[16] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, N.J., 1970.

[17] R.T. Rockafellar. Second-order optimality conditions in nonlinear programming obtained by way of epiderivatives. *Mathematics of Operations Research*, 14:462–484, 1989.

[18] A. Shapiro and M.K.H. Fan. On eigenvalue optimization, 1993. Manuscript.

[19] J. von Neumann and E. Wigner. Über das Verhalten von Eigenwerten bei adiabatischen Prozessen. *Physik. Zeitschr.*, 30:467–470, 1929.

[20] D. Ward. Calculus for parabolic second-order derivatives. Submitted to *Canadian J. Math.*

[21] G.A. Watson. An algorithm for optimal l_2 scaling of matrices. *IMA Journal on Numerical Analysis*, 1991. To appear.

[22] G.A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33, 1992.

[23] R.S. Womersley. Local properties of algorithms for minimizing nonsmooth composite functions. *Mathematical Programming*, 32:69–89, 1985.

[24] E. Yamakawa, M. Fukushima, and T. Ibaraki. An efficient trust region algorithm for minimizing nondifferentiable composite functions. *SIAM Journal on Scientific and Statistical Computing*, 10:562–580, 1989.

[25] X.-J. Ye. On the local properties of algorithms for minimizing nonsmooth composite functions, 1991. Manuscript, Nanjing University.

[26] Y. Yuan. On the superlinear convergence of a trust region algorithm for nonsmooth optimization. *Mathematical Programming*, 31:269–285, 1985.

References

[1] V.I. Arnold. On matrices depending on parameters. *Russian Mathematical Surveys*, 26:29–43, 1971.

[2] J.V. Burke. Descent methods for composite nondifferentiable optimization problems. *Mathematical Programming*, 33:260–279, 1985.

[3] J.V. Burke. Second order necessary and sufficient conditions for convex composite NDO. *Mathematical Programming*, 38:287–302, 1987.

[4] R.W. Chaney. On second derivatives for nonsmooth functions. *Nonlinear Analysis: Theory, Methods and Applications*, 9:1189–1209, 1985.

[5] F. H. Clarke. *Optimization and Nonsmooth Analysis*. John Wiley, New York, 1983.

[6] T.F. Fairgrieve. The application of singularity theory to the computation of the Jordan canonical form. Master's thesis, University of Toronto, 1986.

[7] R. Fletcher. *Practical Methods of Optimization*. John Wiley, Chichester and New York, second edition, 1987.

[8] J.B. Hirriart-Urruty. A new set-valued second order derivative for convex functions. In J.B. Hirriart-Urruty, editor, *Fermat Days 85: Mathematics for Optimization*, pages 157–182, Amsterdam, 1986. North-Holland.

[9] A.D. Ioffe. On some recent developments in the theory of second order optimality conditions. In S.Dolecki, editor, *Optimization*, Lecture Notes in Math. 1405, pages 55–68. Springer-Verlag, 1989.

[10] M.R. Osborne. *Finite Algorithms in Optimization and Data Analysis*. John Wiley, Chichester and New York, 1985.

[11] M.L. Overton. Large-scale optimization of eigenvalues. *SIAM Journal on Optimization*, 2:88–120, 1992.

[12] M.L. Overton and R.S. Sezginer. The largest singular value of $e^X A_0 e^{-X}$ is convex on convex sets of commuting matrices. *IEEE Transactions on Automatic Control*, 35:229–230, 1990.

[13] M.L. Overton and R.S. Womersley. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Mathematical Programming*, 1993. To appear.

[14] M.L. Overton and R.S. Womersley. Second derivatives for optimizing eigenvalues of symmetric matrices. Computer Science Dept. Report 626, Courant Institute of Mathematical Sciences, NYU, 1993. Submitted to *SIAM J. Matrix Anal. Appl.*

[15] R. Poliquin and R.T. Rockafellar. A calculus of epi-derivatives applicable to optimization. *Canadian Journal of Mathematics*. To appear.

$$\begin{aligned} h_3(\widehat{a}) &= \max \{ y^T \widehat{A} z : \ y \in \Re^N, \ z \in \Re^N, \ \|y\| = \|z\| = 1 \} \\ &= \max \{ \mathbf{tr}\ \widehat{A} U \ : \ U = z y^T, \ y \in \Re^N, \ z \in \Re^N, \ \|y\| = \|z\| = 1 \}. \end{aligned}$$

This maximum is achieved by matrices of the form

$$\{ U = \widehat{Q} w w^T \widehat{P}^T : \ w = [u \ \ 0]; \ u \in \Re^t, \ \|u\| = 1 \}, \tag{10}$$

since then

$$\mathbf{tr}\ \widehat{A} U = w^T \widehat{\Sigma} w = \widehat{\sigma}_1.$$

Notice the appearance of the *symmetric* rank-one matrix ww^T. It follows from standard results in convex analysis [16] that $\partial h_3(a)$ is the *convex hull* of (10). The proof is completed by noting that the convex hull of

$$\{ u u^T : \ u \in \Re^t, \ \|u\| = 1 \}$$

is [11, Lemma 1]

$$\{ \ U_{11} \in S\Re^{t \times t}, \ \ \mathbf{tr}\ U_{11} = 1, \ \ U_{11} \geq 0 \}.$$

Alternatively, the theorem may be proved less directly by applying [11, Thm 2] to the maximum eigenvalue of

$$\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$$

noting that the negative copies of the singular values are of no interest.

It is interesting to note that [11, Thm 1] does *not* seem to generalize nicely to the singular value case, because there does not seem to be any simple formula for the convex hull of

$$\{ z y^T : \ \ y, \ \ z \in \Re^N, \ \|y\| = \|z\| = 1 \}.$$

Constructing a structure function for h_3 is similar to the process for h_2. Let

$$b = (\beta, \mathbf{vecskew}\ B, \mathbf{vecskew}\ C, \mathbf{vecdiag}\ D)$$

where B and C are skew-symmetric matrices of order N, and D is a diagonal matrix of order $N - t$. Define $g(b) = \beta$ and

$$\Phi(a, b) = \mathbf{vecsym}\ (e^B \widehat{P}^T (\mathbf{Sym}\ a) \widehat{Q} e^C - \begin{bmatrix} \beta I & 0 \\ 0 & D \end{bmatrix}).$$

As in the case of h_2, it is necessary to introduce further restrictions on b in order to obtain the right dimension count and a regularity condition on Φ. In this case, the appropriate restriction is apparently that the leading t by t blocks of B and C be equal.

to hear of other structured convex functions that arise in applications and have a *genuinely different structure* from those discussed here.

Acknowledgements. Helpful conversations with Jim Burke, Jean-Pierre Haeberly, Mike Osborne and Rob Womersley are gratefully acknowledged.

Appendix.

Suppose that $n = N^2$, define

$$A = \mathbf{Mat}\ a$$

to be the N by N matrix defined by the elements of a, and let

$$a = \mathbf{vecmat}\ A.$$

Let

$$h_3(a) = \text{ the maximum singular value of the matrix } \mathbf{Mat}\ a.$$

Let $\widehat{a}$ be given, with $\widehat{A} = \mathbf{Mat}\ \widehat{a}$. Suppose the singular values of $\widehat{A}$ are

$$\widehat{\sigma}_1 = \cdots = \widehat{\sigma}_t > \widehat{\sigma}_{t+1} \geq \cdots \geq \widehat{\sigma}_N,$$

with $\widehat{P}$ and $\widehat{Q}$ orthogonal matrices whose columns are respectively the left and right singular vectors of $\widehat{A}$, i.e.

$$\widehat{A}\widehat{Q} = \widehat{P}\widehat{\Sigma}$$

where $\widehat{\Sigma} = \mathrm{Diag}(\widehat{\sigma}_i)$. The following result is already known [22], but the simple derivation may be of interest.

Theorem.

$$\partial h_3(\widehat{a}) = \{\mathbf{vecsym}\ \widehat{Q}U\widehat{P}^T : \ \ U \in S\Re^{N\times N}, \ \ U = \begin{bmatrix} U_{11} & 0 \\ 0 & 0 \end{bmatrix},$$

$$U_{11} \in S\Re^{t\times t}, \ \ \mathbf{tr}\ U_{11} = 1, \ \ U_{11} \geq 0\}.$$

Proof: A standard variational result is

Suppose that convergence does take place towards a point x satisfying $0 \in \partial f(\widehat{x})$. In order to guarantee that the convergence rate is second-order, we need the manifold Ω and the associated structure function to be more than just C^1, but rather C^2 with a Lipschitz condition on the second derivative. (If we insist that Ω be C^2 in its original definition, we cannot claim that $\mathbf{T}\Omega(\widehat{a})$ and $\mathbf{aff}\ \partial h$ are, in general, orthogonal complements. A counterexample is obtained by considering a function which is C^1 but not C^2 in a certain direction.)

A Newton step for (8)–(9) linearizes Φ to make first-order approximations to the constraint manifold, and uses a quadratic objective based on the Hessian of $\widetilde{L}(x, b, u)$. Considering again the orthogonal decomposition of the variable space into the tangent space to the constraint manifold and the affine hull of the generalized gradient, we see that *first-order* approximation is used in the *latter* direction while *second-order* information is needed in the *former* direction. Given regularity conditions on the functions $a(x)$ and $\Phi(a, b)$, as well as the usual second-order positive definite condition in a subspace, quadratic convergence of such a Newton method can be expected.

There is one important difficulty which has not yet been discussed, and that is the dependence of the definition of the structure function itself on $\widehat{x}$. In the case of the function h_1, this simply means that an "active set" of indices must be identified. This is a standard approach in nonlinear and minmax programming and is quite acceptable, at least for nondegenerate problems, since it only requires estimating which elements of a are coalescing to the same maximum value. In the case of the function h_2, the multiplicity t of the maximum eigenvalue must be identified: this is also acceptable for the same reason. But more is needed in the case of h_2: the definition of the structure function also requires $\widehat{Q}$, which is assumed to be a set of eigenvectors corresponding to an *exactly* multiple eigenvalue. In practice, however, the eigenvalue will be multiple only at the limit point of the iteration. The best we can do is to define Φ in terms of the eigenvectors of the current matrix iterate. It turns out that quadratic convergence is still obtained, but the convergence analysis is substantially complicated by this difficulty: see [14] for details.

We conclude by soliciting examples of other interesting convex functions. We are aware of many examples related to those discussed here, such as other polyhedral functions [7,10], the maximum singular value of a nonsymmetric matrix (see Appendix), other norm functions [22], diagonal scaling problems [12,21], sums of eigenvalues [13], etc. The ideas discussed in this paper are apparently applicable to all these functions. We would be very interested

function g and point $\widehat{b}$. Given a regularity condition on the function $a(x)$, we can expect to find the same orthogonal decomposition of the variable space into the tangent space to the composite manifold and the affine hull of the composite subdifferential (generalized gradient). Consider the equality-constrained nonlinear program

$$\min_{x \in \Re^m,\ b \in \Re^r} \quad g(b) \tag{8}$$

$$\text{s.t.} \quad \Phi(a(x), b) = 0. \tag{9}$$

The corresponding Lagrangian is

$$\widetilde{L}(x, b, u) = g(b) + \langle u, \Phi(a(x), b) \rangle.$$

We have $\nabla_b \widetilde{L} = \nabla_b L$ and

$$\nabla_x \widetilde{L}(x, b, u) = \nabla_x a \nabla_a L(a(x), b, u) = \nabla_x a \nabla_a \Phi(a(x), b)\ u.$$

Let

$$\widetilde{\Psi}(x) = \{v : \ \exists b,\ u \ \text{ s.t. } \nabla_a \widetilde{L}(x, b, u) = v;\ \ \nabla_b \widetilde{L}(x, b, u) = 0;\ \ \Phi(a(x), b) = 0\}.$$

Then if $\Psi(\widehat{a}) = \mathbf{aff}\ \partial h(\widehat{a})$, we also have $\widetilde{\Psi}(\widehat{x}) = \mathbf{aff}\ \partial f(\widehat{x})$. Given a regularity condition on Φ, a necessary condition for $(\widehat{x}, \widehat{b}, u)$ to locally solve the nonlinear program is that $\nabla_x \widetilde{L}(\widehat{x}, \widehat{b}, u) = 0$, $\nabla_b \widetilde{L}(\widehat{x}, \widehat{b}, u) = 0$, $\Phi(a(\widehat{x}), \widehat{b}) = 0$. If $\widetilde{\Psi}(\widehat{x}) = \mathbf{aff}\ \partial f(\widehat{x})$, this means that $0 \in \mathbf{aff}\ \partial f(\widehat{x})$. We therefore have the standard necessary condition for $\widehat{x}$ to minimize the composite function $f = h \circ a$, *with the exception of the inequality conditions on the dual variables defining* $\partial h(\widehat{a})$ [1].

These ideas suggest that the *Newton step for the nonlinear program* (8)–(9) is the main ingredient needed to construct a second-order minimization algorithm for $h \circ a$. Of course, we want convergence to a point $\widehat{x}$ with $0 \in \partial f(\widehat{x})$, not just $0 \in \mathbf{aff}\ \partial f(\widehat{x})$. The best way to ensure this seems to be to compute approximate dual variable information at every step, check the necessary inequality condition on these quantities ($u \geq 0$ in the case of h_1, $\mathbf{Sym}\ u \geq 0$ in the case of h_2), and, if this condition does *not* hold, conclude that the current manifold (defined by $\widehat{x}$) is *not* optimal, and take a step *away* from the manifold instead of *towards* it. It may not be obvious how to do this: in the case of h_2, see [11, Sec. 3].

[1] It is instructive to compare these formulas and conclusions with the results for the general polyhedral case given by Osborne [10, pp.191-193].

to $\partial \Phi_j / \partial a_i$, evaluated at (a, b). Then

$$\nabla_a L(a, b, u) = \nabla_a \Phi \ u,$$

$$\nabla_b L(a, b, u) = \nabla g + \nabla_b \Phi \ u,$$

and, of course,

$$\nabla_u L(a, b, u) = \Phi.$$

Define the set

$$\Psi(a) = \{d : \ \exists b, \ u \ \text{ s.t. } \nabla_a L(a, b, u) = d; \ \ \nabla_b L(a, b, u) = 0; \ \ \Phi(a, b) = 0\}.$$

Since the definition of L depends on $\widehat{a}$, so does the definition of Ψ. In the case of the ordinary max function h_1, we have $b = \beta$,

$$\nabla_b L(a, b, u) = \frac{\partial L(a, b, u)}{\partial \beta} = 1 - e^T u$$

and

$$\nabla_a L(a, b, u) = \left[\begin{array}{c} I \\ 0 \end{array} \right] u = \left[\begin{array}{c} u \\ 0 \end{array} \right]$$

so we see from (3) that, if a is near enough to $\widehat{a}$, the set $\Psi(a)$ is *precisely* **aff** $\partial h_1(a)$.

Now consider the max eigenvalue function h_2. It is straightforward to show that

$$\nabla_a L(\widehat{a}, \widehat{b}, u) = \mathbf{vecsym} \ \widehat{Q} U \widehat{Q}^T,$$

where $U = \mathbf{Sym} \ u$. (This equation does not hold in a neighborhood of $(\widehat{a}, \widehat{b})$, but only at the point.) The first component of $\nabla_b L(a, b, u)$ is

$$\frac{\partial L(a, b, u)}{\partial \beta} = 1 - \mathbf{tr} \ U_{11},$$

where U_{11} is the leading t by t block of U. Furthermore, it can be shown [14, Thm 5.2] that setting the other components of $\nabla_b L(\widehat{a}, \widehat{b}, u)$ to zero gives the condition that all elements of U outside the U_{11} block must be zero. Therefore, we see from (5) that $\Psi(\widehat{a})$ is precisely **aff** $\partial h_2(\widehat{a})$.

We conjecture that this result, $\Psi(\widehat{a}) =$ **aff** $\partial h(\widehat{a})$, can be stated and proved in a fairly general setting.

Now recall that our objective is to minimize the composite function $f = h \circ a$. Suppose that a point $\widehat{x}$ is given: this defines $\widehat{a} = a(\widehat{x})$, which in turn defines a composite structure function Φ and the associated smooth

where $\widehat{Q}$ is defined in (4) and I denotes the identity matrix of order t. We have $s = N(N+1)/2$. Clearly, Φ is smooth, and letting

$$\widehat{b} = (h_2(\widehat{a}), \mathbf{vecskew}\ 0, (\widehat{\lambda}_{t+1}, \ldots, \widehat{\lambda}_N)),$$

we have $\Phi(\widehat{a}, \widehat{b}) = 0$. Furthermore, $\Phi(a, b) = 0$ implies that $A = \mathbf{Sym}\ a$ is similar to a diagonal matrix whose first t elements are β. Since the eigenvalues of a matrix are a continuous function of the matrix elements, this implies that the maximum eigenvalue of A equals β if a is sufficiently close to $\widehat{a}$.

This choice of structure function has one difficulty: $n+r-s$, the number of variables in (a, b) reduced by the number of equations in Φ, equals $n + 1 - t$, which is not the dimension of $\Omega(\widehat{a})$. This implies that too many variables were introduced in b, or alternatively, that not enough equations were imposed in Φ. The inevitable conclusion is that a desirable regularity condition, namely that the Jacobian of Φ have full rank at $(\widehat{a}, \widehat{b})$, will fail to hold. It turns out [14] that the solution to this difficulty is to restrict the leading t by t block of B to be zero. This reduces r by $t(t-1)/2$, so that $n + r - s = n + 1 - t(t+1)/2$, as desired. An alternative approach to parameterizing $\Omega(\widehat{a})$ for the max eigenvalue function has recently been proposed by [18].

It is no surprise to find that $\mathbf{T}\Omega(\widehat{a})$ and $\mathbf{aff}\ \partial h(\widehat{a})$ are orthogonal complements, since h is smooth when restricted to Ω and, on the other hand, the subdifferential describes exactly directions in which h is nonsmooth. Theorem 13.1 of [16] is highly relevant in this regard. More worthy of note, perhaps, is that for the interesting function h_2, beautifully simple arguments from convex analysis show that $\partial h_2(\widehat{a})$ has the form (5), while equally beautiful and simple arguments from differential geometry show that the tangent space $\mathbf{T}\Omega(\widehat{a})$ is orthogonal to the affine hull of (5). We must, however, be cautious in making general statements. In particular, nothing has yet been said about the uniqueness of Ω: clearly, one should define Ω to have dimension as large as possible. Provided this is done properly, it seems that the orthogonal complementarity of $\mathbf{T}\Omega(\widehat{a})$ and $\mathbf{aff}\ \partial h$ should hold in general.

Let us introduce a Lagrangian

$$L(a, b, u) = g(b) + \langle u, \Phi(a, b) \rangle.$$

Since the definition of the structure function Φ depends on $\widehat{a}$, so does the definition of the Lagrangian. Let $\nabla_a L$ denote the gradient of L with respect to a, etc., with $\nabla_a \Phi$, for example, being the matrix with i, j element equal

Clearly, the conditions on the structure function are satisfied. Note that, in this case, $n + r - s$, the number of variables in a and b reduced by the number of equations in Φ, equals 1, the dimension of Ω.

Now consider h_1, the ordinary max function, with t defined in (1). The manifold $\Omega(\widehat{a})$ is defined by constraining $a_1 = \cdots = a_t$, which is an affine space with dimension $n + 1 - t$ (codimension $t - 1$). As was the case with h_0, we see from (3) that $\Omega(\widehat{a})$ and **aff** $\partial h_1(\widehat{a})$ are orthogonal complements. Let $r = 1$, $b = \beta \in \Re$, $s = t$,

$$g(\beta) = \beta \quad \text{and} \quad \Phi(a, \beta) = \begin{bmatrix} a_1 - \beta \\ \vdots \\ a_t - \beta \end{bmatrix}.$$

Letting $\widehat{\beta} = h_1(\widehat{a})$, we have $\Phi(\widehat{a}, \widehat{\beta}) = 0$. The equation $\Phi(a, \beta) = 0$ implies that $a_1 = \cdots = a_t = \beta$, and hence, if a is close enough to $\widehat{a}$, the max element equals β. Thus, the conditions on the structure function are satisfied. We have $n + r - s = n + 1 - t$, the dimension of $\Omega(\widehat{a})$.

Now consider $h_2(a)$, the maximum eigenvalue of $A =$ **Sym** a. We have $n = N(N+1)/2$, where N is the dimension of A. Define t, the multiplicity of the largest eigenvalue, by (4). The manifold $\Omega(\widehat{a})$ is defined by constraining the t largest eigenvalues of **Sym** a to be equal. This is a *nonlinear* manifold, with dimension $n+1-t(t+1)/2$ (codimension $t(t+1)/2-1$), a fact apparently first observed in [19]. Let $\mathbf{T}\Omega(\widehat{a})$ be the *linear* manifold which is *tangent* to $\Omega(\widehat{a})$ at $\widehat{a}$. Then it can be shown, using elementary techniques from differential geometry described in [1, Sec. 2] or [6, Prop. 2.1.1], that the tangent space $\mathbf{T}\Omega(\widehat{a})$ is orthogonal to the affine space (6), **aff** $\partial h_2(\widehat{a})$.

The best way to parameterize $\Omega(\widehat{a})$ using a structure function is not obvious, but one convenient way is as follows. Let

$$b = (\beta, \mathbf{vecskew}\ B, \mathbf{vecdiag}\ D),$$

where $\beta \in \Re$, B is a *skew-symmetric* matrix of order N, i.e. $B = -B^T$, with **vecskew** B the corresponding vector in $\Re^{N(N-1)/2}$, and D is a *diagonal* matrix of order $N - t$, with **vecdiag** D the corresponding vector in $\Re^{N-t}$. Thus, $b \in \Re^r$, where $r = N(N + 1)/2 - t + 1$. Define $g(b) = \beta$ and

$$\Phi(a, b) = \mathbf{vecsym}\ (e^{-B}\widehat{Q}^T(\mathbf{Sym}\ a)\widehat{Q}e^B - \begin{bmatrix} \beta I & 0 \\ 0 & D \end{bmatrix}),$$

see Burke [2] for a more general treatment, including an interesting historical account of convex composite optimization. Another general treatment is given by [25].

Our *key assumption* is that, given any point $\widehat{a}$, the *local structure* of the convex function h is known. By this we mean that a manifold $\Omega(\widehat{a})$, containing the point $\widehat{a}$ and contained in $\Re^n$, on which h reduces to a smooth function near $\widehat{a}$, is known. More specifically, suppose that $\Omega(\widehat{a})$ is the solution set of a known equation

$$\Phi(a, b) = 0, \tag{7}$$

projected into $\Re^n$. Here the *structure function*

$$\Phi: \quad \Re^n \times \Re^r \to \Re^s \quad \text{ is } C^1,$$

with $\Phi(\widehat{a}, \widehat{b}) = 0$ for some $\widehat{b} \in \Re^r$, and with, for some positive ϵ,

$$\Phi(a, b) = 0 \text{ and } ||a - \widehat{a}|| \leq \epsilon \quad \Rightarrow \quad h(a) = g(b)$$

where

$$g: \quad \Re^r \to \Re \quad \text{ is } C^1.$$

We have introduced the additional variables $b \in \Re^r$ in order to facilitate the description of Ω by the structure function Φ. Our use of the structure function Φ is partly inspired by the notion of *structure functionals* for polyhedral convex functions introduced by Osborne[10] and may be viewed as a generalization of this concept to the nonpolyhedral case.

Before applying these ideas to h_1 and h_2, let $n = 2$ and consider the simple convex function

$$h_0(a) = \frac{1}{2}a_1^2 + |\, a_2 \,|\,.$$

To avoid the trivial case, assume $\widehat{a}_2 = 0$. The subdifferential is then easily seen to be

$$\{a: \quad a_1 = \widehat{a}_1; \quad a_2 \in [-1, 1]\},$$

so

$$\textbf{aff } \partial h_0(\widehat{a}) = \{a: \quad a_1 = \widehat{a}_1\}.$$

The manifold $\Omega(\widehat{a})$ is the linear space $\{a: \quad a_2 = 0\}$. Note particularly that $\Omega(\widehat{a})$ and **aff** $\partial h_0(\widehat{a})$ are *orthogonal complements*. To parameterize Ω in the structure function notation, let $r = 1$, and write $b = \beta$ to emphasize that $b \in \Re$. Let $s = 2$,

$$g(b) = \frac{1}{2}\beta^2 \quad \text{ and } \quad \Phi(a, \beta) = \begin{bmatrix} a_1 - \beta \\ a_2 \end{bmatrix}.$$

case of h_1, there are no second-order effects to consider, since h_1 is polyhedral. But in the case of h_2, we would like to use second-order information as well; the subdifferential does *not* contain this information.

For the composite function $f = h \circ a$, it is well known that a generalized subdifferential, more often known as a generalized gradient, can be obtained by means of a chain rule [5,7]. Composing the subdifferential of the convex function h with the gradient of the smooth function a, we have

$$\partial f(x) = \partial h(a) \circ \nabla a(x)$$

$$= \{v \in \Re^m : \quad v = \langle d, \nabla a(x) \rangle, \text{ for some } d \in \partial h(a)\}.$$

A necessary condition for optimality is then

$$0 \in \partial f(x).$$

Let $\widehat{x}$ be a given point, defining $\widehat{a} = a(\widehat{x})$, and consider $f_1 = h_1 \circ a$, $f_2 = h_2 \circ a$. We have

$$\partial f_1(\widehat{x}) = \{v : \quad v_k = \sum_{i=1}^{t} u_i \frac{\partial a_i}{\partial x_k}, \quad u \in \Re^t, \quad e^T u = 1, \quad u \geq 0\}$$

and

$$\partial f_2(\widehat{x}) = \{v : \quad v_k = \mathbf{tr}\, U \widehat{Q}^T (\mathbf{Sym}\, \frac{\partial a}{\partial x_k}) \widehat{Q}, \quad U = \begin{bmatrix} U_{11} & 0 \\ 0 & 0 \end{bmatrix},$$

$$U_{11} \in S\Re^{t \times t}, \quad \mathbf{tr}\, U_{11} = 1, \quad U_{11} \geq 0\}.$$

But, since the second-order information in h_2 is not contained in the subdifferential, the second-order information in f_2 cannot be obtained from simply composing the subdifferential of h_2 with the Hessian of a.

There has been a great deal of recent work concerning *second-order analytical tools* for *general* convex functions. An excellent overview is given by Burke[3]; other references include [4,8,9,15,17,20]. We take a very different approach here, assuming specifically that a lot is known about the *structure* of the convex function. Work on *numerical methods* for convex composite optimization has, on the other hand, been mostly limited to the case where the convex function h is *polyhedral*: see [7,23,24,25,26]. These methods do not easily generalize to the nonpolyhedral case in a satisfactory way since the subproblems which need to be solved are not tractable in general. However,

the affine space of smallest dimension containing $\partial h(\widehat{a})$. We have **aff** $\partial h(\widehat{a}) = \partial h(\widehat{a})$ if and only if h is differentiable at $\widehat{a}$.

Duality principles which give a concise representation of the subdifferential of h are known for the functions h_1 and h_2. In the first case, let $\widehat{a}$ be a given point and assume without loss of generality that

$$\widehat{a}_1 = \cdots = \widehat{a}_t > \widehat{a}_{t+1} \geq \cdots \geq \widehat{a}_n, \tag{1}$$

for some integer t. Then, as is well known [7],

$$\partial h_1(\widehat{a}) = \{ \begin{bmatrix} u \\ 0 \end{bmatrix} : \quad u \in \Re^t, \quad e^T u = 1, \quad u \geq 0 \}, \tag{2}$$

where $e = [1 \ldots 1]^T$. Consequently

$$\textbf{aff } \partial h_1(\widehat{a}) = \{ \begin{bmatrix} u \\ 0 \end{bmatrix} : \quad u \in \Re^t, \quad e^T u = 1 \}. \tag{3}$$

In the case of h_2, suppose without loss of generality that the eigenvalues of $\widehat{A} = \textbf{Sym } \widehat{a}$ are

$$\widehat{\lambda}_1 = \cdots = \widehat{\lambda}_t > \widehat{\lambda}_{t+1} > \cdots > \widehat{\lambda}_N \tag{4}$$

with $\widehat{Q}$ a matrix whose columns are a corresponding orthonormal set of eigenvectors. Then [11],

$$\partial h_2(\widehat{a}) = \{ \textbf{vecsym } \widehat{Q} U \widehat{Q}^T : \quad U \in S\Re^{N \times N}, \quad U = \begin{bmatrix} U_{11} & 0 \\ 0 & 0 \end{bmatrix},$$

$$U_{11} \in S\Re^{t \times t}, \quad \textbf{tr } U_{11} = 1, \quad U_{11} \geq 0 \}. \tag{5}$$

Here $S\Re^{N \times N}$ denotes the set of real symmetric matrices, **tr** denotes trace and $U_{11} \geq 0$ means that U_{11} is positive semi-definite. It follows that

$$\textbf{aff } \partial h_2(\widehat{a}) = \{ \textbf{vecsym } \widehat{Q} U \widehat{Q}^T : \quad U \in S\Re^{N \times N}, \quad U = \begin{bmatrix} U_{11} & 0 \\ 0 & 0 \end{bmatrix},$$

$$U_{11} \in S\Re^{t \times t}, \quad \textbf{tr } U_{11} = 1 \}. \tag{6}$$

The subdifferential gives all the relevant *first-order* variational information about h, but does not, of course, give second-order information. In the

(continuously differentiable). Thus

$$f: \quad \Re^m \to \Re$$

with

$$f(x) = h(a(x)).$$

In practice, the convex function h often has a rather special structure. The simplest example is

$$h_1(a) = \text{ the maximum element of the vector } a.$$

In this case, h_1 is said to be *polyhedral*, since the epigraph of h_1 (the set of points $(a, \eta) \in \Re^{n+1}$ satisfying $\eta \geq h_1(a)$) is a polyhedron.

Some more interesting examples are obtained by considering convex functions whose argument, while conveniently denoted as a vector $a \in \Re^n$, is really a matrix associated with that vector. If $n = N(N+1)/2$ for some integer N, define

$$A = \mathbf{Sym}\ a$$

to be the N by N symmetric matrix defined by copying the elements of a consecutively into its upper triangle, multiplying the off-diagonal elements by the factor $1/\sqrt{2}$ for convenience, and let

$$a = \mathbf{vecsym}\ A$$

denote the inverse operation, defining the vector a in terms of the elements of A. Let

$$h_2(a) = \text{ the maximum eigenvalue of the symmetric matrix } \mathbf{Sym}\ a.$$

By Rayleigh's variational principle, h_2 is convex, but it is not polyhedral.

First-order optimality conditions for a convex function h are conveniently described in terms of its *subdifferential* [16], defined by

$$\partial h(\hat{a}) = \{d \in \Re^n : \quad h(a) \geq h(\hat{a}) + \langle a - \hat{a}, d \rangle, \quad \forall a \in \Re^n\}.$$

A necessary condition for $\hat{a}$ to minimize h is then

$$0 \in \partial h(\hat{a}).$$

The subdifferential $\partial h(\hat{a})$ is compact and reduces to a single point if and only if h is differentiable at $\hat{a}$, in which case $\partial h(\hat{a}) = \nabla h(\hat{a})$. We shall need to refer to the *affine hull* of the subdifferential,

$$\mathbf{aff}\ \partial h(\hat{a}),$$

Towards Second-Order Methods for Structured Nonsmooth Optimization *

Michael L. Overton † and Xianjian Ye †

NYU Computer Science Dept Report 631
April 1993

Abstract

Structured nonsmooth optimization objectives often arise in a composite form $f = h \circ a$, where h is convex (but not necessarily polyhedral) and a is smooth. We consider the case where the structure of the nonsmooth convex function h is known. Specifically, we assume that, for any given point in the domain of h, a parameterization of a manifold Ω, on which h reduces locally to a smooth function, is given. We discuss two affine spaces: the tangent space to the manifold Ω at a point, and the affine hull of the subdifferential of h at the same point, and explain that these are typically orthogonal complements. We indicate how the construction of locally second-order methods is possible, even when h is nonpolyhedral, provided the appropriate Lagrangian, modeling the structure, is used. We illustrate our ideas with two important convex functions: the ordinary max function, and the max eigenvalue function for symmetric matrices, and we solicit other interesting examples with genuinely different structure from the community.

A minimization objective which often arises in practice is the composite function $f = h \circ a$ where

$$h: \quad \Re^n \to \Re \qquad \text{is convex}$$

and

$$a: \quad \Re^m \to \Re^n \qquad \text{is } C^1$$

*Presented at the Sixth IIMAS-UNAM Workshop on Numerical Analysis and Optimization, Oaxaca, Mexico, January 1992.

†Computer Science Dept., Courant Institute of Mathematical Sciences, New York University. The work of both authors was supported in part by National Science Foundation Grant CCR-9101649.