MATHEMATICS OF OPERATIONS RESEARCH
Vol. 21, No. 3, August 1996
Printed in U.S.A

DERIVATIVES OF SPECTRAL FUNCTIONS

A. S. LEWIS

A *spectral* function of a Hermitian matrix X is a function which depends only on the eigenvalues of X, $\lambda_1(X) \geq \lambda_2(X) \geq \cdots \geq \lambda_n(X)$, and hence may be written $f(\lambda_1(X), \lambda_2(X), \ldots, \lambda_n(X))$ for some symmetric function f. Such functions appear in a wide variety of matrix optimization problems. We give a simple proof that this spectral function is differentiable at X if and only if the function f is differentiable at the vector $\lambda(X)$, and we give a concise formula for the derivative. We then apply this formula to deduce an analogous expression for the Clarke generalized gradient of the spectral function. A similar result holds for real symmetric matrices.

1. Introduction and notation. Optimization problems involving a symmetric matrix variable, X say, frequently involve symmetric functions of the eigenvalues of X in the objective or constraints. Examples include the maximum eigenvalue of X, or $\log(\det X)$ (for positive definite X), or eigenvalue constraints such as positive semidefiniteness. The aim of this paper is to provide a unified, concise and constructive approach to the calculus of such matrix functions. The convex case was covered in Lewis (1996): here we use an independent approach to develop the nonconvex case.

Since the seminal paper of Cullum, Donath and Wolfe (1975), the study of matrix optimization problems (and in particular eigenvalue optimization) has become extremely prominent. A typical constraint is positive semidefiniteness (see for example Fletcher 1985, Shapiro 1985, Wolkowicz 1993, Yang and Vanderbei 1993), and with the modern trend towards interior point methods, it has become popular to incorporate this constraint by a barrier penalty function (involving the eigenvalues), as in Nesterov and Nemirovsky (1994), Alizadeh (1992) and Jarre (1993). A related objective function is used in Fletcher (1991) to give an elegant variational characterization of certain quasi-Newton formulae (see also Wolkowicz 1993). One very common objective function is the maximum eigenvalue (Overton 1988, 1992; Rendl and Wolkowicz 1992; Jarre 1993), or more generally, sums of the largest eigenvalues (Overton and Womersley 1993, Hiriart-Urruty and Ye 1995).

A key step in algorithm development is the investigation of sensitivity results, and hence differentiability questions about the eigenvalues. The standard reference on the effect on eigenvalues of perturbations to a matrix is Kato (1982), which for the most part deals with matrices parametrized by a scalar. By contrast, what are needed in this context are sensitivity results with respect to *matrix* perturbations: the two recent papers Overton and Womersley (1993) and Hiriart-Urruty and Ye (1995) undertake detailed constructive studies of this question. More generally, we may wish to construct generalized gradients: Burke and Overton (1993) is an interesting example, examining the much more difficult question of eigenvalue analysis for

Received March 23, 1994; revised January 30, 1995.
AMS 1991 subject classification. Primary: 26B05; 26B05; Secondary: 15A18, 49K40, 90C31.
OR/MS Index 1978 subject classification. Primary: Mathematics/Matrix functions; Secondary: Programming/Sensitivity.
Key words. Matrix functions, spectral functions, eigenvalues, perturbation, unitarily invariant, differentiability, nonsmooth analysis, Clarke derivative.

0364-765X/96/2103/0576/$01.25

non-Hermitian matrices. For recent, second-order approaches, see Shapiro and Fan (1995) and Overton and Womersley (1995) and the references therein. Interesting applications include Polak and Wardi (1983) and Watson (1991).

Let $\mathscr{H}_n$ denote the real vector space of $n \times n$ Hermitian matrices, endowed with the trace inner product, $\langle X, Y \rangle = \operatorname{tr} XY$, and let $\mathscr{U}_n$ denote the $n \times n$ unitary matrices. A real-valued function F defined on a subset of $\mathscr{H}_n$ is *unitarily invariant* if $F(U^*XU) = F(X)$ for any unitary U. Such functions are called *spectral functions* in Friedland (1981), since clearly $F(X)$ depends only on the set of eigenvalues of X, denoted $\lambda_1(X) \geq \lambda_2(X) \geq \cdots \geq \lambda_n(X)$. (This notation also permits us to consider the function $\lambda: \mathscr{H}_n \to \mathbb{R}^n$.)

Associated with any spectral function F is a symmetric, real-valued function f of n real variables (where by *symmetric* we mean that $f(\mu) = f(P\mu)$ for all $n \times n$ permutation matrices P). Specifically, we define $f(\mu) = F(\operatorname{Diag} \mu)$, where $\operatorname{Diag} \mu$ is the diagonal matrix with diagonal $\mu_1, \mu_2, \ldots, \mu_n$. Thus we see that spectral functions $F(X)$ are exactly those functions of the form $f(\lambda(X))$ for symmetric functions f.

We begin by describing a straightforward approach to answering the question: When is the spectral function $f(\lambda(\cdot)) = (f \circ \lambda)(\cdot)$ differentiable at the Hermitian matrix X? We prove the following result. (A set Ω in $\mathbb{R}^n$ is *symmetric* if $P\Omega = \Omega$ for all $n \times n$ permutation matrices P.)

THEOREM 1.1. *Let the set Ω in $\mathbb{R}^n$ be open and symmetric, and suppose that the function $f: \Omega \to \mathbb{R}$ is symmetric. Then the spectral function $f(\lambda(\cdot))$ is differentiable at the matrix X if and only if f is differentiable at the vector $\lambda(X)$. In this case the gradient of $f \circ \lambda$ at X is*

$$(f \circ \lambda)'(X) = U^*(\operatorname{Diag}(f'(\lambda(X))))U, \tag{1.2}$$

for any unitary matrix U satisfying $X = U^*(\operatorname{Diag}(\lambda(X)))U$.

It is easy to see that f must be differentiable at $\lambda(X)$ whenever $f \circ \lambda$ is differentiable at X, since we can write

$$f(\mu) = f(\lambda(U^*(\operatorname{Diag} \mu)U)),$$

(with U as in the theorem), and apply the chain rule at $\mu = \lambda(X)$. Furthermore, the converse is also straightforward at matrices X with distinct eigenvalues, since then the map $\lambda: \mathscr{H}_n \to \mathbb{R}^n$ is differentiable at X and we can easily apply the chain rule to deduce formula (1.2). The interesting case is when some of the eigenvalues of X coalesce: remarkably the spectral function $f(\lambda(\cdot))$ remains differentiable at X even though the map λ is not. The technique revolves around first establishing the result for a diagonal matrix X, and then extending to the general case by a unitary similarity transformation. In this paper we will be concerned only with first-order results. By contrast, in Tsing, Fan and Verriest (1994) it is shown that $f \circ \lambda$ is analytic at X if and only if f is analytic at $\lambda(X)$, in which case (1.2) holds.

The situation where the function f is convex is considered in rather more generality and with an entirely different approach in Lewis (1996). The following analogous result is Theorem 3.2 in Lewis (1996). In this result, ∂ denotes the convex subdifferential.

THEOREM 1.3. *Suppose that the function $f: \mathbb{R}^n \to (-\infty, +\infty]$ is symmetric, convex and lower semicontinuous. Then the Hermitian matrix Y lies in $\partial(f \circ \lambda)(X)$ if and only if $\lambda(Y)$ lies in $\partial f(\lambda(X))$ and there exists a unitary U with $X = U^*(\operatorname{Diag} \lambda(X))U$ and $Y = U^*(\operatorname{Diag} \lambda(Y))U$.*

In fact when f is convex, and differentiable at the vector $\lambda(X)$ (lying in the interior of its domain), Theorem 1.3 reduces to formula (1.2) (see Lewis 1996).

A nice example is to take the symmetric (convex) function

$$f(\mu) = -\sum_1^n \log \mu_i \quad \text{on } \Omega = \{\mu | \mu_1, \mu_2, \ldots, \mu_n > 0\},$$

which corresponds to the spectral function

$$F(X) = f(\lambda(X)) = -\log(\det X),$$

defined on the set of positive definite matrices X. Formula (1.2) then gives $F'(X) = -X^{-1}$ (which may also be obtained directly).

Our ultimate aim in this paper is to unify Theorems 1.1 and 1.3 by considering the Clarke generalized gradient. For an open set Ω in $\mathbb{R}^n$ and a function $f: \Omega \to \mathbb{R}$, we say that f is *locally Lipschitz* around a point μ in Ω if there is a real constant k with $|f(\nu) - f(\gamma)| \leq k\|\nu - \gamma\|$ for all points ν and γ close to μ. If the set Ω is symmetric and the function f is symmetric and locally Lipschitz around a point μ in Ω then the spectral function $f \circ \lambda$ is locally Lipschitz around the matrix $\operatorname{Diag}\mu$, because each component $\lambda_i(\cdot)$ is locally Lipschitz throughout $\mathscr{H}_n$ (see the next section).

The *directional derivative* of Clarke (1983) is defined for a direction ρ in $\mathbb{R}^n$ by

$$f^o(\mu; \rho) = \limsup_{\nu \to \mu, t \downarrow 0} \frac{f(\nu + t\rho) - f(\nu)}{t},$$

and we say that a vector γ lies in the *Clarke generalized gradient* $\partial f(\mu)$ if $\langle \gamma, \rho \rangle \leq f^o(\mu; \rho)$ for all ρ in $\mathbb{R}^n$. We make analogous definitions for locally Lipschitz functions on $\mathscr{H}_n$.

The set $\partial f(\mu)$ is compact, convex and nonempty. It coincides with the convex subdifferential when f is finite and convex on the open, convex set Ω, and it is exactly $\{f'(\mu)\}$ if f is continuously differentiable (though not necessarily if f' is not continuous at μ): see Clarke (1983) for these and related results. Our main result, the following theorem, thus goes a long way toward unifying Theorems 1.1 and 1.3 (in a fashion that we will make more precise at the end of the paper). Like Theorem 1.1 (which is used in the proof) the result is first established in the diagonal case, and is then extended by unitary similarity.

THEOREM 1.4. *Let the set Ω in $\mathbb{R}^n$ be open and symmetric, and suppose that the Hermitian matrix X has $\lambda(X) \in \Omega$. Suppose that the function $f: \Omega \to \mathbb{R}$ is symmetric, and is locally Lipschitz around the point $\lambda(X)$. Then*

$$\partial(f \circ \lambda)(X) = \{U^*(\operatorname{Diag}\gamma)U | \gamma \in \partial f(\lambda(X)), U \in \mathscr{U}_n, U^*(\operatorname{Diag}\lambda(X))U = X\}.$$

We conclude by observing that the same approach applies to real symmetric matrices X, simply substituting "real orthogonal" for "unitary" wherever appropriate.

2. The differentiable case. For each integer $m = 1, 2, \ldots, n$, define a function $\sigma_m: \mathscr{H}_n \to \mathbb{R}$ by $\sigma_m(X) = \sum_1^m \lambda_i(X)$, the sum of the m largest eigenvalues of the matrix X. It is a well-known result of Fan (1949) that σ_m is convex (see also Horn and Johnson 1985). Our development revolves around the following known fact. We denote the standard basis in $\mathbb{R}^n$ by $e^1, e^2, \ldots, e^n$.

THEOREM 2.1. *For real numbers $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$, if $\mu_m > \mu_{m+1}$ for some m then the function σ_m is differentiable at* Diag μ *with gradient*

$$\sigma_m'(\text{Diag } \mu) = \text{Diag} \sum_{i=1}^{m} e^i.$$

(The condition holds vacuously for $m = n$.)

PROOF. See Corollary 3.10 in Hiriart-Urruty and Ye (1995), or the proof of Corollary 3.10 in Lewis (1996), or formula (3.28) in Overton and Womersley (1993) for example. □

Given two vectors α and μ in $\mathbb{R}^n$, we say that μ *block-refines* α if $\alpha_i = \alpha_j$ whenever $\mu_i = \mu_j$.

LEMMA 2.2. *If μ block-refines α in $\mathbb{R}^n$, and $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$, then the function $\alpha^T\lambda(\cdot)$ is differentiable at* Diag μ *with* $(\alpha^T\lambda)'(\text{Diag } \mu) = \text{Diag } \alpha$.

PROOF. Suppose that

$$\mu_1 = \mu_2 = \cdots = \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} = \cdots = \mu_{k_r}.$$

Since μ block-refines α, there exist reals $\beta_1, \beta_2, \ldots, \beta_r$ with

$$\alpha_i = \beta_j \quad \text{whenever } k_{j-1} < i \leq k_j, \qquad j = 1, 2, \ldots, r,$$

where we set $k_0 = 0$. Defining $\sigma_0 \equiv 0$, we obtain

$$\alpha^T\lambda(X) = \sum_{j=1}^{r} \beta_j \sum_{i=k_{j-1}+1}^{k_j} \lambda_i(X) = \sum_{j=1}^{r} \beta_j\big(\sigma_{k_j}(X) - \sigma_{k_{j-1}}(X)\big).$$

Now applying Theorem 2.1 gives

$$\begin{aligned}(\alpha^T\lambda)'(\text{Diag } \mu) &= \sum_{j=1}^{r} \beta_j \left(\text{Diag} \sum_{i=1}^{k_j} e^i - \text{Diag} \sum_{i=1}^{k_{j-1}} e^i\right) \\ &= \sum_{j=1}^{r} \beta_j \text{ Diag} \sum_{i=k_{j-1}+1}^{k_j} e^i = \text{Diag } \alpha,\end{aligned}$$

as required. □

Henceforth we shall assume that the set Ω is open and symmetric in $\mathbb{R}^n$ and that the function $f\colon \Omega \to \mathbb{R}$ is symmetric.

LEMMA 2.3. *If f is differentiable at a point μ in Ω satisfying $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$, then μ block-refines $f'(\mu)$. Consequently the function $f'(\mu)^T\lambda(\cdot)$ is differentiable at the matrix* Diag μ, *with* $(f'(\mu)^T\lambda)'(\text{Diag } \mu) = \text{Diag}(f'(\mu))$.

PROOF. Suppose that $\mu_i = \mu_j$ for some distinct indices i and j. Let P be the matrix of the permutation which transposes the ith and jth components. Since the function f is symmetric, $f(\gamma) = f(P\gamma)$ for all points γ in the set Ω, so applying the chain rule at $\gamma = \mu$ gives $f'(\mu) = P^Tf'(P\mu)$. Thus $Pf'(\mu) = f'(\mu)$, so that $(f'(\mu))_i = (f'(\mu))_j$. The last statement follows from the previous lemma. □

Since each component of the function $\lambda(\cdot)$ can be written as a difference of two finite, convex functions, $\lambda_i(\cdot) = \sigma_i(\cdot) - \sigma_{i-1}(\cdot)$, it follows that λ is locally Lipschitz (see Clarke 1983). We now prove the key result.

THEOREM 2.4. *If the symmetric function f is differentiable at a point μ in $\Omega \subset \mathbb{R}^n$ satisfying $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$, then the spectral function $f(\lambda(\cdot))$ is differentiable at the matrix* Diag μ *with*

$$(f \circ \lambda)'(\text{Diag } \mu) = \text{Diag}(f'(\mu)).$$

PROOF. Given any real $\epsilon > 0$, since f is differentiable at μ we have

$$|f(\gamma) - f(\mu) - f'(\mu)^T(\gamma - \mu)\| \leq \epsilon\|\gamma - \mu\|,$$

for points γ sufficiently close to μ. Since λ is locally Lipschitz around Diag μ, there is a real constant K with

$$\|\lambda(Y + \text{Diag } \mu) - \mu\| \leq K\|Y\|$$

for all Hermitian Y sufficiently small. Hence

$$|f(\lambda(Y + \text{Diag } \mu)) - f(\mu) - f'(\mu)^T(\lambda(Y + \text{Diag } \mu) - \mu)|$$

$$\leq \epsilon\|\lambda(Y + \text{Diag } \mu) - \mu\| \leq K\epsilon\|Y\|,$$

for all small Y.

We also know from Lemma 2.3 that

$$|f'(\mu)^T\lambda(Y + \text{Diag } \mu) - f'(\mu)^T\mu - \text{tr}(Y \text{ Diag}(f'(\mu)))| \leq \epsilon\|Y\|,$$

for all small Y. Now adding the two previous inequalities and using the triangle inequality gives

$$|f(\lambda(Y + \text{Diag } \mu)) - f(\mu) - \text{tr}(Y \text{ Diag}(f'(\mu)))| \leq (K + 1)\epsilon\|Y\|$$

for all small Y, which completes the proof. □

PROOF OF THEOREM 1.1. As we observed, one direction is easy, so suppose that f is differentiable at the vector $\lambda(X)$, and choose any unitary matrix U with $X = U^*(\text{Diag}(\lambda(X)))U$. Now clearly for all Hermitian Z close to X,

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

Applying Theorem 2.4 and the chain rule at $Z = X$ gives

$$\begin{aligned}(f \circ \lambda)'(X) &= U^*((f \circ \lambda)'(UXU^*))U \\ &= U^*((f \circ \lambda)'(\text{Diag}(\lambda(X))))U \\ &= U^*(\text{Diag}(f'(\lambda(X))))U,\end{aligned}$$

since the adjoint of the linear map $X \mapsto UXU^*$ is just $W \mapsto U^*WU$. □

COROLLARY 2.5. *Theorem* 2.4 *holds without the assumption that* $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$.

PROOF. Let $\overline{\mu}$ be the vector obtained by permuting the components of the vector μ into nonincreasing order, and pick a permutation matrix P with $P\mu = \overline{\mu}$. Since f is symmetric, we know that $f(P\nu) = f(\nu)$ for all points ν close to μ, so applying the

chain rule at $\nu = \mu$ gives $f'(\mu) = P^T f'(P\mu)$, and hence $f'(\overline{\mu}) = Pf'(\mu)$. Now if we set $X = \operatorname{Diag}\mu$ then $\lambda(X) = \overline{\mu}$. Observe that $P(\operatorname{Diag}\mu)P^T = \operatorname{Diag}(P\mu)$, so we can choose $U = P$ in Theorem 1.1, and deduce that

$$(f \circ \lambda)'(\operatorname{Diag}\mu) = P^T(\operatorname{Diag}(f'(\overline{\mu})))P = \operatorname{Diag}(P^T f'(\overline{\mu})) = \operatorname{Diag}(f'(\mu)),$$

as required. □

As an example, let the symmetric function $f\colon \mathbb{R}^n \to \mathbb{R}$ be defined by

$$f(\mu) = \text{the } m\text{th largest element of } \{\mu_1, \mu_2, \ldots, \mu_n\}.$$

This function is differentiable at any point μ for which

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_{m-1} > \mu_m > \mu_{m+1} \geq \mu_{m+2} \geq \cdots \geq \mu_n,$$

with $f'(\mu) = e^m$. Now Theorem 1.1 states that the mth largest eigenvalue $\lambda_m(\cdot)$ is differentiable at the Hermitian matrix X if and only if the eigenvalue $\lambda_m(X)$ has multiplicity one, in which case the gradient $\lambda_m'(X) = uu^*$ for any corresponding normalized eigenvector u.

3. The locally Lipschitz case. Throughout this section we shall suppose that the set Ω in $\mathbb{R}^n$ is symmetric and open, that the point μ in $\mathbb{R}^n$ satisfies $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$, and that the symmetric function $f\colon \Omega \to \mathbb{R}$ is locally Lipschitz around μ.

For a Hermitian matrix Z, the vector $\operatorname{diag} Z$ is the diagonal of Z. The map diag: $\mathscr{H}_n \to \mathbb{R}^n$ may be thought of as the adjoint of the map Diag: $\mathbb{R}^n \to \mathscr{H}_n$. For square matrices $A_1, A_2, \ldots, A_r$, we write the block-diagonal matrix

$$\begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_r \end{pmatrix} = \operatorname{Diag}(A_1, A_2, \ldots, A_r).$$

We denote the set of $n \times n$ (real) doubly-stochastic matrices by $\mathscr{S}_n$.

The following result is elementary: it describes how the Clarke directional derivative defined in the introduction is affected by unitary similarity transformations.

LEMMA 3.1. *For any unitary matrix U and any Hermitian matrices X and Z we have* $(f \circ \lambda)^\circ(X; Z) = (f \circ \lambda)^\circ(U^*XU; U^*ZU)$.

PROOF.

$$\begin{aligned}
(f \circ \lambda)^\circ(X; Z) &= \limsup_{Y \to X,\, t \downarrow 0} \frac{f(\lambda(Y + tZ)) - f(\lambda(Y))}{t} \\
&= \limsup_{Y \to X,\, t \downarrow 0} \frac{f(\lambda(U^*(Y + tZ)U)) - f(\lambda(U^*YU))}{t} \\
&= \limsup_{W \to U^*XU,\, t \downarrow 0} \frac{f(\lambda(W + tU^*ZU)) - f(\lambda(W))}{t} \\
&= (f \circ \lambda)^\circ(U^*XU; U^*ZU),
\end{aligned}$$

as required. □

We define a compact set of $n \times n$ matrices,

$$\mathscr{W}_\mu = \{U \in \mathscr{U}_n | U^*(\text{Diag } \mu)U = \text{Diag } \mu\}. \tag{3.2}$$

THEOREM 3.3. *For any Hermitian matrix Z we have*

$$(f \circ \lambda)^\circ(\text{Diag } \mu; Z) = \max\{f^\circ(\mu; \text{diag}(UZU^*)) | U \in \mathscr{W}_\mu\}. \tag{3.4}$$

PROOF. From page 64 in Clarke (1983), there exists a sequence of Hermitian $X_r \to \text{Diag}\,\mu$ with $f \circ \lambda$ differentiable at each X_r and

$$\langle (f \circ \lambda)'(X_r), Z\rangle \to (f \circ \lambda)^\circ(\text{Diag } \mu; Z),$$

and notice that $\lambda(X_r) \to \mu$. For each $r = 1, 2, \ldots$, there exists a unitary U_r with $U_r^*(\text{Diag}(\lambda(X_r))U_r = X_r$, and so by Theorem 1.1,

$$(f \circ \lambda)'(X_r) = U_r^*(\text{Diag}(f'(\lambda(X_r))))U_r.$$

Since $\mathscr{U}_n$ is compact, there is a subsequence for which $U_{r'} \to U \in \mathscr{U}_n$. But now,

$$U^*(\text{Diag } \mu)U = \lim_{r'} U_{r'}(\text{Diag}(\lambda(X_{r'})))U_{r'} = \lim_{r'} X_{r'} = \text{Diag } \mu,$$

so that $U \in \mathscr{W}_\mu$. Hence

$$\begin{aligned}
(f \circ \lambda)^\circ(\text{Diag } \mu; Z) &= \lim_r \langle U_r^*(\text{Diag}(f'(\lambda(X_r))))U_r, Z\rangle \\
&= \lim_{r'} \langle f'(\lambda(X_{r'})), \text{diag}(U_{r'}ZU_{r'}^*)\rangle \\
&= \lim_{r'} \langle f'(\lambda(X_{r'})), \text{diag}(UZU^*)\rangle \\
&\leq \limsup_{\gamma \to \mu} \langle f'(\gamma), \text{diag}(UZU^*)\rangle \\
&= f^\circ(\mu; \text{diag}(UZU^*)).
\end{aligned}$$

Thus we have proved " $\leq$ " in formula (3.4).

On the other hand, fix a matrix U in $\mathscr{W}_\mu$. Again from page 64 in Clarke (1983), there is a sequence of points $\mu^r \to \mu$ with

$$\begin{aligned}
f^\circ(\mu; \text{diag}(UZU^*)) &= \lim_r \langle f'(\mu^r), \text{diag}(UZU^*)\rangle \\
&= \lim_r \langle \text{Diag}(f'(\mu_r)), UZU^*\rangle \\
&= \lim_r \langle (f \circ \lambda)'(\text{Diag } \mu_r), UZU^*\rangle \\
&\leq \limsup_{Y \to \text{Diag } \mu} \langle (f \circ \lambda)'(Y), UZU^*\rangle \\
&= (f \circ \lambda)^\circ(\text{Diag } \mu; UZU^*) \\
&= (f \circ \lambda)^\circ(\text{Diag } \mu; Z),
\end{aligned}$$

using Corollary 2.5 and Lemma 3.1. □

We next translate Theorem 3.3 into a statement about the Clarke generalized gradient. We define another set of $n \times n$ matrices

$$\mathscr{D}_\mu = \{U^*(\operatorname{Diag}\gamma)U | U \in \mathscr{W}_\mu, \gamma \in \partial f(\mu)\}. \tag{3.5}$$

Notice that since $\partial f(\mu)$ is a compact set in $\mathbb{R}^n$ and $\mathscr{W}_\mu$ is a compact set in $\mathbf{C}^{n\times n}$, and since the map $(\gamma, U) \mapsto U^*(\operatorname{Diag}\gamma)U$ is clearly continuous, it follows that $\mathscr{D}_\mu$ is compact.

COROLLARY 3.6. *The Clarke generalized gradient* $\partial(f \circ \lambda)(\operatorname{Diag}\mu)$ *is the convex hull of the set* $\mathscr{D}_\mu$.

PROOF. The convex hull of $\mathscr{D}_\mu$, and the generalized gradient are both compact, convex sets, so it will suffice to see that the two corresponding support functions are identical (see page 28 in Clarke 1983). The support function of the convex hull of $\mathscr{D}_\mu$, $\operatorname{conv}\mathscr{D}_\mu$, evaluated at a Hermitian matrix Z, is (using Theorem 3.3)

$$\begin{aligned}
\max\{\langle Z, Y\rangle | Y \in \mathscr{D}_\mu\} &= \max\{\langle Z, Y\rangle | Y \in \operatorname{conv}\mathscr{D}_\mu\} \\
&= \max\{\langle Z, U^*(\operatorname{Diag}\gamma)U\rangle | U \in \mathscr{W}_\mu, \gamma \in \partial f(\mu)\} \\
&= \max\{\max\{\langle \operatorname{diag}(UZU^*), \gamma\rangle | \gamma \in \partial f(\mu)\} | U \in \mathscr{W}_\mu\} \\
&= \max\{f^\circ(\mu; \operatorname{diag}(UZU^*)) | U \in \mathscr{W}_\mu\} \\
&= (f \circ \lambda)^\circ(\operatorname{Diag}\mu; Z),
\end{aligned}$$

which is the support function of the required Clarke generalized gradient. □

To continue the proof of Theorem 1.4 we need to show that the set $\mathscr{D}_\mu$ is convex. This fact, which does not seem obvious, is extremely important: it ensures that no convex hull operation is required in the Clarke derivative formula in Theorem 1.4, making this formula much more computationally attractive. It should be noted that if the generalized gradient $\partial f(\mu)$ is replaced in equation (3.5) by an arbitrary convex set then the corresponding set $\mathscr{D}_\mu$ will *not* generally be convex.

The first step is an alternative description of the set $\mathscr{W}_\mu$. Suppose that

$$\mu_1 = \mu_2 = \cdots \mu_{k_1} > \mu_{k_1+1} = \cdots = \mu_{k_2} > \mu_{k_2+1} \cdots = \mu_{k_r}, \tag{3.7}$$

and define $k_0 = 0$.

PROPOSITION 3.8. *The set* $\mathscr{W}_\mu$ *consists of all block-diagonal matrices of the form* $\operatorname{Diag}(U_1, U_2, \ldots, U_r)$, *with* U_j *in* $\mathscr{U}_{k_j - k_{j-1}}$, *for* $j = 1, 2, \ldots, r$.

PROOF. This is standard: matrices in $\mathscr{W}_\mu$ have the form $(u^1, u^2, \ldots, u^n)$, with columns an orthonormal basis of eigenvectors for $\operatorname{Diag}\mu$, and since the standard unit vector e^i is an eigenvector with eigenvalue μ_i, we have $\langle u^j, e^i\rangle = 0$ whenever $\mu_j \neq \mu_i$. The result then follows. □

LEMMA 3.9. *Suppose that for vectors* γ^j *in* $\mathbb{R}^{k_j - k_{j-1}}$ $(j = 1, 2, \ldots, r)$, *the (partitioned) vector* $\gamma = (\gamma^1, \gamma^2, \ldots, \gamma^r)$ *lies in* $\partial f(\mu)$. *Then for any doubly-stochastic matrices* S_j *in* $\mathscr{S}_{k_j - k_{j-1}}$ $(j = 1, 2, \ldots, r)$, *the (partitioned) vector* $(S_1\gamma^1, S_2\gamma^2, \ldots, S_r\gamma^r)$ *also lies in* $\partial f(\mu)$.

PROOF. By Birkhoff's Theorem (a result proved earlier by König, see Chvátal 1983), since the set $\partial f(\mu)$ is convex it will suffice to prove the result when each S_j is a permutation matrix. In this case, define the permutation matrix $P = \mathrm{Diag}(S_1, S_2, \ldots, S_r)$, and note that $P\mu = \mu$. Now since f is symmetric, $f(\nu) = f(P\nu)$ for all points ν close to μ, so by the Chain Rule, Theorem 2.3.10 in Clarke (1983), $\partial f(\mu) = P^T \partial f(P\mu)$. Hence $P\gamma = (S_1\gamma^1, S_2\gamma^2, \ldots, S_r\gamma^r) \in \partial f(\mu)$. □

We can now derive an alternative description of the set $\mathscr{D}_\mu$.

COROLLARY 3.10. *The set $\mathscr{D}_\mu$ consists of all block-diagonal matrices of the form $\mathrm{Diag}(D_1, D_2, \ldots, D_r)$ with D_j in $\mathscr{H}_{k_j - k_{j-1}}$, for $j = 1, 2, \ldots, r$, and with the vector $(\lambda(D_1), \lambda(D_2), \ldots, \lambda(D_r))$ in $\partial f(\mu)$.*

PROOF. This follows easily by applying Proposition 3.8 and Lemma 3.9 (with the S_j chosen as suitable permutation matrices). □

LEMMA 3.11. *For any two $m \times m$ Hermitian matrices D and E and any real α in $[0, 1]$, there is an $m \times m$ doubly-stochastic matrix S with*

$$\lambda(\alpha D + (1 - \alpha)E) = S(\alpha\lambda(D) + (1 - \alpha)\lambda(E)).$$

PROOF. In the Schur partial order, the vector $\alpha\lambda(D) + (1 - \alpha)\lambda(E)$ *majorizes* the vector $\lambda(\alpha D + (1 - \alpha)E)$: in other words (see Marshall and Olkin 1979)

$$\sum_{i=1}^{j} \lambda_i(\alpha D + (1 - \alpha)E) \leq \sum_{i=1}^{j} (\alpha\lambda_i(D) + (1 - \alpha)\lambda_i(E))$$

for $j = 1, 2, \ldots, m$, with equality for $j = m$. This follows from Fan's result that the function $\Sigma_1^j \lambda_i(\cdot)$ is convex (c.f. Friedland 1981). The result now follows from page 11 in Marshall and Olkin (1979). □

THEOREM 3.12. *The set $\mathscr{D}_\mu$ is convex.*

PROOF. Suppose that the matrices A and B belong to $\mathscr{D}_\mu$, and fix a real α in $[0, 1]$. By Corollary 3.10 there are matrices D_j and E_j in $\mathscr{H}_{k_j - k_{j-1}}$ for $j = 1, 2, \ldots, r$ with

$$A = \mathrm{Diag}(D_1, D_2, \ldots, D_r), \qquad B = \mathrm{Diag}(E_1, E_2, \ldots, E_r),$$

and with both the vectors

$$(\lambda(D_1), \lambda(D_2), \ldots, \lambda(D_r)) \quad \text{and} \quad (\lambda(E_1), \lambda(E_2), \ldots, \lambda(E_r))$$

in $\partial f(\mu)$. By Lemma 3.11 there exist matrices S_j in $\mathscr{S}_{k_j - k_{j-1}}$ with

$$\lambda(\alpha D_j + (1 - \alpha)E_j) = S_j(\alpha\lambda(D_j) + (1 - \alpha)\lambda(E_j)), \tag{3.13}$$

for $j = 1, 2, \ldots, r$. Since the set $\partial f(\mu)$ is convex we have that

$$\alpha(\lambda(D_1), \lambda(D_2), \ldots, \lambda(D_r)) + (1 - \alpha)(\lambda(E_1), \lambda(E_2), \ldots, \lambda(E_r)) \in \partial f(\mu).$$

Hence by Lemma 3.9 and (3.13),

$$(\lambda(\alpha D_1 + (1 - \alpha)E_1), \lambda(\alpha D_2 + (1 - \alpha)E_2), \ldots, \lambda(\alpha D_r + (1 - \alpha)E_r)) \in \partial f(\mu),$$

so $\alpha A + (1 - \alpha)B \in \mathscr{D}_\mu$ by Corollary 3.10. □

We have now proved, by Corollary 3.6, that $\partial(f \circ \lambda)(\mathrm{Diag}\,\mu) = \mathscr{D}_\mu$. The general result follows by a simple change of variables.

PROOF OF THEOREM 1.4. Pick a unitary V with $V(\mathrm{Diag}(\lambda(X))V^* = X$. Since $(f \circ \lambda)(V^*YV) = (f \circ \lambda)(Y)$ for all Hermitian Y close to X, we can apply the Chain Rule, Theorem 2.3.10 in Clarke (1983), at $Y = X$ (observing that the linear map $Y \mapsto V^*YV$ is invertible) to deduce that

$$\begin{aligned}
&\partial(f \circ \lambda)(X)\\
&\quad = V(\partial(f \circ \lambda)(V^*XV))V^*\\
&\quad = V(\partial(f \circ \lambda)(\mathrm{Diag}(\lambda(X))))V^*\\
&\quad = \{VU^*(\mathrm{Diag}\,\gamma)UV^* | \gamma \in \partial f(\lambda(X)), U \in \mathscr{U}_n, U^*(\mathrm{Diag}(\lambda(X)))U\\
&\qquad\qquad = \mathrm{Diag}(\lambda(X))\}\\
&\quad = \{W^*(\mathrm{Diag}\,\gamma)W | \gamma \in \partial f(\lambda(X)), W \in \mathscr{U}_n, W^*(\mathrm{Diag}(\lambda(X)))W = X\},
\end{aligned}$$

as required. □

EXAMPLE OF COX AND OVERTON (1994). Let the function $f\colon \mathbb{R}^n \to \mathbb{R}$ be defined by

$$f(\gamma) = m\text{th largest element of } \{\gamma_1, \gamma_2, \ldots, \gamma_n\}.$$

Notice that f is symmetric and locally Lipschitz on $\mathbb{R}^n$, and the corresponding spectral function is given by

$$f(\lambda(X)) = m\text{th largest eigenvalue of } X.$$

Suppose that the point μ in $\mathbb{R}^n$ satisfies (3.7) and that $k_j < m \le k_{j+1}$. Then it is easy to compute (for example using Theorem 2.5.1 in Clarke 1983) that

$$\partial f(\mu) = \mathrm{conv}\{e^i | k_j < i \le k_{j+1}\}.$$

Using this observation, it is a straightforward consequence of Theorem 1.4 that for any Hermitian matrix X,

$$\partial(f \circ \lambda)(X) = \mathrm{conv}\{uu^* | Xu = \lambda_m(X)u, \|u\| = 1\},$$

as observed in Cox and Overton (1994). □

The connection between the convex case, Theorem 1.3, and the locally Lipschitz case, Theorem 1.4, can be made clear by rewriting Theorem 1.3 in the following form. In this result, ∂ denotes the convex subdifferential.

COROLLARY 3.14. *Suppose that the function $f\colon \mathbb{R}^n \to (-\infty, +\infty]$ is symmetric, convex and lower semicontinuous. Then for any Hermitian matrix X, the convex subdifferential of $f \circ \lambda$ at X is*

$$\partial(f \circ \lambda)(X) = \{U^*(\mathrm{Diag}\,\gamma)U | \gamma \in \partial f(\lambda(X)), U \in \mathscr{U}_n, U^*(\mathrm{Diag}\,\lambda(X))U = X\}.$$

PROOF. Suppose first that some Hermitian Y lies in $\partial(f \circ \lambda)(X)$. By Theorem 1.3, choosing $\gamma = \lambda(Y)$ shows that Y lies in the right-hand side above.

Conversely, suppose that for some vector γ in $\partial f(\lambda(X))$ and some unitary U with $U^*(\mathrm{Diag}\,\lambda(X))U = X$, we define a matrix $Y = U^*(\mathrm{Diag}\,\gamma)U$. Let $\bar{\gamma}$ be the vector with components $\gamma_1, \gamma_2, \ldots, \gamma_n$ permuted into nonincreasing order, so that $\lambda(Y) = \bar{\gamma}$. Let the function $f^*\colon \mathbb{R}^n \to (-\infty, +\infty]$ be the *convex conjugate* of f, defined by $f^*(\mu) = \sup_\nu\{\nu^T\mu - f(\nu)\}$. Then it is immediate that f^* is also symmetric, and by definition, $f(\nu) + f^*(\mu) = \nu^T\mu$ if and only if $\mu \in \partial f(\nu)$. Since

$$f(\lambda(X)) + f^*(\gamma) = \gamma^T\lambda(X) \leq \bar{\gamma}^T\lambda(X) \leq f(\lambda(X)) + f^*(\bar{\gamma}) = f(\lambda(X)) + f^*(\gamma),$$

it follows that $\bar{\gamma} \in \partial f(\lambda(X)$, and furthermore there is a permutation matrix P with $P\gamma = \bar{\gamma}$ and $P\lambda(X) = \lambda(X)$, by Lemma 2.1 in Lewis (1993). Now

$$Y = (PU)^*(\mathrm{Diag}(\lambda(Y)))(PU) \quad \text{and} \quad X = (PU)^*(\mathrm{Diag}(\lambda(X)))(PU),$$

so by Theorem 1.3, Y lies in $\partial(f \circ \lambda)(X)$. □

Our main result, Theorem 1.4, coincides with the above corollary when the function f is convex and continuous on a neighbourhood of the point $\lambda(X)$. The corollary is more general in the convex case however, because it applies at boundary points of the domain of f.

To link Theorem 1.4 with the differentiable case, Theorem 1.1, we need one further idea. If the set $\Omega \subset \mathbb{R}^n$ is open then a locally Lipschitz function $f\colon \Omega \to \mathbb{R}$ is *strictly differentiable* at point μ in Ω if it is differentiable there with, for all vectors ζ in $\mathbb{R}^n$,

$$\lim_{\xi \to \mu,\, t \downarrow 0} \frac{f(\xi + t\zeta) - f(\xi)}{t} = f'(\mu)^T\zeta.$$

In fact, a locally Lipschitz f is strictly differentiable at μ exactly when its Clarke generalized gradient is a singleton there, in which case $\partial f(\mu) = \{f'(\mu)\}$. If f is continuously differentiable at μ then it is strictly differentiable there (and locally Lipschitz). For these ideas, see Clarke (1983).

COROLLARY 3.15. *Let the set Ω in $\mathbb{R}^n$ be open and symmetric, and suppose that the function $f\colon \Omega \to \mathbb{R}$ is symmetric and locally Lipschitz. Then $f \circ \lambda$ is strictly differentiable at the matrix X if and only if f is strictly differentiable at the vector $\lambda(X)$.*

PROOF. If $f \circ \lambda$ is strictly differentiable at X, then we have $\partial(f \circ \lambda)(X) = \{(f \circ \lambda)'(X)\}$. Hence by Theorem 1.4, for any γ in $\partial f(\lambda(X))$ and any unitary U with $U^*(\mathrm{Diag}\,\lambda(X))U = X$ we have

$$\|\gamma\|_2^2 = \langle U^*(\mathrm{Diag}\,\gamma)U, U^*(\mathrm{Diag}\,\gamma)U\rangle = \langle (f \circ \lambda)'(X), (f \circ \lambda)'(X)\rangle.$$

Since the right-hand side is constant and $\|\cdot\|_2^2$ is strictly convex, the convex set $\partial f(\lambda(X))$ is a singleton. Hence f is strictly differentiable at $\lambda(X)$.

Conversely, if f is strictly differentiable at $\lambda(X)$ then we have $\partial f(\lambda(X)) = \{f'(\lambda(X))\}$. Now $\partial(f \circ \lambda)(X) = \{(f \circ \lambda)'(X)\}$ follows by applying Theorems 1.1 and 1.4, and hence $f \circ \lambda$ is strictly differentiable at X. □

It is a nice exercise using Theorem 1.1 to show the analogous result that $f \circ \lambda$ is continuously differentiable at X if and only if f is continuously differentiable at $\lambda(X)$. Our last example shows, perhaps not surprisingly, that the exact correspondence between differentiability, continuous differentiability, and strict differentiabil-

ity of $f \circ \lambda$ at X with that of f at $\lambda(X)$, does not extend to Gâteaux differentiability. Recall that f has *Gâteaux derivative* γ at the point μ if for all vectors ν we have

$$\lim_{t \to 0} \frac{f(\mu + t\nu) - f(\mu)}{t} = \gamma^T \nu.$$

EXAMPLE. In $\mathbb{R}^2$, let S be the punctured hyperbola consisting of points $(\mu_1, \mu_2)^T$ distinct from $(1,0)^T$ and $(0,1)^T$ satisfying

$$\mu_1^2 + \mu_2^2 + 3\mu_1\mu_2 - 2\mu_1 - 2\mu_2 + 1 = 0.$$

Define a symmetric function $f\colon \mathbb{R}^2 \to \mathbb{R}$ to take the value 0 on S and 1 otherwise. Then clearly f has Gâteaux derivative zero at the point $(1,0)^T$. Notice however that for any nonzero real t,

$$\lambda\left(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + t\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}\right) \in S,$$

and so

$$\lim_{t \to 0} t^{-1}\left\{f\left[\lambda\left(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + t\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}\right)\right] - f\left[\lambda\left(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\right)\right]\right\}$$

does not exist. Thus $f \circ \lambda$ is not Gâteaux differentiable at the matrix Diag(1, 0).

Acknowledgements. Many thanks to Michael Overton for many helpful suggestions, and to Jon Borwein for suggesting Corollary 3.15. Research partially supported by the Natural Sciences and Engineering Research Council of Canada.

References

Alizadeh, F. (1992). Optimization over the positive definite cone: interior point methods and combinatorial applications. P. Pardolos, Ed., *Advances in Optimization and Parallel Computing*, 1–25. North-Holland, Amsterdam.

Burke, J. and M. L. Overton (1993). On the subdifferentiability of functions of a matrix spectrum: I: Mathematical foundations; II: Subdifferential formulas. F. Giannessi, Ed., *Nonsmooth Optimization: Methods and Applications*, 11–29, Gordon and Breach, Philadelphia.

Chvátal, V. (1983). *Linear Programming*. Freeman, New York.

Clarke, F. H. (1983). *Optimization and Nonsmooth Analysis*. Wiley, New York.

Cox, S. and M. Overton (1994). Private communication.

Cullum, J., W. E. Donath and P. Wolfe (1975). The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices. *Math. Programming Study* **3** 35–55.

Fan, K. (1949). On a theorem of Weyl concerning eigenvalues of linear transformations. *Proceedings of the National Academy of Sciences of U.S.A.* **35** 652–655.

Fletcher, R. (1985). Semi-definite matrix constraints in optimization. *SIAM J. Control and Optimization* **23** 493–513.

____ (1991). A new variational result for quasi-Newton formulae. *SIAM J. Optimization* **1** 18–21.

Friedland, S. (1981). Convex spectral functions. *Linear and Multilinear Algebra* **9** 299–316.

Hiriart-Urruty, J.-B. and D. Ye (1995). Sensitivity analysis of all eigenvalues of a symmetric matrix. *Numerical Math.* **70** 45–72.

Horn, R. A. and C. Johnson (1985). *Matrix Analysis*. Cambridge University Press, Cambridge, U.K.

Jarre, F. (1993). An interior-point method for minimizing the maximum eigenvalue of a linear combination of matrices. *SIAM J. Control and Optimization* **31** 1360–1377.

Kato, T. (1982). *A Short Introduction to Perturbation Theory for Linear Operators*. Springer-Verlag, New York.

Lewis, A. S. (1996). Convex analysis on the Hermitian matrices. *SIAM J. Optimization* **6** 164–177.

Marshall, A. W. and I. Olkin (1979). *Inequalities: Theory of Majorization and Its Applications*. Academic Press, NY.

Nesterov, Y. E. and A. S. Nemirovski (1994). *Interior Point Polynomial Methods in Convex Programming*. SIAM Publications, Philadelphia, PA.

Overton, M. L. (1988). On minimizing the maximum eigenvalue of a symmetric matrix. *SIAM J. Matrix Analysis and Appl.* **9** 256–268.

____ (1992). Large-scale optimization of eigenvalues. *SIAM J. Optimization* **2** 88–120.

____ and R. S. Womersley (1993). Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Math. Programming, Series B* **62** 321–357.

____ and ____ (1995). Second derivatives for optimizing eigenvalues of symmetric matrices. *SIAM J. Matrix Analysis and Its Appl.* **16** 697–718.

Polak, E. and Y. Wardi (1983). A nondifferentiable optimization algorithm for structural problems with eigenvalue inequality constraints. *J. Structural Mechanics* **11** 561–577.

Rendl, F. and H. Wolkowicz (1992). Applications of parametric programming and eigenvalue maximization to the quadratic assignment problem. *Math. Programming* **53** 63–78.

Shapiro, A. (1985). Extremal problems on the set of nonnegative definite matrices. *Linear Algebra and Its Appl.* **67** 7–18.

____ and M. K. H. Fan (1995). On eigenvalue optimization. *SIAM J. Optimization* **3** 552–568.

Tsing, N.-K., M. K. H. Fan, and E. I. Verriest (1994). On analyticity of functions involving eigenvalues. *Linear Algebra and Its Appl.* **207** 159–180.

Watson, G. A. (1991). An algorithm for optimal l_2 scaling of matrices. *IMA J. Numerical Analysis* **11** 481–492.

Wolkowicz, H. (1993). Explicit solutions for interval semidefinite linear programs. Technical Report CORR93-29, Department of Combinatorics and Optimization, University of Waterloo.

Yang, B. and R. J. Vanderbei (1993). The simplest semidefinite programs are trivial. Technical Report, Program in Statistics and Operations Research, Princeton University.

A. S. Lewis: Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1; e-mail: aslewis@orion.uwaterloo.ca