When LP is not a good idea – using structure in polyhedral optimization problems

M.R.Osborne
Mathematical Sciences Institute, Australian National University
ACT 0200, Australia

Abstract

It has been known for almost 50 years [15] that the discrete l_1 approximation problem can be solved by linear programming. However, improved algorithms involve a step which can be interpreted as a line search, and which is not part of the standard solution procedures. This is the simplest example of a class of problems with a structure distinctly more complicated than that of the so-called nondegenerate linear programs. Our aim is to uncover this structure for these more general polyhedral functions and to show that it can be used it to develop what are recognizably algorithms of simplicial type for minimizing them. A key component of this work is a compact description of polyhedral convex functions described in some detail in [11], and this can be applied also in the development of active set type methods in polyhedral function constrained optimization problems. Applications include the development of new algorithms for problems which include problems in statistical estimation and data mining.

1 Introduction

A convex function is the supremum of an affine family:

$$f(\mathbf{x}) = \sup_{i\in\sigma} \mathbf{c}_i^T\mathbf{x} - d_i. \tag{1}$$

If the index set σ is finite then $f(\mathbf{x})$ is polyhedral. The problem of minimizing a polyhedral convex function (PCF) $f(\mathbf{x})$ over a polyhedral set $A\mathbf{x} \geq \mathbf{b}$ can always be written as a linear program (LP):

$$\min_{A\mathbf{x}\geq\mathbf{b}} h;\ h \geq \mathbf{c}_i^T\mathbf{x} - d_i,\ i \in \sigma. \tag{2}$$

Linear programming can be regarded as the simplest example of a PCF minimization problem. Certainly it is the best known as a result of its extensive use in applications. It has the generic form

$$\min_{\mathbf{x}\in X} \mathbf{c}^T\mathbf{x};\ X = \{\mathbf{x} : A\mathbf{x} \geq \mathbf{b}\}$$

where $A : R^p \rightarrow R^n$, $p < n$. Note that it can be written also as

$$\min_{\mathbf{x}} F(\mathbf{x});\ F(\mathbf{x}) = F_1(\mathbf{x}) + F_2(\mathbf{x}).$$
$$F_1(\mathbf{x}) = \mathbf{c}^T\mathbf{x}, \text{ type 1 PCF},$$
$$F_2(\mathbf{x}) = \delta(\mathbf{x} : X), \text{ type 2 PCF}.$$

The type 2 PCF is an indicator function which builds vertical walls in R^{p+1} above the plan of the constraint set in R^p. The Kuhn-Tucker conditions characterize the optimum:

$$\mathbf{c}^T = \mathbf{u}^T A,$$
$$u_i \geq 0,\ u_i(A_{i*}\mathbf{x} - b_i) = 0,\ i = 1, 2, \cdots, n,$$

where "Matlab like" notation is used to identify matrix rows and columns.

This linear program supports a simple picture! This is illustrated in the following figure which shows a corner of the epigraph of the objective function sitting above the constraint set in R^2.

Here three faces of the epigraph intersect at the indicated extreme point $\mathbf{x} \in R^2$, and in this picture each simplex step can move along an edge only to the adjacent extreme point where it hits one of the walls built around the epigraph by the indicator function of the constraint set. This illustrates the point that a line search step is not a part of the basic simplex algorithm. The traditional problem of degeneracy corresponds here to more than three faces intersecting at an extreme point. Thus a degenerate vertex is in this sense overdetermined. Problems arise in naive implementations which select a subset of the active constraints according to some a priori rule in order to generate a descent edge as the resulting direction may immediately violate one of the ignored active constraints.

A rich source of problems possessing an inherently more complex structure arise in discrete estimation. Here we consider algorithms for linear estimation problems which are characterised by:

1 The epigraph of the function is generically degenerate in the sense of linear programming - remember that the problem of minimizing any PCF can always be written as a linear program.

2 There is a well defined set of necessary conditions which describe the problem optimum and which can be taken here as defining an appropriate sense of nondegeneracy.

Let the linear model be

$$\mathbf{r} = A\mathbf{x} - \mathbf{b}. \tag{3}$$

It is assumed that $\text{rank}(A) = p$, and that this suffices to guarantee a bounded optimum. Associated with extreme points of the epigraph are appropriate sets of algebraic conditions specifying the faces that intersect there. Typically these involve a subset of the equations specifying the linear model, and we refer to this subset as the "active set" at $\mathbf{x}_\sigma$ where σ is the index set pointing to the rows $\{A_\sigma, \mathbf{b}_\sigma\}$ corresponding to the active components in the linear model.

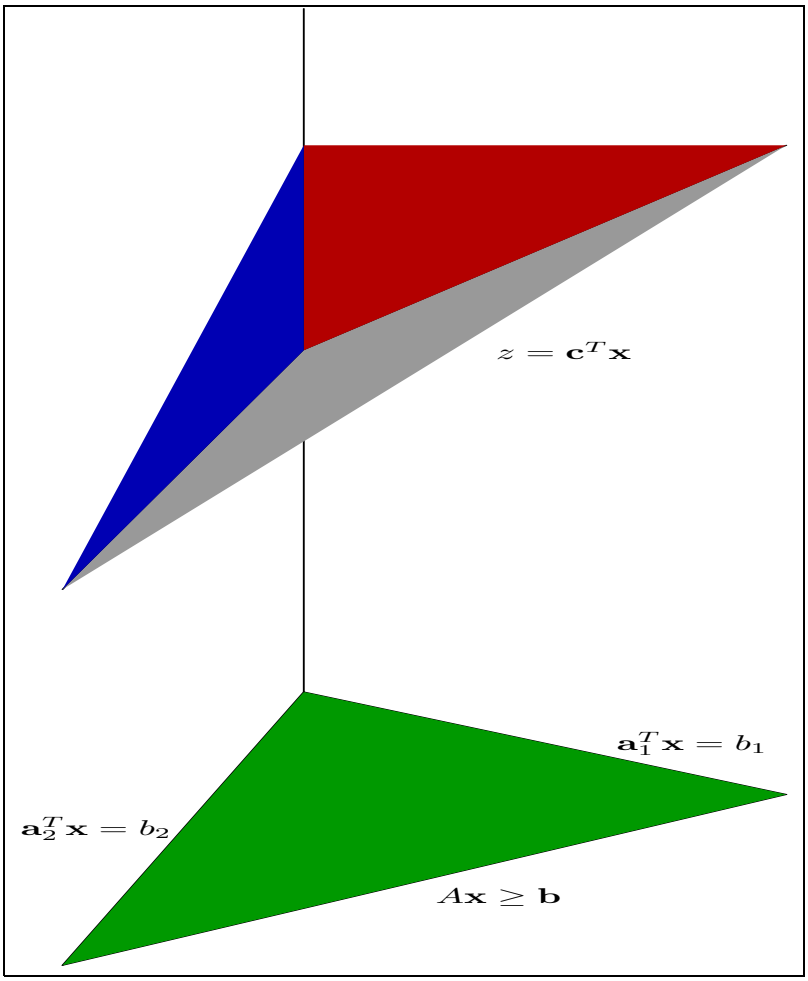

Table 1: nondegenerate linear program

Example 1.1 *l_1 estimation. This estimation problem has a long history and has attracted recent attention because of its robustness properties:*

$$\min_{\mathbf{x}} \sum |r_i|, \quad \mathbf{r} = A\mathbf{x} - \mathbf{b}, \quad A : R^p \rightarrow R^n.$$

This corresponds to a PCF with the defining affine family specified by

$$\begin{aligned} \mathbf{c}_j^T &= \boldsymbol{\theta}_j^T A, \\ d_j &= \boldsymbol{\theta}_j^T \mathbf{b}, \end{aligned}$$

where $\boldsymbol{\theta}_j$, $j = 1, 2, \cdots, 2^n$ has the form of one realization from among the possibilities:

$$[\pm 1, \pm 1, \cdots, \pm 1]^T.$$

The cause of the nonsmoothness of the epigraph stems from the ambiguity in allocating the signs associated with zero residuals. Thus extreme points will be characterized by sets of (at least) p zero residuals. The necessary conditions characterizing the optimum extreme point are:

$$\begin{aligned} 0 &= \sum_{i \in \sigma^C} \theta_i A_{i*} + \sum_{i \in \sigma} u_i A_{i*}, \\ \theta_i &= sign(r_i), \ r_i \neq 0, \\ \sigma &= \{i; r_i = 0\}, \\ |u_i| &\leq 1, i \in \sigma. \end{aligned}$$

In the case $p = 2$ a typical extreme point could be characterized by (say)

$$\begin{aligned} \pm r_1(x_1, x_2) &= 0, \\ \pm r_2(x_1, x_2) &= 0. \end{aligned}$$

Here four faces of the epigraph intersect at each extreme point (LP expects 3). Each face can be picked out by the assignation $\pm 1 \Rightarrow \theta_i$ such that directions into the face satisfy

$$\begin{aligned} \theta_1 A_{1*} \mathbf{t} &= \lambda_1 > 0, \\ \theta_2 A_{2*} \mathbf{t} &= \lambda_2 > 0. \end{aligned}$$

for convex combination of edge directions. Now the ambiguity of signs associated with the zero residuals makes more sense. This resolution is illustrated in figure 2.

This figure shows also that the l_1 problem typically supports a linesearch. For example, the line $r_2 = 0$ defines two edges of the epigraph which join at the extreme point $r_1 = r_2 = 0$ so a search for a minimum in the corresponding direction is possible. What happens when moving through the extreme point in this direction is an increase in directional derivative, and this is illustrated in a one dimensional example in the next figure. The function in figure 3 is

$$f(x) = |x| + |x - 1| + |x - 2| + |x - 3| + |x - 4|$$

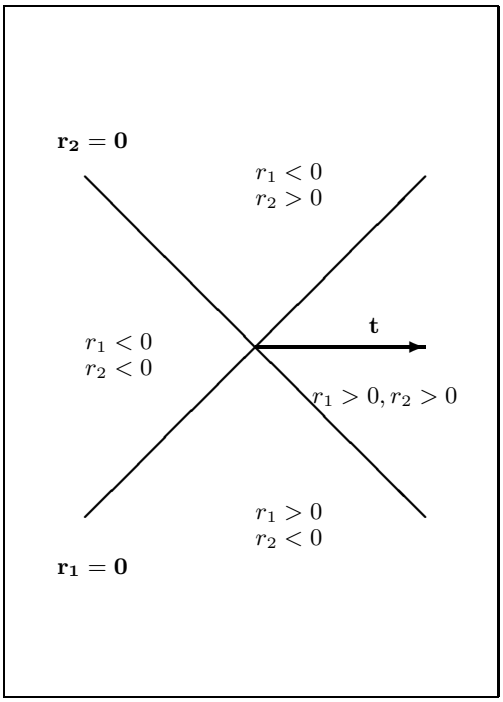

Table 2: four faces of the epigraph intersect at extreme points $\mathbf{x}_\sigma \in R^2$

Example 1.2 *Rank regression [6], [8]. This is an estimation procedure with particularly attractive properties, but unfortunately a tractable numerical algorithm had not been one of them until recently. Let nondecreasing scores w_i summing to zero be given. The estimation problem is*

$$\min_{\mathbf{x}} \sum_{i=1}^{n} w_i r_{\nu(i)}$$

$$w_1 \leq w_2 \cdots \leq w_n, \ \sum_{i=1}^{n} w_i = 0, \ \|\mathbf{w}\| > 0.$$

Here ν is the ranking set pointing to the ordered residuals, and typical scores are the Wilcoxon scores:

$$w_i = \sqrt{12}\left(\frac{i}{n+1} - \frac{1}{2}\right), \ i = 1, 2, \cdots, n.$$

Nonsmoothness of the epigraph is caused by the possible reassigning of scores associated with tied residuals as a result of small perturbations about the extreme point. Here the necessary conditions are distinctly more complicated! The reasons for this relate to additional structural complexity and will motivate subsequent developments.

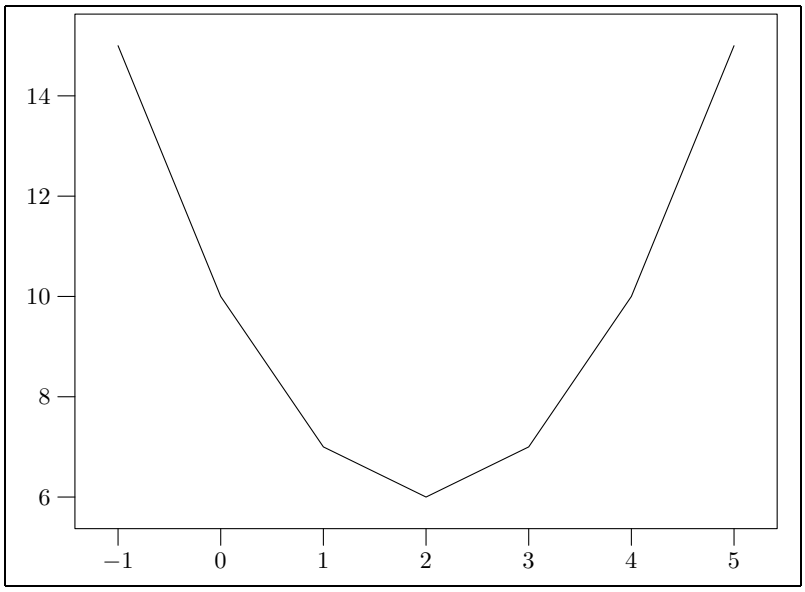

Table 3: Simple l_1 example

It turns out that now 6 faces intersect at each extreme point of the epigraph over R^2 providing a first indication of this additional complexity. Consider the lines characterizing tied residuals with equations

$$\pm(r_2 - r_1) = \pm(r_3 - r_2) = \pm(r_1 - r_3) = 0.$$

The first point to make is that there is a distinctly more serious redundancy:

$$r_1 - r_3 = -r_3 + r_2 - r_2 + r_1.$$

This forces the line given by the third equation to pass through the intersection of the other two - hence the six faces. Again it makes more sense to look at characterizing particular faces by looking at directions into faces as convex combinations of directions along appropriate edges.

$$\theta_{ik}(A_{i*} - A_{k*})\,\mathbf{t} = \lambda_{ik} > 0,$$
$$\theta_{kj}(A_{k*} - A_{j*})\,\mathbf{t} = \lambda_{kj} > 0.$$

The picture that corresponds here to figure 2 is figure 4.

Rank regression has an agreeably high statistical efficiency for a relatively robust estimator. This is mirrored in surprising apparent smoothness of the epigraph. This is illustrated by two classic examples [6].

1. *The first picture in figure 5 is the classical Hubble dataset giving velocity of recession against distance. This small dataset presents strong visual evidence for a linear relation. The second picture gives the graph of the*

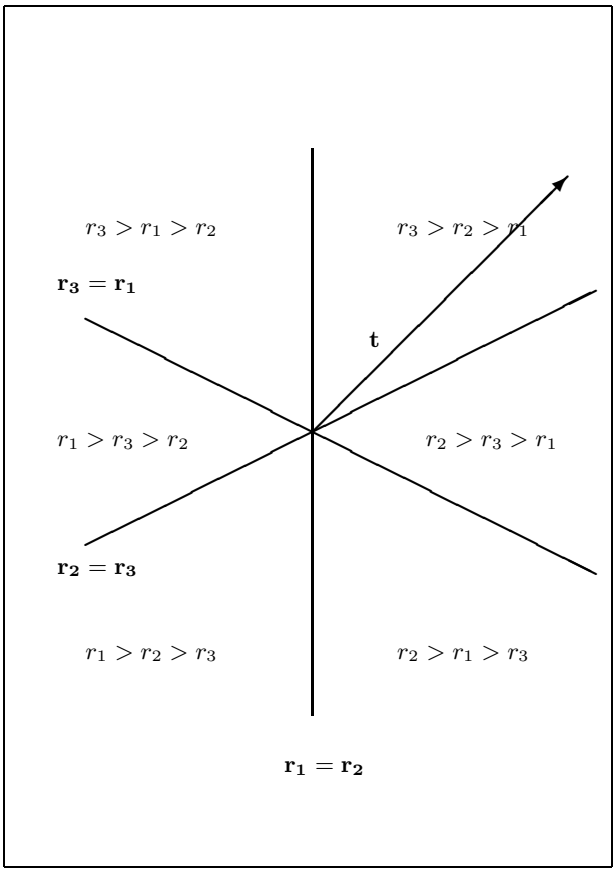

Table 4: six faces of the epigraph intersect at extreme points $\mathbf{x}_\sigma \in R^2$

derivative of the piecewise linear rank regression objective in this case. This is strictly piecewise constant, but note both the apparent smoothness, and the steep linear section in the centre of the picture. Here the reciprocal of the slope is linked to the variance reinforcing the quality of the data:

2. *The dataset plotted in figure 6 is associated with the question if two populations differ only by a constant. Here the observations are on weight gain in newborn babies corresponding to a treatment and control. The general properties of the rank regression estimate are similar, but the evidence for a conclusion is not so strong here, and this is reflected in the smaller slope of the characteristic linear middle section in the graph of the piecewise constant derivative:*

The use of the affine family description of a convex function does not lead to a practical linear programming algorithm for the l_1 problem. However, practical LP algorithms for this problem are well known, and one way of approaching

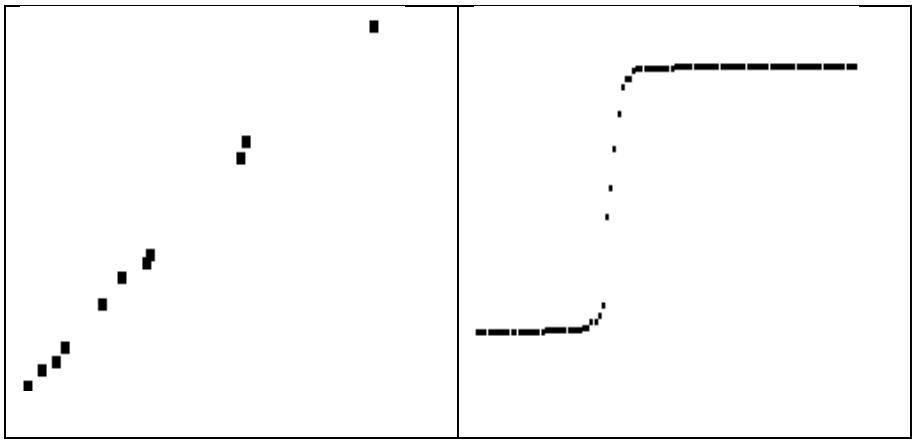

Table 5: Hubble data and Hubble rank statistic

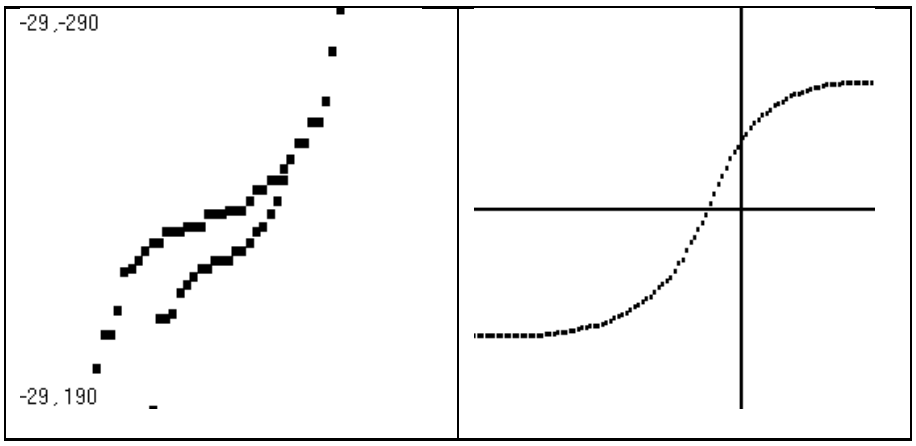

Table 6: Weight data and rank statistic

these is through Fenchel duality [11]. Here this gives:

$$\min_{\mathbf{u}} \mathbf{b}^T\mathbf{u},\ A^T\mathbf{u} = 0,\ -\mathbf{e} \leq \mathbf{u} \leq \mathbf{e}.$$

The approach works also for rank regression, but while the result looks somewhat familiar, and it is an LP, the inequality constraints have both an unusual description and an apparent fearful complexity if the scores are mostly distinct.

$$\min_{\mathbf{u}} \mathbf{b}^T\mathbf{u},\ A^T\mathbf{u} = 0,\ \mathbf{u} \in \text{conv}\,\{\mathbf{w}_i\}$$

where $\mathbf{w}_i$ are all distinct permutations of
$w_1, w_2, \cdots, w_n$. l_1 is actually a limiting case of rank regression corresponding to sign scores [10].

2 Identifying structure

The approach taken is one of seeking a compact characterization of extreme points of the epigraph in order to provide local information concerning the objective. It is a development of ideas originally presented in [10].

Definition 2.1 *The set* $\phi_i(\mathbf{r}),\ i = 1, 2, \cdots, N$ *are* **structure functionals** *for* $f(\mathbf{r}(\mathbf{x}))$ *if each extreme point* $\begin{bmatrix} \mathbf{x}^* \\ f(\mathbf{r}(\mathbf{x}^*)) \end{bmatrix}$ *of* $\operatorname{epi}(f)$ *is determined by a linear system*

$$\phi_i(\mathbf{r}(\mathbf{x}^*)) = 0,\ i \in \sigma \subseteq \{1, 2, \cdots, N\}.$$

where σ *defines the associated active set.*

The rank regression example shows that there can be a natural redundancy among the structure functionals associated with extreme points of the epigraph.

Definition 2.2 Redundancy*: A structure equation* $\phi_s = 0$ *is redundant if*

$$\exists\, \pi \neq \emptyset,\ s \notin \pi\ \ni\ (\phi_i = 0\, \forall\, i \in \pi) \Rightarrow \phi_s = 0$$

identically in r*.*

Consider the rank regression example:

$$\begin{aligned}
&\phi_{12} = r_2 - r_1, \phi_{23} = r_3 - r_2, \phi_{31} = r_1 - r_3.\\
&\phi_{12} = \phi_{23} = 0 \Rightarrow \phi_{31} = \phi_{23} - \phi_{12} = 0,\\
&\phi_{12} = 0 \Rightarrow \phi_{21} = 0.
\end{aligned}$$

Note that multiplication by -1 is significant!

In this example $\phi_{12},\ \phi_{23}$ and $\phi_{23},\ \phi_{31}$provide examples of nonredundant pairs and give structure equations corresponding to independent linear systems (here each of rank 2) over $\mathbf{r} \in R^n$. Say that such nonredundant configurations are obtained by *"allowable reductions"*.

Definition 2.3 Linear independence of structure functionals*: Let* $\mathbf{x}$ *be in the intersection of* k *structure equations pointed to by the index set* σ*. Then these structure functionals are linearly independent relative to the design matrix* A *provided*

$$\operatorname{rank}(V_\sigma) = k = |\sigma| \leq p.$$

where

$$\begin{aligned}
&V_\sigma^T = \Phi_\sigma^T A \in R^p \rightarrow R^k\\
&\Phi_\sigma = \begin{bmatrix} \nabla_r \phi_{\sigma(1)}^T & \cdots & \nabla_r \phi_{\sigma(k)}^T \end{bmatrix} \in R^k \rightarrow R^n.
\end{aligned}$$

Definition 2.4 Non-degeneracy*: This requires for a given active set of structure functionals that each allowable reduction is linearly independent relative to the problem design. For example, in* R^2 *the additional condition* $\phi_{45} = 0$ *adjoined to those above cannot be removed by an allowable reduction and leads to a degeneracy.*

Non-degeneracy is assumed here. Consider now a particular allowable reduction σ_s Let $\mathbf{x} = \mathbf{x}^* + \varepsilon\mathbf{t}$, $\varepsilon > 0$ small enough. Then, using the piecewise linearity of the objective and the linearity of the active structure functionals, this permits the objective to be written in terms of the local structure as

$$f(\mathbf{r}(\mathbf{x})) = f_{\sigma_s}(\mathbf{r}(\mathbf{x})) + \sum_{i=1}^{|\sigma_s|} \omega_i(\mathbf{t})\, \phi_{\sigma_s(i)}(\mathbf{r}(\mathbf{x})), \tag{4}$$

In this case:

1. f_{σ_s} is smooth, $\omega_i(\mathbf{t})$ provides the nonsmooth behaviour.
2. Each distinct realization of $\omega_i(\mathbf{t})$, $i = 1, 2, \cdots, |\sigma|$ characterizes one of the faces of $\mathrm{epi}(f)$ meeting at $\begin{bmatrix} \mathbf{x}^* \\ f(\mathbf{r}(\mathbf{x}^*)) \end{bmatrix}$.

This representation encapsulates the nonsmoothness in the representation of the objective function. An alternative is to provide an explicit description of the faces of the tangent cone $\mathcal{T}$ and here it is useful to introduce a further concept based on the approach exemplified in figure 2 and figure 4.

Definition 2.5 Completeness*: Let $\mathbf{x}^*$ be an extreme point. The structure functional description is complete if for each face $1 \le s \le q$ of the tangent cone $\mathcal{T}(\mathrm{epi}(f), \mathbf{x}^*)$ $\exists\, \sigma_s$, $|\sigma_s| = p$ such that directions into the face*

$$\begin{bmatrix} \mathbf{x}^* + \varepsilon\mathbf{t} \\ f(\mathbf{x}^* + \varepsilon\mathbf{t}) \end{bmatrix} = \begin{bmatrix} \mathbf{x}^* \\ f(\mathbf{x}^*) \end{bmatrix} + \varepsilon \begin{bmatrix} \mathbf{t} \\ f'(\mathbf{x}^* : \mathbf{t}) \end{bmatrix} \tag{5}$$

satisfy

$$V_{\sigma_s}^T \mathbf{t} = \lambda > 0, \tag{6}$$

where V_{σ_s} is nonsingular.

Remark 2.1 *Completeness is closely related to redundancy. For all s the systems*

$$\phi_{\sigma_s(i)}(\mathbf{r}(\mathbf{x})) = 0,\ i = 1, 2, \cdots, p$$

have the same solution $\mathbf{x}^*$. Redundancy requires in addition $q > p + 1$ where $p + 1$ corresponds to the generic vertex of linear programming.

Remark 2.2 *Extreme directions in $\mathcal{T}(\mathrm{epi}(f), \mathbf{x}^*)$ are given by*

$$\mathbf{t}_{\sigma_s}^i = V_{\sigma_s}^{-T} \mathbf{e}_i,\ i = 1, 2, \cdots, p,\ s = 1, 2, \cdots, q. \tag{7}$$

There is an overspecifcation here. Extreme directions (edges) formed by the intersection of adjacent faces are determined by an equation of this form for each face (say σ_s, σ_t) so there must be relations of linear dependence on the set of active structure functionals. What proves to be common is that a particular structure functional in the allowable reductions increases away from zero.

Example 2.1 *l_1 estimation: Here active structure functionals correspond to zero residuals.*

$$\phi_{2i-1} = r_i,\ \phi_{2i} = -r_i,\ N = 2n.$$

Let an extreme point x^* be determined by

$$\phi_{\sigma(i)} = r_i,\ \sigma = \{1, 3, \cdots, 2p-1\}$$

Let $x = x^* + \varepsilon t$ .Then

$$f(\mathbf{r}(\mathbf{x})) = \sum_{|r_i(\mathbf{x}^*)|>0} |r_i| + \sum_{i=1}^{p} \omega_i(\mathbf{t})\, \phi_{\sigma(i)}(\mathbf{r}(\mathbf{x}))$$

For each allowable reduction of the active structure functionals $\omega_i(\mathbf{t})\, \phi_{\sigma(i)}(\mathbf{r}) = |r_i|$, $\omega_i = \pm 1$. Completeness needs to characterize faces by explicitly revealing the associated index sets. For example:

$$r_1 > 0, r_2 > 0, r_3 > 0 \Rightarrow \sigma_s = \{1, 3, 5\}$$
$$r_1 > 0, r_2 < 0, r_3 > 0 \Rightarrow \sigma_s = \{1, 4, 5\}$$

There are 2^p faces intersecting at x^* in the l_1 example. However, differences between the sets of equations determining the extreme directions are pretty trivial in this case.

Example 2.2 *rank regression: Here the structure functionals express the condition for ties in the ranking set:*

$$\phi_{ij} = r_j - r_i,\ 1 \le i \ne j \le n,\ N = n(n-1). \tag{8}$$

This example supports two types of structure functional redundancy:

$$\phi_{ij} = -\phi_{ji},\ \phi_{ik} = \phi_{jk} + \phi_{ij}.$$

Examples of possible structure equations when $p = 3$ are

$$r_1 = r_2 = r_3 = r_4,$$
$$r_1 = r_2, r_3 = r_4 = r_5,$$
$$r_1 = r_2, r_3 = r_4, r_5 = r_6.$$

In the first case $\{\phi_{12}, \phi_{13}, \phi_{14}\}$ form a possible reduced set of structure functionals which specializes the role of r_1. Such a pivotal element is here called an origin. If the tie involves positions $l, l+1, \cdots, l+4$ in the sorted list then the objective function can be written:

$$f(\mathbf{r}) = \sum_{i=5}^{n} w_{\mu(i)} r_i + \left(\sum_{i=l}^{l+4} w_i \right) r_1 + \sum_{i=2}^{4} \omega_{i-1}(\mathbf{t})\, \phi_{1i}.$$

However, not all reduced active sets are equal in the sense that this may not be a good set for completeness. Let $\mathbf{t}$ be into the face $r_1 < r_2 < r_3 < r_4$. To express

this condition at nearby points in this direction using these structure functionals gives

$$r_1 < r_2 < r_3 < r_4 \Rightarrow \phi_{12} > 0, \phi_{13} > \phi_{12}, \phi_{14} > \phi_{13}.$$

The required set must satisfy (6). Thus relaxing this set of structure functionals does not give the right ordering. Here this is $\sigma_s = \{12, 23, 34\}$ *which gives*

$$\phi_{12} > 0, \phi_{23} = \phi_{13} - \phi_{12} > 0, \phi_{34} = \phi_{14} - \phi_{13} > 0.$$

The above equations show this set is related to the previous set by a linear transformation. This information can be used to change the current reduced structure functional basis representation of the non-smooth part of the objective function:

$$\begin{aligned}\sum_{i=1}^{p} \omega_i(\mathbf{t})\, \phi_{\sigma(i)}(\mathbf{r}(\mathbf{x}+\mathbf{t})) &= \mathbf{t}^T V_\sigma \boldsymbol{\omega}(\mathbf{t}), \\ &= \mathbf{t}^T V_\sigma T_s T_s^{-1} \boldsymbol{\omega}(\mathbf{t}), \\ &= \sum_{i=1}^{p} (\omega_s(\mathbf{t}))_i\, \phi_{\sigma_s(i)}(\mathbf{r}(\mathbf{x}+\mathbf{t}))\end{aligned}$$

where $\boldsymbol{\phi}_\sigma^T T = \boldsymbol{\phi}_{\sigma_s}^T$

$$\begin{aligned}&\begin{bmatrix} \phi_{12} & \phi_{13} & \phi_{14} \end{bmatrix} \begin{bmatrix} 1 & -1 & \\ & 1 & -1 \\ & & 1 \end{bmatrix} \\ &= \begin{bmatrix} \phi_{12} & \phi_{23} & \phi_{34} \end{bmatrix}\end{aligned}$$

Solutions of the systems $V_{\sigma_s}^T \mathbf{t}_i^s = \mathbf{e}_i$, $i = 1, 2, 3$,*break ties as follows:*

$$\begin{aligned}\mathbf{t}_1^s &: r_1 < r_2 = r_3 = r_4, \\ \mathbf{t}_2^s &: r_1 = r_2 < r_3 = r_4, \\ \mathbf{t}_3^s &: r_1 = r_2 = r_3 < r_4.\end{aligned}$$

3 Differential properties

Let $f(\mathbf{x}),\ \mathbf{x} \in X$ be convex. The subdifferential $\partial f(\mathbf{x})$ is the set

$$\{\mathbf{v}; f(\mathbf{t}) \geq f(\mathbf{x}) + \mathbf{v}^T(\mathbf{t} - \mathbf{x}), \forall \mathbf{t} \in X\}.$$

The elements $\mathbf{v} \in \partial f(\mathbf{x})$ are called subgradients. They generalise the idea of a gradient vector at points of nondifferentiability of $f(\mathbf{x})$. For example, the vectors $\begin{bmatrix} \mathbf{v} \\ -1 \end{bmatrix}$ give the normals to the supporting hyperplanes to $f(\mathbf{x})$ at points of nondifferentiability, and the subdifferential is the convex hull of gradient vectors at nearby differentiable points. The subdifferential is important for characterizing optima and calculating descent directions in nonsmooth convex

optimization. In particular, $\mathbf{x}$ minimizes $f(\mathbf{x})$ if $0 \in \partial f(\mathbf{x})$. The corresponding definition of the directional derivative is:

$$f'(\mathbf{x}:\mathbf{t}) = \inf_{\lambda>0} \frac{f(\mathbf{x}+\lambda\mathbf{t}) - f(\mathbf{x})}{\lambda}, \tag{9}$$

$$= \max_{\mathbf{v}\in\partial f(\mathbf{x})} \mathbf{v}^T\mathbf{t}. \tag{10}$$

To compute the subdifferential specialize an allowed reduction σ of the active set and make use of the representation (4). The convex hull form of the subdifferential now gives

$$\mathbf{v}^T \in \partial f\left(\mathbf{r}\left(\mathbf{x}\right)\right) \rightarrow \mathbf{v} = \mathbf{f}_g + V_\sigma \mathbf{z}, \tag{11}$$

where

$$\mathbf{f}_g = \nabla_x f_\sigma\left(\mathbf{r}\right)^T \tag{12}$$

is the gradient of smooth part of the objective, and

$$\left(V_\sigma\right)_{*i} = \nabla_x \phi^T_{\sigma(i)} = \left\{\nabla_r \phi_{\sigma(i)} A\right\}^T,\ i = 1, 2, \cdots, |\sigma|, \tag{13}$$

$$\mathbf{z} \in Z_\sigma = \mathrm{conv}(\boldsymbol{\omega}_s,\ s = 1, 2, \cdots, q). \tag{14}$$

The standard inequality (10) for the directional derivative gives

$$Z_\sigma = \left\{\mathbf{z}; \left(\mathbf{f}_g + V_\sigma \mathbf{z}\right)^T \mathbf{t} \leq f'\left(\mathbf{x}^* : \mathbf{t}\right)\right\}. \tag{15}$$

It follows that the constraint set is known if the directional derivative can be computed.

The constraint set Z_σ is polyhedral and has the important property that the extreme points are determined by the extreme directions of $\mathcal{T}\left(\mathrm{epi}\left(f\right), \mathbf{x}^*\right)$. Let $\mathbf{t}_s$ be into a face of $\mathcal{T}$ as illustrated in the examples sketched in figures 2 and 4. Then it can be written as a convex combination of the edge directions. The key calculation is

$$f'\left(\mathbf{x}^* : \mathbf{t}_s\right) = \mathbf{f}_g^T \mathbf{t}_s + \max_{\mathbf{z}\in Z_\sigma} \mathbf{z}^T V_\sigma^T \mathbf{t}_s, \tag{16}$$

$$= \mathbf{f}_g^T \mathbf{t}_s + \max_{\mathbf{z}\in Z_\sigma} \left\{\sum_{i=1}^p \lambda_i \mathbf{z}^T V_\sigma^T \mathbf{t}_i^s\right\},$$

$$\leq \mathbf{f}_g^T \mathbf{t}_s + \sum_{i=1}^p \lambda_i \max_{\mathbf{z}\in Z_\sigma} \mathbf{z}^T V_\sigma^T \mathbf{t}_i^s$$

$$= \sum_{i=1}^p \lambda_i f'\left(\mathbf{x}^* : \mathbf{t}_i^s\right). \tag{17}$$

However, equality between the directional derivative and the component directional derivatives on the edges follows from the linearity of f on the face containing $\mathbf{t}_s$. This shows that $\widehat{\mathbf{z}_s}$ which maximizes (16) also maximizes each of the terms in (17) and is thus a characteristic of the face.

To compute Z_σ it is important to take account of the form of the structure functionals that remain active on an edge leading from the extreme point. Consider, for example, the tie $r_1 = r_2 = r_3 = r_4$ determining an extreme point in a rank regression problem corresponding to $p = 3$. If r_1 leaves the group with the remaining residuals still tied then this determines an edge on which $\phi_{12}, \phi_{13}, \phi_{14}$ all relax away from zero. On this edge $\phi_{23} = \phi_{24} = 0$. Thus it is necessary to relate $\phi_{12}, \phi_{13}, \phi_{14}$ and $\phi_{12}, \phi_{23}, \phi_{24}$. In general this will lead to a relation of the form

$$\left[\begin{array}{cc} \Phi_j & \nabla_r \phi_j^T \end{array} \right] \left[\begin{array}{cc} S_j & \\ \mathbf{s}_j^T & 1 \end{array} \right] = \Phi_\sigma P_j, \tag{18}$$

where:

1. ϕ_j is the structure functional that increases from zero on the edge (here ϕ_{12});
2. Φ_j is the gradient matrix associated with the structure functionals that remain active on the edge (here $\phi_{23},\ \phi_{24}$);
3. the form of the transformation matrix is fixed by the particular mode by which the number of active structure functionals is reduced in the current allowable reduction;
4. P_j is a permutation matrix performing the necessary rearrangements which include switching $\nabla_r \phi_j^T$ to the last column; and
5. the active set condition on the edge is

$$\Phi_j^T A \mathbf{t} = 0. \tag{19}$$

The directional derivative on the edge is given by

$$\begin{aligned} f'(\mathbf{x}^* : \mathbf{t}) &= \mathbf{f}_g^T \mathbf{t} + \max_{\mathbf{z} \in Z_\sigma} \mathbf{z}^T V_\sigma^T \mathbf{t}, \\ &= \mathbf{f}_g^T \mathbf{t} + \max_{\mathbf{z} \in Z_\sigma} \mathbf{z}^T P_j^{-T} \left[\begin{array}{cc} S_j^T & \mathbf{s}_j \\ & 1 \end{array} \right] \left[\begin{array}{c} \Phi_j^T \\ \nabla_r \phi_j \end{array} \right] A \mathbf{t}, \\ &= \mathbf{f}_g^T \mathbf{t} + \max_{\mathbf{z} \in Z_\sigma} \mathbf{z}^T P_j^{-T} \left[\begin{array}{c} \mathbf{s}_j \\ 1 \end{array} \right] \mathbf{v}_j^T \mathbf{t}, \\ &= \mathbf{f}_g^T \mathbf{t} + \left\{ \begin{array}{l} \zeta_j^+ \mathbf{v}_j^T \mathbf{t},\ \mathbf{v}_j^T \mathbf{t} > 0, \\ \zeta_j^- \mathbf{v}_j^T \mathbf{t},\ \mathbf{v}_j^T \mathbf{t} < 0, \end{array} \right. \end{aligned}$$

where the edge condition (19) has been used. This gives the inequalities determining Z_σ in the form

$$\zeta_j^- \leq \left[\begin{array}{cc} \mathbf{s}_j^T & 1 \end{array} \right] P_j^{-1} \mathbf{z} \leq \zeta_j^+, \tag{20}$$

where the bounds are given by

$$\begin{aligned} \zeta_j^+ &= \max_{\mathbf{z} \in Z_\sigma} \mathbf{z}^T P_j^{-T} \left[\begin{array}{c} \mathbf{s}_j \\ 1 \end{array} \right], \\ \zeta_j^- &= \min_{\mathbf{z} \in Z_\sigma} \mathbf{z}^T P_j^{-T} \left[\begin{array}{c} \mathbf{s}_j \\ 1 \end{array} \right]. \end{aligned}$$

This has sneaked in the assumption that both ϕ_j, and $-\phi_j$ are active at $\mathbf{x}^*$. This is not always true (double sided bounds imply redundancy), and the case where this is not true is interesting. Basically it corresponds to the simple "max" case where the defining affine family supports "non degenerate" extreme points having the form

$$\mathbf{c}_{\nu(i)}^T \mathbf{x} - d_{\nu(i)} = C(\mathbf{x}),\ i = 1, 2, \cdots p+1.$$

Suitable structure equations are

$$\phi_i(\mathbf{x}) = \left(\mathbf{c}_{\nu(i)}^T - \mathbf{c}_{\nu(p+1)}^T\right)\mathbf{x} - \left(d_{\nu(i)} - d_{\nu(p+1)}\right) = 0\ i = 1, 2, \cdots, p.$$

In terms of this set the objective function can be written:

$$C(\mathbf{x}) = \mathbf{c}_{\nu(p+1)}^T \mathbf{x} - d_{\nu(p+1)} + \sum_{i=1}^{p} \omega_i(\mathbf{t})\, \phi_{nu(i)}(\mathbf{x}),$$

the weights characterizing the nonsmoothness satisfy

$$\omega_i(\mathbf{t}) = \begin{cases} 1, \mathbf{t} \text{ is into face } i, \\ 0 \text{ otherwise,} \end{cases}$$

and the resulting constraint set is

$$Z = \left\{ \mathbf{z};\ z_i \geq 0,\ i = 1, 2, \cdots, p,\ \sum_{i=1}^{p} z_i \leq 1. \right\}$$

This case includes discrete maximum norm estimation.

Example 3.1 *Rank regression again: In general, at an extreme point there will exist multiple groups of ties and each edge leading from this point is obtained by relaxing a structure functional in one of these groups. This will result in the group splitting into two subgroups one or both of which could be the trivial group containing a single element. Here the case of a splitting into two nontrivial subgroups, one of which is tied to the group origin, is considered, and this has the consequence that a new origin must be found for the second subgroup thereby releasing one degree of freedom. Let the original subgroup be* $V_0 = \begin{bmatrix} \nabla_x \phi_1^T & \cdots & \nabla_x \phi_m^T \end{bmatrix}$, *the subgroup with the same origin be* $V_1 = \begin{bmatrix} \nabla_x \phi_1^T & \cdots & \nabla_x \phi_{k-1}^T \end{bmatrix}$, *with origin residual* r_{m+1}, *and the new subgroup be*

$$V_2 = \begin{bmatrix} \nabla_x (\phi_{k+1} - \phi_k)^T & \cdots & \nabla_x (\phi_m - \phi_k)^T \end{bmatrix}.$$

The relation (18) here has the particular form

$$\begin{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix} & \nabla_x \phi_k^T \end{bmatrix} \begin{bmatrix} S & \\ \mathbf{s}_k^T & 1 \end{bmatrix} = V_0 P, \tag{21}$$

where the above assumptions imply $P = I$ and

$$\begin{bmatrix} S & \\ \mathbf{s}_k^T & 1 \end{bmatrix} = \begin{bmatrix} I & & & 0 \\ & & I & 0 \\ 0 & \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} & 1 \end{bmatrix}. \tag{22}$$

This gives the inequalities

$$\zeta_k^- \leq \sum_{j=k}^{m} z_j \leq \zeta_k^+ \tag{23}$$

for each mode of separation into subgroups (after reordering if necessary).

The procedure used to calculate ζ for a particular splitting of a group compares two computations of $f'(\mathbf{x}^* : \mathbf{t})$. The first uses the form of the subdifferential at the extreme point based on the original reduced active set structure to obtain a lower bound for the directional derivative using (10). In this case the origin contribution is $\left(\sum_{i=1}^{m+1} w_i\right) A_{(m+1)*}\mathbf{t}$ and the contribution from the group before splitting is

$$\sum_{i=1}^{m} z_i \left(A_{i*} - A_{(m+1)*}\right) \mathbf{t} = \left(\sum_{i=1}^{k} z_i\right) \left(A_{k*} - A_{(m+1)*}\right) \mathbf{t}$$

where the calculation requires that allowance be made for terms which vanish on the edge. The second calculation involves the new subgroups, and here only the contributions of the origin terms matter as the terms involving the active structure functionals vanish on the edge as a consequence of (19). Specializing to the case $A_{k*}\mathbf{t} < A_{(m+1)*}\mathbf{t}$ corresponding to the new subgroup changing more slowly on the edge gives the contribution

$$\left(\left(\sum_{i=1}^{k} w_i\right) A_{k*} + \left(\sum_{i=k+1}^{m+1} w_i\right) A_{(m+1)*}\right) \mathbf{t}.$$

This result has the interesting feature that it is independent of $\mathbf{z}$ so that the contribution of the split groups to the directional derivative estimate is already maximised. The general result [4] is

$$\left(\sum_{i=1}^{m-k+1} w_i\right) \leq \sum_{i=k}^{m} z_{\pi(i)} \leq \left(\sum_{i=k+1}^{m+1} w_i\right),$$

where π is any permutation of $1, 2, \cdots, m$. This says that the sum of the k smallest scores must be less than the sums of any k multipliers z_i and these sums must, in turn be less than the sum of the k largest scores for $k = 1, 2, \cdots, m$.

4 Elements of a simplicial algorithm [10], [11]

The basic steps in a simplicial algorithm are:

1. test at the current extreme point to see if $0 \in \partial f$. If this test is satisfied then the current point is optimal;

2. otherwise use the information from this test to determine an edge of the epigraph generating a descent direction;

3. then proceed using a line search to determine the minimum of the objective in this direction. This search will terminate at another extreme point.

The current point is optimal provided

$$\exists \widetilde{\mathbf{z}} \in Z,\ 0 = \mathbf{f}_g + V\widetilde{\mathbf{z}}. \tag{24}$$

If this test is unsuccessful then $\widetilde{\mathbf{z}} \notin Z$, and there exists a violated member of the set of inequalities. Let this be:

$$\zeta_k^- \leq \left[\begin{array}{cc} \mathbf{s}_k^T & 1 \end{array}\right] P_k^{-1}\mathbf{z} \leq \zeta_k^+. \tag{25}$$

This information can now be used to compute a descent direction. Let the transformation generating the edge be:

$$V \to \left[\begin{array}{cc} V_k & \mathbf{v}_k \end{array}\right] \left[\begin{array}{cc} S_k & \\ \mathbf{s}_k^T & 1 \end{array}\right] P_k^{-1}. \tag{26}$$

Then the direction determined by the edge is found by solving the linear system

$$\mathbf{t}^T \left[\begin{array}{cc} V_k & \mathbf{v}_k \end{array}\right] = \theta \mathbf{e}_p^T,\ \theta = \pm 1. \tag{27}$$

Here the choice of θ depends on whether the left or right inequality in (25) is violated. To verify the descent property compute the directional derivative and use the definition of $\widetilde{\mathbf{z}}$, and the active set condition (19). This gives

$$\begin{aligned}
&\sup_{\mathbf{z}\in Z} \mathbf{t}^T(\mathbf{f}_g + V\mathbf{z}) \\
&= \sup_{\mathbf{z}\in Z} \left(-\mathbf{t}^T V\widetilde{\mathbf{z}} + \theta \left[\begin{array}{cc} \mathbf{s}_k^T & 1 \end{array}\right] P_k^{-1}\mathbf{z}\right), \\
&= \sup_{\mathbf{z}\in Z} \left(\theta \left[\begin{array}{cc} \mathbf{s}_k^T & 1 \end{array}\right] P_k^{-1}\left(\mathbf{z} - \widetilde{\mathbf{z}}\right)\right), \\
&= \begin{cases} \left(\zeta_k^+ - \left[\begin{array}{cc} \mathbf{s}_k^T & 1 \end{array}\right] P_k^{-1}\widetilde{\mathbf{z}}\right), & \left[\begin{array}{cc} \mathbf{s}_k^T & 1 \end{array}\right] P_k^{-1}\widetilde{\mathbf{z}} > \zeta_k^+, \\ -\left(\zeta_k^- - \left[\begin{array}{cc} \mathbf{s}_k^T & 1 \end{array}\right] P_k^{-1}\widetilde{\mathbf{z}}\right), & \left[\begin{array}{cc} \mathbf{s}_k^T & 1 \end{array}\right] P_k^{-1}\widetilde{\mathbf{z}} < \zeta_k^-, \end{cases} \\
&< 0.
\end{aligned}$$

A linesearch must now be performed in the descent direction. Preferred methods work with the piecewise constant directional derivative of the objective function, and it is assumed that this can be evaluated economically. It is necessary to have a global solution strategy in order to avoid the potential computational cost of a close inspection of slope changes - for example, there are $O(n^2)$ slope changes on the line generated by each descent edge in the rank regression problem if the scores are distinct. In general the minimum is not

characterized by a zero of the directional derivative. Rather, the desired point occurs at a "crossing point" where the graph of the directional derivative jumps across the axis from negative to positive in the search direction. This behaviour does not sit well with standard root finding algorithms and suitable modifications must be sought.

- Hoare's partitioning algorithm: A linesearch method using this algorithm (the partitioning step in quicksort) has proved popular in the l_1 estimation problem [2]. Here it is only necessary to know the distances from the current point to nonsmooth points in the search direction. The required point is then identified as a weighted median. Hoare's algorithm has been suggested with the partition bound defined by the standard median of three approach. Interestingly, this proves very successful for problems with randomly generated model data, but appears much less satisfactory when the model corresponds to a standard continuous approximation problem.

- Bisection applied to the directional derivative: Here bisection is applied to refine a bracket of the minimum. Also it is necessary to be able to recognise when the bracket contains just one active member. The shifting strategy required to modify the secant algorithm will do. Bisection has optimal properties which ensure that its worst case performance will never be too bad.

- A secant algorithm: The asymptotic linearity evident in figures 5 and 6 suggests use of the secant algorithm to find the crossing point in the rank regression problem. This was first implemented in [9]. As noted above the continuous algorithm needs modification [5]. Here a secant step identifies a new piecewise constant piece of the directional derivative and this is followed by a shifting strategy which identifies the end of this piece closest to the crossing point as in figure 7. This modification ensures that the algorithm is finite. It proves effective in other applications (for example, l_1), but an application of the secant algorithm which includes the shifting step has been given in which the method encounters every constant piece [11]. This example is extremely badly scaled.

It is important that evaluation of $f'(\mathbf{x} : \mathbf{t})$ be no worse than $n\gamma(n)$, where $\gamma(n)$ is a function of slow growth ($\gamma(n)/n = o(1)$.

5 Polyhedral constrained problems [11]

The basic problem to be considered is that of minimizing an objective function subject to a single polyhedral constraint. The polyhedral constraint can provide a compact representation of relatively complicated systems of linear inequalities, especially when these serve to represent a global statement of the constraint structure. The local representations that serve well in the optimization context are again useful. The basic problem statement is

$$\min_{\mathbf{x}\in X} f(\mathbf{x});\ X = \{\mathbf{x};\ \kappa \geq g(\mathbf{x})\}. \tag{28}$$

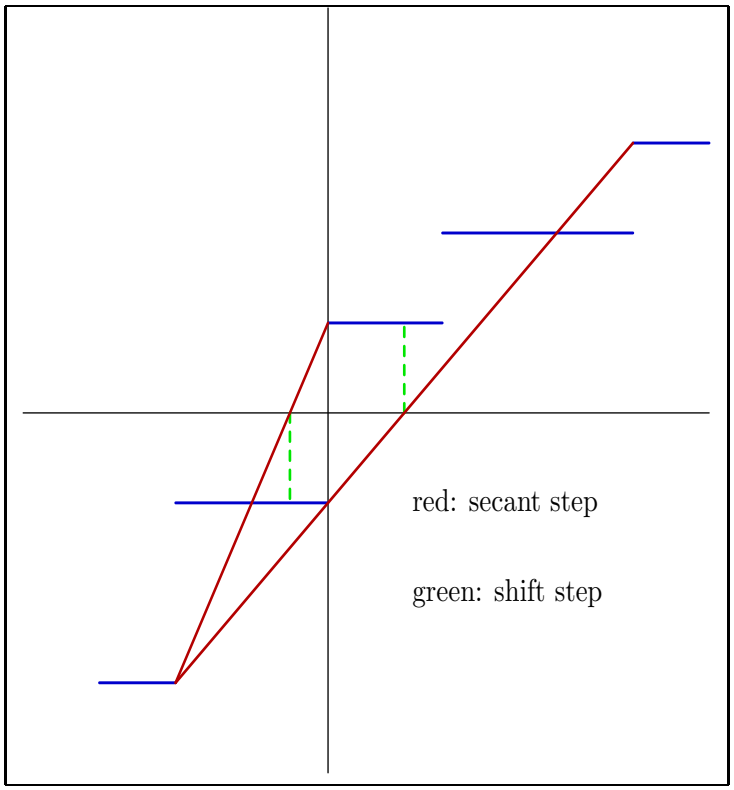

Table 7: Progress in the secant algorithm

The assumptions made here are that $f(\mathbf{x})$ is strictly convex and smooth (typically a positive definite quadratic form), and that $g(\mathbf{x})$ is polyhedral convex. The Kuhn-Tucker conditions for (28) are

$$\nabla f(\mathbf{x}) = -\mu \mathbf{v}^T,\ \mathbf{v}^T \in \partial g(\mathbf{x}),\ \mu \geq 0, \tag{29}$$

where $\mu = \mu(\kappa)$ is the constraint multiplier. Here κ plays the role of a control parameter. As it increases the strength of the constraint is weakened so that

$$\kappa \to \infty,\ \mathbf{x}_\kappa \to \arg\min_{\mathbf{x}} f(\mathbf{x}),\ \mu(\kappa) \to 0.$$

If $\mathbf{x}^*$ is the unconstrained minimum of $f(\mathbf{x})$ then it also solves (28) when $\kappa \geq g(\mathbf{x}^*)$. This gives $\mu(\kappa) = 0$ as the corresponding multiplier value. The limit as κ tends to its lower bound is simplest when $\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} g(\mathbf{x})$ is an isolated (global) minimum. The condition for a nonempty feasible region requires $\kappa \geq g(\widehat{\mathbf{x}})$, and in this case

$$\kappa \to g(\widehat{\mathbf{x}}),\ \mathbf{x}_\kappa \to \widehat{\mathbf{x}},\ \mu(\kappa) \to \mu(g(\widehat{\mathbf{x}}))$$

A closely related problem considers the Lagrangian associated with (28):

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x}), \ \lambda \geq 0. \tag{30}$$

Note that L is strictly convex as a function of $\mathbf{x}$ and that λ is here the control parameter and it could well be set a priori. The necessary conditions are identical with those of the constrained problem (29) when $\lambda = \mu(\kappa)$. Also $\lambda = 0$ when $\mu = 0$. However, the Lagrangian is defined if $\lambda \geq \mu(g(\widehat{\mathbf{x}}))$. If $0 \in \partial g(\widehat{\mathbf{x}})^o$, the interior of $\partial g(\mathbf{x})$, then $\widehat{\mathbf{x}}$ minimizes $L(\mathbf{x}, \lambda)$ for $\lambda \geq \mu(\widehat{\mathbf{x}})$. The argument uses that

$$\mathbf{v}^T \in \partial g(\widehat{\mathbf{x}}) \Rightarrow \frac{\mu}{\lambda}\mathbf{v}^T \in \partial g(\widehat{\mathbf{x}}), \ \lambda > \mu.$$

This follows from the convexity of ∂g because $\frac{\mu}{\lambda}\mathbf{v}$ is on the join of $\mathbf{v}$ and 0.

Applications which lead to polyhedral constrained problems either directly or in Lagrangian form include:

1. The "Lasso" provides a new approach to variable selection [12]. It uses the structure of the l_1 unit ball to force components of the state vector to zero. This is illustrated in figure 8. The constrained problem considered is

$$\min_{\mathbf{x}} \frac{1}{2}\mathbf{r}^T\mathbf{r}; \ \|\mathbf{x}\|_1 \leq \kappa. \tag{31}$$

   Here κ is the control parameter. Small values of κ will introduce bias into the parameter estimates in data analytic applications in addition to controlling the number of variables selected.

2. The Lagrangian form of the same problem has been considered as "Basis pursuit denoising" [3] . The problem statement is

$$\min_{\mathbf{x}} \left\{ \frac{1}{2}\mathbf{r}^T\mathbf{r} + \lambda \|\mathbf{x}\|_1 \right\}. \tag{32}$$

3. A differently structured problem occurs in the literature on machine learning [14] and data mining. The "Support vector regression" formulation is

$$\min_{\mathbf{x}} \left\{ \frac{1}{2}\|\mathbf{x}\|_2^2 + \lambda \sum_{i=1}^{n} |r_i|_\varepsilon \right\}, \tag{33}$$

$$|r|_\varepsilon = \begin{cases} |r| - \varepsilon, |r| \geq \varepsilon, \\ 0, \ |r| < \varepsilon. \end{cases} \tag{34}$$

Here the value of λ controls the trade-off between regularization and bias in the estimation procedure - small values introducing the most bias. The value of ε defines the "ε-insensitive region". Data corresponding to residuals that fall into the interior of this region is effectively ignored.

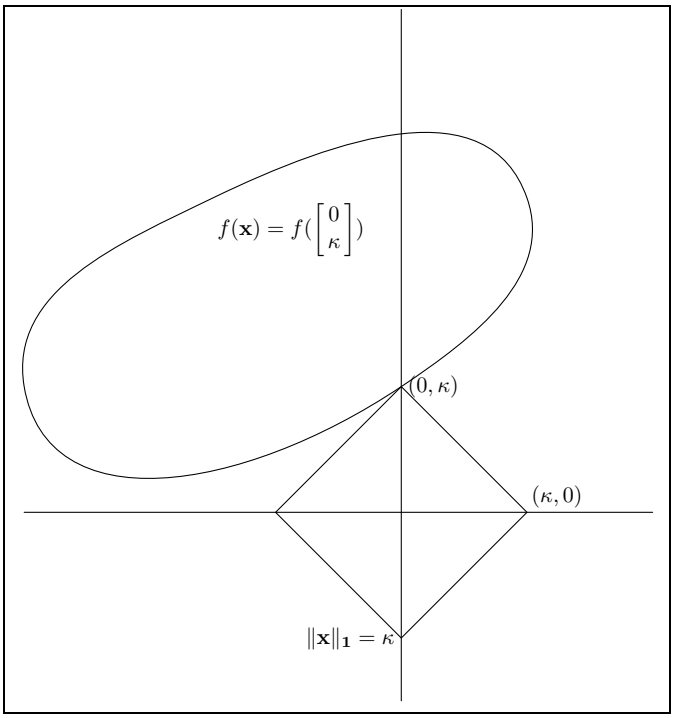

Table 8: A mechanism for variable selection

6 An active set algorithm

The basic components of an active set algorithm are:

1. A local approximation of the problem (typically a quadratic representation of the objective and a linearization of the constraints) is constructed at the current point $\mathbf{x}_0$. This generates a linear subproblem which is solved to give a direction of descent $\mathbf{h}$;

2. This computed direction is used in a linesearch step to generate the next iterate.

There are differences in detail between the algorithms for the constrained and Lagrangian form of the problem. These have to do with the choice of local objective, the constraint set being determined by the active structure functionals in both cases. Here attention is restricted to the Lagrangian form which is somewhat more interesting as there is a constraint contribution to the local objective. The necessary local structure is encapsulated in the compact representation of the subdifferential:

$$\mathbf{v}^T \in \partial g(\mathbf{x}_0) \Rightarrow \mathbf{v} = \mathbf{g}_0 + V_\sigma \mathbf{z},\ \mathbf{z} \in Z_\sigma. \tag{35}$$

Here it should be noted that as curvature in f can be important it is no longer sufficient that the solution be sought among the extreme points of $g(\mathbf{x})$ so that at the optimum $\text{rank}(V_\sigma) = k \leq p$. However, the structure function formalism applies here also; and the representation of the subdifferential as the convex hull of nearby gradients provides an accessible route to this. The subdifferential can be split into components obtained by projecting into the constraint space spanned by the structure functional gradients $\mathbf{v}_i$ and into its orthogonal complement. The derivation of the constraint set is now carried out in the constraint space where the preceding discussion applies. The splitting is particularly simple in the cases corresponding to the lasso and to support vector regression.

The descent direction is generated by solving the quadratic program

$$\min_{V_\sigma^T \mathbf{h}=0} G(\mathbf{x}_0, \mathbf{h}), \tag{36}$$

$$G(\mathbf{x}_0, \mathbf{h}) = \left(\nabla f(\mathbf{x}_0) + \lambda \mathbf{g}_0^T\right) \mathbf{h} + \frac{1}{2} \mathbf{h}^T \nabla^2 f(\mathbf{x}_0) \mathbf{h}. \tag{37}$$

It is assumed that the exact Hessian can be computed readily. This is certainly true for the examples referenced which have quadratic objective functions so that feasible points for (36) satisfy

$$L(\mathbf{x}_0 + \mathbf{h}, \lambda) = L(\mathbf{x}_0, \lambda) + G(\mathbf{x}_0, \mathbf{h}). \tag{38}$$

The immediate region of validity for this derivation of $\mathbf{h}$ corresponds to the region of validity of the compact structure functional representation of $g(\mathbf{x})$ about $\mathbf{x}_0$. This will be called the region of lc-feasibility. It is characterized by:

- the referenced index set σ points to the active structure functionals at $\mathbf{x}_0$;
- the constraints in the quadratic program express the condition that the current active structure is preserved in the computed direction; and
- $\mathbf{g}_0$ is the gradient of the differentiable part of g in the compact representation (4) about $\mathbf{x}_0$.

Lemma 6.1 *Let* $\mathbf{h}$ *minimize* G. *Iff* $\mathbf{h} \neq 0$ *then* $\mathbf{h}$ *is a descent direction for minimizing* $L(\mathbf{x}, \lambda)$.

Proof. By assumption $\mathbf{h}^T \nabla^2 f(\mathbf{x}_0) \mathbf{h} \geq 0$. As $G(\mathbf{x}_0, 0) = 0$ it follows that

$$\mathbf{h} \neq 0 \Rightarrow \min G < 0 \Rightarrow \left(\nabla f(\mathbf{x}_0) + \lambda \mathbf{g}_0^T\right) \mathbf{h} < \mathbf{0}. \tag{39}$$

A direct computation now shows that the directional derivative is negative.

$$\begin{aligned} L'(\mathbf{x}, \lambda : \mathbf{h}) &= \max_{\mathbf{v}^T \in \partial L} \mathbf{v}^T \mathbf{h}, \\ &= \max_{\mathbf{z} \in Z_\sigma} \left\{ \nabla f(\mathbf{x}_0) + \lambda \left(\mathbf{g}_0 + V_\sigma \mathbf{z}\right)^T \right\} \mathbf{h}, \\ &= \left(\nabla f(\mathbf{x}_0) + \lambda \mathbf{g}_0^T\right) \mathbf{h} < \mathbf{0}. \end{aligned}$$

■

The next phase of the computations is simplest when (38) is satisfied corresponding to f a quadratic form. There are two possibilities:

1. either $\mathbf{x}_1 = \mathbf{x}_0 + \mathbf{h}$ is an lc-feasible minimum of the quadratic program; or

2. the the step to the minimum in the direction of $\mathbf{h}$ requires at least one new structure functional to change sign or become active. The situation here differs from the case of purely polyhedral objectives in that the search need not terminate at a new active structure functional.

If $\mathbf{x}_1$ is an lc-feasible minimum then the necessary conditions give

$$0 = \nabla f(\mathbf{x}_1)^T + \lambda \left(\mathbf{g}_0 + V_\sigma \mathbf{z}_1\right), \tag{40}$$

where $\mathbf{z}_1$ is the multiplier vector for the quadratic program. If

$$\mathbf{z}_1 \in Z_\sigma \Rightarrow 0 \in \partial L(\mathbf{x}_1, \lambda)$$

then $\mathbf{x}_1$ is optimal. If not then V_σ does not correspond to the correct active set at the optimum and must be modified. This follows the same pattern as before. It is necessary to :

1. relax an active structure functional associated with a violated constraint (20) on Z_σ;

2. redefine the local linearization.

If ϕ_j is the structure functional deleted from the active set then the index set becomes $\sigma \leftarrow \sigma \diagdown \{j\}$. The corresponding modification of the active set is

$$\begin{aligned}
&\begin{bmatrix} V_j & \mathbf{v}_j \end{bmatrix} \begin{bmatrix} S & \\ \mathbf{s}_j^T & 1 \end{bmatrix} = V_\sigma P_j, \\
&\mathbf{g}_1^j = \mathbf{g}_0 + \zeta_j \mathbf{v}_j, \\
&\zeta_j = \begin{cases} \zeta_j^-, \begin{bmatrix} \mathbf{s}_j^T & 1 \end{bmatrix} P_j^{-1} \mathbf{z}_1 < \zeta_j^-, \\ \zeta_j^+, \begin{bmatrix} \mathbf{s}_j^T & 1 \end{bmatrix} P_j^{-1} \mathbf{z}_1 > \zeta_j^+. \end{cases}
\end{aligned}$$

Lemma 6.2 *The solution of the revised QP is a descent direction which is lc-feasible. Let*

$$\mathbf{h}_j = \arg \min_{V_j^T \mathbf{h} = 0} G\left(\mathbf{x}_1, \mathbf{h}\right).$$

Then $\mathbf{h}_j$ is a descent direction, and is lc-feasible in the sense that the behaviour of ϕ_j is determined by

$$\begin{aligned}
\begin{bmatrix} \mathbf{s}_j^T & 1 \end{bmatrix} P_j^{-1} \mathbf{z}_1 > \zeta_j^+ &\Rightarrow \mathbf{v}_j^T \mathbf{h}_j > 0, \\
\begin{bmatrix} \mathbf{s}_j^T & 1 \end{bmatrix} P_j^{-1} \mathbf{z}_1 < \zeta_j^- &\Rightarrow \mathbf{v}_j^T \mathbf{h}_j < 0.
\end{aligned}$$

Proof. It is necessary to verify first lc-feasibility. The necessary conditions give

$$\nabla^2 f \mathbf{h}_j + \nabla f^T + \lambda \left(\mathbf{g}_1^j + V_j \mathbf{z}_1\right) = 0,\ V_j^T \mathbf{h}_j = 0,$$

so that

$$\mathbf{h}_j^T \nabla^2 f \mathbf{h}_j + \mathbf{h}_j^T \left(\nabla f^T + \lambda \mathbf{g}_0\right) + \lambda \zeta_j \mathbf{h}_j^T \mathbf{v}_j = 0. \tag{41}$$

Also, if follows from (40) that

$$\begin{aligned} 0 &= \mathbf{h}_j^T \left(\nabla f^T + \lambda \mathbf{g}_0 + \lambda V_\sigma \mathbf{z}_1 \right), \\ &= \mathbf{h}_j^T \left(\nabla f^T + \lambda \mathbf{g}_0 \right) + \lambda \left[\begin{array}{cc} \mathbf{s}_j^T & 1 \end{array} \right] P_j^{-1} \mathbf{z}_1 \mathbf{h}_j^T \mathbf{v}_j. \end{aligned} \tag{42}$$

Combining (41) and (42) gives

$$\mathbf{h}_j^T \nabla^2 f \mathbf{h}_j + \lambda \left(\zeta_j - \left[\begin{array}{cc} \mathbf{s}_j^T & 1 \end{array} \right] P_j^{-1} \mathbf{z}_1 \right) \mathbf{h}_j^T \mathbf{v}_j = 0.$$

The result follows from this using the mode of violation of the constraint inequalities (20). The descent property now follows from Lemma 6.1. ■

If $f(\mathbf{x})$ is not a quadratic form then the optimum is approached iteratively using the solution of the locally defined quadratic program as a descent direction at each stage. Note that a line search in this computed descent direction is then used to determine the next iterate. Also there is a line search used in the solution of the current quadratic program. Thus there could be some scope for balancing the computational load between these two component steps in each iteration.

The quadratic program line search is complicated by the occurrence of jump discontinuities in the directional derivatives so that the minimum may occur either at a new active structure functional corresponding to a crossing point or at a zero of the directional derivative forced by the curvature of the objective function. It has proved convenient to work with the directional derivative and to first isolate the minimum to an interval which contains at most one discontinuity. If there are none then the secant algorithm can be applied once on this interval. If there is exactly one then it is straightforward to distinguish between a crossing point and a zero. It is necessary to test first for a crossing point, but if this fails then the zero can be determined by a single secant step on the appropriate subinterval.

7 A homotopy algorithm

In [11] an effective algorithm for implementing the lasso is described. This makes use of the result that the optimal solution trajectory $\mathbf{x}(\kappa)$ is a piecewise linear function of the constraint bound κ in (31). Here the resulting algorithm has much of the character of a stepwise variable selection procedure with the added advantage that the global optimum is obtained for each value of the selection parameter κ. This contrasts with the classical stepwise regression procedure where the local greedy algorithm employed need not produce a global result. The existence of a piecewise linear optimal homotopy path extends to the case where the objective $f(\mathbf{x})$ is a positive definite quadratic form and the constraint $g(\mathbf{x})$ is polyhedral convex.

The problem is considered in Lagrangian form and the multiplier λ is used in the role of homotopy control parameter. Let $\mathbf{x}$ be optimal for the current value of λ. Then the necessary conditions are

$$\nabla f(\mathbf{x})^T + \lambda \left(\mathbf{g}_\sigma + V_\sigma \mathbf{z}_\sigma \right) = 0$$

where σ is the index set pointing to the current (non redundant) set of active structure functionals which correspond to equality constraints here. Assume $\mathbf{z}_\sigma \in Z_\sigma^o$. This, plus continuity of the minimizer $\mathbf{x}(\lambda)$, ensures local differentiability with respect to λ. This gives

$$\nabla^2 f \frac{d\mathbf{x}}{d\lambda} + V_\sigma \frac{d}{d\lambda}(\lambda \mathbf{z}_\sigma) + \mathbf{g}_\sigma = 0, \tag{43}$$

$$V_\sigma^T \frac{d\mathbf{x}}{d\lambda} = 0. \tag{44}$$

Strict convexity plus the nonredundancy assumption implies that the augmented matrix of the quadratic program (36) is nonsingular. Here this gives

$$\begin{bmatrix} \frac{d\mathbf{x}}{d\lambda} \\ \frac{d}{d\lambda}(\lambda \mathbf{z}_\sigma) \end{bmatrix} = - \begin{bmatrix} \nabla^2 f & V_\sigma \\ V_\sigma^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{g}_\sigma \\ 0 \end{bmatrix}. \tag{45}$$

The piecewise linear nature of the solution trajectory follows because the right hand side is independent of λ. The consistency of the assumptions made is verified readily. Adding δ times (43) to the necessary conditions gives

$$\begin{aligned} \nabla f^T + \delta \nabla^2 f \frac{d\mathbf{x}}{d\lambda} + (\lambda + \delta)\mathbf{g}_\sigma + V_\sigma \left(\lambda \mathbf{z} + \delta \frac{d}{d\lambda}(\lambda \mathbf{z}) \right) \\ = \nabla f(\mathbf{x} + \delta \frac{d\mathbf{x}}{d\lambda}) + (\lambda + \delta)\left(\mathbf{g}_\sigma + V_\sigma \mathbf{z}(\lambda + \delta)\right) \\ = 0 \end{aligned}$$

as the Taylor series for $\lambda \mathbf{z}$ terminates after the first derivative term and Z_σ is constant, depending only on the constant derivative information, on any set on which the active structure functional information is preserved exactly. This shows that optimality is preserved under the current structure provided $\mathbf{z}(\lambda + \delta) \in Z_\sigma$.

To determine the behaviour of $L(\mathbf{x}(\lambda), \lambda))$ on the homotopy trajectory consider a displacement δ that does not involve a slope discontinuity. Then

$$\begin{aligned} \Delta L &= L\left(\mathbf{x} + \delta \frac{d\mathbf{x}}{d\lambda}, \lambda + \delta\right) - L(\mathbf{x}, \lambda) \\ &= f\left(\mathbf{x} + \delta \frac{d\mathbf{x}}{d\lambda}\right) - f(\mathbf{x}) + \lambda \left(g\left(\mathbf{x} + \delta \frac{d\mathbf{x}}{d\lambda}\right) - g(\mathbf{x}) \right) + \delta g\left(\mathbf{x} + \delta \frac{d\mathbf{x}}{d\lambda}\right). \end{aligned}$$

This gives

$$\lim_{\delta \to 0} \frac{\Delta L}{\delta} = L'\left(\mathbf{x} : \frac{d\mathbf{x}}{d\lambda}, \lambda\right) + g(\mathbf{x}),$$

where

$$L'\left(\mathbf{x} : \frac{d\mathbf{x}}{d\lambda}, \lambda\right) = \nabla f(\mathbf{x})^T \frac{d\mathbf{x}}{d\lambda} + \lambda g'(\mathbf{x} : \frac{d\mathbf{x}}{d\lambda}).$$

Also

$$L'\left(\mathbf{x} : \frac{d\mathbf{x}}{d\lambda}, \lambda\right) = 0.$$

This follows from the necessary conditions because

$$\nabla f(\mathbf{x})^T \frac{d\mathbf{x}}{d\lambda} = -\lambda \mathbf{g}^T \frac{d\mathbf{x}}{d\lambda},$$
$$= \lambda \frac{d\mathbf{x}}{d\lambda}^T \nabla^2 f \frac{d\mathbf{x}}{d\lambda} > 0,$$

and

$$g'(\mathbf{x} : \frac{d\mathbf{x}}{d\lambda}) = \sup_{\mathbf{z} \in Z_\sigma} \{\mathbf{g} + V_\sigma \mathbf{z}\}^T \frac{d\mathbf{x}}{d\lambda},$$
$$= \mathbf{g}^T \frac{d\mathbf{x}}{d\lambda} < 0.$$

A consequence is that $f(\mathbf{x})$ increases along the homotopy path while $g(\mathbf{x})$ decreases. For fixed λ the minimizer of L also minimizes $\frac{1}{\lambda} f + g$ so the minimizer must tend to the minimizer $\widehat{\mathbf{x}}$ of g for large λ. It follows that $\widehat{\mathbf{x}}$ is the only limit point as $\lambda \to \infty$. Further, it is attained for finite λ. To see this let the necessary conditions for the minimum of g be

$$\widehat{\mathbf{g}} + \widehat{V}\widehat{\mathbf{z}} = 0,\ \widehat{\mathbf{z}} \in \widehat{Z}^o.$$

Necessarily $\widehat{V}$ has its full row rank so that it follows that there is λ_0 large enough, and a ball $S_0(\lambda_0)$ centred on $\widehat{\mathbf{x}}$ such that

$$\widehat{V}^+ \{-\widehat{\mathbf{g}} - \frac{1}{\lambda}\mathbf{x}\} \in \widehat{Z}, \forall \mathbf{x} \in S_0, \forall \lambda > \lambda_0.$$

Thus $\widehat{\mathbf{x}}$ is the unique minimizer of $L(\mathbf{x}, \lambda),\ \lambda > \lambda_0$.

It remains to consider what happens at the end of the current linear section as λ increases. Two types of behavior trigger slope discontinuities:

1 The multiplier vector $\mathbf{z}_\sigma(\lambda)$ reaches a boundary point of Z_σ. In this case a structure functional (say ϕ_j) is about to become inactive. By (20) this implies an equality

$$\left[\ \mathbf{s}_j^T \quad 1\ \right] P_j^{-1} \mathbf{z}_\sigma = \zeta_j^\pm.$$

Updating the necessary conditions gives

$$\nabla f^T + \lambda \left\{ \mathbf{g}_\sigma + \zeta_j^\pm \mathbf{v}_j + V_j \mathbf{z}_- \right\} = 0. \tag{46}$$

It follows from (18) and continuity of $\mathbf{x}(\lambda)$ that the homotopy continues with the reduced constraint set defined by V_j where

$$\left[\ V_j \quad \mathbf{v}_j\ \right] \begin{bmatrix} S_j & \\ \mathbf{s}_j & 1 \end{bmatrix} = V_\sigma P_j. \tag{47}$$

2 A new nonredundant structure functional ϕ_j becomes active. Here the revised necessary conditions give

$$\nabla f^T + \lambda \left\{ \mathbf{g}_\sigma - \zeta_j^\pm \mathbf{v}_j + \left[\ V_\sigma \quad \mathbf{v}_j\ \right] \begin{bmatrix} \mathbf{z}_\sigma \\ \zeta_j^\pm \end{bmatrix} \right\} = 0. \tag{48}$$

The homotopy continues with the updated set of active structure functionals and continuity requires that the modified multiplier vector move off its bound and into the interior of the updated constraint set.

Example 7.1 *The lasso (31), (32). Here:*

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{r}\|^2,\ g(\mathbf{x}) = \sum_{i=1}^{p} |x_i|,$$

and the structure functionals are

$$\phi_{2i-1} = x_i,\ \phi_{2i} = -x_i,\ i = 1, 2, \cdots, p.$$

Let the current point on the homotopy trajectory be characterized by an index set $|\sigma| = k < p$ pointing to the (appropriately signed) nonzero components of $\mathbf{x}$, so it also indicates the inactive structure functionals, and let P_σ be the permutation matrix such that

$$P_\sigma \mathbf{x} = \begin{bmatrix} \mathbf{x}_\sigma \\ 0 \end{bmatrix},\ P_\sigma \mathbf{g} = \begin{bmatrix} \boldsymbol{\theta}_\sigma \\ 0 \end{bmatrix},\ P_\sigma \mathbf{z} = \begin{bmatrix} 0 \\ \mathbf{z}_\sigma \end{bmatrix}. \tag{49}$$

Also define the Cholesky factorization

$$P_\sigma A^T A P_\sigma^T = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \begin{bmatrix} U_1^T & \\ U_{12}^T & U_2^T \end{bmatrix} \begin{bmatrix} U_1 & U_{12} \\ & U_2 \end{bmatrix}, \tag{50}$$

and note that

$$P_\sigma \begin{bmatrix} V & 0 \end{bmatrix} P_\sigma^T = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}.$$

Then (43), (44) are equivalent to

$$\begin{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \\ \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} & 0 \end{bmatrix} \begin{bmatrix} \begin{bmatrix} \frac{d\mathbf{x}_\sigma}{d\lambda} \\ \frac{d\mathbf{x}_c}{d\lambda} \end{bmatrix} \\ \begin{bmatrix} \frac{d}{d\lambda}(\lambda \mathbf{z}_c) \\ \frac{d}{d\lambda}(\lambda \mathbf{z}_\sigma) \end{bmatrix} \end{bmatrix} + \begin{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_\sigma \\ 0 \end{bmatrix} \\ 0 \end{bmatrix} = 0.$$

It follows immediately that the contributions from the fixed state variables $\mathbf{x}_c$, corresponding to the active structure functionals, and multipliers $\mathbf{z}_c$, corresponding to inactive structure functionals, are zero. The result of solving for the remaining quantities is

$$M_{11} \frac{d\mathbf{x}_\sigma}{d\lambda} + \boldsymbol{\theta}_\sigma = 0$$
$$M_{21} \frac{d\mathbf{x}_\sigma}{d\lambda} + \frac{d}{d\lambda}(\lambda \mathbf{z}_\sigma) = 0.$$

Making use of the factorization (50) of M leads to the equations

$$U_1 \frac{d\mathbf{x}_\sigma}{d\lambda} = -U_1^T \boldsymbol{\theta}_\sigma = -\mathbf{w}_\sigma, \tag{51}$$

and

$$U_{12}^T U_1 \frac{d\mathbf{x}_\sigma}{d\lambda} + \frac{d}{d\lambda}(\lambda \mathbf{z}_\sigma) = -U_{12}^T \mathbf{w}_\sigma + \frac{d}{d\lambda}(\lambda \mathbf{z}_\sigma).$$

Simplifying gives

$$\frac{d}{d\lambda}(\lambda \mathbf{z}_\sigma) = U_{12}^T \mathbf{w}_\sigma. \tag{52}$$

Equations (51) and (52) are equivalent to the homotopy equations given in [11] when the relation

$$\frac{d\lambda}{d\kappa} = -\frac{1}{\|\mathbf{w}_\sigma\|^2}$$

(equation (6.23) in [11]) is used.

Example 7.2 *Support vector regression (33). This example finds the residual structure in the constraint while the objective function is a simple function of the state. We have:*

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2,\ g(\mathbf{x}) = \sum_{1=1}^{n} |r_i|_\varepsilon.$$

The active structure functionals correspond to residual values $r_i = \pm\varepsilon$. At the current point let these be referenced by index sets ε_+ and ε_- respectively, the set corresponding to values $|r_i| < \varepsilon$ by σ_0, and the values $r_i > \varepsilon$, $r_i < -\varepsilon$ by σ_+, and σ_- respectively. Then

$$V_{*i} = A_{\varepsilon(i)*}^T,\ \varepsilon = \varepsilon_+ \cup \varepsilon_-, \tag{53}$$

$$\mathbf{g} = \sum_{i \in \sigma_+ \cup \sigma_-} A_{i*}^T \theta_i,\ \theta_i = sgn(r_i). \tag{54}$$

It proves convenient to compute an orthogonal factorization of V. Let

$$V = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} U \\ 0 \end{bmatrix}.$$

Then

$$\frac{d\mathbf{x}}{d\lambda} = -Q_2 Q_2^T \mathbf{g}, \tag{55}$$

$$\frac{d}{d\lambda}(\lambda \mathbf{z}) = -U^{-1} Q_1^T \mathbf{g}. \tag{56}$$

This shows that the trajectory is piecewise linear. Continuity of the state variable is used to patch the pieces together, and up and downdating to take account of changes in active structure functionals follows standard practice.

8 Computational experience

The complexity of the simplex algorithm for linear programming has been analyzed under random and worst case scenarios. However, this work does not directly apply to polyhedral function minimization when there is the possibility of a line search step, and this case is less well understood. In the case of deterministic approximation problems where there is a well defined set of continuously differentiable basis functions then the best results are obtained for discrete maximum norm approximation in the case that the Haar condition (equals nonsingularity of the $p \times p$ minors of the design matrix) holds. These results have been obtained for a dual simplex algorithm applied to a linear programming formulation of the problem and need not apply to other formulations [10]. This method can be considered as a discretization of the classical first algorithm of Rémès for which p-step second order convergence has been proved [13]. This suggests strongly that a complexity estimate will involve $O(p)$ simplex steps and will be effectively independent of the fineness of the discretization which here determines n. A formal argument has been presented in [10] to show that something similar in the sense of predicting a complexity estimate depending on p happens in the discrete l_1 approximation problem, and here the line search is seriously important in an effective algorithm. If the l_1 problem data is randomly generated, and this is the case, for example, if a popular way of generating test problems with known solutions [1] is used, then these results do not apply. However, computational experience suggests the complexity has a dominant dependence on p but with a further factor depending on n which has slow growth [10]. The two cases need not lend themselves to similar line search strategies [11].

Rank regression has been considered here as an example with significant intrinsic complexity. In this case the empirical evidence obtained from computations with simplicial methods involving a line search strategy suggests that the gross indicators such as the total number of descent steps have a dominant dependence on p which is similar to that in the l_1 case. However, the work per step is somewhat greater and more care is now needed in treating the line search because of the potential $O(n^2)$ slope changes in the computed direction. The secant algorithm is strongly favoured because of the asymptotic linearity properties of the rank regression estimator [8]. These offer more if the model is exact because now there are consistency results [7] which show that a good initial approximation can be obtained by iterating a few least squares problems with a common design matrix. If advantage is taken of this property then the computational experience [11] suggests relatively few iterations of the simplicial algorithm are required. Typically it requires p steps to establish a first extreme point but then few more are required to complete convergence.

Polyhedral function constrained problems fall into two groupings depending on the complexity of the constraint formulation. In the lasso this is low as it depends only on the state variable $\mathbf{x}$ and typically $p \ll n$ in standard data sets. For example, for the Iowa wheat data $p = 9,\ n = 33$, while for the Boston housing data $p = 13,\ n = 506$ - both sets can be found readily by standard web

ε	λ	nits	n0	ne
10	10	121	471	13
	1	113	471	10
	.1	92	459	10
1	10	144	135	13
	1	130	135	13
	.1	201	129	12
.1	10	262	16	13
	1	179	14	12
	.1	183	12	11

Table 9: Active set algorithm: Boston housing data

searches. For both these data sets, for the lasso started at $\kappa = 0$, the homotopy algorithm turns out to be clearly the method of choice as it takes exactly p simplicial steps of $O(np)$ operations applied to an appropriately organised data set [12] to compute the solutions for the full range of κ in each case with two more steps being necessary if an intercept term is included in the housing data. This is essentially the minimum number possible, the cost is strictly comparable with the work required to solve the least squares problem for the full data set, and a great deal more information is obtained.

Support vector regression provides an example in which the residual vector in the linear model appears in the polyhedral function constraint. This now contains a number of terms equal to the number of observations so that it is distinctly more complex than in the lasso. The active set algorithm proves reasonably efficient on the Boston housing data set and results are summarized in Table 9. Here the data presented are the number of iterations to convergence (nits), the number of residuals in the ε-insensitive region (n0) and the number of residuals at the ε bound (ne) for a range of values of λ and ε. Each iteration is an $O(np)$ sweep operation on a similar tableau data structure to that used in the lasso [11]. The total work corresponds very roughly to $O(10)$ solutions of the least squares problem for the corresponding design matrix. For comparison, the corresponding values for the Iowa wheat data are given in Table 10. Here the increase in computing cost for the housing data example suggests a stronger dependence on n than in the lasso computations. This would seem to reflect the additional complexity in the polyhedral constraint.

The homotopy algorithm is relatively less favoured in this case. The obvious starting point in the sense that the solution $\mathbf{x} = 0$ is known is $\lambda = 0$. Results for the two data sets are given in Table 11 and Table 12. The number of iterations is very much larger than in the lasso, and only snapshot results are presented, but these make clear that the initial progress is very slow with only very small increments in λ being taken and with few structure functionals being active at any increment point. Note that this is the region where bias can be expected to be maximized so there well could be a message that in this region results are of little interest. Initially all the residuals are of the same sign and (mostly)

ε	λ	nits	n0	ne
10	10	32	17	9
	1	32	18	8
	.1	33	18	6
1	10	31	3	9
	1	26	2	8
	.1	16	0	6
.1	10	54	1	9
	.1	34	0	8
	.1	18	0	5

Table 10: Active set algorithm: Iowa wheat data

bigger than $|\lambda|$ in both examples, and passing through the ε-insensitive region takes a minimum of two simplicial steps, so that $O(n)$ homotopy steps can be anticipated. However, the process of settling down is reflected in the number of iterations and is more convoluted than one involving just a sequence of simple sign changes. Eventually the final progress to the large λ solution minimizing the polyhedral objective is reasonably efficient. For example, for the housing data, around the last 1000 steps are taken in moving λ up to its final value from values which are order 10^{-4} smaller. This corresponds to between 5 and 10 applications of the active set algorithm and, by construction, each step along the homotopy trajectory is optimal for the current λ values.

These results suggest that homotopy may be most useful in some form of post-optimality strategy. For example, the active set algorithm could be used to find starting values for the homotopy, especially starting values avoiding the small initial values of λ. The homotopy algorithm could then be used to provide more local information for decision making purposes.

References

[1] R. H. Bartels, *A penalty linear programming algorithm method using reduced gradient basis -exchange techniques*, Linear Algebra and Applic. **29** (1980), 17–32.

[2] P. Bloomfield and W. L. Steiger, *Least absolute deviations*, Birkhauser, Boston, 1983.

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput. **20** (1998), no. 1, 33–61.

[4] D. I. Clark and M. R. Osborne, *A descent algorithm for minimizing polyhedral convex functions*, SIAM J. Sci. and Stat. Comp. **4** (1983), 757–786.

[5] Karen George and M. R. Osborne, *The efficient computation of linear rank statistics*, J. Comp. Simul. **35** (1990), 227–237.

ε	λ	nits	n0	ne
.1	6.2813 -7	800	7	1
	1.3640 -4	1600	4	5
	1.2205 -2	2400	11	11
	1.7506 -1	3200	14	11
	1.3873 +2	3504	17	13
1	8.4170 -7	900	63	1
	5.6961 -4	1800	81	5
	2.5095 -2	2700	106	11
	8.5303 +0	3600	134	13
	2.6616 +2	3630	137	13
5	3.3052 -7	600	189	1
	3.1050 -5	1200	276	3
	3.7948 -3	1800	318	9
	1.5889 -1	2400	394	11
	6.1290 +2	2592	405	13

Table 11: Homotopy: Boston housing data

ε	λ	nits	n0	ne
1	6.1039 -7	30	0	1
	4.1825 -6	60	0	1
	6.1329 -6	90	1	4
	1.8249 +0	120	2	7
	6.9885 +0	128	3	9
5	4.7748 -7	25	4	0
	1.5381 -6	50	11	1
	2.1717 -2	75	11	1
	7.9804 -1	100	11	8
	4.1176 +0	112	9	9
10	5.3009 -7	30	10	1
	4.1587 -6	60	18	1
	5.7636 -2	90	19	3
	9.9232 -1	120	18	8
	2.0812 +0	128	17	9

Table 12: Homotopy: Iowa wheat data

[6] T. P. Hettmansperger, *Statistical inference based on ranks*, John Wiley, New York, 1984.

[7] T. P. Hettmansperger and J. W. McKean, *A robust alternative based on ranks to least squares in analyzing linear models*, Technometrics **19** (1977), 275–284.

[8] ———, *Robust nonparametric statistical methods*, Arnold, 1998.

[9] J. W. McKean and T. A. Ryan Jr., *An algorithm for obtaining confidence intervals and point estimates based on ranks in the two sample location problem*, Trans. Math. Software **3** (1977), 183–185.

[10] M. R. Osborne, *Finite algorithms in optimization and data analysis*, Wiley, Chichester, 1985.

[11] ———, *Simplicial algorithms for minimizing polyhedral functions*, Cambridge University Press, 2001.

[12] M. R. Osborne, Brett Presnell, and B. A. Turlach, *A new approach to variable selection in least squares problems*, IMA J. Numerical Analysis **20** (2000), 389–403.

[13] M. J. D. Powell, *Approximation theory and methods*, C. U. P., Cambridge, 1981.

[14] V. Vapnik, S. E. Golowich, and A. Smola, *Support vector method for function approximation, regression estimation, and signal processing*, Advances in Neural Information Processing Systems (M. C. Mozer, M. I. Jordan, and T. Petsche, eds.), MIT Press, 1997.

[15] H. M. Wagner, *Linear programming techniques for regression analysis*, J. Amer. Stat. Assoc. **54** (1959), 206–212.