Low-rank optimization with convex constraints
CHRISTIAN GRUSSLER *, ANDERS RANTZER *, AND 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 an approximation of desired lower rank that fulfills the convex constraints and minimizes 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. We instead propose to use the optimal convex minorizer – the closed convex hull – of the 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, we get easily verifiable conditions 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. We also see that our proposed convex relaxation consistently performs better than the nuclear norm heuristic, especially in the matrix completion case.

The expressibility and computational tractability is of great importance for a convex relaxation. We provide a closed-form expression for the proposed convex approximation, and show how to represent it as a semi-definite program. We also show how to compute the proximal operator of the convex approximation. This allows us to use scalable first-order methods to solve convex approximation problems of large size.

Key words. Low-rank approximation, Douglas-Rachford splitting, Matrix Completion, k-support-norm

1. Introduction. The main reason for low-rank approximation theory lies in the idea of studying only the few essential components of an otherwise complex operator. For instance, it is well-known that 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 is 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 $\text{rank}(\hat{N}) < \text{rank}(N)$.

With this concept in mind, one can understand why many areas such as image analysis, model order reduction, multivariate linear regression, etc. desire a low-rank approximation (see [2, 9, 8, 11, 36, 43, 39, 45, 50, 57]). In Sections 5 to 8 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 non-negativity, Hankel-structure, prescribed entries, etc. (see [5, 9, 13, 32, 43, 45, 50, 57]). Only for a few known cases an explicit solution to the constrained low-rank approximation problem can be determined (see [2, 43, 57]). To this end, other concepts based on convex optimization have been developed (see [11, 21, 39, 50]). Many of them rely on nuclear-norm regularization, which allows to incorporate any convex constraint (see Section 2). Nevertheless, the question of optimality is not addressed, unless one aims for a minimum rank solution (see [9, 50]). Besides the nuclear-norm heuristic, other commonly used heuristics, e.g. for element-wise non-negativity are briefly considered in Section 5.

In this work, we study the optimal low-rank approximation problem with a prescribed target rank and convex constraints (see Problem 2). This is a continuation of the authors work [28]. 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, it can be determined by solving a semi-definite pro-
gram (SDP). Even though the approach presented can be linked to the regularization method in [39], we will see that it is essentially regularization parameter free.

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

The paper is organized as follows. In Section 2 we recap the unconstrained low-rank approximation problem and the nuclear-norm heuristic. The main approach is derived and discussed in Section 3 with some computational aspects examined in Section 4. In Sections 5 to 8 some applications are presented that show the usefulness of this approach. Moreover, it illustrates some properties and drawbacks of the methods addressed here. Finally, we draw a conclusion and discuss future research in Section 9.

2. Background. The following notation for real matrices and vectors $X = (x_{ij}) \in \mathbb{R}^{n \times m}$ is used throughout this paper. If $X = X^T$, then we write $X \in \mathcal{S}$. Moreover, if $X$ is positive definite (semi-definite) we write $X > 0$ ($X \succeq 0$). 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 $\sigma_1(X) \geq \ldots \geq \sigma_{\min\{m,n\}}(X)$.

Further we define by $\langle X, Y \rangle := \sum_{i=1}^m \sum_{j=1}^n x_{ij}y_{ij} = \text{trace}(X^T Y)$, $X, Y \in \mathbb{R}^{n \times m}$ the Frobenius inner-product on the Hilbert space $\mathbb{R}^{n \times m}$. Correspondingly, the Frobenius-norm is defined as

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

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 is given in [35].

This work mainly considers the norms that are found in the following Lemma, which is proven in Subsection A.3.

**Lemma 1.** Let $M \in \mathbb{R}^{n \times m}$, $r \in \mathbb{N}$ such that $1 \leq r \leq q := \min\{m,n\}$ and $\mathcal{P}_r$ denote the set of all orthogonal projections of rank $r$. Then

$$\|M\|_r := \sqrt{\sum_{i=1}^r \sigma_i^2(M)} = \max_{P \in \mathcal{P}_r} \left\langle P, MM^T \right\rangle$$

is a unitarily invariant norm with dual-norm

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

Moreover,

$$\|M\|_1 \leq \cdots \leq \|M\|_q = \|M\|_F = \|M\|_r = \|M\|_{r*} \leq \cdots \leq \|M\|_q.$$  

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

An explicit expression of $\|\cdot\|_{r*}$ was first derived in [22]. 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 [21, 50]). In Section 3 we will see that the same holds true for $\|\cdot\|_r$ and their duals.
Recently, the vector version of the $r*$-norm has also appeared under the name "$k$-support-norm" (see [3]). As a result, some authors have adopted that name for the matrix case (see [19, 38, 44]). However, as for other vector/matrix-norm pairings e.g. the $\ell_1$-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.

### 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.

**Problem 1.** Given $N \in \mathbb{R}^{n \times m}$ and $r \in \mathbb{N}$ such that $1 \leq r \leq \min\{m, n\}$, find $X^* \in \mathbb{R}^{n \times m}$ with $\text{rank}(X) \leq r$ such that

$$\min_{X \in \mathbb{R}^{n \times m}, \text{rank}(X) \leq r} \|N - X\| = \|N - X^*\|$$

for some given operator norm $\|\cdot\|$.

In case of the Hilbert-Schmidt norm, the natural operator generalization of the Frobenius-norm, the problem has been solved by Schmidt and generalized by Mirsky to unitarily invariant norms (see [2]). 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

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

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 Problem 1 can be derived as $X^* = \text{svd}_r(X) := \sum_{i=r+1}^{\min\{m,n\}} \sigma_i u_i v_i^T$, 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 $X^*$ is guaranteed, which on the other hand does not account for additional constraints.

In this work we look at the following extension of Problem 1.

**Problem 2.** Given $N \in \mathbb{R}^{n \times m}$, find $M^* \in \mathbb{R}^{n \times m}$ with $\text{rank}(M^*) \leq r$ such that

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

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 Problem 1, Problem 2 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 Problem 1 has a unique solution, it is possible to determine it by solving a semi-definite program.

In the following we often think of $g$ as $g(M) \equiv \chi_C(M)$ where

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

is defined to be the characteristic function of a convex set $C \subset \mathbb{R}^{n \times m}$ – within the area of optimization often called indicator function. Notice that the problem formulation also deals
with cases where \( N = 0 \), which then covers the class of matrix completion problems (see Section 6).

### 2.2. Nuclear-Norm Regularization.

One of the most widely used approaches to approximate a solution to Problem 2 is the so-called nuclear-norm regularization. It borrows techniques from sparse regularized regression or Lasso (see [55]), i.e. estimating a sparse solution \( \hat{x} \) to a linear system of equations \( A\hat{x} \approx b \) by solving

\[
\min_{x} \frac{1}{2} \|Ax - b\|^2_2 + \mu \|x\|_1,
\]

where \( \| \cdot \|_2 \) is the Euclidean norm, \( \|x\|_1 = \sum_{i \geq 1} |x_i| \) and \( \mu \geq 0 \) is a regularization parameter. Sparsity of the singular values is equivalent to the matrix having low rank. Therefore, its generalized matrix version for given \( N \in \mathbb{R}^{n \times m} \) reads

\[
\min_{M} \frac{1}{2} \|N - M\|^2_F + \mu \|M\|_{1^*} + g(M),
\]

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 [20, 21, 50] stimulated a big growth in applying this approach in many different areas (see [20, 21, 45, 50]). However, to get a specific rank, \( \mu \) must be chosen a priori, which is often challenging. Commonly one assumes that the rank, as a function of \( \mu \), looks like a staircase, i.e. a large \( \mu \) decreases the rank too much whereas a small \( \mu \) may leave it too large. In order to find the best possible approximation, one usually likes to keep \( \mu \) as small as possible, which on the other hand could end up in a costly search.

In general, even with the best possible choice of \( \mu \), this heuristic does not return an optimal solution to Problem 2. In particular, also for the simple case \( g = 0 \), one usually cannot not choose \( \mu \) such that the SVD-approximation, as required by Proposition 1, is obtained. Furthermore, there is no certificate for checking whether a solution is a minimizer of Problem 2.

### 3. The \( r^* \) approach.

In the following we consider an attempt of finding an optimal solution to Problem 2. It is a continuation of the authors work [28]. The insights obtained here will serve us later to generalize and improve upon current standard approaches (see Section 6). The main idea is to derive a convex minorizer of the non-convex cost-function in Problem 2 by means of Fenchel-duality (see Subsection A.2). We denote by \( f^* \) and \( f^{**} \) the conjugate and bi-conjugate functionals of \( f : \mathbb{R}^{n \times m} \to \mathbb{R} \cup \{\infty\} \) (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 functionals of \( f(M) := \frac{1}{2} \|N - M\|^2_F + \chi_{\text{rank}(M) \leq r}(M) \) are given by

\[
(5) \quad f^*(D) = \frac{1}{2} \|N + D\|^2_F - \frac{1}{2} \|N\|^2_F,
\]

\[
(6) \quad f^{**}(M) = \frac{1}{2} \|M\|^2_{1^*} - \langle N, M \rangle + \frac{1}{2} \|N\|^2_F.
\]
Problem 2: hull (see Figure 1). This allows us to construct the following dual and bi-dual problem to of $f$.

It is possible to show that $f^*$ is the best convex minorizer of $f$. In fact, it is the best convex minorizer of $f$ (see [33, Theorem 1.3.5]), the closed convex hull (see Figure 1). This allows us to construct the following dual and bi-dual problem to Problem 2:

$$\begin{align*}
(A) & \quad - \min_{D \in \mathbb{R}^{n \times m}} \left[ g^*(-D) + \frac{1}{2} \|N + D\|_F^2 - \frac{1}{2} \|N\|_F^2 \right], \\
(B) & \quad \min_{M \in \mathbb{R}^{n \times m}} \left[ \frac{1}{2} \|M\|_{r_{rs}}^2 - \langle N, M \rangle + \frac{1}{2} \|N\|_F^2 + g(M) \right].
\end{align*}$$

Observe that $f^{**} + g$ is the best convex minorizer of $f + g$ that keeps $g$ intact. Therefore, we propose to use (B) instead of the standard nuclear norm heuristic (4) as a convex proxy to Problem 2. 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).

**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 $1 \leq r \leq \min\{m, n\}$ one has

$$\begin{align*}
(C) & \quad \min_{M \in \mathbb{R}^{n \times m} \atop \text{rank}(M) \leq r} \left[ \frac{1}{2} \|N - M\|_F^2 + g(M) \right] \\
& \quad \geq - \min_{D \in \mathbb{R}^{n \times m}} \left[ g^*(-D) + \frac{1}{2} \|N + D\|_F^2 - \frac{1}{2} \|N\|_F^2 \right] \\
& \quad = \min_{M \in \mathbb{R}^{n \times m} \atop \text{rank}(M) \leq r} \left[ \frac{1}{2} \|M\|_{r_{rs}}^2 - \langle N, M \rangle + \frac{1}{2} \|N\|_F^2 + g(M) \right].
\end{align*}$$

Since the original Problem 2 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 situations without a duality-gap. Next, we present a number of such cases.
We start by providing a posteriori tests if a solution to the convex relaxation in (B) is optimal for Problem 2. The set of minimizers of a functional $f$ over a given set $S$ is denoted by $\text{argmin}_S f$. If $\text{argmin}_S f = \{x^*\}$ is just a singleton, we write $x^* = \text{argmin}_S f$.

**Proposition 3.** Assume that $M^*$ is a minimizer of (B) such that $\text{rank}(M^*) \leq r$. Then,

$$\argmin_{M: \text{svd}(M) \leq r} \left[ \frac{1}{2} \|N - M\|_F^2 + g(M) \right] = \argmin_{M: \text{svd}(M) \leq r} \left[ \frac{1}{2} \|M\|_F^2 - \langle N, M \rangle + \frac{1}{2} \|N\|_F^2 + g(M) \right].$$

**Proof.** The result follows by combining (B) with (3) in Lemma 1.

Thus obtaining a rank-$r$ solution to the convex relaxation problem (B) implies solving the original non-convex problem to optimality. Another way to state this result, that gives additional insight on the solution to Problem 2, is as follows.

**Proposition 4.** Assume that $D^*$ is a solution to (A) and $\sigma_r(N + D^*) \neq \sigma_{r+1}(N + D^*)$ or $\sigma_r(N + D^*) = 0$. Then there is no duality gap in (C) and

$$\text{svd}_r(N + D^*) = \argmin_{M: \text{svd}(M) \leq r} \left[ \frac{1}{2} \|N - M\|_F^2 + g(M) \right].$$

This result provides a simple sufficient condition for the uniqueness of a solution to Problem 2. However, it is not a necessary condition due to a possible duality-gap. A proof of Proposition 4 is given in a more general setting in Proposition 6.

Next let us look at cases for a zero duality-gap that do not necessarily involve solving (A) and (B). We first consider the case where $g = 0$.

**Proposition 5.** Let $N \in \mathbb{R}^{n \times m}$ and $r \in \mathbb{N}$ be such that $1 \leq r \leq \min\{m, n\}$. Then

$$\min_{M: \text{svd}(M) \leq r} \frac{1}{2} \|N - M\|_F^2 = \frac{1}{2} \|N\|_F^2 - \frac{1}{2} \|N\|_r^2 = \min_{M: \text{svd}(M) \leq r} \left[ \frac{1}{2} \|M\|_r^2 - \langle N, M \rangle + \frac{1}{2} \|N\|_F^2 \right]$$

and $\text{svd}_r(N) \in \argmin_{M: \text{svd}(M) \leq r} \left[ \frac{1}{2} \|M\|_r^2 - \langle N, M \rangle \right]$. If $\sigma_r(N) \neq \sigma_{r+1}(N)$ or $\sigma_r = 0$ then

$$\text{svd}_r(N) = \argmin_{M: \text{svd}(M) \leq r} \left[ \frac{1}{2} \|M\|_r^2 - \langle N, M \rangle \right].$$
Proof. Notice that \( g = 0 \) implies \( g^*(D) < \infty \) \( \iff D = 0 \), Then the result follows by invoking (3) in Lemma 1 and Proposition 4.

This is another explanation of Proposition 1 that \( \text{svd}(N) \) is a solution to Problem 1, i.e. Problem 2 with \( g = 0 \). An immediate consequence of this result is the following.

**Corollary 1.** Let \( N \in \mathbb{R}^{n \times m} \) and \( r \in \mathbb{N} \) be such that \( 1 \leq r \leq \min\{m,n\} \). Suppose \( g = \chi_\mathcal{C} \) for some closed convex set \( \mathcal{C} \subset \mathbb{R}^{n \times m} \) such that there exists \( M^* \in \mathcal{C} \) with

\[
M^* \in \operatorname{argmin}_{M \in \mathcal{C}} \frac{1}{2} \|N - M\|_F^2.
\]

Then equality in (C) holds and \( M^* \) is a minimizer of the convex proxy in (B).

This corollary implies that if the constraint set \( \mathcal{C} \) does not affect the solution, then there is a zero duality gap.

We finish our discussion on the solutions to (A) and (B) with a result for which we do not necessarily have a zero duality-gap, but we can say something about the rank of the convex relaxation solution.

**Proposition 6.** Let \( D^* \) and \( M^* \) be solutions to (A) and (B), respectively. Further suppose that an SVD of \( N + D^* \) is given by \( N + D^* = \sum_{i=1}^{\min\{m,n\}} \sigma_i(N + D^*) u_i v_i^T \) with \( \sigma_{r+1}(N + D^*) \neq \sigma_{r+1}(N + D^*) = \cdots = \sigma_{s}(N + D^*) \neq \sigma_{s+1}(N + D^*) \), where \( t = r \) and \( s = \min\{m,n\} - r \) if \( \sigma_1 = \sigma \) and \( \sigma_{\min\{m,n\}} = \sigma_r \), respectively. Then there exists \( T \in \mathbb{R}^{n \times r+s+t} \) with \( T \succeq 0 \), \( \|T\|_1 \leq 1 \) and \( \|T\|_{1s} = t \) such that

\[
M^* = \sum_{i=1}^{r} \sigma_i(N + D^*) u_i v_i^T + \sigma_r(N + D^*) (u_{r-t+1}, \ldots, u_{r+s}) T (v_{r-t+1}, \ldots, v_{r+s})^T.
\]

In particular, \( \operatorname{rank}(M^*) \leq r + s \) and if \( \sigma_r(N + D^*) \neq \sigma_{r+1}(N + D^*) \) or \( \sigma_s(N + D^*) = 0 \) then \( M^* = \text{svd}(N + D^*) \).

**Proof.** If \( D^* \) and \( M^* \) are solutions to (A) and (B), respectively, then by Proposition A.3 it holds that \( f^{*\ast}(M^*) = \langle D^*, M^* \rangle - f^*(D^*) \), where \( f^* \) and \( f^{*\ast} \) are given by (5) and (6). Hence, by Proposition A.4 it follows that

\[
M^* \in \partial_D \frac{1}{2} \|N + D\|_F^2 \bigg|_{D = D^*} = \|N + D^*\|_F \partial_D \|N + D\|_F \bigg|_{D = D^*}
\]

and invoking Proposition A.5 proves the result.

Observe that whenever (B) does not have a unique solution, it must hold by Proposition 6 that \( \sigma_s(N + D^*) = \sigma_{s+1}(N + D^*) \) for all solutions \( D^* \) to (A).

Finally, notice that several extensions of Problem 2 are covered by the preceding results. For instance, one can consider the weighted case

\[
\min_{M \in \mathbb{R}^{n \times m}} \min_{\text{rank}(M) \leq r} \left[ \frac{1}{2} \|W(N - M)\|_F^2 + g(M) \right]
\]

where \( W \in \mathbb{R}^{n \times n} \) and \( \text{rank}(W) = n \). Let \( g(M) := g(W^T M) \), where \( W^T \) denotes the pseudo-inverse of \( W \) (see [35]). Since \( \text{rank}(W) = \text{rank}(W^T M) = \text{rank}(M) \), one can reformulate (7) such that it fits the formulation of Problem 2:

\[
\min_{M \in \mathbb{R}^{n \times m}} \min_{\text{rank}(M) \leq r} \left[ \frac{1}{2} \|W(N - M)\|_F^2 + g(M) \right] = \min_{M \in \mathbb{R}^{n \times m}} \min_{\text{rank}(M) \leq r} \left[ \frac{1}{2} \|W(N - M)\|_F^2 + \bar{g}(M) \right].
\]
Note that \( \|W(N-M)\|^2_F = \text{trace}( (N-M)^T W(N+M)) = \langle N-M, N-M \rangle_{W^TW} \) defines another inner product and norm. A suitable \( W \) may enable us to satisfy the requirements of Proposition 4, where the Frobenius-norm fails to do so.

The insights obtained in this subsection will also serve us to generalize and improve upon current standard approaches (see Section 6).

### 3.1. Geometric interpretation

Assuming that \( g(M) = \chi_{\mathcal{E}}(M) \) for some closed convex set \( \mathcal{C} \subset \mathbb{R}^{n \times m} \), the preceding results offer an insightful geometric interpretation. Notice that \((B)\) is equivalent to

\[
\min_{M \in \mathcal{C}} \|M\|_{rs},
\]

where \( M^* \) is a solution to \((B)\) and \( e := \langle N, M^* \rangle \). All solutions of \((8)\) can be found by studying the set \( B^e_{rs} \cap H \cap \mathcal{C} \) where

\[
B^e_{rs} := \{ X : \|X\|_{rs} \leq \epsilon \},
\]

\[
H := \{ X : \langle N, X \rangle = e \}
\]

and \( \bar{\epsilon} := \min \{ \epsilon \geq 0 : B^\epsilon_{rs} \cap H \cap \mathcal{C} \neq \emptyset \} \). Proposition 4 states, if \( \sigma_r(N+D^*) \neq \sigma_{r+1}(N+D^*) \) then \( B^e_{rs} \cap H \cap \mathcal{C} \) 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 \( r = 1 \).

**Lemma 2.** The set of the extreme points of the unit-ball \( B^1_{rs} \) is

\[
E := \{ X \in \mathbb{R}^{n \times m} : \|X\|_F = 1, \text{ rank}(X) \leq r \}.
\]

Hence, \( B^1_{rs} = \text{conv}(E) \), where \( \text{conv}(\cdot) \) denotes the convex hull.

**Proof.** By the triangle inequality and \((3)\) in Lemma 1 it follows that \( \text{conv}(E) \subset B^1_{rs} \) (see Figure 2a). Moreover, by \((1)\) in Lemma 1 it holds that

\[
\forall N \in \mathbb{R}^{n \times m} : \sup_{M \in \text{conv}(E)} \langle N, M \rangle = \|N\|_r = \sup_{M \in B^1_{rs}} \langle N, M \rangle.
\]

However, since \( \text{conv}(E) \) and \( B^1_{rs} \) are closed sets, this equality holds if and only if \( B^1_{rs} = \text{conv}(E) \). If a point \( \tilde{M} \in E \) is not an extreme point of \( E \), then \( \tilde{M} = \sum \alpha_i M_i, \sum \alpha_i = 1, \alpha_i > 0, M_i \in K \setminus \{ \tilde{M} \} \) for all \( i \). Thus, by the Cauchy-Schwarz inequality, we conclude that \( \langle \tilde{M}, M_i \rangle = 1 \) for all \( i \) and \( \tilde{M} = M_i \), which is a contradiction.

Therefore, a geometric interpretation of \( \sigma_r(N+D^*) \neq \sigma_{r+1}(N+D^*) \) is that the only intersection point of \( H \) and \( B^e_{rs} \cap \mathcal{C} \) is an extreme point of \( B^e_{rs} \) (see Figure 2b). Hence, the case of \( \sigma_r(N+D^*) = \sigma_{r+1}(N+D^*) \neq 0 \) can occur if and only if \( H \) intersects \( B^e_{rs} \cap \mathcal{C} \) at several points (see Figure 2c and Subsection 5.2.2) or if there is a duality gap in \((C)\) (see Figure 2d and Subsection 7.1). Finally notice that one can also use Lemma 2 as a definition of \( \| \cdot \|_{rs} \). This has been done for vectors in [3] by intending to generalize upon the \( \ell_1 \)-norm.

### 3.2. Real-valued r

In the following we will see that allowing \( r \) to be real-valued can be considered a regularization parameter. Unlike typical regularization methods (see Subsections 2.2 and 3.3), 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.

Let \( r \in \mathbb{N} \) be such that \( \sigma_r(N+D^*) = \sigma_{r+1}(N+D^*) \) and \( \text{rank}(M^*_r) > r \), where we define

\[
M^*_r := \left[ \arg\min_{M \in \mathbb{R}^{n \times m}} \frac{1}{2} \|M\|_{rs}^2 - \langle N, M \rangle + g(M) \right], \quad D^*_r := \arg\min_{D \in \mathbb{R}^{n \times m}} \left[ g^*(-D) + \frac{1}{2} \|N+D\|_{rs}^2 \right]
\]
LOW-RANK OPTIMIZATION WITH CONVEX CONSTRAINTS

(a) $B_r^1$ (shaded area) as the convex hull of $E$ (elements $\circ$) with the boundary of $B_{\min(n,m)+1}^1(---)$.

(b) Unique solution: $\{M^*\} = B_{r^*}^2 \cap H \cap \mathcal{C}$ with $\text{rank}(M^*) \leq r$.

(c) Non-unique solutions: $\{M^*_1, M^*_2\} = B_{r^*}^2 \cap M^* \cap \mathcal{H} \cap \mathcal{C}$ with $\text{rank}(M^*_1) \leq r$ and $\text{rank}(M^*_2) \leq r$.

(d) Duality gap: $\{M^*\} = B_{r^*}^2 \cap H \cap \mathcal{C}$ with $\text{rank}(M^*) > r$.

Figure 2: Schematic plots to visualize (8) geometrically.

for $1 \leq r \leq \min\{m,n\}$. Assume that $\frac{1}{2}\|N - M_r^*\|_F^2 + g(M^*_r) > \frac{1}{2}\|N - M^*_{r+1}\|_F^2 + g(M^*_{r+1})$ and $\text{rank}(M^*_r) > \text{rank}(M^*_{r+1})$. Then one may be in the situation that $M^*_r$ is an approximation of small enough rank but poor cost $\|N - M^*_r\|_F^2 + g(M^*_r)$. Furthermore, $\|N - M^*_{r+1}\|_F^2 + g(M^*_{r+1})$ may be acceptable while $\text{rank}(M^*_{r+1})$ is too large. Thus a trade-off between $M^*_r$ and $M^*_{r+1}$ 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

\[
\|\cdot\|_r := \sqrt{\sum_{i=1}^{\lfloor r \rfloor} \sigma_i^2(\cdot) + (r - \lfloor r \rfloor) \sigma_{\lfloor r \rfloor + 1}^2(\cdot)},
\]

where $\lfloor r \rfloor := \max\{z \in \mathbb{Z} : z \leq r\}$ and $\lceil r \rceil := \min\{z \in \mathbb{Z} : z \geq r\}$. Observe that for $r \in \mathbb{N}$ and $\alpha \in [0,1]$ we have

\[
\|\cdot\|^2_{r+\alpha} = (1 - \alpha)\|\cdot\|_r^2 + \alpha\|\cdot\|_{r+1}^2,
\]
which means that $\|\cdot\|_{2,1-\alpha}^2$ is a convex combination of $\|\cdot\|_2^2$ and $\|\cdot\|_{2,1}^2$, thus indicating its usefulness in supplying the desired trade-off solution. Similar to Proposition 6, also with $r \in \mathbb{R}_{\geq 1}$ it remains true that $\text{rank}(M_r^f) \leq |r| + s$ and

$$\sigma_{|r|}(N + D_r^f) = \cdots = \sigma_{|r| + s}(N + D_r^f) > \sigma_{|r| + s + 1}(N + D_r^f).$$

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 (11).

Let us define

$$F(r) := g^*(-D) + \frac{1}{2}\|N + D\|_F^2 + \frac{1}{2}\|N\|_F^2.$$  

Using (2) in Lemma 1, we conclude that $F$ is monotonically decreasing. In conjunction with the piecewise linearity in (10), it follows that $F$ is convex and thus continuous. From Berge’s Maximum Theorem (see [4, p. 116] or [54, Theorem 9.17] for the convex case) it is known that the parameter depending set

$$\mathcal{C}^*(r) := \text{argmin}_{D \in \mathbb{R}^{m \times m}} \left[ g^*(-D) + \frac{1}{2}\|N + D\|_F^2 + \frac{1}{2}\|N\|_F^2 \right]$$

is upper hemicontinuous in $r$. This means that for all $r \in [1, \min\{m, n\}]$ and all $\epsilon > 0$ there exists $\delta > 0$ such that for all $t \geq 1$

$$|t - r| < \delta \Rightarrow \mathcal{C}^*(t) \subset \mathcal{B}_\epsilon(\mathcal{C}^*(r)),$$

where $\mathcal{B}_\epsilon(\mathcal{C}^*(r)) := \{X : \exists D \in \mathcal{C}^*(r) \text{ such that } \|X - D\|_F < \epsilon\}$. For simplicity we assume that $D_r^f$ is unique. By (12) and the continuity of the singular values (see [53, Corollary 4.9]), it follows that a sufficiently small increase of $r$ does not increase $s$. Therefore, as for the nuclear-norm regularization, one often observes $\text{rank}(M_r^f)$, as a function of $t \in [r, r + 1]$, to look like a staircase (see Figure 9b in Section 7). Notice that a similar consideration can be done with

$$F(r) := \frac{1}{2}\|M\|_F^2 - \langle N, M \rangle + \frac{1}{2}\|N\|_F^2 + g(M)$$

and

$$\mathcal{C}^*(r) := \text{argmin}_{M \in \mathbb{R}^{m \times n}} \left[ \frac{1}{2}\|M\|_{F_2}^2 - \langle N, M \rangle + \frac{1}{2}\|N\|_F^2 + g(M) \right].$$

In summary, real-valued $r$ can be considered as a regularization parameter.

### 3.3. Rank Regularization

Similar to the nuclear-norm regularization (see Subsection 2.2), one may directly regularize on the rank, i.e.

$$\min_{M \in \mathbb{R}^{m \times n}} \left[ \frac{1}{2}\|N - M\|_F^2 + \mu \text{rank}(M) + g(M) \right],$$

where $\mu \geq 0$ is a regularization parameter and $g : \mathbb{R}^{m \times n} \rightarrow \mathbb{R} \cup \{\infty\}$ a closed and proper convex function. Since this problem is still non-convex, one needs to find a convex proxy of $f(M) := \frac{1}{2}\|N - M\|_F^2 + \mu \text{rank}(M)$. The conjugate and bi-conjugate functionals of $f$ are given by (see [39])

$$f^*(D) = \frac{1}{2}\|N + D\|_F^2 - \frac{1}{2}\|N\|_F^2 - \frac{1}{2} \sum_{i=1}^{\min\{m, n\}} \min\{2\mu, \sigma_i^2(N + D)\},$$

$$f^{**}(M) = \frac{1}{2}\|M - N\|_F^2 + \frac{1}{2} \sum_{i=1}^{\min\{m, n\}} \left( 2\mu - \max\{0, \sqrt{2\mu - \sigma_i(M)}\} \right)^2.$$
Hence, by Fenchel-duality (see Lemma A.1 and Proposition A.3) it holds that
\begin{equation}
\min_{M \in \mathbb{R}^{n \times m}} [f(M) + g(M)] \geq - \min_{D \in \mathbb{R}^{n \times m}} [f^*(M) + g^*(-D)] = \min_{M \in \mathbb{R}^{n \times m}} [f^{**}(M) + g(M)].
\end{equation}

Assume that there is no duality gap in (C) with solutions $D^*$ and $M^*$ to (A) and (B), respectively. Choosing $\frac{\sigma_2^2(N+D^*)}{2} \geq \mu \geq \frac{\sigma_2^2(N+D^*)}{2}$, it is readily seen that
\begin{equation*}
f^*(D^*) = \frac{1}{2} \|N\|^2_F - \frac{1}{2} \|N + D^*\|^2_F + \mu r = \frac{1}{2} \|N - M^*\|^2_F + \mu r + g(M^*),
\end{equation*}
where the last equality follows by Propositions 2 and 3. Hence,
\begin{equation*}
\frac{1}{2} \|N - M^*\|^2_F + \mu r + g(M^*) \geq - \min_{D \in \mathbb{R}^{n \times m}} [f^*(D) + g^*(-D)] \geq \frac{1}{2} \|N - M^*\|^2_F + \mu r + g(M^*),
\end{equation*}
yielding equality in (14). Thus, this 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 to prefer. Notice that, as in the case of real-valued $r$, Berge’s maximum Theorem can be applied since $f^*$ is continuous in $\mu$. Again, this implies that the rank of the solutions is robust to small perturbations in $\mu$. Nevertheless, there is no staircase behavior as one exhibits for real-valued $r$ and the nuclear-norm regularization (see Figure 9b in Subsection 7.1). Thus, since a good choice of the regularization parameter is usually a priori unknown, one is required to sweep over a large set of possible choices of $\mu$.

The proximal-operator of $f^{**}$ is computable (see [39]) and therefore there are several first order optimization methods (see Subsection 4.2) that can be used to compute a minimizer of (14). Nevertheless, this usually limits one to choices of $g$ that posses a cheaply computable proximal operator of $g$ for even small dimensional examples. It is currently unknown if (13) is SDP-representable.

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-operators 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. We further apply Douglas-Rachford to the original non-convex Problem 2. We show that 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
\begin{equation}
\min_{D \in \mathbb{R}^{n \times m}} \|N + D\|^2_F,
\end{equation}
where $\| \cdot \|_F$ is defined as in (9) and $r \in [1, \min\{m, n\}]$. Let $T \in \mathbb{R}^{n \times n}$ be such that $T \geq (N + D)(N + D)^T$, then $\text{trace}(T) = \sum_{i=1}^n \sigma_i(T)$ and $\sigma_i(T) \geq \sigma_i^2(N + D)$ for all $i$ such that $1 \leq i \leq \min\{m, n\}$ (see [35, Corollary 7.7.4.]). Therefore,
\begin{equation*}
\|N + D\|^2_F \leq \text{trace}(T) - ([r] - r)\sigma_{[r]}(T) - \sum_{i=[r]+1}^n \sigma_i(T) \leq \text{trace}(T) - (n - r)\sigma_n(T),
\end{equation*}
which is equivalent to
\begin{equation}
\|N + D\|^2_F \leq \min_{T \geq (N + D)(N + D)^T} \text{trace}(T) - (n - r)\sigma_n(T).
\end{equation}
In particular, equality in (16) can be achieved with

\[ T^+ := \sum_{i=1}^{r} \sigma_i^2 (N + D)u_iu_i^T + \sigma_{r+1}^2 (N + D) \sum_{i=r+1}^{n} u_iu_i^T \]

where \((N + D) = \sum_{i=1}^{n} \sigma_i (N + D)u_iu_i^T\) is an SVD of \(N + D\). Furthermore, using the Schur-complement condition for positive semi-definiteness of \(T - (N + D)(N + D)^T \succeq 0\) (see [35, Theorem 7.7.7.]), gives that (15) is equivalent to

\[
\begin{aligned}
&\text{minimize} \quad \text{trace}(T) - \gamma (n - r) \\
&\text{subject to} \quad \left( \begin{array}{cc} T & N + D \\ (N + D)^T & I \end{array} \right) \succeq 0, \quad T \succeq \gamma I, \quad D \in \mathbb{R}^{n \times m}.
\end{aligned}
\]

Thus, if \(g\) is SDP-representable, then the dual of this optimization yields an SDP-formulation of (B) (see [7, 50] for \(r = 1\)). We get

\[
\begin{aligned}
&\text{minimize} \quad \frac{1}{2} \text{trace}(W) - \text{trace}(N^T M) + g(M) \\
&\text{subject to} \quad \left( \begin{array}{cc} I - P & M \\ M^T & W \end{array} \right) \succeq 0, \quad P \succeq 0, \\
&\quad \text{trace}(P) = m - r.
\end{aligned}
\]

Assuming that \(\sigma_r (N + D^*) \neq \sigma_{r+1} (N + D^*)\), the unique solution \(M^*\) to Problem 2 can be found directly, without first computing the solution to (A).

### 4.2. Convex Douglas-Rachford

Many SDP-solvers are based on interior point methods (see [48, 56]). These solvers have good convergence properties, but the iteration complexity typically grows unfavorably with 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 [16, 17, 41]). Let us recall the basic concept of this method. We want to determine a solution to

\[
\begin{aligned}
&\text{minimize} \quad f(X) + g(X)
\end{aligned}
\]

where \(f, g : \mathbb{R}^{n \times m} \to \mathbb{R} \cup \{\infty\}\) are closed and proper convex functions. Then the Douglas-Rachford iteration is given by

\[
\begin{aligned}
X^k &= \text{prox}_{\gamma f}(Z^{k-1}), \\
Y^k &= \text{prox}_{\gamma g}(2X^k - Z^{k-1}), \\
Z^k &= Z^{k-1} + \rho (Y^k - X^k),
\end{aligned}
\]

where \(\gamma > 0, \rho \in (0, 2)\) and the proximal-operator is defined as

\[
\text{prox}_{\gamma f}(Z) := \arg\min_X \left( f(X) + \frac{1}{2\gamma} \|X - Z\|_F^2 \right).
\]

It is known that \(X^k\) and \(Y^k\) converge towards a minimizer of (17) (see [16, 17, 41]). A special case of these iterations is the well-known Alternating Direction Methods of Multipliers (ADMM) (see [6, 23, 26]). Note that the Douglas-Rachford splitting algorithm can also be applied to sums of more than two functions \(f\) and \(g\) (see e.g. [14]).
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 (B) it remains to find \( \text{prox}_{\gamma f}(Z) \) with

\[
f(M) := \frac{1}{2} \|M\|_r^2 - \langle N, M \rangle + \frac{1}{2} \|N\|_F^2.
\]

We get

\[
\text{prox}_{\gamma f}(Z) = \arg\min_{M \in \mathbb{R}^{n \times m}} \left( \frac{1}{2} \|M\|_r^2 - \langle N, M \rangle + \frac{1}{2} \|N\|_F^2 + \frac{1}{2\gamma} \|M - Z\|_F^2 \right)
\]

\[
= \arg\min_{M \in \mathbb{R}^{n \times m}} \left( \frac{1}{2} \|M\|_r^2 + \frac{1}{2\gamma} \|M - (\gamma N + Z)\|_F^2 + \langle Z, N \rangle \right)
\]

\[
= \text{prox}_{\frac{\gamma}{2\gamma} \cdot \|\cdot\|_r^2} \left( \gamma N + Z \right).
\]

Using the extended Moreau-decomposition (see [46]) and Theorem 1, it holds that for all \( Z \)

\[
\text{prox}_{\frac{\gamma}{2\gamma} \cdot \|\cdot\|_r^2} \left( \gamma N + Z \right) + \gamma \text{prox}_{\frac{1}{2\gamma} \cdot \|\cdot\|_r^2} \left( \gamma^{-1} Z \right) = Z.
\]

In combination we arrive at

\[
(22) \quad \text{prox}_{\gamma f}(Z) = \gamma N + Z - \gamma \text{prox}_{\frac{1}{2\gamma} \cdot \|\cdot\|_r^2} \left( \gamma N + Z \right).
\]

Therefore, it is sufficient to derive how to compute \( \text{prox}_{\frac{\gamma}{2\gamma} \cdot \|\cdot\|_r^2} \). This is done in Algorithm 1 on page 32 for \( r \in [1, \min\{m, n\}] \). Explanatory derivations can be found in Subsection A.4. For integer-valued \( r \) similar derivations are presented in [19].

Finally, observe that if \( r \in \mathbb{N} \) and

\[
(23) \quad \sigma_r(\gamma N + Z) > (1 + \gamma) \sigma_{r+1}(\gamma N + Z),
\]

it follows from the derivations of \( \text{prox}_{\frac{\gamma}{2\gamma} \cdot \|\cdot\|_r^2} \) (see Subsection A.4 and in particular (45)) that

\[
\text{prox}_{\frac{\gamma}{2\gamma} \cdot \|\cdot\|_r^2} \left( \frac{\gamma N + Z}{\gamma} \right) = \frac{\gamma N + Z}{\gamma} - \frac{1}{1 + \gamma} \text{svd}_r \left( \frac{\gamma N + Z}{\gamma} \right).
\]

Therefore, (22) implies that

\[
(24) \quad \text{prox}_{\gamma f}(Z) = \frac{1}{1 + \gamma} \text{svd}_r \left( \gamma N + Z \right).
\]

We will use this fact in Subsection 4.4 to show a tight relationship with 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 Proposition 6), gives that all limit points \( Z^* = \lim_{k \to \infty} Z_k \) of (20) can be expressed as

\[
(25) \quad Z^* = M^* + \gamma D^*,
\]

where \( D^* \) and \( M^* \) are solutions to (A) and (B), respectively. Given \( M^*, Z^* \) and \( \gamma \), this allows us to determine \( D^* \). Moreover, by inspection of the Douglas-Rachford iterations, it can be shown that several known properties of the standard SVD-approximation remain true, if \( \text{prox}_{\gamma}(X) \) is preserving them.
**Proposition 7.** Let $N$ and $g$ be as in Problem 2. 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 $D^*, M^* \in \mathbb{S}$.

ii. Let $Nv = 0$ and $\text{prox}_g(X)v = 0$ for all $X$ with $Xv = 0$. Then (B) has a solution $M^*$ such that $M^*v = 0$.

In particular, if (B) has a unique solution and there is a zero duality gap in (C), then the solution to Problem 2 preserves these properties.

**Proof.** Using [58, Theorem 2] it can be shown that $\text{prox}_{\frac{1}{2}\|\cdot\|^2_F}^g(N)$ has the same singular vectors as $N$. Therefore, $\text{prox}_{\frac{1}{2}\|\cdot\|^2_F}^g(N)$ 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 7 applies, a few examples will be discussed in Sections 5 to 7.

According to Proposition 4, $\sigma_r(N + D^*) \neq \sigma_{r+1}(N + D^*)$ is a sufficient condition for the uniqueness of a solution to (B). Note that without this uniqueness assumption, a solution to Problem 2 does not necessarily preserve these properties. This can be used to construct non-trivial examples where $\sigma_r(N + D^*) = \sigma_{r+1}(N + D^*)$ (see Subsection 5.2.2).

**4.4. Non-convex Douglas-Rachford (NDR).** Another approach to solve Problem 2 is to directly apply the Douglas-Rachford method to the non-convex problem

$$\min_{M \in \mathbb{R}^{n \times m}} \left[ \frac{1}{2} \|N - M\|^2_F + \chi_{\text{rank}(M) \leq r}(M) + g(M) \right].$$

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 [30, 31, 49]). 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 $\text{prox}_{\tilde{f}(Z)}$ where

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

We get

$$\text{prox}_{\tilde{f}}(Z) = \arg\min_{M \in \mathbb{R}^{n \times m}} \left( \frac{\gamma}{2} \|N - M\|^2_F + \frac{1}{2} \|M - Z\|^2_F \right).$$

This can be written as

$$\arg\min_{M \in \mathbb{R}^{n \times m}} \left( \frac{\gamma + 1}{2} \|M\|^2_F - \langle \gamma N + Z, M \rangle \right).$$

Hence, by Proposition 1

$$\frac{1}{1 + \gamma} \text{svd}_r(\gamma N + Z) \in \text{prox}_{\tilde{f}}(Z).$$

Next let $D^*$ and $M^*$ be solutions to (A) and (B), respectively. If the Douglas-Rachford iterations are applied to (B), then it follows by (25) that $Z^* = \gamma D^* + M^*$ is a limit point to (20).
Assuming that $\sigma_r(N + D^*) \neq \sigma_{r+1}(N + D^*)$, it holds that

$$\sigma_r(\gamma N + Z^*) = \sigma_r(\gamma (N + D^*) + M^*) = (1 + \gamma)\sigma_r(N + D^*) > (1 + \gamma)\sigma_{r+1}(N + D^*).$$

By the continuity of the singular values (see [53, Corollary 4.9]), this allows us to conclude that (23) applies in a sufficiently small neighborhood of $Z^*$. Then (24) implies that for all $Z$ within this neighborhood we get

$$\prox_{\nu f}(Z) = \prox_{\nu f}(Z),$$

where $f(M) := \frac{1}{2}\|M\|_F^2 - \langle N, M \rangle + \frac{1}{2}\|N\|_F^2$. Hence, the convex and non-convex Douglas-Rachford iterations locally coincide.

Furthermore, the Douglas-Rachford iterations cannot escape from this neighborhood, because the sequence $\|Z^* - Z^k\|_F$ of the convex Douglas-Rachford is known to be non-increasing (see [17]). Thus proving local convergence of the non-convex Douglas-Rachford when $\sigma_r(N + D^*) \neq \sigma_{r+1}(N + D^*)$.

Finally, notice that if there is a zero duality-gap in (C), then it follows by Proposition 6 and (25) that the convex and non-convex Douglas-Rachford have limit points that correspond to a solution to Problem 2, even if $\sigma_r(N + D^*) = \sigma_{r+1}(N + D^*)$. We will see in Subsection 5.2.2 that the non-convex Douglas-Rachford iterations can converge to these solutions. However, this may not be the case for all choices of $Z^0$.

5. Non-negative low-rank approximation. A particularly well studied low-rank approximation problem is the case of preserving non-negativity constraints.

**Problem 3.**

$$\min_{M \in \mathbb{R}^{n \times m}_{\geq 0}} \|N - M\|_F^2$$

subject to $M \in \mathbb{R}^{n \times m}_{\geq 0}$

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 2 with $g = \chi_{\mathbb{R}^{n \times m}_{\geq 0}}$. The examples shown here help to illustrate some of the elementary results that have been discussed in the previous sections.

5.1. Algorithms. Since non-negativity is a rather mild constraint, there are several algorithms available, a selection is summarized in the following.

5.1.1. Non-negative matrix factorization (NNMF). The probably most well-known approach to this problem is the so-called non-negative matrix factorization (see [5, 37]). Given $N \in \mathbb{R}^{n \times m}_{\geq 0}$ one intends to find a solution to

$$\min_{L \in \mathbb{R}^{n \times r}_{\geq 0}, R \in \mathbb{R}^{r \times m}_{\geq 0}} \|N - LR\|_F^2.$$

The rank constraint is explicitly taken into account by forming $LR$. To require both, $L$ and $R$ to be non-negative might be very conservative, since we only want the product $LR$ to be non-negative.

5.1.2. Alternating Non-negative Least-Squares (ANLS). NNMF is often solved by alternating projection algorithms such as alternating least-squares [37]. This can also be used for finding a non-negative approximation, i.e. given $N \in \mathbb{R}^{n \times m}_{\geq 0}$ and some $V_0 \in \mathbb{R}^{r \times m}_{\geq 0}$, one
interchangeably solves
\[ U_k := \arg\min_{U_{k-1} \in \mathbb{R}_{\geq 0}^{n \times m}} \| N - U V_{k-1} \|_F^2, \]
\[ V_k := \arg\min_{U_k V \in \mathbb{R}_{\geq 0}^{n \times m}} \| N - U_k V \|_F^2, \]

with \( k \geq 1 \). Numerical experiments indicate that this method also converges to optimality if \( \sigma_r(N + D^*) \neq \sigma_{r+1}(N + D^*) \). Moreover, there are examples where its solution attains the lower-bound of Proposition 2 even though \( \sigma_r(N + D) = \sigma_{r+1}(N + D) \). In our numerical examples this method often reproduces the same solutions as the non-convex Douglas-Rachford algorithm (see Subsection 4.4).

Note that alternating least-squares without constraints converges for almost all initial conditions to a standard SVD-approximation (see [52]). For other constraints than non-negativity, alternating least-squares may not be a feasible choice since it is often unclear how to choose \( V_0 \) or how to project onto the constraint set.

5.1.3. Lift-and-project Algorithm (LP). The idea of the so-called lift-and-project algorithm (see [13]) is to interchangeably perform a standard SVD-approximation of desired rank and project the result orthogonally onto the non-negative orthant, which again increases the rank. Eventually, this method has to converge since the Frobenius-norm is decreased in every step. Naturally, this algorithm always returns the standard SVD-approximation if it is non-negative. Unfortunately, it is difficult to know whether the solution converges to something useful or to zero (see [13]).

Finally note that this method is equivalent to applying the so-called backward-backward algorithm (see [14]) to the non-convex problem (26).

5.2. Examples. In the following we look at examples with a non-negativity constraint in order to illustrate several results that have been discussed throughout this work.

5.2.1. Image compression. A common example in the literature (see [2, 18]) 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 discussed in Subsection 5.1 to the Baboon-image in Figure 3a. A comparison among the relative errors of the methods as well as the normalized lower-bound obtained from (B), is shown in Figure 3b.

By the Perron-Frobenius Theorem (see [35, 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 this problem, since it usually cannot recover standard SVD-approximations. Moreover, we observe that 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, but 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.

5.2.2. Asymmetric optimal approximations. Let \( N \in \mathbb{S} \cap \mathbb{R}_{\geq 0}^{n \times n} \) and \( D^* \) be a solution of (A) corresponding to Problem 3. According to Proposition 4 and Proposition 7 we know that if \( \sigma_r(N + D^*) \neq \sigma_{r+1}(N + D^*) \), then \( \text{svd}_r(N + D^*) \in \mathbb{S} \) is the unique solution to (B)
and Problem 3. In the following we will see that preserving the symmetry may no longer be valid for an optimal non-negative approximation if \( \sigma_r(N + D^*) = \sigma_{r+1}(N + D^*) \).

Consider Problem 3 with \( r = 2 \) and symmetric

\[
N = \begin{pmatrix}
\frac{\sqrt{5} - 1}{2} & 1 & 3 \\
1 & 4 & 1 \\
3 & 1 & \frac{\sqrt{5} - 1}{2}
\end{pmatrix}.
\]

A non-symmetric solution to this is

\[
M^* = \begin{pmatrix}
0 & \frac{\sqrt{5} + 1}{2} & \frac{\sqrt{3} + 3}{2} \\
2 & 3 & \frac{\sqrt{3} + 1}{2} \\
2 & 2 & 0
\end{pmatrix}.
\]

Since \( N \) is symmetric, its singular values are given by the absolute value of its eigenvalues \( \{ \pm \frac{7-\sqrt{5}}{2}, 3 + \sqrt{5} \} \). Then with \( \|N - M^*\|_F = \frac{7-\sqrt{5}}{2} \) and Proposition 1 we conclude that \( M^* \) and \( M^{*T} \) are optimal non-negative rank-2 approximations of \( N \). Thus, by Proposition 5 it follows that \( D^* = 0 \) and \( \sigma_2(N + D^*) = \sigma_3(N + D^*) \). Furthermore, it implies that \( M^* \) and \( M^{*T} \) are solutions to (B).

Since the solution set of a convex problem is convex, all points \( \alpha M^* + (1 - \alpha)M^{*T} \) with \( \alpha \in [0, 1] \) must be solutions to (B). However, rank \( \{ \alpha M^* + (1 - \alpha)M^{*T} \} = 3 \) for all \( \alpha \in (0, 1) \). Thus we cannot expect to numerically find the rank-2 solutions by solving (B) (see Figure 2c).

In particular, let either of the discussed Douglas-Rachford algorithms (see Subsections 4.2 and 4.4) be initialized with \( Z^0 \in S \). Then Proposition 7 implies that they may converge to a symmetric solution, which can be shown to be non-optimal for Problem 3.

Nevertheless, it is interesting to note that under random initialization, NDR and ANLS often converge to an optimal solution.
6. 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 [8, 9, 10, 50, 59]). This leads to the following problem.

**Problem 4.**

\[
\begin{aligned}
\text{minimize} & \quad \text{rank}(M) \\
\text{subject to} & \quad m_{ij} = z_{ij}, \ (i, j) \in \mathcal{I}
\end{aligned}
\]

where \(\mathcal{I}\) is an index set.

One of the most popular methods for solving Problem 4 is the technique introduced in [9]. It states that if \(Z \in \mathbb{R}^{n \times n}\) then with high probability it is a solution to

\[
\begin{aligned}
\text{minimize} & \quad \|M\|_{1^*} \\
\text{subject to} & \quad m_{ij} = z_{ij}, \ (i, j) \in \mathcal{I}
\end{aligned}
\]

if \(\text{card}(\mathcal{I}) \geq Cn^{1.2}\text{rank}(Z)\log(n)\), where \(\text{card}(\mathcal{I})\) denotes the cardinality of \(\mathcal{I}\) and \(C\) is a constant. Similar to that, it has been shown in [50] that (28) 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 is a special case of Proposition 2 with \(r = 1\), because

\[
\begin{aligned}
\min_{M \in \mathbb{R}^{n \times n}} & \quad \frac{1}{2}\|M\|^2_F + g(M) \\
\text{subject to} & \quad m_{ij} = z_{ij}, \ (i, j) \in \mathcal{I}
\end{aligned}
\]

where \(g(M) = \chi_{\mathcal{M}}(M)\) and \(\mathcal{M} := \{M \in \mathbb{R}^{n \times n} : m_{ij} = z_{ij}, \ (i, j) \in \mathcal{I}\}\).

However, we suggest to keep the flexibility of \(r\) as a tuning parameter intact and consider instead

\[
\begin{aligned}
\text{minimize} & \quad \|M\|_{r^*} \\
\text{subject to} & \quad m_{ij} = z_{ij}, \ (i, j) \in \mathcal{I},
\end{aligned}
\]

where it is possible to sweep over real-valued \(r \geq 1\).

6.1. A motivational example. In the following we will see that \(r > 1\) may help to complete matrices where \(r = 1\) fails.

**Proposition 8.** Let \(Z \in \mathbb{R}^{n \times m}\) with \(r = \text{rank}(Z)\) and \(\mathcal{I} = \{(i, j) : z_{ij} \neq 0\}\). Then \(Z\) can be exactly recovered by determining a solution to (B), i.e. \(Z\) is the unique solution to

\[
\begin{aligned}
\text{minimize} & \quad \|M\|_{r^*} \\
\text{subject to} & \quad m_{ij} = z_{ij}, \ (i, j) \in \mathcal{I}
\end{aligned}
\]

**Proof.** The result follows by Corollary 1 and Proposition 3. \(\square\)

Proposition 8 may seem trivial, since we complete matrices with missing zeros only. However, we will show that the nuclear-norm heuristic in (28) cannot provide the same deterministic guarantee. To this end, consider the rank-2 matrices

\[
\begin{align*}
Z_1 &= \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, & Z_2 &= \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}, & Z_3 &= \begin{pmatrix} 0 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 4 \end{pmatrix}
\end{align*}
\]
We would like to recover these matrices under the assumption that the zero entries are the only unknown ones. Then it can be shown, e.g. by Proposition 7, that solving (28) is equivalent to determining

\[
\min_{t \in \mathbb{R}} \|Z_i(t)\|_{1^*}, \quad i = 1, 2, 3
\]

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 $\text{rank}(Z_1(t)) = 1$ if and only if $t = 1$. Hence, by (29) and Corollary 1 it follows that $Z_1(1)$ is the unique solution to (28), which implies that the nuclear-norm heuristic does not recover $Z_1$.

Next we show that non-uniqueness in (28) is another issue that can be avoided with the proposed approach in Proposition 8. Since $Z_2(t)$ is symmetric, it holds that

\[
\|Z_2(t)\|_{1^*} \geq \text{trace}(Z_2(t)) \equiv 5
\]

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 $t \in [0, 2]$. This implies that all of these points are solutions to the nuclear-norm heuristic (28). However, a numerical solver for (28) does not necessarily determine $Z_2$.

Finally, observe that the nuclear-norm heuristic does not always determine the lowest rank solution. It holds that $\text{rank}(Z_3(t)) \geq 2$ with equality if and only if $t = 0$. Moreover, it can be verified that $\|Z_3\|_{1^*} > \|Z_3(0.1)\|_{1^*}$. Thus $Z_3$ is not a solution to (28).

These examples show that additional knowledge about the true rank as well as the minimality in the Frobenius-norm (see (29)) can be utilized with $\|\cdot\|_{r^*}$ to possibly gain better completion. The following subsections will demonstrate the same behavior for a larger academic example and a practical application.

Finally, notice that in view of (4) one may also consider

\[
\text{minimize} \quad \frac{1}{2} \|M\|_F^2 + \mu \|M\|_{1^*}
\]

subject to

\[
m_{ij} = z_{ij}, \quad (i, j) \in \mathcal{I},
\]

where one sweeps over $\mu \geq 0$. Applied to the previous examples, this approach is also able to recover $Z_1$, $Z_2$ and $Z_3$ with $\mu = 0$. Nevertheless, we will see that generally there may not be any $\mu$ that leads to a solution of low rank.

### 6.2. Academic Example

The following example is mainly of academic interest and shows a comparison among (30) and (32). Moreover, we will see that even though Proposition 8 does not apply, our observations from above remain true.

Let $Z = \text{svd}_5(H)$ where $H \in \mathbb{R}^{10 \times 10}$ is a Hankel-matrix of the following structure

\[
H = \begin{pmatrix}
1 & 1 & \cdots & 1 & 1 \\
1 & & & & 0 \\
\vdots & & & & \vdots \\
1 & & & & 0 \\
1 & \cdots & 0 & 0 & 0
\end{pmatrix}.
\]
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 $M^*_r$ of (30) for different integer-valued $r$. The corresponding results for $M^*_\mu$, obtained by sweeping over $\mu \geq 0$ in (32), are presented in Figure 5.

The solution to the nuclear-norm heuristic $M^*_1$ ($r = 1$), gives the worst completion error and full rank. Notice that $n^{1/2}\text{rank}(Z)\log(n) \gg \text{card}(\mathcal{I}) = 22$ which is why the example does not lie within the scope of this method. In contrast, $M^*_5$ ($r = 5$) recovers the true matrix and is a sweet spot among all solutions. Furthermore, there is no $\mu$ such that $\text{rank}(M^*_\mu) < 10$.

![Figure 4: Relative completion error and ranks obtained with (30) for different values of $r$.](image1)

![Figure 5: Relative completion error and ranks obtained with (32) for different values of $\mu$.](image2)

### 6.3. Covariance completion.

Consider

$$\dot{x}(t) = Ax(t) + Bu(t),$$

where $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$ with $m \leq n$ and $u(t)$ is a zero-mean stationary stochastic process. For Hurwitz $A$ and reachable $(A,B)$ it has been shown (see [24, 25]) that the following are equivalent:
i. $X := \lim_{t \to \infty} \mathbb{E}(x(t)x^T(t)) \succeq 0$ is the steady-state covariance matrix of $x(t)$, where $\mathbb{E}(\cdot)$ denotes the expected value.

ii. $\exists H \in \mathbb{R}^{m \times n} : AX + XA^T = -(BH + H^T B^T)$.

iii. $\text{rank} \begin{pmatrix} AX + XA^T & B \\ B^T & 0 \end{pmatrix} = \text{rank} \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}$.

In particular, $H = \frac{1}{2} \mathbb{E}(u(t)u^T(t))B^T$ if $u$ is white noise. In [12, 40, 59, 60, 61] the problem of unknown $B$ and only partially known $X$ has been addressed as follows.

**Problem 5.**

\[
\begin{align*}
\text{minimize} & \quad \text{rank}(M) \\
\text{subject to} & \quad \hat{x}_{ij} = x_{ij}, \quad (i, j) \in \mathcal{I} \\
& \quad AX + \hat{X}A^T = -M \\
& \quad \hat{X} \succeq 0.
\end{align*}
\]

The problem has been solved in the same manner as Problem 4 i.e. by convexifying $\text{rank}(M)$ with $\|M\|_1 \ast$. However, since some practical examples only supply up to $2n$ known entries of specific structure (see [59, 60, 61]), it is not surprising that the quality of completion is often not satisfactory.

Instead, in [29] its generalization as in (30) is considered, i.e.

\[
\begin{align*}
\text{minimize} & \quad \|M\|_r \ast \\
\text{subject to} & \quad \hat{x}_{ij} = x_{ij}, \quad (i, j) \in \mathcal{I} \\
& \quad AX + \hat{X}A^T = -M \\
& \quad \hat{X} \succeq 0,
\end{align*}
\]

where it is possible to sweep over $r \geq 1$. Again, one may also consider

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2}\|M\|_F^2 + \mu \|M\|_1 \ast \\
\text{subject to} & \quad \hat{x}_{ij} = x_{ij}, \quad (i, j) \in \mathcal{I} \\
& \quad AX + \hat{X}A^T = -M \\
& \quad \hat{X} \succeq 0,
\end{align*}
\]

while sweeping over $\mu \geq 0$.

**6.3.1. 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

\[
\Delta T_{ij} \approx -\frac{1}{h^2} (4T_{ij} - T_{i+1,j} - T_{i,j+1} - T_{i-1,j} - T_{i,j-1}),
\]

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 receive a linear system

\[
\dot{x}(t) = \frac{1}{h^2} Ax(t) + \frac{1}{h^2} B\xi(t)
\]
where $A \in \mathbb{R}^{N^2 \times N^2}$ is the Poisson-matrix and $B = [b_{ij}] \in \mathbb{R}^{N^2 \times 4}$ with $b_{ij} = 0$ except for the following cases:

- $b_{i1} := 1$, for $i = 1, 2, \ldots, N$
- $b_{i2} := 1$, for $i = N, 2N, \ldots, N^2$
- $b_{i3} := 1$, for $i = N(N - 1) + 1, N(N - 1) + 2, \ldots, N^2$
- $b_{i4} := 1$, for $i = 1, N + 1, \ldots, N(N - 1) + 1$.

Moreover, let $\xi(t)$ be generated by a low-pass filtered white-noise signal $w(t)$ with unit covariance $E(w(t)w(t)^T) = I$ and

$$\dot{\xi}(t) = -\xi(t) + w(t).$$

As mentioned before, the extended covariance matrix

$$X_e := E(x_e x_e^T) = \begin{pmatrix} X & X_{x\xi} \\ X_{\xi x} & X_{\xi\xi} \end{pmatrix}$$

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 $A_e := \begin{pmatrix} A & B \\ 0 & -I \end{pmatrix}$, $B_e := \begin{pmatrix} 0 \\ I \end{pmatrix}$ 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 7a, where the black lines indicate the known entries. Figure 7b displays the relative completion error of the solutions obtained by (34) and (35) with dependency on $r$ and $\mu$. We observe that the error obtained by (34) with $r = 2$ 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 (35) (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.
Figure 7: Interpolated colormap of the steady-state covariance matrix and relative errors dependent on \( r \) and \( \mu \) obtained by (34) and (35).

(a) Solution obtained by (34) with \( r = 2 \).

(b) Solution obtained by (34) with \( r = 1 \).

(c) Solution obtained by (35) with \( \mu = 4.23 \).

(d) Solution obtained by (35) with \( \mu = 0 \).

Figure 8: Interpolated colormaps of the completed covariance matrices obtained by (34) and (35).
7. Hankel matrices. In the field of system and control, the rank of a Hankel operator/matrix is crucial, since it determines the complexity (order) of a linear system. By that it tells how costly it is to simulate a system or to implement a controller (see [2, 62]). For this reason, much focus was put into areas such as model order reduction. Even though the celebrated Adamyan-Arov-Krein Proposition (see [2, 47]) answers the question of optimal low-rank approximation of infinite dimensional Hankel operators, the following finite dimensional case still remains open.

**Problem 6.**

\[
\min_M \|N - M\|_F^2 \\
\text{subject to } \text{rank}(M) \leq r \\
M \in \mathcal{H},
\]

where \(N \in \mathcal{H} := \{H : H \text{ is Hankel}\}\).

Only the optimal rank-1 approximation in case of the spectral-norm \(\|\cdot\|_1\) has been determined in [1]. Moreover, for so-called linear externally positive systems the problem of non-negativity preserving Hankel-operator approximation has been considered in [27].

7.1. Academic Example. In the following we compare the \(r^*\)-norm framework with the regularization methods in Subsections 2.2 and 3.3 as well as the Lift-and-project algorithm from Subsection 5.1.3. To this end, let \(N \in \mathbb{R}^{10 \times 10}\) be the following Hankel matrix

\[
N = \begin{pmatrix}
1 & 2 & \cdots & 9 & 10 \\
2 & 9 \\
\vdots \\
9 & 2 \\
10 & 9 & \cdots & 2 & 1
\end{pmatrix}
\]

The relative errors together with the relative lower bound are shown in Figure 9a. For \(r = 1, \ldots, 4\) there is a zero duality gap and therefore the lower bound is achieved by the rank-regularization method (see Subsection 3.3) as well as the non-convex Douglas-Rachford and the \(r^*\)-norm. Moreover, even when Proposition 4 cannot guarantee a zero 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 \(\mu\), respectively. The dependency of the rank on these parameters is displayed in Figure 9b.

In contrast to the nuclear-norm regularization and the \(r^*\)-norm, that show the expected staircase behavior, the rank-regularization method seems to exhibit a non-intuitive oscillation, which complicates the search for an optimal \(\mu\), e.g. via a bisection algorithm.

8. Multivariate reduced-rank regression. In multivariate linear regression one wants to estimate a regression matrix \(C \in \mathbb{R}^{n \times m}\) assuming the underlying linear model

\[
Y = CX + E
\]

where \(Y \in \mathbb{R}^{n \times T}\) is a matrix with \(T\) measurements of \(n\) response variables, \(X \in \mathbb{R}^{m \times T}\) are the corresponding predictor variables and \(E \in \mathbb{R}^{m \times T}\) is Gaussian white-noise. Assuming that \(\text{rank}(X) = m < T\) one can determine the well-known least-squares estimator

\[
\hat{C} = YX^T(XX^T)^{-1},
\]
which is a minimizer of \( \min_{\mathcal{C}} \| Y - CX \|_F^2 \). Let \( \hat{c}_k \) and \( y_k \) denote the \( k \)-th row of \( \hat{C} \) and \( Y \), respectively, then

\[
\hat{c}_k = y_k X^T (X X^T)^{-1}
\]

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 \( \text{rank}(C) = r < \min\{m,n\} \) (see [36, 57]). 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 [57]). If \( X \) consists of information that is used to send \( T \) messages \( Y \) over \( r \) channels, then \( BX \) can be considered 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 7.**

\[
\begin{align*}
\minimize_{\tilde{C}} & \quad \| Y - CX \|_F^2 \\
\text{subject to} & \quad \text{rank}(C) \leq r.
\end{align*}
\]

Assuming that \( \text{rank}(X) = m < T \), an explicit solution can be determined as follows. Let \( X = U (\Sigma \ 0) \) \( (V_1 \ V_2)^T \) be an SVD of \( X \) with \( \Sigma \in \mathbb{R}^{m \times m} \), then

\[
\| Y - CX \|_F^2 = \| Y (V_1 \ V_2) - (CU \Sigma \ 0) \|_F^2 = \| Y V_1 - CU \Sigma \|_F^2 + \| Y V_2 \|_F^2.
\]

Hence, Problem 7 reduces to

\[
\begin{align*}
\minimize_{\tilde{C}} & \quad \| Y V_1 - \tilde{C} \|_F^2 \\
\text{subject to} & \quad \text{rank}(\tilde{C}) \leq r.
\end{align*}
\]
By Proposition 1 we know that a minimizer of (37) is given by $\text{svd}_r(YV_1)$ and therefore $\hat{C} = \text{svd}_r(YV_1)\Sigma^{-1}U^T$ is a solution to Problem 7. Observe that Problem 7 can also be stated as

$$\min_M \frac{1}{2} \|Y - M\|_F^2 + \chi_\mathcal{L}(M)$$

subject to $\text{rank}(M) \leq r$,

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 $\text{rank}(X) = m$ then $\text{rank}(M) = \text{rank}(C)$ and solving

$$\min_M \frac{1}{2} \|M\|_F^2 - \langle Y, M \rangle + \chi_\mathcal{L}(M)$$

leads to the same solution as above if $\text{svd}_r(YV_1)$ is unique. This can be shown by considering the dual of (38). By Proposition 2 we get

$$\max_D \frac{1}{2} \|Y + D\|_F^2$$

subject to $DX^T = 0$,

where $D^* = Y(V_1^TV_1 - I)$ is a feasible maximizer such that $\text{svd}_r(Y + D^*) = \text{svd}_r(YV_1V_1^T) = \hat{C}X$ is a solution to (38).

9. 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. Whereas factor based approaches such as ANLS are also regularization free, they depended on its initialization and therefore are less applicable for general convex constraints, which can be handled with regularization approaches. The $r*$-approach combines the benefits of both approaches. Moreover, we have seen that it can be turned into a regularization dependent method, where unlike other approaches the parameter has a clear relationship to the desired rank (see Subsection 3.2). As a result, a generalization of (28) to solve the matrix completion problem has been suggested. Furthermore, we have linked this approach to the rank-regularization method (see Subsection 3.3). Nevertheless, 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. Based on that it was possible to show that several other useful properties known from the SVD-solution may be preserved (see Proposition 7). 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 2. This work is merely a starting point to investigating its power for solving 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 Subsection 4.4 to the known local convergence results in the vector case (see [31]).

The numerical examples 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
It is known that \( f^{**} \) is called its conjugate (dual) functional. Further, the bi-conjugate functional of \( f \) is defined
\begin{equation}
(39)
\end{equation}

In the following we say that \( g \) is a symmetric gauge function if and only if
\begin{itemize}
  \item \( g(\cdot) \) is a norm.
  \item \( \forall x \in \mathbb{R}^n : g(|x|) = g(x) \), where \(|x|\) denotes the element-wise absolute value.
  \item \( g(Px) = g(x) \) for all permutation matrices \( P \) and all \( x \).
\end{itemize}

**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.

**A.2. Convex Optimization.** The following definitions and results from convex optimization (see [33, 42, 51]) are used to prove the main results. In the following we assume that all functionals are defined on a real Hilbert space \( X \) with inner product \( \langle \cdot, \cdot \rangle \). The domain of a functional \( f \) on \( X \) is defined as \( \text{dom} f := \{ x \in X : f(x) < \infty \} \).

**Definition A.1.** Let \( f : X \to \mathbb{R} \cup \{\infty\} \) be a functional with \( \text{dom} f \neq \emptyset \), minorized by an affine functional 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) functional. Further, the bi-conjugate functional of \( f \) is defined as \( f^{**} := (f^*)^* \).

**Definition A.2.** A convex functional \( f : X \to \mathbb{R} \cup \{\infty\} \) with \( \text{dom} f \neq \emptyset \) is
\begin{itemize}
  \item proper if \( \text{dom} f \neq \emptyset \).
  \item closed if the epigraph \( \{ x : f(x) \leq t, x \in \text{dom} f \} \) is a closed set for all \( t \in \mathbb{R} \).
\end{itemize}
It is known that \( f^{**} = f \) if only if \( f \) is a closed and proper convex functional.

**Lemma A.1.** Let \( f, g : X \to \mathbb{R} \cup \{\infty\} \) be functionals as in Definition A.1. Then
\begin{equation}
(39)
\end{equation}
PROPOSITION A.3. Let \( f, g : X \to \mathbb{R} \cup \{\infty\} \) be closed and proper convex functionals. Assume that \( \text{ri}(\text{dom} f) \cap \text{ri}(\text{dom} g) \neq \emptyset \) and \( \text{ri}(\text{dom} f^*) \cap \text{ri}(\text{dom} g^*) \neq \emptyset \), where \( \text{ri}(\cdot) \) denotes the relative interior. Then,

\[
\min_{x \in X} [f(x) + g(x)] = - \min_{x^* \in X^*} [f^*(x^*) + g^*(-x^*)].
\]

Moreover, the minimum on the left is attained at some \( x_0 \) and the minimum on the right by some \( x_0^* \) such that

\[
f^*(x_0^*) = \langle x_0, x_0^* \rangle - f(x_0),
g^*(-x_0^*) = \langle x_0, x_0^* \rangle - g(x_0).
\]

DEFINITION A.3. Let \( f : X \to \mathbb{R} \cup \{\infty\} \) be a functional. Then

\[
\partial f(x_0) := \{ x_0^* \in X : f(x) \geq f(x_0) + \langle x - x_0, x_0^* \rangle \}
\]

is called the subdifferential of \( f \) at \( x_0 \). Moreover, each \( x_0^* \in \partial f(x_0) \) is referred to as a subgradient of \( f \) at \( x_0 \).

PROPOSITION A.4. Let \( f : X \to \mathbb{R} \cup \{\infty\} \) be a closed and proper convex functional. 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. \( f(x_0) = \langle x_0, x_0^* \rangle - f^*(x_0^*) \).

For \( x \in \mathbb{R}^n \) and \( r \in [1, n] \) we define \( \|x\|_r := \sqrt{g_r(x)} \) with

\[
g_r(x) := \max\{x_{i_1}^2 + \cdots + x_{i_r}^2 + (r - |r|)x_{i_r} : 1 \leq i_1 < \cdots < i_r \leq n\}.
\]

The following Lemma on the subgradients of \( \| \cdot \|_r \) has been shown in [15] for \( r \in \mathbb{N} \). We simply extend it to the real-valued case.

LEMA A.2. Let \( r \in [1, n] \), \( \bar{r} := \lceil r \rceil \) and \( \sigma \in \mathbb{R}^r_{\geq 0} \) with

\[
\sigma_1 \geq \cdots > \sigma_{\bar{r}-1} = \cdots = \sigma_r = \cdots = \sigma_{r+t} \geq \cdots \geq \sigma_n,
\]

where \( t = \bar{r} - r \) if \( \sigma_1 = \sigma_r \) and \( \sigma_n = \sigma_{\bar{r}} \), respectively. Then \( \nu \in \partial\|\sigma\|_r \) if and only if

i. \( 1 \leq i \leq \bar{r} - t : \nu_i = \frac{\sigma_i}{\|\sigma\|_r} \),

ii. \( \bar{r} - t + 1 \leq i \leq \bar{r} + s : \nu_i = \frac{\sigma_{\bar{r}+s}}{\|\sigma\|_r} \), with \( 0 \leq \tau_1 \leq 1, \sum_{i=\tau_1+1}^{\bar{r}+s} \tau_i = t - \bar{r} + r \),

iii. \( \bar{r} + s + 1 \leq i \leq n : \nu_i = 0 \).

Proof. Let \( r \in [1, n] \) and \( \sigma \in \mathbb{R}^n_{\geq 0} \) as in (40). Then

\[
\|\sigma\|_r = \max_{\mathcal{J} \subset \{1, \ldots, n\}} g_{\mathcal{J}}(\sigma),
\]

where \( g_{\mathcal{J}}(\sigma) := \sqrt{\sum_{\mathcal{J} \cap \{1, \ldots, n\}} \sigma_i^2 + (r - |r|)\sigma_{\max(\mathcal{J})}^2} \) and \( \text{card}(\mathcal{J}) \) denotes the cardinality of \( \mathcal{J} \). Since \( \|\sigma\|_r \neq 0 \) it follows (see [34, Proposition 4.3.1]) that the subdifferentials of \( \| \cdot \|_r \) evaluated at \( \sigma \) are given by

\[
\partial\|\sigma\|_r = \text{conv} \{ \nabla g_{\mathcal{J}}(\sigma) : \mathcal{J} \subset \{1, \ldots, n\}, \text{card}(\mathcal{J}) = \bar{r}, g_{\mathcal{J}}(\sigma) = \|\sigma\|_r \},
\]

where \( \nabla \) denotes the gradient operator with respect to \( \sigma \). Next we determine the gradient at these points where \( \|\sigma\|_r = g_{\mathcal{J}}(\sigma) \). Then, by assumption (40) it holds that \( \{1, \ldots, \bar{r} - t\} \subset \mathcal{J} \) and therefore
\begin{itemize}
  \item $1 \leq i \leq \tilde{r} - t$: \( \frac{\partial g_{\mathcal{S}}(\sigma)}{\partial \sigma_i} = \frac{\sigma_i}{\|\sigma\|_r} \).
  \item $i \in \mathcal{S} \cap \{ \tilde{r} - t + 1, \ldots, \tilde{r} + s \} \setminus \max(\mathcal{S})$: \( \frac{\partial g_{\mathcal{S}}(\sigma)}{\partial \sigma_i} = \frac{\sigma_i}{\|\sigma\|_r} \).
  \item $i = \max(\mathcal{S})$: \( \frac{\partial g_{\mathcal{S}}(\sigma)}{\partial \sigma_i} = \frac{\bar{r} - |r|}{\|\sigma\|_r} \).
  \item $\tilde{r} + s + 1 \leq i \leq n$: \( \frac{\partial g_{\mathcal{S}}(\sigma)}{\partial \sigma_i} = 0 \).
\end{itemize}

Thus, by (41) it holds that $v \in \partial \|\sigma\|_r$ if and only if
\begin{enumerate}[i.]
  \item $1 \leq i \leq \tilde{r} - t$: $v_i = \frac{\sigma_i}{\|\sigma\|_r}$.
  \item $\tilde{r} - t + 1 \leq i \leq \tilde{r} + s$: $v_i = \sigma_i \frac{\sigma_i}{\|\sigma\|_r}$ with $0 \leq \sigma_i \leq 1$ and $\sum_{t = \tilde{r} - t + 1}^{\tilde{r} + s} \sigma_i = t - \tilde{r} + r$,
  \item $\tilde{r} + s + 1 \leq i \leq n$: $v_i = 0$.
\end{enumerate}
where the last part of the second condition follows from
\[
\sum_{i \in \mathcal{S}} \frac{\partial g_{\mathcal{S}}(\sigma)}{\partial \sigma_i} = (t - \tilde{r} + r) \frac{\sigma_i}{\|\sigma\|_r}.
\]
\[\square\]

From Lemma A.2 and [58, Theorem 2] the following Proposition follows in the same way as in [15] for $r \in \mathbb{N}$.

**Proposition A.5.** Let $A \in \mathbb{R}^{n \times m}$, $r \in [1, \min\{m, n\}]$ and $\tilde{r} := \lceil r \rceil$. Further, let an SVD of $A$ be given by $A = \sum_{i=1}^{\min\{m, n\}} \sigma_i u_i v_i^T$ with $\sigma_{\tilde{r} - t} \neq \sigma_{\tilde{r} - t + 1} = \cdots = \sigma_r = \cdots = \sigma_{\tilde{r} + s} \neq \sigma_{\tilde{r} + s + 1}$, where $t = \tilde{r}$ and $s = \min\{m, n\} - \tilde{r}$ if $\sigma_1 = \sigma_r$ and $\sigma_{\min\{m, n\}} = \sigma_r$, respectively. Then $M \in \partial \|A\|_r$ if and only if
\[
M = \frac{1}{\|A\|_r} \left( \sum_{i=1}^{t-1} \sigma_i u_i v_i^T + \sigma_r (u_{\tilde{r} - t + 1}, \ldots, u_{\tilde{r} + s}) (v_{\tilde{r} - t + 1}, \ldots, v_{\tilde{r} + s})^T \right),
\]
where $T \succeq 0$, $\|T\|_1 = t - \tilde{r} + r$ and $\|T\|_1 \leq 1$. In particular, if $\sigma_r \neq \sigma_{r+1}$ or $\sigma_r = 0$ then rank($M$) $\leq \tilde{r}$.

**A.3. Proof of Lemma 1.**

**Proof.** Let $1 \leq r \leq q := \min\{m, n\}$ and
\[
g(x_1, \ldots, x_q) := \|\text{diag}(x_1, \ldots, x_q)\|_r.
\]
Then $\| \cdot \|_r$ is a unitarily invariant norm by Proposition A.2, because $g$ is a symmetric gauge function. Let $M \in \mathbb{R}^{n \times m}$, then by Corollary A.1
\[
\|M\|_r^2 = \max\{ \langle M^T M, U PV \rangle : U \text{ and } V \text{ are unitary} \},
\]
with $P := \begin{pmatrix} I_r & 0 \\ 0 & 0_{m-r} \end{pmatrix}$. If $M^T M = \sum_{i=1}^{m} \sigma_i(M) u_i u_i^T$ we can define a projection $P_r := \sum_{i=1}^{r} u_i u_i^T$ such that $\|M\|_r^2 = \langle P_r, M^T M \rangle$.

Since $\| \cdot \|_{rs}$ inherits the unitary invariance, we have
\[
\|\Sigma\|_{rs} = \|\Sigma\|_{rs} \leq \max \{ \Sigma X \} \leq \max \sum_{\|X\|_r \leq 1} \sum_{i=1}^{q} \sigma_i(M) \sigma_r(X)
\]
\begin{align*}
&= \max \sum_{\|X\|_r \leq 1} \left[ \sum_{i=1}^{r} \sigma_i(M) \sigma_r(X) + \sigma_r(X) \sum_{i=r+1}^{q} \sigma_i(M) \right],
&\text{with } \Sigma := \text{diag}(\sigma_1(M), \ldots, \sigma_q(M)).
\end{align*}

The last inequality follows by Proposition A.1 and can be attained. Hence,
\[
\|M\|_{rs} = \max \sum_{\|X\|_r \leq 1} \sum_{i=1}^{q} \sigma_i(M) \sigma_r(X) \geq \max \sum_{\|X\|_r \leq 1} \sum_{i=1}^{r} \sigma_i(M) \sigma_i(X) = \sum_{i=1}^{r} \sigma_i^2(M),
\]
with equality if and only if \( \text{rank}(M) \leq r \).

**A.4. Derivation of** \( \text{prox}_{\frac{r}{2}\|\cdot\|^2}(\cdot) \).

\[
\text{prox}_{\frac{r}{2}\|\cdot\|^2}(Z) = \arg\min_X \left( \frac{r}{2}\|X\|^2_r + \frac{1}{2}\|X - Z\|^2_F \right),
\]

which is equivalent to

\[
X^* = \text{prox}_{\frac{r}{2}\|\cdot\|^2}(Z) \iff 0 \in \partial_X \left( \frac{r}{2}\|X\|^2_r + \frac{1}{2}\|X - Z\|^2_F \right)_{|X=X^*} \iff Z - X^* \in \gamma\|X^*\|_r \partial_X \|X\|_r|_{X=X^*}.
\]

Let \( \tilde{r} := \lceil r \rceil \) and an SVD of \( X^* \) be given by \( X^* = \sum_{i=1}^{\min\{m,n\}} \sigma_i(X^*)u_i v_i^T \) such that

\[
\sigma_{\tilde{r}-t}(X^*) > \sigma_{\tilde{r}-t+1}(X^*) = \cdots = \sigma_t(X^*) = \cdots = \sigma_{\tilde{r}+s}(X^*) > \sigma_{\tilde{r}+s+1}(X^*),
\]

where \( t = \tilde{r} \) and \( s = n - \tilde{r} \) if \( \sigma_1(X^*) = \sigma_\tilde{r}(X^*) \) and \( \sigma_{\min\{m,n\}}(X^*) = \sigma_{\tilde{r}}(X^*) \), respectively. Further, let \( U_2 := (u_{\tilde{r}-t+1}, \ldots, u_{\tilde{r}+s}) \) and \( V_2 := (v_{\tilde{r}-t+1}, \ldots, v_{\tilde{r}+s}) \). Then, by Proposition A.5

\[
Z = (1+\gamma) \sum_{i=1}^{\tilde{r}-t} \sigma_i(X^*)u_i v_i^T + \sigma_t(X^*)U_2 (I+\gamma T)V_2^T + \sum_{i=\tilde{r}+s+1}^{\min\{m,n\}} \sigma_i(X^*)u_i v_i^T
\]

with \( \|T\|_1 \leq 1, \|T\|_2 = t - \tilde{r} + r, T \succeq 0 \). Using [58, Theorem 2] it follows that \( Z \) has the same singular vectors as \( X^* \) and therefore \( T = \text{diag}(T_{\tilde{r}-t+1}, \ldots, T_{\tilde{r}+s}) \) can be chosen to be diagonal. This gives

i. \( 1 \leq i \leq \tilde{r} - t : \sigma_i(X^*) = \frac{1}{1+\gamma \sigma_i(Z)} \).

ii. \( \tilde{r} - t + 1 \leq i \leq \tilde{r} + s : \sigma_i(X^*) = \frac{1}{1+\gamma \sigma_i(Z)} \).

iii. \( \tilde{r} + s + 1 \leq i \leq \min\{m,n\} : \sigma_i(X^*) = \sigma_i(Z) \).

Hence, the main task is to determine \( s \geq 0, t \geq 1 \) and \( T \succeq 0 \) such that

\[
\sigma_{\tilde{r}}(X^*) = \frac{\sigma_{\tilde{r}-t+1}(Z)}{1+\gamma T_{\tilde{r}-t+1}} = \cdots = \frac{\sigma_{\tilde{r}+s}(Z)}{1+\gamma T_{\tilde{r}+s}},
\]

where

\[
\sum_{i=1}^{k} T_{\tilde{r} - t + i} = t - \tilde{r} + r \text{ and } T_{\tilde{r} - t + i} \leq 1, 1 \leq i \leq t + s
\]

and

\[
\sigma_{\tilde{r}}(X^*) > \frac{\sigma_{\tilde{r}-t+1}(Z)}{1+\gamma} > \sigma_{\tilde{r}+s+1}(Z).
\]

Next we will show how \( s, t \) and \( T \) can be determined inductively. Clearly, there exists \( T_{\tilde{r}}, \ldots, T_{\tilde{r}+s_0} \) for some \( s_0 \geq 0 \), fulfilling (42) and (43) with \( t = 1 \) and \( s = s_0 \). However, if

\[
\frac{1}{1+\gamma \sigma_{\tilde{r}-t-1}(Z)} \leq \frac{\sigma_{\tilde{r}}(Z)}{1+\gamma T_{\tilde{r}}},
\]

then requirement (44) is violated. Hence, \( t = 0 \) is not a feasible choice and we want to find the smallest possible \( \tilde{r} \) for which this requirement is met after constructing \( T \). Let us assume that with \( t = \tilde{r} - 1 \) and \( s = s_{\tilde{r}-1} \), there is no solution that satisfies all three conditions (42)–(44).
Then one can construct \( T_{\bar{r},t+1}, \ldots, T_{\bar{r},t+\tilde{t}} \) fulfilling (42) and (43) with \( t = \bar{t} \) and \( s = \tilde{t} \), as follows:

Let \( i \geq 2 \) and \( T_{\bar{r},t+1}^{(i-1)}, \ldots, T_{\bar{r},t+i-1}^{(i-1)} \leq 1 \) be determined such that

\[
\frac{\sigma_{\bar{r},t+1}(Z)}{1 + \gamma T_{\bar{r},t+1}^{(i-1)}} = \cdots = \frac{\sigma_{\bar{r},t+i-1}(Z)}{1 + \gamma T_{\bar{r},t+i-1}^{(i-1)}} = \sigma_{\bar{r},t+i}(Z) \quad \text{and} \quad \sum_{j=1}^{i-1} T_{\bar{r},t+j}^{(i-1)} < \bar{r} - \bar{t} + r.
\]

**Case 1:** Assume \( T_{\bar{t},t+i}^{(i)} \) is such that \( \sum_{j=1}^{i} T_{\bar{t},t+j}^{(i)} = \bar{t} = \bar{r} + r \) and

\[
\sigma_{\bar{r},t+i+1}(Z) < \frac{\sigma_{\bar{r},t+i+1}(Z)}{1 + \gamma T_{\bar{r},t+i}^{(i)}} = \frac{\sigma_{\bar{r},t+i}(Z)}{1 + \gamma T_{\bar{r},t+i}^{(i-1)}} \quad \text{for} \quad 1 \leq j \leq i - 1.
\]

Then, \( i = \tilde{t} \) and \( T_{\bar{r},t+i} = T_{\bar{t},t+i}^{(i)} \), \( 1 \leq j \leq i \), where

\[
(1 + \gamma T_{\bar{r},t+i-1}^{(i-1)})(1 + \gamma T_{\bar{r},t+i}^{(i)}) = 1 + \gamma T_{\bar{r},t+i}^{(i-1)} = 1 + \gamma \left( \bar{t} - \bar{r} + r - T_{\bar{r},t+i-1}^{(i)} - \sum_{j=1}^{i-2} T_{\bar{r},t+j}^{(i-1)} \right)
\]

yields that

\[
T_{\bar{r},t+i}^{(i)} = \frac{\bar{t} - \bar{r} + r - \sum_{j=1}^{i-1} T_{\bar{r},t+j}^{(i-1)}}{i + \gamma \sum_{j=1}^{i-1} T_{\bar{r},t+j}^{(i-1)}}
\]

**Case 2:** Assume that there exists \( T_{\bar{t},t+i}^{(i)} \) such that for all \( j \) with \( 1 \leq j \leq i - 1 \)

\[
\sigma_{\bar{r},t+i+1}(Z) = \frac{\sigma_{\bar{r},t+i+1}(Z)}{1 + \gamma T_{\bar{r},t+i}^{(i)}}, \quad \sigma_{\bar{r},t+i}(Z) = \frac{\sigma_{\bar{r},t+i}(Z)}{1 + \gamma T_{\bar{r},t+i}^{(i-1)}} \quad \text{for} \quad 1 \leq j \leq i - 1.
\]

Then \( i < \tilde{t} \) and we can set

\[
T_{\bar{r},t+i}^{(i)} = \gamma^{-1} \left( \frac{\sigma_{\bar{r},t+i}(Z)}{\sigma_{\bar{r},t+i+1}(Z)} - 1 \right).
\]

In both cases

\[
T_{\bar{t},t+i}^{(i)} = \gamma^{-1} \left( (1 + \gamma T_{\bar{r},t+i}^{(i-1)})(1 + \gamma T_{\bar{r},t+i}^{(i)}) - 1 \right), \quad 1 \leq j \leq i - 1.
\]

Eventually this procedure will find \( t,s \) and \( T \) that satisfy (42) – (44). Finally observe that

\[
(45) \quad \frac{\sigma_{\bar{r}}(Z)}{1 + \gamma} > \sigma_{\bar{r}+1}(Z) \quad \Rightarrow \quad s = 0
\]

in which case \( \text{rank}(X^*) = \bar{r} \) and only \( t \) has to be determined. If additionally \( \bar{r} = r \) then \( T \) is the identity matrix and finding \( t \) is redundant.
Algorithm 1: Determine $X = \text{prox}_\gamma \frac{1}{\|\cdot\|_2^2}(Z)$

1: **Input:** Let $\gamma, r > 0$ and $Z \in \mathbb{R}^{n \times m}$ be given and set $\tilde{r} = \lceil r \rceil$ and $s = t = \tilde{T}_r = 0$.
2: Let $Z = \sum_{i=1}^{\min\{m,n\}} \sigma_i(Z)u_i v_i^T$ be an SVD of $Z$.
3: while $(\tilde{r} > t \text{ AND } \sigma_{\tilde{r}-t}(Z) \leq \frac{(1 + \gamma)\sigma_{\tilde{r}+1}(Z)}{1 + \gamma + \tilde{T}_r})$ or $t = 0$ do
4: \hspace{1em} $T_{\tilde{T}_r-t} = 0$
5: \hspace{1em} $t = t + 1$
6: \hspace{1em} $k = \tilde{r} - t$
7: \hspace{2em} while $s \neq k$ do
8: \hspace{3em} $k = k + 1$
9: \hspace{4em} $T_k = \frac{t - \tilde{r} + r - \sum_{j=1}^{k-1} T_j}{t - k + \tilde{r} - \sum_{j=1}^{k-1} T_j}$
10: \hspace{4em} if $\frac{\sigma_k(Z)}{1 + \gamma T_k} \geq \sigma_{k+1}(Z)$ then
11: \hspace{5em} $s = k$
12: \hspace{4em} else
13: \hspace{5em} $T_k = \gamma^{-1} \left( \frac{\sigma_k(Z)}{\sigma_{k+1}(Z)} - 1 \right)$
14: \hspace{5em} end if
15: \hspace{4em} $i = \tilde{r} + t - 1$
16: \hspace{4em} while $i < k$ do
17: \hspace{5em} $T_i = \gamma^{-1}(1 + \gamma T_i)(1 + \gamma T_k) - 1)$
18: \hspace{5em} $i = i + 1$
19: \hspace{4em} end while
20: \hspace{2em} end while
21: \hspace{1em} end while
22: **Output:** $X = \frac{1}{1 + \gamma} \sum_{i=1}^{\tilde{T}_r-t} \sigma_i(Z)u_i v_i^T + \frac{\sigma_{\tilde{T}_r-t}(Z)}{1 + \gamma \tilde{T}_r} \sum_{i=\tilde{T}_r-t+1}^{\tilde{T}_r+t-1} u_i v_i^T + \sum_{i=\tilde{T}_r+t+1}^{\min\{m,n\}} \sigma_i(Z)u_i v_i^T$.

Acknowledgments. We thank Andrey Ghalchak for his useful comments and numerous counter-examples. All authors are members of the LCCC Linnaeus Center and the eL-LIIT Excellence Center at Lund University. The third author is financially supported by the Swedish Foundation for Strategic Research.

REFERENCES

[1] A. Antoulas, *On the approximation of Hankel matrices*, in Operators, Systems and Linear Algebra, Springer, 1997, pp. 17–22.
[2] A. C. Antoulas, *Approximation of large-scale dynamical systems*, vol. 6, SIAM, 2005.
[3] A. Argyriou, R. Foygel, and N. Srebro, *Sparse prediction with the k-support norm*, in Advances in Neural Information Processing Systems, 2012, pp. 1457–1465.
[4] C. Berge, *Topological Spaces: including a treatment of multi-valued functions, vector spaces, and convexity*, Courier Corporation, 1963.
[5] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, *Algorithms and applications for approximate nonnegative matrix factorization*, Computational Statistics & Data Analysis, 52 (2007), pp. 155–173.
[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends in Machine Learning, 3 (2011), pp. 1–122.
[7] S. P. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.
[8] E. J. Candes and Y. Plan, *Matrix completion with noise*, Proceedings of the IEEE, 98 (2010), pp. 925–936.
[9] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Foundations of Computational mathematics, 9 (2009), pp. 717–722.
[10] E. J. Candès and T. Tao, The power of convex relaxation: Near-optimal matrix completion, Information Theory, IEEE Transactions on, 56 (2010), pp. 2053–2080.

[11] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, Rank-sparcity incoherence for matrix decomposition, SIAM Journal on Optimization, 21 (2011), pp. 572–596.

[12] Y. Chen, M. R. Jovanovic, and T. T. Georgiou, State covariances and the matrix completion problem, in Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on, IEEE, 2013, pp. 1702–1707.

[13] M. T. Chu, R. E. Funderlic, and R. J. Plemmons, Structured low rank approximation, Linear algebra and its applications, 366 (2003), pp. 157–172.

[14] P. L. Combettes and J.-C. Pesquet, Proximal splitting methods in signal processing, in Fixed-point algorithms for inverse problems in science and engineering, Springer, 2011, pp. 185–212.

[15] X. V. Doan and S. Vavasis, Finding the largest low-rank clusters with Ky Fan 2-k-norm and l1-norm, SIAM J. Optim., 26 (2016), pp. 274–312.

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

[17] J. Eckstein and D. P. Bertsekas, On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators, Mathematical Programming, 55 (1992), pp. 293–318.

[18] L. Eldén, Matrix methods in data mining and pattern recognition, vol. 4, SIAM, 2007.

[19] A. Eriksson, T. Thanh Pham, T.-J. Chin, and I. Reid, The k-support Norm and Convex Envelopes of Cardinality and Rank, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3349–3357.

[20] M. Fazel, Matrix rank minimization with applications, PhD thesis, Stanford University, 2002.

[21] M. Fazel, H. Hindi, and S. P. Boyd, A rank minimization heuristic with application to minimum order system approximation, in American Control Conference, 2001. Proceedings of the 2001, vol. 6, IEEE, 2001, pp. 4734–4739.

[22] M. Freimer and G. S. Mudholkar, A class of generalizations of Hölder’s inequality, Lecture Notes-Monograph Series, (1984), pp. 59–67.

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

[24] T. T. Georgiou, Spectral analysis based on the state covariance: the maximum entropy spectrum and linear fractional parametrization, IEEE transactions on Automatic Control, 47 (2002), pp. 1811–1823.

[25] T. T. Georgiou, The structure of state covariances and its relation to the power spectrum of the input, Automatic Control, IEEE Transactions on, 47 (2002), pp. 1056–1066.

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

[27] C. Grussler and A. Rantzer, Modified balanced truncation preserving ellipsoidal cone-invariance, in 53rd IEEE Conference on Decision and Control, 2014, pp. 2365–2370.

[28] C. Grussler and A. Rantzer, On optimal low-rank approximation of non-negative matrices, in Decision and Control (CDC), 2015 IEEE 54th Annual Conference on, IEEE, 2015, pp. 5278–5283.

[29] C. Grussler, A. Zare, M. R. Jovanovic, and A. Rantzer, The use of the r* heuristic in covariance completion problems, Submitted to Decision and Control (CDC), 2016 IEEE 55th Annual Conference on, IEEE, 2016, http://www.ece.umn.edu/users/mihailo//papers/gruzarjovranCDC16.pdf.

[30] R. Hesse and D. R. Luke, Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems, SIAM Journal on Optimization, 23 (2013), pp. 2397–2419.

[31] R. Hesse, D. R. Luke, and P. Neumann, Alternating projections and Douglas-Rachford for sparse affine feasibility, Signal Processing, IEEE Transactions on, 62 (2014), pp. 4868–4881.

[32] N. J. Higham, Computing the nearest correlation matrix – a problem from finance, IMA journal of Numerical Analysis, 22 (2002), pp. 329–343.

[33] J.-B. Hiriart-Urruty and C. Lemarchal, Convex analysis and minimization algorithms. II., Advanced theory and bundle methods, Grundlehren der mathematischen Wissenschaften, Springer-Verlag, 1996.

[34] J.-B. Hiriart-Urruty and C. Lemaréchal, Convex analysis and minimization algorithms I: Fundamentals, vol. 305, Springer Science & Business Media, 2013.

[35] R. A. Horn and C. R. Johnson, Matrix Analysis, Cambridge University Press, 2013.

[36] A. J. Izenman, Reduced-rank regression for the multivariate linear model, Journal of multivariate analysis, 5 (1975), pp. 248–264.

[37] J. Kim and H. Park, Fast nonnegative matrix factorization: An active-set-like method and comparisons, SIAM Journal on Scientific Computing, 33 (2011), pp. 3261–3281.

[38] H. Lai, Y. Pan, C. Lu, Y. Tang, and S. Yan, Efficient k-support matrix pursuit, in Computer Vision–ECCV 2014, Springer, 2014, pp. 617–631.

[39] V. Larsson, C. Olsson, E. Bylow, and F. Kahl, Rank minimization with structured data patterns, in Computer Vision–ECCV 2014, Springer, 2014, pp. 250–265.

[40] F. Lin, M. R. Jovanovic, and T. T. Georgiou, An ADMM algorithm for matrix completion of partially
known state covariances, in Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on, IEEE, 2013, pp. 1684–1689.

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

[42] D. G. Luenberger, Optimization by vector space methods, John Wiley & Sons, 1968.

[43] I. Markovsky, Structured low-rank approximation and its applications, Automatica, 44 (2008), pp. 891–909.

[44] A. M. McDonald, M. Pontil, and D. Stamos, New perspectives on k-support and cluster norms, CoRR, abs/1512.08204 (2015), http://arxiv.org/abs/1512.08204.

[45] C. Olsson and M. Oskarsson, A convex approach to low rank matrix approximation with missing data, in Image Analysis, Springer, 2009, pp. 301–309.

[46] N. Parikh and S. P. Boyd, Proximal algorithms., Foundations and Trends in optimization, 1 (2014), pp. 127–239.

[47] J. R. Partington, An Introduction to Hankel Operators, vol. 13, Cambridge University Press, 1988.

[48] D. Peaucelle, D. Henrion, Y. Labit, and K. Taitz, Users guide for SEDUMI interface 1.04, LAAS-CNRS, Toulouse, (2002).

[49] H. M. Phan, Linear convergence of the Douglas-Rachford method for two closed sets. Available: http://arxiv.org/abs/1401.6509, Oct. 2014.

[50] B. Recht, M. Fazel, and P. A. Parrilo, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, SIAM Rev., 52 (2010), pp. 471–501.

[51] R. T. Rockafellar, Convex Analysis, no. 28, Princeton University Press, 1970.

[52] N. Srebro, T. Jaakkola, et al., Weighted low-rank approximations, in ICML, vol. 3, 2003, pp. 720–727.

[53] G. W. Stewart and J.-G. Sun, Matrix perturbation theory, Academic press, 1990.

[54] R. K. Sundaram, A first course in optimization theory, Cambridge university press, 1996.

[55] R. Tibshirani, Regression shrinkage and selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological), (1996), pp. 267–288.

[56] K.-C. Toh, M. J. Todd, and R. H. Tutuncu, SDPT3—a MATLAB software package for semidefinite programming, version 1.3, Optimization methods and software, 11 (1999), pp. 545–581.

[57] R. Velu and G. C. Reinse, Multivariate reduced-rank regression: theory and applications, vol. 136, Optimization methods and software, 11 (2013), pp. 515–520.

[58] G. A. Watson, Characterization of the subdifferential of some matrix norms, Linear algebra and its applications, 170 (1992), pp. 33–45.

[59] A. Zare, Y. Chen, M. R. Jovanović, and T. T. Georgiou, Low-complexity modeling of partially available second-order statistics via matrix completion, arXiv preprint arXiv:1412.3399, (2014).

[60] A. Zare, M. R. Jovanovic, and T. T. Georgiou, Alternating direction optimization algorithms for covariance completion problems, in American Control Conference (ACC), 2015, IEEE, 2015, pp. 515–520.

[61] A. Zare, M. R. Jovanovic, and T. T. Georgiou, Color of turbulence, arXiv preprint arXiv:1602.05105, (2016).

[62] K. Zhou, J. C. Doyle, K. Glover, et al., Robust and optimal control, vol. 40, Prentice hall New Jersey, 1996.