Statistics and Machine Learning via a Modern Optimization Lens

Dimitris Bertsimas

Operations Research Center
Massachusetts Institute of Technology

June 2015
1. Motivation

2. Best Subset Selection
   Joint work with Angie King and Rahul Mazumder.

3. Least Median of Squares Regression

4. An Algorithmic Approach to Linear Regression
   Joint work with Angie King.

5. Conclusions
Continuous optimization methods have historically played a significant role in statistics.

In the last two decades convex optimization methods have had increasing importance: Compressed Sensing, Matrix Completion among many others.

Many problems in statistics and machine learning can naturally be expressed as Mixed integer optimizations (MIO) problems.

MIO in statistics are considered impractical and the corresponding problems intractable.

Heuristics methods are used: Lasso for best subset regression or CART for optimal classification.
Progress of MIO

- Speed up between CPLEX 1.2 (1991) and CPLEX 11 (2007): 29,000 times
- Gurobi 1.0 (2009) comparable to CPLEX 11
- Speed up between Gurobi 1.0 and Gurobi 5.5 (2013): 20 times
- Total speedup: 580,000 times
- A MIO that would have taken 7 years to solve 20 years ago can now be solved on the same 20-year-old computer in less than one second.
- Hardware speedup: $10^{5.5} = 320,000$ times
- Total Speedup: **200 Billion times!**
Research Objectives

- Given the dramatically increased power of MIO, is MIO able to solve key multivariate statistics problems considered intractable a decade ago?

- How do MIO solutions compete with state of the art solutions?

- Building Regression models is an art. Can we algorithmize the process?

- What are the implications on teaching statistics?
Problems

- Best Subset Regression:
  \[
  \min_\beta \frac{1}{2} \| y - X\beta \|^2 \quad \text{subject to} \quad \|\beta\|_0 \leq k
  \]

- Least Median Regression:
  \[
  \min_\beta \text{median}_{i=1,\ldots,n} |y_i - x_i^T \beta|
  \]

- Algorithmic Regression to accommodate Sparsity, Limiting multicollinearity, Categorical variables, Group sparsity, Nonlinear transformations, Robustness, Statistical significance
Best Subset Regression

\[
\min_{\beta} \quad \frac{1}{2} \|y - X\beta\|_2^2 \quad \text{subject to} \quad \|\beta\|_0 \leq k
\]

- Furnival and Wilson (1974) solve it by implicit enumeration, 1eaps routine in R. Cannot scale beyond \( p = 30 \).

\[
\min_{\beta} \quad \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \sum_i |\beta_i|
\]

- Under regularity conditions on \( X \), Lasso leads to sparse models and good predictive performance.
Lasso is a Robustification method not sparsity inducing

- Norm:
  \[ \| \Delta \|_{(q,p)} = \max_{\beta} \frac{\| \Delta \beta \|_p}{\| \beta \|_q} \]

- Uncertainty Set:
  \[ U_{(q,p)} = \{ \Delta \in \mathbb{R}^{m \times n} : \| \Delta \|_{(q,p)} \leq \lambda \} \]

- Robustification - Regularization Equivalence
  \[ \min_{\beta} \max_{\Delta \in U_{(q,p)}} \| y - (X + \Delta)\beta \|_p = \min_{\beta} \| y - X\beta \|_p + \lambda \| \beta \|_q \]
  and

- \( p=2, \ q=1 \): Lasso.
Our approach

- Natural MIO formulation

\[
\begin{align*}
\min_{\beta, z} & \quad \frac{1}{2} \|y - X\beta\|_2^2 \\
\text{subject to} & \quad |\beta_i| \leq M \cdot z_i, \ i = 1, \ldots, p \\
& \quad \sum_{i=1}^{p} z_i \leq k \\
& \quad z_i \in \{0, 1\}, \ i = 1, \ldots, p
\end{align*}
\]

- First Order methods to find good feasible solutions as warm-starts

- Enhance MIO by warm-starts and improved formulation
Best Subset Regression

Parameters

- $\mu := \max_{i \neq j} |\langle x_i, x_j \rangle|$.
- $\mu(m) := \max_{|I|=m} \max_{j \not\in I} \sum_{i \in I} |\langle x_j, x_i \rangle| \leq m\mu$
- $|\langle x(1), y \rangle| \geq |\langle x(2), y \rangle| \ldots \geq |\langle x(p), y \rangle|$
- $M = \min \left\{ \frac{1}{\eta_k} \sqrt{\sum_{i=1}^{k} |\langle x(i), y \rangle|^2}, \frac{1}{\sqrt{\eta_k}} \|y\|_2 \right\}$
- $\eta_k = (1 - \mu(k - 1))$. 
Special Ordered Sets - formulation

\[
\begin{align*}
\min_{\beta, z} & \quad \|y - X\beta\|_2^2 \\
\text{subject to} & \quad (\beta_i, 1 - z_i) : \text{SOS type-1, } i = 1, \ldots, p \\
& \quad \sum_{i=1}^{p} z_i \leq k \\
& \quad z_i \in \{0, 1\}, i = 1, \ldots, p
\end{align*}
\]
Typical behavior of Overall Algorithm

- **Upper Bounds**: Blue line
- **Lower Bounds**: Dashed black line
- **Global Minimum**: Red line

Diabetes Dataset, $n = 350$, $p = 64$, $k = 6$
Typical behavior of Overall Algorithm
First Order Method

Consider

$$\min_{\beta} \quad g(\beta) = \|y - X\beta\|_2^2 \quad \text{subject to} \quad \|\beta\|_0 \leq k$$

$g(\beta)$ convex and

$$\|\nabla g(\beta) - \nabla g(\beta_0)\| \leq \ell \cdot \|\beta - \beta_0\|.$$  

This implies that for all $L \geq \ell$

$$g(\beta) \leq Q(\beta) = g(\beta_0) + \langle \nabla g(\beta_0), \beta - \beta_0 \rangle + \frac{L}{2} \|\beta - \beta_0\|_2^2$$

For the purpose of finding feasible solutions, we propose

$$\min_{\beta} \quad Q(\beta) \quad \text{subject to} \quad \|\beta\|_0 \leq k$$
Solution

- Equivalent to

\[
\min_{\beta} \frac{L}{2} \left\| \beta - \left( \beta_0 - \frac{1}{L} \nabla g(\beta_0) \right) \right\|^2_2 - \frac{1}{2L} \left\| \nabla g(\beta_0) \right\|^2_2 \quad \text{s.t.} \quad \|\beta\|_0 \leq k
\]

- Reducing to

\[
\min_{\beta} \|\beta - u\|^2_2 \quad \text{subject to} \quad \|\beta\|_0 \leq k
\]

- Order: \(|u(1)| \geq |u(2)| \geq \ldots \geq |u(p)|,\)

- Optimal solution is \(\beta^* = H_k(u),\) where \(H_k(u)\) retains the \(k\) largest elements of \(u\) and sets the rest to zero.

\[
(H_k(u))_i = \begin{cases} 
  u_i, & i \in \{(1), \ldots, (k)\} \\
  0, & \text{otherwise}
\end{cases}
\]
First Order Algorithm

1. Initialize with a solution $\beta_0$; $m = 0$.

2. $m := m + 1$.

3. $\tilde{\beta}_{m+1} = H_k (\beta_m - \frac{1}{L} \nabla g(\beta_m))$.

4. Perform a line search: $\lambda_{m+1} = \arg\min_\lambda g(\lambda \tilde{\beta}_{m+1})$

   \[
   \beta_{m+1} = \lambda_{m+1} \tilde{\beta}_{m+1}
   \]

5. Repeat Steps 2-4 until $\|\beta_{m+1} - \beta_m\| \leq \epsilon$. 
Rate of Convergence

- The sequence $g(\beta_m)$ converges to $g(\bar{\beta})$ where
  \[ \bar{\beta} = H_k \left( \bar{\beta} - \frac{1}{L} \nabla g(\bar{\beta}) \right). \]

- After $M$ iterations:
  \[
  \min_{m=0,\ldots,M} \| \beta_{m+1} - \beta_m \|^2 \leq \frac{2 \cdot (g(\beta_0) - g(\bar{\beta}))}{M \cdot (L - \ell)}
  \]

- After $M = O\left(\frac{1}{\epsilon}\right)$ iterations the Algorithm converges.
Properties

- Let $S_m = \{ i : \beta_{m,i} \neq 0, \beta_m = (\beta_{m,1}, \ldots, \beta_{m,p}) \}$ be the support of the solution $\beta_m$.

- The support stabilizes, i.e., there exists an $m^*$ such that $S_{m+1} = S_m$ for all $m \geq m^*$.

- Once the support stabilizes, then the algorithm converges linearly.

- The practical implication is that the algorithm scales to very large problems and converges very fast.
Quality of Solutions for $n > p$

Diabetes data: $n = 350$, $p = 64$.

Relative Accuracy $= (f_{\text{alg}} - f_{\ast})/f_{\ast}$

maximum time 500 secs

<table>
<thead>
<tr>
<th>$k$</th>
<th>First Order Accuracy</th>
<th>First Order Time</th>
<th>MIO Cold Start Accuracy</th>
<th>MIO Cold Start Time</th>
<th>MIO Warm Start Accuracy</th>
<th>MIO Warm Start Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>0.1306</td>
<td>1</td>
<td>0.0036</td>
<td>500</td>
<td>0</td>
<td>346</td>
</tr>
<tr>
<td>20</td>
<td>0.1541</td>
<td>1</td>
<td>0.0042</td>
<td>500</td>
<td>0</td>
<td>77</td>
</tr>
<tr>
<td>49</td>
<td>0.1915</td>
<td>1</td>
<td>0.0015</td>
<td>500</td>
<td>0</td>
<td>87</td>
</tr>
<tr>
<td>57</td>
<td>0.1933</td>
<td>1</td>
<td>0</td>
<td>500</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Quality of Solutions for $n < p$

Synthetic data: $n = 50$, $p = 2000$.  

<table>
<thead>
<tr>
<th>$k$</th>
<th>First Order</th>
<th>MIO Cold</th>
<th>MIO Warm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Start</td>
<td>Start</td>
</tr>
<tr>
<td></td>
<td>Time</td>
<td>Accuracy</td>
<td>Time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNR = 3</td>
<td>4</td>
<td>0.1091</td>
<td>42.9</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.1647</td>
<td>37.2</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.6152</td>
<td>41.1</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.7843</td>
<td>40.7</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.5515</td>
<td>38.8</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.7131</td>
<td>45.0</td>
</tr>
<tr>
<td>SNR = 7</td>
<td>4</td>
<td>0.2708</td>
<td>47.8</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.5072</td>
<td>45.6</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>1.3221</td>
<td>40.3</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.9745</td>
<td>40.9</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.8293</td>
<td>40.5</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>1.1879</td>
<td>44.2</td>
</tr>
</tbody>
</table>
Sparsity Detection for $n = 500, \ p = 100$
Prediction Error = \( \frac{\| X\beta_{alg} - X\beta_{true} \|^2}{\| X\beta_{true} \|^2} \)
Sparsity Detection for $n = 50, \ p = 2000$

![Bar chart showing sparsity detection results for different methods: Lasso, First Order + MIO, First Order Only, and Sparsenet. The x-axis represents the signal-to-noise ratio, and the y-axis shows the number of nonzeros.](image-url)
Prediction Error for $n = 50, p = 2000$
What did we learn?

- For the case $n > p$, MIO+warm-starts finds provably optimal solutions for $n = 1000s$, $p = 100s$ in minutes.

- For the case $n < p$, MIO+warm-starts finds solutions with better prediction accuracy than Lasso or $n = 100s$, $p = 1000s$ in minutes and proving optimality in hours.

- MIO solutions have a significant edge in detecting sparsity, and outperform Lasso in prediction accuracy.

- Modern optimization (MIO+warm-starts) is capable of solving large scale instances.
The Art of Building Regression Models: Current Practice

- Transform variables.
- Pairwise scatterplots, correlation matrix.
- Delete redundant variables.
- Fit full model, delete variables with insignificant $t$-tests. Examine residuals.
- See if additional variables can be dropped/new variables brought in.
- Validate the final model.
Aspirations: From Art to Science

- Propose an algorithm (automated process) to build regression models.

- Approach: Express all desirable properties as MIO constraints.
The initial MIO model

\[
\begin{align*}
\min_{\beta, z} & \quad \frac{1}{2} \| y - X \beta \|^2_2 + \Gamma \| \beta \|_1 \\
\text{s.t.} & \quad z_l \in \{0, 1\}, \quad l = 1, \ldots, p \\
& \quad -M z_l \leq \beta_l \leq M z_l, \quad l = 1, \ldots, p \\
& \quad \sum_{l=1}^{p} z_l \leq k \\
& \quad z_1 = \ldots = z_l \quad \forall i = 1, \ldots, l \in G S_m, \quad \forall m \\
& \quad z_i + z_j \leq 1 \quad \forall (i, j) \in H C \\
& \quad \sum_{i \in T_m} z_i \leq 1 \quad \forall m
\end{align*}
\]

Robustness

Sparsity

Group Sparsity

Pairwise Collinearity

Nonlinear Transform
Controlling Multicollinearity and Statistical Significance

- Given a set $S$, we compute via bootstrap confidence levels of the variables in $S$ as well as the condition number of the model.
- If the condition number is higher than desired or there are variables that are not statistically significant, we add the constraint:

$$\sum_{j \in S} z_j \leq |S| - 1$$

- Iterate.
Contrast with existing practice

- All desired properties are simultaneously enforced.

- MIO does not have to choose which model properties to favor by performing the steps in a certain order.

- MIO is capable of handling datasets with more variables than a modeler can address manually.
## Examples

<table>
<thead>
<tr>
<th>Dataset</th>
<th>n</th>
<th>p</th>
<th>MIO K*</th>
<th>$R^2$</th>
<th>Max Cor</th>
<th>Lasso K*</th>
<th>$R^2$</th>
<th>Max Cor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compact</td>
<td>4096</td>
<td>21</td>
<td>15</td>
<td>0.717</td>
<td>0.733</td>
<td>21</td>
<td>0.725</td>
<td>0.942</td>
</tr>
<tr>
<td>Elevator</td>
<td>8280</td>
<td>18</td>
<td>11</td>
<td>0.808</td>
<td>0.678</td>
<td>15</td>
<td>0.809</td>
<td>0.999</td>
</tr>
<tr>
<td>LPGA 2009</td>
<td>73</td>
<td>11</td>
<td>7</td>
<td>0.814</td>
<td>0.784</td>
<td>10</td>
<td>0.807</td>
<td>0.943</td>
</tr>
<tr>
<td>Airline Costs</td>
<td>15</td>
<td>9</td>
<td>2</td>
<td>0.672</td>
<td>0.501</td>
<td>9</td>
<td>0.390</td>
<td>0.973</td>
</tr>
<tr>
<td>Diabetes</td>
<td>221</td>
<td>64</td>
<td>4</td>
<td>0.334</td>
<td>0.423</td>
<td>14</td>
<td>0.381</td>
<td>0.672</td>
</tr>
<tr>
<td>HIV 2</td>
<td>528</td>
<td>98</td>
<td>11</td>
<td>0.945</td>
<td>0.662</td>
<td>39</td>
<td>0.944</td>
<td>0.760</td>
</tr>
</tbody>
</table>
Effect of Outliers in Regression

- Least Squares (LS) estimator
  \[
  \hat{\beta}^{(LS)} \in \arg\min_{\beta} \sum_{i=1}^{n} r_i^2, \quad r_i = y_i - x_i'\beta
  \]
  is adversely affected by a single outlier and has a limiting Breakdown point of 0 (Dohono & Huber '83; Hampel '75). \((n \to \infty, \text{ and } p \text{ fixed})\)

- The Least Absolute Deviation (LAD) estimator has 0 breakdown point.
  \[
  \hat{\beta}^{(LAD)} \in \arg\min_{\beta} \sum_{i=1}^{n} |r_i|,
  \]

- M-Estimators (Huber 1973) slightly improve the breakdown point
  \[
  \sum_{i=1}^{n} \rho(r_i), \quad \rho(r) \text{ symmetric function}
  \]
Least Median of Squares (LMS) estimator (Rousseeuw (1984))

\[ \hat{\beta}^{(\text{LMS})} \in \arg\min_{\beta} \left( \text{median}_{i=1,...,n} |r_i| \right). \]

LMS highest possible breakdown point of almost 50%.

More generally, Least Quantile of Squares (LQS) estimator:

\[ \hat{\beta}^{(\text{LQS})} \in \arg\min_{\beta} |r(q)|, \]

where, \( r(q) \) is the \( q \)th ordered absolute residual:

\[ |r(1)| \leq |r(2)| \leq \cdots \leq |r(n)|. \]
LMS Computation : State of the art

- LMS problem is NP-hard (Bernholt 2005).

- **Exact** Algorithms
  - Enumeration based, branch and bound, scale like $O(n^p)$.
  - Practically Exact algorithms scale up to $n = 50, p = 5$.

- **Heuristic** Algorithms
  - Based on heuristic subsampling / local searches.
  - Practically scale significantly better, but no guarantees.
Solve the following problem:

\[ \min_{\beta} |r(q)|, \]

where, \( r_i = y_i - x'_i \beta \), \( q \) is a quantile.

Our approach extends to

\[ \min_{\beta} |r(q)|, \text{ subject to } A\beta \leq b \text{ (and/or } \|\beta\|_2^2 \leq \delta) \]
Overview of our approach

1. Write the LMS problem as a MIO.

2. Using first order methods we find feasible solutions which we use as warm-starts.

3. First order solutions serve as warm-starts that enhance running times.
MIO Formulation

Notation:

\[ |r(1)| \leq |r(2)| \leq \ldots \leq |r(n)|. \]

**Step 1:** Introduce binary variables \( z_i, i = 1, \ldots, n \) such that:

\[
z_i = \begin{cases} 
1, & \text{if } |r_i| \leq |r(q)|, \\
0, & \text{otherwise}.
\end{cases}
\]

**Step 2:** Use auxiliary continuous variables \( \mu_i, \overline{\mu}_i \geq 0 \) such that:

\[
|r_i| - \mu_i \leq |r(q)| \leq |r_i| + \overline{\mu}_i, \quad i = 1, \ldots, n,
\]

with the conditions:

If \( |r_i| \geq |r(q)| \), then \( \overline{\mu}_i = 0, \mu_i \geq 0 \),

and if \( |r_i| \leq |r(q)| \), then \( \mu_i = 0, \overline{\mu}_i \geq 0 \).
MIO Formulation

\[
\begin{align*}
\min & \quad \gamma \\
\text{subject to} & \quad |r_i| + \bar{\mu}_i \geq \gamma, \quad i = 1 \ldots, n \\
& \quad \gamma \geq |r_i| - \mu_i, \quad i = 1 \ldots, n \\
& \quad M_uz_i \geq \bar{\mu}_i, \quad i = 1, \ldots, n \\
& \quad M_\ell(1 - z_i) \geq \mu_i, \quad i = 1, \ldots, n \\
& \quad \sum_{i=1}^nz_i = q \\
& \quad \mu_i \geq 0, \quad i = 1, \ldots, n \\
& \quad \bar{\mu}_i \geq 0, \quad i = 1, \ldots, n \\
& \quad z_i \in \{0, 1\}, \quad i = 1, \ldots, n,
\end{align*}
\]

where $\gamma, z_i, \mu_i, \bar{\mu}_i, i = 1, \ldots, n$ are decision variables and $M_u, M_\ell$ are Big-M constants.
Least Median Regression

**SOS-formulation**

\[
\begin{align*}
\min & \quad \gamma \\
\text{subject to} & \quad |r_i| - \gamma = \mu_i - \overline{\mu}_i, \quad i = 1 \ldots, n \\
& \quad \sum_{i=1}^{n} z_i = q \\
& \quad \gamma \geq \overline{\mu}_i, \quad i = 1 \ldots, n \\
& \quad \overline{\mu}_i \geq 0, \quad i = 1 \ldots, n \\
& \quad \mu_i \geq 0, \quad i = 1, \ldots, n \\
& \quad (\overline{\mu}_i, \mu_i) : \text{SOS-1}, \quad i = 1, \ldots, n \\
& \quad (z_i, \mu_i) : \text{SOS-1}, \quad i = 1, \ldots, n \\
& \quad z_i \in \{0, 1\}, \quad i = 1, \ldots, n.
\end{align*}
\]
Typical Evolution of MIO

Alcohol Data; \( (n,p,q) = (44,5,31) \)

Alcohol Data; \( (n,p,q) = (44,7,31) \)
First Order Method

- Writing \( r(q) = y(q) - x'(q)\beta \)

\[
|y(q) - x'(q)\beta| = \sum_{i=1}^{q+1} |y(i) - x'(i)\beta| - \sum_{i=1}^{q} |y(i) - x'(i)\beta|,
\]

\[
H_{q+1}(\beta) - H_q(\beta)
\]

- The function \( H_m(\beta) \) is convex in \( \beta \) (sum of piecewise convex linear functions)

\[
H_m(\beta) := \max_w \sum_{i=1}^{n} w_i |y_i - x'_i\beta|
\]

subject to \( \sum_{i=1}^{n} w_i = m \)

\( 0 \leq w_i \leq 1, \quad i = 1, \ldots, n. \)
Expressing $H_{q+1}(\beta)$ a LO

- $H_{q+1}(\beta)$ can be expressed as

$$H_{q+1}(\beta) := \max_w \sum_{i=1}^{n} w_i |y_i - x_i'\beta|$$

subject to $\sum_{i=1}^{n} w_i = q + 1$

$0 \leq w_i \leq 1, \quad i = 1, \ldots, n.$

- By taking the dual and invoking strong duality, we have:

$$H_{q+1}(\beta) = \min_{\theta, \nu} \theta (q + 1) + \sum_{i=1}^{n} \nu_i$$

subject to $\theta + \nu_i \geq |y_i - x_i'\beta|, \quad i = 1, \ldots, n$

$\nu_i \geq 0, \quad i = 1, \ldots, n.$
Expressing $H_q(\boldsymbol{\beta})$

- $H_q(\boldsymbol{\beta})$ can be expressed as

$$H_m(\boldsymbol{\beta}) := \max_{\mathbf{w}} \sum_{i=1}^{n} w_i |y_i - x_i'\boldsymbol{\beta}|$$

subject to

$$\sum_{i=1}^{n} w_i = m$$

$$0 \leq w_i \leq 1, \quad i = 1, \ldots, n.$$  

- Subgradients:

$$\partial H_q(\boldsymbol{\beta}) = \text{conv} \left\{ \sum_{i=1}^{n} w_i^* \text{sgn}(y_i - x_i'\boldsymbol{\beta})\mathbf{x}_i : \mathbf{w}^* \in \arg\max_{\mathbf{w} \in \mathcal{W}_q} \mathcal{L}(\boldsymbol{\beta}, \mathbf{w}) \right\},$$

- $H_q(\boldsymbol{\beta}) \approx H_q(\boldsymbol{\beta}_k) + \langle \partial H_q(\boldsymbol{\beta}_k), \boldsymbol{\beta} - \boldsymbol{\beta}_k \rangle.$
Combining the pieces.....

- Given a current solution $\beta_k$, we find $\beta_{k+1}$:

$$\min_{\nu, \theta, \beta} \theta(q + 1) + \sum_{i=1}^{n} \nu_i - \langle \partial H_q(\beta_k), \beta \rangle$$

subject to

$$\theta + \nu_i \geq |y_i - x'_i/\beta|, \quad i = 1, \ldots, n$$

$$\nu_i \geq 0, \quad i = 1, \ldots, n.$$ 

- Termination

$$\left( |y(q) - x'_(q)/\beta_k| - |y(q) - x'_(q)/\beta_{k+1}| \right) \leq \text{Tol} \cdot |y(q) - x'_(q)/\beta_k|$$

- The algorithm converges in $O(1/\epsilon)$ to a stationary point (local optimal solution).
Impact of Warm-Starts

Evolution of MIO (cold-start) [top] vs (warm-start) [bottom]

$n = 501, p = 5$, synthetic example
Conclusions

- MIO+warm-starts solves to provable optimality problems of medium ($n = 500$) size problems in under two hours.

- MIO+warm-starts finds high quality solutions for large ($n = 10,000$) scale problems in under two hours outperforming all state of the art algorithms that are publicly available for the LQS problem.

- For problems of this size ($n = 10,000$), MIO does not provide a certificate of optimality in a reasonable amount of time.
Remarks on Complexity:

- A key requirement of a theory is to be positively correlated with empirical evidence.
- Consider the Simplex method and solving the TSP.
- A 200 billion speed up forces us to reconsider what is tractable.
- **Definition:** A problem is tractable if it can be solved for sizes and in times that are appropriate for the application.
- Online trading problems need to be solved in milliseconds.
- Regression problems used for planning need to be solved in minutes or in hours.
- Asymptotic polynomial solvability or NP-hardness is not relevant under this definition.
Remarks on Statistics

- Problems in statistics and machine learning considered intractable a generation ago are now tractable.

- Lasso is a robustness property not a sparsity inducing property as widely believed.

- In comparison to Lasso, MIO provides a significance edge for detecting sparsity as well as an edge on predictive accuracy.

- It is my view that the time to include MIO in statistics curricula has come.

- Correspondingly, I will be teaching a first year graduate level class at MIT: Statistics under a modern optimization lens.