find $M^\star \in \mathbb{R}^{n\times m}$ with $\operatorname{rank}(M^\star) \leq r$ such that

$$\min_{\substack{M\in\mathbb{R}^{n\times m}\\ \operatorname{rank}(M)\leq r}} \left[\frac{1}{2}\|N - M\|_F^2 + g(M)\right] = \frac{1}{2}\|N - M^\star\|_F^2 + g(M^\star),$$

where $g : \mathbb{R}^{n\times m} \to \mathbb{R} \cup \{\infty\}$ is a given closed proper convex function (see Definition A.2). □

Compared to (4), Problem 1 has an additional function g that can be used to add information about the desired solution. Both problems are non-convex due to the rank constraint. Nevertheless, we will see in Section 3 that they can often be solved by convex optimization. In particular, if (4) has a unique solution, it is possible to determine it by solving a semi-definite program. Notice that Problem 1 also deals with cases where $N = 0$ and thus covers the class of matrix completion problems (see Section 7).

In the following we often use $g(M) \equiv \chi_{\mathcal{C}}(M)$, where

$$\chi_{\mathcal{C}}(M) := \begin{cases} 0, & M \in \mathcal{C} \\ \infty, & M \notin \mathcal{C} \end{cases}$$

is defined to be the indicator function of a (convex) set $\mathcal{C} \subset \mathbb{R}^{n\times m}$. We also use $\chi_{\operatorname{rank}(M)\leq r}$ to denote the indicator function of the set of matrices with at most rank r. In the remainder of this paper, it is assumed that $g + \chi_{\operatorname{rank}(M)\leq r}$ is proper.

3. The $r*$-approach

In the following we consider the problem of finding *optimal* solutions to Problem 1. It is a continuation of the authors work [Grussler and Rantzer, 2015]. The insights obtained here will allow us to generalize and improve upon current standard approaches (see Section 7). The main idea is to derive a convex minorizer (under-approximation) of the non-convex cost-function in Problem 1 by means of Fenchel-duality (see Section A.2). We denote by f^* and f^{**} the conjugate and bi-conjugate functions of $f : \mathbb{R}^{n\times m} \to \mathbb{R}\cup\{\infty\}$ (for those unfamiliar with these concepts see Definition A.1).

Theorem 1
Let $N \in \mathbb{R}^{n\times m}$, and $r \in \mathbb{N}$ such that $1 \leq r \leq \min\{m, n\}$. Then the conjugate and bi-conjugate functions of

$$f(M) := \frac{1}{2}\|N - M\|_F^2 + \chi_{\text{rank}(M)\leq r}(M)$$

are given by

$$f^*(D) = \frac{1}{2}\|N + D\|_r^2 - \frac{1}{2}\|N\|_F^2, \tag{5}$$

$$f^{**}(M) = \frac{1}{2}\|M\|_{r*}^2 - \langle N, M\rangle + \frac{1}{2}\|N\|_F^2. \tag{6}$$

for all $D, M \in \mathbb{R}^{n\times m}$. □

Proof Let $N \in \mathbb{R}^{n\times m}$ and $f(M) := \frac{1}{2}\|N - M\|_F^2 + \chi_{\text{rank}(M)\leq r}(M)$. Then,

$$\begin{aligned}
f^*(D) &= \sup_{\substack{M\in\mathbb{R}^{n\times m}\\ \text{rank}(M)\leq r}} \left[\langle D, M\rangle - \frac{1}{2}\|N - M\|_F^2\right] \\
&= \sup_{\substack{M\in\mathbb{R}^{n\times m}\\ \text{rank}(M)\leq r}} -\frac{1}{2}\|N - M + D\|_F^2 + \langle D, N\rangle + \frac{1}{2}\|D\|_F^2 \\
&= -\frac{1}{2}\|N + D\|_F^2 + \frac{1}{2}\|N + D\|_r^2 + \langle D, N\rangle + \frac{1}{2}\|D\|_F^2 \\
&= -\frac{1}{2}\|N\|_F^2 + \frac{1}{2}\|N + D\|_r^2
\end{aligned}$$

where the third equality follows by Proposition 1, because

$$-\frac{1}{2}\|N + D\|_F^2 + \frac{1}{2}\|N + D\|_r^2 = \|\text{diag}(\sigma_{r+1}(N + D), \ldots, \sigma_{\min\{m,n\}}(N + D))\|_F^2.$$

Hence,

$$\begin{aligned} f^{**}(M) &= \sup_{D \in \mathbb{R}^{n \times m}} \left[\langle D, M \rangle + \frac{1}{2}\|N\|_F^2 - \frac{1}{2}\|N + D\|_r^2 \right] \\ &= \sup_{D \in \mathbb{R}^{n \times m}} \left[\langle D - N, M \rangle + \frac{1}{2}\|N\|_F^2 - \frac{1}{2}\|D\|_r^2 \right] \\ &= \frac{1}{2}\|N\|_F^2 - \langle N, M \rangle + \sup_{D \in \mathbb{R}^{n \times m}} \left[\langle D, M \rangle - \frac{1}{2}\|D\|_r^2 \right] \\ &= \frac{1}{2}\|N\|_F^2 - \langle N, M \rangle + \frac{1}{2}\|M\|_{r*}^2, \end{aligned}$$

where the last equality follows by

$$\frac{1}{2}\|\cdot\|_{r*}^2 = \left(\frac{1}{2}\|\cdot\|_r^2 \right)^*,$$

which is for instance shown in [Rockafellar, 1970, Corollary 15.3.1]. □

It is possible to show that f^* and f^{**} are convex (see [Hiriart-Urruty and Lemaréchal, 2013]). Moreover, $f(M) \geq f^{**}(M)$ for all $M \in \mathbb{R}^{n \times m}$, i.e. f^{**} is a convex minorizer of f. In fact, f^{**} it is the largest convex minorizer of f (see [Hiriart-Urruty and Lemaréchal, 1996, Theorem X.1.3.5]), that is, it is the point-wise supremum of all affine functions majorized by f (see Figure 1). This allows us to construct the following dual and bi-dual problem to Problem 1:

$$-\min_{D \in \mathbb{R}^{n \times m}} \left[g^*(-D) + \frac{1}{2}\|N + D\|_r^2 - \frac{1}{2}\|N\|_F^2 \right], \tag{A}$$

$$\min_{M \in \mathbb{R}^{n \times m}} \left[\frac{1}{2}\|M\|_{r*}^2 - \langle N, M \rangle + \frac{1}{2}\|N\|_F^2 + g(M) \right]. \tag{B}$$

Observe that $f^{**} + g$ is the largest convex minorizer of $f + g$ with g as a summand. Therefore, we propose to use (B) instead of the nuclear norm heuristic (see (25) in Section 5.1) as a convex proxy to Problem 1. We will see that it has many interesting properties and that sometimes it can be guaranteed to solve the original non-convex problem. Theorem 1 gives the following duality result through Fenchel-duality (see Lemma A.1 and Proposition A.3).

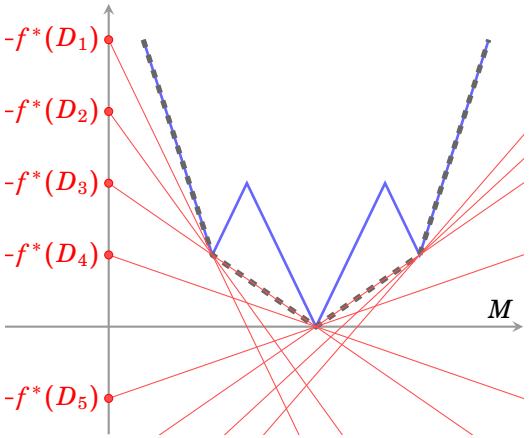


Figure 1. Schematic plot of $f(M)$, $f^{**}(M)$ and tangents through $-f^*(D_i)$.

Proposition 2
Let $N \in \mathbb{R}^{n\times m}$ and $g : \mathbb{R}^{n\times m} \to \mathbb{R}\cup\{\infty\}$ be a closed proper convex function. Then for all $r \in \mathbb{N}$ such that $1 \le r \le \min\{m, n\}$

$$\begin{aligned}
&\min_{\substack{M\in\mathbb{R}^{n\times m}\\ \operatorname{rank}(M)\le r}} \left[\frac{1}{2}\|N-M\|_F^2 + g(M)\right] \\
&\qquad \ge -\min_{D\in\mathbb{R}^{n\times m}} \left[g^*(-D) + \frac{1}{2}\|N+D\|_r^2 - \frac{1}{2}\|N\|_F^2\right] \qquad \text{(C)}\\
&\qquad = \min_{M\in\mathbb{R}^{n\times m}} \left[\frac{1}{2}\|M\|_{r*}^2 - \langle N, M\rangle + \frac{1}{2}\|N\|_F^2 + g(M)\right].
\end{aligned}$$

□

Since the original Problem 1 is non-convex, there is a duality-gap for some choices of g (see Section 7). This is reflected by the inequality in (C). However, there are many situations with no duality-gap. Next, a number of important cases are presented.

In the following, the set of minimizers of a function f over a given set $S \subset \mathbb{R}^{n\times m}$ is denoted by $\operatorname{argmin}_S f$. If $\operatorname{argmin}_S f = \{x^\star\}$ is just a singleton, we write $x^\star = \operatorname{argmin}_S f$.

Proposition 3
Assume that (B) has a minimizer $M^\star$ with $\operatorname{rank}(M^\star) \leq r$. Then,

$$\operatorname*{argmin}_{\substack{M \in \mathbb{R}^{n\times m} \\ \operatorname{rank}(M)\leq r}} \left[\frac{1}{2}\|N - M\|_F^2 + g(M)\right] = \operatorname*{argmin}_{\substack{M \in \mathbb{R}^{n\times m} \\ \operatorname{rank}(M)\leq r}} \left[\frac{1}{2}\|M\|_{r*}^2 - \langle N, M\rangle + \frac{1}{2}\|N\|_F^2 + g(M)\right]. \qquad \square$$

Proof The result follows by combining Proposition 2 with (3) in Lemma 1.□

Thus obtaining a rank-r solution to the convex relaxation problem (B) implies solving the original non-convex problem. Next, this result is restated to provide additional insight on the solution to Problem 1.

Proposition 4
Assume that $D^\star$ is a solution to (A) and $\sigma_r(N + D^\star) \neq \sigma_{r+1}(N + D^\star)$ or $\sigma_r(N + D^\star) = 0$. Then there is no duality gap in (C) and $\operatorname{svd}_r(N + D^*)$ is the unique minimizing argument of Problem 1, i.e.

$$\operatorname{svd}_r(N + D^\star) = \operatorname*{argmin}_{\substack{M \in \mathbb{R}^{n\times m} \\ \operatorname{rank}(M)\leq r}} \left[\frac{1}{2}\|N - M\|_F^2 + g(M)\right].$$

□

The theorem provides a simple sufficient condition for the uniqueness of a solution to Problem 1. However, this is not a necessary condition. A proof of Proposition 4 is given in a more general setting than in Theorem 2, which also allows us to say something about the rank of the solution to the convex relaxation if there is a duality-gap.

Theorem 2
Let $D^\star$ and $M^\star$ be solutions to (A) and (B), respectively. Further, suppose that an SVD of $N + D^\star$ is given by $N + D^\star = \sum_{i=1}^{\min\{m,n\}} \sigma_i u_i v_i^T$ with

$$\sigma_{r-t} \neq \sigma_{r-t+1} = \cdots = \sigma_r = \cdots = \sigma_{r+s} \neq \sigma_{r+s+1},$$

where $t = r$ and $s = \min\{m, n\} - r$ if $\sigma_1 = \sigma_r$ and $\sigma_{\min\{m,n\}} = \sigma_r$, respectively. Then there exists $T \in \mathbb{R}^{s+t\times s+t}$ such that

$$M^\star = \sum_{i=1}^{r-t} \sigma_i u_i v_i^T + \sigma_r \begin{pmatrix} u_{r-t+1} & \dots & u_{r+s} \end{pmatrix} T \begin{pmatrix} v_{r-t+1} & \dots & v_{r+s} \end{pmatrix}^T$$

where $T \succeq 0$, $\|T\|_1 \leq 1$, and $\|T\|_{1*} = t$. In particular, $\operatorname{rank}(M^\star) \leq r + s$, and if $\sigma_r \neq \sigma_{r+1}$ or $\sigma_r = 0$, then $M^\star = \operatorname{svd}_r(N + D^\star)$. □

A proof to this theorem is given in Section A.4. Observe that whenever (B) does not have a unique solution, Proposition 1 and Theorem 2 imply that $\sigma_r(N + D^\star) = \sigma_{r+1}(N + D^\star)$ for all solutions $D^\star$ to (A). Furthermore, Theorem 2 shows that $\operatorname{svd}_r(N)$ with $\sigma_r(N) \neq \sigma_{r+1}(N)$ can be determined by solving a convex problem.

COROLLARY 1
Let $N \in \mathbb{R}^{n\times m}$, and $r \in \mathbb{N}$ be such that $1 \leq r \leq \min\{m, n\}$. Then

$$\begin{aligned}\min_{\substack{M\in\mathbb{R}^{n\times m}\\ \operatorname{rank}(M)\leq r}} \frac{1}{2}\|N - M\|_F^2 &= \frac{1}{2}\|N\|_F^2 - \frac{1}{2}\|N\|_r^2 \\ &= \min_{M\in\mathbb{R}^{n\times m}} \left[\frac{1}{2}\|M\|_{r*}^2 - \langle N, M\rangle + \frac{1}{2}\|N\|_F^2\right]\end{aligned}$$

and

$$\operatorname{svd}_r(N) \in \operatorname*{argmin}_{M\in\mathbb{R}^{n\times m}} \left[\frac{1}{2}\|M\|_{r*}^2 - \langle N, M\rangle\right].$$

If $\sigma_r(N) \neq \sigma_{r+1}(N)$ or $\sigma_r = 0$ then

$$\operatorname{svd}_r(N) = \operatorname*{argmin}_{M\in\mathbb{R}^{n\times m}} \left[\frac{1}{2}\|M\|_{r*}^2 - \langle N, M\rangle\right].$$

□

Proof Since $g = 0$ implies that $g^*(D) < \infty \Leftrightarrow D = 0$, the result follows by Theorem 2. □

Finally, notice that several extensions of Problem 1 are covered by the preceding results. For instance, one can consider the weighted case

$$\min_{\substack{M\in\mathbb{R}^{n\times m}\\ \operatorname{rank}(M)\leq r}} \left[\frac{1}{2}\|W(N - M)\|_F^2 + g(M)\right] \tag{7}$$

where $W \in \mathbb{R}^{l\times n}$ and $\operatorname{rank}(W) = n$. Let $\tilde{g}(\tilde{M}) := g(W^\dagger \tilde{M})$, where $W^\dagger$ denotes the pseudo-inverse of W (see [Horn and Johnson, 2012]). Since $\operatorname{rank}(\tilde{M}) = \operatorname{rank}(W^\dagger \tilde{M}) = \operatorname{rank}(M)$, one can reformulate (7) such that it fits the formulation of Problem 1:

$$\min_{\substack{M\in\mathbb{R}^{n\times m}\\ \operatorname{rank}(M)\leq r}} \left[\frac{1}{2}\|W(N - M)\|_F^2 + g(M)\right] = \min_{\substack{\tilde{M}\in\mathbb{R}^{n\times m}\\ \operatorname{rank}(\tilde{M})\leq r}} \left[\frac{1}{2}\|WN - \tilde{M}\|_F^2 + \tilde{g}(\tilde{M})\right].$$

In particular,

$$\|W(N - M)\|_F^2 = \operatorname{trace}((N - M)^T W^T W(N - M)) =: \langle N - M, N - M\rangle_{W^T W}$$

defines another inner product and norm, and thus a suitable W may enable us to satisfy the requirements of Proposition 4 in situations where the Frobenius norm fails.

3.1 Geometric interpretation

Assuming that $g(M) \equiv \chi_{\mathcal{C}}(M)$ for some closed convex set $\mathcal{C} \subset \mathbb{R}^{n\times m}$, the preceding results have an insightful geometric interpretation. Note that (B) has the same solutions as

$$\min_{\substack{M\in\mathcal{C}\\ \langle N,M\rangle=c}} \|M\|_{r*}, \tag{8}$$

where $c := \langle N, M^\star\rangle$, and $M^\star$ is a solution to (B). The solutions of (8) can be found by studying the set $B^{\bar\varepsilon}_{r*} \cap H \cap \mathcal{C}$ where

$$B^{\varepsilon}_{r*} := \{X \in \mathbb{R}^{n\times m} : \|X\|_{r*} \le \varepsilon\},$$
$$H := \{X \in \mathbb{R}^{n\times m} : \langle N, X\rangle = c\},$$

and

$$\bar\varepsilon := \min\{\varepsilon \ge 0 : B^{\varepsilon}_{r*} \cap H \cap \mathcal{C} \neq \emptyset\}.$$

Proposition 4 states that if $\sigma_r(N + D^\star) \neq \sigma_{r+1}(N + D^\star)$, then $B^{\bar\varepsilon}_{r*} \cap H \cap \mathcal{C}$ consists of a single element. This can also be understood geometrically with the help of the following Lemma, which generalizes the corresponding result for the nuclear norm and $r = 1$ (see [Recht et al., 2010]).

Lemma 2
The set of the extreme points of the unit-ball B^1_{r*} is

$$E := \{X \in \mathbb{R}^{n\times m} : \|X\|_F = 1,\ \operatorname{rank}(X) \le r\}.$$

Hence, $B^1_{r*} = \operatorname{conv}(E)$, where $\operatorname{conv}(\cdot)$ denotes the convex hull. □

Proof By (1) in Lemma 1, it holds that for all $N \in \mathbb{R}^{n\times m}$

$$\sup_{M\in\operatorname{conv}(E)} \langle N, M\rangle = \|N\|_r = \sup_{M\in B^1_{r*}} \langle N, M\rangle. \tag{9}$$

Since $\operatorname{conv}(E)$ and B^1_{r*} are closed convex sets, Lemma A.2 implies that $B^1_{r*} = \operatorname{conv}(E)$. If a point $\bar M \in E$ is not an extreme point of E, then

$$\bar M = \textstyle\sum_i \alpha_i M_i, \quad \text{with} \quad \sum_i \alpha_i = 1,$$

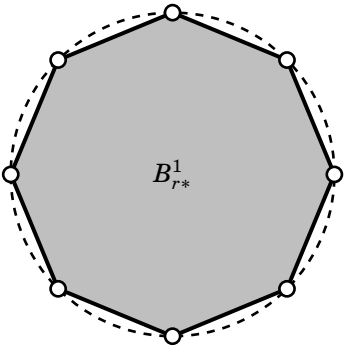
such that

$$M_i \in K \setminus \{\bar M\} \quad \text{and} \quad \alpha_i > 0 \text{ for all } i.$$

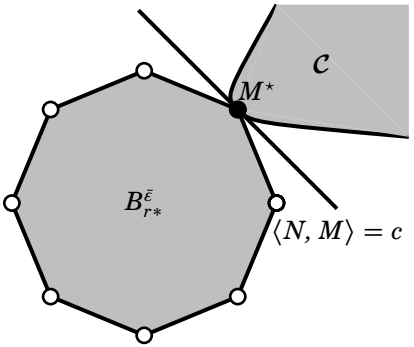
Hence, by the Cauchy-Schwarz inequality we conclude that

$$1 = \langle \bar M, \bar M\rangle = \textstyle\sum_i \alpha_i \langle \bar M, M_i\rangle \le \sum_i \alpha_i = 1.$$

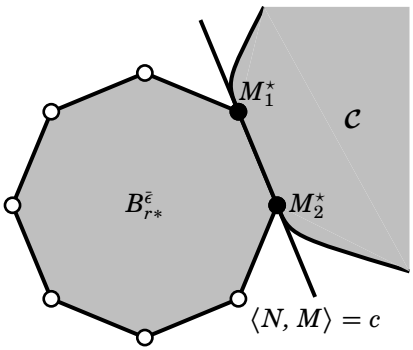
However, this can only be true if $\langle \bar M, M_i\rangle = 1$ for all i. Equivalently, $\bar M = M_i$ by the Cauchy-Schwarz inequality and that is a contradiction. □



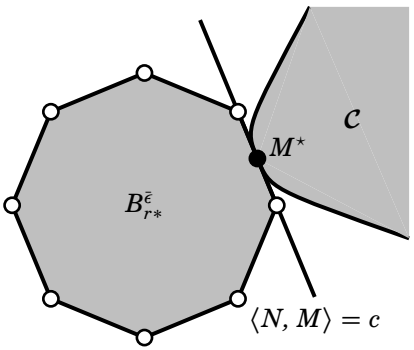
(a) B^1_{r*} (shaded area) as the convex hull of E (elements ∘) with the boundary of $B^1_{\min\{m,n\}*}$(- - -).



(b) Unique solution: $\{M^\star\} = B^{\bar{\varepsilon}}_{r*} \cap H \cap \mathcal{C}$ with $\operatorname{rank}(M^\star) \le r$.



(c) Non-unique solutions: $\operatorname{conv}(\{M^\star_1, M^\star_2\}) = B^{\bar{\varepsilon}}_{r*} \cap H \cap \mathcal{C}$ with $\operatorname{rank}(M^\star_1) \le r$ and $\operatorname{rank}(M^\star_2) \le r$.



(d) Duality gap: $\{M^\star\} = B^{\bar{\varepsilon}}_{r*} \cap H \cap \mathcal{C}$ with $\operatorname{rank}(M^\star) > r$.

Figure 2. Schematic plots to visualize (8) geometrically.

Therefore, a geometric interpretation of $\sigma_r(N + D^\star) \neq \sigma_{r+1}(N + D^\star)$ is that the only intersection point of H and $B^{\bar{\varepsilon}}_{r*} \cap \mathcal{C}$ is an extreme point of $B^{\bar{\varepsilon}}_{r*}$ and $\mathcal{C}$ (see Figure 2(b)). Hence, the case of $\sigma_r(N + D^\star) = \sigma_{r+1}(N + D^\star) \neq 0$ can occur if and only if H intersects $B^{\bar{\varepsilon}}_{r*} \cap \mathcal{C}$ at several points (see Figure 2(c) and Section 6.1) or if there is a duality gap in (C) (see Figure 2(d) and Section 7.3). Finally notice that one can also use Lemma 2 as a definition of $\|\cdot\|_{r*}$. This has been done for vectors in [Argyriou et al., 2012; Bach et al., 2012] in an attempt to generalize the ℓ_1 norm.

3.2 Real-valued r

In the following we will see that allowing r to be real-valued can be considered as a regularization parameter. Unlike typical regularization methods (see Sections 5.1 and 5.2), this parameter has a close relationship to the rank of the corresponding solutions.

It suffices to discuss the case where Proposition 4 does not apply. Therefore, let

$$D_t^\star := \operatorname*{argmin}_{D\in\mathbb{R}^{n\times m}} \left[g^*(-D) + \frac{1}{2}\|N+D\|_t^2 \right],$$

and

$$M_t^\star := \operatorname*{argmin}_{M\in\mathbb{R}^{n\times m}} \left[\frac{1}{2}\|M\|_{t*}^2 - \langle N, M\rangle + g(M) \right].$$

be defined for all $t \in \mathbb{N}$ such that $1 \le t \le \min\{m,n\}$, and assume that there exists $r \in \mathbb{N}$ with

$$\sigma_r(N+D_r^\star) = \sigma_{r+1}(N+D_r^\star) \quad \text{and} \quad \operatorname{rank}(M_r^\star) > r.$$

Furthermore, assume that

$$\frac{1}{2}\|N-M_r^\star\|_F^2 + g(M_r^\star) > \frac{1}{2}\|N-M_{r+1}^\star\|_F^2 + g(M_{r+1}^\star)$$

with

$$\operatorname{rank}(M_{r+1}^\star) > \operatorname{rank}(M_r^\star).$$

Then, on the one hand, one may face the situation that $M_r^\star$ is an approximation of small rank, but poor cost $\|N-M_r^\star\|_F^2 + g(M_r^\star)$. On the other hand, $\|N-M_{r+1}^\star\|_F + g(M_{r+1}^\star)$ may be acceptable, but $\operatorname{rank}(M_{r+1}^\star)$ is too large. Thus a trade-off between $M_r^\star$ and $M_{r+1}^\star$ is desired. This can be achieved by letting r become a non-integer valued in the r norm. The r norm is then defined as

$$\|\cdot\|_r := \sqrt{\sum_{i=1}^{\lfloor r\rfloor} \sigma_i^2(\cdot) + (r-\lfloor r\rfloor)\sigma_{\lceil r\rceil}^2(\cdot)}, \tag{10}$$

where $\lfloor r\rfloor := \max\{z\in\mathbb{Z} : z\le r\}$ and $\lceil r\rceil := \min\{z\in\mathbb{Z} : z\ge r\}$. Observe that for $r\in\mathbb{N}$ and $\alpha\in[0,1]$ we have

$$\|\cdot\|_{r+\alpha}^2 = (1-\alpha)\|\cdot\|_r^2 + \alpha\|\cdot\|_{r+1}^2. \tag{11}$$

This means that $\|\cdot\|_{r+1-\alpha}^2$ is a convex combination of $\|\cdot\|_r^2$ and $\|\cdot\|_{r+1}^2$, and thus indicates its usefulness in supplying the desired trade-off solution. Similar to Theorem 2, it remains true with $r\in\mathbb{R}_{\ge 1}$ that $\operatorname{rank}(M_r^\star) \le \lceil r\rceil + s$ if

$$\sigma_{\lceil r\rceil}(N+D_r^\star) = \cdots = \sigma_{\lceil r\rceil+s}(N+D_r^\star) > \sigma_{\lceil r\rceil+s+1}(N+D_r^\star). \tag{12}$$

Hence, allowing r to assume values in $\mathbb{R}_{\geq 1}$ may allow us to find solutions of both lower rank and lower cost. Next let us have a closer look at the dependency of s on r in (12).

We define

$$F(D, r) := g^*(-D) + \frac{1}{2}\|N + D\|_r^2 + \frac{1}{2}\|N\|_F^2.$$

Using (2) in Lemma 1 and the piecewise linearity in (11), it follows that F is convex. From Berge's Maximum Theorem (see [Berge, 1963, p. 116] or [Sundaram, 1996, Theorem 9.17] for the convex case) it is known that the parameter depending set

$$\mathcal{C}^*(r) := \underset{D \in \mathbb{R}^{n\times m}}{\operatorname{argmin}} \left[g^*(-D) + \frac{1}{2}\|N + D\|_r^2 + \frac{1}{2}\|N\|_F^2 \right]$$

is upper hemicontinuous in r. This means that for all $r \in [1, \min\{m, n\}]$ and all $\varepsilon > 0$ there exists $\delta > 0$ such that for all $t \geq 1$

$$|t - r| < \delta \Rightarrow \mathcal{C}^*(t) \subset \mathcal{B}_\varepsilon\left(\mathcal{C}^*(r)\right), \tag{13}$$

where

$$\mathcal{B}_\varepsilon\left(\mathcal{C}^*(r)\right) := \{X \in \mathbb{R}^{n\times m} : \exists D \in \mathcal{C}^*(r) \text{ such that } \|X - D\|_F < \varepsilon\}.$$

For simplicity assume that $D_r^\star$ is unique. By (13) and the continuity of the singular values (see [Stewart and Sun, 1990, Corollary 4.9]), it follows that a sufficiently small increase of r does not increase s in (12). Therefore, just as in nuclear norm regularization, one often observes that $\operatorname{rank}(M_t^*)$ looks like a staircase as t varies over $[r, r+1]$ (see Figure 9(b) in Section 8.1). Notice, the same observation can be made with

$$F(M, r) := \frac{1}{2}\|M\|_{r*}^2 - \langle N, M\rangle + \frac{1}{2}\|N\|_F^2 + g(M)$$

and

$$\mathcal{C}^*(r) := \underset{M \in \mathbb{R}^{n\times m}}{\operatorname{argmin}} \left[\frac{1}{2}\|M\|_{r*}^2 - \langle N, M\rangle + \frac{1}{2}\|N\|_F^2 + g(M) \right].$$

In summary, real-valued r can be considered as a regularization parameter, similar to the regularization methods in Section 5.

4. Computability

This section is devoted to the computability aspects of the $r*$-approach. We show that the problems (A) and (B) can be formulated as SDPs if g is SDP-representable. Moreover, we compute the proximal mappings of f^* and f^{**}

in Theorem 1. This allows us to solve (A) and (B) using a first order method such as Douglas-Rachford splitting. Further, we apply Douglas-Rachford to the original non-convex Problem 1. If Proposition 4 applies, then its iterates coincide locally with the convex Douglas-Rachford.

4.1 SDP-representations

We start with an SDP-representation of the optimization problem

$$\min_{D\in\mathbb{R}^{n\times m}} \|N+D\|_r^2, \tag{14}$$

where $\|\cdot\|_r$ is defined as in (10) and $r \in [1, \min\{m,n\}]$. Let $T \in \mathbb{R}^{n\times n}$ be such that

$$T \succeq (N+D)(N+D)^T.$$

Then $\sigma_i(T) \geq \sigma_i^2(N+D)$ for all i such that $1 \leq i \leq n$ (see [Horn and Johnson, 2012, Corollary 7.7.4]) and $\operatorname{trace}(T) = \sum_{i=1}^n \sigma_i(T)$. Hence,

$$\begin{aligned}\|N+D\|_r^2 &\leq \operatorname{trace}(T) - (\lceil r\rceil - r)\sigma_{\lceil r\rceil}(T) - \sum_{i=\lceil r\rceil+1}^{n} \sigma_i(T)\\ &\leq \operatorname{trace}(T) - (n-r)\sigma_n(T),\end{aligned}$$

which is equivalent to

$$\|N+D\|_r^2 \leq \min_{T\succeq(N+D)(N+D)^T} \operatorname{trace}(T) - (n-r)\sigma_n(T). \tag{15}$$

In particular, equality in (15) can be achieved with

$$T^\star := \sum_{i=1}^{\lceil r\rceil} \sigma_i^2(N+D)u_iu_i^T + \sigma_{\lceil r\rceil}^2(N+D)\sum_{i=\lceil r\rceil+1}^{n} u_iu_i^T,$$

where $N+D = \sum_{i=1}^n \sigma_i(N+D)u_iv_i^T$ is an SVD of $N+D$. Using the Schur-complement condition for $T-(N+D)(N+D)^T \succeq 0$ (see [Horn and Johnson, 2012, Theorem 7.7.7]) gives that (14) is SDP-representable as

$$\begin{aligned}\underset{D,T,\gamma}{\text{minimize}} \quad & \operatorname{trace}(T) - \gamma(n-r)\\ \text{subject to} \quad & \begin{pmatrix} T & N+D\\ (N+D)^T & I\end{pmatrix} \succeq 0,\ T \succeq \gamma I,\ D\in\mathbb{R}^{n\times m}.\end{aligned}$$

Moreover, if g is SDP-representable, then an SDP-formulation of (B) can be obtained by the dual of this optimization problem. We get

$$\begin{aligned}\underset{M,P,W}{\text{minimize}} \quad & \frac{1}{2}\operatorname{trace}(W) - \operatorname{trace}(N^TM) + g(M)\\ \text{subject to} \quad & \begin{pmatrix} I-P & M\\ M^T & W\end{pmatrix} \succeq 0,\ P\succeq 0,\\ & \operatorname{trace}(P) = n-r.\end{aligned}$$

Assuming that $\sigma_r(N + D^\star) \neq \sigma_{r+1}(N + D^\star)$, the unique solution $M^\star$ to Problem 1 can be found without computing the solution to (A).

4.2 Convex Douglas-Rachford

Many SDP-solvers are based on interior point methods (see [Toh et al., 1999; Peaucelle et al., 2002]). These solvers have good convergence properties, but the iteration complexity typically grows unfavorably with the problem dimension. In order to deal with problems of higher dimensions, it is often more desirable to look at first-order methods such as the Douglas-Rachford splitting algorithm (see [Douglas and Rachford, 1956; Lions and Mercier, 1979; Eckstein and Bertsekas, 1992]). Let us recall the basic concept of this method. We want to determine a solution to

$$\underset{X}{\text{minimize}} \quad f(X) + g(X), \tag{16}$$

where $f, g : \mathbb{R}^{n\times m} \to \mathbb{R} \cup \{\infty\}$ are closed and proper convex functions with intersecting domains. Then the Douglas-Rachford iteration is given by

$$X^k = \text{prox}_{\gamma f}(Z^{k-1}), \tag{17a}$$
$$Y^k = \text{prox}_{\gamma g}(2X^k - Z^{k-1}), \tag{17b}$$
$$Z^k = Z^{k-1} + \rho(Y^k - X^k), \tag{17c}$$

where $\gamma > 0$, $\rho \in (0, 2)$, and the proximal mapping is defined as

$$\text{prox}_{\gamma f}(Z) := \underset{X}{\text{argmin}} \left(f(X) + \frac{1}{2\gamma}\|X - Z\|_F^2 \right). \tag{18}$$

It is known that X^k and Y^k converge towards a minimizer of (16) (see [Douglas and Rachford, 1956; Lions and Mercier, 1979; Eckstein and Bertsekas, 1992]). In fact, the well-known Alternating Direction Methods of Multipliers (ADMM) is a special case of the Douglas-Rachford iteration (see [Glowinski and Marroco, 1975; Gabay and Mercier, 1976; Boyd et al., 2011]). Note that the Douglas-Rachford splitting algorithm can also be applied to sums of more than two functions f and g by using a consensus formulation (see [Combettes and Pesquet, 2011]).

Let g be as in (B), and assume that $\text{prox}_{\gamma g}(X)$ is easy to compute. In order to apply the Douglas-Rachford algorithm to solve (B), it remains to find $\text{prox}_{\gamma f}(Z)$ with

$$f(M) := \frac{1}{2}\|M\|_{r*}^2 - \langle N, M\rangle + \frac{1}{2}\|N\|_F^2$$

for all $M \in \mathbb{R}^{n\times m}$. For $Z \in \mathbb{R}^{n\times m}$, we get

$$\begin{aligned}\operatorname{prox}_{\gamma f}(Z) &= \operatorname*{argmin}_{M\in\mathbb{R}^{n\times m}} \left(\frac{1}{2}\|M\|_{r*}^2 - \langle N, M\rangle + \frac{1}{2}\|N\|_F^2 + \frac{1}{2\gamma}\|M - Z\|_F^2 \right) \\ &= \operatorname*{argmin}_{M\in\mathbb{R}^{n\times m}} \left(\frac{1}{2}\|M\|_{r*}^2 + \frac{1}{2\gamma}\|M - (\gamma N + Z)\|_F^2 + \langle Z, N\rangle \right) \\ &= \operatorname{prox}_{\frac{\gamma}{2}\|\cdot\|_{r*}^2}(\gamma N + Z).\end{aligned} \tag{19}$$

Using the extended Moreau-decomposition (see [Bauschke and Combettes, 2011, Theorem 14.3]) and Theorem 1, it holds that for all Z

$$\operatorname{prox}_{\frac{\gamma}{2}\|\cdot\|_{r*}^2}(Z) + \gamma \operatorname{prox}_{\frac{1}{2\gamma}\|\cdot\|_r^2}\left(\gamma^{-1} Z\right) = Z.$$

In combination with (19), we arrive at

$$\operatorname{prox}_{\gamma f}(Z) = \gamma N + Z - \gamma \operatorname{prox}_{\frac{1}{2\gamma}\|\cdot\|_r^2}\left(\frac{\gamma N + Z}{\gamma}\right). \tag{20}$$

Note that

$$\operatorname{prox}_{\frac{\gamma^{-1}}{2}\|\cdot\|_r^2}(Z) = \operatorname{prox}_{\frac{1}{2\gamma}(Z)\|\cdot\|^2},$$

which is why it is sufficient to derive how to compute $\operatorname{prox}_{\frac{\gamma}{2}\|\cdot\|_r^2}$. This is done in Algorithm 2 on page 126 for $r \in [1, \min\{m, n\}]$. Explanatory derivations can be found in Section A.5. Similar derivations based on the extended Moreau-decomposition are presented in [Eriksson et al., 2015] for integer-valued r.

Finally, observe that if $r \in \mathbb{N}$ and

$$\sigma_r(\gamma N + Z) > \left(1 + \gamma^{-1}\right) \sigma_{r+1}(\gamma N + Z), \tag{21}$$

it follows from the derivations of $\operatorname{prox}_{\frac{\gamma}{2}\|\cdot\|_r^2}$ (see Section A.5 and (51)) that

$$\operatorname{prox}_{\frac{1}{2\gamma}\|\cdot\|_r^2}\left(\frac{\gamma N + Z}{\gamma}\right) = \frac{\gamma N + Z}{\gamma} - \frac{1}{1+\gamma}\operatorname{svd}_r\left(\frac{\gamma N + Z}{\gamma}\right).$$

Therefore, (20) implies that

$$\operatorname{prox}_{\gamma f}(Z) = \frac{1}{1+\gamma}\operatorname{svd}_r\left(\gamma N + Z\right). \tag{22}$$

This fact is used in Section 4.4 to show a tight relationship to the non-convex Douglas-Rachford algorithm.

4.3 Douglas-Rachford limit point properties

A comparison between the Douglas-Rachford limit points and the optimality conditions for (A) and (B) (see Theorem 2) gives that all limit points $Z^\star = \lim_{k\to\infty} Z^k$ of (17c) can be expressed as

$$Z^\star = M^\star + \gamma D^\star, \tag{23}$$

where $D^\star$ and $M^\star$ are solutions to (A) and (B), respectively. Given $M^\star$, $Z^\star$ and γ, this allows us to determine $D^\star$. Moreover, by inspection of the Douglas-Rachford iterations, it can be shown that several known properties of the standard SVD-approximation remain true if they are preserved by $\text{prox}_g(X)$.

Proposition 5
Let N and g be as in Problem 1. Then the following hold:

i. Let $N \in \mathbb{S}$ and $\text{prox}_g(X) \in \mathbb{S}$ for all $X \in \mathbb{S}$. Then (A) and (B) have solutions $D^\star, M^\star \in \mathbb{S}$.

ii. Let $Nv = 0$ and $\text{prox}_g(X)v = 0$ for all X with $Xv = 0$. Then (B) has a solution $M^\star$ such that $M^\star v = 0$.

In particular, the solution to Problem 1 preserves these properties if (B) has a unique solution and there is no duality gap in (C). □

Proof Using [Watson, 1992, Theorem 2] it holds that $\text{prox}_{\frac{\gamma}{2}\|\cdot\|_{r*}^2}(X)$ has the same singular vectors as X. Therefore, $\text{prox}_{\frac{\gamma}{2}\|\cdot\|_{r*}^2}(X)$ preserves these properties and i. and ii. are proven by starting the Douglas-Rachford iterations for (B) with $Z^0 = 0$. The last claim follows with Proposition 3. □

There are numerous reasonable choices of g such that Proposition 5 applies, a few examples will be discussed in Sections 6 to 8.

According to Proposition 4, $\sigma_r(N + D^\star) \neq \sigma_{r+1}(N + D^\star)$ is a sufficient condition for the uniqueness of a solution to (B). Note that without this assumption, a solution to Problem 1 does not necessarily preserve the properties in Proposition 5. This can be used to construct non-trivial examples where $\sigma_r(N + D^\star) = \sigma_{r+1}(N + D^\star)$ (see Section 6.1).

4.4 Non-convex Douglas-Rachford (NDR)

Another approach to solve Problem 1 is to directly apply the Douglas-Rachford method to the non-convex problem

$$\min_{M\in\mathbb{R}^{n\times m}} \left[\frac{1}{2}\|N - M\|_F^2 + \chi_{\text{rank}(M)\leq r}(M) + g(M)\right]. \tag{24}$$

This has the advantage that we are guaranteed to get a solution of desired rank, if the iterates converge. Recently, some local convergence guarantees for the non-convex Douglas-Rachford have appeared in the literature (see [Hesse and Luke, 2013; Hesse et al., 2014; Phan, 2016]). Here, we add to these findings by showing that the non-convex Douglas-Rachford reduces locally to its convex counterpart if Proposition 4 applies. To this end, we start by deriving $\operatorname{prox}_{\gamma \bar{f}(Z)}$ where

$$\bar{f}(M) := \frac{1}{2}\|N - M\|_F^2 + \chi_{\operatorname{rank}(M)\leq r}(M)$$

for all $M \in \mathbb{R}^{n\times m}$. For $Z \in \mathbb{R}^{n\times m}$, we get

$$\begin{aligned}
\operatorname{prox}_{\gamma\bar{f}}(Z) &= \operatorname*{argmin}_{\substack{M\in\mathbb{R}^{n\times m}\\ \operatorname{rank}(M)\leq r}} \left(\frac{\gamma}{2}\|N - M\|_F^2 + \frac{1}{2}\|M - Z\|_F^2 \right) \\
&= \operatorname*{argmin}_{\substack{M\in\mathbb{R}^{n\times m}\\ \operatorname{rank}(M)\leq r}} \left(\frac{\gamma+1}{2}\|M\|_F^2 - \langle \gamma N + Z, M\rangle \right) \\
&= \operatorname*{argmin}_{\substack{M\in\mathbb{R}^{n\times m}\\ \operatorname{rank}(M)\leq r}} \left\| \frac{\gamma N + Z}{\gamma+1} - M \right\|_F^2.
\end{aligned}$$

Hence, by Proposition 1

$$\frac{1}{1+\gamma}\operatorname{svd}_r\left(\gamma N + Z\right) \in \operatorname{prox}_{\gamma\bar{f}}(Z).$$

Next let $D^\star$ and $M^\star$ be solutions to (A) and (B), respectively. If the convex Douglas-Rachford iterations are applied to (B), then it follows by (23) that $Z^\star = \gamma D^\star + M^\star$ is a limit point to (17c). Then, assuming that

$$\sigma_r(N + D^\star) \neq \sigma_{r+1}(N + D^\star),$$

it holds by Theorem 2 that $M^\star = \operatorname{svd}_r(N + D^\star)$, i.e.

$$\sigma_r(M^\star) = \sigma_r(N + D^\star) \quad \text{and} \quad \sigma_{r+1}(M^\star) = 0.$$

Therefore,

$$\begin{aligned}
(1+\gamma^{-1})\sigma_{r+1}(\gamma N + Z^\star) &= (1+\gamma^{-1})\sigma_{r+1}(\gamma(N + D^\star) + M^\star) \\
&= (1+\gamma)\sigma_{r+1}(N + D^\star) \\
&< (1+\gamma)\sigma_r(N + D^\star) \\
&= \sigma_r(\gamma(N + D^\star) + M^\star) \\
&= \sigma_r(\gamma N + Z^\star).
\end{aligned}$$

By the continuity of the singular values (see [Stewart and Sun, 1990, Corollary 4.9]), this allows us to conclude that (21) applies in a sufficiently small neighborhood of $Z^\star$. Thus, (22) implies that for all Z within this neighborhood

$$\text{prox}_{\gamma \bar{f}}(Z) = \text{prox}_{\gamma f}(Z),$$

where $f(M) := \frac{1}{2}\|M\|_{r*}^2 - \langle N, M\rangle + \frac{1}{2}\|N\|_{r*}^2$. As a result, the convex and non-convex Douglas-Rachford iterations locally coincide. Furthermore, there always exists a neighborhood that the Douglas-Rachford iterations cannot escape from, because the sequence $\|Z^\star - Z^k\|_F$ of the convex Douglas-Rachford is known to be non-increasing (see [Eckstein and Bertsekas, 1992]). This proves the *local convergence* of the non-convex Douglas-Rachford if $\sigma_r(N + D^\star) \neq \sigma_{r+1}(N + D^\star)$.

Notice that by Theorem 2 and (23) we can conclude that a zero duality-gap in (C) implies that both the convex and non-convex Douglas-Rachford have limit points corresponding to a solution to Problem 1 (even if $\sigma_r(N + D^\star) = \sigma_{r+1}(N + D^\star)$). We will see in Sections 6 to 8 that the non-convex Douglas-Rachford can converge to these solutions. However, this may not be the case for all choices of Z^0, since $\text{prox}_{\gamma \bar{f}}(Z)$ is not necessarily unique (see Section 6.1). Moreover, Section 7.3 shows that the choice of γ can be crucial for the existence of a limit-point of the non-convex Douglas-Rachford if there is a duality-gap in (C).

Finally, observe that $\text{prox}_{\gamma \bar{f}}(Z)$ only requires the determination of the dominant r singluar values and singluar vectors. Hence, sparse SVD solvers such as in [Liu et al., 2013] can be used to determine a dominant SVD, and to gain more computational speed with large-scale problems. The same holds true for $\text{prox}_{\gamma f}(Z)$, where maybe a larger, but not full, SVD needs to be determined.

5. Other Approaches

In the following we compare the $r*$-approach to other methods for solving Problem 1. These methods will also be used for numerical comparisons throughout the subsequent sections.

5.1 Nuclear Norm Regularization

One of the most widely used methods to approximate a solution to Problem 1 is the so-called nuclear norm regularization. It borrows techniques from sparse regularized regression, commonly called Lasso (see [Tibshirani, 1996]). This method estimates a sparse solution $\hat{x}$ to a linear system

of equations $A\hat{x} \approx b$ by solving

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \mu\|x\|_{\ell_1},$$

where $\|\cdot\|_2$ is the Euclidean norm, $\|x\|_{\ell_1} = \sum_{i\geq 1}|x_i|$, and $\mu \geq 0$ is a regularization parameter. In our case, rather than having a sparse solution, we are interested in having a small number of non-zero singular values. Therefore, for given $N \in \mathbb{R}^{n\times m}$, a corresponding matrix version reads

$$\min_{M\in\mathbb{R}^{n\times m}} \frac{1}{2}\|N - M\|_F^2 + \mu\|M\|_{1*} + g(M), \tag{25}$$

where $g : \mathbb{R}^{n\times m} \to \mathbb{R} \cup \{\infty\}$ is a given closed proper convex function. The simplicity of this convexification, as well as the results in [Fazel et al., 2001; Fazel, 2002; Recht et al., 2010], stimulated a large growth in the application of this method in many different areas (see [Fazel et al., 2001; Fazel, 2002; Olsson and Oskarsson, 2009; Recht et al., 2010]). However, it is often challenging to choose μ a priori in order to obtain a solution of specific rank. Commonly one assumes that the rank as a function of μ looks like a staircase, i.e. a large μ decreases the rank too much, whereas a small μ may leave it too large. In order to find the best possible approximation, one usually tries to keep μ as small as possible, which can result in a costly search.

In general, even with the best possible choice of μ, this heuristic does not return an optimal solution to Problem 1. Even in the simple case $g = 0$, one usually cannot choose μ such that the SVD-approximation is obtained. Finally, there is no certificate for checking whether a solution is a minimizer of Problem 1.

5.2 Rank Regularization

Similar to the nuclear norm regularization, it has been suggested in [Larsson et al., 2014; Larsson and Olsson, 2016] to directly regularize on the rank, i.e. solve

$$\min_{M\in\mathbb{R}^{n\times m}} \left[\frac{1}{2}\|N - M\|_F^2 + \mu\text{rank}(M) + g(M)\right],$$

where $\mu \geq 0$ is a regularization parameter and $g : \mathbb{R}^{n\times m} \to \mathbb{R} \cup \{\infty\}$ a closed and proper convex function. Since this problem is still non-convex, one needs to find a convex proxy of $f(M) := \frac{1}{2}\|N - M\|_F^2 + \mu\text{rank}(M)$. As shown in [Larsson et al., 2014; Larsson and Olsson, 2016], the conjugate

and bi-conjugate functions of f are given by

$$f^*(D) = \frac{1}{2}\|N+D\|_F^2 - \frac{1}{2}\|N\|_F^2 - \frac{1}{2}\sum_{i=1}^{\min\{m,n\}} \min\{2\mu, \sigma_i^2(N+D)\},$$

$$f^{**}(M) = \frac{1}{2}\|M-N\|_F^2 + \frac{1}{2}\sum_{i=1}^{\min\{m,n\}} \left(2\mu - \max\{0, \sqrt{2\mu} - \sigma_i(M)\}^2\right). \quad (26)$$

Hence, by Fenchel-duality (see Lemma A.1 and Proposition A.3) it holds that

$$\begin{aligned} \min_{M\in\mathbb{R}^{n\times m}} [f(M)+g(M)] &\geq -\min_{D\in\mathbb{R}^{n\times m}} [f^*(D)+g^*(-D)] \\ &= \min_{M\in\mathbb{R}^{n\times m}} [f^{**}(M)+g(M)]. \end{aligned} \quad (27)$$

Assume that there is no duality gap in (C) with solutions $D^\star$ and $M^\star$ to (A) and (B), respectively. Choosing $\frac{\sigma_r^2(N+D^\star)}{2} \geq \mu \geq \frac{\sigma_{r+1}^2(N+D^\star)}{2}$, it can be seen that

$$f^*(D^\star) = \frac{1}{2}\|N\|_F^2 - \frac{1}{2}\|N+D^\star\|_r^2 + \mu r = \frac{1}{2}\|N-M^\star\|_F^2 + \mu r + g(M^\star),$$

where the last equality follows by Propositions 2 and 3. Hence,

$$\begin{aligned} \frac{1}{2}\|N-M^\star\|_F^2 + \mu r + g(M^\star) &\geq -\min_{D\in\mathbb{R}^{n\times m}} [f^*(D)+g^*(-D)] \\ &\geq \frac{1}{2}\|N-M^\star\|_F^2 + \mu r + g(M^\star), \end{aligned}$$

yielding equality in (27). This shows that the method obtains the same guaranteed optimal solutions as previously discussed for (A) and (B). Evidently, there is a strong relationship to Propositions 2 and 4. However, if there is a duality-gap, then the solutions may differ from those with non-integer valued $r \in [1, \min\{m,n\}]$, and it is unclear which method yields better results. Moreover, even in the zero-duality gap case, a costly search for μ is required. Finally note that despite the fact that the proximal operator of f^{**} is computable (see [Larsson et al., 2014; Larsson and Olsson, 2016]), it is currently unknown if (26) is SDP-representable. This has the disadvantage that first order methods can be used only. Moreover, even for small dimensional examples, g is required to have a cheaply computable proximal operator.

5.3 Projection-based methods

In the following let $g(M) = \chi_{\mathcal{C}}(M)$ be the indicator function of a closed convex set $\mathcal{C}$. If the projection onto $\mathcal{C}$ is computable, then there are several other heuristics, of which a few are outlined next.

Lift-and-project Algorithm (LP) The idea of the so-called lift-and-project algorithm (see [Chu et al., 2003]) is to interchangeably perform a standard SVD-approximation of desired rank, and project the result orthogonally onto the convex set $\mathcal{C}$, which again increases the rank. By starting with N as the first iterate, one hopes to keep the distance to N small. Naturally, this algorithm always returns the standard SVD-approximation of N if it lies within $\mathcal{C}$. Unfortunately, it is generally difficult to know whether the algorithm converges, and if a possible limit point gives a satisfactory error (see [Chu et al., 2003]). However, if $\mathcal{C}$ is closed and $0 \in \mathcal{C}$, one can show that the Frobenius norm decreases in every step, and the convergence is guaranteed (see [Hiriart-Urruty and Lemaréchal, 2013, p. 118]).

Alternating Least-Squares (ALS) All the approaches considered so far share the drawback that when implemented, their iterates usually need to converge in order to guarantee a feasible solution. The so-called alternating least-squares method is a way of overcoming this drawback by working with iterates that lie in $\mathcal{C}$ and are of desired rank.

Given $V_0 \in \mathbb{R}^{r\times n} \setminus \{0\}$ such that $\{U \in \mathbb{R}^{m\times r} : UV_0 \in \mathcal{C}\} \setminus \{0\} \neq \emptyset$, one interchangeably solves

$$\begin{aligned} U_k &:= \operatorname*{argmin}_{UV_{k-1}\in\mathcal{C}} \|N - UV_{k-1}\|_F^2, \\ V_k &:= \operatorname*{argmin}_{U_kV\in\mathcal{C}} \|N - U_kV\|_F^2, \end{aligned}$$

with $k \geq 1$. Thus the rank constraint is explicitly taken into account by forming U_kV_k. Note that alternating least-squares without constraints converges for almost all V_0 to a standard SVD-approximation (see [Srebro, Jaakkola, et al., 2003]). The results in Section 6 indicate that for certain choices of $\mathcal{C}$, this method often converges to an optimal solution if $\sigma_r(N+D^\star) \neq \sigma_{r+1}(N+D^\star)$. Moreover, there are examples where its solution attains the lower-bound of Proposition 2 even though $\sigma_r(N + D^\star) = \sigma_{r+1}(N + D^\star)$. Nevertheless, in many cases ALS may not be a good choice since it is often unclear how to choose V_0.

6. Non-negative low-rank approximation

A particularly well studied low-rank approximation problem is the case of preserving non-negativity constraints.

Problem 2

$$\begin{aligned} &\text{minimize} && \|N - M\|_F^2 \\ &\text{subject to} && M \in \mathbb{R}^{n\times m}_{\geq 0} \end{aligned}$$

where $\mathbb{R}^{n\times m}_{\geq 0} := \{X \in \mathbb{R}^{n\times m} : x_{ij} \geq 0\}$ and $N \in \mathbb{R}^{n\times m}_{\geq 0}$. □

Note that this is the same as Problem 1 with $g = \chi_{\mathbb{R}^{n\times m}_{\geq 0}}$. Probably the most well-known approach to solving this problem is the so-called non-negative matrix factorization (see [Berry et al., 2007; Kim and Park, 2011]). Given $N \in \mathbb{R}^{n\times m}_{\geq 0}$, one intends to find a solution to

$$\min_{\substack{U\in\mathbb{R}^{n\times r}_{\geq 0},\\ V\in\mathbb{R}^{r\times m}_{\geq 0}}} \|N - UV\|_F^2.$$

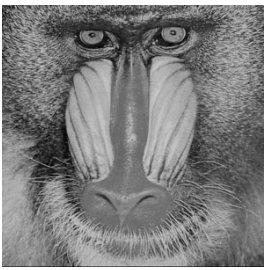
Non-negative matrix factorization (NNMF) is often approximately solved by applying alternating least-squares (see [Kim and Park, 2011] and Section 5.3). However, to require both U and V to be non-negative might be very conservative, since Problem 2 only requires that the product UV is non-negative.

6.1 Examples

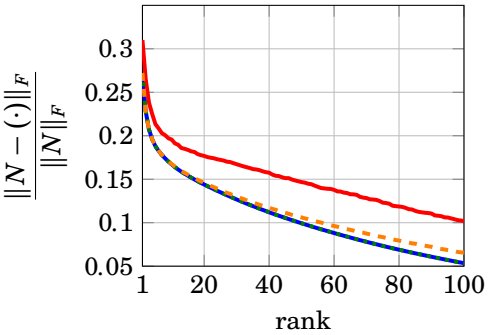
In the following we look at examples with a non-negativity constraint. The purpose is to illustrate several results that have been discussed in the previous sections.

Image compression A common example in the literature (see [Antoulas, 2005; Eldén, 2007]) is to use the SVD for image compression. Given a grey-scale picture, one maps the pixels to a matrix of corresponding grey-scale values, typically integer values in $\{0, \ldots, 255\}$, and performs a low-rank approximation of rank r. If r is sufficiently small, then the factors of the low-rank approximation are cheaper to store than the original matrix. Since the matrix is non-negative, it is very natural to keep this constraint intact.

We apply all the methods that have been discussed so far to the Baboon-image in Figure 3(a). A comparison among the relative errors of the methods, as well as the normalized lower-bound obtained from (B), is shown in Figure 3(b). By the Perron-Frobenius Theorem (see [Horn and Johnson, 2012, Theorem 8.4.4]) the rank-1 standard SVD-approximation is always non-negative. This reveals a major drawback of the nuclear norm heuristic for



(a) Baboon – 298 × 298 pixels.



(b) Relative errors:
- Nuclear norm heuristic
- Non-negative matrix factorization
- Lift-and-project
- $r*$-approach, lower bound, ALS, NDR

Figure 3. Non-negative Baboon-image approximation.

this problem, since it usually cannot recover standard SVD-approximations. Moreover, all the SVD-based methods produce results of similar quality. In fact, alternating least-squares (ALS), non-convex Douglas-Rachford (NDR) and the $r*$-approach give solutions that coincide numerically with the lower-bound, i.e. there is a zero duality gap for all ranks. The errors of the lift-and-project method are only slightly larger, and therefore not visible in this plot. Non-negative matrix factorization (based on alternating least-squares), however, tends to produce larger errors with increasing rank. Overall, the nuclear norm heuristic performs significantly worse than any of the other methods.

Asymmetric optimal approximations Let $N \in \mathbb{S} \cap \mathbb{R}^{n\times n}_{\geq 0}$ and $D^\star$ be a solution of (A) corresponding to Problem 2. By Proposition 4 and Proposition 5, we know that $\sigma_r(N + D^\star) \neq \sigma_{r+1}(N + D^\star)$ implies that $\operatorname{svd}_r(N + D^\star) \in \mathbb{S}$ is the unique solution to (B) and Problem 2. In the following it is shown that preservation of symmetry may no longer be valid for an optimal non-negative approximation if $\sigma_r(N + D^\star) = \sigma_{r+1}(N + D^\star)$.

Consider Problem 2 with $r = 2$ and

$$N = \begin{pmatrix} \frac{\sqrt{5}-1}{2} & 1 & 3 \\ 1 & 4 & 1 \\ 3 & 1 & \frac{\sqrt{5}-1}{2} \end{pmatrix}.$$

A non-symmetric solution is

$$M^\star = \begin{pmatrix} 0 & \frac{\sqrt{5}+1}{2} & \frac{\sqrt{5}+3}{2} \\ 2 & 3 & \frac{\sqrt{5}+1}{2} \\ 2 & 2 & 0 \end{pmatrix}.$$

Indeed, since N is symmetric, its singular values are given by the absolute value of its eigenvalues $\left\{\pm\frac{7-\sqrt{5}}{2}, 3+\sqrt{5}\right\}$. Then since $\|N - M^\star\|_F = \frac{7-\sqrt{5}}{2}$, and by Proposition 1, we conclude that $M^\star$ and $M^{\star T}$ are optimal non-negative rank-2 approximations of N. By Corollary 1 it follows that $D^\star = 0$ and $\sigma_2(N + D^\star) = \sigma_3(N + D^\star)$. Therefore, $M^\star$ and $M^{\star T}$ are solutions to (B).

Since the solution set of a convex problem is convex, all points

$$\alpha M^\star + (1-\alpha) M^{\star T} \text{ with} \alpha \in [0, 1]$$

are solutions to (B). However,

$$\operatorname{rank}\left(\alpha M^\star + (1-\alpha) M^{\star T}\right) = 3 \text{ for all } \alpha \in (0, 1)$$

This shows that we cannot expect to numerically find the rank-2 solutions by solving (B) (see Figure 2(c)). In particular, let either of the discussed Douglas-Rachford algorithms (see Sections 4.2 and 4.4) be initialized with $Z^0 \in \mathbb{S}$. Then Proposition 5 implies that they may converge to a symmetric solution, which can be shown to be non-optimal for Problem 2. Nevertheless, it is interesting that NDR and ALS often converge to an optimal solution under random initialization.

7. Matrix Completion

Assuming that the entries of a matrix are only partially known, the so-called matrix completion problem asks when and how the unknown elements can be recovered. The low-rank assumption turned out to be suitable for theoretical developments, as well as for many practical applications (see [Candès and Recht, 2009; Candès and Plan, 2010; Candès and Tao, 2010; Recht et al., 2010; Zare et al., 2016a]). This leads to the following problem.

Problem 3

$$\begin{aligned} &\text{minimize} && \operatorname{rank}(M) \\ &\text{subject to} && m_{ij} = z_{ij}, \ (i,j) \in \mathcal{I} \end{aligned} \tag{28}$$

where $\mathcal{I}$ is an index set. □

One of the most popular methods for solving Problem 3 is the technique introduced in [Candès and Recht, 2009]. It states that if $Z \in \mathbb{R}^{n\times n}$, then with high probability it is a solution to

$$\begin{aligned} &\text{minimize} && \|M\|_{1*} \\ &\text{subject to} && m_{ij} = z_{ij},\ (i,j) \in \mathcal{I}, \end{aligned} \tag{29}$$

under the additional assumption that $\text{card}(\mathcal{I}) \geq Cn^{1.2}\text{rank}(Z)\log(n)$, where $\text{card}(\mathcal{I})$ denotes the cardinality of $\mathcal{I}$ and C is a constant. Similar to that, it has been shown in [Recht et al., 2010] that (29) is able to detect a lowest rank solution. This means that one does not expect any other matrix of lower rank than Z having those partially known entries. Note that this formulation can be considered as a special case of Proposition 2 with $r = 1$, because

$$\begin{aligned} \min_{\substack{M\in\mathbb{R}^{n\times m}\\ \text{rank}(M)\leq r}} \left[\frac{1}{2}\|M\|_F^2 + g(M)\right] &\geq -\min_{D\in\mathbb{R}^{n\times m}} \left[g^*(-D) + \frac{1}{2}\|D\|_r^2\right] \\ &= \min_{M\in\mathbb{R}^{n\times m}} \left[\frac{1}{2}\|M\|_{r*}^2 + g(M)\right], \end{aligned} \tag{30}$$

where $g(M) = \chi_{\mathcal{M}}(M)$ and $\mathcal{M} := \{M \in \mathbb{R}^{n\times n} : m_{ij} = z_{ij},\ (i,j) \in \mathcal{I}\}$.

We suggest to utilize the flexibility in r and to consider instead

$$\begin{aligned} &\text{minimize} && \|M\|_{r*} \\ &\text{subject to} && m_{ij} = z_{ij},\ (i,j) \in \mathcal{I}, \end{aligned} \tag{31}$$

where it is possible to sweep over real-valued $r \geq 1$. In Sections 7.1 to 7.4 it is seen that this may significantly improve the quality of completion. Finally, let us determine when $Z \in \mathbb{R}^{n\times m}$ is a solution to (31).

THEOREM 3
Let $Z \in \mathbb{R}^{n\times m}$ with $r = \text{rank}(Z)$ and $\mathcal{I} \subset [1,\ldots,n] \times [1,\ldots,m]$. Then Z is a solution to (31) if and only if there exists $D^\star \in \mathbb{R}^{n\times m}$ with $Z = \text{svd}_r(D^\star)$ and $d^\star_{ij} = 0$ for all $(i,j) \notin \mathcal{I}$. □

Proof Let $g(M) = \chi_{\mathcal{M}}(M)$ and $\mathcal{M} := \{M \in \mathbb{R}^{n\times m} : m_{ij} = z_{ij},\ (i,j) \in \mathcal{I}\}$. Then

$$g^*(D) = \sup_{M\in\mathcal{M}} \langle D, M\rangle < \infty \ \Leftrightarrow\ \forall (i,j) \notin \mathcal{I} : d_{ij} = 0.$$

Hence, by Theorem 2, the existence of $D^\star$ such that $Z = \text{svd}_r(D^\star)$ is necessary for Z to be a solution to (31).

Assume that there exists $D^\star \in \mathbb{R}^{n\times m}$ such that $Z = \text{svd}_r(D^\star)$ and $d^\star_{ij} = 0$ for all $(i,j) \notin \mathcal{I}$. Then, by Theorem 2 it follows that $Z \in \partial\frac{1}{2}\|D^\star\|_r^2$. According

to Proposition A.4 this is equivalent to $D^\star \in \partial \frac{1}{2}\|Z\|_{r*}^2$ and therefore for all $\tilde{Z} \in \mathcal{M}$ it holds that

$$\frac{1}{2}\|\tilde{Z}\|_{r*}^2 \geq \frac{1}{2}\|Z\|_{r*}^2 + \langle D^\star, \tilde{Z} - Z\rangle = \frac{1}{2}\|Z\|_{r*}^2.$$

This shows the sufficiency, and concludes the proof. □

7.1 Some motivational examples

Next we want to demonstrate that $r > 1$ may help to complete matrices when $r = 1$ fails. To this end, consider the rank-2 matrices

$$Z_1 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad Z_2 = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad Z_3 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 4 \end{pmatrix}.$$

We would like to recover these matrices under the assumption that the zero entries are the only unknown ones. By Theorem 3 we know that this is possible with (31) and $r = 2$. By Proposition 5, it can shown that solving (29) is equivalent to determining

$$\min_{t \in \mathbb{R}} \|Z_i(t)\|_{1*}, \ i = 1, 2, 3 \tag{32}$$

where

$$Z_1(t) := \begin{pmatrix} t & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad Z_2(t) = \begin{pmatrix} 2 & t & 1 \\ t & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad Z_3(t) = \begin{pmatrix} t & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 4 \end{pmatrix}.$$

First we show that finding the lowest rank solution may not be sufficient to recover the true matrix. In case of Z_1 we get that $\operatorname{rank}(Z_1(t)) = 1$ if and only if $t = 1$. Moreover, for $u := \begin{pmatrix} -1 & 0.5 & 0.5 \end{pmatrix}^T$ it holds that $\|uu^T\|_F < \|Z_1(1)\|_F$ and $Z_1(1)u = 0$. Hence, as required by Theorem 3, $D^\star = Z_1(1) - uu^T$ guarantees that $Z_1(1)$ is the unique solution to (29) and therefore the nuclear norm heuristic does not recover Z_1.

Next we show that non-uniqueness in (29) is another issue that can be avoided with the proposed approach in (31). Since $Z_2(t)$ is symmetric, it holds that

$$\|Z_2(t)\|_{1*} \geq \operatorname{trace}(Z_2(t)) \equiv 5,$$

with equality if and only if $Z_2(t) \succeq 0$. It is readily seen that $Z_2(t) \succeq 0$ if and only if $t \in [0, 2]$, which implies that all of these points are solutions to the nuclear norm heuristic (29). However, a numerical solver for (29) does not necessarily determine Z_2. Instead, it is more likely to obtain a convex combination of these solutions.

Finally, observe that the nuclear norm heuristic does not always determine the lowest rank solution. It holds that $\operatorname{rank}(Z_3(t)) \geq 2$ with equality if and only if $t = 0$. However, it can be numerically verified that $\|Z_3\|_{1*} > \|Z_3(0.1)\|_{1*}$ and thus Z_3 is not a solution to (29).

These examples show that additional knowledge about the true rank, as well as the minimality in the Frobenius norm (see (30)), can be utilized with $\|\cdot\|_{r*}$ to possibly gain a better completion. The following subsections will demonstrate the same behavior for a larger example, and a practical application.

To conclude this subsection, note that in view of (25) one may also consider

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|M\|_F^2 + \mu\|M\|_{1*} \\ &\text{subject to} && m_{ij} = z_{ij},\ (i,j) \in \mathcal{I}, \end{aligned} \tag{33}$$

where one sweeps over $\mu \geq 0$. This is a strategy that has been discussed earlier in [Cai et al., 2010]. Applied to the previous examples, this approach is also able to recover Z_1, Z_2 and Z_3 with $\mu = 0$. Nevertheless, the following example shows that there may not be any μ that leads to a low-rank solution.

7.2 Numerical Example

This example intends to show a numerical comparison among (31) and (33). Let $Z = \operatorname{svd}_5(H)$ where $H \in \mathbb{R}^{10\times 10}$ is a Hankel matrix with the following structure

$$H = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ 1 & & \iddots & & 0 \\ \vdots & \iddots & \iddots & \iddots & \vdots \\ 1 & \iddots & \iddots & & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

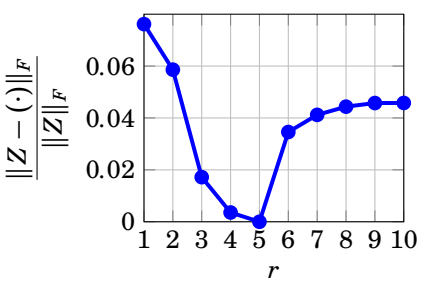
Moreover, let the index-set of the known entries be $\mathcal{I} = \{(i,j) : z_{ij} > 0\}$.

Figure 4 shows the relative completion errors, as well as the obtained ranks of the solutions to (31), for different integer-valued r. The corresponding results obtained by sweeping over $\mu \geq 0$ in (33) are presented in Figure 5.

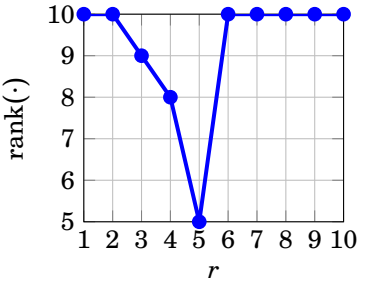
The solution to the nuclear norm heuristic ($r = 1$) gives the worst completion error, and full rank. Notice that

$$n^{1.2}\operatorname{rank}(Z)\log(n) \gg \operatorname{card}(\mathcal{I}) = 78,$$

which is why one cannot expect exact recovery. In contrast, $r = 5$ recovers the true matrix and is a sweet spot among all solutions. In fact, this is guaranteed by Theorem 3 because $\mathcal{I} \subset \{(i,j) : h_{ij} = 0\}$. Finally note that there is no μ such that $\operatorname{rank}(M_\mu^\star) < 10$.

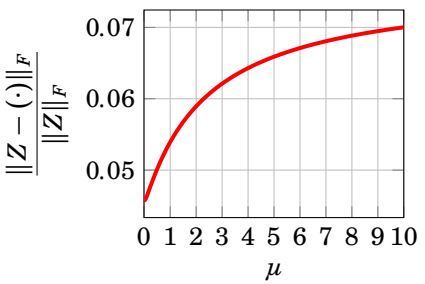


(a) Relative completion errors.

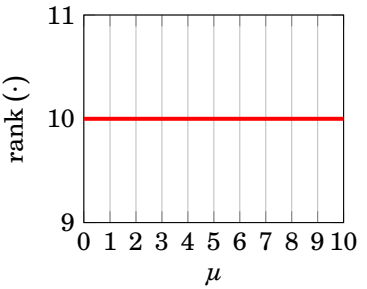


(b) Ranks.

Figure 4. Relative completion error and ranks of the solutions to (31) depending on r.



(a) Relative completion errors.



(b) Ranks.

Figure 5. Relative completion error and ranks of the solution to (33) depending on μ.

7.3 Example: Non-convex Douglas-Rachford

In the following we use Theorem 3 to construct examples where the nuclear norm heuristic, as well as the $r*$-approach, fail to determine a solution to Problem 3. This helps to understand why the non-convex Douglas-Rachford (see Section 4.4) may still be able to find those solutions and that, unlike in the convex Douglas-Rachford, the choice of γ is crucial.

First note that the existence of $D^\star$ in Theorem 3 is equivalent to having an $R \in \mathbb{R}^{n\times m}$ such that

$$R^T Z = 0,\ \ Z R^T = 0,\ \ \sigma_1(R) \leq \sigma_r(Z), \tag{34a}$$

$$z_{ij} + r_{ij} = 0 \text{ for all } (i,j) \notin \mathcal{I}. \tag{34b}$$

Let us define for $t \in [-1, 1]$ the following unitary rank-1 matrix

$$Z := \begin{pmatrix} t \\ \sqrt{1-t^2} \end{pmatrix} \begin{pmatrix} t & \sqrt{1-t^2} \end{pmatrix} = \begin{pmatrix} t^2 & t\sqrt{1-t^2} \\ t\sqrt{1-t^2} & 1-t^2 \end{pmatrix}.$$

Correspondingly, all $R \in \mathbb{R}^{2\times 2}$ that fulfill (34a) are given by

$$R = k \begin{pmatrix} \sqrt{1-t^2} \\ -t \end{pmatrix} \begin{pmatrix} \sqrt{1-t^2} & -t \end{pmatrix} = \begin{pmatrix} 1-t^2 & -t\sqrt{1-t^2} \\ -t\sqrt{1-t^2} & t^2 \end{pmatrix},$$

where $k \in [-1, 1]$. If $\mathcal{I} = \{(1, 2), (2, 1), (2, 2)\}$, it follows that (34b) can be satisfied if and only if $t^2 \leq \frac{1}{2}$. Hence, despite the fact that the solution to Problem 3 is unique, neither the nuclear norm heuristic nor the $r*$-approach is able to determine it if $t^2 > \frac{1}{2}$.

Next let us look at the limit-points of the non-convex Douglas-Rachford. Assume $X^\star$, $Y^\star$ and $Z^\star$ are limit-points to the iterations (17a) – (17c) of the non-convex Douglas-Rachford applied to

$$\min_{\substack{M\in\mathbb{R}^{n\times m} \\ \operatorname{rank}(M)\leq r}} \left[\frac{1}{2}\|M\|_F^2 + g(M) \right],$$

with $g(M) = \chi_{\mathcal{M}}(M)$ and $\mathcal{M} := \{M \in \mathbb{R}^{n\times n} : m_{ij} = z_{ij},\ (i, j) \in \mathcal{I}\}$.

By (17a) and (17c) it follows that $X^\star = \frac{1}{1+\gamma}\operatorname{svd}_r(Z^\star) = Y^\star$, and (17b) implies that

$$x_{ij}^\star - z_{ij}^\star = 0 \text{ for all } (i, j) \notin \mathcal{I}.$$

Equivalently, if $R^\star := Z^\star - \operatorname{svd}_r(Z^\star) = Z^\star - (1 + \gamma)X^\star$, then

$$\gamma x_{ij}^\star + r_{ij}^\star = 0 \text{ for all } (i, j) \notin \mathcal{I}.$$

Therefore, the non-convex Douglas-Rachford has a limit-point at $X^\star \in \mathcal{M}$ if and only if there exists $R \in \mathbb{R}^{n\times m}$ such that

$$R^T X^\star = 0,\ X^\star R^T = 0,\ \sigma_1(R) \leq (1 + \gamma^{-1})\sigma_r(X^\star),$$
$$x_{ij}^\star + r_{ij} = 0 \text{ for all } (i, j) \notin \mathcal{I}.$$

The inequality in the above follows from

$$(1 + \gamma)\sigma_r(X^\star) = \sigma_r(Z^\star) \geq \sigma_1(R^\star) = \gamma\sigma_1(\gamma^{-1}R^\star).$$

Thus, in the non-convex Douglas-Rachford, R is allowed to be $(1 + \gamma^{-1})$-times as large as in (34a). This means that for sufficiently small $\gamma > 0$, all rank-r elements in $\mathcal{M}$ are limit-points to the non-convex Douglas-Rachford. Applied to Z and $\mathcal{I}$ from above, we conclude that for all $t \in [-1, 1]$ there exists $\gamma > 0$ such that the non-convex Douglas-Rachford has a limit-point in Z. Indeed, in numerical computations, the algorithm also converges to Z. Just as in the convex Douglas-Rachford, a poor choice of γ may also prevent the existence of such a limit-point.

7.4 Covariance completion

Consider

$$\dot{x}(t) = Ax(t) + Bu(t),$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $m \leq n$, and $u(t)$ is a zero-mean stationary stochastic process. For Hurwitz A and reachable (A, B), it has been shown (see [Georgiou, 2002a; Georgiou, 2002b]) that the following are equivalent:

i. $X := \lim_{t\to\infty} \mathbf{E}\left(x(t)\, x^T(t)\right) \succeq 0$ is the steady-state covariance matrix of $x(t)$, where $\mathbf{E}(\cdot)$ denotes the expected value.

ii. $\exists H \in \mathbb{R}^{m\times n} : AX + XA^T = -(BH + H^T B^T)$.

iii. $\operatorname{rank}\begin{pmatrix} AX + XA^T & B \\ B^T & 0 \end{pmatrix} = \operatorname{rank}\begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}$.

In particular, $\operatorname{rank}(BH - H^T B^T)$ is an upper bound on the number of input channels, and $H = \frac{1}{2}\mathbf{E}\left(u(t)\, u^T(t)\right) B^T$ when u is white noise.

In [Chen et al., 2013; Lin et al., 2013; Zare et al., 2016a; Zare et al., 2015; Zare et al., 2016b] the problem of unknown B and only partially known X has been addressed by considering the following problem.

Problem 4

$$\begin{aligned} \text{minimize} \quad & \operatorname{rank}(M) \\ \text{subject to} \quad & \hat{x}_{ij} = x_{ij}, \ \ (i,j) \in \mathcal{I} \\ & A\hat{X} + \hat{X}A^T = -M \\ & \hat{X} \succeq 0. \end{aligned} \tag{35}$$

□

The problem has been tackled by convexifying the rank with the nuclear norm. However, since some practical examples only supply up to $2n$ known entries of specific structure (see [Zare et al., 2016a; Zare et al., 2015; Zare et al., 2016b]), it is not surprising that the quality of completion is often not satisfactory.

Instead, in [Grussler et al., 2016] its generalization as in (31) is considered, i.e.

$$\begin{aligned} \text{minimize} \quad & \|M\|_{r*} \\ \text{subject to} \quad & \hat{x}_{ij} = x_{ij}, \ \ (i,j) \in \mathcal{I} \\ & A\hat{X} + \hat{X}A^T = -M \\ & \hat{X} \succeq 0, \end{aligned} \tag{36}$$

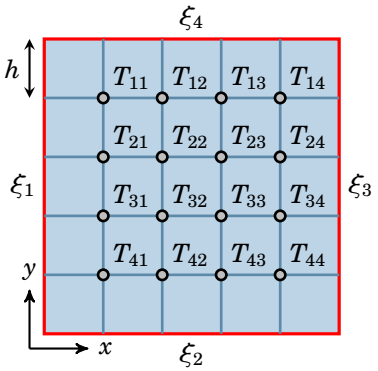


Figure 6. Discretized grid on the unit square with inputs $\xi_1, \ldots, \xi_4$.

where it is possible to sweep over $r \geq 1$. Again, one may also consider

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|M\|_F^2 + \mu\|M\|_{1*} \\ \text{subject to} \quad & \hat{x}_{ij} = x_{ij}, \ (i,j) \in \mathcal{I} \\ & A\hat{X} + \hat{X}A^T = -M \\ & \hat{X} \succeq 0, \end{aligned} \tag{37}$$

while sweeping over $\mu \geq 0$.

Example: Discretized Heat-Equation Let us illustrate these approaches by a numerical comparison. Consider the two-dimensional heat-equation

$$\dot{T} = \Delta T = \frac{\partial^2}{\partial x^2} T + \frac{\partial^2}{\partial y^2} T$$

on the unit-square. Finite difference discretization on a uniform grid with step size $h = \frac{1}{N+1}$ gives

$$\Delta T_{ij} \approx -\frac{1}{h^2}(4\,T_{ij} - T_{i+1,j} - T_{i,j+1} - T_{i-1,j} - T_{i,j-1}),$$

where T_{ij} are the temperatures of the inner grid points as indicated in Figure 6. By letting the boundaries of the unit-square be the inputs, we obtain a linear system

$$\dot{x}(t) = \frac{1}{h^2} A x(t) + \frac{1}{h^2} B\xi(t), \tag{38}$$

where $A \in \mathbb{R}^{N^2 \times N^2}$ is the Poisson-matrix, and $B = [b_{ij}] \in \mathbb{R}^{N^2 \times 4}$. The entries of B are all zeros, except:

$$\begin{aligned} b_{i1} &:= 1, && \text{for } i = 1, 2, \ldots, N \\ b_{i2} &:= 1, && \text{for } i = N, 2N, \ldots, N^2 \\ b_{i3} &:= 1, && \text{for } i = N(N-1)+1, N(N-1)+2, \ldots, N^2 \\ b_{i4} &:= 1, && \text{for } i = 1, N+1, \ldots, N(N-1)+1. \end{aligned}$$

Moreover, let $\xi(t)$ be generated by a low-pass filtered white-noise signal $w(t)$ with unit covariance $\mathbf{E}\left(w(t)w(t)^T\right) = I$, and

$$\dot{\xi}(t) = -\xi(t) + w(t).$$

The extended covariance matrix

$$X_e := \mathbf{E}\left(x_e x_e^T\right) = \begin{pmatrix} X & X_{x\xi} \\ X_{\xi x} & X_\xi \end{pmatrix} \text{ with } x_e := \begin{pmatrix} x(t) \\ \xi(t) \end{pmatrix}$$

is then determined by

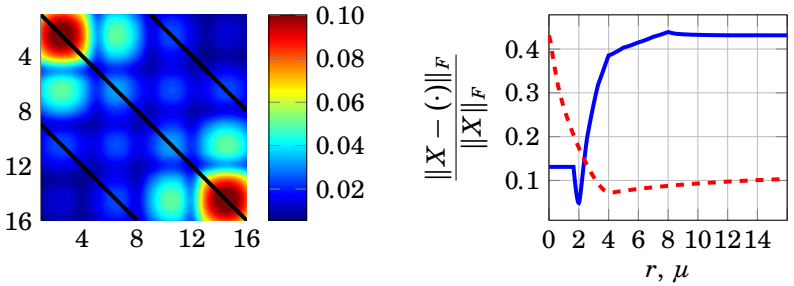
$$A_e X_e + X_e A_e^T = -B_e B_e^T,$$

where

$$A_e := \begin{pmatrix} A & B \\ 0 & -I \end{pmatrix}, \quad B_e := \begin{pmatrix} 0 \\ I \end{pmatrix},$$

and X is the steady-state covariance matrix of $x(t)$.

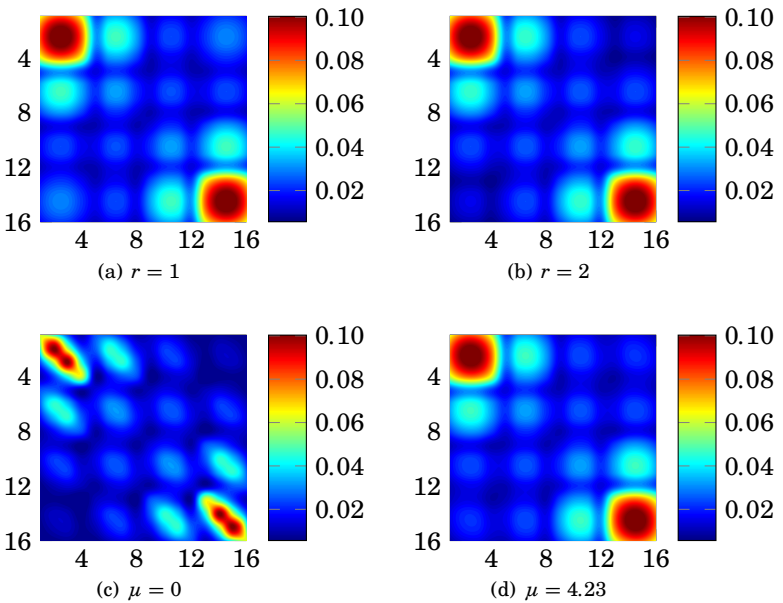
In the following we assume that only the first and third input channels are used, i.e. we remove the second and fourth columns from B and adjust A_e, B_e and $\xi(t)$, accordingly. An interpolated colormap of X is shown in Figure 7(a), where the black lines indicate the known entries. Figure 7(b) displays the relative completion error of the solutions obtained by (36) and (37) with dependency on r and μ. We observe that the error obtained by (36) in $r = 2$ is the smallest, and in fact it is of rank 2. This implies that there is no duality-gap. In contrast, the best solution that originates from (37) (with $\mu = 4.23$) is of rank 3 and has an error that is about 1.5 times as large. Figure 8 illustrates these differences through the interpolated colormaps.



(a) True covariance matrix X with inputs ξ_1 and ξ_3 and ━━ indicating the known entries.

(b) Relative errors:
━━ $r*$-approach dependent on r
- - - nuclear norm heuristic (37) dependent on μ.

Figure 7. Interpolated colormap of the true steady-state covariance matrix X and plot of the relative errors depending on r and μ obtained by (36) and (37).



(a) $r = 1$ (b) $r = 2$

(c) $\mu = 0$ (d) $\mu = 4.23$

Figure 8. Interpolated colormaps of the completed covariance matrices obtained by (36) and (37).

8. Hankel matrices

In the field of system and control, the rank of a Hankel operator/matrix is crucial, because it determines the complexity (order) of a linear system. This determines how costly it is to simulate a system, or to implement controllers designed by a number of standard methods (see [Zhou et al., 1996; Antoulas, 2005]). For these reasons, much focus is put into model order reduction. Even though the celebrated Adamyan-Arov-Krein theorem (see [Partington, 1989; Antoulas, 2005]) answers the question of optimal low-rank approximation of infinite dimensional Hankel operators, the following finite dimensional case is still an open problem.

Problem 5

$$\begin{aligned} \underset{M}{\text{minimize}} \quad & \|N - M\|_F^2 \\ \text{subject to} \quad & \operatorname{rank}(M) \leq r \\ & M \in \mathcal{H} \end{aligned}$$

where $N \in \mathcal{H} := \{H : H \text{ is Hankel}\}$. □

The only solved variant of Problem 5 is the case where $r = 1$, and the Frobenius norm is replaced by the spectral norm (see [Antoulas, 1997]). Moreover, for so-called linear externally positive systems the problem of non-negativity preserving Hankel-operator approximation has been considered in [Grussler and Rantzer, 2014].

8.1 Numerical Example

In the following we compare the $r*$-approach with the regularization methods in Sections 5.1 and 5.2, as well as the lift-and-project algorithm from Section 5.3. To this end, let $N \in \mathbb{R}^{10\times 10}$ be the following Hankel matrix

$$N = \begin{pmatrix} 1 & 2 & \cdots & 9 & 10 \\ 2 & & \iddots & & 9 \\ \vdots & \iddots & \iddots & \iddots & \vdots \\ 9 & & \iddots & & 2 \\ 10 & 9 & \cdots & 2 & 1 \end{pmatrix}.$$

The relative errors together with the relative lower bound are shown in Figure 9(a), where for each method and rank the solution with lowest possible error has been chosen. In case of $r = 1, \ldots, 4$ there is a zero duality gap, and therefore the lower bound is achieved by the rank-regularization method (see Section 5.2) as well as the non-convex Douglas-Rachford and the $r*$-norm. Moreover, even when Proposition 4 cannot guarantee a zero

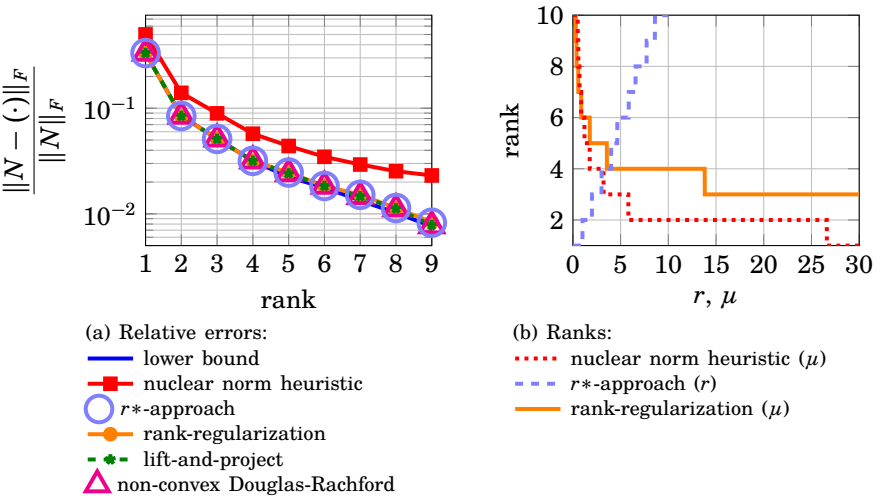


Figure 9. Hankel matrix preservation – relative error and rank dependency on r and μ

duality gap, it appears that those methods and the lift-and-project algorithm are close to the lower bound, and outperform the nuclear norm heuristic. Nonetheless, in order to get these (sub-optimal) solutions, we had to sweep over real-valued r and μ, respectively. The dependency of the rank on these parameters is displayed in Figure 9(b). All regularization methods show the expected staircase behavior.

9. Multivariate Reduced-Rank Regression

In multivariate linear regression one wants to estimate a regression matrix $C \in \mathbb{R}^{n\times m}$ in an underlying linear model

$$Y = CX + E,$$

where $Y \in \mathbb{R}^{n\times K}$ is a matrix with K measurements of n response variables, $X \in \mathbb{R}^{m\times K}$ are the corresponding predictor variables, and $E \in \mathbb{R}^{m\times K}$ is Gaussian white-noise. Assuming that $\operatorname{rank}(X) = m < T$ one can determine the well-known least-squares estimator

$$\hat{C} = YX^T(XX^T)^{-1},$$

which is a minimizer of $\min_C \|Y - CX\|_F^2$. Let $\hat{c}_k$ and y_k denote the k-th row of $\hat{C}$ and Y, respectively. Then

$$\hat{c}_k = y_k X^T(XX^T)^{-1},$$

and therefore $\hat{c}_k$ only depends on the k-th response variable y_k. Hence, the estimator does not account for possible correlations among the response variables.

In order to get estimators that include these correlations, one may restrict oneself to $\operatorname{rank}(C) = r < \min\{m, n\}$ (see [Izenman, 1975; Reinsel and Velu, 1998]). Assuming that $C = AB$, where $A \in \mathbb{R}^{n\times r}$ and $B \in \mathbb{R}^{r\times m}$, a physical interpretation of this assumption on C can be given (see [Reinsel and Velu, 1998]) as follows. If X consists of information that is used to send T messages Y over r channels, then BX can be considered as a code for the information, and ABX the decoded messages which are intended to be close to Y. Hence, given X, Y and r one would like to solve the problem

Problem 6

$$\begin{aligned} \underset{C}{\text{minimize}} \quad & \|Y - CX\|_F^2 \\ \text{subject to} \quad & \operatorname{rank}(C) \leq r. \end{aligned}$$

□

Assuming that $\operatorname{rank}(X) = m < K$, an explicit solution can be determined as follows. Let $X = U \begin{pmatrix} \Sigma & 0 \end{pmatrix} \begin{pmatrix} V_1 & V_2 \end{pmatrix}^T$ be an SVD of X with $\Sigma \in \mathbb{R}^{m\times m}$. Then

$$\|Y - CX\|_F^2 = \|Y \begin{pmatrix} V_1 & V_2 \end{pmatrix} - \begin{pmatrix} CU\Sigma & 0 \end{pmatrix}\|_F^2 = \|YV_1 - CU\Sigma\|_F^2 + \|YV_2\|_F^2.$$

Hence, Problem 6 reduces to

$$\begin{aligned} \underset{\tilde{C}}{\text{minimize}} \quad & \|YV_1 - \tilde{C}\|_F^2 \\ \text{subject to} \quad & \operatorname{rank}(\tilde{C}) \leq r. \end{aligned} \tag{39}$$

By Proposition 1 we know that a minimizer of (39) is given by $\operatorname{svd}_r(YV_1)$, and therefore $\hat{C} = \operatorname{svd}_r(YV_1)\Sigma^{-1}U^T$ is a solution to Problem 6. Observe that Problem 6 can also be stated as

$$\begin{aligned} \underset{M}{\text{minimize}} \quad & \frac{1}{2}\|Y - M\|_F^2 + \chi_{\mathcal{L}}(M) \\ \text{subject to} \quad & \operatorname{rank}(M) \leq r, \end{aligned}$$

where $\mathcal{L} = \{M : M = CX \text{ for some } C \in \mathbb{R}^{n\times m}\}$ and thus fits into the scope of Proposition 2. Indeed, if $\operatorname{rank}(X) = m$, then $\operatorname{rank}(M) = \operatorname{rank}(C)$, and solving

$$\underset{M}{\text{minimize}} \quad \frac{1}{2}\|M\|_{r*}^2 - \langle Y, M\rangle + \chi_{\mathcal{L}}(M) \tag{40}$$

leads to the same solution as above if $\text{svd}_r(YV_1)$ is unique. This can be shown by considering the dual of (40). By Proposition 2 we get

$$\begin{aligned} \underset{D}{\text{maximize}} \quad & \frac{1}{2}\|Y + D\|_r^2 \\ \text{subject to} \quad & DX^T = 0, \end{aligned}$$

where $D^\star = Y(V_1^T V_1 - I)$ is a feasible maximizer such that

$$\text{svd}_r(Y + D^\star) = \text{svd}_r(YV_1V_1^T) = \hat{C}X$$

is a solution to (40).

Finally notice that further convex constraints on C can be added to (40), and by that more classes of regressors can be defined and computed.

10. Discussion and Future Developments

In this work, a method to determine optimal low-rank approximations with convex constraints has been studied. The main benefits of the $r*$-approach are that it is essentially regularization parameter free, gives a certificate of optimality, and does not depend on a particular initialization. This combines the benefits of both factor and regularization based methods. Moreover, we have seen that the $r*$ approach can be turned into a regularization dependent method, where, unlike other approaches, the parameter has a clear relationship to the desired rank (see Section 3.2). As a result, a generalization of (29) to solve the matrix completion problem has been suggested. Furthermore, we have linked this approach to the rank-regularization method (see Section 5.2). The principal advantage here is that the $r*$-norm, in contrast to the rank-regularization method, is known to have an SDP-representation.

Since standard interior-point methods for SDPs are known to have iterations that grow unfavorably with dimension, the Douglas-Rachford splitting algorithm is used to gain computability for problems of larger dimensions. In this setting, it was possible to show that several other useful properties known from the SVD-solution may be preserved (see Proposition 5). Moreover, it allowed us to show local convergence of the non-convex Douglas-Rachford if Proposition 4 applies. This motivates the overall usefulness of the non-convex Douglas-Rachford for solving Problem 1. This work is merely a starting point to investigating its power for the problems considered here. Further developments in this direction are likely to contribute to a better understanding of the duality-gap cases. One could start by linking the results in Section 4.4 to the known local convergence results in the vector case (see [Hesse et al., 2014]).

The numerical examples presented in this paper indicate the superiority of the $r*$-approach and others over the nuclear-norm heuristic. Since the $r*$-approach is as general as the nuclear-norm heuristic, we suggest to use the $r*$-norm heuristic instead. In fact, several other authors (see [Argyriou et al., 2012; Lai et al., 2014; McDonald et al., 2015; Doan and Vavasis, 2016]) have recently used the $r*$-norm to replace the nuclear-norm as a regularizer in (33). However, this neither takes advantage of its own regularization character nor its optimality guarantees. Notice that despite the nice geometric interpretation (see Section 3.1), we were only able to guarantee a zero duality gap in simple cases such as Theorem 3. Investigating this further may lead to more deterministic guarantees.

A. Appendix

A.1 Unitarily invariant norms

The following results can be found e.g. in [Horn and Johnson, 2012].

Proposition A.1
Let $A, B \in \mathbb{R}^{n\times m}$, then

$$\langle A, B\rangle \leq \sum_{i=1}^{\min\{m,n\}} \sigma_i(A)\sigma_i(B). \qquad \square$$

Corollary A.1
Let $A, B \in \mathbb{R}^{n\times m}$ then

$$\sum_{i=1}^{\min\{m,n\}} \sigma_i(A)\sigma_i(B) = \max\{\langle A, UBV\rangle : U \text{ and } V \text{ are unitary}\}. \qquad \square$$

In the following we say that $g : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is a symmetric gauge function if and only if

i. g is a norm.

ii. $\forall x \in \mathbb{R}^n : g(|x|) = g(x)$, where $|x|$ denotes the element-wise absolute value.

iii. $g(Px) = g(x)$ for all permutation matrices P and all x.

Proposition A.2
$\|\cdot\|$ is a unitarily invariant norm on $\mathbb{R}^{n\times m}$ if and only if

$$\|X\| = g(\sigma_1(X), \ldots, \sigma_{\min\{m,n\}}(X)),$$

where g is a symmetric gauge function. $\square$

A.2 Convex Optimization

The following definitions and results from convex optimization (see [Luenberger, 1968; Rockafellar, 1970; Hiriart-Urruty and Lemaréchal, 1996; Bauschke and Combettes, 2011]) are used throughout the paper. In the following we assume that all functions are defined on a real finite-dimensional Hilbert space X with inner product $\langle \cdot, \cdot \rangle$. The *domain* of a function f on X is defined as $\mathrm{dom} f := \{x \in X : f(x) < \infty\}$.

Definition A.1
Let $f : H \to \mathbb{R} \cup \{\infty\}$ be a function with $\mathrm{dom} f \neq \emptyset$, minorized by an affine function i.e. $\exists (x^*, b) \in X \times \mathbb{R} : f(x) \geq \langle x, x^* \rangle - b$ for all $x \in X$. Then,

$$f^*(x^*) := \sup_{x \in X} [\langle x, x^* \rangle - f(x)]$$

is called its conjugate (dual) function. Further, the bi-conjugate function of f is defined as $f^{**} := (f^*)^*$. □

Definition A.2
A convex function $f : H \to \mathbb{R} \cup \{\infty\}$ with $\mathrm{dom} f \neq \emptyset$ is

- *proper* if $\mathrm{dom} f \neq \emptyset$.
- *closed* if the epigraph $\{(t, x) : f(x) \leq (x, t) \in \mathrm{dom} f\}$ is a closed set. □

It is known that $f^{**} = f$ if only if f is a closed and proper convex function.

Lemma A.1
Let $f, g : H \to \mathbb{R} \cup \{\infty\}$ be functions as in Definition A.1. Then

$$\inf_{x \in X} [f(x) + g(x)] \geq - \inf_{x \in X} [f^*(x) + g^*(-x)]. \tag{41}$$

□

Proposition A.3
Let $f, g : H \to \mathbb{R} \cup \{\infty\}$ be closed and proper convex functions. Assume that $\mathrm{ri}(\mathrm{dom} f) \cap \mathrm{ri}(\mathrm{dom} g) \neq \emptyset$ and $\mathrm{ri}(\mathrm{dom} f^*) \cap \mathrm{ri}(\mathrm{dom} g^*) \neq \emptyset$, where $\mathrm{ri}(\cdot)$ denotes the relative interior. Then,

$$\min_{x \in X} [f(x) + g(x)] = - \min_{x^* \in X} [f^*(x^*) + g^*(-x^*)].$$

Moreover, if the minimum on the left is attained at some x_0 and the minimum on the right by some x_0^*, then

$$\begin{aligned} f^*(x_0^*) &= \langle x_0, x_0^* \rangle - f(x_0), \\ g^*(-x_0^*) &= \langle x_0, -x_0^* \rangle - g(x_0). \end{aligned}$$

□

Definition A.3
Let $f : H \to \mathbb{R} \cup \{\infty\}$ be a function. Then

$$\partial f(x_0) := \{x_0^* \in X : f(x) \geq f(x_0) + \langle x - x_0, x_0^* \rangle \text{ for all } x \in X\}$$

is called the subdifferential of f at x_0. □

Proposition A.4
Let $f : H \to \mathbb{R} \cup \{\infty\}$ be a closed and proper convex function. Then the following statements are equivalent:

i. $x_0^* \in \partial f(x_0)$.

ii. $f^*(x_0^*) = \langle x_0, x_0^* \rangle - f(x_0)$.

iii. $x_0 \in \partial f^*(x_0^*)$.

iv. $f(x_0) = \langle x_0, x_0^* \rangle - f^*(x_0^*)$. □

Lemma A.2
Let $\mathcal{C}_1, \mathcal{C}_2 \subset X$ be closed convex sets. Then, $\mathcal{C}_1 = \mathcal{C}_2$ if and only if

$$\sup_{y \in \mathcal{C}_1} \langle x, y \rangle = \sup_{y \in \mathcal{C}_2} \langle x, y \rangle \text{ for all } x \in X.$$

For $x \in \mathbb{R}^n$ and $r \in [1, n]$ we define $\|x\|_r := \sqrt{g_r(x)}$ with

$$g_r(x) := \max\{x_{i_1}^2 + \cdots + x_{i_{\lceil r \rceil - 1}}^2 + (r - \lfloor r \rfloor) x_{i_{\lceil r \rceil}} : 1 \leq i_1 < i_2 < \cdots < i_{\lceil r \rceil} \leq n\}.$$

The following Lemma on the subgradients of $\|\cdot\|_r$ has been shown in [Doan and Vavasis, 2016] for $r \in \mathbb{N}$. We simply extend it to the real-valued case.

Lemma A.3
Let $r \in [1, n]$, $\bar{r} := \lceil r \rceil$ and $\sigma \in \mathbb{R}^n_{\geq 0}$ with $\sigma \neq 0$ and

$$\sigma_1 \geq \cdots > \sigma_{\bar{r}-t+1} = \cdots = \sigma_{\bar{r}} = \cdots = \sigma_{\bar{r}+s} > \cdots \geq \sigma_n, \tag{42}$$

where $t = \bar{r}$ and $s = n - \bar{r}$ if $\sigma_1 = \sigma_{\bar{r}}$ and $\sigma_n = \sigma_{\bar{r}}$, respectively. Then $v \in \partial\|\sigma\|_r$ if and only if

i. $1 \leq i \leq \bar{r} - t$: $v_i = \frac{\sigma_i}{\|\sigma\|_r}$.

ii. $\bar{r} - t + 1 \leq i \leq \bar{r} + s$: $v_i = \tau_i \frac{\sigma_{\bar{r}}}{\|\sigma\|_r}$ with $0 \leq \tau_i \leq 1$, $\sum_{i=\bar{r}-t+1}^{\bar{r}+s} \tau_i = t - \bar{r} + r$.

iii. $\bar{r} + s + 1 \leq i \leq n$: $v_i = 0$.

Moreover, let $\|x\|_{r*}$ be the dual norm to $\|\sigma\|_r$. Then

$$\partial\|0\|_r = \{x \in \mathbb{R}^n : \|x\|_{r*} \leq 1\}.$$ □

Proof Let $r \in [1, n]$ and $\sigma \in \mathbb{R}^n_{\geq 0}$ as in (42). Then

$$\|\sigma\|_r = \max_{\substack{\mathcal{I}\subset\{1,\dots,n\} \\ \operatorname{card}(\mathcal{I})=\bar{r}}} g_{\mathcal{I}}(\sigma),$$

where $g_{\mathcal{I}}(\sigma) := \sqrt{\sum_{i\in\mathcal{I}\setminus\max(\mathcal{I})} \sigma_i^2 + (r - \lfloor r \rfloor)\sigma^2_{\max(\mathcal{I})}}$ and $\operatorname{card}(\mathcal{I})$ denotes the cardinality of $\mathcal{I}$. Since $\|\sigma\|_r \neq 0$ it follows (see [Hiriart-Urruty and Lemaréchal, 2013, Corollary VI.4.4.4]) that the sub-differentials of $\|\cdot\|_r$ evaluated at σ are given by

$$\partial\|\sigma\|_r = \operatorname{conv}\{\nabla g_{\mathcal{I}}(\sigma) : \ \mathcal{I} \subset \{1, \dots, n\}, \ \operatorname{card}(\mathcal{I}) = \bar{r}, \ g_{\mathcal{I}}(\sigma) = \|\sigma\|_r\}, \tag{43}$$

where ∇ denotes the gradient operator with respect to σ. Next we determine the gradient at these points where $\|\sigma\|_r = g_{\mathcal{I}}(\sigma)$. Then, by assumption (42) it holds that $\{1, \dots, \bar{r} - t\} \subset \mathcal{I}$ and therefore

- $1 \leq i \leq \bar{r} - t$: $\frac{\partial g_{\mathcal{I}}(\sigma)}{\partial \sigma_i} = \frac{\sigma_i}{\|\sigma\|_r}$.
- $i \in \mathcal{I} \cap \{\bar{r} - t + 1, \dots, \bar{r} + s\} \setminus \max(\mathcal{I})$: $\frac{\partial g_{\mathcal{I}}(\sigma)}{\partial \sigma_i} = \frac{\sigma_i}{\|\sigma\|_r}$.
- $i = \max(\mathcal{I})$: $\frac{\partial g_{\mathcal{I}}(\sigma)}{\partial \sigma_{\bar{r}}} = \frac{(r - \lfloor r \rfloor)\sigma_i}{\|\sigma\|_r}$.
- $\bar{r} + s + 1 \leq i \leq n$: $\frac{\partial g_{\mathcal{I}}(\sigma)}{\partial \sigma_i} = 0$.

Thus, by (43) we get that $v \in \partial\|\sigma\|_r$ if and only if

i. $1 \leq i \leq \bar{r} - t$: $v_i = \frac{\sigma_i}{\|\sigma\|_r}$,

ii. $\bar{r}-t+1 \leq i \leq \bar{r}+s : v_i = \tau_i \frac{\sigma_{\bar{r}}}{\|\sigma\|_r}$ with $0 \leq \tau_i \leq 1$ and $\sum_{i=\bar{r}-t+1}^{\bar{r}+s} \tau_i = t-\bar{r}+r$,

iii. $\bar{r} + s + 1 \leq i \leq n$: $v_i = 0$,

where the last part of the second condition follows from

$$\sum_{i\in\mathcal{I}} \frac{\partial g_{\mathcal{I}}(\sigma)}{\partial \sigma_i} = (t - \bar{r} + r)\frac{\sigma_{\bar{r}}}{\|\sigma\|_r}.$$

The last claim simply follows by the definition of the dual norm as

$$\begin{aligned}\partial\|0\|_r = \{x_0 \in \mathbb{R}^n : \langle x, x_0\rangle \leq \|x\|_r\} &= \{x_0 \in \mathbb{R}^n : \sup_{\|x\|_r\leq 1} \langle x, x_0\rangle \leq 1\} \\ &= \{x_0 \in \mathbb{R}^n : \|x_0\|_{r*} \leq 1\}. \qquad \square\end{aligned}$$

From Lemma A.3 and [Watson, 1992, Theorem 2] the following Proposition follows in the same way as in [Doan and Vavasis, 2016] for $r \in \mathbb{N}$.

PROPOSITION A.5
Let $A \in \mathbb{R}^{n\times m} \setminus \{0\}$, $r \in [1, \min\{m,n\}]$ and $\bar{r} := \lceil r \rceil$. Further, let an SVD of A be given by $A = \sum_{i=1}^{\min\{m,n\}} \sigma_i u_i v_i^T$ with

$$\sigma_{\bar{r}-t} \neq \sigma_{\bar{r}-t+1} = \cdots = \sigma_{\bar{r}} = \cdots = \sigma_{\bar{r}+s} \neq \sigma_{\bar{r}+s+1},$$

where $t = \bar{r}$ and $s = \min\{m,n\} - \bar{r}$ if $\sigma_1 = \sigma_{\bar{r}}$ and $\sigma_{\min\{m,n\}} = \sigma_{\bar{r}}$, respectively. Then $M \in \partial\|A\|_r$ if and only if

$$M = \frac{1}{\|A\|_r}\left(\sum_{i=1}^{\bar{r}-t} \sigma_i u_i v_i^T + \sigma_{\bar{r}} \begin{pmatrix} u_{\bar{r}-t+1} & \dots & u_{\bar{r}+s} \end{pmatrix} T \begin{pmatrix} v_{\bar{r}-t+1} & \dots & v_{\bar{r}+s} \end{pmatrix}^T\right),$$

where

$$T \succeq 0, \quad \|T\|_{1*} = t + \bar{r} - r, \quad \text{and} \quad \|T\|_1 \leq 1.$$

In particular, if $\sigma_{\bar{r}} \neq \sigma_{\bar{r}+1}$ or $\sigma_{\bar{r}} = 0$ then $\operatorname{rank}(M) \leq \bar{r}$. Moreover,

$$\|0\|_r = \{M \in \mathbb{R}^{n\times m} : \|M\|_{r*}\}. \qquad \square$$

A.3 Proof of Lemma 1

Proof Let $1 \leq r \leq q := \min\{m,n\}$, $M \in \mathbb{R}^{n\times m}$ and the $g : \mathbb{R}^q \to \mathbb{R}_{\geq 0}$ be defined by

$$g(x_1, \dots, x_q) := \|\operatorname{diag}(x_1, \dots, x_q)\|_r.$$

The unitary invariance of $\|\cdot\|_r$ follows by Proposition A.2, because g is a symmetric gauge function. By Corollary A.1 it holds that

$$\sup_{\substack{\|X\|_F=1 \\ \operatorname{rank}(X)\leq r}} \langle X, M\rangle = \sup_{\sum_{i=1}^r \sigma_i(X)=1} \sum_{i=1}^r \sigma_i^2(X)\sigma_i(M) = \|M\|_r.$$

Then the $r*$-norm inherits the unitary invariance of the r-norm and with $\Sigma := \operatorname{diag}(\sigma_1(M), \dots, \sigma_q(M))$ it follows that

$$\begin{aligned}
\|M\|_{r*} = \|\Sigma\|_{r*} &= \max_{\|X\|_r \leq 1} \langle \Sigma, X\rangle \\
&= \max_{\sum_{i=1}^r \sigma_i^2(X)=1} \sum_{i=1}^q \sigma_i(M)\sigma_i(X) \\
&= \max_{\sum_{i=1}^r \sigma_i^2(X)\leq 1} \left[\sum_{i=1}^r \sigma_i(M)\sigma_i(X) + \sigma_r(X) \sum_{i=r+1}^q \sigma_i(M)\right].
\end{aligned}$$

The third equality follows by Corollary A.1. Hence,

$$\|M\|_{1*} = \max_{\sum_{i=1}^{1} s_i^2 = 1} \sum_{i=1}^{q} \sigma_i(M) s_i \geq \cdots \geq \max_{\sum_{i=1}^{q} s_i^2 = 1} \sum_{i=1}^{q} \sigma_i(M) s_i = \|M\|_{q*} = \|M\|_F.$$

Moreover, by the definition of the r-norm

$$\|M\|_F = \|M\|_q \geq \cdots \geq \|M\|_1$$

and therefore (2) is shown. In particular,

$$\|M\|_{r*} = \max_{\sum_{i=1}^{r} s_i^2 = 1} \sum_{i=1}^{q} \sigma_i(M) s_i \geq \|M\|_F \geq \max_{\sum_{i=1}^{r} s_i^2 = 1} \sum_{i=1}^{r} \sigma_i(M) s_i = \|M\|_r. \quad (44)$$

Obviously, $\|M\|_F = \|M\|_r$ if and only $\operatorname{rank}(M) \leq r$, and therefore equality in (44) holds if and only if $\operatorname{rank}(M) \leq r$. Thus the last claim is proven. □

A.4 Proof of Theorem 2

Proof If $D^\star$ and $M^\star$ are solutions to (A) and (B), respectively, then by Proposition A.3 it holds that

$$f^{**}(M^\star) = \langle D^\star, M^\star \rangle - f^*(D^\star),$$

where f^* and f^{**} are given by (5) and (6). Hence, by Proposition A.4 it follows that

$$M^\star \in \partial_D \frac{1}{2}\|N + D\|_r^2 \Big|_{D=D^\star} = \|N + D^\star\|_r \partial_D \|N + D\|_r|_{D=D^\star}$$

and invoking Proposition A.5 proves the result. □

A.5 Derivation of $\operatorname{prox}_{\frac{\gamma}{2}\|\cdot\|_r^2}(\cdot)$

$$\operatorname{prox}_{\frac{\gamma}{2}\|\cdot\|_r^2}(Z) = \operatorname*{argmin}_X \left(\frac{\gamma}{2}\|X\|_r^2 + \frac{1}{2}\|X - Z\|_F^2 \right).$$

which is equivalent to

$$\begin{aligned} X^\star = \operatorname{prox}_{\frac{\gamma}{2}\|\cdot\|_r^2}(Z) \;&\Leftrightarrow\; 0 \in \partial_X \left(\frac{\gamma}{2}\|X\|_r^2 + \frac{1}{2}\|X - Z\|_F^2 \right)\Big|_{X=X^\star} \\ &\Leftrightarrow\; Z - X^\star \in \gamma \|X^\star\|_r \, \partial_X \|X\|_r|_{X=X^\star}. \end{aligned}$$

Let $\bar{r} := \lceil r \rceil$ and an SVD of $X^\star$ be given by $X^\star = \sum_{i=1}^{\min\{m,n\}} \sigma_i(X^\star) u_i v_i^T$ with

$$\sigma_{\bar{r}-t}(X^\star) > \sigma_{\bar{r}-t+1}(X^\star) = \cdots = \sigma_{\bar{r}}(X^\star) = \cdots = \sigma_{\bar{r}+s}(X^\star) > \sigma_{\bar{r}+s+1}(X^\star),$$

where we say that $t = \bar{r}$ and $s = \min\{m,n\} - \bar{r}$ if $\sigma_1(X^\star) = \sigma_{\bar{r}}(X^\star)$ and $\sigma_{\min\{m,n\}}(X^\star) = \sigma_{\bar{r}}(X^\star)$, respectively. Further, let

$$U_2 := \big(u_{\bar{r}-t+1}, \dots, u_{\bar{r}+s}\big) \quad \text{and} \quad V_2 := \big(v_{\bar{r}-t+1}, \dots, v_{\bar{r}+s}\big).$$

By Proposition A.5,

$$Z = (1+\gamma)\sum_{i=1}^{\bar{r}-t} \sigma_i(X^\star) u_i v_i^T + \sigma_{\bar{r}}(X^\star) U_2 (I + \gamma T) V_2^T + \sum_{i=\bar{r}+s+1}^{\min\{m,n\}} \sigma_i(X^\star) u_i v_i^T$$

with $\|T\|_1 \leq 1$, $\|T\|_{1*} = t - \bar{r} + r$, $T \succeq 0$. Using [Watson, 1992, Theorem 2] it follows that Z has the same singular vectors as $X^\star$ and therefore a diagonal $T = \operatorname{diag}(T_{\bar{r}-t+1}, \dots, T_{\bar{r}+s})$ can be chosen. This gives

i. $1 \leq i \leq \bar{r} - t$: $\sigma_i(X^\star) = \dfrac{\sigma_i(Z)}{1+\gamma}$.

ii. $\bar{r} - t + 1 \leq i \leq \bar{r} + s$: $\sigma_{\bar{r}}(X^\star) = \dfrac{\sigma_i(Z)}{1+\gamma T_i}$.

iii. $\bar{r} + s + 1 \leq i \leq \min\{m,n\}$: $\sigma_i(X^\star) = \sigma_i(Z)$.

Hence, the main task is to determine $s \geq 0$, $t \geq 1$ and $T \succeq 0$ such that

$$\sigma_{\bar{r}}(X^\star) = \frac{\sigma_{\bar{r}-t+1}(Z)}{1+\gamma T_{\bar{r}-t+1}} = \cdots = \frac{\sigma_{\bar{r}+s}(Z)}{1+\gamma T_{\bar{r}+s}}, \tag{45}$$

where

$$\sum_{i=1}^{s} T_{\bar{r}-t+i} = t - \bar{r} + r \quad \text{and} \quad T_{\bar{r}-t+i} \leq 1 \text{ for all} i \in \{1, ..., t+s\} \tag{46}$$

and

$$\frac{\sigma_{\bar{r}-t}(Z)}{1+\gamma} > \sigma_{\bar{r}}(X^\star) > \sigma_{\bar{r}+s+1}(Z). \tag{47}$$

Next it is shown how s, t and T can be determined inductively. Clearly, there exists $T_{\bar{r}}, \dots, T_{\bar{r}+s_0}$ for some $s_0 \geq 0$, fulfilling (45) and (46) with $t = 1$ and $s = s_0$. However, if

$$\frac{\sigma_{\bar{r}-1}(Z)}{1+\gamma} \leq \frac{\sigma_{\bar{r}}(Z)}{1+\gamma T_{\bar{r}}},$$

then requirement (47) is violated. Hence, $t = 0$ is not a feasible choice and we want to find the smallest possible t for which this requirement is met

after constructing T. Let us assume that with $t = \tilde{t} - 1$ and $s = \tilde{s}_{\tilde{t}-1}$, there is no solution that satisfies all three conditions (45) – 47.

Then one can construct $T_{\bar{r}-\tilde{t}+1}, \dots, T_{\bar{r}+\tilde{s}_{\tilde{t}}}$ fulfilling (45) and (46) with $t = \tilde{t}$ and $s = \tilde{s}_{\tilde{t}}$, as follows: Let $i \geq 2$ and $T^{(i-1)}_{\bar{r}-\tilde{t}+1}, \dots, T^{(i-1)}_{\bar{r}-\tilde{t}+i-1} \leq 1$ be determined such that

$$\frac{\sigma_{\bar{r}-\tilde{t}+1}(Z)}{1+\gamma T^{(i-1)}_{\bar{r}-\tilde{t}+1}} = \cdots = \frac{\sigma_{\bar{r}-\tilde{t}+i-1}(Z)}{1+\gamma T^{(i-1)}_{\bar{r}-\tilde{t}+i-1}} = \sigma_{\bar{r}-\tilde{t}+i}(Z) \text{ and } \sum_{j=1}^{i-1} T^{(i-1)}_{\bar{r}-\tilde{t}+j} < \tilde{t} - \bar{r} + r.$$

Case 1: Assume that there exists $T^{(i)}_{\bar{r}-\tilde{t}+i}$ such that for all $j \in \{1, ..., i-1\}$

$$\sigma_{\bar{r}-\tilde{t}+i+1}(Z) = \frac{\sigma_{\bar{r}-\tilde{t}+i}(Z)}{1+\gamma T^{(i)}_{\bar{r}-\tilde{t}+i}} = \frac{\sigma_{\bar{r}-\tilde{t}+j}(Z)}{\left(1+\gamma T^{(i-1)}_{\bar{r}-\tilde{t}+j}\right)\left(1+\gamma T^{(i)}_{\bar{r}-\tilde{t}+i}\right)} = \frac{\sigma_{\bar{r}-\tilde{t}+j}(Z)}{1+\gamma T^{(i)}_{\bar{r}-\tilde{t}+j}}.$$

Then $i < \tilde{t} + \tilde{s}_{\tilde{t}}$ and we get

$$T^{(i)}_{\bar{r}-\tilde{t}+i} = \gamma^{-1}\left(\frac{\sigma_{\bar{r}-\tilde{t}+i}(Z)}{\sigma_{\bar{r}-\tilde{t}+i+1}(Z)} - 1\right). \tag{48}$$

Thus,

$$T^{(i)}_{\bar{r}-\tilde{t}+j} = \gamma^{-1}\left((1+\gamma T^{(i-1)}_{\bar{r}-\tilde{t}+j})(1+\gamma T^{(i)}_{\bar{r}-\tilde{t}+i}) - 1\right)$$

for all $j \in \{1, ..., i-1\}$. Since (48) is valid for all $i < \tilde{t} + \tilde{s}_{\tilde{t}}$, it is readily shown that the previous equation can also be written as

$$T^{(i)}_{\bar{r}-\tilde{t}+j} = \gamma^{-1}\left(\frac{\sigma_{\bar{r}-\tilde{t}+j}(Z)}{\sigma_{\bar{r}-\tilde{t}+i+1}(Z)} - 1\right). \tag{49}$$

Case 2: Assume $T^{(i)}_{\bar{r}-\tilde{t}+i}$ is such that $\sum_{j=1}^{i} T^{(i)}_{\bar{r}-\tilde{t}+j} = \tilde{t} - \bar{r} + r$ and

$$\sigma_{\bar{r}-\tilde{t}+i+1}(Z) < \frac{\sigma_{\bar{r}-\tilde{t}+i}(Z)}{1+\gamma T^{(i)}_{\bar{r}-\tilde{t}+i}} = \frac{\sigma_{\bar{r}-\tilde{t}+j}(Z)}{\left(1+\gamma T^{(i-1)}_{\bar{r}-\tilde{t}+j}\right)\left(1+\gamma T^{(i)}_{\bar{r}-\tilde{t}+i}\right)} = \frac{\sigma_{\bar{r}-\tilde{t}+j}(Z)}{1+\gamma T^{(i)}_{\bar{r}-\tilde{t}+j}},$$

for all $j \in \{1, ..., i-1\}$. Then $i = \tilde{t} + \tilde{s}_{\tilde{t}}$ and it follows as in (49) that for all $j \in \{1, ..., i-1\}$

$$\begin{aligned} T^{(i)}_{\bar{r}-\tilde{t}+j} &= \gamma^{-1}\left((1+\gamma T^{(i-1)}_{\bar{r}-\tilde{t}+j})(1+\gamma T^{(i)}_{\bar{r}-\tilde{t}+i}) - 1\right) \\ &= \gamma^{-1}\left(\frac{\sigma_{\bar{r}-\tilde{t}+j}(Z)}{\sigma_{\bar{r}-\tilde{t}+i}(Z)}(1+\gamma T^{(i)}_{\bar{r}-\tilde{t}+i}) - 1\right) \end{aligned} \tag{50}$$

Algorithm 2 Determine $X = \operatorname{prox}_{\frac{\gamma}{2}\|\cdot\|_r^2}(Z)$

Input: Let $\gamma, r > 0$ and $Z \in \mathbb{R}^{n\times m}$ be given and set $\bar{r} = \lceil r \rceil$ and $s = t = 0$.
Let $Z = \sum_{i=1}^{\min\{m,n\}} \sigma_i(Z) u_i v_i^T$ be an SVD of Z.
while $\left(\bar{r} > t \text{ AND } \sigma_{\bar{r}-t}(Z) \leq \frac{(1+\gamma)\sigma_{\bar{r}-t+1}(Z)}{1+\gamma T_{\bar{r}-t+1}}\right)$ **or** $t = 0$ **do**
  $t = t + 1$
  $k = \bar{r} - t$
  while $s \neq k$ **do**
    $k = k + 1$
    $T_k = \frac{t-\bar{r}+r-\gamma^{-1}\sum_{j=1}^{t+k-1}\left(\frac{\sigma_{\bar{r}-t+j}(Z)}{\sigma_{\bar{r}+k}(Z)}-1\right)}{t+k+\sum_{j=1}^{t+k-1}\left(\frac{\sigma_{\bar{r}-t+j}(Z)}{\sigma_{\bar{r}+k}(Z)}-1\right)}$
    if $\frac{\sigma_k(Z)}{1+\gamma T_k} \geq \sigma_{k+1}(Z)$ **then**
      $s = k$
    end if
  end while
end while
Output:

$$X = \frac{1}{1+\gamma}\sum_{i=1}^{\bar{r}-t}\sigma_i(Z)u_i v_i^T + \frac{\sigma_s(Z)}{1+\gamma T_s}\sum_{i=\bar{r}-t+1}^{\bar{r}+s} u_i v_i^T + \sum_{i=\bar{r}+s+1}^{\min\{m,n\}}\sigma_i(Z)u_i v_i^T.$$

Thus $T^{(i)}_{\bar{r}-\tilde{t}+i}$ is left to be determined. To this end, notice that

$$\begin{aligned}(1+\gamma T^{(i-1)}_{\bar{r}-\tilde{t}+i-1})(1+\gamma T^{(i)}_{\bar{r}-\tilde{t}+i}) &= 1+\gamma T^{(i)}_{\bar{r}-\tilde{t}+i-1} \\ &= 1+\gamma\left(\tilde{t}-\bar{r}+r-T^{(i)}_{\bar{r}-\tilde{t}+i}-\sum_{j=1}^{i-2}T^{(i)}_{\bar{r}-\tilde{t}+j}\right).\end{aligned}$$

In conjunction with (50), this yields

$$T^{(i)}_{\bar{r}-\tilde{t}+i} = \frac{\tilde{t}-\bar{r}+r-\sum_{j=1}^{i-1}T^{(i-1)}_{\bar{r}-\tilde{t}+j}}{i+\gamma\sum_{j=1}^{i-1}T^{(i-1)}_{\bar{r}-\tilde{t}+j}} = \frac{\tilde{t}-\bar{r}+r-\gamma^{-1}\sum_{j=1}^{i-1}\left(\frac{\sigma_{\bar{r}-\tilde{t}+j}(Z)}{\sigma_{\bar{r}-\tilde{t}+i}(Z)}-1\right)}{i+\sum_{j=1}^{i-1}\left(\frac{\sigma_{\bar{r}-\tilde{t}+j}(Z)}{\sigma_{\bar{r}-\tilde{t}+i}(Z)}-1\right)}.$$

Thus, $T_{\bar{r}-\tilde{t}+j} = T^{(i)}_{\bar{r}-\tilde{t}+j}$ for all $j \in \{1,\ldots,i-1\}$.

By the injectivity of the proximal mapping, this procedure eventually finds t, s and T that satisfy (45) – (47). Moreover, since (45) – (47) can be checked efficiently, one can perform a binary search over s and t of complexity $\mathcal{O}(n)$ (see [Eriksson et al., 2015]). Hence, the bottle neck of the prox computation is the cost for the SVD. In practice t and s are rather

small, which can be seen by the fact that

$$\frac{\sigma_{\bar{r}}(Z)}{1+\gamma} > \sigma_{\bar{r}+1}(Z) \;\Rightarrow\; s = 0, \tag{51}$$

In this case $\operatorname{rank}(X^\star) = \bar{r}$ and only t has to be determined. If additionally $\bar{r} = r$, then T is the identity matrix and finding t is redundant. In case of high dimensional matrices, a linear search as outlined in Algorithm 2 may advisable, since that allows us to incorporate sparse SVD solvers by computing one singular value at a time it is needed.

References

Antoulas, A. (2005). *Approximation of Large-Scale Dynamical Systems*. SIAM.

Antoulas, A. C. (1997). "On the approximation of hankel matrices". In: *Operators, Systems and Linear Algebra: Three Decades of Algebraic Systems Theory*. Vieweg+Teubner Verlag, pp. 17–22.

Argyriou, A., R. Foygel, and N. Srebro (2012). "Sparse prediction with the k-support norm". In: Pereira, F. et al. (Eds.). *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., pp. 1457–1465.

Bach, F., R. Jenatton, J. Mairal, and G. Obozinski (2012). "Optimization with sparsity-inducing penalties". *Foundations and Trends in Machine Learning* **4**:1, pp. 1–106.

Bauschke, H. H. and P. L. Combettes (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer New York.

Berge, C. (1963). *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity*. Courier Corporation.

Berry, M. W., M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons (2007). "Algorithms and applications for approximate nonnegative matrix factorization". *Computational Statistics & Data Analysis* **52**:1, pp. 155–173.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011). "Distributed optimization and statistical learning via the alternating direction method of multipliers". *Foundations and Trends in Machine Learning* **3**:1, pp. 1–122.

Cai, J.-F., E. J. Candès, and Z. Shen (2010). "A singular value thresholding algorithm for matrix completion". *SIAM Journal on Optimization* **20**:4, pp. 1956–1982.

Candès, E. J. and Y. Plan (2010). "Matrix completion with noise". *Proceedings of the IEEE* **98**:6, pp. 925–936.

Candès, E. J. and T. Tao (2010). "The power of convex relaxation: near-optimal matrix completion". *IEEE Transactions on Information Theory* **56**:5, pp. 2053–2080.

Candès, E. J. and B. Recht (2009). "Exact matrix completion via convex optimization". *Foundations of Computational Mathematics* **9**:6, p. 717.

Chandrasekaran, V., B. Recht, P. A. Parrilo, and A. S. Willsky (2012). "The convex geometry of linear inverse problems". *Foundations of Computational Mathematics* **12**:6, pp. 805–849.

Chandrasekaran, V., S. Sanghavi, P. A. Parrilo, and A. S. Willsky (2011). "Rank-sparsity incoherence for matrix decomposition". *SIAM Journal on Optimization* **21**:2, pp. 572–596.

Chen, Y., M. R. Jovanović, and T. T. Georgiou (2013). "State covariances and the matrix completion problem". In: *52nd IEEE Conference on Decision and Control*, pp. 1702–1707.

Chu, M. T., R. E. Funderlic, and R. J. Plemmons (2003). "Structured low rank approximation". *Linear Algebra and its Applications* **366**, pp. 157–172.

Combettes, P. L. and J.-C. Pesquet (2011). "Proximal splitting methods in signal processing". In: *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer New York, pp. 185–212.

Doan, X. V. and S. Vavasis (2016). "Finding the largest low-rank clusters with Ky Fan 2-k-norm and ℓ_1-norm". *SIAM Journal on Optimization* **26**:1, pp. 274–312.

Douglas, J. and H. H. Rachford (1956). "On the numerical solution of heat conduction problems in two and three space variables". *Transactions of the American Mathematical Society* **82**:2, pp. 421–439.

Eckstein, J. and D. P. Bertsekas (1992). "On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators". *Mathematical Programming* **55**:1, pp. 293–318.

Eldén, L. (2007). *Matrix methods in data mining and pattern recognition*. SIAM.

Eriksson, A., T. T. Pham, T.-J. Chin, and I. Reid (2015). "The k-support norm and convex envelopes of cardinality and rank". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3349–3357.

Fazel, M., H. Hindi, and S. P. Boyd (2001). "A rank minimization heuristic with application to minimum order system approximation". In: *Proceedings of the 2001 American Control Conference.* Vol. 6, pp. 4734–4739.

Fazel, M. (2002). *Matrix Rank Minimization with Applications*. PhD thesis. Stanford University.

Freimer, M. and G. S. Mudholkar (1984). "A class of generalizations of Hölder's inequality". *Lecture Notes-Monograph Series* **5**, pp. 59–67.

Gabay, D. and B. Mercier (1976). "A dual algorithm for the solution of nonlinear variational problems via finite element approximation". *Computers and Mathematics with Applications* **2**:1, pp. 17–40.

Georgiou, T. T. (2002a). "Spectral analysis based on the state covariance: the maximum entropy spectrum and linear fractional parametrization". *IEEE Transactions on Automatic Control* **47**:11, pp. 1811–1823.

Georgiou, T. T. (2002b). "The structure of state covariances and its relation to the power spectrum of the input". *IEEE Transactions on Automatic Control* **47**:7, pp. 1056–1066.

Glowinski, R. and A. Marroco (1975). "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problémes de dirichlet non linéaires". *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique* **9**, pp. 41–76.

Grussler, C. and A. Rantzer (2014). "Modified balanced truncation preserving ellipsoidal cone-invariance". In: *53rd IEEE Conference on Decision and Control (CDC)*, pp. 2365–2370.

Grussler, C. and A. Rantzer (2015). "On optimal low-rank approximation of non-negative matrices". In: *54th IEEE Conference on Decision and Control (CDC)*, pp. 5278–5283.

Grussler, C., A. Zare, M. R. Jovanovic, and A. Rantzer (2016). "The use of the $r*$ heuristic in covariance completion problems". In: *55th IEEE Conference on Decision and Control (CDC)*.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.

Hesse, R., D. R. Luke, and P. Neumann (2014). "Alternating projections and Douglas-Rachford for sparse affine feasibility". *IEEE Transactions on Signal Processing* **62**:18, pp. 4868–4881.

Hesse, R. and D. R. Luke (2013). "Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems". *SIAM Journal on Optimization* **23**:4, pp. 2397–2419.

Higham, N. J. (2002). "Computing the nearest correlation matrix – a problem from finance". *IMA Journal of Numerical Analysis* **22**:3, pp. 329–343.

Hiriart-Urruty, J.-B. and C. Lemaréchal (1996). *Convex analysis and minimization algorithms II*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Hiriart-Urruty, J.-B. and C. Lemaréchal (2013). *Convex analysis and minimization algorithms I: Fundamentals*. Vol. 305. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Horn, R. A. and C. R. Johnson (2012). *Matrix Analysis*. 2nd ed.

Izenman, A. J. (1975). "Reduced-rank regression for the multivariate linear model". *Journal of Multivariate Analysis* **5**:2, pp. 248–264.

Kim, J. and H. Park (2011). "Fast nonnegative matrix factorization: an active-set-like method and comparisons". *SIAM Journal on Scientific Computing* **33**:6, pp. 3261–3281.

Lai, H., Y. Pan, C. Lu, Y. Tang, and S. Yan (2014). "Efficient k-support matrix pursuit". In: *Proccedings in ECCV, Part II*. Springer International Publishing, pp. 617–631.

Larsson, V. and C. Olsson (2016). "Convex low rank approximation". *International Journal of Computer Vision* **120**:2, pp. 194–214.

Larsson, V., C. Olsson, E. Bylow, and F. Kahl (2014). "Rank minimization with structured data patterns". In: *Proceedings in ECCV, Part III*. Springer International Publishing, Cham, pp. 250–265.

Lin, F., M. R. Jovanović, and T. T. Georgiou (2013). "An admm algorithm for matrix completion of partially known state covariances". In: *52nd IEEE Conference on Decision and Control*, pp. 1684–1689.

Lions, P.-L. and B. Mercier (1979). "Splitting algorithms for the sum of two nonlinear operators". *SIAM Journal on Numerical Analysis* **16**:6, pp. 964–979.

Liu, X., Z. Wen, and Y. Zhang (2013). "Limited memory block Krylov subspace optimization for computing dominant singular value decompositions". *SIAM Journal on Scientific Computing* **35**:3, A1641–A1668.

Luenberger, D. G. (1968). *Optimization by Vector Space Methods*. John Wiley & Sons.

Markovsky, I. (2008). "Structured low-rank approximation and its applications". *Automatica* **44**:4, pp. 891–909.

McDonald, A. M., M. Pontil, and D. Stamos (2015). "New Perspectives on k-Support and Cluster Norms". arXiv: 1512.08204.

Olsson, C. and M. Oskarsson (2009). "A convex approach to low rank matrix approximation with missing data". In: *Image Analysis: 16th Scandinavian Conference, SCIA*. Springer Berlin Heidelberg, pp. 301–309.

Partington, J. R. (1989). *An Introduction to Hankel Operators*. Cambridge University Press.

Peaucelle, D., D. Henrion, Y. Labit, and K. Taitz (2002). "User's guide for SEDUMI INTERFACE 1.04". LAAS-CNRS, Toulouse.

Phan, H. M. (2016). "Linear convergence of the Douglas–Rachford method for two closed sets". *Optimization* **65**:2, pp. 369–385.

Recht, B., M. Fazel, and P. A. Parrilo (2010). "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization". *SIAM Review* **52**:3, pp. 471–501.

Reinsel, G. C. and R. Velu (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*. Vol. 136. Lecture Notes in Statistics. Springer New York.

Rockafellar, R. T. (1970). *Convex Analysis*. 28. Princeton University Press.

Srebro, N., T. Jaakkola, et al. (2003). "Weighted low-rank approximations". In: *20th International Conference on Machine Learning (ICML)*. Vol. 3, pp. 720–727.

Stewart, G. W. and J.-g. Sun (1990). *Matrix Perturbation Theory*. Academic press.

Sundaram, R. K. (1996). *A First Course in Optimization Theory*. Cambridge University Press.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society* **58**:1, pp. 267–288.

Toh, K.-C., M. J. Todd, and R. H. Tütüncü (1999). "SDPT3 – a MATLAB software package for semidefinite programming, version 1.3". *Optimization Methods and Software* **11**:1-4, pp. 545–581.

Vidal, R., Y. Ma, and S. S. Sastry (2016). *Generalized Principal Component Analysis*. Vol. 40. Interdisciplinary Applied Mathematics. Springer-Verlag New York.

Watson, G. (1992). "Characterization of the subdifferential of some matrix norms". *Linear Algebra and its Applications* **170**, pp. 33–45.

Zare, A., Y. Chen, M. Jovanović, and T. T. Georgiou (2016a). "Low-complexity modeling of partially available second-order statistics: theory and an efficient matrix completion algorithm". *IEEE Transactions on Automatic Control*. arXiv:1412.3399. DOI: 10.1109/TAC.2016.2595761.

Zare, A., M. R. Jovanović, and T. T. Georgiou (2015). "Alternating direction optimization algorithms for covariance completion problems". In: *2015 American Control Conference (ACC)*, pp. 515–520.

Zare, A., M. R. Jovanović, and T. T. Georgiou (2016b). "Color of turbulence". *Journal of Fluid Mechanics*. arXiv:1602.05105. DOI: 10.1017/jfm.2016.682.

Zhou, K., J. C. Doyle, K. Glover, et al. (1996). *Robust and Optimal Control*. Vol. 40. Prentice Hall.

Acknowledgments

We thank Andrey Ghulchak for his useful comments and numerous counter-examples. All authors are members of the LCCC Linnaeus Center and the eLLIIT Excellence Center at Lund University. The first author is financially supported by the Swedish Research Council through the project 621-2012-5357. The third author is financially supported by the Swedish Foundation for Strategic Research.

Paper IV

Low-Rank Inducing Norms with Optimality Interpretations

Christian Grussler Pontus Giselsson

Abstract

Optimization problems with rank constraints appear in many diverse fields such as control, machine learning and image analysis. Since the rank constraint is non-convex, these problems are often approximately solved via convex relaxations. Nuclear norm regularization is the prevailing convexifying technique for dealing with these types of problem. This paper introduces a family of low-rank inducing norms and regularizers which includes the nuclear norm as a special case. A posteriori guarantees on solving an underlying rank constrained optimization problem with these convex relaxations are provided. We evaluate the performance of the low-rank inducing norms on three matrix completion problems. In all examples, the nuclear norm heuristic is outperformed by convex relaxations based on other low-rank inducing norms. For two of the problems there exist low-rank inducing norms that succeed in recovering the partially unknown matrix, while the nuclear norm fails. These low-rank inducing norms are shown to be representable as semi-definite programs and to have cheaply computable proximal mappings. The latter makes it possible to also solve problems of large size with the help of scalable first-order methods. Finally, it is proven that our findings extend to the more general class of atomic norms. In particular, this allows us to solve corresponding vector-valued problems, as well as problems with other non-convex constraints.

Preprint.

1. Introduction

Many problems in machine learning, image analysis, model order reduction, multivariate linear regression, etc. (see [Izenman, 1975; Antoulas, 2005; Candès and Recht, 2009; Candès and Plan, 2010; Recht et al., 2010; Chandrasekaran et al., 2011; Reinsel and Velu, 1998; Hastie et al., 2015; Larsson and Olsson, 2016; Vidal et al., 2016]), can be posed as a low-rank estimation problems based on measurements and prior information about a data matrix. These estimation problems often take the form

$$\begin{aligned} \underset{M}{\text{minimize}} \quad & f_0(M) \\ \text{subject to} \quad & \operatorname{rank}(M) \leq r, \end{aligned} \tag{1}$$

where f_0 is a proper closed convex function and r is a positive integer that specifies the desired or expected rank. Due to non-convexity of the rank constraint a solution to (1) is known only in a few special cases (see e.g. [Antoulas, 1997; Antoulas, 2005; Reinsel and Velu, 1998]).

A common approach to deal with the rank constraint is to use the nuclear norm heuristic (see [Fazel et al., 2001; Recht et al., 2010]). The idea is to convexify the problem by replacing the non-convex rank constraint with a nuclear norm regularization term. For matrix completion problems, this approach is shown to recover the true low-rank matrix with high probability, provided that enough random measurements are available (see [Candès and Recht, 2009; Recht et al., 2010; Chandrasekaran et al., 2012]). If these assumption are not met, however, the nuclear norm heuristic may fail in producing satisfactory estimates (see [Grussler et al., 2016a; Grussler et al., 2016b]).

This paper introduces a family of low-rank inducing norms as alternatives to the nuclear norm. These norms can be interpreted as the largest convex minorizers of non-convex functions f of the form

$$f := \|\cdot\| + \chi_{\operatorname{rank}(\cdot)\leq r}, \tag{2}$$

where $\|\cdot\|$ is an arbitrary unitarily invariant norm, and $\chi_{\operatorname{rank}(\cdot)\leq r}$ is the indicator function for matrices with rank less than or equal to r. This interpretation motivates the use of low-rank inducing norms in convex relaxations to (1). In particular, assume that f_0 in (1) can be split into the sum of a convex function and unitarily invariant norm, and the solution to the corresponding convex relaxation has rank r. Then this solution also solves the non-convex problem, and thus provides an *a posteriori* optimality guarantee. Furthermore, the choice of norms and target ranks r can be considered as regularization parameters when used in convex relaxations of (1). Compared to the nuclear norm approach, it is shown that this gives additional flexibility which can be exploited to improve the quality of the

estimate. Specifically, the nuclear norm is the largest convex minorizer of f in (2) with $r = 1$, making it a less natural choice than other low-rank inducing norms, because it convexifies constraints that allow for matrices of rank 1, only.

This work particularly focuses on low-rank inducing norms, where the norm in (2) is the Frobenius norm or the spectral norm. We refer to these norms as *low-rank inducing Frobenius norms* and *low-rank inducing spectral norms*, respectively. The low-rank inducing Frobenius norms, also called $r*$ norms, have been previously discussed in the literature (see [Bach et al., 2012; Eriksson et al., 2015; McDonald et al., 2015; Grussler and Rantzer, 2015; Doan and Vavasis, 2016; Grussler et al., 2016a; Grussler et al., 2016b]). In [Bach et al., 2012; Eriksson et al., 2015; McDonald et al., 2015; Doan and Vavasis, 2016], no optimality interpretations are considered, but in previous work we have presented such interpretations for the squared $r*$ norms (see [Grussler and Rantzer, 2015; Grussler et al., 2016a; Grussler et al., 2016b]). In this paper these findings are shown to extend to any function of low-rank inducing norms that is increasing on the nonnegative real numbers. Most importantly, our results hold for linear increasing functions, i.e. the low-rank inducing norm itself. To the best of our knowledge, no other low-rank inducing norms from the proposed family, including low-rank inducing spectral norms, have been proposed in the literature.

For the family of low-rank inducing norms to be useful in practice, they must be suitable for numerical optimization. We show that low-rank inducing Frobenius norms and spectral norms are representable as semi-definite programs (SDP). This allows us to readily formulate and solve small to medium scale problems using standard SDP-solvers (see [Peaucelle et al., 2002; Toh et al., 2004]). Moreover, it is demonstrated that these norms have cheaply computable proximal mappings, comparable with the computational cost for the proximal mapping of the nuclear norm. This allows us to solve large-scale problems involving low-rank inducing norms by means of proximal splitting methods (see [Combettes and Pesquet, 2011; Parikh and Boyd, 2014]). To enable formulations with increasing convex functions, the projection onto their epi-graphs is computed. This extends the proximal mapping computations of the squared $r*$ norm in [Argyriou et al., 2012; Eriksson et al., 2015; Grussler et al., 2016a] to the non-squared case.

The performance of different low-rank inducing norms is evaluated on three matrix completion problems. The evaluation reveals that the choice of low-rank inducing norms has tremendous impact on the ability to complete the covariance matrix. In particular, the nuclear norm is significantly outperformed by the low-rank inducing Frobenius norm, as well as the low-rank inducing spectral norm.

The findings in this work are also valid for the corresponding vector-valued problems by replacing rank with cardinality. This gives rise to optimality interpretations of, e.g., lasso-type and inverse problems (see [Tibshirani, 1996; Hastie et al., 2015; Vidal et al., 2016]). More generally, all low-rank inducing norms lie within the class of so-called atomic norms (see [Chandrasekaran et al., 2012]). It is shown that our optimality interpretations also hold for atomic norms under very mild assumptions. Therefore, these findings provide optimality interpretations for many other problems, such as those listed in [Chandrasekaran et al., 2012, Section 2.2].

The paper is organized as follows. We start by introducing some preliminaries in Section 2. In Section 3, we introduce the class of low-rank inducing norms, and provide optimality interpretations of these in Section 4. In Section 5, computability of low-rank inducing Frobenius and spectral norms is addressed. To support the usefulness of having more low-rank inducing regularizers at our supply, numerical examples are presented in Section 6. The optimality results are extended to the vector case and to atomic norms in Section 7 and conclusions are drawn in Section 8.

2. Preliminaries

The set of reals is denoted by $\mathbb{R}$, the set of real vectors by $\mathbb{R}^n$, and the set of real matrices by $\mathbb{R}^{n\times m}$. Element-wise nonnegative matrices $X \in \mathbb{R}^{n\times m}$ are denoted by $X \in \mathbb{R}^{n\times m}_{\geq 0}$. If symmetric $X \in \mathbb{R}^{n\times n}$ is positive definite (semi-definite), we write $X \succ 0$ ($X \succeq 0$). These notations are also used to describe relations between matrices, e.g., $A \succeq B$ means $A - B \succeq 0$. The non-increasingly ordered singular values of $X \in \mathbb{R}^{n\times m}$, counted with multiplicity, are denoted by $\sigma_1(X) \geq \cdots \geq \sigma_{\min\{m,n\}}(X)$. Furthermore,

$$\langle X, Y\rangle := \sum_{i=1}^{m}\sum_{j=n}^{n} x_{ij}y_{ij} = \operatorname{trace}(X^T Y)$$

defines the Frobennius inner-product for $X, Y \in \mathbb{R}^{n\times m}$. This inner-product gives the *Frobenius norm*

$$\|X\|_F := \sqrt{\operatorname{trace}(X^T X)} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}^2} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(X)},$$

which is a unitarily invariant norm, i.e., $\|UXV\|_F = \|X\|_F$ for all unitary matrices $U, V \in \mathbb{R}^{n\times m}$. For all $x = (x_1, \ldots, x_q) \in \mathbb{R}^q$, we define

$$\ell_1(x) := \sum_{i=1}^{q} |x_i|, \qquad \ell_2(x) := \sqrt{\sum_{i=1}^{q} x_i^2}, \qquad \ell_\infty(x) := \max_i |x_i|, \tag{3}$$

Then the Frobenius norm satisfies $\|X\|_F = \ell_2(\sigma(X))$, where

$$\sigma(X) := (\sigma_1(X), \ldots, \sigma_q(X)).$$

The functions ℓ_1 and ℓ_∞ define the nuclear norm $\|X\|_{\ell_1} := \ell_1(\sigma(X))$ and the spectral norm $\|X\|_{\ell_\infty} := \ell_\infty(\sigma(X)) = \sigma_1(X)$.

For a set $\mathcal{C} \subset \mathbb{R}^{n\times m}$,

$$\chi_{\mathcal{C}}(X) := \begin{cases} 0, & X \in \mathcal{C} \\ \infty, & X \notin \mathcal{C} \end{cases}$$

denotes the so-called *indicator function*. We also use $\chi_{\text{rank}(\cdot)\leq r}$ to denote the indicator function of the set of matrices which have at most rank r.

The following function properties will be used in this paper. The *effective domain* of a function $f : \mathbb{R}^{n\times m} \to \mathbb{R} \cup \{\infty\}$ is defined as

$$\text{dom} f := \{X \in \mathbb{R}^{n\times m} : f(X) < \infty\}$$

and the *epigraph* is defined as

$$\text{epi}(f) := \{(X, t) : f(X) \leq t, X \in \text{dom} f,\ t \in \mathbb{R}\}.$$

Further, f is said to be:

- *proper* if $\text{dom} f \neq \emptyset$.
- *closed* if the epigraph is a closed set.
- *positively homogeneous (of degree 1)* if for all $X \in \text{dom}(f)$ and $t > 0$ it holds that $f(tX) = tf(X)$.
- *nonnegative* if $f(X) \geq 0$ for all $X \in \text{dom}(f)$.
- *coercive* if $\lim_{\|X\|_F \to \infty} f(X) = \infty$.

A function $f : \mathbb{R} \cup \{\infty\} \to \mathbb{R} \cup \{\infty\}$ is called *increasing* if

$$x \leq y \ \Rightarrow\ f(x) \leq f(y) \ \text{ for all } \ x, y \in \text{dom}(f)$$

and if there exist $x, y \in \mathbb{R}$ such that $x < y$ and $f(x) < f(y)$.

The *conjugate (dual) function* f^* of f is defined as

$$f^*(Y) := \sup_{X \in \mathbb{R}^{n\times m}} [\langle X, Y\rangle - f(X)]$$

for all $Y \in \mathbb{R}^{n\times m}$. As long as f is proper and minorized by an affine function, the conjugate f^* is proper, closed and convex (see [Hiriart-Urruty

and Lemaréchal, 2013]). The function $f^{**} := (f^*)^*$ is called the *biconjugate function* of f and can be shown to be a convex minorizer of f, i.e.

$$f(X) \geq f^{**}(X) \text{ for all } X \in \mathbb{R}^{n\times m}.$$

In fact, f^{**} is the point-wise supremum of all affine functions majorized by f and therefore the largest convex minorizer of f. This can equivalently be stated as follows (see [Hiriart-Urruty and Lemaréchal, 1996, Theorem X.1.3.5,Corollary X.1.3.6]).

LEMMA 2.1
Let $f : \mathbb{R}^{n\times m} \to \mathbb{R} \cup \{\infty\}$ be such that f^{**} is proper. Then

$$\operatorname{epi}(f^{**}) = \operatorname{cl}(\operatorname{conv}(\operatorname{epi} f)),$$

where $\operatorname{cl}(\cdot)$ denotes the topological closure of a set and $\operatorname{conv}(\cdot)$ the convex hull. Further, $f^{**} = f$ if and only if f is proper closed and convex. □

Lemma 2.1 implies that for a closed proper, but possibly non-convex function f, it holds that

$$\inf_{X\in\mathbb{R}^{n\times m}} f(X) = \inf_{X\in\mathbb{R}^{n\times m}} f^{**}(X).$$

However, determining the convex function f^{**} is as difficult as minimizing the non-convex function f. Instead, it is common to convexify the problem by splitting the function into $f = f_1 + f_2$, such that f_1^{**} and f_2^{**} can be easily computed. If f_1 is proper, closed and convex, then $f_1 = f_1^{**}$ and $f_1 + f_2^{**}$ is the largest convex minorizer of f that keeps f_1 as a summand. In particular,

$$\inf_{X\in\mathbb{R}^{n\times m}} [f_1(X) + f_2(X)] \geq \inf_{X\in\mathbb{R}^{n\times m}} [f_1(X) + f_2^{**}(X)], \tag{4}$$

which holds with equality if the solution $X^\star$ to the right-hand side problem satisfies $f_2^{**}(X^\star) = f_2(X^\star)$. Then $X^\star$ also solves the non-convex problem on the left-hand side. This motivates the use of our terminology that $f_1 + f_2^{**}$ is the *optimal convex relaxation* of a given splitting $f_1 + f_2$, when f_1 is proper closed and convex.

Finally, if $f : \mathbb{R} \cup \{\infty\} \to \mathbb{R} \cup \{\infty\}$, then the *monotone conjugate* is defined as

$$f^+(y) := \sup_{x\geq 0} [\langle x, y\rangle - f(x)] \quad \text{for all } y \in \mathbb{R}.$$

3. Low-Rank Inducing Norms

This section introduces the family of *low-rank inducing norms*, which includes the nuclear norm as a special case. These can be used as regularizers in optimization problems to promote low-rank solutions. To define them, we

need to characterize the class of unitarily invariant norms in terms of symmetric gauge functions. This characterization can be found in, e.g. [Horn and Johnson, 2012, Theorem 7.4.7.2].

Definition 3.1
A function $g : \mathbb{R}^q \to \mathbb{R}_{\geq 0}$ is a symmetric gauge function if

i. g is a norm.

ii. $\forall x \in \mathbb{R}^q : g(|x|) = g(x)$, where $|x|$ denotes the element-wise absolute value.

iii. $g(Px) = g(x)$ for all permutation matrices $P \in \mathbb{R}^{q \times q}$ and all $x \in \mathbb{R}^q$. □

Proposition 3.1
The norm $\|\cdot\| : \mathbb{R}^{n \times m} \to \mathbb{R}$ is unitarily invariant if and only if

$$\|X\| = g(\sigma_1(X), \ldots, \sigma_{\min\{m,n\}}(X))$$

for all $X \in \mathbb{R}^{n \times m}$, where g is a symmetric gauge function. □

As noted in Section 2, the gauge functions for the Frobenius norm, spectral norm, and nuclear norm are $g = \ell_2$, $g = \ell_\infty$, and $g = \ell_1$, respectively, where ℓ_1, ℓ_2, and ℓ_∞, are defined in (3).

The dual norm of a unitarily invariant norm is also unitarily invariant (see [Horn and Johnson, 2012, Theorem 5.6.39]. Therefore, it has an associated symmetric gauge function. This will be denoted by g^D if the symmetric gauge function of the original norm is denoted by g. More specifically, let $M \in \mathbb{R}^{n \times m}$, $q := \min\{m, n\}$, and $g : \mathbb{R}^q \to \mathbb{R}_{\geq 0}$ be a symmetric gauge function associated with a unitarily invariant norm

$$\|M\|_g := g(\sigma_1(M), \ldots, \sigma_q(M)).$$

The dual of this norm is defined as

$$\|Y\|_{g^D} := \max_{\|M\|_g \leq 1} \langle Y, M \rangle = g^D(\sigma_1(Y), \ldots, \sigma_q(Y)), \tag{5}$$

where the dual gauge function g^D satisfies

$$g^D(\sigma_1(Y), \ldots, \sigma_q(Y)) = \max_{g(\sigma_1(M), \ldots, \sigma_q(M)) \leq 1} \sum_{i=1}^{q} \sigma_i(M)\sigma_i(Y). \tag{6}$$

The low-rank inducing norms will be defined as the dual norm of a rank constrained dual norm in (5). This rank constrained dual norm is defined as

$$\|Y\|_{g^D,r} := \max_{\substack{\operatorname{rank}(M) \leq r \\ \|M\|_g \leq 1}} \langle M, Y \rangle \tag{7}$$

and the corresponding *low-rank inducing norm* as

$$\|M\|_{g,r*} := \max_{\|Y\|_{g^D,r}\leq 1} \langle Y, M\rangle. \tag{8}$$

For $q = \min\{m, n\}$, the rank constraint in (7) is redundant and the dual of the dual becomes the norm itself.

For symmetric gauge functions $g : \mathbb{R}^q \to \mathbb{R}_{\geq 0}$, we denote their truncated symmetric gauge functions by $g(\sigma_1, \ldots, \sigma_r) := g(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$ for any $r \in \{1, \ldots, q\}$. With this notation in mind, some properties of low-rank inducing norms and their duals are stated in the following lemma. A proof is given in Section A.1.

LEMMA 3.1
Let $M, Y \in \mathbb{R}^{n\times m}$, $r \in \mathbb{N}$ be such that $1 \leq r \leq q := \min\{m, n\}$, and $g : \mathbb{R}^q \to \mathbb{R}_{\geq 0}$ be a symmetric gauge function. Then $\|\cdot\|_{g^D,r}$ is a unitarily invariant norm that satisfies

$$\|Y\|_{g^D,r} = g^D(\sigma_1(Y), \ldots, \sigma_r(Y)) \tag{9}$$

Its dual norm $\|\cdot\|_{g,r*}$ satisfies

$$\|M\|_{g,r*} = \max_{g^D(\sigma_1(Y),\ldots,\sigma_r(Y))\leq 1} \left[\sum_{i=1}^{r} \sigma_i(M)\sigma_i(Y) + \sigma_r(Y) \sum_{i=r+1}^{q} \sigma_i(M)\right], \tag{10}$$

and

$$\|M\|_g = \|M\|_{g,q*} \leq \cdots \leq \|M\|_{g,1*}, \tag{11}$$

$$\operatorname{rank}(M) \leq r \;\Rightarrow\; \|M\|_g = \|M\|_{g,r*}. \tag{12}$$

□

This paper particularly focuses on low-rank inducing norms originating from the Frobenius norm and the spectral norm. When the original norm is the Frobenius norm, then $g = \ell_2$. Since the norm is self dual, it satisfies $g^D = \ell_2^D = \ell_2$. The truncated version in (9) (which is denote by $\|\cdot\|_r$ to comply with notation used, e.g., in [Grussler et al., 2016a]) becomes

$$\|Y\|_r := \|Y\|_{\ell_2^D,r} = \sqrt{\sum_{i=1}^{r} \sigma_i^2(Y)}.$$

The corresponding low-rank inducing norm is referred to as the *low-rank inducing Frobenius norm*, and is denoted by

$$\|M\|_{r*} := \|M\|_{\ell_2,r*} = \max_{\|Y\|_r\leq 1} \langle Y, M\rangle.$$

In [Grussler et al., 2016a], this norm is referred to as the $r*$ norm.

If the original norm, instead, is the spectral norm, we have $g = \ell_\infty$. The dual norm is the nuclear (trace) norm (see [Horn and Johnson, 2012, Theorem 5.6.42]), with gauge function $g^D = \ell_1$. The truncated version becomes

$$\|Y\|_{\ell_1,r} := \sum_{i=1}^{r} \sigma_i(Y),$$

and its dual, which we refer to as the *low-rank inducing spectral norm*, is denoted by

$$\|M\|_{\ell_\infty,r*} := \max_{\|Y\|_{\ell_1,r}\leq 1} \langle Y, M\rangle.$$

The nuclear norm is a special case of these low-rank inducing norms, corresponding to $r = 1$.

Proposition 3.2
The nuclear norm satisfies $\|\cdot\|_{\ell_1} = \|\cdot\|_{g,1*}$, where $\|\cdot\|_g$ is any unitarily invariant norm with $g(\sigma_1) = |\sigma_1|$.

A proof to this proposition is found in Section A.1.

Next, we state a result that is the key to our optimality interpretations for low-rank inducing norms in the next section.

Lemma 3.2
Let $B^1_{g,r*} := \{X \in \mathbb{R}^{n\times m} : \|X\|_{g,r*} \leq 1\}$ be the unit low-rank inducing norm ball and let

$$E_{g,r} := \{X \in \mathbb{R}^{n\times m} : \|X\|_g = 1,\ \operatorname{rank}(X) \leq r\}. \tag{13}$$

Then $B^1_{g,r*} = \operatorname{conv}(E_{g,r})$, i.e. all $M \in \mathbb{R}^{n\times m}$ can be decomposed as

$$M = \textstyle\sum_i \alpha_i M_i \quad \text{with} \quad \sum_i \alpha_i = 1,\ \alpha_i \geq 0,$$

where M_i satisfies $\operatorname{rank}(M_i) \leq r$ and

$$\|M_i\|_g = \|M_i\|_{g,r*} = \|M\|_{g,r*}. \qquad \square$$

A proof to this lemma is given in Section A.1. The result is a direct consequence of Lemma 3.1 and extends what is known about the nuclear norm, and the results on low-rank inducing Frobenius norms in [Grussler et al., 2016a].

In many cases, the set $E_{g,r}$ is the set of extreme points to the unit ball $B^1_{g,r*}$. The following result is proven in Section A.1.

PROPOSITION 3.3
Suppose that $\|\cdot\|_g$ satisfies

$$\|\textstyle\sum_i \alpha_i M_i\|_g < \sum_i \alpha_i \|M_i\|_g$$

for all $\alpha_i \in (0, 1)$ such that $\sum_i \alpha_i = 1$, and all $M_i \in \mathbb{R}^{n\times m}$ with $\|M_i\|_g = 1$. Then $E_{g,r}$ in (13) is the set of extreme points to $B^1_{g,r*}$. □

All ℓ_p norms with $1 < p < \infty$ satisfy these assumptions, and therefore the unit balls of their low-rank inducing norms have $E_{g,r}$ as their extreme point sets.

The extreme point sets for the unit balls of the low-rank inducing spectral norms are characterized next.

COROLLARY 3.1
The extreme point set of the unit ball to the low-rank inducing spectral norm $B^1_{\ell_\infty,r*}$ is given by

$$\mathcal{E}_r := \{X \in \mathbb{R}^{n\times m} : \sigma_1(X) = \cdots = \sigma_r(X) = 1 \text{ and } \operatorname{rank}(X) = r\}.$$

□

This result is proven in Section A.1.

We could also use the nuclear norm as a basis for the low-rank inducing norm. By Proposition 3.2, we know that $\|\cdot\|_{\ell_1,1*} = \|\cdot\|_{\ell_1}$. Therefore (11) implies that any low-rank inducing nuclear norm is just the nuclear norm, i.e.,

$$\|\cdot\|_{\ell_1} = \|\cdot\|_{\ell_1,q*} = \cdots = \|\cdot\|_{\ell_1,1*}.$$

Compared to using the low-rank inducing Frobenius and spectral norms, this does not provide us with a richer family of low-rank inducing norms.

4. Optimality Interpretations

In this section, we shown that low-rank inducing norms can be interpreted as the largest convex minorizers, i.e., the biconjugates of non-convex functions of the form (2), where the norm is arbitrary but unitarily invariant. Using this interpretation, we show how to create optimal convex relaxations of rank constrained optimization problems. This yields a posteriori guarantees on when a convex relaxation involving a low-rank inducing norm solves the corresponding rank constrained problem.

The interpretation of low-rank inducing norms follows as a special case of the following more general result.

THEOREM 1
Assume $f : \mathbb{R}_{\geq 0} \to \mathbb{R} \cup \{\infty\}$ is an increasing closed convex function, and let $f_{\text{reg}} := f(\|\cdot\|_g) + \chi_{\text{rank}(\cdot)\leq r}$ with $r \in \mathbb{N}$ such that $1 \leq r \leq \min\{m, n\}$. Then,

$$f^*_{\text{reg}} = f^+(\|\cdot\|_{g^D,r}), \tag{14}$$
$$f^{**}_{\text{reg}} = f(\|\cdot\|_{g,r*}). \tag{15}$$
□

Proof Since $\text{epi}(f(\|\cdot\|_{g,r*}))$ is closed by [Hiriart-Urruty and Lemaréchal, 2013, Proposition IV.2.1.8], it follows by Lemma 2.1 that if

$$\text{epi}(f(\|\cdot\|_{g,r*})) = \text{conv}(\text{epi}(f_{\text{reg}})),$$

then (15) follows.

Let us start by showing that $\text{epi}(f(\|\cdot\|_{g,r*})) \subset \text{conv}(\text{epi}(f_{\text{reg}}))$. Assume that $(M, t) \in \text{epi}(f(\|\cdot\|_{g,r*}))$. By Lemma 3.2,

$$M = \textstyle\sum_i \alpha_i M_i \quad \text{with} \quad \sum_i \alpha_i = 1,\ \alpha_i \geq 0$$

where M_i satisfies

$$\text{rank}(M_i) \leq r, \quad \text{and} \quad \|M_i\|_{g,r*} = \|M\|_{g,r*}.$$

Hence, $(M, t) = \sum_i \alpha_i (M_i, t)$, where

$$t \geq f(\|M\|_{g,r*}) = f(\|M_i\|_{g,r*}) \quad \text{and} \quad \text{rank}(M_i) \leq r.$$

This shows that $(M_i, t) \in \text{epi}(f_{\text{reg}})$, and therefore $(M, t) \in \text{conv}(\text{epi}(f_{\text{reg}}))$.

Conversely, if $(M, t) \in \text{conv}(\text{epi}(f_{\text{reg}}))$, then

$$(M, t) = \textstyle\sum_i \alpha_i (M_i, t_i) \quad \text{with} \quad \sum_i \alpha_i = 1, \alpha_i \geq 0,$$

where M_i satisfies

$$\text{rank}(M_i) \leq r, \quad \text{and} \quad t_i \geq f(\|M_i\|_g) = f(\|M_i\|_{g,r*}),$$

where the equality is due to (12) in Lemma 3.1. Since f is convex and increasing, it holds that the composition $f(\|\cdot\|_{g,r*})$ is convex (see [Hiriart-Urruty and Lemaréchal, 2013, Proposition IV.2.1.8]). Thus,

$$t := \textstyle\sum_i \alpha_i t_i \geq \sum_i \alpha_i f(\|M_i\|_{g,r*}) \geq f\left(\|\sum_i \alpha_i M_i\|_{g,r*}\right) = f\left(\|M\|_{g,r*}\right),$$

which implies that $(M, t) \in \text{epi}(f(\|\cdot\|_{g,r*}))$, and (15) follows. Applying [Rockafellar, 1970, Theorem 15.3] to $f(\|\cdot\|_{g,r*})$ shows (14). □

This result generalizes the corresponding result in [Grussler et al., 2016a], in which the special case $f(x) \equiv x^2$ and $\|\cdot\|_g$ being the Frobenius norm is shown. For linear $f(x) \equiv x$, the biconjugate in (15) reduces to the low-rank inducing norms of Section 3. Therefore, they can be characterized as follows.

COROLLARY 4.1
Let $r \in \mathbb{N}$ be such that $1 \leq r \leq q := \min\{m, n\}$. Then

$$\|\cdot\|_{r*} = \left(\|\cdot\|_F + \chi_{\text{rank}(\cdot)\leq r}\right)^{**},$$
$$\|\cdot\|_{\ell_\infty,r*} = \left(\|\cdot\|_{\ell_\infty} + \chi_{\text{rank}(\cdot)\leq r}\right)^{**},$$

and the nuclear norm satisfies

$$\|\cdot\|_{\ell_1} = \left(\|\cdot\|_g + \chi_{\text{rank}(\cdot)\leq 1}\right)^{**},$$

where $\|\cdot\|_g$ is an arbitrary unitarily invariant norm that satisfies $\|M\|_g = \sigma_1(M)$ for all rank-1 matrices M. □

Proof This follows immediately from Theorem 1, since $\|\cdot\|_{r*} = \|\cdot\|_{\ell_2,r*}$, where $\|\cdot\|_{\ell_2} = \|\cdot\|_F$ is the Frobenius norm, and from Proposition 3.2. □

REMARK 1
This nuclear norm representation differs from the one in [Fazel et al., 2001; Fazel, 2002], where it is shown that $\|\cdot\|_{\ell_1} = \left(\text{rank} + \chi_{B^1_{\ell_\infty}}\right)^{**}$, i.e., it is the convex hull of the rank function restricted to the unit spectral norm ball.□

Using Theorem 1, optimal convex relaxations of rank constrained problems

$$\begin{aligned} \underset{M}{\text{minimize}} \quad & f_0(M) + f(\|M\|_g) \\ \text{subject to} \quad & \text{rank}(M) \leq r, \end{aligned} \tag{16}$$

can be provided, where $f_0 : \mathbb{R}^{n\times m} \to \mathbb{R} \cup \{\infty\}$ is a proper and closed convex function and $f : \mathbb{R}_{\geq 0} \to \mathbb{R} \cup \{\infty\}$ is an increasing and closed convex function. The problem in (16) is equivalent to minimizing $f_0 + f_{\text{reg}}$ with the non-convex f_{reg} defined in Theorem 1. Therefore, the optimal convex relaxation of (16) is given by

$$\underset{M}{\text{minimize}} \quad f_0(M) + f(\|M\|_{g,r*}). \tag{17}$$

Including an additional regularization parameter $\theta \geq 0$ (that can be included in f) yields the following proposition.

PROPOSITION 4.1
Assume that $f_0 : \mathbb{R}^{n\times m} \to \mathbb{R} \cup \{\infty\}$ is a proper closed convex function, and that $r \in \mathbb{N}$ is such that $1 \leq r \leq \min\{m, n\}$. Let $f : \mathbb{R}_{\geq 0} \to \mathbb{R} \cup \{\infty\}$ be an

increasing, proper closed convex function, and let $\theta \geq 0$. Then

$$\inf_{\substack{M \in \mathbb{R}^{n \times m} \\ \operatorname{rank}(M) \leq r}} \left[f_0(M) + \theta f(\|M\|_g) \right] \geq \inf_{M \in \mathbb{R}^{n \times m}} \left[f_0(M) + \theta f(\|M\|_{g,r*}) \right]. \tag{18}$$

If $M^\star$ solves the problem on the right such that $\operatorname{rank}(M^\star) \leq r$, then equality holds, and $M^\star$ is also a solution to the problem on the left. □

Proof The inequality holds since $f(\|\cdot\|_{g,r*}) = f_{\text{reg}}^{**} \leq f_{\text{reg}}$. From Lemma 3.1 it follows that if $\operatorname{rank}(M^\star) \leq r$ then

$$f_{\text{reg}}^{**}(M^\star) = f(\|M^\star\|_{g,r*}) = f(\|M^\star\|_g) = f_{\text{reg}}(M^\star),$$

which implies that the lower bound is attained with $M^\star$ and equality holds.□

Since the nuclear norm is obtained by creating a low-rank inducing norm with $r = 1$, it follows that any nuclear norm regularized problem can be interpreted as an optimal convex relaxation to a non-convex problem of the form (16), with the constraint $\operatorname{rank}(M) \leq 1$.

Proposition 4.1 also covers the results in our previous work [Grussler et al., 2016a], where the matrix approximation problem

$$\begin{aligned} &\min_{\substack{M \in \mathbb{R}^{n \times m} \\ \operatorname{rank}(M) \leq r}} \left[\tfrac{1}{2}\|N - M\|_F^2 + h(M) \right] \\ &\qquad = \min_{\substack{M \in \mathbb{R}^{n \times m} \\ \operatorname{rank}(M) \leq r}} \left[\tfrac{1}{2}\|N\|_F^2 - \langle N, M \rangle + \tfrac{1}{2}\|M\|_F^2 + h(M) \right], \end{aligned}$$

is considered. Letting

$$f_0(\cdot) = \tfrac{1}{2}\|N\|_F^2 - \langle N, \cdot \rangle + h(\cdot), \quad f(x) = \tfrac{1}{2}x^2, \quad \text{and} \quad \|\cdot\|_g = \|\cdot\|_F,$$

the results in [Grussler et al., 2016a] are a special cases of Theorem 1.

5. Computability

This section addresses the computability of convex optimization problems involving low-rank inducing regularizers of the form $f(\|\cdot\|_{g,r*})$. We restrict ourselves to low-rank inducing Frobenius and spectral norm regularizers. A requirement for the optimal convex relaxation problem in (17) to be solved efficiently, is that these regularizers are suitable for numerical optimization.

Assuming that f_0 and f are SDP representable, it is shown that (17) can be solved via semi-definite programming. To be able to solve larger problems using first-order proximal splitting methods (see [Combettes and Pesquet, 2011; Parikh and Boyd, 2014] and references therein), we show how to efficiently compute the proximal mappings of the considered regularizers. The

computational cost of computing these proximal mappings is comparable to the cost of computing the proximal mapping for the nuclear norm, since the cost in all cases is dominated by the singular value decomposition.

In order to deal with increasing convex functions f in (17), the problem is rewritten into the equivalent epigraph form

$$\underset{M,v}{\text{minimize}}\ f_0(M) + f(v) + \chi_{\text{epi}(\|\cdot\|_{g,r*})}(M, v). \tag{19}$$

5.1 SDP representation

The low-rank inducing Frobenius norm and spectral norm

$$\|M\|_{r*} := \max_{\|Y\|_r \le 1} \langle M, Y\rangle = \max_{\|Y\|_r^2 \le 1} \langle M, Y\rangle, \tag{20}$$

$$\|M\|_{\ell_\infty, r*} := \max_{\|Y\|_{\ell_1,r} \le 1} \langle M, Y\rangle, \tag{21}$$

are SDP representable via $\|Y\|_r^2$ and $\|Y\|_{\ell_1,r}$. From [Grussler and Rantzer, 2015; Grussler et al., 2016a], it is known that

$$\begin{aligned} \|Y\|_r^2 = \min_{T,\gamma}\ & \operatorname{trace}(T) - \gamma(n-r) \\ \text{s.t.}\ & \begin{pmatrix} T & Y \\ Y^T & I \end{pmatrix} \succeq 0,\ T \succeq \gamma I. \end{aligned}$$

Similarly, one can verify that

$$\begin{aligned} \|Y\|_{\ell_1,r} = \min_{T_1,T_2,\gamma}\ & \frac{1}{2}[\operatorname{trace}(T_1) + \operatorname{trace}(T_2) - (n+m-2r)\gamma] \\ \text{s.t.}\ & \begin{pmatrix} T_1 & Y \\ Y^T & T_2 \end{pmatrix} \succeq 0,\ T_1, T_2 \succeq \gamma I, \end{aligned}$$

which generalizes the SDP representation of $\|Y\|_{\ell_1,\min\{m,n\}}$ in [Recht et al., 2010]. This implies that

$$\begin{aligned} \|M\|_{r*} = \max_{Y,T,\gamma}\ & \langle M, Y\rangle \\ \text{s.t.}\ & \begin{pmatrix} T & Y \\ Y^T & I \end{pmatrix} \succeq 0,\ T \succeq \gamma I, \\ & \operatorname{trace}(T) - \gamma(n-r) \le 1, \end{aligned}$$

$$\begin{aligned} \|M\|_{\ell_\infty, r*} = \max_{Y,T_1,T_2,\gamma}\ & \langle M, Y\rangle \\ \text{s.t.}\ & \begin{pmatrix} T_1 & Y \\ Y^T & T_2 \end{pmatrix} \succeq 0,\ T_1, T_2 \succeq \gamma I, \\ & \frac{1}{2}[\operatorname{trace}(T_1) + \operatorname{trace}(T_2) - (n+m-2r)\gamma] \le 1, \end{aligned}$$

However, these formulations cannot be used in convex optimization problems with M as a decision variable due to the inner product $\langle M, Y \rangle$. Therefore, we use duality to arrive at

$$\begin{aligned} \|M\|_{r*} = \min_{W_1, W_2, k} \quad & \frac{1}{2}(\operatorname{trace}(W_2) + k) \\ \text{s.t.} \quad & \begin{pmatrix} kI - W_1 & M \\ M^T & W_2 \end{pmatrix} \succeq 0, \ W_1 \succeq 0, \\ & \operatorname{trace}(W_1) = (n-r)k; \end{aligned}$$

$$\begin{aligned} \|M\|_{\ell_\infty, r*} = \min_{W_1, W_2, k} \quad & k \\ \text{s.t.} \quad & \begin{pmatrix} kI - W_1 & M \\ M^T & kI - W_2 \end{pmatrix} \succeq 0, \ W_1, W_2 \succeq 0, \\ & \operatorname{trace}(W_1) + \operatorname{trace}(W_2) = [(n-r) + (m-r)]k. \end{aligned}$$

These formulations can be used to, e.g. solve problems on the epigraph form (19) by enforcing the respective costs to be smaller than or equal to $v \in \mathbb{R}$. This gives constraints of the form $\|M\|_{g,r*} \leq v$, i.e., $(M, v) \in \operatorname{epi}(\|\cdot\|_{g,r*})$. If f and f_0 are SDP representable, then (19) can be solved via semi-definite programming.

5.2 Splitting algorithms

Conventional SDP solvers are often based on interior point methods (see [Toh et al., 1999; Peaucelle et al., 2002]). These have good convergence properties, but the iteration complexity typically grows unfavorably with the problem dimension. This limits their application to small or medium scale problems. First order proximal splitting methods (see e.g. [Combettes and Pesquet, 2011; Parikh and Boyd, 2014]) typically have a lower complexity per iteration, and are thus more suitable for large problems.

These methods require the proximal mapping for all non-smooth parts of the problem to be available. The *proximal mapping* for a proper closed and convex functions $h : \mathbb{R}^{n \times m} \to \mathbb{R} \cup \{\infty\}$ is defined as

$$\operatorname{prox}_{\gamma h}(Z) := \operatorname*{argmin}_X \left(h(X) + \frac{1}{2\gamma} \|X - Z\|_F^2 \right). \tag{22}$$

Applying proximal splitting methods to (19) therefore requires that the proximal mapping of $\chi_{\operatorname{epi}(\|\cdot\|_{g,r*})}$ is readily computable. Since $\chi_{\operatorname{epi}(\|\cdot\|_{g,r*})}$ is an indicator function of the epigraph set, the proximal mapping becomes a projection, which is denoted by $\Pi_{\operatorname{epi}(\|\cdot\|_{g,r*})}$.

The epigraph of a norm is a cone (see [Bauschke and Combettes, 2011, Proposition 10.2]). Appealing to the Moreau-decomposition (see [Bauschke

and Combettes, 2011, Theorem 6.29]), we compute the projection $\Pi_{\text{epi}(\|\cdot\|_{g,r*})}$ via

$$\Pi_{\text{epi}(\|\cdot\|_{g,r*})}(Z, z_v) = (Z, z_v) - \Pi_{(\text{epi}(\|\cdot\|_{g,r*})^\circ}(Z, z_v), \tag{23}$$

where $Z \in \mathbb{R}^{n\times m}$, $z_v \in \mathbb{R}$, and $\Pi_{(\text{epi}(\|\cdot\|_{g,r*})^\circ}$ is projection onto the polar cone (which is the negative dual cone of $\text{epi}(\|\cdot\|_{g,r*})$ by definition).

Algorithms for projecting onto the polar cones of the low-rank inducing Frobenius and spectral norms are derived in Section A.2. In these algorithms, the first step is to perform a singular value decomposition of the prox argument $Z \in \mathbb{R}^{n\times m}$. Then a vector optimization problem of dimension $q := \min\{m, n\}$ needs to be solved. To this end, a nested binary search is applied that only requires the solutions to simple optimization problems with at most $r + 1$ decision variables.

In case of the low-rank inducing Frobenius norm, these problems can be solved explicitly, and results in an overall worst-case complexity of $\mathcal{O}(\log(r)\log(q-r))$ with an additional $\mathcal{O}(q)$ to set up the inner problems and to return the full solution. The cost of the prox computation is therefore dominated by the cost of computing the SVD. For large q one may consider sparse SVD algorithms such as [Liu et al., 2013].

The projection onto the epigraph of the low-rank inducing spectral norm is performed via the projection onto the epigraph of the truncated nuclear norm (modulo a sign flip). Since this requires a third layer in the nested binary search, the worst-case complexity is given by $\mathcal{O}(\log^2(r)\log(q-r)+q)$. In [Wu et al., 2014], another algorithm to project onto the truncated nuclear norm is presented. It uses similar techniques, but performs a linear search for finding the parameters and thus has a higher worst case computational cost.

Finally, note that the detour over the epigraph projection is not needed for all increasing functions. The proximal mapping for the low-rank inducing Frobenius and spectral norms can be derived very similarly to the epigraph case in Section A.2. The proximal mapping for the squared low-rank inducing Frobenius norm is derived in [Eriksson et al., 2015; Grussler et al., 2016a]. Details are omitted for brevity.

6. Examples: Matrix Completion

The matrix completion problem seeks to complete a low-rank matrix based on limited knowledge about its entries. The problem is often posed as

$$\begin{aligned} &\text{minimize} && \text{rank}(X) \\ &\text{subject to} && \hat{x}_{ij} = x_{ij},\ (i,j) \in \mathcal{I}, \end{aligned} \tag{24}$$

where $\mathcal{I}$ denotes the index set of the known entries. Another formulation that fits with the low-rank inducing norms proposed in this paper is

$$\begin{array}{ll} \text{minimize} & \|X\|_g \\ \text{subject to} & \operatorname{rank}(X) \leq r \\ & \hat{x}_{ij} = x_{ij},\ (i,j) \in \mathcal{I}, \end{array} \tag{25}$$

where r is the target rank of the matrix to be completed. In the following, two examples of this form will be convexified using different low-rank inducing norms. That is,

$$\begin{array}{ll} \text{minimize} & \|X\|_{g,r*} \\ \text{subject to} & \hat{x}_{ij} = x_{ij},\ (i,j) \in \mathcal{I}, \end{array} \tag{26}$$

is solved for different low-rank inducing norms $\|\cdot\|_{g,r*}$.

Further, we discuss a covariance completion problem which is a generalization of the problem above. In all problems it will be observed that there are convex relaxations with low-rank inducing norms whose solutions give better completion than the nuclear norm approach, without increasing the rank.

6.1 Example 1

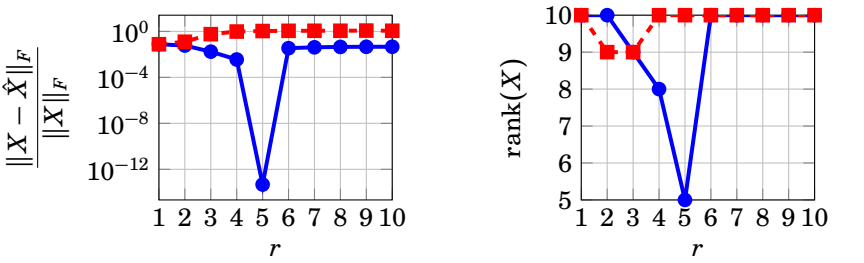
In the first problem, which is taken from [Grussler et al., 2016a], the matrix $\hat{X}$ to be completed is a low-rank approximation of the Hankel matrix

$$H = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ 1 & & \iddots & & 0 \\ \vdots & & \iddots & & \vdots \\ 1 & \iddots & \iddots & & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{pmatrix} \in \mathbb{R}^{10\times 10}. \tag{27}$$

Let the singular value decomposition of H be given by $H = \sum_{i=1}^{10} \sigma_i(H) u_i u_i^T$ and

$$\hat{X} := \sum_{i=1}^{5} \sigma_i(H) u_i u_i^T \quad \text{and } \mathcal{I} := \{(i,j) : \hat{x}_{ij} > 0\},$$

where $\mathcal{I}$ is the index set of known entries. The cardinality of $\mathcal{I}$ is 78, i.e. 22 out of 100 entries are unknown. Figure 1 shows the completion errors and ranks of the completed matrices for different value of r. The nuclear norm ($r = 1$) returns a full rank matrix and gives a worse completion error than all other low-rank inducing Frobenius norms. For $r = 5$, the solution with the low-rank inducing Frobenius norm has rank 5. Given the known entries, this is the matrix of smallest Frobenius norm which has at most rank 5, by Proposition 4.1. As indicated by the small relative error, this matrix



(a) Relative completion errors of:
(26) with $\|\cdot\|_{g,r*} = \|\cdot\|_{r*}$
(26) with $\|\cdot\|_{g,r*} = \|\cdot\|_{\ell_\infty,r*}$

(b) Rank of the solutions to:
(26) with $\|\cdot\|_{g,r*} = \|\cdot\|_{r*}$
(26) with $\|\cdot\|_{g,r*} = \|\cdot\|_{\ell_\infty,r*}$

Figure 1. Example 1: Relative completion error and ranks of the solution to (26) with $\|\cdot\|_{g,r*} = \|\cdot\|_{r*}$ and $\|\cdot\|_{g,r*} = \|\cdot\|_{\ell_\infty,r*}$.

coincides with $\hat{X}$. In fact, this is also verified analytically in [Grussler et al., 2016a, Theorem 3].

Notice that

$$10^{1.2}\text{rank}(\hat{X})\log(10) \gg \text{card}(\mathcal{I}) = 78,$$

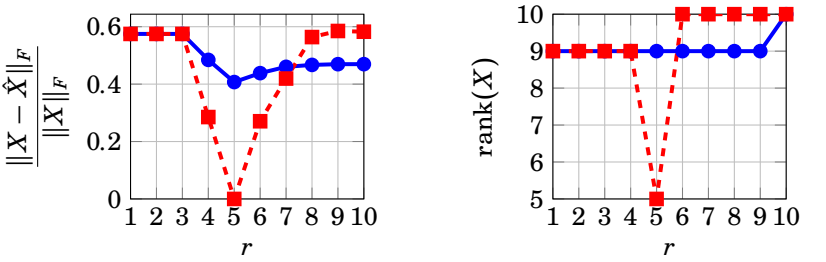
which is why exact completion results for the nuclear norm (see [Candès and Recht, 2009]) do not apply. Furthermore, the low-rank inducing spectral norm shows no improvement in comparison with the nuclear norm.

6.2 Example 2

In the this second example, it as assumed that

$$\hat{X} := \sum_{j=1}^{5} \sigma_j \sum_{i=1}^{5} (H) u_i v_i^T \quad \text{and } \mathcal{I} := \{(i,j) : \hat{x}_{ij} > 0\},$$

where H is given in (27) with the singular value decomposition $H = \sum_{i=1}^{10} \sigma_i(H) u_i v_i^T$. The cardinality of $\mathcal{I}$ is 67, that is, 33 out of 100 entries are unknown. Figure 2 shows the completion errors and ranks of the completed matrices with different value of r. The nuclear norm ($r = 1$) returns a close to full rank matrix with a relative completion error that is among the largest for all r. In this example, the low-rank inducing spectral norms perform significantly better than the low-rank inducing Frobenius norms. In particular, for $r = 5$, the low-rank inducing spectral norm returns a rank 5 solution. Given the known entries, this solution is the matrix of smallest spectral norm of rank at most 5 (see Proposition 4.1). As indicated by the zero completion error, this matrix coincides with $\hat{X}$. Just as in the exact recovery result for the low-rank inducing Frobenius norm in [Grussler et



(a) Relative completion errors of:
(26) with $\|\cdot\|_{g,r*} = \|\cdot\|_{r*}$
(26) with $\|\cdot\|_{g,r*} = \|\cdot\|_{\ell_\infty,r*}$

(b) Rank of the solutions to:
(26) with $\|\cdot\|_{g,r*} = \|\cdot\|_{r*}$
(26) with $\|\cdot\|_{g,r*} = \|\cdot\|_{\ell_\infty,r*}$

Figure 2. Example 2: Relative completion error and ranks of the solution to (26) with $\|\cdot\|_{g,r*} = \|\cdot\|_{r*}$ and $\|\cdot\|_{g,r*} = \|\cdot\|_{\ell_\infty,r*}$.

al., 2016a, Theorem 3], it can be analytically guaranteed that the low-rank inducing spectral norm with $r = 5$ recovers the true matrix. Analogous to the previous example,

$$10^{1.2}\text{rank}(\hat{X})\log(10) \gg \text{card}(\mathcal{I}) = 67,$$

which is why exact completion with the nuclear norm cannot be expected.

In both examples, the nuclear norm neither produces the lowest rank solution, nor recovers the true matrix. In contrast, other low-rank inducing norms succeed in both aspects. This indicates that the richness in the family of low-rank inducing norms should be exploited to achieve satisfactory performance in rank constrained problems. In practical applications, the 'true' matrix is not known, and this comparison cannot be made. However, cross validation techniques can often be used to assess the performance.

6.3 Covariance Completion

In this section, the performance of the low-rank inducing Frobenius and spectral norms is evaluated by means of a covariance completion problem. This is a variation of the matrix completion problems above.

Consider the linear state-space system

$$\dot{x}(t) = Ax(t) + Bu(t),$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $m \leq n$ and $u(t)$ is a zero-mean stationary stochastic process. For Hurwitz A and reachable (A, B), it has been shown (see [Georgiou, 2002a; Georgiou, 2002b]) that the following are equivalent:

i. $X := \lim_{t\to\infty} \mathbf{E}\left(x(t)\,x^T(t)\right) \succeq 0$ is the steady-state covariance matrix of $x(t)$, where $\mathbf{E}(\cdot)$ denotes the expected value.

ii. $\exists H \in \mathbb{R}^{m\times n} : AX + XA^T = -(BH + H^TB^T)$.

iii. $\operatorname{rank}\begin{pmatrix} AX + XA^T & B \\ B^T & 0 \end{pmatrix} = \operatorname{rank}\begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}$.

In particular, $H = \frac{1}{2}\mathbf{E}\left(u(t)\,u^T(t)\right)B^T$ if u is white noise. The problem considered in [Chen et al., 2013; Lin et al., 2013; Zare et al., 2016a; Zare et al., 2015; Zare et al., 2016b] is to reconstruct the partially known covariance matrix X and the input matrix B, via $M = -(BH + H^TB^T)$, where the rank of M sets an upper bound on the rank of B, i.e., the number of inputs. The objective is to keep the rank of M low, while achieving satisfactory completion of X. In [Chen et al., 2013; Lin et al., 2013; Zare et al., 2016a; Zare et al., 2015; Zare et al., 2016b] the problem is addressed by searching for the lowest rank solution:

$$\begin{aligned} \text{minimize} \quad & \operatorname{rank}(M) \\ \text{subject to} \quad & \hat{x}_{ij} = x_{ij},\ (i,j) \in \mathcal{I} \\ & A\hat{X} + \hat{X}A^T = -M \\ & \hat{X} \succeq 0, \end{aligned} \tag{28}$$

where $\mathcal{I}$ denotes set of pairs of indices of known entries. Another option is to search for a low-rank solution, while minimizing the norm of M measured by some unitarily invariant norm. This helps to avoid overfitting, and gives

$$\begin{aligned} \text{minimize} \quad & \|M\|_g \\ \text{subject to} \quad & \operatorname{rank}(M) \leq r \\ & \hat{x}_{ij} = x_{ij},\ (i,j) \in \mathcal{I} \\ & A\hat{X} + \hat{X}A^T = -M \\ & \hat{X} \succeq 0. \end{aligned} \tag{29}$$

The authors in [Chen et al., 2013; Lin et al., 2013; Zare et al., 2016a; Zare et al., 2015; Zare et al., 2016b] convexify the problem by using the nuclear norm. In [Grussler et al., 2016b], a similar problem is instead convexified with the low-rank inducing Frobenius norm. We will also make a comparison with convex relaxations based on low-rank inducing spectral norms. All these convex relaxations are of the form

$$\begin{aligned} \text{minimize} \quad & \|M\|_{g,r*} \\ \text{subject to} \quad & \hat{x}_{ij} = x_{ij},\ (i,j) \in \mathcal{I} \\ & A\hat{X} + \hat{X}A^T = -M \\ & \hat{X} \succeq 0, \end{aligned} \tag{30}$$

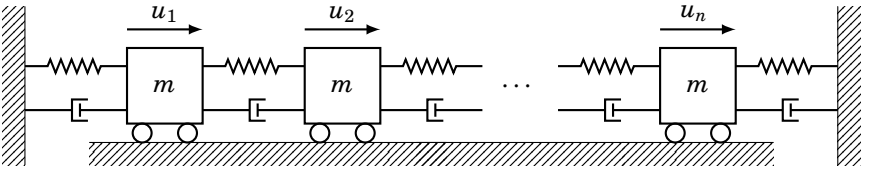


Figure 3. Mass-spring-damper system with n masses and input forces $u_1, \ldots, u_n$.

with the appropriate low-rank inducing norm in the cost.

Mass-spring-damper system The system considered in our example is the so-called *mass-spring-damper system (MSD)* (see [Zare et al., 2015; Grussler et al., 2016b]) with n masses (see Figure 3).

Assuming that the stochastic forcing affects all masses, this yields the following state-space representation

$$\dot{x}(t) = Ax(t) + B\xi(t)$$

with

$$A = \begin{pmatrix} 0 & I \\ -S & -I \end{pmatrix} \in \mathbb{R}^{2n \times 2n}, \quad B = \begin{pmatrix} 0 \\ I \end{pmatrix} \in \mathbb{R}^{2n \times n}.$$

Here, S is a symmetric tridiagonal Toeplitz matrix with 2 on the main diagonal, -1 on the first upper and lower sub-diagonals, and I and 0 stand for the identity and zero matrices of appropriate size. The state vector x consists of the positions and velocities of the masses, $x = (p, v)$. Furthermore, $\xi(t)$ is generated via a low-pass filtered white noise signal $w(t)$ with unit covariance $\mathbf{E}\left(w(t)w(t)^T\right) = I$ as

$$\dot{\xi}(t) = -\xi(t) + w(t).$$

The extended covariance matrix

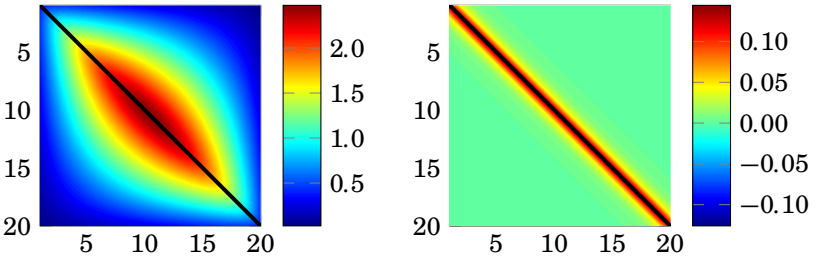
$$X_e := \mathbf{E}\left(x_e x_e^T\right) = \begin{pmatrix} X & X_{x\xi} \\ X_{\xi x} & X_\xi \end{pmatrix} \quad \text{with} \quad x_e := \begin{pmatrix} x(t) \\ \xi(t) \end{pmatrix}$$

is then determined by

$$A_e X_e + X_e A_e^T = -B_e B_e^T,$$

where X is the steady-state covariance matrix of $x(t)$ and

$$A_e := \begin{pmatrix} A & B \\ 0 & -I \end{pmatrix}, \quad B_e := \begin{pmatrix} 0 \\ I \end{pmatrix}.$$



(a) X_{pp} (b) X_{vv}

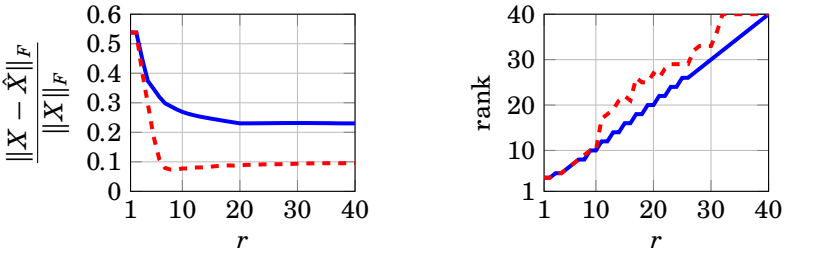
Figure 4. Interpolated colormap of the steady-state covariance matrices X_{pp} and X_{vv} of the positions and the velocities in the MSD system with $n = 20$. —— indicates the available one-point correlations.

In our numerical experiments, we choose $n = 20$ masses and assume that only one-point correlations are available, i.e. the known entries are given by the diagonal of X. The steady-state covariance matrix can be partitioned as

$$X = \begin{pmatrix} X_{pp} & X_{pv} \\ X_{vp} & X_{vv} \end{pmatrix},$$

where X_{pp} and X_{vv} are the covariance matrices of the positions and the velocities, respectively. To visualize the effects of using different low-rank inducing norms in (30), an interpolated colormap of the reconstructed $\hat{X}_{pp}$ and $\hat{X}_{vv}$ is used (see Figure 6). The interpolated colormap of the true covariance matrices is shown in Figure 4, where the black lines indicate the known measured entries.

Figure 5 displays the relative errors and the ranks of the estimates obtained by (30) for different low-rank inducing norms as functions of r. The nuclear norm minimization ($r = 1$), as shown in Figures 6(a) and 6(b), gives the same rank as both the low-rank inducing Frobenius and spectral norms for $r = 2$. However, the latter approaches give better completions. The low-rank inducing spectral norm outperforms the low-rank inducing Frobenius norm for all $r \geq 2$. In particular, $r = 9$ gives the best completion, with a solution of rank 10 (see Figures 6(e) and 6(f)). It is interesting that the solutions to (30) with $r = 10$ for both the low-rank inducing Frobenius and spectral norms are of rank 10. By Proposition 4.1, there are no better feasible rank-10 solutions that minimize the Frobenius and spectral norms respectively. The solution to (30) with the low-rank inducing Frobenius norm and $r = 10$, is shown in Figure 6(c) and 6(d). The solution to the low-rank inducing spectral norm with $r = 10$ looks identical to Figures 6(e) and 6(f).



(a) Relative completion errors of:
(30) with $\|\cdot\|_{g,r*} = \|\cdot\|_{r*}$
(30) with $\|\cdot\|_{g,r*} = \|\cdot\|_{\ell_\infty,r*}$

(b) Rank of the solutions to:
(30) with $\|\cdot\|_{g,r*} = \|\cdot\|_{r*}$
(30) with $\|\cdot\|_{g,r*} = \|\cdot\|_{\ell_\infty,r*}$

Figure 5. Relative errors and ranks of solutions to (30) with $\|\cdot\|_{g,r*} = \|\cdot\|_{r*}$ and $\|\cdot\|_{g,r*} = \|\cdot\|_{\ell_\infty,r*}$.

7. Extensions

7.1 The Vector Case

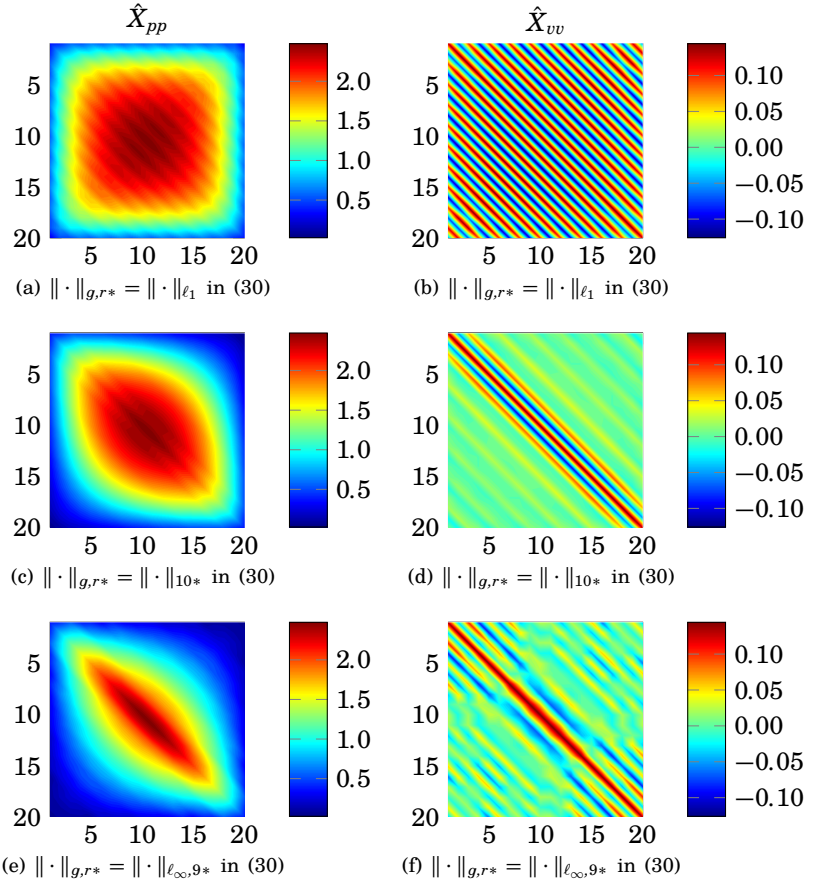
The results in Section 4 translate to the corresponding vector-valued problem, by replacing rank with cardinality, and $\|M\|_g$ with $\|x\|_g := \|diag(x)\|_g$. Therefore, our optimality interpretations, as well as the variety of regularizers, can be applied to problems such as sparse linear regression (see [Tibshirani, 1996; Candès et al., 2006; Argyriou et al., 2012]). The SDP representation and the proximal mapping computations in Section 5 carry over, though here they have lower computational cost. For instance, the required SVD in the prox computations turns into a sorting, which reduces the total complexity.

7.2 Atomic Norms

In [Chandrasekaran et al., 2012], the concept of an atomic norm is introduced. An atomic norm is defined as the gauge function or the Minkowski functional of the convex hull of a set of atoms $\mathcal{A}$ (see [Chandrasekaran et al., 2012])

$$\|x\|_{\mathcal{A}} := \inf\{t > 0 : t^{-1}x \in \text{conv}(\mathcal{A})\}. \tag{31}$$

Despite its name, the atomic norm is not necessarily a norm, but always defines a distance measure. The atoms are used to model properties of a quantity that is to be estimated. The atomic norm is a way of imposing these properties on the solution of an optimization problem. In [Chandrasekaran et al., 2012], examples of atomic sets that naturally appear in different applications are listed. For instance, if $\mathcal{A}$ is the set of rank 1 matrices with unit Frobenius norm, then the resulting atomic norm is the nuclear norm.



(a) $\|\cdot\|_{g,r*} = \|\cdot\|_{\ell_1}$ in (30)
(b) $\|\cdot\|_{g,r*} = \|\cdot\|_{\ell_1}$ in (30)
(c) $\|\cdot\|_{g,r*} = \|\cdot\|_{10*}$ in (30)
(d) $\|\cdot\|_{g,r*} = \|\cdot\|_{10*}$ in (30)
(e) $\|\cdot\|_{g,r*} = \|\cdot\|_{\ell_\infty,9*}$ in (30)
(f) $\|\cdot\|_{g,r*} = \|\cdot\|_{\ell_\infty,9*}$ in (30)

Figure 6. Recovered covariance matrices of positions ($\hat{X}_{pp}$ to the left), and velocities ($\hat{X}_{vv}$ to the right), in the MSD system with $n = 20$ masses resulting from problem (30), with different low-rank inducing norms.

More generally, all low-rank inducing norms in Section 3 can be considered as atomic norms, because Lemma 3.2 implies that

$$\|X\|_{g,r*} = \inf\left\{t > 0 : t^{-1}X \in \operatorname{conv}(E_{g,r})\right\},$$

with $E_{g,r} := \{X \in \mathbb{R}^{n\times m} : \|X\|_g = 1, \operatorname{rank}(X) \leq r\}$.

As presented for the low-rank inducing norms and regularizers in Section 4, this section provides similar optimality interpretations for general atomic norms. It is assumed that the atoms lie within a finite-dimensional real Hilbert space $\mathcal{H}$ with inner product $\langle\cdot,\cdot\rangle$, i.e. $\mathcal{A} \subset \mathcal{H}$. In the following, the definitions of the *conic hull* of $\mathcal{A} \subset \mathcal{H}$

$$\operatorname{cone}(\mathcal{A}) := \{\alpha x : x \in \mathcal{A},\ \alpha \geq 0\},$$

and the *polar gauge function* to (31)

$$\|y\|_{\mathcal{A}}^{\circ} := \inf\{\mu \geq 0 : \langle x, y\rangle \leq \mu\|x\|_{\mathcal{A}} \text{ for all } x \in \mathcal{H}\},$$

are needed. Note that, if the atomic norm in (31) is a norm, then the polar gauge function is equal to the corresponding dual norm Our optimality interpretations will hold if the atomic set denoted by $\mathcal{A}_G$ can be represented as

$$\mathcal{A}_G := \{a \in \operatorname{cone}(\mathcal{A}) : G(a) = 1\}, \tag{32}$$

where $\mathcal{A} \subset \mathcal{H}$, and $G : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$, satisfy the following assumptions.

Assumption 1
The set $\mathcal{A} \subset \mathcal{H}$ is nonempty such that $\operatorname{cone}(\mathcal{A})$ is closed. The function $G : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ is positively homogeneous (of degree 1), proper, closed, convex and nonnegative with $G(a) > 0$ for all $a \in \mathcal{A} \setminus \{0\}$.

Many atomic sets from [Chandrasekaran et al., 2012] satisfy these assumptions. For example, if $\mathcal{A}$ is the set of all permutation matrices, then

$$\|\cdot\|_{\mathcal{A}} = \|\cdot\|_{\mathcal{A}_G} \quad \text{with} \quad G(\cdot) = \|\cdot\|_{\ell_\infty}.$$

Similar constructions apply to the atomic norms that are induced, e.g. by *binary vectors*, *sparse vectors*, *low-rank matrices*, *vectors from lists*, and many more (see [Chandrasekaran et al., 2012])

Using the definition of atomic norms in (31), an explicit expression of the atomic norm associated with $\mathcal{A}_G$ is

$$\|x\|_{\mathcal{A}_G} = \inf\{t > 0 : t^{-1}x \in \operatorname{conv}(\{a \in \operatorname{cone}(\mathcal{A}) : G(a) = 1\})\}. \tag{33}$$

The next theorem gives optimality interpretations of these atomic norms, and generalizes Theorem 1 in the following two aspects:

I. The rank-constraint is generalized to other non-convex constraints.

II. The norms are replaced by more general functions G.

To prove the result, the following lemma is needed.

LEMMA 7.1
Let $\mathcal{A} \subset \mathcal{H}$ and $G : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ satisfy Assumption 1, and let $\mathcal{A}_G$ and $\|\cdot\|_{\mathcal{A}_G}$ be defined as in (32) and (33). Then,

i. $\operatorname{conv}(\mathcal{A}_G)$ is closed and bounded.

ii. $\|x\|_{\mathcal{A}_G} = 0$ if and only if $x = 0$.

iii. $\|x\|_{\mathcal{A}_G} \geq G(x)$ for all $x \in \mathcal{H}$, and $\|x\|_{\mathcal{A}_G} = G(x)$ for all $x \in \operatorname{cone}(\mathcal{A})$.

iv. For all $x \in \operatorname{dom}(\|\cdot\|_{\mathcal{A}_G})$ there exist $x_i \in \operatorname{cone}(\mathcal{A})$ such that

$$x = \textstyle\sum_i \alpha_i x_i, \quad \sum_i \alpha_i = 1, \ \alpha_i \geq 0, \quad \text{and} \quad G(x_i) = \|x\|_{\mathcal{A}_G}.$$

□

Proof Item i: Since $G + \chi_{\operatorname{cone}(\mathcal{A})}$ is coercive, it follows from [Bauschke and Combettes, 2011, Proposition 11.11] that the sub-level set

$$\{a \in \operatorname{cone}(\mathcal{A}) : G(a) \leq 1\}$$

is bounded. Thus the same applies to $\mathcal{A}_G$. Further, convexity of G implies that

$$\{x \in \mathcal{H} : G(x) = 1\}$$

is closed, because, by [Hiriart-Urruty and Lemaréchal, 2013, Proposition VI.1.3.3], it is the boundary of

$$\{x \in \mathcal{H} : G(x) \leq 1\}.$$

Thus, as the intersection of two closed sets is closed,

$$\mathcal{A}_G = \operatorname{cone}(\mathcal{A}) \cap \{x \in \mathcal{H} : G(x) = 1\}$$

is closed. Applying [Hiriart-Urruty and Lemaréchal, 2013, Theorem III.1.4.3] shows that $\operatorname{conv}(\mathcal{A}_G)$ is closed and bounded.

Item ii: This claim follows by [Hiriart-Urruty and Lemaréchal, 2013, Corollary V.1.2.6].

Item iii: Let us introduce the sub-levelset

$$S_G^s := \{x \in \mathcal{H} : G(x) \leq s\},$$

which by the positive homogeneity of G satisfies

$$S_G^s = \{sx \in \mathcal{H} : G(x) \leq 1\}$$

for all $s \geq 0$. By the definition of $\mathcal{A}_G$, it holds that

$$\begin{aligned}\operatorname{conv}(\mathcal{A}_G) &= \operatorname{conv}(\{a \in \operatorname{cone}(\mathcal{A}) : G(a) = 1\}) \\ &\subset \operatorname{conv}(\{a \in \operatorname{cone}(\mathcal{A}) : G(a) \leq 1\}) \\ &\subset \operatorname{conv}(\{a \in \mathcal{H} : G(a) \leq 1\}) \\ &= \{a \in \mathcal{H} : G(a) \leq 1\}) = S_G^1.\end{aligned}$$

This yields that

$$\begin{aligned}\|x\|_{\mathcal{A}_G} &= \inf\{t > 0 : x \in t\operatorname{conv}(\mathcal{A}_G)\} \\ &\geq \inf\{t > 0 : x \in tS_G^1\} \\ &= \inf\{t > 0 : G(x) \leq t\} = G(x)\end{aligned}$$

for all $x \in \mathcal{H}$, and the first claim of this item is proven.

To prove the second claim, let $x \in \operatorname{cone}(\mathcal{A})$. If $x \notin \operatorname{dom}(G)$, the above implies that

$$\|x\|_{\mathcal{A}_G} = G(x) = \infty.$$

Further, Item ii shows that

$$x = 0 \;\Rightarrow\; \|0\|_{\mathcal{A}_G} = G(0) = 0.$$

It remains to show the claim for $x \in \operatorname{dom}(G) \setminus \{0\}$. In this case, we can define $\bar{x} := G(x)^{-1}x$, which satisfies

$$\bar{x} \in \operatorname{cone}(\mathcal{A}) \quad \text{and} \quad G(\bar{x}) = 1,$$

i.e. $\bar{x} \in \mathcal{A}_G \subset \operatorname{conv}(\mathcal{A}_G)$, and therefore

$$\|\bar{x}\|_{\mathcal{A}_G} = \inf\{t > 0 : \bar{x} \in t\operatorname{conv}(\mathcal{A}_G)\} \leq \inf\{t > 0 : \bar{x} \in t\bar{x}\} = 1.$$

That is, $\|x\|_{\mathcal{A}_G} \leq G(x)$, which in conjunction with $\|x\|_{\mathcal{A}_G} \geq G(x)$ proves that

$$\|x\|_{\mathcal{A}_G} = G(x) \text{ for all } x \in \operatorname{cone}(\mathcal{A}).$$

Item iv: Since the claim holds trivially if $x = 0$, it is enough to assume that

$$x \in \operatorname{dom}(\|\cdot\|_{\mathcal{A}_G}) \setminus \{0\}.$$

By Item ii it follows that $\infty > \|x\|_{\mathcal{A}_G} > 0$. Further, Item i and the definition of $\|x\|_{\mathcal{A}_G}$ in (31) imply that

$$\|x\|_{\mathcal{A}_G}^{-1} x \in \operatorname{conv}(\mathcal{A}_G).$$

Thus,

$$\|x\|_{\mathcal{A}_G}^{-1} x = \textstyle\sum_i \alpha_i \bar{x}_i \quad \text{with} \quad \sum_i \alpha_i = 1,\ \alpha_i \geq 0,$$

where $\bar{x}_i$ satisfies

$$\bar{x}_i \in \operatorname{cone}(\mathcal{A}) \quad \text{and} \quad G(\bar{x}_i) = 1.$$

Defining $x_i := \bar{x}_i \|x\|_{\mathcal{A}_G}$, it follows that

$$x = \sum_i \alpha_i x_i \quad \text{with} \quad x_i \in \operatorname{cone}(\mathcal{A}).$$

Finally, the positive homogeneity of G ensures that

$$G(x_i) = G(\|x\|_{\mathcal{A}_G} \bar{x}_i) = \|x\|_{\mathcal{A}_G} G(\bar{x}_i) = \|x\|_{\mathcal{A}_G}. \qquad \square$$

THEOREM 2
Assume $\mathcal{A} \subset \mathcal{H}$ and $G : H \to \mathbb{R} \cup \{\infty\}$ satisfy Assumption 1, and let $\mathcal{A}_G$ and $\|\cdot\|_{\mathcal{A}_G}$ be defined as in (32) and (33). Further, let $f_{\text{reg}} := f(G(\cdot)) + \chi_{\operatorname{cone}(\mathcal{A})}$, where $f : \mathbb{R} \cup \{\infty\} \to \mathbb{R} \cup \{\infty\}$ is an increasing, proper closed convex function. Then,

$$f_{\text{reg}}^* = f^+(\|\cdot\|_{\mathcal{A}_G}^\circ), \tag{34}$$

$$f_{\text{reg}}^{**} = f(\|\cdot\|_{\mathcal{A}_G}). \tag{35}$$

$\square$

Proof Since $\|\cdot\|_{\mathcal{A}_G}$ is a Minkowski functional, it is closed function (see [Luenberger, 1968, Lemma 1 in 5.12]). Thus, $\operatorname{epi}(f(\|\cdot\|_{\mathcal{A}_G})$ is a closed set, and by Lemma 2.1,

$$\operatorname{epi}(f(\|\cdot\|_{\mathcal{A}_G})) = \operatorname{conv}(\operatorname{epi}(f_{\text{reg}})) \tag{36}$$

implies (35).

We start with $\operatorname{conv}(\operatorname{epi}(f_{\text{reg}})) \subset \operatorname{epi}(f(\|\cdot\|_{\mathcal{A}_G}))$. If $(x, t) \in \operatorname{conv}(\operatorname{epi}(f_{\text{reg}}))$, then

$$(x, t) = \sum_i \alpha_i (x_i, t_i) \quad \text{with} \quad \sum_i \alpha_i = 1,\ \alpha_i \geq 0,$$

where x_i satisfies

$$x_i \in \operatorname{cone}(\mathcal{A}), \quad \text{and} \quad t_i \geq f(G(x_i)) = f(\|x_i\|_{\mathcal{A}_G}),$$

and the equality follows by Lemma 7.1(Item iii). Since f is convex and increasing, it holds that the composition $f(\|\cdot\|_{\mathcal{A}_G})$ is convex (see [Hiriart-Urruty and Lemaréchal, 2013, Proposition IV.2.1.8]). Therefore,

$$t := \textstyle\sum_i \alpha_i t_i \geq \sum_i \alpha_i f(\|x_i\|_{\mathcal{A}_G}) \geq f\left(\|\sum_i \alpha_i x_i\|_{\mathcal{A}_G}\right) = f(\|x\|_{\mathcal{A}_G}),$$

and $(x, t) \in \operatorname{epi}(f(\|\cdot\|_{\mathcal{A}_G}))$.

Conversely, let $(x, t) \in \operatorname{epi}(f(\|\cdot\|_{\mathcal{A}_G}))$ with $\|x\|_{\mathcal{A}_G} \neq 0$. Lemma 7.1(Item iv) implies that

$$x = \textstyle\sum_i \alpha_i x_i \quad \text{with} \quad \sum_i \alpha_i = 1,\ \alpha_i \geq 0,$$

where x_i satisfies

$$x_i \in \operatorname{cone}(\mathcal{A}) \quad \text{and} \quad G(x_i) = \|x\|_{\mathcal{A}_G}.$$

Thus, $(x, t) = \sum_i \alpha_i (x_i, t)$ such that

$$t \geq f(\|x\|_{\mathcal{A}_G}) = f(G(x_i)), \quad \text{and} \quad x_i \in \operatorname{cone}(\mathcal{A}).$$

Consequently,

$$(x_i, t) \in \operatorname{epi}(f_{\text{reg}}), \text{ and therefore } (x, t) \in \operatorname{conv}(\operatorname{epi}(f_{\text{reg}})).$$

Lemma 7.1 (Item ii) shows that $(x, t) \in \operatorname{conv}(\operatorname{epi}(f_{\text{reg}}))$ is trivially fulfilled if $\|x\|_{\mathcal{A}_G} = 0$. Finally, (34) can be proven by applying [Rockafellar, 1970, Theorem 15.3] to $f(\|\cdot\|_{\mathcal{A}_G})$. □

Similarly to Section 4, this result gives rise to optimal convex relaxations for atomic norms.

Proposition 7.1
Assume $\mathcal{A} \subset \mathcal{H}$ and $G : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ satisfy Assumption 1, and let $\mathcal{A}_G$ and $\|\cdot\|_{\mathcal{A}_G}$ be defined as in (32) and (33). Further, let $f : \mathbb{R}\cup\{\infty\} \to \mathbb{R}\cup\{\infty\}$ be an increasing, closed convex function, and let $f_0 : H \to \mathbb{R}\cup\{\infty\}$ be closed, proper, and convex. For $\theta \geq 0$, it holds that

$$\inf_{x\in\mathcal{A}} \left[f_0(x) + \theta f(G(x))\right] \geq \inf_{x\in\operatorname{conv}(\mathcal{A})} \left[f_0(x) + \theta f(\|x\|_{\mathcal{A}_G})\right]. \tag{37}$$

If the right-hand side of the inequality is solved by $x^\star \in \mathcal{A}$, then $x^\star$ is a solution to the left-hand side. □

Proof By (4) and Theorem 2 it follows that

$$\inf_{x\in\operatorname{cone}(\mathcal{A})} \left[\tilde{f}_0(x) + \theta f(G(x))\right] \geq \inf_{x\in\mathcal{H}} \left[\tilde{f}_0(x) + \theta f(\|x\|_{\mathcal{A}_G})\right], \tag{38}$$

for any closed and proper convex function $\tilde{f}_0 : H \to \mathbb{R} \cup \{\infty\}$. In particular, let $\tilde{f}_0 = f_0 + \chi_{\text{conv}(\mathcal{A})}$, which is closed by Assumption 1. Then the left-hand side of (38) satisfies

$$\begin{aligned}\inf_{x \in \text{cone}(\mathcal{A})} \left[\tilde{f}_0(x) + \theta f(G(x))\right] &= \inf_{\substack{x \in \text{cone}(\mathcal{A}) \\ x \in \text{conv}(\mathcal{A})}} \left[f_0(x) + \theta f(G(x))\right] \\ &\leq \inf_{x \in \mathcal{A}} \left[f_0(x) + \theta f(G(x))\right],\end{aligned}$$

because $\mathcal{A} \subset \text{cone}(\mathcal{A}) \cap \text{conv}(\mathcal{A})$. The right-hand side of (38) satisfies

$$\inf_{x \in \mathcal{H}} \left[\tilde{f}_0(x) + \theta f(\|x\|_{\mathcal{A}_G})\right] = \inf_{x \in \text{conv}(\mathcal{A})} \left[f_0(x) + \theta f(\|x\|_{\mathcal{A}_G})\right],$$

and (37) is proven. The last claim follows by Lemma 7.1 (Item iii). □

In [Chandrasekaran et al., 2012] exact recovery results are presented for the cases when f_0 is an indicator of an affine set that contains measurement of an observed quantity $x_0 \in \mathcal{H}$. Let $\mathcal{Q} := \{x \in \mathcal{H} : Ax = Ax_0\}$ denote that affine set and let $f_0 = \chi_{\mathcal{Q}}$. Then the recovery problem becomes

$$\underset{x \in \mathcal{Q}}{\text{minimize}} \ \|x\|_{\mathcal{A}_G}.$$

Assume that this problem has a unique solution $x^\star$. In [Chandrasekaran et al., 2012], conditions on the measurement set $\mathcal{Q}$ are stated under which exact recovery $x^\star = x_0$ is guaranteed. The underlying assumption in [Chandrasekaran et al., 2012], is that for small k it holds that

$$x_0 = \sum_{i=1}^{k} c_i a_i \text{ with } c_i \geq 0 \text{ and } a_i \in \mathcal{A}_G.$$

That is, the observed quantity is assumed to be a conic combination of a few atoms. For many examples in [Chandrasekaran et al., 2012, Section 2.2], the assumption holds with $k = 1$ and $c_1 = 1$, i.e., $x_0 = a$ for some $a \in \mathcal{A}$. A notable exception is the case of low rank matrix recovery. In [Chandrasekaran et al., 2012], rank one matrices of unit norm are used as atoms, which yields the nuclear norm as the corresponding atomic norm. Therefore, a conic combination of r atoms is needed to recover a rank-r matrix x_0. By using a low-rank inducing norm $\|\cdot\|_{g,r*}$ instead, the matrix x_0 satisfies $x_0 = a$ for some $a \in \mathcal{A}$, where $\mathcal{A}$ is the set of matrices with rank less than or equal to r. With this atomic set, the problem in [Chandrasekaran et al., 2012] reduces to recover $x_0 = a$, where $a \in \mathcal{A}$. Upon successful recovery, the convex atomic norm minimization problem on the right-hand side of (37) solves the corresponding non-convex problem on its left-hand side.

8. Conclusion

We have proposed a family of low-rank inducing norms and regularizers. These norms are interpreted as the largest convex minorizers of a unitarily invariant norm that is restricted to matrices of at most rank r. One feature of these norms is that optimality interpretations in the form of a posteriori guarantees can be provided. In particular, it can be checked if the solutions to a convex relaxation involving low-rank inducing norms, also solve an underlying rank constrained problem. Our numerical examples indicate that this is useful for, e.g. the so-called matrix completion problem. A suitably chosen low-rank inducing norm yields significantly better completion and/or lower rank than the commonly used nuclear norm approach. This has been demonstrated on the basis of what we called low-rank inducing Frobenius and spectral norms. Both norms have been shown to have cheaply computable proximal mappings, as well as simple SDP representations. As a result, this extends proximal mapping computations that are found, in e.g. [Wu et al., 2014; Eriksson et al., 2015; Grussler et al., 2016a]. Moreover, The class of low-rank inducing norms can be further broadened by using continuous r as discussed in [Grussler et al., 2016a] for the low-rank inducing Frobenius norm. Finally, it has been highlighted that our findings also generalize to atomic norms, and to other non-convex problems.

A. Appendix

A.1 Proofs to Results in Section 3

***Proof to Lemma* 3.1**

Proof Let $1 \leq r \leq q := \min\{m, n\}$, $g : \mathbb{R}^q \to \mathbb{R}_{\geq 0}$ be a symmetric gauge function, $\Sigma_j(M) := \operatorname{diag}(\sigma_1(M), \dots, \sigma_j(M))$ for $M \in \mathbb{R}^{n\times m}$, and $1 \leq j \leq q$. Then for all $Y \in \mathbb{R}^{n\times m}$,

$$\begin{aligned}\|Y\|_{g^D,r} &= \max_{\substack{\operatorname{rank}(M)\leq r\\ \|M\|_g\leq 1}} \langle M, Y\rangle = \max_{\substack{\operatorname{rank}(\Sigma_q(M))\leq r\\ \|\Sigma_q(M)\|_g\leq 1}} \langle \Sigma_q(Y), \Sigma_q(M)\rangle \\ &= \max_{\|\Sigma_r(M)\|_g\leq 1} \langle \Sigma_r(Y), \Sigma_r(M)\rangle = \|\Sigma_r(Y)\|_{g^D},\end{aligned}$$

where the second equality follows by [Horn and Johnson, 2012, Corollary 7.4.1.3(c)]. Further, $\|\cdot\|_{g^D,r}$ is unitarily invariant, since

$$\|\Sigma_r(Y)\|_{g^D} = g^D(\sigma_1(Y), \dots, \sigma_r(Y))$$

defines a symmetric gauge function (see Proposition 3.1). Similarly to the

above, this implies that

$$\begin{aligned}\|M\|_{g,r*} &= \max_{\|Y\|_{g^D,r}\leq 1} \langle M, Y\rangle = \max_{g^D(\sigma_1(Y),\ldots,\sigma_r(Y))\leq 1} \sum_{i=1}^{q} \sigma_i(M)\sigma_i(Y) \\ &= \max_{g^D(\sigma_1(Y),\ldots,\sigma_r(Y))\leq 1} \left[\sum_{i=1}^{r} \sigma_i(M)\sigma_i(Y) + \sigma_r(Y) \sum_{i=r+1}^{q} \sigma_i(M)\right].\end{aligned}$$

It remains to prove (11) and (12). The constraint set for $r+1$ is a superset of the constraint set for r and by the definition of $\|\cdot\|_{g^D,r}$ in (9) it follows that $\|Y\|_{g^D,r} \leq \|Y\|_{g^D,r+1}$. Therefore,

$$\|M\|_{g,r*} = \max_{\|Y\|_{g^D,r}\leq 1} \langle M, Y\rangle \geq \max_{\|Y\|_{g^D,r+1}\leq 1} \langle M, Y\rangle = \|M\|_{g,(r+1)*}.$$

Note that $\|\cdot\|_{g^D} = \|\cdot\|_{g^D,q}$, which implies that $\|\cdot\|_{g,q*} = \|\cdot\|_g$ and thus (11) is proven. The implication in (12) follows from the derived expression for $\|\cdot\|_{g,r*}$, since for rank-r matrices M, $\sigma_i(M) = 0$ for all $i \in \{r+1,\ldots,q\}$. □

***Proof to Proposition* 3.2** By [Horn and Johnson, 2012, Corollary 7.4.1.3(c)] it holds that $g^D(\sigma_1) = \sigma_1$ if and only if $g(\sigma_1) = \sigma_1$. Thus, (10) yields for all $M \in \mathbb{R}^{n\times m}$ that

$$\|M\|_{g,1*} = \max_{\sigma_1(Y)\leq 1} \sigma_r(Y) \sum_{i=1}^{\min\{m,n\}} \sigma_i(M) = \|M\|_{\ell_\infty^D} = \|M\|_{\ell_1},$$

where we use the fact that the dual norm of the spectral norm is the nuclear norm (see [Horn and Johnson, 2012, Theorem 5.6.42]).

***Proof to Lemma* 3.2**

Proof By definition of $\|\cdot\|_{g^D,r}$ in (9) in Lemma 3.1, it holds that for all $Y \in \mathbb{R}^{n\times m}$,

$$\max_{X\in\operatorname{conv}(E_{g,r})} \langle X, Y\rangle = \max_{\substack{\operatorname{rank}(X)\leq r \\ \|X\|_{g^D}\leq 1}} \langle X, Y\rangle = \|Y\|_{g^D,r} = \max_{\|X\|_{g,r*}\leq 1} \langle X, Y\rangle = \max_{X\in B^1_{g,r*}} \langle X, Y\rangle.$$

Since $\operatorname{conv}(E_{g,r})$ and $B^1_{g,r*}$ are closed convex sets, this equality can only be fulfilled if the sets are equal (see [Hiriart-Urruty and Lemaréchal, 2013, Theorem V.3.3.1]).

Next, we prove the decomposition. Since the decomposition trivially holds for $M = 0$, we assume that $M \in \mathbb{R}^{n\times m} \setminus \{0\}$ and define $\bar{M} := \|M\|^{-1}_{g,r*} M$. Then $\bar{M} \in B^1_{g,r*} = \operatorname{conv}(E)$ and therefore be decomposed as

$$\bar{M} = \textstyle\sum_i \alpha_i \tilde{M}_i \quad \text{with} \quad \sum_i \alpha_i = 1,\ \alpha_i \geq 0$$

where all $\tilde{M}_i$ satisfy

$$\|\tilde{M}_i\|_g = \|\tilde{M}_i\|_{g,r*} = 1 \quad \text{and} \quad \operatorname{rank}(\tilde{M}_i) \leq r,$$

where the first equality is from (12) in Lemma 3.1. Defining $M_i := \tilde{M}_i \|M\|_{g,r*}$ gives

$$M = \textstyle\sum_i \alpha_i M_i \quad \text{with} \quad \operatorname{rank}(M_i) \leq r$$

and

$$\|M_i\|_g = \|M_i\|_{g,r*} = \|\|M\|_{g,r*} \tilde{M}_i\|_{g,r*} = \|M\|_{g,r*} \|\tilde{M}_i\|_{g,r*} = \|M\|_{g,r*}.$$

This concludes the proof. □

***Proof to Proposition* 3.3**

Proof Let $\bar{M} = \sum_i \alpha_i M_i$ with $M_i \in E_{g,r}$ and $\alpha_i \in (0, 1)$, $\sum_i \alpha_i = 1$ be a convex combination of points in $E_{g,r}$. Then, by assumption,

$$\|\bar{M}\|_g = \|\textstyle\sum_i \alpha_i M_i\|_g < \sum_i \alpha_i \|M_i\|_g = \sum_i \alpha_i = 1$$

and thus $\bar{M} \notin E_{g,r}$. Since $\operatorname{conv}(E_{g,r}) = B^1_{g,r*}$, this implies that $E_{g,r}$ is the set of extreme points of $B^1_{g,r*}$. □

***Proof to Corollary* 3.1**

Proof Let us start by showing that $\operatorname{conv}(\mathcal{E}_r) = B^1_{\ell_\infty,r*}$. Since $\|\cdot\|_{\ell_1,r}$ and $\|\cdot\|_{\ell_\infty,r}$ are dual norms to each other, it follows by Lemma 3.2 that

$$\|Y\|_{\ell_1,r} = \max_{X \in B^1_{\ell_\infty,r*}} \langle X, Y \rangle = \max_{\substack{\operatorname{rank}(X)=r \\ 1=\sigma_1(X)=\ldots=\sigma_r(X)}} \sum_{i=1}^{r} \sigma_i(X)\sigma_i(Y) = \max_{X \in \operatorname{conv}(\mathcal{E}_r)} \langle X, Y \rangle,$$

where the last two equalities are a result of [Horn and Johnson, 2012, Corollary 7.4.1.3(c)]. However, $\operatorname{conv}(\mathcal{E}_r)$ and $B^1_{\ell_\infty,r*}$ are closed convex sets and therefore this equation can only hold if the sets are identical (see [Hiriart-Urruty and Lemaréchal, 2013, Proposition V.3.3.1]).

It remains to show that no point in $\mathcal{E}_r$ can be constructed as a convex combination of other points in $\mathcal{E}_r$. To this end, note that a necessary condition for $M \in \mathcal{E}_r$ is that

$$\|M\|_F^2 = \sum_{i=1}^{\min\{m,n\}} \sigma_i^2(M) = \sum_{i=1}^{r} \sigma_i^2(M) = r.$$

Let $\bar{M} = \sum_i \alpha_i M_i$ be an arbitrary convex combination with $\alpha_i > 0$ and $\sum_i \alpha_i = 1$, of distinct points $M_i \in \mathcal{E}_r$. By the strict convexity of $\|\cdot\|_F^2$, it holds that

$$\|\bar{M}\|_F^2 = \|\textstyle\sum_i \alpha_i M_i\|_F^2 < \sum_i \alpha_i \|M_i\|_F^2 = r \sum_i \alpha_i = r.$$

Hence, $\bar{M} \notin \mathcal{E}_r$ and this concludes the proof. □

A.2 Derivations to $\Pi_{\text{epi}(\|\cdot\|_{g,r*})}$

Utilizing the Moreau decomposition in (23), we determine the projection onto $\text{epi}(\|\cdot\|_{g,r*})$, by computing a projecting onto the polar cone $(\text{epi}(\|\cdot\|_{g,r*}))^\circ$. The latter is by definition (see [Bauschke and Combettes, 2011, Definition 6.21]) the negative of the dual cone to $\text{epi}(\|\cdot\|_{g,r*})$, i.e.

$$\begin{aligned}(\text{epi}(\|\cdot\|_{g,r*}))^\circ &= -\text{epi}(\|\cdot\|_{g^D,r}) \\ &= \{(-Y,-w) : \|Y\|_{g^D,r} \leq w\} = \{(Y,w) : \|Y\|_{g^D,r} \leq -w\}.\end{aligned}$$

Thus, the projection onto the polar cone becomes

$$\Pi_{(\text{epi}(\|\cdot\|_{g,r*}))^\circ}(Z, z_v) = \underset{\substack{w\in\mathbb{R},\ Y\in\mathbb{R}^{n\times m} \\ w+\|Y\|_{g^D,r}\leq 0}}{\text{argmin}} \frac{1}{2}\left[(w-z_v)^2 + \|Y-Z\|_F^2\right]$$

and we need to solve

$$\begin{aligned}\underset{Y,w}{\text{minimize}} \quad & \frac{1}{2}\left[(w-z_v)^2 + \|Y-Z\|_F^2\right] \\ \text{subject to} \quad & -w \geq \|Y\|_{g^D,r},\ Y \in \mathbb{R}^{n\times m}.\end{aligned} \tag{39}$$

Since the cost and the constraint in (39) are unitarily invariant, it can be shown (see [Watson, 1992; Lewis, 1995]) that $Y^\star$ and Z have a simultaneous SVD, i.e. if $Z = \sum_{i=1}^q \sigma_i(Z) u_i v_i^T$ is an SVD of Z then $Y^\star = \sum_{i=1}^q \sigma_i(Y^\star) u_i v_i^T$ where we define $q := \min\{m, n\}$. Consequently, it is equivalent to consider the vector-valued problem

$$\begin{aligned}\underset{y,w}{\text{minimize}} \quad & \frac{1}{2}\left[(w-z_v)^2 + \sum_{i=1}^q (y_i - z_i)^2\right] \\ \text{subject to} \quad & -w \geq \|\text{diag}(y)\|_{g^D,r},\ y \in \mathbb{R}^q, \\ & y_1 \geq \cdots \geq y_q,\end{aligned} \tag{40}$$

with $z_1 \geq \cdots \geq z_q \geq 0$, $z_i = \sigma_i(Z)$ and $y_i = \sigma_i(Y)$ for $1 \leq i \leq q$.

Remark 2

The unique solution $(y^\star, w^\star)$ fulfills $0 \leq y_i^\star \leq z_i$ for $1 \leq i \leq q$. The upper bound holds, because otherwise $\bar{y}^\star$ with $\bar{y}_i^\star = \min\{z_i, y_i^\star\}$ is a feasible solution with smaller cost. Similarly, the lower bound holds, because otherwise $\bar{y}^\star$ with $\bar{y}_i^\star = \max\{0, y_i^\star\}$ is a feasible solution with smaller cost. Thus, it is not necessary to explicitly restrict y to be nonnegative. □

To solve (40), note that there exists a $t^\star \in \{1, \ldots, r\}$ such that

$$y_{r-t^\star}^\star > y_{r-t^\star+1}^\star = \cdots = y_r^\star, \tag{41}$$

where $t^\star = r$ if $y_1^\star = y_r^\star$. This assumption implies that $y_{r-t^\star} \geq y_{r-t^\star+1}$ is assumed to be inactive and therefore can be removed from (40). Then also the constraints $y_1 \geq \cdots \geq y_{r-t^\star}$ can be removed, because the cost function and the sorting of z ensures that the solution will always fulfill them. This yields the following problem

$$\begin{aligned} \underset{y,w}{\text{minimize}} \quad & \frac{1}{2}\left[(w - z_v)^2 + \sum_{i=1}^{q}(y_i - z_i)^2\right] \\ \text{subject to} \quad & -w \geq \|\text{diag}(y)\|_{g^D,r}, \ \ y \in \mathbb{R}^q, \\ & y_{r-t+1} = \cdots = y_r \geq \cdots \geq y_q. \end{aligned} \tag{42}$$

Thus, solving (40) reduces to finding $t^\star$ such that (42) solves (40). As it is shown later, solving (42) can be done efficiently for the low-rank inducing norms that are considered in this paper. The following lemma shows that $t^\star$ can be found by a binary search over t, where the decision to increase or decrease t is based on the solution of (42).

Lemma A.1
Let $\left(y^{(t)}, w^{(t)}\right)$ denote the solution to (42) depending on t such that $1 \leq t \leq r$. Further let $\left(y^{(t^\star)}, w^{(t^\star)}\right)$ be the solution to (40) such that $y_{r-t^\star}^{(t^\star)} > y_{r-t^\star+1}^{(t^\star)}$ and $y_{r-t^\star}^{(t^\star)} = y_{r-t^\star+1}^{(t^\star)}$ if $t^\star = r$. Then,

i. $t^\star = \min\{\{t : y_{r-t}^{(t)} > y_{r-t+1}^{(t)}\} \cup \{r\}\}$.

ii. If $y_{r-t'}^{(t')} \geq y_{r-t'+1}^{(t')}$ then $y_{r-t}^{(t)} \geq y_{r-t+1}^{(t)}$ for all $t \geq t'$.

iii. If $y_{r-t'}^{(t')} < y_{r-t'+1}^{(t')}$ then $y_{r-t}^{(t)} < y_{r-t+1}^{(t)}$ for all $t \leq t'$.

In particular,

I. $y_{r-t}^{(t)} \geq y_{r-t+1}^{(t)}$ for all $t \geq t^\star$.

II. $y_{r-t}^{(t)} \leq y_{r-t+1}^{(t)}$ for all $t < t^\star$

III. If $t < t^\star$ and $y_{r-t}^{(t)} \leq y_{r-t+1}^{(t)}$ then $\left(y^{(t)}, w^{(t)}\right) = \left(y^{(t^\star)}, w^{(t^\star)}\right)$. □

Proof Throughout this proof, we let $p(t)$ denote the optimal cost of (42) as a function of t. Since adding constraints cannot reduce the optimal cost, p is a nondecreasing function.

Item i: By the same reasoning that led to (42), it holds that

$$y_1^{(t)} \geq \cdots \geq y_{r-t}^{(t)} \text{ for } 1 \leq t \leq r. \tag{43}$$

Using (43), the set $\min\{\{t : y^{(t)}_{r-t} > y^{(t)}_{r-t+1}\} \cup \{r\}\}$ contains all t for which the solution of (42) is feasible for (40). Since p is nondecreasing and $\left(y^{(t^\star)}, w^{(t^\star)}\right)$ is unique, the first claim follows.

Item ii: The second claim is proven by contradiction. Let $(y^{(t')}, w^{(t')})$ be such that $y^{(t')}_{r-t'} \geq y^{(t')}_{r-t'+1}$. Further assume that $y^{(t'+1)}_{r-t'-1} < y^{(t'+1)}_{r-t'}$. In the following, we construct another solution $(\tilde{y}, \tilde{w}) \in \mathbb{R}^{q+1}$ to (42) with $t = t' + 1$, which has a cost that is no larger than $p(t'+1)$. However, (42) has a unique solution due to strong convexity of the cost function. This yields the desired contradiction.

The contradicting solution is constructed as a convex combination $\tilde{w} = (1-\alpha)w^{(t'+1)} + \alpha w^{(t')}$ with $\alpha \in (0, 1]$ and a partially sorted convex combination of $y^{(t')}$ and $y^{(t'+1)}$ with the same α. Let $\hat{y} := (1-\alpha)y^{(t'+1)} + \alpha y^{(t')}$ and let

$$\tilde{y} := (\mathrm{sort}(\hat{y}_1, \ldots, \hat{y}_{r-t'-2}, \hat{y}_{r-t'}), \hat{y}_{r-t'-1}, \hat{y}_{r-t'+1}, \ldots, \hat{y}_q),$$

be the partially sorted convex combination, where $\mathrm{sort}(\cdot)$ denotes sorting in descending order.

To select α, we note that by assumption,

$$y^{(t')}_{r-t'-1} \geq y^{(t')}_{r-t'} \geq y^{(t')}_{r-t'+1} \quad \text{and} \quad y^{(t'+1)}_{r-t'-1} < y^{(t'+1)}_{r-t'} = y^{(t'+1)}_{r-t'+1}.$$

Therefore, there exists an $\alpha \in (0, 1]$ such that

$$\begin{aligned}\tilde{y}_{r-t'} = \hat{y}_{r-t'-1} &= (1-\alpha)y^{(t'+1)}_{r-t'-1} + \alpha y^{(t')}_{r-t'-1} \\ &= (1-\alpha)y^{(t'+1)}_{r-t'+1} + \alpha y^{(t')}_{r-t'+1} = \hat{y}_{r-t'+1} = \tilde{y}_{r-t'+1}.\end{aligned}$$

Since

$$y^{(t')}_{r-t'+1} = \cdots = y^{(t')}_r \quad \text{and} \quad y^{(t'+1)}_{r-t'-1} = \cdots = y^{(t'+1)}_r,$$

it follows that

$$\tilde{y}_{r-t'} = \cdots = \tilde{y}_r.$$

Furthermore, the construction of $\tilde{y}$ as well as the sorting give that

$$\tilde{y}_r \geq \cdots \geq \tilde{y}_q \quad \text{and} \quad \tilde{y}_1 \geq \cdots \geq \tilde{y}_{r-t'-1}.$$

Hence, $\tilde{y}$ satisfies the chain of inequalities in (42) for $t = t' + 1$.

It remains to show that $\tilde{y}$ satisfies the epigraph constraint and that the cost is not higher than $p(t'+1)$. These properties are already fulfilled for $\hat{y}$ being a convex combination of two feasible points with costs $p(t')$ and $p(t'+1)$, respectively, where $p(t') \leq p(t'+1)$. Therefore, it is left to show that the sorting involved in $\tilde{y}$ maintains these properties. First, we show

that sorting of any sub-vector in y does not increase the cost. Suppose that $z_i \geq z_j$, $y_i \leq y_j$, i.e., y is not sorted the same way as z. Then

$$\begin{aligned}\tfrac{1}{2}\left((z_i - y_i)^2 + (z_j - y_j)^2\right) &= (z_i - z_j)(y_j - y_i) + \tfrac{1}{2}\left((z_i - y_j)^2 + (z_j - y_i)^2\right) \\ &\geq \left((z_i - y_j)^2 + (z_j - y_i)^2\right),\end{aligned}$$

and thus the cost is not increased by sorting y or any sub-vector of it. Further, note that the permutation caused by the sorting of the first r elements of y does not influence the epigraph constraint, because $\|\mathrm{diag}(y)\|_{g^D,r}$ is permutation invariant by definition.

Next notice that $\tilde{y}$ is obtained from $\hat{y}$ by first swapping $\hat{y}_{r-t'-1}$ and $\hat{y}_{r-t'}$. From the choice of α, we conclude that

$$\hat{y}_{r-t'} = (1-\alpha)y_{r-t'}^{(t'+1)} + \alpha y_{r-t'}^{(t')} \geq (1-\alpha)y_{r-t'+1}^{(t'+1)} + \alpha y_{r-t'+1}^{(t')} = \hat{y}_{r-t'+1} = \hat{y}_{r-t'-1}.$$

Thus, this swap is a sorting which does neither increase the cost, nor does it violate the epigraph constraint. Analogously, sorting the first $r-t'$ elements of the resulting vector to obtain $\tilde{y}$ has the same effect and therefore we receive the desired contradiction.

Item iii: Suppose that there exist t and t' with $t' > t$ such that $y_{r-t'}^{(t')} < y_{r-t'+1}^{(t')}$ and $y_{r-t}^{(t)} \geq y_{r-t+1}^{(t)}$. Then Item ii shows that $y_{r-t'}^{(t')} \geq y_{r-t'+1}^{(t')}$, which is a contradiction.

Items I to III: The statements follow immediately from Items i to iii. □

In order to solve (42), one can proceed similarly to solving (40). There always exists $s^\star \geq 0$ such that the solution $(y^{(t)}, w^{(t)})$ of (42) satisfies

$$y_{r-t+1}^{(t)} = \cdots = y_{r+s^\star}^{(t)} > y_{r+s^\star+1}^{(t)},$$

where $s^\star = q - r$ if $y_r^{(t)} = y_q^{(t)}$. As before, this allows us to remove the inactive constraint $y_{r+s} \geq y_{r+s+1}$. Then the constraints $y_{r+s+1} \geq \cdots \geq y_q$ become redundant, because any solution fulfills $y_j = z_j$, $j \geq r+s+1$. Finally, we are left with the following reduced optimization problem

$$\begin{aligned}\underset{y,w}{\text{minimize}} \quad & \frac{1}{2}\left[(w - z_v)^2 + \sum_{i=1}^{r+s}(y_i - z_i)^2\right] \\ \text{subject to} \quad & -w \geq \|\mathrm{diag}(y)\|_{g^D,r}, \ y \in \mathbb{R}^q, \\ & y_{r-t+1} = \cdots = y_{r+s}.\end{aligned} \tag{44}$$

For given t, one can perform a binary search on s in (44) in order find $s^\star$. This can be done with the help of the following lemma.

Algorithm A.1 Determine $(Y^\star, w^\star) = \Pi_{(\mathrm{epi}(\|\cdot\|_{g,r^*}))^\circ}(Z, z_v)$, i.e., solve (39)

Input: Let $Z \in \mathbb{R}^{n\times m}$, $z_v \in \mathbb{R}$ and $r \in \mathbb{N}$ such that $1 \le r \le q := \min\{m, n\}$ be given.
Let $Z = \sum_{i=1}^{q} \sigma_i(Z) u_i v_i^T$ be an SVD of Z.
// *Solve* (39) *via the vector problem* (40) *with data* $z = (\sigma_1(Z), \dots, \sigma_q(Z))$ *and* z_v
Set $t_{\min} = 1$, $t_{\max} = r$, and $t = \lceil \frac{t_{\min}+t_{\max}}{2} \rceil$
// *Solve* (40) *via* (42) *and binary search over* t
while $t_{\min} \neq t_{\max}$ **do**
  Set $s_{\min} = 0$, $s_{\max} = q - r$, and $s = \lceil \frac{s_{\min}+s_{\max}}{2} \rceil$
  // *Solve* (42) *via* (44) *and binary search over* s
  while $s_{\min} \neq s_{\max}$ **do**
    Solve (44)
    Update $s_{\min}$, $s_{\max}$, and s using the binary search rules in Lemma A.2
  end while
  Update $t_{\min}$, $t_{\max}$, and t using the binary search rules in Lemma A.1
end while
Output: $(Y^\star, w^\star) = (\sum_{i=1}^{q} y_i u_i v_i^T, w)$ with (y, w) being the last solution to (44).

Lemma A.2
For fixed t with $1 \le t \le r$, let $\left(y^{(t,s)}, w^{(t,s)}\right)$ denote the solution to (44) for different s satisfying $0 \le s \le r-q$. Further let $\left(y^{(t,s^\star)}, w^{(t,s^\star)}\right)$ be the solution to (42) such that $y^{(t,s^\star)}_{r+s^\star} > y^{(t,s^\star)}_{r+s^\star+1}$ and $y^{(t,s^\star)}_{r+s^\star} = y^{(t)}_{r+s^\star+1}$ if $s^\star = q - r$. Then,

i. $s^\star = \min\{\{s : y^{(t,s^\star)}_{r+s^\star} > y^{(t,s^\star)}_{r+s^\star}\} \cup \{q-r\}\}$.

ii. If $y^{(t,s')}_{r+s'} \ge y^{(t,s')}_{r+s'+1}$ then $y^{(t,s)}_{r+s} \ge y^{(t,s)}_{r+s+1}$ for all $s \ge s'$.

iii. If $y^{(t,s')}_{r+s'} < y^{(t,s')}_{r+s'+1}$ then $y^{(t,s)}_{r+s} < y^{(t,s)}_{r+s+1}$ for all $s \le s'$.

In particular,

I. $y^{(t,s)}_{r+s} \ge y^{(t,s)}_{r+s+1}$ for all $s \ge s^\star$.

II. $y^{(t,s)}_{r+s} \le y^{(t,s)}_{r+s+1}$ for all $s < s^\star$.

III. If $s < s^\star$ and $y^{(t,s)}_{r+s} \ge y^{(t,s)}_{r+s+1}$ then $\left(y^{(t,s)}, w^{(t)}\right) = \left(y^{(t,s^\star)}, w^{(t,s^\star)}\right)$. □

Proof The proof goes analogously to the proof of Lemma A.1 and is therefore omitted. □

The nested binary search algorithm to solve (39) via (40) is summarized in Algorithm A.1. The problem that decides how to update the parameters in the nested binary search is (44). In order to solve (44) explicitly, we introduce new variables $\tilde{y}, \tilde{z} \in \mathbb{R}^{r-t+1}$ as

$$\tilde{y}_i = \begin{cases} y_i, & \text{if } 1 \le i \le r-t \\ \sqrt{t+s}\, y_r, & \text{if } i = r-t+1 \end{cases} \qquad \tilde{z}_i = \begin{cases} z_i, & \text{if } 1 \le i \le r-t \\ \frac{1}{\sqrt{t+s}} \sum_{i=r-t+1}^{r+s} z_i, & \text{if } i = r-t+1 \end{cases} \tag{45}$$

This gives

$$\sum_{i=r-t+1}^{r+s} (y_r - z_i)^2 = (\tilde{y}_{r-t+1} - \tilde{z}_{r-t+1})^2 + \sum_{i=r-t+1}^{r+s} z_i^2 - \left(\frac{1}{\sqrt{t+s}} \sum_{i=r-t+1}^{r+s} z_i \right)^2.$$

Since we can ignore the constant terms, we are left with the following projection problem of reduced dimension

$$\begin{aligned} \underset{\tilde{y},w}{\text{minimize}} \quad & \frac{1}{2}\left[(w - z_v)^2 + \sum_{i=1}^{r-t+1} (\tilde{y}_i - \tilde{z}_i)^2 \right] \\ \text{subject to} \quad & -w \ge \|\text{diag}(\tilde{y}_1, \ldots, \tilde{y}_{r-t}, \underbrace{\tfrac{\tilde{y}_{r-t+1}}{\sqrt{s+t}}, \ldots, \tfrac{\tilde{y}_{r-t+1}}{\sqrt{s+t}}}_{t \text{ times}})\|_{g^D,r}, \ \tilde{y} \in \mathbb{R}^{r-t+1}. \end{aligned}$$

Below, it is shown how to explicitly solve this projection problem for $g^D = \ell_2$ and $g^D = \ell_1$ in order to arrive at the epigraph projections of the low-rank inducing Frobenius and spectral norms.

The case $\|\cdot\|_{g^D,r} = \|\cdot\|_r$ In this case, $g^D = \ell_2$ and the projection problem becomes

$$\begin{aligned} \underset{\tilde{y},w}{\text{minimize}} \quad & \frac{1}{2}\left[(w - z_v)^2 + \sum_{i=1}^{r-t+1} (\tilde{y}_i - \tilde{z}_i)^2 \right] \\ \text{subject to} \quad & -w \ge \sqrt{\sum_{i=1}^{r-t} \tilde{y}_i^2 + \frac{t}{s+t} \tilde{y}_{r-t+1}^2}, \ y \in \mathbb{R}^{r-t+1}. \end{aligned}$$

Consequently, the solution $(\tilde{y}^\star, w^\star)$ is the orthogonal projection of $(\tilde{z}, z_v)$ onto the second-order cone

$$K := \left\{ (\tilde{y}, w) \in \mathbb{R}^{r-t+2} : \sqrt{\sum_{i=1}^{r-t} \tilde{y}_i^2 + \frac{t}{s+t} \tilde{y}_{r-t+1}^2} \le -w \right\}. \tag{46}$$

The associated polar cone $K^\circ := \{y : \langle y, x\rangle \leq 0 \text{ for all } x \in K\}$ is then given by (see e.g. [Grussler and Rantzer, 2014])

$$K^\circ := \left\{ (y, p) \in \mathbb{R}^{r-t+2} : \sqrt{\sum_{i=1}^{r-t} \tilde{y}_i^2 + \frac{s+t}{t}\tilde{y}_{r-t+1}^2} \leq p \right\}.$$

This allows us to summarize the following two simple cases:

i. $(\tilde{y}^\star, w^\star) = (\tilde{z}, z_v)$ if and only if $(\tilde{z}, z_v) \in K$, i.e.

$$\sqrt{\sum_{i=1}^{r-t} \tilde{z}_i^2 + \frac{t}{s+t}\tilde{z}_{r-t+1}^2} \leq -z_v,$$

ii. $(\tilde{y}^\star, w^\star) = (0, 0)$ if and only if $(\tilde{z}, z_v) \in K^\circ$, i.e.

$$\sqrt{\sum_{i=1}^{r-t} \tilde{z}_i^2 + \frac{s+t}{t}\tilde{z}_{r-t+1}^2} \leq z_v,$$

where the last statement follows by [Hiriart-Urruty and Lemaréchal, 2013, Proposition III.3.2.3].

Next, it is shown how to compute the projection if $(\tilde{z}, z_v)$ does not belong to either of these cones. By [Bauschke and Combettes, 2011, Proposition 6.46] it holds that $(\tilde{z} - \tilde{y}^\star, z_v - w^\star)$ is an element of the normal cone to the cone K at $(\tilde{y}^\star, w^\star)$. Using the normal cone description in [Hiriart-Urruty and Lemaréchal, 2013, Theorem VI.1.3.5], this implies that

$$(\tilde{z} - \tilde{y}^\star, z_v - w^\star) = \mu \left. \nabla_{(\tilde{y},w)} \sqrt{\sum_{i=1}^{r-t} \tilde{y}_i^2 + \frac{t}{s+t}\tilde{y}_{r-t+1}^2} + w \right|_{(\tilde{y},w)=(\tilde{y}^\star,w^\star)} \tag{47}$$

for some $\mu \geq 0$. Since $(\tilde{z}, z_v) \notin K$ we conclude that the optimal point is on the boundary of the cone K, i.e.

$$-w^\star = \sqrt{\sum_{i=1}^{r-t} \tilde{y}_i^{\star 2} + \frac{t}{s+t}\tilde{y}_{r-t+1}^{\star 2}}. \tag{48}$$

Solving the equations in (47) and using (48) give

$$\begin{aligned}
\tilde{y}_i^\star &= \frac{\tilde{z}_i}{1 - \frac{\mu}{w^\star}}, \quad 1 \leq i \leq r-t, \\
\tilde{y}_{r-t+1}^\star &= \frac{\tilde{z}_{r-t+1}}{1 - \frac{\mu t}{w^\star(s+t)}}, \\
w^\star &= z_v - \mu.
\end{aligned}$$

To characterize the solution, it is left to compute μ. By plugging the solution into (48), diving by $w^\star$ and taking the square, we arrive at

$$1 = \frac{\sum_{i=1}^{r-t} \tilde{z}_i^2}{(2\mu - z_v)^2} + \frac{t}{s+t} \frac{\tilde{z}_{r-t+1}^2}{\left(\mu - z_v + \frac{\mu t}{s+t}\right)^2}.$$

Defining $c_1 := \sum_{i=1}^{r-t} \tilde{z}_i^2 = \sum_{i=1}^{r-t} z_i^2$ and $c_2 := \sqrt{t+s}\tilde{z}_{r-t+1} = \sum_{i=r-t+1}^{r+s} z_i$ this can rewritten as the fourth order polynomial equation

$$[(2\mu - z_v)^2 - c_1][(t+s)(\mu - z_v) + \mu t]^2 - t c_2^2 (2\mu - z_v)^2 = 0, \tag{49}$$

which can be solved explicitly for $\mu \geq 0$. Resubstitution in (45) gives that the solution $\left(y^{(t,s)}, w^{(t,s)}\right)$ to (44) can be expressed as

i. $1 \leq j \leq r-t : y_j^{(t,s)} = \dfrac{z_i(\mu - z_v)}{2\mu - z_v}$,

ii. $r-t+1 \leq j \leq r+s : y_j^{(t,s)} = \dfrac{(\mu - z_v)\sum_{i=r-t+1}^{r+s} z_i}{(s+t)(\mu - z_v) + \mu t}$,

iii. $r+s+1 \leq j \leq q : y_j^{(t,s)} = z_j$,

iv. $w^{(t,s)} = z_v - \mu$,

if $(z, z_v) \notin K \cup K^\circ$.

The case $\|\cdot\|_{g^D,r} = \|\cdot\|_{\ell_1,r}$ The second case is analog to the first case. We would like to solve

$$\begin{aligned} \underset{\tilde{y},w}{\text{minimize}} \quad & \frac{1}{2}\left[(w - z_v)^2 + \sum_{i=1}^{r-t+1} (\tilde{y}_i - \tilde{z}_i)^2\right] \\ \text{subject to} \quad & 0 \geq \sum_{i=1}^{r-t} |\tilde{y}_i| + \frac{t}{\sqrt{t+s}}|\tilde{y}_{r-t+1}| + w, \ y \in \mathbb{R}^{r-t+1}. \end{aligned} \tag{50}$$

Consequently, the solution $(\tilde{y}^\star, w^\star)$ is the orthogonal projection of $(\tilde{z}, z_v)$ onto

$$K := \left\{(\tilde{y}, w) \in \mathbb{R}^{r-t+2} : \sum_{i=1}^{r-t} |\tilde{y}_i| + \frac{t}{\sqrt{t+s}}|\tilde{y}_{r-t+1}| \leq -w\right\}. \tag{51}$$

The polar cone $K^\circ := \{y : \langle y, x\rangle \leq 0 \text{ for all } x \in K\}$ is then given by

$$K^\circ := \left\{(y, p) \in \mathbb{R}^{r-t+2} : \max\left(|y_1|, \ldots, |y_{r-t-2}|, \frac{\sqrt{t+s}}{t}|y_{r-t+1}|\right) \leq p\right\}.$$

Similarly to before, we get the following two simple cases:

i. $(\tilde{y}^\star, w^\star) = (\tilde{z}, z_v)$ if and only if $(\tilde{z}, z_v) \in K$, i.e.

$$\sum_{i=1}^{r-t} \tilde{z}_i + \frac{t}{\sqrt{t+s}} \tilde{z}_{r-t+1} \leq -z_v,$$

ii. $(\tilde{y}^\star, w^\star) = (0, 0)$ if and only if $(\tilde{z}, z_v) \in K^\circ$, i.e.

$$\max\left(\tilde{z}_1, \frac{\sqrt{t+s}}{t} \tilde{z}_{r-t+1}\right) \leq z_v,$$

where it is used that the $\tilde{z}_i$ are nonnegative and decreasingly sorted.

It remains to show how to compute the projection if $(\tilde{z}, z_v)$ does not belong to either of these cones. By [Bauschke and Combettes, 2011, Proposition 6.46] it holds that $(\tilde{z} - \tilde{y}^\star, z_v - w^\star)$ is an element of the normal cone to the cone K at $(\tilde{y}^\star, w^\star)$. Using the normal cone description in [Hiriart-Urruty and Lemaréchal, 2013, Theorem VI.1.3.5], we get

$$(\tilde{z} - \tilde{y}^\star, z_v - w^\star) \in \mu \, \partial_{(\tilde{y},w)} \left(\sum_{i=1}^{r-t} |\tilde{y}_i| + \frac{t}{\sqrt{s+t}} |\tilde{y}_{r-t+1}| + w \right) \Bigg|_{(\tilde{y},w)=(\tilde{y}^\star,w^\star)} \tag{52}$$

for some $\mu \geq 0$. First note that any solution to (50) satisfies $\tilde{y}^\star \geq 0$. The optimality conditions for $y_i^\star = 0$ and $y_i^\star > 0$ become

$$\tilde{y}_i^\star = 0 \Leftrightarrow \tilde{z}_i \in [0, \mu], \qquad \tilde{y}_i^\star > 0 \Leftrightarrow \tilde{y}_i^\star = \tilde{z}_i - \mu$$

for all $i \in \{1, \ldots, r-t\}$. These equivalences also hold for $\tilde{y}_{r-t+1}$ with μ multiplied by $t/\sqrt{s+t}$. Therefore,

$$\begin{aligned} \tilde{y}_i^\star &= \max(\tilde{z}_i - \mu, 0), \ 1 \leq i \leq r-t, \\ \tilde{y}_{r-t+1}^\star &= \max\left(\tilde{z}_{r-t+1} - \frac{t\mu}{\sqrt{t+s}}, 0\right), \\ w^\star &= z_v - \mu. \end{aligned}$$

In order to determine μ, notice that $(\tilde{y}^\star, w^\star)$ lies on the boundary of the cone K in (51), which implies

$$\begin{aligned} 0 &= \sum_{i=1}^{r-t} |\tilde{y}_i^\star| + \frac{t}{\sqrt{t+s}} |\tilde{y}_{r-t+1}^\star| + w^\star \\ &= \sum_{i=1}^{r-t} \max\left(\tilde{z}_i - \mu, 0\right) + \frac{t}{\sqrt{t+s}} \max(\tilde{z}_{r-t+1} - \tfrac{t\mu}{\sqrt{t+s}}, 0\Big) + z_v - \mu. \end{aligned}$$

We denote the solution to this equation by $\mu^\star$, and solve it using a so-called *break point searching algorithm*, as it has been done for similar problems in [Held et al., 1974; Duchi et al., 2008; Condat, 2016]. To this end, let

$$\hat{z} = \left(\tilde{z}_1, \dots, \tilde{z}_j, \frac{t}{\sqrt{t+s}} \tilde{z}_{r-t+1}, \tilde{z}_{j+1}, \dots, \tilde{z}_{r-t} \right),$$

be the vector that sorts $\tilde{z}$ according to the break points of the max expressions, i.e., the index j satisfies $\tilde{z}_j > \frac{\sqrt{t+s}}{t} \tilde{z}_{r-t+1} \geq \tilde{z}_{j+1}$. Defining

$$\alpha = \left(1, \dots, 1, \frac{t^2}{t+s}, 1, \dots, 1 \right)$$

gives that $\mu^\star$ can be found by solving

$$\sum_{i=1}^{r-t+1} \max(\hat{z}_i - \alpha_i \mu, 0) + z_v - \mu = 0. \tag{53}$$

Assuming that we know an index $k = k^\star$ such that

$$\hat{z}_{k^\star+1} - \alpha_{k^\star+1} \mu^\star \leq 0 \qquad \text{and} \qquad \hat{z}_{k^\star} - \alpha_{k^\star} \mu^\star \geq 0, \tag{54}$$

then $\mu^\star$ can be determined from (53) as

$$\mu^\star = \frac{\sum_{i=1}^{k^\star} \hat{z}_i + z_v}{1 + \sum_{i=1}^{k^\star} \alpha_i}. \tag{55}$$

Thus, computing $\mu^\star$ reduces to searching for $k^\star \in \{1, \dots, r-t\}$ for which (55) satisfies (54). This can be done using a binary search, with rules from the following proposition.

Lemma A.3
Let $\mu^\star$ be the solution to (53), let μ_k be the solution to

$$\left(\sum_{i=1}^{r-t+1} \hat{z}_i - \alpha_i \mu \right) + z_v - \mu = 0, \qquad \text{i.e.,} \qquad \hat{\mu}_k = \frac{\sum_{i=1}^{k} \hat{z}_i + z_v}{1 + \sum_{i=1}^{k} \alpha_i}, \tag{56}$$

and let $k^\star$ be such that $\hat{\mu}_{k^\star} = \mu^\star$. Then,

i. $k^\star = \max(\{k : \hat{z}_k - \alpha_k \hat{\mu}_k \geq 0\})$.

ii. If $\hat{z}_k - \alpha_k \hat{\mu}_k \geq 0$, then $\hat{z}_i - \alpha_i \hat{\mu}_i \geq 0$ for all $i \in \{1, \dots, k\}$.

iii. If $\hat{z}_k - \alpha_k \hat{\mu}_k < 0$, then $\hat{z}_i - \alpha_i \hat{\mu}_i < 0$ for all $i \in \{k, \dots, r-t\}$.

In particular,

I. $\hat{z}_k - \alpha_k \hat{\mu}_k \geq 0$ for all $k \in \{1, \ldots, k^\star\}$.

II. $\hat{z}_k - \alpha_k \hat{\mu}_k < 0$, for all $k \in \{k^\star + 1, \ldots, r - t\}$. □

Proof We first show some results needed to prove Items i and ii. Let

$$g_k(\mu) := \sum_{i=1}^{k} \max(\hat{z}_i - \alpha_i \mu, 0) + z_v - \mu,$$

which is strictly decreasing in μ. Let μ_k be the unique solution to the equation

$$g_k(\mu) = 0.$$

For all $\mu \in \mathbb{R}$, we have

$$g_{k-1}(\mu) = g_k(\mu) - \max(\hat{z}_k - \alpha_k \mu, 0) \leq g_k(\mu).$$

Since all g_i are strictly decreasing in μ, we conclude the following facts:

a. $\mu_{k-1} \leq \mu_k$.

b. If $\hat{z}_k - \alpha_k \mu_k \leq 0$, then $g_{k-1}(\mu_k) = g_k(\mu_k) = 0$, hence $\mu_{k-1} = \mu_k$.

Because $\hat{z}$ is sorted according to break points, we conclude that if l and μ are such that $\hat{z}_l - \alpha_l \mu \geq 0$, then also $\hat{z}_i - \alpha_i \mu \geq 0$ for all $i \in \{1, \ldots, l\}$. Therefore, if μ is such that $\hat{z}_k - \alpha_k \mu \geq 0$, we get

$$\sum_{i=1}^{k} \max(\hat{z}_k - \alpha_k \mu, 0) + z_v - \mu = \left(\sum_{i=1}^{k} \hat{z}_k - \alpha_k \mu \right) + z_v - \mu.$$

Hence,

c. If $\hat{z}_k - \alpha_k \mu_k \geq 0$ or $\hat{z}_k - \alpha_k \hat{\mu}_k \geq 0$, then $\hat{\mu}_k = \mu_k$.

Item i: Using Items b and c, we conclude that

$$\hat{\mu}_{k^\star} = \mu_{k^\star} = \mu_{k^\star + 1} = \mu_{r-t+1} = \mu^\star.$$

Item ii: Now, assume that $\hat{z}_k - \alpha_k \hat{\mu}_k \geq 0$. Then, by break point sorting, it holds that $\hat{z}_{k-1} - \alpha_{k-1} \hat{\mu}_k \geq 0$. Using Items a and c, we conclude that

$$0 \leq \hat{z}_{k-1} - \alpha_{k-1} \hat{\mu}_k = \hat{z}_{k-1} - \alpha_{k-1} \mu_k \leq \hat{z}_{k-1} - \alpha_{k-1} \mu_{k-1} = \hat{z}_{k-1} - \alpha_{k-1} \hat{\mu}_{k-1}.$$

Using induction proves the result.
Item iii: Assume, on the contrary, that k is such that $\hat{z}_k - \alpha_k \hat{\mu}_k < 0$ but that there exists $i \in \{k, \ldots, r-t\}$ such that $\hat{z}_i - \alpha_i \hat{\mu}_i \geq 0$. Then, by Item ii, $\hat{z}_k - \alpha_k \hat{\mu}_k \geq 0$ and we have reached the desired contradition.
Items I and II: Follow immediately from Items i to iii. □

Now, that we know how to compute the dual variable $\mu = \mu^\star$, we go back to the original variables in (45), to conclude that the solution $\left(y^{(t,s)}, w^{(t,s)}\right)$ to (44) can be expressed as

i. $1 \leq j \leq r-t : y_j^{(t,s)} = \max(z_j - \mu, 0)$,

ii. $r-t+1 \leq j \leq r+s : y_j^{(t,s)} = \dfrac{1}{\sqrt{t+s}} \max(\sum_{i=r-t+1}^{r+s} z_i - t\mu, 0)$,

iii. $r+s+1 \leq j \leq q : y_j^{(t,s)} = z_j$,

iv. $w^{(t,s)} = z_v - \mu$.

if $(z, z_v) \notin K \cup K^\circ$.

References

Antoulas, A. (2005). *Approximation of Large-Scale Dynamical Systems*. SIAM.

Antoulas, A. C. (1997). "On the approximation of hankel matrices". In: *Operators, Systems and Linear Algebra: Three Decades of Algebraic Systems Theory*. Vieweg+Teubner Verlag, pp. 17–22.

Argyriou, A., R. Foygel, and N. Srebro (2012). "Sparse prediction with the k-support norm". In: Pereira, F. et al. (Eds.). *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., pp. 1457–1465.

Bach, F., R. Jenatton, J. Mairal, and G. Obozinski (2012). "Optimization with sparsity-inducing penalties". *Foundations and Trends in Machine Learning* **4**:1, pp. 1–106.

Bauschke, H. H. and P. L. Combettes (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer New York.

Candès, E. J. and Y. Plan (2010). "Matrix completion with noise". *Proceedings of the IEEE* **98**:6, pp. 925–936.

Candès, E. J., J. Romberg, and T. Tao (2006). "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information". *IEEE Transactions on Information Theory* **52**:2, pp. 489–509.

Candès, E. J. and B. Recht (2009). "Exact matrix completion via convex optimization". *Foundations of Computational Mathematics* **9**:6, p. 717.

Chandrasekaran, V., B. Recht, P. A. Parrilo, and A. S. Willsky (2012). "The convex geometry of linear inverse problems". *Foundations of Computational Mathematics* **12**:6, pp. 805–849.

Chandrasekaran, V., S. Sanghavi, P. A. Parrilo, and A. S. Willsky (2011). "Rank-sparsity incoherence for matrix decomposition". *SIAM Journal on Optimization* **21**:2, pp. 572–596.

Chen, Y., M. R. Jovanović, and T. T. Georgiou (2013). "State covariances and the matrix completion problem". In: *52nd IEEE Conference on Decision and Control*, pp. 1702–1707.

Combettes, P. L. and J.-C. Pesquet (2011). "Proximal splitting methods in signal processing". In: *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer New York, pp. 185–212.

Condat, L. (2016). "Fast projection onto the simplex and the ℓ_1 ball". *Mathematical Programming* **158**:1, pp. 575–585.

Doan, X. V. and S. Vavasis (2016). "Finding the largest low-rank clusters with Ky Fan 2-k-norm and ℓ_1-norm". *SIAM Journal on Optimization* **26**:1, pp. 274–312.

Duchi, J., S. Shalev-Shwartz, Y. Singer, and T. Chandra (2008). "Efficient projections onto the L1-ball for learning in high dimensions". In: *Proceedings of the 25th International Conference on Machine Learning*. 25th International Conference on Machine Learning (ICML), pp. 272–279.

Eriksson, A., T. T. Pham, T.-J. Chin, and I. Reid (2015). "The k-support norm and convex envelopes of cardinality and rank". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3349–3357.

Fazel, M., H. Hindi, and S. P. Boyd (2001). "A rank minimization heuristic with application to minimum order system approximation". In: *Proceedings of the 2001 American Control Conference.* Vol. 6, pp. 4734–4739.

Fazel, M. (2002). *Matrix Rank Minimization with Applications*. PhD thesis. Stanford University.

Georgiou, T. T. (2002a). "Spectral analysis based on the state covariance: the maximum entropy spectrum and linear fractional parametrization". *IEEE Transactions on Automatic Control* **47**:11, pp. 1811–1823.

Georgiou, T. T. (2002b). "The structure of state covariances and its relation to the power spectrum of the input". *IEEE Transactions on Automatic Control* **47**:7, pp. 1056–1066.

Grussler, C. and A. Rantzer (2014). "Modified balanced truncation preserving ellipsoidal cone-invariance". In: *53rd IEEE Conference on Decision and Control (CDC)*, pp. 2365–2370.

Grussler, C. and A. Rantzer (2015). "On optimal low-rank approximation of non-negative matrices". In: *54th IEEE Conference on Decision and Control (CDC)*, pp. 5278–5283.

Grussler, C., A. Rantzer, and P. Giselsson (2016a). "Low-rank optimization with convex constraints". arXiv: 1606.01793.

Grussler, C., A. Zare, M. R. Jovanovic, and A. Rantzer (2016b). "The use of the $r*$ heuristic in covariance completion problems". In: *55th IEEE Conference on Decision and Control (CDC)*.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.

Held, M., P. Wolfe, and H. P. Crowder (1974). "Validation of subgradient optimization". *Mathematical Programming* **6**:1, pp. 62–88.

Hiriart-Urruty, J.-B. and C. Lemaréchal (1996). *Convex analysis and minimization algorithms II*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Hiriart-Urruty, J.-B. and C. Lemaréchal (2013). *Convex analysis and minimization algorithms I: Fundamentals*. Vol. 305. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Horn, R. A. and C. R. Johnson (2012). *Matrix Analysis*. 2nd ed.

Izenman, A. J. (1975). "Reduced-rank regression for the multivariate linear model". *Journal of Multivariate Analysis* **5**:2, pp. 248–264.

Larsson, V. and C. Olsson (2016). "Convex low rank approximation". *International Journal of Computer Vision* **120**:2, pp. 194–214.

Lewis, A. S. (1995). "The convex analysis of unitarily invariant matrix functions". *Journal of Convex Analysis* **2**:1, pp. 173–183.

Lin, F., M. R. Jovanović, and T. T. Georgiou (2013). "An admm algorithm for matrix completion of partially known state covariances". In: *52nd IEEE Conference on Decision and Control*, pp. 1684–1689.

Liu, X., Z. Wen, and Y. Zhang (2013). "Limited memory block Krylov subspace optimization for computing dominant singular value decompositions". *SIAM Journal on Scientific Computing* **35**:3, A1641–A1668.

Luenberger, D. G. (1968). *Optimization by Vector Space Methods*. John Wiley & Sons.

McDonald, A. M., M. Pontil, and D. Stamos (2015). "New Perspectives on k-Support and Cluster Norms". arXiv: 1512.08204.

Parikh, N. and S. Boyd (2014). "Proximal algorithms". *Foundations and Trends in Optimization* **1**:3, pp. 127–239.

Peaucelle, D., D. Henrion, Y. Labit, and K. Taitz (2002). "User's guide for SEDUMI INTERFACE 1.04". LAAS-CNRS, Toulouse.

Recht, B., M. Fazel, and P. A. Parrilo (2010). "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization". *SIAM Review* **52**:3, pp. 471–501.

Reinsel, G. C. and R. Velu (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*. Vol. 136. Lecture Notes in Statistics. Springer New York.

Rockafellar, R. T. (1970). *Convex Analysis*. 28. Princeton University Press.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society* **58**:1, pp. 267–288.

Toh, K. C., R. H. Tutuncu, and M. J. Todd (2004). "On the implementation of SDPT3 (version 3.1) – a MATLAB software package for semidefinite-quadratic-linear programming". In: *2004 IEEE International Conference on Robotics and Automation*, pp. 290–296.

Toh, K.-C., M. J. Todd, and R. H. Tütüncü (1999). "SDPT3 – a MATLAB software package for semidefinite programming, version 1.3". *Optimization Methods and Software* **11**:1-4, pp. 545–581.

Vidal, R., Y. Ma, and S. S. Sastry (2016). *Generalized Principal Component Analysis*. Vol. 40. Interdisciplinary Applied Mathematics. Springer-Verlag New York.

Watson, G. (1992). "Characterization of the subdifferential of some matrix norms". *Linear Algebra and its Applications* **170**, pp. 33–45.

Wu, B., C. Ding, D. Sun, and K.-C. Toh (2014). "On the Moreau–Yosida regularization of the vector k-norm related functions". *SIAM Journal on Optimization* **24**:2, pp. 766–794.

Zare, A., Y. Chen, M. Jovanović, and T. T. Georgiou (2016a). "Low-complexity modeling of partially available second-order statistics: theory and an efficient matrix completion algorithm". *IEEE Transactions on Automatic Control*. arXiv:1412.3399. DOI: 10.1109/TAC.2016.2595761.

Zare, A., M. R. Jovanović, and T. T. Georgiou (2015). "Alternating direction optimization algorithms for covariance completion problems". In: *2015 American Control Conference (ACC)*, pp. 515–520.

Zare, A., M. R. Jovanović, and T. T. Georgiou (2016b). "Color of turbulence". *Journal of Fluid Mechanics*. arXiv:1602.05105. DOI: 10.1017/jfm.2016.682.

Acknowledgments

All authors are members of the LCCC Linnaeus Center and the eLLIIT Excellence Center at Lund University. The first author is financially supported by the Swedish Research Council through the project 621-2012-5357. The second author is financially supported by the Swedish Foundation for Strategic Research.