THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Université Toulouse III Paul Sabatier (UT3 Paul Sabatier)

Discipline ou spécialité :
Domaine mathématiques – Mathématiques appliquées

Présentée et soutenue par
Hai Yen LE
le : 22 mai 2013

Titre :
Approche Variationnelle De La Fonction Rang : Relaxation Convexe ; Sous-différentiation Généralisée ; Régularisation-approximation De Moreau

A Variational Look At The Rank Function : Convex Relaxation ; Generalized Subdifferentiation ; Moreau's Regularization-approximation.

École doctorale :
Mathématiques Informatique Télécommunications (MITT)

Unité de recherche :
UMR 5219

Directeur de thèse :
Jean-Baptiste HIRIART-URRUTY, Université Paul Sabatier à Toulouse

Rapporteurs :
Russell LUKE, Université de Gottingen
Michel THÉRA, Université de Limoges

Autres membres du jury :
Jérôme BOLTE, Université Toulouse 1 Capitole
Abderrahim JOURANI, Université de Bourgogne à Dijon
François MALGOUYRES, Université Paul Sabatier à Toulouse

To my parents.

Acknowledgements

I would never have been able to finish my dissertation without the guidance, the help and support of the kind people around me, to only some of whom it is possible to give particular mention here.

First and foremost, I would like to express my gratitude to my adviser Prof. Jean-Baptiste Hiriart-Urruty, for his continuous support, excellent guidance, caring and patience. I feel very fortunate to have had him as an adviser and a teacher.

My sincere thanks are due to my referess, Prof. Russell Luke and Prof. Michel Théra, for their detailed review and comments. I would also like to thank Prof. Jérôme Bolte, Prof. Abderrahim Jourani, and Prof. François Malgouyres for being a member of the committee.

I am thankful to Prof. Tien Zung Nguyen for his helpful advice about the academic research. I am grateful to all friends who have supported me through the years in Toulouse. In particular, I would like to thank Chinh, Elissar, V.Minh, H.Minh, Ngoc and Lien.

I would like to thank my boyfriend Hung who is also a PhD student in mathematics for many fruitful discussions and his understanding. He was always there cheering me up and stood by me through the good and bad times. My deepest gratitude and love belong to my parents Nhat and Dung, my sister Hoa for their unconditional love and support.

Contents

Introduction

Beside the trace and the determinant, the rank function is one of the most important functions in matrix theory. Its properties have been studied in linear algebra or matrix calculus ([60],[26]), semi-algebraic geometry ([55]), *etc.*

In this dissertation, we consider the rank function *from the variational point of view*. The reason why we are interested in the rank function from this point of view is that the rank function appears as an objective (or constraint) function in various modern optimization problems. Many notions in engineering applications such as the order, complexity, or dimension of a model or design can be expressed as the rank of a matrix. The simplest model that can be expressed as a rank minimization problem or a rank constraint problem is always preferred to a complicated detailed one. More precisely, a low-rank matrix could correspond to a low-order controller for a system, a low-order statistical model and a design with a small number of components. For example:

- **Low rank matrix completion:** We are given a random subset of entries of a matrix and would like to fill in the missing entries such that the resulting matrix has the lowest possible rank. This problem is often encountered in the analysis of incomplete data sets exhibiting an underlying factor model with applications in collaborative filtering, computer vision and control.
- **Rank of a covariance matrix:** From noisy data, we obtain the estimated covariance matrices. Because of the noise, the estimated covariance matrices have full rank (with probability one). We want to find a covariance matrix of low rank such that the error is less than a given tolerance. In this example, a low rank covariance matrix corresponds to a simple explanation or model for the data.

- **Image approximation:** The general problem of image compression is to reduce the amount of data required to represent a digital image or video, and the underlying basis of the reduction process is the removal of redundant data. A two-dimensional image can be associated with a rectangular matrix, and in order to compress the given image, we need to find a low-rank approximation of the associated matrix.

The so-called "rank minimization problem" can be formulated as follows:

$$(\mathcal{P}) \quad \begin{cases} \text{Minimize } f(A) := \text{rank of } A \\ \text{subject to } A \in \mathcal{C}, \end{cases}$$

where $\mathcal{C}$ is a subset of $\mathcal{M}_{m,n}(\mathbb{R})$ (the vector space of m by n real matrices). The constraint set is usually rather "simple" (expressed as linear equalities, for example), the main difficulty lies in the objective function.

The "rank constraint problem" can be formulated as follows:

$$(\mathcal{P}_1) \quad \begin{cases} \text{Minimize } g(A) \\ \text{subject to } A \in \mathcal{C} \text{ and rank } A \leq k, \end{cases}$$

with a rather simple objective function but a fairly complicated constraint set. Both problems $(\mathcal{P})$ and $(\mathcal{P}_1)$ suffer from the same intrinsic difficulty: the occurence of the rank function.

A related (or cousin) problem to $(\mathcal{P})$, actually a special case of $(\mathcal{P})$, stated in $\mathbb{R}^n$ this time, consists in minimizing the so-called counting function $x = (x_1, \ldots, x_n) \in \mathbb{R}^n \longmapsto c(x) :=$ number of nonzero components x_i of x:

$$(\mathcal{Q}) \quad \begin{cases} \text{Minimize } c(x) \\ \text{subject to } x \in S, \end{cases}$$

where S is a subset of $\mathbb{R}^n$. Often $c(x)$ is denoted as $\|x\|_0$, although it is not a norm.

Problem $(\mathcal{Q})$ is always referred to as *cardinality minimization* and is known to be NP-hard. Minimization of the l_1 norm of vectors is a well-known heuristic method for the cardinality minimization problem and widely used in image denoising, sparse approximation, *etc.* Recently, Candès and Tao ([10]) and Donoho ([15])

proposed some conditions under which the l_1 heuristic can be *a priori* guaranteed to yield an optimal solution.

Problems $(\mathcal{P})$ and $(\mathcal{Q})$ are actually equivalent in terms of difficulty. In some particular cases, the rank minimization problems can be solved by using the singular values decomposition or can be reduced to the solution of linear systems. But, in general, the problem $(\mathcal{P})$ is NP-hard and, so, is a challenging nonconvex optimization problem.

Many heuristic algorithms, based on alternating inequalities, linearization and augumented Lagrange methods, have been proposed to solve the problem $(\mathcal{P})$. In 2007, Fazel introduced an heuristic method that minimizes the nuclear norm, *i.e.* the sum of singular values, over the constraint set. And she also provided the theorical support for the use of the nuclear norm: "The convex envelope of the rank function restricted to the unit ball for the spectral norm is the nuclear norm" ([22]). The nuclear norm is not only convex but also continuous, thus numerous efficient methods can be applied to solve the nuclear norm minimization problems. Moreover, Recht, Fazel and Parrilo ([56]) showed that under some suitable conditions ("restricted isometry property"), such a convex relaxation is tight in the case where the constraint set $\mathcal{C}$ is an affine manifold.

In this dissertation, we provide several properties of the rank function from the variational point of view: additional proofs for the closed convex relaxation, the expressions of the general subdifferentials and the Moreau regularization-approximation. The general method that we use to study these properties is based on the relationship between the counting function and the rank function.

In Chapter 1, we recall the definition of the convex envelope (or convex hull) of a function, its properties and the relationship between the original function and its convex envelope. Then, the Fazel's theorem and her proof (via the calculation the biconjugate of the restricted rank function) is presented in Section 1.4. We also provide two new proofs of this theorem. The first proof is based on the relationship between the rank and the counting function, a result of Lewis and the convex hull of the counting function restricted to a ball (Theorem 1.13). The second proof is geometrical, it is obtained by computing the convex hull of the sub-level sets of the rank function (Theorem 1.15).

In Chapter 2, we begin by introducing the definitions and properties of the generalized subdifferentials (the proximal, Fréchet, limiting and Clarke one) of

a lower-semicontinuous function. As a result, all types of generalized subdifferentials of the counting function coincide and an explicit formula of the common subdifferential is given in Theorem 2.9 and Theorem 2.10. Then, thanks to theorems of LEWIS and SENDOV ([48],[49]), we obtain the corresponding generalized subdifferentials of the rank function (Theorem 2.14). All types of generalized subdifferentials of the rank function also coincide. And we observe that the generalized subdifferential of the rank function is always a vector space. Certainly, 0 always belongs to the generalized subdifferential of the rank function at any point. This was foreseen by the fact that "Every point is a local minimizer of the rank function" ([32]). Finally, thanks to an alternate expression of the common subdifferential of the rank function (Prop 2.17), we provide its dimension.

In Chapter 3, we consider another way to approach the rank minimization problem - using smooth or just continuous approximations of the rank function. Two examples of smooth versions of the rank were provided by HIRIART-URRUTY ([29]) in 2010 and ZHAO ([61]) in 2012. We present here the regularization-approximation relying on the so-called Moreau-Yosida technique, widely used in the context of variational analysis. Although the rank function is a bumpy one, it is amenable to such an approximation-regularization process, and we get the explicit forms of the Moreau-Yosida approximation-regularization of the rank and of the restricted rank function in terms of singular values (Theorem 3.6 and Theorem 3.7). We also provide the generalized subdifferentials of this approximation; then thanks to a theorem of JOURANI ([42]), we can retrieve the FRÉCHET subdifferential of the rank function.

In the last Chapter, we study the cp-rank function of completely positive matrices. This function shares several common properties with the rank function such as being lower-semicontinuous, subadditive. Moreover, the convex envelope of the cp-rank function restricted to the unit ball (for the nuclear norm) is also the nuclear norm. Finally, we propose two open questions about the upper bound and the generalized subdifferentials of the cp-rank function.

Summary:

Principal results and publications.

- **Principal results in Chapter 1:**

  Theorem 1.13, Theorem 1.15

  Corresponding papers:

  J.-B.Hiriart-Urruty and H.Y.Le, *Convexifying the set of matrices of bounded rank: applications to the quasiconvexification and convexification of the rank function*, Optimization Letters, Vol 6(5) (2012), 841–849.

  H.Y.Le, *Confexifying the counting function on $\mathbb{R}^p$ for convexifying the rank function on $\mathcal{M}_{m,n}(\mathbb{R})$*, J. of Convex Analysis, Vol 19(2) (2012), 519-524.

- **Principal results in Chapter 2:**

  Theorem 2.14

  Corresponding paper:

  H.Y.Le, *The generalized subdifferentials of the rank function*, Optimization Letters (2012), DOI: 10.1007/s11590-012-0456-x.

- **Principal results in Chapter 3:**

  Theorem 3.6, Theorem 3.7

  Corresponding paper:

  J.-B.Hiriart-Urruty and H.Y.Le, *From Eckart & Young approximation to Moreau envelopes and vice versa*. Preprint 2012. Submitted.

- **Principal results in Chapter 4:**

  Theorem 4.8

Survey paper:

J.-B.Hiriart-Urruty and H.Y.Le, *A variational approach of the rank function*, TOP (Journal of the Spanish Society of Statistics and Operations Research). DOI: 10.1007/s11750-013-0283-y.

List of Figures

Notations

$\mathcal{M}_{m,n}(\mathbb{R})$	The set of real $m \times n$ matrices
$\mathcal{M}_n(\mathbb{R})$	The set of real $n \times n$ matrices
$\mathcal{S}_n(\mathbb{R})$	The set of real symmetric $n \times n$ matrices
A^T	The transpose of A
rank A	The rank of A
tr A	The trace of A
det A	The determinant of A
$\|A\|_F$	The Frobenius norm of A
$\|A\|_{sp}$	The spectral norm of A
$\|A\|_*$	The nuclear norm of A
$\sigma_(A)$	The vector of singular values of A
$\sigma_i(A)$	The ith largest singular value of A
$O(m)$	The set of real orthogonal $m \times m$ matrices
cp-rank A	The cp-rank of A
$\partial^F f$	The Fréchet subdifferential of f
$\partial^L f$	The limiting subdifferential of f
$\partial^V f$	The viscosity subdifferential of f
$\partial^P f$	The proximal subdifferential of f
$\partial^F f$	The Clarke subdifferential of f
epi J	The epigraph of J
dom J	The domain of J
co J	The convex hull of J

$\overline{\text{co}}J$	The closed convex hull of J
RMP	Rank Minimization Problem
RCP	Rank Constraint Problem
CP	Completely positive

Chapter 1

The closed convex relaxation of the rank function

In this chapter, we recall the generalities on the relaxation operation, then introduce the rank minimization problem and the relaxed form of it (given by Fazel *[22]). After that, we use several methods to achieve the* Fazel*'s theorem. At last, thanks to a result of* Recht *et al ([56]), we can understand the link between the original problem and the relaxed form.*

1.1 Generalities on the relaxation operation

In optimization or variational calculus, we usually study the minimization problem expressed as:

$$(\mathcal{P}) \begin{cases} \textit{Minimize } J(x) \\ x \in S \end{cases}$$

where $J : X \to \mathbb{R} \cup \{+\infty\}$ and $S \subset X$.

In general, this problem is very hard because we have no property of J, S and X. Then, it is natural to replace the problem $(\mathcal{P})$ by the relaxed problem which is obtained by substitute $\widehat{J}$ for J or "relax" the constraint set S by enlarging it, or "enrich" the underlying space X.

In a variational context, when we deal with the minimization of $J : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$, one usually considers the closed convex hull of J. In the present approach, we are not going to consider the more general framework, but indeed the way in which we relax in passing from J to its closed convex hull. Henceforth, the context is as follows:

$J : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ is not identically equal to $+\infty$, and it is minorized by some affine function, i.e. for some $(s, b) \in \mathbb{R}^n \times \mathbb{R}$,

$$J(x) \geq \langle s, x \rangle - b \text{ for all } x \in \mathbb{R}^n. \tag{1.1}$$

First of all, we recall some definitions and notations:

- The *domain* of J is the nonempty set:

$$\mathrm{dom} J := \{x \in \mathbb{R}^n : J(x) < +\infty\}.$$

- The *epigraph* of J is the nonempty set:

$$\mathrm{epi} J := \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : r \geq J(x)\}.$$

- The *sub-level set* of J at level $r \in \mathbb{R}$ is defined by:

$$[J \leq r] := \{x \in \mathbb{R}^n : J(x) \leq r\}.$$

- J is said closed if it is lower-semicontinuous everywhere, or if its epigraph is closed or if all its sub-level sets are closed.

- The class of all convex functions is denoted by $\mathrm{Conv}\mathbb{R}^n$ and the class of all closed convex functions is denoted by $\overline{\mathrm{Conv}}\mathbb{R}^n$.

1.1.1 Closed convex hull of a function

Proposition 1.1. *The functions below*

$$\bar{J}_1(x) := \inf \{r : (x, r) \in \overline{\mathrm{co}}\ \mathrm{epi} J\},$$

$$\bar{J}_2(x) := \sup \left\{h(x) : h \in \overline{\mathrm{Conv}}\mathbb{R}^n, h \leq g\right\},$$

$$\bar{J}_3(x) := \sup\{\langle s, y\rangle - b : \langle s, y\rangle - b \leq J(y) \ \textit{for all}\ y \in \mathbb{R}^n\}$$

are closed, convex, and coincide on $\mathbb{R}^n$.

Proof. See [36, page 100]. □

Definition 1.2. The common function $\bar{J}_1 = \bar{J}_2 = \bar{J}_3$ of the Proposition 1.1 is called the *closed convex hull* or *closed convex envelope* of J and is denoted by $\overline{\text{co}}J$.

By definition, we have at least two ways of constructing $\overline{\text{co}}J$:

- the "internal construction": consider all the convex combinations of elements of the epigraph epi$J$ of J, so that co(epi$J$) is built, and then close it; the set $\overline{\text{co}}$(epi$J$) is the epigraph of a function, namely of $\overline{\text{co}}J$.
- the "external construction": consider all the continuous affine functions a_J minorizing J; then $\overline{\text{co}}J = \sup a_J$.

Recall that the Legendre - Fenchel *conjugate* of J is the function J^* defined by:

$$\mathbb{R}^n \ni s \mapsto J^*(s) = \sup\{\langle s, x\rangle - J(x) : x \in \text{dom}J\}.$$

J satisfies (1.1) , and so is J^*. So we can compute the *biconjugate* function of J. For all $x \in \mathbb{R}^n$,

$$J^{**}(x) := (J^*)^*(x) = \sup\{\langle s, x\rangle - J^*(s) : s \in \mathbb{R}^n\}.$$

The function J^{**} turns out to be the closed-convex hull $\overline{\text{co}}J$, *i.e.*

$$J^{**} = \overline{\text{co}}J.$$

If J is lower-semicontinuous and coercive (that is to say, if $\lim_{|x|\to+\infty} J(x) = +\infty$), then:

$$J^{**} = \text{co}J,$$

where co$J$ is the convex hull or convex envelope of J, *i.e.* the largest convex function minorizing J.

1.1.2 Properties

Proposition 1.3. *(From J to $\overline{co}J$)*

(i) The infimal values. *We have*

$$\inf_{x\in\mathbb{R}^n} J(x) = \inf_{x\in\mathbb{R}^n} (\overline{\text{co}}J)(x) \qquad (\textit{an equality in } \mathbb{R}\cup\{-\infty\}). \tag{1.2}$$

(ii) The set of minimizers. *If we denote by* $\text{argmin}J$ *the set of* $x\in\mathbb{R}^n$ *minimizing* J *on* $\mathbb{R}^n$ *(possibly, the empty set), we have that:*

$$\overline{\text{co}}(\text{argmin}J) \subset \text{argmin}(\overline{\text{co}}J). \tag{1.3}$$

Proof. (i) Because $J^{**} = \overline{\text{co}}J$, then $J^* = (\overline{\text{co}}J)^*$. Therefore,

$$\inf_{x\in\mathbb{R}^n} J(x) = -J^*(0) = -(\overline{\text{co}}J)^*(0) = \inf_{x\in\mathbb{R}^n} (\overline{\text{co}}J)(x).$$

(ii) $\overline{\text{co}}J$ is the closed-convex hull of J, then

$$\overline{\text{co}}J \leq J,$$

with (1.2), we infer that

$$\text{argmin}J \subset \text{argmin}(\overline{\text{co}}J).$$

Moreover, since $\text{argmin}(\overline{\text{co}}J)$ is closed and convex, we then have

$$\overline{\text{co}}(\text{argmin}J) \subset \text{argmin}(\overline{\text{co}}J).$$

□

Theorem 1.4. *Let* $J : \mathbb{R}^n \longrightarrow \mathbb{R}\cup\{+\infty\}$ *be differentiable at* $\overline{x}$*. Then the two following statements are equivalent:*

(i) $\overline{x}$ *is a global minimizer of* J

(ii) $\nabla J(\overline{x}) = 0$ *and* $J(\overline{x}) = (\overline{co}J)(\overline{x})$*, where* $\nabla J(x)$ *denotes the gradient of* J *at* x*.*

Proof.

(i)⇒(ii) If $\overline{x}$ is a global minimizer of J, then

$$\nabla J(\overline{x}) = 0.$$

From the fact that $x \in \operatorname{argmin} J$ and the properties (1.3) (Prop 1.3),

$$x \in \operatorname{argmin}(\overline{\text{co}}J).$$

This means that

$$\begin{aligned} \overline{\text{co}}J(\overline{x}) &= \inf_{x\in\mathbb{R}^n} (\overline{\text{co}}J)(x) \\ &= \inf_{x\in\mathbb{R}^n} J(x) \\ &= J(\overline{x}). \end{aligned}$$

(ii)⇒(i) Let $\overline{x}$ satisfy $\nabla J(\overline{x}) = 0$ and $J(\overline{x}) = (\overline{co}J)(\overline{x})$.

For $d \in \mathbb{R}^n$

$$\frac{J(\overline{x}+td) - J(\overline{x})}{t} \longrightarrow \langle \nabla J(\overline{x}), d\rangle \text{ when } t \to 0^+,$$

$$\frac{(\overline{\text{co}}J)(\overline{x}+td) - (\overline{\text{co}}J)(\overline{x})}{t} \longrightarrow (\overline{\text{co}}J)'(\overline{x}, d) \text{ when } t \to 0^+,$$

where $(\overline{\text{co}}J)'(\overline{x}, d)$ stands for the directional derivative of the convex function $\overline{\text{co}}J$.

Moreover,

$$\frac{(\overline{\text{co}}J)(\overline{x}+td) - (\overline{\text{co}}J)(\overline{x})}{t} \leq \frac{J(\overline{x}+td) - J(\overline{x})}{t}.$$

Then,

$$(\overline{\text{co}}J)'(\overline{x}, d) \leq \langle \nabla J(\overline{x}), d\rangle .$$

Hence, $\overline{\text{co}}J$ is differentiable at $\overline{x}$ and $\nabla(\overline{\text{co}}J)(\overline{x}) = 0$. Since $\overline{\text{co}}J$ is convex, $\overline{x}$ is a minimizer of $\overline{\text{co}}J$, we have

$$(\overline{\text{co}}J)(x) \geq \overline{\text{co}}J(\overline{x}) \quad \forall x \in \mathbb{R}^n.$$

Thus,

$$J(x) \geq J(\overline{x}) \quad \forall x \in \mathbb{R}^n.$$

□

Remark 1.5.

1. If the property "$\overline{x}$ is a local minimizer of J" replaces "$\nabla J(\overline{x}) = 0$", in absence of differentiability of J at $\overline{x}$, then the equivalence of Theorem 1.4 breaks down (see Fig 1.1).

2. This theorem will still be true if we replace $\mathbb{R}^n$ by a Hilbert space.

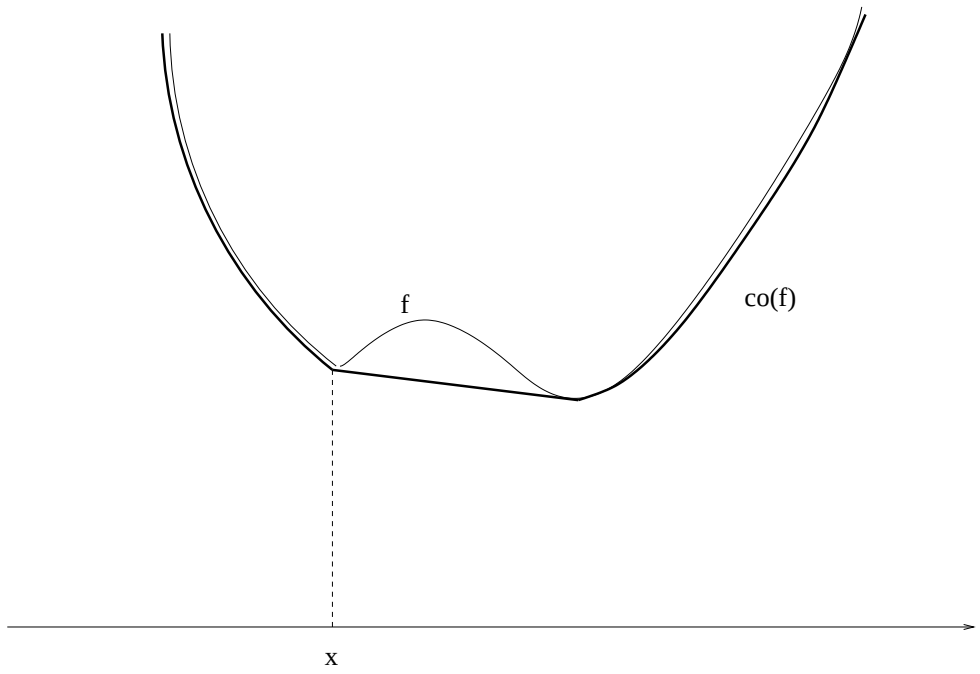


FIGURE 1.1: x is a local minimiser of f
$\overline{\text{co}}f(x) = \text{co}f(x) = f(x)$, but x is not a global minimiser of f.

Now, we will continue with considering some other interesting properties of the closed-convex hull of J

- The *continuity property.* Even if J is the restriction of a $\mathcal{C}^\infty$ function on a compact convex subset C of $\mathbb{R}^n$ (and $+\infty$ out of C), the (convex) function $\overline{co}(J)$ may not be continuous at some boundary point of C.

- The *differentiability property.* If $J : \mathbb{R} \to \mathbb{R}$ is differentiable on $\mathbb{R}$, then so is $\overline{co}J$ (even if $(\overline{co}J)(x) < J(x)$ for all $x \in \mathbb{R}$). There are however $\mathcal{C}^\infty$ functions $J : \mathbb{R}^2 \to \mathbb{R}$ for which $\overline{\text{co}}J$ is no more differentiable on $\mathbb{R}^2$. An example of such a function can be found in [6].

- *Behavior at infinity.* Indeed $\overline{\text{co}}J \leq J$. However, $\overline{\text{co}}J$ ends by "behaving like J at infinity".

  Theorem 1.6. *We have*

$$\liminf_{\|x\| \to +\infty} \frac{J(x) - (\overline{\text{co}}J)(x)}{\|x\|} = 0.$$

  Proof. Since $J(x) \geq (\overline{\text{co}}J)(x)$ for all $x \in \mathbb{R}^n$, the above lim inf is $l \geq 0$ (possibly $+\infty$). Suppose $l > 0$. Therefore, there exist $c > 0$ and $A > 0$ such

that:

$$\inf_{\|y\|\geq\|x\|} \frac{J(y) - (\overline{\text{co}}J)(y)}{\|y\|} \geq c \text{ whenever } \|x\| > A.$$

Thus, $J(x) - (\overline{\text{co}}J)(x) \geq c\|x\|$ when $\|x\| > A$, while $J(x) - (\overline{\text{co}}J)(x) \geq 0$ otherwise. In short,

$$J(x) \geq (\overline{\text{co}}J)(x) + c(\|x\| - A) \text{ for all } x \in \mathbb{R}^n.$$

This comparison result between the two functions above, the one on the right-hand side being convex, yeilds

$$(\overline{\text{co}}J)(x) \geq (\overline{\text{co}}J)(x) + c(\|x\| - A) \text{ for all } x \in \mathbb{R}^n.$$

This does not hold true for $\|x\| > A$. Hence, the hypothesis at the beginning of the proof, $l > 0$, is wrong. □

1.2 The rank minimization problem and the rank constrained problem

The Rank Minimization Problem (RMP) and the Rank constrained Problem (RCP) are optimization problems where the rank function appears respectively in the objective or in the constraints. They can be formulated as follows

$$(RMP) \quad \begin{cases} \text{minimize} & \text{rank } A \\ \text{subject to} & A \in C \end{cases}$$

and

$$(RCP) \quad \begin{cases} \text{minimize} & f(A) \\ \text{subject to} & A \in C \\ & \text{rank } A \leq k \end{cases}.$$

Such problems appear in many areas like: control, statistics, signal processing, computational geometry and combinatorical optimization. Some special cases can be solved with a special algorithm. For example, Eckart and Young found the distance from an abitrary matrix to the set of matrices with rank less than k in 1936 ([28]). But in general, these problems are NP-hard.

Now, we consider some examples of RMP and RCP ([22],[56]).

Low rank matrix completion. In the matrix completion problem, we are given a random subset of entries of a matrix and would like to fill in the missing entries such that the resulting matrix has the lowest possible rank. It is often encountered in the analysis of incomplete data sets exhibiting an underlying factor model with applications in collaborative filtering, computer vision, control.

Suppose that we are presented with a set of triples $(I(i), J(i), S(i))$ for $i = 1, \ldots, k$ and wish to find a matrix with $S(i)$ in the entry corresponding to row $I(i)$ and column $J(i)$ for all i. The matrix completion can be formulated as follows

$$\begin{cases} \text{minimize} & \text{rank } A \\ \text{subject to} & A_{I(i),J(i)} = S(i), \quad \text{for all } i = 1, \ldots, k \end{cases}$$

which is a special case of the rank minimization problems.

Image approximation. A simple and well-known method to compress two-dimensional images can be obtained by using the singular value decomposition. The basic idea is to associate to the given grayscale image a rectangular matrix A, with the entries A_{ij} corresponding to the gray level of the (i, j) pixel. The best rank-k approximation of A is given by

$$X^* = \arg \min_{\text{rank } X \leq k} \|A - X\|$$

where $\|.\|$ is any unitarily invariant norm. By the classical Eckart-Young-Mirsky theorem, the optimal approximation is given by a truncated singular value decomposition of A, *i.e.*, if $A = U\Sigma V^T$, then $X^* = U\Sigma_k V^T$, where the first k diagonal entries of Σ_k are the largest k singular values and the rest of the entries are zero.

Multivariate statistical data analysis. In this example, we have to deal with covariance matrices estimated from noisy data. In fact, the estimated covariance matrix has full rank because of the noise (with probability one). We want to find a covariance matrix Σ with the least rank such that the error is at most equal to a given positive number ϵ

$$\begin{cases} \text{minimize} & \text{rank } \Sigma \\ \text{subject to} & \|\Sigma - \hat{\Sigma}\|_F \leq \epsilon \\ & \Sigma \succeq 0 \\ & \Sigma \in \mathcal{C}, \end{cases}$$

where Σ is the optimization variable, $\hat{\Sigma}$ is the measured covariance matrix, $\mathcal{C}$ is a convex set denoting the prior information or assumptions on Σ, and $\|.\|_F$ denotes the classical Frobenius norm of a matrix (see in the next section). The constraint $\Sigma \succeq 0$ is necessary because Σ is a covariance matrix.

The Frisch problem. Let $x \in \mathbb{R}^n$ be a random vector, with covariance matrix Σ_x. Suppose that we have:

$$y(t) = x(t) + v(t)$$

where the measurement noise v has zero mean, is uncorrelated with x, and has an unknown but diagonal covariance matrix $D = \text{diag } d$. It follows that:

$$\Sigma_y = \Sigma_x + D,$$

where Σ_y denotes the covariance of y. The problem is to identify, from noisy observations, the largest number of linear relations among the underlying data. This corresponds to the minimum rank of Σ_x. We assume that Σ_y can be estimated with high confidence; *i.e.* we consider it known. This problem can be expressed as the following RMP:

$$\begin{cases} \text{minimize} & \text{rank } (\Sigma_y - D) \\ \text{subject to} & \Sigma_y - D \succeq 0 \\ & D \succeq 0 \\ & \text{D diagonal} \end{cases}.$$

Bilinear Matrix Inequality problems. Consider the following problem:

$$\begin{cases} \text{minimize} & c^T x \\ \text{subject to} & C + \sum_{i=1}^m x_i A_i + \sum_{i,j=1}^m x_i x_j B_{ij} \preceq 0, \end{cases}$$

where $x \in \mathbb{R}^n$ is the optimization variable, and $c \in \mathbb{R}^n$ and the symmetric matrices A_i, B_{ij}, C are given. This problem is very general, but also non-convex. We now show that this problem can be considered as a rank-constrained problem. This problem can be expressed as:

$$\begin{cases} \text{minimize} & c^T x \\ \text{subject to} & C + \sum_{i=1}^m x_i A_i + \sum_{i,j=1}^m w_{ij} B_{ij} \preceq 0 \\ & w_{ij} = x_i x_j \text{ for all } i, j = 1, \ldots, m \end{cases}. \tag{1.4}$$

The second constraint can be written as $W = xx^T$. This equality is equivalent to the following one:

$$\text{rank} \begin{bmatrix} W & x \\ x^T & 1 \end{bmatrix} = 1.$$

Therefore, the problem (1.4) is equivalent to:

$$\begin{cases} \text{minimize} & c^T x \\ \text{subject to} & C + \sum_{i=1}^{m} x_i A_i + \sum_{i,j=1}^{m} w_{ij} B_{ij} \preceq 0 \\ & \text{rank} \begin{bmatrix} W & x \\ x^T & 1 \end{bmatrix} \leq 1. \end{cases}$$

Combinatorial optimization problems. Many combinatorial optimization problems can be expressed as rank-constrained problems. Consider the quadratic optimization problem:

$$\begin{cases} \text{minimize} & x^T A_0 x + 2b_0^T x + c_0 \\ \text{subject to} & x^T A_i x + 2b_i^T x + c_i \leq 0 \text{ for all } i = 1, \ldots, L \end{cases}. \tag{1.5}$$

where $x \in \mathbb{R}^k$ is the optimization variable.

Define the new variable X as $X = xx^T$. As shown in the previous example, this can be written as:

$$\text{rank} \begin{bmatrix} X & x \\ x^T & 1 \end{bmatrix} = 1.$$

Note that

$$x^T A_i x = \text{tr } (A_i x x^T) = \text{tr}(A_i X),$$

so that we can write the quadratic terms in the objective function and the constraints in terms of X. Thus, problem (1.5) becomes equivalent to

$$\begin{cases} \text{minimize} & \text{tr}(A_0 X) + 2b_0^T x + c_0 \\ \text{subject to} & \text{tr}(A_i X) + 2b_i^T x + c_i \leq 0 \text{ for all } i = 1, \ldots, L \\ & \text{rank} \begin{bmatrix} X & x \\ x^T & 1 \end{bmatrix} \leq 1 \end{cases}.$$

1.3 The rank function

Let $\mathcal{M}_{m,n}(\mathbb{R})$ be the set of real matrices with m columns and n rows and $p = \min(m,n)$. For a matrix $A \in \mathcal{M}_{m,n}(\mathbb{R})$, the spaces spanned by the columns and rows of a matrix have the same dimension. We call that the rank of matrix A:

$$\begin{array}{rccl} \text{rank}: & \mathcal{M}_{m,n}(\mathbb{R}) & \longrightarrow & \{0,1,\ldots,p\} \\ & A & \longmapsto & \text{rank } A. \end{array}$$

We recall here some basic properties of the rank function in the context of linear algebra or matricial calculus.

Proposition 1.7. *1.* $\text{rank } A = \text{rank } A^T; \text{rank } A = \text{rank } (AA^T) = \text{rank } (A^TA)$.

2. If the product AB can be done,

$$\text{rank } (AB) \leq \min(\text{rank } A, \text{rank } B) \quad (\text{Sylvester inequality}).$$

As a general rule, when the proposed products of matrices can be done,

$$\text{rank } (A_1A_2\ldots A_k) \leq \min_{i=1,\ldots,k}(\text{rank } A_1, \text{rank } A_2, \ldots, \text{rank } A_k).$$

When $m = n$,

$$\text{rank } (A^k) \leq \text{rank } A.$$

3. $\text{rank } A = 0$ *if and only if* $A = 0$*;* $\text{rank } (cA) = \text{rank } A$ *for* $c \neq 0$.

4. $|\text{rank } A - \text{rank } B| \leq \text{rank } (A+B) \leq \text{rank } A + \text{rank } B$.

The only (useful) topological property of the rank function is that it is *lower-semicontinuous*.

Proposition 1.8. *If* $A_\nu \to A$ *in* $\mathcal{M}_{m,n}(\mathbb{R})$ *when* $\nu \to +\infty$*, then*

$$\liminf_{\nu \to +\infty} \text{rank } A_\nu \geq \text{rank } A. \tag{1.6}$$

This is easy to see if one thinks of rank A characterized as the maximal integer r such that the determinant of a (r,r)-submatrix extracted from A is non-null.

Since the rank function is integer-valued, a consequence of the inequality (1.6) is that the rank function does not decrease in a sufficiently small neighborhood of any matrix A.

For $k \in \{0, 1, \ldots, p\}$, consider now the following two subsets of $\mathcal{M}_{m,n}(\mathbb{R})$:

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

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

S_k is the sub-level set (at level k) of the lower-semicontinuous function rank; it is therefore closed. But, apart from the case $k = 0$ (where $S_0 = \Sigma_0 = \{0\}$), what about the topological structure of Σ_k? The answer is given in the following statement.

Theorem 1.9.
- *Σ_p is an open dense subset of $S_p = \mathcal{M}_{m,n}(\mathbb{R})$.*
- *If $k < p$, the interior of Σ_k is empty and its closure is S_k.*

The singular value decomposition of matrices will show how intricate the subsets S_k and Σ_k may be. For example, *if* rank $A = k$, *in any neighborhood of A, there exist matrices of rank $k + 1, k + 2, \ldots, p$.*

From the algebraic geometry viewpoint, S_k is a semi-algebraic variety ([55]). Because it can be defined by the vanishing of all $(k + 1, k + 1)$-minors, it is thus a solution set of polynomial equations. In case $m = n$, its dimension is $(2n - k)k$ and the tangent space to S_k at a matrix A of rank k can be made explicit from a singular value decomposition of matrix A ([55]).

1.4 Fazel's convex relaxation result

The problem

$$\begin{cases} \text{minimize} & \text{rank } A \\ \text{subject to} & A \in \mathcal{C} \end{cases} \tag{1.7}$$

is a non-convex optimization problem, even when $\mathcal{C}$ is a convex constraint set or an affine subspace. As we have seen before, in general, the rank minimization problem (1.7) is NP-hard. In a situation such as problem (1.7) where the objective function is non-convex, it is natural to replace the problem by the relaxed problem which is obtained by subtituting the rank function with its convex envelope. It

is easy to see that the convex envelope of the rank function on the set $\mathcal{M}_{m,n}(\mathbb{R})$ is the zero function. Such a result is useless, so Fazel tried to find the convex envelope of the rank function on the unit ball for the spectral norm. And this turns out to be the nuclear norm. Fazel et al proposed a heuristic method in [22] that minimizes the nuclear norm, the sum of the singular values of a matrix, over the constraint set. The nuclear norm is not only convex but also continuous and can be optimized efficiently. Many algorithms have been proposed to solve the nuclear norm minimization problem.

First of all, we recall a well-known result about the singular value decomposition.

Theorem 1.10 (The singular value decomposition theorem)**.** *For $A \in \mathcal{M}_{m,n}(\mathbb{R})$, there exists a factorization of A of the form*

$$A = U\Sigma V^T$$

where U is an $m \times m$ orthogonal matrix, Σ is an $m \times n$ "diagonal" matrix with nonnegative real numbers on the diagonal, and V an $n \times n$ orthogonal matrix. Such a factorization is called a singular value decomposition of A.

A common convention is to order the diagonal entries Σ_{ii} in decreasing order. In this case, the diagonal matrix Σ is uniquely determined by A (though the matrices U and V are not). The diagonal entries of Σ are known as the singular values of A.

Let $p = \min(m, n)$. For $A \in \mathcal{M}_{m,n}(\mathbb{R})$, let $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_p(A)$ denote the singular values of A, arranged in the decreasing order; if r stands for the rank of A, the first r singular values are non-zero, the remaining ones are zero. And the vector of singular values of A is $\sigma(A) = (\sigma_1(A), \sigma_2(A), \ldots, \sigma_p(A))$.

If ϕ is a norm in $\mathbb{R}^p$, then we can define an associated matrix norm in $\mathcal{M}_{m,n}(\mathbb{R})$ as

$$\|A\|_\phi = \phi(\sigma(A)).$$

By that way, we have three important (classical) matrix norms:

- **Frobenius (or Euclidean) norm**

$$\|A\|_F = \sqrt{\mathrm{tr}(A^T A)} = \sqrt{\sum_{i=1}^{p} \sigma_i^2(A)} = \|\sigma(A)\|_2.$$

- **Nuclear norm**

$$\|A\|_* = \|\sigma(A)\|_1 = \sum_{i=1}^{p} \sigma_i(A).$$

- **Spectral (maximum singular value) norm**

$$\|A\|_{sp} = \|\sigma(A)\|_\infty = \sigma_1(A) = \max_i \sigma_i(A).$$

$\|.\|_F$ is a "smooth" norm since it derives from an inner product on $\mathcal{M}_{m,n}(\mathbb{R})$, namely $\langle\langle A, B\rangle\rangle := tr(A^T B)$. It is therefore its own dual, while the spectral norm and the nuclear norm are mutually dual (one is the dual norm of the other). These are classical results in matricial analysis. For variational characterizations of this duality relationship as semidefinite programs, see [56].

We consider the function $\phi : \mathcal{M}_{m,n}(\mathbb{R}) \longrightarrow \mathbb{R} \cup \{+\infty\}$ defined by

$$\phi(A) := \begin{cases} \text{rank } A & \text{if } \|A\|_{sp} \leq 1 \\ +\infty & \text{otherwise.} \end{cases}$$

Theorem 1.11 (Fazel's theorem)**.** *The convex hull of ϕ is given by*

$$\text{co}(\phi)(A) := \begin{cases} \|A\|_* & \textit{if } \|A\|_{sp} \leq 1 \\ +\infty & \textit{otherwise} \end{cases} ;$$

i.e., on the set $S = \{A \in \mathcal{M}_{m,n}(\mathbb{R}) : \|A\|_{sp} \leq 1\}$, the convex hull of the rank function is $\|A\|_* = \sum_{i=1}^{p} \sigma_i(A)$.

Proof. (Fazel's proof, [22])

Let $J : \mathbb{R}^d \longrightarrow \mathbb{R} \cup \{+\infty\}$ be a function not identically equal to $+\infty$ and minorized by some affine function. As stated in Section 1.1, the biconjugate function J^{**} is the closed convex envelope of J.

Consequently, the biconjugate function of ϕ and the convex envelope of ϕ coincide.

Part 1. *Computing ϕ^*:* The conjugate of the rank function ϕ, on the set of matrices with spectral norm less than or equal to one, is

$$\phi^*(B) = \sup_{\|A\|_{sp} \leq 1} (\text{tr}(B^T A) - \phi(A)), \tag{1.8}$$

where $\langle B, A\rangle = \text{tr}(B^T A)$ is the inner product in $\mathcal{M}_{m,n}(\mathbb{R})$. By Von Neumann's trace theorem ([41]),

$$\text{tr}(B^T A) \leq \sum_{i=1}^{p} \sigma_i(B)\sigma_i(A), \tag{1.9}$$

where $\sigma_i(A)$ denotes the ith largest singular value of A. Given B, equality in (1.9) is achieved if U_A and V_A are chosen equal to U_B and V_B, respectively, where $A = U_A \Sigma_A V_A^T$ and $B = U_B \Sigma_B V_B^T$ are the singular value decompositions of A and B. The term $\phi(A)$ in (1.8) is independent of U_A and V_A, therefore to find the supremum, we pick $U_A = U_B$ and $V_A = V_B$ to maximize the $\text{tr}(B^T A)$ term. Then the next maximization is with respect to the singular values $\sigma_1(A), \ldots, \sigma_p(A)$. It finally follows that

$$\phi^*(B) = \sup_{\|A\|_{sp} \leq 1} \left(\sum_{i=1}^{p} \sigma_i(B)\sigma_i(A) - \text{rank } A \right).$$

If $A = 0$, for all B, we have

$$\sum_{i=1}^{p} \sigma_i(B)\sigma_i(A) - \text{rank } A = 0.$$

If rank $A = r$ for $1 \leq r \leq p$, then

$$\sup_{\substack{\|A\|_{sp} \leq 1 \\ \text{rank } A = r}} \left(\sum_{i=1}^{p} \sigma_i(B)\sigma_i(A) - \text{rank } A \right) = \sum_{i=1}^{r} \sigma_i(B) - r.$$

Hence, $\phi^*(B)$ can be expressed as

$$\phi^*(B) = \max\{0, \sigma_1(B) - 1, \ldots, \sum_{i=1}^{p} \sigma_i(B) - p\}.$$

The largest term in this set is the one that sums all positive $(\sigma_i(B) - 1)$ terms. We conclude that

$$\phi^*(B) = \sum_{i=1}^{p} (\sigma_i(B) - 1)^+,$$

where a^+ denotes the positive part of a, *i.e.* $a^+ = \max\{0, a\}$.

Part 2. *Computing* ϕ^{**} : We now find the conjugate of ϕ^*, defined as

$$\phi^{**}(C) = \sup_B \left(\text{tr}(C^T B) - \phi^*(B) \right).$$

As before, we choose $U_C = U_B$ and $V_C = V_B$ to get

$$\phi^{**}(C) = \sup_B \left(\sum_{i=1}^{p} \sigma_i(B)\sigma_i(C) - \phi^*(B) \right).$$

We consider two cases, $\|C\|_{sp} > 1$ and $\|C\|_{sp} \leq 1$.

If $\|C\|_{sp} > 1$, we can choose $\sigma_1(B)$ large enough so that $\phi^{**}(C) \to \infty$. To see this, note that in

$$\phi^{**}(C) = \sup_B \left(\sum_{i=1}^{p} \sigma_i(B)\sigma_i(C) - \sum_{i=1}^{p} (\sigma_i(B) - 1)^+ \right),$$

the coefficient of $\sigma_1(B)$ is positive.

Now let $\|C\|_{sp} \leq 1$. For $\|B\|_{sp} \leq 1$, then $\phi^*(B) = 0$ and the supremum is achieved for $\sigma_i(B) = 1$ for $i = 1, \ldots, p$, yielding

$$\sup_{\|B\|_{sp} \leq 1} \left(\sum_{i=1}^{p} \sigma_i(B)\sigma_i(C) - \sum_{i=1}^{p} (\sigma_i(B) - 1)^+ \right) = \sum_{i=1}^{p} \sigma_i(C) = \|C\|_*.$$

Now, we show that for $\|B\|_{sp} > 1$, the argument of the sup is always smaller than the value given above. By adding and subtracting the term $\sum_{i=1}^{p} \sigma_i(C)$ and rearranging the terms, we get

$$\begin{aligned}
&\textstyle\sum_{i=1}^{p} \sigma_i(C)\sigma_i(B) - \sum_{i=1}^{r} (\sigma_i(B) - 1) \\
&= \textstyle\sum_{i=1}^{p} \sigma_i(C)\sigma_i(B) - \sum_{i=1}^{r} (\sigma_i(B) - 1) - \sum_{i=1}^{p} \sigma_i(C) + \sum_{i=1}^{p} \sigma_i(C) \\
&= \textstyle\sum_{i=1}^{r} (\sigma_i(B) - 1)(\sigma_i(C) - 1) + \sum_{i=r+1}^{p} (\sigma_i(B) - 1)\sigma_i(C) + \sum_{i=1}^{p} \sigma_i(C) \\
&< \textstyle\sum_{i=1}^{p} \sigma_i(C).
\end{aligned}$$

The last inequality holds because the first two sums on the third line always have a negative value.

In summary, we have shown, if $\|C\|_{sp} \leq 1$,

$$\phi^{**}(C) = \|C\|_*.$$

Thus, on the set S, $\|.\|_*$ is the convex envelope of ϕ. □

Remark 1.12. 1. The convex hull of the rank function on the set $S_R = \{A \in \mathcal{M}_{m,n}(\mathbb{R}) : \|A\|_{sp} \leq R\}$ is $\frac{1}{R}\|A\|_*$.

2. The convex hull of the rank function on the set $S^1 = \{A \in \mathcal{M}_{m,n}(\mathbb{R}) : \|A\|_* \leq 1\}$ is also $\|A\|_*$.

1.5 A first new proof of Fazel's theorem

A special case of the rank minimization problem is minimizing the counting function. Recall that the counting function $c : \mathbb{R}^p \to \mathbb{R}$ is defined as follows:

$$\forall x = (x_1, \ldots, x_p) \in \mathbb{R}^p, c(x) := \text{ the number of i's for which } x_i \neq 0.$$

Sometimes, $c(x)$ is denoted as $\|x\|_0$, a misleading notation since $c(x)$ is not a norm on $\mathbb{R}^p$. Note however that, if $\|x\|_k$ denotes $(\sum_{i=1}^p |x_i|^k)^{1/k}$ as usual, $(\|x\|_k)^k \to c(x)$ when $k \to 0^+$ (but $\|x\|_k$ does not converge to 0 when $k \to 0^+$, as it is stated sometimes). The function c gives rise to the so-called Hamming distance d (used in coding theory), defined on $\mathbb{R}^p$ as:

$$d(x, y) := c(x - y).$$

Our strategy:
We will calculate the convex hull of the counting function restricted to a l^∞-ball of $\mathbb{R}^p$, and we then use it, with a result of A.Lewis, to recover the relaxed form of the rank function.

When dealing with matrices $A \in \mathcal{M}_{m,n}(\mathbb{R})$, we know that:

- for $x = (x_1, \ldots, x_p) \in \mathbb{R}^p, \text{rank}[\text{diag}_{m,n}(x)] = c(x)$ where $\text{diag}_{m,n}(x)$ is a matrix in $\mathcal{M}_{m,n}(\mathbb{R})$ such that all the "non-diagonal" entries are null and $x_1, x_2, \ldots, x_p$ are on the "diagonal";
- for $A \in \mathcal{M}_{m,n}(\mathbb{R}), \text{rank } A = c[\sigma(A)]$, where $\sigma(A) = (\sigma_1(A), \ldots, \sigma_p(A))$ is the vector made up with the singular values $\sigma_i(A)$ of A.

A.Lewis ([46],[47]) showed that the Legendre-Fenchel conjugate of a function of matrice A (satisfying some specific properties) could be obtained by just conjugating some associated function of the singular values of A. Using his results twice, we are able to calculate the Legendre-Fenchel biconjugate of the rank

function (that is the convex hull of the rank function) by calling on the biconjugate of the c function. In doing so, we retrieve FAZEL's relaxation theorem.

1.5.1 The closed convex hull of the counting function

The function c is an integer-valued, subadditive, lower-semicontinous function on $\mathbb{R}^p$. Since $c(\alpha x) = c(x)$ for all $\alpha \neq 0$, there is no hope to get anything interesting by convexifying (*i.e.*, taking the convex hull of) the function c (on the whole space $\mathbb{R}^p$). So, we consider it on some appropriate ball, namely, for $R > 0$:

$$c_R(x) := \begin{cases} c(x) & \text{if } \|x\|_\infty \leq R; \\ +\infty & \text{otherwise.} \end{cases} \tag{1.10}$$

Taking the convex hull and the closed convex hull of c amount to the same here; so we just note $\mathrm{co}(c_R)$ the convexified form of c (*i.e.*, the largest convex function minorizing c_R).

Here is the result of this section.

Theorem 1.13. *We have:*

$$\forall x \in \mathbb{R}^p, \qquad \mathrm{co}(c_R)(x) = \begin{cases} \frac{1}{R}\|x\|_1 & \textit{if } \|x\|_\infty \leq R; \\ +\infty & \textit{otherwise.} \end{cases}$$

Contrary to Section 1.1, we do not go here through the calculate of the LEGENDRE-FENCHEL conjugate of the c_R function.

Proof. The basic properties of the convexifying operation (see [38] for example) show that the domain of $\mathrm{co}(c_R)$, *i.e.* the set on which this function is finite-valued, is just the convex hull of the domain of c_R. So, in our particular instance, the domain of $\mathrm{co}(c_R)$ is that of c_R, which is the convex set $\{x | \|x\|_\infty \leq R\}$.

We therefore have to prove that $\mathrm{co}(c_R)(x) = \frac{1}{R}\|x\|_1$ whenever $\|x\|_\infty \leq R$.

First point. $\mathrm{co}(c_R)(x) \geq \frac{1}{R}\|x\|_1$ for x satisfying $\|x\|_\infty \leq R$.
If $\|x\|_\infty \leq R$,

$$c_R(x) = c(x) \geq \sum_{i=1}^{p} \frac{|x_i|}{\max_i |x_i|} = \frac{1}{\max_i |x_i|} \sum_{i=1}^{p} |x_i| \geq \frac{1}{R}\|x\|_1.$$

Second point. $\frac{1}{R}\|x\|_1 \geq \mathrm{co}(c_R)(x)$ for x satisfying $\|x\|_\infty \leq R$.

Let x satisfy $\|x\|_\infty \leq R$. For such an $x = (x_1, \ldots, x_p)$, we define vectors $y = (y_1, \ldots, y_p)$ according to the following rule:

$$\begin{cases} \text{if } x_i = 0, \text{ then } y_i = 0; \\ \text{if } x_i > 0, \text{ then } y_i = 0 \text{ or } R; \\ \text{if } x_i < 0, \text{ then } y_i = 0 \text{ or } -R; \end{cases} \tag{1.11}$$

In doing so, we get at a "net on a box" $\square_x$ of $\mathbb{R}^p$:

$$\square_x := \{(y_1, \ldots, y_p) |\ y_i \text{ designed according to the rule (1.11)}\}$$
(see Figure 1.1, with $p = 2$).

$\square_x$ has $2^{c(x)}$ points, which are the vertices of a box containing x (this has been done for that!). In other words, x lies in the convex hull of $\square_x$: there exist real numbers $\alpha_1, \ldots, \alpha_k$ and $y_1, \ldots, y_k$ in $\square_x$ such that:

$$\begin{cases} \alpha_i \geq 0 \text{ for all } i \\ \sum_{i=1}^k \alpha_i = 1 \\ x = \sum_{i=1}^k \alpha_i y^i. \end{cases}$$

Consider now an arbitrary convex function h minorizing c_r. Then, due to the convexity of h,

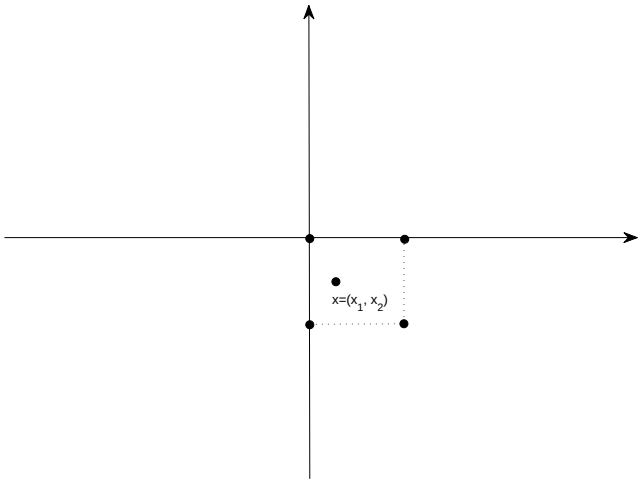
$$h(x) = h(\sum_{i=1}^k \alpha_i y^i) \leq \sum_{i=1}^k \alpha_i h(y^i). \tag{1.12}$$

But, when $y \in \square_x$,

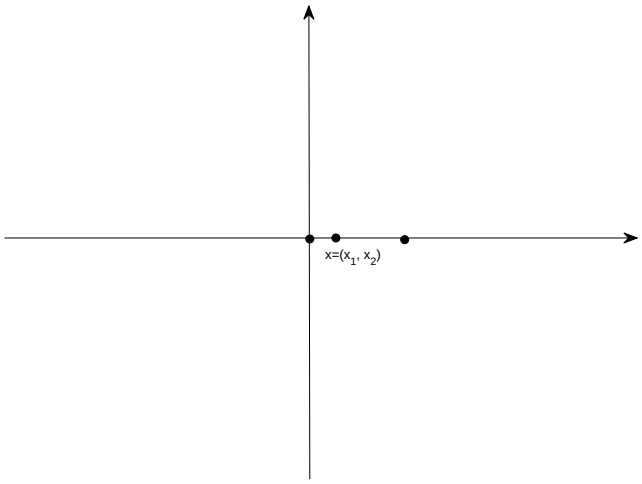
$$\begin{aligned} c_R(y) &= \text{number of j's for which } y_j \neq 0 \\ &= \sum_{\{j | y_j \neq 0\}} \frac{|y_j|}{R} \qquad \text{(because } |y_j| = R \text{ whenever } y_j \neq 0) \\ &= \frac{1}{R} \sum_{\{j | y_j \neq 0\}} |y_j| = \frac{1}{R}\|y\|_1. \end{aligned}$$

So, with all the y^i lying in $\square_x$, we get from (1.12):

$$h(x) \leq \sum_{i=1}^k \alpha_i h(y^i) \leq \sum_{i=1}^k \alpha_i c_r(y^i) = \frac{1}{R} \sum_{i=1}^k \alpha_i \|y^i\|_1. \tag{1.13}$$



(a) $c(x) = 2$



(b) $c(x) = 1$

FIGURE 1.2: $\square_x$ for $x = (x_1, x_2) \in \mathbb{R}^2$

On the other hand, we have

$$x_j = \sum_{i=1}^{k} \alpha_i (y^i)_j \text{ for all } j = 1, \ldots, p.$$

Thus, due to the specific correspondence between the signs of x_j and $(y^i)_j$ (*cf.* (1.11)),

$$|x_j| = \sum_{i=1}^{k} \alpha_i |(y^i)_j| \text{ for all } j = 1, \ldots, p$$

so that:

$$\|x\|_1 = \sum_{i=1}^{k} \alpha_i \|y^i\|_1.$$

Consequently, we derive from (1.13):

$$h(x) \leq \frac{1}{R}\|x\|_1.$$

Finally,

$$\begin{array}{rcl} \mathrm{co}(c_R)(x) & = & \sup\{h(x) |\quad h \text{ convex function minorizing } c_R\} \\ & \leq & \frac{1}{R}\|x\|_1. \end{array}$$

Altogether (First point and Second point), we have proved that $\mathrm{co}(c_R)(x) = \frac{1}{R}\|x\|_1$ whenever $\|x\|_\infty \leq R$.

□

Comment 1: The result of Theorem 1.11 is part of the "folklore" in the areas where minimizing counting function appears (there are numerous papers in signal recovery, compressed sensing, statistics, etc.). We did not find any reference where it was stated in a clear-cut manner. That was the reason for a *direct* proof here.

Comment 2: Another convexification result, similar to Theorem 1.11, easy to prove, is as follows: Consider the function $\|.\|_k$ with $0 < k < 1$ (no more a norm), restricted to the ball $\{x|\ \|x\|_1 \leq 1\}$; then its convex hull is still the l_1 norm $\|.\|_1$ (restricted to the same ball).

1.5.2 The first new proof of Fazel's theorem

Consider the following function on $\mathcal{M}_{m,n}(\mathbb{R})$, it is just the "matricial cousin" of the c_R function:

$$\mathrm{rank}_R(A) := \begin{cases} \text{rank of } A & \text{if } \|A\|_{sp} \leq R; \\ +\infty & \text{otherwise.} \end{cases}$$

We propose here another path to prove Theorem 1.11: apply A.Lewis' fine results (of conjugation), such as displayed in [46],[47]. Let us recall them briefly.

A function $f : \mathbb{R}^p \to \mathbb{R}$ is called *absolutely symmetric* if, for all $x \in \mathbb{R}^p$,

$$f(x_1, \ldots, x_p) = f(\hat{x}_1, \ldots, \hat{x}_p),$$

where $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_p)$ is the vector, built up from $x = (x_1, \ldots, x_p)$, whose components are the $|x_i|$'s arranged in a decreasing order. Associated with f is the function $F : \mathcal{M}_{m,n}(\mathbb{R}) \to \mathbb{R} \cup \{+\infty\}$ defined as follows:

$$\forall A \in \mathcal{M}_{m,n}(\mathbb{R}), F(A) := f[\sigma_1(A), \ldots, \sigma_p(A)].$$

A.Lewis' conjugacy rule is now:

Theorem 1.14. *([46],[47])*

With f satisfying the symmetry property above, we have:

$$\forall A \in \mathcal{M}_{m,n}(\mathbb{R}), F^*(A) = f^*[\sigma_1(A), \ldots, \sigma_p(A)].$$

Proof. (of Theorem 1.11)

From the fact that f is absolutely symmetric, we can easily prove that f^* is also absolutely symmetric. Thus, by applying Lewis' theorem twice, we obtain

$$\forall A \in \mathcal{M}_{m,n}(\mathbb{R}), F^{**}(A) = f^{**}[\sigma_1(A), \ldots, \sigma_p(A)]. \tag{1.14}$$

In our particular setting, we choose:

$$f = c_R, \text{ so that } F = \text{rank}_R.$$

The biconjugate of f (resp. of F) is its (closed) convex hull $\text{co}(c_R)$ (resp. $\text{co}(\text{rank}_R)$). Whence Fazel's theorem follows from (1.14) and Theorem 1.13. □

1.6 The explicit quasi-convex relaxation of the rank function

In this section, we will provide an explicit description of the convex hull of the set of matrices of bounded rank, restricted to balls for the spectral norm. As

applications, we deduce two relaxed forms of the rank function restricted to balls for the spectral norm: one is the quasiconvex hull of this rank function, another is its convex hull, thus retrieving (again) FAZEL's theorem.

For $k \in \{0, 1, \ldots, p\}$ and $R \geq 0$,

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

$$S_k^R := S_k \cap \{A \in \mathcal{M}_{m,n}(\mathbb{R}) | \quad \|A\|_{sp} \leq R\}.$$

For $m = n$, S_k is an algebraic variety of dimension $(2n - k)k$.

Convexifying the set S_k is not of any use since the convex hull of S_k, denoted as co S_k, is the whole space $\mathcal{M}_{m,n}(\mathbb{R})$; indeed this comes from the singular value decomposition technique. Thus:

$$\forall k = 1, \ldots, p \qquad \text{co } S_k = \mathcal{M}_{m,n}(\mathbb{R}).$$

The question becomes of some interest if we add some "moving wall" $\|A\|_{sp} \leq R$, like in the definition of S_k^R. So, we will give an explicit description of co S_k^R. As applications, we deduce two relaxed forms of the following (restricted) rank function:

$$\text{rank}_R(A) := \begin{cases} \text{rank of } A & \text{if } \|A\|_{sp} \leq R, \\ +\infty & \text{otherwise.} \end{cases} \tag{1.15}$$

The first relaxed form is the so-called quasiconvex hull of rank_R, *i.e.*, the largest quasiconvex function minorizing it. Then, as an ultimate step, we retrieve FAZEL's theorem on the convex hull (or biconjugate) of the rank_R function (Theorem 1.11).

1.6.1 Convexifying the set of matrices of bounded rank

Theorem 1.15. *We have:*

$$co\ S_k^R = \{A \in \mathcal{A}_{m,n}(\mathbb{R}) | \quad \|A\|_{sp} \leq R \text{ and } \|A\|_* \leq Rk\}. \tag{1.16}$$

Proof. For either $k = 0$ or $R = 0$ there is nothing to prove. We therefore suppose that k is a positive integer and $R > 0$. Moreover, since rank $(A/R) =$ rank A, and the norms are positively homogeneous functions ($\|A/R\| = \|A\|/R$), it suffices to prove (1.16) for $R = 1$.

First inclusion

$$\text{co } S_k^1 \subset \{A \in \mathcal{A}_{m,n}(\mathbb{R}) |\ \|A\|_{sp} \leq 1 \text{ and } \|A\|_* \leq k\}. \tag{1.17}$$

Let $A \in S_k^1$; by definition of S_k^1, we have $\|A\|_{sp} = \sigma_1(A) \leq 1$ and rank $A \leq k$. Consequently, all the non-zero singular values of A - they are less than k - are majorized by 1; hence

$$\|A\|_* = \sum_{i=1}^{\text{rank } A} \sigma_i(A) \leq k.$$

Since the right-hand side of (1.17) is convex (as an intersection of sub-level sets of two norms), we derive the inclusion (1.17).

Reverse inclusion

$$\text{co } S_k^1 \supset \{A \in \mathcal{A}_{m,n}(\mathbb{R}) |\quad \|A\|_{sp} \leq 1 \text{ and } \|A\|_* \leq k\}. \tag{1.18}$$

This is the tricky part of the proof. We first begin with a technical lemma on a specific convex polyhedron in $\mathbb{R}^p$; its proof can be found in ([30], Exercises V.4 and V.15).

Lemma 1.16. *For $k = 1, \ldots, p$, let*

$$D := \{x = (x_1, \ldots, x_p) \in \mathbb{R}^p |\quad 0 \leq x_i \leq 1 \text{ for all } i, \sum_{i=1}^{p} x_i \leq k\},$$

$$\Omega := \{x = (x_1, \ldots, x_p) \in \mathbb{R}^p |\quad x_i \in \{0, 1\} \text{ for all } i, \sum_{i=1}^{p} x_i = k\}.$$

Then, $D = co\ \Omega$.

This result holds true because k is an integer. A picture in $\mathbb{R}^p$ helps to understand its meaning.

Let now A satisfy $\|A\|_{sp} \leq 1$ and $\|A\|_* \leq k$. Consider a singular value decomposition of A:

$$A = U\Sigma V^T, \tag{1.19}$$

where U and V are orthogonal matrices of appropriate size and Σ, of the same type as A, with $\sigma_1(A), \ldots, \sigma_p(A)$ on the "diagonal" and 0 elsewhere. We write

$$\Sigma = \text{diag}_{m,n}(\sigma_1(A), \ldots, \sigma_p(A)).$$

Because $0 \leq \sigma_i(A) \leq 1$ for all i and $\sum_{i=1}^{p} \sigma_i(A) \leq k$, according to the lemma recalled above, the vector $(\sigma_1(A), \ldots, \sigma_p(A))$ can be expressed as a convex combination of elements in Ω: there exist real numbers $\alpha_1, \ldots, \alpha_q$, vectors $\beta^1, \ldots, \beta^q$ in Ω such that:

$$\begin{cases} \alpha_j \in [0,1] \text{ for all } j, \sum_{j=1}^{p} \alpha_j = 1 \\ (\sigma_1(A), \ldots, \sigma_p(A)) = \sum_{j=1}^{p} \alpha_j \beta^j. \end{cases} \tag{1.20}$$

For $\beta^j = (\beta_1^j, \ldots, \beta_p^j)$, we set

$$Y^j = \text{diag}_{m,n}(\beta_1^j, \ldots, \beta_p^j), B^j = UY^jV^T. \tag{1.21}$$

Because $\beta^j \in \Omega$, we have:

$$\|B^j\|_{sp} = \|Y^j\|_{sp} \leq 1, \text{rank } B^j = \text{rank } Y^j \leq k.$$

Moreover, in view of (1.20) and (1.21), we derive from (1.19):

$$\Sigma = \sum_{j=1}^{q} \alpha_j Y^j, A = \sum_{j=1}^{q} \alpha_j B^j.$$

Hence, A is a convex combination of matrices in S_k^1. □

Remarks

1. Although S_k^R is a fairly complicated set of matrices (due to the definition of S_k), its convex hull is simple: according to (1.16), it is the intersection of two balls, one for the spectral norm, the other one for the nuclear norm. Getting at such an explicit form of co S_k^R is due to the happy combination of these specific norms. If $\|.\|$ were any norm on $\mathcal{M}_{m,n}(\mathbb{R})$ and

   $$\hat{S}_k^R = \{A | \text{ rank } A \leq k \text{ and } \|A\| \leq R\},$$

   due to the equivalence between the norms $\|.\|$ and $\|.\|_{sp}$, we would get with (1.16) an inner estimate and an outer estimate of co $\hat{S}_k^R$.

2. A particular case. Let $R = 1$ and $k = 1$ in the result of Theorem 1.11. We get that

   $$\begin{array}{ll} \text{co } \{A| & \text{rank } A \leq 1 \text{ and } \sigma_1(A) \leq 1\} \\ = \{A| & \sum_{i=1}^{p} \sigma_i(A) \leq 1\}. \end{array} \tag{1.22}$$

Remember that maximizing a linear form (of matrices) on both sets in (1.22) yields the same optimal value.

3. There are quite a few examples where the convex hull of a set of matrices can be expressed explicity. We mention here one of them, a very recent result indeed (see [24],[43]). For $m \leq n$, let

$$T_m^n := \{A \in \mathcal{M}_{m,n}(\mathbb{R}) | \quad A^T A = I_m\}.$$

T_m^n is called the Stiefel manifold. For $m = n$, T_n^n is just the set orthogonal (n, n) matrices. According to [43, p.531] (see also [24]), the support function of T_m^n is $\|.\|_*$, hence:

$$\text{co } T_m^n = \{A| \quad \|A\|_{sp} \leq 1\}. \tag{1.23}$$

1.6.2 The quasiconvex hull of the restricted rank function

Before going futher, we recall some basic facts about quasiconvex functions and the quasiconvexification of functions.

⋆ **Quasi-convexity** *(in the sense used in Optimization and Mathematical Economy, different from the one used in the Calculus of variations)*

Definition 1.17. $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is said to be quasi-convex when:

$$\forall x_1, x_2 \in \mathcal{X}, \forall \lambda \in [0, 1] : \quad f[\lambda x_1 + (1 - \lambda)x_2] \leq \max[f(x_1), f(x_2)].$$

Besides this analytical definition, there is a geometrical characterization. Recall that $[f \leq \alpha] := \{x \in \mathcal{X} : f(x) \leq \alpha\}$ (the sub-level set of f at the level $\alpha \in \mathbb{R}$). $[f \leq \alpha]$ is possibly empty. If $\mu := \inf_{\mathcal{X}} f$ is finite, $[f \leq \mu]$ is the set of (global) minimizers of f on $\mathcal{X}$.

Characterization 1.18. *$f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is quasi-convex if and only if $[f \leq \alpha]$ is convex for all $\alpha \in \mathbb{R}$.*

Remark 1.19.

- For a quasi-convex function, dom$f$ is a convex set, of course.
- Clearly, $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is lower-semicontinuous and quasi-convex on $\mathcal{X}$ if and only if $[f \leq \alpha]$ is closed and convex for all $\alpha \in \mathbb{R}$.

⋆ Constructing a function from the collection of sub-level sets

Let $(T_\alpha)_{\alpha\in\mathbb{R}}$ be a collection of subsets of $\mathcal{X}$ satisfying the property:

$$(\alpha < \beta) \Rightarrow (T_\alpha \subset T_\beta). \tag{1.24}$$

Given a collection $(T_\alpha)_{\alpha\in\mathbb{R}}$ satisfying the property (1.24), we can define a function g as follows:

$$g(x) := \inf\{\alpha : x \in T_\alpha\} \qquad (\inf\emptyset = +\infty \text{ as usual}). \tag{1.25}$$

Example 1.1. *If $(T_\alpha)_{\alpha\in\mathbb{R}}$ is the collection of sub-level sets associated with a function f, i.e. $T_\alpha = [f \le \alpha]$ for all $\alpha \in \mathbb{R}$, then the function g defined from the T_α's as in (1.25) coincides with f.*

But, a collection (T_α) of sets may satisfy the property (1.24) without being a collection of sub-level sets associated with a function. So, a certain "regularization" is necessary beforehand. Let us pose:

$$\forall\alpha \in \mathbb{R} \quad \hat{T}_\alpha := \cap_{\alpha'>\alpha} T_{\alpha'}.$$

Then, $(\hat{T}_\alpha)_\alpha$ does satisfy the property (1.24), and it is the collection of sub-level sets of the function g defined as in (1.25). If the T_α's are convex (closed), then so are the $\hat{T}_\alpha$'s.

Proposition 1.20. *Let $(T_\alpha)_\alpha$ satisfy the property (1.24), and let g be defined from the T_α's as in (1.25). Thus the sub-level sets of g are the $\hat{T}_\alpha$'s. Then:*

(a) If T_α is convex for all $\alpha \in \mathbb{R}$, then g is quasi-convex on $\mathcal{X}$.

(b) If T_α is closed for all $\alpha \in \mathbb{R}$, then g is lower-semicontinuous on $\mathcal{X}$.

⋆ Quasi-convex hull and lower-semicontinuous (or closed) quasi-convex hull of a function

Definition 1.21. The quasi-convex hull f_q of f is the largest quasi-convex function minorizing f. The closed quasi-convex hull $f_{\overline{q}}$ of f is the largest closed quasi-convex function minorizing f.

It is obvious that $f_{\overline{q}} \le f_q \le f$. Also, since any convex function is quasi-convex,

$$\overline{\text{co}} f \le f_{\overline{q}} \le f. \tag{1.26}$$

Proposition 1.22. *(of construction of f_q and $f_{\overline{q}}$ from the sub-level sets of f) We have*

$$\forall x \in \mathcal{X}, f_q(x) = \inf\{\alpha : x \in \mathrm{co}[f \leq \alpha]\}$$
$$f_{\overline{q}}(x) = \inf\{\alpha : x \in \overline{\mathrm{co}}[f \leq \alpha]\}. \tag{1.27}$$

All these results date back to J.-P. Crouzeix's works ([14]).

The explicit form of the quasiconvex hull of the rank function

Theorem 1.23. *The quasiconvex hull $rank_{R,q}$ of the function $rank_R$ is given as follows:*

$$A \mapsto rank_{R,q}(A) = \begin{cases} \lceil \frac{1}{R}\|A\|_* \rceil & \text{if } \|A\|_{sp} \leq R, \\ +\infty & \text{otherwise,} \end{cases} \tag{1.28}$$

wher $\lceil a \rceil$ stands for the smallest integer which is larger than a.

Proof. Since the domain of the function rank_R (*i.e.*, the set of A at which $\mathrm{rank}_R(A)$ is finite-valued) is the (convex compact) ball $\{A| \quad \|A\|_{sp} \leq R\}$, the quasiconvex hull $\mathrm{rank}_{R,q}$ will have the same domain. In short,

$$\mathrm{rank}_{R,q}(A) = +\infty \text{ if } \|A\|_{sp} > R.$$

Let $\alpha \geq 0$. Since the rank is an integer, one obviously has

$$[\mathrm{rank}_R \leq \alpha] = [\mathrm{rank}_R \leq \lfloor \alpha \rfloor],$$

where $\lfloor \alpha \rfloor$ denotes the integer part of α. So, by application of Theorem 1.15,

$$\begin{aligned} \mathrm{co}\,[\mathrm{rank}_R \leq \alpha] &= \mathrm{co}\,[\,\mathrm{rank}_R \leq \lfloor \alpha \rfloor] \\ &= \{A| \quad \|A\|_{sp} \leq R \text{ and } \|A\|_* \leq R\lfloor \alpha \rfloor\}. \end{aligned}$$

Now, following the construction recalled in (1.27), we have: for all A such that $\|A\|_{sp} \leq R$,

$$\begin{aligned} \mathrm{rank}_{R,q}(A) &= \inf\{\alpha| \quad \|A\|_* \leq R\lfloor \alpha \rfloor\} \\ &= \inf\{\alpha| \quad \frac{\|A\|_*}{R} \leq \lfloor \alpha \rfloor\} = \lceil \frac{1}{R}\|A\|_* \rceil. \end{aligned}$$

□

1.6.3 Another proof of Fazel's theorem

Proof. (of Theorem 1.11)

As a further and ultimate step from Theorem 1.23, we easily get at FAZEL's theorem by showing that the convex hull of rank_R is defined by

$$A \mapsto \text{co}(\text{rank}_R)(A) = \begin{cases} \frac{1}{R}\|A\|_* & \text{if } \|A\|_{sp} \leq R, \\ +\infty & \text{otherwise.} \end{cases} \tag{1.29}$$

When $\|A\|_{sp} > R$, there is nothing special to say:

$$\text{rank}_R(A) = \text{co}(\text{rank}_R)(A) = +\infty.$$

We just have to prove that $\text{co}(\text{rank}_R)(A) = \frac{1}{R}\|A\|_*$ whenever $\|A\|_{sp} \leq R$. Consider therefore such an A.

First of all, since any convex function is quasiconvex,

$$\text{co}(\text{rank}_R) \leq \text{rank}_{R,q},$$

thus

$$\text{co}(\text{rank}_R) \leq \text{co}(\text{rank}_{R,q}). \tag{1.30}$$

As in the second part of the proof of Theorem 1.15, we set:

$$\Gamma := \{x = (x_1, \ldots, x_p) \in \mathbb{R}^p | \quad x_i \in \{0, R\} \text{ for all } i\},$$

$$\mathcal{M} := \{X \in \mathcal{M}_{m,n}(\mathbb{R}) | \quad \sigma_i(X) \in \{0, R\} \text{ for all } i\}.$$

Since $(\sigma_1(A), \ldots, \sigma_p(A)) \in \text{co}\ \Gamma$, A lies in $\text{co}\ \mathcal{M}$. There therefore exist real numbers $\alpha_1, \ldots, \alpha_l$, matrices $X_1, \ldots, X_l$ in $\mathcal{M}$ (constructed like the matrices B^j in the proof of Theorem 1.15) such that:

$$\begin{cases} \alpha_i \in [0,1] \text{ for all } i, \sum_{j=1}^{l} \alpha_i = 1 \\ A = \sum_{j=1}^{l} \alpha_j X_j. \end{cases} \tag{1.31}$$

Now, since $X_j \in \mathcal{A}$, it comes from Theorem 1.23 that

$$\text{rank}_{R,q}(X_j) = \lceil \frac{1}{R}\|X_j\|_* \rceil = \frac{1}{R}\|X_j\|_*.$$

Consequently,

$$\begin{aligned} \text{co}(\text{rank}_{R,q})(A) &= \text{co}(\text{rank}_{R,q})(\textstyle\sum_{j=1}^{l} \alpha_j X_j) \\ &\leq \textstyle\sum_{j=1}^{l} \alpha_j \text{co}(\text{rank}_{R,q})(X_j) \\ &\leq \textstyle\sum_{j=1}^{l} \alpha_j \text{rank}_{R,q}(X_j) \\ &= \textstyle\sum_{j=1}^{l} \frac{1}{R}\alpha_j \|X_j\|_* = \frac{1}{R}\|A\|_*. \end{aligned}$$

Thus, $\text{co}(\text{rank}_{R,q})(A) \leq \frac{1}{R}\|A\|_*$.

On the other hand, because $\text{rank}_{R,q}(A) \geq \frac{1}{R}\|A\|_*$ for all $A \in \mathcal{M}_{m,n}(\mathbb{R})$ and $\frac{1}{R}\|.\|_*$ is convex, we have that:

$$\text{co}(\text{rank}_{R,q})(A) \geq \frac{1}{R}\|A\|_*.$$

So we have proved that

$$\text{co}(\text{rank}_{R,q})(A) = \text{co}(\text{rank}_R) = \frac{1}{R}\|A\|_*.$$

□

1.7 Rank vs Nuclear Norm Minimization

The minimization problem of the counting function is a special case of the rank minimization problem. And for a long time, the l^1 norm was used as the relaxed form of the counting function, *i.e.* instead of finding the vectors with the minimum number of nonzero components, we find the minimum l^1 norm solutions. But can we recover the sparsest solution? The same question was raised when the nuclear norm was used as the relaxed form of the rank function.

Many studies concentrated on these questions for the affine minimization (*i.e.* the constraint set is affine) of the counting function and the rank function. Several conditions under which the sparsest can be recovered were proposed. Candès and Tao gave the so-called restricted isometry condition for the vector case ([10]). Another result - the spark condition was proposed by Donoho et al in [17]. Then, based on the idea of the restricted isometry condition for the vector case, Recht et al developped a condition under which the minimum-rank solution can be recovered ([56]).

1.7.1 Spark

Given $A \in \mathcal{M}_{m,n}(\mathbb{R})$, the *spark* of A is the smallest positive integer k such that there exists a set of k columns of A which are linearly dependent. Remember that the rank of A is the largest number of columns of A which are linearly independent. The term spark seems to have been coined by Donoho and Elad in 2003.

Actually, the given definition of spark is a bit uncomplete: If A is of full column rank, *i.e.* if rank $A = n$, there is no set of k columns of A which are linearly dependent. In that case, we should adopt $+\infty$ as for the spark of A (the infimum over the empty set).

The other extreme case is when one column of A is a zero-column: then spark $A = 1$. In short, *if A does not contain any zero-column and is not of full column rank,*

$$2 \leq \text{spark } A \leq \text{rank } A + 1.$$

The spark gives a criterion for the uniqueness of the sparsest possible solution to the equation $u = Av$.

Lemma 1.24 ([17])**.** *If $u = Av_0$ and $\|v_0\|_0 < \text{spark}(A)/2$, then v_0 is the unique sparsest possible solution to the equation $u = Av$.*

1.7.2 Restricted Isometry Property

We consider the affine rank minimization problem

$$\begin{array}{ll} \text{minimize} & \text{rank } X \\ \text{subject to} & \mathcal{A}(X) = b \end{array}$$

where $X \in \mathcal{M}_{m,n}(\mathbb{R})$ and the linear map $\mathcal{A} : \mathcal{M}_{m,n}(\mathbb{R}) \longrightarrow \mathbb{R}^d$ and vector $b \in \mathbb{R}^d$ are given.

Let X_0 be a matrix of rank r satisfing $\mathcal{A}(X_0) = b$ and

$$X^* = \arg\min_X \|X\|_* \quad \text{s.t. } \mathcal{A}(X) = b. \tag{1.32}$$

Definition 1.25. For every $1 \leq r \leq p$, define the r-restricted isometry constant to be the smallest number $\delta_r(\mathcal{A})$ such that

$$(1 - \delta_r(\mathcal{A}))\|X\|_F \leq \|\mathcal{A}(X)\| \leq (1 + \delta_r(\mathcal{A}))\|X\|_F \tag{1.33}$$

holds for all matrices X of rank at most r.

The restricted isometry property (RIP) for sparse vectors was developed by Candès and Tao in [10]. It requires that (1.33) holds with Euclidean norm replacing by the Frobenius norm and rank being replaced by cardinality.

In the next two theorems, we see the power of the restricted isometry property.

Theorem 1.26 ([56])**.** *Suppose that $\delta_{2r} < 1$ for some integer $r \geq 1$. Then X_0 is the only matrix of rank at most r satisfying $\mathcal{A}(X) = b$.*

Theorem 1.27 ([56])**.** *Suppose that $r \geq 1$ is such that $\delta_{5r} < 1/10$. Then $X^* = X_0$.*

Chapter 2

Generalized subdifferentials of the rank function

In this chapter, we calculate the generalized subdifferentials; i.e. *the proximal subdifferential, the* Fréchet *subdifferential, the limiting subdifferential and the* Clarke *subdifferential of the counting function. Then, thanks to theorems of* Lewis *and* Sendov *about the nonsmooth analysis of functions of singular values, we obtain the corresponding generalized subdifferentials of the rank function.*

2.1 Definitions and Properties

In the last decades, nonsmooth analysis has grown rapidly and has come to play a role in functional analysis, optimization, optimal design, mechanics and plasticity, differential equations (as in the theory of viscosity solutions), control theory, and increasingly, in analysis generally (critical point theory, inequalities, fixed point theory, variational methods, *etc.*). One of the most important keys in nonsmooth analysis is the notion of generalized subdifferential. The definitions and properties of generalized subdifferentials have been developed in several works, beginning with the case of locally Lipschitz functions (see in [12],[54]). Then, they have been generalized for lower-semicontinous functions (see in [58],[59]). Because the rank function is lower-semicontinuous (but not locally Lipschitz), we only focus on the lower-semicontinuous case.

The notions of generalized subdifferentials of a lower-semicontinuous function were mostly introduced in the 70s-80s. The FRÉCHET subdifferential can be traced back to BAZARAA and GOODE ([5]). A few years later, the concept of proximal subdifferential was defined by ROCKAFELLAR in [57] (1981). And then, in 1983, GRANDALL and LIONS introduced the concept of viscosity solution of a Hamilton-Jacobi equation ([13]). Two other types of generalized subdifferentials are: the limiting and the CLARKE ones, proposed by MORDUKHOVICH and CLARKE.

We begin by recalling the definitions and some properties of several types of the generalized subdifferentials: FRÉCHET, proximal, viscosity, limiting and CLARKE.

Let $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ be proper, lower-semicontinuous (l.s.c) and $\tilde{x} \in \text{dom} f$, *i.e.* $f(x) < +\infty$.

Definition 2.1. A vector $x^* \in \mathbb{R}^p$ is a *F-subderivative* of f at $\tilde{x}$ if

$$\liminf_{y \to 0} \frac{f(\tilde{x}+y) - f(\tilde{x}) - \langle x^*, y\rangle}{\|y\|} \geq 0. \tag{2.1}$$

The set of all F-subderivatives of f at $\tilde{x}$ is called the FRÉCHET *subdifferential* of f at $\tilde{x}$, and denoted as $\partial^F f(\tilde{x})$.

Definition 2.2. A vector $x^* \in \mathbb{R}^p$ is a *viscosity subderivative* of f at $\tilde{x}$ if there exists a C^1-function $g : \mathbb{R}^p \to \mathbb{R}$ such that $\nabla g(\tilde{x}) = x^*$ and $f - g$ attains a local minimum at $\tilde{x}$. If, in particular,

$$g(x) = \langle x^*, x - \tilde{x}\rangle - \sigma\|x - \tilde{x}\|^2$$

with some positive constant σ, then x^* is called a *proximal subgradient* of f at $\tilde{x}$.

The set of all viscosity subderivatives and proximal subgradients of f at $\tilde{x}$ are called the *viscosity subdifferential* and the *proximal subdifferential* of f at $\tilde{x}$ and denoted as $\partial^V f(\tilde{x})$ and $\partial^P f(\tilde{x})$, respectively.

In a finite dimensional context, the FRÉCHET and the viscosity subdifferentials coincide. And this common subdifferential is also called "regular subdifferential" in some other works (see [48], [49], [58]).

Definition 2.3. A vector $x^* \in \mathbb{R}^p$ is a *limiting subgradient* of f at $\tilde{x}$ if there is a sequence of points x^r in $\mathbb{R}^p$ approaching $\tilde{x}$ with values $f(x^r)$ approaching the finite value $f(\tilde{x})$, and a sequence of y^r in $\partial^F f(x^r)$ approaching x^*.

The set of all limiting subgradients is called the *limiting subdifferential* and denoted as $\partial^L f(\tilde{x})$.

Definition 2.4. The CLARKE *subdifferential* $\partial^C f(\tilde{x})$ of f at $\tilde{x}$ is the set of all $x^* \in \mathbb{R}^p$ such that

$$\forall v \in \mathbb{R}^p, \quad \langle x^*, v\rangle \leq f^0(\tilde{x}, v) := \lim_{\varepsilon \downarrow 0} \limsup_{\substack{y \downarrow_f \tilde{x} \\ t \downarrow 0}} \inf_{w \in v + \varepsilon B} \frac{f(y+tw) - f(y)}{t}, \tag{2.2}$$

where B is the unit ball in $\mathbb{R}^p$ and $y \downarrow_f \tilde{x}$ signifies that y and $f(y)$ converge to $\tilde{x}$ and $f(\tilde{x})$, respectively.

Definition 2.5 (normal cone)**.** A vector $v \in \mathbb{R}^p$ is *normal* to a closed set $\Omega \subset \mathbb{R}^p$ at $\tilde{x} \in \Omega$, written $v \in N_\Omega(\tilde{x})$, if there are sequence $(x^k)_{k\in\mathbb{N}}$ in Ω with $x^k \to_\Omega \tilde{x}$ and $(v^k)_{k\in\mathbb{N}}$ in $\mathbb{R}^p$ with $v^k \to v$ such that

$$\limsup_{\substack{x \to_\Omega x^k \\ x \neq x^k}} \frac{\langle v^k, x - x^k\rangle}{|x - x^k|} \leq 0. \tag{2.3}$$

The vectors v^k satisfy (2.3) as above are FRÉCHET(regular) normals to Ω at x^k and the cone of FRÉCHET normals at x^k is denoted $\hat{N}_\Omega(x^k)$.

Remark 2.6. The limiting subdifferential of f at $\tilde{x}$ can be defined as the set of x^* for which $(x^*, -1)$ lies in the normal cone of epi$f$ at $(\tilde{x}, f(\tilde{x}))$. We can also define the CLARKE subdifferential of f at $\tilde{x}$ as the set of x^* for which $(x^*, -1)$ lies in the closed-convex hull of the normal cone of epi$f$ at $(\tilde{x}, f(\tilde{x}))$ (see [12]). Thus, $\partial^C f(\tilde{x})$ is closed and convex for every $\tilde{x}$ in $\mathbb{R}^p$.

Moreover, since we are in a finite dimensional context, we have the next string of inclusions

$$\partial^P f(\tilde{x}) \subset \partial^V f(\tilde{x}) = \partial^F f(\tilde{x}) \subset \partial^L f(\tilde{x}) \subset \partial^C f(\tilde{x}). \tag{2.4}$$

Proposition 2.7. *(*Local extrema, [59]*) If f attains a local minimum at x, then* 0 *belongs to the proximal subdifferential of f at x.*

Theorem 2.8 (Sum rule, [59]). *Let $f_1, f_2 : \mathbb{R}^p \to \mathbb{R}$ be proper, lower-semicontinuous, f_1 is Fréchet differentiable at x. Then*

$$\partial^F(f_1 + f_2)(\tilde{x}) = \nabla f_1(\tilde{x}) + \partial^F f_2(\tilde{x}).$$

2.2 The generalized subdifferentials of the counting function

Recall that the so-called counting function is defined as follows:

$$\begin{array}{rcl} c : \mathbb{R}^p & \to & \mathbb{R} \\ x & \mapsto & c(x) := \text{number of } i\text{'s such that } x_i \neq 0. \end{array}$$

In the next two theorems, we prove that all the generalized subdifferentials of the counting function coincide and provide a simple formula for the common subdifferential.

Theorem 2.9. *For all $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$*

$$\partial^P c(x) = \partial^V c(x) = \partial^F c(x) = X^{\perp}(x), \tag{2.5}$$

where $X^{\perp}(x) = \{x^* \in \mathbb{R}^p | \quad x_i^* = 0$ for those i such that $x_i \neq 0\}$.

Proof. Let

$$I(x) = \{i \in 1, \ldots, p| \quad x_i = 0\},$$

$$X(x) = \{y \in \mathbb{R}^p| \quad y_i = 0 \quad \text{for all } i \in I(x)\}.$$

It is easy to see that every point in $\mathbb{R}^p$ is a local minimum of the counting function. Thus, there exists a positive δ such that, whenever z is in $B(x, \delta)$, we have

$$c(z) \geq c(x) \tag{2.6}$$

and

$$c(z) = c(x) \Leftrightarrow z \in X(x). \tag{2.7}$$

First step. We prove that

$$\partial^F c(x) \subset X^{\perp}(x). \tag{2.8}$$

By the definition of the FRÉCHET subdifferential, we have

$$\begin{aligned} & x^* \in \partial^F c(x) \\ \Leftrightarrow \quad & \liminf_{y \to 0} \frac{c(x+y) - c(x) - \langle x^*, y \rangle}{\|y\|} \geq 0. \end{aligned}$$

Consequently,

$$\liminf_{\tau \to 0} \frac{c(x+\tau y) - c(x) - \tau \langle x^*, y \rangle}{\tau \|y\|} \geq 0. \tag{2.9}$$

Let $y \in X(x)$. There exists $\varepsilon > 0$ such that

$$\forall \tau \in [0, \varepsilon], \quad x + \tau y \in B(x, \delta).$$

Then, from the fact that $X(x)$ is a vector space and (2.7), we obtain that $c(x + \tau y) = c(x)$ for $\tau \in [0, \varepsilon]$.

Now, (2.9) becomes

$$\liminf_{\tau \to 0} -\frac{\tau \langle x^*, y \rangle}{\tau \|y\|} \geq 0 \quad \text{for all nonzero } y \in X(x).$$

Thus, $\langle x^*, y \rangle \leq 0$ for all $y \in X(x)$. This means that $\langle x^*, y \rangle = 0$ for all $y \in X(x)$ because $X(x)$ is a vector space.

So we have proved that

$$x^* \in X^{\perp}(x).$$

Second step. Now, we prove that

$$X^{\perp}(x) \subset \partial^P c(x). \tag{2.10}$$

Indeed, for $x^* \in X^{\perp}(x)$, we consider the function

$$g(y) = \langle x^*, y - x \rangle - \sigma \|y - x\|^2$$

with $\sigma > 0$.
So,

$$(c - g)(y) = c(y) - \langle x^*, y - x \rangle + \sigma \|y - x\|^2.$$

Certainly, $x^* = 0$ belongs to $\partial^P c(x)$. For $x^* \neq 0$, we set $\xi = \min\{\frac{1}{2\|x^*\|_\infty}; \delta\}$, where $\|x^*\|_\infty = \sup_{\|y\| \leq 1} \langle x^*, y \rangle$.

For $y \in B(x, \xi)$, we have:

- If $y - x \in X(x)$, then $\langle x^*, y - x \rangle = 0$. From $y \in B(x, \xi) \subset B(x, \delta)$, we infer that $c(y) = c(x)$. Hence,

$$(c - g)(y) = \sigma \|y - x\|^2 + c(x) \geq c(x).$$

- If $y - x \notin X(x)$, then $c(y) > c(x)$ or $c(y) \geq c(x) + 1$. Hence,

$$\begin{aligned}
(c - g)(y) &= c(y) - \langle x^*, y - x \rangle + \sigma \|y - x\|^2 \\
&\geq c(x) + 1 - \|y - x\| \langle x^*, \frac{y - x}{\|y - x\|} \rangle + \sigma \|y - x\|^2 \\
&\geq c(x) + 1 - \frac{1}{2\|x^*\|_\infty} \|x^*\|_\infty + \sigma \|y - x\|^2 \\
&= c(x) + \tfrac{1}{2} + \sigma \|y - x\|^2 \\
&> c(x).
\end{aligned}$$

  Thus, $(c - g)$ attains a local minimum at x. So, remembering the definition of $\partial^P f(x)$,

$$x^* \in \partial^P c(x).$$

  We thus have proved that

$$X^\perp(x) \subset \partial^P c(x).$$

From (2.4),(2.8),(2.10), we deduce that

$$\partial^P c(x) = \partial^V c(x) = \partial^F c(x) = X^\perp(x) \quad \forall x \in \mathbb{R}^p.$$

□

In the next theorem, we prove that the Clarke subdifferential of c at x also equals $X^\perp(x)$.

Theorem 2.10. *For all $x \in \mathbb{R}^p$,*

$$\partial^C c(x) = X^\perp(x). \tag{2.11}$$

Proof. Recall that the CLARKE subdifferential of c at x is the set of $x^* \in \mathbb{R}^p$ such that:

$$\forall v \in \mathbb{R}^p, \quad \langle x^*, v\rangle \leq c^0(x,v) = \lim_{\varepsilon \downarrow 0} \limsup_{\substack{y \downarrow_c x \\ t \downarrow 0}} \inf_{w \in v+\varepsilon B} \frac{c(y+tw)-c(y)}{t}.$$

As we have seen in the proof of Theorem 2.9, there exists a positive δ such that, for y in $B(x,\delta)$,

$$c(y) \geq c(x)$$

and

$$c(y) = c(x) \Leftrightarrow y \in X(x).$$

Thus, for y in $B(x,\delta)$

$$y \downarrow_c x \Leftrightarrow \begin{cases} y \to x \\ y \in X(x). \end{cases}$$

Then,

$$c^0(x,v) = \lim_{\varepsilon \downarrow 0} \limsup_{\substack{y \to x \\ y \in X(x) \\ t \downarrow 0}} \inf_{w \in v+\varepsilon B} \frac{c(y+tw)-c(y)}{t}.$$

For $y \in B(x, \frac{\delta}{2})$ and $y \in X(x)$, we have: $c(y) = c(x)$ and

$$\forall t < \frac{\delta}{2}, \quad \begin{cases} c(y+tw) \geq c(x)+1 = c(y)+1 & \text{if } w \notin X(x) \\ c(y+tw) = c(x) = c(y) & \text{if } w \in X(x). \end{cases}$$

Hence, for $t < \frac{\delta}{2}$ and for any w, we have $c(y+tw) \geq c(y)$. So,

$$\inf_{w \in v+\varepsilon B} \frac{c(y+tw)-c(y)}{t} \geq 0.$$

- If $v \in X(x)$, then $c(y+tv) = c(y)$. Thus, for $t < \frac{\delta}{2}, y \in B(x, \frac{\delta}{2})$ and $y \in X(x)$

$$\inf_{w \in v+\varepsilon B} \frac{c(y+tw)-c(y)}{t} = 0.$$

This shows that

$$c^0(x,v) = 0.$$

- If $v \notin X(x)$, then there exists a positive ε such that $B(y,\varepsilon) \cap X(x) = \emptyset$. Thus, for $t < \frac{\delta}{2}, y \in B(x, \frac{\delta}{2})$ and $y \in X(x)$

$$\inf_{w \in v+\varepsilon B} \frac{c(y+tw) - c(y)}{t} = \frac{1}{t}.$$

  This implies that

$$c^0(x,v) = +\infty.$$

We finally obtain

$$c^0(x,v) = \begin{cases} 0 & \text{if } v \in X(x) \\ +\infty & \text{otherwise.} \end{cases}$$

Consequently,

$$\partial^C c(x) = X^\perp(x).$$

□

Remark 2.11. To conclude, all types of generalized subdifferentials coincide and are equal to $X^\perp(x)$. And although not computed directly here, the limiting subdifferential $\partial^L c(x)$, caught between $\partial^F c(x)$ and $\partial^C c(x)$, also equals $X^\perp(x)$.

2.3 The generalized subdifferentials of the rank function

The rank function and the counting function share many common properties. Firstly, we tried to calculate the generalized subdifferentials of the rank function by the same method as the one we used in the above part. But we only deduced an inclusion because of the appearance of singular values. Fortunately, thanks to the works of Lewis and Sendov in [48],[49], we are able to obtain the subdifferentials of the rank function from Theorems 2.9 and 2.10.

Before going further, let us fix some notation:

- $\mathcal{M}_{m,n}(\mathbb{R})$ is the set of real matrices with m columns and n rows.
- For $x \in \mathbb{R}^p$, let $\mathrm{diag}_{m,n}(x)$ denote an $m \times n$ matrix with entries $\mathrm{diag}_{m,n}(x)^{i,i} = x_i$ for all i, and $\mathrm{diag}_{m,n}(x)^{i,j} = 0$ for $i \neq j$.

- $O(n)$ is the set of orthogonal matrices in $\mathcal{M}_n(\mathbb{R})$.
- $O(m,n) = \{(U,V)| \quad U \in O(m), V \in O(n)\}$.
- $O(m,n)^A = \{U \in O(m), V \in O(n)| \quad U\text{diag}_{m,n}(\sigma(A))V^T = A\}$.
- $S_k = \{A| \quad \text{rank } A \leq k\}$.
- For a matrix $A \in \mathcal{M}_{m,n}(\mathbb{R})$, $\sigma(A) = (\sigma_1(A), \ldots, \sigma_p(A))$ denotes the vector of singular values of A.
- $f \circ \sigma(A) = f(\sigma(A))$.

2.3.1 Nonsmooth analysis of singular values

The nonsmoothness of an absolutely symmetric function (*cf.* definition below) of the singular values of a real rectangular matrix was analysed by Lewis and Sendov. They gave simple formula for the generalized subdifferentials of such functions for both the Lipschitz and lower-semicontinuous case. As we said before, we are only interested in the lower-semicontinuous case.

Let $f : \mathbb{R}^p \to \mathbb{R}$ be an *absolutely symmetric* function, *i.e.* satisfying

$$f(x_1, \ldots, x_p) = f(\hat{x}_1, \ldots, \hat{x}_p) \quad \text{for all } x \in \mathbb{R}^p.$$

where $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_p)$ is the vector, built up from $x = (x_1, \ldots, x_p)$, whose components are the $|x_i|$'s arranged in a decreasing order.

Theorem 2.12 ([49]). *If $A \in \mathcal{M}_{m,n}(\mathbb{R})$ and if f is an absolutely symmetric function, lower-semicontinuous around $\sigma(A)$, then $f \circ \sigma$ is lower-semicontinuous around A and*

$$\begin{aligned} \partial^C(f \circ \sigma)(A) &= O(m,n)^A.\text{diag}_{m,n}\partial^C(f(\sigma(A)) \\ &= \{U.\text{diag}_{m,n}(y).V^T| \quad y \in \partial^C(f(\sigma(A)), (U,V) \in O(m,n)^A\} \end{aligned}$$

$$\begin{aligned} \partial^F(f \circ \sigma)(A) &= O(m,n)^A.\text{diag}_{m,n}\partial^F(f(\sigma(A)) \\ &= \{U.\text{diag}_{m,n}(y).V^T| \quad y \in \partial^F(f(\sigma(A)), (U,V) \in O(m,n)^A\} \end{aligned}.$$

Theorem 2.13 ([49]). *If $A \in \mathcal{M}_{m,n}(\mathbb{R})$ and if f is an absolutely symmetric function, lower-semicontinuous around $\sigma(A)$, then the proximal subdifferential of*

any singular value function $f \circ \sigma$ at A is given by the formula

$$\begin{aligned}\partial^P(f\circ\sigma)(A) &= O(m,n)^A.\mathrm{diag}_{m,n}\partial^P(f(\sigma(A))\\ &= \{U.\mathrm{diag}_{m,n}(y).V^T| \quad y\in\partial^P(f(\sigma(A)),(U,V)\in O(m,n)^A\}.\end{aligned}$$

2.3.2 Generalized subdifferentials of the rank function

Theorem 2.14. *All the generalized subdifferentials (proximal,* Fréchet*, viscosity, limiting,* Clarke*) of the rank function coincide. We denote the common subdifferential by $\partial(\mathrm{rank})$. For $A\in\mathcal{M}_{m,n}(\mathbb{R})$, $\partial(\mathrm{rank})(A)$ is constructed as follows:*

- *Consider the matrices $U\in O(m)$ and $V\in O(n)$ such that*

$$U.\mathrm{diag}_{m,n}(\sigma(A)).V^T = A$$

  (in other words, we collect all the orthogonal matrices U and V which give a singular value decomposition of A).

- *Consider the "diagonal" matrices $\mathrm{diag}_{m,n}(x^*)$, where $x^*\in\mathbb{R}^p$ is such that $x_i^*=0$ for all $i=1,\dots,r$ (recall that $r=\mathrm{rank}\ A$).*

- *Then, collect all the matrices of the form $U\mathrm{diag}_{m,n}(x^*)V^T$.*

In a single formula,

$$\begin{aligned}&\partial(\mathrm{rank})(A)\\ &= \{U\mathrm{diag}_{m,n}(x^*)V^T| \quad U\in O(m),V\in O(n) \text{ such that } U.\mathrm{diag}_{m,n}(\sigma(A)).V^T=A,\\ &\qquad\qquad x_i^*=0 \text{ for all } i=1,\dots,r\}.\end{aligned}$$

Proof. It is well-known that

$$\mathrm{rank}\ A = c\circ\sigma(A),$$

and that c is lower-semicontinuous, absolutely symmetric. Moreover, by Theorems 2.9 and 2.10, all the subdifferentials of c coincide and are equal to $X^\perp$. By applying Theorems 2.12 and 2.13, we obtain that the Clarke, Fréchet and the proximal subdifferentials of the rank function are the same. So, all the subdifferentials of

the rank are given by

$$\begin{aligned}\partial(\text{rank})(A) &= O(m,n)^A.\text{diag}_{m,n}\partial c(\sigma(A)) \\ &= O(m,n)^A.\{\text{diag}_{m,n}(x^*)| \quad x_i^* = 0 \text{ for all } i = 1,...r\} \\ &= \{U\text{diag}_{m,n}(x^*)V^T| \quad U \in O(m), V \in O(n) \\ &\quad \text{such that } U.\text{diag}_{m,n}(\sigma(A)).V^T = A, x_i^* = 0 \text{ for all } i = 1,\ldots,r\}.\end{aligned}$$

□

Remark 2.15. The limiting subdifferential of the rank function can be computed in another way: using the relationship between the limiting subdifferential and the normal cone.

Indeed, let A be a matrix in $\mathcal{M}_{m,n}(\mathbb{R})$ and rank $A = r$. A matrix $X \in \mathcal{M}_{m,n}(\mathbb{R})$ is a F-subderivative of the rank function at A if and only if

$$\liminf_{B\to 0} \frac{\text{rank}(A+B) - \text{rank}(A) - \langle X, B\rangle}{\|B\|_F} \geq 0. \tag{2.12}$$

But, the rank of a matrix in a sufficient small neighborhood of A is an integer number at least equal to rank of A. Thus, (2.12) is equivalent to

$$\limsup_{A+B\to_{S_r} B} \frac{\langle X, B\rangle}{\|B\|_F} \leq 0.$$

This means that X is a FRÉCHET normal to S_r at A. From the definitions of the limiting subdifferential and the normal cone, and the fact $S_r \cap B(A,\varepsilon) \subset \Sigma_r$ for ε small enough, we can conclude

$$X \in \partial^L(\text{rank})(A) = N_{S_r}(A).$$

LUKE has proposed an explicit formula for the normal cone to S_k at any matrix A in S_k ([52]). If we consider the case where $k = r = \text{rank } A$, we obtain the same formula for the limiting subdifferential as in Theorem 2.14.

A further property of $\partial(\text{rank})(A)$ is that it is not only a closed convex set but a vector space.

Proposition 2.16. *For $A \in \mathcal{M}_{m,n}(\mathbb{R})$, $\partial(\text{rank})(A)$ is a vector space.*

Proof. For $N \in \partial(\text{rank})(A)$ and $k \in \mathbb{R}$, we first prove that

$$k.N \in \partial(\text{rank})(A).$$

Indeed, for $N \in \partial(\text{rank})(A)$, there exist $(U, V) \in O(m, n)^A$ and $x^* \in \mathbb{R}^p$, with $x_i^* = 0$ for all $i = 1, \ldots, r$, such that

$$N = U\text{diag}_{m,n}(x^*)V^T.$$

Then

$$k.N = k.U\text{diag}_{m,n}(x^*)V^T = U\text{diag}_{m,n}(k.x^*)V^T.$$

This means that $k.N \in \partial(\text{rank})(A)$.

Now, for $N_1, N_2 \in \partial(\text{rank})(A)$, we prove that

$$N_1 + N_2 \in \partial(\text{rank})(A).$$

On the one hand, $2N_1$ and $2N_2$ are also in $\partial(\text{rank})(A)$ as in the first part. On the other hand, the CLARKE subdifferential is always convex. It means that $\partial(\text{rank})(A)$ is convex for all A.

Hence, $N_1 + N_2 = \frac{1}{2}(2N_1 + 2N_2) \in \partial(\text{rank})(A)$. □

An alternate expression of $\partial(\text{rank})(A)$ is possible. The subdifferential of the rank function can be also represented as the tensor product of two vector spaces in $\mathbb{R}^m$ and $\mathbb{R}^n$, as indicated in the following proposition.

Proposition 2.17. *Let $N(A)$ and $N(A^T)$ be the null spaces of matrices A and A^T, respectively. Then*

$$\partial(\textit{rank})(A) \quad = N(A^T) \otimes N(A)$$

where $\otimes$ is the tensor product. In a more detailed form,

$$N(A^T) \otimes N(A) = \left\{ \sum a_{ij}\alpha_i\beta_j^T \middle| \begin{array}{l} (\alpha_i) \textit{ is a basis of } N(A^T) \\ (\beta_j) \textit{ is a basis of } N(A) \end{array} \right\}.$$

Consequentely, the dimension of $\partial(\textit{rank})(A)$ is $(m-r)(n-r)$, where r is the rank of A.

Proof. Let $u_1, \ldots, u_m$ be the columns of U and $v_1, \ldots, v_n$ be the columns of V. Recall that in Theorem 2.14, we already have

$$\begin{aligned}&\partial(\text{rank})(A)\\ &= \{U\text{diag}_{m,n}(x^*)V^T| \quad U \in O(m), V \in O(n) \text{ such that } U.\text{diag}_{m,n}(\sigma(A)).V^T = A,\\ &\qquad\qquad x_i^* = 0 \text{ for all } i = 1, \ldots, r\}.\end{aligned} \tag{2.13}$$

Since the first r components of x^* are zero, then we can rewrite (2.13) as following:

$$\begin{aligned}\partial(\text{rank})(A) = \{\textstyle\sum_{i=r+1}^{p} x_i^*.u_iv_i^T| \quad &x_i^* \in \mathbb{R}, U \in O(m), V \in O(n)\\ &\text{such that } U.\text{diag}_{m,n}(\sigma(A)).V^T = A\}.\end{aligned}$$

From the facts that $U.\text{diag}_{m,n}(\sigma(A)).V^T$ is a singular value decomposition of A and rank $A = r$, we have

$$A = \sum_{i=1}^{r} \sigma_i.u_iv_i^T.$$

Moreover, $\{v_1, \ldots, v_n\}$ is an orthogonal basis of $\mathbb{R}^n$. Hence

$$Av_i = 0 \quad \text{for all } i = r+1, \ldots, n.$$

It means that $\{v_{r+1}, \ldots, v_n\}$ can be any orthogonal basis of $N(A)$.

Similarly, $\{u_{r+1}, \ldots, u_m\}$ can be any orthogonal basis of $N(A^T)$. It implies that $\{u_iv_j^T\}$ is a basis of $N(A^T) \otimes N(A)$.

On another hand, $u_iv_j^T$ is an element of the vector space $\partial(\text{rank})(A)$. Thus,

$$N(A^T) \otimes N(A) \subset \partial(\text{rank})(A).$$

Clearly, $\partial(\text{rank})(A) \subset N(A^T) \otimes N(A)$. So, we obtain

$$N(A^T) \otimes N(A) = \partial(\text{rank})(A).$$

As we know, $N(A)$ and $N(A^T)$ are vector spaces of dimensions $n - r$ and $m - r$, respectively. Then, the tensor product of them is a vector space of dimension $(m - r)(n - r)$. This means that the dimension of $\partial(\text{rank})(A)$ is $(m - r)(n - r)$.

□

We illustrate our results by considering the special case where $m = n = 2$.

Example 2.1. *For $m = n = 2$, we have*

- *If $A = 0$ then $\partial(\text{rank})(0) = \mathcal{M}_{2,2}(\mathbb{R})$.*
- *If* rank $A = 2$ *then $\partial(\text{rank})(A) = \{0\}$.*
- *If* rank $A = 1$ *then $\partial(\text{rank})(A)$ is a vector space of dimension* 1 *(cf. Proposition 2.17), of the form $\{kA_0| \quad k \in \mathbb{R}\}$. An explicit form of A_0 will be given in the proof.*

Proof.

- $A = 0$: By Theorem 2.14, we have

$$\begin{aligned}\partial(\text{rank})(0) &= \{U\text{diag}_{m,n}(x^*)V^T| \quad U, V \in O(2) \text{ such that } U.0.V^T = 0, x^* \in \mathbb{R}^2\}\\ &= \{U\text{diag}_{m,n}(x^*)V^T| \quad U, V \in O(2), x^* \in \mathbb{R}^2\}.\end{aligned}$$

  So, the subdifferential of the rank function at 0 is the set of all real matrices 2×2.

- rank $A = 2$: By Theorem 2.14, we have

$$\partial(\text{rank})(A) = \{U\text{diag}_{m,n}(0,0)V^T| \quad (U, V) \in O(2,2)^A\} = \{0\}.$$

  This occurs on an open dense set of $\mathcal{M}_{2,2}(\mathbb{R})$.

- rank $A = 1$: We have

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{with} \quad \begin{cases} a, b, c, d \in \mathbb{R} \\ a^2 + b^2 + c^2 + d^2 > 0 \\ ad = bc. \end{cases}$$

  Without loss of generality, we assume that $a \neq 0$ and take $\alpha_0, \beta_0 \in (0; \pi)$ such that

$$\cot \alpha_0 = \frac{c}{a} \qquad \cot \beta_0 = \frac{b}{a}.$$

Because rank $A = 1$, the singular values of A are σ_1 and 0. For $(U, V) \in O(2,2)^A$,

$$\begin{aligned} A &= U \begin{pmatrix} \sigma_1 & 0 \\ 0 & 0 \end{pmatrix} V^T \\ &= \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix} \\ &= \sigma_1 \begin{pmatrix} u_{11}v_{11} & u_{12}v_{12} \\ u_{21}v_{21} & u_{22}v_{22} \end{pmatrix}. \end{aligned} \tag{2.14}$$

Using Theorem 2.14, we obtain

$$\begin{aligned} \partial(\text{rank})(A) &= \{U\text{diag}(0,k)V^T | \quad k \in \mathbb{R}, (U,V) \in O(2,2)^A\} \\ &= \{k \begin{pmatrix} u_{12}v_{12} & u_{12}v_{22} \\ u_{22}v_{12} & u_{22}v_{22} \end{pmatrix} | \quad k \in \mathbb{R}, (U,V) \in O(2,2)^A\} \\ &= \{ku_2v_2^T | \quad k \in \mathbb{R}, (U,V) \in O(2,2)^A\}. \end{aligned}$$

Case 1: $\det U = \det V = 1$.
There exist $\alpha, \beta \in$ such that

$$U = \begin{pmatrix} \sin\alpha & \cos\alpha \\ -\cos\alpha & \sin\alpha \end{pmatrix}; \qquad V = \begin{pmatrix} \sin\beta & \cos\beta \\ -\cos\beta & \sin\beta \end{pmatrix}.$$

Then,(2.14) implies that

$$\begin{cases} \cot\alpha = -\cot\alpha_0 \\ \cot\beta = -\cot\beta_0 \end{cases} \Leftrightarrow \begin{cases} \alpha = \begin{bmatrix} -\alpha_0 \\ \pi - \alpha_0 \end{bmatrix} \\ \beta = \begin{bmatrix} -\beta_0 \\ \pi - \beta_0 \end{bmatrix} \end{cases}.$$

Then,

$$\begin{aligned} u_2v_2^T &= \begin{pmatrix} \cos\alpha\cos\beta & \cos\alpha\sin\beta \\ \sin\alpha\cos\beta & \sin\alpha\sin\beta \end{pmatrix} \\ &= \pm \begin{pmatrix} \cos\alpha_0\cos\beta_0 & -\cos\alpha_0\sin\beta_0 \\ -\sin\alpha_0\cos\beta_0 & \cos\alpha_0\cos\beta_0 \end{pmatrix} \\ &= \pm A_0. \end{aligned}$$

Case 2: $\det U = 1, \det V = -1$.
There exist α, β such that

$$U = \begin{pmatrix} \sin\alpha & \cos\alpha \\ -\cos\alpha & \sin\alpha \end{pmatrix}; \qquad V = \begin{pmatrix} \sin\beta & \cos\beta \\ \cos\beta & -\sin\beta \end{pmatrix}.$$

Then,(2.14) implies that

$$\begin{cases} \cot\alpha = -\cot\alpha_0 \\ \cot\beta = \cot\beta_0 \end{cases} \Leftrightarrow \begin{cases} \alpha = \begin{bmatrix} -\alpha_0 \\ \pi - \alpha_0 \end{bmatrix} \\ \\ \beta = \begin{bmatrix} \beta_0 \\ \pi + \beta_0 \end{bmatrix} \end{cases}.$$

Then,

$$\begin{aligned} u_2 v_2^T &= \begin{pmatrix} \cos\alpha\cos\beta & -\cos\alpha\sin\beta \\ \sin\alpha\cos\beta & -\sin\alpha\sin\beta \end{pmatrix} \\ &= \pm \begin{pmatrix} \cos\alpha_0\cos\beta_0 & -\cos\alpha_0\sin\beta_0 \\ -\sin\alpha_0\cos\beta_0 & \cos\alpha_0\cos\beta_0 \end{pmatrix} \\ &= \pm A_0. \end{aligned}$$

By doing the same for the last two cases, we obtain $u_2 v_2^T \in \{\pm A_0\}$ for all $(U, V) \in O(2,2)^A$. We conclude that

$$\partial(\text{rank})(A) = \{kA_0 | \quad k \in \mathbb{R}\},$$

a vector space of dimension 1.

□

Chapter 3

Regularization-Approximation of the rank function

We revisited, in Chapter 1, the relaxed form of the rank function, the nuclear norm. In this chapter, we consider another way to approach the rank minimization problem, using smooth or just continuous approximations R_ε of the rank function, depending on some parameter $\varepsilon > 0$. We propose here two classes of regularization-approximation of the rank function: the first one consists of smooth versions of the rank, the second one relies on the so-called Moreau-Yosida technique, widely used in the context of variational analysis. Then, from the generalized subdifferentials of the Moreau-Yosida approximation of the rank function, we can retrieve the main result of Chapter 2.

3.1 Smooth versions

This section is taken from [34].

Let θ be the function defined by

$$\begin{array}{rcl} \theta : \mathbb{R} & \longrightarrow & \{0,1\} \\ x & \mapsto & \theta(x) = \left\{ \begin{array}{ll} 1 & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{array} \right. . \end{array}$$

Then, the rank function can be presented as

$$\begin{aligned} \text{rank } A &= c[\sigma_1(A), \ldots, \sigma_p(A)] \quad \text{(recall that } c \text{ is the counting function on } \mathbb{R}^p) \\ &= \sum_{i=1}^{p} \theta[\sigma_i(A)]. \end{aligned} \tag{3.1}$$

In order to obtain a smooth regularization-approximation of the rank function, we need to design some smooth approximation of the θ function.

A first example was proposed in [29] by Hiriart-Urruty, it is as following: For $\varepsilon > 0$, let θ_ε be defined as

$$x \in \mathbb{R} \longmapsto \theta_\varepsilon(x) := 1 - e^{-x^2/\varepsilon}. \tag{3.2}$$

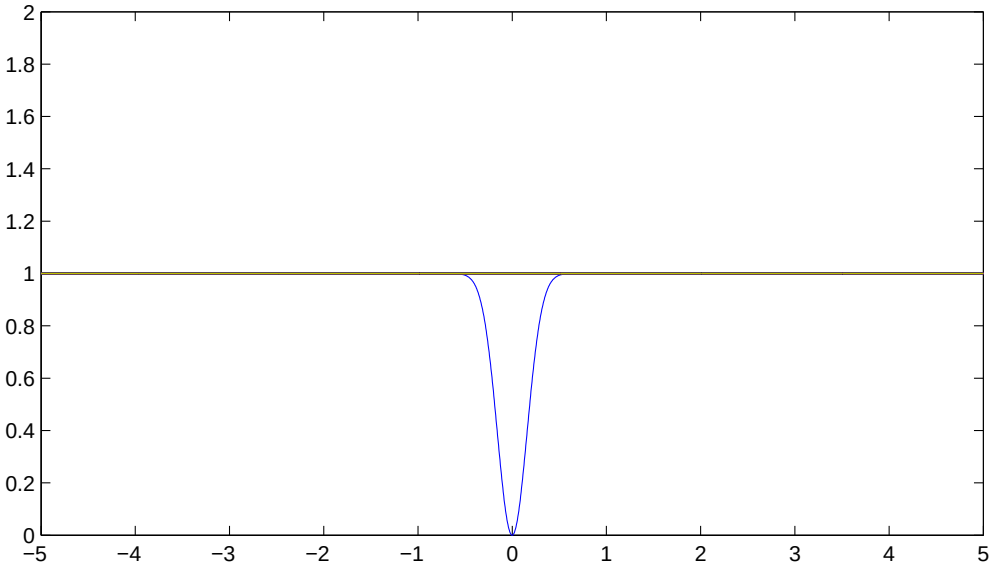


Figure 3.1: $\theta_{0.05}$

The resulting approximation of the rank function is

$$A \in \mathcal{M}_{m,n}(\mathbb{R}) \longmapsto R_\varepsilon(A) := \sum_{i=1}^{p} [1 - e^{-\sigma_i^2(A)/\varepsilon}]. \tag{3.3}$$

An alternate expression of the R_ε function is

$$R_\varepsilon(A) = p - \operatorname{tr}(e^{-A^T A/\varepsilon}). \tag{3.4}$$

Then, R_ε is a C^∞ (even analytic) function of A. The properties of R_ε as an approximation of the rank function are summarized in the statement below.

Theorem 3.1 ([34]). *We have*

(i) $R_\varepsilon(A) \leq \text{rank } A$ *for all* $\varepsilon > 0$.

(ii) The sequence of functions $(R_\varepsilon)_{\varepsilon>0}$ *increases when* ε *decreases, and* $R_\varepsilon(A) \to \text{rank } A$ *for all* A *when* $\varepsilon \to 0$.

(iii) If $A \neq 0$ *and* $r = \text{rank } A$,

$$\text{rank } A - R_\varepsilon(A) \leq \varepsilon \sum_{i=1}^{r} \frac{1}{\sigma_i^2(A)}, \tag{3.5}$$

as also

$$\text{rank } A - R_\varepsilon(A) \leq \varepsilon^2 \sum_{i=1}^{r} \frac{1}{\sigma_i^4(A)}. \tag{3.6}$$

Another proposal for approximating the rank function, a quite recent one, is due to Zhao ([61]). It consists of using, for all $\varepsilon > 0$, the following even approximation of the θ function:

$$x \in \mathbb{R} \longmapsto \tau_\varepsilon(x) := \frac{x^2}{x^2 + \varepsilon}. \tag{3.7}$$

The resulting approximation of the rank function is

$$A \in \mathcal{M}_{m,n}(\mathbb{R}) \longmapsto Z_\varepsilon(A) := \sum_{i=1}^{p} \frac{\sigma_i^2(A)}{\sigma_i^2(A) + \varepsilon}. \tag{3.8}$$

Alternate expressions of the Z_ε function are:

$$\begin{aligned} Z_\varepsilon(A) &= \text{tr}[A(A^TA + \varepsilon I_n)^{-1}A^T] \\ &= n - \varepsilon \text{tr}(A^TA + \varepsilon I_n)^{-1}. \end{aligned}$$

Here also, Z_ε is a C^∞ (even analytic) function of A. The properties of Z_ε as an approximation of the rank function are summarized in the next statement.

Theorem 3.2 (Zhao, [61]). *We have*

(i) $Z_\varepsilon(A) \leq \text{rank } A$ *for all* $\varepsilon > 0$.

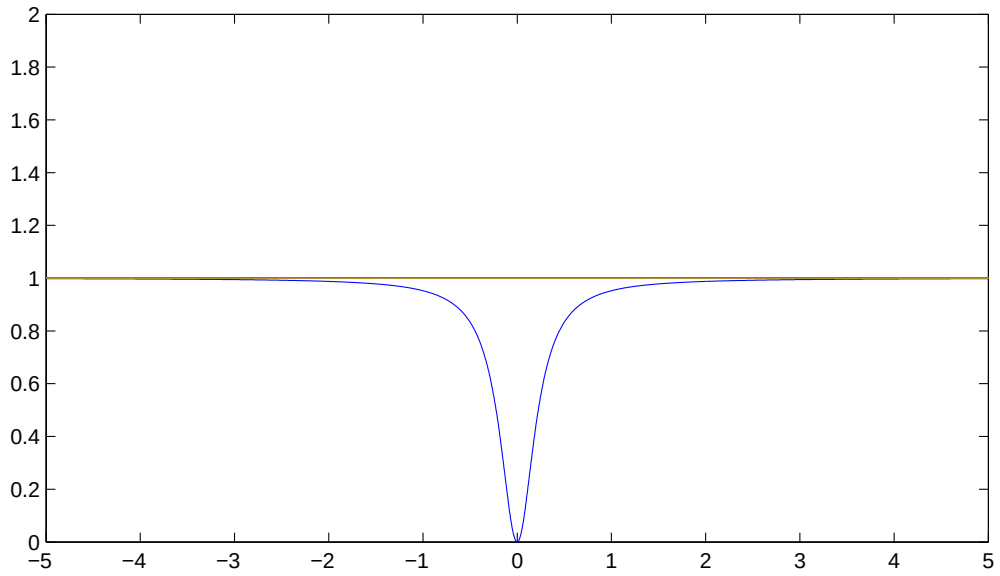


FIGURE 3.2: $\tau_{0.05}$

(ii) *The sequence of functions* $(Z_\varepsilon)_{\epsilon>0}$ *increases when* ε *decreases, and* $Z_\varepsilon(A) \to \operatorname{rank} A$ *for all* A *when* $\varepsilon \to 0$.

(iii) *If* $A \neq 0$ *and* $r = \operatorname{rank} A$,

$$\operatorname{rank} A - Z_\varepsilon(A) = \sum_{i=1}^{r} \frac{\varepsilon}{\sigma_i^2(A) + \varepsilon} \leq \varepsilon \sum_{i=1}^{r} \frac{1}{\sigma_i^2(A)}. \tag{3.9}$$

The use of this function Z_ε (instead of the rank function) in rank minimization problems as well as an application to solving a system of quadratic functions are discussed in ([61], Sections 3 and 4).

Another approximation of the counting function is the so-called scaled and shifted Fermi-Dirac entropy; it is defined and studied in [9].

3.2 Moreau-Yosida approximation

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-Yosida way. Moreover, the Moreau-Yosida approximation of the rank function can be computed explicitly. Let us firstly recall what is known, as a general rule, for

the Moreau-Yosida approximation-regularization technique in a nonconvex context (see [58, Section 1.G] for details, for example).

Let $(E, \|.\|)$ be an Euclidean space and $f : E \longrightarrow \mathbb{R} \cup \{+\infty\}$ be a lower-semicontinuous function, bounded from below on E. For a parameter value $\lambda > 0$, the Moreau-Yosida approximate (or Moreau envelope) function f_λ and proximal set-valued mapping $\text{Prox}_\lambda f$ are defined by

$$f_\lambda(x) := \inf_{u \in E} \{f(u) + \frac{1}{2\lambda}\|x - u\|^2\}, \tag{3.10}$$

$$\text{Prox}_\lambda f(x) := \left\{ \overline{u} \in E | \quad f(\overline{u}) + \frac{1}{2\lambda}\|x - \overline{u}\|^2 = f_\lambda(x) \right\}. \tag{3.11}$$

Then:

(i) f_λ is a finite-valued continuous function on E;

(ii) The sequence of function $(f_\lambda)_{\lambda>0}$ increases when λ decreases, and $f_\lambda(x) \to f(x)$ for all x when $\lambda \to 0$;

(iii) The set $\text{Prox}_\lambda f(x)$ is nonempty and compact;

(iv) The lower bounds of f and f_λ on E are equal:

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

We now apply this process to the rank function (or restricted rank function). The context is therefore as following: $E = \mathcal{M}_{m,n}(\mathbb{R})$, $\|.\|_F$ is the Frobenius-Schur norm and $f : E \longrightarrow \mathbb{R} \cup \{+\infty\}$ is the rank (or restricted rank) function. The Moreau-Yosida approximation rank_λ of the rank function is defined by

$$(\text{rank})_\lambda(A) = \inf_{B \in \mathcal{M}_{m,n}(\mathbb{R})} \left\{ \text{rank } B + \frac{1}{2\lambda}\|A - B\|_F^2 \right\}. \tag{3.12}$$

The Moreau-Yosida approximation of the restricted rank function is defined by

$$(\text{rank}_R)_\lambda(A) = \inf_{\substack{B \in \mathcal{M}_{m,n}(\mathbb{R}) \\ \|B\|_{sp} \leq R}} \left\{ \text{rank } B + \frac{1}{2\lambda}\|A - B\|_F^2 \right\}. \tag{3.13}$$

In a simpler context, we can also define the Moreau-Yosida approximation of the counting (or restricted counting) function as following:

$$c_\lambda(x) = \inf_{y\in\mathbb{R}^p}\left\{c(y) + \frac{1}{2\lambda}\|y-x\|^2\right\},$$

$$(c_R)_\lambda = \inf_{\substack{y\,\in\,\mathbb{R}^p \\ \|y\|_\infty\,\leq\,R}}\left\{c(y) + \frac{1}{2\lambda}\|y-x\|^2\right\}.$$

In the next proposition, we provide the formula for the Moreau-Yosida approximation of the counting function. This result was also observed in Example 5.4 of [1].

Proposition 3.3. *The Moreau-Yosida approximation of index $\lambda > 0$ of the counting function is given by:*

$$c_\lambda(x) = \frac{1}{2\lambda}\left(\|x\|^2 - \sum_{i=1}^{n}(|x_i|^2 - 2\lambda)^+\right).$$

And one element in $\mathrm{Prox}_\lambda(c)(x)$ *is provided by:*

$$y = (y_1, \ldots, y_n)$$

where

$$y_i = \begin{cases} x_i & \text{if } |x_i| > \sqrt{2\lambda}, \\ x_i \text{ or } 0 & \text{if } |x_i| = \sqrt{2\lambda}, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. By definition,

$$c_\lambda(x) = \inf_{y\in\mathbb{R}^n}\{c(y) + \frac{1}{2\lambda}\|y-x\|^2\}. \tag{3.14}$$

Since the counting function takes only integer values, the vector space $\mathbb{R}^n$ is the union of all sets T_k of vector y such that $c(y) = k$ for all $k = 0, 1, 2, \ldots, n$. By fixing the value of $c(y)$ (over the set T_k), the minimal value of the function $c(y) + \frac{1}{2\lambda}\|y-x\|^2$ can be easily computed. Indeed,

- If $k = c(x)$, then it is easy to see that

$$\min_{y\in T_k}\{c(y) + \frac{1}{2\lambda}\|y-x\|^2\} = k.$$

- If $k > c(x)$, then

$$\min_{y \in T_k}\{c(y) + \frac{1}{2\lambda}\|y - x\|^2\} \geq \min_{y \in S_k} c(y) = k.$$

- If $k < c(x)$, then

$$\min_{y \in T_k}\{c(y) + \frac{1}{2\lambda}\|y - x\|^2\} = k + \frac{1}{2\lambda} d(x, T_k)^2,$$

where $d(x, T_k)$ denotes the distance from x to T_k. Let $x^{\downarrow}$ be the vector of components of x being arranged in the non-increasing order of $|x_i|$, *i.e.* $|x_1^{\downarrow}| \geq \cdots \geq |x_n^{\downarrow}|$. Then, the distance from x to T_k is

$$d(x, T_k) = \sqrt{\sum_{i=k+1}^{n} |x_i^{\downarrow}|^2}.$$

So, we can rewrite (3.14) as

$$c_\lambda(x) = \min_{0 \leq k \leq c(x)} \{k + \frac{1}{2\lambda} \sum_{i=k+1}^{n} |x_i^{\downarrow}|^2\}.$$

Because

$$\begin{aligned} k + \frac{1}{2\lambda} \sum_{i=k+1}^{n} |x_i^{\downarrow}|^2 &= \frac{1}{2\lambda} \left[\sum_{i=1}^{n} |x_i^{\downarrow}|^2 - \sum_{i=1}^{k} (|x_i^{\downarrow}|^2 - 2\lambda) \right] \\ &= \frac{1}{2\lambda} \left[\|x\|^2 - \sum_{i=1}^{k} (|x_i^{\downarrow}|^2 - 2\lambda) \right], \end{aligned}$$

then

$$c_\lambda(x) = \frac{1}{2\lambda} \left[\|x\|^2 - \max_{0 \leq k \leq c(x)} \sum_{i=1}^{k} (|x_i^{\downarrow}|^2 - 2\lambda) \right].$$

Recall that

$$|x_1^{\downarrow}| \geq \cdots \geq |x_n^{\downarrow}|,$$

then

$$|x_1^{\downarrow}|^2 - 2\lambda \geq \cdots \geq |x_n^{\downarrow}|^2 - 2\lambda.$$

Hence, among the values of $(|x_1^{\downarrow}|^2 - 2\lambda), \sum_{i=1}^{2}(|x_i^{\downarrow}|^2 - 2\lambda), \ldots, \sum_{i=1}^{c(x)}(|x_i^{\downarrow}|^2 - 2\lambda)$, the largest term is the one that sums all positive $(|x_i^{\downarrow}|^2 - 2\lambda)$.

We conclude that

$$\begin{aligned} c_\lambda(x) &= \frac{1}{2\lambda}\left(\|x^\downarrow\|^2 - \sum_{i=1}^{c(x)}(|x_i|^2 - 2\lambda)^+\right) \\ &= \frac{1}{2\lambda}\left(\|x\|^2 - \sum_{i=1}^{n}(|x_i|^2 - 2\lambda)^+\right). \end{aligned}$$

A vector y is an element of $\mathrm{Prox}_\lambda(c)(x)$ if and only if

$$c_\lambda(x) = c(y) + \frac{1}{2\lambda}\|y - x\|^2.$$

This means that y is a projection of x onto $T_{\tilde{k}}$, where

$$\tilde{k} \in \mathrm{argmax}_{0\leq k\leq c(x)} \sum_{i=1}^{k}(|x_i^\downarrow|^2 - 2\lambda).$$

Hence, we can conclude that

$$y \in \mathrm{Prox}_\lambda(c)(x) \Leftrightarrow \forall i = 1,\ldots,n \quad y_i = \begin{cases} x_i & \text{if } |x_i| > \sqrt{2\lambda}, \\ x_i \text{ or } 0 & \text{if } |x_i| = \sqrt{2\lambda}, \\ 0 & \text{otherwise.} \end{cases}$$

□

The next theorem is a classical result that provides the distance from an arbitrary matrix to the set of matrices of rank at most k. This theorem is usually called the theorem of Eckart- Young or Eckart, Young and Mirsky, but in fact the first one who discovered it is Schmidt. By using this theorem, we can calculate the Moreau-Yosida approximation of the rank function. Conversely, we can get a best approximation of a matrix by a matrix of rank at most k from the Moreau-Yosida approximation of the rank function ([35]).

Theorem 3.4 (Eckart, Young and Mirsky [28]). *Given $A \in \mathcal{M}_{m,n}(\mathbb{R})$ of rank r, we consider the following problem:*

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

Let $U\Sigma_A V^T$ be a singular value decomposition of A with $\Sigma_A = \mathrm{diag}_{m,n}(\sigma_1(A), \ldots, \sigma_p(A))$. Choose $\|.\|$ as either $\|.\|_F$ or $\|.\|_{sp}$. Then

$$A_k := U\Sigma_k V^T,$$

(where Σ_k is obtained from Σ_A by keeping $\sigma_1(A), \ldots, \sigma_k(A)$ and putting 0 in the place of $\sigma_{k+1}(A), \ldots, \sigma_r(A)$) is a solution of the best approximation problem $(\mathcal{A}_k)$. For the Frobenius-Schur norm case, A_k is the unique solution in $(\mathcal{A}_k)$ when $\sigma_k(A) > \sigma_{k+1}(A)$.

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

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

We denote $O(m,n)^A$ the set of (U,V) such that U and V are orthogonal matrices and $U\mathrm{diag}_{m,n}(\sigma_1(A), \ldots, \sigma_p(A))V^T$ is a singular value decomposition of A. Then, in Theorem 1.11, *one solution* of the problem $(\mathcal{A}_k)$ is given by the formula $A_k = U\Sigma_k V^T$, with (U,V) fixed in $O(m,n)^A$. But in fact, *all the solutions* of $(\mathcal{A}_k)$ can be determined by

$$\tilde{U}\Sigma_k \tilde{V}^T,$$

with $(\tilde{U}, \tilde{V}) \in O(m,n)^A$.(see for example [53]). This means that the set of solutions of $(\mathcal{A}_k)$ is

$$\left\{ \tilde{U}\Sigma_k \tilde{V}^T | \quad (\tilde{U}, \tilde{V}) \in O(m,n)^A \right\}.$$

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

$$\tilde{U}_1 \Sigma_k \tilde{V}_1^T = \tilde{U}_2 \Sigma_k \tilde{V}_2^T,$$

for any $(\tilde{U}_1, \tilde{V}_1)$ and $(\tilde{U}_2, \tilde{V}_2)$ in $O(m,n)^A$. Hence, the set of solutions is a singleton, *i.e.* the solution of $(\mathcal{A}_k)$ is unique.

When $\sigma_k = \sigma_{k+1}$ for any $k = 1, 2, \ldots, p-1$, the problem $(\mathcal{A}_k)$ may have *infinitely many* solutions. The formula for the set of solutions is given in [52].

Lemma 3.5. *Let $A \neq 0$ be a matrix in $\mathcal{M}_{m,n}(\mathbb{R})$ and $k <$ rank A be an integer. We denote the set of matrices of rank k by R_k. Then,*

$$d(A, R_k) = d(A, S_k).$$

Proof. The theorem of Eckart-Young says that

$$d(A, S_k) = \sqrt{\sum_{i=k+1}^{\text{rank } A} \sigma_i^2(A)},$$

and a minimizer point is

$$A_k = U \text{diag}_{m,n}\left(\sigma_1(A), \ldots, \sigma_k(A), 0, \ldots, 0\right) V^T.$$

But R_k is a subset of S_k and A_k is contained in R_k, thus the distance from A to R_k exactly equals the distance from A to S_k. □

Theorem 3.6. *We have, for all $A \in \mathcal{M}_{m,n}(\mathbb{R})$ with rank $r \geq 1$:*

(i)

$$(\text{rank})_\lambda(A) = \frac{1}{2\lambda}\|A\|_F^2 - \frac{1}{2\lambda}\sum_{i=1}^{r}[\sigma_i^2(A) - 2\lambda]^+. \tag{3.15}$$

(ii) One minimizer in (3.12), i.e. one element in $\text{Prox}_\lambda(\text{rank})(A)$, is provided by $B := \tilde{U}\Sigma_B\tilde{V}^T$, where:

- *$(\tilde{U}, \tilde{V}) \in O(m,n)^A$, i.e. $\tilde{U}$ and $\tilde{V}$ are orthogonal matrices such that $A = \tilde{U}\Sigma_A\tilde{V}^T$, with $\Sigma_A = \text{diag}_{m,n}[\sigma_i(A), \ldots, \sigma_r(A), 0, \ldots, 0]$ (a singular value decomposition of A with $\sigma_1(A) \geq \cdots \geq \sigma_r(A) > 0$);*
- $$\Sigma_B = \begin{cases} 0 & \text{if } \sigma_1 \leq \sqrt{2\lambda}, \\ \Sigma_A & \text{if } \sigma_r(A) \geq \sqrt{2\lambda}, \\ \text{diag}_{m,n}[\sigma_1(A), \ldots, \sigma_k(A), 0, \ldots, 0] & \text{if there is an integer } k \text{ s.t.} \\ & \sigma_k(A) \geq \sqrt{2\lambda} > \sigma_{k+1}(A). \end{cases} \tag{3.16}$$

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

- If $\sigma_1(A) < \sqrt{2\lambda}$, then $\text{Prox}_\lambda(\text{rank})(A) = \{0\}$.
- If $\sigma_r(A) > \sqrt{2\lambda}$, then $\text{Prox}_\lambda(\text{rank})(A) = \{A\}$.

- If there is k such that $\sigma_k(A) > \sqrt{2\lambda} > \sigma_{k+1}(A)$, then the set $\text{Prox}_\lambda(\text{rank})(A)$ is a singleton and

$$\text{Prox}_\lambda(\text{rank})(A) = \{U\text{diag}_{m,n}[\sigma_1(A), \ldots, \sigma_k(A), 0, \ldots, 0]V^T\}.$$

- Suppose there is k such that $\sigma_k(A) = \sqrt{2\lambda}$. We define

$$k_0 := \min\{k| \quad \sigma_k(A) = \sqrt{2\lambda}\},$$

$$k_1 := \max\{k| \quad \sigma_k(A) = \sqrt{2\lambda}\}.$$

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

$$\tau_i = \sigma_i(A) \quad \text{if } i < k_0, \quad \tau_i = 0 \quad \text{if } i > k_1,$$

$$\tau_i = 0 \text{ or } \sigma_i(A) \quad \text{if } k_0 \leq i \leq k_1.$$

  where k is an integer between k_0 and k_1.

Comments

1. One could wish to express $(\text{rank})_\lambda(A)$ in terms of traces of matrices as this was done for the smoothed versions of the rank function in Section 3.1. Indeed, $A^TA - 2\lambda I_n$ is a matrix whose eigenvalues are $\sigma_1^2(A) - 2\lambda, \ldots, \sigma_r^2(A) - 2\lambda, -2\lambda, \ldots, -2\lambda$. Its projection on the cone $\mathcal{S}_n^+(\mathbb{R})$ of positive semidefinite matrices has eigenvalues $[\sigma_1^2(A) - 2\lambda]^+, \ldots, [\sigma_r^2(A) - 2\lambda]^+, 0, \ldots, 0$ ([39]). Thus, an alternate expression for $(\text{rank})_\lambda(A)$ is:

$$(\text{rank})_\lambda(A) = \frac{1}{2\lambda}\text{tr}(A^TA) - \frac{1}{2\lambda}\text{tr}[P_{\mathcal{S}_n^+(\mathbb{R})}(A^TA - 2\lambda I_n)] \tag{3.17}$$

2. The Moreau-Yosida approximation of the rank function is only continuous (not smooth as the approximations in the first section of this chapter), but for any matrix $A \in \mathcal{M}_{m,n}(\mathbb{R})$ there exist $\lambda(A)$ such that

$$\forall 0 \leq \lambda \leq \lambda(A) \quad \text{rank}_\lambda(A) = \text{rank } A.$$

   Indeed, if $\sqrt{2\lambda} \leq \sigma_r(A)$, then

$$(\text{rank})_{\lambda(A)}(A) = \text{rank } A.$$

This easily comes from (3.15) since $\sigma_i^2(A) - 2\lambda(A) \geq 0$ for all i and $\|A\|_F^2 = \sum_{i=1}^r \sigma_i^2(A)$. Therefore, the general convergence result that is known for the Moreau-Yosida approximates f_λ of f is made much stronger here.

Proof. Using the same method as in Proposition 3.3, we divide the vector space $\mathcal{M}_{m,n}(\mathbb{R})$ into the sets R_k of matrices rank k.

- If $k = \text{rank } A$, then

$$\min_{B \in R_k} \{\text{rank } B + \frac{1}{2\lambda}\|B - A\|_F^2\} = \text{rank } A.$$

- If $k > \text{rank } A$, then

$$\min_{B \in R_k} \{\text{rank } B + \frac{1}{2\lambda}\|B - A\|_F^2\} \geq k > \text{rank } A.$$

- If $k < \text{rank } A$, then

$$\min_{B \in R_k} \{\text{rank } B + \frac{1}{2\lambda}\|B - A\|_F^2\} = k + \frac{1}{2\lambda} d(A, R_k)^2.$$

From Lemma 1, we can replace $d(A, R_k)$ by $d(A, S_k)$. And then, by the theorem of Eckart-Young and Mirsky, we have

$$\min_{B \in R_k} \{\text{rank } B + \frac{1}{2\lambda}\|B - A\|_F^2\} = k + \frac{1}{2\lambda} \sum_{i=k+1}^r \sigma_i^2(A).$$

Hence, the Moreau-Yosida approximation of the rank function can be represented as

$$\text{rank}_\lambda(A) = \min_{0 \leq k \leq r} \{k + \frac{1}{2\lambda} \sum_{i=k+1}^r \sigma_i^2(A)\}.$$

Because

$$\begin{aligned} k + \frac{1}{2\lambda} \sum_{i=k+1}^r \sigma_i^2(A) &= \frac{1}{2\lambda} \left\{ \sum_{i=1}^r \sigma_i^2(A) - \sum_{i=1}^k \left(\sigma_i^2(A) - 2\lambda\right) \right\} \\ &= \frac{1}{2\lambda} \left\{ \|A\|_F^2 - \sum_{i=1}^k \left(\sigma_i^2(A) - 2\lambda\right) \right\}, \end{aligned}$$

we have

$$\text{rank}_\lambda(A) = \frac{1}{2\lambda}\{\|A\|_F^2 - \max_{0\le k\le r} \sum_{i=1}^{k} \big(\sigma_i^2(A) - 2\lambda\big)\}. \tag{3.18}$$

Among the values of $\left(\sigma_1^2(A) - 2\lambda\right), \sum_{i=1}^{2}\left(\sigma_i^2(A) - 2\lambda\right), \ldots, \sum_{i=1}^{r}\left(\sigma_i^2(A) - 2\lambda\right)$, the largest term is the one that sums all positive $(\sigma_i^2(A) - 2\lambda)$ terms.

We conclude that

$$\text{rank}_\lambda(A) = \frac{1}{2\lambda}\{\|A\|_F^2 - \sum_{i=1}^{r}(\sigma_i^2(A) - 2\lambda)^+\}.$$

A matrix B is an element of $\text{Prox}_\lambda(\text{rank})(A)$ if and only if B is a projection of A over $S_{\tilde{k}}$, where

$$\tilde{k} \in \text{argmin}_{0\le k\le r}\{k + \frac{1}{2\lambda}\sum_{i=k+1}^{r}\sigma_i^2(A)\}.$$

This means that

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

On the other hand, since

$$\tilde{k} + \frac{1}{2\lambda}\sum_{i=\tilde{k}+1}^{r}\sigma_i^2(A) = \frac{1}{2\lambda}\{\|A\|_F^2 - \sum_{i=1}^{\tilde{k}}\big(\sigma_i^2(A) - 2\lambda\big)\},$$

we deduce the following:

- If $\sigma_r(A) > \sqrt{2\lambda}$, then

$$\text{argmin}_{0\le k\le r}\{k + \frac{1}{2\lambda}\sum_{i=k+1}^{r}\sigma_i^2(A)\} = \{r\}.$$

- If $\sigma_1(A) < \sqrt{2\lambda}$, then

$$\text{argmin}_{0\le k\le r}\{k + \frac{1}{2\lambda}\sum_{i=k+1}^{r}\sigma_i^2(A)\} = \{0\}.$$

- If there exists k_0 such that $\sigma_{k_0}(A) > \sqrt{2\lambda} > \sigma_{k_0+1}(A)$, then

$$\text{argmin}_{0\leq k\leq r}\{k + \frac{1}{2\lambda}\sum_{i=k+1}^{r}\sigma_i^2(A)\} = \{k_0\}.$$

- If there exists i such that $\sigma_i(A) = \sqrt{2\lambda}$, then

$$\text{argmin}_{0\leq k\leq r}\{k + \frac{1}{2\lambda}\sum_{i=k+1}^{r}\sigma_i^2(A)\} = \{k_0 - 1, \ldots, k_1\},$$

where

$$k_0 := \min\{k| \quad \sigma_k(A) = \sqrt{2\lambda}\},$$
$$k_1 := \max\{k| \quad \sigma_k(A) = \sqrt{2\lambda}\}.$$

Now, thanks to the Theorem of Eckart-Young, we can express the whole set $\text{Prox}_\lambda(\text{rank})(A)$ as in Theorem 3.6. □

As we saw it when considering the relaxed forms of the rank function (in Chapter 1), what is more useful and interesting for applications is the restricted rank function rank_R. The calculations for its Moreau-Yosida approximates or proximal set-valued mappings are a bit more complicate than for the rank itself, of the same vein however. Here is the final and complete result.

Theorem 3.7. *We have, for all $A \in \mathcal{M}_{m,n}(\mathbb{R})$ of rank $r \geq 1$:*

(i)

$$(\mathrm{rank}_R)_\lambda(A) = \frac{1}{2\lambda}\|A\|_F^2 - \frac{1}{2\lambda}\sum_{i=1}^{r}\{\sigma_i^2(A) - [(\sigma_i(A) - R)^+]^2 - 2\lambda\}^+. \quad (3.19)$$

(ii) One minimizer in (3.13), i.e. one element in $\mathrm{Prox}_\lambda(\mathrm{rank}_R)(A)$*, is provided by* $B := \tilde{U}\Sigma_B\tilde{V}^T$ *with* $\Sigma_B = \mathrm{diag}_{m,n}[\sigma_1(B), \ldots, \sigma_p(B)]$*, where* $(\tilde{U}, \tilde{V}) \in O(m,n)^A$*. Here*

- *If* $\sqrt{2\lambda} \geq R$

$$\sigma_i(B) := \begin{cases} R & \text{if } \sigma_i(A) > \frac{2\lambda+R^2}{2R}, \\ 0 \text{ or } R & \text{if } \sigma_i(A) = \frac{2\lambda+R^2}{2R}, \\ 0 & \text{if } \sigma_i(A) < \frac{2\lambda+R^2}{2R}. \end{cases}$$

- *If* $\sqrt{2\lambda} < R$

$$\sigma_i(B) := \begin{cases} R & \text{if } \sigma_i(A) > R, \\ \sigma_i(A) & \text{if } \sqrt{2\lambda} < \sigma_i(A) \leq R, \\ 0 \text{ or } \sigma_i(A) & \text{if } \sqrt{2\lambda} = \sigma_i(A), \\ 0 & \text{if } \sigma_i(A) < \sqrt{2\lambda}. \end{cases}$$

Comments

1. As the positive parameter λ is supposed to approach 0 in the proximal approximation process, the second case of (ii) in the theorem above is more important than the first one.
2. When $\sqrt{2\lambda} < R$ and $\|A\|_{sp} = \max_{i=1,\ldots,p}\sigma_i(A) \leq R$, both Moreau-Yosida approximates $(\mathrm{rank})_\lambda$ and $(\mathrm{rank}_R)_\lambda$ coincide at A.

Proof. In order to find the minimal value of the function rank $B + \frac{1}{2\lambda}\|A - B\|_F^2$ over the ball $\{\|B\|_{sp} \leq R\}$, we divide the ball into the intersections of it with the sets of matrices with fixed rank.

By the Theorem of Eckart-Young, we know exactly the distance from a matrix A to the set of matrices rank k. And now, we try to find the distance from A to

the intersection of the set of matrices rank k and a ball for the spectral norm, *i.e.*

$$S_{k,R} = \{B \in \mathcal{M}_{m,n}(\mathbb{R}) | \quad \text{rank } B = k, \|B\|_{sp} \leq R\}.$$

For $B \in S_{k,R}$,

$$\|A - B\|_F^2 = \|A\|_F^2 - 2\langle A, B\rangle + \|B\|_F^2.$$

Because $\langle A, B\rangle \leq \sum_{i=1}^{p} \sigma_i(A)\sigma_i(B)$ and rank $B = k$, then

$$\|A - B\|_F^2 \geq \sum_{i=1}^{p} \sigma_i^2(A) - 2\sum_{i=1}^{k} \sigma_i(A)\sigma_i(B) + \sum_{i=1}^{k} \sigma_i^2(B). \tag{3.20}$$

Combining this with the condition $\|B\|_{sp} \leq R$, we obtain

$$\min_{B \in S_{k,R}} \|A - B\|_F^2 = \sum_{i=1}^{k} \{[\sigma_i(A) - R]^+\}^2 + \sum_{i=k+1}^{p} \sigma_i^2(A).$$

Thus,

$$\min_{B \in S_{k,R}} \{\text{rank } B + \frac{1}{2\lambda}\|A - B\|_F^2\} = k + \frac{1}{2\lambda}\sum_{i=1}^{k} \{[\sigma_i(A) - R]^+\}^2 + \frac{1}{2\lambda}\sum_{i=k+1}^{p} \sigma_i^2(A). \tag{3.21}$$

Equality holds in (3.20) if and only if B has a singular value decompostion $B = U_B \Sigma_B V_B^T$ where (U_B, V_B) is an element of $O(m,n)^A$. Thus, B is a projection of A on $S_{k,R}$ if and only if

$$B = \tilde{U}\Sigma_k^R \tilde{V}^T,$$

where $(\tilde{U}, \tilde{V}) \in O(m,n)^A$ and $\Sigma_k^R = \text{diag}_{m,n}(\sigma_1, \ldots, \sigma_p)$ with

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

Of course, in the case where $\sigma_k(A) > \sigma_{k+1}(A)$ the projection is unique.

The right-hand side of equation (3.21) can be expressed as

$$\frac{1}{2\lambda}\|A\|_F^2 - \frac{1}{2\lambda}\sum_{i=1}^{k} \{\sigma_i^2(A) - [(\sigma_i(A) - R)^+]^2 - 2\lambda\}.$$

Hence, the Moreau-Yosida approximation of the restricted rank function is given by

$$(\text{rank}_R)_\lambda(A) = \frac{1}{2\lambda}\left\{\|A\|_F^2 - \max_{0\le k\le r}\left[\sum_{i=1}^{k}\{\sigma_i^2(A) - [(\sigma_i(A)-R)^+]^2 - 2\lambda\}\right]\right\}. \tag{3.22}$$

Consider now the function

$$\begin{array}{rccl} f: & [0;+\infty) & \longrightarrow & \mathbb{R} \\ & x & \mapsto & f(x) = x^2 - [(x-R)^+]^2 - 2\lambda. \end{array}$$

More precisely, if $x \le R$ then $f(x) = x^2 - 2\lambda$ and if $x \ge R$, $f(x) = 2Rx - R^2 - 2\lambda$. Hence, f is continuous and increasing. Moreover,

$$f(x) = 0 \Leftrightarrow \left[\begin{array}{ll} x = \sqrt{2\lambda} & \text{if } \sqrt{2\lambda} < R, \\ x = \frac{R^2+2\lambda}{2R} & \text{otherwise.} \end{array}\right.$$

Then,

- If $\sqrt{2\lambda} < R$
$$f(x) \ge 0 \Leftrightarrow x \ge \sqrt{2\lambda}. \tag{3.23}$$
- If $\sqrt{2\lambda} > R$
$$f(x) \ge 0 \Leftrightarrow x \ge \frac{R^2+2\lambda}{2R}. \tag{3.24}$$

From the properties of f, $\max_{0\le k\le r}\left[\sum_{i=1}^{k}\{\sigma_i^2(A) - [(\sigma_i(A)-R)^+]^2 - 2\lambda\}\right]$ is the one that sums all positive $\{\sigma_i^2(A) - [(\sigma_i(A)-R)^+]^2 - 2\lambda\}$ terms.

To conclude, the Moreau-Yosida approximation of the restricted rank function is given by

$$\begin{array}{c} \min_{\|B\|_{sp}\le R}\{\text{rank } B + \frac{1}{2\lambda}\|A-B\|_F^2\} \\ = \frac{1}{2\lambda}\|A\|_F^2 - \frac{1}{2\lambda}\sum_{i=1}^{r}\{\sigma_i^2(A) - [(\sigma_i(A)-R)^+]^2 - 2\lambda\}^+. \end{array}$$

From (3.23), (3.24) and the projections of A onto $S_{k,R}$, we can determine the whole set $\text{Prox}_\lambda(\text{rank}_R)(A)$ as in Theorem 3.7. □

The pictures below show, in the one dimensional case, the behaviour of c_λ and $(c_R)_\lambda$ as $\lambda \to 0$.

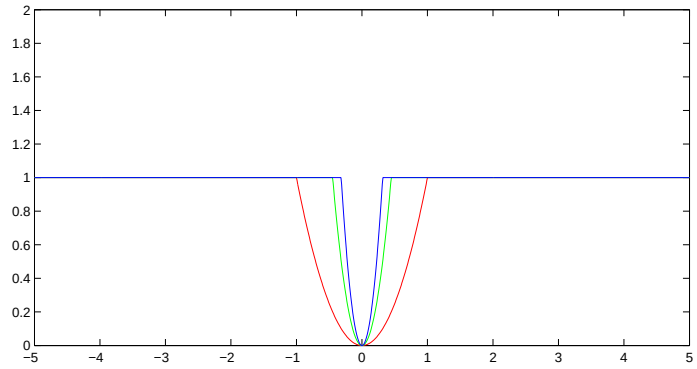


FIGURE 3.3: The Moreau-Yosida approximations of the rank.

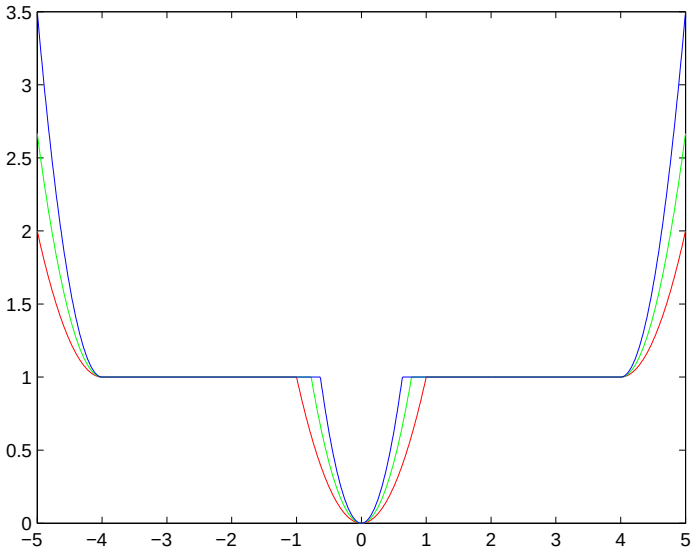


FIGURE 3.4: The Moreau-Yosida approximations of the restricted rank.

We know from Chapter 1 that the convex relaxed form of the restricted rank function rank_R is

$$\text{co}(\text{rank}_R) = \psi_R(A) := \begin{cases} \frac{1}{R}\|A\|_* & \text{if } \|A\|_{sp} \leq R, \\ +\infty & \text{otherwise.} \end{cases}$$

It is interesting to calculate explicitly the Moreau-Yosida approximations $(\psi_R)_\lambda$ of ψ_R, and to compare them with those of rank_R in Theorem 3.7. Here we are in a more familiar convex framework, so that calculations are easier to carry out. Since the proximal set-valued mapping $\text{Prox}_\lambda \psi_R$ is actually single-valued on $\mathcal{M}_{m,n}(R)$,

we adopt the notation

$$\mathrm{Prox}_{\lambda}\psi_R(A) = \{\mathrm{prox}_{\lambda}\psi_R(A)\}.$$

Theorem 3.8. *Let U and V be orthogonal matrices such that $A = U\Sigma_A V^T$, with $\Sigma_A = \mathrm{diag}_{m,n}(\sigma_1(A), \ldots, \sigma_r(A), 0, \ldots, 0)$ (a singular value decomposition of A with $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_r(A) > 0$). We set*

$$p_{\lambda}^{R}(A) = (y_1, \ldots, y_p),$$

with

$$y_i = \begin{cases} R & \text{if } \sigma_i(A) \geq \frac{\lambda}{R} + R, \\ \sigma_i(A) - \frac{\lambda}{R} & \text{if } \frac{\lambda}{R} \leq \sigma_i(A) < \frac{\lambda}{R} + R, \\ 0 & \text{if } \sigma_i(A) < \frac{\lambda}{R}. \end{cases} \tag{3.25}$$

Then, the proximal mapping and the Moreau envelope of ψ_R are described as following.

(i) Proximal mapping:

$$\mathrm{prox}_{\lambda}\psi_R(A) = U\mathrm{diag}_{m,n}(y_1, \ldots, y_p)V^T; \tag{3.26}$$

(ii) Moreau envelope:

We define, for $t \in \mathbb{R}$,

$$f_{\frac{\lambda}{R}}^{i}(t) := t^2 - 2\sigma_i(A)t + 2\frac{\lambda}{R}|t|,$$

and for $x = (x_1, \ldots, x_p)$

$$f_{\frac{\lambda}{R}}(x) := \sum_{i=1}^{p} f_{\frac{\lambda}{R}}^{i}(x_i).$$

Then,

$$(\psi_R)_{\lambda}(A) = \frac{1}{2\lambda}\|A\|_F^2 - \frac{1}{2\lambda} f_{\frac{\lambda}{R}}[p_{\lambda}^{R}(A)]. \tag{3.27}$$

Moreover, for A such that $\|A\|_{sp} \leq R$,

$$(\psi_R)_{\lambda}(A) = \frac{1}{2\lambda}\|A\|_F^2 - \frac{1}{2\lambda}\|p_{\lambda}^{R}(A)\|^2. \tag{3.28}$$

The picture below shows, in the one dimensional case, the behaviour of $(\psi_R)_\lambda$ when $\lambda \to 0$, as well as how it compares with $(\text{rank}_R)_\lambda$. It also illustrates the following fact: *the convex hull (or closed convex hull) of* $(\text{rank}_R)_\lambda$ *is exactly* $(\psi_R)_\lambda$.

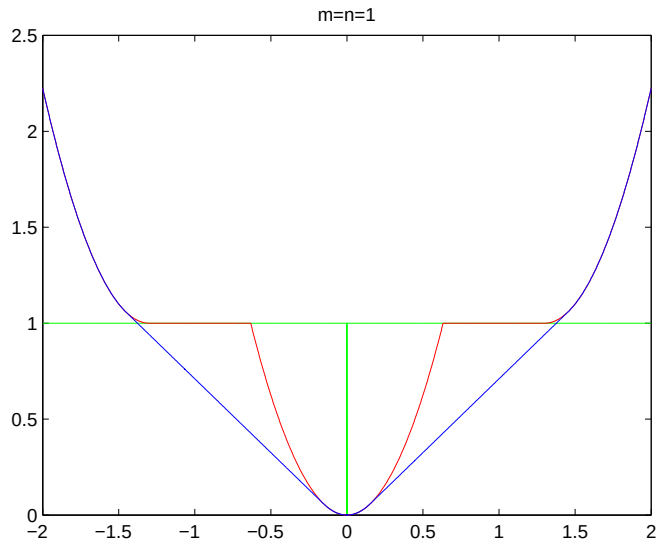


FIGURE 3.5: The Moreau-Yosida approximations of the restricted rank and nuclear norm.

The case where $R = 1$ deserves some additional comments. Recalling that

$$\psi_1(A) = \begin{cases} \|A\|_* & \text{if } \|A\|_{sp} \leq 1, \\ +\infty & \text{otherwise,} \end{cases}$$

we have

$$(\psi_1)_\lambda(A) = \frac{1}{2}\|A\|_F^2 - \frac{1}{2} f_\lambda[p_\lambda^1(A)], \tag{3.29}$$

where $p_\lambda^1(A) = (y_1, \ldots, y_p)$, with

$$y_i = \begin{cases} 1 & \text{if } \sigma_i(A) \geq \lambda + 1, \\ \sigma_i(A) - \lambda & \text{if } \lambda \leq \sigma_i(A) < \lambda + 1, \\ 0 & \text{if } \sigma_i(A) < \lambda. \end{cases} \tag{3.30}$$

In short,

$$y_i = [\sigma_i(A) - \lambda]^+ - [\sigma_i(A) - (\lambda + 1)]^+ \text{ for all } i = 1, \ldots, p,$$

so that

$$\text{prox}_\lambda \psi_1(A) = U \text{diag}_{m,n}\{([\sigma_i(A) - \lambda]^+ - [\sigma_i(A) - (\lambda + 1)]^+)_i\} V^T, \tag{3.31}$$

$$
\begin{aligned}
(\psi_1)_\lambda(A) &= \frac{1}{2\lambda}\|A\|_F^2 - \frac{1}{2\lambda}\sum_{i=1}^{p} f_\lambda^i(y_i) \\
&= \frac{1}{2\lambda}\sum_{i=1}^{r}\sigma_i^2(A) - \frac{1}{2\lambda}\sum_{i=1}^{r} f_\lambda^i([\sigma_i(A)-\lambda]^+ - [\sigma_i(A)-(\lambda+1)]^+) \qquad (3.32)
\end{aligned}
$$

These formulas (3.31) and (3.32) should be put side by side with the expressions of $(\|.\|_*)_\lambda$ and $\mathrm{prox}_\lambda(\|.\|_*)$, such as given in [50] for example:

$$
\mathrm{prox}_\lambda(\|.\|_*)(A) = U\mathrm{diag}_{m,n}[\sigma_i(A)-\lambda]^+V^T, \qquad (3.33)
$$

$$
\begin{aligned}
(\|.\|_*)_\lambda(A) &= \frac{1}{2}\|A\|_F^2 - \frac{1}{2}\sum_{i=1}^{r}\{([\sigma_i(A)-\lambda]^+)_i\}^2. \\
&= \frac{1}{2}\sum_{i=1}^{r}\sigma_i^2(A) - \frac{1}{2}\sum_{i=1}^{r}\{[\sigma_i(A)-\lambda]^+\}^2. \qquad (3.34)
\end{aligned}
$$

As expected, since $\|.\|_* \leq \psi_1$, one has $(\|.\|_*)_\lambda \leq (\psi_1)_\lambda$ for all $\lambda > 0$. Also, for λ small enough, namely for $\lambda \leq \sigma_r(A)$,

$$
(\|.\|_*)_\lambda(A) = (\psi_1)_\lambda(A).
$$

Note however that the convex relaxed form of $(\mathrm{rank})_\lambda$ is not $(\|.\|_*)_\lambda$; as said before, to compare the relaxed form of the rank function with $\|.\|_*$, as well as their corresponding Moreau-Yosida regularized forms, one has to consider their restricted versions on balls $\{A| \quad \|A\|_{sp\leq R}\}$.

The formulas (3.33) and (3.34) are used for designing a proximal point algorithm scheme for nuclear norm minimization ([50]).

3.3 The generalized subdifferentials of the Moreau-Yosida approximation

A.Jourani studied in [42] the limit superior of the Fréchet subdifferentials of the Moreau-Yosida envelopes and he proved that in a Asplund space (*i.e.* a Banach space on which every continuous convex function is Fréchet subdifferentiable on a dense set of points), the Fréchet subdifferential of a function can be obtained from the Fréchet subdifferential of its Moreau envelopes.

Let X be an Asplund space and X^* be the dual space of X equipped with the weak-star topology w^*.

We say that a function f on X is bounded from below by a *negative quadratic form* if and only if

$$\exists c > 0, \exists \overline{x} \in X \text{ such that } f(x) \geq -c(\|x - \overline{x}\|^2 + 1) \quad \text{for all } x \in X.$$

Theorem 3.9 ([42]). *Let f be a lower-semicontinuous real-valued extended function on X. Suppose that f is bounded from below by a negative quadratic form. Then, for all x_0 such that $f(x_0) < \infty$,*

$$\partial^F f(x_0) = seq - \limsup_{\substack{\lambda \to 0^+ \\ u \to x_0 \\ f_\lambda(u) \to f(x_0)}} \partial^F f_\lambda(u),$$

where

$$seq - \limsup_{\substack{\lambda \to 0^+ \\ u \to x_0 \\ f_\lambda(u) \to f(x_0)}} \partial^F f_\lambda(u) = \{x^* \in X^* | \quad \exists \text{ sequence } \lambda_k \to 0^+, u_k \to x_0, f_{\lambda_k}(u_k) \to f(x_0)$$
$$\text{and } u_k^* \to x^* \text{ with } u_k^* \in \partial^F f_{\lambda_k}(u_k) \text{ for all } k = 1, 2, \dots \}.$$

We can use this result and the explicit formula of the Moreau-Yosida approximation of the rank function to retrieve the Fréchet subdifferential of the rank function (as in Theorem 2.9, Chapter 2). Before going into details, we recall here two theorems that provide calculus rules for the Fréchet subdifferential of a function.

Theorem 3.10 (Sum rule, [59]). *Let $f_1, f_2 : \mathbb{R}^p \to \mathbb{R}$ be proper, lower-semicontinuous. If f_1 is Fréchet differentiable at $\tilde{x}$, then*

$$\partial^F (f_1 + f_2)(\tilde{x}) = \nabla f_1(\tilde{x}) + \partial^F f_2(\tilde{x}).$$

Theorem 3.11 (Separable functions, [58]). *Let $f(x) = f(x_1) + \cdots + f(x_q)$ for lower-semicontinuous functions $f_i : \mathbb{R}^{p_i} \longrightarrow \mathbb{R} \cup \{-\infty; +\infty\}$, where $x \in \mathbb{R}^p$ is expressed as $(x_1, \dots, x_q)$ with $x_i \in \mathbb{R}^{p_i}$. Then, at any point $\overline{x} = (\overline{x}_1, \dots, \overline{x}_q)$ where f is finite, one has*

$$\partial^F (\overline{x}) = \partial^F f_1(\overline{x}_1) \times \cdots \times \partial^F f_q(\overline{x}_q).$$

As usual, we begin with calculating the Fréchet subdifferential of the Moreau envelope of the counting function.

Theorem 3.12. *Let x be a vector in $\mathbb{R}^p$ such that*

$$x_1 \geq x_2 \geq \cdots \geq x_p \geq 0.$$

The Fréchet subdifferential of c_λ at x can be expressed as follows:

- *If $x_1 < \sqrt{2\lambda}$, then*

$$\partial^F c_\lambda(x) = \{\nabla c_\lambda(x)\} = \left\{\left(\frac{x_1}{\lambda}, \ldots, \frac{x_p}{\lambda}\right)\right\}.$$

- *If $x_p > \sqrt{2\lambda}$, then*

$$\partial^F c_\lambda(x) = \{\nabla c_\lambda(x)\} = \{(0, \ldots, 0)\}.$$

- *If there exists k such that $x_k > \sqrt{2\lambda} > x_{k+1}$, then*

$$\partial^F c_\lambda(x) = \{\nabla c_\lambda(x)\} = \left\{\left(0, \ldots, 0, \frac{x_{k+1}}{\lambda}, \ldots, \frac{x_p}{\lambda}\right)\right\}.$$

- *If there exists k such that $x_k = \sqrt{2\lambda}$, then*

$$\partial^F c_\lambda(x) = \emptyset.$$

Lemma 3.13. *For $\lambda > 0$, we define h as follows*

$$\begin{array}{rccc} h: & \mathbb{R} & \to & \mathbb{R} \\ & x & \mapsto & -\frac{1}{2\lambda}(x^2 - 2\lambda)^+. \end{array}$$

Then

$$\partial^F h(x) = \begin{cases} \{0\} & \text{if } x^2 < 2\lambda \\ \{-\frac{x}{\lambda}\} & \text{if } x^2 > 2\lambda \\ \emptyset & \text{if } x^2 = 2\lambda \end{cases}.$$

Proof. By definition,

$$h(x) = \begin{cases} 0 & \text{if } x^2 < 2\lambda \\ -\frac{1}{2\lambda}(x^2 - 2\lambda) & \text{if } x^2 \geq 2\lambda \end{cases}.$$

Thus, the function h is differentiable at any $x \notin \left\{-\sqrt{2\lambda}; \sqrt{2\lambda}\right\}$ and

$$h'(x) = 0 \quad \text{if } x^2 < 2\lambda,$$

$$h'(x) = -\frac{x}{\lambda} \quad \text{if } x^2 > 2\lambda.$$

For $x = -\sqrt{2\lambda}$, $h(x) = 0$. Then, $x^* \in \partial^F h(-\sqrt{2\lambda})$ if and only if

$$\liminf_{y \to 0} \frac{h(y - \sqrt{2\lambda}) - x^* y}{|y|} \geq 0.$$

This is equivalent to

$$\liminf_{y \to 0^+} \frac{h(y - \sqrt{2\lambda}) - x^* y}{|y|} \geq 0, \tag{3.35}$$

and

$$\liminf_{y \to 0^-} \frac{h(y - \sqrt{2\lambda}) - x^* y}{|y|} \geq 0. \tag{3.36}$$

When $y \to 0^+$, the value of h at $y - \sqrt{2\lambda}$ is 0. Thus (3.35) becomes

$$x^* \leq 0.$$

When $y \to 0^-$, the value of h at $y - \sqrt{2\lambda}$ is $-\frac{1}{2\lambda}(y^2 - 2\sqrt{2\lambda}y)$. Thus (3.36) becomes

$$x^* \geq \frac{\sqrt{2\lambda}}{\lambda} > 0.$$

This means that $\partial^F h(-\sqrt{2\lambda})$ has no element or

$$\partial^F h(-\sqrt{2\lambda}) = \emptyset.$$

We can also prove that

$$\partial^F h(\sqrt{2\lambda}) = \emptyset.$$

□

Proof. (of Theorem 3.12)

The Moreau-Yosida approximation of the counting function is given by (*cf.* Proposition 3.3):

$$c_\lambda(x) = \frac{1}{2\lambda}\|x\|^2 - \frac{1}{2\lambda}\sum_{i=1}^{p}(x_i^2 - 2\lambda)^+.$$

We can rewrite c_λ as the sum of two functions c^1 and c^2 where $c^1(x) = \frac{1}{2\lambda}\|x\|^2$ and $c^2(x) = -\frac{1}{2\lambda}\sum_{i=1}^p (x_i^2 - 2\lambda)^+$.

It is easy to see that c^1 is a smooth function and $\nabla c^1(x) = \frac{x}{\lambda}$. Because $c^2(x) = -\frac{1}{2\lambda}\sum_{i=1}^p (x_i^2 - 2\lambda)^+ = \sum_{i=1}^p h(x_i)$, the Fréchet subdifferentials of c^2 at x can be presented as the product of the ones of h at x_i (*cf.* Theorem 3.11). By applying Theorem 3.10 for two functions c^1 and c^2, we obtain

$$\partial^F c_\lambda(x) = \nabla c^1 + \partial^F c^2.$$

Thus,

$$\partial^F c_\lambda(x) = \frac{x}{\lambda} + \prod_{i=1}^p \partial^F h(x_i).$$

For $x = (x_1, \ldots, x_p)$ such that $x_1 \geq \cdots \geq x_p \geq 0$ and $\lambda > 0$, we consider two cases:

- If there exists k such that $x_k > \sqrt{2\lambda} > x_{k+1}$, then c_λ is differentiable at x and
$$\partial^F c_\lambda(x) = \{\nabla c_\lambda(x)\} = \{(0, \ldots, 0, \frac{x_{k+1}}{\lambda}, \ldots, \frac{x_p}{\lambda})\}.$$
- If there exists k satisfying
$$x_k = \sqrt{2\lambda},$$
then by using Lemma 3.13, we have
$$\partial^F c_\lambda(x) = \emptyset.$$

□

The Moreau-Yosida approximation of the counting function is absolutely symmetric and continuous. Hence, by using Theorem 2.10 (of Lewis and Sendov) and the fact that (*cf.* Theorem 3.6)

$$\mathrm{rank}_\lambda(A) = c_\lambda \circ \sigma(A),$$

we obtain the following theorem.

Theorem 3.14. *The generalized subdifferentials of* rank_λ *at a matrix* A *is given as below.*

- *If there exists k such that $\sigma_k(A) \geq \sqrt{2\lambda} \geq \sigma_{k+1}(A)$, then* rank_λ *is differentiable at A and*

$$\partial^F \text{rank}_\lambda(A) = \{\nabla \text{rank}_\lambda(x)\} = \{U \text{diag}(0, \ldots, 0, \frac{x_{k+1}}{\lambda}, \ldots, \frac{x_p}{\lambda}) V^T\}.$$

- *If there exist k such that $\sigma_k(A) = \sqrt{2\lambda}$, then*

$$\partial^F \text{rank}_\lambda(A) = \emptyset.$$

Another way to find the Fréchet subdifferential of the rank function

Let $x = (x_1, \ldots, x_p)$ be a vector in $\mathbb{R}^p$ satisfies

$$x_1 \geq \cdots \geq x_p \geq 0.$$

Thanks to Theorem 3.9, we have

$$\partial^F c(x) = seq - \limsup_{\substack{\lambda \to 0^+ \\ u \to x \\ c_\lambda(u) \to c(x)}} \partial^F c_\lambda(u),$$

(see Theorem 3.9 for the definition of $seq - \limsup$).

Let $\{\lambda_k\}_k$ be a sequence that converges to 0 and $\{u^k\}_k$ be a sequence that converges to x.

Let $r = c(x)$. For $\epsilon > 0$ small, there exist K_1 and K_2 such that

$$\forall k \geq K_1 \quad \forall i = 1, \ldots, r, \quad u_i^k \geq x_i - \epsilon,$$

$$\forall k \geq K_2, \quad \sqrt{2\lambda_k} < x_r - \epsilon.$$

Then, if $K_0 = \max(K_1, K_2)$, we have

$$\forall k \geq K_0 \quad \forall i = 1, \ldots, r, \quad u_i^k > \sqrt{2\lambda_k}.$$

By Theorem 3.12, we obtain

$$\forall k \geq K_0 \quad \partial^F c_{\lambda_k}(x_k) \subset \{0\}^r \times \mathbb{R} \times \cdots \times \mathbb{R}.$$

Hence, $\partial^C c(x) \subset \{0\}^r \times \mathbb{R} \times \cdots \times \mathbb{R}$.

On the other hand, any vector in $\mathbb{R}^p$ whose these first r components are 0 belongs to $\partial^F c(x)$.

Indeed, for $a = (0, \ldots, 0, a_{r+1}, \ldots, a_p)$ and $\lambda_k \to 0^+$, we take

$$y_k = (x_1, \ldots, x_r, \lambda_k a_{r+1}, \ldots, \lambda_k a_p) \to x.$$

Because $\lambda_k \to 0$,there exists K_3 such that

$$\forall k \geq K_3 \quad \forall i = r+1, \ldots, p \quad |\lambda_k a_i| < \sqrt{2\lambda_k}.$$

Then, by using Theorem 3.12, for all $k \geq K_3$

$$\partial^F c_{\lambda_k}(y_k) = a.$$

So that, $a \in \partial^F c(x)$. Hence,

$$\partial^F c(x) = \{0\}^r \times \mathbb{R} \times \cdots \times \mathbb{R}.$$

Now, thanks to Theorem 2.12, we can retrieve the subdifferentials of the rank function.

Acknowledgment. We would like to thank Prof. A.Jourani (University of Bourgogne, Dijon) for drawing our attention to this possible way of getting at the Fréchet generalized subdifferential of the rank function (ALEL meeting in Castro-Urdiales, June 2011).

Chapter 4

The cp-rank function revisited

In this chapter, we revisit a notion whose definition resembles that of the rank, the cp-rank function. It is defined for completely positive matrices, a specific class of positive matrices. We recall here the definition and some properties of the cp-rank function. And then, we provide its convex relaxed form and list some open questions concerning it.

4.1 Definition and Properties

4.1.1 Definition

Let $\mathcal{S}_n(\mathbb{R})$ be the set of real square symmetric matrices of dimension $n \times n$. Recall that a matrix A is positive semidefinite if it can be decomposed as $A = BB^T$ where B is a real matrix. The rank of a positive definite matrix A can be defined as the smallest number of columns of B in such a factorization.

Definition 4.1. A real square symmetric (elementwise) nonnegative matrix A in $\mathcal{S}_n(\mathbb{R})$ is *completely positive* (CP) if it can be factorized as $A = BB^T$ where B is a real *nonnegative* matrix. The smallest number of columns of B in such a factorization is then called *the cp-rank of* A and is denoted by cp-rank A.

If A is a square symmetric matrix which is not CP, we say by convention that A has a cp-rank equal to $+\infty$, written cp-rank $A = +\infty$. Also by convention, cp-rank of the zero matrix is zero.

Remark 4.2. If $A = BB^T$, then A can be represented as the sum of the matrices $b_i b_i^T$, where the b_i's are the columns of B. Hence, the cp-rank of A is also the minimal number of summands in a rank 1-presentation of A, $A = \sum_{i=1}^{k} b_i b_i^T$, with $b_i \geq 0$ for all i (a vector of $\mathbb{R}^n$ with nonnegative components).

The set of all completely positive matrices is a *closed convex cone* in $\mathcal{S}_n(\mathbb{R})$. We denote it by $CP_n(\mathbb{R})$. Moreover,

$$CP_n(\mathbb{R}) = \text{conv}\{xx^T : x \in \mathbb{R}^n_+\}.$$

The positive polar (or dual) cone $CP_n^*(\mathbb{R})$ of $CP_n(\mathbb{R})$ is defined by

$$CP_n^*(\mathbb{R}) := \{S \text{ a symmetric } n \times n \text{ matrix } : \langle S, X\rangle \geq 0 \text{ for all } X \in CP_n(\mathbb{R})\}.$$

It can be proved that $CP_n^*(\mathbb{R})$ coincides with the cone of *copositive matrices*, namely

$$C_n(\mathbb{R}) = \{S \text{ a symmetric } n \times n \text{ matrix } : x^T S x \geq 0 \text{ for all } x \in \mathbb{R}^n_+\}.$$

More information about the cone of completely positive matrices and copositive matrices can be found in several references, for example in [40].

4.1.2 Properties

Proposition 4.3. *If A is an $n \times n$ completely positive matrix, then*

$$\text{cp-rank } A \geq \text{rank } A. \tag{4.1}$$

In some cases, the cp-rank of A is equal to the rank of A, for example: when the rank of A is less than or equal to 2 or when $n \leq 3$, *etc.* But in general, the inequality in (4.1) is strict.

Example 4.1.

$$A = \begin{pmatrix} 6 & 3 & 3 & 0 \\ 3 & 5 & 1 & 3 \\ 3 & 1 & 5 & 3 \\ 0 & 3 & 3 & 6 \end{pmatrix}$$

Here, rank $A = 3$ *while* cp-rank $A = 4$.

The cp-rank however enjoys some properties similar to those of the rank (see Chapter 1).

Proposition 4.4. *If A and B are $n \times n$ completely positive matrices, then*

(i) cp-rank $(A + B) \leq$ cp-rank $A +$ cp-rank B.

(ii) cp-rank $(kA) =$ cp-rank A for every positive real number k.

Proposition 4.5. *Suppose that $\{A_m\}_m$ is a sequence of completely positive matrices in $\mathcal{S}_n(\mathbb{R})$, and that*

$$A = \lim_{m\to\infty} A_m.$$

Then

$$\text{cp-rank } A \leq \liminf_{m\to\infty} \text{cp-rank } A_m.$$

This means that the cp-rank function is lower-semicontinuous.

Proof. Suppose that

$$k = \liminf_{m\to\infty} \text{cp-rank } A_m.$$

We can extract from $\{A_m\}_m$ a subsequence where each A_m has a cp-rank equal to k. Indeed, according to the definition of k, there exists a subsequence $\{A_{m_q}\}_q$ of $\{A_m\}_m$ such that $k = \lim_{q\to\infty} \text{cp-rank } A_{m_q}$. Hence, for Q large enough and $q \geq Q$,

$$k - \frac{1}{2} < \text{cp-rank } A_{m_q} < k + \frac{1}{2}.$$

From the fact that cp-rank only takes integer values, we can infer from above that cp-rank $A_{m_q} = k$ for every $q \geq Q$.

Hence, for $q \geq Q$, $A_{m_q} = (a^{i,j}_{m_q})_{i,j=1,\ldots,n}$ can be factorized as $A_{m_q} = B_{m_q} B^T_{m_q}$, where B_{m_q} is a real nonnegative matrix with dimension $k \times n$. Let $b^1_{m_q}, \ldots, b^k_{m_q}$ denote the columns of B_{m_q}.

We already know that $A = \lim_{q\to\infty} A_{m_q}$, where $A = (a^{ij})_{i,j=1,\ldots,n}$. Then,

$$a^{ii} = \lim_{q\to\infty} a^{ii}_{m_q}.$$

Moreover, $a^{ii}_{m_q} = \|b^i_{m_q}\|^2$. Thus, $\lim_{q\to\infty} \|b^i_{m_q}\|^2 = a^{ii}$. This means that the sequence of vectors $\{b^i_{m_q}\}_q$ is bounded for all i. There then exists a subsequence of $\{b^i_{m_q}\}$ converges to a vector b^i of $\mathbb{R}^n$.

Now, let B be the matrix with dimension $k \times n$, defined by $B = (b^1, \ldots, b^k)$. It is easy to see that $A = BB^T$.

So, by the definition of cp-rank itself, we conclude cp-rank $A \leq k$.

□

One of the most interesting questions concerning the cp-rank is to find an upper bound for the cp-rank of completely positive matrices of a given rank r. In 1983, Hannay and Laffey showed that the maximal cp-rank of a CP matrix of rank r is less than or equal to $r(r+1)/2$ ([27]). Then, this upper bound was improved by Barioli and Berman in [4]: they proved that the maximal cp-rank of a CP matrix of rank r is equal to $r(r+1)/2 - 1$ for $r \geq 2$.

Theorem 4.6 ([4]). *For every rank r completely positive matrix A, $r \geq 2$*

$$\text{cp-rank } A \leq \frac{r(r+1)}{2} - 1.$$

Theorem 4.7 ([4]). *For every $r \geq 2$ there exists a completely positive matrix A with* $\text{rank}A = r$ *and* $\text{cp-rank}A = r(r+1)/2 - 1$.

4.2 The convex relaxed form of the cp-rank

In this section, we compute the (convex) relaxed form of the cp-rank function. From Proposition 4.4 (ii), it is easy to see that the convex hull of the cp-rank function on the whole space $\mathcal{S}_n(\mathbb{R})$ is the zero function. So, like for the rank function, we restrict it to some appropriate ball. Let us consider:

$$A \in \mathcal{S}_n(\mathbb{R}) \mapsto \psi(A) := \begin{cases} \text{cp-rank } A & \text{if } A \text{ is CP and } \|A\|_* \leq 1. \\ +\infty & \text{otherwise.} \end{cases}$$

Theorem 4.8. *The convex hull (or closed convex hull) of ψ is*

$$A \in \mathcal{S}_n(\mathbb{R}) \mapsto \hat{\psi} = \begin{cases} \|A\|_* & \text{if } A \text{ is CP and } \|A\|_* \leq 1. \\ +\infty & \text{otherwise.} \end{cases}$$

Proof. The domain of the function $\hat{\psi}$, *i.e.* the set of matrices where it is finite-valued, and that of ψ are equal: it is the compact convex set

$$CP_n(\mathbb{R} \cap \{A \in \mathcal{S}_n(\mathbb{R}) | \quad \|A\|_* \leq 1\}.$$

So, if $A \in \mathcal{S}_n(\mathbb{R})$ lies out of the above set, the function ψ and $\hat{\psi}$ coincide at A, their common value is $+\infty$.

Now, let A be chosen completely positive, with $\|A\|_* \leq 1$. Firstly, if A is the zero matrix, it is clear that

$$\mathrm{co}(\psi)(0) = 0 = \hat{\psi}(0).$$

Secondly, let us assume that $A \neq 0$. We have to prove that

$$\mathrm{co}(\psi)(A) = \|A\|_*.$$

Because

$$\text{cp-rank} \geq \text{rank} \geq \|.\|_* \text{ on } \{A \in \mathcal{S}_n(\mathbb{R}) | \quad \|A\|_* \leq 1\},$$

and the function $\hat{\psi}$ is closed and convex, we get the first inequality

$$\mathrm{co}(\psi) \geq \hat{\psi}. \tag{4.2}$$

Now, because A is CP, we can decompose A as $A = \sum_{i=1}^{k} b_i b_i^T$, where $b_i \neq 0$ and $b_i \in \mathbb{R}_+^n$. By setting $c_i = \frac{b_i}{\|b_i\|}$ and $\alpha_i = \|b_i\|^2$, we obtain

$$A = \sum_{i=1}^{k} \alpha_i c_i c_i^T. \tag{4.3}$$

Then

$$\text{tr } A = \sum_{i=1}^{k} \alpha_i \mathrm{tr}(c_i c_i^T) = \sum_{i=1}^{k} \alpha_i \quad (\text{because } \mathrm{tr}(c_i c_i^T) = 1).$$

Hence,

$$0 \leq \sum_{i=1}^{k} \alpha_i \leq 1.$$

We complete the decomposition (4.3) with the zero matrix

$$A = \sum_{i=1}^{k} \alpha_i c_i c_i^T + (1 - \sum_{i=1}^{k} \alpha_i)0.$$

By using the convexity of $\mathrm{co}(\psi)$ and the fact that $\psi(0) = 0$, we have

$$\mathrm{co}(\psi)(A) \leq \sum_{i=1}^{k} \alpha_i \mathrm{co}(\psi)(c_i c_i^T) = \sum_{i=1}^{k} \alpha_i.$$

This means that

$$\mathrm{co}(\psi)(A) \leq \|A\|_*. \tag{4.4}$$

From (4.2) and (4.4), we deduce

$$\mathrm{co}(\psi) = \hat{\psi}.$$

□

Therefore, the rank and the cp-rank are two functions that share several common properties: they are lower semi-continuous, subadditive, they take only integer values. Moreover, on the set $\{A | \|A\|_* \leq 1\}$, they also have the same relaxed form, namely the nuclear norm.

4.3 Open questions

4.3.1 The DJL conjecture

By Theorem 4.6, if A is a $n \times n$ completely positive matrix, then

$$\text{cp-rank } A \leq \frac{n(n-1)}{2} - 1. \tag{4.5}$$

But is there any better upper bound on the cp-rank of $n \times n$ matrices? Drew, Johnson and Loewy proved that cp-rank $A \leq \frac{n^2}{4}$ for every completely positive matrix of order $n \geq 4$ whose graph is triangle free ([18]). The fact that the bound $\frac{n^2}{4}$ was also valid for all other known cases led the authors to wonder whether this holds for every completely positive matrix of order $n \geq 4$.

Conjecture (The DJL Conjecture) If A is an $n \times n$ completely positive matrix, with $n \geq 4$, then

$$\text{cp-rank } A \leq \frac{n^2}{4}. \tag{4.6}$$

The conjecture was proved by Berman and Shaked-Monderer ([7]) for matrices whose comparison matrices are M-matrices, and by Loewy and Tam ([51]) for 5×5 matrices whose graph is not complete. But then, for the first time, Barioli announced (in [3]) an example of 7×7 completely positive matrix of rank 5 and cp-rank 14. Such a matrix is a counter-example to the DJL Conjecture.

4.3.2 The generalized subdifferentials

We computed explicitly the generalized subdifferentials of the rank function as in chapter 2. The same question could be posed for the cp-rank function:to calculate explicitly the generalized subdifferentials of the cp-rank function. We have been unable to provide an answer to such a question. The main reason is that, contrary to the rank function (which is the number of nonzero singular values), the cp-rank of A cannot be deduced from the singular values of A.

Acknowledgment. We would like to thank Prof. I.Bomze (University of Vienna) for drawing our attention to the similarities between the rank function and the cp-rank function (AFG'11 meeting in Toulouse, September 2011).

Bibliography

[1] H. Attouch, J. Bolte, and B.F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[2] D. Azé and J.-B. Hiriart-Urruty. *Analyse variationnelle et optimisation: Eléments de cours, exercices et corrigés*. Editions Cépaduès, 2010.

[3] F. Barioli. Completely positive matrices of small and large acute sets of vectors, 2002. Talk at the 10th ILAS conference in Auburn.

[4] F. Barioli and A. Berman. The maximal cp-rank of rank k completely positive matrices. *Linear Algebra and its Applications*, 363(0):17–33, 2003.

[5] M.S. Bazaraa, J.J. Goode, and M.Z. Nashed. On the cones of tangents with applications to mathematical programming. *Journal of Optimization Theory and Applications*, 13:389–426, 1974.

[6] J. Benoist and J.-B Hiriart-Urruty. What is the subdifferential of the closed convex hull of a function? *SIAM J.Math Anal.*, 27(6):1661–1679, 1996.

[7] A. Berman and N. Shaked-Monderer. Remarks on completey positive matrices. *Linear and Multilinear Algebra*, 44(2):149–163, 1998.

[8] A. Berman and N. Shaked-Monderer. *Completely positive matrices*. World Scientific, 2003.

[9] M. Borwein and R. Luke. Entropic regularization of the l_0 function. *Springer Optimization and Its Applications*, Fixed-point algorithms for inverse problems in science and engineering:65–92, 2011.

[10] E.J. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203 – 4215, December 2005.

[11] E.J. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080, 2010.

[12] F.H. Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[13] M.G. Crandall and P.-L Lions. Viscosity solutions of Hamilton-Jacobi equations. *Transactions of the American Mathematical Society*, 277(1):1–42, 1983.

[14] J.-P. Crouzeix. *Contributions à l'étude des fonctions quasiconvexes*. Thèse de Doctorat ès Sciences, Université de Clermont-Ferrand II, 1977.

[15] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289 –1306, 2006.

[16] D.L Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[17] D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6 – 18, 2006.

[18] J.-H. Drew, C.R. Johnson, and R. Loewy. Completely positive matrices associated with m-matrices. *Linear and Multilinear Algebra*, 37(4):303–310, 1994.

[19] D. Drusvyatskiy, A.D. Ioffe, and A.S. Lewis. The dimension of semialgebraic subdifferential graphs. *Nonlinear Analysis.*, 75(3):1231–1245, 2012.

[20] D. Drusvyatskiy and A.S. Lewis. Semi-algebraic functions have small subdifferentials. *To appear in special issue of Mathematical Programming Ser.B in honor of C.Lemaréchal*, 2013.

[21] M. Dür. Copositive programming - a survey. In *Recent Advances in Optimization and its Applications in Engineering*, pages 3–20. Springer Berlin Heidelberg, 2010.

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

[23] G. Fung and O. Mangasarian. Equivalence of minimal l^0 and l^p norm solutions of linear equalities, inequalities and linear programs for sufficiently small p. *Journal of Optimization Theory and Applications*, 151(1):1–10, 2011.

[24] K.A. Gallivan and P.-A. Absil. Note on the convex hull of the Stiefel manifold. *Working note*, 2010.

[25] D. Ge, X. Jiang, and Y. Ye. A note on the complexity of l^p minimization. *Mathematical Programming*, 129(2):285–299, 2011.

[26] G.H. Golub and C.F. Van Loan. *Matrix Computations. Third edition.* Johns Hopkins University Press, Baltimore, MD, 1996.

[27] J. Hannah and T.J. Laffey. Nonnegative factorization of completely positive matrices. *Linear Algebra and its Applications*, 55(0):1–9, December 1983.

[28] N.J. Higham. Matrix nearness problems and applications. *Applications of Matrix Theory*, pages 1–27, 1989.

[29] J.-B. Hiriart-Urruty. Deux questions de rang. *Working Note, Université Paul Sabatier*, 2009.

[30] J.-B. Hiriart-Urruty. *Optimisation et analyse convexe: exercices et problèmes corrigés, avec rappels de cours.* EDP Sciences, Paris, 2009.

[31] J.-B. Hiriart-Urruty. *Bases, outils et principes pour l'analyse variationnelle.* Springer, 2012.

[32] J.-B. Hiriart-Urruty. When only global optimization matters. *Journal of Global Optimization, DOI 10.1007/s10898-011-9826-7*, 2012.

[33] J.-B. Hiriart-Urruty and H.Y. Le. Convexifying the set of matrices of bounded rank; Applications to the quasiconvexification and convexification of the rank function. *Optimization Letters*, 6(5):841–849, 2012.

[34] J.-B. Hiriart-Urruty and H.Y. Le. A variational approach of the rank function. *TOP (Journal of the Spanish Society of Statistics and Operations Research)*, DOI: 10.1007/s11750-013-0283-y (2013).

[35] J.-B. Hiriart-Urruty and H.Y. Le. From Eckart-Young approximation to Moreau envelopes and vice versa. *Preprint*, Submitted 2012.

[36] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis.* Springer, 2001.

[37] J.-B. Hiriart-Urruty and Lemaréchal, C. *Convex analysis and minimization algorithms I.Fundamentals*, volume 305. Springer-Verlag, 1993.

[38] J.-B. Hiriart-Urruty and Lemaréchal, C. *Convex analysis and minimization algorithms II.Advanced theory and bundle methods*, volume 306. Springer-Verlag, 1993.

[39] J.-B. Hiriart-Urruty and J. Malick. A fresh Variational-Analysis look at the positive semidefinite matrices world. *Journal of Optimization Theory and Applications*, 153(3):551–557, 2012.

[40] J.-B. Hiriart-Urruty and A. Seeger. A variational approach to copositive matrices. *SIAM Review*, 52(4):593, 2010.

[41] R.A. Horn and C.R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1994.

[42] A. Jourani. Limit superior of subdifferentials of uniformly convergent functions. *Positivity*, 3(1):33–47, 1999.

[43] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.*, 11:517–553, 2010.

[44] H.Y. Le. Convexifying the counting function on $\mathbb{R}^p$ for convexifying the rank function on $\mathcal{M}_{m,n}(\mathbb{R})$. *Journal of Convex Analysis*, 19(2), 2012.

[45] H.Y. Le. Generalized subdifferentials of the rank function. *Optimization Letters, DOI 10.1007/s11590-012-0456-x*, 2012.

[46] A.S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1):173–183, 1995.

[47] A.S. Lewis. Convex analysis on the hermitian matrices. *SIAM Journal on Optimization*, 6:164–177, 1996.

[48] A.S. Lewis and H.S. Sendov. Nonsmooth analysis of singular values. Part I: theory. *Set-Valued Analysis*, 13(3):213–241, 2005.

[49] A.S Lewis and H.S Sendov. Nonsmooth analysis of singular values. Part II: applications. *SetValued Analysis*, 13(3):243–264, 2005.

[50] Y.-J. Liu, D. Sun, and K.-C. Toh. An implementable proximal point algorithmic framework for nuclear norm minimization. *Mathematical Programming*, 133(1-2):399–436, January 2011.

[51] R. Loewy and B.-S. Tam. CP rank of completely positive matrices of order 5. *Linear algebra and its applications*, 363:161–176, 2003.

[52] D. R Luke. Prox-regularity of rank constraint sets and implications for algorithms. *Journal of Mathematical Imaging and Vision, DOI 10.1007/s10851-012-0406-3*, 2012.

[53] I. Markovsky. *Low Rank Approximation*. Springer, 2012.

[54] B.S. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*. Springer, 2010.

[55] P.A. Parrilo. The convex algebraic geometry of rank minimization, 2009. Plenary talk at the International Symposium on Mathematical Programming, Chicago.

[56] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed Minimum-Rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471, 2010.

[57] R. T. Rockafellar. Proximal subgradients, marginal values, and augmented lagrangians in nonconvex optimization. *Mathematics of Operations Research*, 6(3):424–436, 1981.

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

[59] W. Schirotzek. *Nonsmooth analysis*. Springer, 2007.

[60] G.W. Stewart. *Matrix Algorithms, Volume 1:Basic Decompositions*. SIAM, 1998.

[61] Y-B. Zhao. Approximation theory of matrix rank minimization and its application to quadratic equations. *Linear Algebra and Its Applications*, pages 77–93, October 2012.

Abstract

In this dissertation, we consider the rank function *from the variational point of view*. The reason why we are interested in this function is that it appears as an objective (or constraint) function in various modern optimization problems, such as: low rank matrix completion, multivariate statistical data analysis, compressed sensing, etc. In some particular cases, the rank minimization problems can be solved by using the singular value decomposition of matrices or can be reduced to the solution of linear systems. But in general, the rank minimization problems is known to be NP-hard.

We provide here several properties of the rank function from the variational point of view: additional proofs for its closed convex relaxation, the expressions of its generalized subdifferentials and the explicit expression of its Moreau regularization-approximation form. Then, in the last chapter, we revisit a notion whose definition resembles that of the rank, the cp-rank function.

Keywords: the rank function; convex relaxation; generalized subdifferential; Moreau regularization-approximation; the cp-rank function.

Résumé

Dans ce mémoire de thèse, nous étudions la fonction rang *du point de vue variationnel*. La raison pour laquelle nous nous intéressons à cette fonction est qu'elle apparaît comme une fonction objectif (ou comme fonction contrainte) dans divers problèmes d'optimisation moderne, par exemple: complétion de matrices, analyse de données statistiques, acquisition parcimonieuse de données, etc. Dans certains cas particuliers, les problèmes de minimisation de la fonction rang peuvent être résolus en utilisant la décomposition en valeurs singulières. Mais, en géneral, les problèmes de minimisation de la fonction rang sont "NP-difficiles".

Nous proposons ici quelques propriétés de la fonction rang du point de vue variationnel: des démonstrations supplémentaires pour son enveloppe convexe fermée (restreinte à des boules spectrales), les expressions des sous-différentiels généralisés et la régularisation-approximation au sens de Moreau. Puis, dans le dernier chapitre, nous revenons sur une notion dont la définition ressemble à celle de la fonction rang, la fonction cp-rang.

Mots-clés: la fonction rang: relaxation convexe; sous-différentiel généralisé; régularisation-approximation de Moreau; la fonction cp-rang.