Polyhedral function constrained optimization problems

M.R.Osborne
Mathematical Sciences Institute
Australian National University

Abstract: Recently polyhedral functions have proved distinctly useful in expressing selection criteria in various model building techniques. Here they play the role of a constraint on an estimation problem. While they can always be replaced by an appropriate family of linear constraints the result can be a very large constraint set. Compact representations are available and their use is illustrated by developing both active set and homotopy algorithms. Their use is illustrated using some well known data sets.

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

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

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’

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

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

Let $g(\mathbf{x})$ be polyhedral convex (supremum of a finite affine family). Non-smooth points $\mathbf{x}^*$ of the epigraph characterized by vanishing of certain linear functions ("structure functionals")

$$\phi_i(\mathbf{x}^*) = 0, \quad i \in \sigma.$$

This characterization is not unique. Each face $1 \leq s \leq q$ of tangent cone $\mathcal{T}$ at $\mathbf{x}^*$ is characterized by a particular reduced set σ_s with property that directions $\mathbf{t}$ into face satisfy

$$V_s^T \mathbf{t} = \boldsymbol{\lambda} > 0, \quad V_s = \nabla \boldsymbol{\phi}_{\sigma_s}^T.$$

$g(\mathbf{x})$ has local representation

$$g(\mathbf{x}) = g_s(\mathbf{x}) + \sum_{i \in \sigma_s} w_i^s \phi_i(\mathbf{x}),$$

and its subdifferential - convex hull of gradients at nearby diffferentiable points - is given by

$$\mathbf{v} = \mathbf{g}_s + V_s \mathbf{z}_s, \quad \mathbf{z}_s \in Z_s = \text{conv}\{\mathbf{w}^s\}.$$

Edges of $\mathcal{T}$ found by dropping particular ϕ_i. Each relation has form

$$\left[\begin{array}{cc} \nabla \boldsymbol{\phi}_i^T & \nabla \phi_i^T \end{array} \right] \left[\begin{array}{cc} S_i^s & \\ \mathbf{s}_i^s & 1 \end{array} \right] = V_s P_i.$$

Edge condition is $\nabla \boldsymbol{\phi}_i \mathbf{t} = 0$, P_i is a permutation matrix.

Edges of $\mathcal{T}$ generate extreme points of Z_s which has representation as a system of linear inequalities, one for each ϕ_i.

$$\zeta_i^- \leq \left[\begin{array}{cc} \mathbf{s}_i^T & 1 \end{array} \right] \leq \zeta_i^+$$

ζ_i^-, ζ_i^+ computed from directional derivative of $g(\mathbf{x})$.

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

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

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

$\mathbf{x}$ (in particular $\mathbf{x} + \mathbf{h}$) is lc-feasible provided:

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

Linesearch ends either with new active structure functional or zero derivative of directional derivative.

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 $0 \in \partial L(\mathbf{x}_0, \lambda)$, $\mathbf{z}_0 \in Z_\sigma$. Otherwise it is necessary to :

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

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

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

Homotopy approach: Assume $\mathbf{x}, \lambda$ are optimal, that an index set σ points to the active structure functionals, and $\mathbf{z}_\sigma \in Z^o_\sigma$. 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 equations for $\mathbf{z}_\sigma$ and $\mathbf{x}$:

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

where S is the oblique projection onto the column space of V_σ. $\mathbf{x}$ and $\lambda \mathbf{z}_\sigma$ are piecewise linear and continuous in λ.

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