Bilevel programming programs: introduction, reformulation and partial calmness

Jane J. Ye

University of Victoria, Canada

Lecture 2 at the Forum on Developments and Origins of Operations Research

November 25, 2021

Organizers: The Mathematical Programming Branch of OR Society of China

& Southern University of Science and Technology

Outline

- Introduction to bilevel programs (background and applications)
- Reformulations as single level optimization problems
- Partial calmness

Bilevel programs

$$(BP) \quad \min_{x,y} \quad F(x,y)$$
$$s.t. \quad y \in S(x)$$

where $S(x)$ denotes the set of solutions of the lower level problem:

$$(P_x) \quad \min_{y \in Y(x)} \quad f(x,y).$$

where $Y(x) := \{y | g(x,y) \leq 0\}$.

- In this talk we assume all functions F, f and g are smooth.

Bilevel Programs

Suppose that for each x, the lower level problem (P_x) has a unique solution $y(x)$. Then by substituting $y(x)$ into the upper level, the bilevel program becomes an one-level optimization problem

$$\min_x F(x, y(x)).$$

If $y(x)$ is a "nice" function of x, then perhaps the above problem can be solved.
But if the lower level problem has multiple solutions, then there are two versions of the bilevel program: optimistic and pessimistic.

- Optimistic: $\min_{x,y}\{F(x,y) : y \in S(x)\}$.
- Pessimistic: $\min_x \max_{y \in S(x)} F(x,y)$.

In this talk we only deal with the optimistic case.

Applications in economics

- The first formulation of a simplier case of the bilevel program was introduced by Stackelberg (1934). Hence it is known as a Stackelberg game in economic game theory.
- The classical principal-agent/moral hazard problem in economics is a bilevel program: This is the situation where the principal can only observe the outcome of the agent's action but not the action itself. How can the principal design a contract in order to maximize the expected utility subject to the optimizing behavior of the agent?
- Nobel prize has been awarded twice for study of the moral hazard problem. Vickrey and Mirrlees shared the 1996 Nobel prize in economics which was awarded for their fundamental contributions to the economic theory of incentives under asymmetric information. Holmström and Hart shared the 2016 Nobel prize in economics which was awarded for their fundamental contributions to contract theory.

Applications in machine learning

- The bilevel program was first introduced to the optimization community by Bracken and McGill (1973).
- It was first introduced to the model selection in machine learning by Bennett, Hu, Ji, Kunapuli and Pang in 2006.
- Recently there are more and more work on hyper-parameter learning via bilevel optimization:

$$\min_{\theta,\lambda} F(\theta)$$

$$\text{s.t. } \theta \in \arg \underbrace{\min_{\theta'} f(\theta') + \sum_{i=1}^{r} \lambda_i P_i(\theta')}_{\text{lower level training problem}},$$

where $P_i(\theta)$ are penalty functions.

Bilevel programs where the lower level is unconstrained

Bilevel program: lower level unconstrained case

$$(BP) \quad \min F(x,y) \quad s.t. \quad y \text{ solves } \min_{y'} f(x,y').$$

If f is convex and smooth in y, then (BP) is equivalent to the single-level problem:

$$\begin{aligned} (SP): \quad & \min \quad F(x,y) \\ & s.t. \quad 0 = \nabla_y f(x,y) \end{aligned}$$

If around point $(\bar{x},\bar{y})$, the solution of the lower level $S(x) = \{y(x)\}$ is a singleton, then locally around $(\bar{x},\bar{y})$, (BP) is equivalent to $\min_x F(x,y(x))$. If one can either solve $y(x)$ explicitly or have an expression for $\nabla y(x)$, then one may solve the problem (BP) numerically. But these assumptions are usually not satisfied.

The first order approach for nonconvex lower level case

- Consider the case where f is not convex in

$$BP \qquad \min F(x,y) \quad s.t. \quad y \text{ solves } \min_{y'} f(x,y').$$

The first order approach replaces the lower level problem by its first order condition:

$$SP: \qquad \begin{array}{ll} \min & F(x,y) \\ s.t. & 0 = \nabla_y f(x,y) \end{array}$$

- Question: Must an optimal solution of BP always be a stationary point of SP?
- If yes, we may use SP to find an candidate for the optimal solution of BP. If not, then we may not be able to find a solution of BP by solving SP.
- The answer is no!

Mirrlees' example

$$\text{(BP)} \qquad \begin{array}{ll} \min & F(x,y) := (x-2)^2 + (y-1)^2 \\ s.t. & y \text{ solves } \min\limits_y f(x,y) := -xe^{-(y+1)^2} - e^{-(y-1)^2}. \end{array}$$

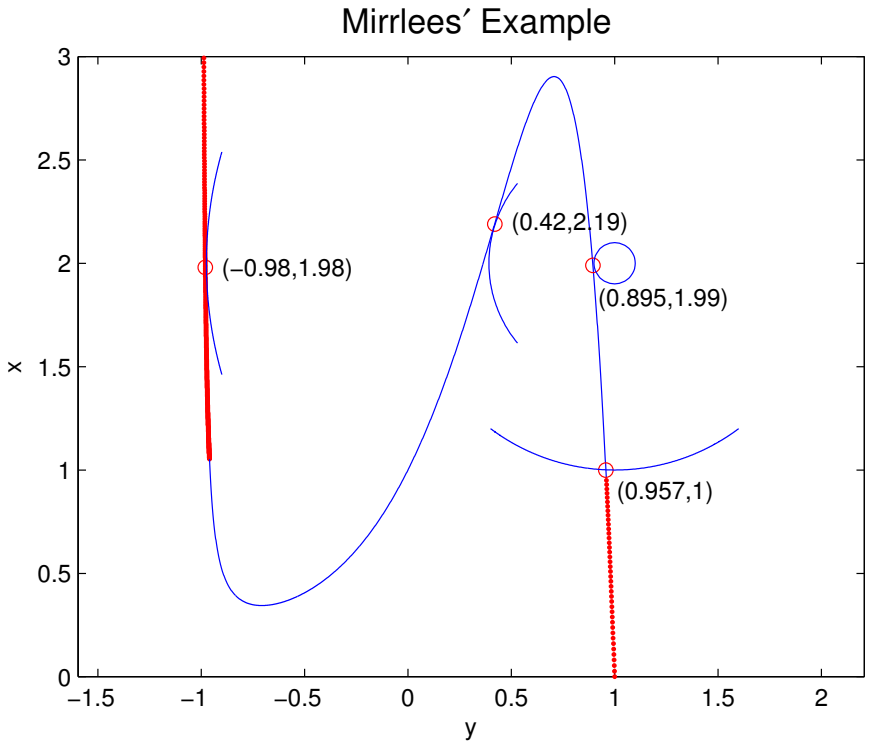
The first order condition for minimization of the lower level objective function with respect to y is

$$x(y+1)e^{-(y+1)^2} + (y-1)e^{-(y-1)^2} = 0.$$

Hence, each x and its stationary point of the lower level problem are related by the equation

$$x = \frac{1-y}{1+y} e^{4y},$$

which is a smooth and connected curve.



- $(1, 0.957)$ is the unique optimizer for (BP) while the three stationary points are $(1.99, 0.895), (2.19, 0.42), (1.98, -0.98)$.
- The optimal solution of (BP) is not a stationary point of (SP)!

Implicit function reformulation

- Implicit function reformulation (Outrata 1990, Dempe 1992): If $S(x) = \{y(x)\}$ locally around $(\bar{x}, \bar{y})$, then locally

$$\begin{array}{ll} \min\limits_{x,y} & F(x,y) \\ s.t. & y \in S(x) \end{array} \iff \min_{x} \quad F(x, y(x))$$

- To derive a KKT condition or a numerical algorithm, the solution map $y(x)$ needs to be at least Lipschitz continuous. Hence very strong conditions such as LICQ and the strong second order sufficient condition for the lower level problem is required.
- Taking derivative respect to x and applying the chain rule to $0 = \nabla_y f(x,y)$, we have

$$0 = \nabla^2_{xy} f(x,y) + \nabla^2_{yy} f(x,y) \nabla y(x).$$

So $\nabla y(x) = -\nabla^2_{yy} f(x,y)^{-1} \nabla^2_{xy} f(x,y)$ if $\nabla^2_{yy} f(x,y)$ is invertible.

The value function approach: lower level unconstrained case

- Define the value function $V(x) := \inf_y \{f(x,y)\}$.
- The following single level problem is equivalent to (BP):

$$(VP) \qquad \min_{x,y} \quad F(x,y)$$
$$s.t. \quad f(x,y) - V(x) \leq 0.$$

- Difficulty 1: The value function $V(x)$ is an implicitly defined function. In general it is nonsmooth.
- Difficulty 2: Even if all defining functions are smooth, (VP) is still a nonsmooth optimization problem. Even if all functions including the value function are Lipschitz continuous, the usual constraint qualification such as the nonsmooth MFCQ never hold (JY and Zhu 1995).

Sensitivity analysis of the value function: unconstrained case

- Suppose the lower level problem is unconstrained:

$$(P_x): \qquad \min_y f(x,y)$$

and f is convex in y.

- The first order condition $0 = \nabla_y f(x,y)$ is necessary and sufficient for $y \in S(x)$.
- Let $\bar{y} \in S(\bar{x})$. If the Hessian matrix $\nabla^2_{yy} f(x,y)$ is nonsingular at $(\bar{x}, \bar{y})$, then by the classical implicit function theorem, the solution mapping $S(x) = \{y(x)\}$ is single-valued around $\bar{x}$ and $y(x)$ is a C^1 function around $\bar{x}$. Hence

$$V(x) = f(x, y(x)) \quad \text{for all } x \text{ around } \bar{x}.$$

- So by the chain rule,

$$\nabla V(\bar{x}) = \nabla_x f(\bar{x}, \bar{y})) + \underbrace{\nabla_y f(\bar{x}, \bar{y})}_{=0} \nabla y(\bar{x}) = \nabla_x f(\bar{x}, \bar{y}).$$

Sensitivity analysis of the value function: Danskin's Theorem

Suppose the lower level problem is:

$$(P_x): \qquad \min_{y \in Y} f(x, y),$$

where Y is a compact set. Then by Danskin's Theorem, the value function is Lipschitz continuous and the Clarke subdifferential of the value function is

$$\partial^c V(x) = \mathrm{co}\{\nabla_x f(x, y) : y \in S(x)\}.$$

The nonsmooth MFCQ for (VP) fails at each point of the feasible region!

$$(VP) \qquad \min_{x,y} \quad F(x,y)$$
$$s.t. \quad f(x,y) - V(x) \leq 0$$

For any feasible solution $(\bar{x}, \bar{y})$, one always have $f(\bar{x}, \bar{y}) - V(\bar{x}) = 0$. In this case the nonsmooth LICQ is the same as the nonsmooth MFCQ and is equivalent to

$$0 \notin \nabla f(\bar{x}, \bar{y}) - \partial^c V(\bar{x}) \times \{0\}.$$

But a feasible solution $(\bar{x}, \bar{y})$ of (VP) must be a solution to the following problem:

$$\min_{x,y} \{f(x,y) - V(x)\}.$$

By the optimality condition, we must have

$$0 \in \nabla f(\bar{x}, \bar{y}) - \partial^c V(\bar{x}) \times \{0\}.$$

The value function approach may fail!

- Why?

$$\begin{array}{ll} \min F(x,y) & \Longleftrightarrow \min \quad F(x,y) \\ \text{s.t. } y \in \arg\min_y f(x,y) & \quad s.t. \quad f(x,y) - V(x) \leq 0. \end{array}$$

- If $V(x)$ is Lipschitz continuous at $\bar{x}$, then KKT condition is

$$\begin{aligned} 0 &\in \nabla_x F(\bar{x},\bar{y}) + \mu(\nabla_x f(\bar{x},\bar{y}) - \partial^c V(\bar{x})) \\ 0 &= \nabla_y F(\bar{x},\bar{y}) + \mu \underbrace{\nabla_y f(\bar{x},\bar{y})}_{=0} \end{aligned}$$

  which is true only if $\nabla_y F(\bar{x},\bar{y}) = 0$.

- It does not hold for Mirrlees' problem since $F(x,y) = (x-2)^2 + (y-1)^2, F_y(x,y) = 2(y-1), \bar{y} = 0.957$!

The combined approach (JY and Zhu, 2010) for the lower level unconstrained case

- $$(CP): \quad \min \quad F(x,y)$$
$$s.t. \quad f(x,y) - V(x) \leq 0$$
$$0 = \nabla_y f(x,y)$$

The KKT system holds for (CP) if there are $\mu \geq 0, \beta$ s.t.

$$\begin{aligned} 0 &\in \nabla_x F(\bar{x},\bar{y}) + \mu(\nabla_x f(\bar{x},\bar{y}) - \partial^c V(\bar{x})) \\ &\quad + \nabla^2_{xy} f(\bar{x},\bar{y})^T \beta, \\ 0 &= \nabla_y F(\bar{x},\bar{y}) + \underbrace{\nabla^2_{yy} f(\bar{x},\bar{y})^T}_{\geq 0} \beta, \end{aligned}$$

- The optimal solution for Mirrlees' problems is a KKT point of (CP) with $\mu = 2.05, \beta = 0.04918$.

The combined approach with second order condition (Ma, Yao, JY and Zhu, 2021)

Adding the necessary optimality condition of the lower level problem, (BP) is obviously equivalent to the combined program with the second order condition:

$$
\begin{array}{lll}
(CPSOC) & \min & F(x,y) \\
& s.t. & f(x,y) - V(x) \leq 0, \\
& & 0 = \nabla_y f(x,y), \\
& & \nabla^2_{yy} f(x,y) \in \mathbb{S}^m_+,
\end{array}
$$

where $\mathbb{S}^m_+$ is the positive semi-definite matrix cone. It is obvious that

$$\text{KKT for } (VP) \Longrightarrow \text{ KKT for } (CP) \Longrightarrow \text{ KKT for } (CPSOC).$$

Bilevel programs where the lower level is constrained

The MPEC approach for bilevel program with an inequality constrained lower level problem

- For each x, if the lower level problem

$$(P_x) : \min_y f(x,y) \text{ s.t. } g(x,y) \leq 0$$

is convex and the Slater condition holds, then it is necessary and sufficient for the KKT condition

$$\exists \lambda \geq 0, 0 = \nabla_y f(x,y) + \nabla_y g(x,y)^T \lambda, 0 \leq \lambda \perp -g(x,y) \geq 0$$

to hold at a solution $y \in S(x)$.

- It is popular to solve the mathematical program with equilibrium/complementarity constraints (MPEC/MPCC):

$$\begin{aligned} (MPCC) \qquad \min_{x,y,\lambda} \quad & F(x,y) \\ s.t. \quad & 0 = \nabla_y f(x,y) + \nabla_y g(x,y)^T \lambda, \\ & 0 \leq \lambda \perp -g(x,y) \geq 0. \end{aligned}$$

instead.

Three issues arised using this approach:

- Issue (a): The approach is only applicable if the lower leve problem are convex: counter example: Mirrlees' example.
- Issue (b): The condition

$$0 \leq \lambda \perp -g(x,y) \geq 0$$

is a complementarity constraint. If we treat it as equality and inequality constraints

$$\lambda \geq 0, g(x,y) \leq 0, \langle \lambda, g(x,y) \rangle = 0,$$

then the usual constraint qualification such as Mangasarian Fromovitz constraint qualification (MFCQ) never hold.

- Issue (c): If the lower level has multiple multipliers, then a local solution of MPCC may not recover a local solution of the original bilevel program (Dempe and Dutta (2012)).

The value function approach

- The following single level problem is equivalent to (BP):

$$(VP) \qquad \min_{x,y} \quad F(x,y)$$
$$s.t. \quad f(x,y) - V(x) \leq 0,$$
$$g(x,y) \leq 0,$$

where $V(x) := \inf_y \{f(x,y) : g(x,y) \leq 0\}$ is the value function.

- Difficulty 1: The value function $V(x)$ is an implicitly defined function. In general it is nonsmooth.
- Difficulty 2: Even if all defining functions are smooth, (VP) is still a nonsmooth optimization problem. Even if all functions including the value function are Lipschitz continuous, the usual constraint qualification such as the nonsmooth MFCQ never hold (JY and Zhu 1995).

Sensitivity analysis of value functions: constrained with unique solution and unique multiplier case

- If $S(x) = \{y(x)\}$ and $y(x)$ is C^1, then by the chain rule,

$$\nabla V(x) = \nabla_x f(x, y(x)) + \nabla_y f(x, y(x)) \nabla y(x).$$

- Suppose KKT condition holds and the multiplier is unique and is a smooth function $\lambda(x)$. Then by differentiating the complementary slackness condition, $g(x, y)^T \lambda = 0$ we can get

$$0 = \nabla_x g(x, y(x))^T \lambda(x) + \underbrace{\nabla_y g(x, y(x))^T \lambda(x)}_{=-\nabla_y f(x, y(x))} \nabla y(x).$$

- Hence

$$\nabla V(x) = \nabla_x f(x, y(x)) + \nabla_x g(x, y(x))^T \lambda(x).$$

Sensitivity analysis of value functions: constrained case

- Let $\mathrm{KT}(x, y)$ denotes the set of KKT multipliers for problem (P_x) at y.
- By Gauvin (1979), if MFCQ holds at each $y \in S(x)$ and the feasible region is uniformly bounded, then the value function is Lipschitz continuous at x and

$$\partial^c V(x) \subseteq co \bigcup_{y \in S(x), \lambda \in \mathrm{KT}(x,y)} \{\nabla_x f(x, y) + \nabla_x g(x, y)^T \lambda\},$$

  where $\mathrm{KT}(x, y)$ denotes the set of KKT multipliers at $y \in S(x)$.
- Shaper estimates under weaker assumptions are given in Guo, Lin, JY and Zhang (2014).

NNAMCQ for (VP) fails at each point of the feasible region!

Any feasible solution $(\bar{x}, \bar{y})$ of (VP) must be a solution to the following problem:

$$\min_{x,y} f(x,y) - V(x) \quad \text{s.t.} g(x,y) \leq 0.$$

By the optimality condition, we must have $\eta \geq 0$ such that

$$0 \in \partial(f - V)(\bar{x}, \bar{y}) + \nabla g(\bar{x}, \bar{y})^T \eta, \quad g(\bar{x}, \bar{y})^T \eta = 0.$$

This means that $(1, \eta)$ is a nonzero abnormal multiplier for (VP). So NNAMCQ for (VP) fails.

The combined approach

- Combined program with KKT condition (JY-Zhu (2010)):

$$(CP) \qquad \min_{x,y,u} F(x,y)$$
$$\text{s.t. } f(x,y) - V(x) \le 0,$$
$$\nabla_y f(x,y) + u\nabla_y g(x,y) = 0,$$
$$g(x,y) \le 0,\ u \ge 0,\ u^T g(x,y) = 0.$$

- Combined program with Fritz John condition or Bouligand (B)-condition (Ke, Yao, JY and Zhang, 2021) or the second order condition (Ma, Yao, JY and Zhang, 2021).
- By B-condition, we mean $0 \in \nabla_y f(x,y) + \widehat{\mathcal{N}}_{Y(x)}(y)$.
- It is easier for the resulting KKT condition to hold than the one based on the value function approach.

Since the value function constraint $f(x,y) - V(x) \leq 0$ is actually an equality constraint (the strict inequality $f(x,y) - V(x) < 0$ never hold),

NNAMCQ/nonsmooth MFCQ fails for (VP) and (CP)!

Partial calmness condition

- Definition (JY and Zhu (1995)): (BP) is said to be partially calm at $(\bar{x}, \bar{y})$ if $(\bar{x}, \bar{y})$ is a local solution to the partially penalized problem for some $\mu \geq 0$:

$$\begin{aligned} \min_{x,y} \quad & F(x,y) + \mu(f(x,y) - V(x)) \\ s.t. \quad & g(x,y) \leq 0. \end{aligned}$$

- JY and Zhu (1995) proved that if the lower level objective function and the constraints are jointly linear then BP is partially calm.
- Some sufficient conditions for partial calmness have been derived such as uniform weak sharp minimum, uniform parametric error bounds (JY, Zhu and Zhu 1997 and JY 1998), Directional quasi-/pseudo-normality (Bai, JY and Zhang 2019), Relaxed constant positive linear dependence constraint qualification (Xu and JY 2020).

Partial calmness condition

Basic features of the partial calmness condition:

- the partial calmness condition allows one to partially penalize the value function constraint $f(x,y) - V(x) \leq 0$ to the objective function. Consequently, the usual constraint qualifications can be applied to the rest of the constraints.
- proposed for the value function reformulation (JY-Zhu (1995)), the combined program with KKT condition (JY-Zhu (2010)), and the combined program with FJ condition and B-condition (Ke-Yao-JY-Zhang (2021)) and the combined program with second order condition (Ma, JY, Yao and Zhang (2021)).

How stringent is the partial calmness condition?

- Recently in Ke-Yao-JY-Zhang (2021), we have shown that at least for the case where x is one-dimensional, the partial calmness for the combined program is a generic condition while the one for the value function reformulation is not.

Semi-infinite programming reformulation

$$y \in S(x) \Longleftrightarrow g(x,y) \leq 0 \text{ and } f(x,z) - f(x,y) \geq 0 \quad \forall z \in Y(x)$$

- When all functions are polynomials and KKT condition holds at each $y \in S(x)$, we can find a multiplier of the lower level problem as a polynomial or rational function of (x,y), denoted by $\lambda(x,y)$.
- The bilevel program is equivalent to the generalized SIP:

$$\begin{aligned}(SIP) \qquad \min_{x,y} \; & F(x,y) \\ \text{s.t. } & f(x,z) - f(x,y) \geq 0 \quad \forall z \in Y(x), \\ & \nabla_y f(x,y) + \lambda(x,y)\nabla_y g(x,y) = 0, \\ & g(x,y) \leq 0, \; \lambda(x,y) \geq 0, \; \lambda(x,y)^T g(x,y) = 0.\end{aligned}$$

- Based on this reformulation recently we have proposed a numerical algorithm to globally solve the polynomial bilevel program in Nie, Wang, JY and Zhong, (2021) .

Reference

- JY and D.L. Zhu 1995, *Optimality conditions for bilevel programming problems*, Optimization **33**, 9-27.
- JY, D.L. Zhu and Q.J. Zhu 1997, *Exact penalization and necessary optimality conditions for generalized bilevel programming problems*, SIOPT **7**, 481-507.
- JY 1998, *New uniform parametric error bounds*, JOTA **98**, 197-219.
- JY and D.L. Zhu 2010, *New necessary optimality conditions for bilevel programs by combining MPEC and the value function approach*, SIOPT **20**, 1885-1905.
- L. Guo, G-H. Lin, JY and J. Zhang 2014, *Sensitivity analysis of the value function for parametric mathematical programs with equilibrium constraints*, SIOPT **24**, 1206-1237.

Reference

- R. Ke, W. Yao, JY and J. Zhang 2021, *Generic property of the partial calmness condition for bilevel programming problems*, to appear in SIOPT, arXiv (2017.14469).
- J. Nie, L. Wang, JY and S. Zhong 2021, *A Lagrange multiplier expression method for bilevel polynomial optimization*, SIOPT **31**, 2368-2395.
- X. Ma, W. Yao, JY and J. Zhang 2021, *Combined approach with second-order optimality conditions for bilevel programming problems*, arXiv (2018.00179).
- K. Bai, JY and J. Zhang 2019, *Directional quasi-/pseudo-normality as sufficient conditions for metric subregularity,* SIOPT **29**, 2625-2649.
- M. Xu and JY 2020, *Relaxed constant positive linear dependence constraint qualification and its application to bilevel programs*, JGO **78**, 181-205.

- Thank You -