LUND UNIVERSITY

On optimal low-rank approximation of non-negative matrices

Grussler, Christian; Rantzer, Anders

Published in:
2015 IEEE 54th Annual Conference on Decision and Control (CDC)

DOI:
10.1109/CDC.2015.7403045

2015

Link to publication

Citation for published version (APA):
Grussler, C., & Rantzer, A. (2015). On optimal low-rank approximation of non-negative matrices. In *2015 IEEE 54th Annual Conference on Decision and Control (CDC)* (pp. 5278-5283). IEEE - Institute of Electrical and Electronics Engineers Inc.. https://doi.org/10.1109/CDC.2015.7403045

Total number of authors:
2

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Download date: 08. Dec. 2025

2015 IEEE 54th Annual Conference on Decision and Control (CDC)
December 15-18, 2015. Osaka, Japan

On optimal low-rank approximation of non-negative matrices

Christian Grussler and Anders Rantzer

Abstract— **For low-rank Frobenius-norm approximations of matrices with non-negative entries, it is shown that the Lagrange dual is computable by semi-definite programming. Under certain assumptions the duality gap is zero. Even when the duality gap is non-zero, several new insights are provided.**

I. INTRODUCTION

Low-rank approximation of matrices and more generally of Hankel-Schmidt operators is considered an established tool in many areas such as image analysis, model order reduction or system identification (cf. [1], [3], [5], [12], [13], [15]). The main idea is to truncate less important parts in order to find an underlying structure of a given data matrix. For unitarily invariant norms the optimal low-rank approximation problem can be solved by performing a singular value decomposition (SVD) (see Section III).

However, these solutions usually do not fulfil any further desired structural constraints such as non-negativity, Hankel-structure, etc. (cf. [2], [3], [4], [8], [12], [13], [15]). Only in a few known cases an explicit solution to the constrained low-rank approximation problem can be determined e.g. Hankel-norm approximation (cf. [1], [14]). To this end, new concepts based on convex optimization have been developed (cf. [3], [5], [13], [15]), many among them relying on the so-called nuclear-norm heuristic which allows to incorporate any convex constraint (see also Section VI). In general, solutions determined by this heuristic are satisfactory, however the question of optimality remains and has been addressed e.g. in [15] for minimum rank solutions among affine constraints.

In previous work [7], the authors have developed a model reduction method which preserves input-output positivity i.e. the non-negativity of the approximating Hankel-operator. This work intends to address the problem of approximating a general non-negative matrix under the preservation of non-negativity (see Problem 2 and Section IV).

Besides the nuclear-norm heuristic, the most well-known approach to this problem is the so-called non-negative matrix factorization (NNMF) (see Section VI). Since each of the presented methods in this work has its benefits and downsides, we will conclude this work with a comparison between them in Sections VII and VIII. Nevertheless, all available solution approaches can only guarantee a local optima.

In contrast, we will show that a globally optimal solution to our non-convex problem can often be determined by convex optimization. (see Sections IV and V). As a byproduct of our derivations we will prove that this is also true for the unconstrained problem. Moreover, we derive several notable generalizations of concepts which are often used in the context of the nuclear-norm heuristic (see Section II and V). Finally, in Section VIII we are going to observe that the so-called alternating least squares algorithm (see Section VI) usually converges to an optimal solution.

II. PRELIMINARIES

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

If $X = X^T$, then we write $X \succ 0$ ($X \succeq 0$) if X is positive definite (semi-definite). We also use these notations to describe the relation between two matrices, e.g. $A \succeq B$ means $A - B \succeq 0$. Furthermore, we denote by $\sigma_1(X) \geq \cdots \geq \sigma_{\min\{m,n\}}(X)$ the non-increasingly ordered singular values of X counted with multiplicity. $\mathbb{R}^{n\times m}$ is equipped with the Frobenius inner product $\langle X,Y\rangle := \sum_{i=1}^{m}\sum_{j=n}^{n} x_{ij}y_{ji} = \text{trace}(X^TY),\ X,Y \in \mathbb{R}^{n\times m}$ and Frobenius-norm

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

The Frobenius-norm is unitarily invariant, i.e. $\|TXU\|_F = \|X\|_F$ for all unitary $T \in \mathbb{U}_n$ and $U \in \mathbb{U}_m$. A complete characterization of all unitarily invariant norms is given in [9]. However, this work only considers the norms which are found in the following Lemma, where $\mathscr{P}_r$ stands for the set of all orthogonal projections of rank r.

Lemma 1: Let $M \in \mathbb{R}^{n\times m}$ and $1 \leq r \leq q := \min\{m,n\}$, then

$$\|M\|_r := \sqrt{\sum_{i=1}^{r} \sigma_i^2(M)} = \sqrt{\max_{P\in\mathscr{P}_r} \langle P, M^TM\rangle} \tag{1}$$

is a unitarily invariant norm with dual-norm

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

Moreover,

- $\|M\|_1 \leq \cdots \leq \|M\|_q = \|M\|_F = \|M\|_{q*} \leq \cdots \leq \|M\|_{1*}$
- $\text{rank}(M) \leq r$ if and only if $\|M\|_r = \|M\|_F = \|M\|_{*r}$
- If $\text{rank}(M) > r$ then $\|M\|_{r*} > \|\hat{M}\|_r$ for any sub-matrix $\hat{M} \in \mathbb{R}^{p\times q}$ of M obtained by deleting columns and rows.

A proof of this lemma can be found in the appendix. Notice, $\|M\|_1 = \sigma_1(M)$ equals the spectral norm and its dual norm $\|M\|_{1*} = \sum_{i=1}^{\min\{m,n\}} \sigma_i(M)$ equals the Nuclear-norm.

This work was supported by the ELLIIT Excellence Center and by the Swedish Research Council through the LCCC Linnaeus Center. The authors are members of the LCCC Linnaeus Center and the eLLIIT Excellence Center at Lund University.

The authors are with the Department of Automatic Control, Lund University, Box 118, 22100 Lund, Sweden `{christiang,rantzer}@control.lth.se`

978-1-4799-7886-1/15/$31.00 ©2015 IEEE

It is well known that these norms can be reformulated into semi-definite programs (SDPs) (cf. [5], [15]). In Section V we will see that the same holds true for all $\|\cdot\|_r$ as for their duals.

III. LOW-RANK APPROXIMATION

Now, 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 $A \in \mathbb{R}^{n\times m}$ and $1 \leq r \leq \min\{m,n\}$, find $X^* \in \mathbb{R}^{n\times m}$ with $\operatorname{rank}(X^*) \leq r$ such that

$$\inf_{\operatorname{rank}(X)\leq r} \|A-X\| = \|A-X^*\|$$

for some given operator norm $\|\cdot\|$.
In case of the Hilbert-Schmidt norm, which is the natural operator generalization of the Frobeninus-norm, the problem has been solved by Schmidt and was generalized by Mirsky to unitarily invariant norms (cf. [1])

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

$$\inf_{\operatorname{rank}(X)\leq r} \|A-X\| = \|\operatorname{diag}(\sigma_{r+1}(A),\dots,\sigma_n(A))\|,$$

for any unitarily invariant norm $\|\cdot\|$.
Hence, if an SVD of A is given by $A = \sum_{i=1}^{\min\{m,n\}} \sigma_i u_i v_i^T$, then an optimal solution is $X^* = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, which we will 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 $\sigma_r(A) = \sigma_{r+1}(A)$. Nevertheless, if the chosen norm is the Frobenius-norm and $\sigma_r(A) \neq \sigma_{r+1}(A)$, then there is a unique solution.

The main problem of this work is the following:

Problem 2: Given $N \in \mathbb{R}^{n\times m}_{\geq 0}$ and $1 \leq r \leq \min\{m,n\}$ find $M^* \in \mathbb{R}^{n\times m}_{\geq 0}$ with $\operatorname{rank}(M^*) \leq r$ such that

$$\inf_{\substack{M\in\mathbb{R}^{n\times m}_{\geq 0},\\ \operatorname{rank}(M)\leq r}} \|N-M\|_F = \|N-M^*\|_F.$$

Clearly, Problem 1 and 2 are non-convex due to the rank constraint. Nevertheless, we will see in the following two sections that both problems can often be solved by convex optimization.

IV. MAIN RESULT

Problem 2 is usually solved by approximating the optimal solution through heuristics (see Section VI). In the following we want to elaborate on the *optimal* solution.

Here is our main result:

Theorem 1: Let $N \in \mathbb{R}^{n\times m}_{\geq 0}$ then

$$\inf_{\substack{M\in\mathbb{R}^{n\times m}_{\geq 0},\\ \operatorname{rank}(M)\leq r}} \|N-M\|_F^2 \geq \max_{D\in\mathbb{R}^{n\times m}_{\geq 0}} \|N\|_F^2 - \|N+D\|_r^2 \qquad (2)$$

If the maximum on the right is achieved by $D^* \in \mathbb{R}^{n\times m}_{\geq 0}$ and $\sigma_r(N+D^*) \neq \sigma_{r+1}(N+D^*)$, then the infimum on the left equals the maximum on the right. Moreover, the minimizer on the left is uniquely determined by the unique optimal rank r Frobenius-norm approximation of $N+D^*$.

Two different proofs of this theorem can be found in the appendix. The second one relies mostly on differentiability of $\|N+D^*+D\|_r^2$ in D, which breaks down if $\sigma_r(N+D^*) = \sigma_{r+1}(N+D^*)$. The first proof is more revealing, because it shows, if $\sigma_r(N+D^*) = \dots = \sigma_k(N+D^*) > \sigma_{k+1}(N+D^*)$, then the standard rank-r SVD-approximation of $N+D^*$ is non-negative. Moreover, the proofs do not rely on the fact that $N \in \mathbb{R}^{n\times m}_{\geq 0}$. However, if $N \in \mathbb{R}^{n\times m}_{\geq 0}$ then there is a significantly higher chance for $\sigma_r(N+D^*) > \sigma_{r+1}(N+D^*)$ to hold as we will see in Section VIII.

V. EQUIVALENT REFORMULATIONS

Let us derive several reformulations of Theorem 1 in order to gain insightful geometric ideas and to prove computability.

A. SDP-reformulations

We start with an SPD-reformulation of

$$\min_{D\in\mathbb{R}^{n\times m}} \|N+D\|_r^2. \qquad (3)$$

Let $T \succeq (N+D)(N+D)^T$ and $q := \min\{m,n\}$, then

$$\|N+D\|_r^2 \leq \operatorname{trace}(T) - \sum_{i=r+1}^{q} \sigma_i(T) \leq \operatorname{trace}(T) - (q-r)\sigma_q(T).$$

Hence,

$$\|N+D\|_r^2 \leq \min_{T\succeq(N+D)(N+D)^T} \operatorname{trace}(T) - (q-r)\sigma_q(T)$$

and equality can evidently be achieved. Using the Schur-complement (cf. [9]) shows that (3) is equivalent to

$$\begin{array}{ll} \text{minimize} & \operatorname{trace}(T) - \gamma(\min\{m,n\} - r) \\ \text{subject to} & \begin{pmatrix} T & N+D \\ (N+D)^T & I \end{pmatrix} \succeq 0,\ T \succeq \gamma I,\ D \in \mathbb{R}^{n\times m}_{\geq 0}. \end{array}$$

By taking the Lagrange dual of this expression we can also determine the solution to its dual

$$\begin{array}{ll} \text{minimize} & \operatorname{trace}(W) - 2\operatorname{trace}(N^T M) \\ \text{subject to} & \begin{pmatrix} I-P & M \\ M^T & W \end{pmatrix} \succeq 0,\ M \in \mathbb{R}^{n\times m}_{\geq 0}, P \succeq 0, \\ & \operatorname{trace}(P) = \min\{n,n\} - r. \end{array}$$

Under the assumptions of Theorem 1, an optimal solution M^* to the dual problem is equal to the unique optimal solution to Problem 2, which does not require a standard SVD-approximation of $N+D^*$.

B. Gauge-dual

If D^* is an optimal solution to $\min_{D\in\mathbb{R}^{n\times m}_{\geq 0}} \|N+D\|_r$ then by Proposition 5 in the appendix

$$\|N+D^*\|_r = \max_{\substack{M_d\in\mathbb{R}^{n\times m}_{\geq 0}\\ \|M_d\|_{r*}\leq 1}} \langle N, M_d\rangle.$$

This can be reformulated into

$$\frac{1}{\|N+D^*\|_r} = \min_{\substack{M_d\in\mathbb{R}^{n\times m}_{\geq 0}\\ \langle N,M_d\rangle=1}} \|M_d\|_{r*}. \qquad (4)$$

The right-side of (4) is sometimes referred to as a gauge dual to $\min_{D\in\mathbb{R}^{n\times m}_{\geq 0}} \|N+D\|_r$ (cf. [6]). Consequently, the optimal solution to Problem 2 can be found by studying the set $B_{r*}\cap H\cap \mathbb{R}^{n\times m}_{\geq 0}$ where

$$\begin{aligned} B_{r*} &:= \{X : \|X\|_{r*} \leq \frac{1}{\|N+D^*\|_r}\}, \\ H &:= \{X : \langle N,X\rangle = 1\}. \end{aligned}$$

Theorem 1 states, if $\sigma_r(N+D^*) > \sigma_{r+1}(N+D^*)$ then $B_{r*}\cap H\cap \mathbb{R}^{n\times m}_{\geq 0}$ consists of a single element. This can also be understood geometrically with the help of the following lemma.

Lemma 2:

$$\{X\in\mathbb{R}^{n\times m} : \|X\|_{r*}\leq 1\} = \operatorname{conv}(K)$$

with $K := \{X\in\mathbb{R}^{n\times m} : \|X\|_r = 1, \operatorname{rank}(X)\leq r\}$ and $\operatorname{conv}(\cdot)$ denoting the convex hull.

Proof: By Lemma 1 we know that $\operatorname{conv}(K)\subset\{X\in\mathbb{R}^{n\times m} : \|X\|_{r*}\leq 1\}$. Moreover, by first assertion in Lemma 1

$$\sup_{M\in\operatorname{conv}(K)} \langle N,M\rangle = \|N\|_r = \sup_{\|M\|_{r*}\leq 1} \langle N,M\rangle$$

for all $N\in\mathbb{R}^{n\times m}$. However, since both sets are closed, equality holds if and only if $\{X\in\mathbb{R}^{n\times m} : \|X\|_{r*}\leq 1\} = \operatorname{conv}(K)$ by Theorem 17.1 in [16]. ■

This gives us the geometric interpretation that the hyperplane H intersects $B_{r*}\cap\mathbb{R}^{n\times m}_{\geq 0}$ at a non-negative extremal point of B_{r*}. Hence, $\sigma_r(N+D^*) = \sigma_{r+1}(N+D^*) > 0$ can occur if and only if H intersects $B_{r*}\cap\mathbb{R}^{n\times m}_{\geq 0}$ at several points, i.e. either the solution to Problem 2 is non-unique or there is a duality-gap in (2). Furthermore, it is interesting to notice that neglecting the non-negativity in (4) leads to the following theorem.

Theorem 2: Let $A\in\mathbb{R}^{n\times m}$, $1\leq r\leq\min\{m,n\}$ and X^* be a solution to the convex problem

$$\min_{\langle A,X\rangle = 1} \|X\|_{r*}.$$

If $\sigma_r(A) > \sigma_{r+1}(A)$, then $\operatorname{rank}(X^*) = r$ and there exists $c>0$ such that

$$\inf_{\operatorname{rank}X = r} \|A-X\| = \|A-cX^*\|$$

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

This implies, if $N\in\mathbb{R}^{n\times m}_{\geq 0}$ and $\sigma_r(N) > \sigma_{r+1}(N) = 0$, then the gauge dual (4) returns the solution $\frac{N}{\|N\|_F^2}$ and it holds that one can choose $D^* = 0$.

VI. OTHER METHODS

In order to put our result in the context of earlier work, we recall two of the most commonly used solution approaches to Problem 2. Additionally, we consider two other heuristics which appear to work very well for this particular problem.

A. Non-negative matrix factorization

The first solution approach which comes to ones mind is the well-known non-negative matrix factorization (NNMF), i.e. given $N\in\mathbb{R}^{n\times m}_{\geq 0}$ find

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

Requiring non-negative factors is a much stronger assumption than our original problem. In contrast to Theorem 1, it is mostly unknown how to determine an optimal solution for this problem. Moreover, even if the standard SVD-approximation of N is non-negative, it does not necessarily have a NNMF. In addition, all algorithms depend on a choice of initialization.

B. Convex relaxation

The second approach borrows techniques from sparse regularized regression or Lasso (cf. [17]), which aims to estimate a sparse solution $\hat{x}$ to a linear system of equations $A\hat{x}\approx b$ by solving

$$\min_x \frac{1}{2}\|Ax-b\|^2 + \gamma\|x\|_1,$$

where $\|\cdot\|$ can be any norm, $\|x\|_1 = \sum_{i\geq 1}|x_i|$ and $\gamma > 0$. Self-evidently, sparsity of the singular values is equivalent to a low rank. Hence, for given $N\in\mathbb{R}^{n\times m}$, its matrix version reads

$$\min_M \frac{1}{2}\|N-M\|^2 + \gamma\|M\|_{1*}, \tag{5}$$

where in this paper $\|\cdot\| = \|\cdot\|_F$. Obviously, this formulation allows to add any convex constraint such as non-negativity of M, which is why variants of this approach have been used extensively (cf. [5], [13], [15]). The limiting factor here is the need to know γ a priori – whereas a large γ decreases the rank too much, a small γ may leave it too large. Moreover, in order to find the best approximation, one usually likes to keep γ as small as possible, which on the other hand could end up in a costly search, since each optimization requires to solve an SDP.

Again, even without any further constraints and the smallest possible γ, this heuristic usually does not return a solution which is comparable to the standard SVD-approximation.

C. Lift-and-project Algorithm

The idea behind the so-called lift-and-project algorithm 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. Unlike the previous methods, this algorithm will always return the standard SVD-approximation if it is non-negative. Unfortunately, it is difficult to prove whether the final result will be non-zero. In our numerical experiments we never encountered zero results.

D. Alternating Non-negative Least-Squares

NNMF is often solved by alternating projection algorithms such as alternating non-negative least-squares [11]. This can also be used for our problem, i.e. given $N \in \mathbb{R}_{\geq 0}^{n\times m}$ and some $V_0 \in \mathbb{R}_{\geq 0}^{r\times n}$, one interchangeably solves

$$\begin{aligned} U_k &:= \underset{UV_{k-1}\in\mathbb{R}_{\geq 0}^{n\times m}}{\operatorname{argmin}} \|N - UV_{k-1}\|_F^2, \\ V_k &:= \underset{U_kV\in\mathbb{R}_{\geq 0}^{n\times m}}{\operatorname{argmin}} \|N - U_kV\|_F^2, \end{aligned}$$

where $k \geq 1$. Unlike NNMF, numerical experiments indicate a relatively fast convergence. The limit points usually approach the global optimum as derived in Theorem 1. Moreover, there are examples where its solution fulfils the lower-bound of Theorem 1 even though $\sigma_r(N+D_*) = \sigma_r(N+D_*)$. The authors hope to give a mathematical analysis for this in a future publication.

VII. DISCUSSION

In the previous sections we have discussed several solution approaches to determine

$$\inf_{\substack{M\in\mathbb{R}_{\geq 0}^{n\times m},\\ \operatorname{rank}(M)\leq r}} \|N - M\|_F.$$

Whereas our method only provides a solution of desired rank if the requirements of Theorem 1 are fulfilled, the heuristic methods always obtain a suboptimal solution. Fortunately, numerical experiments with randomly generated non-negative matrices indicate that the requirements of Theorem 1 are usually met and a guaranteed optimal solution can be found. According to Theorem 2, this also includes most of the non-negative standard SVD-approximations which not necessarily can be recovered by the first two heuristics.

Nevertheless, as mentioned in Section V, $\sigma_r(N+D^*) = \sigma_{r+1}(N+D^*)$ may imply that there is a non-unique solution to Problem 2. Indeed, one can readily find such examples e.g. among non-negative symmetric matrices, which have an optimal non-negative low-rank approximation that is not symmetric. Therefore, it would be interesting to extend Theorem 1 to unitarily invariant norms which benefit from multiple singular values. Further notice that our method mainly relies on solving a single SDP, whereas the heuristic in (5) requires the solution of several SDPs. Computationally wise, the alternating projection based methods are the most efficient approaches discussed here.

Finally, Lemma 2 and Theorem 2 are similar to some of the ideas in [15], where one tries to find the lowest rank solution among linear constrains, i.e. minimization of $\|\cdot\|_{1*}$. Instead, if one can afford a solution of certain rank r, then Theorem 2 and Lemma 2 suggest to minimize over $\|\cdot\|_{r*}$. This relation may also explain why our method seems to work for all unstructured, randomly generated examples as well as why the alternating least squares approach converges to optimality. Together, both methods seem to supply good lower and upper bounds on our problem, even in case of a duality-gap.

VIII. EXAMPLE

Now, let us look at the performance of the discussed methods based on an example. In our comparison we use the method given in [11]) for the NNMF, which is known to perform very well. Moreover, we give the convex relaxation algorithm (Lasso) the benefit of a nearly optimal γ. As mentioned in the discussion, for randomly generated dense matrices there appears little chance to get $\sigma_r(N+D^*) = \sigma_{r+1}(N+D^*)$, therefore we consider an example where this fact is present intentionally. To this end we choose the following random binary matrix

$$N = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

for which any of the standard SVD-approximations of rank $r > 1$ has $7\% - 30\%$ negative elements and therefore a simple projection onto $\mathbb{R}_{\geq 0}^{10\times 10}$ always increases the rank. The results of the discussed methods are summarized in Fig. 1.

By the Frobenius-Perron Theorem (cf. [9]) the standard SVD-approximation of rank 1 always has a non-negative factorization, which reveals the clear downside of the nuclear-norm heuristic for this problem. Furthermore, we observe that there has been a double singular value of $N+D^*$ in Theorem 1 for $r=2$ and $r=5$ – therefore no optimal solution is obtained by the SDP-formulations. However, note that the iterative non-negative least-squares algorithm always reproduces the optimal solution and even fulfils the lower bound in case of $r=2$ and $r=5$. Finally, we observe that though the standard SVD-approximations are far away from being non-negative, their relative errors are very close to those of the optimal non-negative approximations.

IX. CONCLUSION

In this work, a new method to determine optimal non-negative low-rank approximations of non-negative matrices is presented. It appears that the non-negativity constraint has little effect on the approximation error. Additionally, our result supplies a lower bound on the non-negative matrix factorization problem, which makes it a useful tool besides the usual benchmark test for NNMF-algorithms. Apart from the main result, this work has presented several new insights into the convexification of low-rank approximation problems.

REFERENCES

[1] A. C. Antoulas, *Approximation of large-scale dynamical systems*. SIAM, 2005, vol. 6.

[2] 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*, vol. 52, no. 1, pp. 155–173, 2007.

[3] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, “Rank-sparsity incoherence for matrix decomposition,” *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

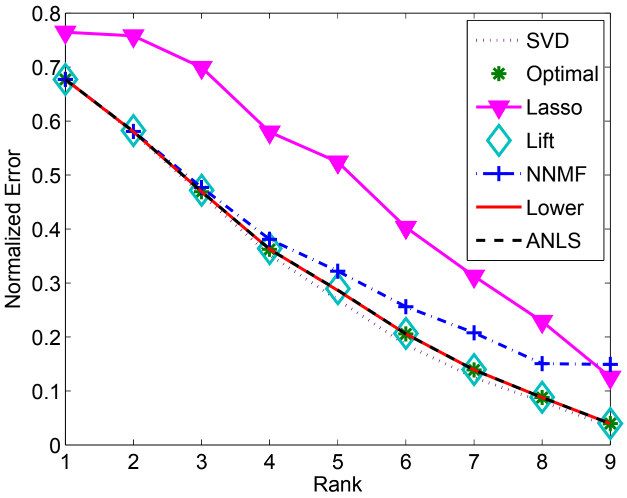

Fig. 1: Normalized $\|\cdot\|_F$-error
SVD: Standard SVD-approximation (without sign-constraints)
Optimal: Solution by Theorem 1
Lasso: Convex Relaxation
Lift: Lift-and-project Algorithm
NNMF: Non-negative matrix factorization
Lower: Lower bound by Theorem 1
ANLS: Alternating non-negative least-squares

[4] M. T. Chu, R. E. Funderlic, and R. J. Plemmons, “Structured low rank approximation,” *Linear algebra and its applications*, vol. 366, pp. 157–172, 2003.

[5] 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.

[6] R. M. Freund, “Dual gauge programs, with applications to quadratic programming and the minimum-norm problem,” *Mathematical Programming*, vol. 38, no. 1, pp. 47–67, 1987.

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

[8] N. J. Higham, “Computing the nearest correlation matrix – a problem from finance,” *IMA journal of Numerical Analysis*, vol. 22, no. 3, pp. 329–343, 2002.

[9] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2012.

[10] T. Kato, *A short introduction to perturbation theory for linear operators*. Springer, 1982.

[11] J. Kim and H. Park, “Fast nonnegative matrix factorization: An active-set-like method and comparisons,” *SIAM Journal on Scientific Computing*, vol. 33, no. 6, pp. 3261–3281, 2011.

[12] I. Markovsky, “Structured low-rank approximation and its applications,” *Automatica*, vol. 44, no. 4, pp. 891–909, 2008.

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

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

[15] B. Recht, M. Fazel, and P. A. Parrilo, “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,” *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

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

[17] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

X. APPENDIX

A. *Unitarily invariant norms*

The following results can be found e.g. in [9].

Proposition 2: Let $A,B\in\mathbb{R}^{n\times m}$, then

$$\langle A,B\rangle \le \sum_{i=1}^{\min\{m,n\}} \sigma_i(A)\sigma_i(B).$$

Corollary 1: Let $A,B\in\mathbb{R}^{n\times m}$ then

$$\sum_{i=1}^{\min\{m,n\}} \sigma_i(A)\sigma_i(B) = \max\{\langle A,TBU\rangle : T\in\mathbb{U}_n \text{ and } U\in\mathbb{U}_m\}.$$

In the following we say that $g(\cdot):\mathbb{R}^n\to\mathbb{R}_{\ge 0}$ is a symmetric gauge function if and only if

1) $g(\cdot)$ is a norm.
2) $\forall x\in\mathbb{R}^n : g(|x|) = g(x)$.
3) $g(Px) = g(x)$ for all permutation matrices P and all x.

Proposition 3: $\|\cdot\|$ is a unitarily invariant norm on $\mathbb{R}^{n\times m}$ if and only if $\|X\| = g(\sigma_1(X),\dots,\sigma_{\min\{m,n\}}(X))$, where g is a symmetric gauge function.

B. *Convex Optimization*

In Section IV, the following elementary convex optimization results (cf. [16]) are used.

Proposition 4: Let K_1 and K_2 be convex sets in a Hilbert space H with inner product $\langle\cdot,\cdot\rangle$. Moreover, assume K_1 has some interior points and K_2 contains non of these interior points. Then there exists an $x^*\in H$ such that

$$\sup_{x\in K_1}\langle x,x^*\rangle \le \inf_{x\in K_2}\langle x,x^*\rangle.$$

Proposition 5: Let H be a Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and $n\in H$. Moreover, let the distance of n from the convex set $K\subset H$ be measured in some norm $\|\cdot\|$, then

$$\inf_{m\in K}\|n-m\| = \max_{\substack{\|m^*\|_*\le 1\\ m^*\in H}} [\langle n,m^*\rangle - \sup_{m\in K}\langle m,m^*\rangle],$$

where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$. If the maximum on the right is achieved by some $m_0^*\in H$ and the infimum on the left by some $m_0\in K$ then $\langle n-m_0,-m_0^*\rangle = \|m_0^*\|_*\|n-m_0\|$.

Proposition 6: Let $K\subset H$ be a convex cone in a real Hilbert space H with inner product $\langle\cdot,\cdot\rangle$. Moreover, let f be a differentiable convex function on H. Then a necessary and sufficient condition that $d_0\in K$ minimizes f over K is

$$\begin{aligned}\forall d\in K : \langle\nabla f(d_0),d\rangle &\ge 0,\\ \langle\nabla f(d_0),d_0\rangle &= 0,\end{aligned}$$

where $\nabla f(d_0)$ denotes the gradient of f evaluated at d_0.

C. *Proof of Lemma 1 and Theorem 1*

Proof **(Lemma 1)**: Let $1\le r\le q:=\min\{m,n\}$ and

$$g(x_1,\dots,x_q) := \|\mathrm{diag}(x_1,\dots,x_q)\|_r.$$

Then $\|\cdot\|_r$ is a unitarily invariant norm by Proposition 3, because g is a symmetric gauge function. Now, let $M\in\mathbb{R}^{n\times m}$, then

$$\|M\|_r^2 = \max\{\langle M^TM,TPU\rangle : T,U\in\mathbb{R}^{m\times m} \text{ are unitary}\},$$

with $P := \operatorname{blkdiag}(I_r, 0_{m-r})$ by Corollary 1. If $M^TM = \sum_{i=1}^m \sigma_i(M)u_iu_i^T$ we can define a projection $P_r := \sum_{i=1}^r u_iu_i^T$ such that $\|M\|_r^2 = \langle P_r, M^TM\rangle$.

Since $\|\cdot\|_{*r}$ inherits the unitary invariance, it follows with $\Sigma := \operatorname{diag}(\sigma_1(M), \dots, \sigma_{\min\{m,n\}}(M))$ that

$$\begin{aligned}\|M\|_{*r} = \|\Sigma\|_{*r} &= \max_{\|X\|_r \le 1} \langle \Sigma, X\rangle \le \max_{\sum_i^r \sigma_i^2(X)=1} \sum_{i=1}^q \sigma_i(M)\sigma_i(X)\\ &= \max_{\sum_i^r \sigma_i^2(X)\le 1} \left[\sum_{i=1}^r \sigma_i(M)\sigma_i(X) + \sigma_r(X)\sum_{i=r+1}^q \sigma_i(M)\right],\end{aligned}$$

where the last inequality follows by Proposition 2 and clearly it can be attained. Hence,

$$\|M\|_{*r} = \max_{\sum_i^r s_i^2=1} \sum_{i=1}^q \sigma_i(M)s_i \ge \max_{\sum_i^r s_i^2=1} \sum_{i=1}^r \sigma_i(M)s_i = \sum_{i=1}^r \sigma_i^2(M),$$

with equality if and only if $\operatorname{rank}(M) \le r$. The last assertion follows in the same manner. ■

***Proof* (Theorem 1)**: Let $N \in \mathbb{R}_{\ge 0}^{n\times m}$ with $\operatorname{rank}(N) > r$, then it must hold that

$$\begin{aligned}&\min_{\substack{M\in\mathbb{R}_{\ge 0}^{n\times m}\\ \operatorname{rank}(M)\le r}} \|N-M\|_F^2 \ge \max_{D\in\mathbb{R}_{\ge 0}^{n\times m}} \min_{\substack{M\in\mathbb{R}^{n\times m}\\ \operatorname{rank}(M)\le r}} \|N-M\|_F^2 - 2\langle D, M\rangle\\ &= \max_{D\in\mathbb{R}_{\ge 0}^{n\times m}} \min_{\substack{M\in\mathbb{R}^{n\times m}\\ \operatorname{rank}(M)\le r}} \|N-M+D\|_F^2 - 2\langle D,N\rangle - \|D\|_F^2\\ &= \max_{D\in\mathbb{R}_{\ge 0}^{n\times m}} \|N+D\|_F^2 - \|N+D\|_r^2 - 2\langle D,N\rangle - \|D\|_F^2\\ &= \max_{D\in\mathbb{R}_{\ge 0}^{n\times m}} \|N\|_F^2 - \|N+D\|_r^2.\end{aligned}$$

Hence, we need to derive $\min_{D\in\mathbb{R}_{\ge 0}^{n\times m}} \|N+D\|_r = \|N+D^*\|_r$. We assume that $N+D^* = \sum_{i=1}^n \bar{\sigma}_i\bar{u}_i\bar{v}_i^T$ with $D^* \ne 0$ and $\bar{\sigma}_r > \bar{\sigma}_{r+1}$. Applying Proposition 5 from the appendix gives

$$\|N+D^*\|_r = \max_{\substack{M_d\in\mathbb{R}_{\ge 0}^{n\times m}\\ \|M_d\|_{r*}\le 1}} \langle N, M_d\rangle = \langle N, M_d^*\rangle$$

with $M_d^* \in \mathbb{R}_{\ge 0}^{n\times m}$ and $\langle D^*, M_d^*\rangle = 0$. We show now that $\operatorname{rank}(M_d^*) = r$.

First, notice that $\{X : \langle X, M_d^*\rangle = 0\}$ defines a supporting, separating hyperplane between $\mathbb{R}_{\le 0}^{n\times m}$ and the set $B := \{X \in \mathbb{R}^{n\times m} : \|N+X\|_r \le \|N+D^*\|_r\}$, which is why

$$0 = \sup_{X\in\mathbb{R}_{\le 0}^{n\times m}} \langle X, M_d^*\rangle \le \inf_{X\in B} \langle X, M_d^*\rangle \le \langle D^*, M_d^*\rangle = 0 \qquad (6)$$

by Proposition 4. Since, $\bar{\sigma}_r > \bar{\sigma}_{r+1}$ it follows that $\tilde{X} + D^* \in B$ for all $\tilde{X} = \sum_{i=r+1}^{\min\{m,n\}} \delta_i\bar{u}_i\bar{v}_i$ with $\max_i |\delta_i| < \bar{\sigma}_r - \bar{\sigma}_{r+1}$ and $\delta_i \in \mathbb{R}$. Hence, if

$$\sum_{i=1}^{\min\{m,n\}} \bar{u}_i^T M_d^* \bar{v}_i = \begin{pmatrix} \tilde{M}_{11} & \tilde{M}_{12}\\ \tilde{M}_{21} & \tilde{M}_{22}\end{pmatrix},$$

then the diagonal entries of $\tilde{M}_{22} \in \mathbb{R}^{n-r\times m-r}$ need to be zero by (6). Consequently,

$$\begin{aligned}\|N+D^*\|_r &= \max_{\substack{M_d\in\mathbb{R}_{\ge 0}^{n\times m}\\ \|M_d\|_{r*}\le 1}} \langle N, M_d\rangle = \langle \Sigma_1, \tilde{M}_{11}\rangle\\ &\le \|\Sigma_1\|_r\|\tilde{M}_{11}\|_r = \|N+D^*\|_r\|\tilde{M}_{11}\|_r \qquad (7)\end{aligned}$$

which is why by Lemma 1 $\|\tilde{M}_{11}\|_r = \|M_d^*\|_{r*} = \|M_d^*\|_F = \|M_d^*\|_r = 1$ and $\operatorname{rank}(M_d^*) = r$. We conclude that

$$\sum_{i=1}^{\min\{m,n\}} \bar{u}_i^T M_d^* \bar{v}_i = \begin{pmatrix} \tilde{M}_{11} & 0_{r,m-r}\\ 0_{n-r,r} & 0_{n-r,m-r}\end{pmatrix}$$

and $M^* := M_d^*\|N+D^*\|_r \in \mathbb{R}_{\ge 0}^{n\times m}$ with $\operatorname{rank}(M^*) = r$ fulfils

$$\|N-M^*\|_F^2 = \|N\|_F^2 - \|N+D^*\|_r^2.$$

Clearly, any other minimizer M_2^* also has to fulfil that $\langle D^*, M_2^*\rangle = 0$ and therefore by the previous derivations

$$\|N - M_2^* + D^*\|_F^2 = \min_{\substack{M\in\mathbb{R}^{n\times m}\\ \operatorname{rank}(M)\le r}} \|N + D^* - M\|_F^2,$$

which has a unique minimizer under our assumption. ■

An alternative proof of Theorem 1 can be obtained by using Proposition 6 together with the following lemma.

Lemma 3: Let $N \in \mathbb{R}^{n\times m}$ with singular value decomposition $N = \sum_{i=1}^{\min\{m,n\}} \sigma_iu_iv_i^T$ and $\sigma_r > \sigma_{r+1}$. Then for $D \in \mathbb{R}^{n\times m}$ and all $k = 1, \dots, n$ and $l = 1, \dots, m$

$$\left[\frac{\partial}{\partial d_{kl}} \|N+D\|_r^2\right]_{k,l} = \left[2\sum_{i=1}^r \sigma_iu_i^TE_{kl}v_i\right]_{k,l} = 2\sum_{i=1}^r \sigma_iu_iv_i^T$$

where E_{kl} is the matrix with a one in the (k,l)-entry and zeros elsewhere.

Proof: Assume $N \in \mathbb{R}^{n\times m}$ with singular value decomposition $N = \sum_{i=1}^{\min\{m,n\}} \sigma_iu_iv_i^T = \sum_{j=1}^p \sigma_{n_j} \sum_{i=n_j+1}^{n_{j+1}} u_iv_i^T$ and $\sigma_1 = \sigma_{n_1} > \sigma_{n_2} > \dots > \sigma_{n_p} = \sigma_n$ and $\sigma_r > \sigma_{r+1}$. Since $(N+\varepsilon E_{kl})(N+\varepsilon E_{kl})^T$ is symmetric and analytic in ε it is known (cf. Chapter 2, Theorem 5.6. in [10]) that the repeated eigenvalues as well as the orthonormal eigenvectors of $(N+\varepsilon E_{kl})(N+\varepsilon E_{kl})^T$ are analytic in ε. Hence,

$$(N+\varepsilon E_{kl})(N+\varepsilon E_{kl})^T = \sum_{i=1}^{\min\{m,n\}} \sigma_i(\varepsilon)^2u_i(\varepsilon)u_i(\varepsilon)^T,$$

where $u_i(\varepsilon) = u_i + \varepsilon\frac{du_i}{d\varepsilon} + \mathscr{O}(\varepsilon^2)$ are the perturbed right-singular vectors. Therefore, $1 = u_i(\varepsilon)^Tu_i(\varepsilon) = 1 + 2\varepsilon u_i^T\frac{du_i}{d\varepsilon} + \mathscr{O}(\varepsilon^2) \Rightarrow u_i^T\frac{du_i}{d\varepsilon} = \mathscr{O}(\varepsilon)$. Together with

$$\sum_{i=n_j+1}^{n_{j+1}} u_i^TNN^Tu_i = \sigma_{n_j}^2(n_{j+1} - n_j)$$

it follows that

$$\begin{aligned}\sigma_{n_j}(\varepsilon)^2(n_{j+1}-n_j) &= \sum_{i=n_j+1}^{n_{j+1}} u_i(\varepsilon)^T(N+\varepsilon E_{kl})(N+\varepsilon E_{kl})^Tu_i(\varepsilon)\\ &= \sigma_{n_j}^2(n_{j+1}-n_j) + 2\varepsilon\sum_{i=n_j+1}^{n_{j+1}} u_i^TE_{kl}N^Tu_i + \mathscr{O}(\varepsilon^2)\\ &= \sigma_{n_j}^2(n_{j+1}-n_j) + 2\varepsilon\sigma_{n_j}\sum_{i=n_j+1}^{n_{j+1}} u_i^TE_{kl}v_i + \mathscr{O}(\varepsilon^2).\end{aligned}$$

Since all sums are independent of a particular choice of u_i and v_i this implies that

$$\frac{\partial}{\partial d_{kl}} \sum_{i=n_j+1}^{n_j} \sigma_i^2(N+D) = 2\sigma_{n_j}\sum_{i=n_j+1}^{n_{j+1}} u_i^TE_{kl}v_i$$

is well-defined, even if $n_j > 1$. Due to the assumption that $\sigma_r > \sigma_{r+1}$, this concludes the proof. ■

Note that, if $\sigma_r(N) = \sigma_{r+1}(N)$, then one may only finds subdifferentials.