

Nonsmooth Analysis of Singular Values. Part II: Applications [★]

ADRIAN S. LEWIS¹ and HRISTO S. SENDOV²

¹*Department of Combinatorics & Optimization, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. e-mail: aslewis@math.uwaterloo.ca*

²*Department of Mathematics, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada. e-mail: hssendov@cecm.sfu.ca*

(Received: 4 April 2003; in final form: 29 October 2004)

Abstract. In this work we continue the nonsmooth analysis of absolutely symmetric functions of the singular values of a real rectangular matrix. Absolutely symmetric functions are invariant under permutations and sign changes of its arguments. We extend previous work on subgradients to analogous formulae for the proximal subdifferential and Clarke subdifferential when the function is either locally Lipschitz or just lower semicontinuous. We illustrate the results by calculating the various subdifferentials of individual singular values. Another application gives a nonsmooth proof of Lidskii's theorem for weak majorization.

Mathematics Subject Classifications (2000): Primary 90C31, 15A18; secondary 49K40, 26B05.

Key words: nonsmooth analysis, singular values, regular subdifferential, limiting subdifferential, proximal subdifferential, Clarke subdifferential, lower semicontinuous, Lidskii.

1. Introduction

This paper is a continuation of our work in [10], where we began a systematic study of the nonsmooth properties of functions of the singular values of a rectangular matrix. There we gave simple formulae for the regular subdifferential, the limiting subdifferential, and the horizon subdifferential, of such functions and illustrated the results with several applications.

To make the development as self contained as possible, in the next section we have stated all results from the first paper that are needed in the proofs here.

The development that follows has four main parts. We begin by discussing absolutely symmetric functions of singular values that are locally Lipschitz and show that the main formula from [10] is preserved for the Clarke subdifferential as well. Next we relax that assumption and require the functions involved to be only lower semicontinuous. The independent development for the Lipschitz case is interesting in its own right: It deepens the analogies with the work of Lewis in [9], as well as extending and generalizing the convexity results there. We need

[★] Research supported by NSERC.

some of these convexity results later in the third part, where we are interested in the individual singular values of a real rectangular matrix. The last part deals with another application of our theory. We derive, through elementary nonsmooth analysis, a famous theorem in matrix perturbation analysis: Lidskii's theorem for weak majorization between the vectors of singular values of perturbed rectangular matrices. The results described here were first investigated in the second author's dissertation [15].

2. Definitions and Preliminary Results

Given a function $f: \mathbb{R}^n \rightarrow [-\infty, +\infty]$ we say vector $y \in \mathbb{R}^n$ is a *regular subgradient* of f at x if $f(x) < \infty$ and

$$f(x+z) \geq f(x) + \langle y, z \rangle + o(z) \quad \text{as } z \rightarrow 0.$$

The set of all regular subgradients at x is denoted by $\hat{\partial}f(x)$ and called the *regular subdifferential*.

A vector $y \in \mathbb{R}^n$ is a (*limiting*) *subdifferential* of f at x if $f(x) < \infty$ and there is a sequence of points x^r in E approaching x with values $f(x^r)$ approaching the finite value $f(x)$, and a sequence of regular subgradients y^r in $\hat{\partial}f(x^r)$ approaching y . The set of all limiting subgradients is denoted $\partial f(x)$. In case when $f(x) = \infty$ we set $\hat{\partial}f(x) = \partial f(x) = \emptyset$. The reader can verify that $\partial f(x)$ and $\hat{\partial}f(x)$ are always closed sets and that $\hat{\partial}f(x)$ is convex.

If the function f is locally Lipschitz around x , convex combinations of subgradients are called *Clarke subgradients*. The set of Clarke subgradients is the *Clarke subdifferential* $\partial^c f(x)$. (This definition is equivalent to the standard one in [2] – see for example Theorem 2 in [5].)

Henceforth we will assume that n and m are natural numbers and $n \leq m$. Let $M_{n,m}$ denote the Euclidean space of $n \times m$ real matrices, with inner product $\langle X, Y \rangle = \text{tr } X^T Y$. Simpler, M_n will denote $M_{n,n}$. By $O(n)$ we will denote the group of $n \times n$ orthogonal matrices, and the product $O(n) \times O(m)$ will be denoted by $O(n, m)$. One of the main objects of this paper is the class of *singular value functions*. These are functions $F: M_{n,m} \rightarrow [-\infty, +\infty]$ with the invariance property

$$F(U_n^T X U_m) = F(X) \quad \text{for all } (U_n, U_m) \in O(n, m) \text{ and } X \in M_{n,m}.$$

When (U_n, U_m) varies freely over $O(n, m)$, in the product $U_n^T X U_m$ only the singular values are invariant. Thus it is not surprising that F can be expressed as the composition $F(X) = (f \circ \sigma)(X)$, where $\sigma(X)$ are the singular values of X and $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is such that

$$f(x_1, x_2, \dots, x_n) = f(|x_{\pi(1)}|, |x_{\pi(2)}|, \dots, |x_{\pi(n)}|),$$

for any permutation π . We will call such functions *absolutely symmetric*. In this way the singular value functions are in one-to-one correspondence with the absolutely symmetric functions. Throughout we will assume without loss of general-

ity that the singular values of X are ordered nonincreasingly, that is,

$$\sigma_1(X) \geq \sigma_2(X) \geq \dots \geq \sigma_n(X).$$

We would like to note that analogous results to those we present in this work hold also for the space of $n \times m$ complex matrices with the inner product $\langle X, Y \rangle = \text{Re}(\text{tr } X^*Y)$, where X^* denotes transposition and complex conjugation. With this inner product the complex matrices turn into an Euclidean space over the *reals*. Orthogonal matrices below become unitary, but the functions with matrix argument are still (extended) real valued.

We will use the following notation throughout:

- $\mathbb{R}_\downarrow^n = \{x \in \mathbb{R}^n \mid x_1 \geq x_2 \geq \dots \geq x_n\}$.
- $\mathbb{R}_+^n = \{x \in \mathbb{R}^n \mid x_i \geq 0, i = 1, \dots, n\}$.
- $\mathbb{R}_\downarrow^+ = \mathbb{R}_\downarrow^n \cap \mathbb{R}_+^n$.
- $|x| = (|x_1|, |x_2|, \dots, |x_n|)$, for $x \in \mathbb{R}^n$.
- \bar{x} denotes the vector with the same entries as $x \in \mathbb{R}^n$ ordered in nonincreasing order, that is, $\bar{x}_1 \geq \bar{x}_2 \geq \dots \geq \bar{x}_n$.
- $\hat{x} = |\bar{x}|$.
- $P(n)$ the set of all $n \times n$ permutation matrices.
- $P_{(-)}(n)$ the set of all $n \times n$ matrices that have only one nonzero entry in every row and column, which is ± 1 (we will call them *signed permutation matrices*).
- E will stand for any finite dimensional Euclidean space and $O(E)$ will denote the group of its orthogonal transformations.
- For $(U_n, U_m) \in O(n, m)$ and $X \in M_{n,m}$ we denote $(U_n, U_m).X = U_n^T X U_m$, the action of (U_n, U_m) on X .
- For $x \in \mathbb{R}^n$, $\text{Diag } x \in M_{n,m} \cup M_n$ will denote the matrix with vector x on its main diagonal and zeros elsewhere. The dimensions on $\text{Diag } x$ will be clear from the context. For $X \in M_{n,m} \cup M_n$, by $\text{diag } X$ we will denote the vector in \mathbb{R}^n of diagonal entries of X .

Finally we will need the following preliminary results.

SUBGRADIENT INVARIANCE THEOREM. *If $f: E \rightarrow [-\infty, +\infty]$ is invariant under a subgroup G of $O(E)$, then any point x in E and transformation g in G satisfy $\partial f(gx) = g\partial f(x)$. Corresponding result holds for the regular subdifferential.*

SYMMETRICITY THEOREM. *If $Y \in M_{n,m}$ is a regular or a limiting subgradient of a singular value function F at $X \in M_{n,m}$, then $X^T Y$ and $Y^T X$ are symmetric matrices.*

(See the theorem with the same name in [10].)

ORDER INEQUALITY. *For any $x, y \in \mathbb{R}^n$ we have $x^T y \leq \bar{x}^T \bar{y}$ with equality iff $\exists Q \in P(n)$ such that $Qx = \bar{x}$ and $Qy = \bar{y}$.*

(See, for example, [6].)

ABSOLUTE ORDER INEQUALITY. For any $x, y \in \mathbb{R}^n$ we have $x^T y \leq \hat{x}^T \hat{y}$ with equality iff $\exists P_{(-)} \in P_{(-)}(n)$ such that $P_{(-)}x = \hat{x}$ and $P_{(-)}y = \hat{y}$.

(For a direct proof see [10], or [7] for generalizations.)

SIMULTANEOUS RECTANGULAR CONJUGACY THEOREM. For any vectors $x, y, u,$ and v in \mathbb{R}^n , there is an element (U_n, U_m) in $O(n, m)$ such that $\text{Diag } x = U_n^T (\text{Diag } u) U_m$ and $\text{Diag } y = U_n^T (\text{Diag } v) U_m$ iff there is a matrix $P_{(-)}$ in $P_{(-)}(n)$ with $x = P_{(-)}u$ and $y = P_{(-)}v$.

(See the proposition with the same name in [10].)

SINGULAR VALUES DERIVATIVE THEOREM. Any x in \mathbb{R}_+^n and $M \in M_{n,m}$ satisfy $\text{diag } M \in \text{conv}(P_{(-)}(n)_x \sigma'(\text{Diag } x; M))$.

(See the proposition with the same name in [10].)

VON NEUMANN'S TRACE THEOREM. Any $X, Y \in M_{n,m}$ satisfy the inequality $\text{tr } X^T Y \leq \sigma(X)^T \sigma(Y)$. Equality holds iff there is $(U_n, U_m) \in O(n, m)$ such that $X = U_n^T (\text{Diag } \sigma(X)) U_m$ and $Y = U_n^T (\text{Diag } \sigma(Y)) U_m$.

(See [10] or the original proof in [18].)

We are also going to need the main result from [10]:

THEOREM 2.1 (Subgradients). The (limiting) subdifferential of a singular value function $f \circ \sigma$ at $X \in M_{n,m}$ is given by the formula

$$\partial(f \circ \sigma)(X) = O(n, m)^X \cdot \text{Diag } \partial f(\sigma(X)), \quad (1)$$

where

$$O(n, m)^X = \{(U_n, U_m) \in O(n, m) : (U_n, U_m) \cdot \text{Diag } \sigma(X) = X\}.$$

The regular subgradients satisfy corresponding formula.

We define $O(n, m)_X = \{(U_n, U_m) \in O(n, m) : (U_n, U_m) \cdot X = X\}$, which is the stabilizer of X in $O(n, m)$ under the defined action. Clearly for any $(U_n, U_m) \in O(n, m)^X$ we have the relationship

$$(U_n, U_m) O(n, m)_{\text{Diag } \sigma(X)} = O(n, m)^X.$$

3. Clarke Subgradients – the Lipschitz Case

One can easily see that f is locally Lipschitz around $\sigma(X)$ if and only if $F = f \circ \sigma$ is locally Lipschitz around X , and in this section we will assume that this is the case. It is important to notice that we have the following extension. The proof follows immediately from the definitions.

THEOREM 3.1 (Subgradient Invariance & Symmetricity). *If the function f is locally Lipschitz around x then both the Subgradient Invariance Theorem and the Symmetricity Theorem, stated in the previous section, can be extended to cover the Clarke subdifferential case.*

If X is an $n \times n$ square symmetric matrix (that is $X \in S(n)$) then $\lambda(X)$ will denote its eigenvalues arranged in nonincreasing order. The following lemma whose proof can be found in [9, Lemma 3], is needed later.

LEMMA 3.2. *For any vector w in \mathbb{R}_\downarrow^n , the function $w^T \lambda$ is convex on $S(n)$, and any vector x in \mathbb{R}_\downarrow^n satisfies $\text{Diag } w \in \partial(w^T \lambda)(\text{Diag } x)$.*

The proof of the next lemma is elementary and uses the fact that the sum of the k -largest eigenvalues or the k -largest singular values is a sublinear function, see [3, Corollary 4.3.18] and [3, Example 7.4.24].

LEMMA 3.3. (i) *For any vector w in \mathbb{R}_\downarrow^n the function $w^T \lambda$ is sublinear.*
 (ii) *For any vector w in \mathbb{R}_\downarrow^n the function $w^T \sigma$ is sublinear.*

A subset C of E is *invariant* under a subgroup, G , of $O(n)$ if $gC = C$ for all transformations g in G . If the function $f: \mathbb{R}^n \rightarrow [-\infty, +\infty]$ is absolutely symmetric then the regular subdifferential of f at a point x in \mathbb{R}^n is a convex set, invariant under the stabilizer $P_{(-)}(n)_x$ by the Subgradient Invariance Theorem.

Given a partitioning of the set $\{1, 2, \dots, n\}$, into $r + 1$ blocks I_1, I_2, \dots, I_{r+1} , of one or several consecutive integers we, write any vector y in \mathbb{R}^n in the form

$$y = \bigoplus_{l=1}^{r+1} y^l, \quad \text{where } y^l \in \mathbb{R}^{|I_l|} \text{ for each } l.$$

For matrices U^l in $M_{|I_l|}$ for each $1 \leq l \leq r$, and U^{r+1} in either $M_{|I_{r+1}|}$, $M_{|I_{r+1}|+m-n}$, or $M_{|I_{r+1}|, |I_{r+1}|+m-n}$, we write $\text{Diag}(U^l)$ for the block diagonal matrix

$$\begin{pmatrix} U^1 & 0 & \dots & 0 \\ 0 & U^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & U^{r+1} \end{pmatrix}.$$

It is clear that $\text{Diag}(U^l)$ will be either an $n \times n$, $m \times m$ square or an $n \times m$ rectangular matrix, depending on the dimensions of U^{r+1} , and it will be clear from the context which is the case.

Suppose we are given the following subgroups of $P_{(-)}(n)$ and $O(n, m)$ respectively:

$$\begin{aligned} \tilde{P}(n) &= \{\text{Diag}(P^l) : P^l \in P(|I_l|), 1 \leq l \leq r \text{ and } P^{r+1} \in P_{(-)}(|I_{r+1}|)\}, \\ \tilde{O}(n, m) &= \{(\text{Diag}(U^l), \text{Diag}(V^l)) : U^l = V^l \in O(|I_l|), 1 \leq l \leq r \text{ and} \\ &\quad U^{r+1} \in O(|I_{r+1}|); V^{r+1} \in O(|I_{r+1}| + m - n)\}. \end{aligned}$$

Notice that $\tilde{P}(n)$ is the group defined by the property: $\tilde{P}(n)x = x$ for all $x \in \mathbb{R}^n$ such that $x_i = x_j \Leftrightarrow i, j \in I_l$ for some l and $x_i = 0 \Leftrightarrow i \in I_{r+1}$.

LEMMA 3.4 (Sum of Invariant Sets). *If the sets $C, D \subset \mathbb{R}^n$ are convex and invariant under the group $\tilde{P}(n)$ then*

$$\tilde{O}(n, m).Diag C + \tilde{O}(n, m).Diag D = \tilde{O}(n, m).Diag(C + D).$$

Proof. Diagonalizing each block for $1 \leq l \leq r$ and applying the singular value decomposition theorem to the last, $(r + 1)$ st, block proves the equality

$$\tilde{O}(n, m).Diag C = \left\{ \text{Diag}(X^l) : \bigoplus_{l=1}^r \lambda(X^l) \oplus \sigma(X^{r+1}) \in C \right\}. \quad (2)$$

Let

$$X = \text{Diag}(X^l) \in \tilde{O}(n, m).Diag C,$$

and

$$Y = \text{Diag}(Y^l) \in \tilde{O}(n, m).Diag D.$$

We wish to show

$$X + Y \in \tilde{O}(n, m).Diag(C + D),$$

or equivalently, by identity (2),

$$\bigoplus_{l=1}^r \lambda(X^l + Y^l) \oplus \sigma(X^{r+1} + Y^{r+1}) \in C + D.$$

Since identity (2) shows $\bigoplus_{l=1}^r \lambda(X^l) \oplus \sigma(X^{r+1})$ lies in the convex set C and $\bigoplus_{l=1}^r \lambda(Y^l) \oplus \sigma(Y^{r+1})$ lies in the convex set D , it suffices to show

$$\begin{aligned} & \bigoplus_{l=1}^r \lambda(X^l + Y^l) \oplus \sigma(X^{r+1} + Y^{r+1}) \\ & \in \text{conv} \left(\tilde{P}(n) \left(\bigoplus_{l=1}^r \lambda(X^l) \oplus \sigma(X^{r+1}) \right) \right) + \\ & \quad + \text{conv} \left(\tilde{P}(n) \left(\bigoplus_{l=1}^r \lambda(Y^l) \oplus \sigma(Y^{r+1}) \right) \right). \end{aligned}$$

If this fails then there is a separating hyperplane separating the point from the set. That is, there exists a vector $z = \bigoplus_l z^l$ satisfying

$$\left\langle z, \bigoplus_{l=1}^r \lambda(X^l + Y^l) \oplus \sigma(X^{r+1} + Y^{r+1}) \right\rangle$$

$$\begin{aligned}
 &> \max \left\langle z, \text{conv} \left(\tilde{P}(n) \left(\bigoplus_{l=1}^r \lambda(X^l) \oplus \sigma(X^{r+1}) \right) \right) \right\rangle + \\
 &\quad + \text{conv} \left(\tilde{P}(n) \left(\bigoplus_{l=1}^r \lambda(Y^l) \oplus \sigma(Y^{r+1}) \right) \right) \Big\rangle \\
 &= \max \left\langle z, \tilde{P}(n) \left(\bigoplus_{l=1}^r \lambda(X^l) \oplus \sigma(X^{r+1}) \right) \right\rangle + \\
 &\quad + \max \left\langle z, \tilde{P}(n) \left(\bigoplus_{l=1}^r \lambda(Y^l) \oplus \sigma(Y^{r+1}) \right) \right\rangle.
 \end{aligned}$$

But then the (Absolute) Order Inequality and Lemma 3.3 show

$$\begin{aligned}
 &\sum_{l=1}^r \langle z^l, \lambda(X^l + Y^l) \rangle + \langle z^{r+1}, \sigma(X^{r+1} + Y^{r+1}) \rangle \\
 &> \sum_{l=1}^r \langle \bar{z}^l, \lambda(X^l) \rangle + \langle \widehat{z}^{r+1}, \sigma(X^{r+1}) \rangle + \sum_{l=1}^r \langle \bar{z}^l, \lambda(Y^l) \rangle + \langle \widehat{z}^{r+1}, \sigma(Y^{r+1}) \rangle \\
 &= \sum_{l=1}^r \langle \bar{z}^l, \lambda(X^l) + \lambda(Y^l) \rangle + \langle \widehat{z}^{r+1}, \sigma(X^{r+1}) + \sigma(Y^{r+1}) \rangle \\
 &\geq \sum_{l=1}^r \langle \bar{z}^l, \lambda(X^l + Y^l) \rangle + \langle \widehat{z}^{r+1}, \sigma(X^{r+1} + Y^{r+1}) \rangle \\
 &\geq \sum_{l=1}^r \langle z^l, \lambda(X^l + Y^l) \rangle + \langle z^{r+1}, \sigma(X^{r+1} + Y^{r+1}) \rangle,
 \end{aligned}$$

which is a contradiction. □

COROLLARY 3.5 (Convex Invariant Sets). *If the set $C \subset \mathbb{R}^n$ is convex and invariant under the group $\tilde{P}(n)$ then the set of matrices $\tilde{O}(n, m). \text{Diag } C$ is convex.*

Proof. We just have to apply the above lemma to the sets

$$C_1 = \lambda C, \quad D_1 = (1 - \lambda)C,$$

where λ is a number in $[0, 1]$. □

LEMMA 3.6. *If the set $C \subset \mathbb{R}^n$ is invariant under the group $\tilde{P}(n)$, then the following equality holds*

$$\text{conv}(\tilde{O}(n, m). \text{Diag } C) = \tilde{O}(n, m). \text{Diag}(\text{conv } C).$$

Proof. It is clear that $\tilde{O}(n, m). \text{Diag } C \subset \tilde{O}(n, m). \text{Diag}(\text{conv } C)$, and the later set is convex because of Corollary 3.5. Consequently

$$\text{conv}(\tilde{O}(n, m). \text{Diag } C) \subseteq \tilde{O}(n, m). \text{Diag}(\text{conv } C).$$

The opposite inclusion is trivial. □

Recently, an independent result by Tam and Hill, covering the result below, appeared in [16]. They consider invariant functions, called *orbital*, in the context of semisimple Lie group theory. We offer a direct approach that first appeared in the second author’s thesis [15].

THEOREM 3.7 (Clarke Subgradients). *The Clarke subdifferential of a locally Lipschitz singular value function $f \circ \sigma$ at a matrix X in $M_{n,m}$ is given by the formula*

$$\partial^c(f \circ \sigma)(X) = O(n, m)^X \cdot \text{Diag } \partial^c f(\sigma(X)), \tag{3}$$

where

$$O(n, m)^X = \{(U_n, U_m) \in O(n, m) : (U_n, U_m) \cdot \text{Diag } \sigma(X) = X\}.$$

Proof. Assume first $X = \text{Diag } x$ for a vector x in \mathbb{R}_{\neq}^n . After that the general case will follow easily by the Subgradient Invariance Theorem. Let

$$\begin{aligned} x_1 &= \dots = x_{k_1} > x_{k_1+1} = \dots = x_{k_2} > x_{k_2+1} \dots = x_{k_r} > x_{k_r+1} \\ &= \dots = x_{k_{r+1}} = 0, \end{aligned}$$

where $k_{r+1} = n$. Partition the set $\{1, 2, \dots, n\}$ into $r+1$ blocks: $I_1 = \{1, 2, \dots, k_1\}$, $I_2 = \{k_1 + 1, \dots, k_2\}, \dots, I_{r+1} = \{k_r + 1, \dots, k_{r+1}\}$.

We are going to compute the group $O(n, m)^{\text{Diag } x}$ (it is a group since $x \in \mathbb{R}_{\neq}^n$). If (U_n, U_m) is in $O(n, m)^{\text{Diag } x}$, then we have

$$\begin{aligned} (\text{Diag } x)(\text{Diag } x)^T U_n &= U_n (\text{Diag } x)(\text{Diag } x)^T, \\ (\text{Diag } x)^T (\text{Diag } x) U_m &= U_m (\text{Diag } x)^T (\text{Diag } x), \end{aligned}$$

which shows that $U_n = \text{Diag}(U^l)$, where $U^l \in O(|I_l|)$ for $1 \leq l \leq r + 1$, and $U_m = \text{Diag}(V^l)$, where $V^l \in O(|I_l|)$ for $1 \leq l \leq r$, and $V^{r+1} \in O(|I_{r+1}| + m - n)$. Now from the identity

$$U_n^T (\text{Diag } x) = (\text{Diag } x) U_m^T$$

one sees that $U^l = V^l$ for each $1 \leq l \leq r$. So we obtain

$$O(n, m)^{\text{Diag } x} = \tilde{O}(n, m). \tag{4}$$

Since x is invariant under the group $\tilde{P}(n)$ the convex set $\partial^c f(x)$ is also invariant under $\tilde{P}(n)$, by the Subgradient Invariance Theorem. Corollary 3.5 now shows that the set $\tilde{O}(n, m) \cdot \text{Diag } \partial^c f(x)$ is convex.

The Subgradient Theorem 2.1 now gives us

$$\partial^c(f \circ \sigma)(\text{Diag } x) = \text{conv } \partial(f \circ \sigma)(\text{Diag } x) = \text{conv}(\tilde{O}(n, m) \cdot \text{Diag } \partial f(x)).$$

Using the easily established fact

$$\tilde{O}(n, m) \cdot \text{Diag } \partial f(x) \subseteq \tilde{O}(n, m) \cdot \text{Diag } \partial^c f(x)$$

and the convexity of the right-hand side, we see that

$$\text{conv}(\tilde{\text{O}}(n, m).\text{Diag } \partial f(x)) \subseteq \tilde{\text{O}}(n, m).\text{Diag } \partial^c f(x).$$

On the other hand from $\partial^c f(x) = \text{conv } \partial f(x)$ one can immediately see that the reverse inclusion holds as well:

$$\begin{aligned} \tilde{\text{O}}(n, m).\text{Diag } \partial^c f(x) &= \tilde{\text{O}}(n, m).\text{Diag}(\text{conv } \partial f(x)) \\ &= \tilde{\text{O}}(n, m).\text{conv}(\text{Diag } \partial f(x)) \\ &\subseteq \text{conv}(\tilde{\text{O}}(n, m).\text{Diag } \partial f(x)) \\ &= \text{conv } \partial(f \circ \sigma)(\text{Diag } x) = \partial^c(f \circ \sigma)(\text{Diag } x). \end{aligned}$$

The result follows. \square

For completeness we would like to state and prove the Clarke version of the Diagonal Subgradients Corollary in [10]. (Diagonal Subgradients Corollary in [10] states that the result below holds for regular and limiting subgradients.)

COROLLARY 3.8 (Diagonal Clarke Subgradients). *For any vectors x and y in \mathbb{R}^n and any singular value function $f \circ \sigma$,*

$$y \in \partial^c f(x) \Leftrightarrow \text{Diag } y \in \partial^c(f \circ \sigma)(\text{Diag } x).$$

Proof. If the function f is Lipschitz around $\sigma(X)$ and y is a Clarke subgradient at x , then y is a convex combination of limiting subgradients $y^i \in \partial f(x)$. By the Diagonal Subgradients Theorem for limiting subgradients in [10], each matrix $\text{Diag } y^i$ is a subgradient of $f \circ \sigma$ at X , and since $\text{Diag } y$ is a convex combination of these matrices, $\text{Diag } y$ must be a Clarke subgradient.

To see the reverse implication choose a diagonal matrix $\text{Diag } y \in \partial^c(f \circ \sigma)(\text{Diag } x)$. Then the Clarke Subgradients Theorem above shows the existence of an element (U_n, U_m) in $\text{O}(n, m)$ and a vector z in $\partial^c f(\hat{x})$ such that $\text{Diag } y = (U_n, U_m).\text{Diag } z$ and $\text{Diag } x = (U_n, U_m).\text{Diag } \hat{x}$. By the Simultaneous Rectangular Conjugacy Theorem, there is a matrix $P_{(-)}$ in $P_{(-)}(n)$ with $y = P_{(-)}z$ and $x = P_{(-)}\hat{x}$, and the result follows from the Subgradient Invariance Theorem. \square

COROLLARY 3.9 (Strict Differentiability). *If f is Lipschitz around $\sigma(X)$, then $f \circ \sigma$ is strictly differentiable at X if and only if f strictly differentiable at $\sigma(X)$.*

Proof. In the Lipschitz case f is strictly differentiable at x if and only if the Clarke subdifferential is a singleton. By the above theorem and the fact that the Clarke subdifferential is a convex set this happens if and only if $\partial^c(f \circ \sigma)(X)$ is a singleton (since a convex set with a constant norm is a singleton).

4. Clarke Subgradients – the Lower Semicontinuous Case

A function f is called *lower semicontinuous* if its graph

$$\text{epi } f = \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leq \alpha\}$$

is a closed subset of \mathbb{R}^{n+1} . Let $C \subset \mathbb{R}^n$ and $x \in C$. A vector v is a *regular normal* to C at x , written $v \in \hat{N}_C(x)$, if

$$\overline{\lim}_{\substack{z \rightarrow x \\ z \in C}} \frac{\langle v, z - x \rangle}{\|z - x\|} \leq 0.$$

A vector v is a *normal* to C at x , written $v \in N_C(x)$, if there is a sequence of points x^r in C approaching x , and a sequence of regular normals v^r in $\hat{N}_C(x^r)$ approaching v . Notice that $N_C(x)$ is a closed cone. The set of *Clarke subgradients* of a function f at x , $\bar{\partial} f(x)$, is defined by

$$\partial^c f(x) = \{v \mid (v, -1) \in \text{cl conv } N_{\text{epi } f}(x, f(x))\},$$

and is called the *Clarke subdifferential*. It can be shown (see [14, Theorem 9.13(b) and Theorem 8.49]) that if f is locally Lipschitz around x then this definition coincides with the definition given at the beginning, that is why we use the same notation for the subdifferential, ∂^c , as in the locally Lipschitz case. If f is lower semicontinuous around x then we have the formula (see [14, Theorem 8.9]):

$$\begin{aligned} N_{\text{epi } f}(x, f(x)) \\ = \{\lambda(v, -1) \mid v \in \partial f(x), \lambda > 0\} \cup \{(v, 0) \mid v \in \partial^\infty f(x)\}. \end{aligned} \quad (5)$$

The following lemma can be found in [12, Proposition 2.6], we include a proof for completeness.

LEMMA 4.1. *If f is lower semicontinuous around x we have the representation*

$$\partial^c f(x) = \text{cl}(\text{conv } \partial f(x) + \text{conv } \partial^\infty f(x)).$$

In particular when the cone $\partial^\infty f(x)$ doesn't contain lines we have (see also [14, Theorem 8.49]) the simpler formula

$$\partial^c f(x) = \text{conv } \partial f(x) + \text{conv } \partial^\infty f(x).$$

Proof. Define the sets

$$\begin{aligned} K_1 &= \{(v, 0) \mid v \in \partial^\infty f(x)\}, \\ K_2 &= \{\lambda(v, -1) \mid v \in \partial f(x), \lambda > 0\}, \end{aligned}$$

and

$$L = \{x \in \mathbb{R}^{n+1} \mid x_{n+1} = -1\}.$$

Then by (5) we get

$$\text{conv } N_{\text{epi } f}(x, f(x)) = \text{conv } K_1 + \text{conv } K_2, \tag{6}$$

and by the definition of the set L

$$(\text{conv } K_1 + \text{conv } K_2) \cap L = \{(v, -1) \mid v \in \text{conv } \partial^\infty f(x) + \text{conv } \partial f(x)\}. \tag{7}$$

Let us see on the other hand that the following equality holds:

$$(\text{cl conv } N_{\text{epi } f}(x, f(x))) \cap L = \text{cl}(\text{conv } N_{\text{epi } f}(x, f(x)) \cap L). \tag{8}$$

Indeed, take a point $(v, -1)$ in $(\text{cl conv } N_{\text{epi } f}(x, f(x))) \cap L$. So there is a sequence (v^r, α^r) in $\text{conv } N_{\text{epi } f}(x, f(x))$, approaching $(v, -1)$. For big enough r , we have $\alpha^r < 0$. Then

$$\left(\frac{v^r}{|\alpha^r|}, \frac{\alpha^r}{|\alpha^r|} \right) = \left(\frac{v^r}{|\alpha^r|}, -1 \right)$$

is in $\text{conv } N_{\text{epi } f}(x, f(x)) \cap L$, approaching $(v, -1)$. So $(v, -1)$ is in $\text{cl}(\text{conv } N_{\text{epi } f}(x, f(x)) \cap L)$. The opposite inclusion is clear.

So putting (6), (7), and (8) together

$$\begin{aligned} \{(v, -1) \mid v \in \partial^c f(x)\} &= (\text{cl conv } N_{\text{epi } f}(x, f(x))) \cap L \\ &= \text{cl}\{(v, -1) \mid v \in \text{conv } \partial^\infty f(x) + \text{conv } \partial f(x)\} \\ &= \{(v, -1) \mid v \in \text{cl}(\text{conv } \partial^\infty f(x) + \text{conv } \partial f(x))\}, \end{aligned}$$

and we are done. In the other case, we have that the cone $\partial^\infty f(x)$ doesn't contain lines if and only if $N_{\text{epi } f}(x, f(x))$ doesn't contain lines. Since when a cone doesn't contain lines and is closed, so too is its convex hull (see [14, Theorem 3.15]), we get

$$\text{cl conv } N_{\text{epi } f}(x, f(x)) = \text{conv } N_{\text{epi } f}(x, f(x))$$

and the second formula becomes clear. □

Let (U_n, U_m) be an arbitrary, fixed element of the set $O(n, m)^X$. Then the representation $O(n, m)^X = (U_n, U_m)O(n, m)_{\text{Diag } \sigma(X)}$ holds, where the symbol $O(n, m)_{\text{Diag } \sigma(X)}$ denotes the stabilizer of the matrix $\text{Diag } \sigma(X)$ in the group $O(n, m)$. Notice that the matrices in the stabilizer $O(n, m)_{\text{Diag } \sigma(X)}$ have the same structure as those in the set $\tilde{O}(n, m)$ in Lemma 3.4 and Corollary 3.5. Let now f be an absolutely symmetric function. Clearly f is lower semicontinuous if and only if $f \circ \sigma$ is lower semicontinuous. Using (in this order) Lemma 4.1, Theorem 2.1, Lemma 3.6, Corollary 3.5, Lemma 3.4, a simple limiting argument using the fact that the set $O(n, m)^X$ is compact (when exchanging it with 'cl'), and using

everywhere the above representation, we get:

$$\begin{aligned}
& \partial^c(f \circ \sigma)(X) \\
&= \text{cl}(\text{conv } \partial^\infty(f \circ \sigma)(X) + \text{conv } \partial(f \circ \sigma)(X)) \\
&= \text{cl}(\text{conv } \mathbf{O}(n, m)^X \cdot \text{Diag } \partial^\infty f(\sigma(X)) + \text{conv } \mathbf{O}(n, m)^X \cdot \text{Diag } \partial f(\sigma(X))) \\
&= \text{cl}(\mathbf{O}(n, m)^X \cdot \text{conv Diag } \partial^\infty f(\sigma(X)) + \mathbf{O}(n, m)^X \cdot \text{conv Diag } \partial f(\sigma(X))) \\
&= \text{cl}(\mathbf{O}(n, m)^X \cdot (\text{conv Diag } \partial^\infty f(\sigma(X)) + \text{conv Diag } \partial f(\sigma(X)))) \\
&= \mathbf{O}(n, m)^X \cdot \text{cl}(\text{conv Diag } \partial^\infty f(\sigma(X)) + \text{conv Diag } \partial f(\sigma(X))) \\
&= \mathbf{O}(n, m)^X \cdot \text{Diag } \text{cl}(\text{conv } \partial^\infty f(\sigma(X)) + \text{conv } \partial f(\sigma(X))) \\
&= \mathbf{O}(n, m)^X \cdot \text{Diag } \partial^c(f(\sigma(X))).
\end{aligned}$$

This proves the following theorem.

THEOREM 4.2. *If $X \in M_{n,m}$ and f is an absolutely symmetric function and lower semicontinuous around $\sigma(X)$, then $f \circ \sigma$ is lower semicontinuous around X and*

$$\partial^c(f \circ \sigma)(X) = \mathbf{O}(n, m)^X \cdot \partial^c(f(\sigma(X))),$$

where

$$\mathbf{O}(n, m)^X = \{(U_n, U_m) \in \mathbf{O}(n, m) : (U_n, U_m) \cdot \text{Diag } \sigma(X) = X\}.$$

5. Proximal Subgradients

In this section we show that the formula in Theorem 4.2 also holds for *proximal subgradients* of singular value functions.

DEFINITION 5.1 (Proximal Subgradients). A vector y is called a *proximal subgradient* of a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ at x , a point where $f(x)$ is finite, if there exist $\rho > 0$ and $\delta > 0$ such that

$$f(x + z) \geq f(x) + \langle y, z \rangle - \frac{1}{2}\rho \|z\|^2 \quad \text{when } \|z\| \leq \delta.$$

The set of all proximal subgradients will be denoted with $\partial_p f(x)$.

It is clear from the definition that

$$\partial_p f(x) \subseteq \hat{\partial} f(x). \tag{9}$$

LEMMA 5.2 (Proximal Subgradients Invariance). *Suppose the function $f: E \rightarrow [-\infty, +\infty]$ (E is an inner product space) is invariant under a subgroup G of $\mathbf{O}(E)$, then any point x in E and transformation g in G satisfy $\partial_p f(gx) = g \partial_p f(x)$.*

Proof. Suppose first $y \in \partial_p f(x)$, so there is a $\rho > 0$ such that all z in E sufficiently close to 0 satisfy $f(x + z) \geq f(x) + \langle y, z \rangle - \frac{1}{2}\rho\|z\|^2$. Using the invariance of f we get

$$\begin{aligned} f(gx + z) &= f(x + g^{-1}z) \geq f(x) + \langle y, g^{-1}z \rangle - \frac{1}{2}\rho\|g^{-1}z\|^2 \\ &= f(gx) + \langle gy, z \rangle - \frac{1}{2}\rho\|z\|^2, \end{aligned}$$

so $gy \in \partial_p f(gx)$. One can easily see that $\partial_p f(gx) = g\partial_p f(x)$. □

5.1. A PRELIMINARY RESULT

Our aim in this auxiliary section will be to prove the identity

$$\sigma(X + M) = \sigma(X) + \sigma'(X; M) + O(\|M\|^2)$$

and as an added bonus we will obtain an expression for $\sigma'(X; M)$. First of all from [3, Theorem 4.3.1] we have that

$$\lambda(X + M) = \lambda(X) + O(\|M\|). \tag{10}$$

We will use the following notation and results from [17]. If A is an $n \times n$ symmetric matrix, its eigenvalues are all real and we can arrange them in nonincreasing order

$$\begin{aligned} \lambda_1(A) \cdots &\geq \lambda_{i-1}(A) > \lambda_i(A) = \cdots = \lambda_j(A) > \lambda_{j+1}(A) \\ &\geq \cdots \lambda_n(A), \end{aligned}$$

where $i \leq l \leq j$ and $\lambda_l(A)$ is the l -th largest eigenvalue of A (counting multiplicity of each of them). The following proposition is an easy consequence of Equation (10) and Proposition 1.4 in [17].

PROPOSITION 5.3. *Let $A \in S(n)$ and $U \in O(n)$ so that*

$$U^T A U = \text{Diag}(\lambda_1(A), \dots, \lambda_n(A)) \quad (U = [u_1, \dots, u_n]).$$

If we set $U_1 := [u_i, \dots, u_j]$ then

$$\lambda_l(A + E) = \lambda_l(A) + \lambda_{l-i+1}(U_1^T E U_1) + O(\|E\|^2).$$

Fix $X \in M_{n,m}$, let $M \in M_{n,m}$ be a perturbation matrix, and

$$X = V^T (\text{Diag } \sigma(X)) W$$

be the singular value decomposition of X . Define

$$A := \begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix}, \quad E := \begin{pmatrix} 0 & M \\ M^T & 0 \end{pmatrix}.$$

It is well known (see [3, Theorem 7.3.7]) that the eigenvalues of the matrix A are $(\sigma_1(X), \dots, \sigma_n(X), 0, \dots, 0, -\sigma_n(X), \dots, -\sigma_1(X))$ with $m - n$ zeros in the middle. Let $U \in M_{n+m}$ be the orthogonal matrix that gives the ordered spectral decomposition of A , that is

$$U^T A U = \text{Diag}(\sigma_1(X), \dots, \sigma_n(X), 0, \dots, 0, -\sigma_n(X), \dots, -\sigma_1(X)).$$

We apply the above proposition to the l -th eigenvalue of A , $1 \leq l \leq n$, using the matrices A , E , and U to get

$$\begin{aligned} \sigma_l(X + M) &= \lambda_l(A + E) = \lambda_l(A) + \lambda_{l-i+1}(U_1^T E U_1) + O(\|E\|^2) \\ &= \sigma_l(X) + \lambda_{l-i+1}(U_1^T E U_1) + O(\|M\|^2). \end{aligned}$$

In particular we get that

$$\sigma'(X; M) = \lambda_{l-i+1}(U_1^T E U_1).$$

5.2. PROXIMAL SUBGRADIENTS

Following the standard reduction ideas we first prove a simpler version of the theorem we want.

LEMMA 5.4 (Diagonal Proximal Subgradients). *For any vectors x in \mathbb{R}_\pm^n , y in \mathbb{R}^n and any singular value function $f \circ \sigma$ we have*

$$y \in \partial_p f(x) \Leftrightarrow \text{Diag } y \in \partial_p (f \circ \sigma)(\text{Diag } x).$$

Proof. Suppose first that $\text{Diag } y$ is a proximal subgradient. Then there are $\rho > 0$ and $\delta > 0$ such that for all vectors z in \mathbb{R}^n such that $\|z\| < \delta$ we have

$$\begin{aligned} f(x + z) &= (f \circ \sigma)(\text{Diag } x + \text{Diag } z) \\ &\geq (f \circ \sigma)(\text{Diag } x) + \text{tr}(\text{Diag } y)(\text{Diag } z) - \frac{1}{2}\rho\|\text{Diag } z\|^2 \\ &= f(x) + \langle y, z \rangle - \frac{1}{2}\rho\|z\|^2, \end{aligned}$$

so $y \in \partial_p f(x)$. (In this case we didn't use that $x \in \mathbb{R}_\pm^n$.)

In the opposite direction, let $y \in \partial_p f(x)$. By Lemma 5.2, every element of the finite set $P_{(-)}(n)_x y$ is a proximal subgradient of f at x . We consider the support function of the convex hull of this set (which we denote by Λ),

$$\delta_\Lambda^*(z) = \max\{z^T P_{(-)} y : P_{(-)} \in P_{(-)}(n)_x\}, \quad \text{for all } z \text{ in } \mathbb{R}^n.$$

This function is sublinear, with global Lipschitz constant $\|y\|$. The definition of proximal subgradients implies that there are numbers $\rho > 0$ and $\delta > 0$ such that for all vectors z in \mathbb{R}^n satisfying $\|z\| < \delta$ we have

$$f(x + z) \geq f(x) + \delta_\Lambda^*(z) - \frac{1}{2}\rho\|z\|^2. \tag{11}$$

On the other hand using the result from the previous subsection, sufficiently small matrices Z in $M_{m,n}$ must satisfy

$$\|\sigma(\text{Diag } x + Z) - x - \sigma'(\text{Diag } x; Z)\| \leq K \|Z\|^2.$$

Therefore by inequality (11), together with the Lipschitzness of δ_Λ^* and σ , we get

$$\begin{aligned} f(\sigma(\text{Diag } x + Z)) &= f(x + (\sigma(\text{Diag } x + Z) - x)) \\ &\geq f(x) - \frac{1}{2}\rho \|\sigma(\text{Diag } x + Z) - x\|^2 + \\ &\quad + \delta_\Lambda^*(\sigma'(\text{Diag } x; Z) + [\sigma(\text{Diag } x + Z) - x - \sigma'(\text{Diag } x; Z)]) \\ &\geq f(x) + \delta_\Lambda^*(\sigma'(\text{Diag } x; Z)) - (\frac{1}{2}\rho + K\|y\|)\|Z\|^2. \end{aligned}$$

Recall that by the Singular Value Derivatives Theorem we have

$$\text{diag } Z \in \text{conv}(P_{(-)}(n)_x \sigma'(\text{Diag } x; Z)). \quad (12)$$

Since the polytope Λ is invariant under the group $P_{(-)}(n)_x$, so is its support function, so

$$\delta_\Lambda^*(P_{(-)} \sigma'(\text{Diag } x; Z)) = \delta_\Lambda^*(\sigma'(\text{Diag } x; Z)),$$

for any matrix $P_{(-)}$ in $P_{(-)}(n)_x$. The convexity of δ_Λ^* , its invariance property, and relation (12), imply that

$$\delta_\Lambda^*(\text{diag } Z) \leq \delta_\Lambda^*(\sigma'(\text{Diag } x; Z)).$$

We continue the chain of inequalities above:

$$\begin{aligned} f(\sigma(\text{Diag } x + Z)) &\geq f(x) + \delta_\Lambda^*(\text{diag } Z) - (\frac{1}{2}\rho + K\|y\|)\|Z\|^2 \\ &\geq f(x) + y^T \text{diag } Z - (\frac{1}{2}\rho + K\|y\|)\|Z\|^2 \\ &= f(x) + \langle \text{Diag } y, Z \rangle - (\frac{1}{2}\rho + K\|y\|)\|Z\|^2, \end{aligned}$$

and the result follows. \square

We are now ready to prove again the formula that pervades the whole paper in the case of proximal subdifferentials.

THEOREM 5.5 (Proximal Subgradients). *The proximal subdifferential of any singular value function $f \circ \sigma$ at a matrix X in $M_{n,m}$ is given by the formula*

$$\partial_p(f \circ \sigma)(X) = \text{O}(n, m)^X \cdot \text{Diag } \partial_p f(\sigma(X)),$$

where

$$\text{O}(n, m)^X = \{(U_n, U_m) \in \text{O}(n, m) : (U_n, U_m) \cdot \text{Diag } \sigma(X) = X\}.$$

Proof. For any vector y in $\partial_p f(\sigma(X))$, the Diagonal Proximal Subgradients Lemma 5.4 shows

$$\text{Diag } y \in \partial_p(f \circ \sigma)(\text{Diag } \sigma(X)),$$

and now, for any element (U_n, U_m) in $O(n, m)^X$, from the Proximal Subgradients Invariance Lemma 5.2 we get

$$(U_n, U_m) \cdot \text{Diag } y \in \partial_p(f \circ \sigma)((U_n, U_m) \cdot \text{Diag } \sigma(X)) = \partial_p(f \circ \sigma)(X),$$

and we are done with showing the inclusion ‘ \supseteq ’. We now show the opposite inclusion ‘ \subseteq ’. Let $Y \in \partial_p(f \circ \sigma)(X)$. Because $\partial_p(f \circ \sigma)(X) \subseteq \hat{\partial}(f \circ \sigma)(X) \subseteq \partial(f \circ \sigma)(X)$, the Symmetricity Theorem implies that $X^T Y = Y^T X$ and $Y^T X = X^T Y$. This means that the rectangular matrices X and Y can be simultaneously diagonalized by one and the same orthogonal pair (U_n, U_m) (see [10]). We get that

$$Y = U_n^T (\text{Diag } P_{(-)} \sigma(Y)) U_m, \quad X = U_n^T (\text{Diag } \sigma(X)) U_m,$$

for some element (U_n, U_m) in $O(n, m)$, and some $P_{(-)}$ in $P_{(-)}(n)$. Consequently $(U_n, U_m) \in O(n, m)^X$. Lemma 5.2 shows that

$$\text{Diag } P_{(-)} \sigma(Y) \in \partial_p(f \circ \sigma)(\text{Diag } \sigma(X)).$$

Finally the Diagonal Proximal Subgradients Lemma 5.4 gives us

$$P_{(-)} \sigma(Y) \in \partial_p f(\sigma(X)).$$

Thus the matrix Y belongs to the set $O(n, m)^X \cdot \text{Diag } \partial_p f(\sigma(X))$. □

6. Absolute Order Statistics and Individual Singular Values

In this section we want to present a useful application of the different variations of the Subgradients Theorems. We are going to calculate the proximal, regular, limiting, horizon, and Clarke subdifferentials of an individual singular value $\sigma_k(\cdot)$. The availability of such formulas indicated the potential of this approach in matrix perturbation theory.

We start by defining the absolutely symmetric function corresponding to the r -th singular value. The k th *absolute order statistic* $\varphi_k: \mathbb{R}^n \rightarrow \mathbb{R}$ is defined to be

$$\varphi_k(x) = k\text{th largest element of } \{|x_1|, |x_2|, \dots, |x_n|\}$$

(or in other words $\varphi_k(x) = (\hat{x})_k$). It clearly satisfies the relation $\varphi_k(x) = \sigma_k(\text{Diag } x)$. To apply the Subgradient Theorem, note that $\sigma_k = \varphi_k \circ \sigma$. Thus we must first compute the subdifferential of φ_k . We define the function $\text{sign}(x)$ as

$$\text{sign}(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ -1, & \text{if } x < 0. \end{cases}$$

Let $\{e_1, \dots, e_n\}$ be the standard basis in \mathbb{R}^n .

PROPOSITION 6.1. *At any point x in \mathbb{R}^n , the regular subgradients of the k th absolute order statistic are described by*

$$\hat{\partial}\varphi_k(x) = \begin{cases} \text{conv}\{\pm e_i \mid |x_i| = \varphi_k(x)\}, & \text{if } \varphi_{k-1}(x) > \varphi_k(x) = 0, \\ \text{conv}\{(\text{sign}(x_i))e_i \mid |x_i| = \varphi_k(x)\}, & \text{if } \varphi_{k-1}(x) > \varphi_k(x) \neq 0, \\ \emptyset, & \text{otherwise,} \end{cases}$$

and moreover $\partial^\infty\varphi_k(x) = \{0\}$, and $\partial_p\varphi_k(x) = \hat{\partial}\varphi_k(x)$.

Proof. Define the set of indices $I = \{i \mid |x_i| = \varphi_k(x)\}$, and consider several cases.

If the inequality $\varphi_{k-1}(x) > \varphi_k(x)$ holds then clearly, close to the point x , the function φ_k is given by $w \in \mathbb{R}^n \mapsto \max_{i \in I} |w_i|$. The subdifferential at x of this second function (which is convex) is $\text{conv}\{\pm e_i \mid |x_i| = \varphi_k(x)\}$ if $\varphi_k(x) = 0$ or is $\text{conv}\{(\text{sign}(x_i))e_i \mid |x_i| = \varphi_k(x)\}$ if $\varphi_k(x) \neq 0$. (See [13, Theorem 23.8] together with [1, Problem 3.2.13].)

On the other hand, in the case $\varphi_{k-1}(x) = \varphi_k(x)$, suppose y is regular subgradient, and so satisfies

$$\varphi_k(x + z) \geq \varphi_k(x) + y^T z + o(z), \quad \text{as } z \rightarrow 0.$$

Here we consider two subcases whose argumentation slightly differ from one another.

Assume first that $\varphi_{k-1}(x) = \varphi_k(x) = 0$. For any index i in I , all small positive δ satisfy $\varphi_k(x + \delta e_i) = \varphi_k(x)$ and $\varphi_k(x - \delta e_i) = \varphi_k(x)$, from which we deduce $y_i = 0$ for each i in I . But also

$$\varphi_k\left(x + \delta \sum_{i \in I} e_i\right) = \varphi_k(x) + \delta, \quad \text{and} \quad \varphi_k\left(x - \delta \sum_{i \in I} e_i\right) = \varphi_k(x) + \delta,$$

which leads to the contradiction $\sum_{i \in I} y_i = 1$. So $\hat{\partial}\varphi_k(x) = \emptyset$.

Second, suppose we have $\varphi_{k-1}(x) = \varphi_k(x) > 0$. For any index i in I , all small positive δ satisfy $\varphi_k(x + \delta(\text{sign}(x_i))e_i) = \varphi_k(x)$, from which we deduce $(\text{sign}(x_i))y_i \leq 0$, but also

$$\varphi_k\left(x - \delta \sum_{i \in I} (\text{sign}(x_i))e_i\right) = \varphi_k(x) - \delta,$$

which leads to the contradiction $\sum_{i \in I} (\text{sign}(x_i))y_i \geq 1$. Again we must have had $\sum_{i \in I} y_i = 1$.

The horizon subdifferential is easy to check since φ_k is Lipschitz. For the last claim we use the fact that for any function $\partial_p f(x) \subseteq \hat{\partial} f(x)$ with equality whenever f is convex. \square

For a vector y in \mathbb{R}^n we define the support of y to be

$$\text{supp } y = \{i \mid y_i \neq 0\}.$$

The number of elements in this set is then $|\text{supp } y|$. It will help to think that the structure of the vector $(\varphi_1(x), \varphi_2(x), \dots, \varphi_n(x))$ is given by

$$\begin{aligned} \varphi_1(x) &= \dots = \varphi_{k_1}(x) > \dots > \varphi_{k_{l-1}+1}(x) = \dots = \varphi_k(x) = \dots = \varphi_{k_l}(x) \\ &> \dots \varphi_{k_r}(x) \geq 0 \quad (k_0 = 0, k_r = n), \end{aligned}$$

so that $\varphi_k(x)$ is somewhere in the l th block of equal entries.

THEOREM 6.2 (*k*th Absolute Ordered Statistic). *The Clarke subdifferential of the kth absolute ordered statistic φ_k at a point x in \mathbb{R}^n is given by*

$$\partial^c \varphi_k(x) = \begin{cases} \text{conv}\{\pm e_i \mid |x_i| = \varphi_k(x)\}, & \text{if } \varphi_k(x) = 0, \\ \text{conv}\{(\text{sign}(x_i))e_i \mid |x_i| = \varphi_k(x)\}, & \text{otherwise,} \end{cases}$$

whereas the (limiting) subdifferential is given by

$$\begin{aligned} \partial \varphi_k(x) &= \{y \in \partial^c \varphi_k(x) \mid |\text{supp } y| \leq \alpha\}, \quad \text{where} \\ \alpha &= 1 - k + |\{i \mid |x_i| \geq \varphi_k(x)\}|. \end{aligned} \tag{13}$$

Regularity holds if and only if $\varphi_{k-1}(x) > \varphi_k(x)$.

Remark 6.3. Notice that α is equal to the number of elements in the same block as $\varphi_k(x)$ after $\varphi_k(x)$, including $\varphi_k(x)$. In other words, with the notation introduced right before the theorem we can get the expression $\alpha = k_l - k + 1$.

Proof. We begin by proving Equation (13). Every vector z in a small enough neighbourhood around x will have the property that $\hat{z}_i = \hat{z}_j \Rightarrow \hat{x}_i = \hat{x}_j$ for all i and j . That is why by using Proposition 6.1 one can easily see that for all z in that neighbourhood, $\partial \varphi_k(z)$ is contained in the set in the right-hand side of Equation (13). Because this set is closed, after taking limits we see that $\partial \varphi_k(x)$ is contained in it as well.

We now show the opposite inclusion. Take a vector y in the right-hand side of (13) and an index set J such that

$$\begin{aligned} |J| &= n - \alpha, \\ j \in J &\Rightarrow y_j = 0, \\ \{i \mid |x_i| \neq \varphi_k(x)\} &\subseteq J. \end{aligned}$$

It can easily be seen that for small enough δ we have

$$\varphi_{k-1}\left(x + \delta \sum_{i \in J} (\text{sign}(x_i))e_i\right) > \varphi_k\left(x + \delta \sum_{i \in J} (\text{sign}(x_i))e_i\right) = \varphi_k(x).$$

Finally using Proposition 6.1 we see that, depending on the case considered,

$$y \in \left\{ \begin{array}{l} \text{conv}\{\pm e_i \mid i \notin J\} \\ \text{conv}\{(\text{sign}(x_i))e_i \mid i \notin J\} \end{array} \right\} = \hat{\partial} \varphi_k\left(x + \delta \sum_{i \in J} (\text{sign}(x_i))e_i\right),$$

whence by taking limits we conclude that $y \in \partial\varphi_k(x)$. The formulas for the Clarke case follow by taking convex hulls. The regularity claim follows by Proposition 6.1. \square

Finally the subdifferentials of the singular value function $\sigma_k(X)$ are given by the following corollary.

COROLLARY 6.4 (Singular Value Subgradients). *The Clarke subdifferential of the k th singular value σ_k at a matrix X in $M_{n,m}$ is given by*

$$\partial^c \sigma_k(X) = \text{conv}\{vw^T \mid \|v\| = \|w\| = 1, Xw = \sigma_k(X)v, X^T v = \sigma_k(X)w\}$$

whereas the (limiting) subdifferential is given by

$$\begin{aligned} \partial \sigma_k(X) &= \{Y \in \partial^c \sigma_k(X) \mid \text{rank } Y \leq \alpha\}, \quad \text{where} \\ \alpha &= 1 - k + |\{i \mid \sigma_i(X) \geq \sigma_k(X)\}|. \end{aligned}$$

Regularity holds if and only if $\sigma_{k-1}(X) > \sigma_k(X)$.

Proof. We will only deduce the formula for the Clarke subdifferential. The limiting one and the condition for regularity will follow easily.

Fix a matrix X . For any pair $(V, W) \in O(n, m)^X$ we have that $X = V^T(\text{Diag } \sigma(X))W$ is the (ordered) singular value decomposition of X , where we suppose $V^T = [v_1, \dots, v_n]$ and $W^T = [w_1, \dots, w_m]$. We first consider the case when $\sigma_k(X) > 0$. For any index i , such that $\sigma_i(X) = \sigma_k(X)$, using $VX = (\text{Diag } \sigma(X))W$ we can express the i th row on both sides: $\sigma_i(X)w_i^T = v_i^T X$. Then

$$V^T(\text{Diag } e_i)W = v_i w_i^T.$$

By Theorem 3.7 we get

$$\partial^c \sigma_k(X) = (U_n, U_m)O(n, m)_{\text{Diag } \sigma(X)}.(\text{Diag conv}\{e_i \mid \sigma_i(X) = \sigma_k(X)\}),$$

where (U_n, U_m) is a fixed element of $O(n, m)^X$. The set $\{e_i \mid \sigma_i(X) = \sigma_k(X)\}$ is clearly invariant under the subgroup, $\tilde{P}(n)$, of $P_{(-)}(n)$ that stabilizes $\sigma(X)$. Then by Lemma 3.6 and recalling that $O(n, m)_{\text{Diag } \sigma(X)} = \tilde{O}(n, m)$ we obtain

$$\begin{aligned} \partial^c \sigma_k(X) &= (U_n, U_m)\text{conv } \tilde{O}(n, m).(\text{Diag}\{e_i \mid \sigma_i(X) = \sigma_k(X)\}) \\ &= \text{conv } O(n, m)^X.(\text{Diag}\{e_i \mid \sigma_i(X) = \sigma_k(X)\}) \\ &= \text{conv}\{v_i w_i^T \mid \sigma_i(X) = \sigma_k(X), (V, W) \in O(n, m)^X\}. \end{aligned}$$

Suppose now $\sigma_k(X) = 0$. If, as above, $(V, W) \in O(n, m)^X$ then the only restrictions on v_k and w_k are: $\|v_k\| = \|w_k\| = 1, X^T v_k = X w_k = 0$. Thus

$$\begin{aligned} \partial^c \sigma_k(X) &= (U_n, U_m)\text{conv } \tilde{O}(n, m).(\text{Diag}\{\pm e_i \mid \sigma_i(X) = \sigma_k(X)\}) \\ &= \text{conv } O(n, m)^X.(\text{Diag}\{\pm e_i \mid \sigma_i(X) = 0\}) \\ &= \text{conv}\{\pm v_i w_i^T \mid \sigma_i(X) = 0, (V, W) \in O(n, m)^X\}. \end{aligned}$$

The stated formula now follows. \square

A formula for the regular subdifferential of a singular value can also easily be obtained using Proposition 6.1 and the considerations above.

COROLLARY 6.5. *The Clarke subdifferential of the k th singular value σ_k at 0 is given by*

$$\begin{aligned}\partial^c \sigma_k(0) &= \text{conv}\{vw^T \mid v \in \mathbb{R}^n, w \in \mathbb{R}^m, \|v\| = \|w\| = 1\} \\ &= \left\{ Y \in M_{n,m} \mid \sum_{i=1}^n \sigma_i(Y) \leq 1 \right\},\end{aligned}$$

whereas the (limiting) subdifferential at 0 is given by

$$\begin{aligned}\partial \sigma_k(0) &= \{Y \in \partial^c \sigma_k(0) \mid \text{rank } Y \leq n - k + 1\} \\ &= \left\{ Y \in M_{n,m} \mid \sum_{i=1}^n \sigma_i(Y) = 1 \text{ and } \text{rank } Y \leq n - k + 1 \right\}.\end{aligned}$$

Proof. It is clear from the previous corollary that

$$\partial^c \sigma_k(0) = \text{conv}\{vw^T \mid v \in \mathbb{R}^n, w \in \mathbb{R}^m, \|v\| = \|w\| = 1\}.$$

The equivalence with the second expression (which is just the unit ball for the Schatten 1-norm) is an easy exercise, and well-known. \square

7. Lidskii's Theorem for Weak Majorization – via Nonsmooth Analysis

This section parallels and extends the techniques in [8] where the original form of Lidskii's theorem, about the vector of eigenvalues of perturbed symmetric matrices, was proved using tools from nonsmooth analysis.

The form of Lidskii's theorem (for weak majorization) in which we are interested here states (see [4, Theorem 3.4.5]) that any matrices X and Y in $M_{n,m}$ satisfy

$$|\sigma(X + Y) - \sigma(X)| \prec_w \sigma(Y).$$

The symbol \prec_w denotes *weak majorization*: for two vectors x and y in \mathbb{R}^n we say that y weakly majorizes x , and write $x \prec_w y$ if $\sum_{i=1}^k \bar{x}_i \leq \sum_{i=1}^k \bar{y}_i$ for $k = 1, 2, \dots, n$. Clearly $x \prec_w y$ if and only if $P_1 x \prec_w P_2 y$ (for any permutation matrices P_1 and P_2).

In this section we show how this form of Lidskii's theorem can be easily derived from the results obtained in the paper. We need an equivalent characterization of weak majorization.

LEMMA 7.1. *Let x and y be any two vectors in \mathbb{R}^n , then the following conditions are equivalent*

- (i) $|x| \prec_w |y|$;
- (ii) $x \in \text{conv}(P_{(-)}(n)y)$;
- (iii) for every vector w in \mathbb{R}^n we have $w^T x \leq \hat{w}^T \hat{y}$.

Proof. The equivalence of (i) and (ii) is the content of [11, Theorem 1.2]. Suppose now (ii) holds, then for all w in \mathbb{R}^n ,

$$w^T x \leq \max_{P_{(-)} \in P_{(-)}(n)} (w^T P_{(-)} y) = \hat{w}^T \hat{y}.$$

If (iii) holds but $x \notin \text{conv}(P_{(-)}(n)y)$, then there is a separating hyperplane, that is, there is a vector z in \mathbb{R}^n such that

$$z^T x > \max_{P_{(-)} \in P_{(-)}(n)} (z^T P_{(-)} y) = \hat{z}^T \hat{y},$$

a contradiction. □

Fix w in \mathbb{R}^n and consider the absolutely symmetric function defined by

$$f(x) = w^T \hat{x}. \tag{14}$$

The function f is clearly Lipschitz. If x has coordinates with distinct absolute values, then f is differentiable at x and $\nabla f(x) = P_{(-)} w$ for some $P_{(-)} \in P_{(-)}(n)$. The set of all such vectors x (whose entries have distinct absolute values) has a complement in \mathbb{R}^n with measure zero. On the other hand we have the following theorem (see [2, Theorem 2.5.1]).

THEOREM 7.2 (Intrinsic Clarke Subdifferential). *Let the function f be Lipschitz near x , and suppose S is any set of Lebesgue measure 0 in \mathbb{R}^n . Then*

$$\partial^c f(x) = \text{conv}\{\lim \nabla f(x_i) \mid x_i \rightarrow x, x_i \notin S\}.$$

(It is well known that if f is Lipschitz in a neighbourhood of x then f is differentiable almost everywhere in that neighbourhood.)

From this theorem we get that the function defined in (14) satisfies

$$\partial^c f(x) \subset \text{conv}(P_{(-)}(n)w).$$

We need another theorem [2, Theorem 2.3.7].

THEOREM 7.3 (Mean-Value Theorem). *Let x and y be vectors in \mathbb{R}^n , and suppose that f is Lipschitz on an open set containing the line segment $[x, y]$. Then there exists a point u in (x, y) such that*

$$f(x) - f(y) \in \langle \partial^c f(u), x - y \rangle.$$

We have that $w^T \sigma(\cdot) = (f \circ \sigma)(\cdot)$ is Lipschitz, so there is a matrix U in $M_{n,m}$, between the matrices X and $X + Y$, and a matrix T in $\partial^c(w^T \sigma)(U)$ such that:

$$w^T(\sigma(X + Y) - \sigma(X)) = \text{tr}(T^T Y) \leq \sigma(T)^T \sigma(Y),$$

where the last inequality is von Neumann's Trace Theorem. On the other hand applying formula (3) and the above inclusion we get

$$\sigma(T) \in \text{conv}(P_{(-)}(n)w).$$

Consequently $\sigma(T)^T \sigma(Y) \leq \hat{w}^T \sigma(Y)$. We have thus shown that for every vector w in \mathbb{R}^n we have

$$w^T(\sigma(X + Y) - \sigma(X)) \leq \hat{w}^T \sigma(Y).$$

Lidskii's theorem follows from Lemma 7.1.

An independent work by Tam and Hill, covering this version of Lidskii's theorem, appeared in [16]. Their considerations are in the context of semisimple Lie group theory. Our direct and simpler approach first appeared in the second author's thesis [15].

References

1. Borwein, J. M. and Lewis, A. S.: *Convex Analysis and Nonlinear Optimization*, Springer, New York, 2000.
2. Clarke, F. H.: *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.
3. Horn, R. A. and Johnson, C. R.: *Matrix Analysis*, 2nd edn, Cambridge University Press, 1985.
4. Horn, R. A. and Johnson, C. R.: *Topics in Matrix Analysis*, Cambridge University Press, 1991. Paperback edition with corrections, 1994.
5. Ioffe, A. D.: Approximate subdifferentials and applications. I: The finite dimensional theory, *Trans. Amer. Math. Soc.* **281** (1984), 389–416.
6. Lewis, A. S.: Convex analysis on the Hermitian matrices, *SIAM J. Optim.* **6** (1996), 164–177.
7. Lewis, A. S.: Group invariance and convex matrix analysis, *SIAM J. Matrix Anal.* **17**(4) (1996), 927–949.
8. Lewis, A. S.: Lidskii's theorem via nonsmooth analysis, *SIAM J. Matrix Anal. Appl.* **21** (1999), 379–381.
9. Lewis, A. S.: Nonsmooth analysis of eigenvalues, *Math. Programming* **84** (1999), 1–24.
10. Lewis, A. S. and Sendov, H. S.: Nonsmooth analysis of singular values, Part I: Theory, *Set-Valued Anal.* (2005), 213–241.
11. Markus, A. S.: The eigen- and singular values of the sum and product of linear operators, *Uspekhi Mat. Nauk* **19**(4) (1964), 93–123. *Russian Math. Surveys* **19** (1964), 92–120.
12. Mordukhovich, B. S.: *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988.
13. Rockafellar, R. T.: *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
14. Rockafellar, R. T. and Wets, R. J.-B.: *Variational Analysis*, Springer-Verlag, Berlin, 1998.
15. Sendov, H. S.: *Variational Spectral Analysis*, University of Waterloo, PhD Thesis, 2000, <http://etd.uwaterloo.ca/etd/hssendov2000.pdf>.
16. Tam, T.-Y. and Hill, W. C.: Derivatives of orbital functions, an extension of Berezin–Gel'fand's theorem and applications, Preprint, <http://web6.duc.auburn.edu/~tamtiny/gb2.pdf>.
17. Torik, M.: Second-order directional derivatives of all eigenvalues of a symmetric matrix, *Nonlinear Anal., Ser. A Theory, Methods* **46**(8) (2001).
18. von Neumann, J.: Some matrix inequalities and metrization of matrix-space, *Tomsk University Review* **1** (1937), 286–300. In: *Collected Works*, Vol. IV, Pergamon, Oxford, 1962, pp. 205–218.