Machine Learning under a Modern Optimization Lens

Dimitris Bertsimas
MIT

February 2020

Outline

Papers

1. **Sparse high dimensional regression: Exact scalable algorithms and phase transitions**
   Joint work with Bart van Parys, to appear Annals of Statistics, 2019

2. **Stable Regression**
   Joint work with Ivan Paskov, under review JMLR, 2019.

3. **Interpretable Matrix Completion**
   Joint work with Michael Li, under review, Operations Research 2019.

Motivation

- Continuous optimization methods have historically played a significant role in ML/statistics.
- In the last two decades convex optimization methods have had increasing importance: Compressed Sensing, Matrix Completion among many others.
- Many problems in ML/statistics an 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 6.5 (2015): 48.7 times
- Total speedup 1991-2015: 1,400,000 times
- A MIO that would have taken 16 days to solve 25 years ago can now be solved on the same 25-year-old computer in less than one second.
- Hardware speed: 93.0 PFlop/s in 2016 vs 59.7 GFlop/s in 1993 1,600,000 times
- Total Speedup: **2.2 Trillion times!**
- A MIO that would have taken 71,000 years to solve 25 years ago can now be solved in a modern computer in less than one second.

Research Objectives

- Given the dramatically increased power of MIO, is MIO able to solve key ML/statistics problems considered intractable a decade ago?
- How do MIO solutions compete with state of the art solutions?
- Randomization is the method of choice in a variety of problems in ML/Statistics. Bootstrap, selecting training-validation sets, Randomized clinical trials, random forests. Can optimization play a role?
- What are the implications on teaching ML/statistics?

New Book


The book provides an original treatment of machine learning (ML) using convex, robust and mixed integer optimization that leads to solutions to central ML problems at large scale that can be found in seconds/minutes, can be certified to be optimal in minutes/hours, and outperform classical heuristic approaches in out-of-sample experiments.
Structure of the book:
• Part I covers robust, sparse, nonlinear, holistic regression and extensions.
• Part II contains optimal classification and regression trees.
• Part III outlines prescriptive ML methods.
• Part IV shows the power of optimization over randomization in design of experiments, exceptional responders, stable regression and the bootstrap.
• Part V describes unsupervised methods in ML: optimal missing data imputation and interpretable clustering.
• Part VI develops matrix ML methods: sparse PCA, sparse inverse covariance estimation, factor analysis, matrix and tensor completion.
• Part VII demonstrates how ML leads to interpretable optimization.
Philosophical principles of the book:
• Interpretability in ML is materially important in real world applications.
• Practical tractability not polynomial solvability leads to real world impact.
• NP-hardness is an opportunity not an obstacle.
• ML is inherently linked to optimization not probability theory.
• Data represents an objective reality; models only exist in our imagination.
• Optimization has a significant edge over randomization.
• The ultimate objective in the real world is prescription, not prediction.
DIMITRIS BERTSIMAS is the Boeing Professor of Operations Research, the co-director of the Operations Research Center and the faculty director of the Master of Business Analytics at the Massachusetts Institute of Technology. He is a member of the National Academy of Engineering, an INFORMS fellow, recipient of numerous research and teaching awards, supervisor of 72 completed and 25 current doctoral theses, and co-founder of ten analytics companies.
JACK DUNN is a co-founding partner of Interpretable AI, a leader of interpretable methods in artificial intelligence. He has a Ph.D. in Operations Research from the Massachusetts Institute of Technology. In his doctoral dissertation he developed optimal classification and regression trees, an important part of this book.
Dynamic Ideas LLC
BERTSIMAS • DUNN
Machine Learning under a Modern Optimization Lens
MACHINE LEARNING UNDER A MODERN OPTIMIZATION LENS
DIMITRIS BERTSIMAS
JACK DUNN

Sparse Regression

- Problem with regularization

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta}} & \frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|_2^2+\frac{1}{2\gamma}\left\|\boldsymbol{\beta}\right\|_2^2 \\ \text{s.t.} & \left\|\boldsymbol{\beta}\right\|_0\leq k, \end{array}$$

- Rewrite $\beta_i \rightarrow \beta_i s_i$. Define $\mathbf{S} = diagonal(\mathbf{s})$.
- $S_k := \{\mathbf{s} \in \{0,1\}^p \ : \ \mathbf{e}'\mathbf{s} \leq k\}$

$$\min_{\mathbf{s}\in S_k}\left[\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\ \frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\mathbf{S}\boldsymbol{\beta}\right\|_2^2+\frac{1}{2\gamma}\sum_{i=1}^{p}s_i\beta_i^2\right].$$

- Solution:

$$\begin{array}{ll} \min & c(\mathbf{s})=\frac{1}{2}\mathbf{y}'\left(\mathbb{I}_n+\gamma\sum_j s_j\mathbf{K}_j\right)^{-1}\mathbf{y} \\ \text{s.t.} & \mathbf{s}\in S_k. \end{array}$$

- $\mathbf{K}_j := \mathbf{X}_j\mathbf{X}_j'$
- Binary convex optimization problem

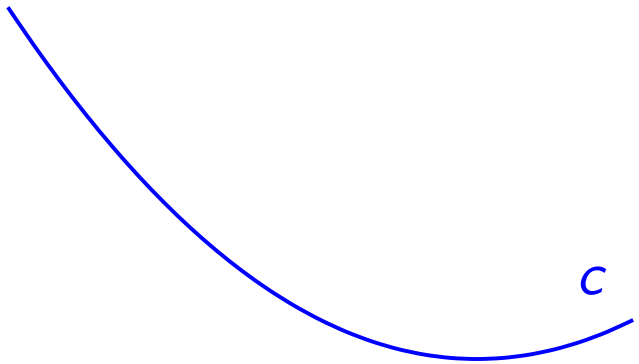
Using Convexity

By convexity of c, for any $\mathbf{s}, \bar{\mathbf{s}} \in S_k$,

$$c(\mathbf{s}) \geq c(\bar{\mathbf{s}}) + \langle \nabla c(\bar{\mathbf{s}}), \mathbf{s} - \bar{\mathbf{s}} \rangle$$

Therefore,

$$c(\mathbf{s}) = \max_{\bar{\mathbf{s}} \in S_k} c(\bar{\mathbf{s}}) + \langle \nabla c(\bar{\mathbf{s}}), \mathbf{s} - \bar{\mathbf{s}} \rangle$$



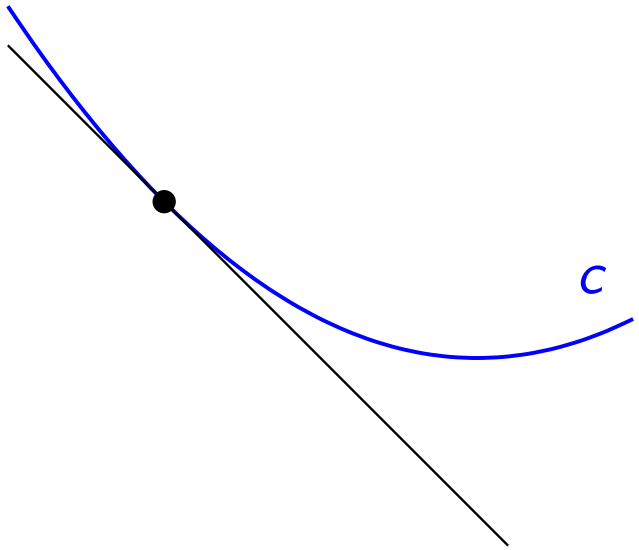
Using Convexity

By convexity of c, for any $\mathbf{s}, \bar{\mathbf{s}} \in S_k$,

$$c(\mathbf{s}) \geq c(\bar{\mathbf{s}}) + \langle \nabla c(\bar{\mathbf{s}}), \mathbf{s} - \bar{\mathbf{s}} \rangle$$

Therefore,

$$c(\mathbf{s}) = \max_{\bar{\mathbf{s}} \in S_k} c(\bar{\mathbf{s}}) + \langle \nabla c(\bar{\mathbf{s}}), \mathbf{s} - \bar{\mathbf{s}} \rangle$$



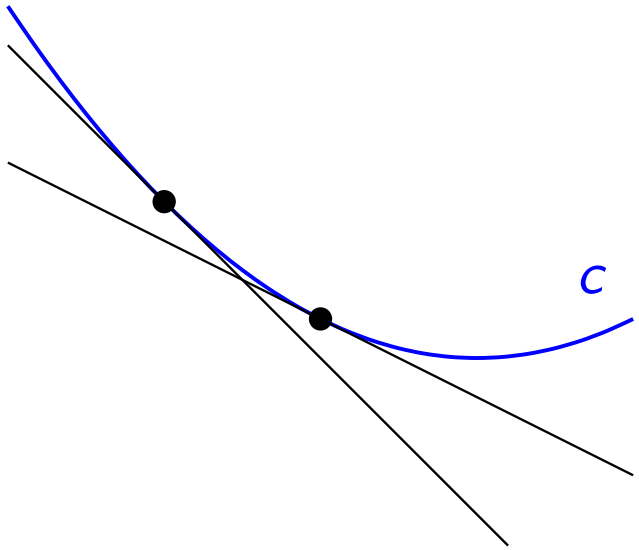
Using Convexity

By convexity of c, for any $\mathbf{s}, \bar{\mathbf{s}} \in S_k$,

$$c(\mathbf{s}) \geq c(\bar{\mathbf{s}}) + \langle \nabla c(\bar{\mathbf{s}}), \mathbf{s} - \bar{\mathbf{s}} \rangle$$

Therefore,

$$c(\mathbf{s}) = \max_{\bar{\mathbf{s}} \in S_k} c(\bar{\mathbf{s}}) + \langle \nabla c(\bar{\mathbf{s}}), \mathbf{s} - \bar{\mathbf{s}} \rangle$$



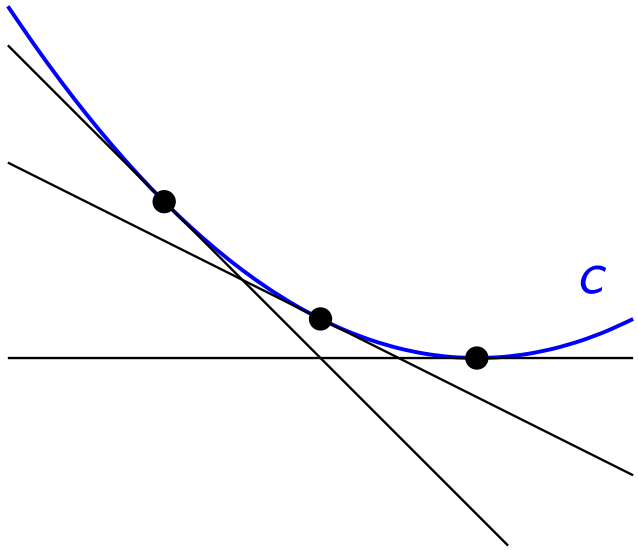
Using Convexity

By convexity of c, for any $\mathbf{s}, \bar{\mathbf{s}} \in S_k$,

$$c(\mathbf{s}) \geq c(\bar{\mathbf{s}}) + \langle \nabla c(\bar{\mathbf{s}}), \mathbf{s} - \bar{\mathbf{s}} \rangle$$

Therefore,

$$c(\mathbf{s}) = \max_{\bar{\mathbf{s}} \in S_k} c(\bar{\mathbf{s}}) + \langle \nabla c(\bar{\mathbf{s}}), \mathbf{s} - \bar{\mathbf{s}} \rangle$$



A Cutting Plane Algorithm

$$c(\mathbf{s}) = \max_{\bar{\mathbf{s}} \in S_k} c(\bar{\mathbf{s}}) + \langle \nabla c(\bar{\mathbf{s}}), \mathbf{s} - \bar{\mathbf{s}} \rangle$$

This leads to a cutting plane algorithm:

1. Pick some $\mathbf{s}_1 \in S_k$ and set $C_1 = \{\mathbf{s}_1\}$.
2. For $t \geq 1$, solve

$$\min_{\mathbf{s} \in S_k} \left[\max_{\bar{\mathbf{s}} \in C_t} c(\bar{\mathbf{s}}) + \langle \nabla c(\bar{\mathbf{s}}), \mathbf{s} - \bar{\mathbf{s}} \rangle \right].$$

3. If solution $\mathbf{s}_t^*$ to Step 2 has $c(\mathbf{s}_t^*) > \max_{\bar{\mathbf{s}} \in C_t} c(\bar{\mathbf{s}}) + \langle \nabla c(\bar{\mathbf{s}}), \mathbf{s}_t^* - \bar{\mathbf{s}} \rangle$, then set $C_{t+1} := C_t \cup \{\mathbf{s}_t^*\}$ and go back to Step 2.

Scalability and Phase Transitions

Cutting plane algorithm can be faster than Lasso.

		Exact T [s]			Lasso T [s]		
		$n = 10$k	$n = 20$k	$n = 100$k	$n = 10$k	$n = 20$k	$n = 100$k
$k = 10$	$p = 50$k	21.2	34.4	310.4	69.5	140.1	431.3
	$p = 100$k	33.4	66.0	528.7	146.0	322.7	884.5
	$p = 200$k	61.5	114.9	NA	279.7	566.9	NA
$k = 20$	$p = 50$k	15.6	38.3	311.7	107.1	142.2	467.5
	$p = 100$k	29.2	62.7	525.0	216.7	332.5	988.0
	$p = 200$k	55.3	130.6	NA	353.3	649.8	NA
$k = 30$	$p = 50$k	31.4	52.0	306.4	99.4	220.2	475.5
	$p = 100$k	49.7	101.0	491.2	318.4	420.9	911.1
	$p = 200$k	81.4	185.2	NA	480.3	884.0	NA

Phase Transitions

- $\mathbf{Y} = \mathbf{X}\beta_{\text{true}} + \mathbf{E}$ where $\mathbf{E}$ is zero mean noise uncorrelated with the signal $\mathbf{X}\beta_{\text{true}}$.

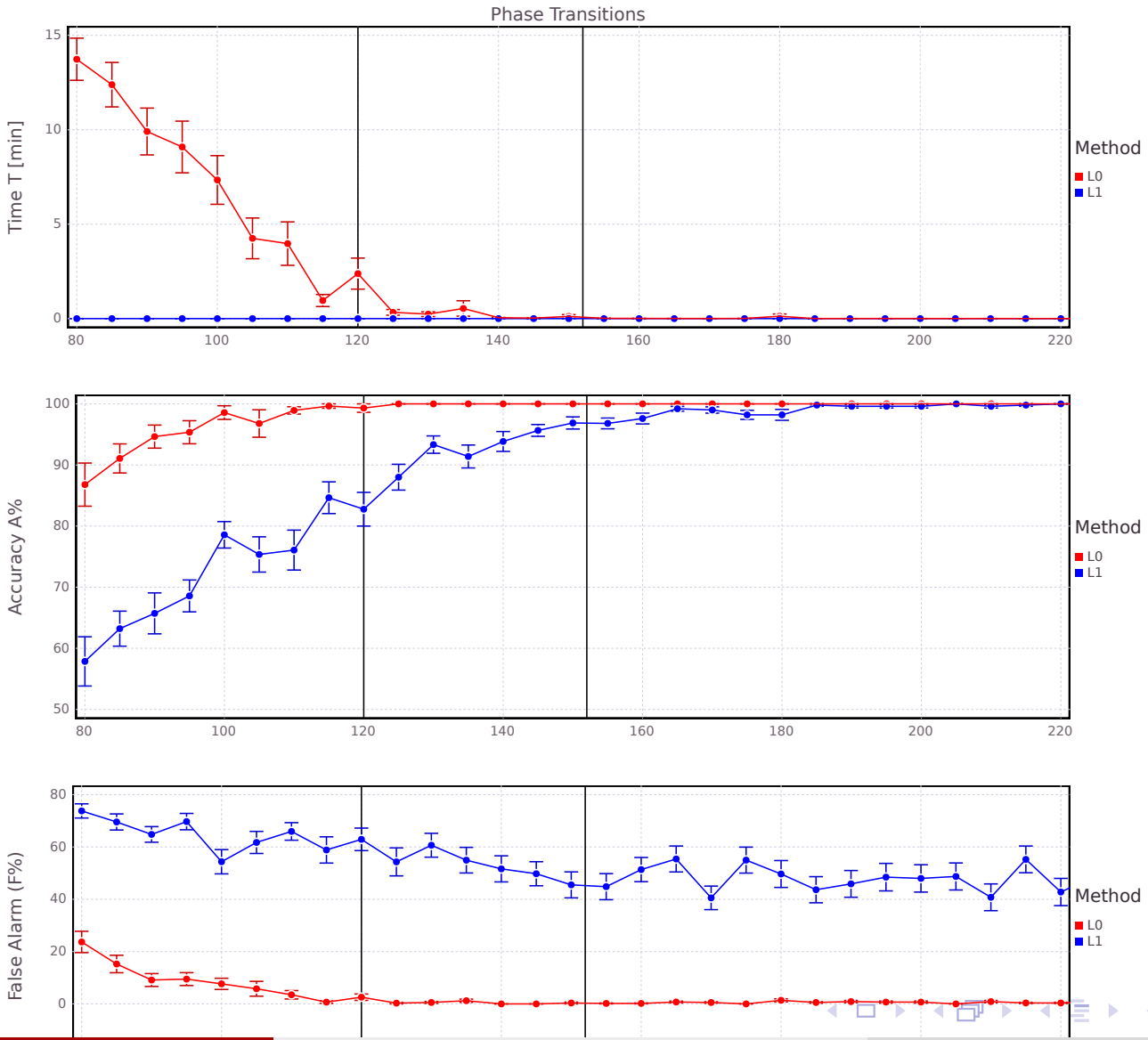
- Accuracy and false alarm rate of a certain solution $\beta^\star$

$$A\% := 100 \times \frac{|\text{supp}(\beta_{\text{true}}) \cap \text{supp}(\beta^\star)|}{k}$$

$$F\% := 100 \times \frac{|\text{supp}(\beta^\star) \setminus \text{supp}(\beta_{\text{true}})|}{|\text{supp}(\beta^\star)|}.$$

- Perfect support recovery occurs only then when $\beta^\star$ tells the whole truth ($A\% = 100$) and nothing but the truth ($F\% = 0$).

Phase Transitions



Remark on Complexity

- Traditional complexity theory suggests that the difficulty of a problem increases with dimension.
- Sparse regression problem has the property that for small number of samples n, the dual approach takes a large amount of time to solve the problem, but most importantly the optimal solution does not recover the true signal.
- However, for a large number of samples n, dual approach solves the problem extremely fast and recovers 100% of the support of the true regressor β_{true}.

Traditional Randomization Approach

- The randomization paradigm— Tuckey, 1968:
- From the given data, a random subset is placed to the side — test set.
- On the remaining data, we randomly split it into training and validation sets.
- After potentially several iterations of this process, the final accuracy is reported on the test set.
- The β coefficients of the regression are not stable often.

Diabetes Data Set

$n = 350$ patients and $p = 55$:

- 10 baseline variables x_i (age, sex, cholesterol levels, etc.)
- Second-order interactions $x_i \cdot x_j$ for $i < j$
- Predicting hemoglobin measure in one year

Diabetes Data Set

$n = 350$ patients and $p = 55$:

- 10 baseline variables x_i (age, sex, cholesterol levels, etc.)
- Second-order interactions $x_i \cdot x_j$ for $i < j$
- Predicting hemoglobin measure in one year

We use ordinary least squares on original data.
Randomly Select different training sets.
Linear coefficients become:

	Age	Sex	LDL	HDL	$\cdots$
Original data	0.05	-0.20	2.91	-2.75	$\cdots$

Diabetes Data Set

$n = 350$ patients and $p = 55$:

- 10 baseline variables x_i (age, sex, cholesterol levels, etc.)
- Second-order interactions $x_i \cdot x_j$ for $i < j$
- Predicting hemoglobin measure in one year

We use ordinary least squares on original data.
Randomly Select different training sets.
Linear coefficients become:

	Age	Sex	LDL	HDL	$\cdots$
Original data	0.05	-0.20	2.91	-2.75	$\cdots$
Perturbed data	0.05	-0.20	-2.62	2.18	$\cdots$

Stable Regression

- How do you train for exams?

Stable Regression

- How do you train for exams?

$$\min_{\beta} \max_{z \in \mathcal{Z}} \sum_{i=1}^{n} z_i |y_i - \beta^T \mathbf{x}_i| + \lambda \sum_{i=1}^{p} \Gamma(\beta_i)$$

$$\text{with} \quad \mathcal{Z} = \left\{ \mathbf{z} \in \{0,1\}^n : \sum_{i=1}^{n} z_i = k \right\},$$

- $\Gamma(\cdot)$ can be Lasso or elastic net or ridge reguralization.

Stable Regression continued

- At an optimal solution each z_i will be equal to either 0 or 1, with the interpretation that if $z_i = 1$, then point $(\mathbf{x}_i, y_i)$ is assigned to the training set, otherwise, it it is assigned to the validation set.
- k indicates the desired proportion between the size of the training and validations sets.
- Problem equivalent to optimizing over the convex hull of $\mathcal{Z}$

$$\text{conv}(\mathcal{Z}) = \left\{ \mathbf{z} : \sum_{i=1}^{n} z_i = k,\ 0 \leq z_i \leq 1,\ \forall i \in [n] \right\}.$$

- Problem is equivalent to

$$\min_{\beta} \max_{z \in \text{conv}(\mathcal{Z})} \sum_{i=1}^{n} z_i |y_i - \beta^T \mathbf{x}_i| + \lambda \sum_{i=1}^{p} \Gamma(\beta_i).$$

An Efficient Algorithm

- Linear optimization dual of the inner maximization problem

$$\max_{\mathbf{z}} \sum_{i=1}^{n} z_i |y_i - \boldsymbol{\beta}^T \mathbf{x}_i| \quad \text{s.t.} \quad \sum_{i=1}^{n} z_i = k, \quad 0 \leq z_i \leq 1, \ \forall i \in [n]$$

- Dual

$$\min_{\theta, \mathbf{u}} \ k\theta + \sum_{i=1}^{n} u_i \quad \text{s.t.} \quad \theta + u_i \geq |y_i - \boldsymbol{\beta}^T \mathbf{x}_i|, \quad u_i \geq 0, \ \forall i \in [n].$$

- Substituting this minimization problem back into the outer minimization:

$$\min_{\boldsymbol{\beta}, \theta, \mathbf{u}} \ k\theta + \sum_{i=1}^{n} u_i + \lambda \sum_{i=1}^{p} \Gamma(\beta_i)$$

$$\text{s.t. } \theta + u_i \geq y_i - \boldsymbol{\beta}^T \mathbf{x}_i, \quad \theta + u_i \geq -(y_i - \boldsymbol{\beta}^T \mathbf{x}_i), \quad u_i \geq 0, \ \forall i \in [n].$$

MSE for randomized and optimization approaches for Lasso regression

Datasets			Randomization			Optimization		
Name	n	p	50/50	60/40	70/30	50/50	60/40	70/30
Abalone	4177	8	5.33	5.40	5.40	5.17	5.27	5.32
Auto MPG	392	7	12.72	12.65	12.62	12.04	12.15	12.42
Comp Hard	209	6	6889.83	6907.53	7194.27	6433.21	6571.57	7069.26
Concrete	103	7	77.62	74.43	70.90	62.14	64.78	65.20
Ecoli	336	7	1.66	1.62	1.63	1.60	1.58	1.59
Forest Fi.	517	12	3927.89	4124.78	3974.49	3886.07	4101.03	3962.84
Glass	214	9	1.35	1.36	1.28	1.32	1.35	1.28
Housing	506	13	28.24	27.23	28.05	27.20	26.52	27.58
Space Sh.	23	4	0.50	0.46	0.41	0.34	0.41	0.37
WPBC	683	10	0.60	0.58	0.61	0.54	0.54	0.58

Prediction standard deviation for randomized and optimization approaches for Lasso regression

Datasets			Randomization			Optimization		
Name	n	p	50/50	60/40	70/30	50/50	60/40	70/30
Abalone	4177	8	0.75	0.77	0.74	0.67	0.70	0.69
Auto MPG	392	7	4.11	4.07	4.24	3.77	3.86	4.15
Comp Hard	209	6	9464.71	9628.97	9689.87	9401.39	9643.33	9837.88
Concrete	103	7	37.65	35.41	32.60	22.82	23.96	26.65
Ecoli	336	7	0.48	0.46	0.45	0.45	0.44	0.43
Forest Fi.	517	12	7371.71	7591.80	7401.56	7343.27	7576.11	7393.83
Glass	214	9	0.68	0.68	0.63	0.60	0.63	0.62
Housing	506	13	13.57	13.00	13.66	13.35	12.85	13.42
Space Sh.	23	4	0.71	0.52	0.51	0.44	0.52	0.48
WPBC	683	10	0.71	0.68	0.74	0.60	0.63	0.69

Coefficients standard deviation for randomized and optimization approaches for Lasso regression

Datasets			Randomization			Optimization		
Name	n	p	50/50	60/40	70/30	50/50	60/40	70/30
Abalone	4177	8	0.066	0.077	0.075	0.070	0.071	0.073
Auto MPG	392	7	0.004	0.003	0.003	0.003	0.003	0.004
Comp Hard	209	6	0.005	0.007	0.008	0.005	0.005	0.004
Concrete	103	7	0.002	0.001	0.001	0.001	0.001	0.001
Ecoli	336	7	0.030	0.029	0.028	0.028	0.027	0.028
Forest Fi.	517	12	0.003	0.005	0.004	0.001	0.002	0.003
Glass	214	9	0.003	0.003	0.003	0.003	0.025	0.003
Housing	506	13	0.014	0.014	0.013	0.015	0.014	0.015
Space Sh.	23	4	0.002	0.001	0.001	0.001	0.001	0.001
WPBC	683	10	0.000	0.000	0.000	0.000	0.000	0.000

Support Recovery

- Stable support recovery

$$\min_{\beta} \max_{z} \sum_{i=1}^{n} z_i |y_i - \beta^T \mathbf{x}_i| + \lambda \sum_{i=1}^{p} \Gamma(\beta_i)$$

$$\text{s.t.} \sum_{i=1}^{n} z_i = k, \quad \sum_{i=1}^{p} \delta_i = s,$$

$$|\beta_i| \leq M\delta_i, \ \delta_i \in \{0,1\}, \ \forall i \in [p], \quad 0 \leq z_i \leq 1, \ \forall i \in [n].$$

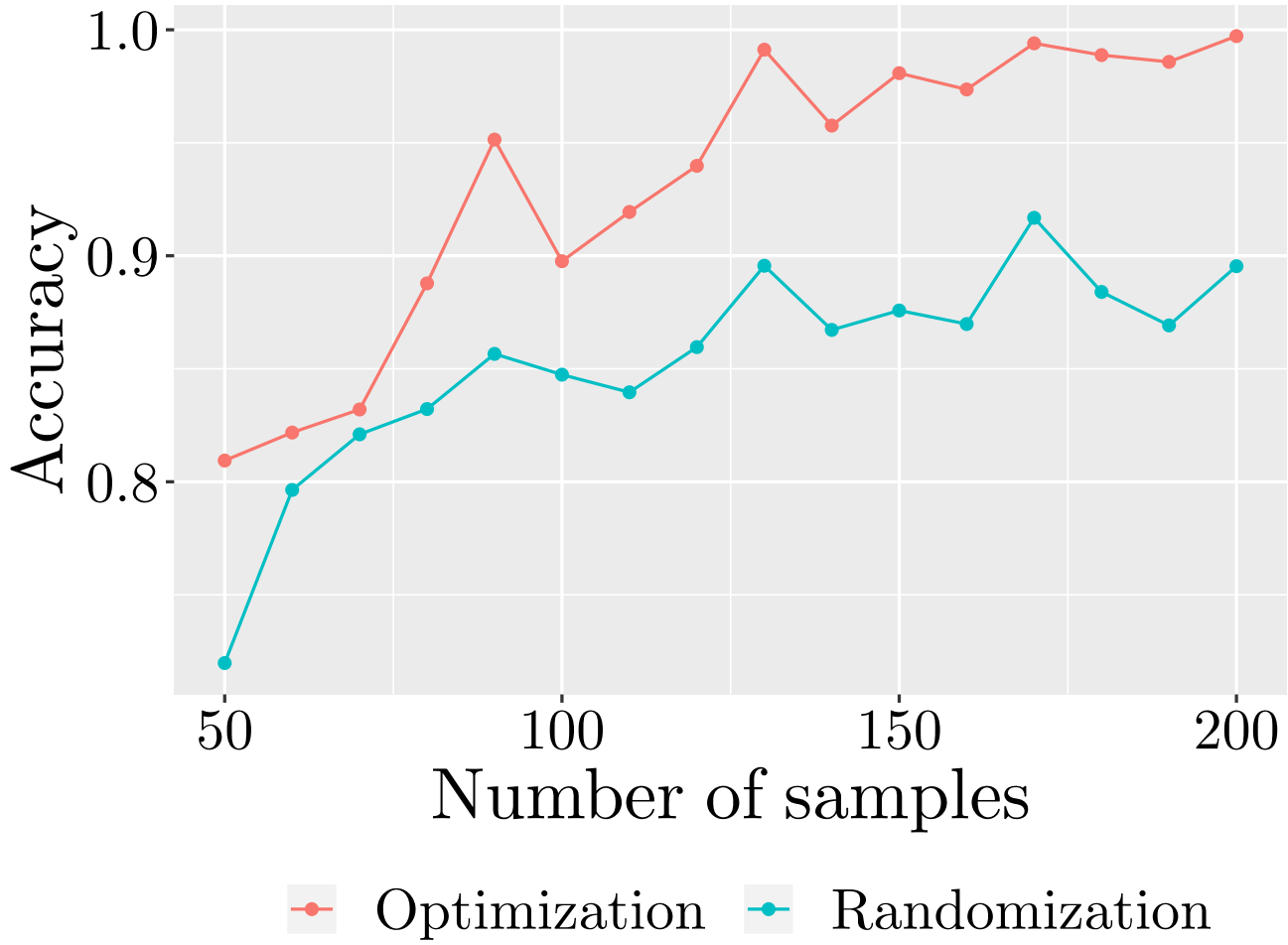
- Randomization to train:

$$\min_{\beta} \sum_{i \in A_{train}} |y_i - \beta^T \mathbf{x}_i| + \lambda \sum_{i=1}^{p} \Gamma(\beta_i)$$

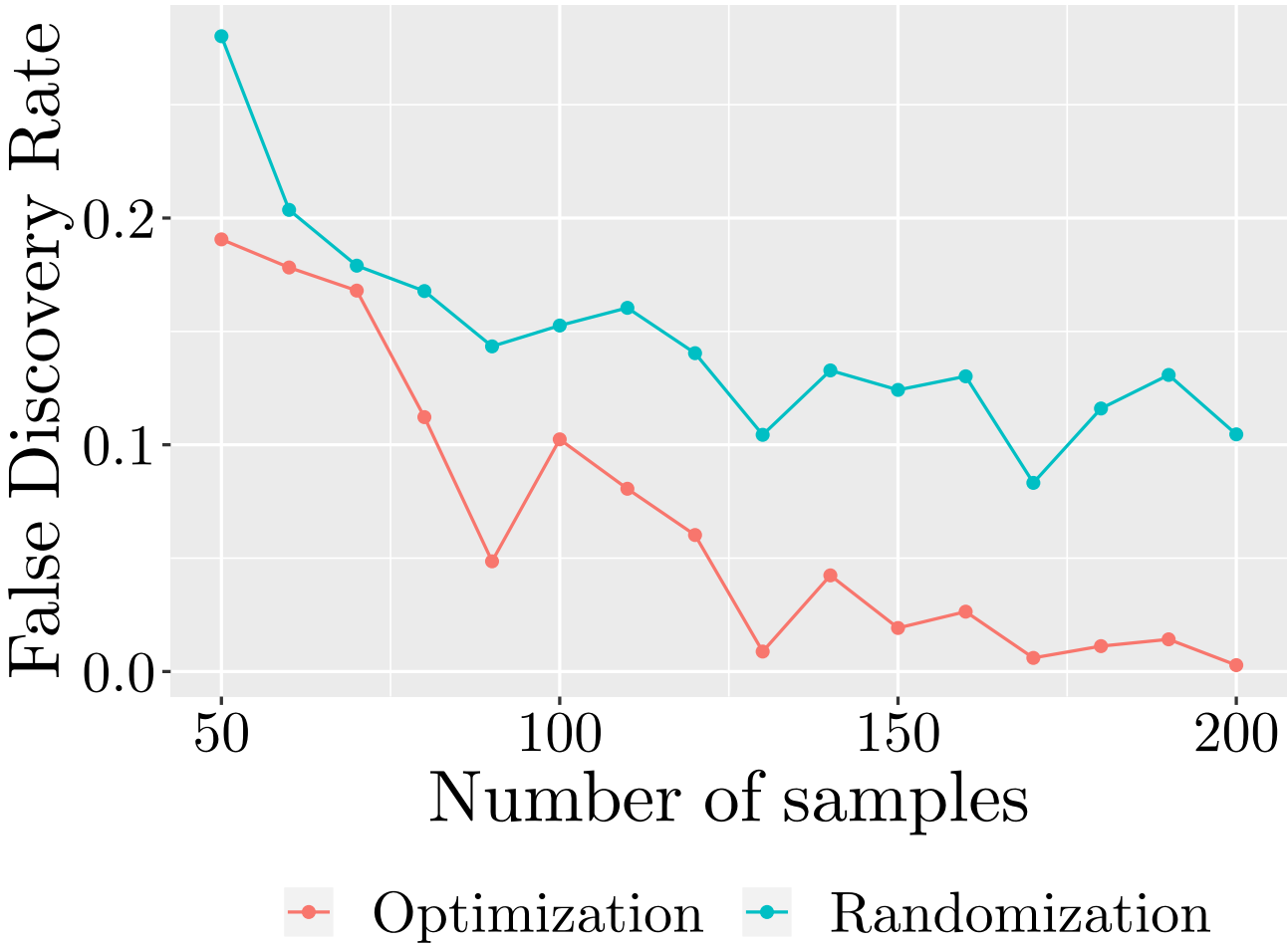
$$\text{s.t.} \sum_{i=1}^{p} \delta_i = s, \ |\beta_i| \leq M\delta_i, \ \delta_i \in \{0,1\}, \ \forall i \in [p],$$

$$0 \leq z_i \leq 1, \ \forall i \in [n].$$

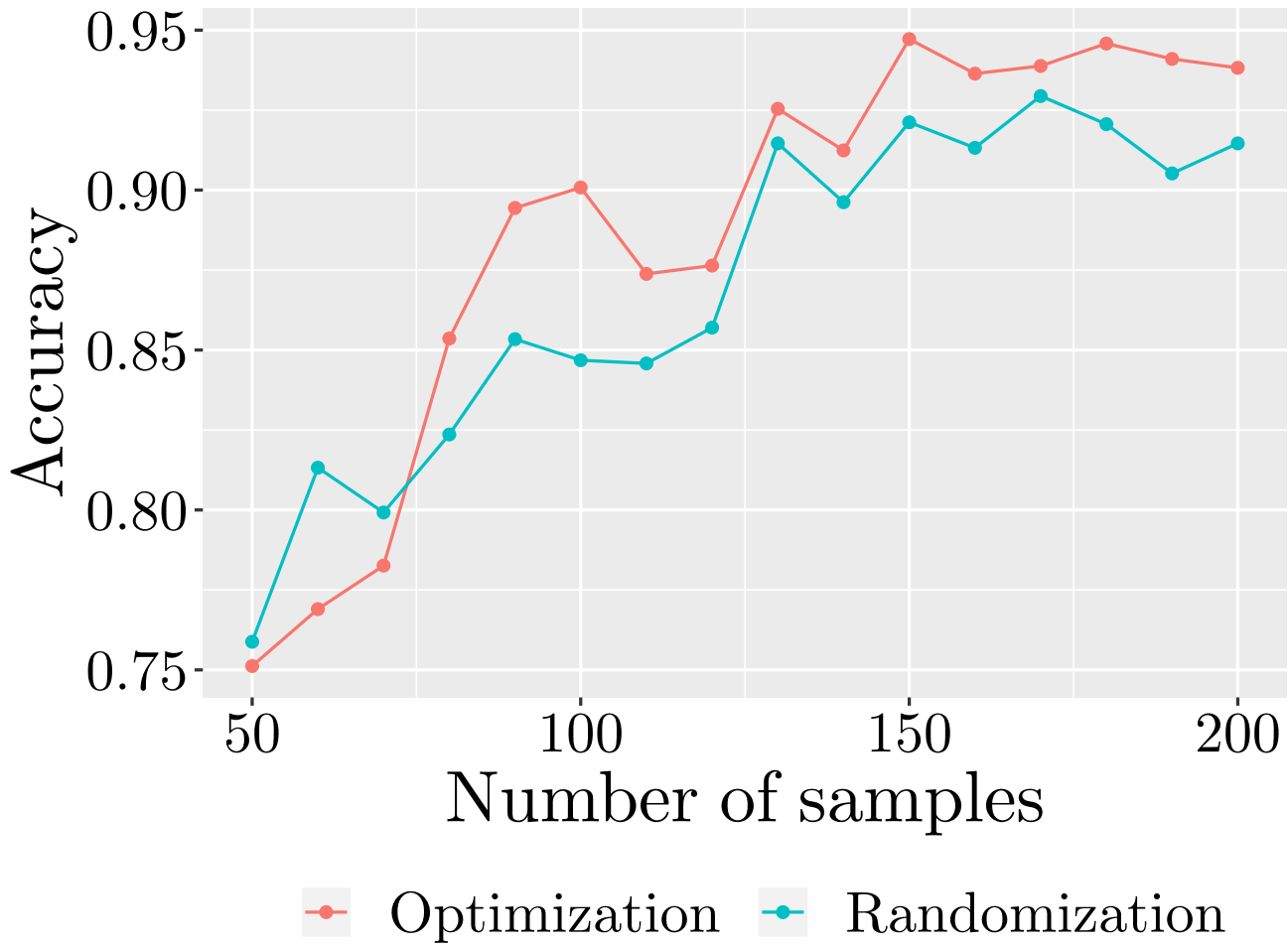
Support recovery accuracy rate for randomized and optimization–known support



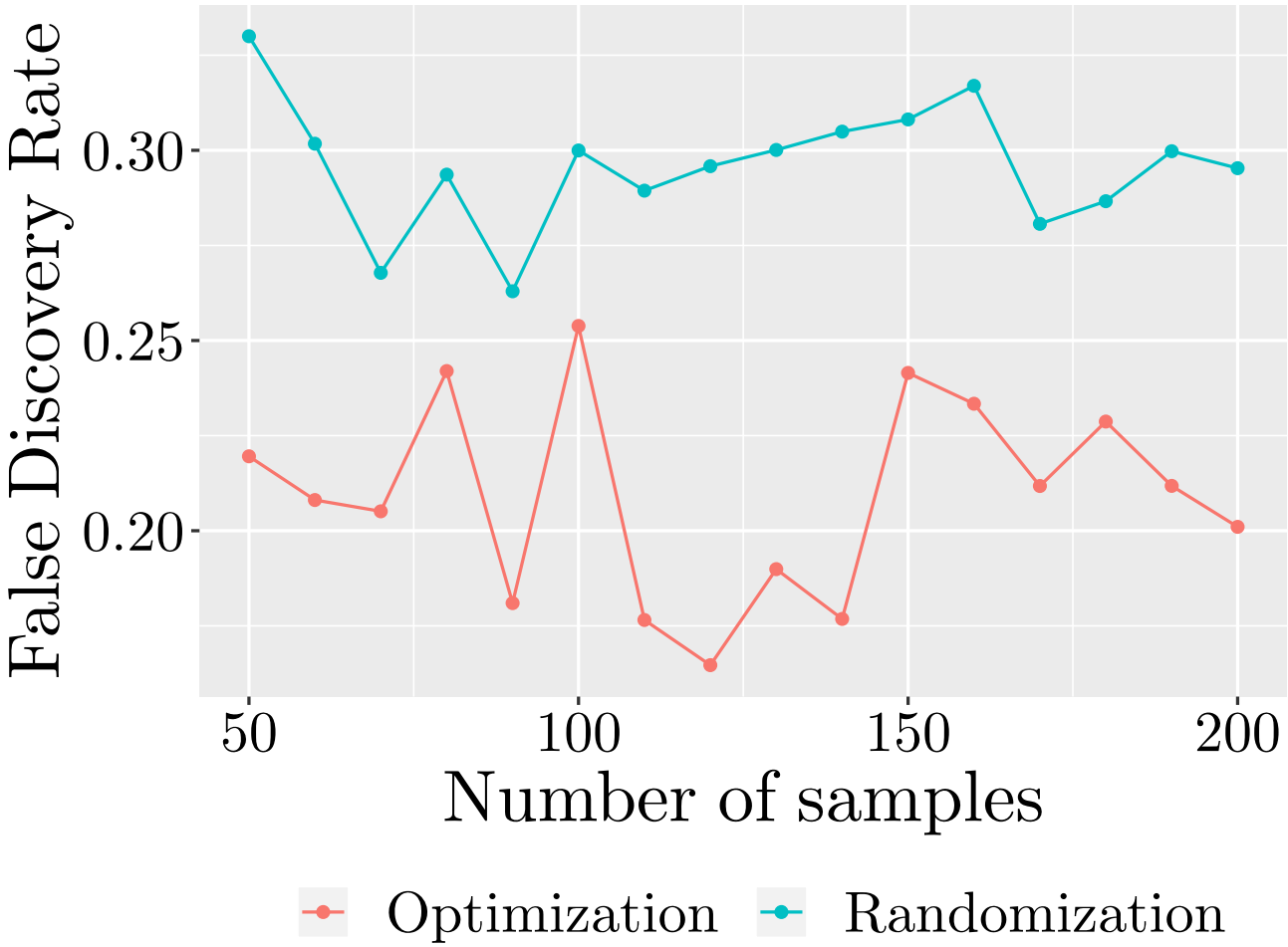
Support recovery false discovery rate for randomized and optimization–known support



Support recovery accuracy rate for randomized and optimization–unknown support



Support recovery false discovery rate for randomized and optimization–unknown support



Extensions of Sparsity

- Sparse Classification
- Matrix completion with and without side information
- Tensor Completion
- Sparse Inverse Covariance estimation
- Factor Analysis
- Sparse PCA

Extensions of Randomization vs Optimization

- Optimal Design of Experiments
- Identifying Exceptional Responders
- Bootstrap
- Stable Trees

Summary

- We can solve sparse regression problems with $n = 100,000$s and $p = 100,000$ to provable optimality in minutes.
- MIO solutions have a significant edge in detecting sparsity, and outperform Lasso in prediction accuracy.
- Optimization has a significant Edge over Randomization.
- Need to reconsider deeply rooted beliefs on complexity and need for relaxations (Lasso) to solve sparse regression problems.
- New book and a class.