Introduction: Why Optimization?

Geoff Gordon & Ryan Tibshirani
Optimization 10-725 / 36-725

Where this course fits in

In many ML/statistics/engineering courses, you learn how to:

translate  into $\min f(x)$

Question/idea *Optimization problem*

In this course, you'll learn that $\min f(x)$ is not the end of the story, i.e., you'll learn

- Algorithms for solving $\min f(x)$, and how to choose between them
- How knowledge of algorithms for $\min f(x)$ can influence the choice of translation
- How knowledge of algorithms for $\min f(x)$ can help you understand things about the problem

Optimization in statistics

A huge number of statistics problems can be cast as optimization problems, e.g.,

- Regression
- Classification
- Maximum likelihood

But a lot of problems cannot, and are based directly on algorithms or procedures, e.g.,

- Clustering
- Correlation analysis
- Model assessment

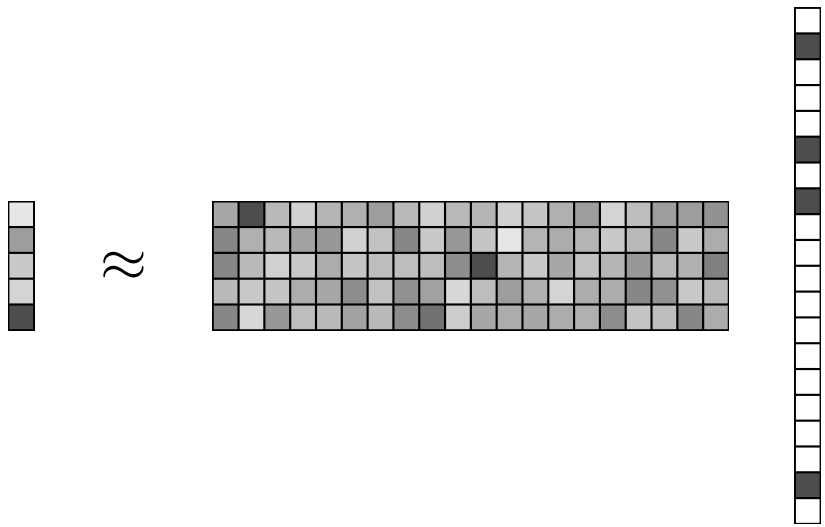
Not to say one camp is better than the other ... but if you *can* cast something as an optimization problem, it is often worthwhile

Sparse linear regression

Given response $y \in \mathbb{R}^n$ and predictors $A = (A_1, \ldots A_p) \in \mathbb{R}^{n \times p}$.
We consider the model

$$y \approx Ax$$

But $n \ll p$, and we think many of the variables $A_1, \ldots A_p$ could be unimportant. I.e., we want many components of x to be zero



E.g., size of tumor $\approx$ linear combination of genetic information, but not all gene expression measurements are relevant

Three methods

Solving the usual linear regression problem

$$\min_{x \in \mathbb{R}^n} \|y - Ax\|^2$$

would return a dense x (and not well-defined if $p > n$).

We want a sparse x. How? Three methods:

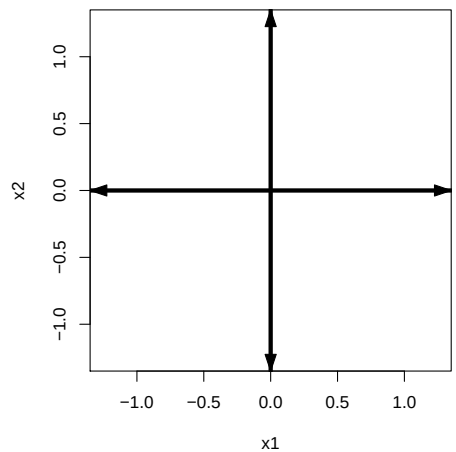
- **Best subset selection** – nonconvex optimization problem
- **Forward stepwise regression** – algorithm
- **Lasso** – convex optimization problem

Best subset selection

Natural idea, we solve

$$\min_{x \in \mathbb{R}^p} \|y - Ax\|^2 \text{ subject to } \|x\|_0 \leq k$$

where $\|x\|_0 =$ number of nonzero components in x, nonconvex "norm"



$$\{x \in \mathbb{R}^2 : \|x\|_0 \leq 1\}$$

- Problem is NP-hard
- In practice, solution cannot be computed for $p \gtrsim 40$
- Very little is known about properties of solution

Forward stepwize regression

Also natural idea: start with $x = 0$, then

- Find variable j such that $|A_j^T y|$ is largest (note: if variables have been centered and scaled, then $A_j^T y = \mathrm{cor}(A_j, y)$)
- Update x_j by regressing y onto A_j, i.e., solve

$$\min_{x_j \in \mathbb{R}} \|y - A_j x_j\|^2$$

- Now find variable $k \neq j$ such that $|A_k^T r|$ is largest, where $r = y - A_j x_j$ (i.e., $|\mathrm{cor}(A_k, r)|$ is largest)
- Update x_j, x_k by regressing y onto A_j, A_k
- Repeat

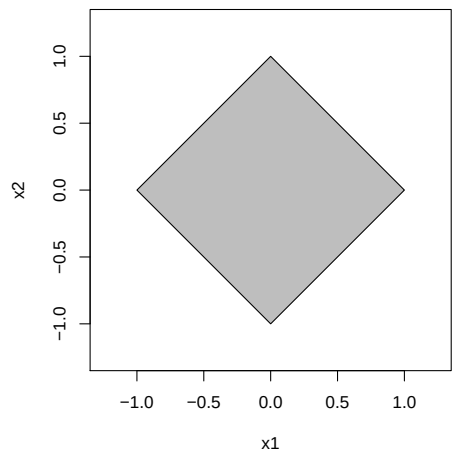
Some properties of this estimate are known, but not many; proofs are (relatively) complicated

Lasso

We solve

$$\min_{x \in \mathbb{R}^p} \|y - Ax\|^2 \text{ subject to } \|x\|_1 \leq t$$

where $\|x\|_1 = \sum_{i=1}^p |x_i|$, a convex norm



$$\{x \in \mathbb{R}^2 : \|x\|_1 \leq 1\}$$

- Delivers exact zeros in solution – lower t, more zeros
- Problem is convex and readily solved
- Many properties are known about the solution

Comparison

	# of Google Scholar hits	# of algorithms	Properties known
Best subset selection	2274	1 (brute force)	Little
Forward stepwise regression	7207	1 (itself)	Some
Lasso	13,100[1]	≥ 10	Lots

[1] I searched for 'lasso + statistics' because 'lasso' resulted in nearly 8 times as many hits. I also tried to be fair, and search for best subset selection and forward stepwise regression under their alternative names. On August 27, 2010.