RAIRO-Oper. Res. 47 (2013) 299–310
DOI: 10.1051/ro/2013040

RAIRO Operations Research
www.rairo-ro.org

FROM ECKART AND YOUNG APPROXIMATION TO MOREAU ENVELOPES AND *VICE VERSA*

Jean-Baptiste Hiriart-Urruty[1] and Hai Yen Le[1]

Abstract. In matricial analysis, the theorem of Eckart and Young provides a best approximation of an arbitrary matrix by a matrix of rank at most r. In variational analysis or optimization, the Moreau envelopes are appropriate ways of approximating or regularizing the rank function. We prove here that we can go forwards and backwards between the two procedures, thereby showing that they carry essentially the same information.

Keywords. Eckart and Young theorem, moreau envelopes, rank minimization problems.

Mathematics Subject Classification. 15A, 46N10, 65K10, 90C.

1. From Eckart and Young theorem to Moreau envelopes

1.1. Eckart and Young Theorem

Let $\mathcal{M}_{m,n}(\mathbb{R})$ be equipped with the usual inner product

$$\langle\langle U, V\rangle\rangle := \text{trace of } U^T V \quad \left(\text{tr}\left(U^T V\right) \text{ in short}\right),$$

and the associated norm

$$\|.\|_F = \sqrt{\langle\langle ., .\rangle\rangle},$$

Received 6 June, 2013. Accepted 21 June, 2013.

[1] Institute of Mathematics, Paul Sabatier University, 118 Route de Narbonne, 31400 Toulouse, France. jbhu@math.univ-toulouse.fr; hyle@math.univ-toulouse.fr; http://www.math.univ-toulouse.fr/~jbhu/

Article published by EDP Sciences

sometimes called the Frobenius or Frobenius-Schur norm. If $p := \min(m, n)$, for $k \in \{0, 1, \ldots, p\}$, we may define

$$S_k := \{M \in \mathcal{M}_{m,n}(\mathbb{R}) | \quad \text{rank } M \leq k\},$$
$$\varSigma_k := \{M \in \mathcal{M}_{m,n}(\mathbb{R}) | \quad \text{rank } M = k\}.$$

If $\max(m, n) \geq 2$, $\varSigma_k$ is a smooth and connected manifold of dimension $k(m+n-k)$ ([1], p. 140). Apart from the case $k = 0$ (where $S_0 = \varSigma_0 = \{0\}$) and the case $k = p$ (see below), $\varSigma_k$ has no specific topological property. As for S_k, it enjoys some nicer mathematical properties. Firstly, it is closed as the sublevel-set (at level k) of the rank function, a lower-semicontinuous one; secondly, since it is characterized by the vanishing of all $(k+1, k+1)$-minors of A, it is a solution set of polynomial equations, thus a so-called semi-algebraic variety. The link between S_k and $\varSigma_k$ is made clear in the following results:

(i) $S_p = \mathcal{M}_{m,n}(\mathbb{R})$ and $\varSigma_p$ is an open dense subset of S_p;
(ii) if $k < p$, the interior of $\varSigma_k$ is empty while its closure is S_k.

Given $A \in \mathcal{M}_{m,n}(\mathbb{R})$ of rank r and an integer $k \leq r$, what could be said about the matrices in S_k closest to A? Observe firstly that this best approximation problem makes sense since we have defined a distance (*via* the Frobenius norm) on $\mathcal{M}_{m,n}(\mathbb{R})$. However, even if the *existence* of best approximants does not offer any difficulty (remember that $\|.\|$ is a continuous function and S_k is a closed subset), the question of *uniqueness* as well as that of an *explicit form* of best approximants remain posed. It turns out that there is a beautiful theorem answering these questions.

Before going further, we recall a technique of decomposition of matrices which is central in numerical matricial analysis and in statistics: the so-called singular value decomposition (SVD). Here it is: *Given $A \in \mathcal{M}_{m,n}(\mathbb{R})$, there is an (m, m) orthogonal matrix U, an (n, n) orthogonal matrix V, a "pseudo-diagonal" matrix D*[2] *of the same size as A, such that $A = UDV$.*

The matrix D is a sort of skeleton of A: all the "non-diagonal" entries of D are zero; on the "diagonal" are the *singular values* $\sigma_1, \sigma_2, \ldots, \sigma_p$ of A, that are the square roots of the eigenvalues of $A^T A$ (or AA^T). By definition, all the σ_i's are nonnegative, and exactly r of them (if $r = \text{rank } A$) are positive. By changing the ordering in columns or rows in U and V, and without loss of generality, we can suppose that

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0. \tag{1.1}$$

U and V are orthogonal matrices of appropriate sizes (so that the product UDV of matrices can be performed). Of course, these U and V are not unique.

[2] $D = [d_{ij}]$ "pseudo-diagonal" means that $d_{ij} = 0$ for $i \neq j$. One also uses the notation $\text{diag}_{m,n}(\sigma_1, \ldots, \sigma_p)$ for D.

The best approximation problem that we consider now is as follows: Given $A \in \mathcal{M}_{m,n}(\mathbb{R})$ of rank r and $k < r$,

$$(\mathcal{A}_k) \begin{cases} \text{Minimize } \|A - M\|_F \\ M \in S_k. \end{cases}$$

This problem is solved in the following theorem.

Theorem 1.1 (Eckart and Young, 1936)**.** *Let $0 \neq A \in \mathcal{M}_{m,n}(\mathbb{R})$ of rank r and let $1 \leq k < r$. Let $A = UDV$ be a singular value decomposition of A. Then*

$$A_k := UD_kV,$$

(*where D_k is obtained from D by keeping $\sigma_1, \ldots, \sigma_k$ and putting 0 in the place of $\sigma_{k+1}, \ldots, \sigma_r$*) *is a solution of the best approximation problem $(\mathcal{A}_k)$. Such a solution is unique when $\sigma_k > \sigma_{k+1}$.*

The optimal value in $(\mathcal{A}_k)$ is

$$\min_{M \in S_k} \|A - M\|_F = \sqrt{\sum_{i=k+1}^{r} \sigma_i^2}.$$

This theorem is a classical result in numerical matricial analysis, usually bearing the name of Eckart and Young. From the historical viewpoint, there is however some discussion about the naming of Theorem 1.1; according to Stewart ([8]), the mathematician E. Schmidt should be credited for having derived this approximation theorem, while studying integral equations, in a publication which dates back to 1907. Moreover, Mirsky (1960) showed that the A_k defined in Theorem 1.1 is a minimizer in problem $(\mathcal{A}_k)$ for any unitary invariant norm (a norm $\|.\|$ on $\mathcal{M}_{m,n}(\mathbb{R})$ is called unitary invariant if $\|UAV\| = \|A\|$ for any orthogonal pair of matrices U and V). See also [2] for references and additional comments on it. So, to be complete, we should call Theorem 1.1 the Schmidt-Eckart-Young-Mirsky theorem. For the sake of brevity, we nevertheless stand by the usual appellation (in papers as well as in textbooks) which is “Eckart and Young theorem”.

Let us go back to the approximation or optimization result itself. Indeed, since $\|.\|_F$ is derived from an inner product, the objective function in $(\mathcal{A}_k)$ (taking its square actually, $\|A - M\|_F^2$) is convex and smooth. However, due to the nonconvexity of the constraint set S_k, the optimization problem $(\mathcal{A}_k)$ is non-convex. It is therefore surprising that one could provide (*via* the Eckart and Young theorem) an explicit form of a solution of this problem. In short, since the S_k's are the sublevel-sets of the rank function, *one always has at one's disposal a “projection” of (an arbitrary) matrix A on the sublevel-sets S_k.*

Some comments are in order here to explain why $(\mathcal{A}_k)$ has a unique solution when $\sigma_k > \sigma_{k+1}$. Let us denote by $O(m,n)_A$ the set of pairs (U, V) of orthogonal matrices appearing in a singular value decomposition $U\text{diag}_{m,n}[\sigma_1, \ldots, \sigma_p]V$ of A. Then, as stated in Theorem 1.1, *one solution* of the problem $(\mathcal{A}_k)$ is given by

$A_k = UD_kV$, with (U, V) arbitrary chosen in $O(m, n)_A$. But actually, *all the optimal solutions* of $(\mathcal{A}_k)$ are given by matrices $\tilde{U}D_k\tilde{V}$, where $(\tilde{U}, \tilde{V}) \in O(m, n)_A$. In short, the solution set of problem $(\mathcal{A}_k)$ is

$$\left\{\tilde{U}D_k\tilde{V} \mid \quad \left(\tilde{U}, \tilde{V}\right) \in O(m, n)_A\right\}.$$

When $\sigma_k > \sigma_{k+1}$, it can easily be proved that

$$\tilde{U}_1 D_k \tilde{V}_1 = \tilde{U}_2 D_k \tilde{V}_2$$

whenever $(\tilde{U}_1, \tilde{V}_1)$ and $(\tilde{U}_2, \tilde{V}_2)$ are taken in $O(m, n)_A$. Hence, in that case, all the solutions of $(\mathcal{A}_k)$ coalesce to just one.

But, if $\sigma_k = \sigma_{k+1}$ for example, the problem $(\mathcal{A}_k)$ has infinitely many solutions. In spite of a thorough search, we could not find the description of the whole solution set of $(\mathcal{A}_k)$ in textbooks on numerical matricial analysis.

Note that all the solutions of the approximation problem $(\mathcal{A}_k)$ are of rank k exactly. As a consequence, under the assumptions of Theorem 1.1, we also have solved the problem of the projection of A on the manifold Σ_k. We will use a by-product of this result in the following form: If $0 \neq A \in \mathcal{M}_{m,n}(\mathbb{R})$ is of rank r and if $1 \leq k < r$,

$$\operatorname{dist}(A, \Sigma_k) = \operatorname{dist}(A, S_k) = \sqrt{\sum_{i=k+1}^{r} \sigma_i^2}. \tag{1.2}$$

1.2. Towards Moreau envelopes

We begin with some historical comments. In 1962–1963, so exactly 50 years ago, the French mechanician-mathematician J.-J.Moreau introduced a way of regularizing and approximating a convex function, called *prox-regularization* ([5], [6]). This was an example of the so-called inf-convolution or epigraphic addition of two functions. The process has some resemblances with the (exterior) penalization of a function with a squared norm term or with the Tikhonov regularization in matricial analysis; it is however different: starting with an arbitrary convex function, the objective was to define, in a "variational way", a regularized version of it which is convex and smooth. The process has been very successful since and one cannot count the number of works on the so called *prox-methods* in convex minimization. Indeed, the date 1962–1963 marks the birth of modern convex analysis and optimization.

In subsequent efforts by several mathematicians, the prox-regularization process has been extended to *nonconvex* functions. We have to rely on this (nonconvex) setting, since the function at stake here, the rank function, does not enjoy any convexity property, by far.

Although the rank function is a "bumpy" one, it is lower-semicontinuous and bounded from below; it therefore can be approximated-regularized in the so-called Moreau and Yosida way. This technique, very much in vogue in the context of variational analysis or optimization, gives rise to continuous approximations of the original function; they are called its *Moreau envelopes*. Surprisingly enough, the rank function is amenable to such an approximation-regularization process, and we get at the end *explicit* forms of the Moreau envelopes in terms of singular matrices. For that purpose, Eckart and Young theorem (Sect. 1.1) will be instrumental.

Let us firstly recall what is known, as a general rule, for the Moreau-Yosida approximation-regularization technique in a *non-convex* context (see [7], Sect. 1.G for example).

Let $(E, \|.\|)$ be an Euclidean space, let $f : E \longrightarrow \mathbb{R}$ be a lower-semicontinuous function, bounded from below on E. For a parameter value $\varepsilon > 0$, the *Moreau envelope* (or Moreau-Yosida approximate) function f_ε and the (so-called) *proximal set-valued mapping* $\mathrm{Prox}_\varepsilon f$ are defined by:

$$f_\varepsilon(x) := \inf_{u \in E} \left\{ f(u) + \frac{1}{\varepsilon}\|x-u\|^2 \right\},$$

$$\mathrm{Prox}_\varepsilon f(x) := \left\{ \overline{u} \in E | \quad f(\overline{u}) + \frac{1}{\varepsilon}\|x-\overline{u}\|^2 = f_\varepsilon(x) \right\}.$$

Then:

(i) f_ε is a finite-valued continuous function on E, minimizing f on E;
(ii) the sequence of functions $(f_\varepsilon)_{\varepsilon>0}$ increases when ε decreases, and $f_\varepsilon(x) \longrightarrow f(x)$ for all $x \in E$;
(iii) the set $\mathrm{Prox}_\varepsilon f(x)$ is nonempty and compact for all $x \in E$;
(iv) the lower bounds of f and f_ε on E are equal:

$$\inf_{x \in E} f(x) = \inf_{x \in E} f_\varepsilon(x).$$

We now apply this process to the rank function. The context is therefore as following: $E = \mathcal{M}_{m,n}(\mathbb{R})$, the norm is the Frobenius norm $\|.\|_F$ and $f : \mathcal{M}_{m,n}(\mathbb{R}) \longrightarrow \mathbb{R}$ is the rank function. By definition,

$$(\mathrm{rank})_\varepsilon (A) = \inf_{M \in \mathcal{M}_{m,n}(\mathbb{R})} \left\{ \mathrm{rank}\ M + \frac{1}{\varepsilon}\|A-M\|_F^2 \right\}, \tag{1.3}$$

$$\mathrm{Prox}_\varepsilon (\mathrm{rank})(A) = \left\{ \overline{M} \in \mathcal{M}_{m,n}(\mathbb{R}) | \quad \mathrm{rank}\ \overline{M} + \frac{1}{\varepsilon}\|A-\overline{M}\|_F^2 = (\mathrm{rank})_\varepsilon(A) \right\}. \tag{1.4}$$

Here is the main theorem in this subsection. It was announced in our concomitent survey paper [3], Section 8.2.

Theorem 1.2. *Let $0 \neq A \in \mathcal{M}_{m,n}(\mathbb{R})$ of rank r and $\varepsilon > 0$. Then:*

(i)

$$(\mathrm{rank})_\varepsilon(A) = \frac{1}{\varepsilon}\|A\|_F^2 - \frac{1}{\varepsilon}\sum_{i=1}^{r}\left[\sigma_i^2(A) - \varepsilon\right]^+ . \tag{1.5}$$

(ii) One minimizer in (1.3), i.e. one element in $\mathrm{Prox}_\varepsilon(\mathrm{rank})(A)$*, is provided by* $B := \tilde{U}\Sigma_B\tilde{V}$*, where*

- $(\tilde{U}, \tilde{V}) \in O(m,n)_A$*, i.e.* $\tilde{U}$ *and* $\tilde{V}$ *are orthogonal matrices such that* $A = \tilde{U}\Sigma_A\tilde{V}$*, with* $\Sigma_A = \mathrm{diag}_{m,n}[\sigma_1(A), \dots, \sigma_r(A), 0, \dots, 0]$ *(a singular value decomposition of* A *with* $\sigma_1(A) \geq \dots \geq \sigma_r(A) > \sigma_{r+1}(A) = \dots = \sigma_p(A) = 0$*);*
-

$$\Sigma_B = \begin{cases} 0 & \text{if } \max_i \sigma_i(A) = \sigma_1(A) \leq \sqrt{\varepsilon}, \\ \Sigma_A & \text{if } \min_{\{i|\sigma_i(A)>0\}} \sigma_i(A) = \sigma_r(A) \geq \sqrt{\varepsilon}, \\ \mathrm{diag}_{m,n}[\sigma_1(A), \dots, \sigma_k(A), 0, \dots, 0] & \text{if there is an integer } k \\ & \text{such that } \sigma_k(A) \geq \sqrt{\varepsilon} > \sigma_{k+1}(A). \end{cases}$$

If $A = 0$, which amounts to having $r = 0$, $(\mathrm{rank})_\varepsilon(A) = 0$, so that the formula (1.5) is still valid (with the usual rule that summing over the empty set gets at 0).

We may complete the result *(ii)* in the theorem above by determining the whole set $\mathrm{Prox}_\varepsilon(\mathrm{rank})(A)$. Indeed, we have four cases to consider:

- If $\max_i \sigma_i(A) = \sigma_1(A) < \sqrt{\varepsilon}$, then $\mathrm{Prox}_\varepsilon(\mathrm{rank})(A) = \{0\}$.
- If $\min_{\{i|\sigma_i(A)>0\}} \sigma_i(A) = \sigma_r(A) > \sqrt{\varepsilon}$, then $\mathrm{Prox}_\varepsilon(\mathrm{rank})(A) = \{A\}$.
- If there is an integer k such that $\sigma_k(A) > \sqrt{\varepsilon} > \sigma_{k+1}(A)$, then

$$\mathrm{Prox}_\varepsilon(\mathrm{rank})(A) = \left\{U\mathrm{diag}_{m,n}[\sigma_1(A), \dots, \sigma_k(A), 0, \dots, 0]V\right\}.$$

In all the three cases above, $\mathrm{Prox}_\varepsilon(\mathrm{rank})$ is single-valued at A.

- Suppose there is an integer k such that $\sigma_k(A) = \sqrt{\varepsilon}$. We define

$$k_0 := \min\{k| \quad \sigma_k(A) = \sqrt{\varepsilon}\},$$

$$k_1 := \max\{k| \quad \sigma_k(A) = \sqrt{\varepsilon}\}.$$

Then, $\mathrm{Prox}_\varepsilon(\mathrm{rank})(A)$ is the set of matrices of the form $\tilde{U}\mathrm{diag}_{m,n}(\tau_1, \dots, \tau_p)\tilde{V}$, where $(\tilde{U}, \tilde{V}) \in O(m,n)_A$ and

$$\tau_i = \begin{cases} \sigma_i(A) & \text{if } i \leq k \\ 0 & \text{otherwise} \end{cases},$$

where k is an integer between k_0 and k_1.

Before going into the proof of Theorem 1.2, a couple of comments are in order.

Comment 1. There are other ways to express $(\mathrm{rank})_\varepsilon(A)$, different from (although equivalent to) the one in (1.5). For example, taking into account the relation $\|A\|_F^2 = \sum_{i=1}^r \sigma_i^2(A)$, we get at

$$(\mathrm{rank})_\varepsilon(A) = \sum_{i=1}^{r} \min\left[1, \frac{\sigma_i^2(A)}{\varepsilon}\right]. \tag{1.6}$$

Suppose now that one wishes to express $(\mathrm{rank})_\varepsilon(A)$ in terms of traces of matrices, without any (explicit) reference to the singular values of A. Indeed, $A^TA - \varepsilon I_n$ is a symmetric matrix whose eigenvalues are $\sigma_1^2(A) - \varepsilon, \ldots, \sigma_r^2(A) - \varepsilon, -\varepsilon, \ldots, -\varepsilon$. Its projection on the closed convex cone $\mathcal{S}_n^+(\mathbb{R})$ of positive semidefinite symmetric matrices has eigenvalues

$$\left[\sigma_1^2(A) - \varepsilon\right]^+, \ldots, \left[\sigma_r^2(A) - \varepsilon\right]^+, 0, \ldots, 0 \quad \text{(see [2] or [4]).}$$

Thus, an alternate expression for $(\mathrm{rank})_\varepsilon(A)$ is:

$$(\mathrm{rank})_\varepsilon(A) = \frac{1}{\varepsilon}\mathrm{tr}\left(A^TA\right) - \frac{1}{\varepsilon}\mathrm{tr}\left[P_{\mathcal{S}_n^+(\mathbb{R})}\left(A^TA - \varepsilon I_n\right)\right]. \tag{1.7}$$

Comment 2. If ε is small enough, say if $\varepsilon \leq \sigma_r^2(A)$, then

$$(\mathrm{rank})_\varepsilon(A) = \mathrm{rank}\ A. \tag{1.8}$$

This easily comes from (1.5) since, in that case, $\sigma_i^2(A) - \varepsilon \geq 0$ for all $i = 1, 2, \ldots, r$ and $\sum_{i=1}^r \sigma_i^2(A) = \|A\|_F^2$. Therefore, the general convergence result that is known for the Moreau envelopes f_ε of f (recalled at the beginning of Sect. 1.2) is made much stronger here: $f_\varepsilon(A) = f(A)$ for ε small enough!

It may be destabilizing to accept that, for ε small enough, the formulas (1.5) or (1.7) for $(\mathrm{rank})_\varepsilon(A)$ produce an integer!

Let us end this comment with a trap in which one could fall: Yes, σ_r is a continuous function of A; but one cannot secure that, for ε small enough, $(\mathrm{rank})_\varepsilon(B) = \mathrm{rank}\ B$ for B in a neighborhood of A; this is due to the fact that, in the required threshold, $\varepsilon \leq \sigma_r^2(A)$, the quantity r (=rank of A) depends also on A.

Proof. (of Theorem 1.2)

To find the lower bound over the whole space $\mathcal{M}_{m,n}(\mathbb{R})$ in (1.3), we divide $\mathcal{M}_{m,n}(\mathbb{R})$ into the "strata" Σ_k, and calculate the individual lower bounds

$$\inf_{M \in \Sigma_k} \left\{\mathrm{rank}\ M + \frac{1}{\varepsilon}\|M - A\|_F^2\right\}$$

over Σ_k, for $k = 0, 1, \ldots, p$. We therefore have three different situations to consider: when $k = r(= \text{rank } A)$, when $k > r$, and when $k < r$. Here are the corresponding results:

- Let $k = r(= \text{rank } A)$. Since $A \in \Sigma_k$, we get immediately
$$\min_{M \in \Sigma_k} \left\{ \text{rank } M + \frac{1}{\varepsilon} \|M - A\|_F^2 \right\} = \text{rank } A = r. \tag{1.9}$$
- Let $k > r$. Then
$$\text{rank } M + \frac{1}{\varepsilon} \|M - A\|_F^2 \geq k > r \quad \text{for all } M \in \Sigma_k,$$
so that
$$\min_{M \in \Sigma_k} \left\{ \text{rank } M + \frac{1}{\varepsilon} \|M - A\|_F^2 \right\} \geq k > r. \tag{1.10}$$
- Let $k < r$. Then
$$\min_{M \in \Sigma_k} \left\{ \text{rank } M + \frac{1}{\varepsilon} \|M - A\|_F^2 \right\} = k + \frac{1}{\varepsilon} \left[\text{dist}\,(A, \Sigma_k) \right]^2.$$
But, in that case, we have observed that $[\text{dist}(A, \Sigma_k)]^2 = \sum_{i=k+1}^{r} \sigma_i^2(A)$ (*cf.* (1.2) and the comment preceding it). In short,
$$\min_{M \in \Sigma_k} \left\{ \text{rank } M + \frac{1}{\varepsilon} \|M - A\|_F^2 \right\} = k + \frac{1}{\varepsilon} \sum_{i=k+1}^{r} \sigma_i^2(A). \tag{1.11}$$

By collecting the results (1.9), (1.10) and (1.11), we get at

$$\begin{aligned}(\text{rank})_\varepsilon(A) &= \min_{k=0,\ldots,p} \min_{M \in \Sigma_k} \left\{ \text{rank } M + \frac{1}{\varepsilon} \|M - A\|_F^2 \right\} \\ &= \min_{k=0,\ldots,r} \left\{ k + \sum_{i=k+1}^{r} \frac{\sigma_i^2(A)}{\varepsilon} \right\} \\ &= \min_{k=0,\ldots,r} \left\{ \frac{\|A\|_F^2}{\varepsilon} + \sum_{i=1}^{k} \frac{1}{\varepsilon} \left[\varepsilon - \sigma_i^2(A) \right] \right\} \end{aligned} \tag{1.12}$$

(with the convention that $\sum_{i=r+1}^{r} \frac{\sigma_i^2(A)}{\varepsilon} = 0$).

Three cases are now to be considered:

- **Case 1:** $\sigma_i^2(A) \leq \varepsilon$ for all $i = 1, \ldots, r$. Then $k^* = 0$ is a solution in the minimization problem (1.12) and the value in (1.12) is
$$\sum_{i=1}^{r} \frac{\sigma_i^2(A)}{\varepsilon} = \frac{1}{\varepsilon} \|A\|_F^2.$$
Therefore, the matrix $B = 0$ is a minimizer in (1.3), *i.e.*, one element in $\text{Prox}_\varepsilon(\text{rank})(A)$.

- **Case 2:** $\sigma_i^2(A) \geq \varepsilon$ for all $i = 1, \ldots, r$. Then $k^* = r$ is a solution in the minimization problem (1.12) and the optimal value in (1.12) is r. Therefore, the matrix $B = A$ is a minimizer in (1.3), *i.e.* one element in $\mathrm{Prox}_\varepsilon(\mathrm{rank})(A)$.
- **Case 3** (the standard one): There is an integer $k_0 \in \{1, \ldots, r-1\}$ such that $\sigma_{k_0}^2(A) > \varepsilon > \sigma_{k_0+1}^2(A)$. Then $k^* = k_0$ is a solution in the minimization problem (1.12) and the optimal value in (1.12) is

  $$k_0 + \sum_{i=k_0+1}^{r} \frac{\sigma_i^2(A)}{\varepsilon}.$$

  A matrix B is a minimizer in (1.3), *i.e.* is an element of $\mathrm{Prox}_\varepsilon(\mathrm{rank})(A)$, when

  $$\mathrm{rank}\ B + \frac{1}{\varepsilon}\|A - B\|_F^2 = k_0 + \frac{1}{\varepsilon} \sum_{i=k_0+1}^{r} \frac{\sigma_i^2(A)}{\varepsilon}.$$

  The theorem of Eckart and Young tells us that such a B is provided by a "projection" of A on S_{k_0}.

  To summarize all the cases, the optimal value in (1.12) is

  $$(\mathrm{rank})_\varepsilon(A) = \sum_{i=1}^{r} \min\left[1, \frac{\sigma_i^2(A)}{\varepsilon}\right],$$

  the alternate form (1.6) of the expression (1.5), while a solution B in (1.3) is as announced in the statement (ii) of Theorem 1.2.

The case of multiple solutions k^* in the minimization problem (1.12) is handled similarly to the Case 3 above. If there is a k such that $\sigma_k^2(A) = \varepsilon$, then the solution set in the minimization problem (1.12) is $\{k_0 - 1, \ldots, k_1\}$, where

$$k_0 = \min\left\{k | \quad \sigma_k^2(A) = \varepsilon\right\},$$

$$k_1 = \max\left\{k | \quad \sigma_k^2(A) = \varepsilon\right\}.$$

Then, again by using the Eckart and Young theorem, all the minimizing matrices B in (1.3) are those described in the comments following Theorem 1.2. □

2. From Moreau envelopes to Eckart and Young theorem

Here we start with the (unconstrained) minimization problem

$$(\mathcal{P}_\varepsilon) \quad \begin{cases} \text{Minimize } \left\{\mathrm{rank}\ M + \frac{1}{\varepsilon}\|A - M\|_F^2\right\} \\ M \in \mathcal{M}_{m,n}(\mathbb{R}) \end{cases}.$$

The rank function is lower-semicontinuous and bounded from below, the function $\|A - M\|_F^2$ goes to infinity as $\|M\|_F$ goes to infinity, thus $(\mathcal{P}_\varepsilon)$ indeed has solutions. The question is: How does this minimization process help to solve our best approximation problem $(\mathcal{A}_k)$? Said otherwise: If M_ε is a minimizer in $(\mathcal{P}_\varepsilon)$, what do we know about its rank? Could we tune the parameter ε so that M_ε be of a prescribed rank k? We answer these questions by following the return path of the one followed in Section 1.

To begin with, we prove an easy technical lemma concerning Moreau envelopes (*cf.* beginning of Sect. 1.2). The result we are going to present is known in the convex case (*i.e.* when f is convex); it also holds true for non-convex f.

Lemma 2.1. *Let f_ε be the Moreau envelope of f obtained with the parameter $\varepsilon > 0$. Let $x \in E$ and let $\overline{u}$ be an element of $Prox_\varepsilon f(x)$. Then $\overline{u}$ is a "projection" of x on the sublevel-set of f at level $f(\overline{u})$. In other words,*

$$\|x-\overline{u}\| \leq \|x-u\| \text{ for all } u \text{ satisfying } f(u) \leq f(\overline{u}).$$

Proof. By definition of $\mathrm{Prox}_\varepsilon f(x)$, to have $\overline{u} \in \mathrm{Prox}_\varepsilon f(x)$ means

$$f(\overline{u}) + \frac{1}{\varepsilon}\|x-\overline{u}\|^2 \leq f(u) + \frac{1}{\varepsilon}\|x-u\|^2 \text{ for all } u \in E. \tag{2.1}$$

Choose u satisfying $f(u) \leq f(\overline{u})$. The above inequality then yields

$$\|x-\overline{u}\| \leq \|x-u\|.$$

Thus, the announced result is proved. □

We now go back to our initial problem $(\mathcal{A}_k)$; we wish to prove Eckart and Young theorem with the help of Moreau envelopes. Let therefore $0 \neq A \in \mathcal{M}_{m,n}(\mathbb{R})$ of rank r and let $1 \leq k < r$. With this integer k given, how to choose the tuning parameter ε? We distinguish two cases.

Theorem 2.2. *Suppose that $\sigma_k(A) > \sigma_{k+1}(A)$. Choose ε such that*

$$\sigma_k(A) > \sqrt{\varepsilon} > \sigma_{k+1}(A).$$

Then, $(\mathcal{P}_\varepsilon)$ has a unique solution M_ε. This matrix M_ε is of rank k, it is the "projection" of A on S_k, that is the unique solution of $(\mathcal{A}_k)$. Moreover

$$\|A-M_\varepsilon\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2(A).$$

Proof. We read the proof of Theorem 1.2 backwards. We have

$$(\mathrm{rank})_\varepsilon(A) = \mathrm{rank}\ M_\varepsilon + \frac{1}{\varepsilon}\|M_\varepsilon - A\|_F^2 \tag{2.2}$$

$$= \min_{l=0,\ldots,r}\left\{l + \frac{1}{\varepsilon}\sum_{i=l+1}^{r}\sigma_i^2(A)\right\}. \tag{2.3}$$

With the choice of ε that we have made, k is the unique solution in (2.3). So, in (2.2),

- rank $M_\varepsilon = k$ and $\|M_\varepsilon - A\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2(A)$;
- according to Lemma 2.1, M_ε is a "projection" of A on the sublevel-set of the rank function at level rank $M_\varepsilon = k$ (*i.e.*, a solution of $(\mathcal{A}_k)$). □

The case where $\sigma_k(A) = \sigma_{k+1}(A)$ is a little more subtle to treat; indeed, in that case, there are several integers which solve (2.3) and the corresponding matrix solutions in (2.2) all do not have the same rank.

Theorem 2.3. *Suppose that $\sigma_k(A) = \sigma_{k+1}(A)$ and choose $\sqrt{\varepsilon}$ as its common value. Denote*

$$k_0 = \min\left\{i| \quad \sigma_i(A) = \sqrt{\varepsilon}\right\},$$

$$k_1 = \max\left\{i| \quad \sigma_i(A) = \sqrt{\varepsilon}\right\}.$$

Then, any integer between k_0 and k_1 is a solution of (2.3), so that the solution matrices M_ε in (2.2) have a rank between k_0 and k_1.

For $k \in \{k_0, \ldots, k_1\}$, the solution matrix M_ε of rank k in (2.2) is a "projection" of A on S_k, that is to say, a solution of $(\mathcal{A}_k)$.

Proof. The same, *mutatis mutandis*, as that of Theorem 2.2. □

3. By way of conclusion

The Eckart and Young theorem allowed us to calculate explicity the Moreau envelopes of the rank function, an objective which was not obvious at all, due to bumpy behavior of this function; various expressions of these Moreau envelopes have been provided (formulas (1.5), (1.6), (1.7), Theorem 1.2).

Conversely, if we want to get a best approximation of A of rank at most k, we could get at it by solving the unconstrained minimization problem

$$\text{Minimize}_M \left\{ \text{rank } M + \frac{1}{\varepsilon}\|A - M\|_F^2 \right\},$$

where the parameter $\varepsilon > 0$ is tuned in function of k (Theorem 2.2, Theorem 2.3).

Acknowledgements. We thank D.O'Leary (University of Maryland) for having pointed out a flaw in the description of the projection set on S_k in the case where σ_k is not distinguishable from σ_{k+1}, and for providing additional references like [8]. We also thank the various readers who, after the first circulation of this note, provided comments and suggested improvements.

References

[1] U. Helmke and J.B. Moore, *Optimization and Dynamical Systems*. Spinger Verlag (1994).

[2] N. Higham, *Matrix nearness problems and applications*, in M.J.C Gover and S. Barnett, eds., *Applications of Matrix Theory*. Oxford University Press (1989) 1–27.

[3] J.-B. Hiriart-Urruty and H.Y. Le, A variational approach of the rank function. TOP (2013) DOI: 10.1007/s11750-013-0283-y.

[4] J.-B. Hiriart-Urruty and J. Malick, A fresh variational analysis look at the world of the positive semidefinite matrices. *J. Optim. Theory Appl.* **153** (2012) 551–577.

[5] J.-J. Moreau, *Fonctions convexes duales et points proximaux dans un espace hilbertien.* (French) C. R. Acad. Sci. Paris **255** (1962) 2897–2899 (Reviewer: I.G. Amemiya) 46.90.

[6] J.-J. Moreau, *Propriétés des applications "prox"*. C. R. Acad. Sci. Paris **256** (1963) 1069–1071.

[7] R.T. Rockafellar and R.J.-B. Wets, *Variational analysis.* Springer (1998).

[8] G.W. Stewart, *Matrix algorithms, Basic decompositions*, Vol. I. Society for Industrial and Applied Mathematics, Philadelphia, PA (1998).