Polyhedral constrained optimization

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

M.R.Osborne
Mathematical Sciences Institute
Australian National University

Abstract: Typically, polyhedral constraints on constrained optimization problems have been replaced by equivalent families of linear constraints. However, this will always expand the problem size, and in extreme cases is impractical. The polyhedral constraint can be used directly given a compact, general method of describing its geometry. The simplest problem has the form

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

where $f(\mathbf{x})$ is strictly convex, $g(\mathbf{x})$ polyhedral convex. The Lagrangian form

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

is of independent interest. Several applications are summarized. Active set and homotopy algorithms are described briefly.

Problems: Assume a linear model

$$\mathbf{r} = A\mathbf{x} - \mathbf{b}, \quad A : R^p \to R^n, n > p.$$

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'

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

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

Polyhedral convex functions:

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

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).

An important feature that can contribute significantly to algorithmic complexity is the potential for the intrinsic presence of redundant elements in the set of structure functionals!

Polyhedral function example 1: l_1 norm

$$g(\mathbf{x}) = \sum |r_i| \quad \mathbf{r} = A\mathbf{x} - \mathbf{b} \quad A : R^p \to R^n.$$

corresponding to the affine representation

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

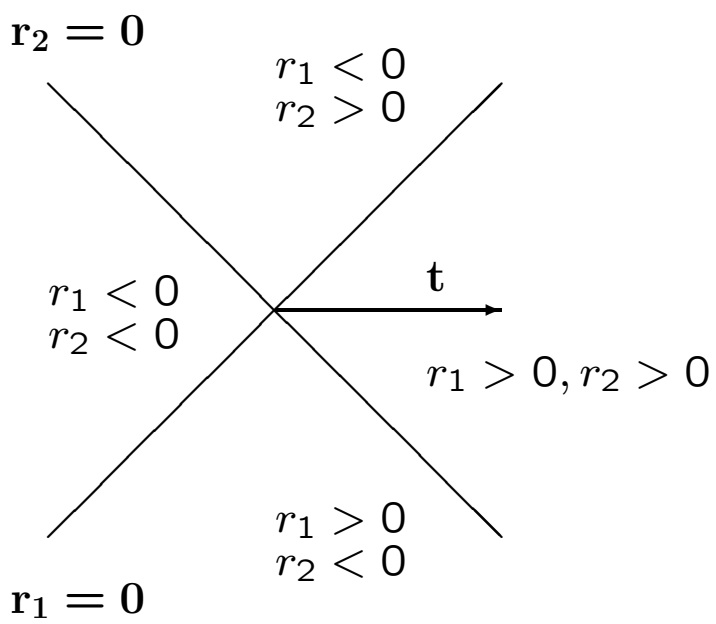
Note apparent redundancy when $r_i = 0$. In case $p = 2$ extreme points characterized by (say)

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

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 $\mathbf{x} \in R^2$

Polyhedral function 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:

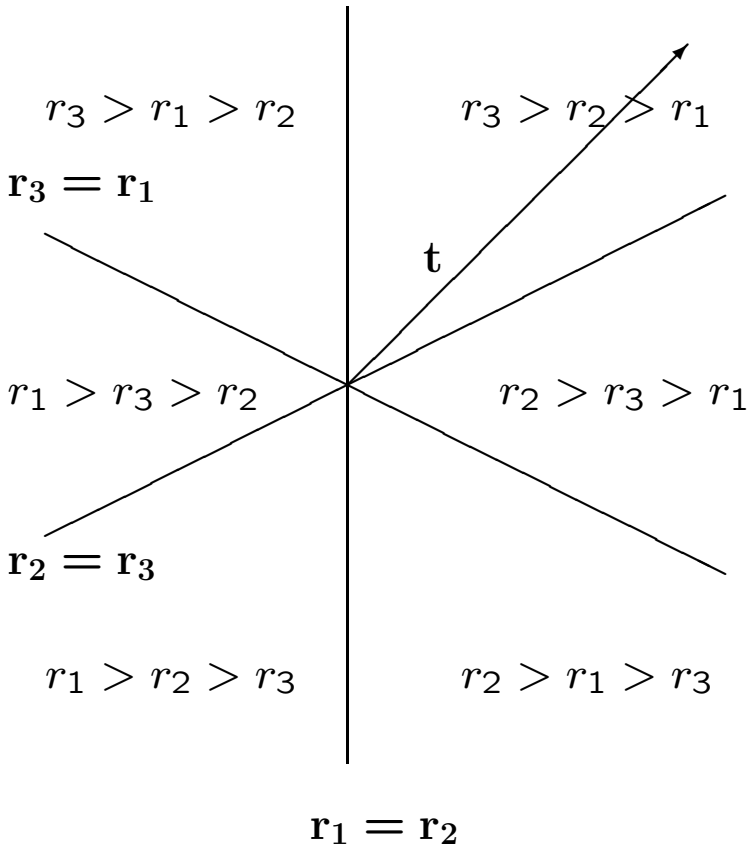
$$\pm (r_2 - r_1) = \pm (r_3 - r_2) = \pm (r_1 - r_3) = 0.$$

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 $\mathbf{x} \in R^2$

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!
ϕ_{12}, ϕ_{23} and ϕ_{23}, ϕ_{31}examples of nonredundant pairs. Say: get nonredundant configurations by *allowable reductions*.
Linear independence: Given set of structure functionals

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

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

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

1. f_σ smooth, $\omega_i(\mathbf{t})$ provides 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 $\begin{bmatrix} \mathbf{x}^* \\ f(\mathbf{r}(\mathbf{x}^*)) \end{bmatrix}$.

An alternative is to use the property displayed in the examples to develop a local description. Simplest case is an extreme point but results apply in other cases as well. Aim is to characterize the individual faces at a particular extreme point.

Completeness: For each face s, $1 \leq s \leq q$ of $T\left(\text{epi}\left(f\right), \mathbf{x}^*\right)$ 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

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

Completeness closely linked to redundancy. $\forall\ s$ system

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

has same solution $\mathbf{x}^*$. Redundancy requires the further condition $q > p + 1$.

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

These are determined by the extreme directions $\boldsymbol{\lambda}$ in R^{p+} for each face. That is by

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

This is typically an overspecification. Edges formed by the intersection of adjacent faces (say σ_s, σ_t) are determined by an equation of this form for each face . *What is common is that a particular structure functional in the allowable reductions increases away from zero.*

Example l_1 function: 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 function

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 $p = 3$.

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

In 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.

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

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

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

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

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

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:

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

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

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

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

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.

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

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

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

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

Kuhn-Tucker conditions

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

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

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

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'

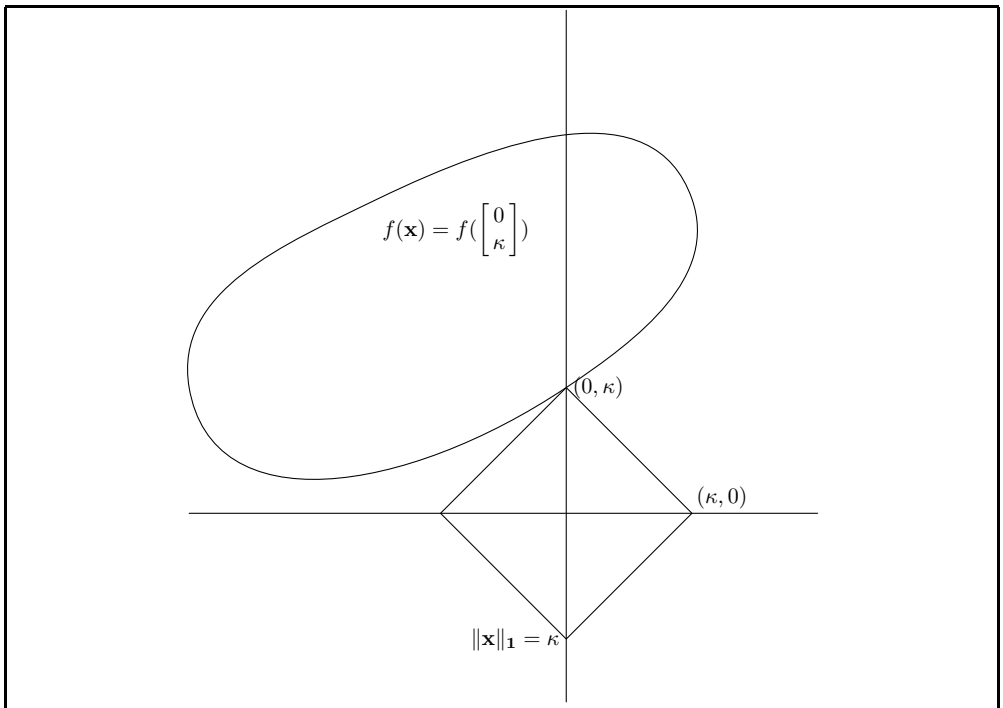
$$\min\left\{\frac{1}{2}\mathbf{r}^T\mathbf{r}+\lambda\|\mathbf{x}\|_1\right\}.$$

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 SIAM meeting interesting source of problems such as support vector regression (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.



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 $L(\mathbf{x} + \gamma \mathbf{h}, \lambda)$.

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