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

Main reference: ”Simplicial Algorithms for Minimizing Polyhedral Functions”, CUP 2001.

M.R.Osborne
Mathematical Sciences Institute
Australian National University

Abstract: It has been known for 50 years that the discrete l_1 approximation problem can be solved by linear programming (L.P.). However, improved algorithms involve a step which can be interpreted as a line search, and which is not part of the standard simplex algorithm. This is the simplest example of a class of problems with a structure distinctly more complicated than that of the standard nondegenerate LP. Our aim is to uncover this structure for these more general polyhedral functions and to use it to develop what are recognizably simplicial type algorithms. It is necessary to generalise what is meant by degeneracy, and a consequence is that a typical feature is a line search step.

A convex function is the supremum of an affine family:

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

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

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

Example 0: Linear programming problem

This could 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 form

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

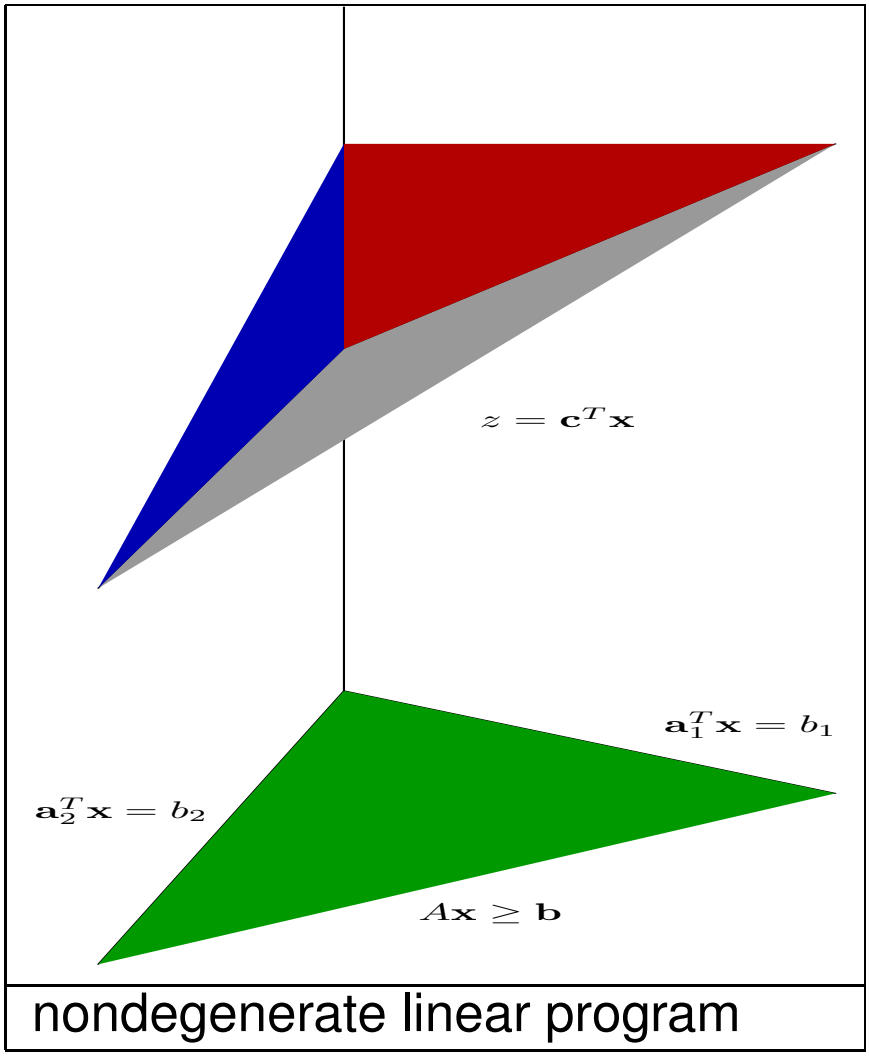
where $A : R^p \rightarrow R^n$, $p < n$. Note that it can be written also as

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

The Kuhn-Tucker conditions characterize the optimum:

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

The linear program supports a simple picture!



nondegenerate linear program

Note that three faces of the epigraph intersect at each extreme point $\mathbf{x} \in R^2$. The case of degeneracy corresponds here to more than three faces intersecting at an extreme point.

Problems which arise in discrete estimation Let the linear model be

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

Here the estimation problem has the form

$$min_{\mathbf{x}} F(\mathbf{r}),$$

where $F(\cdot)$ is a seminorm and polyhedral

We consider algorithms for linear estimation problems which are characterised by:

1 The epigraph of $F(\mathbf{r}(\mathbf{x}))$ is generically degenerate in the sense of linear programming.

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.

It is assumed that 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. 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 active subset.

Example 1: l_1 estimation

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

corresponding to

$$\mathbf{c}_j^T = [\pm 1, \pm 1, \cdots, \pm 1]A, \; j = 1, 2, \cdots$$

Note apparent redundancy when $r_i = 0$. The necessary conditions are:

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

The nondegeneracy condition is

$$|\sigma| = p.$$

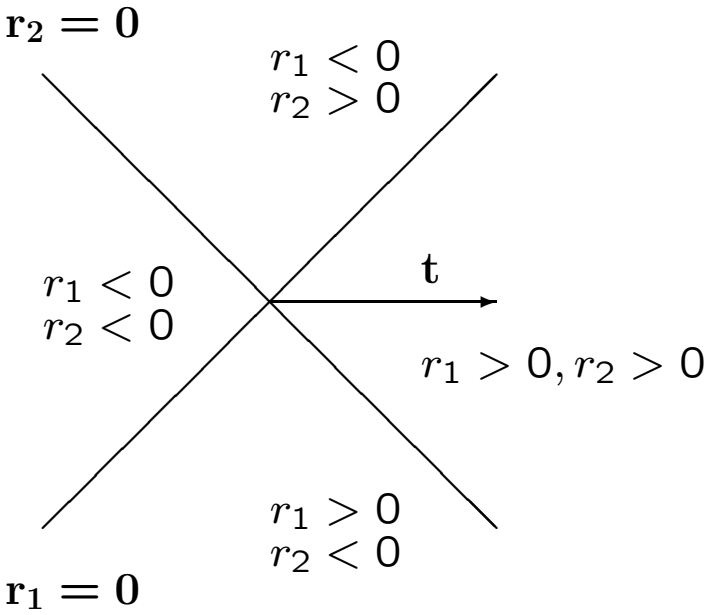
In case $p = 2$ extreme points characterized by (say)

$$\pm r_1(x_1, x_2) = 0,$$
$$\pm r_2(x_1, x_2) = 0.$$

Four faces of epigraph intersect at each extreme point (LP expect 3). $\pm \Rightarrow \theta_i$ such that directions into faces of epigraph satisfy

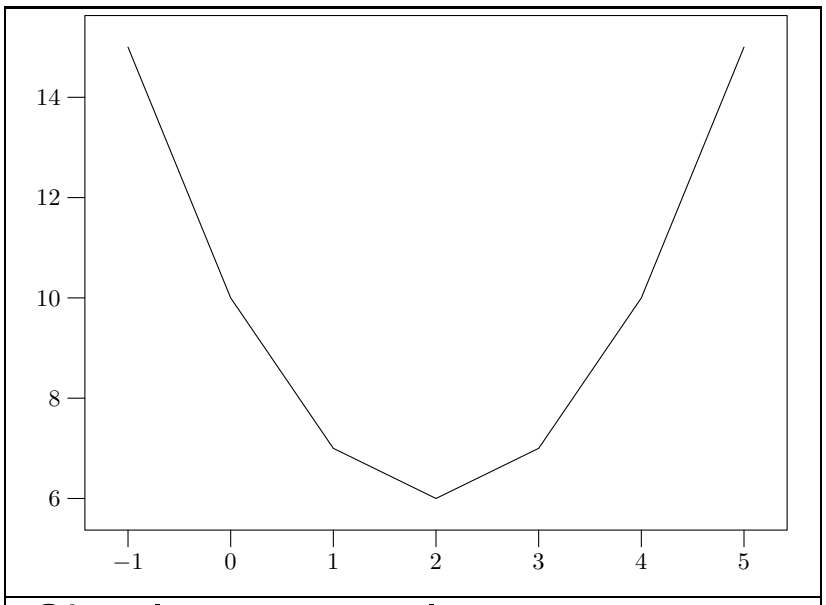
$$\theta_1 A_{1*}\mathbf{t} = \lambda_1 > 0,$$
$$\theta_2 A_{2*}\mathbf{t} = \lambda_2 > 0.$$

for convex combination of edge directions. This convention permits each face to be specified unambiguously!



four faces of epigraph intersect at extreme points $\mathbf{x} \in R^2$

The l_1 problem typically supports a linesearch.



Simple l_1 example

function in figure is

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

Example 2: rank regression.
Let scores w_i nondecreasing and summing to 0 be given - for example $w_i = \sqrt{12}\left(\frac{i}{n+1} - \frac{1}{2}\right),\ i = 1, 2, \cdots, n$

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

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

ν ranking set. Nonsmoothness has its origin in the reordering of scores associated with tied residuals. Here the objective is a seminorm.

The necessary conditions are distinctly more complicated!

6 faces of epigraph intersect at extreme points of epigraph over R^2. Consider the equations characterizing ties:

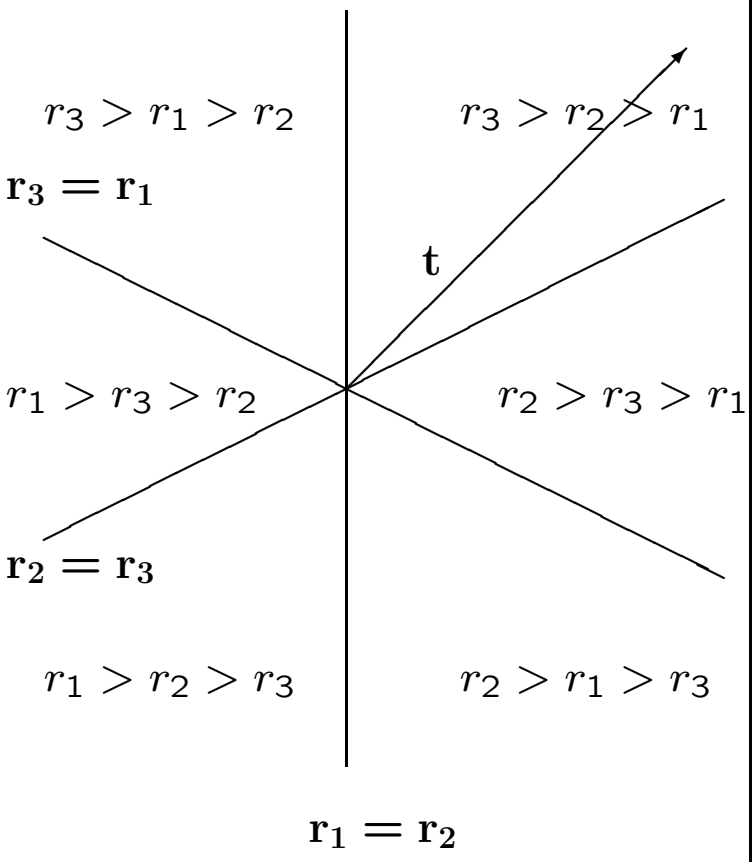
$$\pm (r_2 - r_1) = \pm (r_3 - r_2) = \pm (r_1 - r_3) = 0.$$

Serious redundancy:

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

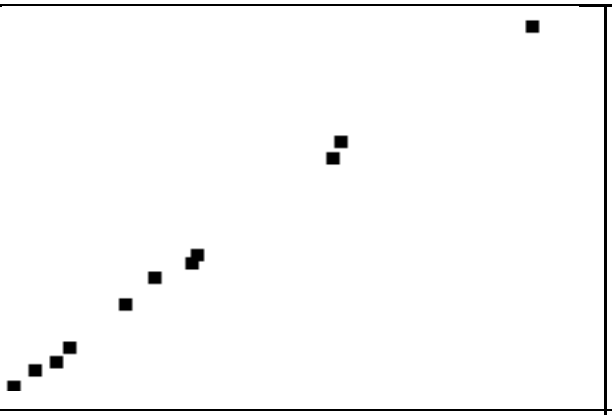
This implies the third line must pass through intersection of first two. Again redundancy in the association of edges and faces can be resolved by looking at directions into faces as convex combinations of directions along edges.

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

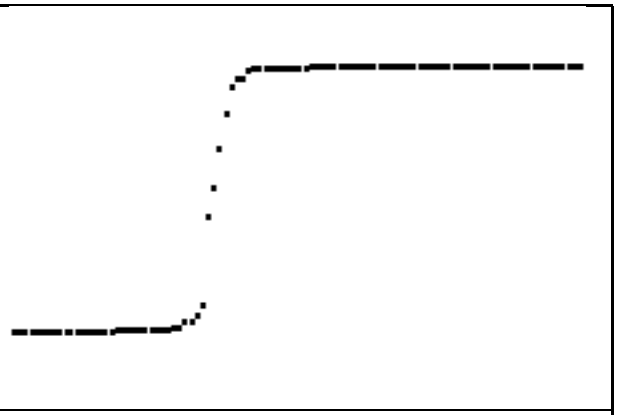


six faces of epigraph intersect at extreme points $\mathbf{x} \in R^2$

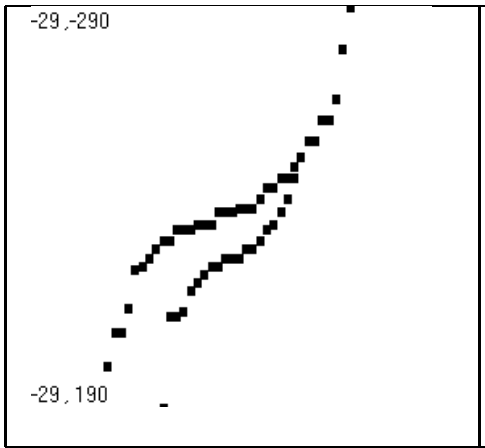
Rank regression examples:



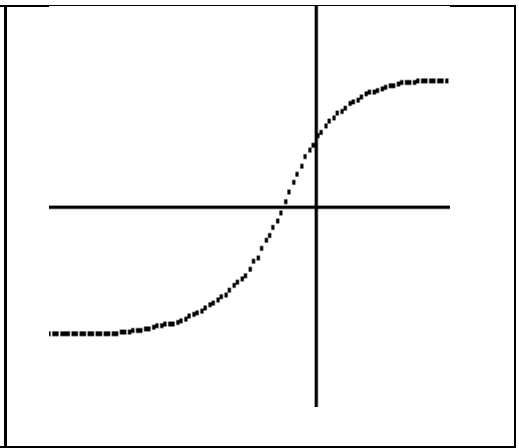
Hubble data



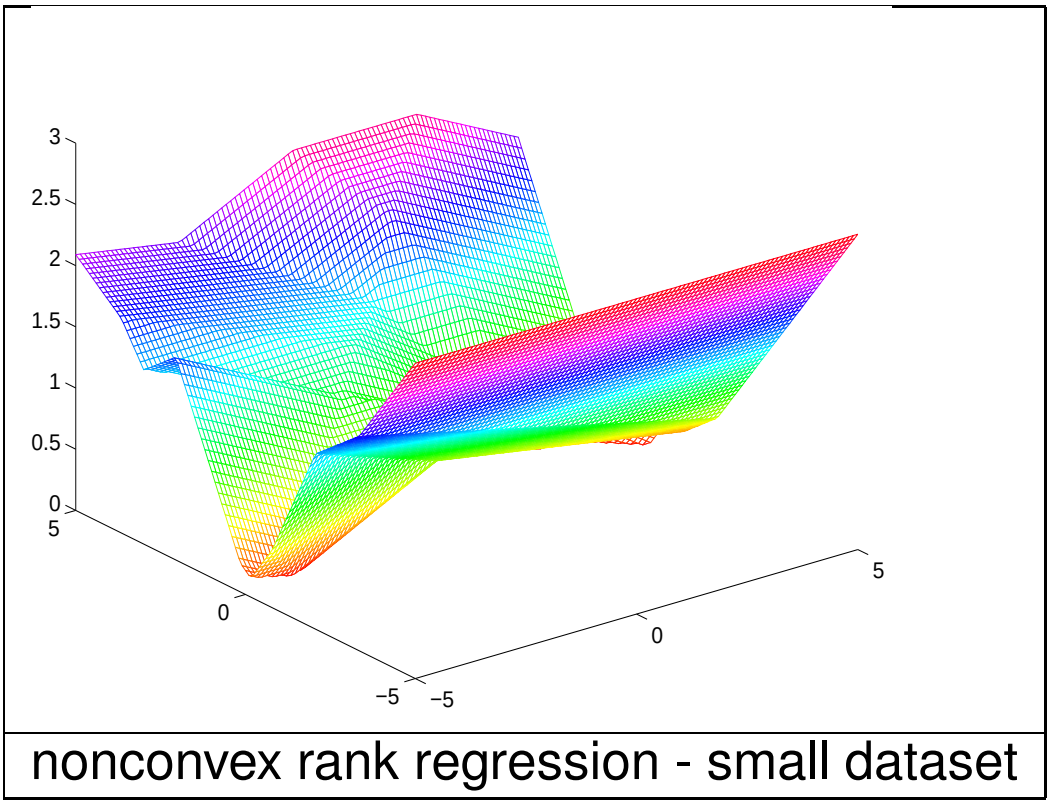
Hubble rank statistic



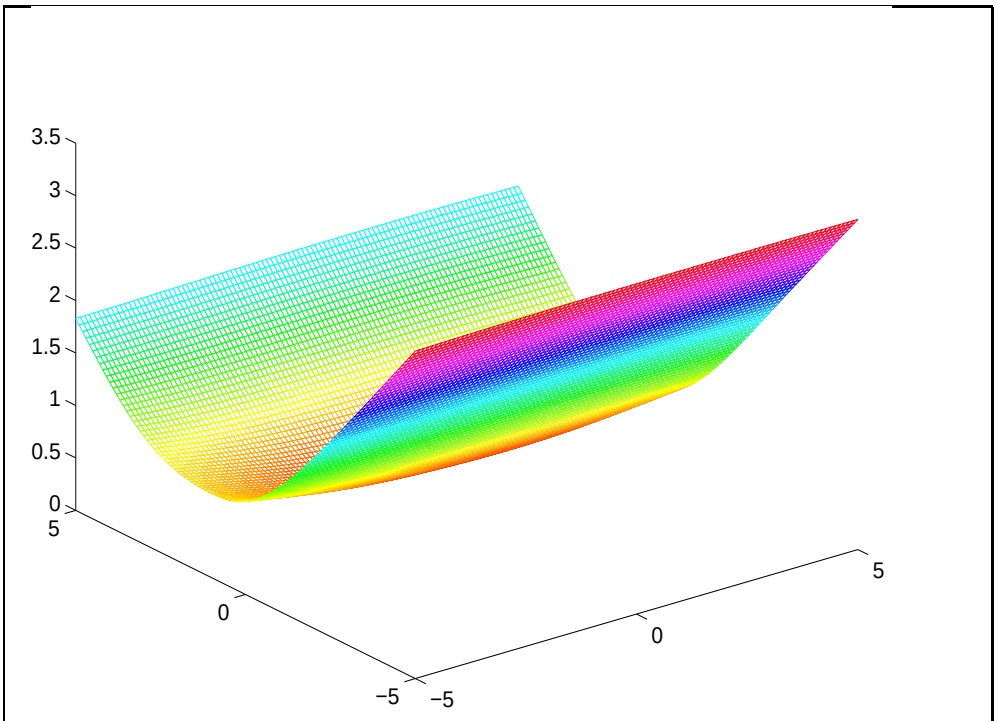
Weight data



Asymptotic linearity



nonconvex rank regression - small dataset



nonconvex rank regression - big dataset

Duality: l_1: Fenchel dual not too bad

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

rank regression: Fenchel dual looks familiar, but ...

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

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.

Type 1 PCF:

$$f(\mathbf{x}) = \max_{1 \le i \le m} \mathbf{c}_i^T \mathbf{x} - d_i.$$

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

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

where σ defines the active set (of structure functionals).

We have already seen examples where the set of structure functionals contains redundant elements!

Redundancy: 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 rank regression example

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

multiplication by -1 is significant!
$\phi_{12},\ \phi_{23}$ and $\phi_{23},\ \phi_{31}$examples of nonredundant pairs.
Say: get nonredundant configurations by *allowable reductions*.
Linear independence: Given set of structure functionals

$$\text{rank}\,(V_\sigma) = k = |\sigma| \leq p.$$
$$V_\sigma^T = \Phi_\sigma^T A \in R^p \rightarrow R^k$$
$$\Phi_\sigma = \left[\ \nabla_r \phi_{\sigma(1)}^T \quad \cdots \quad \nabla_r \phi_{\sigma(k)}^T\ \right] \in R^k \rightarrow R^n.$$

Nondegeneracy: Each allowable reduction of active set is linearly independent.

To generate a compact local representation specialise one of the allowable reductions. Let $\mathbf{x} = \mathbf{x}^* + \varepsilon \mathbf{t}$, $\varepsilon > 0$ small enough. Then, using piecewise linearity of the objective, rearranging gives

$$f\left(\mathbf{r}\left(\mathbf{x}\right)\right) = f_\sigma\left(\mathbf{r}\left(\mathbf{x}\right)\right) + \sum_{i=1}^{|\sigma|} \omega_i\left(\mathbf{t}\right) \phi_{\sigma(i)}\left(\mathbf{r}\left(\mathbf{x}\right)\right),$$

1. f_σ smooth, $\omega_i\left(\mathbf{t}\right)$ provides nonsmooth behaviour.

2. Each distinct realization of $\omega_i\left(\mathbf{t}\right)$, $i = 1, 2, \cdots, |\sigma|$ characterizes one of the faces of $\operatorname{epi}\left(f\right)$ meeting at $\begin{bmatrix} \mathbf{x}^* \\ f\left(\mathbf{r}\left(\mathbf{x}^*\right)\right) \end{bmatrix}$.

An alternative is to use the property displayed in the examples to develop a local description by characterizing the individual faces at a particular extreme point.
Completeness: For each face s, $1 \leq s \leq q$ of $\mathcal{T}\left(\text{epi}\left(f\right), \mathbf{x}^*\right)$ there exists σ_s such that directions into the face

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

satisfy

$$V_{\sigma_s}^T \mathbf{t} = \lambda > 0.$$

Note that $\forall\ s$ system

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

has same solution $\mathbf{x}^*$.

Extreme directions in $\mathcal{T}\left(\text{epi}\left(f\right),\mathbf{x}^{*}\right)$

These are given by

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

There is redundancy here if $q>p+1$. Edges formed by the intersection of adjacent faces (say σ_s,σ_t) are determined by an equation of this form for each face and there is potential here for overspecification. *What characterizes an edge unambiguously is that a particular structure functional in the allowable reductions increases away from zero*.

Example l_1 estimation: 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

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

For each allowable reduction of the active structure functionals $\omega_i\left(\mathbf{t}\right) \phi_{\sigma(i)}\left(r_i\right) = |r_i|\,,\ \omega_i = \pm 1.$

Completeness needs finer structure. For face

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

Differences between sets of equations for extreme directions are pretty trivial in this case.

There are 2^p faces intersecting at x^*.

Example: rank regression

Structure is in ties

$$\phi_{ij} = r_j - r_i, 1 \leq i \neq j \leq n, N = n\,(n-1)\,.$$

Redundancies

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

Possible structure equations when $p = 3$.

$$\begin{aligned} r_1 &= r_2 = r_3 = r_4, \\ r_1 &= r_2, r_3 = r_4 = r_5, \\ r_1 &= r_2, r_3 = r_4, r_5 = r_6. \end{aligned}$$

In first case $\sigma_1 = \{\phi_{12}, \phi_{13}, \phi_{14}\}$ possible set of structure functionals - specializes r_1 (origin!).

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

Redundancy v's completeness!

Not a good set for completeness. If $\mathbf{t}$ is into face $r_1 < r_2 < r_3 < r_4$

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

Relaxed structure functionals do not give right ordering. Right set is ($\sigma_s = \{12, 23, 34\}$)

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

Changing structure functional basis changes representation of non-smooth part of function.

$$\begin{aligned}\sum_{i=1}^{p} \omega_i(\mathbf{t})\,\phi_{\sigma_1(i)}(\mathbf{x}+\mathbf{t}) &= \mathbf{t}^T V_{\sigma_1}\boldsymbol{\omega}(\mathbf{t}),\\ &= \mathbf{t}^T V_{\sigma_1} S_s S_s^{-1}\boldsymbol{\omega}(\mathbf{t}),\\ &= \sum_{i=1}^{p} (\omega_s(\mathbf{t}))_i\,\phi_{\sigma_s(i)}(\mathbf{x}+\mathbf{t})\end{aligned}$$

where $\boldsymbol{\phi}_{\sigma_1}^T S_s = \boldsymbol{\phi}_{\sigma_s}^T$

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

Solution of system $V_{\sigma_s}^T \mathbf{t}_i^s = \mathbf{e}_i,\ i = 1,2,3$,breaks ties

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

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

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

Also subdifferential is convex hull of gradient vectors at nearby differentiable points.
Subgradient $\mathbf{v}$ generalises idea of a gradient vector at points of nondifferentiability of $f(\mathbf{x})$.
Subdifferential is important for characterizing optima and calculating descent directions in nonsmooth convex optimization.
directional derivative:

$$\begin{aligned} f'(\mathbf{x} : \mathbf{t}) &= \inf_{\lambda > 0} \frac{f(\mathbf{x} + \lambda \mathbf{t}) - f(\mathbf{x})}{\lambda}, \\ &= \max_{\mathbf{v} \in \partial f(\mathbf{x})} \mathbf{v}^T \mathbf{t}. \end{aligned}$$

Optimality: $\mathbf{x}$ minimizes $f(\mathbf{x})$ if $0 \in \partial f(\mathbf{x})$.

Recall

$$f\left(\mathbf{r}\left(\mathbf{x}\right)\right)=f_{\sigma}\left(\mathbf{r}\left(\mathbf{x}\right)\right)+\sum_{i=1}^{|\sigma|}\omega_{i}\left(\mathbf{t}\right)\phi_{\sigma(i)}\left(\mathbf{r}\left(\mathbf{x}\right)\right),$$

Implies a representation of subdifferential:

$$\mathbf{v}^{T}\in\partial f\left(\mathbf{r}\left(\mathbf{x}\right)\right)\rightarrow\mathbf{v}=\mathbf{f}_{g}+V_{\sigma}\mathbf{z}.$$

$$\mathbf{f}_{g}=\nabla_{x}f_{\sigma}\left(\mathbf{r}\right)^{T}:\ \text{gradient of smooth part.}$$

$$\left(V_{\sigma}\right)_{*i}=\nabla_{x}\phi_{\sigma(i)}^{T}=\left\{\nabla_{r}\phi_{\sigma(i)}A\right\}^{T},i=1,2,\cdots,|\sigma|,$$

$$\mathbf{z}\in Z_{\sigma}=\mathsf{conv}(\boldsymbol{\omega}_{s},\ s=1,2,\cdots,q).$$

Standard inequality for directional derivative gives

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

Constraint set known if directional derivative known.

Role of extreme directions

Extreme points of Z_σ are determined by the extreme directions associated with the edges of $\mathcal{T}\left(\text{epi}\left(f\right), \mathbf{x}^*\right)$. Key calculation is

$$\begin{aligned} f'\left(\mathbf{x}^* : \mathbf{t}_s\right) &= \mathbf{f}_g^T \mathbf{t}_s + \max_{\mathbf{z} \in Z_\sigma} \mathbf{z}^T V_\sigma^T \mathbf{t}_s, \\ &= \mathbf{f}_g^T \mathbf{t}_s + \max_{\mathbf{z} \in Z_\sigma} \left\{ \sum_{i=1}^{p} \lambda_i \mathbf{z}^T V_\sigma^T \mathbf{t}_i^s \right\}, \\ &\leq \mathbf{f}_g^T \mathbf{t}_s + \sum_{i=1}^{p} \lambda_i \max_{\mathbf{z} \in Z_\sigma} \mathbf{z}^T V_\sigma^T \mathbf{t}_i^s \\ &= \sum_{i=1}^{p} \lambda_i f'\left(\mathbf{x}^* : \mathbf{t}_i^s\right). \end{aligned}$$

It uses the linearity of $f(\mathbf{x})$ on the faces of $\mathcal{T}$ twice.

Computation of Z_σ :

Start with tie $r_1 = r_2 = r_3 = r_4$. If r_1 leaves group on edge then $\phi_{12}, \phi_{13}, \phi_{14}$ all relax. On the edge $\phi_{23} = \phi_{24} = 0$ have to relate $\phi_{12}, \phi_{13}, \phi_{14}$ and $\phi_{12}, \phi_{23}, \phi_{24}$. In general

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

where active set condition on edge is $\Phi_j^T A\mathbf{t} = 0$.

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

This gives the inequalities determining Z_σ in the form

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

Here the bounds are given by

$$\zeta_j^+ = \max_{\mathbf{z}\in Z_\sigma} \mathbf{z}^T P_j^{-T} \begin{bmatrix} \mathbf{s}_j \\ 1 \end{bmatrix},$$

$$\zeta_j^- = \min_{\mathbf{z}\in Z_\sigma} \mathbf{z}^T P_j^{-T} \begin{bmatrix} \mathbf{s}_j \\ 1 \end{bmatrix}.$$

This has sneaked in the assumption that the relaxation could involve either ϕ_j, or $-\phi_j$. Case where this is not true interesting. Consider "non degenerate" vertex having form

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

Structure equations

$$\phi_i(\mathbf{x}) = \left(\mathbf{c}_i^T - \mathbf{c}_{p+1}^T\right)\mathbf{x} - \left(d_i - d_{p+1}\right) = 0.$$

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

$\omega_i(\mathbf{t}) = 1$, $\mathbf{t}$ into i'th face, 0 otherwise,

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

Rank regression: Multiple groups of ties,relaxed structure functional corresponds to one group splitting into two subgroups, and new origin must be found for one of these subgroups.

Original subgroup: $V_0 = \left[\begin{array}{ccc} \nabla_x \phi_1^T & \cdots & \nabla_x \phi_m^T \end{array} \right]$, new, same origin : $V_1 = \left[\begin{array}{ccc} \nabla_x \phi_1^T & \cdots & \nabla_x \phi_{k-1}^T \end{array} \right]$, new, origin:

$$V_2 = \left[\begin{array}{ccc} \nabla_x \left(\phi_{k+1} - \phi_k \right)^T & \cdots & \nabla_x \left(\phi_m - \phi_k \right)^T \end{array} \right].$$

Relation

$$\left[\begin{array}{cc} \left[\begin{array}{cc} V_1 & V_2 \end{array} \right] & \nabla_x \phi_k^T \end{array} \right] \left[\begin{array}{cc} S & \\ \mathbf{s}_k^T & 1 \end{array} \right] = V_0 P,$$

$$\left[\begin{array}{cc} S & \\ \mathbf{s}_k^T & 1 \end{array} \right] = \left[\begin{array}{ccc} I & & 0 \\ & I & 0 \\ 0 & \left[\begin{array}{ccc} 1 & \cdots & 1 \end{array} \right] & 1 \end{array} \right].$$

Obtain, for each mode of separation into subgroups, inequalities

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

To calculate ζ:

Compare two computations of $f'(\mathbf{x}^* : \mathbf{t})$ using first the original group structure, and then the subgroup splitting on the edge. In the first case the origin contribution is $\left(\sum_{i=1}^{m+1} w_i\right) A_{(m+1)*}\mathbf{t}$ and the contribution from a group before splitting is a bound for

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

where the calculation requires that terms which vanish on the edge be grouped. For the new subgroups, only the origin terms contribute as active structure functional terms vanish. The result is

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

when $A_{k*}\mathbf{t} < A_{(m+1)*}\mathbf{t}$. General result is

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

Testing for optimality:

Optimum requires $\exists \widetilde{\mathbf{z}} \in Z$, $0 = \mathbf{f}_g + V\widetilde{\mathbf{z}}$. If $\widetilde{\mathbf{z}} \notin Z$ then there exists a violated member of the set of inequalities

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

Can now generate a descent direction.

Let

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

Then the directional derivative is

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

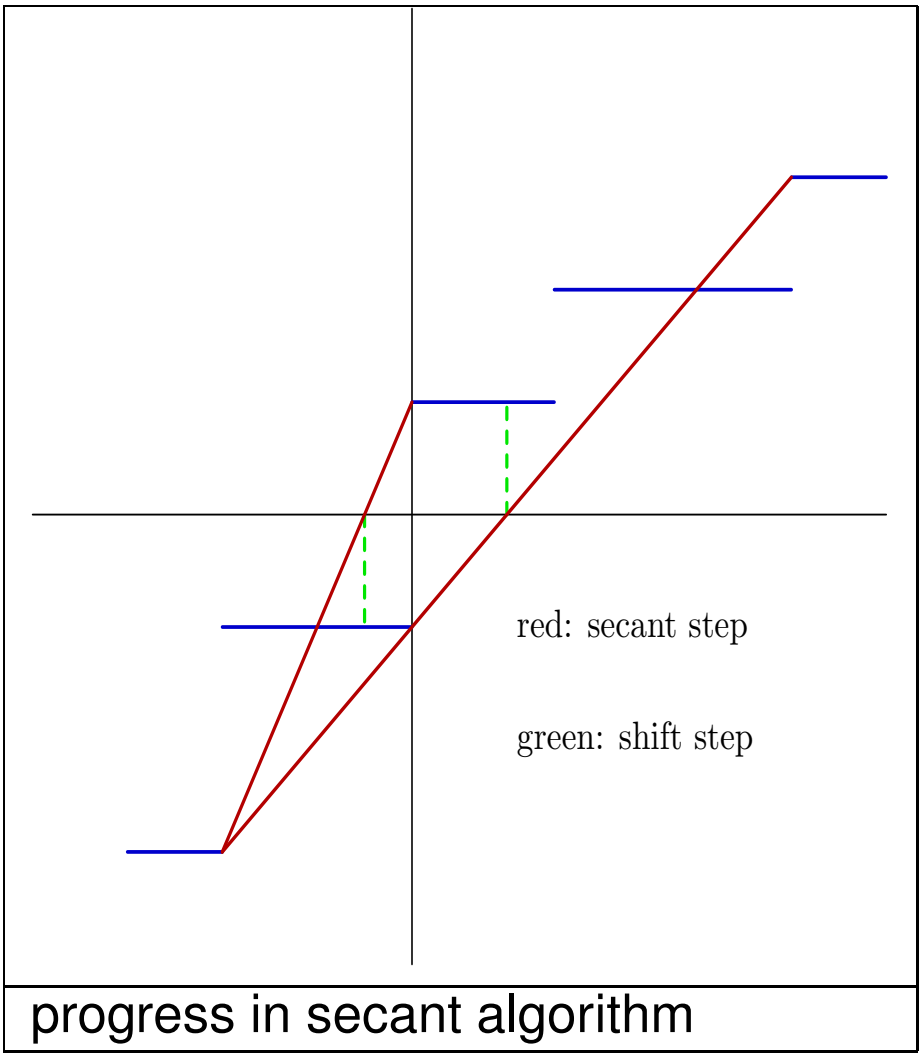
Linesearch: Descent step relaxes off one structure functional while remainder defining edge stay active (simplicial step). Experience is that linesearch in this direction is profitable. Linesearch must terminate at new active structure functional.

l_1: Only necessary to know distances to nonsmooth points in search direction. Required point is a weighted median. Hoare's partitioning algorithm suggested with partition bound defined by standard median of three.

General: Bisection applied to directional derivative to refine bracket of minimum. Explicit computation when bracket contains just one active member.

Statistical estimation: Asymptotic linearity results suggest use of secant algorithm to find axis crossing point of piecewise constant directional derivative. Shifting strategy important.

These are all partitioning methods. Important that evaluation of $f'(\mathbf{x} : \mathbf{t})$ is no worse than $n\gamma(n)$, $\gamma(n)$ of slow growth.



progress in secant algorithm

Polyhedral constrained problems:

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

Here $f(\mathbf{x})$ strictly convex and smooth (typically a quadratic form), and $g(\mathbf{x})$ is polyhedral convex. Assume

$$\widehat{\mathbf{x}} = \arg\ \min_{\mathbf{x}} g(\mathbf{x}) \Rightarrow \kappa \geq g(\widehat{\mathbf{x}})$$

is isolated (global) minimum. Related problem considers the Lagrangian form:

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

Kuhn-Tucker conditions

$$\nabla f(\mathbf{x}) = -\mu \mathbf{v}^T,\ \mathbf{v}^T \in \partial g(\mathbf{x}).$$
$$\kappa \to g(\widehat{\mathbf{x}}),\ \mathbf{x}^* \to \widehat{\mathbf{x}},\ \mu(\mathbf{x}^*) \to \mu(\widehat{\mathbf{x}}),$$
$$\kappa \to \infty, \mathbf{x}^* \to \arg \min_{\mathbf{x}\in \mathrm{eff}(g)} f(\mathbf{x}), \mu(\mathbf{x}^*) \to 0.$$

If $\lambda \geq \mu(\widehat{\mathbf{x}})$, $0 \in \partial g(\widehat{\mathbf{x}})^o$ then $\widehat{\mathbf{x}}$ minimizes $L(\mathbf{x}, \lambda)$. The argument uses that if

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

Problems:

1. ‘Lasso’ provides a new approach to variable selection

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

2. ‘Basis pursuit denoising’

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

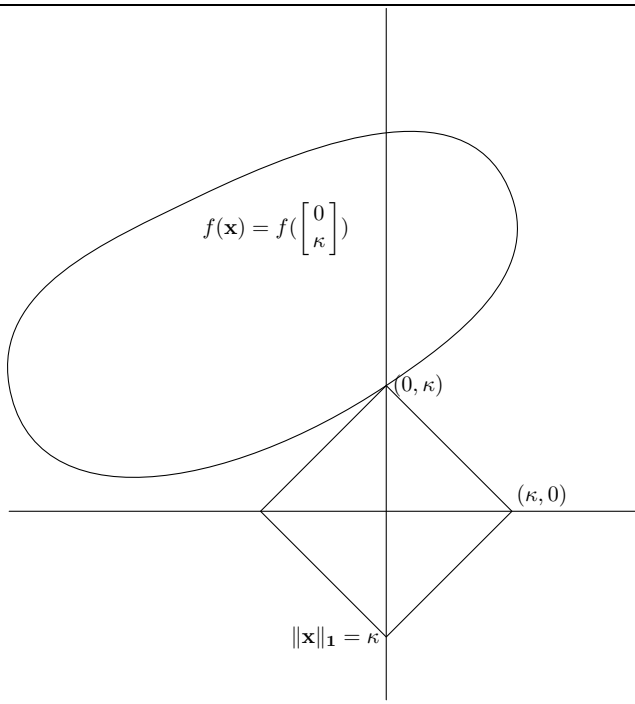
3. ‘Support vector regression’

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

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

Text of Mangasarian's talk on optimization methods in data mining at Atlanta 1999 SIAM meeting interesting source of problems (check his website. He was interested in formulating problems as linear or quadratic programs using inequality representations of polyhedral terms. Interest here is in treating polyhedral functions directly.

The following figure motivates the use of the lasso in variable selection.



A mechanism for variable selection

Basic algorithm: Let

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

Generate direction by solving quadratic program

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

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

lc-feasibility defines region of validity:

- given σ points to active structure functionals
- $\mathbf{g}_g$ is gradient of differentiable part of g.

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

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

$$L'(\mathbf{x} : \mathbf{h}, \lambda) = \max_{\mathbf{v}^T \in \partial L} \mathbf{v}^T \mathbf{h},$$

$$= \max_{\mathbf{z} \in Z_\sigma} \left\{\nabla f(\mathbf{x}_0) + \lambda (\mathbf{g}_g + V_\sigma \mathbf{z})^T\right\}\mathbf{h},$$

$$= \left(\nabla f(\mathbf{x}_0) + \lambda \mathbf{g}_g^T\right)\mathbf{h} < \mathbf{0}.$$

Descent component of active set method:

- compute $\mathbf{h}$ by minimizing $G(\mathbf{x}_0, \mathbf{h})$;
- if $\mathbf{x}_0 + \mathbf{h}$ is an lc-feasible minimum of $L(\mathbf{x}, \lambda)$ then stop;
- else perform linesearch on $L(\mathbf{x} + \gamma\mathbf{h}, \lambda)$, stop at new active structure functional.

If $\mathbf{h} = \mathbf{0}$ lc-feasible minimum then $\exists\, \mathbf{z}_0$

$$\nabla f(\mathbf{x}_0) + \lambda (\mathbf{g}_g + V_\sigma \mathbf{z}_0)^T = 0.$$

$\mathbf{x}_0$ optimal if $\mathbf{z}_0 \in \partial L(\mathbf{x}_0, \lambda)$.

Otherwise it is necessary to :

1. relax an active structure functional associated with a violated constraint;

2. redefine the local linearization.

To update the structure relations ($\sigma \leftarrow \sigma \setminus \{j\}$)

$$\begin{bmatrix} V_j & \mathbf{v}_j \end{bmatrix} \begin{bmatrix} S & \\ \mathbf{s}_j^T & 1 \end{bmatrix} = V_\sigma P_j,$$

$$\mathbf{g}_g^j = \mathbf{g}_g + \zeta_j \mathbf{v}_j,$$

$$\zeta_j = \begin{cases} \zeta_j^-, \begin{bmatrix} \mathbf{s}_j^T & 1 \end{bmatrix} \mathbf{z}_0 < \zeta_j^-, \\ \zeta_j^+, \begin{bmatrix} \mathbf{s}_j^T & 1 \end{bmatrix} \mathbf{z}_0 > \zeta_j^+. \end{cases}$$

Revised QP gives descent direction which is lc-feasible.
Let

$$\mathbf{h}_j = \arg \min_{V_j^T \mathbf{h}=0} G(\mathbf{x}_0, \mathbf{h}).$$

Then $\mathbf{h}$ is a descent direction, and is lc-feasible in the sense that

$$\begin{aligned} \mathbf{v}_j^T \mathbf{h}_j > 0, \left[\begin{array}{cc} \mathbf{s}_j^T & 1 \end{array} \right] \mathbf{z}_0 > \zeta_j^+, \\ < 0, \left[\begin{array}{cc} \mathbf{s}_j^T & 1 \end{array} \right] \mathbf{z}_0 < \zeta_j^-. \end{aligned}$$

The necessary conditions give

$$\begin{aligned} &\nabla^2 f \mathbf{h}_j + \nabla f^T + \lambda \mathbf{g}_g^j + V_j \mathbf{z} = 0, V_j^T \mathbf{h}_j = 0 \\ &\Rightarrow \mathbf{h}_j^T \left(\nabla f^T + \lambda \mathbf{g}_g^j \right) = -\mathbf{h}_j^T \nabla^2 f \mathbf{h}_j < 0. \\ &\mathbf{h}_j^T \nabla^2 f \mathbf{h}_j + \mathbf{h}_j^T \left(\nabla f^T + \lambda \mathbf{g}_g \right) + \lambda \zeta_j \mathbf{h}_j^T \mathbf{v}_j = 0 \end{aligned}$$

Also

$$\begin{aligned} 0 &= \mathbf{h}_j^T \left(\nabla f^T + \lambda \mathbf{g}_g \right) + \lambda V_\sigma \mathbf{z}_0 \\ &= \mathbf{h}_j^T \left(\nabla f^T + \lambda \mathbf{g}_g \right) + \lambda \left[\begin{array}{cc} \mathbf{s}_j^T & 1 \end{array} \right] \mathbf{z}_0 \mathbf{h}_j^T \mathbf{v}_j \\ &\Rightarrow \mathbf{h}_j^T \nabla^2 f \mathbf{h}_j + \lambda \left(\zeta_j - \left[\begin{array}{cc} \mathbf{s}_j^T & 1 \end{array} \right] \mathbf{z}_0 \right) \mathbf{h}_j^T \mathbf{v}_j = 0 \end{aligned}$$

Homotopy approach: Assume $\mathbf{x}, \lambda$ are optimal, and that an index set σ points to the active structure functionals. Differentiating the necessary conditions wrt λ gives

$$\nabla^2 f \frac{d\mathbf{x}}{d\lambda} + \lambda V_\sigma \frac{d\mathbf{z}}{d\lambda} = -\left(\mathbf{g} + V_\sigma \mathbf{z}\right),$$
$$V_\sigma^T \frac{d\mathbf{x}}{d\lambda} = 0.$$

This system can now be used to obtain a differential equation for $\mathbf{z}$:

$$\lambda \frac{d\mathbf{z}}{d\lambda} + \mathbf{z} = \mathbf{a},$$
$$\mathbf{a} = -\left(V_\sigma^T (\nabla^2 f)^{-1} V_\sigma\right)^{-1} V_\sigma^T (\nabla^2 f)^{-1} \mathbf{g}.$$

The general solution is

$$\mathbf{z} = \lambda^{-1}\mathbf{c} + \mathbf{a},$$

where $\mathbf{c}$ is chosen at each updating of σ to ensure continuity. $\mathbf{x}$ is piecewise linear in λ and satisfies $\frac{d\mathbf{x}}{d\lambda} = -(\nabla^2 f)^{-1}(I - S)\,\mathbf{g}$, where S is the oblique projection onto the column space of V_σ.

Trajectory slope discontinuities There are two causes for a slope discontinuity in the piecewise linear $\mathbf{x}$ trajectory.

1. The multiplier vector $\mathbf{z}_\sigma(\lambda)$ reaches a boundary point of Z_σ. This implies an equality

$$\left[\begin{array}{cc} \mathbf{s}_j^T & 1 \end{array}\right] P_j^{-1} \mathbf{z}_\sigma = \zeta_j^{\pm}$$

This corresponds to a reduced constraint set defined by V_j and revised necessary conditions:

$$\left[\begin{array}{cc} V_j & \mathbf{v}_j \end{array}\right] \left[\begin{array}{cc} S_j & \\ \mathbf{s}_j & 1 \end{array}\right] = V_\sigma P_j,$$

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

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

$$\nabla f^T + \lambda \left\{ \mathbf{g}_\sigma - \zeta_j^{\pm} \mathbf{v}_j + \left[\begin{array}{cc} V_\sigma & \mathbf{v}_j \end{array}\right] \left[\begin{array}{c} \mathbf{z}_\sigma \\ \zeta_j^{\pm} \end{array}\right] \right\} = 0.$$

Examples We consider both the lasso and support vector regression applied to the Iowa wheat data (p=9,n=33), and Boston housing data (p=13, n=506).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 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 effective:

ε	λ	nits	n0	ne	nits	n0	ne
10	10	121	471	13	32	17	9
	1	113	471	10	32	18	8
	.1	92	459	10	33	18	6
1	10	144	135	13	31	3	9
	1	130	135	13	26	2	8
	.1	201	129	12	16	0	6
.1	10	262	16	13	54	1	9
	1	179	14	12	34	0	8
	.1	183	12	11	18	0	5

Active set: housing data, wheat data

The homotopy algorithm is relatively less favoured in this case. The obvious starting point in the sense that the solution $\mathbf{x} = 0$, $\lambda = 0$ is known. A characteristic is a slow beginning with repeated changes in little evident structure.

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

Homotopy: Iowa wheat data

In the housing data something needs to be done to escape the small values of λ. The active set algorithm could be useful here.

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

Homotopy: Boston housing data