

# Optimization Along the Path

**Jacob Bien**

Department of Data Sciences and Operations

March 20, 2019

Operator Splitting Methods in Data Analysis Workshop  
Flatiron Institute

# Talk Outline

1 Statistics Background

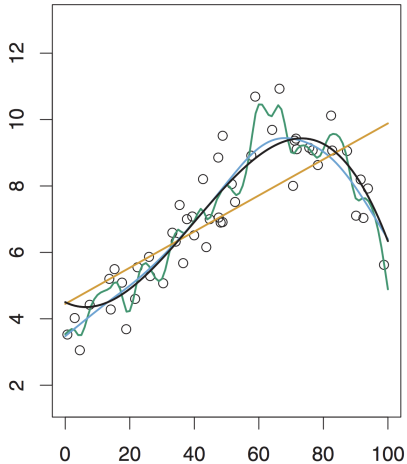
2 Approaches That Exploit the Path

# Statistical Inference in a Slide

Underlying Model  $\rightarrow$  Observed Data  $\rightarrow$  Computed Estimator

**Goal:** Estimator takes data and tries to infer the underlying model.

# Bias-Variance Tradeoff



This figure from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

# Regularization: Navigating the Tradeoff

Define estimator  $\hat{\beta}$  as a solution to

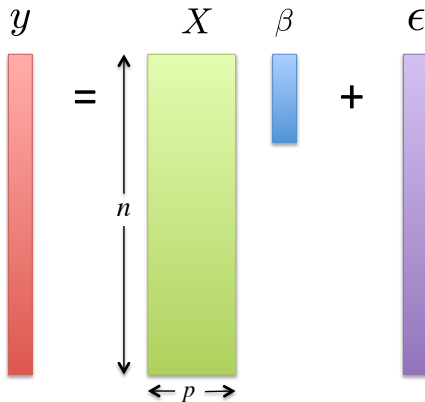
$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \underbrace{f(\beta)}_{\text{data fitting term}} + \underbrace{\lambda \cdot \Omega(\beta)}_{\text{regularization term}}$$

- small  $f$  means  $\beta$  describes the observed data well
- $\Omega$  controls the complexity of the estimator (problem-specific structure)
- $\lambda > 0$  controls tradeoff

# The Linear Model

## Classical setting

- Observe  $n$  observations  $(y_i, x_i)$ , where  $x_i \in \mathbb{R}^p$ .
- $n > p$ .

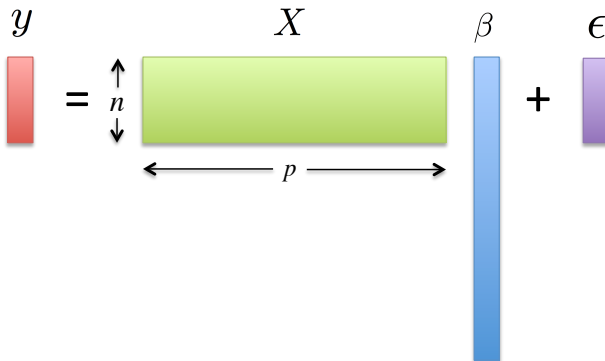


# The Linear Model

## High-dimensional setting

- Observe  $n$  observations  $(y_i, x_i)$ , where  $x_i \in \mathbb{R}^p$ .
- $p \gg n$

*...an intractable situation?*



# The Linear Model

High-dimensional setting

What if true model is actually simpler than it seems?

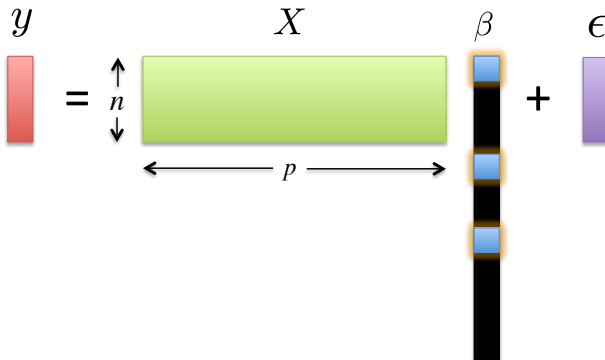
The diagram illustrates the linear model equation  $y = X\beta + \epsilon$ . On the left, a red vertical rectangle represents the response vector  $y$ . To its right is an equals sign. Next is a green rectangle representing the design matrix  $X$ . A vertical double-headed arrow to the left of  $X$  is labeled  $n$ , indicating its height. A horizontal double-headed arrow below  $X$  is labeled  $p$ , indicating its width. To the right of  $X$  is a blue vertical rectangle representing the coefficient vector  $\beta$ . To the right of  $\beta$  is a plus sign. Finally, on the far right, a purple vertical rectangle represents the error term  $\epsilon$ .

...actually a low-dimensional model

# The Linear Model

High-dimensional setting

What if true model is actually simpler than it seems?



...actually a low-dimensional model

# The Lasso

The **Lasso** – an estimator defined by the optimization problem

[Tibshirani, 1996]

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2} \|y - X\beta\|^2 + \lambda \|\beta\|_1$$

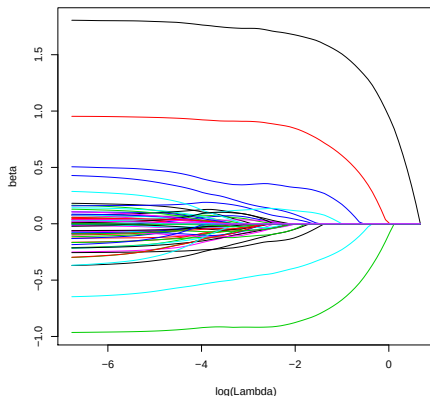
- Penalty  $\implies$  sparsity (feature selection)
- Fit versus parsimony tradeoff

## What value of $\lambda$ to use?

- In theoretical results, it depends on unknowns.
- Choice even depends on your *goal* (support recovery vs. prediction)

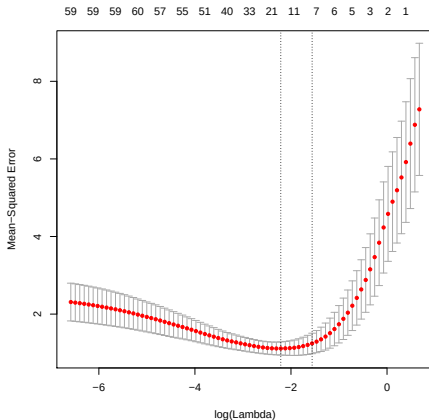
## In practice

- Solve problem for many  $\lambda$  values.
- Cross validation: choose a  $\lambda$  that makes good predictions on held out data.



## In practice

- Solve problem for many  $\lambda$  values.
- Cross validation: choose a  $\lambda$  that makes good predictions on held out data.



# The Solution Path

Statisticians interested in whole **solution path** parameterized by  $\lambda$ :

$$\mathcal{P} = \{\hat{\beta}_\lambda : \lambda > 0\}$$

where  $\hat{\beta}_\lambda$  is the unique<sup>1</sup> solution to

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} f(\beta) + \lambda \cdot \Omega(\beta).$$

or will settle for a finite subset  $\Lambda$ :

$$\{\hat{\beta}_\lambda : \lambda \in \Lambda\} \subset \mathcal{P}$$

---

<sup>1</sup>Although not generally true, holds in many realistic problem instances.

# Examples

Often care about “when” variables enter the model...

$$\hat{\lambda}_j := \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}.$$

- Ask whether after  $k$ th variable added, have all “true” variables been identified, e.g. Lockhart et al. [2014], G’Sell et al. [2016], Fithian et al. [2015].
- Use  $\hat{\lambda}_j$  as a test statistic, e.g., Bien et al. [2015]
- Compare the  $\hat{\lambda}_j$  for real variables as opposed to “knockoff” versions [Barber et al., 2015].
- “False Discoveries Occur Early on the Lasso Path” [Su et al., 2017]

# Examples

Often care about “when” variables enter the model...

$$\hat{\lambda}_j := \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}.$$

- Ask whether after  $k$ th variable added, have all “true” variables been identified, e.g. Lockhart et al. [2014], G’Sell et al. [2016], Fithian et al. [2015].
- Use  $\hat{\lambda}_j$  as a test statistic, e.g., Bien et al. [2015]
- Compare the  $\hat{\lambda}_j$  for real variables as opposed to “knockoff” versions [Barber et al., 2015].
- “False Discoveries Occur Early on the Lasso Path” [Su et al., 2017]

# Examples

Often care about “when” variables enter the model...

$$\hat{\lambda}_j := \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}.$$

- Ask whether after  $k$ th variable added, have all “true” variables been identified, e.g. Lockhart et al. [2014], G’Sell et al. [2016], Fithian et al. [2015].
- Use  $\hat{\lambda}_j$  as a test statistic, e.g., Bien et al. [2015]
- Compare the  $\hat{\lambda}_j$  for real variables as opposed to “knockoff” versions [Barber et al., 2015].
- “False Discoveries Occur Early on the Lasso Path” [Su et al., 2017]

# Examples

Often care about “when” variables enter the model...

$$\hat{\lambda}_j := \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}.$$

- Ask whether after  $k$ th variable added, have all “true” variables been identified, e.g. Lockhart et al. [2014], G’Sell et al. [2016], Fithian et al. [2015].
- Use  $\hat{\lambda}_j$  as a test statistic, e.g., Bien et al. [2015]
- Compare the  $\hat{\lambda}_j$  for real variables as opposed to “knockoff” versions [Barber et al., 2015].
- “False Discoveries Occur Early on the Lasso Path” [Su et al., 2017]

# Examples

Often care about “when” variables enter the model...

$$\hat{\lambda}_j := \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}.$$

- Ask whether after  $k$ th variable added, have all “true” variables been identified, e.g. Lockhart et al. [2014], G’Sell et al. [2016], Fithian et al. [2015].
- Use  $\hat{\lambda}_j$  as a test statistic, e.g., Bien et al. [2015]
- Compare the  $\hat{\lambda}_j$  for real variables as opposed to “knockoff” versions [Barber et al., 2015].
- “False Discoveries Occur Early on the Lasso Path” [Su et al., 2017]

# The General Problem

If our goal is to solve a *sequence* of convex problems, how does this affect our approach to optimization?

Is it necessarily true that the best algorithm for getting  $\hat{\beta}_\lambda$  will also be the best for getting  $\{\hat{\beta}_\lambda : \lambda \in \Lambda\}$ ?

**Example:** Do some methods make better use of warm starts than others?

# The General Problem

If our goal is to solve a *sequence* of convex problems, how does this affect our approach to optimization?

Is it necessarily true that the best algorithm for getting  $\hat{\beta}_\lambda$  will also be the best for getting  $\{\hat{\beta}_\lambda : \lambda \in \Lambda\}$ ?

**Example:** Do some methods make better use of warm starts than others?

Would we use different algorithms if instead of focusing on

$$\|\hat{\beta}_\lambda^t - \hat{\beta}_\lambda\| \leq g_\lambda(t),$$

our goal were

$$\sup_{\lambda \in \Lambda} \|\hat{\beta}_\lambda^t - \hat{\beta}_\lambda\| \leq \tilde{g}_\Lambda(t)?$$

For finite  $\Lambda$ , in a total of  $t \cdot |\Lambda|$  steps, we could easily get

$$\sup_{\lambda \in \Lambda} \|\hat{\beta}_\lambda^t - \hat{\beta}_\lambda\| \leq \sup_{\lambda \in \Lambda} g_\lambda(t)$$

but can we get

$$\tilde{g}_\Lambda(t) \leq \sup_{\lambda \in \Lambda} g_\lambda(t)?$$

(Warm starts would be an example since they make  $g_\lambda(t)$  smaller.)

Would we use different algorithms if instead of focusing on

$$\|\hat{\beta}_\lambda^t - \hat{\beta}_\lambda\| \leq g_\lambda(t),$$

our goal were

$$\sup_{\lambda \in \Lambda} \|\hat{\beta}_\lambda^t - \hat{\beta}_\lambda\| \leq \tilde{g}_\Lambda(t)?$$

For finite  $\Lambda$ , in a total of  $t \cdot |\Lambda|$  steps, we could easily get

$$\sup_{\lambda \in \Lambda} \|\hat{\beta}_\lambda^t - \hat{\beta}_\lambda\| \leq \sup_{\lambda \in \Lambda} g_\lambda(t)$$

but can we get

$$\tilde{g}_\Lambda(t) \leq \sup_{\lambda \in \Lambda} g_\lambda(t)?$$

(Warm starts would be an example since they make  $g_\lambda(t)$  smaller.)

Would we use different algorithms if instead of focusing on

$$\|\hat{\beta}_\lambda^t - \hat{\beta}_\lambda\| \leq g_\lambda(t),$$

our goal were

$$\sup_{\lambda \in \Lambda} \|\hat{\beta}_\lambda^t - \hat{\beta}_\lambda\| \leq \tilde{g}_\Lambda(t)?$$

For finite  $\Lambda$ , in a total of  $t \cdot |\Lambda|$  steps, we could easily get

$$\sup_{\lambda \in \Lambda} \|\hat{\beta}_\lambda^t - \hat{\beta}_\lambda\| \leq \sup_{\lambda \in \Lambda} g_\lambda(t)$$

but can we get

$$\tilde{g}_\Lambda(t) \leq \sup_{\lambda \in \Lambda} g_\lambda(t)?$$

(Warm starts would be an example since they make  $g_\lambda(t)$  smaller.)

# What's Typically Done by Statisticians...

(unless  $\Omega$  is particularly simple)

- ① Write optimization algorithm for single  $\lambda$  case using, say, ADMM.
- ② Solve along grid from largest  $\lambda$  to smallest, using warm starts.

# Talk Outline

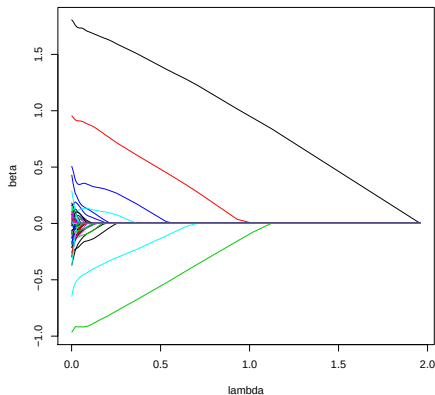
1 Statistics Background

2 Approaches That Exploit the Path

## Getting $\mathcal{P}$ Itself

For lasso, LARS algorithm (homotopy method) [Efron et al., 2004, Osborne et al., 2000]

- Exploits piecewise-linearity of lasso path.
- Solves for the “knots” in the path (worst case  $3^p$  knots)



Also developed for the generalized lasso [Tibshirani, 2011] and any problem in which  $f$  is piecewise quadratic and  $\Omega$  is piecewise linear [Rosset and Zhu, 2007].

glmnet for  $\{\hat{\beta}_\lambda : \lambda \in \Lambda\}$

- Coordinate descent
- Decreasing grid of  $\lambda$  values
- Warm starts
- **Key:** Active set (working set) - highly effective, exploiting sparsity

## Screening Rules

Suppose you have solved  $\hat{\beta}(\lambda')$  and are about to solve for  $\hat{\beta}(\lambda)$  ( $\lambda' > \lambda$ ). Can you cheaply determine that  $\hat{\beta}_j(\lambda) = 0$  before actually starting to solve the problem for that  $\lambda$ ?

“Sequential Dual Polytope Projection” [Wang et al., 2012]

$$|X_j^T(y - X\hat{\beta}(\lambda'))| \leq \lambda' - (\lambda'/\lambda - 1)\|X_j\|_2\|y\|_2 \implies \hat{\beta}_j(\lambda) = 0.$$

# “A General Framework for Fast Stagewise Algorithms”

Tibshirani [2015]

Focuses on approximate solution path of

$$\hat{\beta}(s) \in \arg \min_{\beta} f(\beta) \quad \text{s.t.} \quad \Omega(\beta) \leq s.$$

**Algorithm:** Start with  $\hat{\beta}^0 := \hat{\beta}(s_0)$ .

For  $k = 1, 2, \dots$ :

$$\hat{\beta}^k \leftarrow \hat{\beta}^{k-1} + \hat{\delta}$$

where

$$\hat{\delta} \in \arg \min_{\delta} \langle \nabla f(\hat{\beta}^{k-1}), \delta \rangle \quad \text{s.t.} \quad \Omega(\delta) \leq \epsilon.$$

# “A General Framework for Fast Stagewise Algorithms”

Tibshirani [2015]

Focuses on approximate solution path of

$$\hat{\beta}(s) \in \arg \min_{\beta} f(\beta) \quad \text{s.t.} \quad \Omega(\beta) \leq s.$$

**Algorithm:** Start with  $\hat{\beta}^0 := \hat{\beta}(s_0)$ .

For  $k = 1, 2, \dots$ :

$$\hat{\beta}^k \leftarrow \hat{\beta}^{k-1} + \hat{\delta}$$

where

$$\hat{\delta} \in \arg \min_{\delta} \langle \nabla f(\hat{\beta}^{k-1}), \delta \rangle \quad \text{s.t.} \quad \Omega(\delta) \leq \epsilon.$$

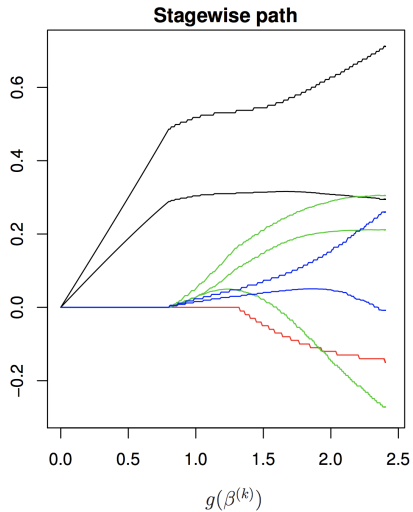
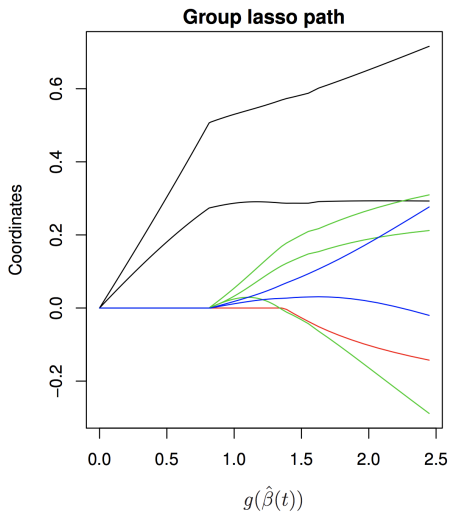


Figure from Tibshirani [2015]

Other Ideas?

# The Obvious: Proximal Gradient Descent + Warm Starts

Suppose  $\lambda_0 > \lambda_1 > \lambda_2 > \dots$ .

Initialize  $\hat{\beta}_{\lambda_0}^T \leftarrow 0$ .

For  $k = 1, \dots, |\Lambda|$ :

$$\hat{\beta}_{\lambda_k}^0 \leftarrow \hat{\beta}_{\lambda_{k-1}}^T$$

For  $t = 1, 2, \dots, T$ :

$$\hat{\beta}_{\lambda_k}^t \leftarrow \text{Prox}_{s\lambda_k\Omega}(\hat{\beta}_{\lambda_k}^{t-1} - s\nabla f(\hat{\beta}_{\lambda_k}^{t-1}))$$

## The Speculative: Interchange loops

For  $t = 1, 2, \dots, T$ :

For  $k = 1, \dots, |\Lambda|$ :

$$\hat{\gamma} \leftarrow \sum_{k' < k} w_{kk'} \hat{\beta}_{\lambda_{k'}}^t + \sum_{k' \geq k} w_{kk'} \hat{\beta}_{\lambda_{k'}}^{t-1}$$

$$\hat{\beta}_{\lambda_k}^t \leftarrow \text{Prox}_{s\lambda_k\Omega}(\hat{\gamma} - s\nabla f(\hat{\gamma}))$$

Did I leave time for discussion?

Hope so!

- Rina Foygel Barber, Emmanuel J Candès, et al. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.
- Jacob Bien, Noah Simon, and Robert Tibshirani. Convex hierarchical testing of interactions. *The Annals of Applied Statistics*, 9(1):27–42, 2015.
- B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.
- William Fithian, Jonathan Taylor, Robert Tibshirani, and Ryan Tibshirani. Selective sequential model selection. *arXiv preprint arXiv:1512.02565*, 2015.
- Max Grazier G'Sell, Stefan Wager, Alexandra Chouldechova, and Robert Tibshirani. Sequential selection procedures and false discovery rate control. *Journal of the royal statistical society: series B (statistical methodology)*, 78(2):423–444, 2016.
- Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.
- Michael R Osborne, Brett Presnell, and Berwin A Turlach. On the lasso and its dual. *Journal of Computational and Graphical statistics*, 9(2):319–337, 2000.
- Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, pages 1012–1030, 2007.
- Weijie Su, Małgorzata Bogdan, Emmanuel Candès, et al. False discoveries occur early on the lasso path. *The Annals of Statistics*, 45(5):2133–2150, 2017.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996.
- Ryan J Tibshirani. A general framework for fast stagewise algorithms. *The Journal of Machine Learning Research*, 16(1):2543–2588, 2015.
- Ryan Joseph Tibshirani. *The solution path of the generalized lasso*. Stanford University, 2011.
- Jie Wang, Binbin Lin, Pinghua Gong, Peter Wonka, and Jieping Ye. Lasso screening rules via dual polytope projection. *arXiv preprint arXiv:1211.3966*, 2012.