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.

> $x = taxable \ activity$ $t = tax \ rate \ on \ activity$ $t \in T$ $F(t, x) = amount \ of \ taxes \ received$



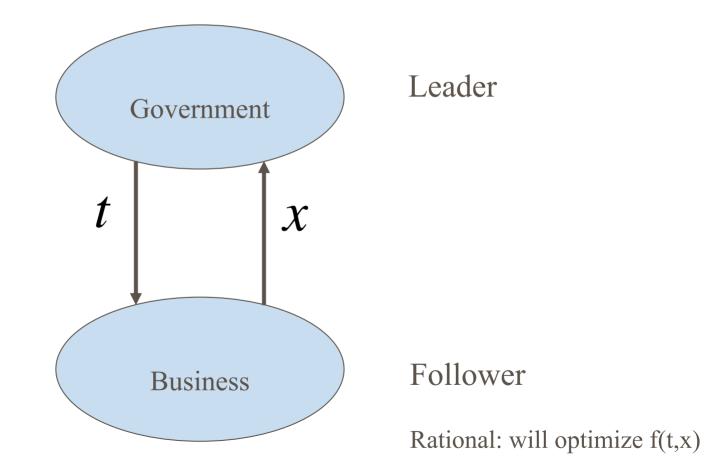
How to set the tax rate?

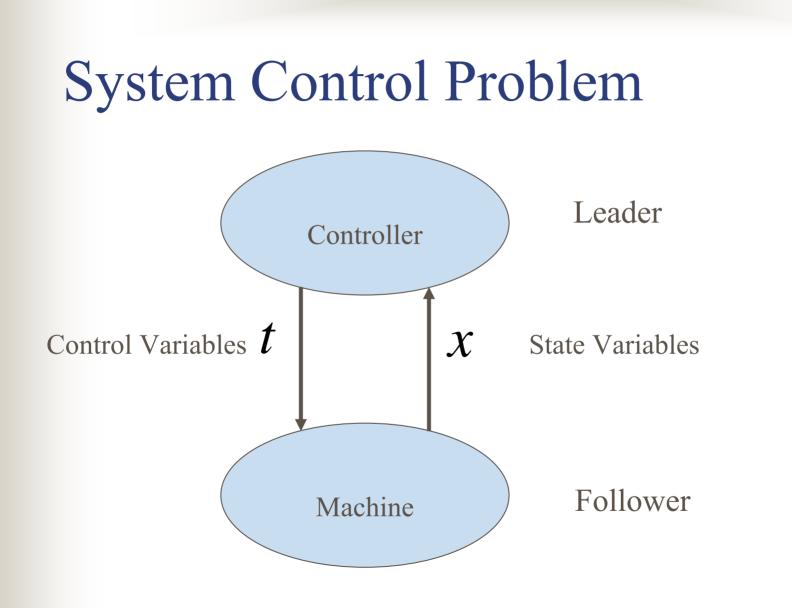
Business wants maximize profit for given tax rate.

 $x = taxable \ activity$ $x \in X$ $t = given \ tax \ rate$ f(t, x) = profit



Assumptions – Stackelberg Game (1952)





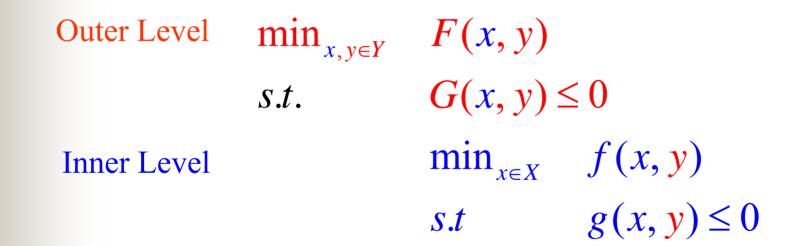
Tax Policy as Bilevel Program $\max_{t \in T, x}$ F(t, x)s.t. $x \in \arg \max_{x' \in X} f(t, x')$

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



Machine Learning Example

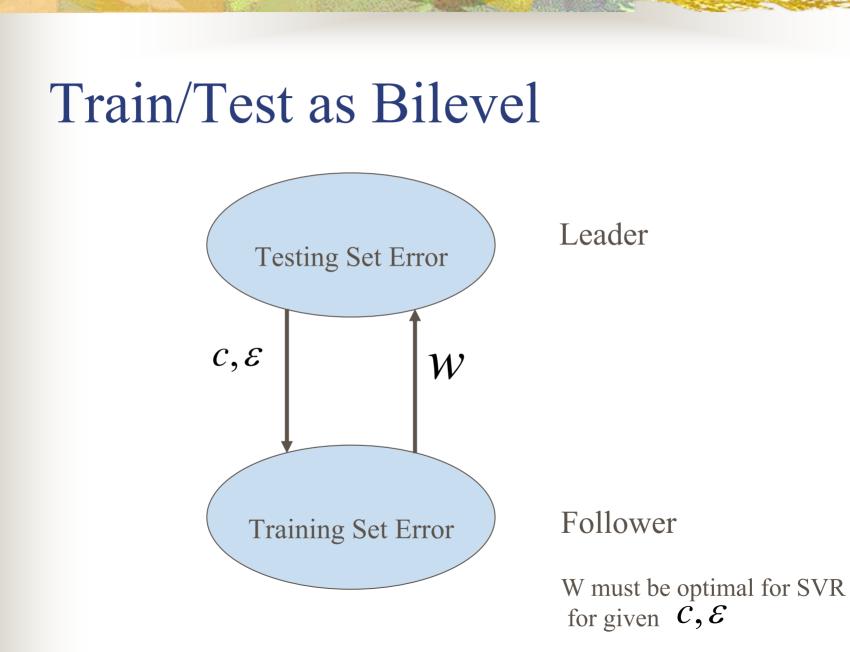
Given

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

Determine c, ϵ such that generalization as measured by the test set is optimized

Grid Search Approach Define grid over C, ε Optimize model on train for each C, ε Select value of C, ε that yields best testing set error C

3



Bilevel Model

Leader: optimize mean absolute testing error by controling c,ɛ Follower: optimizes w using SVR on training data

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

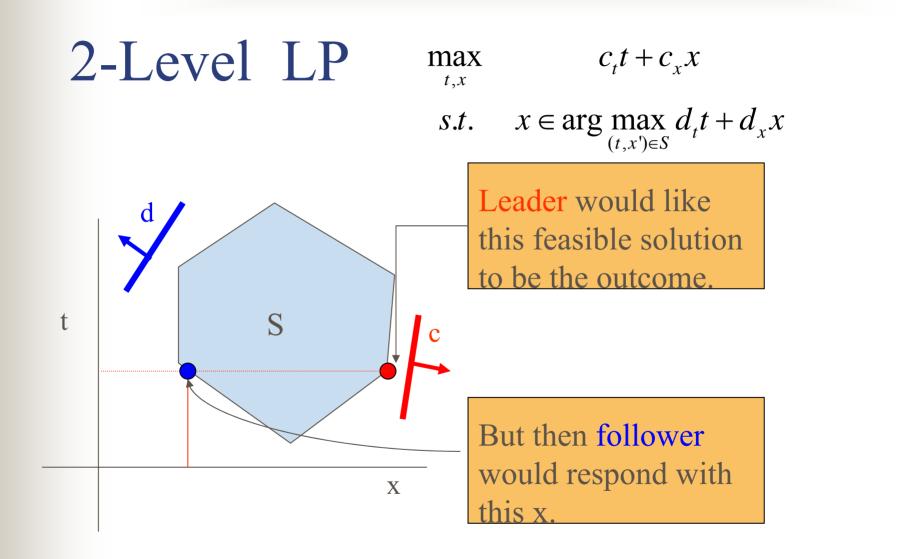
s.t.
$$\min_{w} C \sum_{i \in train} \max(|x_i \cdot w - y_i| - \varepsilon, 0) + \frac{1}{2} ||w||^2$$

Train/Test Linear Bilevel Model

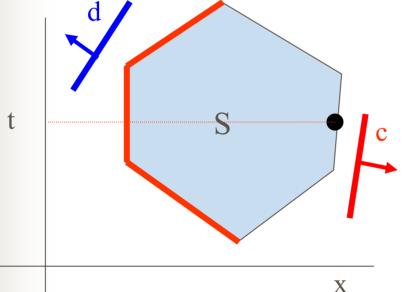
$$\begin{split} \min_{c,\varepsilon,z,w} & \sum_{i \in test} z_i \\ s.t. & -z_i \leq x_i \cdot w - y \leq z_i \quad i \in test \\ & \min_w \quad C \sum_{i \in train} (\xi_i + \xi_i^*) + \frac{1}{2} ||w||^2 \\ & x_i \cdot w - y_i \leq \varepsilon + \xi_i \quad i \in train \\ s.t. & -x_i \cdot w + y_i \leq \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \end{split}$$

Outline

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



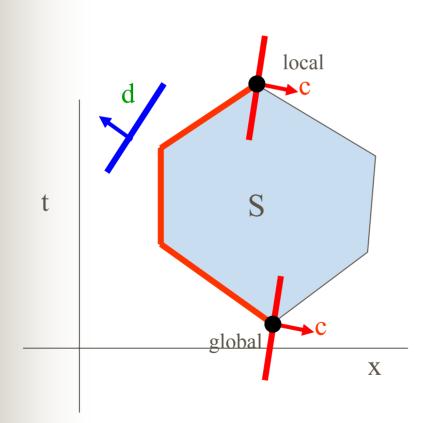
Reaction Set



The leader can examine reaction of follower for each feasible choice of t.

This forms the reaction set S(t)

Optimal Solution



Equivalent problem

 $\max_{t,x} \quad c_t t + c_x x$

- s.t. $(t, x) \in S(t)$ S(t) nonconvex,
- 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

 $\min_{x} f(x, y)$

Inner

s.t $g(x, y) \leq 0$ $g(x, y) \leq 0$ primal feasibility $\nabla_{\mathbf{x}} f(x, y) + \lambda' \nabla_{\mathbf{x}} f(x, y) = 0$ dual feasibility **KKT** $\lambda > 0$ $\lambda \perp g(x, y)$ complementariy *i.e.* $\lambda_i g_i(x, y) = 0$ i = 1, ..., m

If convex problem and (x^*, y^*, λ^*) is KKT point, then x*,y* is globally optimal.

KKT Optimality Conditions of Inner Problem

Inner	$\min_{x} f(x, y)$	
	s.t $g(x, y) \leq 0$	
	\downarrow	
KKT	$g(x, y) \le 0$	primal feasibility
	$\nabla_{\mathbf{x}} f(x, y) + \lambda' \nabla_{\mathbf{x}} f(x, y) = 0$	dual feasibility
	$\lambda \ge 0$	
	$\lambda_{\perp} g(x, y)$	complementariy
	<i>i.e.</i> $\lambda_i g_i(x, y) = 0$	

C()

If (x^*, y^*) is optimal and **Constraint Qualification** satisfied, then KKT point (x^*, y^*, λ^*) 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 $\min_{x,y\in Y}$ F(x,y)s.t. $G(x,y) \le 0$ Inner Level $\lambda \ge 0 \perp g(x,y) \le 0$ KKT $\nabla_x f(x,y) + \lambda' \nabla_x f(x,y) = 0$

Combinatorial Global Search

- For each of m equilibrium constraints $\lambda_i g_i(x, y) = 0 \Leftrightarrow \begin{cases} \lambda_i = 0 \\ or \\ g_i(x, y) = 0 \end{cases}$
- 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 $0 \le \mathbf{a} \perp \mathbf{b} \ge 0 \Leftrightarrow \begin{cases} \mathbf{a}, \mathbf{b} \ge 0 \\ \mathbf{a}'\mathbf{b} = 0 \end{cases}$

to "soft" inexact constraints $0 \le \mathbf{a} \perp_{tol} \mathbf{b} \ge 0 \Leftrightarrow \begin{cases} \mathbf{a}, \mathbf{b} \ge 0 \\ \mathbf{a}'\mathbf{b} \le tol \end{cases}$

tol is some user-defined tolerance.

Relaxed Bilevel Program as NLP

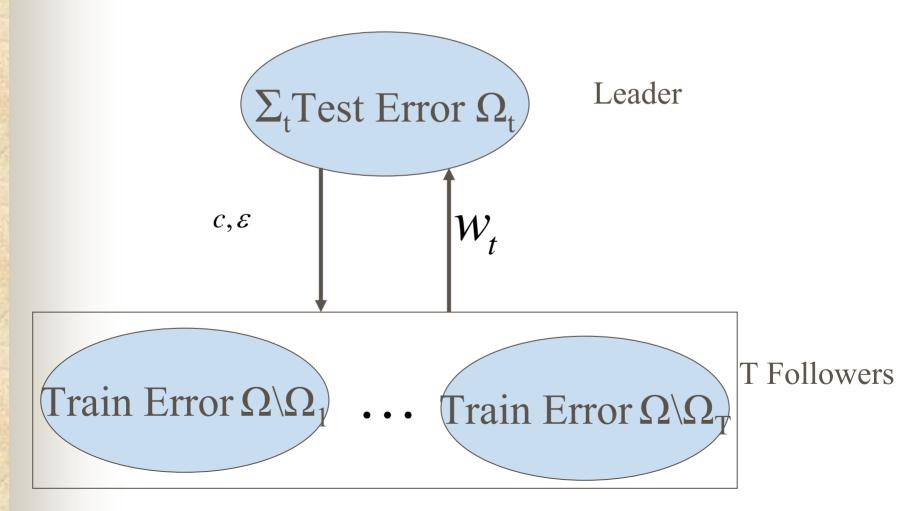
Outer Level $\min_{x,y\in Y}$ F(x,y)s.t. $G(x,y) \leq 0$ $\lambda \geq 0$ $g(x,y) \leq 0$ Inner Level $\lambda_i g_i(x,y) \leq tol$ i = 1,...,m $\nabla_x f(x,y) + \lambda' \nabla_x f(x,y) = 0$

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



CV as Bilevel Optimization (Bennett et al 2006)

Bilevel Program for *T* folds

$$\min_{C,\varepsilon} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{|\Omega_t|} \sum_{i \in \Omega_t} |\mathbf{x}'_i \mathbf{w}^t - y_i| \qquad \text{Outer-level validation problem}$$

s.t. $\mathbf{w}^t \in \arg\min_{\mathbf{w}} \left\{ C \sum_{j \in \overline{\Omega}_t} \max\left(|\mathbf{x}'_j \mathbf{w} - y_j| - \varepsilon, 0 \right) + \frac{1}{2} \|\mathbf{w}\|_2^2 \right\}$

t = 1, ..., T

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

Benefit: More Design Variables

Add feature box constraint: $-\overline{\mathbf{w}} \leq \mathbf{w} \leq \overline{\mathbf{w}}$ in the inner-level problems.

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

s.t.
$$\mathbf{w}^t \in \operatorname*{arg\,min}_{-\overline{\mathbf{w}} \le \mathbf{w} \le \overline{\mathbf{w}}} \left\{ C \sum_{j \in \overline{\Omega}_t} \max\left(|\mathbf{x}_j' \mathbf{w} - y_j| - \varepsilon, 0 \right) + \frac{1}{2} ||\mathbf{w}||_2^2 \right\}$$

Inner-level Problem for *t*-th Fold

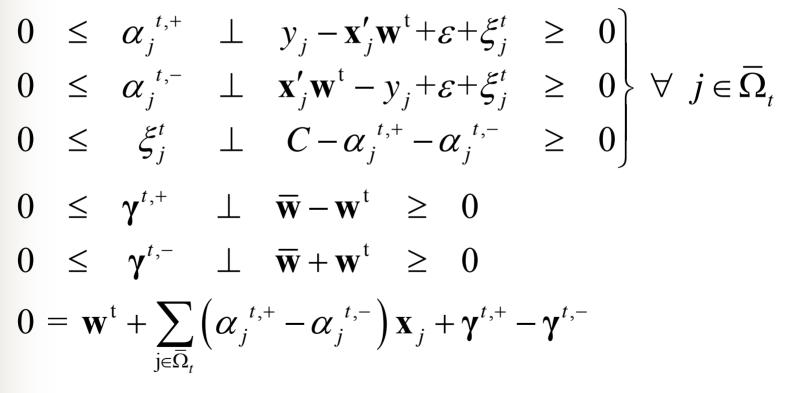
$$\min_{\mathbf{w}^{t} \leq \mathbf{w}} \left\{ C \sum_{j \in \overline{\Omega}_{t}} \max\left(\left\| \mathbf{x}_{j}^{*} \mathbf{w}^{t} - y_{j} \right\| - \varepsilon, 0 \right) + \frac{1}{2} \left\| \mathbf{w}^{t} \right\|_{2}^{2} \right\}$$

$$\lim_{\mathbf{w}^{t}, \xi^{t}} \frac{1}{2} \left\| \mathbf{w}^{t} \right\|_{2}^{2} + C \sum_{j \in \overline{\Omega}_{t}} \xi_{j}^{t}$$
s.t. $\xi_{j}^{t} \geq \mathbf{x}_{j}^{*} \mathbf{w}^{t} - y_{j} - \varepsilon$
 $\xi_{j}^{t} \geq y_{j} - \mathbf{x}_{j}^{*} \mathbf{w}^{t} - \varepsilon$
 $\xi_{j}^{t} \geq 0$
 $-\overline{\mathbf{w}} \leq \mathbf{w}^{t} \leq \overline{\mathbf{w}}$

13

ALL AND ALL

Inner problem optimality conditions for fixed $C, \varepsilon, \overline{w}$



where $\mathbf{a} \perp \mathbf{b} \Leftrightarrow \mathbf{a'b} = 0$

Bilevel Problem as MPEC

$$\min_{\mathbf{w}^{t}, \bar{\mathbf{w}}, C, \varepsilon} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{|\Omega_{t}|} \sum_{i \in \Omega_{t}} |\mathbf{x}_{i}' \mathbf{w}^{t} - y_{i}|$$

s.t. for $t = 1, ..., T$

 P_i Replace *T* inner-level problems with corresponding optimality conditions

 $\begin{array}{rclcrcl}
0 &\leq & \alpha_{j}^{t,+} &\perp & y_{j} - \mathbf{x}_{j}' \mathbf{w}^{t} + \varepsilon + \xi_{j}^{t} &\geq & 0 \\
0 &\leq & \alpha_{j}^{t,-} &\perp & \mathbf{x}_{j}' \mathbf{w}^{t} - y_{j} + \varepsilon + \xi_{j}^{t} &\geq & 0 \\
0 &\leq & \xi_{j}^{t} &\perp & C - \alpha_{j}^{t,+} - \alpha_{j}^{t,-} &\geq & 0
\end{array} \quad \forall \quad j \in \overline{\Omega}_{t} \\
0 &\leq & \mathbf{y}^{t,+} &\perp & \overline{\mathbf{w}} - \mathbf{w}^{t} &\geq & 0
\end{array}$

$$0 \leq \boldsymbol{\gamma}^{t,-} \perp \boldsymbol{\overline{w}} + \boldsymbol{w}^{t} \geq 0$$

 $0 = \mathbf{w}^{t} + \sum_{j \in \overline{\Omega}_{t}} \left(\alpha_{j}^{t,+} - \alpha_{j}^{t,-} \right) \mathbf{x}_{j} + \mathbf{\gamma}^{t,+} - \mathbf{\gamma}^{t,-}$

Relaxed Bilevel CV as NLP

$$\min_{\mathbf{w}^{t}, \bar{\mathbf{w}}, C, \varepsilon} \quad \frac{1}{T} \sum_{t=1}^{T} \frac{1}{|\Omega_{t}|} \sum_{i \in \Omega_{t}} |\mathbf{x}_{i}' \mathbf{w}^{t} - y_{i}|$$

s.t. for $t = 1, ..., T$

s.t.

Replace T inner-level problems with corresponding optimality conditions

Computational Experiments: DATA

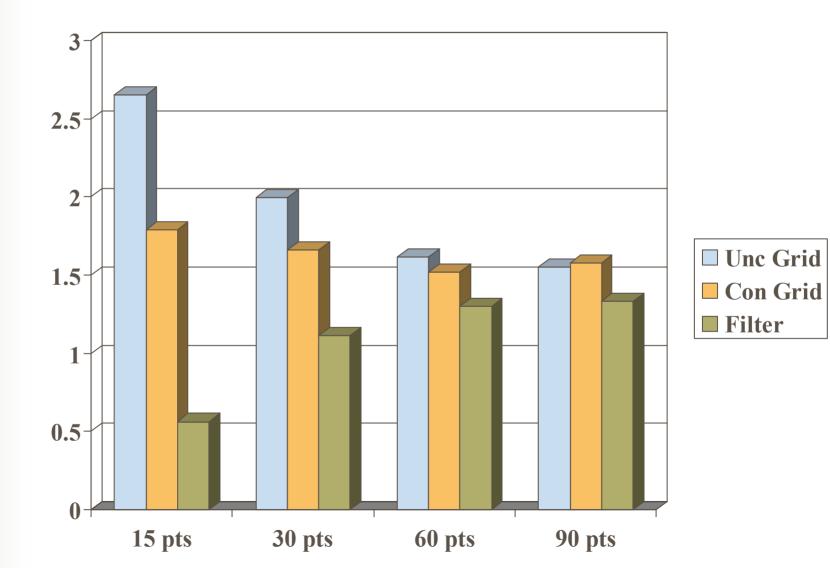
Synthetic

- (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

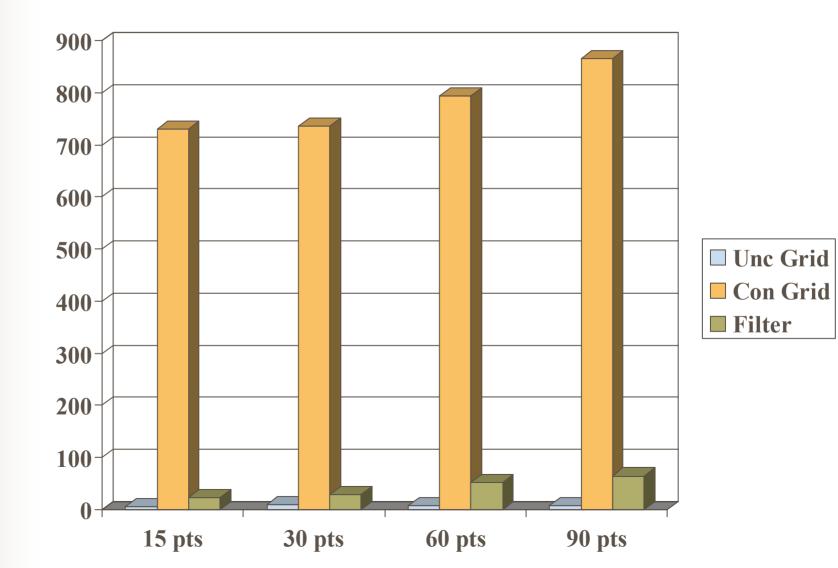
Cross-validation Methods Compared

- Unconstrained Grid:
 - Try 3 values each for C,ε
- Constrained Grid:
 - Try 3 values each for C, ε , and
 - $\{0, 1\}$ for each component of $\overline{\mathbf{w}}$
- Bilevel/FILTER: Nonlinear program solved using off-the-shelf SQP algorithm, FILTER via NEOS

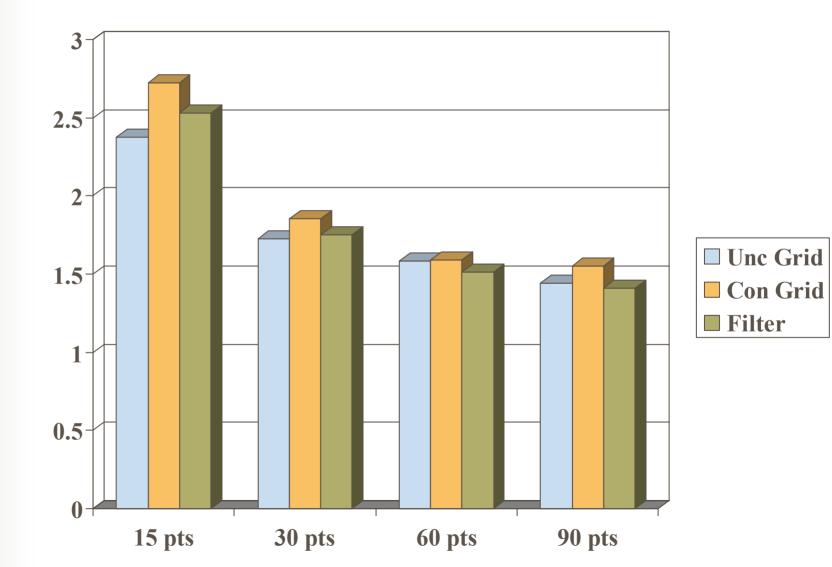
15-D Data: Objective Value



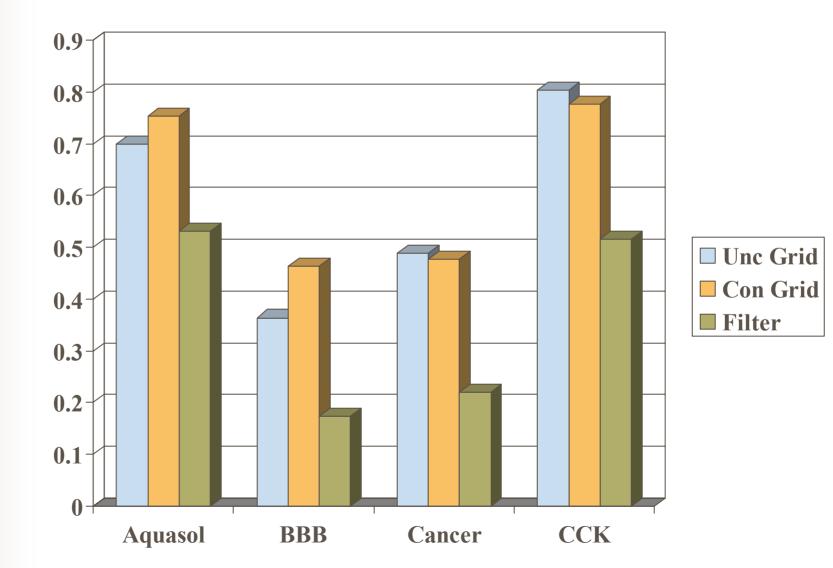
15-D Data: Computational Time



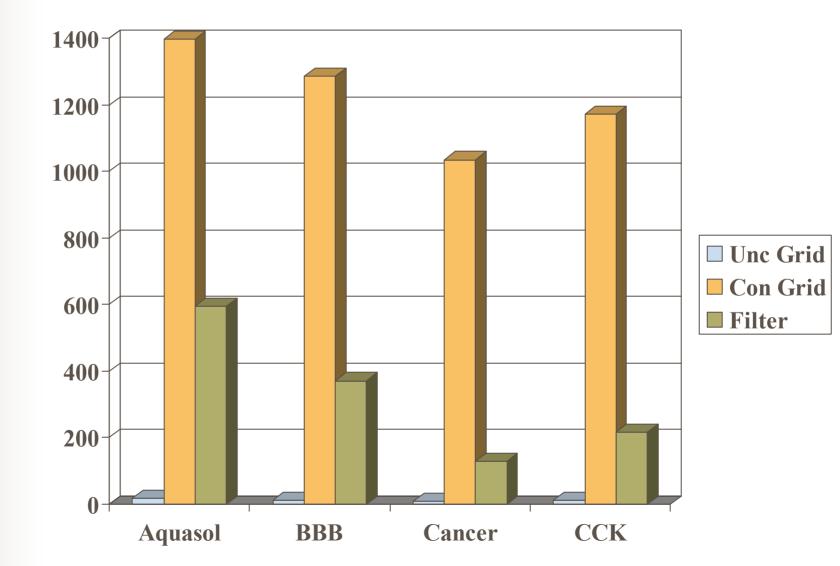
15-D Data: TEST MAD



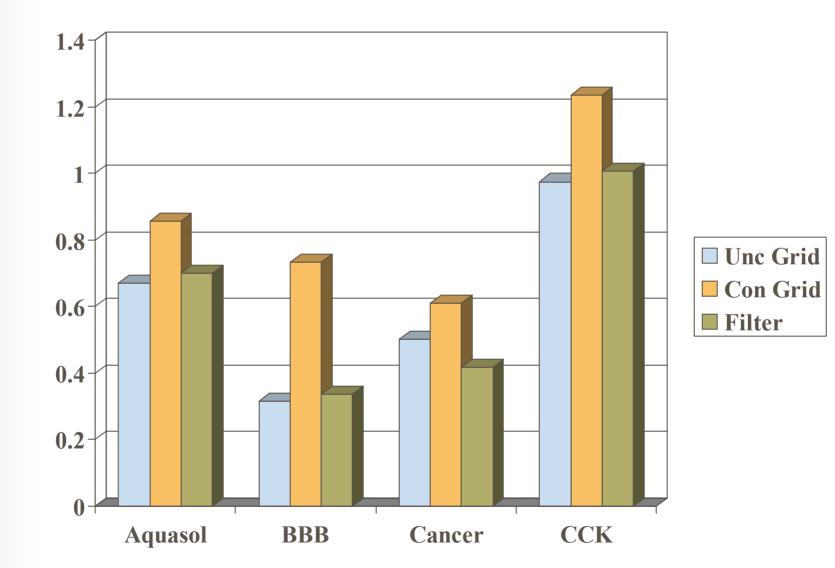
QSAR Data: Objective Value



QSAR Data: Computation Time



QSAR Data: TEST MAD



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.