Nonsmooth analysis of singular values. Part I: Theory

Adrian S. Lewis*and Hristo S. Sendov†

January 25, 2005

Abstract

The singular values of a rectangular matrix are nonsmooth functions of its entries. In this work we study the nonsmooth analysis of functions of singular values. In particular we give simple formulae for the regular subdifferential, the limiting subdifferential, and the horizon subdifferential, of such functions. Along the way to the main result we give several applications and in particular derive von Neumann's trace inequality for singular values.

Key words and phrases: nonsmooth analysis, singular values, regular subdifferential, limiting subdifferential, horizon subdifferential, von Neumann trace inequality, simultaneous diagonalization.

AMS 1991 Subject Classification.
Primary 90C31, 15A18;
Secondary 49K40, 26B05.

1 Introduction

The singular values of a rectangular matrix are natural analogues of the eigenvalues of a square matrix. In this work we are interested in the first-order

*Department of Combinatorics & Optimization, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. Email: aslewis@math.uwaterloo.ca. Research supported by NSERC.

†Department of Mathematics, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada. Email: hssendov@cecm.sfu.ca. Research supported by NSERC.

behaviour of functions of the singular values of a rectangular matrix variable. The singular values, like the eigenvalues, are not smooth functions of the entries of the matrix. Hence, in order to gain insight into their behaviour we resort to "generalized gradients" (which we refer to as "subdifferentials"). Clarke introduced the notion of generalized gradient in [2] and [3]; thorough accounts of more recent developments may be found in [4] and [16].

The main result of this work gives formulae for the regular subdifferential, limiting subdifferential, and horizon subdifferential of singular value functions. Those are the composition of a symmetric and sign invariant function with the singular values of a rectangular matrix: $f \circ \sigma$. A nonsmooth analyst may propose to approach the problem of characterizing the subdifferential of $f \circ \sigma$ using the nonsmooth chain rule. A matrix analyst may notice that every singular value is the difference of two convex functions and the subdifferentials of the latter are easier to describe. Since we are interested in the more general question about functions of the singular values, both approaches will end up using the nonsmooth chain rule which has the form

$$\partial(f \circ \sigma)(X) \subset \cup\{\partial(y^T\sigma)(X) \,|\, y \in \partial f(\sigma(X))\}.$$

There are two potential problems with this formula. First it is an inclusion and the conditions for equality will unnecessary restrict our generality. Second, even in the cases when we can establish equality it is not clear whether and how the union of the sets on the right hand side can be transformed into the simple formula resulting from our approach.

We follow the terminology and notation of [16]. The paper closely imitates and extends the development in [13]. There are obvious patterns between the notation, techniques, and results there and here which suggest that there is a general theoretic framework that encompasses them all. One possible unifying path increasingly receiving attention lately uses properties of semisimple Lie groups and their associated Lie algebras (see [14], [19], [18]).

The results described here were first investigated in the second author's dissertation [17]. In Part II of this paper we extend the results to obtain analogous formulae for the proximal subdifferential and Clarke subdifferential when the function is either locally Lipschitz or just lower semicontinuous. We use them to calculate the subdifferentials of individual singular values. Another application gives a nonsmooth proof of Lidskii's theorem for weak majorization of singular values.

2 The limiting subdifferential

Definition 2.1 (Regular Subgradient) Given a Euclidean space E (by which we mean, a finite-dimensional real inner-product space), a function $f : E \to [-\infty, +\infty]$, and a point x in E at which f is finite, an element y of E is a *regular subgradient* of f at x if it satisfies

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

As usual, $o(\cdot)$ denotes a real-valued function defined on a neighbourhood of the origin in E, and satisfying $\lim_{z \to 0} \|z\|^{-1} o(z) = 0$. The set of regular subgradients is denoted $\hat{\partial} f(x)$ and is called the *regular subdifferential*. It is easy to show that it is always closed and convex.

This definition is just one-sided version of the classical (Fréchet) derivative. A weakness of this natural concept of subdifferential is that even for well-behaved functions f it may be empty as it is for example for the function $f(x) = -|x|$ at $x = 0$. The idea of the limiting subdifferential enhances regular subdifferential by gathering information from the regular subdifferentials at points near x as well.

Definition 2.2 (Limiting Subdifferential) An element y of E is a *limiting subgradient* if 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 subgradients is the *limiting subdifferential* $\partial f(x)$.

Definition 2.3 (Horizon Subgradient) An element y of E is a *horizon subgradient* if there is a sequence of points x^r in E approaching x with values $f(x^r)$ approaching the finite value $f(x)$, a sequence of reals t_r decreasing to 0, and a sequence of regular subgradients y^r in $\hat{\partial} f(x^r)$ such that $t_r y^r \to y$.

The set of horizon subgradients is denoted $\partial^\infty f(x)$. If $f(x)$ is infinite then the sets $\partial f(x)$ and $\hat{\partial} f(x)$ are defined to be empty, and $\partial^\infty f(x)$ to be $\{0\}$. The reader can verify that $\partial f(x)$ and $\hat{\partial} f(x)$ are always closed sets, and we have the inclusion $(\hat{\partial} f(x))^\infty \subset \partial^\infty f(x)$ (where C^∞ denotes the recession cone of a closed convex set). It is easy to see that $\partial^\infty f(x)$ is a cone and if f is Lipschitz around the point x then $\partial^\infty f(x) = \{0\}$.

Definition 2.4 (Subdifferential Regularity) If the function f is finite at the point x with at least one subgradient there then it is *regular* at x if it is lower semicontinuous near x, every subgradient is regular, that is $\hat{\partial} f(x) = \partial f(x)$, and furthermore

$$\partial^\infty f(x) = (\hat{\partial} f(x))^\infty.$$

This definition just says that the set epi $f = \{(x, \alpha) \mid \alpha \geq f(x)\}$ is *(Clarke) regular* at the point $(x, f(x))$: see [16, Corollary 8.11] for the justification.

Definition 2.5 (Tangent Cone) Let L be a subset of the space E, and fix a point x in E. An element d of E belongs to the *regular tangent cone* to L at x, written $T_L(x)$, if

$$\frac{x^r - x}{t_r} \to d,$$

for some sequence x^r in L approaching x and a sequence t_r decreasing to 0.

Definition 2.6 (Negative Polar Cone) The *(negative) polar* of a subset H of E is the set

$$H^- = \{y \in E : \langle x, y \rangle \leq 0 \ \ \forall x \in H\}.$$

The proof of the following easy and standard result can be found in [13, Proposition 1].

Proposition 2.7 (Normal Cone) *Given a function $f : E \to [-\infty, +\infty]$ and a point x in E, any regular subgradient of f at x is polar to the tangent cone of the level set $\mathcal{L} = \{z \in E : f(z) \leq f(x)\}$ at x; that is*

$$\hat{\partial} f(x) \subset (T_{\mathcal{L}}(x))^-.$$

In this paper we are interested in functions that are invariant under certain orthogonal transformations of the space E. A linear transformation g on the space E is *orthogonal* if it preserves the inner product:

$$\langle gx, gy \rangle = \langle x, y \rangle \text{ for all elements } x \text{ and } y \text{ of } E.$$

Such linear transformations form the *orthogonal* group $O(E)$. A function f on E is *invariant* under a subgroup G of $O(E)$ if $f(gx) = f(x)$ for all points x in E and transformations g in G.

In the following proposition, $f'(\cdot;\cdot)$ denotes the usual directional derivative:

$$f'(x;z) = \lim_{t \downarrow 0} \frac{f(x+tz) - f(x)}{t}, \text{ (when well-defined)}$$

for elements x and z of E.

The following needed result is Proposition 2 in [13].

Proposition 2.8 (Subgradient Invariance) *If the function $f : E \to [-\infty, +\infty]$ 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 results hold for regular, horizon, and (if f is Lipschitz around x) Clarke subgradients, and f is regular at the point gx if and only if it is regular at x. Furthermore, for any element z of E, the directional derivative $f'(gx; gz)$ exists if and only if $f'(x; z)$ does, and in this case the two are equal.*

This section ends with a lemma which is useful in the later analysis of regularity. (See [13, Lemma 1].)

Lemma 2.9 (Recession) *For any nonempty closed convex subset C of E, closed subgroup H of $O(E)$, and transformation g in $O(E)$, the set gHC is closed, and if it is also convex then its recession cone is $gH(C^\infty)$.*

3 The normal space

Throughout the whole paper 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 $\langle X, Y \rangle = \operatorname{tr} X^T Y$. It is easily seen that analogous results to those we present in this work hold for the space of $n \times m$ complex matrices with the inner product $\langle X, Y \rangle = \operatorname{Re}(\operatorname{tr} X^* Y)$, 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.

The main goal of this section is to give a straightforward proof that for a fixed $X \in M_{n,m}$ the set

$$\{U_n^T X U_m \,|\, U_n, U_m \text{ – orthogonal } \}$$

is a smooth manifold and to characterize its tangent and normal spaces at every point. To do this precisely we need a little of differential geometry and the results stated below will be needed only in this section.

If M is a smooth manifold and $m \in M$, then $T_M(m)$ will denote the tangent space to M at the point m. The next three results are respectively Proposition 4.5.1, Proposition 12.9.4, Proposition 13.3.1, and Proposition 13.3.2 in [1].

Lemma 3.1 (Manifold Sum) *Let M and M' be smooth manifolds, and let p, p' denote the projections of $M \times M'$ onto M, M' respectively then the function*

$$\lambda : T_{M\times M'}(a, a') \mapsto T_M(a) \oplus T_{M'}(a')$$

defined by $w \mapsto (dp, dp')w$ is a linear isomorphism.

Theorem 3.2 (Quotient Manifold) *If H is a closed subgroup of a Lie group G then either H is open in G (and the quotient set topology on G/H is discrete) or the quotient G/H admits a differentiable structure such that the natural surjection*

$$\begin{aligned} \pi : G &\to G/H \\ g &\mapsto gH \end{aligned}$$

has rank equal to the dimension of G/H at every point; that is, the linear map $d\pi$ between the tangent spaces is onto.

All quotient manifolds below have the differential structure described in Theorem 3.2.

Theorem 3.3 (Orbit Submanifold) *Suppose G is a Lie group that also acts on the Hausdorff manifold M and satisfies the natural conditions*

$$\begin{aligned} G \times M &\to M \\ (g, m) &\mapsto gm \end{aligned}$$

is differentiable and $g_1(g_2m) = (g_1g_2)m$ for all $g_1, g_2 \in G$ and $m \in M$. If the stabilizer G_m is not an open subgroup of G, then the mapping

$$\begin{aligned} \phi_m : G/G_m &\to M, \text{ defined by} \\ g(G_m) &\mapsto gm, \text{ for } g \text{ in } G, \end{aligned}$$

is is one-to-one and has rank equal to the dimension of G/G_m at every point. Moreover, the orbit Gm in M can be given the structure of a submanifold of M diffeomorphic to G/G_m under ϕ_m.

Let $O(n)$ be the Lie group of $n \times n$ real orthogonal matrices, and let $O(n,m)$ denote the Cartesian product $O(n) \times O(m)$, which is also a Lie group. An easy calculation shows that the tangent space to $O(n)$ at the identity matrix I, is just the subspace of skew-symmetric matrices, $A(n)$. Consequently from Lemma 3.1 we see that $T_{O(n,m)}(I_n, I_m) = A(n) \times A(m)$.

Consider the action of the group $O(n,m)$ on the space $M_{n,m}$ defined by

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

For a fixed matrix X in $M_{n,m}$, the orbit

$$O(n,m).X = \{U_n^T X U_m \,:\, (U_n, U_m) \in O(n,m)\}$$

is just the set of $n \times m$ matrices with the same singular values as X. Here is then the key fact.

Theorem 3.4 (Normal Space) *The orbit O(n,m).X is a submanifold of the space $M_{n,m}$, with tangent space*

$$(1) \qquad T_{O(n,m).X}(X) = \{XZ_m - Z_n X \,:\, Z_n \in A(n) \text{ and } Z_m \in A(m)\}$$

and normal space

$$(2) \qquad (T_{O(n,m).X}(X))^{\perp} = \{Y \in M_{n,m} \,:\, X^T Y \text{ and } XY^T \text{ symmetric}\}.$$

Proof. **Part I. The tangent space.** Consider the stabilizer

$$O(n,m)_X = \{(U_n, U_m) \in O(n,m) : U_n^T X U_m = X\}$$

and the bijection ϕ between the sets $O(n,m)/O(n,m)_X$ and $O(n,m).X$ defined by:

$$(U_n, U_m)(O(n,m)_X) \mapsto U_n^T X U_m, \text{ for } (U_n, U_m) \text{ in } O(n,m).$$

Clearly $O(n,m)_X$ is a closed subgroup of $O(n,m)$ (it is closed under limit operations). So from Theorem 3.3 it follows that the map ϕ is a diffeomorphism,

and hence its differential $d\phi$ is an isomorphism between the corresponding tangent spaces

$$T_{O(n,m)/O(n,m)_X}((I_n, I_m)O(n,m)_X) \text{ and } T_{O(n,m).X}(X)$$

Consider, on the other hand, the quotient map

$$\begin{aligned}\pi :& O(n,m) \to O(n,m)/O(n,m)_X, \text{ defined by} \\ & (U_n, U_m) \mapsto (U_n, U_m)(O(n,m)_X), \text{ for all } (U_n, U_m) \text{ in } O(n,m).\end{aligned}$$

Theorem 3.2 implies that its differential

$$d\pi : T_{O(n,m)}(I_n, I_m) \to T_{O(n,m)/O(n,m)_X}((I_n, I_m)O(n,m)_X)$$

is onto. Now consider a third map

$$\begin{aligned}\psi :& O(n,m) \to O(n,m).X, \text{ defined by} \\ & (U_n, U_m) \mapsto U_n^T X U_m, \text{ for all } (U_n, U_m) \text{ in } O(n,m).\end{aligned}$$

Since $\psi = \phi \circ \pi$, the chain rule gives $d\psi = d\phi \circ d\pi$, that is

$$(d\psi)T_{O(n,m)}(I_n, I_m) = T_{O(n,m).X}(X).$$

But as we noted above $T_{O(n,m)}(I_n, I_m) = A(n) \times A(m)$. Now we show that $(d\psi)(Z_n, Z_m) = XZ_m - Z_nX$. Define the map

$$\begin{aligned}&\Phi : M_n \times M_m \to M_{n,m} \\ &\Phi(U, V) = U^T X V,\end{aligned}$$

where M_n, M_m, and $M_{n,m}$ have their standard differential structure. Let $d\Phi$ be its differential at (I_n, I_m). Then because $T_{M_n}(M) = M_n$ for each $M \in M_n$ it is easy to see that

$$\begin{aligned}&d\Phi : M_n \times M_m \to M_{n,m} \\ &d\Phi(U, V) = U^T X + XV.\end{aligned}$$

We have that $O(n) \times O(m)$ is a submanifold of $M_n \times M_m$, so the tangent space $T_{O(n)\times O(m)}(I_n, I_m)$ is isomorphic to a vector subspace of $T_{M_n \times M_m}(I_n, I_m)$. Also the end of Theorem 3.3 implies that the tangent space $T_{O(n,m).X}(X)$

is isomorphic to a vector subspace of $T_{M_{n,m}}(X)$. Let i be the natural injection of $O(n) \times O(m)$ into $M_n \times M_m$, and let j be the natural injection of $O(n,m).X$ into $M_{n,m}$. Then from the definitions $j \circ \psi = \Phi \circ i$. So $dj \circ d\psi = d\Phi \circ di$, but $(di)(Z_n, Z_m) = (Z_n, Z_m)$ for each (Z_n, Z_m) in $A(n) \times A(m)$, and dj is the identity on $T_{O(n,m).X}(X)$. Thus, we obtain $(d\psi)(Z_n, Z_m) = (d\Phi)(Z_n, Z_m) = Z_n^T X + X Z_m = X Z_m - Z_n X$, as we claimed.

Part II. The normal space. If a matrix Y in $M_{n,m}$ satisfies $X^T Y = Y^T X$, and $XY^T = YX^T$, then for any matrices $Z_n \in A(n)$, and $Z_m \in A(m)$ we have

$$\begin{aligned}\langle Y, XZ_m - Z_n X\rangle &= \operatorname{tr} Y^T(XZ_m - Z_n X)\\ &= \operatorname{tr} Y^T X Z_m - \operatorname{tr} Y^T Z_n X\\ &= \operatorname{tr} Y^T X Z_m - \operatorname{tr} XY^T Z_n.\end{aligned}$$

We will show now that $\operatorname{tr} Y^T X Z_m$=0. Indeed,

$$\operatorname{tr} Y^T X Z_m = \operatorname{tr} Z_m^T X^T Y = -\operatorname{tr} Z_m X^T Y = -\operatorname{tr} Z_m Y^T X = -\operatorname{tr} Y^T X Z_m.$$

Analogously we get $\operatorname{tr} XY^T Z_n = 0$, consequently $Y \in (T_X O(n,m).X)^{\perp}$.

Conversely suppose that $\operatorname{tr} Y^T(XZ_m - Z_n X) = 0$ for all $Z_n \in A(n)$ and $Z_m \in A(m)$. For each $Z_n \in A(n)$ we have

$$\operatorname{tr} Y^T Z_n X = \operatorname{tr} XY^T Z_n = \operatorname{tr} (XY^T Z_n)^T = \operatorname{tr} Z_n^T Y X^T = -\operatorname{tr} Z_n Y X^T,$$

that is

$$\operatorname{tr} XY^T Z_n = -\operatorname{tr} Z_n Y X^T.$$

Let $Z_m = 0$. Then our assumption becomes $\operatorname{tr} XY^T Z_n = 0$ and consequently we have $\operatorname{tr} Z_n Y X^T = 0$ and so is their difference:

$$\operatorname{tr} (XY^T Z_n - Z_n Y X^T) = 0.$$

Choosing $Z_n = XY^T - YX^T$ gives

$$\begin{aligned}0 &= \operatorname{tr}\big(XY^T(XY^T - YX^T) - (XY^T - YX^T)YX^T\big)\\ &= \operatorname{tr}\big(XY^T(XY^T - YX^T)\big) - \operatorname{tr}\big(YX^T(XY^T - YX^T)\big)\\ &= \operatorname{tr}(XY^T - YX^T)(XY^T - YX^T) = -\operatorname{tr}(XY^T - YX^T)^T(XY^T - YX^T),\end{aligned}$$

whence $XY^T = YX^T$. Analogously by choosing first $Z_n = 0$ and then $Z_m = Y^T X - X^T Y$ we obtain $X^T Y = Y^T X$. ■

Throughout the entire paper all vectors are considered to be column vectors unless stated otherwise. We denote the cone of vectors x in $\mathbb{R}^n$ satisfying $x_1 \geq x_2 \geq ... \geq x_n$ by $\mathbb{R}^n_{\downarrow}$. We denote the standard basis in $\mathbb{R}^n$ by $e^1, e^2, ..., e^n$. For any vector x in $\mathbb{R}^n$ we denote by $\bar{x}$ the vector with the same entries as x ordered in nonincreasing order. Let $P(n)$ denote the set of all $n \times n$ permutation matrices. (Those matrices that have only one nonzero entry in every row or column, which is 1.) Let $P_{(-)}(n)$ denote the set of all $n \times n$ signed permutation matrices. (Those matrices that have only one nonzero entry in every row or column, which is ± 1.) If $P_{(-)} \in P_{(-)}(n)$ then we will denote by $|P_{(-)}|$ the permutation matrix obtained from $P_{(-)}$ by taking the absolute values of its entries. If x is a vector in $\mathbb{R}^n$ then $|x|$ will denote the vector $(|x_1|, |x_2|, ..., |x_n|)^T$ and x^2 will denote the vector $(x_1^2, ..., x_n^2)^T$. Finally if $x, y \in \mathbb{R}^n$ then $x \cdot y = (x_1 y_1, ..., x_n y_n)$. We will need the following standard lemma in our proofs (see [10]).

Lemma 3.5 *Any vectors x and y in $\mathbb{R}^n$ we have the inequality*

$$x^T y \leq \bar{x}^T \bar{y}.$$

Equality holds if and only if some matrix Q in $P(n)$ satisfies $Qx = \bar{x}$ and $Qy = \bar{y}$.

4 Singular Values

Analogously to the eigenvalue decomposition of a symmetric matrix via an orthogonal transformation, any rectangular matrix can also be diagonalized via an orthogonal transformation on $M_{n,m}$. We state the precise result below. For the proof the reader may refer either to [7, Theorem 7.3.5] or to [8, Theorem 3.1.1].

For any matrix X, with $X^{i,j}$ we denote its (i,j)-th entry. For any vector x in $\mathbb{R}^n$ let $\operatorname{Diag} x$ denote the matrix with entries $(\operatorname{Diag} x)^{i,i} = x_i$ for all i, and $(\operatorname{Diag} x)^{i,j} = 0$ for $i \neq j$. We want to turn the readers attention to the fact that sometimes $\operatorname{Diag} x$ will denote an $n \times m$ matrix, sometimes $n \times n$ and sometimes $m \times m$ (this in case $x \in \mathbb{R}^m$), but there will be no confusion because the context will make clear which is the case.

Theorem 4.1 (Singular Value Decomposition) *Let $X \in M_{n,m}$ ($n \leq m$). There are positive real numbers $\sigma(X) := (\sigma_1(X), \sigma_2(X), ..., \sigma_n(X))^T$*

in nonincreasing order $\sigma_1(X) \geq \sigma_2(X) \geq ... \geq \sigma_n(X)$*, and square orthogonal (unitary if* X *is complex) matrices* U_n *and* U_m *such that*

$$X = U_n^T \big(\text{Diag}\, \sigma(X)\big) U_m.$$

The entries of the vector $\sigma(X) = (\sigma_1(X), \sigma_2(X), ..., \sigma_n(X))^T$ are called the singular values of X. The numbers $\{\sigma_1(X), \sigma_2(X), ..., \sigma_n(X)\}$ are the nonnegative square roots of the eigenvalues of XX^T and thus are uniquely determined. For convenience and without loss of generality we have assumed that they are ordered nonincreasingly.

Definition 4.2 We say that two matrices X and Y in $M_{n,m}$ have a *simultaneous ordered singular value decomposition* if there is an element (U_n, U_m) in $O(n, m)$ such that $X = U_n^T(\text{Diag}\, \sigma(X))U_m$ and $Y = U_n^T(\text{Diag}\, \sigma(Y))U_m$.

We need to introduce more notation that will be used in the proof of the next lemma. Let M be a matrix in $M_{n,m}$, and $1 \leq i_1 < i_2 < ... < i_r \leq n$, $1 \leq j_1 < j_2 < ... < j_s \leq m$ be given numbers. Then $M(i_1, i_2, ..., i_r;\, j_1, j_2, ..., j_s)$ will denote the minor of M (with dimensions $r \times s$) obtained at the intersection of the rows with indexes $i_1, i_2, ..., i_r$, and columns with indexes $j_1, j_2, ..., j_s$. If v is a vector in $\mathbb{R}^n$ then we will use similar notation to denote a subvector of v. That is, a subvector of v formed by the entries with indexes $1 \leq i_1 < i_2 < ... < i_r \leq n$ will be denoted by $v(i_1, i_2, ..., i_r)$. Finally $M(i; \cdot)$ will denote the row of M with index i (these are row vectors), and $M(\cdot; i)$ will denote the column of M with index i. The following lemma gives a necessary and sufficient condition for two matrices to "almost" have a simultaneous ordered singular value decomposition. For a necessary and sufficient condition for simultaneous ordered singular value decomposition see Theorem 4.6.

Lemma 4.3 *Two matrices* Y *and* Z *in* $M_{n,m}$ *satisfy* $Z^TY = Y^TZ$ *and* $ZY^T = YZ^T$ *if and only if there exists an element* (U_n, U_m) *in* $O(n, m)$ *and a signed permutation matrix* $P_{(-)}$ *in* $P_{(-)}(n)$ *such that*

$$Y = U_n^T(\text{Diag}\, P_{(-)}\sigma(Y))U_m, \quad Z = U_n^T(\text{Diag}\, \sigma(Z))U_m. \tag{3}$$

Before we prove the result we need to comment on it.

Remark 4.4 There are four interesting variations of this kind of problem that appear in the literature. Given a set of complex rectangular matrices one

may ask when they can be simultaneously diagonalized with unitary matrices (U_n, U_m) where the resulting diagonal matrices are allowed to have complex entries, and a second variation asks when the same diagonalization can be performed with a pair of orthogonal matrices. For both those questions we refer the reader to Theorem 1 and Theorem 4 in [5]. The third form of the problem is the one we need and formulated above: when a set of real rectangular matrices can be simultaneously diagonalized with orthogonal pair (U_n, U_m) (see [20] who also credits the original result to Wiegmann [22]). The proof we present here is somewhat different and the reduction steps in it make the main idea quite transparent. The final fourth variation asks when the set of matrices have simultaneous (ordered) singular value decomposition. (That is, the resulting diagonal matrices have real, nonnegative (ordered) diagonal.) A necessary and sufficient condition for the fourth problem is given by von Neumann in [21]. We address that question below in Theorem 4.6 by giving a variational proof of this result.

Proof. In one direction the lemma is clear. In the other direction, suppose first that $n = m$ and Y and Z are nonsingular. We will divide the proof into several reduction stages. It is well known that the eigenvalues of $Y^T Z$ are the same as the eigenvalues of ZY^T counting multiplicities. Then because they are both symmetric, there are two orthogonal matrices A and B in $O(n)$ such that $Y^T Z = A^T \Lambda A$ and $ZY^T = B^T \Lambda B$. Consequently $Y^T Z = (A^T B)(ZY^T)(B^T A)$. We make the substitution: $\breve{Y} = (A^T B)Y$ and $\breve{Z} = (A^T B)Z$. Then we have

$$\breve{Y}^T \breve{Z} = Y^T Z = (A^T B)(ZY^T)(B^T A) = \breve{Z}\breve{Y}^T,$$

that is $\breve{Y}^T$ and $\breve{Z}$ commute. Hence $\breve{Y}$ and $\breve{Z}^T$ commute as well. Next, because $\breve{Y}^T$ and $\breve{Z}$ commute with the symmetric matrix $\breve{Y}^T \breve{Z}$ it follows that every eigenspace of $\breve{Y}^T \breve{Z}$ is invariant under $\breve{Y}^T$ and $\breve{Z}$. Thus if V_n is an orthogonal matrix in $O(n)$, whose columns are eigenvectors of $\breve{Y}^T \breve{Z}$ so that all eigenvectors corresponding to the same eigenvalues occur one after another, then both $V_n^T \breve{Y}^T V_n$ and $V_n^T \breve{Z} V_n$ must be block diagonal (recall that eigenvectors corresponding to different eigenvalues are orthogonal):

$$V_n^T \breve{Y}^T V_n = \mathrm{Diag}\,(\breve{Y}_1^T, \breve{Y}_2^T, ..., \breve{Y}_l^T), \quad V_n^T \breve{Z} V_n = \mathrm{Diag}\,(\breve{Z}_1, \breve{Z}_2, ..., \breve{Z}_l),$$

where $\breve{Y}_i^T$,$\breve{Z}_i \in M_{n_i}$, $1 \le n_i \le n$, $n_1 + n_2 + \cdots n_l = n$, and each $\breve{Y}_i^T \breve{Z}_i = \breve{Z}_i \breve{Y}_i^T = \lambda_i I_{n_i}$, where $\lambda_1, \lambda_2, ..., \lambda_l$ are the distinct (all of them are nonzero)

eigenvalues of the symmetric matrix $\breve{Y}^T\breve{Z}$. For each i choose a singular value decomposition $\breve{Z}_i = R_i^T D_i S_i$ (R_i, S_i - orthogonal, D_i - diagonal), and observe $\breve{Y}_i^T = S_i^T(\lambda_i D_i^{-1})R_i$. Note that the absolute values of the diagonal entries of $\lambda_i D_i^{-1}$ are the singular values of $\breve{Y}_i^T$. So we reduced Y and Z to l pairs of matrices $\breve{Y}_i$ and $\breve{Z}_i$ that satisfy (3). Clearly the singular values of Z are the same as the singular values of $\breve{Z}$ and are the union of diagonal entries of $D_1, ..., D_l$. Let P be a permutation matrix in $P(n)$ such that $\mathrm{Diag}\,\sigma(Z) = P^T\mathrm{Diag}\,(D_1, ..., D_l)P$. Then retracing back the reductions one sees that the lemma holds in the case when $n = m$ and the matrices Y, Z are nonsingular. In fact, decomposition (3) holds with

$$U_n^T = B^T A V_n \mathrm{Diag}\left(R_1^T, ..., R_l^T\right)P, \quad U_m = P^T\big(\mathrm{Diag}\,(S_1, ..., S_l)\big)V_n^T.$$

We now consider the general case $n \leq m$. First we observe that the symmetric matrices Y^TY and Z^TZ commute. Indeed

$$\begin{aligned}(Z^TZ)(Y^TY) &= Z^T(YZ^T)Y = (Z^TY)(Z^TY)\\ &= (Y^TZ)(Y^TZ) = Y^T(ZY^T)Z = (Y^TY)(Z^TZ).\end{aligned}$$

Analogously one sees that the pair of symmetric matrices YY^T and ZZ^T also commute. It is well known that the eigenvalues of Y^TY are the same as the eigenvalues of YY^T plus $m-n$ additional zeros. Hence there is a matrix V_m in $O(m)$ and a matrix V_n in $O(n)$ that simultaneously diagonalize the above two pairs respectively (Recall that for any matrix Y, the eigenvalues of YY^T are the singular values of Y squared and similarly for Y^T.):

$$V_n^T(YY^T)V_n = \mathrm{Diag}\,\sigma^2(Y), \quad V_m^T(Y^TY)V_m = \mathrm{Diag}\,(\sigma^2(Y)^T, \underbrace{0, ..., 0}_{m-n})^T,$$

$$V_n^T(ZZ^T)V_n = \mathrm{Diag}\,P_n\sigma^2(Z), \quad V_m^T(Z^TZ)V_m = \mathrm{Diag}\,P_m(\sigma^2(Z)^T, \underbrace{0, ..., 0}_{m-n})^T,$$

where P_n is a permutation matrix in $P(n)$, and P_m is in $P(m)$. Now we make the substitution:

$$\hat{Y} = V_n^TYV_m, \qquad \hat{Z} = V_n^TZV_m.$$

Observe that:

$$\hat{Y}^T\hat{Z} = V_m^TY^TZV_m = V_m^TZ^TYV_m = \hat{Z}^T\hat{Y},$$

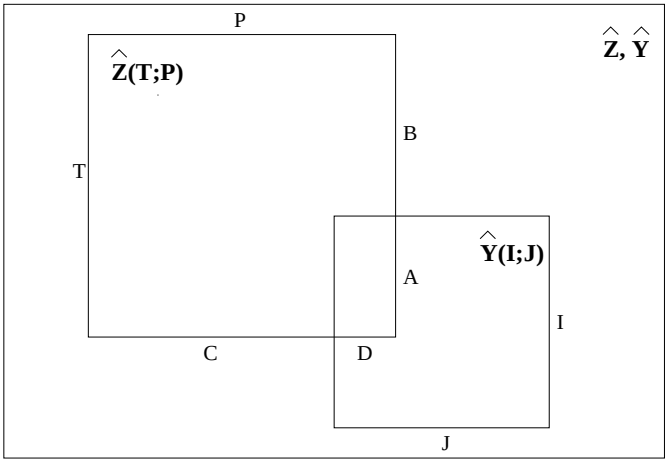

Figure 1: The sets I,J,T,P and A,B,C,D.

and similarly one checks that $\hat{Y}\hat{Z}^T = \hat{Z}\hat{Y}^T$. Moreover we have that

$$\hat{Y}\hat{Y}^T = \operatorname{Diag}\sigma^2(Y), \quad \hat{Y}^T\hat{Y} = \operatorname{Diag}\,(\sigma^2(Y)^T, \underbrace{0, ..., 0}_{m-n})^T \tag{4}$$

and

$$\hat{Z}\hat{Z}^T = \operatorname{Diag} P_n\sigma^2(Z), \quad \hat{Z}^T\hat{Z} = \operatorname{Diag} P_m(\sigma^2(Z)^T, \underbrace{0, ..., 0}_{m-n})^T. \tag{5}$$

Next, we investigate the structure of the matrices $\hat{Y}$ and $\hat{Z}$. Let the ranks of $\hat{Y}$ and $\hat{Z}$ be k and l respectively, and let $\hat{Y}(i_1, ..., i_k;\, j_1, ..., j_k)$ and $\hat{Z}(t_1, t_2, ..., t_l;\, p_1, p_2, ..., p_l)$ be nonsingular minors. Let $I = \{i_1, i_2, ..., i_k\}$, $J = \{j_1, j_2, ..., j_k\}$, $T = \{t_1, t_2, ..., t_l\}$, $P = \{p_1, p_2, ..., p_l\}$. Equation (4) tells us that the rows and the columns of $\hat{Y}$ are mutually orthogonal. If we take a row, r_i of $\hat{Y}$, such that $i \notin I$ then r_i is a linear combination of rows with indexes from the set I. Multiplying this linear combination by r_i gives that $r_i^T r_i = 0$. Similar argument for the columns imply that all the entries of $\hat{Y}$ that don't belong to the minor $\hat{Y}(i_1, ..., i_k;\, j_1, ..., j_k)$ are zero. The same arguments apply to $\hat{Z}$.

Let $A = I \cap T$, $B = T \backslash I$, $C = P \backslash J$ and $D = P \cap J$, see Figure 1. Take an index i in the set B. From the above paragraph we have that the i-th row of $\hat{Y}$ is the zero vector: $\hat{Y}(i; \cdot) = 0$. So we get $\hat{Y}(i; \cdot)\hat{Z}(x; \cdot)^T = 0$ for all $1 \leq x \leq n$.

Using the relationship $\hat{Y}\hat{Z}^T = \hat{Z}\hat{Y}^T$ we get that $\hat{Z}(i;\cdot)\hat{Y}(x;\cdot)^T = 0$ for all $1 \leq x \leq n$. So in particular the vector $\hat{Z}(i;\cdot)(J)$ (that is, the subvector of the i-th row of $\hat{Z}$ formed from the entries with indexes in J) is orthogonal to all the vectors $\hat{Y}(x;\cdot)(J)$ for all $x \in I$. But the last set of vectors form the nonsingular minor of $\hat{Y}$. So $\hat{Z}(i;\cdot)(J) = 0$. We already knew that $\hat{Z}(i;\cdot)(J\backslash D) = 0$ so what we get in addition is that $\hat{Z}(i;\cdot)(D) = 0$, and this applies for every i in B. So all the entries of the submatrix $\hat{Z}(B;D)$ of the nonsingular minor $\hat{Z}(T;P)$, are zero. Completely analogously but now choosing an index from the set C and using the relationship $\hat{Y}^T\hat{Z} = \hat{Z}^T\hat{Y}$ one sees that all the entries of the submatrix $\hat{Z}(A;C)$ of the nonsingular minor $\hat{Z}(T;P)$, are zero.

Next, we want to show that $|A| = |D|$ and $|C| = |B|$. Suppose $|C| < |B|$, so the submatrix $\hat{Z}(B;C)$ has linearly dependent rows. But then the rows of $\hat{Z}(B;P)$ are linearly dependent and this contradicts that fact that $\hat{Z}(T;P)$ is nonsingular. Suppose now $|C| > |B|$, so the columns of $\hat{Z}(B;C)$ are linearly dependent, and so will be the columns of $\hat{Z}(T;C)$ and we get again the same contradiction. So $|C| = |B|$, and because $|A| + |B| = l$ and $|C| + |D| = l$ we obtain that $|A| = |D|$ as well. In summary, we proved that the nonsingular minor of $\hat{Z}$ is block diagonal:

$$\hat{Z}(T;P) = \text{Diag}\left(\hat{Z}(B;C), \hat{Z}(A;D)\right).$$

Completely analogously we obtain the same result for $\hat{Y}$. That is the nonsingular minor of $\hat{Y}$ is block diagonal:

$$\hat{Y}(I;J) = \text{Diag}\left(\hat{Y}(A;D), \hat{Y}(I\backslash A; J\backslash D)\right).$$

Now, because $\hat{Y}\hat{Z}^T = \hat{Z}\hat{Y}^T$ and $\hat{Y}^T\hat{Z} = \hat{Z}^T\hat{Y}$ one easily sees that

$$\begin{aligned}
&\hat{Y}(A;D)\hat{Z}(A;D)^T = \hat{Z}(A;D)\hat{Y}(A;D)^T, \quad \text{and} \\
&\hat{Y}(A;D)^T\hat{Z}(A;D) = \hat{Z}(A;D)^T\hat{Y}(A;D)
\end{aligned}$$

Moreover $\hat{Y}(A;D)$, $\hat{Z}(A;D)$ are square and nonsingular. So from the first part of the proof they have simultaneous singular value decompositions as described in the lemma. Next, we find (four) orthogonal matrices that give the singular value decomposition of $\hat{Y}(I\backslash A; J\backslash D)$ and $\hat{Z}(B;C)$ and because $(I\backslash A)\cap B = \emptyset$ and $(J\backslash D)\cap C = \emptyset$ it is not difficult to see how we can obtain the singular value decomposition described in the lemma. ■

In what follows, for a vector x in $\mathbb{R}^n$, we write $\hat{x}$ for the vector in $\mathbb{R}^n$ with the same entries as $|x|$ arranged in nonincreasing order. Note that $\sigma(\mathrm{Diag}\, x) = \hat{x}$. The following lemma follows as a particular case of the more general framework in [12, Theorem 2.2, Example 7.2], we give a direct proof here.

Lemma 4.5 *For any vectors x and y in $\mathbb{R}^n$ we have the inequality*

$$x^T y \leq \hat{x}^T \hat{y} \tag{6}$$

with equality if and only if there is a signed permutation matrix $P_{(-)}$ in $P_{(-)}(n)$ such that $P_{(-)}x = \hat{x}$ and $P_{(-)}y = \hat{y}$.

Proof. It is clear that the inequality holds since

$$x^T y \leq |x|^T |y| \leq \hat{x}^T \hat{y},$$

where the last inequality follows from Lemma 3.5. The condition for equality in one direction is clear too. Now suppose we have equalities above. Because $|x|^T|y| = \hat{x}^T\hat{y}$, from Lemma 3.5, there is a permutation matrix Q in $P(n)$ such that $Q|x| = \hat{x}$ and $Q|y| = \hat{y}$.

Let I be the $n \times n$ identity matrix. The fact that we have the equality $x^T y = |x|^T|y|$ makes it possible to assign signs to the entries of the identity matrix I so that if $I_{(-)}$ is the so-formed matrix, we have $I_{(-)}x = |x|$ and $I_{(-)}y = |y|$. Indeed, for every index i, $1 \leq i \leq n$, we assign the signs as follows:
if $x_i = 0$ and $y_i = 0$ set $I^{i,i}_{(-)} = 1$;
if $x_i = 0$ and $y_i \neq 0$ set $I^{i,i}_{(-)} =$sign (y_i);
if $x_i \neq 0$ and $y_i = 0$ set $I^{i,i}_{(-)} =$sign (x_i);
if $x_i \neq 0$ and $y_i \neq 0$, in order for the equality to hold we must have sign $(x_i) =$ sign (y_i), so set $I^{i,i}_{(-)} =$ sign (x_i). We have that $QI_{(-)}x = \hat{x}$ and $QI_{(-)}y = \hat{y}$; let $P_{(-)} = QI_{(-)}$. ■

The Normal Space Theorem (3.4) will be extremely useful to us in the following sections. However we can immediately demonstrate its importance by recalling the variational proof of an inequality essentially due to von Neumann [8, p. 182]. The following theorem may also be viewed as a necessary and sufficient condition for two matrices to have a simultaneous ordered singular value decomposition.

Theorem 4.6 (Von Neumann's Trace Theorem) *Any matrices X and Y in $M_{n,m}$ satisfy the inequality* $\operatorname{tr} X^T Y \leq \sigma(X)^T \sigma(Y)$. *Equality holds if and only if X and Y have a simultaneous ordered singular value decomposition*

Proof. For fixed X and Y, consider the optimization problem

$$\alpha = \sup_{Z \in O(n,m).X} \operatorname{tr} Y^T Z. \tag{7}$$

Observe first that there is an element (U_n, U_m) in $O(n,m)$ satisfying $Y = U_n^T(\operatorname{Diag}\sigma(Y))U_m$, and then choosing $Z = U_n^T(\operatorname{Diag}\sigma(X))U_m$ shows that $\alpha \geq \sigma(X)^T\sigma(Y)$.

Next, since the orbit $O(n,m).X$ is compact, problem (7) has an optimal solution, $Z = Z_0$ say, and any such Z_0 by stationarity must satisfy

$$Y \perp T_{O(n,m).X}(Z_0) \ \ (= T_{O(n,m).Z_0}(Z_0)).$$

The Normal Space Theorem now shows that the matrices Y and Z_0 satisfy $Z_0^T Y = Y^T Z_0$ and $Z_0 Y^T = Y Z_0^T$. Then by Lemma 4.3, there is an element (U_n, U_m) in $O(n,m)$, and a signed permutation matrix $P_{(-)}$ in $P_{(-)}(n)$ such that

$$Y = U_n^T(\operatorname{Diag} P_{(-)}\sigma(Y))U_m, \quad Z_0 = U_n^T(\operatorname{Diag}\sigma(Z_0))U_m. \tag{8}$$

Hence using Lemma 3.5 we get

$$\begin{aligned}\alpha = \operatorname{tr} Y^T Z_0 &= \sigma(Z_0)^T P_{(-)}\sigma(Y) \leq \sigma(Z_0)^T |P_{(-)}|\sigma(Y) \\ &\leq \sigma(Z_0)^T\sigma(Y) = \sigma(X)^T\sigma(Y) \leq \alpha.\end{aligned}$$

Hence we can conclude that $\alpha = \sigma(X)^T\sigma(Y)$ and, using Lemma 4.5, there exists a signed permutation matrix R in $P_{(-)}(n)$ such that $RP_{(-)}\sigma(Y) = \sigma(Y)$ and $R\sigma(Z_0) = \sigma(Z_0)$. Plugging this into equations (8) we get that

$$Y = U_n^T\big(\operatorname{Diag} R^T\sigma(Y)\big)U_m, \quad Z_0 = U_n^T\big(\operatorname{Diag} R^T\sigma(Z_0)\big)U_m.$$

But

$$\big(\operatorname{Diag} R^T\sigma(Y)\big) = R^T(\operatorname{Diag}\sigma(Y)) \begin{pmatrix} |R| & 0 \\ 0 & I_{m-n,m-n} \end{pmatrix},$$

and there is a similar equation involving Z_0. The theorem follows. ■

This section ends with two simple linear-algebraic results which are useful later. The first is Proposition 3 in [13].

Proposition 4.7 (Simultaneous Square Conjugacy) *For any vectors x, y, u, v in $\mathbb{R}^n$, there is a matrix U in $O(n)$ with $\operatorname{Diag} x = U^T(\operatorname{Diag} u)U$ and $\operatorname{Diag} y = U^T(\operatorname{Diag} v)U$ if and only if there is a matrix P in $P(n)$ with $x = Pu$ and $y = Pv$.*

Proposition 4.8 (Simultaneous Rectangular Conjugacy) *For any vectors x, y, u, and v in $\mathbb{R}^n$, there is an element (U_n, U_m) in $O(n,m)$ with $\operatorname{Diag} x = U_n^T(\operatorname{Diag} u)U_m$ and $\operatorname{Diag} y = U_n^T(\operatorname{Diag} v)U_m$ if and only if there is a matrix $P_{(-)}$ in $P_{(-)}(n)$ with $x = P_{(-)}u$ and $y = P_{(-)}v$.*

Proof. In one direction the proof is easy. In the other direction we divide it into four steps. First we note that

$$(\operatorname{Diag} x)(\operatorname{Diag} x)^T = U_n^T(\operatorname{Diag} u)(\operatorname{Diag} u)^T U_n$$

$$(\operatorname{Diag} y)(\operatorname{Diag} y)^T = U_n^T(\operatorname{Diag} v)(\operatorname{Diag} v)^T U_n$$

So from Proposition 4.7, there is a permutation matrix P_1 in $P(n)$ such that

$$x^2 = P_1 u^2, \text{ and } y^2 = P_1 v^2.$$

This implies that the number of zero entries in vector u is equal to the number of zero entries in vector x, and the permutation is such that if $P_1 e^i = e^j$ then $|u_i| = |x_j|$ and $|v_i| = |y_j|$.

Second we have that

$$(\operatorname{Diag} x)(\operatorname{Diag} x)^T = U_n^T(\operatorname{Diag} u)(\operatorname{Diag} u)^T U_n$$

$$(\operatorname{Diag} x)(\operatorname{Diag} y)^T = U_n^T(\operatorname{Diag} u)(\operatorname{Diag} v)^T U_n$$

Again according to the previous proposition, there is a permutation matrix P_2 in $P(n)$ such that

$$x^2 = P_2 u^2 \text{ and } x \cdot y = P_2(u \cdot v).$$

Third, let π_1 and π_2 be the permutations corresponding to the permutation matrices P_1 and P_2, that is, $P_j e^i = e^{\pi_j(i)}$ for all $j = 1, 2$ and $i = 1, ..., n$. We use π_1 and π_2 to form a new permutation π (with corresponding permutation matrix P) in the following way:

$$\pi(i) = \begin{cases} \pi_1(i) & \text{if } u_i = 0 \\ \pi_2(i) & \text{if } u_i \neq 0. \end{cases}$$

Because P_2 also maps the zero entries of u one-to-one onto the zero entries of x, the above construction is well defined.

In the last step we show that we can turn P into a signed permutation matrix $P_{(-)}$ with the desired properties and such that $|P_{(-)}| = P$. Suppose $\pi(i) = j$ (this of course means $P^{j,i} = 1$), then
If $u_i = 0$ and $v_i = 0$ then we set $P^{j,i}_{(-)} = P^{j,i} = 1$.
If $u_i = 0$ and $v_i \neq 0$ then set $P^{j,i}_{(-)} = \text{sign}\,(v_i)\text{sign}\,(y_j)$.
If $u_i \neq 0$ and $v_i = 0$ then set $P^{j,i}_{(-)} = \text{sign}\,(u_i)\text{sign}\,(x_j)$.
If $u_i \neq 0$ and $v_i \neq 0$ then set again $P^{j,i}_{(-)} = \text{sign}\,(u_i)\text{sign}\,(x_j)$.

It is easily verified that $x = P_{(-)}u$ and $y = P_{(-)}v$. ■

5 Simultaneous Diagonalization

The reader can easily check the following elementary statement using the singular value decomposition theorem.

Proposition 5.1 (Orthogonally Invariant & Absolutely Symmetric) *The following two properties of a function $F : M_{n,m} \to [-\infty, +\infty]$ are equivalent:*

(i) F is orthogonally invariant; that is, any matrices X in $M_{n,m}$, U_n in $O(n)$, and U_m in $O(m)$ satisfy $F(U_n^T X U_m) = F(X)$.

(ii) $F = f \circ \sigma$ for some absolutely symmetric function $f : \mathbb{R}^n \to [-\infty, +\infty]$ that is, any vector x in $\mathbb{R}^n$ and matrix P in $P_{(-)}(n)$ satisfy $f(Px) = f(x)$.

Definition 5.2 (Singular Value Function) A *singular value function* is an extended-real-value function defined on $M_{n,m}$ of the form $f \circ \sigma$ for an absolutely symmetric function $f : \mathbb{R}^n \to [-\infty, +\infty]$.

Theorem 5.3 (Symmetricity) *If a matrix Y in $M_{n,m}$ is a regular, a limiting, or a horizon subgradient of a singular value function F at a matrix X in $M_{n,m}$, then X and Y satisfy $X^T Y = Y^T X$ and $Y^T X = X^T Y$.*

Proof. Take first $Y \in \hat{\partial} F(X)$ to be a regular subgradient. The orthogonal invariance property of the singular value functions implies that the orbit

$O(n,m).X$ is contained in the level set $\mathcal{L} = \{Z \in M_{n,m} \,|\, F(Z) \leq F(X)\}$ of F at X. Then using the Normal Cone Proposition (2.7) we get

$$Y \in \big(T_{\mathcal{L}}(X))\big)^- \subset \big(T_{O(n,m).X}(X)\big)^- = \big(T_{O(n,m).X}(X)\big)^{\perp}.$$

Since by the Normal Space Theorem (3.4) the tangent cone T_X of the orbit $O(n,m).X$ at X is a linear space. Thus we get $X^TY = Y^TX$ and $Y^TX = X^TY$.

Next, let Y be a limiting subgradient of F at X. By the definition, there is a sequence of matrices X_r in $M_{n,m}$ approaching X with a corresponding sequence of regular subgradients Y_r in $\hat{\partial}F(X_r)$, approaching Y. By the above paragraph we have

$$X^TY = \lim_r X_r^T Y_r = \lim_r Y_r^T X_r = Y^TX.$$

The relationship $Y^TX = X^TY$ is similar.

If Y is a horizon subgradient then there are sequences Y_r approaching Y and reals t_r decreasing to 0 such that t_rY_r approaches Y. Thus, together with the sequence X_r in $M_{n,m}$ approaching X we have

$$X^TY = \lim_r X_r^T t_r Y_r = \lim_r t_r Y_r^T X_r = Y^TX. \qquad \blacksquare$$

The above theorem together with Lemma 4.3 show that if a matrix Y is a subgradient of some singular value function F at the matrix X, (where $X, Y \in M_{n,m}$) then X and Y can be simultaneously diagonalized:

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

where (U_n, U_m) is in $O(n,m)$, and $P_{(-)}$ is a signed permutation matrix in $P_{(-)}(n)$. Using the Subgradient Invariance Proposition (2.8) applied to the space $M_{n,m}$ with the action of the group $O(n,m)$, we see that the matrix $\operatorname{Diag} P_{(-)}\sigma(Y)$ must be a subgradient at $\operatorname{Diag}\sigma(X)$. All this shows how we can simplify the problem of characterizing the nonsmooth subdifferentials of a singular value function. We can see that it is enough to consider only the case when X and Y are both diagonal (by that we mean $X_{i,j} = 0$ if $i \neq j$). We make all these observations precise in the following sections. In the next proposition we show the easy inclusion.

Proposition 5.4 *Any vectors x and y in $\mathbb{R}^n$, and singular value function $f \circ \sigma$ satisfy*

$$\operatorname{Diag}(y) \in \partial(f \circ \sigma)(\operatorname{Diag} x) \Rightarrow y \in \partial f(x).$$

Corresponding results hold for regular and horizon subgradients.

Proof. We show first that the claim holds when $\operatorname{Diag} y$ is a regular subgradient of $f \circ \sigma$ at $\operatorname{Diag} x$. For vectors z in $\mathbb{R}^n$ close to the origin we have

$$\begin{aligned} f(x+z) &= f(|x+z|) \\ &= (f \circ \sigma)(\operatorname{Diag} x + \operatorname{Diag} z) \\ &\geq (f \circ \sigma)(\operatorname{Diag} x) + \operatorname{tr}(\operatorname{Diag} y)^T(\operatorname{Diag} z) + o(\operatorname{Diag} z) \\ &= f(|x|) + y^T z + o(z) \\ &= f(x) + y^T z + o(z), \end{aligned}$$

whence $y \in \hat{\partial} f(x)$.

Next, if $\operatorname{Diag} y \in \partial(f \circ \sigma)(\operatorname{Diag} x)$, then there is a matrix sequence X_r in $M_{n,m}$ approaching $\operatorname{Diag} x$, with $(f \circ \sigma)(X_r)$ approaching $(f \circ \sigma)(\operatorname{Diag} x)$, and a sequence of regular subgradients Y_r in $\hat{\partial}(f \circ \sigma)(X_r)$ approaching $\operatorname{Diag} y$. By Theorem 5.3 there is a sequence of elements (U_n^r, U_m^r) of $O(n,m)$ and a sequence of matrices $P_{(-)}^r$ in $P_{(-)}(n)$ such that

$$X_r = (U_n^r)^T\big(\operatorname{Diag} P_{(-)}^r \sigma(X_r)\big)U_m^r \text{ and } Y_r = (U_n^r)^T\big(\operatorname{Diag} \sigma(Y_r)\big)U_m^r \tag{9}$$

for every r. The Subgradient Invariance Proposition (2.8) now shows that $\operatorname{Diag}\sigma(Y_r) \in \hat{\partial}(f \circ \sigma)\big(\operatorname{Diag} P_{(-)}^r \sigma(X_r)\big)$. Therefore by the first paragraph $\sigma(Y_r) \in \hat{\partial} f(P_{(-)}^r \sigma(X_r))$.

The groups $O(n,m)$ and $P_{(-)}(n)$ are compact. So without loss of generality we can assume that (U_n^r, U_m^r) approaches an element (U_n, U_m) in $O(n,m)$ and $P_{(-)}^r$ approaches $P_{(-)}$ in $P_{(-)}(n)$. Moreover because $P_{(-)}(n)$ is a discrete group the elements of the sequence $P_{(-)}^r$ will be equal to $P_{(-)}$ for big enough r's. Hence from equation (9), taking the limit and rearranging we get

$$\begin{aligned} U_n(\operatorname{Diag} x)U_m^T &= \operatorname{Diag}(P_{(-)}\sigma(\operatorname{Diag} x)), \quad \text{and} \\ U_n(\operatorname{Diag} y)U_m^T &= \operatorname{Diag}\sigma(\operatorname{Diag} y). \end{aligned} \tag{10}$$

Since $P_{(-)}^r \sigma(X_r)$ approaches $P_{(-)}\sigma(\operatorname{Diag} x)$, with $f(P_{(-)}^r \sigma(X_r)) = f(\sigma(X_r))$ approaching $f(\sigma(\operatorname{Diag} x)) = f(P_{(-)}\sigma(\operatorname{Diag} x))$, and $\sigma(Y_r) \in \hat{\partial} f(P_{(-)}^r \sigma(X_r))$ approaching $\sigma(\operatorname{Diag} y)$, then $\sigma(\operatorname{Diag} y)$ belongs to $\partial f(P_{(-)}\sigma(\operatorname{Diag} x))$.

Combining Equation (10) and Proposition 4.8, there exists a signed permutation matrix $\hat{P}_{(-)}$ such that $x = \hat{P}_{(-)}P_{(-)}\sigma(\operatorname{Diag} x)$, $y = \hat{P}_{(-)}\sigma(\operatorname{Diag} y)$. Applying the Subgradient Invariance Proposition (2.8) again, this time to the space $\mathbb{R}^n$ with the group $P_{(-)}(n)$, we get that y belongs to $\partial f(x)$ as we claimed.

In the case when $\operatorname{Diag} y$ is a horizon subgradient, the calculations are analogous. ■

6 Directional derivatives of singular values

As we said before Proposition 5.4, the opposite inclusion to the one stated there is the more difficult one. It is our goal in this section to show that. Once we show the opposite inclusion for regular subgradients, most of the goal will be achieved. Thus the difficulty is in showing that for vectors x and y in $\mathbb{R}^n$ and a singular value function $f \circ \sigma$ we have

$$y \in \hat{\partial} f(x) \Rightarrow \operatorname{Diag} y \in \hat{\partial}(f \circ \sigma)(\operatorname{Diag} x). \tag{11}$$

We need to state two more propositions. The first is obtained by combining Theorem 4.3 with Example 7.6 in [12]. The second is Theorem 3.1 in [9].

Proposition 6.1 (Characterization Of Convexity) *Let the function $f : \mathbb{R}^n \to (-\infty, +\infty]$ be absolutely symmetric. Then the corresponding singular value function $f \circ \sigma$ is convex on $M_{n,m}$ if and only if f is convex.*

Proposition 6.2 (Gradient Formula) *If a function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is convex and absolutely symmetric, then the corresponding convex, orthogonally invariant function $f \circ \sigma$ is differentiable at the matrix X if and only if f is differentiable at $\sigma(X)$. In this case*

$$\nabla(f \circ \sigma)(X) = U_n^T\big(\operatorname{Diag} \nabla f(\sigma(X))\big)U_m,$$

for any matrices U_n in $O(n)$ and U_m in $O(m)$ with $X = U_n^T\big(\operatorname{Diag}\sigma(X)\big)U_m$.

For each integer $k = 0, 1, 2, ..., n$ we define the function $S_k : M_{n,m} \to \mathbb{R}$ by $S_k(M) = \sum_{i=1}^{k} \sigma_i(M)$, the sum of the k largest singular values of the matrix M. For convenience we define $S_0 = 0$. It is well known result of Fan that S_k

is convex (even sublinear) function on $M_{n,m}$ (see for example Corollary 3.4.4 in [8]). Another way to see this is by using Proposition 6.1. We define a new symbol $\mathbb{R}^n_{\downarrow} := (\mathbb{R}^n_{\downarrow} \cap \mathbb{R}^n_{+})$. To simplify the notation in the following several lemmas, if x is a vector from $\mathbb{R}^n$ but the indexing refers to the element x_{n+1}, then we will assume that $x_{n+1} = 0$.

Lemma 6.3 *The function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = \sum_{i=1}^{k} \hat{x}_i$ $(k \leq n)$ is differentiable at any point $\mu \in \mathbb{R}^n_{\downarrow}$ such that $\mu_k > \mu_{k+1}$, and its derivative is*

$$\nabla f(\mu) = \sum_{i=1}^{k} e^i.$$

Proof. Set $v := \sum_{i=1}^{k} e^i$. For all vectors x with sufficiently small norm we have $f(\mu + x) = \sum_{i=1}^{k}(\mu_k + x_k)$. So for all sufficiently small vectors $x \neq 0$, $\frac{f(\mu+x)-f(\mu)-\langle v,x\rangle}{\|x\|} = 0$. Consequently $\nabla f(\mu) = \sum_{i=1}^{k} e^i$. ■

Lemma 6.4 *Fix an integer k, $1 \leq k \leq n$. For any real vector x in $\mathbb{R}^n$ such that $\hat{x}_k > \hat{x}_{k+1}$ the function S_k is differentiable at* $\operatorname{Diag} x$ *with gradient*

$$\nabla S_k(\operatorname{Diag} x) = U_n^T \left(\operatorname{Diag} \sum_{i=1}^{k} e^i \right) U_m,$$

where U_n, U_m are any orthogonal matrices such that $\operatorname{Diag} x = U_n^T (\operatorname{Diag} \hat{x}) U_m$.

Note 6.5 Of course one can choose the matrices U_n and U_m in such a way that U_n is a signed permutation matrix, $P_{(-)}$, and U_m is the block diagonal matrix $\operatorname{Diag}(|P_{(-)}|, I_{m-n,m-n})$. In particular if $x \in \mathbb{R}^n_{\downarrow}$ we can take $U_n = I_n$ and $U_m = I_m$.

Proof. The function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(y) = \sum_{i=1}^{k} \hat{y}_i$ is easily seen to be absolutely symmetric and convex. From Lemma 6.3 it is also differentiable at the point $\sigma(\operatorname{Diag} x) = \hat{x}$. So by Proposition 6.2 it follows that $f \circ \sigma$ is differentiable at $\operatorname{Diag} x$. But $(f \circ \sigma)(M) = S_k(M)$ for each M in $M_{n,m}$, so S_k is differentiable at $\operatorname{Diag} x$ and the formula for its gradient follows from Proposition 6.2 and Lemma 6.3. ■

Lemma 6.6 *For any vector w in $\mathbb{R}^n_{\downarrow}$, the function $w^T\sigma$ is convex, and any vector x in $\mathbb{R}^n_{\downarrow}$ satisfies* $\operatorname{Diag} w \in \partial(w^T\sigma)(\operatorname{Diag} x)$.

Proof. The absolutely symmetric continuous function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(z) = w^T \hat{z}$ is convex because it is the maximum of a family of convex (linear in this case) functions

$$f(z) = \max \{w^T P_{(-)} z : P_{(-)} \in P_{(-)}(n)\},$$

by Lemma 4.5. Then by Proposition 6.1 we obtain that $f \circ \sigma$ is convex. To prove the claim about the gradient it is enough to show that any matrix Z in $M_{n,m}$ satisfies

$$\operatorname{tr}(\operatorname{Diag} w)(Z - \operatorname{Diag} x) \leq w^T \sigma(Z) - w^T x,$$

or in other words, $\operatorname{tr}(\operatorname{Diag} w) Z \leq w^T \sigma(Z)$. This inequality follows from von Neumann's Theorem (4.6) ■

For any vector x in $\mathbb{R}^n$, we denote by $P_{(-)}(n)_x$ the stabilizer of x in the group $P_{(-)}(n)$, that is

$$P_{(-)}(n)_x = \{P_{(-)} \in P_{(-)}(n) : P_{(-)} x = x\}.$$

The following lemma is an extension and generalization for singular values of Lemma 5.3 in [11].

Lemma 6.7 *If x is a vector in $\mathbb{R}^n_{\downarrow}$, and w is a vector in $\mathbb{R}^n$ such that the stabilizer $P_{(-)}(n)_x$ is a subgroup of $P_{(-)}(n)_w$, then the function $w^T \sigma(\cdot)$ is differentiable at* $\operatorname{Diag} x$ *with*

$$\nabla(w^T \sigma)(\operatorname{Diag} x) = \operatorname{Diag} w.$$

Proof. Suppose that the structure of vector x is

$$x_1 = ... = x_{k_1} > x_{k_1+1} = ... = x_{k_2} > ... > x_{k_r+1} = ... = x_{k_{r+1}} = 0, \ (k_{r+1} = n).$$

(The proof of the lemma is the same even if $x_n > 0$.) Since the stabilizer $P_{(-)}(n)_x$ is a subgroup of $P_{(-)}(n)_w$, there exist reals β_1, β_2,...,β_r,β_{r+1} with

$$w_i = \beta_j \text{ whenever } k_{j-1} < i \leq k_j, \ j = 1, 2, ..., r,$$

where $\beta_{r+1} = 0$ and we set $k_0 = 0$. We obtain

$$w^T\sigma(X) = \sum_{j=1}^{r+1} \beta_j \sum_{i=k_{j-1}+1}^{k_j} \sigma_i(X) = \sum_{j=1}^{r+1} \beta_j \big(S_{k_j}(X) - S_{k_{j-1}}(X)\big).$$

Let $P^1_{(-)} = I_n$ and $P^2 = I_m$ the identity matrices of the indicated dimension. Then applying Lemma 6.4 and the note after it gives

$$\begin{aligned}
\nabla(w^T\sigma)(\operatorname{Diag} x) &= \sum_{j=1}^{r+1} \beta_j I_n^T \bigg(\operatorname{Diag} \sum_{i=1}^{k_j} e^i - \operatorname{Diag} \sum_{i=1}^{k_{j-1}} e^i\bigg) I_m \\
&= \bigg(\sum_{j=1}^{r} \beta_j \ \operatorname{Diag} \sum_{i=k_{j-1}+1}^{k_j} e^i\bigg) \\
&= \operatorname{Diag} w,
\end{aligned}$$

as required. ■

The following theorem is crucial for the proof of implication (11). Notice that the adjoint of the linear map $\operatorname{Diag}\colon \mathbb{R}^n \to M_{n,m}$ is the map $\operatorname{diag}\colon M_{n,m} \to \mathbb{R}^n$, taking a matrix M to a vector with components $M_{i,i}$ $(1 \le i \le n)$.

Theorem 6.8 (Singular Value Derivatives) *Any vector x in $\mathbb{R}^n_\downarrow$ and matrix M in $M_{n,m}$ satisfy*

$$\operatorname{diag} M \in \operatorname{conv}\big(P_{(-)}(n)_x \sigma'(\operatorname{Diag} x; M)\big). \tag{12}$$

Proof. Assume first that $x_n = 0$. Suppose again that the structure of the vector $x \in \mathbb{R}^n_\downarrow$ is

$$x_1 = ... = x_{k_1} > x_{k_1+1} = ... = x_{k_2} > ... > x_{k_r+1} = ... = x_{k_{r+1}} = 0, \ (k_{r+1} = n).$$

The indexes define a partitioning of the integers $\{1, 2, ..., n\}$ into consecutive blocks

$$I_1 = \{1, 2..., k_1\},\ I_2 = \{k_1+1, k_1+2, ..., k_2\}, ..., I_{r+1} = \{k_r+1, k_r+2, ..., k_{r+1}\}$$

Thus $x_i = x_j$ if and only if the indices i and j belong to the same block and $x_i \in I_{r+1}$ if and only if $x_i = 0$. We are also going to say that an entry of x

belongs to a particular block if its index is in that block. With respect to these blocks, write any vector y in $\mathbb{R}^n$ in the form

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

The stabilizer $P_{(-)}(n)_x$ consists of matrices permuting the entries of x in a block I_i, (for every fixed i, $1 \leq i \leq r$) among themselves (without sign changes) and permuting the entries of x belonging to the block I_{r+1} among themselves (with possible sign changes).

Assume that relation (12) fails. Then there exists a hyperplane separating $\operatorname{diag} M$ from $\operatorname{conv}\big(P_{(-)}(n)_x \sigma'(\operatorname{Diag} x; M)\big)$. That is, some vector y in $\mathbb{R}^n$ satisfies

$$y^T \operatorname{diag} M > y^T P_{(-)} \sigma'(\operatorname{Diag} x; M), \text{ for all } P_{(-)} \text{ in } P_{(-)}(n)_x. \tag{13}$$

Let $\tilde{y}$ denote the vector $\oplus_{i=1}^r \overline{y^i} \oplus \widehat{y^{r+1}}$. There is a vector v in $\mathbb{R}^n$ with equal components within every block I_i $(1 \leq i \leq r)$ and $v_j = 0$ whenever $j \in I_{r+1}$ (that is, $P_{(-)}(n)_x$ is a subgroup of $P_{(-)}(n)_v$) so that $v+\tilde{y}$ lies in $\mathbb{R}^n_{\Downarrow}$. Lemma 6.6 shows that

$$\operatorname{Diag}(v + \tilde{y}) \in \partial\big((v + \tilde{y})^T \sigma\big)(\operatorname{Diag} x),$$

which in turn means that for any T in $M_{n,m}$ and a real t, using the definition of a convex subgradient for the matrix $\operatorname{Diag} x + tT$

$$\operatorname{tr}\big((tT)^T(\operatorname{Diag}(v + \tilde{y}))\big) \leq \big((v + \tilde{y})^T \sigma\big)(\operatorname{Diag} x + tT) - \big((v + \tilde{y})^T \sigma\big)(\operatorname{Diag} x).$$

Dividing by t and letting it go to 0^+ we arrive at

$$\operatorname{tr}\big(T^T(\operatorname{Diag}(v + \tilde{y}))\big) \leq (v + \tilde{y})^T \sigma'(\operatorname{Diag} x; T), \tag{14}$$

for any matrix T in $M_{n,m}$. On the other hand, Lemma 6.7 shows that

$$\operatorname{tr}\big(T^T(\operatorname{Diag} v)\big) = v^T \sigma'(\operatorname{Diag} x; T). \tag{15}$$

Subtracting equation (15) from inequality (14) gives

$$\operatorname{tr}\big(T^T(\operatorname{Diag} \tilde{y})\big) \leq \tilde{y}^T \sigma'(\operatorname{Diag} x; T). \tag{16}$$

If we set $\operatorname{diag} M =: w = \oplus_r w^r$, then there is a matrix Q in $P_{(-)}(n)_x$ satisfying

$$\operatorname{diag}\left(Q^T M \begin{pmatrix} |Q| & 0 \\ 0 & I_{m-n,m-n} \end{pmatrix}\right) = \oplus_{i=1}^r \overline{w^i} \oplus \widehat{w^{r+1}}.$$

Choosing the matrix T in inequality (16) to be $T = Q^T M \begin{pmatrix} |Q| & 0 \\ 0 & I_{m-n,m-n} \end{pmatrix}$ and using Lemma 3.5 repeatedly and Lemma 4.5 shows

$$\begin{aligned} y^T w &\leq \left(\oplus_{i=1}^r \overline{y^i} \right)^T \left(\oplus_{i=1}^r \overline{w^i} \right) + \widehat{y^{r+1}}^T \widehat{w^{r+1}} \\ &= \operatorname{tr} \left(T^T (\operatorname{Diag} \tilde{y}) \right) \\ &\leq \tilde{y}^T \sigma'(\operatorname{Diag} x; T) \\ &= \tilde{y}^T \sigma'(\operatorname{Diag} x; M). \end{aligned}$$

In the last equality we used the Subgradient Invariance Proposition (2.8) and the fact that Q is in $P_{(-)}(n)_x$. But now choosing the matrix $P_{(-)} \in P_{(-)}(n)_x$ in inequality (13) so that $P_{(-)}^T y = \tilde{y}$ gives a contradiction.

Assume now $x_n > 0$. Then the reader can verify that the proof works again if we consider that the block I_{r+1} is empty. That is, we write any vector y in $\mathbb{R}^n$ in the form

$$y = \oplus_{i=1}^r y^i, \text{ where } y^i \in \mathbb{R}^{|I_i|} \text{ for each } 1 \leq i \leq r.$$

The stabilizer $P_{(-)}(n)_x$ consists of matrices of permutations fixing each block I_i, $(1 \leq i \leq r)$. The vector $\tilde{y}$ denotes $\oplus_{i=1}^r \overline{y^i}$. There is a vector v in $\mathbb{R}^n$ with equal components within every block I_i, $(1 \leq i \leq r)$ so that...and so on. We just omit the "r+1"-part of each vector until the end of the proof. ■

Another result that we will need is that the singular value map σ can be directionally expanded in a first order series. This expansion is uniform in the perturbation matrix. In other words we have the following lemma.

Lemma 6.9 *Given a matrix X in $M_{n,m}$, small matrices M in $M_{n,m}$ satisfy*

$$\sigma(X + M) = \sigma(X) + \sigma'(X; M) + o(M).$$

Proof. The above uniform first order directional expansion is true for any convex function [6, Lemma VI.2.1.1]. In our case σ_i is the difference of the two convex functions $\sum_{j=1}^i \sigma_j$ and $\sum_{j=1}^{i-1} \sigma_j$ (see Lemma 6.6). So it is true for σ_i as well. ■

Finally we prove the implication (11). Notice though, that first we require x to be in $\mathbb{R}^n_{\downarrow}$. In the corollary that follows we remove this condition.

Theorem 6.10 *For any vectors x in $\mathbb{R}^n_{\downarrow}$ and y in $\mathbb{R}^n$, and any singular value function $f \circ \sigma$,*

$$y \in \hat{\partial} f(x) \Rightarrow \operatorname{Diag} y \in \hat{\partial}(f \circ \sigma)(\operatorname{Diag} x).$$

Proof. The orbit of y under the action of the stabilizer $P_{(-)}(n)_x$ of x contains only regular subgradients in $\hat{\partial} f(x)$, (this follows from the Subgradient Invariance Proposition (2.8)). In other words we have $P_{(-)}(n)_x y \subset \hat{\partial} f(x)$. Denote the convex hull of this orbit by Λ. Then the support function of λ is given by

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

The support function is clearly sublinear (convex and positively homogeneous). It is also globally Lipschitz with constant $\|y\|$.

Fix a real $\epsilon > 0$. For any $P_{(-)} \in P_{(-)}(n)_x$ the definition of regular subgradients implies, for small vectors z in $\mathbb{R}^n$,

$$f(x+z) \geq f(x) + \langle P_{(-)} y, z \rangle - \epsilon \|z\|. \tag{17}$$

Thus using the finiteness of $P_{(-)}(n)_x$ we can conclude that for vectors $z \in \mathbb{R}^n$ in a smaller neighbourhood around the origin we have

$$f(x+z) \geq f(x) + \delta^*_\Lambda(z) - \epsilon \|z\|. \tag{18}$$

On the other hand, using the previous lemma (6.9), small matrices Z in $M_{n,m}$ must satisfy

$$\|\sigma(\operatorname{Diag} x + Z) - x - \sigma'(\operatorname{Diag} x; Z)\| \leq \epsilon \|Z\|,$$

and hence, by inequality (18),

$$\begin{aligned}
f\big(\sigma(\operatorname{Diag} x + Z)\big) &= f\big(x + (\sigma(\operatorname{Diag} x + Z) - x)\big) \\
&\geq f(x) - \epsilon \|\sigma(\operatorname{Diag} x + Z) - x\| \\
&\quad + \delta^*_\Lambda\big(\sigma'(\operatorname{Diag} x; Z) + [\sigma(\operatorname{Diag} x + Z) - x - \sigma'(\operatorname{Diag} x; Z)]\big) \\
&\geq f(x) - \epsilon \|Z\| \\
&\quad + \delta^*_\Lambda\big(\sigma'(\operatorname{Diag} x; Z) + [\sigma(\operatorname{Diag} x + Z) - x - \sigma'(\operatorname{Diag} x; Z)]\big) \\
&\geq f(x) + \delta^*_\Lambda\big(\sigma'(\operatorname{Diag} x; Z)\big) - (1 + \|y\|)\epsilon \|Z\|.
\end{aligned}$$

In the second inequality we used the Lipschitz property of σ together with the assumption $x \in \mathbb{R}^n_{\downarrow}$, that is, $\sigma(\operatorname{Diag} x) = x$. In the last inequality we used the Lipschitz property of the support function δ^*_Λ. Recall that the Singular Value Derivatives Theorem (6.8) implies

$$\operatorname{diag} Z \in \operatorname{conv}\big(P_{(-)}(n)_x \sigma'(\operatorname{Diag} x; Z)\big). \tag{19}$$

The support function $\delta^*_\Lambda(z)$ is invariant under the stabilizer $P_{(-)}(n)_x$ acting on the argument z since the set Λ is invariant. Thus

$$\delta^*_\Lambda\big(P_{(-)}\sigma'(\operatorname{Diag} x; Z)\big) = \delta^*_\Lambda\big(\sigma'(\operatorname{Diag} x; Z)\big),$$

for any matrix $P_{(-)}$ in $P_{(-)}(n)_x$. This combined with the convexity of δ^*_Λ and relation (19), demonstrates

$$\delta^*_\Lambda(\operatorname{diag} Z) \leq \delta^*_\Lambda(\sigma'(\operatorname{Diag} x; Z)).$$

Continuing the argument above we have

$$\begin{aligned} f(\sigma(\operatorname{Diag} x + Z)) &\geq f(x) + \delta^*_\Lambda(\operatorname{diag} Z) - (1 + \|y\|)\epsilon\|Z\| \\ &\geq f(x) + y^T \operatorname{diag} Z - (1 + \|y\|)\epsilon\|Z\| \\ &= f(x) + \langle \operatorname{Diag} y, Z\rangle - (1 + \|y\|)\epsilon\|Z\|, \end{aligned}$$

where the number ϵ was arbitrary. The result follows. ■

Corollary 6.11 (Diagonal Subgradients) *For any vectors x and y in $\mathbb{R}^n$ and any singular value function $f \circ \sigma$,*

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

Corresponding results hold for regular and horizon subgradients.

Proof. Again we will first show the corollary in the case when y is a regular subgradient. Let $P_{(-)}$ be a signed permutation matrix in $P_{(-)}(n)$ such that $\hat{x} = P_{(-)}x$. By the Subgradient Invariance Proposition (2.8) the assumption $y \in \hat{\partial} f(x)$ implies $P_{(-)}y \in \hat{\partial} f(P_{(-)}x)$. We now apply Theorem 6.10 to get

$$P_{(-)}(\operatorname{Diag} y) \begin{pmatrix} |P^T_{(-)}| & 0 \\ 0 & I_{m-n,m-n} \end{pmatrix} = \operatorname{Diag}(P_{(-)}y) \in \hat{\partial}(f \circ \sigma)(\operatorname{Diag}(P_{(-)}x))$$

$$= \hat{\partial}(f \circ \sigma) \left(P_{(-)}(\operatorname{Diag} x) \begin{pmatrix} |P_{(-)}| & 0 \\ 0 & I_{m-n,m-n} \end{pmatrix} \right),$$

Apply again the Subgradient Invariance Proposition to get the result.

In the limiting subdifferential case, $y \in \partial f(x)$, there is a sequence of vectors x^r in $\mathbb{R}^n$ approaching x, with $f(x^r)$ approaching $f(x)$, and a sequence of regular subgradients $y^r \in \hat{\partial} f(x^r)$ approaching y. Clearly $\operatorname{Diag} x^r$ approaches $\operatorname{Diag} x$ with $f(\sigma(\operatorname{Diag} x^r))$ approaching $f(\sigma(\operatorname{Diag} x))$, and by the above argument, each matrix $\operatorname{Diag} y^r$ is a regular subgradient of $f \circ \sigma$ at $\operatorname{Diag} x^r$. Since $\operatorname{Diag} y^r$ approaches $\operatorname{Diag} y$, the result follows. The horizon subgradient case is almost identical. ■

7 The main result

The hard part is over. We now present the main result of the paper giving an easy-to-use and easy-to-remember formula for the subdifferential of a singular value function in terms of the subdifferential of the corresponding absolutely symmetric function. The theorem just builds on the reduced case given in Corollary 6.11.

Theorem 7.1 (Subgradients) *The limiting subdifferential of a singular value function $f \circ \sigma$ at a matrix X in $M_{n,m}$ is given by the formula*

$$\partial(f \circ \sigma)(X) = O(n,m)^X.\operatorname{Diag} \partial f(\sigma(X)), \tag{20}$$

where

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

The sets of regular and horizon subgradients satisfy corresponding formulae.

Proof. The Diagonal Subgradients Corollary (6.11) shows that for any vector y in $\partial f(\sigma(X))$ we have

$$\operatorname{Diag} y \in \partial(f \circ \sigma)(\operatorname{Diag} \sigma(X))$$

Now, for any element (U_n, U_m) of $O(n,m)$ such that $U_n^T(\operatorname{Diag} \sigma(X))U_m = X$, the Subgradient Invariance Proposition (2.8) implies

$$U_n^T(\operatorname{Diag} y)U_m \in \partial(f \circ \sigma)(U_n^T(\operatorname{Diag} \sigma(X))U_m) = \partial(f \circ \sigma)(X).$$

All this shows the inclusion $\partial(f \circ \sigma)(X) \supseteq O(n,m)^X.\mathrm{Diag}\,\partial f(\sigma(X))$.

For the opposite inclusion, take a subgradient Y in $\partial(f \circ \sigma)(X)$. By the Symmetricity Theorem (5.3) it satisfies the relationships: $Y^T X = X^T Y$ and $YX^T = XY^T$. Hence by Lemma 4.3 there exists an element (U_n, U_m) in $O(n,m)$ and a signed permutation matrix $P_{(-)}$ in $P_{(-)}(n)$ such that

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

Then the Subgradient Invariance Proposition (2.8) shows

$$\mathrm{Diag}\,P_{(-)}\sigma(Y) \in \partial(f \circ \sigma)(\mathrm{Diag}\,\sigma(X)),$$

whence $P_{(-)}\sigma(Y) \in \partial f(\sigma(X))$, by the Diagonal Subgradient Corollary. The arguments for regular and horizon subgradients are similar. ■

Note 7.2 Analogous result also holds for the Clarke subgradients - see [15].

Corollary 7.3 (Unique Regular Subgradients) *A singular value function $f \circ \sigma$ has a unique regular subgradient at a matrix X in $M_{n,m}$ if and only if f has a unique regular subgradient at $\sigma(X)$.*

Proof. If $f \circ \sigma$ has a unique regular subgradient at a matrix X then clearly f has a unique regular subgradient at the vector $\sigma(X)$.

To show the opposite, suppose f has unique regular subgradient y at $\sigma(X)$. Then by the subdifferential formula (20) we get that every matrix in the convex set $\hat{\partial}(f \circ \sigma)(X) \neq \emptyset$ has the same norm, namely $\|y\|$, and therefore this set has just one element. ■

Corollary 7.4 (Fréchet Differentiability) *A singular value function $f \circ \sigma$ is Fréchet differentiable at a matrix X in $M_{n,m}$ if and only if f is Fréchet differentiable at $\sigma(X)$.*

Proof. A function h is Fréchet differentiable at a point if and only if both h and $-h$ have unique regular subgradients there. Thus this corollary follows from Corollary 7.3 ■

Corollary 7.5 (Regularity) *Suppose the absolute symmetric function f is finite at $\sigma(X)$ (for a matrix X in $M_{n,m}$). Then the singular value function $f \circ \sigma$ is regular at X if and only if f is regular at $\sigma(X)$.*

Proof. Recall that $f \circ \sigma$ is lower semicontinuous around X if and only if f is lower semicontinuous around $\sigma(X)$.

The definition of regularity [16, Corollary 8.11] states that f is regular at $\sigma(X)$ if and only if if it is lower semicontinuous around $\sigma(X)$ and the following conditions hold

$$\partial f(\sigma(X)) = \hat{\partial} f(\sigma(X)) \neq \emptyset, \text{ and} \tag{21}$$

$$(\hat{\partial} f(\sigma(X)))^\infty = \partial^\infty f(\sigma(X)). \tag{22}$$

On the other hand $f \circ \sigma$ is regular at X if and only if it is lower semicontinuous around X and the following conditions hold

$$\partial(f \circ \sigma)(X) = \hat{\partial}(f \circ \sigma)(X) \neq \emptyset, \text{ and} \tag{23}$$

$$(\hat{\partial}(f \circ \sigma)(X))^\infty = \partial^\infty (f \circ \sigma)(X). \tag{24}$$

By formula (20) and its regular analogue, condition (21) implies condition (23). Conversely, by the Subgradient Invariance Proposition (2.8), condition (23) is equivalent to

$$\partial(f \circ \sigma)(\operatorname{Diag} \sigma(X)) = \hat{\partial}(f \circ \sigma)(\operatorname{Diag} \sigma(X)),$$

and condition (21) follows by the Diagonal Subgradient Corollary (6.11).

Notice that the set of regular subgradients is always closed and convex. Thus, the regular subgradients version of formula (20) states that the sets on both sides of the equality are convex. This allows us to apply the Recession Lemma (2.9) to obtain the second equality below, and assuming that (22) holds, we get

$$\begin{aligned}
(\hat{\partial}(f \circ \sigma)(X))^\infty &= [O(n,m)^X.\operatorname{Diag} \hat{\partial} f(\sigma(X))]^\infty \\
&= O(n,m)^X.[\operatorname{Diag} \hat{\partial} f(\sigma(X))]^\infty \\
&= O(n,m)^X.\operatorname{Diag} [\hat{\partial} f(\sigma(X))]^\infty \\
&= O(n,m)^X.\operatorname{Diag} \partial^\infty f(\sigma(X)) \\
&= \partial^\infty (f \circ \sigma)(X).
\end{aligned}$$

So condition (22) implies condition (24), by the horizon version of formula (20) used in the last equality.

On the other hand, by the Subgradient Invariance Proposition (2.8), condition (24) is equivalent to

$$(\hat{\partial}(f \circ \sigma)(\operatorname{Diag} \sigma(X)))^\infty = \partial^\infty (f \circ \sigma)(\operatorname{Diag} \sigma(X)).$$

Using the Diagonal Subgradients Corollary again and the above equality we obtain

$$\begin{aligned}\operatorname{Diag}(\hat{\partial} f(\sigma(X)))^{\infty} &= (\operatorname{Diag}\hat{\partial} f(\sigma(X)))^{\infty} \\ &= (\hat{\partial}(f\circ\sigma)(\operatorname{Diag}\sigma(X))\cap \operatorname{Diag}\mathbb{R}^n)^{\infty} \\ &= (\hat{\partial}(f\circ\sigma)(\operatorname{Diag}\sigma(X)))^{\infty}\cap \operatorname{Diag}\mathbb{R}^n \\ &= \partial^{\infty}(f\circ\sigma)(\operatorname{Diag}\sigma(X))\cap \operatorname{Diag}\mathbb{R}^n \\ &= \operatorname{Diag}\partial^{\infty} f(\sigma(X)).\end{aligned}$$

Condition (22) follows. ■

Corollary 7.6 (Strict Differentiability) *A singular value function $f\circ\sigma$ is strictly differentiable at a matrix X in $M_{n,m}$ if and only if the function f is strictly differentiable at $\sigma(X)$.*

Proof. Theorem 9.18 in [16] states that a function f is strictly differentiable at $\sigma(X)$ if and only if it is continuous there and both f and $-f$ are regular at $\sigma(X)$ Thus the corollary follows by the Regularity Corollary (7.5) just proved. ■

The Subgradients Theorem (7.1) can be written in graphical form. The graph of the subdifferential is the set

$$\operatorname{Graph}\partial f = \{(x,y)\in\mathbb{R}^n\times\mathbb{R}^n : y\in\partial f(x)\}.$$

Define a binary operation $*: O(n,m)\times(\mathbb{R}^n\times\mathbb{R}^n)\to M_{n,m}\times M_{n,m}$ by

$$(U_n,U_m)*(x,y) = ((U_n,U_m).\operatorname{Diag}x, (U_n,U_m).\operatorname{Diag}y).$$

Corollary 7.7 (Subdifferential Graphs) *The graph of the subdifferential of a singular value function $f\circ\sigma$ is given by the formula*

$$\operatorname{Graph}\partial(f\circ\sigma) = O(n,m)*\operatorname{Graph}\partial f.$$

Analogous formulae hold for the subdifferentials $\hat{\partial}$, ∂^{∞}.

Proof. We first show that the left hand side is contained in the set on the right. Suppose the pair (X,Y) is in Graph $\partial(f\circ\sigma)$. This happens exactly when $Y\in\partial(f\circ\sigma)(X)$. Using the Subgradients Theorem (7.1), this implies

that there is a vector y in $\partial(f(\sigma(X))$ and an element (U_n, U_m) in $O(n,m)^X$ satisfying $Y = (U_n, U_m).\mathrm{Diag}\, y$. Hence $(X, Y) = (U_n, U_m).(\sigma(X), y)$.

For the converse inclusion take a pair of vectors (x, y) in Graph ∂f and an element (U_n, U_m) in $O(n,m)$. Since y lies in $\partial f(x)$ we have $\mathrm{Diag}\, y \in \partial(f \circ \sigma)(\mathrm{Diag}\, x)$, by the Diagonal Subgradients Corollary (6.11). The Subgradient Invariance Proposition implies $(U_n, U_m).\mathrm{Diag}\, y \in \partial(f \circ \sigma)((U_n, U_m).\mathrm{Diag}\, x)$, or in other words $(U_n, U_m) * (x, y) \in$Graph $\partial(f \circ \sigma)$. The arguments for the other subdifferentials are analogous. ■

The regular subgradients of a convex function are exactly the usual convex subgradients. It is also known that in the case of an absolutely symmetric function f, f is convex if and only if $f \circ \sigma$ is. (See [12, Theorem 4.3 and Example 7.5].) With this notes in mind the following corollary is easily deduced from the Subgradients Theorem (7.1). An independent proof can be found in [9, Corollary 2.5].

Corollary 7.8 (Convex Subgradients) *Let the function f be absolutely symmetric and convex. Consider the corresponding convex singular value function $f \circ \sigma$. The matrix Y is a (convex) subgradient of $f \circ \sigma$ at X if and only if $\sigma(Y)$ is a (convex) subgradient of f at $\sigma(X)$ and the two matrices X and Y admit simultaneous ordered singular value decomposition.*

References

[1] F. BRICKELL and R.S. CLARK. *Differential Manifolds; an Introduction.* Van nostrand Reinhold company, London, first edition, 1970.

[2] F.H. CLARKE. *Necessary conditions for nonsmooth problems in optimal control and the calculus of variations.* University of Washington, 1973. Ph.D. Thesis.

[3] F.H. CLARKE. *Optimization and Nonsmooth Analysis.* Wiley, New York, 1983.

[4] F.H. CLARKE, YU.S. LEDYAEV, R.J. STERN, and P.R. WOLENSKI. *Nonsmooth Analysis and Control Theory.* Springer-Verlag, New York, Inc., 1998.

[5] P.M. GIBSON. Simultaneous diagonalization of rectangular complex matrices. *Linear Algebra and Its Applications*, 9:45–53, 1974.

[6] J.-B. HIRIART-URRUTY and C. LEMARÉCHAL. *Convex Analysis and Minimization Algorithms I: Fundamentals.* Number 305 in Grundlehren der Mathematischen Wissenschaften. Springer-Verlag, Berlin, 1993.

[7] R.A. HORN and C.R. JOHNSON. *Matrix Analysis.* Cambridge University Press, second edition, 1985.

[8] R.A. HORN and C.R. JOHNSON. *Topics in Matrix Analysis.* Cambridge University Press, first edition, 1991. Paperback edition with corrections, 1994.

[9] A.S. LEWIS. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1-2):173–183, 1994.

[10] A.S. LEWIS. Convex analysis on the Hermitian matrices. *SIAM Journal on Optimization*, 6:164–177, 1996.

[11] A.S. LEWIS. Derivatives of spectral functions. *Mathematics of Operations Research*, 21:576–588, 1996.

[12] A.S. LEWIS. Group invariance and convex matrix analysis. *SIAM Journal on Matrix Analysis*, 17(4):927–949, 1996.

[13] A.S. LEWIS. Nonsmooth analysis of eigenvalues. *Mathematical Programming*, 84:1–24, 1999.

[14] A.S. LEWIS. Convex analysis on Cartan subspaces. *Nonlinear Analysis, Theory, Methods and Applications*, 42:813–820, 2000.

[15] A.S. LEWIS and H.S. SENDOV. Nonsmooth analysis of singular values: Applications. 2002. Personal Communications.

[16] R.T. ROCKAFELLAR and R.J.-B. WETS. *Variational Analysis.* Springer-Verlag, Berlin, 1998.

[17] H.S. SENDOV. *Variational Spectral Analysis.* University of Waterloo, 2000, http://etd.uwaterloo.ca/etd/hssendov2000.pdf. Ph.D. Thesis.

[18] TIN-YAU TAM and W.C. HILL. Derivatives of orbital functions, an extension of Berezin-Gel'fand's theorem and applications. http://web6.duc.auburn.edu/ tamtiny/gb2.pdf, preprint.

[19] T.Y. TAM. An extension of a result of Lewis. *Electronic Journal of Linear Algebra*, 5:1–10, 1999.

[20] R.C. THOMPSON. On Pearl's paper "A decomposition theorem for matrices. *Canadian Mathematical Bulletin*, 12 (6):805–808, 1969.

[21] J. von NEUMANN. Some matrix inequalities and metrization of matric-space. *Tomsk University Review*, 1:286–300, 1937. In: *Collected Works*, Pergamon, Oxford, 1962, Volume IV, 205-218.

[22] S. WIEGMANN. Some analogues of the generalized principle axis transformation. *Bulletin of the American Mathematical Society*, 54:905–908, 1948.