Trust Region Approaches for Nonlinear Bilevel Programming

Nitish Srivastava
MTH506: Optimization Term Paper
Instructor: Dr Joydeep Dutta

November 4, 2010

Abstract

Bilevel programs form a class of hierarchical optimzation problems in which the constraint set is not explicit but is defined in terms of another optimization problem. Optimization problems in several real-life domains can be expressed as bilevel programs, making such problems ubiquitous. In this term paper, we analyze a trust region approach for solving a special case of this problem.

1 Introduction

Bilevel programming problems are hierarchical optimization problems where the constraints of one problem (the so-called upper level problem or the leader problem) are defined in part by a second parametric optimization problem (the lower level problem or the follower problem). The general form of such problems is

$$\begin{aligned}
(BLPP) \quad \min \quad & F(x,y) && (1.1)\\
\text{s.t.} \quad & G_i(x,y) \leq 0, i = 1,2,\ldots,m_1 && (1.2)\\
& y \in \underset{y'}{\operatorname{argmin}} f(x,y') && (1.3)\\
& \text{s.t. } g_j(x,y') \leq 0 j = 1,2,\ldots,m_2 && (1.4)
\end{aligned}$$

Where all functions are defined from $\mathbb{R}^{n_1+n_2} \to \mathbb{R}$.

Bilevel programs have been studied extensively because many problems such as in economics (Stackelberg-Nash games), engineering problems involving equilibrium constraints, network design problems and classification model selection in machine learning can be formulated as bilevel programs.

Bilevel programming is inherently a hard problem. The simple case when both the upper and lower constraint functions (G_i, g_j's) are linear was shown to be NP-Hard in [3] (Jeroslow, 1985). Later, even the problem of checking local optimality for bilevel programs was show to be NP-hard [5](Vicente, 1994). Therefore, in order to solve it, slightly restricted versions of the problem have been posed. In this paper, we present an application of the trust region method to solving this problem. The algorithm is due to Marcotte, Savard and Zhu [4]

(2001) which is for solving BLPP with a a nonlinear objective function constrained over a polyhedron at the upper level and a linear variational inquality at the lower level. Their method combines the use of the trust region approach with linesearch and guarantees convergence to a strong stationary point. In this paper, we study this method and analyze how trust region method has been effectively used.

The rest of the paper is structured as follows. In section 2 we give the relationship between *Mathematical Programs with Equilibrium Constraints* (MPECs) and bilevel programs, which is used in formulating the problem by Marcotte et. al. [4]. Section 3 describes the proof for optimality of the method and section 4 gives the convergence analysis.

2 Relationship of bilevel programs with MPECs

The most general form of an MPEC is

$$\begin{aligned} \min_{x,y} \quad & F(x,y) \\ s.t. \quad & (x,y) \in Z \text{ and } y \in \mathcal{S}(x), \end{aligned}$$

where $Z \subseteq \mathbb{R}^{n_1+n_2}$ is a non-empty closed set and $\mathcal{S}(x)$ is the solution set of the variational inequality

$$y \in \mathcal{S} \Leftrightarrow y \in C(x) \text{ and } (y-v)^T\psi(x,y) \leq 0 \forall v \in C(x) \tag{2.1}$$

defined over the closed convex set $C(x) \subseteq \mathbb{R}^{n_2}$.

MPECs can be seen as bilevel program where the lower level problem is replaced by a variational inequality. Given a function $\psi : \mathbb{R}^n \to \mathbb{R}^n$ and $C \subseteq \mathbb{R}^n$, convex; $x^* \in C$ is said to be a solution of the variational inequality $\mathrm{VI}(\psi, C)$ if it satisfies

$$(x^* - x)^T\psi(x^*) \leq 0 \forall x \in C$$

One way to understand this equivalence is to see that if $\psi(x,.)$ is such that

$$\psi(x,y) = \nabla_y f(x,y)$$

for some $f : \mathbb{R}^{n_1+n_2} \to \mathbb{R}$. Then the variational inequality becomes

$$(y-v)^T\nabla_y(x,y) \leq 0 \forall x \in C$$

If f is convex, this defines the set of points satisfying the stationary point conditions of the following problem

$$\begin{aligned} \min_y \quad & f(x,y) \\ \text{s.t.} \quad & y \in C(x) \end{aligned}$$

If $C(x) = \{y : g_i(x,y) \leq 0\}$ then the above problem reduces to BLPP. The converse reduction is also true. The above arguments give a flavour of the kind of relation between bilevel programs and MPECs. For a more detailed and rigorous study refer to Colson et al. [1].

3 Trust region formulation for bilevel programs

This section describes the work done by Marcotte et. al. [4]. They formulate the program as

$$\min_{x \in X, y \in Y(x)} \quad f(x, y) \tag{3.1}$$

$$\text{s.t.} \quad \langle F(x, y), y - y' \rangle \leq 0 \forall y' \in Y(x) \tag{3.2}$$

where the mapping F is strongly monotone w.r.t y, X is polyhedral and $Y(x) = \{y : Ax + By \geq b\}$. By the arguments given in section 2, we can see that this problem which is formulated as an MPEC is an instance of a bilevel program.

The trust region method for solving this is based on the idea of making linear approximation of f and F by their first order Taylor series expansions $\bar{f}$ and $\bar{F}$ around each iterate $(\bar{x}_k, \bar{y}_k)$ in the trust region.

$$\min_{x \in X, y \in Y(x)} \quad \bar{f}(x, y) \tag{3.3}$$

$$\text{s.t.} \quad \langle \bar{F}(x, y), y - y' \rangle \leq 0 \forall y' \in Y(x) \tag{3.4}$$

$$||x - \bar{x}_k|| \leq \epsilon_k \tag{3.5}$$

Let the solution to the above problem be $(\bar{x}_k^*, \bar{y}_k^*)$. Let $y(x)$ and $\bar{y}(x)$ be the optimal reaction functions.

$$\rho_k = \frac{f(\bar{x}_k, y(\bar{x}_k)) - f(\bar{x}_k^*, y(\bar{x}_k^*))}{\bar{f}(\bar{x}_k, \bar{y}(\bar{x}_k)) - \bar{f}(\bar{x}_k^*, \bar{y}(\bar{x}_k^*))}$$

The algorithm then does the following
if $\rho_k \geq 2/3$ then $\bar{x}_{k+1}$ is set to $\bar{x}_k^*$ and the trust region size is expanded $(\epsilon_{k+1} = 2\epsilon_k)$.
if $1/3 \leq \rho_k < 2/3$ then $\bar{x}_{k+1}$ is set to $\bar{x}_k^*$ but the trust region size is not changed $(\epsilon_{k+1} = \epsilon_k)$.
else $\bar{x}_{k+1}$ is set to $\bar{x}_k$ and the trust region size is reduced $(\epsilon_{k+1} = \epsilon_k/2)$.

This process is then repeated till convergence.

In each intermediate problem, the lower-level problem can be substituted by its KKT optimality conditions (which are necessary and sufficient here because the problem has been linearized). The intermediate optimization problem for the lower level variational inequality is

$$\begin{aligned}
\min_x \quad & \bar{f}(x, y) \\
\text{s.t} \quad & x \in X \\
& ||x - \bar{x}|| \leq \epsilon_k \\
& Ax + By \geq b \\
& \pi B = \bar{F}(x, y) \\
& \pi \geq 0 \\
& \pi(Ax + By - b) = 0
\end{aligned}$$

This reduces the problem to a mixed integer program which can be solved by branch-and-bound techniques because the constraint set is a polyhedron and some vertex must lie in the solution set.

3.1 Change in the algorithm to guarantee convergence

Using the above algorithm optimality of the limit point of the iterates (x^*) can be proved in the Clarke sense. Marcotte et. al., in their algorithm make a change so that a stronger result can be proved - that the limit of the iterates is a B-Stationary point, i.e., there are no descent directions at the point x^*. To do this we need that the trust region radius ϵ_k must stay bounded away from 0. To ensure this, the following change is made to the last condition of the algorithm (i.e. when $\rho_k \leq 1/3$).

$$
\begin{aligned}
\text{if } \rho_k \leq 1/3 \text{ then} \qquad v_k &\in \{\arg \min_{v \in \{2^j \epsilon_k : 1 \leq j \leq \lceil -\log(\epsilon_k) \rceil\}} f(\bar{x}_v, y(\bar{x}_v))\} \\
\bar{x}_{k+1} &\in \ \text{argmin} \ \{f(\bar{x}_{v_k}, y(\bar{x}_{v_k})), f(\bar{x}_k, y(\bar{x}_k))\} \\
\epsilon_{k+1} &= \epsilon_k
\end{aligned}
$$

To understand the intuition behind this modification, first note that if x_k is not a stationary point, then there must exist a trust region radius ϵ such that $\bar{x}_k$ improves on x_k, even though the model function is not a good approximation of the actual objective function. Therefore by doubling the trust region radius a few times (fintely many), we can move away from teh non-stationary points.

This modification allows us to prove a stronger convergence and optimality result. Since the proof is quite involved, we refer the reader to Marcotte et. al. [4] for explicit details, instead of reproducing the proofs here. However, in order to get an understanding of the main ideas used, we give a brief sketch. Firstly, we lay down the key assumptions which allow the proof to hold. These assumptions are not too harsh, just basic smoothness and Lipschitz continuity properties, except the last one which is a strong condition and limits the application of this technique to larger class of problems.

- The set X is compact
- f and ∇f are Lispchitz continuous on $Z = \{(x, y) : x \in X, y \in Y(x)\}$
- F and Jocobian of F are Lipschitz continuous on Z.
- F is uniformly strongly monotone in y on Z.

Firstly, it can be shown that $y(x)$ and $\bar{y}(x)$ are Lipschitz continuous where they are defined and are also *directionally differentiable.* The proof uses a result by Dafermos [2] which states that the reaction functions will be Lipschitz continuous if the projection operator $p_z(x) = \text{proj}_{Y(x)}(z)$ is Lipschitz continuous, for fixed z. Once we have these two properties, we use them to show that the difference between the global minimum within the trust region and the true minimum at any iteration k is of the order of ϵ_k^2, i.e., (omitting the subscript k to increase readability)

$$\bar{f}(\bar{x}^*, \bar{y}^*) - f(x^*, y(x^*)) \leq \mu\epsilon^2$$

where μ is a constant depending only on the Lipschitz constants of the different functions involved. Recall that $(x^*, y(x^*))$ denotes the true minimzer of the bilevel program in the trust region and $(\bar{x}^*, \bar{y}^*)$ is the minimzer of the model problem over the trust region.

The proof by Marcotte et. al. [4] goes on to say that the difference between the best value of the cost function ($\tilde{f}(x) = f(x, y(x))$) within the trust region and the actual value given by the model is of the order ϵ^2. This allows them to prove that
If $\tilde{f}$ admits a feasible descent direction at $\tilde{x}$, then there exists an ϵ such that $\tilde{f}(\bar{x}_{\bar{\epsilon}}) < \tilde{f}(\tilde{x})$ whenever $||\tilde{x} - \bar{x}|| \le \epsilon/2$ and $\bar{\epsilon} \in [\epsilon, 2\epsilon]$.

4 Proof of convergence

In this section we prove that the trust region method converges to a *strongly stationary* point. The proof is abstracted from the work of Marcotte et al. [4]. A point x is said to be strongly stationary for a function g if there do not exist $\alpha > 0$ and a sequence $\{x^i\}_{i\in I}$ converging to x such that

$$g(x^i) < g(x) - \alpha||x - x^i||$$

In general the set of strong stationary points could be a strict subset of points that satisfy the stationary point conditions w.r.t the Clarke subdifferential, but in this case it has been shown that g is directionally differentiable and hence the two sets are same. If the trust region algorithm stops in fintely many steps, then the model agrees with the true objective function which can happen only at a B-stationary point. In order to prove convergence for the infinte case (${x^k}_{k\in N}$), we show that any accumulation point is strongly stationary. Consider a convergent subsequence $\{x_{k_i}\}_{i\in I}$. There are two cases.

- Case 1: $\{\epsilon_{k_i}\}_{i\in I}$ has a non-zero lower bound.
  Let $y_i(x)$ be the solution of the lower level linearized variational inequalitycorresponding to the upper level vector x^i and let $f_i(x) = f(x, y_i(x))$. Since $\{\epsilon_{k_i}\}$ does not converge to zero, $\rho_k \ge 1/3$ must happen infinitely many times as $k \to \infty$, because the value of ϵ would be halved each time $\rho \le 1/3$ and for the sequence ϵ_{k_i} to not converge this halving cannot happen infintely unless compensated by doubling infinitely. That is,

  $$\tilde{f}(x^{k_i}) - \tilde{f}(x^{k_i+1}) \ge \frac{1}{3}\left(\bar{f}(x^{k_i}) - \bar{f}(\bar{x}^{k_i})\right)$$

  If $x*$ is not B-stationary for $\tilde{f}$, then it cannot be B-stationary for $f_*(x) = f(x, y_*(x))$ where $y_*(x)$ satisfies

  $$\langle F(x^*, y(x^*)) + F'_x(x^*, y(x^*))(x - x^*) + F'_y(x^*, y(x^*))(y_*(x) - y(x^*)), y_*(x) - y\rangle \le 0 \forall y \in Y(x)$$

  Therefore for any $\epsilon \in (0, \mu]$, $\exists \hat{x}$ such that $||\hat{x} - x^*|| \le \epsilon$ and

  $$f(\hat{x}, y_*(\hat{x})) < f(x^*, y(x^*))$$

  Using results proved while developing the optimality conditions, we have that $||y_*(x) - y_{k_i}(x)|| \to 0$ and $f(\hat{x}, y_{k_i}(\hat{x})) \to f(\hat{x}, y_*(\hat{x}))$.

  $$\begin{aligned} f_{k_i}(\bar{x}^{k_i}) &\le f_{k_l}(\hat{x}) \\ f_{k_i}(x^{k_i}) - f_{k_i}(\bar{x}^{k_i}) &\ge f_{k_i}(x^{k_i}) - f_{k_l}(\hat{x}) \\ &\ge \frac{1}{2}\left(f(x^*, y(x^*)) - f(\hat{x}, y_*(\hat{x}))\right) \end{aligned}$$

Hence the series $\sum \left(\tilde{f}(x^{k_i}) - \tilde{f}(x_{k_i+1})\right)$ should diverge. This gives us the contradiction because the algorithm generates a decraesing sequence, that is

$$\sum_{i=1}^{\infty} \left(\tilde{f}(x^{k_i}) - \tilde{f}(x^{k_i+1})\right) \leq \sum_{i=1}^{\infty} \left(\tilde{f}(x^{k_i}) - \tilde{f}(x^{k_{i+1}})\right) = \tilde{f}(x^{k_1}) - \tilde{f}(x^*) < \infty$$

Hence the assumption that x^* is not B-stationary is wrong.

- Case 2: $\lim_{i\to\infty} \epsilon_{k_i} = 0$
  Proof is again by contardiction. Let x^* not be a B=stationary point. Choose ϵ as in the last optimality result proof.Let i be an index such that $||x^{k_i} - x^*|| \leq \epsilon/2$, j be such that it satisfies $\epsilon \leq 2^j \epsilon_{k_i} = \epsilon' \leq 2\epsilon$. Then we have that $\tilde{f}(\bar{x}_{\epsilon'} \leq f(x^*)$ which is a contradiction of the fact that x^* is a stationary point of the sequence of iterates.

5 Conclusion

In this paper, we give a trust region algorithm for solving bilevel programs expressed as an MPEC. The proof of convergence of this algorithm guarantees that the sequence of iterates converges to a strongly stationary point. The only drawback of the algorithm is that it involves an expensive linesearch in each iteration where $\rho_k \leq 1/3$ but the authors of the algorithm claim that such iterations will not be too many and do not adversely affect the performance of the algorithm.

References

[1] Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of Operations Research*, 153:235–256, 2007. 10.1007/s10479-007-0176-2.

[2] Stella Dafermos. Sensitivity analysis in variational inequalities. *Mathematics of Operations Research*, 13(3):pp. 421–434, 1988.

[3] Robert G. Jeroslow. The polynomial hierarchy and a simple model for competitive analysis. *Mathematical Programming*, 32:146–164, 1985. 10.1007/BF01586088.

[4] P. Marcotte, G. Savard, and D. L. Zhu. A trust region algorithm for nonlinear bilevel programming. *Operations Research Letters*, 29(4):171 – 179, 2001.

[5] L. Vicente, G. Savard, and J. Júdice. Descent approaches for quadratic bilevel programming. *J. Optim. Theory Appl.*, 81:379–399, May 1994.