Optim Lett (2017) 11:1783–1795
DOI 10.1007/s11590-017-1146-5

ORIGINAL PAPER

Convex envelopes for fixed rank approximation

Fredrik Andersson[1] · **Marcus Carlsson**[1] · **Carl Olsson**[1]

Received: 5 February 2016 / Accepted: 21 April 2017 / Published online: 5 May 2017

Abstract A convex envelope for the problem of finding the best approximation to a given matrix with a prescribed rank is constructed. This convex envelope allows the usage of traditional optimization techniques when additional constraints are added to the finite rank approximation problem. Expression for the dependence of the convex envelope on the singular values of the given matrix is derived and global minimization properties are derived. The corresponding proximity operator is also studied.

Keywords Convex envelope · Rank constraint · Approximation · Proximity operator

1 Introduction

Let $\mathbb{M}_{m,n}$ denote the Hilbert space of complex $m \times n$-matrices equipped with the Frobenius (Hilbert–Schmidt) norm. The Eckart–Young–Schmidt theorem [4, 14] provides a solution to the classical problem of approximating a matrix by another matrix with a prescribed rank, i.e.,

$$\begin{aligned} &\min \|A - F\|^2 \\ &\text{subject to } \operatorname{rank} A \leq K, \end{aligned} \tag{1.1}$$

✉ Fredrik Andersson
fa@maths.lth.se

Marcus Carlsson
mc@maths.lth.se

Carl Olsson
calle@maths.lth.se

[1] Centre for Mathematical Sciences, Lund University, Box 118, 22100 Lund, Sweden

by means of a singular value decomposition of F and keeping only the K largest singular values. However, if additional constraints are added then there will typically not be an explicit expression for the best approximation.

Let $g(A) = 0$ describe the additional constraints (for instance imposing a certain matrix structure on A), and consider

$$\begin{aligned} &\min \|A - F\|^2 \\ &\text{subject to } \operatorname{rank} A \leq K, \quad g(A) = 0. \end{aligned} \tag{1.2}$$

The problem (1.2) can be reformulated as minimizing

$$\begin{aligned} &\mathcal{I}(A) = \mathcal{R}_K(A) + \|A - F\|^2 \\ &\text{subject to } g(A) = 0, \quad \text{where } \mathcal{R}_K(A) = \begin{cases} 0 & \operatorname{rank} A \leq K, \\ \infty & \text{otherwise.} \end{cases} \end{aligned} \tag{1.3}$$

For instance, if g describes the condition that A is a Hankel matrix and F is the Hankel matrix generated by some vector f, then the minimization problem above is related to that of approximating f by K exponential functions [9]. This particular case of (1.3) was for instance studied in [1].

Standard (e.g. gradient based) optimization techniques do no work on (1.3) due to the highly discontinuous behavior of the rank function. A popular approach is to relax the optimization problem by replacing the rank constraint with a nuclear norm penalty, i.e. to consider the problem

$$\begin{aligned} &\min_A \mu_K \|A\|_* + \|A - F\|^2 \\ &\text{subject to } g(A) = 0. \end{aligned} \tag{1.4}$$

where $\|A\|_* = \sum_j \sigma_j(A)$ and the parameter μ_K is varied until the desired rank K is obtained.

In contrast to $\mathcal{R}_K(A)$ the nuclear norm $\|A\|_*$ is a convex function, and hence (1.4) is much easier to solve than (1.3). In fact, the nuclear norm is the convex envelope of the rank function restricted to matrices with operator norm ≤ 1 [5] which motivates the replacement of $\mathcal{R}_K(A)$ with $\mu_K \|A\|_*$ (for a suitable choice of μ_K).

However, the solutions obtained by solving this relaxed problem are different and exhibit bias and other undesirable side-effects (compared with the originally sought solution), because the contribution of the (convex) misfit term $\|A - F\|^2$ is not used. In [10,11] it was suggested to incorporate the misfit term and work with the l.s.c. convex envelopes of

$$\mu \operatorname{rank}(A) + \|A - F\|^2, \tag{1.5}$$

and

$$\mathcal{I}(A) = \mathcal{R}_K(A) + \|A - F\|^2, \tag{1.6}$$

respectively for the problem of low-rank and fixed rank approximations, (where l.s.c. refers to lower semi-continuous). The superior performance of using this relaxation approach in comparison to the nuclear norm approach was verified for several examples in [10,11], where also efficient optimization algorithms for the corresponding restricted minimization problems are presented. In [2] these functionals are studied in a common framework called the $\mathcal{S}$-transform. Grussler and Rantzer [6] consider optimization of (1.6) over non-negative matrices. They derive the Lagrange dual of the problem and show that the resulting relaxation can be optimized using semidefinite programming. Furthermore, they derive sufficient conditions (in terms of the dual variables) for the relaxation to be tight. The use of semidefinite programming does however limit the approach to moderate scale problems.

For the l.s.c convex envelope of (1.5) it turns out that there are simple explicit formulas acting on each of the singular values of F individually. In this paper we present explicit expressions for the l.s.c. convex envelope of (1.6) in terms of the singular values $(\alpha_j)_{j=1}^{\min(m,n)}$ of A, as well as detailed information about global minimizers. More precisely, in Theorem 1 we show that the l.s.c. convex envelope of (1.6) is given by

$$\mathcal{I}^{**}(A) = \frac{1}{k_*}\left(\sum_{j>K-k_*} \alpha_j\right)^2 - \sum_{j>K-k_*} \alpha_j^2 + \|A-F\|^2. \tag{1.7}$$

where k_* is a particular value between 1 and K [see (2.1)]. This article also contains further information on how the l.s.c. convex envelope can be used in optimization problems. Since (1.7) is finite at all points it is also continuous, so we will sometimes write “convex envelope” instead of “l.s.c. convex envelope”.

The second main result of this note is Theorem 2, where the global minimizers of (1.7) are found. In case the Kth singular value of F (denoted ϕ_K) has multiplicity one, then the minimizer of (1.7) is unique and coincides with that of (1.6), given by the Eckart–Young–Schmidt theorem. If ϕ_K has multiplicity M and is constant between sub-indices $J \leq K \leq L$, it turns out that the singular values α_j of global minimizers A, in the range $J \leq j \leq L$ lie on a certain simplex in $\mathbb{R}^M$. We refer to Sect. 3, in particular (3.3), for further details.

Many optimization routines for solving the convex envelope counterpart of (1.3) involve computing the so called proximal operator, i.e. the operator

$$A \mapsto \underset{A}{\operatorname{argmin}}\, \mathcal{I}^{**}(A) + \rho\|A-F\|^2, \quad \rho > 0.$$

In Sect. 4 we investigate the properties of this operator. In particular we show that it is a contraction with respect to the Frobenius norm and show that the proximal operator coincides with the solution of (1.1) whenever F has a sufficient gap between the Kth and $K+1$th singular value.

Since the submission of this article the two related papers [8] and [7] have appeared. In [8] the convex envelope of (1.6) and its proximal operator are computed. In [7] these results are generalized to arbitrary unitarily invariant norms when $F = 0$.

2 Fenchel conjugates and the l.s.c. convex envelope

The Fenchel conjugate, also called the Legendre transform [13, Section 26], of a function f is defined by

$$f^*(B) = \sup_A \langle A, B\rangle - f(A).$$

Note that $\mathbb{M}_{m,n}$ becomes a real Hilbert space with the scalar product

$$\langle A, B\rangle = \mathsf{Re}\sum_{i,j} a_{i,j}\overline{b_{i,j}},$$

and that for any function $f : \mathbb{M}_{m,n} \to \mathbb{R}$ that only depends on the singular values, we have that the maximum of $\langle A, B\rangle - f(A)$ with respect to A is achieved for a matrix A with the same Schmidt-vectors (singular vectors) as B, by von-Neumann's inequality [12]. More precisely, denote the singular values of A, B by α, β and denote the singular value decomposition by $A = U_A\Sigma_\alpha V_A^*$, where Σ_α is a diagonal matrix of length $N = \min(m, n)$. We then have:

Proposition 1 *For any $A, B \in \mathbb{M}_{m,n}$ we have $\langle A, B\rangle \leq \sum_{j=1}^N \alpha_j\beta_j$ with equality if and only if the singular vectors can be chosen such that $U_A = U_B$ and $V_A = V_B$.*

See [3] for a discussion regarding the proof and the original formulation of von Neumann.

Proposition 2 *Let $\mathcal{I}(A) = \mathcal{R}_K(A) + \|A - F\|^2$ [see (1.3)]. For its Fenchel conjugate it then holds that*

$$\mathcal{I}^*(B) = \sum_{j=1}^{K}\big(\sigma_j\,(F + B/2)\big)^2 - \|F\|^2.$$

Proof $\mathcal{I}^*(B)$ is the supremum over A of the expression

$$\langle A, B\rangle - \mathcal{R}_K(A) - \|A - F\|^2 = 2\left\langle A, F + \frac{B}{2}\right\rangle - \mathcal{R}_K(A) - \|A\|^2 - \|F\|^2.$$

If we fix the singular values of A, then the last three terms are independent of the singular vectors. By Proposition 1 it follows that the maximum value is attained for a matrix A which has the same singular vectors as $F + \frac{B}{2}$. We denote $\sigma_j\left(F + \frac{B}{2}\right)$ by γ_j and $\sigma_j(A)$ by α_j, and write $\mathcal{R}_K(\alpha)$ in place of $\mathcal{R}_K(A)$ (since the singular vectors are irrelevant for this functional). Combining the above results gives

$$\mathcal{I}^*(B) = \sup_{\alpha_1\geq\alpha_2\cdots\geq\alpha_N} -\mathcal{R}_K(\alpha) - \sum_{j=1}^{N}(\alpha_j - \gamma_j)^2 + \sum_{j=1}^{N}\gamma_j^2 - \|F\|^2.$$

It is optimal to choose $\alpha_j = \gamma_j$ for $1 \leq j \leq K$ and $\alpha_j = 0$ otherwise. Hence,

$$\mathcal{I}^*(B) = -\sum_{j=K+1}^{N} \gamma_j^2 + \sum_{j=1}^{N} \gamma_j^2 - \|F\|^2 = \sum_{j=1}^{K} \left(\sigma_j\left(F + B/2\right)\right)^2 - \|F\|^2.$$

□

The computation of $\mathcal{I}^{**}$ is a bit more involved.

Theorem 1 *Given a positive non-increasing sequence $\alpha = (\alpha_j)_{j=1}^N$, the sequence*

$$\sum_{j>K-k} \alpha_j - k\alpha_{K+1-k}, \quad k = 1, \ldots, K \tag{2.1}$$

is also non-increasing for $k = 1, \ldots, K$. Let k_* be the largest value such that (2.1) is non-negative. The l.s.c. convex envelope of $\mathcal{I}(A) = \mathcal{R}_K(A) + \|A - F\|^2$ then equals

$$\mathcal{I}^{**}(A) = \frac{1}{k_*}\left(\sum_{j>K-k_*} \alpha_j\right)^2 - \sum_{j>K-k_*} \alpha_j^2 + \|A - F\|^2. \tag{2.2}$$

Note that $\mathcal{I}(A) \geq \|A - F\|^2$ and $\|A - F\|^2$ is convex and continuous in A. Since $\mathcal{I}^{**}(A)$ is the largest l.s.c. convex lower bound on $\mathcal{I}(A)$ we therefore have $\mathcal{I}^{**}(A) \geq \|A - F\|^2$ which shows that

$$\frac{1}{k_*}\left(\sum_{j>K-k_*} \alpha_j\right)^2 - \sum_{j>K-k_*} \alpha_j^2 \geq 0. \tag{2.3}$$

Proof We again employ the notation $\sigma_j\left(F + \frac{B}{2}\right) = \gamma_j$. For the bi-conjugate it then holds that

$$\begin{aligned}
\mathcal{I}^{**}(A) &= \sup_B \langle A, B\rangle - \sum_{j=1}^{K} \gamma_j^2 + \|F\|^2 = \sup_B 2\left\langle A, F + \frac{B}{2}\right\rangle \\
&\quad - \sum_{j=1}^{K} \gamma_j^2 + \|A - F\|^2 - \|A\|^2 \\
&= \sup_{\gamma_1 \geq \gamma_2 \cdots \geq \gamma_N} 2\sum_{j=1}^{N} \alpha_j \gamma_j - \sum_{j=1}^{K} \gamma_j^2 + \|A - F\|^2 - \|A\|^2 \\
&= \sup_{\gamma_1 \geq \gamma_2 \cdots \geq \gamma_N} \left(2\sum_{j=K+1}^{N} \alpha_j \gamma_j - \sum_{j=1}^{K} (\gamma_j - \alpha_j)^2\right) + \sum_{j=1}^{K} \alpha_j^2 + \|A - F\|^2 - \|A\|^2 \\
&= \sup_{\gamma_1 \geq \gamma_2 \cdots \geq \gamma_N} \left(2\sum_{j=K+1}^{N} \alpha_j \gamma_j - \sum_{j=1}^{K} (\gamma_j - \alpha_j)^2\right) - \sum_{j=K+1}^{N} \alpha_j^2 + \|A - F\|^2
\end{aligned}$$

where the third identity follows by Proposition 1 and analogous considerations as those in Proposition 2. Let the supremum be attained at the point γ^*. If we hold γ_K^* fixed and consider the supremum over the remaining variables, we get $\gamma_j^* = \max\{\alpha_j, \gamma_K^*\}$. By inspection of the above expression it is also clear that $\gamma_K^* \geq \alpha_K$ (in particular, $\gamma_j^* = \gamma_K^*$ for $j > K$). Introducing

$$f(t) = 2\sum_{j=K+1}^{N} \alpha_j t - \sum_{j=1}^{K} (\max(0, t-\alpha_j))^2$$

we conclude that

$$\mathcal{I}^{**}(A) = \sup_{t \geq \alpha_K} f(t) - \sum_{j=K+1}^{N} \alpha_j^2 + \|A-F\|^2. \tag{2.4}$$

The function f is clearly differentiable with derivative

$$f'(t) = 2\sum_{j=K+1}^{N} \alpha_j - 2\sum_{j=1}^{K} (\max(0, t-\alpha_j)),$$

which is a non-increasing function of t. In particular, the sequence $(f'(\alpha_{K+1-k}))_{k=1}^K$ is non-increasing and up to a factor of 2 it equals (2.1), which proves the first claim in the theorem. Moreover $f'(\alpha_K) = \sum_{j=K+1}^N \alpha_j$ and $\lim_{t\to\infty} f'(t) = -\infty$, whereby it follows that f has a maximum in (α_K, ∞) at a point t_* where $f'(t_*) = 0$. It also follows that k_* is the largest integer k such that $f'(\alpha_{K+1-k}) \geq 0$, and hence t_* lies in the interval $[\alpha_{K+1-k_*}, \alpha_{K-k_*})$, (with the convention $\alpha_0 = \infty$ in case $k_* = K$). In this interval we have

$$f'(t) = 2\left(\sum_{j=K+1-k_*}^{N} \alpha_j - k_* t,\right)$$

whereby it follows that

$$t_* = \frac{\sum_{j=K+1-k_*}^N \alpha_j}{k_*}.$$

Moreover,

$$f(t_*) = 2\sum_{j=K+1}^{N} \alpha_j t_* - \sum_{j=K+1-k_*}^{K} (t_* - \alpha_j)^2 = 2\sum_{j=K+1-k_*}^{N} \alpha_j t_* - k_* t_*^2 - \sum_{j=K+1-k_*}^{K} \alpha_j^2 = k_* t_*^2 - \sum_{j=K+1-k_*}^{K} \alpha_j^2$$

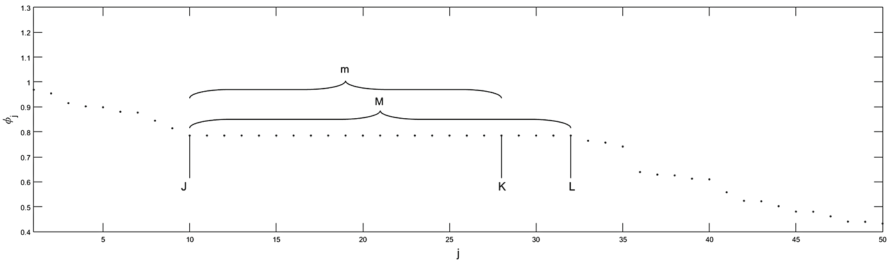

Fig. 1 Illustration of the notation used in Theorem 2

Returning to (2.4) we conclude that

$$\mathcal{I}^{**}(A) = k_* t_*^2 - \sum_{j=K+1-k_*}^{K} \alpha_j^2 - \sum_{j=K+1}^{N} \alpha_j^2 + \|A - F\|^2.$$

which equals (2.2), and the proof is complete. □

3 Global minimizers

We now consider global minimizers of $\mathcal{I}$ and $\mathcal{I}^{**}$. Given a sequence $(\phi_n)_{n=1}^N$ we recall that Σ_ϕ denotes the corresponding diagonal matrix. We introduce the notation $\tilde{\phi}$ for the sequence ϕ truncated at K, i.e.

$$\tilde{\phi}_j = \begin{cases} \phi_j & \text{if } 1 \leq j \leq K, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

Recall the Eckart-Young-Schmidt theorem, which can be rephrased as follows;

The elements of $\underset{A}{\operatorname{argmin}}\, \mathcal{I}(A)$ *are all matrices of the form* $A_* = U\Sigma_{\tilde{\phi}}V^*$, *where* $U\Sigma_\phi V^*$ *is any singular value decomposition of* F. *In particular,* A_* *is unique if and only if* $\phi_K \neq \phi_{K+1}$.

Obviously, a global minimizer of $\mathcal{I}$ is a global minimizer of $\mathcal{I}^{**}$, but the converse need not be true. It is not hard to see that, in case ϕ_K has multiplicity one, the minimizer of $\mathcal{I}$ is also the (unique) minimizer of $\mathcal{I}^{**}$. The general situation is more complicated.

Theorem 2 *Let* $K \in \mathbb{N}$ *be given, let* F *be a fixed matrix and let* ϕ *be its singular values. Let* ϕ_J *(respectively* ϕ_L*) be the first (respectively last) singular value that equals* ϕ_K, *and set* $M = L + 1 - J$ *(that is, the multiplicity of* ϕ_K*). Finally set* $m = K + 1 - J$, *(that is, the multiplicity of* $\tilde{\phi}_K$*). Figure 1 illustrates the setup.*

The global minimum of $\mathcal{I}$ *and* $\mathcal{I}^{**}$ *both equal* $\sum_{j>K} \phi_j^2$ *and the elements of*

$$\underset{A}{\operatorname{argmin}}\, \mathcal{I}^{**}(A) \tag{3.2}$$

are all matrices of the form $A_* = U\Sigma_\alpha V^*$, *where* $U\Sigma_\phi V^*$ *is any singular value decomposition of* F, *and* α *is a non-increasing sequence satisfying:*

$$\begin{cases} \alpha_j = \phi_j, & \text{for } 1 \le j < J, \\ \alpha_j \le \phi_K, & \text{for } J \le j \le L \text{ and } \sum_{j=J}^{L} \alpha_j = \phi_K m, \\ \alpha_j = 0, & \text{for } j > L. \end{cases} \tag{3.3}$$

Also, the minimal rank of such an A_* *is* K *and the maximal rank is* L. *In particular,* A_* *is unique if and only if* $\phi_K \neq \phi_{K+1}$.

Proof The fact that the minimum value of $\mathcal{I}$ and $\mathcal{I}^{**}$ coincide follows immediately since $\mathcal{I}^{**}$ is the l.s.c. convex envelope of $\mathcal{I}$, and the fact that this value is $\sum_{j>K} \phi_j^2$ follows by the Eckart-Young-Schmidt theorem.

Suppose first that A_* is as stated in (3.3). We first prove that $k_* = m$. Evaluating the testing condition (2.1) for $k = m + 1$ gives

$$\sum_{j>K-(m+1)} \alpha_j - (m+1)\alpha_{K+1-(m+1)} = \sum_{j=J}^{N} \alpha_j - m\alpha_{J-1} = m(\phi_K - \phi_{J-1}) < 0 \tag{3.4}$$

(where we use (3.3) in the last identity) so $k_* \le m$. But on the other hand, the testing condition for $k = m$ is

$$\sum_{j>K-m} \alpha_j - m\alpha_{K+1-m} = \sum_{j=J}^{N} \alpha_j - m\alpha_J = m(\phi_K - \alpha_J) \ge 0$$

so we must have $m = k_*$. With (3.3) in mind we get that $\sum_{j>K-k_*} \alpha_j = \sum_{j=J}^{N} \alpha_j = m\phi_K$ and then

$$\begin{aligned} \mathcal{I}^{**}(A_*) = \frac{1}{m}(m\phi_K)^2 - \sum_{j=J}^{N} \alpha_j^2 + \sum_{j=J}^{N} (\alpha_j - \phi_j)^2 = m\phi_K^2 \\ - \sum_{j=J}^{N} (2\alpha_j\phi_j - \phi_j^2) = m\phi_K^2 \\ - \sum_{j=J}^{L} 2\alpha_j\phi_K + \sum_{j=J}^{N} \phi_j^2 = m\phi_K^2 - 2\left(\sum_{j=J}^{L} \alpha_j\right)\phi_K + \sum_{j=J}^{N} \phi_j^2 \\ = -m\phi_K^2 + \sum_{j=J}^{N} \phi_j^2 = \sum_{j=K+1}^{N} \phi_j^2 \end{aligned}$$

since $m\phi_K^2 = \sum_{j=J}^{K} \phi_j^2$. This proves that A_* is a solution to (3.2).

Conversely, let A_* be a solution to (3.2). The only part of the expression (2.2) for $\mathcal{I}^{**}(A_*)$ that depends on the singular vectors of A_* is $\|A_* - F\|^2$. By expanding

$\|A_*\|^2 - 2\langle A_*, F\rangle + \|F\|^2$ and invoking Proposition 1, it follows that we can choose matrices U and V such that $A_* = U\Sigma_\alpha V^*$ and $F = U\Sigma_\phi V^*$ are singular value decompositions of A_* and F respectively. Set $\tilde{F} = U\Sigma_{\tilde{\phi}} V^*$ and note that $\tilde{F}$ also is a minimizer of (3.2), by the first part of the proof. Since $\mathcal{I}^{**}$ is the l.s.c. convex envelope of $\mathcal{I}$, it follows that all matrices

$$A(t) = \tilde{F} + t(A_* - \tilde{F}), \quad 0 \le t \le 1,$$

are solutions of (3.2). Set

$$\epsilon = \alpha - \tilde{\phi}, \tag{3.5}$$

and note that $A(t) = U\Sigma_{\tilde{\phi}+t\epsilon}V^*$, i.e. the singular values of $A(t)$ equals $\tilde{\phi} + t\epsilon = t\alpha + (1-t)\tilde{\phi}$ which is non-increasing for all $t \in [0, 1]$, being the weighted mean of two non-increasing sequences. Since $\tilde{F}$ satisfies (3.3), it follows that $A(t)$ satisfies (3.3) if and only if

$$\begin{cases} \epsilon_j = 0, & 1 \le j < J, \\ t\epsilon_j \le 0 & \text{for } j = J, \dots, K; \quad \text{and } \sum_{j=J}^{L} t\epsilon_j = 0, \\ \epsilon_j = 0, & j > L. \end{cases}$$

This is independent of t, and hence it suffices to prove (3.3) for some fixed $A(t_0)$ in order for $A_* = A(1)$ to satisfy (3.3) as well. In other words we may assume that ϵ in (3.5) is arbitrarily small [by redefining A_* to equal $A(t_0)$]. With this at hand, we evaluate the testing condition for k_* [recall (2.1)] at $k = m + 1$;

$$\sum_{j>K-(m+1)} \alpha_j - (m+1)\alpha_{K+1-(m+1)} = \sum_{j>J-2} \alpha_j - (m+1)\alpha_{J-1} = \sum_{j=J}^{N} \alpha_j - m\alpha_{J-1}.$$

This expression is certainly strictly negative if $\alpha = \tilde{\phi}$, [by the calculation (3.4)], and hence it is also strictly negative for α sufficiently close to $\tilde{\phi}$. Since we have already argued that it is no restriction to make this assumption, we conclude that $k_* \le m$.

With this at hand, we have

$$\begin{aligned} \mathcal{I}^{**}(A_*) &= \frac{1}{k_*}\Bigg(\sum_{j>K-k_*} \alpha_j\Bigg)^2 - \sum_{j>K-k_*} \alpha_j^2 + \sum_{j=1}^{N} (\alpha_j - \phi_j)^2 \\ &= \frac{1}{k_*}\Bigg(\sum_{j>K-k_*} \alpha_j\Bigg)^2 - \sum_{j>K-k_*} \alpha_j^2 + \sum_{j>K-k_*} (\alpha_j - \phi_j)^2 + \sum_{j=1}^{K-k_*} (\alpha_j - \phi_j)^2. \end{aligned}$$

Upon omitting the last term we get

$$\mathcal{I}^{**}(A_*) \ge \frac{1}{k_*}\Bigg(\sum_{j>K-k_*} \alpha_j\Bigg)^2 - 2\sum_{j>K-k_*} \alpha_j\phi_j + \sum_{j>K-k_*} \phi_j^2 \ge$$

$$\frac{1}{k_*}\Bigg(\sum_{j>K-k_*}\alpha_j\Bigg)^2 - 2\phi_K\sum_{j>K-k_*}\alpha_j + \sum_{j>K-k_*}\phi_j^2. \tag{3.6}$$

Moreover, since $\sum_{j>K-k_*}\phi_j^2 = k_*\phi_K^2 + \sum_{j=K+1}^N \phi_j^2$, this can be further simplified to

$$\mathcal{I}^{**}(A_*) = \frac{1}{k_*}\Bigg(\sum_{j>K-k_*}\alpha_j - k^*\phi_K\Bigg)^2 + \sum_{j=K+1}^N \phi_j^2 \geq \sum_{j=K+1}^N \phi_j^2. \tag{3.7}$$

Since the right hand side equals the global minimum of $\mathcal{I}^{**}$, we must have equality in all the above inequalities. The first one in (3.6) is equal if and only if $\alpha_j = \phi_j$ for $j = 1\ldots, K-k_*$, the second one if and only if $\alpha_j = 0$ for $j > L$, leading us to $\sum_{j>K-k_*}\alpha_j = \sum_{j=K-k_*+1}^L \alpha_j$. Since we need inequality in (3.7) as well, this in turn gives $\sum_{j=K-k_*+1}^L \alpha_j = k^*\phi_K$. As $\alpha_j = \phi_j = \phi_K$ for $j = J,\ldots, K-k_*$, this implies

$$\sum_{j=J}^L \alpha_j = ((K-k_*) - J + 1)\phi_K + k_*\phi_K = m\phi_K. \tag{3.8}$$

To verify (3.3) it remains to verify that $\alpha_j \leq \phi_K$ for $J \leq j \leq L$. If $k_* < m$, this is immediate since α by definition is a non-increasing sequence and $\alpha_J = \phi_K$ in this case. Otherwise, by (2.1) for $k_* = m$ we get

$$0 \leq \sum_{j>K-m}\alpha_j - m\alpha_{K+1-m} = \sum_{j=J}^L \alpha_j - m\alpha_J = m(\phi_K - \alpha_J),$$

where we used (3.8) in the last identity. Thus $\alpha_j \leq \alpha_J \leq \phi_K$ for $J \leq j \leq L$, which concludes the converse part of the proof.

Finally, the uniqueness statement is immediate. Clearly we can pick α_j in accordance with (3.3) to get L non-zero entries, but not more, so the maximal possible rank is L. In order to have as few non-zero entries as possible, the condition $\sum_{j=J}^L \alpha_j = \phi_K m$ together with $\alpha_j \leq \phi_K$ for $J \leq j \leq L$ clearly forces at least m non-zero entries in $J \leq j \leq L$, so the minimal possible rank is $J - 1 + m = K$. □

4 The proximal operator

As argued in the introduction, many optimization routines for solving

$$\begin{gathered}\min \mathcal{I}^{**}(A)\\ \text{subject to } g(A) = 0,\end{gathered}$$

i.e. the l.s.c. convex envelope counterpart of (1.3), require efficient computation of the proximal operator, which we now address.

Theorem 3 *Let $F = U\Sigma_\phi V^*$ be given. The solution of*

$$\operatorname*{argmin}_A \mathcal{I}^{**}(A) + \rho\|A - F\|^2, \tag{4.1}$$

is of the form $A = U\Sigma_\alpha V^*$ where α has the following structure; there exists natural numbers $k_1 \leq K \leq k_2$ and a real number $s > \phi_{k_2}$ such that

$$\alpha_j = \begin{cases} \phi_j, & j < k_1 \\ \phi_j - \frac{s-\phi_j}{\rho}, & k_1 \leq j \leq k_2 \\ 0, & j > k_2 \end{cases} \tag{4.2}$$

The appropriate value of s is found by minimizing the convex function

$$\sum_{j=1}^{K} \left(\max(\phi_j, s) - \phi_j\right)^2 + \sum_{j=K+1}^{N} \left(\min(\phi_j, \frac{s}{1+\rho}) - \phi_j\right)^2. \tag{4.3}$$

in the interval $[\phi_K, (1+\rho)\phi_{K+1}]$. Given such an s, k_1 is the smallest index ϕ with $\phi_{k_1} < s$ and k_2 last index with $\phi_{k_2} > \frac{s}{1+\rho}$. In particular, α is a non-increasing sequence and $\alpha \leq \phi$. In other words, the proximal operator is a contraction.

The theorem can be deduced by working directly with the expression for $\mathcal{I}^{**}$, but it turns out that it is easier to follow the approach in [10] which is based on the minimax theorem and an analysis of the simpler functional $\mathcal{I}^*$. Note in particular that the proximal operator (given by Theorem 3) reduce to the "Eckart–Young approximation" (3.1) if $\phi_K \geq (1+\rho)\phi_{K+1}$.

Proof Note that $\mathcal{I}^{**}(0)+\rho\|0-F\|^2 = (1+\rho)\|F\|^2$, and that $\mathcal{I}^{**}(A)+\rho\|A-F\|^2 \geq (1+\rho)\|F\|^2$ whenever A is outside the compact convex set $\mathcal{C} = \{A : \|A-F\| \leq \|F\|\}$ [recall (2.3)]. This combined with Proposition 2 and some algebraic simplifications shows that

$$\begin{aligned}
&\min_A \mathcal{I}^{**}(A) + \rho\|A - F\|^2 = \min_{A\in\mathcal{C}} \mathcal{I}^{**}(A) + \rho\|A - F\|^2 \\
&\quad = \min_{A\in\mathcal{C}} \max_B \langle A, B\rangle - \mathcal{I}^*(B) + \rho\|A - F\|^2 \\
&\quad = \min_{A\in\mathcal{C}} \max_B \langle A, B\rangle - \sum_{j=1}^{K} \left(\sigma_j\left(F + \frac{B}{2}\right)\right)^2 + \|F\|^2 + \rho\|A - F\|^2 \\
&\quad = \min_{A\in\mathcal{C}} \max_Z 2\langle A, Z\rangle - 2\langle A, F\rangle - \sum_{j=1}^{K} \left(\sigma_j\,(Z)\right)^2 + \|F\|^2 + \rho\|A - F\|^2 \\
&\quad = \min_{A\in\mathcal{C}} \max_Z \rho\left\|A - \frac{(1+\rho)F - Z}{\rho}\right\|^2 - \frac{1}{\rho}\|Z - (1+\rho)F\|^2 + (1+\rho)\|F\|^2 \\
&\qquad - \sum_{j=1}^{K} \left(\sigma_j(Z)\right)^2,
\end{aligned}$$

where $Z = F + \frac{B}{A}$. Let us denote the function in the last line by $f(A, Z)$, and note that by construction it is convex in A and concave in Z. By Sion's minimax theorem [15] the order of max and min can be switched (giving the relation $A = ((1+\rho)F - Z)/\rho$), and the above min max thus equal

$$\min_{A\in\mathcal{C}} \max_{Z} f(A, Z) = \max_{Z} \min_{A\in\mathcal{C}} f(A, Z) = \max_{Z} \; -\frac{1}{\rho}\|Z - (1+\rho)F\|^2 - \sum_{j=1}^{K} \zeta_j^2, \tag{4.4}$$

where $\zeta_j = \sigma_j(Z)$. The maximum is clearly obtained at some finite matrix $Z = Z_*$. Set

$$A_* = ((1+\rho)F - Z_*)/\rho. \tag{4.5}$$

For these points it then holds

$$\min_{A\in\mathcal{C}} f(A, Z) \le f(A_*, Z_*) \le \max_{Z} f(A, Z).$$

The latter inequality together with the previous calculations imply that $\mathcal{I}^{**}(A_*) + \rho\|A_* - F\|^2 \le \mathcal{I}^{**}(A) + \rho\|A - F\|^2$. Thus A_* given by (4.5) solves the original problem (4.1) as long as Z_* is a maximizer of (4.4). By Proposition 1 it follows that the appropriate Z shares singular vectors with F, so the problem reduces to that of minimizing

$$\begin{aligned}\operatorname*{argmin}_{\zeta} \sum_{j=1}^{N}(\zeta_j - (1+\rho)\phi_j)^2 + \rho\sum_{j=1}^{K}\zeta_j^2 &= \operatorname*{argmin}_{\zeta}\,(1+\rho)\sum_{j=1}^{K}(\zeta_j - \phi_j)^2 \\ &+ \sum_{j=K+1}^{N}(\zeta_j - (1+\rho)\phi_j)^2.\end{aligned}$$

The unconstrained minimization (i.e. ignoring that the singular values need to be non-increasing) of this is $\zeta_j = \phi_j$ for $j \le K$ and $\zeta_j = (1+\rho)\phi_j$ for $j > K$. It is not hard to see (see the appendix of [10] for more details) that the constrained minimization has the solution

$$\zeta_j = \begin{cases} \max(\phi_j, s), & j \le K \\ \min((1+\rho)\phi_j, s), & j > K \end{cases} \tag{4.6}$$

where s is a parameter between ϕ_K and $(1+\rho)\phi_{K+1}$. Inserting this in the previous expression gives (4.3) and the appropriate value of s is easily found. Let k_1 resp. k_2 be the first resp. last index where s shows up in ζ. Formula (4.2) is now an easy consequence of (4.6). □

5 Conclusions

We have analyzed and derived expressions for how to compute the l.s.c. convex envelope corresponding to the problem of finding the best approximation to a given matrix with a prescribed rank. These expressions work directly on the singular values.

Acknowledgements This research is partially supported by the Swedish Research Council, Grants Nos. 2011-5589, 2012-4213 and 2015-03780; and the Crafoord Foundation. We are also indebted to an anonymous referee for significantly improving the manuscript.

References

1. Andersson, F., Carlsson, M., Tourneret, J.-Y., Wendt, H.: A new frequency estimation method for equally and unequally spaced data. IEEE Trans. Signal Process. **62**(21), 5761–5774 (2014)
2. Carlsson, M.: On convexification/optimization of functionals including an l2-misfit term. arXiv preprint arXiv:1609.09378 (2016)
3. de Sá, E.M.: Exposed faces and duality for symmetric and unitarily invariant norms. Linear Algebra Appl. **197**, 429–450 (1994)
4. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psychometrika **1**(3), 211–218 (1936)
5. Fazel, Maryam: Matrix rank minimization with applications. PhD thesis, Stanford University (2002)
6. Grussler, C., Rantzer, A.: On optimal low-rank approximation of non-negative matrices. In: 2015 54th IEEE Conference on Decision and Control (CDC), pp. 5278–5283 (2015)
7. Grussler, C., Giselsson, P.: Low-rank inducing norms with optimality interpretations. CoRR arXiv:1612.03186 (2016)
8. Grussler, C., Rantzer, A., Giselsson, P.: Low-rank optimization with convex constraints. CoRR arXiv:1606.01793 (2016)
9. Kronecker, L.: Zur Theorie der Elimination einer Variabeln aus zwei algebraischen Gleichungen. Königliche Akad. der Wissenschaften (1881)
10. Larsson, V., Olsson, C.: Convex envelopes for low rank approximation. In: Energy Minimization Methods in Computer Vision and Pattern Recognition, pp. 1–14. Springer, Berlin (2015)
11. Larsson, V., Olsson, C., Bylow, E., Kahl, F.: Rank minimization with structured data patterns. In: Computer Vision—ECCV 2014, pp. 250–265. Springer, Berlin (2014)
12. Mirsky, L.: A trace inequality of John von Neumann. Monatshefte für Mathematik **79**(4), 303–306 (1975)
13. Rockafellar, R.T.: Convex Analysis. Princeton university press, Princeton (2015)
14. Schmidt, E.: Zur Theorie der linearen und nichtlinearen Integralgleichungen. III. Teil. Mathematische Annalen **65**(3), 370–399 (1908)
15. Sion, M., et al.: On general minimax theorems. Pac. J. Math. **8**(1), 171–176 (1958)