Variational Spectral Analysis

by

Hristo Stoyanov Sendov

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Combinatorics and Optimization

Waterloo, Ontario, Canada, 2000

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

We present results on smooth and nonsmooth variational properties of *symmetric* functions of the eigenvalues of a real symmetric matrix argument, as well as *absolutely symmetric* functions of the singular values of a real rectangular matrix. Such results underpin the theory of optimization problems involving such functions. We answer the question of when a symmetric function of the eigenvalues allows a quadratic expansion around a matrix, and then the stronger question of when it is twice differentiable. We develop simple formulae for the most important nonsmooth subdifferentials of functions depending on the singular values of a real rectangular matrix argument and give several examples. The analysis of the above two classes of functions may be generalized in various larger abstract frameworks. In particular, we investigate how functions depending on the eigenvalues or the singular values of a matrix argument may be viewed as the composition of symmetric functions with the roots of *hyperbolic polynomials*. We extend the relationship between hyperbolic polynomials and *self-concordant barriers* (an extremely important class of functions in contemporary interior point methods for convex optimization) by exhibiting a new class of self-concordant barriers obtainable from hyperbolic polynomials.

Acknowledgements

I would like to state my deepest gratitude to my Ph.D. supervisor Adrian S. Lewis. Without his excellent mathematical knowledge and professional guidance, this work would not have existed in its present form. I am grateful to him for introducing me to the many areas of research treated in this thesis. I am extremely thankful to him for his professionalism, patience, and friendship. His wisdom and attitude will always be a guide to me. Thank you for believing in me.

I want to thank my committee members: Michael Overton, Levent Tunçel, Henry Wolkowicz, and Kirsten Morris for reading my work and providing many useful comments that improved it tremendously.

I want to thank Heinz Bauschke and Jonathan Borwein for useful suggestions on Chapter 4.

I would like to thank all my teachers for making me the person I am today.

I also like to thank my parents, sister, relatives, and all my good friends for the strength and support, the inspirations, and the good times they are giving me. I love you.

Contents

List of Figures

Chapter 1

Introduction

In this work we focus on nonsmooth analysis of the singular values of a general linear transformation between finite dimensional linear spaces, differentiability properties of the eigenvalues of a finite dimensional real symmetric linear operator and related matters. More precisely we will deal with *spectral* and *singular value* functions, that is, symmetric functions of the eigenvalues and absolutely symmetric functions of the singular values. (See Definition 6.3.2 and [52, Definition 4.2].) Even though the eigenvalues and the singular values are invariants of two seemingly different classes of matrices, symmetric and rectangular, there are some connections between the two sets of numbers. For example if X is an $n \times m$ rectangular matrix (say $n \leq m$) the singular values of X together with their negatives and a few additional zeros are precisely the eigenvalues of the larger symmetric matrix

$$R(X) = \begin{pmatrix} 0 & X \\ X^T & 0 \end{pmatrix}. \tag{1.1}$$

Also the square roots of the eigenvalues of XX^T are precisely the singular values of X, so in the case when X is a square symmetric matrix the singular values are just the absolute values of the eigenvalues. Despite these connections some results about the nonsmooth behaviour of singular values are not obvious consequences of the corresponding results for eigenvalues, as will be explained later in the introduction.

The spectrum of a general symmetric matrix can behave in extremely complicated ways. Generally when the entries of the matrix depend on free parameters, the difficulties increase with the number of parameters. The perturbation theory of the spectrum of a symmetric matrix depending on one parameter is laid out in detail in the now classical book by T. Kato [41]. In contrast we consider the eigenvalues of a matrix X while X varies freely over the Euclidean space of $n \times n$ real symmetric matrices S^n, and respectively the singular values of a free $n \times m$ real matrix from the Euclidean space $M_{n,m}$. We denote the eigenvalues of $X \in S^n$ (counting multiplicities) by $\lambda_1(X) \geq \lambda_2(X) \geq ... \geq \lambda_n(X)$, and the singular values of $X \in M_{n,m}$ $(n \leq m)$ by $\sigma_1(X) \geq \sigma_2(X) \geq ... \geq \sigma_n(X)$. It is well known that at matrices X that have repeated eigenvalues, say $\lambda_1(X) = \lambda_2(X)$, these eigenvalues are nondifferentiable with respect to X. That is, in order for λ_i to be differentiable at X we must have $\lambda_{i-1}(X) > \lambda_i(X) > \lambda_{i+1}(X)$. This realization brings us to the first important question that we must clarify: is there a better way of defining the n eigenvalue functions (maybe not by ordering them decreasingly) so that we do not loose smoothness? This question is emphasized by the following example. Consider the matrix

$$T(x) = \begin{pmatrix} 0 & x \\ x & 0 \end{pmatrix}$$

depending on one parameter x. At every x its set of eigenvalues is given by $\{\lambda_1(T(x)), \lambda_2(T(x))\} = \{|x|, -|x|\}$ where the functions λ_1 and λ_2 are clearly non-smooth at 0 (where we have repeated eigenvalues). On the other hand at every point x the set of the eigenvalues is also given by $\{\mu_1(x), \mu_2(x)\}$, where $\mu_1(x) = x$, $\mu_2(x) = -x$ for every x, and now these functions are smooth. This question has been completely answered by Rellich in [77] and the answer depends heavily on the degrees of freedom.

Theorem 1.0.1 (Rellich, 1953). *Assume $T(x)$ is $n \times n$ symmetric and continuously differentiable in an interval $I \subset \mathbb{R}$ of x. Then there exist n continuously differentiable functions $\mu_n(x)$ on I that represent the eigenvalues (counting multiplicities) of $T(x)$.*

More surprisingly, the above result is optimal in the sense that even if $T(x)$ is $\mathcal{C}^\infty$ in x the $\mu_n(x)$ need not be $\mathcal{C}^2$, see [93]. But in a final twist if $T(x)$ symmetric and analytic on an interval, then the $\mu_n(x)$ may also be chosen to be analytic on this interval. An equivalent of Rellich's theorem, when the matrix T depends on two or more parameters, is impossible. Consider for example the matrix

$$T(x) = \begin{pmatrix} x_1 & x_2 \\ x_2 & -x_1 \end{pmatrix},$$

where $x \in \mathbb{R}^2$, and assume that there is a neighbourhood U in $\mathbb{R}^2$ around 0 such that for every point $x \in U$ the set of eigenvalues of $T(x)$ is given by the smooth (at least differentiable) functions $\{\mu_1(x), \mu_2(x)\}$. Clearly for every x in $\mathbb{R}^2$ the eigenvalues of $T(x)$ are $\{\|x\|, -\|x\|\}$. Fix a nonzero point $\bar{x} \in U$ and without loss

of generality suppose that $\mu_1(\bar{x}) = \|\bar{x}\|$ and $\mu_2(\bar{x}) = -\|\bar{x}\|$. Now, take an arbitrary second nonzero point $\hat{x} \in U$ and connect it to $\bar{x}$ with smooth curve γ avoiding the origin. Moving from $\bar{x}$ towards $\hat{x}$ along γ both $\mu_1(x)$ and $\mu_2(x)$ will vary smoothly and neither will become 0. So their signs will stay the same, that is $\mu_1(\hat{x}) > 0$ and $\mu_2(\hat{x}) < 0$, and consequently $\mu_1(\hat{x}) = \|\hat{x}\|$ and $\mu_2(\hat{x}) = -\|\hat{x}\|$. Remembering that $\hat{x}$ was arbitrary we see that the last two equalities must hold for every $\hat{x}$ in U, but this is a contradiction because these functions are not smooth at the origin.

These difficulties suggest why in our discussion of the differentiability properties of eigenvalues and singular values we are going to use the broad theory of nonsmooth analysis. The fact that we are considering symmetric functions of the spectrum is not a restriction because $\lambda_i = \phi \circ \lambda$, where

$$\begin{aligned} &\phi(x) : \mathbb{R}^n \to \mathbb{R} \\ &\quad x \mapsto i^{\text{th}} \text{ largest element of } \{x_1, ..., x_n\}, \end{aligned}$$

and we have a similar expression for the i^{th} singular value (see Section 6.8). So nonsmooth results for such functions immediately have equivalents for the individual eigen- or singular values.

Why would somebody interested in optimization be interested in functions of the spectrum of linear operators? Some of the first concrete applications of the perturbation theory of eigenvalues were in quantum mechanics [80], [42] where results like those obtained in Section 4.4 were well known. The following two inequalities are essentially due to John von Neumann [90], who also made fundamental

contributions to quantum theory:

$$\operatorname{tr} X^T Y \leq \lambda(X)^T \lambda(Y), \text{ for any } X, Y \in S^n,$$

$$\operatorname{tr} X^T Y \leq \sigma(X)^T \sigma(Y), \text{ for any } X, Y \in M_{n,m}.$$

(Using the relationship between the eigenvalues and singular values described in the beginning one can see that each inequality quickly follows from the other.) For contemporary proofs of these inequalities, using an optimization approach, as well as necessary and sufficient conditions for equality see [52, Theorem 3.5] and Theorem 6.2.9.

More recently, spectrally defined functions have started coming up in various areas of applied variational mathematics: optimality criteria in experimental design theory [75], [83], barrier functions in matrix optimization [67], [48], matrix updates in quasi-Newton methods [22], [94], semidefinite programming [11], potential energy densities for isotopic elastic materials [16], etc. For a comprehensive account of the role of eigenvalues and spectral functions in modern optimization the reader may refer to [55]. The following are just a few examples of spectral functions with their corresponding symmetric functions that researchers in the above areas encounter. We start with an important function from convex analysis, [78, pp. 68,148-149].

$$X \in S^n \mapsto F(X) = \log(\operatorname{tr} e^X),$$
$$x \in \mathbb{R}^n \mapsto f(x) = \log(e^{x_1} + \cdots + e^{x_n}).$$

Next is the largest eigenvalue function, having the *first order statistic* (see [33] for

an explanation of the name) as its corresponding symmetric function:

$$X \in S^n \mapsto F(X) = \lambda_1(X),$$
$$x \in \mathbb{R}^n \mapsto f(x) = \max\{x_1, ..., x_n\}.$$

The following spectral function arises in the theory of optimal experimental design, [75]:

$$X \in S^n \mapsto F(X) = \begin{cases} \operatorname{tr} X^{-1}, & \text{if } X \text{ is positive definite} \\ +\infty, & \text{otherwise,} \end{cases}$$
$$x \in \mathbb{R}^n \mapsto f(x) = \begin{cases} \frac{1}{x_1} + \cdots + \frac{1}{x_n}, & \text{if } x_1 > 0, ..., x_n > 0 \\ +\infty, & \text{otherwise.} \end{cases}$$

The following spectral function is fundamental to the development in [68]: it is the standard self-concordant barrier on the convex cone of positive semidefinite matrices, and its corresponding symmetric function is the standard self-concordant barrier on the positive orthant of $\mathbb{R}^n$:

$$\begin{aligned} X \in S^n \mapsto F(X) &= -\log\det(X), \\ x \in \mathbb{R}^n \mapsto f(x) &= -\sum_{i=1}^{n} \log(x_i). \end{aligned} \tag{1.2}$$

The square of the Frobenius (Euclidean) norm of a symmetric matrix with corresponding symmetric function - the square of the Euclidean norm in $\mathbb{R}^n$ is an obvious

example:

$$X \in S^n \mapsto F(X) = \|X\|_2^2,$$
$$x \in \mathbb{R}^n \mapsto f(x) = x_1^2 + \cdots x_n^2.$$

The last example is update formulae for Quasi-Newton algorithms [72, p. 227]:

$$X \in S^n \mapsto F(X) = \begin{cases} \frac{\mathrm{tr}\,(X)}{n \det(X)}, & \text{if } X \text{ is positive definite} \\ +\infty, & \text{otherwise,} \end{cases}$$
$$x \in \mathbb{R}^n \mapsto f(x) = \begin{cases} \frac{\sum_{i=1}^n x_i}{n \prod_{i=1}^n x_i}, & \text{if } x_i > 0 \text{ for all } i \\ +\infty, & \text{otherwise.} \end{cases}$$

A big part of our work deals with the differentiability properties of functions F on the real vector space of symmetric matrices that are *orthogonally invariant*:

$$F(U^T A U) = F(A), \text{ for all } A \text{ symmetric and } U \text{ orthogonal.}$$

One can easily see ([49, Proposition 4.1]) that every orthogonally invariant function is the composition of a *symmetric* function on $\mathbb{R}^n$ and the eigenvalues of the matrix argument:

$$F(A) = (f \circ \lambda)(A),$$

where $\lambda(A) = (\lambda_1(A), ..., \lambda_n(A))$. As we mentioned above we call such functions F *spectral*. The spectral functions F are in one-to-one correspondence with the symmetric functions f. A lot of research in recent years shows that properties of

f are inherited by F and vice versa. The list is long. Let F and f be a pair of a spectral function and its corresponding symmetric function, and let C be a symmetric set in $\mathbb{R}^n$. Then, for example:

1. F is lower semicontinuous (l.s.c.) at A if and only if f is at $\lambda(A)$, [48].

2. F is l.s.c. and convex if and only if f is, [18], [48].

3. The symmetric function corresponding to the Fenchel conjugate of F is the Fenchel conjugate of f, [82], [48]. (A similar statement holds for the recession function of F, [82].)

4. F is pointed, has good asymptotic behaviour or is a barrier function on the set $\lambda^{-1}(C)$ if and only if f is on C, [82].

5. F is Lipschitz around A if and only if f is such around $\lambda(A)$, [49]

6. F is (continuously) differentiable at A if and only if f is at $\lambda(A)$, [49].

7. F is strictly differentiable at A if and only if f is at $\lambda(A)$, [49], [52]. (But this correspondence doesn't carry over for the Gâteaux derivative.)

8. If f is l.s.c. and convex then F is twice epi-differentiable at A relatively to Ω if and only if f is twice epi-differentiable at $\lambda(A)$ relative to $\lambda(\Omega)$, [86], where Ω is an arbitrary epi-gradient.

9. F is a polynomial of the entries of A if and only if f is a polynomial. This is a consequence of the Chevalley Restriction Theorem, [92, p. 143].

10. $F \in \mathcal{C}^\infty$ at $A \Leftrightarrow f \in \mathcal{C}^\infty$ at $\lambda(A)$, [17].

11. F is analytic at A if and only if f is at $\lambda(A)$, [88].

On the other hand a variety of smooth and nonsmooth objects of F can be expressed in terms of the corresponding objects of f. For example, a description of the convex subdifferential of F is given in [48]; the Clarke subdifferential is given in [49],[52]; the regular, approximate, and horizon subdifferentials are given in [52]; the second order epi-derivative of a convex F is given in [86].

The results we present in Chapter 4 and Chapter 5 stay in some sense (mathematically) between the results in points 6 and 10 from the above list. Indeed, in Chapter 4 we show that F is twice differentiable at A if and only if f is twice differentiable at $\lambda(A)$, and then we show even more, that the Hessian of F is continuous at A if and only if the Hessian of f is continuous at $\lambda(A)$, that is, $F \in \mathcal{C}^2 \Leftrightarrow f \in \mathcal{C}^2$. We also give a concise and easy-to-use formula for the Hessian (see Theorem 4.2.2 and Theorem 4.2.3), while the results in [88] are rather implicit.

Second order differentiability is important for optimization because of many reasons. A few of its applications are Newton's method, second order necessary optimality conditions, second order sufficient optimality conditions, and modern interior point methods.

Several authors have recently been concerned with second order spectral analysis. For example, A. Seeger, in a related work, expressed his doubts that the $\mathcal{C}^2$-property of f is inherited by F, (see the end of Section 11 in [82]). Also, H. Bauschke and J. Borwein, in [5], pose a conjecture about the joint convexity of the Bregman distance associated with a spectral function, and in their opinion the $\mathcal{C}^2$-property of the spectral function and the form of its Hessian will play a crucial

role for solving it. The results in Chapter 4 are a necessary step towards answering another conjecture posed by L. Tunçel [55]: "if the function f is a self-concordant barrier, is the same true of the spectral function F?". An example supporting the conjecture is (1.2). One reason why answering this conjecture may be interesting is given in [89, Chapter 8]. It is shown there that the spectral barrier, F, from 1.2 has the same barrier parameter as f. If this property is 'approximately' preserved in general then one will be able to obtain self-concordant barrier functions with 'small' parameters on sets with high dimension using the existing lower dimensional examples. (It is well known that the barrier parameter directly affects the speed of convergence of the underlying interior point method.)

Next, in Chapter 5 we treat a related question and show that a spectral function F has quadratic expansion at A if and only if f has one at $\lambda(A)$. Many functions have quadratic expansions. For example a theorem of Alexandrov [1] states that every finite, convex function on an open subset of $\mathbb{R}^n$ has quadratic expansion at almost every point. Notice that it is not necessary for a function to be twice differentiable in order to have quadratic expansion. For example the function

$$f(x) = \begin{cases} x^3 \sin(1/x), & \text{if } x \neq 0 \\ 0, & \text{if } x = 0 \end{cases} \tag{1.3}$$

has quadratic expansion around $x = 0$ but is not twice differentiable there. On the other hand being twice differentiable at x implies having quadratic expansion at x.

Concluding the topic of differentiability properties of spectral functions we give a final glimpse at a part of the picture up to this moment. We present schematically

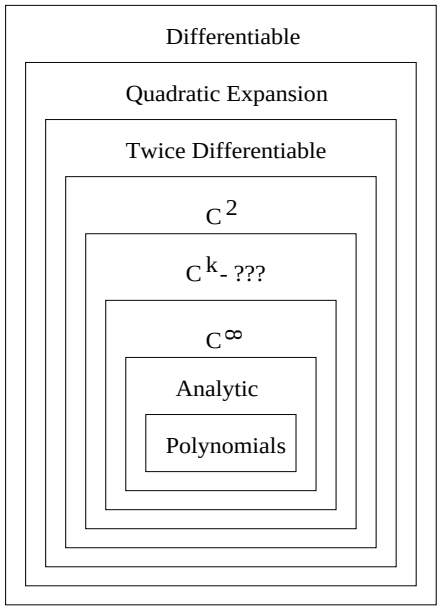

Figure 1.1: Some differentiability properties

on Figure 1.1, a 'chain' of gradually weaker properties that are carried from the symmetric function to its spectral equivalent and vice versa. The reader should refer to the list above for a full account. (Note that the property of being $\mathcal{C}^1$ is not in the 'chain' because it can not be fitted between the property of being differentiable and the property of having quadratic expansion. For the figure we tried to select properties that will make the 'chain' as long as possible.)

Another major theme in our work is the nonsmooth analysis of singular values. In particular we consider the composition of an absolutely symmetric function (see Definition 6.3.2) with the singular value map. We call such functions on a rectangular matrix argument *singular value* functions. One of the first results about singular value functions is the characterization, by von Neumann in [90], of all

unitary invariant norms, that is, norms $\|\cdot\|$ on $M_{n,m}$ such that

$$\|UXV\| = \|X\|, \text{ for all } X \in M_{n,m} \text{ and all unitary matrices } U \in M_n,\ V \in M_m.$$

He showed that such norms are precisely the compositions of an absolutely symmetric function that is also a norm on $\mathbb{R}^n$ (such functions are knows as *symmetric gauge* functions) with the singular value map.

The singular values are strongly connected with some matrix-optimization problems. For example, if we want to find the nearest rank k matrix to a given matrix X with respect to a given orthogonally invariant norm, $\|\cdot\|$, then we form the singular value decomposition of X, $X = U^T\Sigma V$ and let $Y = U^T\Lambda V$, where $\Sigma = \text{Diag}\,(\sigma_1(X), \sigma_2(X), ..., \sigma_n(X))$, and $\Lambda = \text{Diag}\,(0, ..., 0, -\sigma_{k+1}(X), ..., -\sigma_n(X))$. The matrix $X+Y$ is the nearest rank k matrix to X, [36, Section 7.4]. In particular, if $\|\cdot\|$ is the spectral norm on $M_{n,m}$ (that is, $\|X\| = \sqrt{\lambda_1(X^TX)} = \sigma_1(X)$), then $\sigma_{k+1}(X)$ is the distance between X and the nearest rank k matrix. Another curious minimization fact, that holds for every unitary invariant norm, is:

$$\|\sigma(X) - \sigma(Y)\| = \min\{\|X - U^TYV\| \,|\, U \in M_n, V \in M_m \text{ orthogonal }\}.$$

(It is easy to prove that the left hand side above is greater than or equal to the right hand side.)

In Chapter 6 we derive the main tools from nonsmooth analysis for singular value functions (see [79], [65], [39], [40]). It can be viewed as a continuation of [47]. Its development follows closely that in [52] and in the process we derive some

of the background tools from [49] in the context of singular values. We go a few steps further than [52], the additional results being the formula for the Clarke subdifferential when the singular value function is only lower semicontinuous and the formula for the proximal subdifferential. For a treatment of related singular values topics see [81].

One may ask if it would be easier to calculate the subdifferentials of each σ_i and then apply the chain rule to $f \circ \sigma$. One will need to apply Theorem 10.49 from [79] which in our context says:

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

Similar formulae hold for the regular, and horizon subdifferentials as well. The problems with this formula are, first, it is not clear whether calculating each $\partial(y^T\sigma)(X)$ will be a simpler task, and second, it is only a one-sided inclusion. The conditions for equality require strong assumptions. In our derivations we dispense with these assumptions throughout, to arrive at compact, closed form expressions that do not seem easy to derive from the above formula even when it holds with equality.

One may think that another way of deducing the results in this chapter may be as corollaries of the corresponding results in [52], using the connection between the eigenvalues and the singular values given in (1.1). One may decide to consider the function

$$(\tilde{f} \circ \lambda) \circ R(X),$$

where $\tilde{f}$ is appropriate modification of our absolute symmetric function f. But

whatever the choice of $\tilde{f}$ is, the difficulties listed in the above paragraph may haunt us here too. Finally, even if we overcome all of these difficulties, the nice algebraic structure stemming from the singular value decomposition, and so nicely evident in our formulae of the subdifferentials of $f \circ \sigma$ (see Theorem 6.5.1 and its analogies), may be unrecognizably obscured.

We now steer towards the work done in Chapter 2, where we investigate a unifying framework for some of the results in this thesis. As we mentioned earlier, for a symmetric gauge g (necessarily convex) and a symmetric, convex f on $\mathbb{R}^n$ the composite functions

$$X \in M_{n,m} \mapsto g(\sigma(X)), \tag{1.4}$$

$$X \in S^n \mapsto f(\lambda(X)) \tag{1.5}$$

are convex. (For $g(\sigma(X))$ this is due to von Neumann and for $f(\lambda(X))$ this is due to Davis.) Not only the convexity of g and f is preserved after the composition, but some important convex analytic notions for the composition are easily expressed through the corresponding notions for g and f. Thus for example, the Fenchel conjugate of the function (1.5) is given elegantly by

$$(f \circ \lambda)^* = f^* \circ \lambda,$$

and the analogous result for $g \circ \sigma$ was shown by von Neumann. These analogies between the two classes of functions are not accidental. In [50] Lewis gives a set of axioms and abstractly derives the convexity properties of a special invariant class

of functions that generalizes both (1.4) and (1.5). Then in [53] he uses semisimple Lie theory and the Kostant convexity theorem to generalize these properties again. In Chapter 2 we give a surprising new approach towards uniting the above type of convexity results via properties of the roots of *hyperbolic polynomials*.

The theory of hyperbolic polynomials has its origins in partial differential equations, and is connected with the well-posedness of the Cauchy problem. We briefly give here a few historical notes about this problem. For more information see Sections 12.3-12.6 in [34], [35]. Let $p : \mathbb{R}^n \to \mathbb{R}$ be a homogeneous polynomial of degree m. To every p corresponds a partial differential operator $p(D)$, obtained from p by replacing x_k with $\frac{-i\partial}{\partial x_k}$. For example, to the polynomial

$$p(x_1, ..., x_n) = x_1^2 - \sum_{k=2}^{n} x_k^2$$

corresponds the operator

$$p(D) = -\frac{\partial^2}{\partial x_1^2} + \sum_{k=2}^{n} \frac{\partial^2}{\partial x_k^2}.$$

Then the Cauchy problem is formulated as follows.

Definition 1.0.2 (Cauchy Problem). *Is there a solution u (a distribution, generalized function) to the equation*

$$p(D)u = f,$$

with support $\operatorname{supp}(u) \subseteq H$ *for a given function* $f \in \mathcal{C}_0^\infty(H)$, *where* $H = \{x \in$

$\mathbb{R}^n | \langle x, d \rangle \geq 0\}$, *and* $d \neq 0$ *is a direction in* $\mathbb{R}^n$ *such that* $p(d) \neq 0$*?*

It turns out that the Cauchy problem has a solution (in fact unique) for any such f if and only if p is a hyperbolic polynomial, defined below:

Definition 1.0.3. *A homogeneous polynomial* $p : \mathbb{R}^n \to \mathbb{R}$ *is called hyperbolic with respect to a direction* $d \in \mathbb{R}^n$ *if* $p(d) \neq 0$ *and the polynomial*

$$t \mapsto p(x + td),$$

has only real roots for any x*.*

The roots $\lambda_1(x) \geq \lambda_2(x) \geq \cdots \geq \lambda_m(x)$ of $t \mapsto p(x - td)$ are called roots or eigenvalues of the hyperbolic polynomial. The name eigenvalues comes from the fact that $p(X) = \det(X)$, $X \in S^n$ is a hyperbolic polynomial and its roots are the eigenvalues of X.

In Chapter 2 we use a result by Gårding [24], saying that the largest root, $\lambda_1(x)$, is always a convex function of x, to prove a generalization of Davis's theorem, that any symmetric convex function of the roots $\lambda(x)$ of a hyperbolic polynomial is convex. This result then allows us to derive many elegant inequalities in a unified fashion. A Fenchel conjugation formula that subsumes the corresponding formulae for (1.4) and (1.5), is also presented. There is a long section on examples, and for each example we go in detail over every property of the hyperbolic polynomials that interests us. Finally in Section 2.6.8 we use one particular hyperbolic polynomial to rederive von Neumann's singular value example (1.4).

In 1988, Nesterov and Nemirovskii developed a general, polynomial time frame-

work for convex programming problems, presented in their monograph [68]. This framework for interior point methods relies on the notion of *self-concordant barrier functions* (see the definition in Section 3.1). These functions are special, convex penalty functions which intricately regulate their own behaviour and growth. One of the most important results in Nesterov and Nemirovskii [68] is that a self-concordant barrier function exists for every open convex set. They construct such a function, called the *universal barrier*. The parameter ϑ (on which every self-concordant function depends) in their construction has magnitude big-O of the dimension of the domain space. Because ϑ plays an important role for the convergence speed of the underlying interior point method the question of finding computable barrier functions with small parameters is of fundamental interest.

In Chapter 3 we investigate a relationship between the hyperbolic polynomials and self-concordant barriers. Every hyperbolic polynomial $p(x)$ with roots $\lambda_i(x)$ has an associated closed convex *hyperbolicity cone* which is defined as

$$\{x \in \mathbb{R}^n | \lambda_i(x) \geq 0 \text{ for all } i\}.$$

(Actually the convexity of the above cone is equivalent to the convexity of $\lambda_1(x)$ - the largest root of $p(x)$.) Güler was the first to observe the connection between hyperbolic polynomials and convex optimization. He showed [25] that the hyperbolicity cone is a good environment for the modern interior point algorithms [68] with a natural self-concordant barrier on it, $-\log p(x)$, with parameter m - the degree of homogeneity of p.

A crucial example of a self-concordant barrier in contemporary optimization

is the function $-\log\det(\cdot)$, which is an m-self-concordant barrier for the cone of $m \times m$ symmetric positive definite matrices, a set of dimension $m(m+1)/2$ (see [68]). The main result in Chapter 3 is that $-m\log(p(x)-1)$ is a 'shifted' m^2-self-concordant barrier on a corresponding subset of the hyperbolicity cone of p. As a consequence we get for example, that $-m\log(\det(\cdot)-1)$ is a 'shifted' m^2-self-concordant barrier on a corresponding subset of the positive definite cone. Even though our function, $-m\log(p(x)-1)$, seems 'close' to Güler's, $-\log p(x)$, our proof turns out to be a lot more complicated than the proof of Theorem 4.1 in [27]. Furthermore, in the last section of this chapter we show that our result cannot be deduced as an elementary consequence (in some sense) of Güler's result, that $-\log p(x)$ is a self-concordant barrier.

Another way to look at spectral and singular value functions is as functions on a symmetric matrix argument, or rectangular matrix argument respectively, invariant under a closed group of orthogonal transformations of the linear space S^n, or $M_{n,m}$ respectively: that is, for all X in the domain of F we have

$$F \text{ - spectral function } \Leftrightarrow F(U^TXU) = F(X),\ \forall U \in O(n),$$
$$F \text{ - singular value function } \Leftrightarrow F(U_n^TXU_m) = F(X),\ \forall (U_n, U_m) \in O(n) \times O(m).$$

In Chapter 7 we treat a class of functions having a different invariant property. We consider functions on $\mathbb{R}^n \times \mathbb{R}$ invariant under orthogonal transformations $(U, 1)$, that is, for all (x,t) in the domain of such a function we have

$$g : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$$

$$g(Ux,t) = g(x,t),\ \forall U \in O(n).$$

We call functions having this property *Lorentz invariant* functions because $(U,1)$ are all the orthogonal transformations that preserve the Lorentz cone, $\{(x,t) \in \mathbb{R}^n \times \mathbb{R} | t \geq \|x\|\}$. Such functions can be decomposed as $g = f \circ \beta$, where $f : \mathbb{R}^2 \to \mathbb{R}$ is symmetric and

$$\beta(x,t) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^2$$
$$\beta(x,t) = \frac{1}{\sqrt{2}}(t + \|x\|, t - \|x\|).$$

The mapping β allows several interpretations. It may either be viewed as the "eigenvalue" map of the roots of a hyperbolic polynomial, when the direction of hyperbolicity is taken to be $(\sqrt{2}, 0, ..., 0)$, see Example 2.6.5, or it can be viewed as the eigenvalue map of an element in the *Jordan algebra of quadratic forms* with respect to a certain Jordan frame, see [95, Example 8.3.12].

For Lorentz invariant functions we derive all the smooth and nonsmooth properties that interested us in the previous chapters. We want to emphasize that the interest here is not necessarily that the results are crucial in their own right, but to draw out the algebraic analogies with the earlier results. These analogies suggest that a unified setting should exist. Deeper investigations into a generalization using Jordan algebras may be a point of a future research, see Chapter 8.

In conclusion we would like to say that Chapter 2 is based on a joint paper with H. Bauschke, O. Güler and A. Lewis [6], to appear in the Canadian Journal of Mathematics. A paper based on Chapter 3 is submitted to Mathematical Program-

ming, Series A, [57]. A paper based on Chapter 4 is submitted to SIAM Journal of Matrix Analysis, [56]. A paper based on Chapter 5 is submitted to Linear Algebra and Its Applications, [58]. Papers based on Chapter 6 and Chapter 7 are still in preparation for submission.

Chapter 2

Hyperbolic Polynomials

2.1 Notation

We write $\mathbb{R}^m_{++}$ (resp. $\mathbb{R}^m_+$) for the set $\{u \in \mathbb{R}^m : u_i > 0, \forall i\}$ (resp. $\{u \in \mathbb{R}^m : u_i \geq 0, \forall i\}$. The *closure* (resp. *boundary*, *convex hull*, *linear span*) of a set S is denoted $\operatorname{cl} S$ (resp. $\operatorname{bd} S$, $\operatorname{conv} S$, $\operatorname{span} S$). A *cone* is a nonempty set that contains every nonnegative multiple of all its members; it thus always contains 0. If $u \in \mathbb{R}^m$ then by either $\bar{u}$ or $u_\downarrow$ we will denote the vector u with its coordinates arranged decreasingly; also, $\bar{U} = U_\downarrow := \{\bar{u} : u \in U\}$, for every subset U of $\mathbb{R}^m$. If $u \in \mathbb{R}^m$, then $|u|$ will denote $(|u_1|, ..., |u_m|)$. The *transpose* of a matrix (or vector) A is denoted A^T. The *identity* matrix or map is written I. Suppose Y is an arbitrary Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and $h : Y \to [-\infty, +\infty]$ is convex, then h^* (resp. ∂h, ∇h, $\operatorname{dom} h$) stands for the *Fenchel conjugate* (resp. *subdifferential map*, *gradient map*, *domain*) of h. (Rockafellar's monograph [78] is the standard reference for these notions from convex analysis.) Higher order derivatives are denoted by

$\nabla^k h$. If $U \subseteq X$, then the *positive polar cone* is $U^+ := \{x \in X : \langle x, U\rangle \geq 0\}$. If A is a linear operator between Euclidean spaces, then its *transpose* is written A^T. The *range* of a map λ is denoted by $\operatorname{ran}\lambda$. Finally, if A, B are two subsets of X, then $d(A,B) := \inf\{\|a-b\| : a \in A,\, b \in B\}$ is the *distance* between A and B.

2.2 Background

We assume throughout the this chapter that

X is a finite-dimensional real vector space.

This section contains a selection of important facts on hyperbolic polynomials from Gårding's fundamental work [24], and a deep inequality on elementary symmetric functions.

For all missing proofs and references the reader should refer to our paper [6].

2.2.1 Hyperbolic polynomials and eigenvalues

Definition 2.2.1 (Homogeneous Polynomial). *Suppose p is a nonconstant polynomial on X and m is a positive integer. Then p is* **homogeneous of degree** *m, if $p(tx) = t^m p(x)$, for all $t \in \mathbb{R}$ and every $x \in X$.*

Definition 2.2.2 (Hyperbolic Polynomial). *Suppose that p is a homogeneous polynomial of degree m on X and $d \in X$ with $p(d) \neq 0$. Then p* **is hyperbolic with respect to** *d, if the polynomial $t \mapsto p(x+td)$ (where t is a scalar) has only real zeros, for every $x \in X$.*

Definition 2.2.3 ("Eigenvalues and Trace"). *Suppose p is hyperbolic with respect to $d \in X$ of degree m. Then for every $x \in X$, we can write*

$$p(x + td) = p(d) \prod_{i=1}^{m} (t + \lambda_i(x))$$

and assume without loss of generality that $\lambda_1(x) \geq \lambda_2(x) \geq \cdots \geq \lambda_m(x)$. The corresponding map $X \to \mathbb{R}^m_\downarrow : x \mapsto (\lambda_1(x), \ldots, \lambda_m(x))$ is denoted by λ and called the **eigenvalue map (with respect to p and d)**. *We say that $\lambda_i(x)$ is the i^{th}* **largest eigenvalue of x (with respect to p and d)** *and define* **the sum of the k largest eigenvalues** *by $\sigma_k := \sum_{i=1}^{k} \lambda_i$, for every $1 \leq k \leq m$. The function σ_m is called the* **trace**.

The eigenvalues $\{\lambda_i(x)\}$ are thus the roots of the polynomial $t \mapsto p(x - td)$. It follows readily that the trace σ_m is linear (see also the paragraph following Proposition 2.2.19).

Unless stated otherwise, we assume throughout the chapter that

p is a hyperbolic polynomial of degree m with respect to d,
with eigenvalue mapλ and $\sigma_k := \sum_{i=1}^{k} \lambda_k$,

for every $1 \leq k \leq m$. The notions "eigenvalues" and "trace" are well-motivated by the the following example.

The Hermitian matrices. Let X be the real vector space of the $m \times m$ Hermitian matrices and $p := \det$. Then p is hyperbolic of degree m with respect to $d := I$ and λ maps $x \in X$ to its eigenvalues, arranged decreasingly. Thus for every $1 \leq k \leq m$, the function σ_k is indeed the sum of the k largest eigenvalues and σ_m is the

(ordinary) trace.

As we go, we will point out what some of the results become in the important case of the *Hermitian matrices*. Details and further examples are provided in Section 2.6.

We now introduce the notion of isomorphic triples, which will simplify the analysis of homogeneous polynomials in Section 2.6 considerably.

Definition 2.2.4. *Suppose p (resp. q) is a homogeneous polynomial on X (resp. Y) and $d \in X$ (resp. $e \in Y$). If there exists a linear one-to-one map Φ from X onto Y with $p = q \circ \Phi$ and $\Phi(d) = e$, then we say that the triple (X, p, d) is* **isomorphic** *to (Y, q, e) (by Φ), and we write $(X, p, d) \simeq (Y, q, e)$.*

It is clear that the binary operation $\simeq$ defines an equivalence relation on all triples. The following basic properties are easy to verify.

Proposition 2.2.5. *Suppose (X, p, d) is isomorphic to (Y, q, e) by Φ. Then:*

1. *The degrees of p and q coincide.*
2. *p is hyperbolic with respect to d if and only if q is hyperbolic with respect to e.*
3. *If p (resp. q) is hyperbolic with respect to d (resp. e) with corresponding eigenvalue map λ (resp. μ), then $\lambda = \mu \circ \Phi$.*

Many examples of hyperbolic polynomials can be obtained as described below.

Proposition 2.2.6.

1. *If q is hyperbolic with respect to the same d, then so is pq.*

2. *If $m > 1$, then $q(x) := \frac{d}{dt}p(x+td)\big|_{t=0} = \nabla p(x)[d]$ is hyperbolic with respect to d.*

3. *If Y is a subspace of X and $d \in Y$, then the restriction $p|_Y$ is hyperbolic with respect to d.*

The technique of Proposition 2.2.6.(2) has a higher order analog, see Proposition 2.2.19 below. Given a hyperbolic polynomial on $\mathbb{R}^n$, we can construct a related one on $\mathbb{R}^{n-1}$ as follows.

Proposition 2.2.7. *Suppose p is hyperbolic with respect to $d \in \mathbb{R}^n$ with eigenvalue map λ. Assume that $d_i \neq 0$ and define q on $\mathbb{R}^{n-1}$ by*

$$q(y_1, \ldots, y_{n-1}) = p\big(y_1, \ldots, y_{n-1}, \tfrac{y_i}{d_i}d_n\big).$$

Then q is hyperbolic with respect to $e := (d_1, \ldots, d_{n-1})$ and its eigenvalue map μ satisfies $\mu(y_1, \ldots, y_{n-1}) = \lambda\big(y_1, \ldots, y_{n-1}, \tfrac{y_i}{d_i}d_n\big)$.

Proof. Straightforward. □

The following property of the eigenvalues is well-known [24, Equation (2)] and easily verified.

Proposition 2.2.8. *For all $r, s \in \mathbb{R}$ and every $1 \leq i \leq m$:*

$$\lambda_i(rx + sd) = \begin{cases} r\lambda_i(x) + s, & \text{if } r \geq 0; \\ r\lambda_{m+1-i}(x) + s, & \text{otherwise.} \end{cases}$$

It follows that the eigenvalue map λ is *positively homogeneous* ($\lambda(tx) = t\lambda(x)$, for all $t \geq 0$ and every $x \in X$) and continuous (the zeros of a polynomial are continuous with respect to the coefficients; see, for instance, [73, Appendix A]).

Gårding showed that the largest eigenvalue map is *sublinear*, that is, positively homogeneous and convex.

Theorem 2.2.9 (Gårding). *The largest eigenvalue map λ_1 is sublinear.*

We continue with the example we started on page 23

The Hermitian matrices (continued). It is well-known that the largest eigenvalue map is convex in this case; see, for instance, [32].

2.2.2 Hyperbolicity cone

Definition 2.2.10 (Hyperbolicity Cone). *The* **hyperbolicity cone of p with respect to** *d, written $C(d)$ or $C(p, d)$, is the set $\{x \in X : p(x + td) \neq 0, \forall t \geq 0\}$.*

We can write the hyperbolicity cone in terms of the eigenvalue map as follows.

Proposition 2.2.11. *$C(d) = \{x \in X : \lambda_m(x) > 0\}$. Hence $C(d)$ is an open convex cone that contains d with closure $\operatorname{cl} C(d) = \{x \in X : \lambda_m(x) \geq 0\}$. If $c \in C(d)$, then p is hyperbolic with respect to c and $C(c) = C(d)$.*

Proof. Gårding [24, Section 2]. □

Remark 2.2.12. *Note that $\lambda_m(x) > 0$ if and only if $\lambda_1(-x) < 0$ by Proposition 2.2.8. Hence Gårding's result (Theorem 2.2.9) implies the convexity of $C(d)$. In fact, the two results are equivalent. To see why, suppose first $C(d)$ is a convex*

cone. Fix x and y in X and observe that $x - \lambda_m(x)d$ and $y - \lambda_m(y)d$ both belong to $\mathrm{cl}\, C(d)$. *By assumption,* $(x+y) - (\lambda_m(x) + \lambda_m(y))d \in \mathrm{cl}\, C(d)$. *On the other hand, the smallest t such that $(x+y) + td$ belongs to* $\mathrm{cl}\, C(d)$ *is $-\lambda_m(x+y)$. Altogether, $\lambda_m(x) + \lambda_m(y) \leq \lambda_m(x+y)$ and the concavity of λ_m (or convexity of λ_1) follows.*

Definition 2.2.13 (Complete Hyperbolic Polynomial). *p is* **complete** *if*

$$\{x \in X : \lambda(x) = 0\} = \{0\}.$$

The following result, which follows easily from Proposition 2.2.5.(3), considers the concepts just introduced for isomorphic triples.

Proposition 2.2.14. *Suppose (X, p, d) is isomorphic to (Y, q, e) by Φ. Then:*

1. *$C(q, e) = \Phi(C(p, d))$.*
2. *p is complete if and only if q is.*

Proposition 2.2.15. *Suppose p is hyperbolic with respect to d, with corresponding eigenvalue map λ and hyperbolicity cone $C(d)$. Then*

$$\begin{aligned}\{x \in X : \lambda(x) = 0\} &= \{x \in X : x + C(d) = C(d)\} \\ &= \{x \in X : p(tx + y) = p(y), \forall y \in X, \forall t \in \mathbb{R}\}.\end{aligned}$$

Consequently, $\{x \in X : \lambda(x) = 0\} = \mathrm{cl}\, C(d) \cap (-\mathrm{cl}\, C(d))$. *Therefore, p is complete if and only if $cl C(d)$ is a pointed cone.*

It is always possible to find a restriction of p that is complete: indeed, $d \notin \{x \in X : \lambda(x) = 0\}$; consequently, if Y is any subspace of X which contains d and is

algebraically complemented to $\{x \in X : \lambda(x) = 0\}$, then $p|_Y$ is hyperbolic with respect to d (Proposition 2.2.6.(3)) and complete.

Example 2.2.16. *We let $X = \mathbb{R}^n$, $p(x) = \sum_j x_j$ and $d = (1, 1, \ldots, 1)$ in X. Then p is hyperbolic with respect to d of degree $m = 1$ and $\lambda(x) = \frac{1}{n}\sum_{j=1}^n x_j$. It follows that p is complete only when $n = 1$.*

The Hermitian matrices (continued). The hyperbolicity cone of $p = \det$ with respect to $d = I$ is the set of all positive definite matrices. The polynomial $p = \det$ is complete, since every nonzero Hermitian matrix has at least one nonzero eigenvalue.

2.2.3 Elementary symmetric functions

Definition 2.2.17 (Symmetric Function). *A function f on $\mathbb{R}^m$ is* **symmetric**, *if $f(u_1, ..., u_m) = f(u_{\pi(1)}, ..., u_{\pi(m)})$, for all permutations π of $\{1, \ldots, m\}$ and every $u \in \mathbb{R}^m$.*

Definition 2.2.18 (Elementary Symmetric Functions). *For any given integer $k = 1, 2, \ldots, m$, the map $E_k : \mathbb{R}^m \to \mathbb{R}$ defined by $u \mapsto \sum_{i_1 < \cdots < i_k} \prod_{l=1}^k u_{i_l}$ is called the k^{th}* **elementary symmetric function** *on $\mathbb{R}^m$. We also set $E_0 := 1$.*

Proposition 2.2.19. *For every $x \in X$ and all $t \in \mathbb{R}$,*

$$p(x + td) = p(d) \prod_{i=1}^{m} (t + \lambda_i(x)) = p(d) \sum_{i=0}^{m} E_i(\lambda(x)) t^{m-i}$$

and for every $0 \leq i \leq m$,

$$p(d)E_i(\lambda(x)) = \frac{1}{(m-i)!}\nabla^{m-i}p(x)[\underbrace{d, d, \ldots, d}_{m-i \text{ times}}].$$

If $1 \leq i \leq m$, *then* $E_i \circ \lambda$ *is hyperbolic with respect to* d *of degree* i.

Proposition 2.2.19 gives a very transparent proof of the linearity of trace: indeed, $\sigma_m = E_1 \circ \lambda$ is a homogeneous (hyperbolic) polynomial of degree 1 and hence linear.

We also note that the elementary symmetric functions themselves are hyperbolic:

Example 2.2.20. *Let* $X = \mathbb{R}^m$ *and* $d = (1, 1, \ldots, 1) \in \mathbb{R}^m$. *Then for every* $1 \leq k \leq m$, *the* k^{th} *elementary symmetric function* E_k *is hyperbolic of degree* k *with respect to* d.

2.2.4 An inequality in elementary symmetric functions

The following inequality was discovered independently by McLeod [62] and by Bullen and Marcus [13, Theorem 3]. We are interested in it mainly because of the two corollaries that follow it.

Proposition 2.2.21. *(McLeod, 1959; Bullen and Marcus, 1961) Suppose* $1 \leq k \leq l \leq m$ *and* $u, v \in \mathbb{R}^m_{++}$. *Set* $q := (E_l/E_{l-k})^{1/k}$. *Then*

$$q(u+v) > q(u) + q(v),$$

unless u *and* v *are proportional or* $k = l = 1$, *in which case we have equality.*

Bullen and Marcus's proof relies on an inequality by Marcus and Lopes ([59, Theorem 1], which is the case $k = 1$ in Proposition 2.2.21. (Proofs can also be found in [7, Theorem 1.16], [14, Section V.4], and [64, Section VI.5].)

We record two interesting consequences of Proposition 2.2.21.

Corollary 2.2.22. *(Marcus and Lopes's [59, Theorem 2]) The function $-E_m^{1/m}$ is sublinear on $\mathbb{R}_+^m$, and it vanishes on $\operatorname{bd}\mathbb{R}_+^m$.*

Recall that a function h is called *logarithmically convex*, if $\log(h)$ is convex. The function q in Proposition 2.2.21 is concave ("strictly modulo rays"), which yields logarithmic and strict convexity of $1/q$:

Proposition 2.2.23. *Suppose q is a function defined on $\mathbb{R}_{++}^m$. Consider the following properties:*

(i) the range of q is contained in $(0, +\infty)$;

(ii) $q(ru) = rq(u)$, for all $r > 0$ and every $u \in \mathbb{R}_{++}^m$;

(iii) $q(u + v) \geq q(u) + q(v)$, for all $u, v \in \mathbb{R}_{++}^m$;

(iv) if $u, v \in \mathbb{R}_{++}^m$ with $q(u + v) = q(u) + q(v)$, then $v = \rho u$, for some $\rho > 0$.

Suppose q satisfies (i)–(iii). Then $1/q$ is logarithmically convex. If furthermore (iv) holds, then $1/q$ is strictly convex.

Corollary 2.2.24. *Suppose $1 \leq k \leq l \leq m$. Then the function $(E_{l-k}/E_l)^{1/k}$ is symmetric, positively homogeneous, and logarithmically convex. Moreover, the function is strictly convex on $\mathbb{R}_{++}^m$ unless $l = 1$ and $m \geq 2$.*

2.3 Convexity

This section is the core of the chapter and that is why we are going to include the proofs of the main results here.

2.3.1 Sublinearity of the sum of the largest eigenvalues

Theorem 2.3.1. *Suppose q is a homogeneous symmetric polynomial of degree n on $\mathbb{R}^m$, hyperbolic with respect to $e := (1, 1, \dots, 1) \in \mathbb{R}^m$, with eigenvalue map μ. Then*

$$q \circ \lambda$$

is a hyperbolic polynomial of degree n with respect to d and its eigenvalue map is $\mu \circ \lambda$.

Proof. For simplicity, write $\tilde{p}$ for $q \circ \lambda$.

Step 1: $\tilde{p}$ is a polynomial on X. Indeed, since $q(y)$ is a symmetric polynomial on $\mathbb{R}^m$, it is (by, e.g., [38, Proposition V.2.20.(ii)]) a polynomial in $E_1(y), \dots, E_m(y)$. On the other hand, by Proposition 2.2.19, $E_i \circ \lambda$ is hyperbolic with respect to d of degree i, for $1 \leq i \leq m$. Altogether, $\tilde{p}(x) = q(\lambda(x))$ is a polynomial on X.

Step 2: $\tilde{p}$ is homogeneous of degree n. Indeed, since q is symmetric and homogeneous, and in view of Proposition 2.2.8, we obtain $\tilde{p}(tx) = q(\lambda(tx)) = t^n \tilde{p}(x)$, for all $t \in \mathbb{R}$ and every $x \in X$.

Step 3: $\tilde{p}(d) \neq 0$. Again using Proposition 2.2.8, we have $\tilde{p}(d) = q(\lambda(d)) = q(e) \neq 0$.

Step 4: $\tilde{p}$ is hyperbolic with respect to d. Using once more Proposition 2.2.8, we write for every $x \in X$ and all $t \in \mathbb{R}$:

$$\tilde{p}(x+td) = q(\lambda(x+td)) = q(\lambda(x)+te) = q(e)\prod_{k=1}^{n}(t+\mu_k(\lambda(x))).$$

□

The next example is easy to check.

Example 2.3.2. *Fix $1 \leq k \leq m$, set $e := (1,1,\ldots,1) \in \mathbb{R}^m$, and let*

$$q(u) := \prod_{1 \leq i_1 < i_2 < \cdots < i_k \leq m} \sum_{l=1}^{k} u_{i_l}.$$

Then q is a homogeneous symmetric polynomial on $\mathbb{R}^m$ of degree $\binom{m}{k}$, hyperbolic with respect to e, and its eigenvalues are $\{\frac{1}{k}\sum_{l=1}^{k} u_{i_l} : 1 \leq i_1 < i_2 < \cdots < i_k \leq m\}$. In particular, the largest eigenvalue of q is the weighted sum of the k largest components of u.

We now present our main result, the generalization of Theorem 2.2.9: the sum of the largest eigenvalues is sublinear. This readily implies local Lipschitzness of each eigenvalue map (see also [91]).

Corollary 2.3.3. *For every $1 \leq k \leq m$, the function σ_k is sublinear and λ_k is locally Lipschitz.*

Proof. Fix $1 \leq k \leq m$, define q as in Example 2.3.2, and consider $\tilde{p} := q \circ \lambda$. By Theorem 2.3.1 and Example 2.3.2, the largest eigenvalue of $\tilde{p}$ is equal to $\frac{1}{k}\sigma_k(x)$. Now Theorem 2.2.9 yields the sublinearity of σ_k. Finally, recall that every convex function is locally Lipschitz ([78, Theorem 10.4]), hence so is each σ_i. So λ_1 is

locally Lipschitz. If $k \geq 2$, then $\lambda_k = \sigma_k - \sigma_{k-1}$ is — as the difference of two locally Lipschitz functions — locally Lipschitz, too. □

The Hermitian matrices (continued). Here it is well known that the sum of the k largest eigenvalues is a convex function and that the k^{th} largest eigenvalue map is locally Lipschitz; see, for instance, [32].

Remark 2.3.4. *Consider the polynomial $\tilde{p}$ in the proof of Corollary 2.3.3 in the context of the Hermitian matrices. Then*

$$(-1)^{\binom{m}{k}}\tilde{p}(x - \tfrac{t}{k}I) = \det(tI - \Delta_k(x)),$$

where $\Delta_k(x)$ denotes the k^{th} additive compound of x. (See [61, Section 19.F] for more on compound matrices.)

Corollary 2.3.5. *The function $w^T\lambda(\cdot)$ is sublinear, for every $w \in \mathbb{R}^m_\downarrow$.*

Proof. Write $w^T\lambda = \sum_{i=1}^m w_i\lambda_i = w_m\sigma_m + \sum_{i=1}^{m-1}(w_i - w_{i+1})\sigma_i$ and then apply Corollary 2.3.3. □

Note that we can rewrite Corollary 2.3.5 quite artificially as $w^T(\lambda(x+y) - \lambda(x)) \leq \bar{w}^T\lambda(y)$, for all $x, y \in X$ and $w \in \mathbb{R}^m_\downarrow$. It would be interesting to find out about the following generalization:

Open Problem 2.3.6 (Lidskii's Theorem). *Decide whether or not*

$$w^T(\lambda(x+y) - \lambda(x)) \leq \bar{w}^T\lambda(y), \quad \textit{for all } x, y \in X \textit{ and } w \in \mathbb{R}^m \ .$$

If this condition is satisfied, then we say that *Lidskii's theorem holds* for the triple (X, p, d). Lidskii's theorem, in the case when x and y are symmetric matrices, and λ is the map of their eigenvalues, is a central result in matrix perturbation theory, see [9, Section III.4].

The condition means that the vector $\lambda(y)$ "majorizes" the vector $\lambda(x+y)-\lambda(x)$, for all $x, y \in X$; see [61, Proposition 4.B.8]. (The interested reader is referred to [61] for further information on majorization.)

The Hermitian matrices (continued). Lidskii's theorem does hold for the Hermitians. A recent and very complete reference is Bhatia's [9]; see also [51] for a new proof rooted in nonsmooth analysis.

In Section 2.6, we point out that Lidskii's theorem holds for all our examples. It will be convenient to have the following simple result ready:

Proposition 2.3.7. *Suppose (X, p, d) is isomorphic to (Y, q, e). Then Lidskii's theorem holds for (X, p, d) if and only if it does for (Y, q, e).*

2.3.2 Convexity of composition

Proposition 2.3.8. *Suppose $f : \mathbb{R}^m \to [-\infty, +\infty]$ is convex and symmetric. Suppose further $u, v \in \mathbb{R}^m_\downarrow$ and $u - v \in (\mathbb{R}^m_\downarrow)^+$. Then $f(u) \geq f(v)$. Moreover: if f is strictly convex on* $\operatorname{conv}\{(u_{\pi(1)}, ..., u_{\pi(m)}) : \pi$ *is a permutation of* $\{1, \ldots, m\}\}$ *and $u \neq v$, then $f(u) > f(v)$.*

Proof. Imitate the proof of [50, Theorem 3.3] and consider [50, Example 7.1]. See also [61, 3.C.2.c on page 68]. □

Theorem 2.3.9 (Convexity). *Suppose* $x, y \in X$, $\alpha \in (0,1)$, *and* $f : \mathbb{R}^m \to [-\infty, +\infty]$ *is convex and symmetric. Then*

$$f(\lambda(\alpha x + (1-\alpha)y)) \leq f(\alpha\lambda(x) + (1-\alpha)\lambda(y))$$

and hence the composition $f \circ \lambda$ *is convex. If* f *is strictly convex and* $\alpha\lambda(x) + (1-\alpha)\lambda(y) \neq \lambda(\alpha x + (1-\alpha)y)$, *then* $f(\lambda(\alpha x + (1-\alpha)y)) < f(\alpha\lambda(x) + (1-\alpha)\lambda(y))$.

Proof. (See also [50, Proof of Theorem 4.3].) Fix an arbitrary $w \in \mathbb{R}^m_\downarrow$. Set $u := \alpha\lambda(x) + (1-\alpha)\lambda(y)$ and $v := \lambda(\alpha x + (1-\alpha)y)$. Then both u and v belong to $\mathbb{R}^m_\downarrow$. By Corollary 2.3.5, $w^T\lambda$ is convex on X. Therefore, $w^T\lambda(\alpha x + (1-\alpha)y) \leq \alpha w^T\lambda(x) + (1-\alpha)w^T\lambda(y)$; equivalently, $w^T(u-v) \geq 0$. It follows that $u - v \in (\mathbb{R}^m_\downarrow)^+$. By Proposition 2.3.8, $f(u) \geq f(v)$, which is the second displayed statement. The convexity of $f \circ \lambda$ follows. Finally, the "If" part is implied by the above and the "Moreover" part of Proposition 2.3.8. □

The Hermitian matrices (continued). In this case, the convexity of the composition is attributed to *Davis* [18]; see also [48, Corollary 2.7].

Another consequence is Gårding's inequality; see [25, Lemma 3.1].

Corollary 2.3.10 (Gårding's Inequality). *Suppose* $p(d) > 0$. *Then function* $x \mapsto -(p(x))^{1/m}$ *is sublinear on the hyperbolicity cone* $C(d)$, *and it vanishes on its boundary.*

Proof. By Corollary 2.2.22, the function $-E_m^{1/m}$ is sublinear and symmetric on $\mathbb{R}^m_+$. Hence, by Theorem 2.3.9, the function $x \mapsto -(E_m(\lambda(x))^{1/m}$ is sublinear on

$\{x \in X : \lambda(x) \geq 0\} = \mathrm{cl}C(d)$. The result follows, since $p(x) = p(d)E_m(\lambda(x))$, for every $x \in X$. □

The Hermitian matrices (continued). Corollary 2.3.10 implies the *Minkowski Determinant Theorem*: $\sqrt[m]{\det(x+y)} \geq \sqrt[m]{\det x} + \sqrt[m]{\det y}$, whenever $x, y \in X$ are positive semi-definite.

Corollary 2.3.11. *Suppose $x, y \in X$. Then:*

1. $\|\lambda(x+y)\| \leq \|\lambda(x) + \lambda(y)\|$.

2. $\|\lambda(x+y)\|^2 - \|\lambda(x)\|^2 - \|\lambda(y)\|^2 \leq 2\langle\lambda(x), \lambda(y)\rangle$.

Moreover, equality holds in 1 or 2 if and only if $\lambda(x+y) = \lambda(x) + \lambda(y)$.

Proof. (1):Let $w := \lambda(x+y) \in \mathbb{R}^m_\downarrow$. Then, using Corollary 2.3.5 and the Cauchy-Schwarz inequality in $\mathbb{R}^m$, we estimate

$$\begin{aligned}\|\lambda(x+y)\|^2 = w^T\lambda(x+y) &\leq w^T(\lambda(x) + \lambda(y)) \\ &\leq \|w\|\|\lambda(x) + \lambda(y)\| = \|\lambda(x+y)\|\|\lambda(x) + \lambda(y)\|.\end{aligned}$$

The inequality follows. The condition for equality follows from the condition for equality in the Cauchy-Schwarz inequality.

(2): The condition is equivalent to *(1)*. □

2.4 Making X Euclidean

So far X has been an arbitrary vector space. We are free to define a norm on it as we wish. To be absolutely precise then, the hyperbolic polynomials, $p(x)$, on X

have to be viewed as polynomials in n linear functionals ($x_i = x_i(x),\ \ i = 1, 2, ..., n$) on X.

Definition 2.4.1. *Define* $\|\cdot\| : X \to [0, +\infty) : x \mapsto \|\lambda(x)\|$ *and*

$$\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R} : (x, y) \mapsto \tfrac{1}{4}\|x + y\|^2 - \tfrac{1}{4}\|x - y\|^2.$$

Theorem 2.4.2. *Suppose p is complete. Then X equipped with $\langle \cdot, \cdot \rangle$ is a Euclidean space with induced norm $\|\cdot\|$.*

Proof. We have

$$\|x\|^2 = \|\lambda(x)\|^2 = \sum_{i=1}^{m} \lambda_i(x)^2 = (E_1(\lambda(x)))^2 - 2E_2(\lambda(x)).$$

Propositions 2.2.8 and 2.2.19 imply that $\|\cdot\|^2$ is a homogeneous polynomial of degree 2 on X. Since $\|\cdot\| \geq 0$ and p is complete, Corollary 2.3.11 says that the equality above indeed defines a norm. Because $\|\cdot\|^2$ is a homogeneous polynomial of degree 2 on X this norm originates trivially from an inner product. The formula for the inner product follows from the Polarization Identity in linear algebra: $\langle x, y \rangle = \frac{1}{4}\|x + y\|^2 - \frac{1}{4}\|x - y\|^2$. □

Remark 2.4.3. *The Euclidean norm $\|\cdot\|$ defined in Definition 2.4.1 is precisely the* Hessian norm *used in interior point methods and thus well-motivated. To see this, assume that p is complete and recall that the* hyperbolic barrier function *is defined by $F(x) := -\ln(p(x))$. The Hessian norm at x is then given by*

$$\|x\|_d^2 := \nabla^2 F(d)[x, x] := \frac{\partial^2}{\partial t^2} F(tx + d)\bigg|_{t=0}.$$

For t positive and sufficiently small, we have $p(tx+d) = p(d)\prod_{i=1}^{m}(1+t\lambda_i(x))$ and hence (after taking logarithms)

$$F(d+tx) = F(d) - \sum_{i=1}^{m} \ln(1+t\lambda_i(x)).$$

Expand the left (resp. right) side of this equation into a Taylor (resp. log) series. Then compare coefficients of t^2 to conclude $\nabla^2 F(d)[x,x]/2! = \|\lambda(x)\|^2/2$. Thus $\|\cdot\|_d = \|\cdot\|$. (It looks as if the right hand side is independent of the direction d, but this is not the case since the eigenvalues λ implicitly depend on it.) Further information can be found in [25]; see, in particular, [25, equation 16].

The norm constructed above has the pleasant property that any isomorphism to another triple is actually an isometry:

Proposition 2.4.4. *Suppose p is complete and the triple (X,p,d) is isomorphic to the triple (Y,q,e) by Φ. Then Φ is an isometry from X onto Y.*

Proposition 2.4.5 (Sharpened Cauchy-Schwarz). *Suppose p is complete. The following inequality then holds*

$$\langle x,y\rangle \le \langle \lambda(x),\lambda(y)\rangle \le \|x\|\|y\|, \qquad \text{for all } x,y \in X.$$

For necessary and sufficient conditions for equality see [6, Theorem 6.6].

Proof. By the Cauchy-Schwarz inequality in $\mathbb{R}^m$ and Corollary 2.3.11.(ii),

$$2\langle \lambda(x),\lambda(y)\rangle \ge \|\lambda(x+y)\|^2 - \|\lambda(x)\|^2 - \|\lambda(y)\|^2$$

$$= \|x+y\|^2 - \|x\|^2 - \|y\|^2$$
$$= 2\langle x, y\rangle.$$

□

The Hermitian matrices (continued). The inner product on the Hermitian matrices is precisely what one would expect: $\langle x, y\rangle = \operatorname{trace}(xy)$. The sharpening of the Cauchy-Schwarz inequality is due to *von Neumann*; see [48, Theorem 2.2] and the discussion therein.

We can now refine Theorem 2.3.9.

Theorem 2.4.6 (Strict Convexity). *Suppose p is complete and the function $f : \mathbb{R}^m \to [-\infty, +\infty]$ is strictly convex and symmetric. Then the composition $f \circ \lambda$ is strictly convex on X.*

Theorem 2.4.6 can be used to recover transparently a recent result by Krylov (see [45, Theorem 6.4.(ii)]).

Corollary 2.4.7. *Suppose $p(d) > 0$. Then each of the following functions is convex on the hyperbolicity cone $C(d)$:*

$$-\ln p, \quad \ln \frac{E_{m-1} \circ \lambda}{E_m \circ \lambda}, \quad \frac{E_{m-1} \circ \lambda}{E_m \circ \lambda}.$$

If p is complete, then each of these functions is strictly convex.

Krylov's result is closely related to parts of Güler's recent work on hyperbolic barrier functions. It suggests a simple approach to Güler's result [25, Theorem 6.1] stated below. The functions F and g below play a crucial role in interior-point

methods as they allow the construction of long-step interior-point methods using the hyperbolic barrier function F.

Corollary 2.4.8. *Suppose $p(d) > 0$ and c belongs to the hyperbolicity cone $C := C(d)$. Define*

$$F : C \to \mathbb{R} : x \mapsto -\ln(p(x)) \quad \textit{and} \quad g : C \to \mathbb{R} : x \mapsto -(\nabla F(x))(c).$$

Then F and g are convex on C. If p is complete, then both F and g are strictly convex.

The Hermitian matrices (continued). The statement on F corresponds to strict convexity of the function $x \mapsto -\ln\det(x)$ on the cone of positive semi-definite Hermitian matrices; this result is due to *Fan* [21].

Remark 2.4.9. *It is worthwhile to point out that Krylov [45] and Güler derived their results from hyperbolic function theory whereas we here "piggyback" on inequalities in elementary symmetric functions. The latter approach is far more elementary.*

2.5 Convex calculus

In this section we present the convex calculus results for hyperbolic polynomials from [6]. We include them for completeness of the exposition, but for brevity we omit the proofs and the details. For definitions of Fenchel conjugate and convex subgradients see the last chapter.

Definition 2.5.1 (Isometric Hyperbolic Polynomial). *We say p is* isometric (with respect to d)*, if for every $y, z \in X$, there exists $x \in X$ such that*

$$\lambda(x) = \lambda(z) \quad \textit{and} \quad \lambda(x+y) = \lambda(x) + \lambda(y).$$

Isometricity depends only on equivalence classes of triples:

Proposition 2.5.2. *Suppose (X, p, d) is isomorphic to (Y, q, e). Then p is isometric if and only if q is.*

It is clear that if p is isometric, then $\operatorname{ran}\lambda$ is a closed convex cone contained in $\mathbb{R}^m_\downarrow$. Examples shows that the range of λ may be nonconvex in general [6].

The Hermitian matrices (continued). Here $\operatorname{ran}\lambda = \mathbb{R}^m_\downarrow$ and and it is easy to see that $p = \det$ is isometric.

Theorem 2.5.3 (Fenchel Conjugacy). *Suppose that $f : \mathbb{R}^m \to (-\infty, +\infty]$ is symmetric. Then $(f \circ \lambda)^* \leq f^* \circ \lambda$. If p is isometric and $f(P_{\operatorname{ran}\lambda} u) \leq f(u)$, for every $u \in (\operatorname{dom} f)_\downarrow$, then $(f \circ \lambda)^* = f^* \circ \lambda$.*

The assumption that $f(P_{\operatorname{ran}\lambda} u) \leq f(u)$, for every $u \in (\operatorname{dom} f)_\downarrow$ is important: in Section 2.6, we present an isometric hyperbolic polynomial and a convex symmetric function f with $(f \circ \lambda)^* \neq f^* \circ \lambda$.

Corollary 2.5.4. *Suppose p is isometric and $f : \mathbb{R}^m \to (-\infty, +\infty]$ is symmetric. Suppose one of the following conditions holds:*

1. $(\operatorname{dom} f) \cap \mathbb{R}^m_\downarrow \subseteq \operatorname{ran}\lambda$.

2. $\operatorname{ran}\lambda = \mathbb{R}^m_\downarrow$.

3. *f is convex and $P_{\operatorname{ran}\lambda} u \in \operatorname{conv}\{(u_{\pi(1)}, ..., u_{\pi(m)}) : \pi$ permutes $\{1, \dots, m\}\}$, for every $u \in (\operatorname{dom} f) \cap \mathbb{R}^m_\downarrow$.*

Then $(f \circ \lambda)^* = f^* \circ \lambda$.

Theorem 2.5.5 (Subgradients). *Suppose p is isometric, $\operatorname{ran}\lambda = \mathbb{R}^m_\downarrow$, and $f : \mathbb{R}^m \to (-\infty, +\infty]$ is convex and symmetric. Let $x, y \in X$. Then*

$$y \in \partial(f \circ \lambda)(x) \text{ if and only if } \lambda(y) \in \partial f(\lambda(x)) \text{ and } \langle x, y\rangle = \langle \lambda(x), \lambda(y)\rangle.$$

Consequently, $\lambda\big[\partial(f \circ \lambda)(x)\big] = \partial f(\lambda(x))$.

The Hermitian matrices (continued). Theorem 2.5.5 corresponds to [48, Theorem 3.2].

Corollary 2.5.6 (Differentiability). *Suppose p is isometric, $\operatorname{ran}\lambda = \mathbb{R}^m_\downarrow$, and $f : \mathbb{R}^m \to (-\infty, +\infty]$ is convex and symmetric. Let $x, y \in X$. Then $f \circ \lambda$ is differentiable at x and $y = \nabla(f \circ \lambda)(x)$ if and only if f is differentiable at $\lambda(x)$ and $\{y' \in X : \lambda(y') = \nabla f(\lambda(x)), \langle x, y'\rangle = \langle \lambda(x), \lambda(y')\rangle\} = \{y\}$.*

Corollary 2.5.7 (Variational Description of σ_k). *Let p be isometric, and suppose $\operatorname{ran}\lambda = \mathbb{R}^m_\downarrow$. Let $1 \le k \le m$. Then for every $x \in X$,*

$$\sigma_k(x) = \max_{y:\lambda(y)\ge 0, \sigma_m(y)=k, \lambda_1(y)\le 1} \langle x, y\rangle$$

and $\partial\sigma_k(x) = \{y \in X : \langle x, y\rangle = \sigma_k(x), \lambda(y) \ge 0, \sigma_m(y) = k, \lambda_1(y) \le 1\}$.

The Hermitian matrices (continued). Corollary 2.5.7 is a direct generalization of the variational formulations due to *Rayleigh* and *Ky Fan*; see [32, Section 2] for more details.

2.6 Examples of hyperbolic polynomials

2.6.1 $\mathbb{R}^n$

Consider the vector space

$$X = \mathbb{R}^n,$$

the polynomial

$$p(x) = \prod_{i=1}^{n} x_i,$$

and the direction

$$d = (1, 1, ..., 1).$$

Then p is a hyperbolic and complete with eigenvalue map

$$\lambda(x) = x_{\downarrow}.$$

The induced norm and inner product in X are just the standard Euclidean ones in $\mathbb{R}^n$. We have $\operatorname{ran}\lambda = \mathbb{R}^n_{\downarrow}$ and p is isometric. In this case the sharpened Cauchy-Schwarz inequality (Proposition 2.4.5) reduces to the well-known Hardy-Littlewood-

Pólya inequality (see [28, Chapter X]).

$$x^T y \leq x_\downarrow^T y_\downarrow$$

and [6, Theorem 6.6] gives necessary and sufficient conditions for equality, which in this case holds if and only if vectors x and y can be simultaneously ordered with the same permutation. Since $\operatorname{ran}\lambda = \mathbb{R}^n_\downarrow$, Corollary 2.5.4 shows that for every symmetric function $f : \mathbb{R}^n \to (-\infty, +\infty]$ we have

$$(f \circ \lambda)^* = f^* \circ \lambda.$$

Also Lidskii's Theorem holds, because $\lambda(x)$ is the ordered set of eigenvalues of the symmetric matrix $\operatorname{Diag}(x)$ (see [9, page 69]).

2.6.2 Hermitian matrices

In this section we summarize the example we have followed throughout the chapter so far. Consider the vector space H^n (of $n \times n$ Hermitian matrices), and denote the ordered eigenvalues of a matrix $x \in H^n$ by $\tilde{\lambda}_1(x) \geq \tilde{\lambda}_2(x) \geq ... \geq \tilde{\lambda}_n(x)$. In the case of Hermitian matrices, the Frobenius [36, page 291] norm can be defined by $\|x\|_F = \|\tilde{\lambda}(x)\|$, where the last norm is the standard Euclidean norm in $\mathbb{R}^n$. Let

$$X = H^n,$$

the polynomial be

$$p(x) = \det x,$$

and the direction be

$$d \;=\; I.$$

Then p is a hyperbolic and complete with eigenvalue map

$$\lambda(x) \;=\; \tilde{\lambda}(x).$$

The induced norm and inner product in X are given by

$$\begin{aligned} \|x\|^2 \;&=\; \|x\|_F^2, \\ \langle x, y \rangle \;&=\; \operatorname{tr} xy. \end{aligned}$$

Clearly we have $\operatorname{ran} \lambda \;=\; \mathbb{R}^n_\downarrow$ and p is isometric. In this case the sharpened Cauchy-Schwarz inequality (Proposition 2.4.5) reduces to Fan's inequality:

$$\operatorname{tr} x^T y \;\leq\; \tilde{\lambda}(x)^T \tilde{\lambda}(y)$$

and equality holds if and only if the matrices x and y can be simultaneously unitarily diagonalized (with eigenvalues in decreasing order), which is due to Theobald. (For the conditions for equality see for example [6, Theorem 6.6] or [52].) Since $\operatorname{ran} \lambda = \mathbb{R}^n_\downarrow$, Corollary 2.5.4 implies that for every symmetric function $f : \mathbb{R}^n \to (-\infty, +\infty]$ we have

$$(f \circ \lambda)^* \;=\; f^* \circ \lambda.$$

It is well known that Lidskii's theorem holds in this case, see [9, Section III.4].

Note that there is an entirely analogous example on the space of n by n real symmetric matrices.

2.6.3 Singular values

Consider the vector space $M_{n,m}$ (of n by m real matrices). We assume $n \leq m$ and denote the singular values of a matrix x in $M_{n,m}$ by $\sigma_1(x) \geq \sigma_2(x) \geq ... \geq \sigma_n(x)$. The Frobenius norm [36, page 291 & page 421] is defined by $\|x\|_F = \|\sigma(x)\|$, where the last norm is the standard Euclidean norm in $\mathbb{R}^n$, and $\sigma(x) = (\sigma_1(x), \sigma_2(x), ..., \sigma_n(x))$. Now consider the vector space

$$X = M_{n,m} \times \mathbb{R},$$

In order to study the singular values we consider the polynomial

$$p(x, \alpha) = \det(\alpha^2 I_n - xx^T) \quad (x \in M_{n,m}, \alpha \in \mathbb{R}),$$

and the direction

$$d = (0, 1).$$

Then p is a hyperbolic and complete polynomial, with eigenvalue map

$$\lambda(x, \alpha) = (\alpha + \sigma_1(x), \alpha + \sigma_2(x), ..., \alpha - \sigma_2(x), \alpha - \sigma_1(x)).$$

The induced norm and inner product are given by

$$\|(x,\alpha)\|^2 = 2n\alpha^2 + 2\|x\|_F^2,$$
$$\langle (x,\alpha),(x,\beta)\rangle = 2n\alpha\beta + 2\mathrm{tr}\, x^T y,$$

for (x,α) and (y,β) in X. Clearly we have $\mathrm{ran}\,\lambda \subset \mathbb{R}^{2m}_{\downarrow}$. Also it is easy to see, using the Singular Value Decomposition Theorem [36, Theorem 7.3.5] that p is isometric. Notice that in this case the sharpened Cauchy-Schwarz inequality (Proposition 2.4.5) reduces to

$$\mathrm{tr}\, x^T y \;\leq\; \sigma(x)^T\sigma(y),$$

and Theorem 6.6 in [6] shows equality holds if and only if x and y have a simultaneous 'ordered' singular value decomposition (that is, there are unitary matrices U and V such that $x = U(\mathrm{Diag}\,\sigma(x))V$ and $y = U(\mathrm{Diag}\,\sigma(y))V$). This is the classical result known as 'von Neumann's Lemma' (see for example [37, page 182]). (For a different proof of von Neumann's result see Theorem 6.2.9.)

Note that when $n = 1$ we get the Lorentz Cone example which is discussed below. An analogous example can be obtained by considering the vector space $X = \mathbb{C}_{n,m} \times \mathbb{R}$.

We now show that for some functions in the singular value case we have $(f\circ\lambda)^* \neq f^*\circ\lambda$. Equality in this case seems to depend on much more algebraic structure, see [53], and Corollary 2.5.4. Consider the symmetric function

$$f(u) \;=\; \max_{1\leq i\leq n} u_i.$$

Then

$$f^*(v) = \begin{cases} 0, & \sum_{i=1}^n v_i = 1,\ v_i \geq 0 \\ +\infty, & \text{otherwise.} \end{cases}$$

Now let $n = 2$. Then $\operatorname{ran}\lambda = \{\alpha e + (\beta, \gamma, -\gamma, -\beta) \,|\, \beta \geq \gamma \geq 0\}$. Let $v = \frac{1}{4}(3, 1, 1, -1) \in \operatorname{ran}\lambda$. Let $y \in X$ be such that $\lambda(y) = v$. It is straightforward to check that $\langle \lambda(z), \lambda(y) \rangle = \lambda_1(z)\ \forall\, z \in \mathrm{X}$. It follows from the sharpened Cauchy-Schwarz inequality (Proposition 2.4.5) that $\langle z, y \rangle \leq \lambda_1(z)\ \forall\, z \in X$. Then

$$(f \circ \lambda)^*(y) = \lambda_1^*(y) = \sup_{z \in X} \{\langle z, y \rangle - \lambda_1(z)\} = 0.$$

On the other hand clearly

$$(f^* \circ \lambda)(y) = f^*(v) = +\infty.$$

Finally we show that Lidskii's theorem holds for this example. For each $w \in \mathbb{R}^n$, let us denote the coordinates of the vector $\bar{w}$ by $\bar{w} = (w_{[1]}, ..., w_{[n]})$. We say that vector y *weakly majorizes* vector x, both in $\mathbb{R}^n$, if the following inequalities hold

$$\sum_{i=1}^k x_{[i]} \leq \sum_{i=1}^k y_{[i]}, \quad \text{for all} \quad k = 1, ..., n.$$

We denote the above relationship by $x \prec_w y$. We also say that a matrix P is *partial permutation matrix* if it has at most one nonzero entry in each row and column, and these nonzero entries (if any) are all 1. A well know result is the following theorem.

Theorem 2.6.1. *If $x \prec_w y$, then x is a convex combination of vectors $P_i y$, for some partial permutation matrices P_i.*

Proof. Combine Theorem 3.2.6 and Theorem 3.2.10 from [37]. □

We also need Theorem 3.4.5 from [37]:

Theorem 2.6.2. *For any matrices x and y in $M_{n,m}$ $(n \leq m)$ the vector $\sigma(y)$ weakly majorizes $|\sigma(x+y) - \sigma(x)|$.*

(For more on weak majorization and a proof of the above theorem using tools from nonsmooth analysis see Section 6.9.)

In order to show Lidskii's theorem for the roots of the hyperbolic polynomial in this example, we have to show that for all $(x, \alpha), (y, \beta) \in X$

$$w^T \left(\lambda(x+y, \alpha+\beta) - \lambda(x, \alpha)\right) \leq \bar{w}^T \lambda(y, \beta) \quad \forall w \in \mathbb{R}^{2n}.$$

This is equivalent to

$$w^T \left(\big(\sigma(x+y), (-\sigma(x+y))_\downarrow\big) - \big(\sigma(x), (-\sigma(x))_\downarrow\big)\right) \leq w_\downarrow^T \big(\sigma(y), (-\sigma(y))_\downarrow\big),$$

for all $w \in \mathbb{R}^{2n}$. This in turn is equivalent to

$$\begin{aligned}(w_1 - w_{2n})(\sigma_1(x+y) - \sigma_1(x)) + (w_2 - w_{2n-1})(\sigma_2(x+y) - \sigma_2(x)) + \cdots \\ + (w_n - w_{n+1})(\sigma_n(x+y) - \sigma_n(x)) \\ \leq (w_{[1]} - w_{[2n]})\sigma_1(y) + (w_{[2]} - w_{[2n-1]})\sigma_2(y) + \cdots + (w_{[n]} - w_{[n+1]})\sigma_n(x).\end{aligned}$$

for each $w \in \mathbb{R}^{2n}$. Let

$$\begin{aligned} U &:= (w_{[1]} - w_{[2n]}, w_{[2]} - w_{[2n-1]}, ..., w_{[n]} - w_{[n+1]}) \\ V &:= (w_1 - w_{2n}, w_2 - w_{2n-1}, ..., w_n - w_{n+1}) \\ \gamma &:= (\sigma_1(x+y) - \sigma_1(x), \sigma_2(x+y) - \sigma_2(x), ..., \sigma_n(x+y) - \sigma_n(x)) \\ \delta &:= (\sigma_1(y), \sigma_2(y), ..., \sigma_n(y)). \end{aligned}$$

Clearly $U \in \mathbb{R}^n_{\downarrow\downarrow} := \mathbb{R}^n_{\downarrow} \cap \mathbb{R}^n_{+}$. Then U is a linear combination with positive coefficients of the vectors $t_i = (\underbrace{1, ..., 1}_{i \text{ times}}, 0, ..., 0) \in \mathbb{R}^n$, that is

$$U = \sum_{j=1}^{n} \alpha_j t_j, \quad \alpha_j \geq 0 \ \forall j.$$

Moreover, it can easily be checked that U weakly majorizes V, so by Theorem 2.6.1 we can write

$$V = \sum_i \beta_i (P_i U), \ \ \beta_i \geq 0, \ \ \sum_i \beta_i = 1,$$

with each P_i a partial permutation matrix. From Theorem 2.6.2 we have that $(P_i t_j)^T \gamma \leq (P_i t_j)^T |\gamma| \leq t_j^T \delta$ for all i and j. Then

$$\begin{aligned} V^T \gamma &= \sum_i \sum_j \beta_i \alpha_j (P_i t_j)^T \gamma \leq \sum_i \sum_j \beta_i \alpha_j t_j^T \delta \\ &= \Big(\sum_i \beta_i \Big) \Big(\sum_j \alpha_j t_j^T \Big) \delta = \Big(\sum_i \beta_i \Big) U^T \delta = U^T \delta, \end{aligned}$$

which is what we want.

2.6.4 Absolute reordering

Consider the vector space

$$X = \mathbb{R}^n \times \mathbb{R}.$$

Let the polynomial be

$$p(x, \alpha) = \prod_{i=1}^{n} (\alpha^2 - x_i^2),$$

and the direction be

$$d = (0, 1).$$

Then p is a hyperbolic and complete with eigenvalue map

$$\lambda(x, \alpha) = ((|x|)_\downarrow, (-|x|)_\downarrow) + \alpha e,$$

where $|x| = (|x_1|, |x_2|, \ldots, |x_n|)$, and $e = (1, 1, \ldots, 1) \in \mathbb{R}^{2n}$. If $\|x\|_2$ denotes the standard Euclidean norm in $\mathbb{R}^n$, then the induced norm and inner product in X are given by

$$\begin{aligned} \|(x, \alpha)\|^2 &= 2\|x\|_2^2 + 2n\alpha^2, \\ \langle (x, \alpha), (y, \beta) \rangle &= 2 \sum_{i=1}^{n} x_i y_i + 2n\alpha\beta. \end{aligned}$$

Clearly $\operatorname{ran} \lambda \subset \mathbb{R}^{2n}_\downarrow$ and it is not difficult to see again that p is isometric. In this case the sharpened Cauchy-Schwarz inequality (Proposition 2.4.5) reduces to

the well-known inequality (see [50, section 7])

$$x^T y \leq (|x|_\downarrow)^T |y|_\downarrow$$

and Theorem 6.6 in [6] shows equality holds if and only if $|x|_\downarrow = P_{(-)}x$ and $|y|_\downarrow = P_{(-)}y$ can be simultaneously ordered with the same *signed permutation matrix*: a permutation matrix in which some of the nonzero entries may be multiplied by -1. (For a direct proof of the above inequality see Lemma 6.2.8.)

Note that the similarities with the previous example are not accidental. This example corresponds to the subspace $(\text{Diag}\,\mathbb{R}^n) \times \mathbb{R}$ of $M_{n,m} \times \mathbb{R}$. So we can immediately see that for some functions f we have $(f \circ \lambda)^* \neq f^* \circ \lambda$. Also, because $|x|_\downarrow = \sigma(\text{Diag}(x))$, one sees, from the corresponding part in the previous example, that Lidskii's Theorem holds.

2.6.5 Lorentz cone

Let the vector space be

$$X = \mathbb{R}^n,$$

and the polynomial be

$$p(x) = x^T A x = x_1^2 - x_2^2 - \cdots - x_n^2,$$

where $A = \text{Diag}(1, -1, -1, ..., -1) \in M_n$ ($n \times n$ real matrices). Let the direction be

$$d = (d_1, d_2, ..., d_n) \in X \text{ such that } d_1^2 > d_2^2 + \cdots + d_n^2.$$

Then p is a hyperbolic and complete with eigenvalue map

$$\lambda(x) = \left(\frac{x^T Ad + \sqrt{D(x)}}{p(d)}, \frac{x^T Ad - \sqrt{D(x)}}{p(d)} \right), \tag{2.1}$$

where $D(x) = (x^T Ad)^2 - p(x)p(d)$ is the discriminant of $p(x + td)$ considered as a quadratic polynomial in t. (The fact that $D(x) \geq 0$ for each x, and so that $p(x)$ is hyperbolic, is the well-known Aczel inequality, see [63, p.57].) The induced norm and inner product are given by

$$\|x\|^2 = 2\frac{2(x^T Ad)^2 - p(x)p(d)}{p(d)^2}, \quad \text{and}$$
$$\langle x, y \rangle = \frac{4(x^T Ad)(y^T Ad) - 2(x^T Ay)p(d)}{p(d)^2},$$

for x and y in X.

We now show that the mapping $\lambda : X \to \mathbb{R}^2_\downarrow$ is onto. Indeed, fix $(t_1, t_2) \in \mathbb{R}^2_\downarrow$, and let l be an arbitrary, fixed nonzero vector from $\{d\}^\perp \subset X$. (The reader can easily verify that $l \in \{d\}^\perp$ if and only if $l^T Ad = 0$.) Set

$$\alpha := \frac{1}{2}(t_1 + t_2), \quad \text{and} \quad v := \sqrt{-\frac{p(d)}{p(l)}} \left(\frac{t_1 - t_2}{2} \right) l. \tag{2.2}$$

Then we have $\lambda(\alpha d + v) = (t_1, t_2)$. Above we have to make sure that $p(l) < 0$. Indeed, because the discriminant of $p(x)$ is always nonnegative we get that $p(l) \leq 0$. If $p(l) = 0$, then this together with $l^T Ad = 0$, and $d^t Ad > 0$ gives us the three relations: $l_1^2 = \tilde{l}^T \tilde{l}$; $d_1 l_1 = \tilde{d}^T \tilde{l}$; $d_1^2 > \tilde{d}^T \tilde{d}$, where we have used the notation $\tilde{x} = (x_2, ..., x_n)$, and the dot product in the relations is the usual one in $\mathbb{R}^{n-1}$.

Notice that $\tilde{l} \neq 0$ since otherwise $l = 0$. Then from the Cauchy-Schwarz inequality we get

$$|d_1 l_1|^2 = |\tilde{d}^T \tilde{l}|^2 \leq |\tilde{d}^T \tilde{d}||\tilde{l}^T \tilde{l}| < d_1^2 l_1^2,$$

which is a contradiction.

We now show that p is isometric. Fix two vectors y, z in X. Let $(t_1, t_2) := \lambda(z)$, and $y = ad + l$, where $a \in \mathbb{R}$ and $l \in \{d\}^\perp$. Define α and v as in Equation (2.2) and set $x := \alpha d + v$. Then the above paragraph shows that $\lambda(x) = \lambda(z)$. So we only have to show that $\lambda(x + y) = \lambda(x) + \lambda(y)$. In order to do that it is enough, by Equation (2.1), to show that $\sqrt{D(x+y)} = \sqrt{D(x)} + \sqrt{D(y)}$. We easily compute that $D(x) = \left(\frac{t_1 - t_2}{2}\right)^2 p(d)^2$ and $D(y) = -(l^T A l) p(d)$ and the rest follows quickly.

Notice that in this case the sharpened Cauchy-Schwarz inequality (Proposition 2.4.5) becomes

$$(x^T A d)(y^T A d) - (x^T A y) p(d) \ \leq \ \sqrt{D(x) D(y)},$$

and Theorem 6.6 in [6] gives the necessary and sufficient condition for equality. Let us show one interesting equivalent form of this sharpened Cauchy-Schwarz inequality.

Corollary 2.6.3 (Sharpened Cauchy-Schwarz). *Let $x, y, d \in \mathbb{R}^n$ and $d_1^2 > d_2^2 + \cdots d_n^2$, then*

$$\sqrt{D(x+y)} \leq \sqrt{D(x)} + \sqrt{D(y)},$$

where $D(x)$ is defined on top of the previous page.

Proof. Because both sides are positive, we can raise the inequality to the second power and substitute the definition of $D(x)$ from the previous page. After canceling several terms we end up exactly with what we originally called sharpened Cauchy-Schwarz inequality, see above. □

Note 2.6.4. *Note that the inequality in the last corollary may be viewed as a measure of how the gap in Aczel's inequality behaves under perturbation. See earlier in this subsection for the definition of Aczel's inequality.*

That Lidskii's Theorem holds for the polynomial $p(x)$ in the direction $f = (1, 0, ..., 0) \in \mathbb{R}^n$ is clear from the corresponding discussion in Section 2.6.3. For arbitrary direction d such that, $d_1^2 > d_2^2 + \cdots d_n^2$, any $w \in \mathbb{R}^2$, and x, $y \in \mathbb{R}^n$ we must show

$$w^T(\lambda(x+y) - \lambda(x)) \leq \bar{w}^T \lambda(y).$$

Using Formula (2.1) for the eigenvalues we see that we have to prove equivalently that

$$w^T(\sqrt{D(x+y)} - \sqrt{D(x)}, -\sqrt{D(x+y)} + \sqrt{D(x)})^T \leq \bar{w}^T(\sqrt{D(y)}, -\sqrt{D(y)})^T. \tag{2.3}$$

We consider two cases.

Case 1. If $w_1 \geq w_2$ then $w = \bar{w}$ and inequality (2.3) becomes

$$(w_1 - w_2)(\sqrt{D(x)} + \sqrt{D(y)} - \sqrt{D(x+y)}) \geq 0.$$

This is immediate from Corollary 2.6.3.

Case 2. If $w_1 < w_2$ then $w = (w_1, w_2)$ and $\bar{w} = (w_2, w_1)$ and inequality (2.3) becomes

$$(w_2 - w_1)(\sqrt{D(y)} + \sqrt{D(x+y)} - \sqrt{D(x)}) \geq 0.$$

Notice that $D(y) = D(-y)$ and use Corollary 2.6.3.

This finally proves that Lidskii's Theorem holds for the roots of the Lorentz hyperbolic polynomial.

2.6.6 Standard hyperbolic triples

We note that if Y is a subspace of H^s (for some positive integer s), $d \in Y$ and $d \succ 0$, then $q(y) := \det y$ is a hyperbolic polynomial over Y with respect to the direction d. Indeed $q(y + td) = \det(d) \det(d^{-\frac{1}{2}} y d^{-\frac{1}{2}} + tI)$ and all the eigenvalues of $d^{-\frac{1}{2}} y d^{-\frac{1}{2}}$ are real numbers because it is a hermitian matrix. The triples of this type, (Y, q, d), will be called **standard hyperbolic triples**.

Many of our examples are isomorphic to a standard hyperbolic triple. For the example in Section 2.6.1, consider the map $\phi(x) = \mathrm{Diag}(x)$. Then clearly $p(x) = \det \phi(x)$. For the example in Section 2.6.2 it is clear. For the example in Section 2.6.4 the following map gives the isomorphism:

$$(x, \alpha) \mapsto \begin{pmatrix} \alpha & x_1 & \dots & 0 & 0 \\ x_1 & \alpha & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \alpha & x_n \\ 0 & 0 & \dots & x_n & \alpha \end{pmatrix}.$$

In general though it is not true that every hyperbolic triple is isomorphic to a standard hyperbolic triple. In Section 2.6.3 take $n = 1$, $m = 4$, which also produces a hyperbolic polynomial of the type discussed in Section 2.6.5. That is, consider for example $X = \mathbb{R}^5$,

$$p(x) = x_0^2 - x_1^2 - x_2^2 - x_3^2 - x_4^2, \quad d = (1, 0, 0, 0, 0).$$

Suppose there is a linear isomorphism $\phi : X \to Y \subset H^s$, such that $p(x) = \det \phi(x)$, and $\phi(d) \succ 0$. Because p is homogeneous of degree 2 we have $t^2 p(x) = p(tx) = \det \phi(tx) = \det t\phi(x) = t^s \det \phi(x)$. Hence we see that $s = 2$. By the linearity of ϕ, there are vectors $a, b, c, f \in \mathbb{R}^5$ such that for every $x \in \mathbb{R}^5$ we have

$$p(x) = \det \begin{pmatrix} a^T x & b^T x + i c^T x \\ b^T x - i c^T x & f^T x \end{pmatrix}.$$

There is a nonzero vector $x \in \mathbb{R}^5$ such that $x_0 = 0$, and $x \perp \operatorname{span}\{a, b, c\}$. So $0 \neq -\|x\|^2 = p(x) = \det \phi(x) = 0$, a contradiction.

We need the following fact on two occasions below.

Proposition 2.6.5. *If A and C are symmetric matrices such that $\det(A) \neq 0$ then*

$$\det \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} = \det(A) \det(C - B^T A^{-1} B).$$

The example in Section 2.6.3 is 'almost' isomorphic to a standard hyperbolic

triple. Indeed, consider the mapping $\phi : M_{n,m} \times \mathbb{R} \to H^{n+m}$ defined by:

$$(x, \alpha) \longmapsto \begin{pmatrix} \alpha I_m & x^T \\ x & \alpha I_n \end{pmatrix},$$

then $\alpha^{m-n} p(x, \alpha) = \det \phi(x, \alpha)$. (This equality holds also in the case when $\alpha = 0$. One needs to consider the two cases $n = m$ and $n < m$ separately.)

Finally, consider a slight variation of the example in Section 2.6.3, the hyperbolic polynomial

$$p(x, \alpha) = \det(\alpha^2 I_m - x^T x)$$

with respect to $d = (0, 1)$, where again $x \in X = M_{n,m} \times \mathbb{R}$. Then the mapping $\Phi : M_{n,m} \times \mathbb{R} \to H^{2m}$ defined by:

$$(x, \alpha) \longrightarrow \begin{pmatrix} \alpha I_{m-n} & 0 & 0 \\ 0 & \alpha I_n & x \\ 0 & x^T & \alpha I_m \end{pmatrix}$$

gives an isomorphism between (X, p, d) and a standard hyperbolic triple. The fact that $p(x, \alpha) = \det \Phi(x, \alpha)$ follows from the identity:

$$\det \begin{pmatrix} \alpha I_{m-n} & 0 & 0 \\ 0 & \alpha I_n & x \\ 0 & x^T & \alpha I_m \end{pmatrix} = \det(\alpha^2 I_m - x x^T).$$

Again, when $\alpha = 0$ the conclusion of the above identity still holds, one just needs to consider the two cases $n = m$ and $n < m$ separately.

Of course the counterexample above doesn't disprove the conjecture made in [46], which concerns polynomials in three variables:

Conjecture 2.6.1. *Every $p(x_1, x_2, x_3)$ hyperbolic with respect to $(0, 0, 1)$, can be expressed as $p(x_1, x_2, x_3) = \det(x_1 A + x_2 B + x_3 I)$ for some symmetric matrices A and B.*

It is worth mentioning that the above conjecture holds when the polynomial is in only two variables. Indeed, suppose $p(x_1, x_2)$ is homogeneous of degree n and hyperbolic with respect to $(0, 1)$. So the polynomial in t, $t \mapsto p(x_1, x_2 + t)$ has only real roots, for every (x_1, x_2). If we let $(x_1, x_2) = (1, 0)$ we can see that $p(1, t) = \prod_{i=1}^{n}(t + a_i)$, where $\{a_i\}$ are real numbers independent of x_1 and x_2. Using that $p(x_1, x_2 + t) = x_1^n p(1, (x_2 + t)/x_1)$ and letting $t = 0$ we see that $p(x_1, x_2) = \prod_{i=1}^{n}(x_2 + a_i x_1)$. The statement is now clear.

2.6.7 The degree 2 case

In this section we show that every complete hyperbolic polynomial of degree two is isometric. Let the vector space be

$$X = \mathbb{R}^n.$$

We will assume that $p(x)$ is a hyperbolic polynomial of degree two with respect to a vector d. Without loss of generality, we write

$$p(x) \;=\; x^T A x,$$

where $A \in H^n$. Proposition 2.2.5 implies that if $S : X \to X$ is a nonsingular linear transformation, then $q(y) := p(Sy)$ is hyperbolic with respect to $l = S^{-1}d$. The next lemma follows also from the fact that p is complete if and only if its hyperbolicity cone is pointed, see Proposition 2.2.15.

Lemma 2.6.6. *If $p(x) = x^T A x$ is hyperbolic, then p is complete if and only if A is nonsingular.*

Proof. Because of Proposition 2.2.15, the linearity space of $p(x)$ in our case is

$$\begin{aligned}
&\{x \in X : (tx+y)^T A(tx+y) = y^T A y,\ \forall y \in X, \forall t \in \mathbb{R}\} \\
&\quad = \{x \in X : x^T A x\, t^2 + 2x^T A y\, t = 0\ \forall y \in X, \forall t \in \mathbb{R}\} \\
&\quad = \{x \in X : x^T A x = 0 \text{ and } x^T A y = 0\ \forall y \in X\} \\
&\quad = \{x \in X : Ax = 0\} = \{0\},
\end{aligned}$$

if and only if A is nonsingular. □

Proposition 2.2.14 now says that if $p(x)$ is a complete hyperbolic polynomial with respect to d, and $S : X \to X$ is a nonsingular linear transformation, then $q(y) := p(Sy)$ is also a complete hyperbolic polynomial with respect to $l = S^{-1}d$.

Lemma 2.6.7. *Let $p(x) = x^T Ax$ be a complete, hyperbolic polynomial, with respect to d of degree two. Then the symmetric matrix A is nonsingular and has exactly $(n-1)$ eigenvalues of one sign, and 1 eigenvalue with the opposite sign.*

Proof. The nonsingularity of A follows from the previous lemma. Now, because $p(x)$ is hyperbolic with respect to d, we have that the discriminant of the quadratic function

$$t \mapsto (x + td)^T A(x + td),$$

$(d^T Ax)^2 - (d^T Ad)(x^T Ax)$ is nonnegative $\forall x \in X$. This inequality implies two things. First A cannot be positive definite because then the Cauchy-Schwarz inequality for the scalar product defined by A contradicts the nonnegativity of the discriminant. Similarly, A cannot be negative definite. Without loss of generality we can assume that that $d^T Ad > 0$, so for every x in the $(n-1)$-dimensional orthogonal complement (with respect to the usual inner product) of the vector Ad we have $0 \geq x^T Ax$. This implies that A has at least $(n-1)$ nonpositive eigenvalues, but none of them can be zero, so A has $(n-1)$ strictly negative eigenvalues. The last eigenvalue must be strictly positive, because A cannot be negative semidefinite. The case $d^T Ad < 0$ is handled analogously. □

Now, Proposition 2.5.2 says that if $p(x)$ is an isometric, complete hyperbolic polynomial with respect to d, and $S : X \to X$ is a nonsingular linear transformation, then $q(y) := p(Sy)$ is also an isometric, complete, hyperbolic polynomial with respect to $l = S^{-1}d$.

Let $p(x) = x^T Ax$ be isometric with respect to d. Without loss of generality we can assume that $p(d) > 0$. By Sylvester's theorem (see for example [36], Theorem 4.5.8), there exists a nonsingular transformation $x = Sy$ of the variable x such that $q(y) := p(Sy)$ has the form: $q(y) = y_1^2 - y_2^2 - \cdots y_n^2$. Moreover, from the above, $q(y)$ is hyperbolic with respect to $S^{-1}d$. Because the subsection about the Lorentz cone showed that $q(y) = y_1^2 - y_2^2 - \cdots - y_n^2$ is isometric with respect to any d in a hyperbolicity cone of q, and $C(q,l) = S^{-1}(C(p,d))$ we have answered the question about isometricity for the whole class of hyperbolic polynomials of degree two.

2.6.8 Unitarily invariant norms

In this section we derive a well known theorem of von Neumann about unitarily invariant norms as a consequence of the convexity results in this chapter.

In 1937, von Neumann [90] gave a famous characterization of unitarily invariant matrix norms (that is, norms f on $\mathbb{C}^{m\times n}$ satisfying $f(uxv) = f(x)$ for all unitary matrices u and v and matrices x in $\mathbb{C}^{m\times n}$). His result states that such norms are precisely the functions of the form $g \circ \sigma$, where the components of the map

$$x \in \mathbb{C}^{m\times n} \mapsto \sigma(x) \in \mathbb{R}^m$$

are the singular values $\sigma_1(x) \geq \sigma_2(x) \geq \ldots \geq \sigma_m(x)$ of x (assuming $m \leq n$) and g is a norm on $\mathbb{R}^m$, that is invariant under sign changes and permutations of components. Proof of this can be found also in [36, Theorem 7.4.24].

Lemma 2.6.8. *For $x, y, \omega \in \mathbb{R}^m$, such that $\omega_1 \geq \omega_2 \geq \ldots \geq \omega_m \geq 0$, and $\lambda \in [0,1]$,*

we have

$$\langle \omega, |\lambda x + (1-\lambda)y|_\downarrow \rangle \leq \langle \omega, \lambda|x|_\downarrow + (1-\lambda)|y|_\downarrow \rangle.$$

Proof. Apply Corollary 2.3.5, with $w = (\omega_1, \omega_2, ..., \omega_m, 0, ..., 0) \in \mathbb{R}^{2m}_\downarrow$, to the roots of the hyperbolic polynomial given in Section 2.6.4 □

Now define $H : \mathbb{R}^{2n} \to \mathbb{R}^n$ by

$$H(u) = \tfrac{1}{2}(v_1 + v_2, v_3 + v_4, \ldots, v_{2n-1} + v_{2n}),$$

where $v = |u|_\downarrow$.

Lemma 2.6.9. *For $u, v \in \mathbb{R}^{2n}$, $z \in \mathbb{R}^n$ such that $z_1 \geq z_2 \geq ... \geq z_n \geq 0$, and $\lambda \in [0,1]$ we have*

$$\langle z, H(\lambda u + (1-\lambda)v) \rangle \leq \langle z, \lambda H(u) + (1-\lambda)H(v) \rangle.$$

Proof. Apply Lemma 2.6.8 with $m = 2n$ and $\omega_{2i-1} = \omega_{2i} = z_i$. □

Now suppose $g : \mathbb{R}^n \mapsto (-\infty, +\infty]$ is convex and absolutely symmetric (that is, $g(x) = g(|x|_\downarrow), \forall x$).

Lemma 2.6.10. $g(H(\lambda u + (1-\lambda)v)) \leq \lambda g(H(u)) + (1-\lambda)g(H(v))$.

Proof. Apply Theorem 3.3 from [50] to Lemma 2.6.9. □

Now define $f : \mathbb{R}^{2n} \mapsto (-\infty, +\infty]$ by $f(u) = g(H(u))$.

Lemma 2.6.11. *The function f is absolutely symmetric and convex.*

Proof. Notice that $H(|u|_\downarrow) = H(u)$. Consequently,

$$f(|u|_\downarrow) = g(H(|u|_\downarrow)) = g(H(u)) = f(u), \forall\, u.$$

So f is absolutely symmetric. The convexity follows from Lemma 2.6.10. □

Theorem 2.6.12 (Von Neumann). *The function $g \circ \sigma$ is convex.*

Proof. Using Section 2.6.3 where $X = M_{n,m} \times \mathbb{R}$, $p(x, \alpha) = \det(\alpha^2 I - xx^T)$, and $d = (0, 1)$, we have that $\lambda(x, 0) = (\sigma_1(x), ..., \sigma_n(x), -\sigma_n(x), ..., -\sigma_1(x))$. So $H(\lambda(x, 0)) = \sigma(x)$. Then finally $g(\sigma(x)) = f(\lambda(x, 0))$, which, because of Theorem 2.3.9, is convex in x. □

Chapter 3

Self-concordant barriers for hyperbolic means

In this chapter we demonstrate an application of hyperbolic polynomials in convex optimization. (The necessary background on hyperbolic polynomials was given in Chapter 2.) Our main result here will be to show how one can construct a class of self-concordant barriers using hyperbolic polynomials. We begin with necessary background about self-concordant barriers. Section 3.3 contains the main result. Some examples and applications in convex optimization conclude the chapter.

3.1 Self-Concordant barriers

We begin by giving the definition of a self-concordant barrier function. Let E be a finite-dimensional real vector space and Q be an open nonempty convex subset of E. A function $F : Q \to \mathbb{R}$ is called a *self-concordant barrier* if it is three times

continuously differentiable, convex and satisfies the conditions

$$|D^3F(x)[h,h,h]| \leq 2\,(D^2F(x)[h,h])^{3/2}, \tag{3.1}$$

$$F(x^r) \to \infty \quad \text{for any sequence } x^r \to x \in \operatorname{bd} Q, \text{ and} \tag{3.2}$$

$$|DF(x)[h]| \leq \sqrt{\vartheta}\,(D^2F(x)[h,h])^{1/2}, \tag{3.3}$$

for all $h \in E$, $x \in Q$. Here $\vartheta \geq 1$ is a fixed constant depending on the function F only, and $D^kF(x)[h,...,h] = \frac{d^k}{dt^k}F(x+th)\Big|_{t=0}$ is the k-th directional derivative at x along the direction h. The constant ϑ is called the *parameter* of the barrier function: smaller parameters ensure that the interior point method using F runs faster. For short we call F a *ϑ-self-concordant barrier*.

If in addition $\operatorname{cl} Q$ is a cone and instead of conditions (3.1), (3.2), and (3.3) the function F satisfies conditions (3.1), (3.2), and

$$F(tx) = F(x) - \vartheta \log(t), \text{ for all } x \in Q,\ t > 0, \tag{3.4}$$

we say F is a *ϑ-normal barrier*. In fact conditions (3.1), (3.2), and (3.4) imply condition (3.3), see [68, Corollary 2.3.2].

Note 3.1.1. *Observe that if F is ϑ-self-concordant then kF is $k\vartheta$-self-concordant for any constant $k \geq 1$.*

3.2 Hyperbolic polynomials & hyperbolicity cone

1. Hyperbolic Polynomials. In this chapter we investigate further properties

of hyperbolic polynomials. The reader should consult Section 2.2 for the necessary definitions and background results. There is only one difference in notation. If p is hyperbolic with respect to d, that is, the polynomial $t \mapsto p(x + td)$ (where t is a scalar) has only real zeros for every $x \in E$, the negatives of these roots will be denoted by $t_i(x, d) = t_i(x)$, and then we can write

$$p(x + td) = p(d) \prod_{i=1}^{m} (t + t_i(x, d)).$$

For convenience we state briefly our main examples from the previous chapter that we will follow up with the present developments.

(a) $E = \mathbb{R}^n$. The polynomial

$$p(x) = \prod_{i=1}^{n} x_i$$

is hyperbolic with respect to the direction $d = (1, ..., 1)$. (cf. Section 2.6.1.)

(b) $E = \mathbb{R}^n$. The polynomial

$$p(x) = x_1^2 - \sum_{i=2}^{n} x_i^2$$

is hyperbolic with respect to the direction $d = (1, 0, ..., 0)$. (cf. Section 2.6.5.)

(c) $E = S^n$ (the set of $n \times n$ symmetric matrices). The polynomial

$$p(X) = \det X$$

is hyperbolic with respect to the direction $d = I$. (cf. Section 2.6.2.)

(d) $E = M_{p,q} \times \mathbb{R}$ (where $M_{p,q}$ is the space of $p \times q$ real matrices, and we assume

$q \leq p$). The polynomial

$$p(X, r) = \det(X^T X - r^2 I_q) \qquad (X \in M_{p,q},\ r \in \mathbb{R})$$

is hyperbolic with respect to the direction $d = (0, 1)$. (cf. Section 2.6.3.)

2. Hyperbolicity cone. Recall that the *hyperbolicity cone of p with respect to d*, written $C(p, d)$, is the set $\{x \in E : p(x + td) \neq 0,\ \forall t \geq 0\}$. In other words

$$C(p, d) = \{x \in E : t_i(x) > 0,\ 1 \leq i \leq m\}.$$

From now on the hyperbolicity cone will be denoted $C(p)$. We now return to the examples in the previous subsection and identify the hyperbolicity cone in each case.

(a) The hyperbolicity cone is the interior of the positive orthant:

$$\{x \in \mathbb{R}^n : x_i > 0,\ 1 \leq i \leq n\}.$$

(b) The hyperbolicity cone is the Lorenz cone:

$$\left\{x \in \mathbb{R}^n : \sqrt{x_2^2 + \cdots x_n^2} < x_1\right\}$$

(c) The hyperbolicity cone is the cone, S^n_{++}, of $n \times n$ symmetric positive definite matrices.

(d) The hyperbolicity cone is the interior of the operator norm epigraph

$$\{(X, r) \in M_{p,q} \times \mathbb{R} : |\sigma_i(X)| < r,\ 1 \leq i \leq q\},$$

where $\sigma_1(X), ..., \sigma_q(X)$ are the singular values of the matrix X [6, Section 7.3].

3.3 A shifted self-concordant barrier

We begin with a trivial lemma.

Lemma 3.3.1. *For any real numbers $t_1, ..., t_m$, the following inequality holds:*

$$\left| \sum_{i=1}^{m} t_i^3 \right| \leq \left(\sum_{i=1}^{m} t_i^2 \right)^{3/2}.$$

The next theorem is our key result in this section.

Theorem 3.3.2. *Let p be a hyperbolic polynomial (homogeneous of degree m) with hyperbolicity cone $C(p)$. Let $a \geq 0$ be a real number and*

$$C_{>a}(p) = \{x \in C(p) : p(x) > a\}.$$

Then the function

$$f(x) = -m \log(p(x) - a)$$

is an m^2-self-concordant barrier on the set $C_{>a}(p)$.

Proof. The case $a = 0$ was proved in [25]. Notice also that condition (3.2) holds trivially.

Step 0. For $x \in C_{>a}(p)$ and $h \in \mathbb{R}^n$, we can write

$$p(x+th) = t^m p\left(h + \frac{1}{t}x\right) = t^m p(x) \prod_{i=1}^{m} \left(\frac{1}{t} + t_i(h,x)\right) = p(x) \prod_{i=1}^{m} (1 + tt_i).$$

What is important is that the roots $t_i = t_i(h,x)$ do not depend on the variable t. Differentiating both sides of the above representation we get the directional derivative of $p(x)$ in the direction of h, which is used below repeatedly:

$$\frac{d}{dt} p(x+th) = p(x+th) \sum_{i=1}^{m} \frac{t_i}{1+tt_i}.$$

Step 1. Observe that in the case $a \neq 0$ we only need to prove self-concordance for $a = 1$, because we can make the linear substitution $x = a^{1/m} y$ to obtain

$$f(a^{1/m} y) = -m \log(p(y) - 1) - m \log(a).$$

(See for example [68, p.148].) So we assume from now on that $a = 1$.

We now compute the directional derivatives of f along the direction h, using the representation from above

$$f(x+th) = -m \log\left(p(x) \prod_{i=1}^{m} (1+tt_i) - 1\right).$$

For short we introduce the notation

$$\alpha = p(x) - 1, \quad C_1 = \sum_{i=1}^{m} t_i, \quad C_2 = \sum_{i=1}^{m} t_i^2, \quad C_3 = \sum_{i=1}^{m} t_i^3, \tag{3.5}$$

and observe that in our situation, for $x \in C_{>1}(p)$, we have $\alpha > 0$. Elementary calculation shows

$$\begin{aligned} Df(x)[h] &= -\frac{m(\alpha+1)}{\alpha}C_1, \\ D^2f(x)[h,h] &= \frac{m(\alpha+1)}{\alpha^2}C_1^2 + \frac{m(\alpha+1)}{\alpha}C_2, \text{ and} \\ D^3f(x)[h,h,h] &= -\frac{m(\alpha+1)(\alpha+2)}{\alpha^3}C_1^3 - \frac{3m(\alpha+1)}{\alpha^2}C_1C_2 - \frac{2m(\alpha+1)}{\alpha}C_3. \end{aligned}$$

We want to prove that inequalities (3.1) and (3.3) hold for every $h \in \mathbb{R}^n$ and $x \in C_{>1}(p)$.

Step 2. We start with inequality (3.3), which in the new notation is

$$\left|\frac{m(\alpha+1)}{\alpha}C_1\right| \leq m\left(\frac{m(\alpha+1)}{\alpha^2}C_1^2 + \frac{m(\alpha+1)}{\alpha}C_2\right)^{1/2}.$$

After squaring both sides and dividing by $\frac{m^2(\alpha+1)}{\alpha}$ we get

$$\frac{(\alpha+1)}{\alpha}C_1^2 \leq \frac{m}{\alpha}C_1^2 + mC_2,$$

so we want to show

$$\frac{\alpha+1-m}{\alpha}C_1^2 \leq mC_2.$$

The Cauchy-Schwarz inequality gives us $C_1^2 \leq mC_2$ so since $m \geq 1$ we obtain

$$\frac{\alpha+1-m}{\alpha}C_1^2 \leq m\frac{\alpha+1-m}{\alpha}C_2 \leq mC_2,$$

as required.

Step 3. Now we turn our attention to inequality (3.1). With the new notation, this is

$$m\left|\frac{(\alpha+1)(\alpha+2)C_1^3}{\alpha^3}+\frac{3(\alpha+1)C_1C_2}{\alpha^2}+\frac{2(\alpha+1)C_3}{\alpha}\right| \leq 2\left(\frac{m(\alpha+1)}{\alpha^2}C_1^2+\frac{m(\alpha+1)}{\alpha}C_2\right)^{3/2}.$$

We multiply both sides by $\frac{\alpha^3}{m(\alpha+1)}$ to get

$$\left|(\alpha+2)C_1^3+3\alpha C_1C_2+2\alpha^2C_3\right| \leq 2\sqrt{m(\alpha+1)}\left(C_1^2+\alpha C_2\right)^{3/2}.$$

Since this inequality is homogeneous of degree 3 in the vector $(t_1, t_2, ..., t_m)$, we may assume without loss of generality that $C_1 = \pm 1$. We distinguish two cases.

Step 3.*a*. Suppose we have $C_1 = +1$. The inequality becomes

$$\left|2+\alpha+3\alpha C_2+2\alpha^2C_3\right| \leq 2\sqrt{m\alpha+m}\,(1+\alpha C_2)^{3/2}.$$

We now square both sides and expand:

$$4 + \alpha^2 + 9\alpha^2C_2^2 + 4\alpha^4C_3^2 + 4\alpha + 12\alpha C_2 + 8\alpha^2C_3 + 6\alpha^2C_2 + 4\alpha^3C_3 + 12\alpha^3C_2C_3 \leq 4m\alpha + 12m\alpha^2C_2 + 12m\alpha^3C_2^2 + 4m\alpha^4C_2^3 + 4m + 12m\alpha C_2 + 12m\alpha^2C_2^2 + 4m\alpha^3C_2^3.$$

Regrouping gives

$$\begin{aligned} 0 \;&\leq\; (4mC_2^3 - 4C_3^2)\alpha^4 + (4mC_2^3 + 12mC_2^2 - 4C_3 - 12C_2C_3)\alpha^3 \\ &+\; (12mC_2^2 + 12mC_2 - 6C_2 - 8C_3 - 9C_2^2 - 1)\alpha^2 \\ &+\; (12mC_2 + 4m - 12C_2 - 4)\alpha + (4m - 4). \end{aligned} \tag{3.6}$$

We show now that all the coefficients are positive. Using Lemma 3.3.1 and the fact $m \geq 1$, $C_2 \geq \frac{1}{m}$ this becomes clear for the coefficients of α^4, α and the constant term. Further, for the coefficient of α^3 using Lemma 3.3.1 we have

$$\begin{aligned} 4mC_2^3 + 12mC_2^2 - 4C_3 - 12C_2C_3 &\geq 4mC_2^3 + 12mC_2^2 - 4C_2^{3/2} - 12C_2^{5/2} \\ &= C_2^{3/2}(4mC_2^{3/2} + 12mC_2^{1/2} - 4 - 12C_2). \end{aligned}$$

Consider the polynomial $q(s) := 4ms^3 - 12s^2 + 12ms - 4$. Its derivative $q'(s) = 12(ms^2 - 2s + m)$ is nonnegative, so q is increasing. Using the fact that $\frac{1}{\sqrt{m}} \leq C_2^{1/2}$ we get

$$\begin{aligned} q\big(C_2^{1/2}\big) \geq q\left(\frac{1}{\sqrt{m}}\right) &= \frac{4\sqrt{m}}{m} - \frac{12}{m} + \frac{12m\sqrt{m}}{m} - \frac{4m}{m} \\ &= \frac{4(\sqrt{m}-1) + 8(m\sqrt{m}-1) + 4m(\sqrt{m}-1)}{m} \geq 0, \end{aligned}$$

which shows that the coefficient of α^3 is positive. For the coefficient of α^2, using Lemma 3.3.1, we have

$$12mC_2^2 + 12mC_2 - 6C_2 - 8C_3 - 9C_2^2 - 1$$

$$\geq 12mC_2^2 + 12mC_2 - 6C_2 - 8C_2^{3/2} - 9C_2^2 - 1$$
$$= 9(m-1)C_2^2 + 6(m-1)C_2 + (mC_2 - 1) + C_2(3mC_2 - 8C_2^{1/2} + 5m).$$

The quadratic polynomial $3ms^2 - 8s + 5m$ is strictly positive in the case when $m \geq 2$, and the fact that $C_2 \geq \frac{1}{m}$ then implies that the last coefficient above is positive. In the case when $m = 1$ we have $C_2 = 1$ and one immediately sees that the coefficient of α^2 is actually zero. The fact that all coefficients of the quadratic polynomial on the right hand side of inequality (3.6) are positive implies that the inequality holds for all $\alpha \geq 0$, which is what we wanted to prove.

Step 3.*b*. Suppose on the other hand we have $C_1 = -1$. The inequality becomes

$$\left|(-2) - \alpha - 3\alpha C_2 + 2\alpha^2 C_3\right| \leq 2\sqrt{m\alpha + m}\,(1 + \alpha C_2)^{3/2}.$$

Again we square both sides and expand to obtain

$$4 + \alpha^2 + 9\alpha^2 C_2^2 + 4\alpha^4 C_3^2 + 4\alpha + 12\alpha C_2 - 8\alpha^2 C_3 + 6\alpha^2 C_2 - 4\alpha^3 C_3 - 12\alpha^3 C_2 C_3 \leq 4m\alpha + 12m\alpha^2 C_2 + 12m\alpha^3 C_2^2 + 4m\alpha^4 C_2^3 + 4m + 12m\alpha C_2 + 12m\alpha^2 C_2^2 + 4m\alpha^3 C_2^3.$$

Regrouping gives

$$0 \leq (4mC_2^3 - 4C_3^2)\alpha^4 + (4mC_2^3 + 12mC_2^2 + 4C_3 + 12C_2C_3)\alpha^3$$
$$+ (12mC_2^2 + 12mC_2 - 6C_2 + 8C_3 - 9C_2^2 - 1)\alpha^2$$
$$+ (12mC_2 + 4m - 12C_2 - 4)\alpha + (4m - 4).$$

Now, if $C_3 > 0$ then we can see analogously (even more simply than in Step 3.a) that all coefficients of the quadric polynomial are positive. If $C_3 < 0$ then we use Lemma 3.3.1 to obtain $C_3 \geq -C_2^{3/2}$ and again proceed as in Step 3.a. □

3.4 Examples

Following our examples from Section 3.2, we obtain the following applications of the main result.

(a) For any natural number m the function

$$f(x_1, ..., x_m) = -m \log \left(\prod_{i=1}^{m} x_i - 1 \right)$$

is an m^2-self-concordant barrier on the set

$$\left\{ x \in \mathbb{R}^m : \prod_{i=1}^{m} x_i > 1,\ x_i > 0,\ 1 \leq i \leq m \right\}.$$

In particular when $m = 2$ this result follows from Proposition 5.3.2 in [68].

(b) The function

$$f(x, y) = -2 \log(y^2 - \|x\|^2 - 1)$$

is a 4-self-concordant barrier on the set

$$\left\{ (y, x) \in \mathbb{R} \times \mathbb{R}^{n-1} : y \geq \sqrt{\|x\|^2 + 1} \right\}.$$

This result can also be found in [68]. (See the proof of Proposition 5.4.3 and

make the linear substitution $t \to z - 1$, $y \to z + 1$ in the function Ψ.) In fact, [68] proves that $-\log(y^2 - \|x\|^2 - 1)$ is a 2-self-concordant barrier on the same set.

(c) A more interesting example is the function

$$f(X) = -m \log(\det X - 1),$$

which is an m^2-self-concordant barrier on the set

$$\{X \in S^m_{++} : \det X > 1\}.$$

(d) The function

$$f(X, r) = -2q \log(\det (X^T X - r^2 I_q) - 1)$$

is a $(2q)^2$-self-concordant barrier on the set

$$\{(X, r) \in M_{p,q} \times \mathbb{R} : \det (X^T X - r^2 I_q) > 1 \,\&\, |\sigma_1(X)| < r\}.$$

3.5 Application: hyperbolic means

A *hyperbolic mean* is a function of the form $p(x)^{1/m}$, where p is a hyperbolic polynomial of degree m, and the domain is the hyperbolicity cone $C(p)$. Hyperbolic means are positively homogeneous and concave [25, Lemma 3.1]. Examples include

the geometric mean $(\prod_{i=1}^m x_i)^{1/m}$, and the function

$$X \in S^m_{++} \mapsto (\det X)^{1/m}.$$

A natural approach to applying interior point methods to convex programs involving hyperbolic means is to use a self-concordant barrier for the *hypograph* of the mean, the convex cone

$$H(p) = \{(x,t) \in \mathbb{R}^n \times \mathbb{R} : x \in C(p),\, 0 < t^m < p(x)\}.$$

The following result provides such a barrier.

Theorem 3.5.1. *For a suitable positive real μ (for example $\mu = 400$), if p is a hyperbolic polynomial of degree m then*

$$(x,t) \mapsto -\mu m \left(\log \left(\frac{p(x)}{t^m} - 1 \right) + 2m \log t \right)$$

is a $2\mu m^2$-normal barrier for the hypograph, H(p), of the hyperbolic mean.

Proof. Apply Proposition 5.1.4 in [68] to Theorem 2.2. □

As a simple-minded illustration, suppose we want to solve the problem

$$\begin{array}{ll} \sup & p(x)^{\frac{1}{m}} + \langle c, x \rangle \\ \text{s.t.} & Ax = b \\ & x \in C(p), \end{array}$$

for some linear map A and given b and c. Rewrite this problem in the equivalent form

$$\begin{aligned} \sup \quad & t + \langle c, x \rangle \\ \text{s.t.} \quad & t < p(x)^{\frac{1}{m}} \\ & Ax = b \\ & x \in C(p), \end{aligned}$$

and finally into the form

$$\begin{aligned} \max \quad & \langle \tilde{c}, \tilde{x} \rangle \\ \text{s.t.} \quad & \tilde{A}\tilde{x} = b \\ & \tilde{x} \in H(p), \end{aligned}$$

where $\tilde{c} := (c, 1)$, $\tilde{x} := (x, t)$, $\tilde{A}(x, t) := Ax$. We have an easily computable self-concordant (logarithmically homogeneous) barrier for the cone $H(p)$, so we can design an interior point algorithm to solve this hyperbolic mean maximization problem. Using this result we can as well easily model convex programs with constraints involving hyperbolic means, since $x \in C(p)$ satisfies an inequality of the form

$$\langle c, x \rangle - p(x)^{1/m} < b$$

if and only if there exists positive real t satisfying

$$\langle c, x \rangle - t < b, \quad t^m < p(x).$$

In [68, p.239], Nesterov and Nemirovskii show how to model convex programs involving the geometric mean or $(\det(\cdot))^{1/m}$ by semidefinite programming. It is inter-

esting to compare their approach to this idea. Their approach involves additional variables ($O(m^2)$ variables to model $\det(\cdot)^{1/m}$, for example), whereas this idea is direct and applies to any hyperbolic mean. On the other hand, extremely efficient algorithms are now available for semidefinite programming (see for example [2], [85]).

3.6 Relationship with Güler's result

As we mentioned above, in [25] Güler proved that $-\log(q(x))$ is an n-self-concordant barrier on $C(q)$ for any hyperbolic polynomial q of degree n. (Güler attributes the observation to Renegar.) In this concluding section we want to show that our result cannot be deduced by an affine restriction of this fact. In other words we want to show that we cannot take a self-concordant barrier of the above type, restrict it to an affine subspace and obtain the self-concordance of $-m\log(p(x)-1)$.

Consider the following special case of Theorem 3.3.2:

$$-3\log(x^3-1) \text{ is self-concordant on } (1,+\infty).$$

To deduce this from [25] we would need a hyperbolic polynomial q with respect to d with hyperbolicity cone $C(q)$ and vectors a and b such that

$$(x^3-1)^3 = q(a+xb), \text{ for all } x \in \mathbb{R}, \text{ and}$$
$$1 < x \in \mathbb{R} \Leftrightarrow a+xb \in C(q).$$

When $x = 0$ we immediately get $q(a) = -1$. We can also conclude that $b \in \mathrm{cl}\, C(q)$ which is a closed convex cone. Since $d \in C(q)$, an open convex cone, we have for all small enough real $\epsilon > 0$, that $b + \epsilon d \in C(q)$, so the polynomial q is hyperbolic with respect to $b + \epsilon d$ as well. That is, for all small enough $\epsilon > 0$ the polynomial (in x) $q(a + x(b + \epsilon d))$ has only real, nonzero roots. Clearly if $q(a + xb) = (x^3 - 1)^3$ then $n \geq 9$. We divide both sides of this equality by x^n, and setting $t := 1/x$ obtain

$$q(at + b) = t^{n-9} - 3t^{n-6} + 3t^{n-3} - t^n = t^{n-9}(1 - t^3)^3.$$

Using the fact that $q(a + x(b + \epsilon d))$ has nonzero roots and applying the same substitution as above we get that the polynomial (in t) $t \mapsto q(at + b + \epsilon d)$ has only real roots. Now, for ϵ close to zero, the degree of the polynomial $q(at + b + \epsilon d)$ is constant, and so its roots approach the roots of $q(at + b)$ as ϵ approaches zero. This is a contradiction with the fact that $q(at + b)$ has a complex root.

3.7 An alternative approach

Our approach up to here originated with [57]. A subsequent approach, [66] uses more sophisticated theory to obtain a broader version of Theorem 3.3.2. Here we describe briefly the details. Let Q be an open, pointed, convex cone and let the function $F : Q \to \mathbb{R}$ satisfy conditions (3.1), (3.2), and (3.3). We need the following definition [68, Definition 5.1.2].

Definition 3.7.1. *Let β be nonnegative real. A function $\mathcal{A} : Q \to \mathbb{R}$ is called* **β-compatible with the barrier** *F if*

(i) *$\mathcal{A}$ is C^3 on Q.*

(ii) *$\mathcal{A}$ is concave with respect to clQ.*

(iii) *For all $x \in Q$, $h \in E$, we have*

$$D^3\mathcal{A}(x)[h,h,h] \leq -3\beta D^2\mathcal{A}(x)[h,h]\sqrt{D^2F(x)[h,h]}.$$

We also need the following result, a special case of [68, Proposition 5.1.7].

Theorem 3.7.2. *Assume $\mathcal{A}$ is β-compatible with F, with $\beta \geq 1$. Then the function*

$$\Psi(x) = \beta^2\{-\log(1+\mathcal{A}(x)) + F(x)\}$$

is a $\beta^2(\vartheta+1)$-self-concordant barrier on the domain $\{x \in Q | \mathcal{A}(x) > -1\}$.

A calculation shows that $\mathcal{A}(x) := -e^{F(x)}$ is a $\sqrt{\vartheta+20}$-compatible with F, so setting $\beta = \sqrt{\vartheta+20}$ we have that

$$\Psi(x) = -(\vartheta+20)\log(e^{-F(x)} - 1),$$

is a $(\vartheta+20)(\vartheta+1)$-self-concordant barrier on the domain $\{x \in Q | F(x) < 0\}$. When p is a hyperbolic polynomial of degree m and $F(x) = -\log(p(x))$ we have $\vartheta = m$ and the above result follows from Theorem 3.3.2 using Note 3.1.1 with $k = (\vartheta+20)/\vartheta$. (In fact Theorem 3.3.2 does a bit better.) We conclude with the equivalent of Theorem 3.5.1.

Theorem 3.7.3. *For a suitable positive real μ (for example $\mu = 400$), if $-\log H(x)$ is a ϑ-normal barrier on Q then*

$$(x,t) \mapsto -\mu^2(\vartheta+20)\left(\log\left(\frac{H(x)}{t^\vartheta}-1\right)+2(\vartheta+1)\log(t)\right),$$

is a $2\mu^2(\vartheta+20)(\vartheta+1)$-normal barrier on the domain $\left\{(x,t)|0<t^\vartheta<H(x)\right\}$.

Proof. Notice first that $H(tx) = t^\vartheta H(x)$ for all $x \in Q$, $t > 0$. Then let $F(x) := -\log H(x)$ in the above paragraph and apply Proposition 5.1.4 in [68] to the function $\Psi(x)$. □

We would like to comment that the constant $\mu = 400$, in Theorem 3.5.1 and Theorem 3.7.3 can be improved using the results in [23].

Chapter 4

Twice Differentiable Spectral Functions

In this chapter we show that a symmetric function f is twice differentiable at the point $\lambda(A)$ if and only if the corresponding spectral function $f \circ \lambda$ is twice differentiable at A. Moreover we will show that $f \in \mathcal{C}^2$ around $\lambda(A)$ if and only if $(f \circ \lambda) \in \mathcal{C}^2$ around A.

4.1 Notation and preliminary results

In what follows S^n will denote the Euclidean space of all $n \times n$ symmetric matrices with inner product $\langle A, B \rangle = \operatorname{tr}(AB)$ and for $A \in S^n$, $\lambda(A) = (\lambda_1(A), ..., \lambda_n(A))$ will be the vector of its eigenvalues ordered in nonincreasing order. (All vectors in this and the following chapters are assumed to be column vectors unless stated otherwise.) By $O(n)$ we will denote the set of all $n \times n$ orthogonal matrices. For

any vector x in $\mathbb{R}^n$, Diag x will denote the diagonal matrix with the vector x on the main diagonal, and $\bar{x}$ will denote the vector with the same entries as x ordered in nonincreasing order, that is $\bar{x}_1 \geq \bar{x}_2 \geq \cdots \geq \bar{x}_n$. Let $\mathbb{R}^n_\downarrow$ denote the set of all vectors x in $\mathbb{R}^n$ such that $x_1 \geq x_2 \geq \cdots \geq x_n$. Let also the operator diag: $S^n \to \mathbb{R}^n$ be defined by diag $(A) = (a_{11}, ..., a_{nn})$. In this chapter $\{M_m\}_{m=1}^\infty$ will denote a sequence of symmetric matrices converging to 0, and $\{U_m\}_{m=1}^\infty$ will denote a sequence of orthogonal matrices. We describe sets in $\mathbb{R}^n$ and functions on $\mathbb{R}^n$ as *symmetric* if they are invariant under coordinate permutations. Thus $f : \mathbb{R}^n \to \mathbb{R}$ will denote a function, defined on an open symmetric set, with the property

$$f(x) = f(Px) \text{ for any permutation matrix } P \text{ and any } x \in \text{domain } f.$$

We denote the gradient of f by ∇f or f', and the Hessian by $\nabla^2 f$ or f''. Vectors are understood to be column vectors, unless stated otherwise. Whenever we denote by μ a vector in $\mathbb{R}^n_\downarrow$ we make the convention that

$$\mu_1 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots \mu_{k_r}, \qquad (k_0 = 0,\ k_r = n).$$

We define a corresponding partition

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

and we call these sets *blocks*. We denote the standard basis in $\mathbb{R}^n$ by $e^1, e^2, ..., e^n$, and e is the vector with all entries equal to 1. We also define corresponding matrices

$$X_l := [e^{k_{l-1}+1}, ..., e^{k_l}], \text{ for all } l = 1, ..., r,$$

For an arbitrary matrix A, A^i will denote its i-th row (a row vector), and $A^{i,j}$ will denote its (i,j)-th entry.

Definition 4.1.1 ([49]). *We say that the vector $\mu \in \mathbb{R}^n$* **block refines** *the vector $b \in \mathbb{R}^n$ if $\mu_i = \mu_j$ implies $b_i = b_j$ for all $i,j \in \{1, ..., n\}$. Equivalently*

$$P\mu = \mu \;\Rightarrow\; Pb = b \quad \text{for all } P \in P(n).$$

(In all of our preliminary results the matrix A will be a diagonal matrix, Diag μ.) We need the following result.

Lemma 4.1.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a symmetric function, twice differentiable at the point $\mu \in \mathbb{R}^n_\downarrow$, and let P be a permutation matrix such that $P\mu = \mu$. Then*

1. $\nabla f(\mu) = P^T \nabla f(\mu)$, *and*
2. $\nabla^2 f(\mu) = P^T \nabla^2 f(\mu) P$.

In particular we have the representation

$$\nabla^2 f(\mu) = \begin{pmatrix} a_{11}E_{11} + b_{k_1}J_1 & a_{12}E_{12} & \cdots & a_{1r}E_{1r} \\ a_{21}E_{21} & a_{22}E_{22} + b_{k_2}J_2 & \cdots & a_{2r}E_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1}E_{r1} & a_{r2}E_{r2} & \cdots & a_{rr}R_{rr} + b_{k_r}J_r \end{pmatrix},$$

where the E_{uv} are matrices of dimensions $|I_u| \times |I_v|$ with all entries equal to one, $(a_{ij})_{i,j=1}^r$ is a real symmetric matrix, $b := (b_1, ..., b_n)$ is a vector which is block refined by μ, and J_u is an identity matrix of the same dimensions as E_{uu}.

Proof. Just apply twice the chain rule to the equality $f(\mu) = f(P\mu)$ in order to get parts 1 and 2. To deduce the block structure of the Hessian, consider the block structure of permutation matrices P such that $P\mu = \mu$: then, when we permute the rows and the columns of the Hessian in the way defined by P, it must stay unchanged. □

Using the notation of this lemma, we define the matrix

$$B := \nabla^2 f(\mu) - \operatorname{Diag} b = (a_{ij}E_{ij})_{i,j=1}^r. \tag{4.1}$$

Note 4.1.3. *We make the convention that if the i-th diagonal block in the above representation has dimensions 1×1 then we set $a_{ii} = 0$ and $b_{k_i} = f''_{k_i k_i}(\mu)$. Otherwise the value of b_{k_i} is uniquely determined as the difference between a diagonal and an off-diagonal element of this block. Note also that the matrix B and the vector b depend on the point μ and the function f.*

Lemma 4.1.4. *For $\mu \in \mathbb{R}^n_\downarrow$ and a sequence of symmetric matrices $M_m \to 0$ we have that*

$$\lambda(\operatorname{Diag}\mu + M_m)^T = \mu^T + \big(\lambda(X_1^T M_m X_1)^T, ..., \lambda(X_r^T M_m X_r)^T\big) + o(\|M_m\|).$$

Proof. Combine Lemma 5.10 in [52] and Theorem 3.12 in [32]. □

The following is our main technical tool.

Lemma 4.1.5. *Let $\{M_m\}$ be a sequence of symmetric matrices converging to 0, such that $M_m/\|M_m\|$ converges to M. Let μ be in $\mathbb{R}^n_\downarrow$ and $U_m \to U \in O(n)$ be a sequence of orthogonal matrices such that*

$$\operatorname{Diag}\mu + M_m = U_m\big(\operatorname{Diag}\lambda(\operatorname{Diag}\mu + M_m)\big)U_m^T, \quad \textit{for all} \quad m = 1, 2, ... \tag{4.2}$$

Then the following properties hold.

1. *The orthogonal matrix U has the form*

$$U = \begin{pmatrix} V_1 & 0 & \cdots & 0 \\ 0 & V_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_r \end{pmatrix},$$

   where V_l is an orthogonal matrix with dimensions $|I_l| \times |I_l|$ for all l.

2. *If $i \in I_l$ then*

$$\lim_{m\to\infty} \frac{1 - \sum_{p\in I_l} (U_m^{i,p})^2}{\|M_m\|} = 0.$$

3. *If i and j do not belong to the same block then*

$$\lim_{m\to\infty} \frac{(U_m^{i,j})^2}{\|M_m\|} = 0.$$

4. *If $i \in I_l$ then*

$$V_l^i \left(\operatorname{Diag} \lambda(X_l^T M X_l)\right) (V_l^i)^T = M^{i,i}.$$

5. *If $i, j \in I_l$, and $p \notin I_l$ then*

$$\lim_{m\to\infty} \frac{U_m^{i,p} U_m^{j,p}}{\|M_m\|} = 0.$$

6. *For any indices $i \neq j$ such that $i, j \in I_l$,*

$$\lim_{m\to\infty} \frac{\sum_{p\in I_l} U_m^{i,p} U_m^{j,p}}{\|M_m\|} = 0.$$

7. *For any indices $i \neq j$ such that $i, j \in I_l$,*

$$V_l^i \left(\operatorname{Diag} \lambda(X_l^T M X_l)\right) (V_l^j)^T = M^{i,j}.$$

8. *For any three indices i, j, p in distinct blocks,*

$$\lim_{m\to\infty} \frac{U_m^{i,p} U_m^{j,p}}{\|M_m\|} = 0.$$

9. For any two indices i, j such that $i \in I_l$, $j \in I_s$, where $l \neq s$,

$$\lim_{m\to\infty}\left(\mu_{k_l}\frac{\sum_{p\in I_l}U_m^{i,p}U_m^{j,p}}{\|M_m\|}+\mu_{k_s}\frac{\sum_{p\in I_s}U_m^{i,p}U_m^{j,p}}{\|M_m\|}\right)=M^{i,j}.$$

Proof. 1. After taking the limit in equation (4.2) we are left with

$$(\operatorname{Diag}\mu)U = U(\operatorname{Diag}\mu).$$

The described representation of the matrix U follows.

2. Let us denote

$$h_m = \left(\lambda(X_1^T M_m X_1)^T, ..., \lambda(X_r^T M_m X_r)^T\right)^T. \tag{4.3}$$

We use Lemma 4.1.4 in equation (4.2) to obtain

$$\operatorname{Diag}\mu + M_m = U_m(\operatorname{Diag}\mu)U_m^T + U_m(\operatorname{Diag}h_m)U_m^T + o(\|M_m\|),$$

and the equivalent form

$$U_m^T(\operatorname{Diag}\mu)U_m + U_m^T M_m U_m = \operatorname{Diag}\mu + \operatorname{Diag}h_m + o(\|M_m\|).$$

We now divide both sides of these equations by $\|M_m\|$ and rearrange:

$$\frac{\operatorname{Diag}\mu - U_m(\operatorname{Diag}\mu)U_m^T}{\|M_m\|} = -\frac{M_m}{\|M_m\|} + \frac{U_m(\operatorname{Diag}h_m)U_m^T}{\|M_m\|} + o(1), \tag{4.4}$$

and

$$\frac{\mathrm{Diag}\,\mu - U_m^T(\mathrm{Diag}\,\mu)U_m}{\|M_m\|} = \frac{U_m^T M_m U_m}{\|M_m\|} - \frac{\mathrm{Diag}\,h_m}{\|M_m\|} - o(1). \qquad (4.5)$$

Notice that the right hand sides of these equations converge to a finite limit as m increases to infinity. If we call the matrix limit of the right hand side of the first equation L, then clearly the limit of the second equation is $-U^T L U$.

We are now going to prove parts 2 and 3 together inductively, by dividing the orthogonal matrix U_m into the same block structure as U. We begin by considering the first row of blocks of U_m.

Let i be an index in the first block, I_1. Then the limit of the (i,i)-th entry in the matrix at the left hand side of equation (4.4) is

$$\lim_{m\to\infty} \frac{\left(\mu_{k_1}\left(1 - \sum_{p\in I_1}\left(U_m^{i,p}\right)^2\right) - \sum_{s=2}^{r}\mu_{k_s}\sum_{p\in I_s}\left(U_m^{i,p}\right)^2\right)}{\|M_m\|} = L^{i,i}. \qquad (4.6)$$

Now recall that

$$L^{i,i} = -M^{i,i} + V_1^i(\mathrm{Diag}\,\lambda(X_1^T M X_1))(V_1^i)^T,$$

and because V_1 is an orthogonal matrix, notice that

$$\begin{aligned}\sum_{i\in I_1} L^{i,i} &= -\mathrm{tr}\,(X_1^T M X_1) + \sum_{i\in I_1} V_1^i(\mathrm{Diag}\,\lambda(X_1^T M X_1))(V_1^i)^T \\ &= -\mathrm{tr}\,(X_1^T M X_1) + \sum_{i\in I_1}\lambda_i(X_1^T M X_1)\sum_{j\in I_1}(V_1^{j,i})^2\end{aligned}$$

$$\begin{aligned} &= -\mathrm{tr}\,(X_1^T M X_1) + \sum_{i\in I_1} \lambda_i (X_1^T M X_1) \\ &= 0. \end{aligned}$$

We now sum equation (4.6) over all i in I_1 to get

$$\lim_{m\to\infty} \frac{\left(\mu_{k_1}\left(|I_1| - \sum_{i,p\in I_1}\left(U_m^{i,p}\right)^2\right) - \sum_{s=2}^{r}\mu_{k_s}\sum_{i\in I_1,\,p\in I_s}\left(U_m^{i,p}\right)^2\right)}{\|M_m\|} = 0.$$

Notice here, that the coefficients in front of the μ_{k_l}, $l = 1, 2, ..., r$ in the numerator sum up to zero. That is,

$$|I_1| - \sum_{i,p\in I_1}\left(U_m^{i,p}\right)^2 - \sum_{s=2}^{r}\sum_{i\in I_1,\,p\in I_s}\left(U_m^{i,p}\right)^2 = 0.$$

So let us choose a number α such that

$$(\mu + \alpha e)_{k_1} > 0 > (\mu + \alpha e)_{k_1+1},$$

and add α to every coordinate of the vector μ thus "shifting" it. The coordinates of the shifted vector that are in the first block are strictly bigger than zero, and the rest are strictly less than zero. By our comment above, the last limit remains true if we "shift" μ in this way. If we rewrite the last limit for the "shifted" vector, because all summands are positive, we immediately see that we must have

$$\lim_{m\to\infty} \frac{|I_1| - \sum_{i,p\in I_1}\left(U_m^{i,p}\right)^2}{\|M_m\|} = 0$$

and

$$\lim_{m\to\infty} \frac{\sum_{i\in I_1,\, p\in I_s} \left(U_m^{i,p}\right)^2}{\|M_m\|} = 0, \quad \text{for all } \ s = 2, .., r.$$

The first of these limits can be written as

$$\lim_{m\to\infty} \frac{\sum_{i\in I_1} \left(1 - \sum_{p\in I_1} \left(U_m^{i,p}\right)^2\right)}{\|M_m\|} = 0,$$

and because all the summands are positive, we conclude that

$$\lim_{m\to\infty} \frac{1 - \sum_{p\in I_1} \left(U_m^{i,p}\right)^2}{\|M_m\|} = 0, \quad \text{for all } i \in I_1.$$

The second of the limits implies immediately that

$$\lim_{m\to\infty} \frac{\left(U_m^{i,p}\right)^2}{\|M_m\|} = 0, \ \text{ for any } i \in I_1,\, p \notin I_1.$$

Thus we proved part 2 for $i \in I_1$ and part 3 for the cases specified above.

Here is a good place to say a few more words about the idea of the proof. As we said, we divide the matrix U_m into blocks complying with the block structure of the vector μ (exactly as in part 1 for the matrix U). We proved part 2 and 3 for the elements in the first row of blocks of this division. What we are going to do now is prove the same thing for the first *column* of blocks. In order to do this we fix an index i in I_1 and consider the (i,i)-th entry in

the matrix at the left hand side of equation (4.5), and take the limit:

$$\lim_{m\to\infty} \frac{\mu_{k_1}\left(1-\sum_{p\in I_1}\left(U_m^{p,i}\right)^2\right)-\sum_{s=2}^{r}\mu_{k_s}\sum_{p\in I_s}\left(U_m^{p,i}\right)^2}{\|M_m\|}$$

$$= -(U^T L U)^{i,i}. \qquad (4.7)$$

Using also the block-diagonal structure of the matrix U, we again have

$$\sum_{i\in I_1}(U^T L U)^{i,i} = \sum_{i\in I_1} L^{i,i} = 0.$$

So we proceed just as before in order to conclude that

$$\lim_{m\to\infty} \frac{1-\sum_{p\in I_1}\left(U_m^{p,i}\right)^2}{\|M_m\|} = 0, \quad \text{for all } i\in I_1,$$

and

$$\lim_{m\to\infty} \frac{\left(U_m^{p,i}\right)^2}{\|M_m\|} = 0, \ \text{ for any } i\in I_1,\ p\notin I_1. \qquad (4.8)$$

We are now ready for the second step of our induction. Let i be an index in I_2. Then the limit of the (i,i)-th entry in the matrix at the left hand side of equation (4.4) is

$$\lim_{m\to\infty} \frac{1}{\|M_m\|}\Bigg(-\mu_{k_1}\sum_{p\in I_1}\left(U_m^{i,p}\right)^2 + \mu_{k_2}\Big(1-\sum_{p\in I_2}\left(U_m^{i,p}\right)^2\Big) -$$

$$\sum_{s=3}^{r} \mu_{k_s} \sum_{p \in I_s} \left(U_m^{i,p}\right)^2 \Bigg) = L^{i,i}.$$

Analogously as above we have

$$\sum_{i \in I_2} L^{i,i} = 0,$$

so summing the above limit over all i in I_2 we get

$$\lim_{m\to\infty} \frac{1}{\|M_m\|} \Bigg(- \mu_{k_1} \sum_{i \in I_2, p \in I_1} \left(U_m^{i,p}\right)^2 + \mu_{k_2} \Big(|I_2| - \sum_{i,p \in I_2} \left(U_m^{i,p}\right)^2 \Big) - \sum_{s=3}^{r} \mu_{k_s} \sum_{i \in I_2, p \in I_s} \left(U_m^{i,p}\right)^2 \Bigg) = 0.$$

We know from (4.8) that

$$\lim_{m\to\infty} \frac{\sum_{i \in I_2, p \in I_1} \left(U_m^{i,p}\right)^2}{\|M_m\|} = 0.$$

So now we choose a number α such that

$$(\mu + \alpha e)_{k_2} > 0 > (\mu + \alpha e)_{k_2+1}$$

and as before exchange μ with its shifted version. Just as before we conclude that

$$\lim_{m\to\infty} \frac{1 - \sum_{p \in I_2} \left(U_m^{i,p}\right)^2}{\|M_m\|} = 0, \text{ for all } i \in I_2,$$

and

$$\lim_{m\to\infty} \frac{(U_m^{i,p})^2}{\|M_m\|} = 0, \text{ for any } i \in I_2,\ p \notin I_2.$$

We repeat the same steps for the second column of blocks in the matrix U_m and so on inductively until we exhaust all the blocks. This completes the proof of parts 2 and 3.

4. For the proof of this part, one needs to consider the (i,i)-th entry of the right hand side of equation (4.4). Because the diagonal of the left hand side converges to zero (by 2 and 3), taking the limit proves the statement in this part.

5. This part follows immediately from part 3.

6. Taking the limit in equation (4.4) gives

$$\lim_{m\to\infty} -\sum_{s\neq l} \mu_{k_s} \frac{\sum_{p\in I_s} U_m^{i,p} U_m^{j,p}}{\|M_m\|} - \mu_{k_l} \frac{\sum_{p\in I_l} U_m^{i,p} U_m^{j,p}}{\|M_m\|} = L^{i,j},$$

where $L^{i,j}$ is the (i,j)-th entry of the limit of the right hand side of equation (4.4). Note that the coefficients of μ_{k_s} again sum up to zero:

$$\sum_{s\neq l} \sum_{p\in I_s} U_m^{i,p} U_m^{j,p} + \sum_{p\in I_l} U_m^{i,p} U_m^{j,p} = 0,$$

because U_m is an orthogonal matrix. Now by part 5 we have

$$0 = \lim_{m\to\infty} -\sum_{s\neq l} \frac{\sum_{p\in I_s} U_m^{i,p} U_m^{j,p}}{\|M_m\|} = \lim_{m\to\infty} \frac{\sum_{p\in I_l} U_m^{i,p} U_m^{j,p}}{\|M_m\|},$$

as required, and moreover $L^{i,j} = 0$.

7. The statement of this part is the detailed way of writing the fact, proved in the previous part, that $L^{i,j} = 0$.

8. This part follows immediately from part 3. (In fact the expression in part 8 is identical to the one in part 5, re-iterated with different index conditions for later convenience.)

9. We again take the limit of the (i,j)-th entry of the matrices on both sides of equation (4.4).

$$\lim_{m\to\infty}\Bigg(-\sum_{t\neq l,s}\mu_{k_t}\frac{\sum_{p\in I_t}U_m^{i,p}U_m^{j,p}}{\|M_m\|}-\mu_{k_l}\frac{\sum_{p\in I_l}U_m^{i,p}U_m^{j,p}}{\|M_m\|}- \mu_{k_s}\frac{\sum_{p\in I_s}U_m^{i,p}U_m^{j,p}}{\|M_m\|}\Bigg)=L^{i,j}.$$

By part 8 we have that all but the l-th and the s-th summand above converge to zero. On the other hand

$$\begin{aligned} L^{i,j} &= \lim_{m\to\infty}\left(-\frac{M_m}{\|M_m\|}+\frac{U_m(\operatorname{Diag}h_m)U_m^T}{\|M_m\|}\right)^{i,j} \\ &= -M^{i,j}+U^i\left(\lim_{m\to\infty}\frac{\operatorname{Diag}h_m}{\|M_m\|}\right)(U^j)^T \\ &= -M^{i,j}, \end{aligned}$$

because U^i and U^j are rows in different blocks and $(\operatorname{Diag}h_m)/\|M_m\|$ converges to a diagonal matrix. □

Now we have all the tools to prove the main result of the chapter.

4.2 Twice differentiable spectral functions

In this section we prove that a symmetric function f is twice differentiable at the point $\lambda(A)$ if and only if the corresponding spectral function $f \circ \lambda$ is twice differentiable at the matrix A.

Recall that the Hadamard product of two matrices $A = [A^{i,j}]$ and $B = [B^{i,j}]$ of the same size is the matrix of their elementwise product $A \circ B = [A^{i,j}B^{i,j}]$. Let the symmetric function $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable at the point $\mu \in \mathbb{R}^n_\downarrow$, where

$$\mu_1 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots \mu_{k_r}, \qquad (k_0 = 0,\ k_r = n).$$

We define the vector $b(\mu) = (b_1(\mu), ..., b_n(\mu))$ as in Lemma 4.1.2. Specifically, for any index i, (say $i \in I_l$ for some $l \in \{1, 2, ..., r\}$) we define

$$b_i(\mu) = \begin{cases} f''_{ii}(\mu), & \text{if } |I_l| = 1. \\ f''_{pp}(\mu) - f''_{pq}(\mu), & \text{for any } p \neq q \in I_l. \end{cases}$$

Lemma 4.1.2 guarantees that the second case of this definition doesn't depend on the choice of p and q. We also define the matrix $\mathcal{A}(\mu)$:

$$\mathcal{A}^{i,j}(\mu) = \begin{cases} 0, & \text{if } \ i = j. \\ b_i(\mu), & \text{if } \ i \neq j \text{ but } i, j \in I_l. \\ \frac{f'_i(\mu) - f'_j(\mu)}{\mu_i - \mu_j}, & \text{otherwise.} \end{cases} \tag{4.9}$$

For simplicity, when the argument is understood by the context, we will write just

b_i and $\mathcal{A}^{i,j}$. The following lemma is Theorem 1.1 in [49].

Lemma 4.2.1. *Let $A \in S^n$ and suppose $\lambda(A)$ belongs to the domain of the symmetric function $f : \mathbb{R}^n \to \mathbb{R}$. Then f is differentiable at the point $\lambda(A)$ if and only if $f \circ \lambda$ is differentiable at the point A. In that case we have the formula*

$$\nabla(f \circ \lambda)(A) = U\big(\operatorname{Diag} \nabla f(\lambda(A))\big)U^T,$$

for any orthogonal matrix U satisfying $A = U\big(\operatorname{Diag} \lambda(A)\big)U^T$.

We recall some standard notions about twice differentiability. Consider a function F from S^n to $\mathbb{R}$. Its gradient at any point A (when it exists) is a linear functional on the Euclidean space S^n, and thus can be identified with an element of S^n, which we denote $\nabla F(A)$. Thus ∇F is a map from S^n to S^n. When this map is itself differentiable at A we say F is *twice differentiable* at A. In this case we can interpret the Hessian $\nabla^2 F(A)$ as a symmetric, bilinear function from $S^n \times S^n$ into $\mathbb{R}$. Its value at a particular point $(H, Y) \in S^n \times S^n$ will be denoted $\nabla^2 F(A)[H, Y]$. In particular, for fixed H, the function $\nabla^2 F(A)[H, \cdot]$ is again a linear functional on S^n, which we consider an element of S^n, for brevity denoted by $\nabla^2 F(A)[H]$. When the Hessian is continuous at A we say F is *twice continuously differentiable* at A. In that case the following identity holds:

$$\nabla^2 F(A)[H, H] = \frac{d^2}{dt^2} F(A + tH)\bigg|_{t=0}.$$

The next theorem is a preliminary version of our main result.

Theorem 4.2.2. *The symmetric function $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable at the point $\mu \in \mathbb{R}^n_\downarrow$ if and only if $f \circ \lambda$ is twice differentiable at the point* $\operatorname{Diag}\mu$. *In that case the Hessian is given by*

$$\nabla^2(f \circ \lambda)(\operatorname{Diag}\mu)[H] \;=\; \operatorname{Diag}\left(\nabla^2 f(\mu)[\operatorname{diag} H]\right) + \mathcal{A} \circ H. \tag{4.10}$$

Hence

$$\nabla^2(f \circ \lambda)(\operatorname{Diag}\mu)[H, H] = \nabla^2 f(\mu)[\operatorname{diag} H, \operatorname{diag} H] + \langle \mathcal{A}, H \circ H \rangle,$$

where $\mathcal{A}$ is defined in (4.9).

Proof. It is easy to see that f must be twice differentiable at the point μ whenever $f \circ \lambda$ is twice differentiable at $\operatorname{Diag}\mu$ because by restricting $f \circ \lambda$ to the subspace of diagonal matrices we get the function f. So the interesting case is the other direction. Let f be twice differentiable at the point $\mu \in \mathbb{R}^n_\downarrow$ and suppose on the contrary that either $f \circ \lambda$ is not twice differentiable at the point $\operatorname{Diag}\mu$, or equation (4.10) fails. Define a linear operator Δ by

$$\Delta(H) := \operatorname{Diag}\left((\nabla^2 f(\mu)(\operatorname{diag} H)\right) + \mathcal{A} \circ H.$$

(Lemma 4.2.1 tells us that $f \circ \lambda$ is at least differentiable around $\operatorname{Diag}\mu$.) So, for this linear operator Δ there is an $\epsilon > 0$ and a sequence of symmetric matrices $\{M_m\}_{m=1}^{\infty}$

converging to 0 such that

$$\frac{\|\nabla(f\circ\lambda)(\operatorname{Diag}\mu+M_m)-\nabla(f\circ\lambda)(\operatorname{Diag}\mu)-\Delta(M_m)\|}{\|M_m\|}>\epsilon$$

for all $m=1,2,....$ Without loss of generality we may assume that the sequence $\{M_m\}_{m=1}^{\infty}$ is such that $M_m/\|M_m\|$ converges to a matrix M, because some subsequence of $\{M_m\}_{m=1}^{\infty}$ surely has this property. Let $\{U_m\}_{m=1}^{\infty}$ be a sequence of orthogonal matrices such that

$$\operatorname{Diag}\mu+M_m=U_m\big(\operatorname{Diag}\lambda(\operatorname{Diag}\mu+M_m)\big)U_m^T,\quad\text{for all}\quad m=1,2,....$$

Without loss of generality we may assume that $U_m\to U\in O(n)$, or otherwise we will just take subsequences of $\{M_m\}_{m=1}^{\infty}$ and $\{U_m\}_{m=1}^{\infty}$. The above inequality shows that for every m there corresponds a pair (or more precisely at least one pair) of indices (i,j) such that

$$\frac{\big|\big(\nabla(f\circ\lambda)(\operatorname{Diag}\mu+M_m)-\operatorname{Diag}\nabla f(\mu)-\Delta(M_m)\big)^{i,j}\big|}{\|M_m\|}>\frac{\epsilon}{n}.\tag{4.11}$$

So at least for one pair of indices, call it again (i,j), we have infinitely many numbers m for which (i,j) is the corresponding pair, and because if necessary we can again take a subsequence of $\{M_m\}_{m=1}^{\infty}$ and $\{U_m\}_{m=1}^{\infty}$ we may assume without loss of generality that there is a pair of indices (i,j) for which the last inequality holds for all $m=1,2,....$ Define the symbol h_m again by equation (4.3). Notice that using Lemma 4.2.1, Lemma 4.1.4, and the fact that ∇f is differentiable at μ,

we get

$$\begin{aligned}\nabla(f\circ\lambda)(\operatorname{Diag}\mu+M_m) &= U_m\big(\operatorname{Diag}\nabla f(\lambda(\operatorname{Diag}\mu+M_m))\big)U_m^T\\ &= U_m\big(\operatorname{Diag}\nabla f(\mu+h_m+o(\|M_m\|))\big)U_m^T\\ &= U_m(\operatorname{Diag}(\nabla f(\mu)+\nabla^2 f(\mu)h_m+o(\|M_m\|)))U_m^T\\ &= U_m(\operatorname{Diag}\nabla f(\mu))U_m^T+U_m(\operatorname{Diag}(\nabla^2 f(\mu)h_m))U_m^T+o(\|M_m\|).\end{aligned} \tag{4.12}$$

We consider three cases. In every case we are going to show that the left hand side of inequality (4.11) actually converges to zero, which contradicts the assumption.

Case I. If $i=j$, then using equation (4.12) the left hand side of inequality (4.11) is less that or equal to

$$\frac{|U_m^i\big(\operatorname{Diag}\nabla f(\mu)\big)(U_m^i)^T-f_i'(\mu)|}{\|M_m\|}+\frac{|U_m^i\big(\operatorname{Diag}\nabla^2 f(\mu)h_m\big)(U_m^i)^T-\big(\nabla^2 f(\mu)(\operatorname{diag}M_m)\big)_i|}{\|M_m\|}+o(1).$$

We are going to show that each summand approaches zero as m goes to infinity. Assume that $i\in I_l$ for some $l\in\{1,...,r\}$. Using the fact that the vector μ block refines the vector $\nabla f(\mu)$ (Lemma 4.1.2, part 1) the first term can be written as

$$\frac{1}{\|M_m\|}\left|f_{k_l}'(\mu)\Big(1-\sum_{p\in I_l}\big(U_m^{i,p}\big)^2\Big)-\sum_{s:s\neq l}f_{k_s}'(\mu)\sum_{p\in I_s}\big(U_m^{i,p}\big)^2\right|.$$

We apply now Lemma 4.1.5 parts 2 and 3 to the last expression.

We now concentrate on the second term above. Using the notation of equation (4.1) (that is, $\nabla^2 f(\mu) = B + \operatorname{Diag} b$) this term is less than or equal to

$$\frac{|U_m^i\big(\operatorname{Diag}(Bh_m)\big)(U_m^i)^T - \big(B(\operatorname{diag} M_m)\big)_i|}{\|M_m\|} + \frac{|U_m^i\big(\operatorname{Diag}((\operatorname{Diag} b)h_m)\big)(U_m^i)^T - \big((\operatorname{Diag} b)(\operatorname{diag} M_m)\big)_i|}{\|M_m\|}. \tag{4.13}$$

As m approaches infinity, we have that $U_m^i \to U^i$. We define the vector h to be:

$$h := \lim_{m\to\infty} \frac{h_m}{\|M_m\|} = \big(\lambda(X_1^T M X_1)^T, ..., \lambda(X_r^T M X_r)^T\big)^T.$$

So taking limits, expression (4.13) turn into:

$$|U^i\big(\operatorname{Diag}(Bh)\big)(U^i)^T - \big(B(\operatorname{diag} M)\big)_i| + |U^i\big(\operatorname{Diag}((\operatorname{Diag} b)h)\big)(U^i)^T - \big((\operatorname{Diag} b)(\operatorname{diag} M)\big)_i|.$$

We are going to investigate each absolute value separately and show that they are both actually equal to zero. For the first, we use the block structure of the matrix B (see Lemma 4.1.2) and the block structure of the vector h to obtain

$$(Bh)_j = \sum_{s=1}^{r} a_{qs} \operatorname{tr}(X_s^T M X_s), \quad \text{when } j \in I_q.$$

Using the fact that $i \in I_l$ and that V_l is orthogonal we get

$$U^i\big(\operatorname{Diag}(Bh)\big)(U^i)^T = \big(V_l^i X_l^T\big)\big(\operatorname{Diag}(Bh)\big)\big(X_l(V_l^i)^T\big)$$

$$\begin{aligned} &= V_l^i\big(X_l^T(\mathrm{Diag}\,(Bh))X_l\big)(V_l^i)^T \\ &= \Big(\sum_{s=1}^{r} a_{ls}\mathrm{tr}\,(X_s^T M X_s)\Big)\Big(\sum_{s=1}^{|I_l|}(V_l^{i,s})^2\Big) \\ &= \sum_{s=1}^{r} a_{ls}\mathrm{tr}\,(X_s^T M X_s) \\ &= (B\mathrm{diag}\,M)_i, \end{aligned}$$

which shows that the first absolute value is zero. For the second absolute value, we use the block structure of the vector b, to write

$$(\mathrm{Diag}\,b)h = \big(b_{k_1}\lambda(X_1^T M X_1)^T, ..., b_{k_r}\lambda(X_r^T M X_r)^T\big)^T.$$

In the next to the last equality below we use part (4) of Lemma 4.1.5:

$$\begin{aligned} U^i\big(\mathrm{Diag}\,((\mathrm{Diag}\,b)h)\big)(U^i)^T &= \big(V_l^i X_l^T\big)\big(\mathrm{Diag}\,((\mathrm{Diag}\,b)h)\big)\big(X_l(V_l^i)^T\big) \\ &= V_l^i\big(X_l^T(\mathrm{Diag}\,((\mathrm{Diag}\,b)h))X_l\big)(V_l^i)^T \\ &= V_l^i\big(\mathrm{Diag}\,b_{k_l}\lambda(X_l^T M X_l)\big)(V_l^i)^T \\ &= b_{k_l}M^{i,i} \\ &= \big((\mathrm{Diag}\,b)(\mathrm{diag}\,M)\big)_i. \end{aligned}$$

We can see now that the second absolute value is also zero.

Case II. If $i \neq j$ but $i, j \in I_l$ for some $l \in \{1, 2, ...r\}$, then using equation (4.12)

the left hand side of inequality (4.11) becomes

$$\frac{|U_m^i\big(\operatorname{Diag}\nabla f(\mu)\big)(U_m^j)^T + U_m^i\big(\operatorname{Diag}\left(\nabla^2 f(\mu)h_m\right)\big)(U_m^j)^T - b_{k_l}M_m^{i,j}|}{\|M_m\|} + o(1).$$

Using the fact that μ block refines vector $\nabla f(\mu)$, we can write the first summand in the absolute value as

$$\frac{1}{\|M_m\|}\Big(\sum_{s\neq l} f'_{k_s}(\mu)\sum_{p\in I_s} U_m^{i,p}U_m^{j,p} + f'_{k_l}(\mu)\sum_{p\in I_l} U_m^{i,p}U_m^{j,p}\Big).$$

We use parts 5 and 6 of Lemma 4.1.5 to conclude that this expression converges to zero. We are left with

$$\frac{|U_m^i\big(\operatorname{Diag}\left(\nabla^2 f(\mu)h_m\right)\big)(U_m^j)^T - b_{k_l}M_m^{i,j}|}{\|M_m\|}.$$

Substituting above $\nabla^2 f(\mu) = B + \operatorname{Diag} b$ we get

$$\frac{|U_m^i\big(\operatorname{Diag}\left(Bh_m\right)\big)(U_m^j)^T + U_m^i\big(\operatorname{Diag}\left((\operatorname{Diag} b)h_m\right)\big)(U_m^j)^T - b_{k_l}M_m^{i,j}|}{\|M_m\|}.$$

Recall the notation from Lemma 4.1.2 used to denote the entries of the matrix B. Then the limit of the first summand above can be written as

$$\begin{aligned}\lim_{m\to\infty}\frac{|U_m^i\big(\operatorname{Diag}\left(Bh_m\right)\big)(U_m^j)^T|}{\|M_m\|} &= |U^i\big(\operatorname{Diag}\left(Bh\right)\big)(U^j)^T| \\ &= \sum_{s=1}^{r}\Big(\Big(\sum_{l=1}^{r} a_{sl}\operatorname{tr}\left(X_l^T M X_l\right)\Big)\sum_{p\in I_s} U^{i,p}U^{j,p}\Big) \\ &= 0,\end{aligned}$$

because clearly $\sum_{p\in I_s} U^{i,p}U^{j,p} = 0$ for all $s \in \{1,2,...r\}$. We are left with the following limit

$$\lim_{m\to\infty} \frac{|U_m^i\big(\mathrm{Diag}\,((\mathrm{Diag}\,b)h_m)\big)(U_m^j)^T - b_{k_l}M_m^{i,j}|}{\|M_m\|} \\ = |U^i\big(\mathrm{Diag}\,((\mathrm{Diag}\,b)h)\big)(U^j)^T - b_{k_l}M^{i,j}|.$$

Using Lemma 4.1.5 part 7 we observe that the last absolute value is zero.

Case III. If $i \in I_l$ and $j \in I_s$, where $l \neq s$, then using equation (4.12), the left hand side of inequality (4.11) becomes (up to $o(1)$)

$$\frac{|U_m^i\big(\mathrm{Diag}\,\nabla f(\mu)\big)(U_m^j)^T + U_m^i\big(\mathrm{Diag}\,\nabla^2 f(\mu)h_m\big)(U_m^j)^T - \frac{f'_{k_l}(\mu)-f'_{k_s}(\mu)}{\mu_{k_l}-\mu_{k_s}}M_m^{i,j}|}{\|M_m\|}.$$

We start with the second term above. Its limit is

$$\lim_{m\to\infty} \frac{U_m^i\big(\mathrm{Diag}\,(\nabla^2 f(\mu)h_m)\big)(U_m^j)^T}{\|M_m\|} = U^i\big(\mathrm{Diag}\,(\nabla^2 f(\mu)h)\big)(U^j)^T = 0,$$

because in our case, U^i has nonzero coordinates where the entries of U^j are zero. We are left with

$$\lim_{m\to\infty} \left| \frac{U_m^i\big(\mathrm{Diag}\,\nabla f(\mu)\big)(U_m^j)^T}{\|M_m\|} - \frac{f'_{k_l}(\mu) - f'_{k_s}(\mu)}{\mu_{k_l}-\mu_{k_s}} \frac{M_m^{i,j}}{\|M_m\|} \right|. \tag{4.14}$$

We expand the first term in this limit.

$$\frac{U_m^i\big(\mathrm{Diag}\,\nabla f(\mu)\big)(U_m^j)^T}{\|M_m\|} = f'_{k_l}(\mu)\frac{\sum_{p\in I_l} U_m^{i,p}U_m^{j,p}}{\|M_m\|} + f'_{k_s}(\mu)\frac{\sum_{p\in I_s} U_m^{i,p}U_m^{j,p}}{\|M_m\|} + \sum_{t\neq l,s} f'_{k_t}(\mu)\frac{\sum_{p\in I_t} U_m^{i,p}U_m^{j,p}}{\|M_m\|}.$$

Using Lemma 4.1.5 part 8 we see that the third summand above converges to zero as m goes to infinity. Part 9 of the same lemma tells us that

$$\lim_{m\to\infty}\frac{M_m^{i,j}}{\|M_m\|} = \lim_{m\to\infty}\left(\mu_{k_l}\frac{\sum_{p\in I_l} U_m^{i,p}U_m^{j,p}}{\|M_m\|} + \mu_{k_s}\frac{\sum_{p\in I_s} U_m^{i,p}U_m^{j,p}}{\|M_m\|}\right).$$

In order to abbreviate the formulae we introduce the following notation

$$\beta_m^l := \frac{\sum_{p\in I_l} U_m^{i,p}U_m^{j,p}}{\|M_m\|}, \qquad \text{for all} \quad l = 1, 2, ..., r.$$

Substituting everything in (4.14) we get the following equivalent limit:

$$\lim_{m\to\infty}\left|\left(f'_{k_l}(\mu)\beta_m^l + f'_{k_s}(\mu)\beta_m^s\right) - \frac{f'_{k_l}(\mu) - f'_{k_s}(\mu)}{\mu_{k_l} - \mu_{k_s}}\left(\mu_{k_l}\beta_m^l + \mu_{k_s}\beta_m^s\right)\right|.$$

Simplifying we get

$$\lim_{m\to\infty}(\beta_m^l + \beta_m^s)\frac{f'_{k_s}(\mu)\mu_{k_l} - f'_{k_l}(\mu)\mu_{k_s}}{\mu_{k_l} - \mu_{k_s}}.$$

Notice now that

$$\sum_{l=1}^{r}\beta_m^l = 0, \quad \text{for all } m,$$

because U_m is an orthogonal matrix and the numerator of the above sum is the product of its i-th and the j-th row. Next, Lemma 4.1.5, part 8 says that

$$\lim_{m\to\infty} \sum_{t\neq l,s} \beta_m^t = 0,$$

so

$$\lim_{m\to\infty} (\beta_m^l + \beta_m^s) = 0,$$

which completes the proof. □

We are finally ready to give and prove the full version of our main result.

Theorem 4.2.3. *Let A be an $n \times n$ symmetric matrix. The symmetric function $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable at the point $\lambda(A)$ if and only if the spectral function $f \circ \lambda$ is twice differentiable at the matrix A. Moreover in this case the Hessian of the spectral function at the matrix A is*

$$\nabla^2(f \circ \lambda)(A)[H] = W\big(\mathrm{Diag}\,\big(\nabla^2 f(\lambda(A))\mathrm{diag}\,\tilde{H}\big) + \mathcal{A} \circ \tilde{H}\big)W^T,$$

where W is any orthogonal matrix such that $A = W\big(\mathrm{Diag}\,\lambda(A)\big)W^T$, $\tilde{H} = W^T H W$, and $\mathcal{A} = \mathcal{A}(\lambda(A))$ is defined by equation (4.9). Hence

$$\nabla^2(f \circ \lambda)(A)[H, H] = \nabla^2 f(\lambda(A))[\mathrm{diag}\,\tilde{H}, \mathrm{diag}\,\tilde{H}] + \langle \mathcal{A}, \tilde{H} \circ \tilde{H}\rangle.$$

Proof. Let W be an orthogonal matrix which diagonalizes A in an ordered fashion, that is

$$A = W\big(\mathrm{Diag}\,\lambda(A)\big)W^T.$$

Let M_m be a sequence of symmetric matrices converging to zero, and let U_m be a sequence of orthogonal matrices such that

$$\operatorname{Diag}\lambda(A) + W^T M_m W = U_m\big(\operatorname{Diag}\lambda(\operatorname{Diag}\lambda(A) + W^T M_m W)\big)U_m^T.$$

Then using Lemma 4.2.1 we get

$$\begin{aligned}
&\nabla(f\circ\lambda)(A+M_m)\\
&\quad= \nabla(f\circ\lambda)\big(W(\operatorname{Diag}\lambda(A) + W^T M_m W)W^T\big)\\
&\quad= \nabla(f\circ\lambda)\big(WU_m(\operatorname{Diag}\lambda(\operatorname{Diag}\lambda(A) + W^T M_m W))U_m^T W^T\big)\\
&\quad= WU_m\big(\operatorname{Diag}\nabla f(\lambda(\operatorname{Diag}\lambda(A) + W^T M_m W)))\big)U_m^T W^T.
\end{aligned}$$

We also have that

$$\nabla(f\circ\lambda)(A) = W\big(\operatorname{Diag}\nabla f(\lambda(A))\big)W^T,$$

and $W^T M_m W \to 0$, as m goes to infinity. Because W is an orthogonal matrix we have $\|WXW^T\| = \|X\|$ for any matrix X. It is now easy to check the result by Theorem 4.2.2. □

4.3 Continuity of the Hessian

Suppose now the symmetric function $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable in a neighbourhood of the point $\lambda(A)$ and its Hessian is continuous at the point $\lambda(A)$. Then

$f \circ \lambda$ as we saw above will be twice differentiable in a neighbourhood of the point A, and in this section we are going to show that $\nabla^2(f \circ \lambda)$ is also continuous at the point A.

We define a basis, $\{H_{ij}\}$, on the space of symmetric matrices. If $i \neq j$ all the entries of the matrix H_{ij} are zeros, except the (i,j)-th and (j,i)-th, which are one. If $i = j$ we have one only on the (i,i)-th position. It suffices to prove that the Hessian is continuous when applied to any matrix of the basis. We begin with a lemma treating, in some sense, all special cases at once.

Lemma 4.3.1. *Let $\mu \in \mathbb{R}^n_\downarrow$ be such that*

$$\mu_1 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots \mu_{k_r}, \qquad (k_0 = 0,\ k_r = n).$$

and let the symmetric function $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable at the point μ. Let $\{\mu^m\}_{m=1}^\infty$ be a sequence of vectors in $\mathbb{R}^n$ converging to μ. Then

$$\lim_{m\to\infty} \nabla^2(f \circ \lambda)(\operatorname{Diag}\mu^m) = \nabla^2(f \circ \lambda)(\operatorname{Diag}\mu).$$

Proof. For every m there is a permutation matrix P_m such that $P_m^T \mu^m = \overline{\mu^m}$. (See the beginning of Section 4.1 for the meaning of the bar above a vector.) But there are finitely many permutation matrices (namely $n!$) so we can form $n!$ subsequences of $\{\mu^m\}$ such that any two vectors in a particular subsequence can be ordered in descending order by the same permutation matrix. If we prove the lemma for every such subsequence we will be done. So without loss of generality we may assume that $P^T \mu^m = \overline{\mu^m}$ for every m, and some fixed permutation matrix P. Clearly for

all large enough m, we are going to have

$$\mu^m_{k_1} > \mu^m_{k_1+1}, \quad \mu^m_{k_2} > \mu^m_{k_2+1}, \cdots, \mu^m_{k_{r-1}} > \mu^m_{k_{r-1}+1},$$

Consequently the matrix P is block-diagonal with permutation matrices on the main diagonal, and dimensions matching the block structure of μ, so $P\mu = \mu$. Consider now the block structure of the vectors $\{\overline{\mu^m}\}$. Because there are finitely many different block structures, we can divide this sequence into subsequences such that the vectors in a particular subsequence have the same block structure. If we prove the lemma for each subsequence we will be done. So without loss of generality we may assume that the vectors $\{\overline{\mu^m}\}$ have the same block structure for every m. Next, using the formula for the Hessian in Theorem 4.2.3 we have

$$\begin{aligned}&\nabla^2(f\circ\lambda)(\mathrm{Diag}\,\mu^m)[H_{ij}] = \\ &\quad P\big(\mathrm{Diag}\,\big(\nabla^2 f(\overline{\mu^m})\mathrm{diag}\,(P^T H_{ij} P)\big) + \mathcal{A}(\overline{\mu^m})\circ(P^T H_{ij} P)\big)P^T,\end{aligned}$$

and Lemma 4.1.2 together with Theorem 4.2.2 give us

$$\begin{aligned}\nabla^2(f\circ\lambda)(\mathrm{Diag}\,\mu)[H_{ij}] &= \mathrm{Diag}\,\big(\nabla^2 f(\mu)\mathrm{diag}\,H_{ij}\big) + \mathcal{A}(\mu)\circ H_{ij} \\ &= P\big(\mathrm{Diag}\,\big(\nabla^2 f(\mu)\mathrm{diag}\,(P^T H_{ij} P)\big) + \\ &\qquad \mathcal{A}(\mu)\circ(P^T H_{ij} P)\big)P^T.\end{aligned}$$

These equations show that without loss of generality it suffices to prove the lemma only in the case when all vectors $\{\mu^m\}$ are ordered in descending order, that is, the

vectors μ^m all block refine the vector μ. In that case we have

$$\nabla^2(f \circ \lambda)(\mathrm{Diag}\,\mu^m)[H_{ij}] = \mathrm{Diag}\left(\nabla^2 f(\mu^m)\mathrm{diag}\,H_{ij}\right) + \mathcal{A}(\mu^m) \circ H_{ij},$$

and

$$\nabla^2(f \circ \lambda)(\mathrm{Diag}\,\mu)[H_{ij}] = \mathrm{Diag}\left(\nabla^2 f(\mu)\mathrm{diag}\,H_{ij}\right) + \mathcal{A}(\mu) \circ H_{ij}.$$

We consider four cases.

Case I. If $i = j$ then

$$\begin{aligned}\lim_{m\to\infty} \nabla^2(f \circ \lambda)(\mathrm{Diag}\,\mu^m)[H_{ij}] &= \lim_{m\to\infty} \mathrm{Diag}\left(\nabla^2 f(\mu^m)e^i\right) \\ &= \mathrm{Diag}\left(\nabla^2 f(\mu)e^i\right) \\ &= \nabla^2(f \circ \lambda)(\mathrm{Diag}\,\mu)[H_{ij}],\end{aligned}$$

just because $\nabla^2 f(\cdot)$ is continuous at μ.

Case II. If $i \neq j$, but belong to the same block for μ^m, then i, j will be in the same block of μ as well and we have

$$\begin{aligned}\lim_{m\to\infty} \nabla^2(f \circ \lambda)(\mathrm{Diag}\,\mu^m)[H_{ij}] &= \lim_{m\to\infty} b_i(\mu^m)H_{ij} \\ &= b_i(\mu)H_{ij} \\ &= \nabla^2(f \circ \lambda)(\mathrm{Diag}\,\mu)[H_{ij}],\end{aligned}$$

again because $\nabla^2 f(\cdot)$ is continuous at μ.

Case III. If i and j belong to different blocks of μ^m but to the same block of μ, then

$$\lim_{m\to\infty} \nabla^2(f\circ\lambda)(\mathrm{Diag}\,\mu^m)[H_{ij}] \;=\; \lim_{m\to\infty} \frac{f_i'(\mu^m)-f_j'(\mu^m)}{\mu_i^m-\mu_j^m}H_{ij},$$

and

$$\nabla^2(f\circ\lambda)(\mathrm{Diag}\,\mu)[H_{ij}] = b_i(\mu)H_{ij}.$$

So we have to prove that

$$\lim_{m\to\infty} \frac{f_i'(\mu^m)-f_j'(\mu^m)}{\mu_i^m-\mu_j^m} = f_{ii}''(\mu)-f_{ij}''(\mu).$$

(See the definition of $b_i(\mu)$ in the beginning of Section 4.2.) For every m we define the vectors $\dot{\mu}^m$ and $\ddot{\mu}^m$ coordinatewise as follows

$$\dot{\mu}_p^m = \begin{cases} \mu_p^m, & p\neq i \\ \mu_j^m, & p=i \end{cases}, \qquad \ddot{\mu}_p^m = \begin{cases} \mu_p^m, & p\neq i,j \\ \mu_j^m, & p=i \\ \mu_i^m, & p=j. \end{cases}$$

Because $\mu_i=\mu_j$ we conclude that both sequences $\{\dot{\mu}^m\}_{m=1}^\infty$ and $\{\ddot{\mu}^m\}_{m=1}^\infty$ converge to μ, because $\{\mu^m\}_{m=1}^\infty$ does so. Below we are applying the mean value theorem

twice:

$$\begin{aligned}
\frac{f_i'(\mu^m) - f_j'(\mu^m)}{\mu_i^m - \mu_j^m} &= \frac{f_i'(\mu^m) - f_i'(\dot{\mu}^m) + f_i'(\dot{\mu}^m) - f_j'(\mu^m)}{\mu_i^m - \mu_j^m} \\
&= \frac{(\mu_i^m - \mu_j^m) f_{ii}''(\xi^m) + f_i'(\dot{\mu}^m) - f_j'(\mu^m)}{\mu_i^m - \mu_j^m} \\
&= f_{ii}''(\xi^m) + \frac{f_i'(\dot{\mu}^m) - f_i'(\ddot{\mu}^m) + f_i'(\ddot{\mu}^m) - f_j'(\mu^m)}{\mu_i^m - \mu_j^m} \\
&= f_{ii}''(\xi^m) + \frac{(\mu_j^m - \mu_i^m) f_{ij}''(\eta^m) + f_i'(\ddot{\mu}^m) - f_j'(\mu^m)}{\mu_i^m - \mu_j^m} \\
&= f_{ii}''(\xi^m) - f_{ij}''(\eta^m),
\end{aligned}$$

where ξ^m is a vector between μ^m and $\dot{\mu}^m$, and η^m is a vector between $\dot{\mu}^m$ and $\ddot{\mu}^m$. Consequently $\xi^m \to \mu$, and $\eta^m \to \mu$. Notice that vector $\ddot{\mu}^m$ is obtained from μ^m by swapping the i-th and the j-th coordinate. Then using the first part of Lemma 4.1.2 we see that $f_i'(\ddot{\mu}^m) = f_j'(\mu^m)$. Finally we just have to take the limit above and use again the continuity of the Hessian of f at the point μ.

Case IV. If i and j belong to different blocks of μ^m and to different blocks of μ, then

$$\begin{aligned}
\lim_{m\to\infty} \nabla^2(f \circ \lambda)(\operatorname{Diag} \mu^m)[H_{ij}] &= \lim_{m\to\infty} \frac{f_i'(\mu^m) - f_j'(\mu^m)}{\mu_i^m - \mu_j^m} H_{ij} \\
&= \frac{f_i'(\mu) - f_j'(\mu)}{\mu_i - \mu_j} H_{ij} \\
&= \nabla^2(f \circ \lambda)(\operatorname{Diag} \mu)[H_{ij}],
\end{aligned}$$

because $\nabla f(\cdot)$ is continuous at μ and the denominator is never zero. □

Now we are ready to prove the main result of this section.

Theorem 4.3.2. *Let A be an $n \times n$ symmetric matrix. The symmetric function $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable at the point $\lambda(A)$ if and only if the spectral function $f \circ \lambda$ is twice continuously differentiable at the matrix A.*

Proof. We know that $f \circ \lambda$ is twice differentiable at A if and only if f is twice differentiable at $\lambda(A)$, so what is left to prove is the continuity of the Hessian. Suppose that f is twice continuously differentiable at $\lambda(A)$ and that $f \circ \lambda$ is not twice continuously differentiable at A. That is, the Hessian $\nabla^2(f \circ \lambda)$ is not continuous at A. Take a sequence, $\{A_m\}_{m=1}^{\infty}$, of symmetric matrices converging to A such that for some $\epsilon > 0$ we have

$$\|\nabla^2(f \circ \lambda)(A_m) - \nabla^2(f \circ \lambda)(A)\| > \epsilon,$$

for all m. Let $\{U_m\}_{m=1}^{\infty}$ be a sequence of orthogonal matrices such that

$$A_m = U_m \big(\operatorname{Diag} \lambda(A_m)\big) U_m^T.$$

Without loss of generality we may assume that $U_m \to U$, where U is orthogonal and then

$$A = U \big(\operatorname{Diag} \lambda(A)\big) U^T.$$

(Otherwise we take subsequences of $\{A_m\}$ and $\{U_m\}$.) Using the formula for the

Hessian given in Theorem 4.2.3 and Lemma 4.3.1 we can easily see that

$$\lim_{m\to\infty} \nabla^2(f\circ\lambda)(A_m)[H] = \nabla^2(f\circ\lambda)(A)[H],$$

for every symmetric H. This is a contradiction.

The other direction follows from the chain rule after observing

$$f(x) = (f\circ\lambda)(\mathrm{Diag}\, x).$$

This completes the proof. □

4.4 Example and Conjecture

As an example, suppose we require the second directional derivative of the function $f\circ\lambda$ at the point A in the direction B. That is, we want to find the second derivative of the function

$$g(t) = (f\circ\lambda)(A+tB),$$

at $t=0$. Let W be an orthogonal matrix such that $A = W(\mathrm{Diag}\,\lambda(A))W^T$. Let $\tilde{B} = W^TBW$. We differentiate twice:

$$g''(t) \;=\; \nabla^2(f\circ\lambda)(A+tB)[B,B].$$

Using Lemma 4.2.1 and Theorem 4.2.3 at $t=0$ we get

$$g(0) \;=\; f(\lambda(A))$$

$$\begin{aligned} g'(0) &= \operatorname{tr}\left(\tilde{B}\operatorname{Diag}\nabla f(\lambda(A))\right) \\ g''(0) &= \nabla^2(f\circ\lambda)(\lambda(A))[\operatorname{diag}\tilde{B},\operatorname{diag}\tilde{B}] + \langle \mathcal{A}, \tilde{B}\circ\tilde{B}\rangle \\ &= \sum_{i,j=1}^{n} f''_{ij}(\lambda(A))(\tilde{B}^{i,i})(\tilde{B}^{j,j}) + \sum_{\substack{i\neq j \\ \lambda_i=\lambda_j}} b_i(\tilde{B}^{i,j})^2 \\ &\quad + \sum_{\substack{i,j \\ \lambda_i\neq\lambda_j}} + \sum_{\substack{i,\,j \\ \lambda_i\,\neq\,\lambda_j}} \frac{f'_i(\lambda(A)) - f'_j(\lambda(A))}{\lambda_i(A)-\lambda_j(A)}(\tilde{B}^{i,j})^2, \end{aligned}$$

In principle, if the function f is analytic, this second directional derivative can also be computed using the implicit formulae from [88]. Some work shows that the answers agree.

As a final illustration, consider the classical example of the power series expansion of a simple eigenvalue. In this case we consider the function f given by

$$f(x) = \bar{x}_k := \text{the } k\text{-th largest entry in } x,$$

and the matrix

$$A = \operatorname{Diag}\mu,$$

where $\mu \in \mathbb{R}^n_{\downarrow}$ and

$$\mu_{k-1} > \mu_k > \mu_{k+1}.$$

Then we have

$$f'(\mu) = e^k, \quad \text{and} \quad f''(\mu) = 0,$$

so for the function $g(t) = \lambda_k(\operatorname{Diag}\mu + tB)$ our results show the following formulae

(familiar in perturbation theory and quantum mechanics):

$$\begin{aligned} g(0) &= \mu_k \\ g'(0) &= B^{k,k} \\ g''(0) &= \sum_{j\neq k} \frac{1}{\mu_k - \mu_j}(B^{k,j})^2 + \sum_{i\neq k} \frac{-1}{\mu_i - \mu_k}(B^{i,k})^2 \\ &= 2\sum_{j\neq k} \frac{1}{\mu_k - \mu_j}(B^{k,j})^2. \end{aligned}$$

This agrees with the result in [41, p. 92]. As we will see in the next chapter, this result can be written using the notion of the Moore-Penrose generalized inverse.

We conclude with the following natural conjecture.

Conjecture 4.4.1. *A spectral function $f \circ \lambda$ is k-times differentiable at the matrix A if and only if its corresponding symmetric function f is k-times differentiable at the point $\lambda(A)$. Moreover, $f \circ \lambda$ is $\mathcal{C}^k$ if and only if f is $\mathcal{C}^k$.*

Chapter 5

Quadratic expansions of spectral functions

In this chapter we relax the assumptions from Chapter 4. We assume that the symmetric function f has a quadratic expansion at the point $\lambda(A)$ and we show that this happens if and only if $f \circ \lambda$ has a quadratic expansion at A. Notice that having a quadratic expansion is a weaker property than being twice differentiable.

5.1 Notation and definitions

We use the notation from the previous chapters. The following definition explains the main property that interests us here.

Definition 5.1.1. *We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ has a* **weak quadratic expansion** *at the point x if there exists a vector $\nabla f(x)$ and a symmetric matrix*

$\nabla^2 f(x)$ *such that for small* $h \in \mathbb{R}^n$

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2}\langle h, \nabla^2 f(x) h \rangle + o(\|h\|^2),$$

and a **strong quadratic expansion** *at the point* x *if*

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2}\langle h, \nabla^2 f(x) h \rangle + O(\|h\|^3).$$

The vector h *is called a* **perturbation vector**.

A few comments on this definition are necessary. Clearly having strong quadratic expansion implies the weak quadratic expansion. We want to alert the reader that a function may not be twice differentiable at the point x but still possesses a strong quadratic expansion at that point. (See, for example, (1.3) in the Introduction.) It is clear that if the function has quadratic expansion at the point x then it is differentiable at x and its gradient is the vector $\nabla f(x)$ from the above definition. If the function has weak quadratic expansion, then there is a unique vector $\nabla f(x)$, and a short elementary argument shows that there is a unique symmetric matrix $\nabla^2 f(x)$ (*the Hessian*) for which the expansion holds. There is a slight abuse of notation when we call $\nabla^2 f(x)$ the Hessian of f, but no danger of confusion exists because when f is twice differentiable at x the symmetric matrix $\nabla^2 f(x)$ is exactly the Hessian. Finally, another way to write the strong quadratic expansion of a function f, consistent with [68], is

$$f(x+h) = f(x) + \nabla f(x)[h] + \frac{1}{2}\nabla^2 f(x)[h,h] + O(\|h\|^3). \tag{5.1}$$

We give some less common notation which will be used throughout the chapter. It is taken directly from [87]. We are interested in quadratic expansions of matrix functions $f \circ \lambda$ around a matrix A. Let $H \in S^n$ be the perturbation matrix. We assume "block structure" of the vector $\lambda(A)$ given by (cf. page 84)

$$\begin{aligned}\lambda_1(A) = \cdots = \lambda_{k_1}(A) > \cdots > \lambda_{k_{l-1}+1}(A) = \cdots = \lambda_m(A) = \cdots = \lambda_{k_l}(A) \\ > \cdots \lambda_{k_r}(A), \quad (k_0 = 0,\ k_r = n).\end{aligned}$$

That is, the eigenvalue $\lambda_m(A)$ lies in the l'th block of equal eigenvalues. Let $X = [x^1, ..., x^n]$ be an orthogonal matrix such that $X^T A X = \operatorname{Diag} \lambda(A)$ (so x^i is a unit eigenvector corresponding to $\lambda_i(A)$) and let

$$X_l = [x^{k_{l-1}+1}, ..., x^{k_l}].$$

Let $U_l = [v^1, ..., v^{k_l - k_{l-1}}]$ be a $(k_l - k_{l-1}) \times (k_l - k_{l-1})$ orthogonal matrix such that

$$U_l^T (X_l^T H X_l) U_l = \operatorname{Diag} \lambda(X_l^T H X_l).$$

Set $H_l := X_l^T H X_l$, $1 \le l \le r$, and suppose

$$\begin{aligned}\lambda_1(H_l) = \cdots = \lambda_{t_{l,1}}(H_l) > \cdots > \lambda_{t_{l,j-1}+1}(H_l) = \cdots \lambda_{m-k_{l-1}}(H_l) \cdots \\ = \lambda_{t_{l,j}}(H_l) > \cdots \lambda_{t_{l,s_l}}(H_l), \quad (t_{l,0} = 0,\ t_{l,s_l} = k_l - k_{l-1})\end{aligned}$$

Finally let

$$U_{l,j} = [v^{t_{l,j-1}+1}, ..., v^{t_{l,j}}].$$

We should point out that $X_l = X_l(A, m)$, and $U_{l,j} = U_{l,j}(A, H, X, m)$ but from now on we will write only X_l and $U_{l,j}$ to simplify the notation.

By $A^\dagger$ we denote the Moore-Penrose generalized inverse of the matrix A. For more information on the topic see [84, p.102]. But for our needs, because we will be working only with symmetric matrices, the concept can be quickly explained. First, $(\mathrm{Diag}\, x)^\dagger_{i,j}$ is equal to $1/x_i$ if $i = j$ and $x_i \neq 0$, and is 0 otherwise. Second, for any orthogonal matrix U, that diagonalizes A, we have $A^\dagger = (U\mathrm{Diag}\,\lambda(A)U^T)^\dagger := U(\mathrm{Diag}\,\lambda(A))^\dagger U^T$.

5.2 Supporting results

Let A be in S^n and its eigenvalues have the following block structure

$$\lambda_1(A) = \cdots = \lambda_{k_1}(A) > \lambda_{k_1+1}(A) = \cdots = \lambda_{k_2}(A) > \lambda_{k_2+1}(A) \cdots\cdots \lambda_{k_r}(A),$$

where $k_r = n$. All our results in this chapter rest on the fact that for every block $l = 1, ..., r$, the following two functions have quadratic expansions at A:

$$\sigma_{k_l}(\cdot) = \sum_{i=1}^{k_l} \lambda_i(\cdot)$$

$$S_l(\cdot) = \sum_{i=k_{l-1}+1}^{k_l} \lambda_i^2(\cdot).$$

We are going to give three justifications of this fact and two of them will show that these functions are even analytic at A. For every index $m = 1, ..., n$ and every block

$l = 1, ..., r$ define the functions

$$f_m(x) = \sum_{i=1}^{m} \bar{x}_i$$
$$s_l(x) = \sum_{i=k_{l-1}+1}^{k_l} \bar{x}_i^2.$$

The function f_m is the sum of the m largest entries in x. The functions f_m and $s_l(x)$ are symmetric. (A function f is *symmetric* if $f(x) = f(Px)$ for any permutation matrix P. We denote the set of all $n \times n$ permutation matrices by $P(n)$.) It is clear that if the point x is such that $\bar{x}_m > \bar{x}_{m+1}$ then f_m is linear near x. In particular, for points x near $\lambda(A)$ the functions $f_{k_l}(x)$ and $s_l(x)$ are both polynomials in the entries of x. Notice also that

$$\sigma_{k_l}(\cdot) = (f_{k_l} \circ \lambda)(\cdot)$$
$$S_l(\cdot) = (s_l \circ \lambda)(\cdot).$$

The first justification comes from our result in Theorem 4.2.3.

Theorem 5.2.1. *The symmetric function $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable at the point $\lambda(A)$ if and only if $f \circ \lambda$ is twice differentiable at the point A.*

The second justification is from [88, Theorem 2.1].

Theorem 5.2.2. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a function analytic at the point $\lambda(A)$ for some A in S^n. Suppose also $f(Px) = f(x)$ for every permutation matrix, P, for which $P\lambda(A) = \lambda(A)$. Then the function $f \circ \lambda$ is analytic at A.*

For the third justification we use the standard algebraic fact that every symmetric polynomial in several variables can be written as a polynomial in the elementary symmetric functions. We also use the following result [3]. Until the end of this section only, $\lambda_i(X)$ will denote an arbitrary eigenvalue of a matrix X, not necessarily the i'th largest one.

Theorem 5.2.3 (Arnold 1971). *Suppose that the matrix $A \in \mathbb{C}^{n\times n}$ has q eigenvalues $\lambda_1(A), ..., \lambda_q(A)$ (counting multiplicities) in an open set $\Omega \subset \mathbb{C}$, and the other $n-q$ eigenvalues are not in Ω. Then for all matrices X in a neighbourhood of A there are holomorphic mappings $S : \mathbb{C}^{n\times n} \to \mathbb{C}^{q\times q}$ and $T : \mathbb{C}^{n\times n} \to \mathbb{C}^{(n-q)\times(n-q)}$ such that*

$$X \text{ is similar to } \begin{pmatrix} S(X) & 0 \\ 0 & T(X) \end{pmatrix},$$

and $S(A)$ has eigenvalues $\lambda_1(A), ..., \lambda_q(A)$.

Using Arnold's theorem we can prove that in fact the functions σ_{k_l} and S_l are holomorphic around A.

Theorem 5.2.4. *For every symmetric polynomial $p : \mathbb{C}^q \to \mathbb{C}$, the function $(p \circ \lambda)(S(X))$ is analytic around A.*

Proof. It suffices to prove the theorem in the case of an elementary symmetric polynomial, since any symmetric polynomial is a polynomial in the elementary symmetric functions (see for example [38, Proposition V.2.20.(ii)]). First we show that $(p \circ \lambda)(S(X))$ is holomorphic around A by using Arnold's theorem. By continuity of the eigenvalues, for every $i = 1, ..., n$ we can define functions

$\lambda_i : \mathcal{C}^{n\times n} \to \mathbb{C}$ such that for matrices X near A, if $\{\lambda_i(X)\}_{i=1}^n$ are the eigenvalues of X then $\{\lambda_i(X)\}_{i=1}^q$ are the eigenvalues of $S(X)$. So the elementary symmetric functions of $\lambda_1(X), ..., \lambda_q(X)$ are the coefficients of the characteristic polynomial $\det(\lambda I - S(X))$. Consequently they are holomorphic around A. Finally, we consider the case when A is a real symmetric matrix, and we restrict ourselves to a neighbourhood of real symmetric matrices around A. Because for matrices X in this neighbourhood the values of $(p \circ \lambda)(S(X))$ are real, one can easily see that the holomorphic expansion around A reduces to an analytic (real) expansion. □

5.3 Quadratic expansion of spectral functions

Our goal in this section is to prove the main result of the chapter. Not surprisingly the form of the Hessian is the same as the one given in Theorem 4.2.3.

Theorem 5.3.1 (Quadratic Expansion). *The symmetric function $f : \mathbb{R}^n \to \mathbb{R}$ has a strong quadratic expansion at the point $x = \lambda(Y)$ ($Y \in S^n$) if and only if $f \circ \lambda$ has a strong quadratic expansion at Y, and in that case*

$$\begin{aligned}
\nabla(f \circ \lambda)(Y)[H] &= \operatorname{tr}\left(\tilde{H}\operatorname{Diag}\nabla f(\mu)\right) \\
\nabla^2(f \circ \lambda)(Y)[H,H] &= \sum_{p,q=1}^{n} \tilde{h}_{pp} f''_{pq}(\mu)\tilde{h}_{qq} + \\
&\qquad \sum_{\substack{p\neq q \\ \mu_p=\mu_q}} b_p \tilde{h}_{pq}^2 + \sum_{p,q:\mu_p\neq\mu_q} \frac{f'_p(\mu) - f'_q(\mu)}{\mu_p - \mu_q}\tilde{h}_{pq}^2,
\end{aligned}$$

where $\mu = \lambda(Y)$, $\tilde{H} = U^T H U$, $Y = U(\operatorname{Diag}\mu)U^T$, U orthogonal, and the vector b is defined in Lemma 5.3.7. The analogous result holds for the weak quadratic

expansion.

We will only talk about strong quadratic expansions in this chapter: the development for the weak version is analogous. We need the following result from [87, Remark 6].

Lemma 5.3.2. *Every eigenvalue, $\lambda_m(Y)$, of a symmetric matrix, Y, has the following expansion in the direction of the symmetric matrix H:*

$$\begin{aligned}\lambda_m(Y+tH) = \lambda_m(Y) + t\lambda_{m-k_{l-1}}(X_l^T H X_l) \\ + \frac{t^2}{2}\lambda_{m-k_{l-1}-t_{l,j-1}}\left(2U_{l,j}^T X_l^T H(\lambda_m(Y)I - Y)^\dagger H X_l U_{l,j}\right) + O(t^3),\end{aligned} \tag{5.2}$$

where the meaning of X_l and $U_{l,j}$ is explained in the previous section.

Next we give a technical lemma that will allow us to cut down on the notation. We use Definition 4.1.1.

Lemma 5.3.3. *Let $\mu \in \mathbb{R}^n$ be such that*

$$\mu_1 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots \mu_{k_r}, \qquad (k_0 = 0,\ k_r = n),$$

and let the vector $b \in \mathbb{R}^n$ be block refined by μ. Let $H \in S^n$ be an arbitrary matrix and $X_i = [e^{k_{i-1}+1}, ..., e^{k_i}]$ for every $i = 1, ..., r$. Then we have the identities:

$$\langle H, b_{k_l}(\mu_{k_l} I - \operatorname{Diag}\mu)^\dagger H X_l X_l^T\rangle = \sum_{p=k_{l-1}+1}^{k_l} \sum_{\substack{q=1 \\ \mu_q \neq \mu_p}}^{n} \frac{b_p}{\mu_p - \mu_q} h_{pq}^2.$$

$$\left\langle H, \sum_{i=1}^{r} b_{k_i}(\mu_{k_i} I - \operatorname{Diag}\mu)^\dagger H X_i X_i^T \right\rangle = \sum_{p,q:\mu_p > \mu_q} \frac{b_p - b_q}{\mu_p - \mu_q} h_{pq}^2.$$

Proof. The product $X_i X_i^T$ is an $n \times n$ matrix with zero entries, except $(X_l X_l^T)^{p,p} = 1$ for $p = k_{i-1} + 1, ..., k_i$. Thus the columns of $H X_i X_i^T$ are zero vectors, except the columns with indexes $p = k_{i-1} + 1, ..., k_i$ which are equal to the corresponding columns of H. The matrix $b_{k_i}(\mu_{k_i} I - \operatorname{Diag} \mu)^\dagger$ is equal to

$$\operatorname{Diag}\left(\frac{b_{k_i}}{\mu_{k_i} - \mu_1}, ..., \frac{b_{k_i}}{\mu_{k_i} - \mu_{k_{i-1}}}, 0, ..., 0, \frac{b_{k_i}}{\mu_{k_i} - \mu_{k_i+1}}, ..., \frac{b_{k_i}}{\mu_{k_i} - \mu_{k_r}}\right).$$

Consequently we have

$$\begin{aligned}\langle H, b_{k_i}(\mu_{k_i} I - \operatorname{Diag} \mu)^\dagger H X_i X_i^T \rangle &= \sum_{p=k_{i-1}+1}^{k_i} \sum_{\substack{q=1 \\ \mu_q \neq \mu_p}}^{n} \frac{b_{k_i}}{\mu_{k_i} - \mu_q} h_{qp}^2 \\ &= \sum_{p=k_{i-1}+1}^{k_i} \left(\sum_{q=1}^{k_{i-1}} \frac{-b_p}{\mu_q - \mu_p} h_{qp}^2 + \sum_{q=k_i+1}^{n} \frac{b_p}{\mu_p - \mu_q} h_{qp}^2 \right),\end{aligned}$$

and the two identities can now be easily obtained. □

Our first goal is to find a formula for the Hessian of σ_{k_l}, $1 \leq l \leq r$. We denote the standard basis in $\mathbb{R}^n$ by e^1,e^2,...,e^n. As a byproduct in the following lemma we derive a formula for the derivative of the function σ_{k_l} at the point $\operatorname{Diag} \mu$. This formula appeared many times in the literature: see for example Corollary 3.10 in [32], or the proof of Corollary 3.3 in [48]. The expression for the Hessian is also known, see Formula (3.28) in [74] or [19] for a differential geometry argument, here we present yet another way of deriving it.

Lemma 5.3.4. *For a real vector $\mu \in \mathbb{R}^n$, such that*

$$\mu_1 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots \mu_{k_r}, \qquad (k_0 = 0,\ k_r = n),$$

the function

$$\sigma_{k_l}(\cdot) = \sum_{i=1}^{k_l} \lambda_i(\cdot)$$

is analytic at the matrix $\operatorname{Diag}\mu$ *with first and second derivatives satisfying*

$$\begin{aligned}
\nabla\sigma_{k_l}(\operatorname{Diag}\mu)[H] &= \operatorname{tr}\left(H \operatorname{Diag}\sum_{i=1}^{k_l} e^i\right) \\
\nabla^2\sigma_{k_l}(\operatorname{Diag}\mu)[H,H] &= 2\sum_{p=1}^{k_l}\sum_{q=k_l+1}^{n} \frac{h_{qp}^2}{\mu_p - \mu_q} \\
&= \operatorname{tr}\left(2H\sum_{i=1}^{l}(\mu_{k_i} I - \operatorname{Diag}\mu)^\dagger H X_i X_i^T\right),
\end{aligned}$$

where $X_i = [e^{k_{i-1}+1}, ..., e^{k_i}]$.

Proof. The fact that σ_{k_l} is analytic at the point $\operatorname{Diag}\mu$ follows from Section 5.2. Next, summing equations (5.2) with $Y = \operatorname{Diag}\mu$, for $m = 1, ..., k_l$ and using the fact that $X = I$ (so $X_i = [e^{k_{i-1}+1}, ..., e^{k_i}]$), we get

$$\begin{aligned}
\sigma_{k_l}(\operatorname{Diag}\mu + tH) &= \sum_{i=1}^{k_l} \lambda_i(\operatorname{Diag}\mu + tH) = \sigma_{k_l}(\operatorname{Diag}\mu) + t\sum_{i=1}^{l} \operatorname{tr}\left(X_i^T H X_i\right) \\
&\quad + \frac{t^2}{2}\sum_{i=1}^{l}\sum_{j=1}^{s_i}\sum_{v=1}^{t_{i,j}-t_{i,j-1}} \lambda_v\left(2U_{i,j}^T X_i^T H\left(\mu_{k_i} I - \operatorname{Diag}\mu\right)^\dagger H X_i U_{i,j}\right) + O(t^3) \\
&= \sigma_{k_l}(\operatorname{Diag}\mu) + t\langle \operatorname{Diag}\sum_{i=1}^{k_l} e^i, H\rangle
\end{aligned}$$

$$+\frac{t^2}{2}\sum_{i=1}^{l}\sum_{j=1}^{s_i}\operatorname{tr}\left(2U_{i,j}^T X_i^T H(\mu_{k_i}I-\operatorname{Diag}\mu)^\dagger HX_iU_{i,j}\right)+O(t^3).$$

We concentrate on the double sum above.

$$\begin{aligned}
&\sum_{i=1}^{l}\sum_{j=1}^{s_i}\operatorname{tr}\left(2U_{i,j}^T X_i^T H(\mu_{k_i}I-\operatorname{Diag}\mu)^\dagger HX_iU_{i,j}\right)= \\
&\qquad=\sum_{i=1}^{l}\sum_{j=1}^{s_i}\operatorname{tr}\left(2X_i^T H(\mu_{k_i}I-\operatorname{Diag}\mu)^\dagger HX_iU_{i,j}U_{i,j}^T\right) \\
&\qquad=\sum_{i=1}^{l}\operatorname{tr}\left(2X_i^T H(\mu_{k_i}I-\operatorname{Diag}\mu)^\dagger HX_i\sum_{j=1}^{s_i}U_{i,j}U_{i,j}^T\right) \\
&\qquad=\sum_{i=1}^{l}\operatorname{tr}\left(2X_i^T H(\mu_{k_i}I-\operatorname{Diag}\mu)^\dagger HX_i\right) \\
&\qquad=\operatorname{tr}\left(2H\sum_{i=1}^{l}(\mu_{k_i}I-\operatorname{Diag}\mu)^\dagger HX_iX_i^T\right) \\
&\qquad=\sum_{p=1}^{k_l}\sum_{\substack{q=1\\ \mu_q\neq\mu_p}}^{n}\frac{2}{\mu_p-\mu_q}h_{qp}^2 \\
&\qquad=2\sum_{p=1}^{k_l}\sum_{q=k_l+1}^{n}\frac{h_{qp}^2}{\mu_p-\mu_q}.
\end{aligned}$$

The next to the last equality follows from Lemma 5.3.3, with $b=(2,...,2)$, while the last equality after canceling all terms with opposite signs. By the uniqueness of the Hessian in the quadratic expansion of a function, we conclude that the last expression above must be indeed the Hessian. □

Note 5.3.5. *Notice that the Hessian above is a positive semidefinite quadratic form. This is not a surprise since a well known result of Fan [21] says that* σ_m

is a convex function for all $m = 1, ..., n$.

Lemma 5.3.6. *For a real vector* $\mu \in \mathbb{R}^n$, *such that*

$$\mu_1 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots \mu_{k_r} \qquad (k_0 = 0,\ k_r = n),$$

the function

$$S_l(\cdot) = \sum_{m=k_{l-1}+1}^{k_l} \lambda_m^2(\cdot)$$

is analytic at the matrix $\operatorname{Diag}\mu$, *with first and second derivatives satisfying*

$$\begin{aligned}
\nabla S_l(\operatorname{Diag}\mu)[H] &= 2\mu_{k_l} \operatorname{tr}\left(H \operatorname{Diag} \sum_{i=k_{l-1}+1}^{k_l} e^i\right) \\
\nabla^2 S_l(\operatorname{Diag}\mu)[H,H] &= 2 \sum_{p,q=k_{l-1}+1}^{k_l} h_{qp}^2 + 4 \sum_{p=k_{l-1}+1}^{k_l} \sum_{\substack{q=1 \\ \mu_p \neq \mu_q}}^{n} \frac{\mu_p}{\mu_p - \mu_q} h_{qp}^2 \\
&= \langle H, 2X_l X_l^T H X_l X_l^T + 4\mu_{k_l}(\mu_{k_l} I - \operatorname{Diag}\mu)^\dagger H X_l X_l^T \rangle,
\end{aligned}$$

where $X_l = [e^{k_{l-1}+1}, ..., e^{k_l}]$.

Proof. The analyticity of $S(\cdot)$ at the point $\operatorname{Diag}\mu$ follows from Section 5.2. Next, summing the squares of equations (5.2) with $Y = \operatorname{Diag}\mu$, for $m = 1, ..., k_l$ and using the fact that $X = I$ (so $X_i = [e^{k_{i-1}+1}, ..., e^{k_i}]$), we get

$$\begin{aligned}
\sum_{m=k_{l-1}+1}^{k_l} \lambda_m^2(\operatorname{Diag}\mu + tH) &= \sum_{m=k_{l-1}+1}^{k_l} \Bigg(\mu_{k_l} + t\lambda_{m-k_{l-1}}(X_l^T H X_l) \\
&\quad + \frac{t^2}{2} \lambda_{m-k_{l-1}-t_{l,j-1}}\left(2U_{l,j}^T X_l^T H (\mu_{k_l} I - \operatorname{Diag}\mu)^\dagger H X_l U_{l,j}\right) + O(t^3) \Bigg)^2 \\
&= (k_l - k_{l-1})\mu_{k_l}^2 + t^2 \sum_{m=k_{l-1}+1}^{k_l} \lambda_{m-k_{l-1}}^2(X_l^T H X_l)
\end{aligned}$$

$$+ 2t\mu_{k_l} \sum_{m=k_{l-1}+1}^{k_l} \lambda_{m-k_{l-1}}(X_l^T H X_l)$$
$$+ t^2 \mu_{k_l} \sum_{j=1}^{s_l} \sum_{v=1}^{t_{l,j}-t_{l,j-1}} \lambda_v\big(2U_{l,j}^T X_l^T H(\mu_{k_l} I - \operatorname{Diag}\mu)^\dagger H X_l U_{l,j}\big) + O(t^3).$$

We recall the fact that for every symmetric $n \times n$ matrix Q we have

$$\sum_{i=1}^{n} \lambda_i^2(Q) = \langle Q, Q\rangle.$$

We use this fact to evaluate the second summand in the formula above.

$$\sum_{m=k_{l-1}+1}^{k_l} \lambda_{m-k_{l-1}}^2(X_l^T H X_l) = \langle X_l^T H X_l, X_l^T H X_l\rangle = \langle H, X_l X_l^T H X_l X_l^T\rangle.$$

Observe as in Lemma 5.3.4 that for the fourth summand in the formula above we have

$$\sum_{j=1}^{s_l} \sum_{v=1}^{t_{l,j}-t_{l,j-1}} \lambda_v\big(2U_{l,j}^T X_l^T H(\mu_{k_l} I - \operatorname{Diag}\mu)^\dagger H X_l U_{l,j}\big)$$
$$= \sum_{j=1}^{s_l} \operatorname{tr}\big(2U_{l,j}^T X_l^T H(\mu_{k_l} I - \operatorname{Diag}\mu)^\dagger H X_l U_{l,j}\big)$$
$$= \operatorname{tr}\big(2X_l^T H(\mu_{k_l} I - \operatorname{Diag}\mu)^\dagger H X_l\big).$$

Substituting everything in the original formula we get

$$\sum_{m=k_{l-1}+1}^{k_l} \lambda_m^2(\operatorname{Diag}\mu + tH) = (k_l - k_{l-1})\mu_{k_l}^2 + t^2\langle H, X_l X_l^T H X_l X_l^T\rangle +$$

$$2t\mu_{k_l}\langle \operatorname{Diag} \sum_{i=k_{l-1}+1}^{k_l} e^i, H\rangle + t^2\mu_{k_l}\langle H, 2(\mu_{k_l} I - \operatorname{Diag}\mu)^\dagger H X_l X_l^T\rangle + O(t^3)$$

$$= (k_l - k_{l-1})\mu_{k_l}^2 + 2t\mu_{k_l}\langle \operatorname{Diag} \sum_{i=k_{l-1}+1}^{k_l} e^i, H\rangle \; +$$

$$\frac{t^2}{2}\langle H, 2X_l X_l^T H X_l X_l^T + 4\mu_{k_l}(\mu_{k_l} I - \operatorname{Diag}\mu)^\dagger H X_l X_l^T\rangle + O(t^3).$$

Using Lemma 5.3.3, with $b = 4\mu$, we conclude that

$$\nabla^2 S_l(\operatorname{Diag}\mu)[H,H] = 2\sum_{p,q=k_{l-1}+1}^{k_l} h_{qp}^2 + 4\sum_{p=k_{l-1}+1}^{k_l}\sum_{\substack{q=1\\ \mu_p\neq\mu_q}}^{n} \frac{\mu_p}{\mu_p - \mu_q} h_{qp}^2.$$

By the uniqueness of the Hessian in the quadratic expansion of a function, we conclude that the last expression above must be indeed the Hessian. □

The lemma below is a repetition of Lemma 4.1.2. The proof given there doesn't apply here because we cannot differentiate twice. That is why for completeness we repeat the whole bit.

Lemma 5.3.7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a symmetric function having quadratic expansion at the point $\mu \in \mathbb{R}^n_\downarrow$, where*

$$\mu_1 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots \mu_{k_r}, \qquad (k_0 = 0,\ k_r = n),$$

and let P be a permutation matrix such that $P\mu = P$. Then

1. *$\nabla f(\mu) = P^T \nabla f(\mu)$, and*
2. *$\nabla^2 f(\mu) = P^T \nabla^2 f(\mu) P$.*

In particular we can write

$$\nabla^2 f(\mu) = \begin{pmatrix} a_{11}E_{11} + b_{k_1} I_1 & a_{12}E_{12} & \cdots & a_{1r}E_{1r} \\ a_{21}E_{21} & a_{22}E_{22} + b_{k_2} I_2 & \cdots & a_{2r}E_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1}E_{r1} & a_{r2}E_{r2} & \cdots & a_{rr}R_{rr} + b_{k_r} I_r \end{pmatrix},$$

where each E_{uv} *is a* $(k_u - k_{u-1}) \times (k_v - k_{v-1})$ *matrix of all ones,* $(a_{ij})_{i,j=1}^{r,r}$ *is a real symmetric matrix,* $b := (b_1, ..., b_n)$ *is a real vector which is block refined by* μ*, and* I_u *is a square identity matrix of the same dimensions as* E_{uu}*. We also define the following matrix*

$$A := \nabla^2 f(\mu) - \operatorname{Diag} b = (a_{ij}E_{ij})_{i,j=1}^{r}.$$

Before we give the proof, some comments about the above representation are necessary.

1. We make the convention that if the i-th diagonal block in the above representation has dimensions 1×1 then we set $a_{ii} = 0$ and $b_{k_i} = f''_{k_i k_i}(\mu)$. Otherwise the value of b_{k_i} is uniquely determined as the difference between a diagonal and an off-diagonal element of this block.

2. Note that the matrix A and the vector b depend on the point around which we form the quadratic expansion (in this case μ) and on the function f.

Proof. We have

$$f(\mu + h) = f(\mu) + \langle \nabla f(\mu), h \rangle + \frac{1}{2}\langle h, \nabla^2 f(\mu) h \rangle + O(\|h\|^3).$$

Let P be a permutation matrix such that $P\mu = \mu$. Then

$$\begin{aligned} f(P(\mu + h)) &= f(\mu) + \langle \nabla f(\mu), Ph \rangle + \frac{1}{2}\langle Ph, \nabla^2 f(\mu) Ph \rangle + O(\|Ph\|^3) \\ &= f(\mu) + \langle P^T \nabla f(\mu), h \rangle + \frac{1}{2}\langle h, (P^T \nabla^2 f(\mu) P)h \rangle + O(\|h\|^3). \end{aligned}$$

Using the fact that f is symmetric gives us that $f(P(\mu+h)) = f(\mu+h)$ so $\nabla f(\mu) = P^T \nabla f(\mu)$. Subtracting the above two equalities we obtain

$$\nabla^2 f(\mu) = P^T \nabla^2 f(\mu) P, \quad \forall P \in P(n) \text{ s.t. } P\mu = \mu. \tag{5.3}$$

The claimed block structure of $\nabla^2 f(\mu)$ is now easy to check. □

Note 5.3.8. *Observe that equation (5.3) holds for arbitrary $\mu \in \mathbb{R}^n$.*

Lemma 5.3.9. *The vector μ block refines $\nabla^2 f(\mu)\mu$.*

Proof. Suppose $P\mu = \mu$. Then using twice Equation (5.3) and the above note, we get

$$P\nabla^2 f(\mu)\mu \;\; = P(P^T \nabla^2 f(\mu) P)\mu = \nabla^2 f(\mu) P\mu = \nabla^2 f(\mu)\mu. \qquad \square$$

Lemma 5.3.10. *Let $\mu \in \mathbb{R}^n_\downarrow$ be such that*

$$\mu_1 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots \mu_{k_r} \quad (k_0 = 0,\ k_r = n).$$

Suppose μ block-refines a vector $b \in \mathbb{R}^n$. Then $b^T\lambda$ is analytic at the matrix Diag μ

with quadratic expansion:

$$b^T\lambda(\mathrm{Diag}\,\mu + H) = b^T\mu + \langle \mathrm{Diag}\,b, H\rangle + \sum_{p,q:\mu_p>\mu_q} \frac{b_p - b_q}{\mu_p - \mu_q} h_{qp}^2 + O(\|H\|^3).$$

Proof. Because the vector μ block-refines the vector b there exist reals b'_1,b'_2,...,b'_r with

$$b_j = b'_i \text{ whenever } k_{i-1} + 1 \le j \le k_i,\ i = 1, 2, ..., r.$$

We obtain

$$b^T\lambda(\cdot) = \sum_{i=1}^{r} b'_i \sum_{j=k_{i-1}+1}^{k_i} \lambda_j(\cdot) = \sum_{i=1}^{r} b'_i\big(\sigma_{k_i}(\cdot) - \sigma_{k_{i-1}}(\cdot)\big).$$

Now applying Lemma 5.3.4 and Lemma 5.3.3 gives the result. □

Lemma 5.3.11. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a symmetric function having quadratic expansion at the point $\mu \in \mathbb{R}^n_\downarrow$, where*

$$\mu_1 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots \mu_{k_r} \quad (k_0 = 0,\ k_r = n).$$

Then the following matrix functions on S^n,

1. $F(\cdot) := \nabla f(\mu)^T\lambda(\cdot)$,
2. $H(\cdot) := \mu^T\nabla^2 f(\mu)\lambda(\cdot)$,
3. $G(\cdot) := \lambda(\cdot)^T\nabla^2 f(\mu)\lambda(\cdot)$,

have quadratic expansions at the matrix $\mathrm{Diag}\,\mu$.

Proof. Later we will need the formulae, giving the quadratic expansions of these functions, derived in the following proof. Notice that the first two parts follow immediately from the previous two lemmas. So we can write, up to $O(\|H\|^3)$,

$$F(\operatorname{Diag}\mu + H) \approx \nabla f(\mu)^T\mu + \langle \operatorname{Diag}\nabla f(\mu), H\rangle + \sum_{p,q:\mu_p>\mu_q} \frac{f'_p(\mu) - f'_q(\mu)}{\mu_p - \mu_q} h_{qp}^2,$$

$$H(\operatorname{Diag}\mu + H) \approx \mu^T\nabla^2 f(\mu)\mu + \langle \operatorname{Diag}\nabla^2 f(\mu)\mu, H\rangle + \sum_{p,q:\mu_p>\mu_q} \frac{(\mu^T\nabla^2 f(\mu))_p - (\mu^T\nabla^2 f(\mu))_q}{\mu_p - \mu_q} h_{qp}^2.$$

(iii) Because of the block structure of $\nabla^2 f(\mu)$ described in Lemma 5.3.7, we have

$$\lambda(\cdot)^T\nabla^2 f(\mu)\lambda(\cdot) = \sum_{i,j=1}^{r} a_{ij}(\sigma_{k_i}(\cdot) - \sigma_{k_{i-1}}(\cdot))(\sigma_{k_j}(\cdot) - \sigma_{k_{j-1}}(\cdot)) + \sum_{l=1}^{r} b_{k_l} S_l(\cdot),$$

where the matrix $(a_{ij})_{i,j=1}^r$, vector b, and $S_l(\cdot)$ are defined in Lemma 5.3.7 and Lemma 5.3.6. Now by Lemma 5.3.4

$$\begin{aligned}\sigma_{k_l}(\operatorname{Diag}\mu + H) - \sigma_{k_{l-1}}(\operatorname{Diag}\mu + H) &= \sum_{i=k_{l-1}+1}^{k_l} \mu_i + \langle \operatorname{Diag} \sum_{i=k_{l-1}+1}^{k_l} e^i, H\rangle \\ &\quad + \frac{1}{2}\langle H, 2(\mu_{k_l} I - \operatorname{Diag}\mu)^\dagger H X_l X_l^T\rangle + O(\|H\|^3) \\ &= \sum_{i=k_{l-1}+1}^{k_l} \mu_i + \sum_{i=k_{l-1}+1}^{k_l} h_{ii} + \sum_{i=k_{l-1}+1}^{k_l} \langle H, (\mu_{k_l} I - \operatorname{Diag}\mu)^\dagger H e^i (e^i)^T\rangle + O(\|H\|^3).\end{aligned}$$

We can evaluate the first summand in the above representation of the function $G(\cdot)$.

$$\begin{aligned}
&\sum_{i,j=1}^{r} a_{ij}(\sigma_{k_i}(\operatorname{Diag}\mu + H) - \sigma_{k_{i-1}}(\operatorname{Diag}\mu + H)) \\
&\qquad\qquad \times (\sigma_{k_j}(\operatorname{Diag}\mu + H) - \sigma_{k_{j-1}}(\operatorname{Diag}\mu + H)) \\
&= \mu^T A\mu + (\operatorname{diag} H)^T A(\operatorname{diag} H) + 2\mu^T A(\operatorname{diag} H) \\
&\qquad + 2\langle H, \sum_{i,j=1}^{n} \mu_i A^{i,j}(\mu_j I - \operatorname{Diag}\mu)^\dagger H e^j (e^j)^T\rangle + O(\|H\|^3) \\
&= \mu^T A\mu + 2\langle \operatorname{Diag} A\mu, H\rangle + \langle H, \operatorname{Diag} A(\operatorname{diag} H)\rangle \\
&\qquad + 2\langle H, \sum_{i,j=1}^{n} \mu_i A^{i,j}(\mu_j I - \operatorname{Diag}\mu)^\dagger H e^j (e^j)^T\rangle + O(\|H\|^3),
\end{aligned}$$

where $\operatorname{diag}\colon S^n \to \mathbb{R}^n$ defined by $\operatorname{diag}(H) = (h_{11}, ..., h_{nn})$ is the conjugate operator of $\operatorname{Diag}\colon \mathbb{R}^n \to S^n$. On the other hand Lemma 5.3.6 gives us:

$$\begin{aligned}
&\sum_{l=1}^{r} b_{k_l} S_l(\operatorname{Diag}\mu + H) = \sum_{l=1}^{r} b_{k_l}\Bigg((k_l - k_{l-1})\mu_{k_l}^2 + 2\mu_{k_l}\langle \operatorname{Diag} \sum_{i=k_{l-1}+1}^{k_l} e^i, H\rangle \\
&\qquad + \langle H, X_l X_l^T H X_l X_l^T + 2\mu_{k_l}(\mu_{k_l} I - \operatorname{Diag}\mu)^\dagger H X_l X_l^T\rangle \Bigg) + O(\|H\|^3) \\
&= \mu^T(\operatorname{Diag} b)\mu + 2\langle \operatorname{Diag}(\operatorname{Diag} b)\mu, H\rangle + \langle H, \sum_{l=1}^{r} b_{k_l} X_l X_l^T H X_l X_l^T\rangle \\
&\qquad + \langle H, 2\sum_{i,j=1}^{n} \mu_i (\operatorname{Diag} b)^{i,j}(\mu_j I - \operatorname{Diag}\mu)^\dagger H e^j (e^j)^T\rangle + O(\|H\|^3).
\end{aligned}$$

Adding these two formulae together we finally get:

$$\begin{aligned}
\lambda(\operatorname{Diag}\mu + H)^T \nabla^2 f(\mu)\lambda(\operatorname{Diag}\mu + H) &= \mu^T \nabla^2 f(\mu)\mu + 2\langle \operatorname{Diag}\nabla^2 f(\mu)\mu, H\rangle \\
&+ \langle H, \operatorname{Diag} A(\operatorname{diag} H)\rangle + \langle H, \sum_{l=1}^{r} b_{k_l} X_l X_l^T H X_l X_l^T\rangle
\end{aligned}$$

$$
\begin{aligned}
&\quad + \langle H, 2\sum_{j=1}^{n} \left(\mu^T \nabla^2 f(\mu)\right)_j (\mu_j I - \operatorname{Diag}\mu)^\dagger H e^j (e^j)^T \rangle + O(\|H\|^3) \\
&= \mu^T \nabla^2 f(\mu)\mu + 2\langle \operatorname{Diag}\nabla^2 f(\mu)\mu, H\rangle + \langle H, \operatorname{Diag} A(\operatorname{diag} H)\rangle \\
&\quad + \sum_{p,q:\mu_p=\mu_q} b_p h_{pq}^2 + 2\sum_{p,q:\mu_p>\mu_q} \frac{(\mu^T\nabla^2 f(\mu))_p - (\mu^T\nabla^2 f(\mu))_q}{\mu_p - \mu_q} h_{qp}^2 + O(\|H\|^3).
\end{aligned}
$$

In the last equality we used Lemma 5.3.9 and Lemma 5.3.3. □

Now we are ready to prove a preliminary case of Theorem 5.3.1, namely, that it holds at $X = \operatorname{Diag}\mu$, $(\mu \in \mathbb{R}^n_\downarrow)$ and to give a formula for the Hessian of $f \circ \lambda$ at that point. The results for the gradient of $f \circ \lambda$ that we will obtain along the way were first obtained in [49].

Theorem 5.3.12. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a symmetric function having quadratic expansion at the point $\mu \in \mathbb{R}^n_\downarrow$, where*

$$\mu_1 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots \mu_{k_r} \quad (k_0 = 0,\ k_r = n).$$

Then $f \circ \lambda$ has quadratic expansion at the point $\operatorname{Diag}\mu$, with

$$
\begin{aligned}
\nabla(f\circ\lambda)(\operatorname{Diag}\mu)[H] &= \operatorname{tr}\left(H \operatorname{Diag}\nabla f(\mu)\right) \\
\nabla^2(f\circ\lambda)(\operatorname{Diag}\mu)[H,H] &= \sum_{p,q=1}^{n} h_{pp} f''_{pq}(\mu) h_{qq} \\
&\quad + \sum_{\substack{p\neq q \\ \mu_p=\mu_q}} b_p h_{pq}^2 + \sum_{p,q:\mu_p\neq\mu_q} \frac{f'_p(\mu) - f'_q(\mu)}{\mu_p - \mu_q} h_{pq}^2
\end{aligned}
$$

(with b defined by Lemma 5.3.7).

Note 5.3.13. *Corollary 5.3.14 will show that the requirement that $\mu \in \mathbb{R}^n_\downarrow$ can be omitted. For a matrix representation of the above formula combine equation (5.4) below, and the first identity in Lemma 5.3.3.*

Proof. We are given that

$$f(x) = f(\mu) + \nabla f(\mu)^T(x-\mu) + \frac{1}{2}(x-\mu)^T\nabla^2 f(\mu)(x-\mu) + O(\|x-\mu\|^3),$$

so after letting $x = \lambda(\operatorname{Diag}\mu + H)$ and using the fact that

$$\lambda(\operatorname{Diag}\mu + H) = \lambda(\operatorname{Diag}\mu) + O(\|H\|)$$

we get

$$\begin{aligned}(f\circ\lambda)(\operatorname{Diag}\mu + H) &= f(\mu) + \nabla f(\mu)^T\lambda(\operatorname{Diag}\mu + H) - \nabla f(\mu)^T\mu \\ &+ \frac{1}{2}\lambda(\operatorname{Diag}\mu + H)^T\nabla^2 f(\mu)\lambda(\operatorname{Diag}\mu + H) - \mu^T\nabla^2 f(\mu)\lambda(\operatorname{Diag}\mu + H) \\ &+ \frac{1}{2}\mu^T\nabla^2 f(\mu)\mu + O(\|H\|^3).\end{aligned}$$

Substituting the three expressions in the proof of Lemma 5.3.11 we obtain

$$\begin{aligned}(f\circ\lambda)(\operatorname{Diag}\mu + H) &= (f\circ\lambda)(\operatorname{Diag}\mu) + \langle \operatorname{Diag}\nabla f(\mu), H\rangle \\ &+ \frac{1}{2}\langle H, \operatorname{Diag} A(\operatorname{diag} H) + \sum_{l=1}^{r} b_{k_l} X_l X_l^T H X_l X_l^T\rangle \qquad (5.4) \\ &+ \sum_{p,q:\mu_p>\mu_q} \frac{f_p'(\mu) - f_q'(\mu)}{\mu_p - \mu_q} h_{qp}^2 + O(\|H\|^3).\end{aligned}$$

Recall that $X_l = [e^{k_{l-1}+1}, ..., e^{k_l}]$. In order to obtain the representation given in the

theorem one has to use the definition of A and $b = (b_1, ..., b_n)$ given in Lemma 5.3.7 and the note that follows it. □

Proof of Theorem 5.3.1. Suppose f has quadratic expansion at the point $\lambda(Y)$, and choose any orthogonal matrix $U = [u^1 \dots u^n]$ that gives the ordered spectral decomposition of Y, $Y = U\big(\operatorname{Diag}\lambda(Y)\big)U^T$. Here we actually have $A = A(\lambda(Y))$ and $b_i = b_i(\lambda(Y))$. While in formula (5.4) we had $A = A(\mu)$ and $b_i = b_i(\mu)$, to make the formulae here easier to read we will write again simply A and b_i. Then we have, using Formula (5.4) and some easy manipulations,

$$\begin{aligned}(f\circ\lambda)(Y+H) &= (f\circ\lambda)(\operatorname{Diag}\lambda(Y) + U^THU)\\ &= (f\circ\lambda)(Y) + \langle \operatorname{Diag}\nabla f(\lambda(Y)), U^THU\rangle\\ &+ \frac{1}{2}\langle U^THU, \operatorname{Diag} A(\operatorname{diag} U^THU) + \sum_{l=1}^{r} b_{k_l} X_l X_l^T U^T HU X_l X_l^T\rangle\\ &+ \sum_{\substack{p,q\\ \lambda_p(Y)>\lambda_q(Y)}} \frac{f_p'(\lambda(Y)) - f_q'(\lambda(Y))}{\lambda_p(Y) - \lambda_q(Y)}((U^THU)^{qp})^2 + O(\|H\|^3),\end{aligned}$$

where $X_l = [e^{k_{l-1}+1}, ..., e^{k_l}]$. □

Corollary 5.3.14. *Theorem 5.3.12 holds for arbitrary $\mu \in \mathbb{R}^n$, where*

$$b(\mu) := Pb(\bar{\mu}), \tag{5.5}$$

and P is a permutation matrix, such that $P^T\mu = \bar{\mu}$.

Proof. Pick a permutation matrix P such that $P^T\mu = \bar{\mu}$ and let π be the permutation associated with it, that is $\bar{\mu} = (\mu_{\pi(1)}, ..., \mu_{\pi(n)})$, or in other words $Pe^i = e^{\pi(i)}$.

We have that f has quadratic expansion at the point μ, that is

$$f(\mu + h) = f(\mu) + \langle \nabla f(\mu), h \rangle + \frac{1}{2}\langle h, \nabla^2 f(\mu) h \rangle + O(\|h\|^3).$$

Using the fact that f is symmetric we obtain

$$\begin{aligned} f(\bar{\mu} + P^T h) &= f(P^T(\mu + h)) = f(\mu + h) \\ &= f(\mu) + \langle \nabla f(\mu), h \rangle + \frac{1}{2}\langle h, \nabla^2 f(\mu) h \rangle + O(\|h\|^3) \\ &= f(\bar{\mu}) + \langle P^T \nabla f(\mu), P^T h \rangle + \frac{1}{2}\langle P^T h, P^T \nabla^2 f(\mu) P P^T h \rangle + O(\|P^T h\|^3). \end{aligned}$$

So f has quadratic expansion at the point $\bar{\mu}$ as well, and we have the relationships:

$$\begin{aligned} \nabla f(\bar{\mu}) &= P^T \nabla f(\mu) \\ \nabla^2 f(\bar{\mu}) &= P^T \nabla^2 f(\mu) P. \end{aligned} \tag{5.6}$$

We have $\operatorname{Diag}\mu = P(\operatorname{Diag}\bar{\mu})P^T$. Applying Theorem 5.3.1 with $Y = \operatorname{Diag}\mu$ and $U = P$, and using Equations (5.6) and (5.5) we get

$$\begin{aligned} \nabla^2(f \circ \lambda)(\operatorname{Diag}\mu)[H, H] &= \sum_{p,q=1}^{n} h_{\pi(p)\pi(p)} f''_{pq}(\bar{\mu}) h_{\pi(q)\pi(q)} \\ &\quad + \sum_{\substack{p \neq q \\ \bar{\mu}_p = \bar{\mu}_q}} b_p(\bar{\mu}) h^2_{\pi(p)\pi(q)} + \sum_{\bar{\mu}_p \neq \bar{\mu}_q} \frac{f'_p(\bar{\mu}) - f'_q(\bar{\mu})}{\bar{\mu}_p - \bar{\mu}_q} h^2_{\pi(p)\pi(q)} \\ &= \sum_{p,q=1}^{n} h_{pp} f''_{pq}(\mu) h_{qq} + \sum_{\substack{p \neq q \\ \mu_p = \mu_q}} b_p(\mu) h^2_{pq} + \sum_{\mu_p \neq \mu_q} \frac{f'_p(\mu) - f'_q(\mu)}{\mu_p - \mu_q} h^2_{pq}. \end{aligned}$$

The invariance of the formula for the gradient is shown in a similar manner. See also [49]. □

5.4 Strongly convex functions

As we mentioned in the introduction, a symmetric function f is convex if and only if $f \circ \lambda$ is convex. The analogous result also holds for essential strict convexity [48, Corollary 3.5]. Here we study yet a stronger property. Specificly, in this section we show that if a symmetric, convex function f has a quadratic expansion at the point $x = \lambda(Y)$ then the symmetric matrix $\nabla^2 f(x)$ is positive *definite*, if and only if the same is true for the bilinear operator $\nabla^2(f \circ \lambda)(Y)$.

Lemma 5.4.1. *If a function $f : \mathbb{R}^n \to \mathbb{R}$ is symmetric, strictly convex, and differentiable at the point μ*

$$\mu_1 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots \mu_{k_r}, \qquad (k_r = n).$$

then its gradient satisfies

$$\frac{f'_p(\mu) - f'_q(\mu)}{\mu_p - \mu_q} > 0 \qquad \text{for all } p, q \text{ such that } \mu_p \neq \mu_q.$$

Proof. Because f is strictly convex and differentiable at μ, for every $x \in \mathbb{R}^n$ $(\mu \neq x)$ we have that (see for example [76, Theorem 2.3.5])

$$\langle \nabla f(\mu), x - \mu \rangle < f(x) - f(\mu).$$

Suppose $\mu_p \neq \mu_q$. Let P be the permutation matrix that transposes p and q only. Then we have

$$(f'_q(\mu) - f'_p(\mu))(\mu_p - \mu_q) = \langle \nabla f(\mu), P\mu - \mu \rangle \quad < \quad f(P\mu) - f(\mu) = 0. \qquad \square$$

Lemma 5.4.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a symmetric function having quadratic expansion at μ, where*

$$\mu_1 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots \mu_{k_r}, \qquad (k_r = n).$$

If the Hessian $\nabla^2 f(\mu)$ is positive definite then the vector $b = (b_1, ..., b_n)$, defined in Lemma 5.3.7, has strictly positive entries.

Proof. It is well known that every principal minor in a positive definite matrix is positive definite. Fix an index $1 \leq i \leq n$. If $\mu_{i-1} > \mu_i > \mu_{i+1}$ then from the representation of the matrix $\nabla^2 f(\mu)$ in Lemma 5.3.7 and the note after it, it is clear that $b_i > 0$. Suppose now that i is in a block of length at least 2. Then some principal minor of $\nabla^2 f(\mu)$ of the form

$$\begin{pmatrix} a + b_i & a \\ a & a + b_i \end{pmatrix}$$

is positive definite, and the result follows. $\square$

Theorem 5.4.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a symmetric, strictly convex function having*

quadratic expansion at μ

$$\mu_1 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots \mu_{k_r}, \qquad (k_r = n).$$

Then the symmetric matrix $\nabla^2 f(\mu)$ *is positive definite if and only if the bilinear operator* $\nabla^2(f \circ \lambda)(Diag\,\mu)$ *is positive definite.*

Note 5.4.4. *In fact by Alexandrov's Theorem, if a function is convex it has quadratic expansion at almost every point of its domain [1]. For a proof of Alexandrov's Theorem in English see [20, Theorem 1, Section 6.4].*

Proof. Suppose first that the symmetric matrix $\nabla^2 f(\mu)$ is positive definite. Take a symmetric matrix $H \neq 0$. Then we have

$$\sum_{p,q=1}^{n} h_{pp} f''_{pq}(\mu) h_{qq} \geq 0,$$

because $\nabla^2 f(\mu)$ is positive definite,

$$2 \sum_{l=1}^{r} b_{k_l} \sum_{k_{i-1} < p < q \leq k_i} h_{pq}^2 \geq 0,$$

follows from Lemma 5.4.2, and

$$2 \sum_{p,q:\mu_p > \mu_q} \frac{f'_p(\mu) - f'_q(\mu)}{\mu_p - \mu_q} h_{pq}^2 \geq 0,$$

which follows from Lemma 5.4.1. Now because $H \neq 0$ at least one of the above inequalities will be strict.

In the other direction the argument is easy: take $H = \operatorname{Diag} x$, for $0 \neq x \in \mathbb{R}^n$ in the formula for $\nabla^2(f \circ \lambda)(\operatorname{Diag} \mu)$ given in Theorem 5.3.12 to get immediately $x^T \nabla^2 f(\mu) x > 0$. □

Theorem 5.4.5. *If $f : \mathbb{R}^n \to \mathbb{R}$ is a symmetric, strictly convex function having quadratic expansion at the point Y, then $\nabla^2 f(\lambda(Y))$ is positive definite if and only if $\nabla^2(f \circ \lambda)(Y)$ is.*

Proof. The proof of this theorem is now clear since $\nabla^2(f \circ \lambda)(Y)$ is positive definite if and only if $\nabla^2(f \circ \lambda)(\operatorname{Diag} \lambda(Y))$ is. □

5.5 Examples

Notice that examples analogous to those below can also be addressed using the theory in the previous chapter.

Example 5.5.1. Let g be a function on a scalar argument. Consider the following *separable* symmetric function with its corresponding spectral function:

$$f(x_1, ..., x_n) = \sum_{i=1}^{n} g(x_i)$$
$$(f \circ \lambda)(Y) = \sum_{i=1}^{n} g(\lambda_i(Y)).$$

Then if g has quadratic expansion at the points $x_1, ..., x_n$ so does f at $x = (x_1, ..., x_n)$ and we have

$$\nabla f(x) = (g'(x_1), ..., g'(x_n))^T,$$

$$\nabla^2 f(x) = \mathrm{Diag}\,(g''(x_1), ..., g''(x_n)),$$
$$b(x) = (g''(x_1), ..., g''(x_n))^T.$$

Suppose g has quadratic expansion at each entry of the vector $\mu \in \mathbb{R}^n_\downarrow$ that satisfies

$$\mu_1 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots \mu_{k_r}, \qquad (k_r = n).$$

Then Theorem 5.3.12 says that

$$\begin{aligned}\nabla^2(f \circ \lambda)(\mathrm{Diag}\,\mu)[H,H] &= \sum_{p=1}^{n} g''(\mu_p)h_{pp}^2 + \sum_{\substack{p\neq q\\ \mu_p=\mu_q}} g''(\mu_p)h_{pq}^2 \\ &\qquad + \sum_{p,q:\mu_p\neq\mu_q} \frac{g'(\mu_p)-g'(\mu_q)}{\mu_p-\mu_q}h_{pq}^2 \\ &= \sum_{p,q:\mu_p=\mu_q} g''(\mu_p)h_{pq}^2 + \sum_{p,q:\mu_p\neq\mu_q} \frac{g'(\mu_p)-g'(\mu_q)}{\mu_p-\mu_q}h_{pq}^2.\end{aligned}$$

Let us define the following notation consistent with [9, Section V.3]. For any differentiable function h defined on a subset of $\mathbb{R}$ define the 'divided difference'

$$h^{[1]}(\alpha,\beta) = \begin{cases} \frac{h(\alpha)-h(\beta)}{\alpha-\beta}, & \text{if } \alpha \neq \beta \\ h'(\alpha), & \text{if } \alpha = \beta. \end{cases}$$

If Λ is a diagonal matrix with diagonal entries α_1,...,α_n, we denote by $h^{[1]}(\Lambda)$ the $n \times n$ symmetric matrix matrix whose (i,j)-entry is $h^{[1]}(\alpha_i, \alpha_j)$.

Using this notation, for the function $h = g'$, we clearly have

$$\nabla^2(f \circ \lambda)(\mathrm{Diag}\,\mu)[H,H] = \langle H, h^{[1]}(\mathrm{Diag}\,\mu) \circ H\rangle,$$

$$\nabla^2(f \circ \lambda)(Y)[H,H] = \langle U^T H U, h^{[1]}(\operatorname{Diag}\lambda(Y)) \circ (U^T H U)\rangle, \tag{5.7}$$

where $Y = U(\operatorname{Diag}\lambda(Y))U^T$, U orthogonal, and $X \circ Y = (x_{ij}y_{ij})_{i,j=1}^n$ is the Hadamard product of matrices X and Y.

Let us extend the domain of the function h to include a subset of the symmetric matrices in the following way. If $\Lambda = \operatorname{Diag}(\alpha_1, ..., \alpha_n)$ is a diagonal matrix whose entries are in the domain of h, we define $h(\Lambda) = \operatorname{Diag}(h(\alpha_1), ..., h(\alpha_n))$. If Y is a symmetric matrix with eigenvalues $\alpha_1, ..., \alpha_n$ in the domain of h, we choose an orthogonal matrix U such that $Y = U\Lambda U^T$ and define $h(Y) = Uh(\Lambda)U^T$. (Notice that the definition of $h(Y)$ doesn't depend on the choice of the orthogonal matrix U.) In this way we can define $h(Y)$ for all symmetric matrices with eigenvalues in the domain of h. Then the formulae for the gradient in Theorem 5.3.1 says that for $h = g'$ we have

$$\nabla(f \circ \lambda)(Y) = h(Y).$$

Thus Equations (5.7) are just the formulae for the derivative ∇h given in Theorem V.3.3 in [9].

Example 5.5.2. Now we specialize the above example even more. The following spectral function finds many applications in semidefinite programming. Consider the symmetric and strictly convex function and corresponding spectral function:

$$f : x \in \mathbb{R}^n_{++} \mapsto -\sum_{i=1}^n \log x_i,$$

$$f \circ \lambda : A \in S^n_{++} \mapsto -\ln \mathrm{Det}\,(A).$$

(Here S^n_{++} denotes the set of all positive definite symmetric matrices.) Then Theorem 5.3.12 says that for $\mu \in \mathbb{R}^n_\downarrow$ such that

$$\mu_1 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots \mu_{k_r}, \qquad (k_r = n),$$

we have

$$\begin{aligned}
\nabla^2(f \circ \lambda)(\mathrm{Diag}\,\mu)[H,H] &= \sum_{p=1}^{n} \frac{h_{pp}^2}{\mu_p^2} + \sum_{\substack{p \neq q \\ \mu_p = \mu_q}} \frac{h_{pq}^2}{\mu_p^2} + \sum_{p,q:\mu_p \neq \mu_q} \frac{h_{pq}^2}{\mu_p \mu_q} \\
&= \sum_{p,q=1,1}^{n,n} \frac{h_{pq}^2}{\mu_p \mu_q} \\
&= \mathrm{tr}\,((\mathrm{Diag}\,\mu)^{-1} H (\mathrm{Diag}\,\mu)^{-1} H).
\end{aligned}$$

The last equality may easily be verified. In general, for an arbitrary symmetric matrix A, we get

$$\nabla^2(f \circ \lambda)(A)[H,H] = \mathrm{tr}\,(A^{-1} H A^{-1} H).$$

This agrees with the standard formula for the second derivative of the function $-\ln \mathrm{Det}\,(A)$. (See for example [68, Proposition 5.4.5].) Moreover the result in Section 5.4 tells us that

$$A \succ 0 \;\text{ implies }\; \mathrm{tr}\,(A^{-1} H A^{-1} H) > 0 \;\text{ for all }\; 0 \neq H \in S^n,$$

and this result is trivial since $\mathrm{tr}\,(A^{-1} H A^{-1} H) = \mathrm{tr}\,\big((A^{-1/2} H A^{-1/2})^2\big)$. ($A \succ 0$

means that the matrix A is positive definite.)

The reader can refer to Section 4.4 for more examples.

5.6 The Eigenvalues of $\nabla^2(f \circ \lambda)$

A natural question one may ask is: Is there any relationship between the eigenvalues of $\nabla^2 f(\lambda(Y))$ and those of $\nabla^2(f \circ \lambda)(Y)$? This section shows that locally such a relationship will be quite weak, although more globally they are closely related. Let Y be a symmetric matrix such that

$$\lambda_1(Y) = \cdots = \lambda_{k_1}(Y) > \cdots > \lambda_{k_{l-1}+1}(Y) = \cdots = \lambda_m(Y) = \cdots = \lambda_{k_l}(Y)$$
$$> \cdots \lambda_{k_r}(Y), \quad (k_0 = 0,\ k_r = n).$$

Using the representation given in Theorem 5.3.1 and Corollary 5.3.14 one can easily see that the $\frac{n(n+1)}{2}$ eigenvalues of $\nabla^2(f \circ \lambda)(Y)$ are

- $\{\lambda_i\big(\nabla^2 f(\lambda(Y))\big) | i = 1, ..., n\}$. (These are just the eigenvalues of $\nabla^2 f(\lambda(Y))$ with the same multiplicities.)
- b_{k_l} is an eigenvalue for every $l = 1, ..., r$ with multiplicity $(k_l - k_{l-1})(k_l - k_{l-1} - 1)/2$.
- $\frac{f'_{k_l}(\lambda(Y)) - f'_{k_s}(\lambda(Y))}{\lambda_{k_l}(Y) - \lambda_{k_s}(Y)}$ is an eigenvalue with multiplicity $(k_l - k_{l-1})(k_s - k_{s-1})$ for every ordered pair $(\lambda_{k_l}(Y), \lambda_{k_s}(Y))$ such that $\lambda_{k_l}(Y) > \lambda_{k_s}(Y)$.

So we can immediately conclude that

$$\begin{aligned} \lambda_{\max}\big(\nabla^2(f \circ \lambda)(Y)\big) &\geq \lambda_{\max}\big(\nabla^2 f(\lambda(Y))\big) \\ \lambda_{\min}\big(\nabla^2 f(\lambda(Y))\big) &\geq \lambda_{\min}\big(\nabla^2(f \circ \lambda)(Y)\big). \end{aligned} \tag{5.8}$$

We are going to show now that the above inequalities may be strict.

Example 5.6.1. Consider the convex function

$$f(x,y) := \frac{x^2+y^2}{4} + \frac{\cos 2x + \cos 2y}{8},$$

and the point

$$\mu = (2\pi, \pi) \in \mathbb{R}^2_{\downarrow}.$$

Then

$$\nabla f(x,y) = \begin{pmatrix} \frac{x}{2} - \frac{\sin 2x}{4} \\ \frac{y}{2} - \frac{\sin 2y}{4} \end{pmatrix}, \quad \nabla^2 f(x,y) = \begin{pmatrix} \sin^2 x & 0 \\ 0 & \sin^2 y \end{pmatrix}.$$

Using the representation in Theorem 5.3.12 we get

$$\nabla^2 f(\mu) = 0, \quad \nabla^2(f \circ \lambda)(\operatorname{Diag}\mu)[H,H] = h_{12}^2,$$

where

$$H = \begin{pmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{pmatrix}.$$

Then clearly

$$\lambda_{\max}\big(\nabla^2(f \circ \lambda)(\operatorname{Diag}\mu)\big) = 1 > \lambda_{\max}\big(\nabla^2 f(\mu)\big) = 0.$$

In order to demonstrate a strict inequality between the smallest eigenvalues one needs to consider the function $-f(x, y)$ at the same point μ.

Even though we may not have equalities in (5.8) at a particular matrix Y, if we consider the eigenvalues of $\nabla^2 f(\lambda(Y))$ and $\nabla^2(f \circ \lambda)(Y)$ as Y varies over an orthogonally invariant (see below) convex set, we can see that they vary within the same bounds. More precisely we have the following theorem. To make its proof precise, we need the main result from the previous chapter and [48] saying that: A symmetric function f is $\mathcal{C}^2$ if and only if $f \circ \lambda$ is, and f is convex if and only if $f \circ \lambda$ is.

Theorem 5.6.2. *Let C be a convex and symmetric subset of $\mathbb{R}^n$, and let $f : C \to \mathbb{R}$ be a symmetric, $\mathcal{C}^2$ function. Then*

$$\min_{y \in C} \lambda_{\min}(\nabla^2 f(y)) = \min_{Y \in \lambda^{-1}(C)} \lambda_{\min}(\nabla^2(f \circ \lambda)(Y)). \tag{5.9}$$

Proof. The following implications are easy to see.

$$\begin{aligned} \lambda_{\min}(\nabla^2 f(y)) \geq \alpha, \ \ \forall y \in C & \\ \Leftrightarrow\ & f - \frac{\alpha}{2}\|\cdot\|^2 \quad \text{convex} \\ \Leftrightarrow\ & \left(f - \frac{\alpha}{2}\|\cdot\|^2\right) \circ \lambda \quad \text{convex} \end{aligned}$$

$$\Leftrightarrow f \circ \lambda - \frac{\alpha}{2}\|\cdot\|_2^2 \quad \text{convex}$$

$$\Leftrightarrow \lambda_{\min}(\nabla^2(f \circ \lambda)(Y)) \geq \alpha, \ \ \forall Y \in C.$$

□

Remark 5.6.3. *If we multiply both sides of Equation (5.9) by* -1 *we will get*

$$\max_{y \in C} \lambda_{\max}(\nabla^2 f(y)) = \max_{Y \in \lambda^{-1}(C)} \lambda_{\max}(\nabla^2(f \circ \lambda)(Y)).$$

Chapter 6

Nonsmooth analysis of singular values

The singular values of a rectangular matrix have many properties analogous to the eigenvalues of a square matrix. In this chapter we are interested in the first order 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. That is why in order to gain insight into their behaviour we need to use the tools of the nonsmooth variational analysis [79].

We give formulae for the approximate subdifferential, Clarke subdifferential (in both cases when the underlying function is Lipschitz or just lower semicontinuous), horizon subdifferential, regular subdifferential, and proximal subdifferential of functions of singular values. We also give several applications of the developed theory. We compute the subdifferentials of σ_k - the k-th largest singular value of a matrix. Finally, we show how Lidskii's theorem for singular values follows easily from the

nonsmooth theory.

We follow the terminology and notation of [79], and the whole chapter closely follows the analogous development for eigenvalues in [52]. There are obvious parallels between the notation, techniques, and results there and here which suggest that there is a general theoretic framework that encompasses them both. (See Chapter 7 for another class of functions that may be part of the general theoretic framework.)

For convenience we state the singular value decomposition theorem. (For details and more results, see [36, Chapter 3].)

Theorem 6.0.4 (Singular Value Decomposition). *Let $A \in M_{n,m}(\mathbb{C})$ be given and $q = \min\{n, m\}$. There is a matrix $\Sigma = (\sigma_{ij}) \in M_{n,m}(\mathbb{R})$ with $\sigma_{ij} = 0$ for all $i \neq j$, and $\sigma_{11} \geq \sigma_{22} \geq \ldots \sigma_{qq} \geq 0$, and two unitary matrices $V \in O(n)$ and $W \in O(m)$ such that $A = V\Sigma W^*$. If $A \in M_{n,m}(\mathbb{R})$, then V and W may be taken to be real orthogonal.*

The numbers $\sigma_{11} \geq \sigma_{22} \geq \ldots \sigma_{qq} \geq 0$ are unique for the matrix A and are called *singular values of A*.

In this chapter we consider only real matrices. There are completely analogous results for complex matrices.

6.1 The approximate subdifferential

This section gives the relevant background of nonsmooth analysis.

Definition 6.1.1 (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 a one-sided version of the classical (Fréchet) derivative. A weakness this natural concept of subdifferential possesses is that even for well-behaved functions f it may be empty, see for example Proposition 6.8.1. The idea of the approximate subdifferential enhances the regular subdifferential by gathering information from the regular subdifferentials at points near x as well.

Definition 6.1.2 (Approximate Subgradient). *A vector* y *of* E *is an* **(approximate) 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 *(approximate) subdifferential* $\partial f(x)$.

Definition 6.1.3 (Horizon Subgradient). *A vector* 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* $\hat{\partial} f(x^r)$ *for which* $t_r y^r$ *approaches* 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).

Definition 6.1.4 (Clarke Regularity, Corollary 8.11 [79]). *If the function f is finite at the point x with at least one subgradient there then it is* **(Clarke) 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*

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

Definition 6.1.5 (Clarke Subgradients). *For a function f which 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 [15] - see for example [39, Theorem 2].)

Definition 6.1.6 (Contingent 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* **contingent cone** *to L at x, written $K(L|x)$, if either $d = 0$ or there is a sequence (x^r) in L approaching x with $\|x^r - x\|^{-1}(x^r - x)$ approaching $\|d\|^{-1}d$.*

Definition 6.1.7 (Negative Polar Cone). *The* **(negative) polar** *of a subset H of E is the set*

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

We use the following easy and standard result later.

Proposition 6.1.8 (Normal Cone). *Given a function $f : E \to [-\infty, +\infty]$ and a point x^0 in E, any regular subgradient of f at x^0 is negative polar to the contingent cone of the level set $L = \{x \in E : f(x) \leq f(x^0)\}$ at x^0; that is*

$$\hat{\partial} f(x^0) \subset (K(L|x^0))^-.$$

Proof. See [52, Proposition 2.1]. □

In this chapter 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.

Proposition 6.1.9 (Subgradient Invariance). *If $f : E \to [-\infty, +\infty]$ 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.

Proof. See [52, Proposition 2.2]. □

This section ends with a lemma which is useful in the later analysis of regularity. For its proof see [52, Lemma 2.3].

Lemma 6.1.10 (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 $gH(C^\infty)$.*

6.2 The normal space

We need a little bit of differential geometry. Definitions for the relevant notions in this section can be found in the following two elementary introductions into the subject [12], [4].

If M is a differential manifold and $m \in M$, then $T_m M$ will denote the tangent space to M at the point m.

Lemma 6.2.1 (Manifold Sum). *Let M and M' be differential manifolds, and let p, p' denote the projections of $M \times M'$ onto M, M' respectively. Then the function*

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

defined by $w \mapsto (dp, dp')w$ *is an isomorphism.*

Proof. See [12, Proposition 4.5.1]. □

Theorem 6.2.2 (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* G/H *admits the structure of a quotient manifold of* G*.*

Proof. See [12, Proposition 12.9.4]. □

Theorem 6.2.3 (Orbit Submanifold). *Suppose* G *is a Lie transformation group on a Hausdorff manifold* M*. If the stabilizer* G_m *is not an open subgroup of* G*, then the mapping*

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

is an imbedding of the quotient manifold G/G_m *into* M*. Moreover, the orbit* Gm *in* M *can be given the structure of a submanifold of* M *diffeomorphic to* G/G_m *under* ϕ_m*.*

Proof. See [12, Proposition 13.3.1 & Proposition 13.3.2]. □

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 6.2.1 we see that $T_{(I_n, I_m)}O(n, m) = A(n) \times A(m)$.

Throughout the whole chapter we will assume that n and m are natural numbers and $n \leq m$. Consider the action of the group $O(n,m)$ on the Euclidean space $M_{n,m}$ (of $n \times m$ real matrices, with the inner product $\langle X, Y \rangle = \operatorname{tr} X^T Y$), 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. (For related results see [52, Theorem 3.1], [8, Proposition 14.1].)

Theorem 6.2.4 (Normal Space). *The orbit $O(n,m).X$ is a submanifold of the space $M_{n,m}$, with tangent space*

$$T_X O(n,m).X = \{XZ_m - Z_n X \,:\, Z_n \in A(n) \text{ and } Z_m \in A(m)\} \tag{6.1}$$

and normal space

$$(T_X O(n,m).X)^{\perp} = \{Y \in M_{n,m} \,:\, X^T Y \text{ and } XY^T \text{ symmetric}\}. \tag{6.2}$$

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

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

It is well known that there is a 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 6.2.3 it follows that the map ϕ is a diffeomorphism, and hence its differential $d\phi$ is an isomorphism between the corresponding tangent spaces

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

Consider, on the other hand, the quotient map

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

Then Theorem 6.2.2 tells us that π is a submersion, and this implies that its differential $d\pi$

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

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

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

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

$$\Phi : M_n \times M_m \to M_{n,m}$$
$$\Phi(U,V) = U^T XV,$$

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 $T_M(M_n) = M_n$ for each $M \in M_n$ it is easy to see that

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

We have that $O(n) \times O(m)$ is a submanifold of $M_n \times M_m$, so the tangent space $T_{(I_n,I_m)}(O(n) \times O(m))$ is isomorphic to a vector subspace of $T_{(I_n,I_m)}(M_n \times M_m)$. Also the end of Theorem 6.2.3 implies that the tangent space $T_X(O(n,m).X)$ is isomorphic to a vector subspace of $T_X(M_{n,m})$. If ι is the natural injection of $O(n) \times O(m)$ into $M_n \times M_m$, then from the definitions $\psi = \Phi \circ \iota$. So $d\psi = d\Phi \circ d\iota$, but $(d\iota)(Z_n, Z_m) = (Z_n, Z_m)$ for each (Z_n, Z_m) in $A(n) \times A(m)$, and we obtain $(d\psi)(Z_n, Z_m) = (d\Phi)(Z_n, Z_m) = Z_n^T X + XZ_m = XZ_m - Z_nX$, as we claimed.

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

$$\begin{aligned}\langle Y, XZ_m - Z_nX\rangle &= \operatorname{tr} Y^T(XZ_m - Z_nX)\\ &= \operatorname{tr}(Y^TXZ_m) - \operatorname{tr}(Y^TZ_nX)\\ &= \operatorname{tr}(Y^TXZ_m) - \operatorname{tr}(XY^TZ_n).\end{aligned}$$

We will show now that $\operatorname{tr}(Y^TXZ_m)$=0. Indeed

$$\begin{aligned}\operatorname{tr}(Y^TXZ_m) = \operatorname{tr}(Y^TXZ_m)^T = \operatorname{tr}(Z_m^TX^TY) = -\operatorname{tr}(Z_mX^TY)\\ = -\operatorname{tr}(Z_mY^TX) = -\operatorname{tr}(Y^TXZ_m),\end{aligned}$$

so $\operatorname{tr}(Y^TXZ_m)$=0. Analogously we get $\operatorname{tr}(XY^TZ_n) = 0$, so $Y \in (T_XO(n,m).X)^\perp$.

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

$$\operatorname{tr}(Y^TZ_nX) = \operatorname{tr}(XY^TZ_n) = \operatorname{tr}(XY^TZ_n)^T = \operatorname{tr}(Z_n^TYX^T) = -\operatorname{tr}(Z_nYX^T),$$

that is

$$\operatorname{tr}(XY^TZ_n) = -\operatorname{tr}(Z_nYX^T).$$

Let $Z_m = 0$. Then our assumption becomes $\operatorname{tr}(XY^TZ_n) = 0$ and consequently we have $\operatorname{tr}(Z_nYX^T) = 0$ and so is their difference:

$$\operatorname{tr}(XY^TZ_n - Z_nYX^T) = 0.$$

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

$$\begin{aligned} 0 &= \mathrm{tr}\left(XY^T(XY^T - YX^T) - (XY^T - YX^T)YX^T\right) \\ &= \mathrm{tr}\left(XY^T(XY^T - YX^T)\right) - \mathrm{tr}\left(YX^T(XY^T - YX^T)\right) \\ &= \mathrm{tr}\,(XY^T - YX^T)(XY^T - YX^T) = -\mathrm{tr}\,(XY^T - YX^T)^T(XY^T - YX^T), \end{aligned}$$

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

Throughout the entire chapter 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 $x \cdot y = (x_1y_1, ..., x_ny_n)$. We will need the following standard lemma in our proofs (see [48]).

Lemma 6.2.5. *Any vectors x and y in $\mathbb{R}^n$ satisfy 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 $Qy = \bar{y}$.

For any matrix $X \in M_{n,m}$, we denote by $X^{i,j}$ its (i,j)-th entry, and by $\sigma_1(X) \geq \sigma_2(X) \geq ... \geq \sigma_n(X)$ its singular values, also we define the vector $\sigma(X) = (\sigma_1(X), \sigma_2(X), ..., \sigma_n(X))^T$. If the matrix $X \in M_{n,n}$ is symmetric, then we denote by $\lambda_1(X) \geq \lambda_2(X) \geq ... \geq \lambda_n(X)$ its eigenvalues, and define the vector $\lambda(X) = (\lambda_1(X), \lambda_2(X), ..., \lambda_n(X))^T$. 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 draw 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.

Definition 6.2.6 (Simultaneous Decomposition). *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 $X = U_n^T(\operatorname{Diag}\sigma(X))U_m$ and $Y = U_n^T(\operatorname{Diag}\sigma(Y))U_m$.*

We need to introduce more notation that will be used only 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 6.2.9.

Lemma 6.2.7. *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*

$$Y = U_n^T(\operatorname{Diag} P_{(-)}\sigma(Y))U_m, \quad Z = U_n^T(\operatorname{Diag}\sigma(Z))U_m. \tag{6.3}$$

Proof. In the "if" direction the result is clear. For the converse, 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^TZ are just 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 $Y^TZ = A^T\Lambda A$ and $ZY^T = B^T\Lambda B$, where $\Lambda = \operatorname{Diag}\lambda(Y^TZ)$. Consequently $Y^TZ = (A^TB)(ZY^T)(B^TA)$. We make the substitution: $\breve{Y} = (A^TB)Y$ and $\breve{Z} = (A^TB)Z$. Then we have

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

that is $\breve{Y}^T$ and $\breve{Z}$ commute. Hence also $\breve{Y}$ and $\breve{Z}^T$ commute. Next, because $\breve{Y}^T$ and $\breve{Z}$ commute with the symmetric matrix $\breve{Y}^T\breve{Z}$ it follows that every eigenspace (all eigenvectors corresponding to one fixed eigenvalue) 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 $V_n^T\breve{Y}^TV_n$ and $V_n^T\breve{Z}V_n$ must be block diagonal (recall that eigenvectors corresponding to different eigenvalues are orthogonal):

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

where $\breve{Y}_i^T$,$\breve{Z}_i \in M_{n_i}$, $1 \leq n_i \leq n$, $n_1 + n_2 + \cdots n_l = n$, 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 (6.3). 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. Decomposition (6.3) holds with

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

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 just 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 (for any matrix Y, the eigenvalues of YY^T are the singular values of Y squared):

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

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

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:

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

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

$$\hat{Y}\hat{Y}^T = \mathrm{Diag}\,\sigma^2(Y), \quad \hat{Y}^T\hat{Y} = \mathrm{Diag}\,(\sigma^2(Y)^T, \underbrace{0, ..., 0}_{m-n})^T \tag{6.4}$$

and

$$\hat{Z}\hat{Z}^T = \mathrm{Diag}\,P_n\sigma^2(Z), \quad \hat{Z}^T\hat{Z} = \mathrm{Diag}\,P_m(\sigma^2(Z)^T, \underbrace{0, ..., 0}_{m-n})^T. \tag{6.5}$$

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 respectively $\hat{Z}(t_1, t_2, ..., t_l; p_1, p_2, ..., p_l)$ 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 (6.4) 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 6.1. 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 $1 \leq x \leq n$. But because of 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

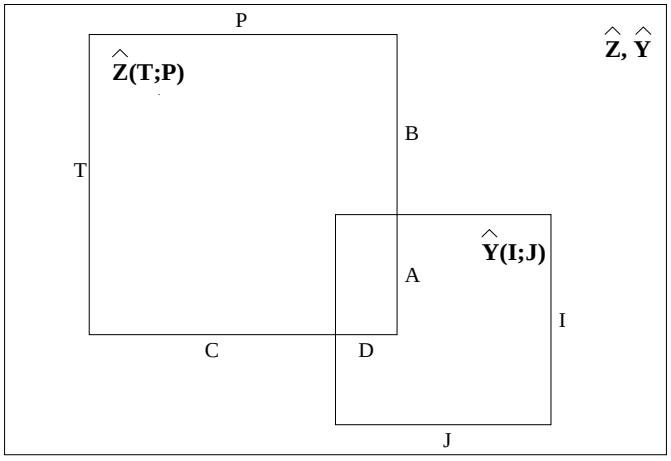

Figure 6.1: The sets I,J,T,P and A,B,C,D.

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 $|C| = |B|$. If $|C| < |B|$, then 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. If now $|C| > |B|$, then the columns of $\hat{Z}(B;C)$ are linearly dependent, and so will be the columns of $\hat{Z}(T;C)$, so 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:

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

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

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

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 decomposition 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 [50, Theorem 2.2 & Example 7.2]: we give a direct proof here.

Lemma 6.2.8. *For any vectors x and y in $\mathbb{R}^n$ we have the inequality*

$$x^T y \leq \hat{x}^T \hat{y}. \tag{6.6}$$

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 6.2.5. 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 6.2.5, 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 nonzero entries of I so that if $I_{(-)}$ is the so-formed matrix, we have $I_{(-)}x = |x|$ and $I_{(-)}y = |y|$. 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 $I^{i,i}_{(-)} = \text{sign}\,(x_i)$. We have that $QI_{(-)}x = \hat{x}$ and $QI_{(-)}y = \hat{y}$; let $P_{(-)} = QI_{(-)}$. □

The Normal Space Theorem (6.2.4) will be extremely useful to us in the following sections. However we can immediately demonstrate its importance by deriving next a famous inequality due to von Neumann [37, p. 182]. The following theorem may be viewed as a necessary and sufficient condition for two matrices to have a simultaneous ordered singular value decomposition.

Theorem 6.2.9 (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

$$\alpha = \sup_{Z \in O(n,m).X} \operatorname{tr} Y^T Z. \tag{6.7}$$

Observe first that there is an element (U_n, U_m) in $O(n,m)$ satisfying the equality $Y = U_n^T (\operatorname{Diag} \sigma(Y)) U_m$. 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 (6.7) has an optimal solution, $Z = Z_0$ say, and any such Z_0 by stationarity must satisfy

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

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 6.2.7, there is an element (U_n, U_m) in

$O(n,m)$, and a signed permutation matrix $P_{(-)}$ in $P_{(-)}(n)$ such that

$$Y = U_n^T(\operatorname{Diag} P_{(-)}\sigma(Y))U_m, \quad Z_0 = U_n^T(\operatorname{Diag}\sigma(Z_0))U_m. \tag{6.8}$$

Hence using Lemma 6.2.5 we get

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

Thus we can conclude that $\alpha = \sigma(X)^T\sigma(Y)$ and, using Lemma 6.2.8, 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 (6.8) 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

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

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.

Proposition 6.2.10 (Simultaneous Square Conjugacy). *For any vectors x, y, u, v in $\mathbb{R}^n$, there is a matrix U in $O(n)$ with*

$$\operatorname{Diag} x = U^T(\operatorname{Diag} u)U \quad \textit{and} \quad \operatorname{Diag} y = U^T(\operatorname{Diag} v)U$$

if and only if there is a matrix P in $P(n)$ with $x = Pu$ and $y = Pv$.

Proof. See [52, Proposition 3.8]. □

Proposition 6.2.11 (Simultaneous Rectangular Conjugacy). *For vectors x, y, u, and v in $\mathbb{R}^n$, there is an element (U_n, U_m) in $O(n,m)$ with*

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

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 6.2.10, 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 the vector u is equal to the number of zero entries in the 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

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

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

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 matches 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 $|P_{(-)}| = P$. If $\pi(i) = j$ (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 $y = P_{(-)}v$. □

6.3 Simultaneous Diagonalization

Proposition 6.3.1. (Orthogonally Invariant & Absolutely Symmetric) *The following two properties of a function $F : M_{n,m} \to [-\infty, +\infty]$ are equivalent:*

1. *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 $F(U_n^T X U_m) = F(X)$.*

2. *$F = f \circ \sigma$ 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 $f(Px) = f(x)$).*

Proof. Elementary. □

As we discussed in the Introduction, the singular value functions are important in various areas.

Definition 6.3.2 (Singular Value Function). *A* **singular value function** *is an extended-real-valued 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 6.3.3 (Symmetricity). *If a matrix Y in $M_{n,m}$ is a subgradient or a horizon subgradient of a singular value function at a matrix X in $M_{n,m}$, then X and Y satisfy $X^T Y = Y^T X$ and $Y^T X = X^T Y$. Furthermore, if the singular value function is Lipschitz around X, and Y is a Clarke subgradient there, then again $X^T Y = Y^T X$ and $Y^T X = X^T Y$.*

Proof. Call the singular value function F, and assume first that the subgradient Y is regular. By the Normal Cone Proposition (6.1.8), the constancy of F on the

orbit $O(n,m).X$ shows

$$\begin{aligned} Y &\in \big(K(\{Z : F(Z) \leq F(X)\}|X)\big)^- \\ &\subset \big(K(O(n,m).X|X)\big)^- = \big(T_X(O(n,m).X)\big)^\perp. \end{aligned}$$

The result follows from the Normal Space Theorem (6.2.4).

Next, let Y be an (approximate) 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

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

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

If Y is a horizon subgradient then there is a sequence Y_r approaching Y and real numbers 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

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

Using Definition 6.1.5, when the singular value function is locally Lipschitz then any Clarke subgradient is a convex combination of subgradients, and since every subgradient satisfies the two properties in the theorem, so must any convex combination. □

Hence if a matrix Y in $M_{n,m}$ is a subgradient of some singular value function at

the matrix X in $M_{n,m}$ then by Lemma 6.2.7 we get that

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

for some element (U_n, U_m) in $O(n,m)$, and some $P_{(-)}$ in $P_{(-)}(n)$. Consequently, by the Subgradient Invariance Proposition (6.1.9) applied to the space $M_{n,m}$ with the action of the group $O(n,m)$, the matrix $\operatorname{Diag} P_{(-)}\sigma(Y)$ must be a subgradient at $\operatorname{Diag}\sigma(X)$. Consequently in order to characterize when a matrix Y is a subgradient of a singular value function at a matrix X, it is enough to consider the case when X and Y are both diagonal (by which we mean $X_{i,j} = 0$ if $i \neq j$). In one direction this is not too hard, and we show it below.

Proposition 6.3.4. *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. As in the previous theorem we show first that the claim holds when $\operatorname{Diag} y$ is a regular subgradient of $f \circ \sigma$ at $\operatorname{Diag} x$. For small vectors z in $\mathbb{R}^n$ we obtain

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

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

Next, if $\operatorname{Diag} y \in \partial(f \circ \sigma)(\operatorname{Diag} x)$, then there is a sequence of matrices X_r in $M_{n,m}$ approaching $\operatorname{Diag} x$, with $f(\sigma(X_r))$ approaching $f(\sigma(\operatorname{Diag} x))$, and a sequence of regular subgradients Y_r in $\hat{\partial}(f \circ \sigma)(X_r)$ approaching $\operatorname{Diag} y$. By Theorem 6.3.3 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

$$X_r = (U_n^r)^T \big(\operatorname{Diag} P_{(-)}^r \sigma(X_r)\big) U_m^r \text{ and } Y_r = (U_n^r)^T \big(\operatorname{Diag} \sigma(Y_r)\big) U_m^r \tag{6.9}$$

for every r. The Subgradient Invariance Proposition (6.1.9) now shows that

$$\operatorname{Diag} \sigma(Y_r) \in \hat{\partial}(f \circ \sigma)\big(\operatorname{Diag} P_{(-)}^r \sigma(X_r)\big),$$

whence by the first part $\sigma(Y_r) \in \hat{\partial} f(P_{(-)}^r \sigma(X_r))$.

The groups $O(n,m)$ and $P_{(-)}$ 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 (6.9), taking the limit and rearranging we get

$$\begin{aligned} U_n(\operatorname{Diag} x)U_m^T &= \operatorname{Diag}\left(P_{(-)}\sigma(\operatorname{Diag} x)\right), \quad \text{and} \\ U_n(\operatorname{Diag} y)U_m^T &= \operatorname{Diag} \sigma(\operatorname{Diag} y). \end{aligned} \tag{6.10}$$

Since $P_{(-)}^r \sigma(X_r)$ approaches $P_{(-)}\sigma(\operatorname{Diag} x)$, with $f(P_{(-)}^r \sigma(X_r)) = f(\sigma(X_r))$ ap-

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

Combining Equation (6.10) and Proposition 6.2.11, 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 (6.1.9) 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.

□

6.4 Directional derivatives of singular values

The aim of this section is to prove the reverse implication of the one stated in Proposition 6.3.4. The main difficulty is to show that for vectors x and y in $\mathbb{R}^n$ and a singular value function $f \circ \sigma$ we have

$$y \in \hat{\partial} f(x) \Rightarrow \operatorname{Diag} y \in \hat{\partial}(f \circ \sigma)(\operatorname{Diag} x). \tag{6.11}$$

After that, to prove the same implication for the (approximate) subdifferential will be easy. We need two propositions whose proofs can be found in [47, Corollary 2.6 and Theorem 3.1]. One may also want to compare the following two results with Theorem 2.3.9 and Corollary 2.5.6 respectively.

Proposition 6.4.1 (Characterization Of Convexity). *Suppose that the function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is absolutely symmetric. Then the corresponding singular value function $f \circ \sigma$ is convex and lower semicontinuous on $M_{n,m}$ if and only if f*

is convex and lower semicontinuous.

Proposition 6.4.2 (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 a convex (even sublinear) function (see also [36]). One can see this also using Proposition 6.4.1. We define a new symbol $\mathbb{R}^n_{\downarrow} := (\mathbb{R}^n_{\downarrow} \cap \mathbb{R}^n_+)$. To simplify the notation in the following few lemmas, if x is a vector from $\mathbb{R}^n$ we will define $x_{n+1} = 0$.

Lemma 6.4.3. *The function $f : \mathbb{R}^n \to (-\infty, +\infty)$ 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 $f(\mu+x) = \sum_{i=1}^k \mu_i + x_i$. 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 6.4.4. *Fix an integer k, $1 \le k \le n$. 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 and U_m are any orthogonal matrices such that $\operatorname{Diag} x = U_n^T (\operatorname{Diag} \hat{x}) U_m$.

Note 6.4.5. *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* $\operatorname{Diag}(|P_{(-)}|, I_{m-n,m-n})$.

Proof. The function $f : \mathbb{R}^n \to (-\infty, +\infty)$ defined by $f(y) = \sum_{i=1}^{k} \hat{y}_i$ is easily seen to be absolutely symmetric and convex. From Lemma 6.4.3 it is also differentiable at the point $\sigma(\operatorname{Diag} x) = \hat{x}$. So by Proposition 6.4.2 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 6.4.2 and Lemma 6.4.3. □

Lemma 6.4.6. *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 6.2.8. Then by Proposition 6.4.1 we obtain that $f \circ \sigma$ is convex. To prove the claim about the subgradient 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 (6.2.9). □

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

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

Lemma 6.4.7. *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 6.4.4 gives

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

as required. □

The following theorem, which will be used in proving implication (6.11), gives information about the directional derivatives of singular values. 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 6.4.8 (Singular Value Derivatives). *Any vector x in $\mathbb{R}^n_{\downarrow}$ and matrix M in $M_{n,m}$ satisfy*

$$\operatorname{diag} M \in \operatorname{conv}\big(P_{(-)}(n)_x \sigma'(\operatorname{Diag} x; M)\big). \tag{6.12}$$

Proof. Assume first that $x_n = 0$. Partition the set of integers $\{1, 2, ..., n\}$ into

consecutive blocks I_1,I_2,...,I_r,I_{r+1}, so that $x_i = x_j$ if and only if the indices i and j belong to the same block. Let us agree that $x_i \in I_{r+1}$ if and only if $x_i = 0$. We are 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 (6.12) 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

$$y^T \operatorname{diag} M > y^T P_{(-)} \sigma'(\operatorname{Diag} x; M), \text{ for all } P_{(-)} \text{ in } P_{(-)}(n)_x. \tag{6.13}$$

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 6.4.6 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 any 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

$$\operatorname{tr}\big(T^T(\operatorname{Diag}(v+\tilde{y}))\big) \leq (v+\tilde{y})^T\sigma'(\operatorname{Diag} x; T), \tag{6.14}$$

for any matrix T in $M_{n,m}$. On the other hand, Lemma 6.4.7 shows that

$$\operatorname{tr}\big(T^T(\operatorname{Diag} v)\big) = v^T\sigma'(\operatorname{Diag} x; T). \tag{6.15}$$

Subtracting equation (6.15) from inequality (6.14) gives

$$\operatorname{tr}\big(T^T(\operatorname{Diag}\tilde{y})\big) \leq \tilde{y}^T\sigma'(\operatorname{Diag} x; T). \tag{6.16}$$

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

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

Choosing the matrix T in inequality (6.16) to be $T = Q^T M \begin{pmatrix} |Q| & 0 \\ 0 & I_{m-n,m-n} \end{pmatrix}$ and using Lemma 6.2.5 repeatedly and Lemma 6.2.8 shows

$$y^T w \leq \big(\oplus_{i=1}^r \overline{y^i}\big)^T \big(\oplus_{i=1}^r \overline{w^i}\big) + \widehat{y^{r+1}}^T \widehat{w^{r+1}}$$

$$
\begin{aligned}
&= \operatorname{tr}\left(T^T(\operatorname{Diag}\tilde{y})\right) \\
&\leq \tilde{y}^T\sigma'(\operatorname{Diag}x;T) \\
&= \tilde{y}^T\sigma'(\operatorname{Diag}x;M).
\end{aligned}
$$

In the last equality we used the Subgradient Invariance Proposition (6.1.9). But now choosing the matrix $P_{(-)} \in P_{(-)}(n)_x$ in inequality (6.13) 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 think that the block I_{r+1} is empty. □

Another result that we will need is that the singular value map σ can be expanded in a first order series, and this expansion stays valid when the direction varies freely. In other words we have the following lemma.

Lemma 6.4.9. *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 first order expansion is true for any continuous convex function. For a proof of this fact see [31, 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 6.4.6). So it is true for σ_i as well. □

Finally we prove the implication (6.11). Notice though, that we require x to be in $\mathbb{R}^n_{\downarrow}$. In the corollary that follows we remove this condition.

Theorem 6.4.10. *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. By the Subgradient Invariance Proposition (6.1.9), every element of the finite set $P_{(-)}(n)_x y$ is a regular subgradient of f at x. The convex hull of this set, which we denote by Λ, has support function given by

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

This function is sublinear, with global Lipschitz constant $\|y\|$.

Fix a real $\epsilon > 0$. The definition of regular subgradients implies, for small vectors z in $\mathbb{R}^n$,

$$f(x+z) \geq f(x) + \delta^*_\Lambda(z) - \epsilon\|z\|. \tag{6.17}$$

On the other hand, using the previous lemma (6.4.9), 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 (6.17),

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

$$\geq f(x) + \delta^*_\Lambda\big(\sigma'(\operatorname{Diag} x; Z)\big) - (1 + \|y\|)\epsilon\|Z\|,$$

using the Lipschitz property of σ and the Lipschitzness of δ^*_Λ. The Singular Value Derivatives Theorem (6.4.8) implies

$$\operatorname{diag} Z \in \operatorname{conv}\big(P_{(-)}(n)_x \sigma'(\operatorname{Diag} x; Z)\big). \tag{6.18}$$

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

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

for any matrix $P_{(-)}$ in $P_{(-)}(n)_x$. This combined with the convexity of δ^*_Λ and relation (6.18), demonstrates

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

Continuing the argument above we have

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

and since ϵ was arbitrary, the result follows. □

Corollary 6.4.11 (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. If f is Lipschitz around $\sigma(X)$ then the implication '$\Rightarrow$' also holds for Clarke subgradients.

Proof. We prove only the implication '$\Rightarrow$', because the opposite direction is Proposition 6.3.4. Again we first show it in the case when y is a regular subgradient. Fixing a matrix $P_{(-)}$ in $P_{(-)}(n)$ satisfying $\hat{x} = P_{(-)}x$, the assumption $y \in \hat{\partial} f(x)$ implies $P_{(-)}y \in \hat{\partial} f(P_{(-)}x)$, by the Subgradient Invariance Proposition (6.1.9). Now we can apply the previous result:

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

and the result follows by applying the Subgradient Invariance Proposition again.

Now suppose $y \in \partial f(x)$, so 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. Hence $\operatorname{Diag} x^r$ 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.

If the function f is Lipschitz around $\sigma(X)$ and y is a Clarke subgradient at x, then y is a convex combination of subgradients $y^i \in \partial f(x)$. Since by the above argument each matrix $\operatorname{Diag} y^i$ is a subgradient of $f \circ \sigma$ at X, and $\operatorname{Diag} y$ is a convex combination of these matrices, $\operatorname{Diag} y$ must be a Clarke subgradient. □

Note 6.4.12. *We prove the converse implication '$\Leftarrow$' in the Clarke case in Section 6.6.*

6.5 The main result

We present the main result of the chapter in this section. It is an easy formula describing the subgradients of any singular value function in terms of its underlying absolutely symmetric function. The proof reduces the general case to the diagonal case developed in the previous section.

Theorem 6.5.1 (Subgradients). *The (approximate) subdifferential of a singular value function $f \circ \sigma$ at a matrix X in $M_{n,m}$ is given by the formula*

$$\partial(f \circ \sigma)(X) = O(n,m)^X.\operatorname{Diag} \partial f(\sigma(X)), \tag{6.19}$$

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. For any vector y in $\partial f(\sigma(X))$, the Diagonal Subgradients Corollary (6.4.11) shows

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

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 (6.1.9) 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 showed the inclusion $\partial(f \circ \sigma)(X) \supseteq O(n,m)^X.\operatorname{Diag}\partial f(\sigma(X))$.

For the opposite inclusion, take a subgradient Y in $\partial(f \circ \sigma)(X)$. By the Symmetricity Theorem (6.3.3) it satisfies the relationships: $Y^TX = X^TY$ and $YX^T = XY^T$. Hence by Lemma 6.2.7 there exists an element (U_n, U_m) in $O(n,m)$ and a signed permutation matrix $P_{(-)}$ in $P_{(-)}(n)$ such that

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

Then the Subgradient Invariance Proposition (6.1.9) shows

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

whence $P_{(-)}\sigma(Y) \in \partial f(\sigma(X))$, by the Diagonal Subgradient Corollary. Thus the matrix Y belongs to the right-hand-side set above, as required. The arguments for regular and horizon subgradients are similar. □

Note 6.5.2. *Same result holds for Clarke subgradient - see Section 6.6. In the case*

when f is lower semicontinuous see Section 6.7.

Corollary 6.5.3 (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. Suppose f has unique regular subgradient y at $\sigma(X)$. Then by the subdifferential formula (6.19) we get that every matrix in the nonempty convex set $\hat{\partial}(f \circ \sigma)(X)$ has the same norm, namely $\|y\|$, and therefore this set is a singleton. The converse is obvious. □

When f is Lipschitz around $\sigma(X)$, then $f \circ \sigma$ is strictly differentiable at X if and only if f strictly differentiable at $\sigma(X)$. The proof follows from the above corollary and Note 6.5.2, because in the Lipschitz case f is strictly differentiable at x if and only if $\partial^c f(x) = \{\phi\}$, that is, if and only if the Clarke subdifferential is a singleton. In that case ϕ is the strict derivative of f at x [10, Exercise 6.4.7].

Corollary 6.5.4 (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. This follows immediately from Corollary 6.5.3, since a function h is Fréchet differentiable at a point if and only if both h and $-h$ have unique regular subgradients there. □

Corollary 6.5.5 (Regularity). *Suppose the absolutely symmetric function f is finite at $\sigma(X)$ (for a matrix X in $M_{n,m}$) with $\partial f(\sigma(X)) \neq \emptyset$. Then the singular*

value function $f \circ \sigma$ is (Clarke) regular at X if and only if f is (Clarke) regular at $\sigma(X)$.

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

Definition 6.1.4 says that if $\partial f(\sigma(X)) \neq \emptyset$, then f is regular at $\sigma(X)$ if and only if it is lower semicontinuous around $\sigma(X)$ and the following conditions hold

$$\partial f(\sigma(X)) = \hat{\partial} f(\sigma(X)), \text{ and} \tag{6.20}$$

$$(\hat{\partial} f(\sigma(X)))^{\infty} = \partial^{\infty} f(\sigma(X)). \tag{6.21}$$

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

$$\partial(f \circ \sigma)(X) = \hat{\partial}(f \circ \sigma)(X), \text{ and} \tag{6.22}$$

$$(\hat{\partial}(f \circ \sigma)(X))^{\infty} = \partial^{\infty}(f \circ \sigma)(X). \tag{6.23}$$

By formula (6.19) and its regular analogue, condition (6.20) implies condition (6.22). Conversely, by the Subgradient Invariance Proposition (6.1.9), condition (6.22) is equivalent to

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

and condition (6.20) follows by the Diagonal Subgradient Corollary (6.4.11).

Applying the Recession Lemma (6.1.10) to the regular version of formula (6.19),

noting that the set of regular subgradients is always closed and convex, and assuming that (6.21) holds, implies that

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

So condition (6.21) implies condition (6.23), by the horizon version of formula (6.19) used in the last equality.

On the other hand, by the Subgradient Invariance Proposition (6.1.9), condition (6.23) is equivalent to

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

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

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

Condition (6.21) follows. □

Corollary 6.5.6 (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. Strict differentiability of f at $\sigma(X)$ is equivalent by [79, Thm 9.18] to continuity in a neighbourhood and regularity of both f and $-f$ at $\sigma(X)$. The result follows by the Regularity Corollary (6.5.5). □

The Subgradients Theorem (6.5.1) 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 6.5.7 (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 $\hat{\partial}$, ∂^∞, and (in the locally Lipschitz case) ∂^c.

Proof. Suppose first that the pair of matrices (X, Y) lies in $\operatorname{Graph} \partial(f \circ \sigma)$. This happens exactly when $Y \in \partial(f \circ \sigma)(X)$. Using the Subgradients Theorem (6.5.1),

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 $(X,Y) = (U_n, U_m).(\sigma(X), y)$.

Conversely, for a pair of vectors (x,y) in $\mathrm{Graph}\,\partial f$ and an element (U_n, U_m) in $O(n,m)$, y lies in $\partial f(x)$, whence $\mathrm{Diag}\, y \in \partial(f \circ \sigma)(\mathrm{Diag}\, x)$, by the Diagonal Subgradients Corollary (6.4.11). The Subgradient Invariance Proposition now implies $(U_n, U_m).\mathrm{Diag}\, y \in \partial(f \circ \sigma)((U_n, U_m).\mathrm{Diag}\, x)$, or in other words $(U_n, U_m) * (x,y) \in \mathrm{Graph}\,\partial(f \circ \sigma)$. The arguments for the other subdifferentials are exactly analogous. □

The regular subgradients of a convex function are exactly the usual convex subgradients [79, Proposition 8.12]. 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 [50, Theorem 4.3 and Example 7.5].) With this in mind the following corollary is easily deduced from the Subgradients Theorem. An independent proof can be found in [47, Corollary 2.5].

Corollary 6.5.8 (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.*

6.6 Clarke subgradients - the Lipschitz case

As we said in Note 6.5.2 the Subgradients Theorem (6.5.1) can be extended word by word to the case of the Clarke subdifferential. The problem is the missing converse

in the Diagonal Subgradients Corollary (6.4.11). In this section we fill this gap. We need some notation and a few lemmas.

If X is a 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 is similar to the proof of Lemma 6.4.6 and can be found in [52, Lemma 5.2], is needed below.

Lemma 6.6.1. *For any vector w in $\mathbb{R}^n_\downarrow$, the function $w^T\lambda$ is convex on $S(n)$, and any vector x in $\mathbb{R}^n_\downarrow$ satisfies* $\operatorname{Diag} w \in \partial(w^T\lambda)(\operatorname{Diag} x)$.

The following lemma should be compared to Lemmas 6.4.6, Lemma 6.6.1, and Corollary 2.3.5. Its proof is immediate.

Lemma 6.6.2. *1. For any vector w in $\mathbb{R}^n_\downarrow$ the function $w^T\lambda$ is sublinear.*

2. For any vector w in $\mathbb{R}^n_{\not\downarrow}$ the function $w^T\sigma$ is sublinear.

A subset C of the Euclidean space E is *invariant* under a subgroup of $O(E)$ if $gC = C$ for all transformations g in G. If the function $f : \mathbb{R}^n \to [-\infty, +\infty]$ 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 Proposition (6.1.9)).

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

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

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

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

It is clear that $\mathrm{Diag}\,(U^l)$ will be either an $n \times n$ 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:

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

Notice, for any vector y in $\mathbb{R}^n$ satisfying $y_i = y_j \ \Leftrightarrow i,j$ in some I_l, and $y_i = 0 \ \Leftrightarrow \ i \in I_{r+1}$, the group $\tilde{P}(n)$ is the stabilizer in $P_{(-)}(n)$ of y, and the group $\tilde{O}(n,m)$ is the stabilizer in $O(n,m)$ of $\mathrm{Diag}\, y$. (See Equation 6.26.)

Lemma 6.6.3 (Sum Of Invariant Sets). *If the sets C, $D \subset \mathbb{R}^n$ are convex and*

invariant under the group $\tilde{P}(n)$ *then*

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

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

$$\tilde{O}(n,m).\mathrm{Diag}\, C = \{\mathrm{Diag}\,(X^l) : \oplus_{l=1}^{r} \lambda(X^l) \oplus \sigma(X^{r+1}) \in C\}. \tag{6.24}$$

Let

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

We wish to show

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

or equivalently, by identity (6.24),

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

Since identity (6.24) shows $\oplus_{l=1}^{r} \lambda(X^l) \oplus \sigma(X^{r+1})$ lies in the convex set C and $\oplus_{l=1}^{r} \lambda(Y^l) \oplus \sigma(Y^{r+1})$ lies in the convex set D, it suffices to show

$$\oplus_{l=1}^{r} \lambda(X^l + Y^l) \oplus \sigma(X^{r+1} + Y^{r+1}) \in$$
$$\mathrm{conv}\,(\tilde{P}(n)(\oplus_{l=1}^{r} \lambda(X^l) \oplus \sigma(X^{r+1}))) + \mathrm{conv}\,(\tilde{P}(n)(\oplus_{l=1}^{r} \lambda(Y^l) \oplus \sigma(Y^{r+1}))).$$

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

$$\begin{aligned}
&\langle z,\ \oplus_{l=1}^r \lambda(X^l + Y^l) \oplus \sigma(X^{r+1} + Y^{r+1})\rangle \\
&\qquad > \max\langle z,\ \mathrm{conv}\,(\tilde{P}(n)(\ \oplus_{l=1}^r \lambda(X^l) \oplus \sigma(X^{r+1}))) \\
&\qquad\qquad + \mathrm{conv}\,(\tilde{P}(n)(\oplus_{l=1}^r \lambda(Y^l) \oplus \sigma(Y^{r+1})))\rangle \\
&\qquad = \max\langle z,\ \tilde{P}(n)(\oplus_{l=1}^r \lambda(X^l) \oplus \sigma(X^{r+1}))\rangle \\
&\qquad\qquad + \max\langle z,\ \tilde{P}(n)(\oplus_{l=1}^r \lambda(Y^l) \oplus \sigma(Y^{r+1}))\rangle.
\end{aligned}$$

But then Lemmas 6.2.5, 6.2.8 and 6.6.2 show

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

which is a contradiction. □

Corollary 6.6.4 (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 $\tilde{O}(n,m).\mathrm{Diag}\, C$ is convex.*

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

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

where λ is a number in $[0,1]$. □

Lemma 6.6.5. *If the set $C \subset \mathbb{R}^n$ is invariant under the group $\tilde{P}(n)$, then the following equality holds*

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

Proof. It is clear that $\tilde{O}(n,m).\mathrm{Diag}\, C \subset \tilde{O}(n,m).\mathrm{Diag}\,(\mathrm{conv}\, C)$, and the latter set is convex because of Corollary 6.6.4. Consequently

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

The opposite inclusion is trivial. □

Theorem 6.6.6 (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*

$$\partial^c (f \circ \sigma)(X) = O(n,m)^X.\mathrm{Diag}\, \partial^c f(\sigma(X)), \tag{6.25}$$

where

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

Proof. Assume first $X = \mathrm{Diag}\,x$ for a vector x in $\mathbb{R}^n_{\downarrow}$. After that the general case will follow easily by the Subgradient Invariance Proposition (6.1.9). Let

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

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

We are going to compute the group $O(n,m)^{\mathrm{Diag}\,x}$. If (U_n, U_m) is in $O(n,m)^{\mathrm{Diag}\,x}$, then we have

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

which shows that U_n =Diag (U^l), where $U^l \in O(|I_l|)$ for $1 \leq l \leq r+1$, and U_m =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

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

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

$$O(n,m)^{\mathrm{Diag}\ x} = \tilde{O}(n,m). \tag{6.26}$$

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 Proposition (6.1.9). Corollary 6.6.4 now shows that the set $\tilde{O}(n,m).\mathrm{Diag}\,\partial^c f(x)$ is convex.

The Subgradient Theorem (6.5.1) now gives us

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

Using the easily established fact

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

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

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

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

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

The result follows. □

For completeness we would like to state and prove the Clarke version of the Diagonal Subgradient Corollary.

Corollary 6.6.7 (Diagonal Clarke Subgradients). *For any vectors x and y in $\mathbb{R}^n$ and any singular value function $f \circ \sigma$,*

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

Proof. We already know that the implication '$\Rightarrow$' holds, and was proved in the Diagonal Subgradients Corollary (6.4.11). To see the reverse implication choose a diagonal matrix $\operatorname{Diag} y \in \partial^c (f \circ \sigma)(\operatorname{Diag} x)$. Then the Clarke Subgradients Theorem above shows the existence of an element (U_n, U_m) in $O(n,m)$ and a vector z in $\partial^c f(\hat{x})$ such that $\operatorname{Diag} y = (U_n, U_m).\operatorname{Diag} z$ and $\operatorname{Diag} x = (U_n, U_m).\operatorname{Diag} \hat{x}$. By the Simultaneous Rectangular Conjugacy Proposition (6.2.11), 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 Proposition (6.1.9). □

6.7 Clarke subgradients - the lower semicontinuous case

In this section we extend our previous result on Clarke subgradients to the non-Lipschitz case.

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

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

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 $\limsup_{z \to 0} \frac{\langle v, x+z \rangle}{|x+z|} \leq 0$. 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. The set of *Clarke subgradients* of a function f at x, $\partial^c f(x)$, is defined by

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

and is called *Clarke subdifferential*. It can be shown that if f is locally Lipschitz around x then this definition coincides with the definition given at the beginning, so there will be no danger of confusion. (See [79, Theorem 9.13 (b) and Theorem 8.49].) If f is lower semicontinuous around x then we have the formula (see [79, Theorem 8.9]):

$$N_{\operatorname{epi} f}(x, f(x)) = \{\lambda(v, -1) \mid v \in \partial f(x),\ \lambda > 0\} \cup \{(v, 0) \mid v \in \partial^\infty f(x)\}. \tag{6.27}$$

Notice that this cone is closed.

Lemma 6.7.1. *If f is lower semicontinuous around x we have the representation*

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

In particular when the cone $\partial^\infty f(x)$ doesn't contain lines we have simpler

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

Proof. Define the sets

$$\begin{aligned} K_1 &= \{(v,0) \,|\, v \in \partial^\infty f(x)\}, \\ K_2 &= \{\lambda(v,-1) \,|\, v \in \partial f(x),\ \lambda > 0\}, \quad \text{and} \\ L &= \{x \in \mathbb{R}^{n+1} \,|\, x_{n+1} = -1\}. \end{aligned}$$

Then by (6.27) we get

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

and by the definition of the set L

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

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

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

Indeed, take a point $(v,-1)$ in $(\operatorname{cl} \operatorname{conv} N_{\operatorname{epi} f}(x, f(x))) \cap L$. So there is a sequence (v^r, α^r) in $\operatorname{conv} N_{\operatorname{epi} f}(x, f(x))$, 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 $\operatorname{conv} N_{\operatorname{epi} f}(x, f(x)) \cap L$, approaching

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

So

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

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 $N_{\operatorname{epi} f}(x,f(x))$ doesn't contain lines. Then by [79, Theorem 3.15]

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

and the second formula becomes clear. □

We now prove a proposition that resembles Proposition 6.4.1 and exhibits another property of absolutely symmetric functions that is preserved after composition with σ.

Proposition 6.7.2 (Characterization of Sublinearity). *Suppose the function $f : \mathbb{R}^n \to (-\infty,+\infty]$ is absolutely symmetric. Then the corresponding singular value function is sublinear and lower semicontinuous on $M_{n,m}$ if and only if f is sublinear and lower semicontinuous.*

Proof. One way to prove this proposition is to use Proposition 6.4.1 and the fact that f is sublinear if and only if it is convex and positively homogeneous, plus the fact that σ is positively homogeneous. □

Let (U_n°, U_m°) be an arbitrary, fixed element of the set $O(n,m)^X$. Then the representation $O(n,m)^X = (U_n^\circ, U_m^\circ)\, O(n,m)_{\mathrm{Diag}\,\sigma(X)}$ holds. (Recall that $O(n,m)_{\mathrm{Diag}\,\sigma(X)}$ denotes the stabilizer of the matrix $\mathrm{Diag}\,\sigma(X)$ in the group $O(n,m)$.) Notice that the matrices in the stabilizer $O(n,m)_{\mathrm{Diag}\,\sigma(X)}$ have the same structure as those in the set $\tilde{O}(n,m)$ in Lemma 6.6.3 and Corollary 6.6.4. Let now f be an absolutely symmetric function. Then f is lower semicontinuous if and only if $f \circ \sigma$ is lower semicontinuous. Using (in this order) Lemma 6.7.1, Theorem 6.5.1, Lemma 6.6.5, Corollary 6.6.4, Lemma 6.6.3, 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:

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

This proves the following theorem.

Theorem 6.7.3. *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

$$\partial^c(f \circ \sigma)(X) = O(n,m)^X . \partial^c(f(\sigma(X)),$$

where

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

Analogous argument proves the corresponding result for spectral functions of symmetric matrices - a result left unproven in [52].

6.8 Absolute order statistics & individual singular values

In this section we want to present a useful application of the Subgradient Theorem (6.5.1). We are going to calculate the approximate and Clarke subdifferentials of an individual singular value $\sigma_k(\cdot)$. The availability of such formulas may be useful in further research 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* $\varphi_k : \mathbb{R}^n \to \mathbb{R}$ is defined to be

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

(or in other words $\varphi_k(x) = (\hat{x})_k$). It clearly satisfies the relation $\varphi_k(x) = \sigma_k(\mathrm{Diag}\, x)$. 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 $\operatorname{sign}(x)$ as

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

Proposition 6.8.1. *At any point x in $\mathbb{R}^n$, the regular subgradients of the k^{th} absolute order statistic are described by*

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

and $\partial^\infty \varphi_k(x) = \{0\}$.

Proof. Define the set of indices $I = \{i : |x_i| = \varphi_k(x)\}$, 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 $\operatorname{conv}\{\pm e^i : |x_i| = \varphi_k(x)\}$ if $\varphi_k(x) = 0$ or is $\operatorname{conv}\{(\operatorname{sign}(x_i))e^i : |x_i| = \varphi_k(x)\}$ if $\varphi_k(x) \neq 0$. (See [78, Theorem 23.8].)

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

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

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

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

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 see, since φ_k is Lipschitz. □

For a vector y in $\mathbb{R}^n$ we define

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

The number of elements in this set is then $|\text{supp}\, y|$.

Theorem 6.8.2 (k^{th} Absolute Order Statistic). *The Clarke subdifferential of*

the k^{th} absolute order statistic φ_k at a point x in $\mathbb{R}^n$ is given by

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

whereas the (approximate) subdifferential is given by

$$\partial \varphi_k(x) = \{y \in \partial^c \varphi_k(x) : |\operatorname{supp} y| \leq \alpha\}, \text{ where} \tag{6.28}$$
$$\alpha = 1 - k + |\{i : |x_i| \geq \varphi_k(x)\}|.$$

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

Proof. We begin by proving Equation (6.28). 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 6.8.1 one can easily see that for all z in that neighbourhood $\hat{\partial}\varphi_k(z)$ is contained in the set in the right hand side of Equation (6.28). 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 (6.28) and an index set J such that

$$|J| = n - \alpha,$$
$$j \in J \;\Rightarrow\; y_j = 0,$$
$$\{i : |x_i| > \varphi_k(x)\} \cup \{i : |x_i| < \varphi_k(x)\} \subseteq J.$$

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

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

Finally using Proposition 6.8.1 we see that

$$y \in \left\{ \begin{array}{c} \operatorname{conv}\{\pm e^i : i \notin J\} \\ \operatorname{conv}\{(\operatorname{sign}(x_i))e^i : i \notin J\} \end{array} \right\} = \hat{\partial}\varphi_k\left(x+\delta\sum_{i\in J}(\operatorname{sign}(x_i))e^i\right),$$

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 6.8.1.

□

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

Corollary 6.8.3 (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*

$$\partial^c\sigma_k(X) = \left\{ \begin{array}{ll} \operatorname{conv}\{\pm uv^T : (u,v) \in \Sigma_k(X)\}, & \text{if } \sigma_k(X) = 0 \\ \operatorname{conv}\{uv^T : (u,v) \in \Sigma_k(X)\}, & \text{otherwise ,} \end{array} \right.$$

where

$$\Sigma_k(X) = \{(u,v) \in \mathbb{R}^n \times \mathbb{R}^m \,|\, \|u\| = \|v\| = 1, XX^Tu = \sigma_k^2(X)u, X^TXv = \sigma_k^2(X)v\}.$$

On the other hand the (approximate) subdifferential is given by

$$\partial\sigma_k(X) = \{Y \in \partial^c\sigma_k(X) : \operatorname{rank} Y \leq \alpha\}, \text{ where}$$
$$\alpha = 1 - k + |\{i : \sigma_i(X) \geq \sigma_k(X)\}|.$$

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

Proof. First we deduce the formula for the Clarke subdifferential. Fix a matrix X. Suppose first that $\sigma_k(X) = 0$. By Theorem 6.6.6 we get

$$\partial^c\sigma_k(X) = (U_n^\circ, U_m^\circ)O(n,m)_{\operatorname{Diag}\sigma(X)}.\big(\operatorname{Diag}\operatorname{conv}\{\pm e^i : \sigma_i(X) = \sigma_k(X)\}\big),$$

where (U_n°, U_m°) is a fixed element of $O(n,m)^X$. The set $\{\pm e^i : \sigma_i(X) = \sigma_k(X)\}$ is clearly invariant under the subgroup, $\tilde{P}(n)$, of $P_{(-)}(n)$ that stabilizes $\sigma(X)$. Then by Lemma 6.6.5 and recalling the $O(n,m)_{\operatorname{Diag}\sigma(X)} = \tilde{O}(n,m)$ we obtain

$$\begin{aligned}\partial^c\sigma_k(X) &= (U_n^\circ, U_m^\circ)\operatorname{conv}\tilde{O}(n,m).\big(\operatorname{Diag}\{\pm e^i : \sigma_i(X) = \sigma_k(X)\}\big)\\ &= \operatorname{conv} O(n,m)^X.\big(\operatorname{Diag}\{\pm e^i : \sigma_i(X) = \sigma_k(X)\}\big)\\ &= \operatorname{conv}\{\pm uv^T : \Sigma_k(X)\}.\end{aligned}$$

The case $\sigma_k(X) > 0$ is analogous, keeping in mind that $\operatorname{sign}\sigma_i(X) = 1$ for all i. The (approximate) subdifferential formula and the condition for regularity also follow easily now. □

6.9 Lidskii's theorem for weak majorization - via nonsmooth analysis

Lidskii's theorem (for weak majorization) states (see [37, 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 $k = 1, 2, ..., n$. 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 chapter. We need an equivalent characterization of weak majorization.

Lemma 6.9.1. *If x and y be any two vectors in $\mathbb{R}^n$, then the following conditions are equivalent:*

1. $|x| \prec_w |y|$;
2. $x \in \mathrm{conv}\,(P_{(-)}(n)y)$;
3. *for every vector w in $\mathbb{R}^n$ we have $w^T x \leq \hat{w}^T \hat{y}$.*

Proof. The equivalence of (1) and (2) is the content of [60, Theorem 1.2]. Suppose

now (2) 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 (3) holds but $x \notin \operatorname{conv}(P_{(-)}(n)y)$, 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

$$f(x) = w^T \hat{x}. \tag{6.29}$$

The function f is clearly Lipschitz. If x has coordinates all nonzero 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 are nonzero with distinct absolute values) has a complement in $\mathbb{R}^n$ with measure zero. On the other hand we have the following theorem (see [15, Theorem 2.5.1]).

Theorem 6.9.2 (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*

$$\partial^c f(x) = \operatorname{conv}\{\lim \nabla f(x_i) \,|\, x_i \to x,\, x_i \notin S\}.$$

(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 (6.29) satisfies

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

We need another theorem, [15, Theorem 2.3.7].

Theorem 6.9.3 (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, then there is a matrix Q in $M_{n,m}$, between the matrices X and $X + Y$, and a matrix T in $\partial^c(w^T\sigma)(Q)$ such that:

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

where the last inequality is the von Neumann's theorem (6.2.9). On the other hand applying formula (6.25) and the above inclusion we get

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

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

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

Lidskii's theorem follows from Lemma 6.9.1.

6.10 Proximal subgradients

In this section we show that the formula in the main result of this chapter also holds in the case of *proximal subgradients*.

Definition 6.10.1 (Proximal Subgradients). *A vector y is called a* **proximal subgradient** *of a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ at x, a point where $f(x)$ is finite, if there exist $\rho > 0$ and $\delta > 0$ such that*

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

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

It is clear from the definition that

$$\partial_p f(x) \subseteq \hat{\partial} f(x). \tag{6.30}$$

Lemma 6.10.2 (Proximal Subgradients Invariance). *If the function $f : E \to [-\infty, +\infty]$ (E is an Euclidean space) is invariant under a subgroup G of $O(E)$, 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 $f(x+z) \ge f(x) + \langle y, z\rangle - \frac{1}{2}\rho\|z\|^2$. 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 $\partial_p f(gx) = g\partial_p f(x)$. □

6.10.1 A preliminary result

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

$$\sigma(X+M) = \sigma(X) + \sigma'(X;M) + O(\|M\|^2). \tag{6.31}$$

Note that this identity is more powerful than the one in Lemma 6.4.9 in the sense that it implies the one in Lemma 6.4.9. First of all from [36, Theorem 4.3.1] we have that

$$\lambda(X+M) = \lambda(X) + O(\|M\|). \tag{6.32}$$

We will use the following notation and results from [87]. If A is $n \times n$ symmetric matrix, its eigenvalues are all real and we can arrange them in nonincreasing order

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

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 (6.32) and Proposition 1.4 in [87].

Proposition 6.10.3. *Let $A \in S(n)$ and $U \in O(n)$ be such that*

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

If we set $U_1 := [u_i, ..., u_j]$ 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. Let the singular value decomposition of X be $X = V^T\big(\operatorname{Diag}\sigma(X)\big)W$. Let

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

It is well known (see [36, Theorem 7.3.7]) that the eigenvalues of the matrix A are $(\sigma_1(X), ..., \sigma_n(X), 0, ..., 0, -\sigma_n(X), ..., -\sigma_1(X))$ with $m - n$ zeros in between. Set $S = \operatorname{Diag}\sigma(X) \in M_{n,n}$ and choose orthogonal U with

$$U^T A U = \operatorname{Diag}(S, 0, -S).$$

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

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

Formula (6.31) now follows.

6.10.2 Proximal subgradients

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

Lemma 6.10.4 (Diagonal Proximal Subgradients). *For any vectors x in $\mathbb{R}^n_\downarrow$, y in $\mathbb{R}^n$ and any singular value function $f \circ \sigma$ we have*

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

Proof. Suppose first that $\operatorname{Diag} y$ 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

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

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

In the opposite direction, let $y \in \partial_p f(x)$. By Lemma 6.10.2, 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).

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

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

$$f(x+z) \geq f(x) + \delta^*_\Lambda(z) - \frac{1}{2}\rho\|z\|^2. \tag{6.33}$$

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

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

Therefore by inequality (6.33), together with the Lipschitzness of δ^*_Λ and σ, we get

$$\begin{aligned} f(\sigma(\operatorname{Diag} x + Z)) &= f(x + (\sigma(\operatorname{Diag} x + Z) - x)) \\ &\geq f(x) - \frac{1}{2}\rho\|\sigma(\operatorname{Diag} x + Z) - x\|^2 \\ &\quad + \delta^*_\Lambda(\sigma'(\operatorname{Diag} x; Z) + [\sigma(\operatorname{Diag} x + Z) - x - \sigma'(\operatorname{Diag} x; Z)]) \\ &\geq f(x) + \delta^*_\Lambda(\sigma'(\operatorname{Diag} x; Z)) - \left(\frac{1}{2}\rho + K\|y\|\right)\|Z\|^2, \end{aligned}$$

Recall that by the Singular Value Derivatives Theorem (6.4.8) we have

$$\operatorname{diag} Z \in \operatorname{conv}\big(P_{(-)}(n)_x \sigma'(\operatorname{Diag} x; Z)\big). \tag{6.34}$$

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

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

for any matrix $P_{(-)}$ in $P_{(-)}(n)_x$. The convexity of δ^*_Λ, its invariance property, and relation (6.34), imply that

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

We continue the chain of inequalities above:

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

so the result follows. □

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

Theorem 6.10.5. *The proximal subdifferential of any singular value function $f \circ \sigma$ at a matrix X in $M_{n,m}$ is given by the formula*

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

where

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

Note 6.10.6. *It is also worth mentioning that much the same argument proves the corresponding result for spectral functions of symmetric matrices - a result left unproven in [52].*

Proof. For any vector y in $\partial_p f(\sigma(X))$, the Diagonal Proximal Subgradients Lemma (6.10.4) shows

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

and now, for any element (U_n, U_m) in $O(n,m)^X$, from the Proximal Subgradients Invariance Lemma (6.10.2) we get

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

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 $\partial_p(f \circ \sigma)(X) \subseteq \hat{\partial}(f \circ \sigma)(X) \subseteq \partial(f \circ \sigma)(X)$, Theorem 6.3.3 shows that $X^T Y = Y^T X$ and $Y^T X = X^T Y$ and then

by Lemma 6.2.7 we get that

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

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 6.10.2 shows that

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

Finally the Diagonal Proximal Subgradients Lemma (6.10.4) gives us

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

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

Chapter 7

Lorentz Invariant Functions

In this final chapter, we derive all the major results from the previous chapters but this time for functions that are invariant under linear orthogonal transformations preserving the Lorentz cone. We call such functions Lorentz invariant. Our motivation for considering such functions originates in [68, Proposition 5.4.3]. Lorentz invariant functions are the composition of a symmetric function on two variables and the eigenvalues of the hyperbolic polynomial $p(x) = x_0^2 - x_1^2 - \cdots - x_n^2$. There are clear similarities between Lorentz invariant functions, eigenvalue functions and singular value functions, which suggest that there is a broader framework (see Chapter 8 for possible ideas) capturing all these examples.

7.1 Notation

We are going to denote the set of all orthogonal $n \times n$ matrices by $O(n)$. Let the function $g(x,t)$ be defined on an open subset of $\mathbb{R}^n \times \mathbb{R}$, taking values in $\mathbb{R}$.

Let the function $f(a,b)$ be defined on an open subset of $\mathbb{R}^2$. We will think of all n-dimensional vectors as column vectors, and the inner product of two $(n+1)$-dimensional vectors, (x,t) and (y,r) will naturally be

$$\langle (x,t),(y,r)\rangle = x^T y + tr.$$

Throughout the entire chapter we assume that

$$g(Ux,t) = g(x,t), \quad \text{for all } U \in O(n), \tag{7.1}$$

and

$$f(a,b) = f(b,a). \tag{7.2}$$

We call a function g with property (7.1) *Lorentz invariant* because it is invariant under the linear orthogonal transformations preserving the *Lorentz cone* $\{(x,t) \in \mathbb{R}^n \times \mathbb{R} | t \geq \|x\|\}$. Functions f with property (7.2) are called *symmetric*. Clearly the domain of f must be a *symmetric subset* of $\mathbb{R}^2$. (A subset A of $\mathbb{R}^2$ is symmetric if $(a,b) \in A \Rightarrow (b,a) \in A$.) We also define the following function

$$\beta(x,t) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^2$$
$$\beta(x,t) = \frac{1}{\sqrt{2}}(t + \|x\|, t - \|x\|).$$

The following lemma is easily established.

Lemma 7.1.1. *(***Lorentz Invariant Functions***) The next two properties of a*

function $g : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ *are equivalent:*

1. g *is Lorentz invariant.*

2. $g = f \circ \beta$ *for some symmetric function* $f : \mathbb{R}^2 \to \mathbb{R}$.

If $g = f \circ \beta$ we say that f is the symmetric function corresponding to g. It is easily seen that the correspondence $g \leftrightarrow f$ is one-to-one and in order to extract the corresponding symmetric function f, given g, we set

$$f(a,b) = g\left(\frac{a-b}{\sqrt{2}}, 0, ..., 0, \frac{a+b}{\sqrt{2}}\right). \tag{7.3}$$

Property (7.1) assures us that $f(a,b)$ is symmetric.

The aim of this chapter is to establish how a variety of properties of the function f are transfered to the function g and vice versa. Every one of the following sections deals with one particular property. We conclude this section with another elementary fact.

Lemma 7.1.2. *If* f *is lower semicontinuous then so is* $f \circ \beta$.

7.2 Fenchel conjugation

For a function $F : \mathbb{R}^n \to (-\infty, +\infty]$, the *Fenchel conjugate* $F^* : \mathbb{R}^n \to [-\infty, +\infty]$ is the function

$$F^*(y) = \sup_{x \in \mathbb{R}^n} \{x^T y - F(x)\}.$$

It is well known that F^* is lower semicontinuous and convex (see [78]). In this section we prove the following formula.

Theorem 7.2.1. *Let $f : \mathbb{R}^2 \to (-\infty, +\infty]$ be a symmetric function. Then*

$$(f \circ \beta)^* = f^* \circ \beta. \tag{7.4}$$

Proof. Let $y \neq 0$. From the definition we have

$$\begin{aligned}
(f \circ \beta)^*(y,r) &= \sup_{(x,t)\in\mathbb{R}^{n+1}} \{\langle (y,r),(x,t)\rangle - (f\circ\beta)(x,t)\} \\
&= \sup_{(a,b)\in\mathbb{R}^2} \sup_{\substack{(x,\,t)\ \text{s.t.} \\ t+\|x\| = \sqrt{2}a \\ t-\|x\| = \sqrt{2}b}} \{\langle (y,r),(x,t)\rangle - f(a,b)\} \\
&= \sup_{(a,b)\in\mathbb{R}^2} \left\{ \left\langle (y,r), \left(\frac{y}{\|y\|}\frac{a+b}{\sqrt{2}}, \frac{a-b}{\sqrt{2}} \right) \right\rangle - f(a,b) \right\} \\
&= \sup_{(a,b)\in\mathbb{R}^2} \left\{ \|y\|\frac{a+b}{\sqrt{2}} + r\frac{a-b}{\sqrt{2}} - f(a,b) \right\} \\
&= \sup_{(a,b)\in\mathbb{R}^2} \left\{ \left\langle \left(\frac{r+\|y\|}{\sqrt{2}}, \frac{r-\|y\|}{\sqrt{2}} \right), (a,b) \right\rangle - f(a,b) \right\} \\
&= (f^* \circ \beta)(y,r).
\end{aligned}$$

The case when $y = 0$ is clear. □

An alternative proof of this theorem uses Theorem 5.5 and the example in Section 7.5 in [6]. One may also deduce the above result from Corollary 2.5.4 and Example 2.6.5.

7.3 Convexity

The following theorem is the main result of this section.

Theorem 7.3.1. *Let $f : \mathbb{R}^2 \to (-\infty, +\infty]$ be symmetric, convex and lower semicontinuous, and $f \circ \beta$ be its corresponding Lorentz function. Then $f \circ \beta$ is convex and lower semicontinuous.*

Proof. If $f \equiv +\infty$ then $f \circ \beta \equiv +\infty$ and the theorem is clear. Suppose f assumes some finite values. Then since $f > -\infty$ we have that $f^{**} = f$ (see [78, Theorem 12.2]). Also since f^* is symmetric because of [78, Corollary 12..3.1], using (7.4), we have

$$f \circ \beta = f^{**} \circ \beta = (f^* \circ \beta)^*.$$

Consequently $f \circ \beta$ is the conjugate of the function $f^* \circ \beta$, so it is convex and lower semicontinuous. □

Note 7.3.2. *The proof of above theorem can be also deduced from Theorem 3.9 and the example in Section 7.5 in [6]. An alternative way would be using Theorem 2.3.9 and Example 2.6.5.*

7.4 Convex subdifferentials

Let $f : \mathbb{R}^2 \to (-\infty, +\infty]$. For every point (a, b) such that $f(a, b) < +\infty$ we define the *subdifferential* of f at (a, b),

$$\partial f(a, b) = \{(a', b') \in \mathbb{R}^2 | f(a, b) + f^*(a, b) = \langle (a, b), (a', b') \rangle\}.$$

The set $\partial f(a, b)$ is a singleton $\{(a', b')\}$ if and only if f is differentiable at the point (a, b) with gradient $\nabla f(a, b) = (a', b')$ (see [78, Theorem 25.1]). If f is convex then

also

$$\partial f(a,b) = \{v | f(c,d) - f(a,b) \geq \langle v, (c,d) - (a,b)\rangle, \forall (c,d)\}.$$

The following result gives a formula for the subgradient of the composition $f \circ \beta$.

Theorem 7.4.1. *Suppose $f : \mathbb{R}^2 \to (-\infty, +\infty]$ is symmetric, convex, and lower semicontinuous. Then $(y,r) \in \partial(f \circ \beta)(x,t)$ if and only if $\beta(y,r) \in \partial f(\beta(x,t))$ and $x^T y = \|x\|\|y\|$.*

Proof. Suppose first $(y,r) \in \partial(f \circ \beta)(x,t)$. Then using formula (7.4) we get

$$\begin{aligned}
x^T y + rt &= \langle (y,r),(x,t)\rangle \\
&= (f \circ \beta)(x,t) + (f \circ \beta)^*(y,r) \\
&= (f \circ \beta)(x,t) + (f^* \circ \beta)(y,r) \\
&= f\left(\frac{t+\|x\|}{\sqrt{2}}, \frac{t-\|x\|}{\sqrt{2}}\right) + f^*\left(\frac{r+\|y\|}{\sqrt{2}}, \frac{r-\|y\|}{\sqrt{2}}\right) \\
&\geq \frac{1}{2}\big((t+\|x\|)(r+\|y\|) + (t-\|x\|)(r-\|y\|)\big) \\
&= rt + \|x\|\|y\|.
\end{aligned}$$

So we must have equality above. This means two things: (a) $\beta(y,r) \in \partial f(\beta(x,t))$ and (b) $x^T y = \|x\|\|y\|$. In the other direction the proof is clear by reversing the steps above. □

The above result is also a particular case of Theorem 2.5.5 applied to Example 2.6.5.

7.5 Differentiability

In this section we prove that f is differentiable if and only if $f \circ \beta$ is.

Theorem 7.5.1. *Let f be symmetric and defined on an open symmetric subset of $\mathbb{R}^2$. Then f is differentiable at the point $\beta(x,t)$ if and only if $f \circ \beta$ is differentiable at (x,t). In that case we have the formulae*

$$\nabla_x (f \circ \beta)(x,t) = \begin{cases} \frac{f_1'(\beta(x,t)) - f_2'(\beta(x,t))}{\sqrt{2}\|x\|} x, & \text{if } x \neq 0 \\ 0, & \text{if } x = 0, \end{cases}$$

and

$$\frac{d}{dt}(f \circ \beta)(x,t) = \frac{1}{\sqrt{2}}(f_1'(\beta(x,t)) + f_2'(\beta(x,t))).$$

Proof. Suppose first that f is differentiable at the point $\beta(x,t)$. If $x \neq 0$ the theorem and the formulae are trivial and follow from the chain rule. So let us assume now that $x = 0$. Let $h = (h_1, h_2) \in \mathbb{R}^n \times \mathbb{R}$ and

$$d := \left(0, ..., 0, \frac{1}{\sqrt{2}}\big(f_1'(t/\sqrt{2}, t/\sqrt{2}) + f_2'(t/\sqrt{2}, t/\sqrt{2})\big)\right) \in \mathbb{R}^n \times \mathbb{R}.$$

Then

$$\lim_{h \to 0} \frac{|(f \circ \beta)((0,t) + (h_1,h_2)) - (f \circ \beta)((0,t)) - d^T h|}{\|h\|} =$$
$$\lim_{h \to 0} \frac{|f(\beta(h_1, t + h_2)) - f(\beta(0,t)) - h_2(f_1'(\beta(0,t)) + f_2'(\beta(0,t)))/\sqrt{2}|}{\|h\|}.$$

The fact that f is differentiable at $\beta(0,t) = (t/\sqrt{2}, t/\sqrt{2})$ gives us

$$f(\beta(h_1, t+h_2)) = f(\beta(0,t)) + f_1'(\beta(0,t))\frac{h_2 + \|h_1\|}{\sqrt{2}} + f_2'(\beta(0,t))\frac{h_2 - \|h_1\|}{\sqrt{2}} + o(\|h\|).$$

Using the fact that for a symmetric function f, $f_1'(\beta(0,t)) = f_2'(\beta(0,t))$ and substituting above we see that the limit is zero, that is, $\nabla(f \circ \beta)(0,t) = d$.

The proof in the other direction is easy using formula (7.3). □

7.6 Continuity of the gradient

Theorem 7.6.1. *Let f be symmetric and defined on an open symmetric subset of $\mathbb{R}^2$. Then $f \circ \beta$ is continuously differentiable at the point (x,t) if and only if f is continuously differentiable at $\beta(x,t)$.*

Proof. Suppose that f is continuously differentiable at $\beta(x,t)$. The theorem is clear if $x \neq 0$. So suppose $x = 0$. Let $\{(x_n, t_n)\}$ be a sequence of points approaching $(0,t)$. We need only prove that $\nabla(f \circ \beta)(x_n, t_n)$ approaches $\nabla(f \circ \beta)(0,t)$. We consider two particular cases. The general case easily follows by combining these two cases.

Case 1. If $x_n = 0$ for all n. Then using the formula in Theorem 7.5.1 we obtain

$$\begin{aligned}\lim_{n\to\infty} \nabla(f \circ \beta)(0,t_n) &= \lim_{n\to\infty} \left(0, ..., 0, \frac{1}{\sqrt{2}}\big(f_1'(\beta(0,t_n)) + f_2'(\beta(0,t_n))\big)\right) \\ &= \left(0, ..., 0, \frac{1}{\sqrt{2}}\big(f_1'(\beta(0,t)) + f_2'(\beta(0,t))\big)\right)\end{aligned}$$

$$= \nabla(f \circ \beta)(0,t),$$

by the continuity of ∇f at $\beta(0,t)$.

Case 2. If $x_n \neq 0$ for all n. Using again the formula in Theorem 7.5.1 for the derivative with respect to t we obtain

$$\begin{aligned}\lim_{n\to\infty}(f\circ\beta)'_t(x_n,t_n) &= \lim_{n\to\infty}\frac{1}{\sqrt{2}}\big(f'_1(\beta(x_n,t_n))+f'_2(\beta(x_n,t_n))\big)\\ &= \frac{1}{\sqrt{2}}\big(f'_1(\beta(0,t))+f'_2(\beta(0,t))\big)\\ &= (f\circ\beta)'_t(0,t).\end{aligned}$$

Now, for the derivative with respect to x^i we get

$$\lim_{n\to\infty}(f\circ\beta)'_{x^i}(x_n,t_n) = \lim_{n\to\infty}\frac{x_n^i}{\sqrt{2}\|x_n\|}\big(f'_1(\beta(x_n,t_n))-f'_2(\beta(x_n,t_n))\big) = 0,$$

because $x_n^i/\|x_n\|$ is bounded, and the continuity of ∇f at $\beta(0,t)$ gives us

$$\lim_{n\to\infty}(f'_1(\beta(x_n,t_n))-f'_2(\beta(x_n,t_n)) = f'_1(\beta(0,t))-f'_2(\beta(0,t)) = 0.$$

(The last equality follows from the fact that f is symmetric.)

The opposite direction of the theorem is easy. □

7.7 The "decomposition" functions

This is a supplementary section in which we define the functions d_z and d_z^* and summarize some of their properties which we will use frequently. We call them

decomposition functions because they describe how the subgradients of $f \circ \beta$ are composed from (or decomposed into) subgradients of f.

Definition 7.7.1. *For every nonzero vector z in $\mathbb{R}^n$ we define the map*

$$d_z : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^2,$$
$$d_z(y,t) = \frac{1}{\sqrt{2}}\left(t + \frac{z^T y}{\|z\|}, t - \frac{z^T y}{\|z\|}\right).$$

In cases when the direction (y,t) is fixed and clear from the context we will denote $d_z(y,t)$ for short by d_z.

Definition 7.7.2. *For every nonzero vector z in $\mathbb{R}^n$ we define the map*

$$d_z^* : (\mathbb{R}^n \setminus \{0\}) \times \mathbb{R}^2 \to \mathbb{R}^n \times \mathbb{R},$$
$$d_z^*(a,b) = \left(\frac{z}{\|z\|}\frac{a-b}{\sqrt{2}}, \frac{a+b}{\sqrt{2}}\right)$$

The following lemma gives some properties of the maps d_z and d_z^* that we will use.

Lemma 7.7.3. *Let z and w be nonzero vectors in $\mathbb{R}^n$.*

1. *The maps $d_z(\cdot)$ and $d_z^*(\cdot)$ are linear and conjugate to each other.*

2. *For every point (γ_1, γ_2) in $\mathbb{R}^2$*

$$d_w d_z^*(\gamma_1, \gamma_2) = \frac{1+\delta}{2}(\gamma_1, \gamma_2) + \frac{1-\delta}{2}(\gamma_2, \gamma_1),$$

where $\delta = \frac{w^T z}{\|w\|\|z\|} \in [-1,1]$. *In particular when* $w = z$ *we have*

$$d_z d_z^*(\gamma_1, \gamma_2) = (\gamma_1, \gamma_2).$$

3. *For every point* (y, r) *in* $\mathbb{R}^n \times \mathbb{R}$ *such that* $y = az$ *for some* $a \in \mathbb{R}$

$$d_z^* d_z(y, r) = (y, r).$$

Proof. Let z be a nonzero vector in $\mathbb{R}^n$, $(y, r) \in \mathbb{R}^n \times \mathbb{R}$, and $(a, b) \in \mathbb{R}^2$. Then

$$\begin{aligned}\langle d_z(y,r),(a,b)\rangle &= \left\langle \frac{1}{\sqrt{2}}\left(r + \frac{z^T y}{\|z\|}, r - \frac{z^T y}{\|z\|}\right),(a,b)\right\rangle \\ &= \frac{a+b}{\sqrt{2}} r + \frac{a-b}{\sqrt{2}} \frac{z^T y}{\|z\|} \\ &= \langle (y,r), d_z^*(a,b)\rangle.\end{aligned}$$

The second and the third part are easy. □

Lemma 7.7.4. *Let* A *and* B *be symmetric subsets of* $\mathbb{R}^2$. *The sets*

$$\mathcal{D}(A) = \{d_z^*(\gamma_1, \gamma_2) | (\gamma_1, \gamma_2) \in A,\, 0z \neq 0\},$$
$$\mathcal{C}(A) = \{(y, r) | d_z(y, r) \in A,\, \forall z \neq 0\},$$

satisfy the following properties.

1. *If* A *is convex then*

   (a) *If* (x, t) *is in* $\mathcal{D}$, *then* $(\delta x, t)$ *is in* $\mathcal{D}$ *for every* $\delta \in [-1, 1]$.

(b) $\mathcal{D}$ *is a convex set.*

(c) $\mathcal{D} = \mathcal{C}$.

(d) If B *is also convex, then* $\operatorname{cl}(\mathcal{D}(A) + \mathcal{D}(B)) = \operatorname{cl}\mathcal{D}(A + B)$.

2. For any A *we have*

(a) $\operatorname{conv}\mathcal{D}(A) = \mathcal{D}(\operatorname{conv} A)$.

(b) $\mathcal{D}(\operatorname{cl} A) = \operatorname{cl}\mathcal{D}(A)$.

Proof. **Part 1a**. Let $(x,t) = d_z^*(\gamma_1,\gamma_2)$ for some (γ_1,γ_2) in A, and $z \neq 0$. Because the set A is symmetric, (γ_2,γ_1) is in A. Because A is convex, we get that for every $\alpha \in [0,1]$ the vector $(\alpha\gamma_1 + (1-\alpha)\gamma_2, \alpha\gamma_2 + (1-\alpha)\gamma_1)$ is in A. Thus

$$\begin{aligned} d_z^*\big(\alpha\gamma_1 + (1-\alpha)\gamma_2, \alpha\gamma_2 + (1-\alpha)\gamma_1\big) &= \left(\frac{z}{\|z\|}\frac{\gamma_1-\gamma_2}{\sqrt{2}}(2\alpha-1), \frac{\gamma_1+\gamma_2}{\sqrt{2}}\right) \\ &= (x(2\alpha-1), t) \in \mathcal{D}, \end{aligned}$$

for all $\alpha \in [0,1]$. We now have to set $\delta := 2\alpha - 1$ for $\alpha \in [0,1]$.

Part 1b. Notice that for any two points (γ_1,γ_2) and (δ_1,δ_2) in A and $\mu \in [0,1]$, we have $(\mu\gamma_1 + (1-\mu)\delta_1, \mu\gamma_2 + (1-\mu)\delta_2)$ is in A and so for every $z \neq 0$

$$\left(\frac{z}{\|z\|}\frac{\mu(\gamma_1-\gamma_2) + (1-\mu)(\delta_1-\delta_2)}{\sqrt{2}}, \frac{\mu(\gamma_1+\gamma_2) + (1-\mu)(\delta_1+\delta_2)}{\sqrt{2}}\right) \in \mathcal{D}. \qquad (7.5)$$

Take two points, (x_1,t_1) and (x_2,t_2) in $\mathcal{D}$, and a number $\mu \in (0,1)$. We want to show that $(\mu x_1 + (1-\mu)x_2, \mu t_1 + (1-\mu)t_2)$ is also in $\mathcal{D}$. Suppose

$$(x_1,t_1) = d_{z_1}^*(\gamma_1,\gamma_2), \qquad (x_2,t_2) = d_{z_2}^*(\delta_1,\delta_2)$$

for some (γ_1, γ_2) and (δ_1, δ_2) in A, $z_1 \neq 0$ and $z_2 \neq 0$. Set

$$z_\mu := \mu \frac{\gamma_1 - \gamma_2}{\sqrt{2}} \frac{z_1}{\|z_1\|} + (1-\mu) \frac{\delta_1 - \delta_2}{\sqrt{2}} \frac{z_2}{\|z_2\|},$$

and notice that

$$\|z_\mu\| \leq \mu \frac{|\gamma_1 - \gamma_2|}{\sqrt{2}} + (1-\mu) \frac{|\delta_1 - \delta_2|}{\sqrt{2}}.$$

Then

$$\mu(x_1, t_1) + (1-\mu)(x_2, t_2) = \left(z_\mu, \frac{\mu(\gamma_1 + \gamma_2) + (1-\mu)(\delta_1 + \delta_2)}{\sqrt{2}} \right).$$

If $z_\mu = 0$ then from (7.5) and part 1a with $\delta = 0$ we see that

$$\mu(x_1, t_1) + (1-\mu)(x_2, t_2) \in \mathcal{D}.$$

Suppose now $z_\mu \neq 0$. Choose one of the points (γ_1, γ_2), (γ_2, γ_1) in A, and one of the points (δ_1, δ_2), (δ_2, δ_1) in A so that inclusion (7.5) and part 1a now say that for all $z \neq 0$ and $\delta \in (0, 1)$

$$\left(\frac{z}{\|z\|} \frac{\mu|\gamma_1 - \gamma_2| + (1-\mu)|\delta_1 - \delta_2|}{\sqrt{2}} \delta, \frac{\mu(\gamma_1 + \gamma_2) + (1-\mu)(\delta_1 + \delta_2)}{\sqrt{2}} \right) \in \mathcal{D}.$$

Let δ now be a number in $(0, 1)$ such that

$$\frac{\mu|\gamma_1 - \gamma_2| + (1-\mu)|\delta_1 - \delta_2|}{\sqrt{2}} \delta = \|z_\mu\|.$$

Putting it all together we obtain

$$\mu(x_1,t_1)+(1-\mu)(x_2,t_2)=\left(\frac{z_\mu}{\|z_\mu\|}\|z_\mu\|,\frac{\mu(\gamma_1+\gamma_2)+(1-\mu)(\delta_1+\delta_2)}{\sqrt{2}}\right)\in\mathcal{D}.$$

This shows that $\mathcal{D}$ is a convex set.

Part 1c. Suppose $(y,r)\in\mathcal{C}$. If $y=0$ then clearly $(y,r)\in\mathcal{D}$. If $y\neq 0$, set $(\gamma_1,\gamma_2):=d_y(y,r)\in A$. Then by Lemma 7.7.3 part 3 $(y,r)=d_y^*d_y(y,r)=d_y^*(\gamma_1,\gamma_2)$. So $\mathcal{C}\subseteq\mathcal{D}$.

Suppose now $(y,r)\in\mathcal{D}$. That is $(y,r)=d_z^*(\gamma_1,\gamma_2)$ for some (γ_1,γ_2) in A and some $z\neq 0$. Let $\hat{z}$ be an arbitrary nonzero vector and set $\delta:=\frac{z^T\hat{z}}{\|z\|\|\hat{z}\|}\in[-1,1]$. Then by Lemma 7.7.3 part 2 we have

$$d_{\hat{z}}(y,r)=d_{\hat{z}}d_z^*(\gamma_1,\gamma_2)=\frac{1+\delta}{2}(\gamma_1,\gamma_2)+\frac{1-\delta}{2}(\gamma_2,\gamma_1)\in A,$$

because A is symmetric and convex. So $\mathcal{D}\subseteq\mathcal{C}$.

Part 1d. By part 1b we have that both $\mathcal{D}(A)+\mathcal{D}(B)$ and $\mathcal{D}(A+B)$ are convex sets. It is clear that the latter set is contained in the former. So

$$\mathrm{cl}\,(\mathcal{D}(A)+\mathcal{D}(B))\supseteq\mathrm{cl}\,\mathcal{D}(A+B).$$

Then, in order to show that they are equal it suffices to show that the support function of the first set is not larger than the support function of the second set.

$$\max_{z_1,z_2\neq 0}\{\langle(x,t),(d_{z_1}^*(\gamma_1,\gamma_2)+d_{z_2}^*(\delta_1,\delta_2))\rangle|(\gamma_1,\gamma_2)\in A,\,(\delta_1,\delta_2)\in B\}$$
$$=\max\{\langle(x,t),(d_x^*(\gamma_1,\gamma_2)+d_x^*(\delta_1,\delta_2))\rangle|(\gamma_1,\gamma_2)\in A,\,(\delta_1,\delta_2)\in B\}$$

$$= \max\{\langle (x,t), d^*_x(\gamma_1+\delta_1, \gamma_2+\delta_2)\rangle | (\gamma_1,\gamma_2)\in A,\ (\delta_1,\delta_2)\in B\}$$

$$= \max\{\langle (x,t), d^*_x(\gamma_1,\gamma_2)\rangle | (\gamma_1,\gamma_2)\in A+B\}.$$

(Actually the two support functions are equal.)

Part 2a. If $(x,t)\in \operatorname{conv}\mathcal{D}(A)$, then there exist points (γ_1^i,γ_2^i) in A, nonzero vectors $z_i\in\mathbb{R}^n$, and nonnegative numbers α_i, $i=1,...,k$, satisfying $\sum_{i=1}^k \alpha_i = 1$ such that

$$(x,t) = \alpha_1 d^*_{z_1}(\gamma_1^1,\gamma_2^1) + \cdots + \alpha_k d^*_{z_k}(\gamma_1^k,\gamma_2^k).$$

Let z be an arbitrary nonzero vector in $\mathbb{R}^n$. Define $\delta_{iz} := \frac{z_i^T z}{\|z_i\|\|z\|}\in[-1,1]$. Then by Lemma 7.7.3 part 1 and part 2 we get

$$d_z(x,t) = \alpha_1\frac{1+\delta_{1z}}{2}(\gamma_1^1,\gamma_2^1) + \alpha_1\frac{1-\delta_{1z}}{2}(\gamma_2^1,\gamma_1^1) + \cdots$$
$$\cdots + \alpha_k\frac{1+\delta_{kz}}{2}(\gamma_1^k,\gamma_2^k) + \alpha_k\frac{1-\delta_{kz}}{2}(\gamma_2^k,\gamma_1^k).$$

Consequently $d_z(x,t)\in\operatorname{conv}A$ for every $z\neq 0$. So $(x,t)\in\mathcal{C}(\operatorname{conv}A) = \mathcal{D}(\operatorname{conv}A)$, by part 1c. The opposite inclusion $\mathcal{D}(\operatorname{conv}A)\subseteq \operatorname{conv}\mathcal{D}(A)$ is easy.

Part 2b. Let $\{d_{x_r}(\gamma_1^r,\gamma_2^r)\}$ be a sequence in $\mathcal{D}(A)$ approaching a vector (z,s). Since the unit sphere in $\mathbb{R}^n$ is compact, we can find a subsequence r' such that $x_{r'}/\|x_{r'}\|$ converges to a unit vector x. For this subsequence we have $|\gamma_1^r-\gamma_2^r|\to\sqrt{2}\|z\|$ and $\gamma_1^r+\gamma_2^r\to\sqrt{2}s$. Consequently $\{(\gamma_1^r,\gamma_2^r)\}$ is bounded so there is a subsequence r'' for which $\{(\gamma_1^{r''},\gamma_2^{r''})\}\to(\gamma_1,\gamma_2)\in\operatorname{cl}A$. So $\{d_{x_{r''}}(\gamma_1^{r''},\gamma_2^{r''})\}$ approaches $\{d_x(\gamma_1,\gamma_2)\}$ which is in $\mathcal{D}(\operatorname{cl}A)$. This shows that for an arbitrary set A we have the inclusion $\mathcal{D}(\operatorname{cl}A)\supseteq\operatorname{cl}\mathcal{D}(A)$. The opposite inclusion is easy. □

7.8 Clarke directional derivative & subdifferential - the Lipschitz case

Suppose in this section that the function f is *Lipschitz* near x, that is, there exists a scalar K such that the following holds

$$|f(x'') - f(x')| \leq K\|x'' - x'\|, \quad \text{for all } x'', x' \text{ close to } x.$$

For Lipschitz functions the *Clarke directional derivative* [15] is defined for a direction v at the point x to be

$$f^\circ(x; v) = \limsup_{y \to x;\, \lambda \downarrow 0} \frac{f(y + \lambda v) - f(y)}{\lambda}.$$

The difference quotient above, for y close to x and λ to 0, is bounded above by $K|v|$, so $f^\circ(x; v)$ is well defined and finite.

A property of the Clarke directional derivative that we will need and may be found in [15, p. 64] is that for every pair $(x; v)$

$$f^\circ(x; v) = \limsup_{y \to x} \{\langle \nabla f(y), v \rangle | y \text{ is s.t. } \nabla f(y) \text{ exists}\}.$$

In other words, there exists a sequence $\{x_n\}$ approaching x such that f is differentiable at each x_n and

$$\langle \nabla f(x_n), v \rangle \to f^\circ(x; v). \tag{7.6}$$

The *Clarke subdifferential* $\partial^c f(x)$ is defined as follows

$$\partial^c f(x) = \{\xi | \langle v, \xi \rangle \leq f^\circ(x; v) \text{ for all } v\}.$$

The set $\partial^c f(x)$ is compact, nonempty and convex. If f is convex and finite on a neighbourhood of x then $\partial^c f(x) = \partial f(x)$, and if f is continuously differentiable at x then $\partial^c f(x) = \{\nabla f(x)\}$. In this sense the Clarke generalized gradient unifies these two properties.

Now let us return to our symmetric, bivariable function f, which we now require to be Lipschitz. We are going to find a formula expressing the Clarke subdifferential of $f \circ \beta$ in terms of the Clarke subdifferential of f.

The following lemma is elementary and shows how the Clarke directional derivative of $f \circ \beta$ changes under Lorentz orthogonal transformations of the argument and the direction.

Lemma 7.8.1. *Let (x,t) be a point in the domain of $f \circ \beta$, (y,r) be a direction, and U be a orthogonal matrix. Then*

$$(f \circ \beta)^\circ((x,t);(y,r)) = (f \circ \beta)^\circ((Ux,t);(Uy,r)).$$

Proof.

$$\begin{aligned}
(f \circ \beta)^\circ((x,t);(y,r)) &= \limsup_{(z,s)\to(x,t);\,\lambda\downarrow 0} \frac{f(\beta((z,s)+\lambda(y,r))) - f(\beta(z,s))}{\lambda} \\
&= \limsup_{(z,s)\to(x,t);\,\lambda\downarrow 0} \frac{f(\beta((Uz,s)+\lambda(Uy,r))) - f(\beta(Uz,s))}{\lambda} \\
&= (f \circ \beta)^\circ((Ux,t);(Uy,r)).
\end{aligned}$$

□

Theorem 7.8.2 (Clarke Directional Derivative). *Let (x,t) be a point in the domain of $f \circ \beta$, (y,r) be a direction. Then if $x = 0$*

$$(f \circ \beta)^\circ((0,t);(y,r)) = \max\{f^\circ(\beta(0,t); d_z(y,r)) | 0 \neq z \in \mathbb{R}^n\}. \tag{7.7}$$

Note 7.8.3. *For the case when $x \neq 0$ see Corollary 7.8.6.*

Proof. We have that there is a sequence of points $\{(x_n,t_n)\}$ approaching $(0,t)$ such that

$$(f \circ \beta)^\circ((x,t);(y,r)) = \lim_{n\to\infty} \langle \nabla(f \circ \beta)(x_n,t_n),(y,r)\rangle.$$

In order to evaluate $\nabla(f \circ \beta)$ using Theorem 7.5.1 we need to know whether x_n is zero or not. That is why we consider two subcases and the general situation follows easily from them.

Subcase 1.a Suppose $x_n = 0$ for all n. Denote

$$\beta_n := \beta(0,t_n).$$

Recall that $f_1'(\beta_n) = f_2'(\beta_n)$. Fix an arbitrary vector $0 \neq z \in \mathbb{R}^n$. Then we have

$$\begin{aligned}&(f \circ \beta)^\circ((0,t);(y,r))\\ &\quad = \lim_{n\to\infty} \langle \nabla(f \circ \beta)(x_n,t_n),(y,r)\rangle\end{aligned}$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \left\langle \left(0, \frac{f_1'(\beta_n) + f_2'(\beta_n)}{\sqrt{2}}\right), (y, r)\right\rangle \\
&= \lim_{n\to\infty} \langle \nabla f(\beta_n), \beta(0, r)\rangle \\
&= \lim_{n\to\infty} \left\langle \nabla f(\beta_n), \beta(0, r) + \left(\frac{z^T y}{\sqrt{2}\|z\|}, -\frac{z^T y}{\sqrt{2}\|z\|}\right)\right\rangle \\
&\le f^\circ(\beta(0, t); d_z(y, r))).
\end{aligned}
$$

Subcase 1.b Suppose $x_n \neq 0$ for all n and $\lim_{n\to\infty} x_n/\|x_n\| = z/\|z\|$. Set

$$\beta_n := \beta(x_n, t_n).$$

Then, we have

$$
\begin{aligned}
&(f \circ \beta)^\circ((0, t);(y, r)) \\
&\qquad = \lim_{n\to\infty} \langle \nabla(f \circ \beta)(x_n, t_n), (y, r)\rangle \\
&\qquad = \lim_{n\to\infty} \left\langle \left(\frac{f_1'(\beta_n) - f_2'(\beta_n)}{\sqrt{2}\|x_n\|} x_n, \frac{f_1'(\beta_n) + f_2'(\beta_n)}{\sqrt{2}}\right), (y, r)\right\rangle \\
&\qquad = \lim_{n\to\infty} \frac{f_1'(\beta_n) - f_2'(\beta_n)}{\sqrt{2}\|x_n\|} x_n^T y + \frac{f_1'(\beta_n) + f_2'(\beta_n)}{\sqrt{2}} r \\
&\qquad = \lim_{n\to\infty} f_1'(\beta_n)\left(\frac{r}{\sqrt{2}} + \frac{x_n^T y}{\sqrt{2}\|x_n\|}\right) + f_2'(\beta_n)\left(\frac{r}{\sqrt{2}} - \frac{x_n^T y}{\sqrt{2}\|x_n\|}\right) \\
&\qquad = \lim_{n\to\infty} \left\langle \nabla f'(\beta_n), \left(\frac{r}{\sqrt{2}} + \frac{z^T y}{\sqrt{2}\|z\|}, \frac{r}{\sqrt{2}} - \frac{z^T y}{\sqrt{2}\|z\|}\right)\right\rangle \\
&\qquad \le f^\circ(\beta(0, t); d_z(y, r)).
\end{aligned}
$$

All this shows that if $x = 0$ then

$$(f \circ \beta)^\circ((0, t); (y, r)) \le \sup\{f^\circ(\beta(0, t); d_z(y, r) | 0 \neq z \in \mathbb{R}^n\}.$$

To show the opposite inequality, fix a nonzero vector $z \in \mathbb{R}^n$. There is a sequence of points $\{(a_n, b_n)\}$ approaching $\beta(0,t)$ such that

$$f^\circ(\beta(0,t); d_z(y,r)) = \lim_{n\to\infty} \langle \nabla f(a_n, b_n), d_z(y,r)\rangle.$$

There is an infinite subsequence of $\{(a_n, b_n)\}$ that satisfies one of the three possibilities

1. $a_{n'} = b_{n'}$ for all n'.
2. $a_{n'} > b_{n'}$ for all n'.
3. $a_{n'} < b_{n'}$ for all n'.

For this subsequence we still have

$$f^\circ(\beta(0,t); d_z) = \lim_{n'\to\infty} \langle \nabla f(a_{n'}, b_{n'}), d_z\rangle.$$

So without loss of generality we may assume that $\{(a_n, b_n)\}$ satisfies one of the three possibilities and we consider three separate cases.

Subcase 2.a Suppose $a_n = b_n$ for all n. Recall that in this case we have $f_1'(a_n, a_n) = f_2'(a_n, a_n)$. So

$$\begin{aligned} f^\circ(\beta(0,t); d_z) &= \lim_{n\to\infty} \langle \nabla f(a_n, a_n), d_z(y,r)\rangle \\ &= \lim_{n\to\infty} \frac{f_1'(a_n, a_n) + f_2'(a_n, a_n)}{\sqrt{2}} r \\ &= \lim_{n\to\infty} \langle \nabla (f\circ\beta)(0, a_n), (y,r)\rangle \\ &\le (f\circ\beta)^\circ((0,t);(y,r)). \end{aligned}$$

Subcase 2.b Suppose $a_n > b_n$ for all n. Define the sequence of vectors in $\mathbb{R}^n$:

$$z_n := \left(\frac{a_n - b_n}{2}, 0, ..., 0\right),$$

(notice that $\|z_n\| = (a_n - b_n)/2$) and let U be an orthogonal matrix such that

$$\lim_{n\to\infty} \frac{Uz_n}{\|z_n\|} = \frac{z}{\|z\|}. \tag{7.8}$$

Then

$$\begin{aligned}
f^\circ(\beta(0,t); d_z(y,r)) &= \lim_{n\to\infty} \langle \nabla f(a_n, b_n), d_z \rangle \\
&= \lim_{n\to\infty} \left\langle \frac{f_1'(a_n,b_n) - f_2'(a_n,b_n)}{\sqrt{2}\|z\|} z, y \right\rangle + \frac{f_1'(a_n,b_n) + f_2'(a_n,b_n)}{\sqrt{2}} r. \\
&= \lim_{n\to\infty} \left\langle \frac{f_1'(a_n,b_n) - f_2'(a_n,b_n)}{\sqrt{2}\|z_n\|} z_n, U^T y \right\rangle + \frac{f_1'(a_n,b_n) + f_2'(a_n,b_n)}{\sqrt{2}} r \\
&= \lim_{n\to\infty} \left\langle \nabla(f\circ\beta)\left(z_n, \frac{a_n+b_n}{2}\right), (U^T y, r) \right\rangle \\
&\le (f\circ\beta)^\circ((0,t); (U^T y, r)) \\
&= (f\circ\beta)^\circ((0,t); (y,r)),
\end{aligned}$$

where in the last equality we used Lemma 7.8.1.

Subcase 2.c Suppose $a_n < b_n$ for all n. This case is analogous to the previous one but there are few minor differences. Define the sequence of vectors in $\mathbb{R}^n$:

$$z_n := \left(\frac{b_n - a_n}{2}, 0, ..., 0\right),$$

(notice that $\|z_n\| = (b_n - a_n)/2$) and let U be an orthogonal matrix satisfying (7.8).

(In the fourth equality below we use the fact that $f_1'(a_n, b_n) = f_2'(b_n, a_n)$.) Then

$$\begin{aligned}
f^\circ(\beta(0,t);d_z(y,r)) &= \lim_{n\to\infty} \langle \nabla f(a_n,b_n), d_z(y,r)\rangle \\
&= \lim_{n\to\infty} \left\langle \frac{f_1'(a_n,b_n) - f_2'(a_n,b_n)}{\sqrt{2}\|z\|} z, y \right\rangle + \frac{f_1'(a_n,b_n) + f_2'(a_n,b_n)}{\sqrt{2}} r. \\
&= \lim_{n\to\infty} \left\langle \frac{f_1'(a_n,b_n) - f_2'(a_n,b_n)}{\sqrt{2}\|z_n\|} z_n, U^T y \right\rangle + \frac{f_1'(a_n,b_n) + f_2'(a_n,b_n)}{\sqrt{2}} r \\
&= \lim_{n\to\infty} \left\langle \nabla(f\circ\beta)\left(-z_n, \frac{a_n+b_n}{2}\right), (U^T y, r) \right\rangle \\
&\le (f\circ\beta)^\circ((0,t);(U^T y, r)) \\
&= (f\circ\beta)^\circ((0,t);(y,r)),
\end{aligned}$$

where again in the last equality we used Lemma 7.8.1. □

We now turn our attention to the problem of characterizing the Clarke subgradient, $\partial^c(f\circ\beta)(x,t)$. We need a lemma whose proof is straightforward.

Lemma 7.8.4. *If $x \neq 0$ then the mapping $\beta(x,t)$ is strictly differentiable and its strict derivative, $\nabla_s\beta(x,t)$, is d_x. That is*

$$\lim_{(x',t')\to(x,t),\, \mu\downarrow 0} \frac{\beta((x',t')+\mu(y,r)) - \beta(x',t)}{\mu} = \langle d_x, (y,r)\rangle = d_x(y,r).$$

Theorem 7.8.5. *The Clarke subgradient at the point (x,t) of any Lorentz invariant function $f\circ\beta$, locally Lipschitz around the point (x,t), is given by the formulae*

1. *if $x \neq 0$ then*

$$\partial^c(f\circ\beta)(x,t) = \{d_x^*(\gamma_1,\gamma_2) | (\gamma_1,\gamma_2) \in \partial^c f(\beta(x,t))\};$$

2. if $x = 0$ *then*

$$\partial^c(f \circ \beta)(0,t) = \{d_z^*(\gamma_1,\gamma_2) | (\gamma_1,\gamma_2) \in \partial^c f(\beta(0,t)),\ z \neq 0\}.$$

Proof. **Case 1** When $x \neq 0$, then by Lemma 7.8.4, β is strictly differentiable at (x,t) with strict derivative d_x. Moreover, d_x is a surjective linear map. So we can apply the chain rule for the Clarke subdifferential [15, Theorem 2.3.10], which in our situation holds with equality:

$$\partial^c(f \circ \beta)(x,t) = \partial^c f(\beta(x,t)) \circ d_x.$$

Now, if $(v,p) \in \partial^c(f \circ \beta)(x,t)$ and $(y,r) \in \mathbb{R}^n \times \mathbb{R}$, then there is a subgradient $(\gamma_1,\gamma_2) \in \partial^c f(\beta(x,t))$ such that

$$\langle (v,p),(y,r)\rangle = ((\gamma_1,\gamma_2) \circ d_x)(y,r) = \langle (\gamma_1,\gamma_2), d_x(y,r)\rangle = \langle d_x^*(\gamma_1,\gamma_2),(y,r)\rangle,$$

by Lemma 7.7.3. So

$$\partial^c(f \circ \beta)(x,t) \subseteq \{d_x^*(\gamma_1,\gamma_2) | (\gamma_1,\gamma_2) \in \partial^c f(\beta(x,t))\},$$

the other inclusion is now clear.

Case 2 Let us denote first

$$\mathcal{D} := \{d_z^*(\gamma_1,\gamma_2) | (\gamma_1,\gamma_2) \in \partial^c f(\beta(0,t)),\ z \neq 0\}.$$

We are going to prove the second part of the theorem in two steps. First we will show that $\partial^c(f \circ \beta)(x,t) = \operatorname{conv} \mathcal{D}$ and next that $\operatorname{conv} \mathcal{D} = \mathcal{D}$.

Two closed convex sets are equal whenever their support functions are the same. The support function for the set $\operatorname{conv} \mathcal{D}$ is

$$\begin{aligned}
&\max\{\langle (y,r),(z,s)\rangle | (z,s) \in \operatorname{conv} \mathcal{D}\} \\
&\qquad = \max\{\langle (y,r),(z,s)\rangle | (z,s) \in \mathcal{D}\} \\
&\qquad = \max\left\{ \frac{z^T y}{\|z\|} \frac{\gamma_1 - \gamma_2}{\sqrt{2}} + \frac{\gamma_1 + \gamma_2}{\sqrt{2}} r | (\gamma_1,\gamma_2) \in \partial^c f(\beta(0,t)),\, z \neq 0 \right\} \\
&\qquad = \max\{\langle d_z(y,r),(\gamma_1,\gamma_2)\rangle | (\gamma_1,\gamma_2) \in \partial^c f(\beta(0,t)),\, z \neq 0\} \\
&\qquad = \max\{\max\{\langle d_z(y,r),(\gamma_1,\gamma_2)\rangle | (\gamma_1,\gamma_2) \in \partial^c f(\beta(0,t))\} \,|\, z \neq 0\} \\
&\qquad = \max\{f^\circ(\beta(0,t); d_z(y,r)) \,|\, z \neq 0\} \\
&\qquad = (f \circ \beta)^\circ((0,t);(y,r)),
\end{aligned}$$

which is the support function of the Clarke subdifferential at the point $(0,t)$ (see [15, Proposition 2.1.2]). The last equality above follows from Theorem 7.8.2. So

$$\operatorname{clconv} \mathcal{D} = \partial^c(f \circ \beta)(x,t),$$

because $\partial^c(f \circ \beta)(x,t)$ is a closed set [15, Proposition 2.1.2]. The fact that f is a symmetric function implies that $\partial^c f(\beta(0,t))$ is symmetric set (use [15, Theorem 2.3.10]). The fact that $\operatorname{conv} \mathcal{D} = \mathcal{D}$, follows from Lemma 7.7.4 part 1b, and $\mathcal{D}$ is closed by the same lemma, part 2b. □

Corollary 7.8.6 (Clarke Directional Derivative, cont.). *Let (x,t) be a point*

in the domain of $f \circ \beta$, (y, r) be a direction. Then if $x \neq 0$,

$$(f \circ \beta)^\circ((x,t);(y,r)) = f^\circ(\beta(x,t); d_x(y,r)).$$

Proof. Use again [15, Proposition 2.1.2] - the fact that $(f \circ \beta)^\circ((x,t);(y,r))$ is the support function of $\partial^c(f \circ \beta)(x,t)$. □

7.9 Second order differentiability

In this section, let f be twice differentiable at the point (a,b). This means that f is differentiable in a neighbourhood of this point and the first derivative, ∇f, is differentiable again at (a,b). The question that we are going to answer now is whether $g := f \circ \beta$ is twice differentiable at any point (x,t) such that $\beta(x,t) = (a,b)$. Clearly, when $x \neq 0$ elementary calculus shows that g is twice differentiable. It turns out that this is always the case and we prove the following theorem.

Theorem 7.9.1. *f is twice differentiable at $\beta(x,t)$ if and only if $g := f \circ \beta$ is twice differentiable at (x,t). In that case we have*

1. *If $x \neq 0$ then*

$$\begin{aligned}
g''_{x_i x_j}(x,t) &= \frac{x_i x_j}{2\|x\|^2}(f''_{11} - f''_{12} - f''_{21} + f''_{22}) + \frac{\delta_{ij}\|x\|^2 - x_i x_j}{\sqrt{2}\|x\|^3}(f'_1 - f'_2), \\
g''_{tx_i}(x,t) &= \frac{x_i}{2\|x\|}(f''_{11} - f''_{12} + f''_{21} - f''_{22}), \\
g''_{x_i t}(x,t) &= \frac{x_i}{2\|x\|}(f''_{11} + f''_{12} - f''_{21} - f''_{22}), \\
g''_{tt}(x,t) &= \frac{1}{2}(f''_{11} + f''_{12} + f''_{21} + f''_{22}),
\end{aligned}$$

where all second derivatives of f are evaluated at $\beta(x,t)$, and δ_{ij} is 1 if $i=j$ and 0 otherwise,

2. *If $x=0$, then*

$$\begin{aligned} g''_{x_i x_j}(0,t) &= \begin{cases} \frac{1}{2}(f''_{11}-f''_{12}-f''_{21}+f''_{22}), & \text{if } i=j, \\ 0 & \text{otherwise,} \end{cases} \\ g''_{tx_i}(0,t) &= 0, \\ g''_{x_i t}(0,t) &= 0, \\ g''_{tt}(0,t) &= \frac{1}{2}(f''_{11}+f''_{12}+f''_{21}+f''_{22}). \end{aligned}$$

Proof. The verification of part *(1)* is straightforward. Denote

$$\begin{aligned} H_{ii} &:= \frac{1}{2}(1,-1)\nabla^2 f(\beta(0,t)) \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \text{for } i=1,...,n, \\ H_{tt} &:= \frac{1}{2}(1,1)\nabla^2 f(\beta(0,t)) \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \\ H &:= \begin{pmatrix} H_{11} & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & H_{nn} & 0 \\ 0 & \dots & 0 & H_{tt} \end{pmatrix}. \end{aligned}$$

Let $h^T := (h_1, h_2) := (h_1^1, ..., h_1^n, h_2)$ be a vector in $\mathbb{R}^n \times \mathbb{R}$. Using Theorem 7.5.1

we form the difference quotient

$$\lim_{h\to 0}\frac{\|\nabla g(h_1,t+h_2)-\nabla g(0,t)-Hh\|}{\|h\|},$$

and are going to show that the limit is 0. We consider each coordinate separately. Two cases are necessary: one for the coordinates from 1 to n and one for the $(n+1)^{\text{st}}$ coordinate.

Case a. Suppose $i\in\{1,...,n\}$. Then the difference quotient becomes

$$\lim_{h\to 0}\frac{|g_i'(h_1,t+h_2)-g_i'(0,t)-H_{ii}h_1^i|}{\|h\|}.$$

We use Theorem 7.5.1 to evaluate the derivatives g_i'. Notice that if $h_1=0$ the limit is obviously 0. So suppose that $h_1\neq 0$. Then the limit becomes

$$\lim_{h\to 0}\frac{|\frac{h_1^i}{\sqrt{2}\|h_1\|}(f_1'(\beta(h_1,t+h_2))-f_2'(\beta(h_1,t+h_2)))-\frac{h_1^i}{2}(f_{11}''-f_{12}''-f_{21}''+f_{22}'')|}{\|h\|},$$

where the second derivatives of f are evaluated at $\beta(0,t)$. Because f_1' and f_2' exist in a neighbourhood of $\beta(0,t)$ and are differentiable at $\beta(0,t)$ we have

$$\begin{aligned}f_1'(\beta(h_1,t+h_2)) &= f_1'(\beta(0,t))+f_{11}''(\beta(0,t))\frac{h_2+\|h_1\|}{\sqrt{2}}\\ &\quad + f_{12}''(\beta(0,t))\frac{h_2-\|h_1\|}{\sqrt{2}}+o(\|h\|)\end{aligned}\tag{7.9}$$

$$\begin{aligned}f_2'(\beta(h_1,t+h_2)) &= f_2'(\beta(0,t))+f_{21}''(\beta(0,t))\frac{h_2+\|h_1\|}{\sqrt{2}}\\ &\quad + f_{22}''(\beta(0,t))\frac{h_2-\|h_1\|}{\sqrt{2}}+o(\|h\|)\end{aligned}\tag{7.10}$$

Because f is symmetric we have that at the point $\beta(0,t)$ $f_1' = f_2'$, $f_{12}'' = f_{21}''$, and $f_{11}'' = f_{22}''$. Substituting the two expansions into the last limit shows that it is indeed 0.

Case b. Suppose $i = n+1$ Then the difference quotient becomes

$$\lim_{h\to 0} \frac{|g_t'(h_1, t+h_2) - g_t'(0,t) - H_{tt}h_2|}{\|h\|}.$$

The arguments are analogous to the previous case. We again use Theorem 7.5.1 to evaluate the derivative g_t' and then expansions (7.9) and (7.10). □

7.10 Continuity of the Hessian

Theorem 7.10.1. *f is twice continuously differentiable at $\beta(x,t)$ if and only if $g := f \circ \beta$ is at (x,t).*

Proof. This is clearly the case when $x \neq 0$. We are going to show that for any sequence of vectors (x_n, t_n) approaching $(0,t)$, $\nabla^2 g(x_n, t_n)$ approaches $\nabla^2 g(0,t)$. Considering $\nabla^2 g(0,t)$ as a matrix, we are going to prove the convergence for each entry. We again consider two cases and the general situation follows easily from them.

Case I. Suppose $x_n = 0$ for all n. This case is actually quite trivial and follows directly from the continuity of $\nabla^2 f$ at the point $\beta(0,t)$.

Case II. Suppose $x_n \neq 0$ for all n. First, directly from the continuity of $\nabla^2 f$

at the point $\beta(0,t)$ and the formulae given in Theorem 7.9.1 we have

$$\begin{aligned}\lim_{n\to\infty} g''_{x_i t}(x_n,t_n) &= \lim_{n\to\infty} g''_{t x_i}(x_n,t_n) = 0\\ \lim_{n\to\infty} g''_{tt}(x_n,t_n) &= g''_{tt}(0,t).\end{aligned}$$

So the interesting part is to prove that $\lim_{n\to\infty} g''_{x_i x_j}(x_n,t_n) = g''_{x_i x_j}(0,t)$. Denote

$$\begin{aligned}\beta^n_{+-} &:= \frac{1}{\sqrt{2}}(t_n+\|x_n\|, t_n-\|x_n\|) = \beta(x_n,t_n),\\ \beta^n_{++} &:= \frac{1}{\sqrt{2}}(t_n+\|x_n\|, t_n+\|x_n\|),\\ \beta^n_{-+} &:= \frac{1}{\sqrt{2}}(t_n-\|x_n\|, t_n+\|x_n\|).\end{aligned}$$

Because f is symmetric $f'_1(\beta^n_{-+}) - f'_2(\beta^n_{+-}) = 0$. First consider the limit (applying the mean value theorem):

$$\begin{aligned}&\lim_{n\to\infty} \frac{1}{\sqrt{2}\|x_n\|}\big(f'_1(\beta(x_n,t_n)) - f'_2(\beta(x_n,t_n))\big)\\ &\qquad = \lim_{n\to\infty} \frac{1}{\sqrt{2}\|x_n\|}\big(f'_1(\beta^n_{+-}) - f'_1(\beta^n_{++}) + f'_1(\beta^n_{++}) - f'_1(\beta^n_{-+})\big)\\ &\qquad = \lim_{n\to\infty}\left(- f''_{12}\left(\frac{t_n+\|x_n\|}{\sqrt{2}}, \nu(n)\right) + f''_{11}\left(\mu(n), \frac{t_n+\|x_n\|}{\sqrt{2}}\right)\right),\end{aligned}$$

where $\nu(n)$ and $\mu(n)$ are numbers between $\frac{t_n-\|x_n\|}{\sqrt{2}}$ and $\frac{t_n+\|x_n\|}{\sqrt{2}}$. We can now evaluate the above limit using the continuity of $\nabla^2 f$:

$$\begin{aligned}&\lim_{n\to\infty} \frac{1}{\sqrt{2}\|x_n\|}\big(f'_1(\beta(x_n,t_n)) - f'_2(\beta(x_n,t_n))\big)\\ &\qquad = \frac{1}{2}\big(f''_{11}(\beta(0,t)) - f''_{12}(\beta(0,t)) - f''_{21}(\beta(0,t)) + f''_{22}(\beta(0,t))\big).\end{aligned}$$

Finally, using the formula for $g''_{x_i x_j}$ given in Theorem 7.9.1 we can immediately conclude that

$$\begin{aligned}
&\lim_{n\to\infty} g''_{x_i x_j}(x_n, t_n) \\
&\quad = \frac{\delta_{ij}}{2}\big(f''_{11}(\beta(0,t)) - f''_{12}(\beta(0,t)) - f''_{21}(\beta(0,t)) + f''_{22}(\beta(0,t))\big) \\
&\quad = g''_{x_i x_j}(0,t).
\end{aligned}$$

□

7.11 Positive definite Hessian

We begin with a simple lemma and the main result of this section follows next.

Lemma 7.11.1. *If f, defined on an open subset of $\mathbb{R}^2$, is a strictly convex and symmetric function and $a > b$ then $f'_1(a,b) > f'_2(a,b)$.*

Theorem 7.11.2. *If f is twice differentiable then $\nabla^2 f$ is positive definite at the point $\beta(x,t)$ if and only if $\nabla^2(f \circ \beta)$ is positive definite at (x,t).*

Proof. We use the formulae in Theorem 7.9.1 to give a matrix representation of the Hessian of $f \circ \beta$. We define the following $2 \times (n+1)$ matrix

$$X := \frac{1}{\sqrt{2}} \begin{pmatrix} \frac{x}{\|x\|} & -\frac{x}{\|x\|} \\ 1 & 1 \end{pmatrix},$$

and the $(n+1) \times (n+1)$ matrix

$$M := \frac{1}{\sqrt{2}\|x\|} \begin{pmatrix} I_n - \frac{xx^T}{\|x\|^2} & 0 \\ 0 & 0 \end{pmatrix},$$

where I_n is the $n \times n$ identity matrix.

Case I. When $x \neq 0$ the Hessian of $f \circ \beta$ can be written as

$$\nabla^2(f \circ \beta)(x,t) = X\nabla^2 f(\beta(x,t))X^T + M\nabla f(\beta(x,t)) \begin{pmatrix} 1 \\ \\ -1 \end{pmatrix}.$$

For any nonzero vector (y,r) we have

$$(y,r)\big(\nabla^2(f \circ \beta)(x,t)\big)(y,r)^T = \frac{1}{2} d_x(y,r)\left(\nabla^2(f \circ \beta)(x,t)\right) d_x(y,r)^T + \frac{1}{\sqrt{2}\|x\|^3}\left(\|y\|^2\|x\|^2 - (x^T y)^2\right)\big(f_1'(\beta(x,t)) - f_2'(\beta(x,t))\big).$$

Now using the Lemma we can see that the above expression is strictly positive.

Case II. In the case when $x = 0$, then the Hessian of $f \circ \beta$ is a diagonal matrix and the fact that it is positive definite can be easily seen.

In the other direction the proof is also easy: one has to consider vectors y that are collinear to x. □

7.12 The regular and proximal subdifferentials

For the definitions of the regular and the proximal subgradients refer to Section 6.1 and Section 6.10 respectively.

Let now f be our symmetric function on $\mathbb{R}^2$ and $g := f \circ \beta$. We are going to give formulae for $\hat{\partial} g(x,t)$ in terms of $\hat{\partial} f$. The next lemma lists a few properties of the map $\beta(x,t)$. By $\mathbb{R}^n_{\geq}$ we denote the cone of vectors x in $\mathbb{R}^n$ satisfying $x_1 \geq x_2 \geq \ldots \geq x_n$.

Lemma 7.12.1. *1. For any vector w in $\mathbb{R}^2_{\geq}$ the function $w^T\beta$ is convex and any point (x,t) in $\mathbb{R}^n \times \mathbb{R}$ satisfies $d^*_x(w) \in \partial(w^T\beta)(x,t)$.*

2. The directional derivative $\beta'((x,t);(y,r))$ is given by

$$\beta'((x,t);(y,r)) = \begin{cases} d_x(y,r), & \text{if } x \neq 0 \\ \beta(y,r), & \text{if } x = 0. \end{cases}$$

3. The map β is Lipschitz with global constant 1.

4. Given a point (x,t) in $\mathbb{R}^n \times \mathbb{R}$, small points (z,s) satisfy

$$\beta((x,t)+(z,s)) = \beta(x,t) + \beta'((x,t);(z,s)) + O(\|(z,s)\|^2).$$

Proof. 1. The convexity is elementary. To check the second half we need to verify that

$$w^T\beta(y,r) - w^T\beta(x,t) \geq \langle d^*_x(w_1,w_2), (y-x, r-t)\rangle$$

which expanded and simplified is equivalent to

$$\frac{w_1 - w_2}{\sqrt{2}}(\|y\| - \|x\|) \geq \frac{x^T(y-x)}{\|x\|}\frac{w_1 - w_2}{\sqrt{2}}.$$

After cancelation, the last inequality follows from the Cauchy-Schwarz inequality.

2. This part is a straightforward verification.

3. For any points (x,t) and (z,s) we have

$$\begin{aligned}
&\|\beta((x,t)+(z,s))-\beta(x,t)\| \\
&\quad=\frac{1}{\sqrt{2}}\|(t+s+\|x+z\|,t+s-\|x+z\|)-(t+\|x\|,t-\|x\|)\| \\
&\quad=\frac{1}{\sqrt{2}}\|(s+\|x+z\|-\|x\|,s-(\|x+z\|-\|x\|))\| \\
&\quad=\sqrt{s^2+(\|x+z\|-\|x\|)^2} \\
&\quad\le\sqrt{s^2+\|z\|^2} \\
&\quad=\|(z,s)\|.
\end{aligned}$$

4. Suppose first that $x\neq 0$. Then using part 2 of this lemma and several times the Cauchy-Schwarz inequality we get

$$\begin{aligned}
&\|\beta((x,t)+(z,s))-\beta(x,t)-\beta'((x,t);(z,s))\|^2 \\
&\qquad=\frac{1}{2}\left\|(\|x+z\|-\|x\|-\frac{x^Tz}{\|x\|},-\|x+z\|+\|x\|+\frac{x^Tz}{\|x\|})\right\|^2 \\
&\qquad=\left(\|x+z\|-\|x\|-\frac{x^Tz}{\|x\|}\right)^2 \\
&\qquad=O(\|z\|^4)=O(\|(z,s)\|^4),
\end{aligned}$$

where the next to the last equality holds since $\nabla\|\cdot\|(x)=\frac{x}{\|x\|}$.

The case $x=0$ is easy.

□

Let L be a subset or $\mathbb{R}^m$ and fix a point x in $\mathbb{R}^m$. An element d belongs to the *contingent cone* to L at x, denoted $K(L|x)$, if either $d=0$ or there is a sequence

$\{x_r\}$ in L approaching x with $(x_r - x)/\|x_r - x\|$ approaching $d/\|d\|$. The *negative polar* of a subset H or $\mathbb{R}^m$ is the set

$$H^- = \{y \in \mathbb{R}^m | \langle x, y\rangle \leq 0 \ \forall x \in H\}.$$

We use the following lemmas from [52, Proposition 2.1, Proposition 2.2].

Lemma 7.12.2. *Given a function $f : \mathbb{R}^m \to [-\infty, +\infty]$ and a point x^0 in $\mathbb{R}^m$, any regular subgradient of f at x^0 is polar to the contingent cone of the level set $L = \{x \in E : f(x) \leq f(x^0)\}$ at x^0; that is*

$$\hat{\partial} f(x^0) \subset (K(L|x^0))^-.$$

Lemma 7.12.3. *If the function $f : \mathbb{R}^m \to [-\infty, +\infty]$ is invariant under a subgroup G of $O(m)$, then any point x in $\mathbb{R}^m$ and transformation g in G satisfy $\hat{\partial} f(gx) = g\hat{\partial} f(x)$. Corresponding results hold for the proximal, approximate, horizon and Clarke subgradients (see next sections).*

We define the action of the orthogonal group $O(n)$ on $\mathbb{R}^n \times \mathbb{R}$ by

$$U.(x, t) = (Ux, t), \text{ for every } U \in O(n).$$

For a fixed point (x, t) in $\mathbb{R}^n \times \mathbb{R}$ we define the orbit

$$O(n).(x, t) = \{(Ux, t) | U \in O(n)\}.$$

If $x \neq 0$, this orbit is just a $n - 1$ dimensional sphere with radius $\|x\|$ at level t in

$\mathbb{R}^n \times \mathbb{R}$. So it is a $n-1$ dimensional manifold and one can easily calculate that its tangent and normal spaces at the point (x,t) are

$$T_{(x,t)}(O(n).(x,t)) = \{(y,0)|y^T x = 0\}$$
$$N_{(x,t)}(O(n).(x,t)) = \{(ax,b)|(a,b) \in \mathbb{R}^2\}.$$

If $x = 0$ then

$$T_{(0,t)}(O(n).(0,t)) = \{0\}$$
$$N_{(0,t)}(O(n).(0,t)) = \mathbb{R}^{n+1}.$$

Now, using these observations and Lemma 7.12.2 we can say some more about $\hat{\partial}(f \circ \beta)(x,t)$ in the case when $x \neq 0$.

Lemma 7.12.4. *If $x \neq 0$ and $(y,r) \in \hat{\partial}(f \circ \beta)(x,t)$ then $(y,r) = (ax,r)$ for some $a \in \mathbb{R}$.*

Proof.

$$\begin{aligned}(y,r) \in \hat{\partial}(f \circ \beta)(x,t) \Rightarrow \\ (y,r) &\in (K(\{(z,s)|(f \circ \beta)(z,s) \leq (f \circ \beta)(x,t)\}|(x,t)))^- \\ &\subset (K(O(n).(x,t)|(x,t)))^- \\ &= N_{(x,t)}(O(n).(x,t)).\end{aligned}$$

□

The following is the main theorem of this section.

Theorem 7.12.5. *The regular subdifferential of any Lorentz invariant function $f \circ \beta$ at the point (x,t) is given by the formulae:*

1. *If $x \neq 0$ then*

$$\hat{\partial}(f \circ \beta)(x,t) = \{d_x^*(\gamma_1,\gamma_2) | (\gamma_1,\gamma_2) \in \hat{\partial} f(\beta(x,t))\};$$

2. *If $x = 0$ then*

$$\hat{\partial}(f \circ \beta)(0,t) = \{d_z^*(\gamma_1,\gamma_2) | (\gamma_1,\gamma_2) \in \hat{\partial} f(\beta(0,t)),\, z \neq 0\}.$$

Similar formulae hold as well for the proximal subdifferential.

Proof. **Case 1**. This case follows immediately from the chain rule [79, Exercise 10.7].

Case 2 Let $x = 0$. Suppose $(y,r) \in \hat{\partial}(f \circ \beta)(0,t)$, let $z := (z_1,z_2) \in \mathbb{R}^2$ be small, and let w be an arbitrary nonzero vector. Then

$$\begin{aligned} f(\beta(0,t) + (z_1,z_2)) &= (f \circ \beta)\left((0,t) + \left(\frac{w}{\|w\|}\frac{z_1 - z_2}{\sqrt{2}}, \frac{z_1 + z_2}{\sqrt{2}}\right)\right) \\ &\geq (f \circ \beta)(0,t) + \frac{w^T y}{\|w\|}\frac{z_1 - z_2}{\sqrt{2}} + r\frac{z_1 + z_2}{\sqrt{2}} + o(\|z\|) \\ &= f(\beta(0,t)) + \langle d_w(y,r), (z_1,z_2)\rangle + o(\|z\|). \end{aligned}$$

Consequently $d_w(y,r) \in \hat{\partial} f(\beta(0,t))$ for all $w \neq 0$.

In the opposite direction suppose that $d_w(y,r) \in \hat{\partial} f(\beta(0,t))$ for all $w \neq 0$. If

$y = 0$ then for any point (z, s) close to 0 we have

$$\begin{aligned}(f \circ \beta)((0,t)+(z,s)) &\\ &= f(\beta(0,t) + (\beta((0,t)+(z,s)) - \beta(0,t)))\\ &\geq f(\beta(0,t)) + \langle d_w(0,r), (\beta((0,t)+(z,s)) - \beta(0,t))\rangle + o(\|(z,s)\|)\\ &= f(\beta(0,t)) + rs + o(\|(z,s)\|)\\ &= (f \circ \beta)(0,t) + \langle (0,r),(z,s)\rangle + o(\|(z,s)\|).\end{aligned}$$

so $(0, r) \in \hat{\partial}(f \circ \beta)(0, t)$.

If $y \neq 0$ then for $w = y$ we have $d_y(y, r) \in \hat{\partial} f(\beta(0, t))$. Let (z, s) be a point close to 0. Then

$$\begin{aligned}(f \circ \beta)((0,t)+(z,s)) &\\ &= f(\beta(0,t) + (\beta((0,t)+(z,s)) - \beta(0,t)))\\ &\geq f(\beta(0,t)) + \langle d_y(y,r), (\beta((0,t)+(z,s)) - \beta(0,t))\rangle + o(\|(z,s)\|)\\ &= f(\beta(0,t)) + \|y\|\|z\| + rs + o(\|(z,s)\|)\\ &\geq (f \circ \beta)(0,t) + \langle (y,r),(z,s)\rangle + o(\|(z,s)\|).\end{aligned}$$

Consequently $(y, r) \in \hat{\partial}(f \circ \beta)(0, t)$. So we showed that

$$\hat{\partial}(f \circ \beta)(0,t) = \{(y,r) | d_z(y,r) \in \hat{\partial} f(\beta(0,t)), \forall z \neq 0\}.$$

The stated version follows from Lemma 7.7.4 part 1c.

The proof for the proximal subdifferential is essentially identical. □

7.13 The approximate and horizon subdifferential

For the definitions of the approximate and the horizon subgradients refer to Section 6.1.

Theorem 7.13.1. *The approximate subdifferential of any Lorentz invariant function $f \circ \beta$ at the point (x,t) is given by the formulae:*

1. *If $x \neq 0$ then*

$$\partial(f \circ \beta)(x,t) = \{d_x^*(a,b) | (a,b) \in \partial f(\beta(x,t))\};$$

2. *If $x = 0$ then*

$$\partial(f \circ \beta)(0,t) = \{d_z^*(a,b) | (a,b) \in \partial f(\beta(0,t)),\ z \neq 0\}.$$

Similar formulae hold for the horizon subgradient.

Proof. **Part I. $\mathbf{x \neq 0}$.** This case follows immediately from the chain rule [79, Exercise 10.7].

Part II. $\mathbf{x = 0}$. Suppose $(y,r) \in \partial(f \circ \beta)(0,t)$. By definition, there is a sequence of points (x_q, t_q) approaching $(0,t)$ with $(f \circ \beta)(x_q, t_q)$ approaching $(f \circ \beta)(0,t)$, and a sequence of regular subgradients (y_q, r_q) approaching (y,r) such that $(y_q, r_q) \in \hat{\partial}(f \circ \beta)(x_q, t_q)$.

Case II.1.a. Suppose $x_q = 0$ for all q. Then Theorem 7.12.5 says that $(y_q, r_q) = d^*_{z_q}(a_q, a_q)$ such that $(a_q, a_q) \in \hat{\partial} f(\beta(0, t_q))$, for some $z_q \neq 0$. Because (y_q, r_q) approaches (y, r) we get that $y = 0$ and $a_q \to a := r/\sqrt{2}$. So $(0, r) = (0, \sqrt{2}a) = d^*_z(a, a)$ for any $z \neq 0$ and $(a, a) \in \partial f(\beta(0, t))$.

Case II.1.b. Suppose $x_q \neq 0$ for all q. Then Theorem 7.12.5 says that $(y_q, r_q) = d^*_{x_q}(a_q, b_q)$ such that $(a_q, b_q) \in \hat{\partial} f(\beta(x_q, t_q))$. Let us choose a subsequence q' for which $x_{q'}/\|x_{q'}\|$ converges to a unit vector z. Then we have that $|a_{q'} - b_{q'}|$ approaches $\sqrt{2}\|y\|$ and $a_{q'} + b_{q'}$ approaches $\sqrt{2}r$, that is, $(a_{q'}, b_{q'})$ is bounded sequence so if necessary we may choose a convergent subsequence q''. Then $(a_{q''}, b_{q''}) \to (a, b) \in \partial f(\beta(0, t))$ and $(y, r) = d^*_z(a, b)$.

Case II.1.c. Suppose the sequence x_q has infinitely many elements that are equal to 0 and infinitely many elements that are not equal to 0. Let $\{x_q\} = \{x_{q'}\} \cup \{x_{q''}\}$, where $x_{q'} \neq 0$ and $x_{q''} = 0$. We now choose any of the subsequences q' or q'' and apply the corresponding subcase above.

Suppose finally that $(y, r) = d^*_z(a, b)$ for some $(a, b) \in \partial f(\beta(0, t))$ and some $z \neq 0$. By the definition of approximate subgradients there is a sequence (c_q, d_q) approaching $\beta(0, t)$, with $f(c_q, d_q)$ approaching $f(\beta(0, t))$, and a sequence of regular subgradients (a_q, b_q) approaching (a, b) and such that $(a_q, b_q) \in \hat{\partial} f(c_q, d_q)$. We have three possible cases.

Case II.2.a. Suppose first that there is an infinite subsequence q' such that $c_{q'} > d_{q'}$ for all q'. Then $d^*_z(c_{q'}, d_{q'})$ approaches $d^*_z(\beta(0, t)) = (0, t)$, with $f(c_{q'}, d_{q'}) = (f \circ \beta)(d^*_z(c_{q'}, d_{q'}))$ approaching $f(\beta(0, t)) = (f \circ \beta)(0, t)$ and regular subgradients $(a_{q'}, b_{q'}) \in \hat{\partial} f(\beta(d^*_z(c_{q'}, d_{q'})))$. If we set $z_{q'} := \frac{z}{\|z\|} \frac{c_{q'} - d_{q'}}{\sqrt{2}}$, then Theorem 7.12.5 says

that $d^*_{z_{q'}}(a_{q'}, b_{q'}) \in \hat{\partial}(f \circ \beta)(d^*_z(c_{q'}, d_{q'}))$. Notice that $z_{q'}/\|z_{q'}\|$ converges to $z/\|z\|$, so $d^*_{z_{q'}}(a_{q'}, b_{q'})$ approaches $d^*_z(a, b) = (y, r)$, so (y, r) is in $\partial(f \circ \beta)(0, t)$.

Case II.2.b. There is an infinite subsequence q' such that $c_{q'} < d_{q'}$ for all q'. Let us repeat, in a slightly different way, what we know. We have that $(y, r) = d^*_{-z}(b, a)$ where $(b, a) \in \partial f(\beta(0, t))$ (see Lemma 7.12.3) and $z \neq 0$. We are given also that the sequence $(d_{q'}, c_{q'})$ approaches $\beta(0, t)$, with $f(d_{q'}, c_{q'})$ approaching $f(\beta(0, t))$, and the sequence of regular subgradients $(b_{q'}, a_{q'})$ approaches (b, a) and is such that $(b_{q'}, a_{q'}) \in \hat{\partial} f(d_{q'}, c_{q'})$ (by Lemma 7.12.3 again). It is clear now that this case is like the previous one.

Case II.2.c. Suppose finally that there is an infinite subsequence q' such that $c_{q'} = d_{q'}$ for all q'. Then $d^*_z(c_{q'}, d_{q'})$ approaches $d^*_z(\beta(0, t)) = (0, t)$, with $f(c_{q'}, d_{q'}) = (f \circ \beta)(d^*_z(c_{q'}, d_{q'}))$ approaching $f(\beta(0, t)) = (f \circ \beta)(0, t)$ and regular subgradients $(a_{q'}, b_{q'}) \in \hat{\partial} f(\beta(d^*_z(c_{q'}, d_{q'})))$. But then by Theorem 7.12.5 we have that $d^*_z(a_{q'}, b_{q'}) \in \hat{\partial}(f \circ \beta)(0, \sqrt{2} d_{q'}) = \hat{\partial}(f \circ \beta)(d^*_z(c_{q'}, d_{q'}))$ and approaches $d^*_z(a, b)$, and we are done.

The proof of the formulae for the horizon subgradient is analogous. □

7.14 Clarke subgradients - the lower semicontinuous case

For the definition and notation of the Clarke subgradient for a lower semicontinuous function refer to Section 6.7. Recall that if h is lower semicontinuous around $\bar{x}$ then

we have the formula (see [79][Theorem 8.9]):

$$N_{\mathrm{epi}\, h}(\bar{x}, h(\bar{x})) = \{\lambda(v,-1) \,|\, v \in \partial h(\bar{x}),\ \lambda > 0\} \cup \{(v,0) \,|\, v \in \partial^\infty h(\bar{x})\}.$$

We need Lemma 6.7.1 and we restate it below for convenience.

Lemma 7.14.1. *If h is lower semicontinuous around $\bar{x}$ we have the representation*

$$\partial^c h(\bar{x}) = \mathrm{cl}\,(\mathrm{conv}\, \partial h(\bar{x}) + \mathrm{conv}\, \partial^\infty h(\bar{x})).$$

In particular when the cone $\partial^\infty h(\bar{x})$ is pointed we have simpler

$$\partial^c h(\bar{x}) = \mathrm{conv}\, \partial h(\bar{x}) + \mathrm{conv}\, \partial^\infty h(\bar{x}).$$

Clearly f is lower semicontinuous if and only if $f \circ \beta$ is such. As may be expected we have the following theorem.

Theorem 7.14.2. *The Clarke subdifferential of any lower semicontinuous, Lorentz invariant function $f \circ \beta$ at the point (x,t) is given by the formulae:*

1. *If $x \neq 0$ then*

$$\partial^c (f \circ \beta)(x,t) = \{d_x^*(a,b) | (a,b) \in \partial^c f(\beta(x,t))\};$$

2. *If $x = 0$ then*

$$\partial^c (f \circ \beta)(0,t) = \{d_z^*(a,b) | (a,b) \in \partial^c f(\beta(0,t)),\ z \neq 0\}.$$

Proof. Suppose first that $x = 0$. Let $A := \partial f(\beta(x,t))$ and $B := \partial^\infty f(\beta(x,t))$. Using Lemma 7.7.4 and Lemma 7.14.1 we get

$$\begin{aligned}
\partial^c(f \circ \beta)(x,t) &= \operatorname{cl}(\operatorname{conv} \partial(f \circ \beta)(x,t) + \operatorname{conv} \partial^\infty(f \circ \beta)(x,t)) \\
&= \operatorname{cl}(\operatorname{conv} \mathcal{D}(A) + \operatorname{conv} \mathcal{D}(B)) \\
&= \operatorname{cl}(\mathcal{D}(\operatorname{conv} A) + \mathcal{D}(\operatorname{conv} B)) \\
&= \operatorname{cl} \mathcal{D}(\operatorname{conv} A + \operatorname{conv} B) \\
&= \mathcal{D}(\operatorname{cl}(\operatorname{conv} A + \operatorname{conv} B)) \\
&= \mathcal{D}(\partial^c f(\beta(x,t)).
\end{aligned}$$

The case $x \neq 0$ is analogous. □

We end the chapter with a conjecture analogous to the one made by L. Tunçel, [55].

Conjecture 7.14.1. *If $f : \mathbb{R}^2 \to \bar{\mathbb{R}}$ is symmetric and μ-self-concordant barrier, is the same true for $f \circ \beta$?*

An example supporting the conjecture is:

Example 7.14.3. *The function*

$$f(a,b) = -\ln a - \ln b$$

is a 2-self-concordant barrier on $\mathbb{R}^2$, and so is

$$(f \circ \beta)(x,t) = -\ln(t^2 - \|x\|^2) + \ln 2$$

on $\mathbb{R}^{n+1}$. *See [68, Proposition 5.4.3].*

Chapter 8

Future Research

The following is a list of ideas, written to serve mainly as a stimulus for future research. The order of the items is somewhat indicative of the degree of interest we have in these questions. We were told, during the time of writing the corrections of this work, that A. Nemirovskii, already answered positively one of the questions posed in item (2) below, about the self-concordancy of $f \circ \beta$.

1. How do we compute $\nabla^3(f \circ \lambda)$?

2. Can we use the result from (1) to prove or disprove L. Tunçel's conjecture, [55]. A question even more general is: Is a symmetric self-concordant barrier of the roots $\lambda(x)$ of a hyperbolic polynomial, self-concordant for any hyperbolic polynomial? In particular, if $f(x, y)$ is a symmetric, self-concordant barrier on $\mathbb{R}^2$ is the same true for $f \circ \beta$, where β is the map defined in Chapter 7? Finally, an even more restricted question is: If f is a symmetric, self-concordant barrier on $\mathbb{R}$, is the same true for $f(\|x\|)$ on $\mathbb{R}^n$?

3. Knowing the results from Chapter 4 and (1), can one see a pattern for $\nabla^k(f \circ \lambda)$ and prove the conjecture posed at the end of Chapter 4? (Maybe some ideas from [17] will be helpful.)

4. Will the result from (1) be useful to prove the questions about Bregman distance (a kind of generalized metric used in proximal algorithms) formulated by Bauschke and Borwein in [5]?

5. Given a set valued mapping $S : \mathbb{R}^n \Rightarrow \mathbb{R}^m$ which is symmetric, what can we say about relating *graphical derivatives* $DS(\lambda(X), \bar{y})$ and $D(S \circ \lambda)(X, \bar{y})$. How about relating the *coderivatives* $D^*S(\lambda(X), \bar{y})$ and $D^*(S \circ \lambda)(X, \bar{y})$? (See [79, p.324] for the definitions.)

6. Try to extend the group invariance/Eaton triple setting [50], (using Niezgoda's papers [69], [70], [71], showing the relevant subgroup is always a reflection group) to the nonconvex case.

7. Investigate further the properties of the class of self-concordant barrier functions defined in Chapter 3 to see if they can be used in the more practical *long-step* interior point methods. (Follow the development in [25], [27].)

8. Theorem 3.3.2 says that for every hyperbolic polynomial $p(x)$ of degree m, $-m \log(p(x) - a)$ is m^2-self-concordant barrier. A natural question is, what is the optimal, that is, the minimal parameter θ for which this function is a self-concordant barrier. Clearly $m \leq \theta \leq m^2$.

9. Compose the *universal barrier* function, [68], on a symmetric convex set C,

with the eigenvalue map λ, and compare its properties with those of the universal barrier function on $\lambda^{-1}(C)$, [26].

10. Think about Lidskii's conjecture for the roots of hyperbolic polynomials, as formulated in Open Problem 2.3.6. (Maybe the techniques in [41], [43], and [44] will be useful.)

11. What can one say about the C^2 properties of $f \circ \lambda$ when λ is the eigenvalue map of a hyperbolic polynomial (f symmetric)?

12. See if the results in Chapter 6 and Chapter 7 can be generalized in the Jordan algebra framework or within the framework developed by Raphael A. Hauser in [29] and [30].

13. Attempt the three questions at the end of [54].

14. Finally, there is always Conjecture 2.6.1 to keep one busy.

Index of Notation

- $\emptyset$: the empty set
- $\mathbb{R}$: the real numbers
- $\bar{\mathbb{R}}$: the extended real numbers
- $\mathbb{N}$: the natural numbers
- $\mathbb{R}^n$: the n-dimensional real vector space
- $\mathbb{R}^n_{\downarrow}$: the cone of all vectors $x \in \mathbb{R}^n$ satisfying $x_1 \geq x_2 \geq ... \geq x_n$
- $\mathbb{R}^n_+$: the cone of vectors with positive entries
- $\mathbb{R}^n_{++}$: cone of vectors with strictly positive entries
- $\mathbb{R}^n_{\downarrow +}$: $= \mathbb{R}^n_{\downarrow} \cap \mathbb{R}^n_+$
- X^T: the transpose of matrix X
- X^*: the conjugate of matrix X
- $X^\dagger$: Moore-Penrose generalized inverse of matrix X
- $X^{i,j}$: the (i,j)-entry of matrix X
- $X \succeq 0$: matrix X is positive semidefinite
- $X \succ 0$: matrix X is positive definite
- $A(n)$: $n \times n$ real skew-symmetric matrices
- H^n: $n \times n$ Hermitian matrices
- I, I_n: $n \times n$ identity matrix
- S^n: $n \times n$ real symmetric matrices
- S^n_+: $n \times n$ real symmetric positive semidefinite matrices
- S^n_{++}: $n \times n$ real symmetric positive definite matrices

- M_n: square $n \times n$ real matrices
- $M_{n,m}$: rectangular $n \times m$ real matrices
- $O(n)$: $n \times n$ real orthogonal matrices
- $O(n,m)$: Cartesian product $O(n) \times O(m)$
- $P(n)$: $n \times n$ permutation matrices
- $P_{(-)}(n)$: $n \times n$ signed permutation matrices
- $\lambda(X)$: the vector of eigenvalues of $X \in S^n$ ordered in nonincreasing order
- $\sigma(X)$: the vector of singular values of $X \in M_{n,m}$ ordered in nonincreasing order
- $O(n,m).X = \{U_n^T X U_m | (U_n, U_m) \in O(n,m)\}$ orbit of $X \in M_{n,m}$ under the action of the group $O(n,m)$
- $O(n,m)_X = \{(U_n, U_m) \in O(n,m) | U_n^T X U_m = X\}$ stabilizer of $X \in M_{n,m}$ in the group $O(n,m)$.
- $e^1, ..., e^n$: the standard basis of $\mathbb{R}^n$
- e: the all 1's vector
- $|x|$: absolute value of $x \in \mathbb{R}$
- $|x| = (|x_1|, ..., |x_n|)$: vector $x \in \mathbb{R}^n$ with its entries replaced by absolute values
- $|X| = (|x_{ij}|)$: matrix $X \in M_{m,n}$ with its entries replaced by absolute values
- $\bar{x}$: the vector with the same entries as x ordered in decreasing order
- $x_\downarrow$: the same as $\bar{x}$
- $x_{[i]}$: the i-th coordinate of vector $\bar{x}$
- $\hat{x}$: the vector in $\mathbb{R}^n$ with the same entries as $|x|$ ordered in nonincreasing order
- $x^2 = (x_1^2, ..., x_n^2)$: $x \in R^n$

- $x \cdot y = (x_1 y_1, ..., x_n y_n)$: $x, y \in \mathbb{R}^n$
- $\oplus_{i=1}^r x^i$: direct sum of vectors x^i
- $\|x\|$: Euclidean norm for $x \in \mathbb{R}^n$
- $\langle x, y \rangle$: the canonical inner product, $x, y \in \mathbb{R}^n$
- $x \prec y$: vector y majorizes x, that is, $\sum_{i=1}^k \bar{x}_i \leq \sum_{i=1}^k \bar{y}_i$, for $k = 1, ..., n-1$, and $\sum_{i=1}^n \bar{x}_i = \sum_{i=1}^n \bar{y}_i$
- $x \prec_w y$: vector y weakly majorizes x, that is, $\sum_{i=1}^k \bar{x}_i \leq \sum_{i=1}^k \bar{y}_i$, for $k = 1, ..., n$
- $\langle X, Y \rangle = \operatorname{tr} X^T Y$: canonical inner product, $X, Y \in M_{n,m}$
- $X \circ Y = (x_{ij} y_{ij})_{i,j=1}^n$: Hermitian product of matrices
- $\mathcal{C}^1$: continuously differentiable
- $\mathcal{C}^k$: k-times continuously differentiable
- C^-: negative polar cone of set C
- C^+: positive polar cone of set C
- $\bar{C} := \{\bar{c} : c \in C\}$
- $C_\downarrow$: the same as $\bar{C}$
- $C \setminus D = \{x \in C | x \notin D\}$: relative complement
- $d(C, D) = \inf\{\|c - d\| : c \in C, d \in D\}$: the distance between sets C and D
- $\operatorname{cl} C$: closure of set C
- $\operatorname{int} C$: interior of set C
- $\operatorname{bd} C$: the boundary of the set C
- $\operatorname{conv} C$: convex hull of set C
- $\operatorname{span} C$: the linear span of the vectors in C
- $\operatorname{Diag} x$: the matrix with vector x on the main diagonal and zeros everywhere else
- $\operatorname{diag} X$: the vector formed by the main diagonal entries of matrix X in $M_{m,n}$

- $\operatorname{dom} f$: domain of function f
- $\operatorname{epi} f = \{(x, \alpha) | \alpha \geq f(x)\}$: epigraph of function f
- $\operatorname{Graph} f = \{(x, y) | y \in f(x)\}$: the graph of (multi) function f
- $\operatorname{rank} X$: rank of matrix X
- $\operatorname{sign}(x)$: the sign of the number x
- $\operatorname{supp} y = \{i : y_i \neq 0\}$: support of vector y
- $\operatorname{tr} X$: trace of matrix $X \in M_n$
- C^∞: horizon cone
- $d\pi$: differential of map π
- δ^*_Λ: the support function of set Λ
- $f \circ g$: composition of functions f and g
- f^*: Fenchel conjugate of f
- f^{**}: Fenchel biconjugate of function f
- $\nabla f(x)$: gradient of function f at point x
- $\nabla^2 f(x)$: Hessian of function f at point x
- $\nabla^k f(x)$: k-th derivative of f
- $f'_i(x)$: the i-th partial derivative of f at x
- $f'(x; y)$: directional derivative of f at x in the direction of y
- f''_{ij}: the (i, j)-th second partial derivative
- $\partial f(x)$: the (approximate) subdifferential at point x
- $\hat{\partial} f(x)$: the regular subdifferential at point x
- $\partial^\infty f(x)$: the horizon subdifferential at point x
- $\bar{\partial} f(x)$: the Clarke subdifferential at the point x for a lower semicontinuous function f

- $\partial^c f(x)$: the Clarke subdifferential at point x for a locally Lipschitz function f
- $\partial_p f(x)$: the proximal subdifferential at point x
- $T_C(x)$: tangent cone to set C at point x
- $N_C(x)$: normal cone to set C at point x
- $\hat{N}_C(x)$: regular normal cone to set C at point x

Bibliography

[1] A .D. ALEXANDROV. Almost everywhere existence of the second differential of a convex function and some properties of convex surfaces connected with it. *Leningrad State University Annals [Uchenye Zapiski], Mathematical Series*, 6:3–35, 1939. Russian.

[2] F. ALIZADEH, J.-P.A. HAEBERLY, and M. L. OVERTON. Primal-dual interior-point methods for semidefinite programming: Convergence rates, stability and numerical results. *SIAM Journal on Optimization*, 8(3):746–768, 1998.

[3] V.I. ARNOLD. On matrices depending on parameters. *Russian Mathematical Surveys*, 26:29–43, 1971.

[4] L. AUSLANDER and R.E. MACKENZIE. *Introduction to Differential Manifolds*. McGraw-Hill Series in Higher Mathematics. McGraw-Hill Book Company, Inc., 1963.

[5] H.H. BAUSCHKE and J.M. BORWEIN. Joint and separate convexity of the bregman distance, 2000. Personal communication.

[6] H.H. BAUSCHKE, O. GÜLER, A.S. LEWIS, and HR.S. SENDOV. Hyperbolic polynomials and convex analysis. *University of Waterloo Research Report*, CORR 98-29, June 1998. to appear in the Canadian Journal of Mathematics.

[7] E.F. BECKENBACH and R. BELLMAN. *Inequalities*. Springer-Verlag, 1961.

[8] R. BHATIA. *Perturbation Bounds for Matrix Eigenvalues*. Pitman Research Notes in Mathematics. John Wiley & Sons, Inc., New York, 1987.

[9] R. BHATIA. *Matrix Analysis*. Springer-Verlag, NY, 1997.

[10] J.M. BORWEIN and A.S. LEWIS. *Convex Analysis and Nonlinear Optimization*. Springer-Verlag, New York, 2000.

[11] S. BOYD, L. EL GHAOUI, E. FERON, and V. BALAKRISHNAN. *Linear Matrix Inequalities in System and Control Theory*, volume 15 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, 1994.

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

[13] P. BULLEN and M. MARCUS. Symmetric means and matrix inequalities. *Proceedings of the American Mathematical Society*, 12:285–290, 1961.

[14] P.S. BULLEN, D.S. MITRINOVIĆ, and P.M. VASIĆ. *Means and Their Inequalities*. Mathematics and its applications. East European Series. D. Reidel Publishing Company, Dordrecht, Holland, 1987.

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

[16] A. CURNIER, Q.C. HE, and P. ZYSSET. Conewise linear elastic materials. *J. Elasticity*, 37:1–38, 1995.

[17] J. DADOK. On the C^∞ Chevalley's theorem. *Advances in Mathematics*, 44:121–131, 1982.

[18] C. DAVIS. All convex invariant functions of hermitian matrices. *Archiv der Mathematik*, 8:276–278, 1957.

[19] A. EDELMAN, T.ARIAS, and S.T. SMITH. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. (electronic).

[20] L.C. EVANS and R.F. GARIEPY. *Measure Theory and Fine Properties of Functions.* Studies in Advanced Mathematics. CRC Press, Inc., 1992.

[21] K. FAN. On a theorem of Weyl concerning eigenvalues of linear transformations ii. *Proceedings of the National Academy of Sciences of the United States of America*, 36:31–35, 1950.

[22] R. FLETCHER. A new variational result for quasi-Newton formulae. *SIAM Journal on Optimization*, 1:18–21, 1991.

[23] R.W. FREUND, F. JARRE, and S. SCHAIBLE. On self-concordant barrier functions for conic hulls and fractional programming. *Mathematical Programming*, 74(no. 3, Ser. A):237–246, 1996.

[24] L. GÅRDING. An inequality for hyperbolic polynomials. *Journal of Mathematics and Mechanics*, 8(6):957–965, 1959.

[25] O. GÜLER. Hyperbolic polynomials and interior point methods for convex programming. *Mathematics of Operations Research*, 22(2):350–377, 1997.

[26] O. GÜLER. On the self-concordance of the universal barrier function. *SIAM Journal on Optimization*, 7(2):295–303, 1997.

[27] O. GÜLER. Hyperbolic polynomials and interior point methods for convex programming. Technical Report TR95-40, University of Maryland, November 1995.

[28] G. HARDY, J.E. LITTLEWOOD, and G. POLYA. *Inequalities*. Cambridge University Press, second edition, 1989.

[29] R.A. HAUSER. Self-scaled barrier functions: decomposition and classification. Technical Report DAMTP 1999/NA13, Department of Applied Mathematics and Theoretical Physics, Silver Street, Cambridge, England CB3 9EW, 1999.

[30] R.A. HAUSER. Self-scaled barriers for semidefinite programming. Technical Report DAMTP 2000/NA02, Department of Applied Mathematics and Theoretical Physics, Silver Street, Cambridge, England CB3 9EW, 2000.

[31] J.-B. HIRIART-URRUTY and C. LEMARÉCHAL. *Convex Analysis and Minimization Algorithms I.* Springer-Verlag, Berlin, 1993.

[32] J.-B. HIRIART-URRUTY and D. YE. Sensitivity analysis of all eigenvalues of a symmetric matrix. *Numerische Mathematik*, 70:45–72, 1992.

[33] R.V. HOGG and E.A. TANIS. *Probability and Statistical Inference.* Prentice Hall, fifth edition, 1997.

[34] L. HÖRMANDER. *Analysis of Linear Partial Differential Operators I.* Springer-Verlag, New-York, Berlin, 1983.

[35] L. HÖRMANDER. *Analysis of Linear Partial Differential Operators II.* Springer-Verlag, New-York, Berlin, 1984.

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

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

[38] T.W. HUNGERFORD. *Algebra*, volume 73 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1974.

[39] A.D. IOFFE. Approximate subdifferentials and applications. I: The finite dimentional theory. *Transactions of the American Mathematical Society*, 281:389–416, 1984.

[40] A.D. IOFFE. Approximate subdifferentials and applications. II. *Mathematika*, 33(1):111–128, 1986.

[41] T. KATO. *A Short Introduction to Perturbation Theory for Linear Operators.* Springer-Verlag, Berlin, 1976.

[42] E.C. KEMBLE. *The Fundamental Principles of Quantum Mechanics*. Dover, New York, 1958.

[43] K. KNOPP. *Theory of Functions*, volume 1. Dover, New York, 1945.

[44] K. KNOPP. *Theory of Functions*, volume 2. Dover, New York, 1947.

[45] N.V. KRYLOV. On the general notion of fully nonlinear second-order elliptic equations. *Transactions of the American Mathematical Society*, 347(3):857–895, March 1995.

[46] P.D. LAX. Differential equations, difference equations and matrix theory. *Comm. Pure Appl. Math.*, 11:175–194, 1958.

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

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

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

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

[51] A.S. LEWIS. Lidskii's theorem via nonsmooth analysis. *SIAM Journal on Matrix Analysis and Applications*, 21(2):379–381, 1999.

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

[53] A.S. LEWIS. Convex analysis on Cartan subspaces. *Nonlinear Analysis, Theory, Methods and Applications*, 42(no. 5, Ser. A: Theory Methods):813–820, 2000.

[54] A.S. LEWIS. Nonsmooth duality, sandwich, and squeeze theorems. *SIAM Journal on Control and Optimization*, 38(2):613–626, 2000.

[55] A.S. LEWIS and M.L. OVERTON. Eigenvalue optimization. *Acta Numerica*, 5:149–190, 1996.

[56] A.S. LEWIS and H.S. SENDOV. Twice differentiable spectral functions. Technical Report CORR 2000-15, University of Waterloo, 2000. submitted to SIAM Journal of Matrix Analysis.

[57] A.S. LEWIS and H.S. SENDOV. Self-concordant barriers for hyperbolic means. Technical Report CORR 99-31, University of Waterloo, August 1999. submitted to Mathematical Programming, Series A.

[58] A.S. LEWIS and H.S. SENDOV. Quadratic expansions of spectral functions. Technical Report CORR 2000-41, University of Waterloo, August 2000. submitted to Linear Algebra and Applications.

[59] M. MARCUS and L. LOPES. Inequalities for symmetric functions and Hermitian matrices. *Canad. J. Math.*, 9:305–312, 1957.

[60] A.S. MARKUS. The eigen- and singular values of the sum and product of linear operators. *Uspehi Matematicheskih Nauk*, 19(4):93–123, 1964. Russian Math. Surveys 19, 92-120 (1964).

[61] A.W. MARSHALL and I. OLKIN. *Inequalities: Theory of Majorization and Its Applications*, volume 143 of *Mathematics in Science and Engineering*. Academic Press, New York, 1979.

[62] J.B. MCLEOD. On four inequalities in symmetric functions. *Proceedings of the Edinburgh Mathematical Society (Series II)*, 11:211–219, 1958-1959.

[63] D.S. MITRINOVIĆ. *Analytic Inequalities*. Springer-Verlag, New York, 1970. In cooperation with P. M. Vasić. Die Grundlehren der mathematischen Wisenschaften, Band 1965.

[64] D.S. MITRINOVIĆ, J.E. PEČARIĆ, and A.M. FINK. *Classical and New Inequalities in Analysis*, volume 61. Kluwer, 1993.

[65] B.S. MORDUKHOVICH. Maximum principle in the problem of time optimal response with nonsmooth constraints. *Journal of Applied Mathematics and Mechanics*, 40:960–969, 1976.

[66] A.S. NEMIROVSKII. Private communication, August 1999.

[67] Y.E. NESTEROV and A.S. NEMIROVSKII. *Optimization Over Positive Semidefinite Matrices: Mathematical Background and User's Manual*. USSR Acad. Sci. Center Econ & Math. Inst., Moscow, 1990.

[68] Y.E. NESTEROV and A.S. NEMIROVSKII. *Interior-Point Polynomial Algorithms in Convex Programming*, volume 13 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, 1994.

[69] M. NIEZGODA. Group majorization and Schur type inequalities. *Linear Algebra and its Applications*, 268:9–30, 1998.

[70] M. NIEZGODA. On Schur-Ostrowski type theorems for group majorizations. *Journal of Convex Analysis*, 1:81–105, 1998.

[71] M. NIEZGODA. An application of Eaton triples. *Linear Algebra and its Applications*, 2000. to appear.

[72] J. NOCEDAL. Theory of algorithms for unconstrained optimization. *Acta Numerica*, pages 199–242, 1992.

[73] A.M. OSTROWSKI. *Solution of Equations in Euclidean and Banach Spaces*. Academic Press, New York, third edition edition, 1973.

[74] M.L. OVERTON and R.S. WOMERSLEY. Second derivatives for eigenvalue optimization. *SIAM Journal on Matrix Analysis and its Applications*, 16(3):697–718, 1995.

[75] A. PAZMAN. *Foundations of Optimum Experimental Design*. Reidel, Boston, MA, 1986.

[76] A. L. PERESSINI, F.E. SULLIVAN, and Jr. J.J. UHL. *The Mathematics of Nonlinear Programing*. Undergraduate Texts in Mathematics. Springer-Verlag, NY, 1988.

[77] F. RELLICH. *Perturbation Theory of Eigenvalue Problems.* Lecture Notes. New York University, 1953.

[78] R.T. ROCKAFELLAR. *Convex Analysis.* Princeton University Press, Princeton, N.J., 1970.

[79] R.T. ROCKAFELLAR and R.J.-B. WETS. *Variational Analysis.* Springer-Verlag, Berlin, 1996.

[80] L.I. SCHIFF. *Quantum Mechanics.* McGraw-Hill, New York, 1955.

[81] A. SEEGER. Sensitivity analysis of nondifferentiable sums of singular values of rectangular matrices. *Numerical Functional Analysis and Optimization*, 16(1 & 2):247–260, 1995.

[82] A. SEEGER. Convex analysis of spectrally defined matrix functions. *SIAM Journal on Optimization*, 7(3):679–696, 1997.

[83] A. SEEGER and T. HOANG. On conjugate functions, subgradient, and directional derivatives of a class of optimality criteria in experimental design. *Statistics*, 22:349–368, 1991.

[84] G. W. STEWARD and J. SUN. *Matrix Perturbation Theory.* Academic Press, 1990.

[85] M.J. TODD, K.C. TOH, and R.H. TÜTÜNCÜ. On the Nesterov-Todd direction in semidefinite programming. *SIAM Journal on Optimization*, 8(3):769–796, 1998.

[86] M. TORKI. Second-order epi-differentiability and subdifferentiability of spectral convex functions. *Journal of Mathematical Analysis and Applications*, 1999. Personal Communication.

[87] M. TORKI. Second-order directional derivatives of all eigenvalues of a symmetric matrix. *Nonlinear Analysis, Theory, Methods and Applications*, 2000. To appear.

[88] N.-K. TSING, M. K. H. FAN, and E. I. VERRIEST. On analyticity of functions involving eigenvalues. *Linear Algebra and its Applications*, 207:159–180, 1994.

[89] L. TUNÇEL. Convex Optimization: Barrier Functions and Interior-Point Methods, 1998. Unpublished Manuscript.

[90] 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.

[91] S. WAKABAYASHI. Remarks on hyperbolic polynomials. *Tsukuba Journal of Mathematics*, 10(1):17–28, 1986.

[92] G. WARNER. *Harmonic Analysis on Semi-Simple Lie Groups.* Springer-Verlag, Berlin-Heidelberg-New York, 1972.

[93] W. WASOW. On the spectrum of hermitian matrix-valued functions. *Resultate der Math.*, 2:206–214, 1979.

[94] H. WOLKOWICZ. Measures for symmetric rank-one updates. *Mathematics of Operations Research*, 19:815–830, 1994.

[95] H. WOLKOWICZ, R. SAIGAL, and L. VANDENBERGHE, editors. *Handbook of Semidefinite Programming - Theory, Algorithms and Applications.* International Series in Operations Research and Optimization. Kluwer Academic Pub., 2000.