Polyhedral function constrained optimization problems

M. R. Osborne*

(Received 25 October 2004, revised 1 March 2005)

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. Whereas they can always be replaced by an appropriate family of linear constraints, the resulting set can be a very large. Compact representations are available and their use is illustrated by developing both active set and homotopy algorithms for the general polyhedral constrained problem. These are illustrated using some well known data sets.

*Mathematical Sciences Institute, Australian National University, ACT 0200, Australia. mailto:Mike.Osborne@maths.anu.edu.au

See http://anziamj.austms.org.au/V46/CTAC2004/Osbo for this article, Published April 22, 2005. ISSN 1446-8735

Contents

1 Introduction

The simplest form of polyhedral constrained optimization problem is

$$\min_{\mathbf{x}\in X} f(\mathbf{x})\,; \quad X = \{\mathbf{x} : \kappa \geq g(\mathbf{x})\}\,. \tag{1}$$

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

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})\,. \tag{2}$$

Note that L is strictly convex for all $\lambda \geq 0$ and hence has an unique minimum.

Remark 1 To relate the Lagrange multiplier for (1) where it can be considered as a function of κ with the value of λ in (2) where it can be assigned a priori [1] assume

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

is an isolated (global) minimum of $g(\mathbf{x})$ so that $\kappa \geq g(\widehat{\mathbf{x}})$ is a necessary condition on (1). The Kuhn–Tucker conditions for (1) are

$$\exists \left\{\mathbf{v}^T \in \partial g(\mathbf{x})\,,\ \mu \geq 0\right\} \ni \nabla f(\mathbf{x}) = -\mu \mathbf{v}^T. \tag{3}$$

Now, as $\kappa \to g(\widehat{\mathbf{x}})$, both the computed solution $\mathbf{x}^* \to \widehat{\mathbf{x}}$, and $\mu(\mathbf{x}^*) \to \mu(\widehat{\mathbf{x}})$, while as $\kappa \to \infty$, then $\mathbf{x}^* \to \arg\min_{\mathbf{x}\in \mathrm{eff}(g)} f(\mathbf{x})$, and $\mu(\mathbf{x}^*) \to 0$. Here $\mathrm{eff}(g) = \{\mathbf{x} : g(\mathbf{x} < \infty\}$. The interesting result is that if $\lambda \geq \mu(\widehat{\mathbf{x}})$, $0 \in \partial g(\widehat{\mathbf{x}})^o$ (set interior) then $\widehat{\mathbf{x}}$ minimizes $L(\mathbf{x}, \lambda)$. The argument uses

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

Several recent papers have considered optimisation problems having this form in a modelling context. Here

$$\mathbf{r} = \mathbf{y} - A\mathbf{x}$$

where $\mathbf{y}$ is a vector of noisy observations, $A : R^p \to R^n$ is the design matrix, and $\mathbf{x}$ is a vector of parameters to be estimated from the observed data. There is interest in the case $p > n$ in variable selection problems.

1. The 'lasso' [6, 4] provides a new approach to variable selection. The constrained optimization problem is

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

The 'extended lasso' seeks a common set of predictor variables from a class of p possibilities in order to model k species on the basis of n observations on each species by considering the constrained problem

$$\min_{\mathbf{x}} \sum_{i=1}^{n}\sum_{j=1}^{k}\left(r_i^j\right)^2; \quad \sum_{m=1}^{p}\max_{1\leq j\leq k}|x_m^j| \leq \kappa.$$

This problem is work in progress by Turlach, Venables, and Wright. It is proving to be amenable to similar techniques to the 'lasso'.

2. The corresponding Lagrangian form of the lasso is

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

It has been considered in 'basis pursuit denoising' [1].

3. A somewhat more complex polyhedral constraint occurs in 'support vector regression' [7]. This problem is

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

where

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

These problems exhibit explicit dependence on a parameter (either κ or λ). This provides a mechanism for developing homotopy methods to completely describe the solution path, and it is observed that in lasso like problems the work required is little more than that for a single active set minimization. This has attracted considerable recent interest [2, 5, 8, e.g.]. However, support vector regression provides an example which shows that not all problems of this class can be solved so economically. Thus an active set method proves to be an important part of a general tool kit.

The advantage in our approach, which concentrates on describing the local structure of $g(\mathbf{x})$ polyhedral convex [3], is that it avoids the potentially very large constraint sets that follow from the familiar representation as the supremum of a finite affine family. Non-smooth points $\mathbf{x}^*$ of the epigraph are characterized by the vanishing of certain linear functions or "structure functionals", characteristic of g, pointed to by an index set σ:

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

This characterization typically contains redundant equations and an efficient set is obtained by considering the tangent cone $\mathcal{T}$ which at each non smooth point inherits the polyhedral structure. Each (plane) face s is characterized by a particular reduced set written in vector form $\boldsymbol{\phi}^s$ with components ϕ_i pointed to by $\sigma_s \subset \sigma$. The defining properties are that directions $\mathbf{t}$ into this face satisfy

$$V_s^T \mathbf{t} = \boldsymbol{\lambda} > 0\,, \quad V_s = (\nabla \boldsymbol{\phi}^s)^T\,,$$

and that V_s has full column rank. Let $\mathbf{x} = \mathbf{x}^* + \mathbf{t}$. Then piecewise linearity permits the local representation

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

where $g_s(\mathbf{x})$ is smooth, and nonsmoothness is captured in the coefficients $w_i^s(\mathbf{t})$ of the structure functional terms. The subdifferential at $\mathbf{x}^*$ is

$$\mathbf{v} = \mathbf{g}_s + V_s \mathbf{z}_s, \quad \mathbf{g}_s = \nabla g_s^T, \quad \mathbf{z}_s \in Z_s = \operatorname{conv}\{\mathbf{w}^s(\mathbf{t})\}.$$

Each edges of $\mathcal{T}$ is found by dropping a particular component ϕ_i from $\boldsymbol{\phi}^s$. Each relation has the form

$$\begin{bmatrix} \nabla \boldsymbol{\phi}_i^T & \nabla \phi_i^T \end{bmatrix} \begin{bmatrix} S_i^s & \mathbf{0} \\ \mathbf{s}_i^s & 1 \end{bmatrix} = V_s P_i,$$

where the edge condition is $\nabla \boldsymbol{\phi}_i \mathbf{t} = 0$, and P_i is a permutation matrix. Edges of $\mathcal{T}$ generate the extreme points of the subdifferential constraint set Z_s which has an explicit representation

$$\zeta_i^- \le \begin{bmatrix} \mathbf{s}_i^T & 1 \end{bmatrix} P_i^{-1} \mathbf{z} \le \zeta_i^+, \quad i \in \sigma_s. \tag{4}$$

The bounds ζ_i^- and ζ_i^+ can be computed when the directional derivative of $g(\mathbf{x})$ is available [3].

2 An active set algorithm

The terminology is intended to indicates that active structure functionals play a similar role to active constraints in standard optimization problems. This analogy is extended here by the development of what is essentially an SQP algorithm for the Lagrangian form of the problem. Let the subdifferential based on a particular face specification be

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

The algorithm generates a descent direction by solving the quadratic program subproblem

$$\min_{V_\sigma^T \mathbf{h}=0} G(\mathbf{x}_0, \mathbf{h}) \,, \tag{5}$$

where

$$G(\mathbf{x}_0, \mathbf{h}) = \left(\nabla f(\mathbf{x}_0) + \lambda \mathbf{g}_g^T\right) \mathbf{h} + \frac{1}{2} \mathbf{h}^T \nabla^2 f \mathbf{h} \,. \tag{6}$$

The subproblem (5) is compatible with the local active structure provided:

- the given σ points to a basis set of active structure functionals, and
- relative to this structure $\mathbf{g}_g$ is the gradient of the differentiable part of g.

Points where this local representation of the problem holds are said to be lc-feasible.

The solution of (5) generates a descent direction. Let $\mathbf{h}$ minimize G. Iff $\|\mathbf{h}\| \neq 0$, then $\mathbf{h}$ is a descent direction for minimizing $L(\mathbf{x}, \lambda)$. First note the result that

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

This is used in the calculation of the directional derivative:

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

The basic steps of the algorithm when $\mathbf{h} \neq 0$ are:

- 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 a line search on $L(\mathbf{x_0} + \gamma\mathbf{h}, \lambda)$.

The line search stops either at a new active structure functional which must then be added to the active set, or at a point where the directional derivative vanishes, and both possibilities need to be considered.

The alternative situation corresponds to $\mathbf{h} = 0$. If this is an lc-feasible minimum then there exists $\mathbf{z}_0$ such that

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

If $0 \in \partial L(\mathbf{x}_0, \lambda)$, $\mathbf{z}_0 \in Z_\sigma$ then $\mathbf{x}_0$ is optimal. 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 ($\sigma \leftarrow \sigma \backslash \{j\}$) use

$$\begin{bmatrix} V_j & \mathbf{v}_j \end{bmatrix} \begin{bmatrix} S & \mathbf{0} \\ \mathbf{s}_j^T & 1 \end{bmatrix} = V_\sigma P_j ,$$

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

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

The key result is that the revised QP gives a descent direction which is lc-feasible for the revised active set. Let

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

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

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

where the inequalities indicate the manner in which the deleted structure functional departs from 0. These results follow from the necessary conditions defining the new descent direction. In outline:

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

Also

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

3 A homotopy approach

This was considered first for the lasso in [4]. Assume $\mathbf{x}, \lambda$ are optimal, that an index set σ points to the active structure functionals, and that the multiplier vector $\mathbf{z}_\sigma \in Z_\sigma^o$, the interior of the constraint set Z_σ. Differentiating the necessary conditions with respect to λ gives

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

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

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

The corresponding equation for $\mathbf{x}$ is

$$\frac{d\mathbf{x}}{d\lambda} = -(\nabla^2 f)^{-1}\,(I - S)\,\mathbf{g}\,,$$

where S is the oblique projection onto the column space of V_σ. The right hand sides are locally constant so that $\mathbf{x}$ and $\lambda\mathbf{z}_\sigma$ are piecewise linear and continuous in λ.

There are two causes for slope discontinuities in the piecewise linear optimal 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:
$$\left[\begin{array}{cc} V_j & \mathbf{v}_j \end{array}\right] \left[\begin{array}{cc} S_j & \mathbf{0} \\ \mathbf{s}_j & 1 \end{array}\right] = V_\sigma P_j\,,$$
$$\nabla f^T + \lambda\left\{\mathbf{g}_\sigma + \zeta_j^\pm \mathbf{v}_j + V_j\mathbf{z}_j\right\} = 0\,.$$

2. A new nonredundant structure functional ϕ_j becomes active. Here the revised necessary conditions give
$$\nabla f^T + \lambda\left\{\mathbf{g}_\sigma - \zeta_j^\pm\mathbf{v}_j + \left[\begin{array}{cc} V_\sigma & \mathbf{v}_j \end{array}\right]\left[\begin{array}{c} \mathbf{z}_\sigma \\ \zeta_j^\pm \end{array}\right]\right\} = 0\,.$$

Updating to take account of these structural changes is carried out in the same manner as in the active set algorithm.

TABLE 1: Active set results: housing data, wheat data

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

4 Examples

We consider both the lasso and support vector regression optimization problems applied to two well known data sets, the Iowa wheat data ($p = 9$, $n = 33$), and the Boston housing data ($p = 13$, $n = 506$). For the lasso, for both data sets, the homotopy algorithm started at $\kappa = 0$ turns out to be clearly the method of choice. Here it takes exactly p updating steps of $\mathcal{O}(np)$ operations applied to an appropriately organized data set to compute the solutions for the full range of κ in each case, while just two more steps are 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. It is also very competitive with the cost of the active set lasso algorithm for a single value of κ especially when a significant number of the variables are selected. Thus the active set algorithm is of interest mainly when answering questions for a specific value of κ.

Support vector regression provides an example in which the residual vec-

TABLE 2: Homotopy: Iowa wheat 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 3: Homotopy: Boston housing data

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

tor 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 for both data sets. Results are given in Table 1. Here nits is the number of iterations, n0 the number of residuals at zero level, and ne the number satisfying $|r| = \epsilon$. The homotopy algorithm is relatively less favoured for support vector regression. The obvious starting point for both data sets is $\mathbf{x} = 0$, $\lambda = 0$ in the sense that the solution is known. Tables 2 and 3 show a slow beginning with repeated changes in the active set and little evident structure until λ is increased away from 0 significantly. In the homotopy algorithm applied to the housing data in particular something needs to be done to escape the small values of λ (see Table 3). The active set algorithm could be useful in probing the range of λ to find suitable starting points for the homotopy here.

References

[1] 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. http://www.siam.org/journals/sissc/20-1/30401.html C197, C198

[2] Trevor Hastie, Saharon Rosset, Rob Tibshirani, and Ji Zhu, *The entire regularization path for the support vector machine*, Statistics Department Technical Report, Stanford University, 2003. http://jmlr.csail.mit.edu/papers/volume5/hastie04a/hastie04a.pdf. C199

[3] M. R. Osborne, *Simplicial algorithms for minimizing polyhedral functions*, Cambridge University Press, 2001. C199, C200

[4] 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.
http://imanum.oupjournals.org/cgi/content/abstract/20/3/389. C198, C203

[5] Saharon Rosset, and Ji Zhu, *Piecewise linear regularised solution paths*, Statistics Department Technical Report, Stanford University, 2003.
http://www-stat.stanford/%7Esaharon/papers/piecewise.ps. C199

[6] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. S. S. B* **58** (1996), no. 1, 267–288.
http://citeseer.ist.psu.edu/tibshirani94regression.html. C198

[7] 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.
http://citeseer.ist.psu.edu/vapnik96support.html C199

[8] Ji Zhu, Saharon Rosset,Trevor Hastie, and Rob Tibshirani, 1-norm support vector machines, *Advances in Neural Information Processing Systems* 16, MIT Press, 2004.
http://www-stat.stanford/%7Esaharon/papers/svmL1.ps C199