Bilevel Mathematical Programming and Machine Learning

Kristin P. Bennett

includes joint work with Jing Hu, Gautam Kunapuli and Jong-Shi Pang

Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, NY
Outline

- Intro to bilevel/multilevel programming
  - Example
  - General formulation
  - Ubiquitous Bilevel Machine Learning Problem
- Properties of Bilevel Programs
- Cross-Validation and Bilevel Programming
How to set the tax rate?

Government wants to set tax rate to maximize revenue.

\[ x = \text{taxable activity} \]
\[ t = \text{tax rate on activity} \]
\[ t \in T \]
\[ F(t, x) = \text{amount of taxes received} \]
How to set the tax rate?

Business wants maximize profit for given tax rate.

\[
x = \text{taxable activity}
\]

\[
x \in X
\]

\[
t = \text{given tax rate}
\]

\[
f(t, x) = \text{profit}
\]
Assumptions — Stackelberg Game (1952)

Government (Leader)

Business (Follower)

Rational: will optimize $f(t,x)$
System Control Problem

Controller

Control Variables $t$

Machine

State Variables $x$

Leader

Follower
Tax Policy as Bilevel Program

$$\max_{t \in T, x} F(t, x)$$

$$s.t. \quad x \in \arg \max_{x' \in X} f(t, x')$$

Two levels of optimization: constraints are themselves mathematical programs.
Tax Policy with regulation

$$\max_{t \in T, x} F(t, x)$$

s.t. \( x \in \arg \max_{(t, x) \in S} f(t, x') \)

Lower level constraints can depend on \( t \) too
General Bilevel Program
Bracken and McGill 1973, Bard 1999

Outer Level
\[ \min_{x, y \in Y} \quad F(x, y) \]
\[ \text{s.t.} \quad G(x, y) \leq 0 \]

Inner Level
\[ \min_{x \in X} \quad f(x, y) \]
\[ \text{s.t.} \quad g(x, y) \leq 0 \]
Machine Learning Example

Given

- Training Set
- Testing Set
- Linear Support Vector Regression Problem with parameters $c, \epsilon$

Determine $c, \epsilon$ such that generalization as measured by the test set is optimized
Grid Search Approach

Define grid over \( C, \varepsilon \)

Optimize model on train for each \( C, \varepsilon \)

Select value of \( C, \varepsilon \) that yields best testing set error
Train/Test as Bilevel

Leader

\[ c, \varepsilon \]

\[ W \]

Follower

W must be optimal for SVR for given \( c, \varepsilon \)
Bilevel Model

Leader: optimize mean absolute testing error by controlling $c, \varepsilon$

Follower: optimizes $w$ using SVR on training data

\[
\min_{c, \varepsilon, w} \sum_{i \in \text{test}} | x_i \cdot w - y_i |
\]

\[
s.t. \quad \min_w \quad C \sum_{i \in \text{train}} \max(| x_i \cdot w - y_i | - \varepsilon, 0) + \frac{1}{2} \| w \|^2
\]
Train/Test Linear Bilevel Model

\[ \min_{c, \varepsilon, z, w} \sum_{i \in \text{test}} z_i \]

s.t. \[ -z_i \leq x_i \cdot w - y \leq z_i \quad i \in \text{test} \]

\[ \min_w C \sum_{i \in \text{train}} (\xi_i + \xi_i^*) + \frac{1}{2} \| w \|^2 \]

s.t. \[ x_i \cdot w - y_i \leq \varepsilon + \xi_i \quad i \in \text{train} \]

\[ -x_i \cdot w + y_i \leq \varepsilon + \xi_i^* \quad i \in \text{train} \]

s.t. \[ \xi_i, \xi_i^* \geq 0 \]
Outline

- Intro to bilevel/multilevel programming
- Properties of Bilevel Programs
  - 2 level LP example
  - MPEC reformulation
- Cross-Validation and Bilevel Programming
2-Level LP

\[
\begin{align*}
\max_{t,x} & \quad c_t t + c_x x \\
\text{s.t.} & \quad x \in \arg \max_{(t,x') \in S} d_t t + d_x x
\end{align*}
\]

Leader would like this feasible solution to be the outcome.

But then follower would respond with this \( x \).
The leader can examine reaction of follower for each feasible choice of $t$. This forms the reaction set $S(t)$.
Optimal Solution

- Equivalent problem
  \[
  \max_{t, x} \quad c_t t + c_x x \\
  \text{s.t.} \quad (t, x) \in S(t)
  \]

- S(t) nonconvex, nonsmooth
- S(t) may not be connected.
- Even LP case is NP-Hard
- Optimality conditions difficult to define
KKT Optimality Conditions of Inner Problem

**Inner**

\[
\min_x f(x, y) \\
\text{s.t.} \quad g(x, y) \leq 0
\]

\[
\Leftrightarrow \quad g(x, y) \leq 0
\]

**KKT**

\[
\nabla_x f(x, y) + \lambda' \nabla_x f(x, y) = 0
\]

\[
\lambda \geq 0
\]

\[
\lambda \perp g(x, y)
\]

\[
i.e. \quad \lambda_i g_i(x, y) = 0 \quad i = 1, \ldots, m
\]

If convex problem and \((x^*, y^*, \lambda^*)\) is KKT point, then \(x^*, y^*\) is globally optimal.
KKT Optimality Conditions of Inner Problem

Inner

\[
\min_x f(x, y) \\
\text{s.t.} \quad g(x, y) \leq 0
\]

\[\iff\]

\[g(x, y) \leq 0 \quad \text{primal feasibility}\]

\[\nabla_x f(x, y) + \lambda' \nabla_x f(x, y) = 0 \quad \text{dual feasibility}\]

KKT

\[\lambda \geq 0\]

\[\lambda \perp g(x, y) \quad \text{complementarity}\]

\[i.e. \quad \lambda_i g_i(x, y) = 0\]

If \((x^*, y^*)\) is optimal and \textbf{Constraint Qualification} satisfied, then KKT point \((x^*, y^*, \lambda^*)\) exists.
Bilevel Optimality Conditions

- CQ usually not satisfied.
- Frequently no KKT points.
- Non-smoothness is the problem.
- Problem is inherently combinatorial.
Key Transformation

- KKT for the inner level training problems are necessary and sufficient
- Replace lower level problems by their KKT Conditions
- Problem becomes a Mathematical Programming Problem with Equilibrium Constraints (MPEC)
Bilevel Program as MPEC

Outer Level
\[ \min_{x,y \in Y} \quad F(x, y) \]
\[ \text{s.t.} \quad G(x, y) \leq 0 \]

Inner Level
\[ \lambda \geq 0 \perp g(x, y) \leq 0 \]
\[ \nabla_x f(x, y) + \lambda' \nabla_x f(x, y) = 0 \]

KKT
Combinatorial Global Search

- For each of m equilibrium constraints
  \[ \lambda_i g_i(x, y) = 0 \iff \begin{cases} \lambda_i = 0 \\ \text{or} \\ g_i(x, y) = 0 \end{cases} \]

- Find global solutions by trying \(2^m\) possibilities.
- For LPs and convex QPs, subproblems are LPs.
- Use Integer Programming/Global Optimization techniques to dramatically improve efficiency.
Alternatively Relax MPEC to NLP

Relax “hard” equilibrium constraints

\[ 0 \leq \mathbf{a} \perp \mathbf{b} \geq 0 \iff \left\{ \begin{array}{l} \mathbf{a}, \mathbf{b} \geq 0 \\ \mathbf{a}'\mathbf{b} = 0 \end{array} \right\} \]

to “soft” inexact constraints

\[ 0 \leq \mathbf{a} \perp_{\text{tol}} \mathbf{b} \geq 0 \iff \left\{ \begin{array}{l} \mathbf{a}, \mathbf{b} \geq 0 \\ \mathbf{a}'\mathbf{b} \leq \text{tol} \end{array} \right\} \]

tol is some user-defined tolerance.
Relaxed Bilevel Program as NLP

Outer Level
\[
\min_{x, y \in Y} \quad F(x, y)
\]
\[
s.t. \quad G(x, y) \leq 0
\]
\[
\lambda \geq 0
\]

Inner Level
\[
g(x, y) \leq 0
\]
\[
\lambda_i g_i(x, y) \leq tol \quad i = 1, \ldots, m
\]
\[
\nabla_x f(x, y) + \lambda' \nabla_x f(x, y) = 0
\]

Nonconvex but nicer. Has KKT points. SQP algorithms such as Filter work well.
Outline

- Intro to bilevel/multilevel programming
- Properties of Bilevel Programs
- Cross-Validation and Bilevel Programming
Bilevel T-fold Cross-Validation

\[ \sum_t \text{Test Error } \Omega_t \]

\[ c, \varepsilon \]

\[ \mathcal{W}_t \]

Train Error \( \Omega \setminus \Omega_t \) \ldots Train Error \( \Omega \setminus \Omega_T \)

Leader

T Followers
CV as Bilevel Optimization (Bennett et al 2006)

- Bilevel Program for $T$ folds

$$\min_{C, \varepsilon} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{|\Omega_t|} \sum_{i \in \Omega_t} |x_i^t w^t - y_i|$$

$s.t. \quad w^t \in \arg \min_w \left\{ C \sum_{j \in \tilde{\Omega}_t} \max \left( |x_j^t w - y_j| - \varepsilon, 0 \right) + \frac{1}{2} \|w\|_2^2 \right\}$

$t = 1, \ldots, T$

- Prior Approaches: Golub et al., 1979, Generalized Cross-Validation for one parameter in Ridge Regression
Benefit: More Design Variables

Add feature box constraint: \(-\bar{w} \leq w \leq \bar{w}\) in the inner-level problems.

\[
\min_{\bar{w}, C, \varepsilon} \quad \frac{1}{T} \sum_{t=1}^{T} \frac{1}{|\Omega_t|} \sum_{i \in \Omega_t} |x_i'w^t - y_i| \\
\text{s.t.} \quad w^t \in \arg \min \left\{ C \sum_{j \in \Omega_t} \max \left( |x_j'w - y_j| - \varepsilon, 0 \right) + \frac{1}{2} \|w\|^2 \right\}
\]
Inner-level Problem for $t$-th Fold

$$\min_{-\bar{w} \leq w^t \leq \bar{w}} \left\{ C \sum_{j \in \Omega_t} \max \left( \left| x'_j w^t - y_j \right| - \varepsilon, 0 \right) + \frac{1}{2} \| w^t \|_2^2 \right\}$$

$$\uparrow$$

$$\min_{w^t, \xi^t} \frac{1}{2} \| w^t \|_2^2 + C \sum_{j \in \Omega_t} \xi^t$$

s.t. $$\xi^t \geq x'_j w^t - y_j - \varepsilon$$

$$\xi^t \geq y_j - x'_j w^t - \varepsilon$$

$$\xi^t \geq 0$$

$$-\bar{w} \leq w^t \leq \bar{w}$$
Inner problem optimality conditions for fixed $C, \varepsilon, \bar{w}$

\[
\begin{align*}
0 & \leq \alpha_{j,t,+} \perp y_j - x_j'w^t + \varepsilon + \xi_j^t \geq 0 \\
0 & \leq \alpha_{j,t,-} \perp x_j'w^t - y_j + \varepsilon + \xi_j^t \geq 0 \\
0 & \leq \xi_j^t \perp C - \alpha_{j,t,+} - \alpha_{j,t,-} \geq 0 \\
0 & \leq \gamma^{t,+} \perp \bar{w} - w^t \geq 0 \\
0 & \leq \gamma^{t,-} \perp \bar{w} + w^t \geq 0 \\
0 & = w^t + \sum_{j \in \bar{\Omega}_t} (\alpha_{j,t,+} - \alpha_{j,t,-}) x_j + \gamma^{t,+} - \gamma^{t,-}
\end{align*}
\]

where $a \perp b \leftrightarrow a'b = 0$
Bilevel Problem as MPEC

$$\begin{align*}
\min_{\mathbf{w}^t, \overline{\mathbf{w}}, \mathbf{C}, \varepsilon} & \quad \frac{1}{T} \sum_{t=1}^{T} \frac{1}{|\Omega_t|} \sum_{i \in \Omega_t} \left| \mathbf{x}_i' \mathbf{w}^t - y_i \right| \\
\text{s.t.} & \quad \text{for } t = 1, \ldots, T \\
0 & \leq \alpha_{j,t,+} \quad \perp \quad y_j - \mathbf{x}_j' \mathbf{w}^t + \varepsilon + \xi_j^t \quad \geq \quad 0 \\
0 & \leq \alpha_{j,t,-} \quad \perp \quad \mathbf{x}_j' \mathbf{w}^t - y_j + \varepsilon + \xi_j^t \quad \geq \quad 0 \\
0 & \leq \xi_j^t \quad \perp \quad \mathbf{C} - \alpha_{j,t,+} - \alpha_{j,t,-} \quad \geq \quad 0 \\
0 & \leq \gamma_{t,+} \quad \perp \quad \overline{\mathbf{w}} - \mathbf{w}^t \quad \geq \quad 0 \\
0 & \leq \gamma_{t,-} \quad \perp \quad \overline{\mathbf{w}} + \mathbf{w}^t \quad \geq \quad 0 \\
0 = & \mathbf{w}^t + \sum_{j \in \overline{\Omega}_t} \left( \alpha_{j,t,+} - \alpha_{j,t,-} \right) \mathbf{x}_j + \gamma_{t,+} - \gamma_{t,-}
\end{align*}$$

Replace $T$ inner-level problems with corresponding optimality conditions.
Relaxed Bilevel CV as NLP

\[
\min_{w^t, \bar{w}, C, \varepsilon} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\Omega_t} \sum_{i \in \Omega_t} \left| x'_i w^t - y_i \right|
\]

s.t. for \( t = 1, \ldots, T \)

\[
0 \leq \alpha_{j,t,+} \downarrow_{tol} y_j - x'_j w^t + \varepsilon + \xi_j^t \geq 0
\]

\[
0 \leq \alpha_{j,t,-} \downarrow_{tol} x'_j w^t - y_j + \varepsilon + \xi_j^t \geq 0 \quad \forall \ j \in \bar{\Omega}_t
\]

\[
0 \leq \xi_j \downarrow_{tol} C - \alpha_{j,t,+} - \alpha_{j,t,-} \geq 0
\]

\[
0 \leq \gamma_{t,+} \downarrow_{tol} \bar{w} - w^t \geq 0
\]

\[
0 \leq \gamma_{t,-} \downarrow_{tol} \bar{w} + w^t \geq 0
\]

\[
0 = w^t + \sum_{j \in \bar{\Omega}_t} \left( \alpha_{j,t,+} - \alpha_{j,t,-} \right) x_j + \gamma_{t,+} - \gamma_{t,-}
\]

Replace \( T \) inner-level problems with corresponding optimality conditions.
Computational Experiments: DATA

Synthetic
- (5,10,15)-D Data with Gaussian and Laplacian noise and (3,7,10) relevant features.
- NLP: 3-fold CV
- Results: 30 to 90 train, 1000 test points, 10 trials

QSAR/Drug Design
- 4 datasets, 600+ dimensions reduced to 25 top principal components.
- NLP: 5-fold CV
- Results: 40 – 100 train, rest test, 20 trials
Cross-validation Methods Compared

- **Unconstrained Grid:**
  
  Try 3 values each for $C, \varepsilon$

- **Constrained Grid:**
  
  Try 3 values each for $C, \varepsilon$, and
  
  $\{0, 1\}$ for each component of $\bar{w}$

- **Bilevel/FILTER:** Nonlinear program solved using off-the-shelf SQP algorithm, FILTER via NEOS
15-D Data: Objective Value

The graph shows the objective value for different data points (15 pts, 30 pts, 60 pts, 90 pts) with three categories: Unc Grid, Con Grid, and Filter. The y-axis represents the value ranging from 0 to 3, while the x-axis represents the number of data points.
15-D Data: Computational Time
15-D Data: TEST MAD

- Unc Grid
- Con Grid
- Filter

- 15 pts
- 30 pts
- 60 pts
- 90 pts
QSAR Data: Objective Value

![Bar chart showing objective values for Aquasol, BBB, Cancer, and CCK]
QSAR Data: Computation Time

![Graph showing computation time for Aquasol, BBB, Cancer, and CCK with different grid types.](image-url)
QSAR Data: TEST MAD

![Bar chart showing data for Aquasol, BBB, Cancer, and CCK with categories Unc Grid, Con Grid, and Filter.]
Machine Learning as Bilevel Programming

- New capacity offers new possibilities:
  - Outer level objectives?
  - Inner level problem?
  - classification, ranking, semi-supervised,
    missing values, kernel selection, variable selection, …
- Special purpose algorithms being developed for greater efficiency, scalability, robustness

This work was supported by Office of Naval Research Grant N00014-06-1-0014.