J. Austral. Math. Soc. Ser. B 28 (1986), 57–74

CONCISE REPRESENTATION OF GENERALISED GRADIENTS

M. R. OSBORNE[1], S. A. PRUESS[2] AND R. S. WOMERSLEY[3]

(Received 16 July 1985; revised 11 November 1985)

Abstract

Computing the generalised gradient directly using its standard definition can involve forming the convex hull of a very large number of vectors. Here an alternative concise parametrization is developed for the generalised gradient of the signed rank regression family of objective functions, a class of piecewise linear functions which includes both convex and nonconvex members. The approach uses the geometry of the epigraph explicitly and this suggests extensions to more general functions. A nondegeneracy condition is assumed which is natural in optimization problems.

1. Introduction

It is roughly ten years since Clarke introduced his concept of the generalised gradient of a locally Lipschitz function and since then considerable effort has been expended in deriving and classifying the similar constructs of possible interest. An excellent account of these developments is given in Rockafellar [9], and Clarke's work is described in detail in his 1983 book [3]. At about the same time, growing interest in the calculation of solutions to nondifferentiable optimization problems was evidenced by the publication of Mathematical Programming Study 3 (Balinski and Wolfe [1]). Thus a stimulating blend of application and technique was available, and the last decade has seen substantial progress. In particular we note the work of Fletcher ([4], Chapter 14) and Womersley [11] in

[1]Department of Statistics, IAS, Australian National University, Canberra 2601
[2]Department of Mathematics and Statistics, University of New Mexico
[3]School of Mathematics, Univ. of New South Wales, Kennington 2033

providing a unified approach to mathematical programming problems and composite nondifferentiable optimization problems. Womersley's approach makes extensive use of results developed in this paper.

Here we consider the important question of finding a representation of the generalised gradient that is appropriate from both the analytic and computational point of view. Our calculations are carried out for a particular family of nondifferentiable (actually piecewise linear) functions containing both convex and nonconvex examples which is of interest in developing robust rank based statistical estimation procedures. This family of functions has the general form of a piecewise linear, composite function

$$F(\mathbf{x}) = h(\mathbf{r}) = \sum_{i=1}^{n} \eta_i |\mathbf{r}|_{\mu(i)} \tag{1.1}$$

where

$$\mathbf{r} = M\mathbf{x} - \mathbf{f},$$

$M: R^p \to R^n$ is assumed to have rank p, $|\mathbf{r}|$ is the vector with components $|r_i|$, $i = 1, 2, \ldots, n$, μ is an index set ranking the components of $|\mathbf{r}|$ (we adopt the convention that the ranking is in increasing order of magnitude), and the η_i, $i = 1, 2, \ldots, n$ are scores which specify the particular realisation of $h(\mathbf{r})$. This function is important in statistical estimation as it provides the extension to the regression problem of the signed rank estimator of location (Randles and Wolfe [7]). Provided the η_i are nondecreasing functions of i then $F(\mathbf{x})$ is convex, but the resistance of the estimator to the effects of extreme outliers can be increased by reducing the size of the scores weighting the large residuals. If the scores then redescend, the resulting function is no longer convex. The problem that $F(\mathbf{x})$ is no longer convex is shared by the corresponding M-estimator (Huber [5]), and interest in studying it at all derives in part from evidence that the redescending M-estimators are of value. In fact the rank based estimators start with an advantage in not requiring an estimate of scale to be computed simultaneously.

EXAMPLE 1.1. In Figure 1, plots of F against x are given for the particular case

$$\left(\sum_{i=1}^{10} w_i \right) F(x) = \sum_{i=1}^{10} w_i |r|_{\mu(i)},$$

where

$$\begin{aligned} r_i &= 1 + xt_i - e^{t_i}, \\ w_i &= t_i(Y - t_i), \\ t_i &= .1 * i, \qquad i = 1,2,\ldots 10, \end{aligned}$$

for $Y = 2, 1, .9, .85, .8, .75, .7$. The scores η_i are nondecreasing for $Y = 2$, but redescend for $Y < 2$. They start becoming negative as Y is reduced below $Y = 1$. The loss of convexity as Y is decreased is clear.

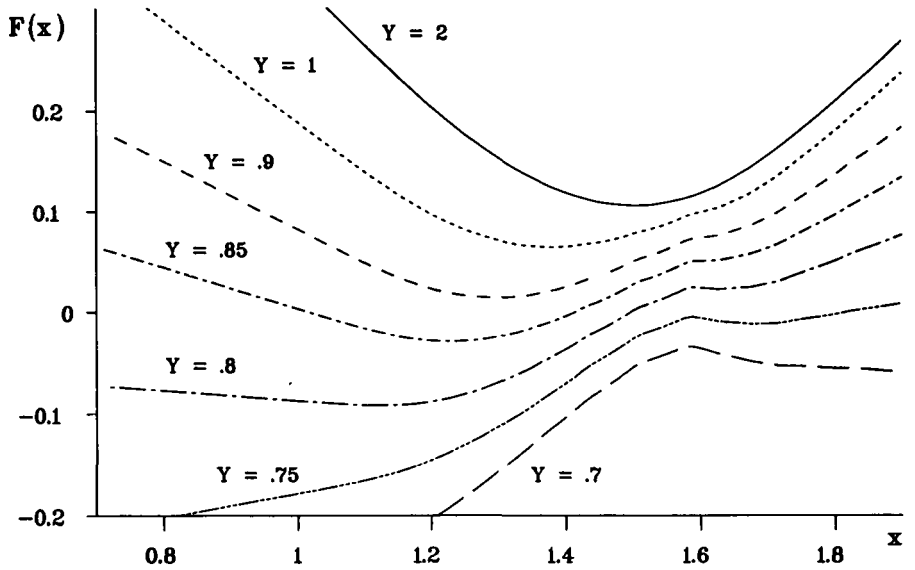

FIGURE 1. Example of a redescending signed rank function

REMARK 1.1. An interesting feature of Figure 1 is the apparent smoothness of this strictly piecewise linear curve. This derives from the property that, if the η_i are all distinct, then, whenever the graphs of the moduli of two residuals cross, the scores associated with them exchange, and the gradient of $F(\mathbf{x})$ is discontinuous. But there are as many as n^2 corners to the graph of $F(\mathbf{x})$, and this leads to a much smoother picture than that which would be obtained if all the η_i had been the same (the case of l_1 estimation which gives just n corners in the location problem). This apparent greater smoothness is closely related to the higher efficiencies attainable by rank estimates. But it is also associated with the difficulties in computing them, which has largely served as a deterrent to their use up till now.

Here the object of primary interest is the generalised gradient $\partial F(\mathbf{x})$ of the locally Lipschitz function F. It is defined by (for example, Clarke [3])

$$\partial F(\mathbf{x}) = \text{conv}\{\mathbf{z}^T \in R^p; \exists \text{ sequence } \{\mathbf{x}^{(k)}\} \tag{1.2}$$

such that

(i) $\{\mathbf{x}^{(k)}\} \to \mathbf{x}$,

(ii) $\nabla F(\mathbf{x}^{(k)})$ exists for all k and

(iii) $\nabla F(\mathbf{x}^{(k)}) \to \mathbf{z}^T$.}

It has the following properties.

(i) $\partial F(\mathbf{x})$ is a nonempty, compact, convex set in R^p.

https://doi.org/10.1017/S0334270000005191 Published online by Cambridge University Press

(ii) $\partial F(\mathbf{x})$ is upper semicontinuous. That is if $\{\mathbf{x}^{(k)}\} \to \mathbf{x}$, $\mathbf{z}^{(k)T} \in \partial F(\mathbf{x}^{(k)})$, and $\mathbf{z}^{(k)} \to \mathbf{z}$, then $\mathbf{z}^T \in \partial F(\mathbf{x})$.

(iii) The generalised gradient of the composite function $F(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{r}(\mathbf{x}))$ where f, $\mathbf{r}$ are smooth (the notation is chosen to stress that (1.1) is an important special case) is related to the generalised gradients of the component functions by

$$\partial F(\mathbf{x}) \subseteq \operatorname{conv}\{\mathbf{z}^T \in R^p;\ \mathbf{z}^T = \nabla f(\mathbf{x}) + \mathbf{w}^T M,\ \mathbf{w}^T \in \partial h(\mathbf{r}),\quad M = \nabla \mathbf{r}(\mathbf{x})\}. \tag{1.3}$$

(iv) The generalized gradient of the max function

$$F(\mathbf{x}) = \max_i \{f_i(\mathbf{x}),\quad i = 1,2,\ldots,m\}$$

is

$$\partial F(\mathbf{x}) = \operatorname{conv}\{\nabla f_i(\mathbf{x}),\ i \in \sigma,\ \sigma = \{i;\ f_i(\mathbf{x}) = F(\mathbf{x})\}\}. \tag{1.4}$$

It is important to note that (1.3) can be a strict inclusion, although this cannot happen in the convex case. However, in the nonconvex case it provides one source for a potential difficulty. This occurs when the set that is easy to construct may be larger than the set that is strictly appropriate for generating descent directions for minimization and for testing for optimality. This problem is considered in Section 3.

A second quantity of importance is an appropriate form of directional derivative for locally Lipschitz functions. Here two possibilities are considered. They are equivalent in the convex case.

(i) The usual one-sided directional derivative

$$F'(\mathbf{x}:\mathbf{t}) = \lim_{\alpha^{(k)} \to 0+} \frac{F(\mathbf{x} + \alpha^{(k)}\mathbf{t}) - F(\mathbf{x})}{\alpha^{(k)}}. \tag{1.5}$$

(ii) The generalised directional derivative in the sense of Clarke

$$F^0(\mathbf{x}:\mathbf{t}) = \limsup_{\substack{\mathbf{x}^{(k)} \to \mathbf{x} \\ \alpha^{(k)} \to 0+}} \frac{F(\mathbf{x}^{(k)} + \alpha^{(k)}\mathbf{t}) - F(\mathbf{x}^{(k)})}{\alpha^{(k)}}. \tag{1.6}$$

Relevant properties of F^0 include the following (Clarke [3]).

(i) $F^0(\mathbf{x}:\mathbf{t})$ is a positively homogeneous convex function of $\mathbf{t} \in R^p$.

(ii) $\partial F(\mathbf{x}) = \{\mathbf{v}^T \in R^p;\ F^0(\mathbf{x}:\mathbf{t}) \geqslant \mathbf{v}^T\mathbf{t},\ \forall \mathbf{t} \in R^p\}$.

(iii) $F^0(\mathbf{x}:\mathbf{t}) = \max_{\mathbf{v} \in \partial F(\mathbf{x})} \mathbf{v}^T\mathbf{t}$.

(iv) $F(\mathbf{x} + \mathbf{t}) \leqslant F(\mathbf{x}) + F^0(\mathbf{x}:\mathbf{t}) + o(\|t\|)$.

(v) $0 \in \partial F(\mathbf{x})$ (int $\partial F(\mathbf{x})$) if and only if

$$F^0(\mathbf{x}:\mathbf{t}) \geqslant 0 (> 0), \qquad \forall \mathbf{t} \in R^p,\ \mathbf{t} \neq 0.$$

Clearly F given by (1.1) is locally Lipschitz. To compute its generalised gradient using (1.2), it is necessary to have a means for referencing the tied and zero residuals which given rise to the nondifferentiability. We assume that there

https://doi.org/10.1017/S0334270000005191 Published online by Cambridge University Press

are n_g distinct groups of ties plus a group of zero residuals at $\mathbf{x}$, and we reference the ith group of ties by an index set ν_i, $|\nu_i| = k_i + 1$, $i = 1,2,\ldots n_g$, and the zero group by an index set ν_0, $|\nu_0| = k_0$. To reference individual elements it is convenient to assume an order imposed on each of the ν_i (so that $\nu_i(j)$ is the jth element of ν_i under this order). To set up the sequence $\{\mathbf{x}^{(k)}\}$ we must consider displacements to points $\mathbf{x} + \varepsilon\mathbf{t}$ at which F is differentiable. That is we must consider displacements $\varepsilon\mathbf{t}$, $\varepsilon > 0$ small enough, which break ties and remove zeros. This gives

$$\nabla F(\mathbf{x} + \varepsilon\mathbf{t}) = \mathbf{g}^T + \sum_{i=0}^{n_g} \sum_{j \in \nu_i} \eta_{\chi(j)} \nabla |\mathbf{r}|_j \tag{1.7}$$

where $\chi(j)$ is an operation which associates the correct score with r_j at $\mathbf{x} + \varepsilon\mathbf{t}$, and where $\mathbf{g}$ is the gradient of the differentiable part of F at $\mathbf{x}$. Because F is piecewise linear we can write (1.2), for $\varepsilon > 0$ small enough, as

$$\nabla F(\mathbf{x}) = \operatorname*{conv}_{\mathbf{t}} \{\nabla F(\mathbf{x} + \varepsilon\mathbf{t})\} \tag{1.8}$$

where the convex hull is taken over all $\mathbf{t}$ which produce distinct values of ∇F. Except in degenerate situations, $\mathbf{t}$ can be chosen to give all possible assignations of scores to residuals in the groups of ties, and all possible assignations both of sign and score to the residuals in the group of zeros. Thus the number of distinct vectors generating the convex hull representation (1.8) in the case of distinct scores is

$$2^{k_0} k_0! \prod_{i=1}^{n_g} (k_i + 1)!$$

REMARK 1.2. This calculation highlights two important defects of the convex hull form of the generalised gradient (1.8). These are:

(1) The convex hull form can be based on a very large number of vectors, so that the corresponding parametrization is very unwieldly; and

(2) The representation concentrates on points adjacent to $\mathbf{x}$ where F is smooth. Thus it is not giving direct insight into the structure of F which is contained in the tied and zero residuals.

Our aim is to develop a representation which overcomes both these objections. Thus we require it to be more concise (involving no more parameters than the natural dimension of the object described), and to reflect the natural geometry of the epigraph of F, by stressing the connection between the extreme points, edges, facets etc. on the one hand and the tied and zero residuals on the other. This representation is developed first for the convex members of our family of piecewise linear functions in the next section, and then extended to the nonconvex

https://doi.org/10.1017/S0334270000005191 Published online by Cambridge University Press

case in Section 3. Piecewise linearity is used explicitly. In particular, the property that the generalised gradient is constant on edges and facets of the epigraph of F (epi F is the set of points $\begin{bmatrix}\mathbf{x}\\ \pi\end{bmatrix}$ with $\pi \geqslant F(\mathbf{x})$) is often used rather than a more elaborate limiting argument based on upper semicontinuity of $\partial F(\mathbf{x})$. But the availability of this alternative suggests a means for extending the results to more general classes of functions.

In the convex case the results are closely related to those given in Osborne [6], where it is shown also that the representation is the right one for characterising the minimum of F and for use with descent methods of minimization, and where projected and reduced gradient algorithms are given which are directly applicable to our main example. However, this example is of independent interest (and not only for the reason that it is more complicated than those previously attempted). The extension to nonconvex problems is new. We believe that the use of generalised gradients simplifies our argument, but the convex case can be treated without this apparatus. Suitable references include Rockafellar [8] and Osborne [6].

2. Representation in the convex case

When F is convex (corresponding to the case in which the scores in (1.1) are nondecreasing) then the generalised gradient is also called the subdifferential. Here we construct a compact parametrization of the subdifferential of (1.1).

The important observation is that the points of nondifferentiability of F are determined by certain basic structural elements: (a) tied residuals, and (b) zero residuals. Making use of the index notation defined above in equation (1.7), we can summarise the connected sets of points such that $h(\mathbf{r})$ has a particular structure and is compatible with a particular ordering μ by means of the *structure equations*

$$\phi_{\nu_i(j)}(\mathbf{r}) \triangleq |\mathbf{r}|_{\nu_i(j+1)} - |\mathbf{r}|_{\nu_i(1)} = 0, \tag{2.1a}$$

$$j = 1, 2, \ldots, k_i, \qquad i = 1, 2, \ldots, n_g,$$

and

$$\phi_{\nu_0(j)}(\mathbf{r}) \triangleq r_{\nu_0(j)} = 0, \qquad j = 1, 2, \ldots, k_0. \tag{2.1b}$$

The ϕ_j are called *structure functionals*. Note that the set of structure functionals is irreducible in the sense that no smaller number expresses the correct configuration of ties and zeros and thus the nature of the nondifferentiability at $\mathbf{r}(\mathbf{x})$.

https://doi.org/10.1017/S0334270000005191 Published online by Cambridge University Press

REMARK 2.1. The definition of $\phi_{\nu_i(j)}$, $i \geqslant 1$ involves a specialised residual which we refer to as the *origin* of the ith group of tied residuals. Independence of the choice of origin is an important aspect of our results which is not discussed here. It is demonstrated for related problems in Clark and Osborne [2] and Osborne [6].

REMARK 2.2. It is convenient to assume that the sets of points satisfying the groups of structure equations have the obvious dimensions, and that independent displacements can be made from the current point to break individual ties or remove individual zeros. This is a form of nondegeneracy condition, and it can be expressed as a condition on the rank of a matrix. Let

$$V = \left[V_0 | V_1 | \cdots | V_{n_g}\right], \tag{2.2}$$

where the submatrices are derived from the different groups of structure functionals and have columns defined by

$$\begin{aligned} \kappa_j(V_i) &= \nabla_x \phi_{\nu_i(j)}(\mathbf{r})^T \qquad (2.3)\\ &= \theta_{\nu_i(j+1)} \kappa_{\nu_i(j+1)}(M^T) - \theta_{\nu_i(1)} \kappa_{\nu_i(1)}(M^T), \qquad i > 0,\\ &= \kappa_{\nu_0(j)}(M^T), \qquad i = 0, \qquad (2.4) \end{aligned}$$

where $\theta_j = \operatorname{sgn}(r_j)$. Then our nondegeneracy condition is

$$\operatorname{rank}(V) = \sum_{i=0}^{n_g} k_i = k \leqslant p. \tag{2.5}$$

Degeneracy has implications for the development of algorithms. An appropriate treatment in the context of descent methods is given in [6]. It is shown that the assumption is natural and involves no real restriction in practice.

Now let $\mathbf{w} \in \partial_r h(\mathbf{r})$, where the notation indicates the generalised gradient with respect to $\mathbf{r}$. Then, if $\mathbf{h}_i$, $i = 1,2,\ldots,m$, are the extreme points of $\partial_r h(\mathbf{r})$,

$$\mathbf{w} = \sum_{i=1}^{m} \lambda_i \mathbf{h}_i, \qquad \lambda_i \geqslant 0, \qquad \sum_{i=1}^{m} \lambda_i = 1. \tag{2.6}$$

The problem with this representation is that m can be very large. However, an alternative representation can be motivated by noting that (1.1) can be rewritten in terms of the structure functionals at $\mathbf{r}(\mathbf{x})$ to give

$$\begin{aligned} h(\mathbf{r}) = &\sum_{j \in \nu^c} \eta_{\chi(j)} |\mathbf{r}|_j + \sum_{i=1}^{n_g} \left(\sum_{j \in \nu_i} \eta_{\chi(j)} \right) |\mathbf{r}|_{\nu_i(1)}\\ &+ \sum_{i=1}^{n_g} \sum_{j \in \nu_i} \eta_{\chi(j)} \phi_j(\mathbf{r}) + \sum_{j \in \nu_0} \eta_{\chi(j)} \theta_j \phi_j(\mathbf{r}) \qquad (2.7) \end{aligned}$$

https://doi.org/10.1017/S0334270000005191 Published online by Cambridge University Press

where $\chi(j)$ is defined after (1.7), θ_j associates the correct sign with r_j, ν^c is the set of indices not contained in ν_i, $i = 0,1,\ldots,n_g$, and the part of $h(\mathbf{r})$ that does not involve the structure functionals is differentiable at $\mathbf{r}(\mathbf{x})$. The argument leading to (1.8) now shows that $\mathbf{w}$ can be written (to simplify notation where possible, we assume the index sets in (2.1) are mapped into $1,2,\ldots,k$ and define the matrix Φ by $\kappa_i(\Phi) = \nabla_r \phi_i^T$, $i = 1,2,\ldots,k$)

$$\mathbf{w} = \mathbf{h}^* + \sum_{i=1}^{k} u_i \kappa_i(\Phi) = \mathbf{h}^* + \Phi\mathbf{u}, \tag{2.8}$$

where $\mathbf{h}^*$ is the gradient of the differentiable part of h at $\mathbf{r}(\mathbf{x})$ and is given by

$$\mathbf{h}^* = \sum_{j\in\nu^c} \eta_{\chi(j)}\theta_j \mathbf{e}_j + \sum_{i=1}^{n_g}\left(\sum_{j\in\nu_i} \eta_{\chi(j)}\right)\theta_{\nu_i(1)}\mathbf{e}_{\nu_i(1)}. \tag{2.9}$$

It remains to determine the range of allowable values of $\mathbf{u}$. Specialising $\mathbf{w}$ to be $\mathbf{h}_i$, $i = 1,2,\ldots,m$ we obtain

$$\mathbf{h}_i = \mathbf{h}^* + \Phi\mathbf{u}^{(i)}, \qquad i = 1,2,\ldots,m. \tag{2.10}$$

The convex hull formula for ∂F now gives $\mathbf{u} \in U$ where the polyhedral convex set

$$U = \operatorname{conv}\{\mathbf{u}^{(i)}, \qquad i = 1,2,\ldots,m\} \subset R^k \tag{2.11}$$

is called the *constraint set associated with the subdifferential.*

REMARK 2.3 It is the property that U is necessarily polyhedral which is special to piecewise linear functions and makes this representation particularly attractive in this case.

To obtain $\partial_x F$ from $\partial_r h$ we use the chain rule (1.3)

$$\partial_x F = \partial_r h M$$

to obtain

$$\mathbf{z}^T \in \partial F(\mathbf{x}) \Leftrightarrow \mathbf{z} = \mathbf{g} + V\mathbf{u}, \tag{2.12}$$

where

$$\mathbf{g} = M^T\mathbf{h}^*,$$

$V = M^T\Phi$ is just (2.2), and $\mathbf{u} \in U$. The formula (2.12) is our concise parametrization of the subdifferential. The key to its utility is the specification of U, and we now show that U can be determined starting from a knowledge of $\mathbf{h}^*$ and Φ.

REMARK 2.4. It turns out that U depends on the geometry of epi F, and it is a useful first step to relate the differential properties to the basic descriptive quantities including extreme points, edges, facets, etc. The link is provided by the set of normal directions at $\mathbf{x}$

$$N(\mathbf{x}) \triangleq \left\{\begin{bmatrix}\mathbf{z}\\ -1\end{bmatrix}, \mathbf{z}^T \in \partial F(\mathbf{x})\right\}. \tag{2.13}$$

https://doi.org/10.1017/S0334270000005191 Published online by Cambridge University Press

Basic properties include

$$\dim(N) = \operatorname{rank}(V) + 1, \tag{2.14}$$

and

$$\dim(N) = p + 1 \Rightarrow \mathbf{x} \text{ is an extreme point,}$$
$$\dim(N) = p \Rightarrow \mathbf{x} \in E, \text{ an edge, and}$$
$$\dim(N) = p - 1 \Rightarrow \mathbf{x} \in G, \text{ a facet.}$$

The key tool in finding U is the directional derivative of F which is unambiguously defined in the convex case. Its utility stems directly from property (iii) of F^0. Given $\mathbf{t}$ let $\mathbf{z}^T \in \partial F(\mathbf{x})$. Then, using (2.12),

$$\mathbf{t}^T\mathbf{z} = \mathbf{t}^T\mathbf{g} + \mathbf{t}^T V\mathbf{u} \leqslant \mathbf{t}^T\mathbf{g} + \max_{\mathbf{u} \in U} \mathbf{t}^T V\mathbf{u} = F'(\mathbf{x} : \mathbf{t}) \tag{2.15}$$

This has the form of a linear inequality on $\mathbf{u} \in U$ and the set of all such inequalities provides the required description of the constraint set as a consequence of property (ii) of F^0. But U is polyhedral convex by (2.11) so we expect to be able to specialise the set of $\mathbf{t}$ required. In fact, it is necessary to consider only directions in the (relative) edges corresponding to connected sets of points containing $\mathbf{x}$ and having normal sets of dimension k, where $k + 1$ is the dimension of $N(\mathbf{x})$. To be specific, let $\mathbf{t}$ be a direction in the facet G_{ij} (connected set with normal set having dimension $k - 1$) having bounding edges E_i, E_j at $\mathbf{x}$. Then we can write

$$\mathbf{t} = \alpha\mathbf{t}_i + \beta\mathbf{t}_j$$

where $\mathbf{t}_i$, $\mathbf{t}_j$ are directions in E_i, E_j respectively, and α, $\beta \geqslant 0$ as a consequence of the convexity of F. Now linearity gives

$$F'(\mathbf{x} : \mathbf{t}) = \alpha F'(\mathbf{x} : \mathbf{t}_i) + \beta F'(\mathbf{x} : \mathbf{t}_j).$$

But this is compatible with the maximization expressed in (2.15) only if the particular $\mathbf{u}$ maximising (2.15) for $\mathbf{t}$ achieves the maximum for each of $\mathbf{t}_i$, $\mathbf{t}_j$ separately. Thus the inequality for $\mathbf{t}$ is a consequence of the inequalities for $\mathbf{t}_i$, $\mathbf{t}_j$ provided α, $\beta \geqslant 0$, showing that higher dimensional configurations than edges contribute no new information, provided epi F is convex.

To characterize the edges at $\mathbf{x}$, we turn to the alternative characterization of the geometry of epi F in terms of structure functionals. In particular, it follows from (2.14) that the different ways of reducing the dimension of $N(\mathbf{x})$ by 1 each correspond to removing one of the structure equations holding at $\mathbf{x}$. For the signed rank regression function this means either:

(a) relaxing one structure functional in a group of ties. The most general situation corresponds to the group splitting into two subgroups in the edge. One subgroup can retain the original origin, but a new origin must be found for the other subgroup, and the structure equation deleted expresses the tie at $\mathbf{x}$ between the two subgroup origins.

https://doi.org/10.1017/S0334270000005191 Published online by Cambridge University Press

(b) a subgroup of zero residuals relaxing to a nonzero group of tied residuals. An origin must be found for the subgroup, and this choice defines the structure equation deleted. In both cases the requirement to choose new subgroup origins on the edge E_q (say) means that the set of structure functionals on E_q may differ from that at $\mathbf{x}$. However, both must give a description of the same situation at $\mathbf{x}$ (that is, describe the same configuration of ties and zeros). To express this, let the structure functionals at $\mathbf{x}$ be $\phi_1, \phi_2, \ldots, \phi_k$, the functional being relaxed be ϕ_q, and the set of structure functionals on E_q be $\phi_1^{(q)}, \ldots, \phi_{k-1}^{(q)}$, with corresponding gradient matrix $\Phi^{(q)}$. The new situation that obtains on E_q (except at $\mathbf{x}$) is that the term involving ϕ_q in (2.7) becomes differentiable. It follows by an argument similar to that leading to (2.8) that the representations of ∂F at $\mathbf{x}$ and on E_q must be related by

$$\mathbf{h}_q = \mathbf{h}^* + \zeta_q \boldsymbol{\kappa}_q(\Phi), \tag{2.16a}$$

and

$$\left[\Phi^{(q)} | \boldsymbol{\kappa}_q(\Phi)\right] \left[\begin{array}{c|c} S_q & 0 \\ \hline \mathbf{s}_q^T & 1 \end{array}\right] = \Phi P \tag{2.16b}$$

where $\mathbf{h}_q^T$ is the gradient of the differentiable part of h on E_q, where ζ_q, S_q, $\mathbf{s}_q$ are defined by the transformation of the structure functionals, and P is a permutation matrix which takes account of any column rearrangement. To determine the direction $\mathbf{t}$ in the edge, note that it must satisfy

$$\phi_i^{(q)}(\mathbf{r} + \lambda M\mathbf{t}) = 0, \qquad i = 1,2,\ldots,k-1 \tag{2.17}$$

for $\lambda > 0$ small enough. Differentiating with respect to λ gives

$$\boldsymbol{\kappa}_i(\Phi^{(q)})^T M\mathbf{t} = \boldsymbol{\kappa}_i(V^{(q)})^T \mathbf{t} = 0, \qquad i = 1,2,\ldots,k-1. \tag{2.18}$$

Also $\phi_q(\mathbf{r} + \lambda\mathbf{t}) \neq 0$ so that (setting $\mathbf{v}_q = \boldsymbol{\kappa}_q(V)$)

$$\boldsymbol{\kappa}_q(\Phi)^T M\mathbf{t} = \mathbf{v}_q^T \mathbf{t} \neq 0. \tag{2.19}$$

REMARK 2.5. The following calculation expresses ζ_q in terms of the transformation of the structure functionals by calculating $F'(\mathbf{x}:\mathbf{t})$ on E_q.

$$\mathbf{z}^T \in \partial F(E_q) \Leftrightarrow \mathbf{z} = \mathbf{g}_q + V^{(q)}\mathbf{u}_q, \qquad \mathbf{u}_q \in U^{(q)},$$

so that, as $\mathbf{t}$ in E_q satisfies (2.18), (2.19),

$$F'(\mathbf{x}:\mathbf{t}) = \mathbf{g}_q^T \mathbf{t}. \tag{2.20}$$

Also $\mathbf{z}^T \in \partial F(E_q) \Rightarrow \mathbf{z}^T \in \partial F(\mathbf{x})$ by the upper semi continuity of ∂F. In particular, $\partial F(E_q) \cap \partial F(\mathbf{x})$ defines a subset $\bar{U}$ of U. In addition, from (2.16)

$$V = V^{(q)}\left[S_q | 0\right] P^{-1} + \mathbf{v}_q \left[\mathbf{s}_q^T | 1\right] P^{-1}. \tag{2.21}$$

https://doi.org/10.1017/S0334270000005191 Published online by Cambridge University Press

Thus, as $F'(\mathbf{x}:\mathbf{t})$ is independent of the choice of $\mathbf{z} \in \partial F(E_q)$,

$$\begin{aligned} F'(\mathbf{x}:\mathbf{t}) &= \mathbf{g}^T\mathbf{t} + \mathbf{u}^T V^T\mathbf{t}, \qquad \mathbf{u} \in \overline{U} \\ &= \mathbf{g}^T\mathbf{t} + \mathbf{v}_q^T\mathbf{t}\left[\mathbf{s}_q^T|1\right] P^{-1}\mathbf{u}, \qquad \mathbf{u} \in \overline{U}. \end{aligned} \tag{2.22}$$

It follows that

$$\left[\mathbf{s}_q^T|1\right] P^{-1}\mathbf{u} = \zeta_q,$$

independent of the particular choice of $\mathbf{u} \in \overline{U}$.

It is an immediate consequence of (2.16a) and the chain rule (compare also (2.22)) that

$$\mathbf{g}_q = \mathbf{g} + \zeta_q \mathbf{v}_q. \tag{2.23}$$

The desired inequality for U now follows. From (2.22)

$$\mathbf{g}^T\mathbf{t} + \left(\mathbf{v}_q^T\mathbf{t}\right)\left[\mathbf{s}_q^T|1\right] P^{-1}\mathbf{u} \leqslant F'(\mathbf{x}:\mathbf{t}), \qquad \forall \mathbf{u} \in U.$$

But (2.23) gives

$$F'(\mathbf{x}:\mathbf{t}) = \mathbf{g}^T\mathbf{t} + \zeta_q\left(\mathbf{v}_q^T\mathbf{t}\right),$$

so that

$$\operatorname{sgn}\left(\mathbf{v}_q^T\mathbf{t}\right)\left\{\zeta_q - \left[\mathbf{s}_q^T|1\right] P^{-1}\mathbf{u}\right\} \geqslant 0, \qquad \forall \mathbf{u} \in U. \tag{2.24}$$

It should be noted that (2.18), (2.19) need not specify $\mathbf{t}$ uniquely. However, (2.24) is independent of the manner in which the specification of $\mathbf{t}$ is completed.

There are two cases to consider in deriving these inequalities for the signed rank regression function.

Case (a) Let $\phi_1, \ldots, \phi_k$ characterize a group of tied residuals pointed to by an index set which is specialized to ν. We set $\nu = \nu_1 \cup \nu_2$ where ν_1 points to the subgroup tied to the origin of ν and ν_2 points to the new subgroup. Then the new structure functionals are

$$\begin{aligned} \phi_i^{(q)} &= |\mathbf{r}|_{\nu_2(i+1)} - |\mathbf{r}|_{\nu_2(1)} \\ &= |\mathbf{r}|_{\nu_2(i+1)} - |\mathbf{r}|_{\nu_1(1)} - \left(|\mathbf{r}|_{\nu_2(1)} - |\mathbf{r}|_{\nu_1(1)}\right) \\ &= \phi_{\xi(i)} - \phi_q, \qquad i = 1,2,\ldots|\nu_2| - 1, \end{aligned} \tag{2.25}$$

where ξ maps ν_2 back into ν. Thus the mapping (2.16) is

$$\left[\Phi_1|\Phi_2^{(q)}|\kappa_q(\Phi)\right]\begin{bmatrix} I_1 & 0 & 0 \\ 0 & I_2 & 0 \\ 0 & \mathbf{e}^T & 1 \end{bmatrix} = \Phi P, \tag{2.26}$$

https://doi.org/10.1017/S0334270000005191 Published online by Cambridge University Press

where Φ_1 corresponds to the unchanged structure functionals, I_1 and I_2 are unit matrices of appropriate dimension, and $\mathbf{e}$ is a vector of ones. It remains to estimate ζ_q. The argument is familiar. We write (1.7) in terms of the $\nabla\phi_i$ by adding and subtracting origin terms and this shows a contribution to $\mathbf{g}$ at $\mathbf{x}$ from the group of ties given by

$$\Delta\mathbf{g} = \Big(\sum_{i\in\nu}\eta_{\chi(i)}\Big)\theta_{\nu(1)}\boldsymbol{\kappa}_{\nu(1)}(M^T).$$

Applying this argument on E_q, and noting that there are contributions from both subgroups, we obtain

$$\begin{aligned}\Delta\mathbf{g}_q = \Delta\mathbf{g}_1 + \Delta\mathbf{g}_2 &= \Big(\sum_{i\in\nu_1}\eta_{\chi(i)}\Big)\theta_{\nu_1(1)}\boldsymbol{\kappa}_{\nu_1(1)}(M^T)\\ &+\Big(\sum_{i\in\nu_2}\eta_{\chi(i)}\Big)\theta_{\nu_2(1)}\boldsymbol{\kappa}_{\nu_2(1)}(M^T)\\ &= \Delta\mathbf{g} + \Big(\sum_{i\in\nu_2}\eta_{\chi(i)}\Big)\mathbf{v}_q.\end{aligned}$$

This expression defines ζ_q, but there are still two possibilities, because if $\mathbf{v}_q^T\mathbf{t} > 0$ then the second subgroup grows in magnitude relative to the first, while if $\mathbf{v}_q^T\mathbf{t} < 0$ then this behaviour is reversed. As scores are assigned according to rank we have (setting $\chi(\nu_1(1)) = l$, $s = l + k$)

$$\zeta_q^+ = \sum_{j=s-|\nu_2|+1}^{s}\eta_j, \qquad \mathbf{v}_q^T\mathbf{t} > 0, \tag{2.27a}$$

$$\zeta_q^- = \sum_{j=l}^{l+|\nu_2|-1}\eta_j, \qquad \mathbf{v}_q^T\mathbf{t} < 0. \tag{2.27b}$$

Inserting the appropriate quantities from (2.26), (2.27) into (2.24), noting that all possible ways of splitting up a group of ties into two subgroups are allowed and that $\mathbf{v}_q^T\mathbf{t}$ can be chosen to have either sign in (2.19), we obtain the system of inequalities

$$\sum_{j=1}^{J}\eta_{\chi(\nu(1))+J-1} \leqslant \sum_{j=1}^{J}u_{\omega(J)} \leqslant \sum_{j=1}^{J}\eta_{\chi(\nu(k+1))-j+1} \tag{2.28}$$

for $J = 1,2,\ldots,k+1$, and ω any selection of J indices from $\nu(1)$, $\nu(2)\cdots\nu(k+1)$.

Case (b) Let ϕ_1, $\phi_2\cdots\phi_k$ characterize a group of zero residuals. Again ν is the associated index set and the splitting into subgroups defined by the partition

https://doi.org/10.1017/S0334270000005191 Published online by Cambridge University Press

$\nu = \nu_1 \cup \nu_2$. The structure functionals associated with the second subgroup are

$$\begin{aligned}\phi_i^{(q)} &= |\mathbf{r}|_{\nu_2(i+1)} - |\mathbf{r}|_{\nu_2(1)}, \\ &= \theta_{\xi(i)}\phi_{\xi(i)} - \theta_q\phi_q,\end{aligned} \tag{2.29}$$

and it is important to note that $\mathbf{t}$ can be chosen in E to give any selected pattern of signs for the $\theta_j = \operatorname{sgn}(r_j(\mathbf{x} + \varepsilon\mathbf{t}))$, $j \in \nu_2$, $\varepsilon > 0$ small enough by satisfying the appropriate system of equations (2.18), (2.19). Identifying terms in (2.16) gives

$$S_q = \left[\begin{array}{c|c} I_1 & \\ \hline & \operatorname{diag}\{\theta_{\xi(i)}, i = 1, 2, \ldots |\nu_2| - 1\} \end{array}\right], \tag{2.30a}$$

$$\mathbf{s}_q^T = \theta_q\left[\theta_{\xi(1)}, \ldots, \theta_{\xi(|\nu_2|-1)}\right]. \tag{2.30b}$$

Also, because the subgroup of nonzero ties must be associated with the larger scores, the argument leading to (2.27) gives

$$\zeta_q = \theta_q \sum_{j=1}^{|\nu_2|} \eta_{\chi(\nu(k+1))-j+1}. \tag{2.31}$$

Substituting from (2.30), (2.31) into (2.24) and noting that all combinations of signs are possible for θ_j, θ_q we obtain the inequalities

$$\sum_{j=1}^{J} |u_{\omega(j)}| \leqslant \sum_{j=1}^{J} \eta_{\chi(\nu(k+1))-j+1} \tag{2.32}$$

for $J = 1,2,\ldots, k+1$, and ω any selections of J indices from $\nu(1), \nu(2), \ldots, \nu(k+1)$. In the case of a single zero residual this corresponds to the usual result that $\partial|x| = [-1, 1]$ when $x = 0$.

REMARK 2.6 All the results needed to put together the concise representation of $\partial F(\mathbf{x})$ have now been derived. The general form is given in (2.12). To construct $\mathbf{g}$ requires $\mathbf{h}^*$ specified in (2.9), V is given in (2.2), (2.3), and the inequalities (2.28) and (2.32) specify the constraint set U.

3. Representation in the nonconvex case

Here we consider the manner in which the approach used in the previous section extends to the non-convex case corresponding to redescending η_i in (1.1). Certainly some problems occur, and the referee suggested that this section should be called 'difficulties in the nonconvex case', but we prefer to be more sanguine. Some progress can be made as the ideas of structure functional and normal set can be employed to characterize the geometry of epi F. The key requirement is a

https://doi.org/10.1017/S0334270000005191 Published online by Cambridge University Press

suitable construct to replace the directional derivative, and for our purposes this is provided for locally Lipschitz functions by Clarke's *generalised directional derivative*.

REMARK 3.1. F^0 is computable for the signed rank regression function (1.1). The only complexity occurs in the groups of ties where the procedure is as follows (for each group):

(i) rank the $\eta_{\chi(j)}$, $j \in \nu_i$,

(ii) rank the $\theta_j \mathbf{t}^T \kappa_j(M^T)$, $j \in \nu_i$, and

(iii) accumulate the products of corresponding terms.

This construction works in the nondegenerate case because then there is a displacement to an adjacent point which achieves any desired order for the residuals within the group, in particular an ordering which matches the terms in (i) and (ii).

It would seem that the development of the previous section could be paralleled, in particular to specify the constraint set U. But there is one problem because $\mathbf{t}$, a direction in a facet, may not be expressible as a combination of directions in the bounding edges with *positive* weights. Thus the edges may not completely specify $\partial F(\mathbf{x})$ so that using just the constraints derived from the edges may give too large a set.

EXAMPLE 3.1. This positive-weight condition is essential. Geometrically it ensures that epi F does not have reentrant corners. If this condition does not hold then it is easy to give an example where the inclusion is proper. Consider

$$\begin{aligned} F(x, y) &= \max\{x, -x - y\}, \qquad y \geq 0, \\ &= \max\{x, -x + y\}, \qquad y < 0, \end{aligned}$$

in which the graph of F is a piecewise linear saddle (see Figure 2). Then the convex hull form of the generalised gradient at $x = 0$, $y = 0$ is given by

$$\partial F(0)^T = \operatorname{conv}\left\{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}-1\\-1\end{bmatrix}, \begin{bmatrix}-1\\1\end{bmatrix}\right\}.$$

A compact representation is obtained by selecting structure functionals which describe the discontinuities in gradient for $y \geq 0$ and $y \leq 0$ respectively. An appropriate choice is

$$\phi_1 = -2x - y, \qquad \phi_2 = -2x + y.$$

Setting

$$\partial F(0)^T = \begin{bmatrix}1\\0\end{bmatrix} + \begin{bmatrix}-2 & -2\\-1 & 1\end{bmatrix}\mathbf{u}, \qquad \mathbf{u} \in U,$$

we can obtain a representation of U by evaluating the $\mathbf{u}_i$ expressing the extreme

https://doi.org/10.1017/S0334270000005191 Published online by Cambridge University Press

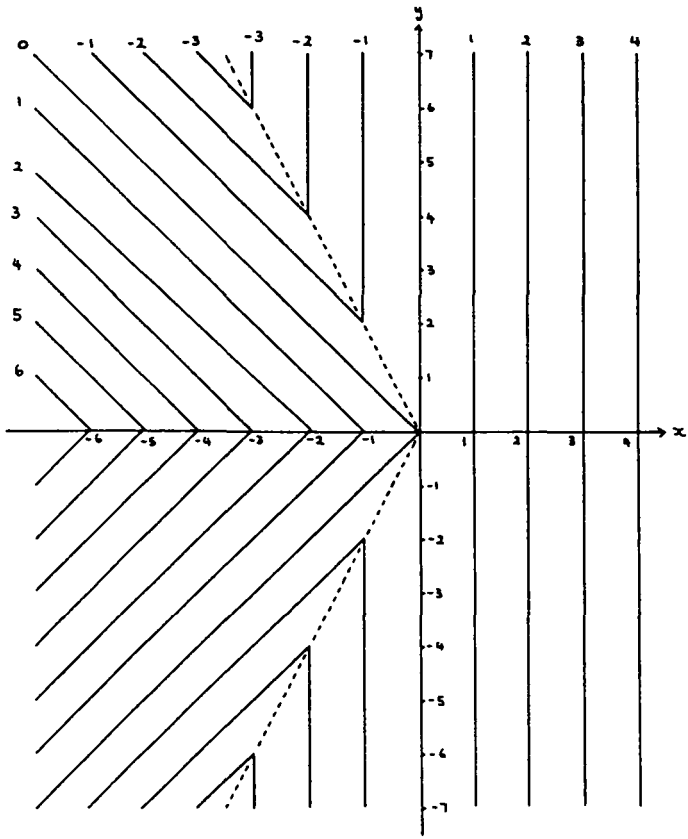

FIGURE 2. Contours of Example 3.1

points in the convex hull form. This gives

$$\begin{bmatrix}1\\0\end{bmatrix} \to \mathbf{u}_1 = \begin{bmatrix}0\\0\end{bmatrix},$$

$$\begin{bmatrix}-1\\-1\end{bmatrix} \to \mathbf{u}_2 = \begin{bmatrix}1\\0\end{bmatrix},$$

$$\begin{bmatrix}-1\\1\end{bmatrix} \to \mathbf{u}_3 = \begin{bmatrix}0\\1\end{bmatrix}.$$

The edges $E_i \in R^3$ at $F = 0$, $x = 0$, $y = 0$ and the corresponding directions $\mathbf{t}_i$ are

$$E_1 = \{F = -x, y = 0\}, \qquad \mathbf{t}_1 = \begin{bmatrix}-1\\0\end{bmatrix},$$

$$E_2 = \{F = x, -2x - y = 0\}, \qquad \mathbf{t}_2 = \begin{bmatrix}-1/2\\1\end{bmatrix}$$

$$E_3 = \{F = x, -2x + y = 0\}, \qquad \mathbf{t}_3 = \begin{bmatrix}-1/2\\-1\end{bmatrix}.$$

https://doi.org/10.1017/S0334270000005191 Published online by Cambridge University Press

Property (iii) of the generalised directional derivative now gives the constraints determined by the edges:

$$F^0(0:\mathbf{t}_1) = 1 \geqslant -1 + 2(u_1 + u_2),$$
$$F^0(0:\mathbf{t}_2) = 3/2 \geqslant -1/2 + 2u_2,$$
$$F^0(0:\mathbf{t}_3) = 3/2 \geqslant -1/2 + 2u_1.$$

Clearly these inequalities specify a large set than the convex hull of $\mathbf{u}_i$, $i = 1,2,3$ determined above.

Also the use of F^0 does lead to results which can be counter intuitive if the connection with the directional derivative is stressed. Necessarily we have

$$F'(\mathbf{x}:\mathbf{t}) \leqslant F^0(\mathbf{x}:\mathbf{t}) \tag{3.1}$$

and strict inequality can hold. The above calculations suffice to show this. For example we have

$$F^0(0:\mathbf{t}_2) = 3/2 > F'(0:\mathbf{t}_2) = -\tfrac{1}{2},$$
$$F^0(0:\mathbf{t}_3) = 3/2 > F'(0:\mathbf{t}_3) = -\tfrac{1}{2}.$$

Note that $0 \in \partial F(0)$ but 0 is a stationary point which is neither a maximum nor a minimum. Here this behaviour is a special case of property (v) of F^0. The difficulty comes about because, although (1.2) shows clearly enough that ∂F is the right kind of set to be considering in the sense that it is likely to know about descent directions at the current point, in certain circumstances it is too large a set.

The directional derivative F' is a more precise quantity for verifying descent directions. The class of functions for which $F^0 = F'$ is called *regular*. It inherits most of the nice properties of the convex case. In particular, it is a consequence of property (v) of F^0 that for piecewise linear regular functions $0 \in \partial F(\mathbf{x}) \Rightarrow \mathbf{x}$ is a local minimum in the sense that F is nondecreasing along straight lines at $\mathbf{x}$. But if F is not regular at $\mathbf{x}$ then $0 \in \partial F(\mathbf{x})$ gives a more general class of stationary point at $\mathbf{x}$ and further work is required to characterise it. The above discussion shows that concave functions cannot be regular at points of nondifferentiability, and figure 1 shows that redescending signed rank estimators cannot be regular everywhere. But a further difficulty is introduced because now the chain rule for generalised gradients (property (iii)) is only an inclusion, and the inclusion can be strict [12]. Thus we again have the problem that the convenient specification may lead to too big a set if F is not regular.

Clearly regularity is a desirable property so that criteria for regularity are important. One such is that if F can be represented in a neighbourhood of $\mathbf{x}$ as the maximum over a finite number of smooth functions then F is regular. Such functions are called locally max. Let

$$F(\mathbf{x}) = \max_i \left(f_i(\mathbf{x}), \qquad i = 1,2,\ldots m\right) \tag{3.2}$$

https://doi.org/10.1017/S0334270000005191 Published online by Cambridge University Press

and

$$\sigma = \{i;\ f_i(\mathbf{x}) = F(\mathbf{x})\}.$$

Then

$$\begin{aligned} F'(\mathbf{x}:\mathbf{t}) &= \frac{F(\mathbf{x}+\alpha\mathbf{t}) - F(\mathbf{x})}{\alpha} + o(1) \\ &\geqslant \mathbf{z}^T\mathbf{t} + 0(\alpha)(\mathbf{z} \in \partial F(\mathbf{x})) \\ &\geqslant F^0(\mathbf{x}:\mathbf{t}) + 0(\alpha), \end{aligned}$$

as $\partial F(\mathbf{x}) = \operatorname{conv}\{\nabla f_i(\mathbf{x}),\ i \in \sigma\}$ and $\max_i\{f_i(\mathbf{x}+\alpha\mathbf{t})\}$ can be attained for $i \notin \sigma$ if $\alpha > 0$. The regularity of $F(\mathbf{x})$ now follows from (3.1).

REMARK 3.2 For locally max functions the obvious structure functionals are

$$\phi_i = f_{\sigma(i+1)} - f_{\sigma(1)}, \qquad i = 1,2,\ldots|\sigma| - 1.$$

Taking

$$\mathbf{g}^T = \nabla f_{\sigma(1)}(\mathbf{x}),$$

$$\kappa_i(V) = \nabla\phi_i, \qquad i = 1,2,\ldots,|\sigma| - 1$$

we obtain the representation of the generalised gradient as

$$\mathbf{z}^T \in \partial F(\mathbf{x}) \Rightarrow \mathbf{z} = \mathbf{g} + V\mathbf{u}.$$

In this case it is not necessary to assume nondegeneracy in order to characterise U as this representation must be equivalent to the convex hull representation (1.4), so that

$$U = \left\{\mathbf{u};\ u_i \geqslant 0, i = 1,2,\ldots|\sigma| - 1,\ \sum_{i=1}^{|\sigma|-1} u_i = 1\right\}.$$

Optimality properties of locally max functions have been discussed in [10].

The results given in this Section are fragmentary but they do suggest a number of interesting problems for further research. One investigation we hope to undertake involves the modification of our reduced-gradient algorithm for minimizing (1.1) in the convex case. Our aim is to gather empirical data concerning its behaviour in the nonconvex case for score functions which make sense in practical estimation problems (this is not the case for the score functions chosen to produce figure 1 when $Y < 1$). Our expectation is that reentrant behaviour of epi F is unlikely in a neighbourhood of the global minimum if the estimation procedure is sensible. Also, if epi F is not reentrant at x then locally F can be represented as the maximum of its linear pieces so that it must be regular at x (and this argument clearly has some range of validity even if F is not piecewise linear). In these circumstances our concise formula for ∂F is exact, not just an inclusion.

https://doi.org/10.1017/S0334270000005191 Published online by Cambridge University Press

References

[1] M. L. Balinski and P. Wolfe, Nondifferentiable optimization, *Math. Programming Stud.* 3 (North Holland, 1975).
[2] D. I. Clark and M. R. Osborne, "A descent algorithm for minimizing polyhedral convex functions," *SIAM J. Sci. Statist. Comput.* 4 (1983), 757–786.
[3] F. H. Clarke, *Optimisation and nonsmooth analysis* (Wiley, New York, 1983).
[4] R. Fletcher, *Practical methods of optimisation*, 2 (Wiley, Chichester, 1981).
[5] P. J. Huber, *Robust statistics* (Wiley, New York, 1981).
[6] M. R. Osborne, *Finite algorithms in optimisation and data analysis* (Wiley, Chichester, 1985).
[7] R. H. Randles and D. A. Wolfe, *Introduction to the theory of nonparametric statistics* (Wiley, New York, 1979).
[8] R. T. Rockafellar, *Convex analysis* (Princeton University Press, 1970).
[9] R. T. Rockafellar, *The theory of subgradients and its application to problems of optimization: Convex and non convex functions* (Heldermann-Verlag, Berlin, 1981).
[10] R. S. Womersley, "Optimality conditions for piecewise smooth functions", *Math. Programming Stud.* 17 (1982), 13–27.
[11] R. S. Womersley, "Local properties of algorithms for minimizing nonsmooth composite functions", *Math. Programming* 32 (1985), 69–89.
[12] R. S. Womersley, "Censored discrete l_1 approximation", *SIAM J. Sci. Statist. Comput.*, to appear.