Case Study 3: fMRI Prediction

LASSO Review,
Fused LASSO,
Parallel LASSO Solvers

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
April 28th, 2015

LASSO Regression

- **LASSO:** least absolute shrinkage and selection operator
- New objective:

$$\min_{\beta} \underbrace{\sum_{i=1}^{N} (y^i - (\beta_0 + \beta^T x^i))^2}_{RSS(\beta)} + \lambda \|\beta\|_1$$

L_1 penalty

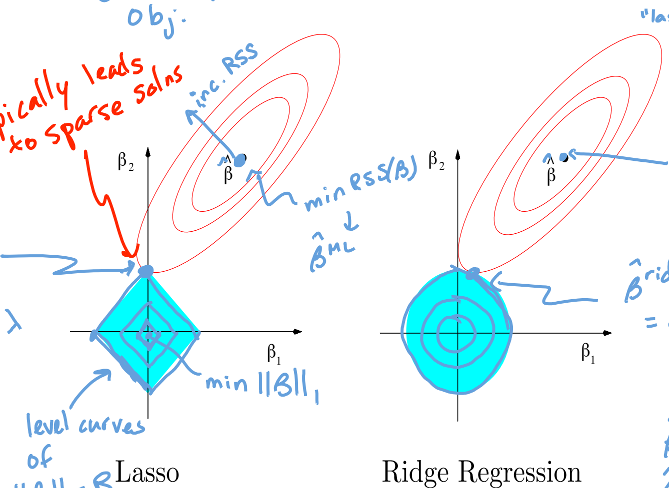
$$\Updownarrow$$

$$\min_{\beta} RSS(\beta) \quad s.t. \quad \|\beta\|_1 \leq B$$

Geometric Intuition for Sparsity

Overall obj.: $F(\beta) = RSS(\beta) + \lambda \|\beta\|$ ← 1 or 2 norm
"lasso" "ridge"

typically leads to sparse solns



$\hat{\beta}_\lambda^{lasso}$ = min $F(\beta)$ for some λ

$\hat{\beta}_1 = 0$
$\hat{\beta}_2$ = sthg.

$\hat{\beta}^{ridge}$ = min $F(\beta)$ for some λ
⇓
$\hat{\beta}_1$ = sthg
$\hat{\beta}_2$ = sthg

Soft Threshholding

If we hold $\beta_1^*, \dots, \beta_{j-1}^*, \beta_{j+1}^*, \dots, \beta_p^*$ fixed then this must hold:

strength of penalty

$c_j \propto corr(\underline{x}_j, r_{-j})$
all examples of feature j covariate
residual from model w/o using jth covariate

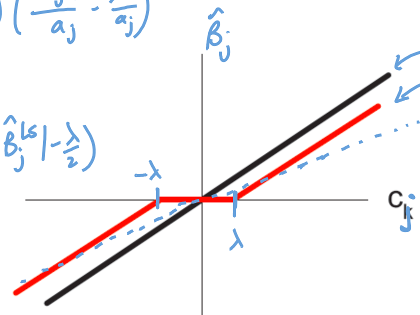
$$\hat{\beta}_j = \begin{cases} (c_j + \lambda)/a_j & c_j < -\lambda \\ 0 & c_j \in [-\lambda, \lambda] \\ (c_j - \lambda)/a_j & c_j > \lambda \end{cases}$$

$$= \text{sign}\left(\frac{c_j}{a_j}\right)\left(\frac{|c_j|}{a_j} - \frac{\lambda}{a_j}\right)$$

If $X^T X = I$

$\hat{\beta}_j^{lasso} = \text{Sign}(\hat{\beta}_j^{LS})\left(|\hat{\beta}_j^{LS}| - \frac{\lambda}{2}\right)$

$\hat{\beta}_j^{ridge} = \frac{\hat{\beta}^{LS}}{1+\lambda}$

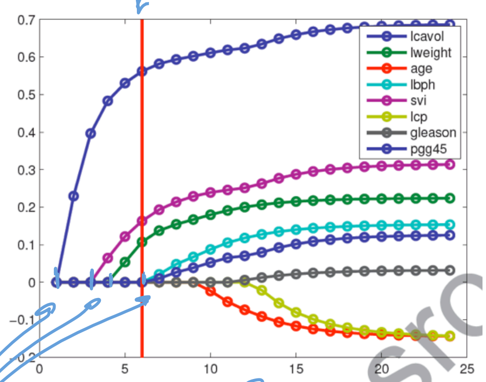


From
Kevin Murphy
textbook

In LASSO, all coeff. are shrunk relative to $\hat{\beta}^{LS}$

LASSO Coefficient Path

CV soln



From
Kevin Murphy
textbook

$\|\beta\|_1 \le B$

B

inc. λ

only a few critical points where coef. enter soln

©Emily Fox 2015 5

LASSO Example

CV soln

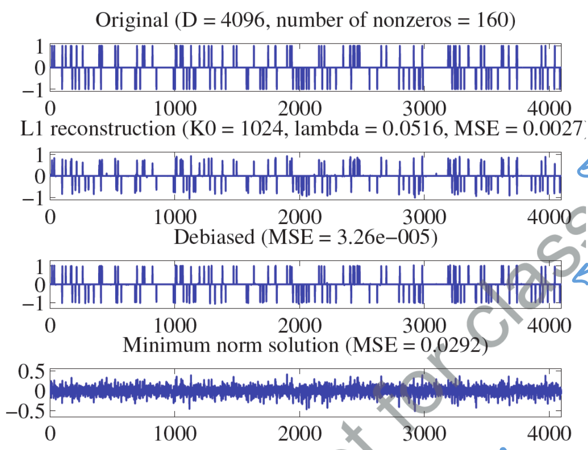
	Term	Least Squares	Ridge	Lasso
$\hat\beta_0$	Intercept	2.465	2.452	2.468
$\hat\beta_1$	lcavol	0.680	0.420	0.533
⋮	lweight	0.263	0.238	0.169
	age	−0.141	−0.046	
	lbph	0.210	0.162	0.002
	svi	0.305	0.227	0.094
	lcp	−0.288	0.000	
	gleason	−0.021	0.040	
$\hat\beta_p$	pgg45	0.267	0.133	

not in the model

shrunk, but non-zero

©Emily Fox 2015 6

Debiasing



From Kevin Murphy textbook

all coeff. shrink
→ bias

some people:

1. Use lasso to find support

2. Run LS regression just w/ selected cov.

⇒ remove bias for this reduced model

LS est.

Sparsistency

- Typical Statistical Consistency Analysis:
  – Holding model size (p) fixed, as number of samples (N) goes to infinity, estimated parameter goes to true parameter

est. param $\hat{\theta} \longrightarrow \theta^*$ true param ?

- Here we want to examine $p >> N$ domains
- Let both model size p and sample size N go to infinity!
  – Hard case: $\boxed{N = k \log p}$

N grows slowly relative to p

Sparsistency

- Rescale LASSO objective by *N*:

$$\min_{\beta} \frac{1}{N} RSS(\beta) + \lambda_N \sum_j |\beta_j|$$

- Theorem (Wainwright 2008, Zhao and Yu 2006, …):
  – Under some constraints on the design matrix *X*, if we solve the LASSO regression using

$$\lambda_N > \frac{2}{\gamma}\sqrt{\frac{2\sigma^2 \log p}{N}}$$

Then for some c_1>0, the following holds with at least probability

$$1 - 4\exp(-c_1 N \lambda_N^2) \longrightarrow 1$$

- The LASSO problem has a ~~unique solution~~ with support contained within the true support $S(\hat{\beta}) \subseteq S(\beta^*)$
- If $\min_{j \in S(\beta^*)} |\beta_j^*| > c_2 \lambda_N$ for some c_2>0, then $\boxed{S(\hat{\beta}) = S(\beta^*)}$

coeff large enough relative to penalty

Case Study 3: fMRI Prediction

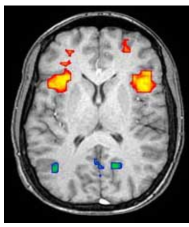
Fused LASSO

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
April 28^{th}, 2015

fMRI Prediction Subtask

- **Goal:** Predict semantic features from fMRI image

Learning S: images → semantic features



X^i

Features of word

y^i

$$X^i = \begin{bmatrix} X_1^i \\ \vdots \\ X_{20,000}^i \end{bmatrix} \xrightarrow{S} \begin{bmatrix} y_1^i \\ \vdots \\ y_d^i \end{bmatrix}$$

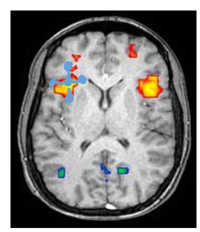
$\in \mathbb{R}^{20,000}$

d = # of semantic features

$y^i \in \mathbb{R}^d$

Simplification: consider predicting each $\underset{\text{feature}}{y_j^i}$ separately

Fused LASSO



- Might want coefficients of neighboring voxels to be similar

  discover regions of importance

- How to modify LASSO penalty to account for this?

- Graph-guided fused LASSO
  - Assume a 2d lattice graph connecting neighboring pixels in the fMRI image
  - Penalty:

$$\|y - X\beta\|_2^2 + \lambda_1 \sum_j |\beta_j| + \lambda_2 \sum_{(s,t)\in E} |\beta_s - \beta_t|$$

has edge in graph

penalizing coeff. having different weights

Generalized LASSO

- Assume a structured linear regression model:

$$\|y - X\beta\|_2^2 + \lambda \|D\beta\|_1$$

$D \in \mathbb{R}^{m \times p}$

- If D is invertible, then get a new LASSO problem if we substitute

$$\beta = D^{-1}\beta^{new} \longrightarrow \|y - \underbrace{XD^{-1}}_{\tilde{X}}\beta^{new}\|_2^2 + \lambda \|\beta^{new}\|_1$$

$\tilde{X}$ new design matrix

- Otherwise, not equivalent

  - solve for $\hat{\beta}^{new}$
  - set $\hat{\beta} = D^{-1}\hat{\beta}^{new}$

- For solution path, see Ryan Tibshirani and Jonathan Taylor, "The Solution Path of the Generalized Lasso." Annals of Statistics, 2011.

Generalized LASSO

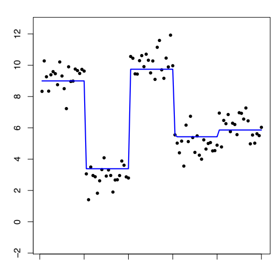
"signal approximation" scenario $X = I$

- associate a unique x^i w/ each y^i

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2}\|y - \beta\|_2^2 + \lambda\|D\beta\|_1$$

Let $D = \begin{bmatrix} -1 & 1 & 0 & 0 & \ldots \\ 0 & -1 & 1 & 0 & \ldots \\ 0 & 0 & -1 & 1 & \ldots \\ \vdots & & & & \end{bmatrix}$. This is the 1d fused lasso.

$$\lambda \sum_j |\beta_j - \beta_{j-1}|$$



encourages piecewise constant

Generalized LASSO

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^n}{\text{argmin}} \ \frac{1}{2}\|y - \beta\|_2^2 + \lambda\|D\beta\|_1$$

Suppose D gives "adjacent" differences in β:

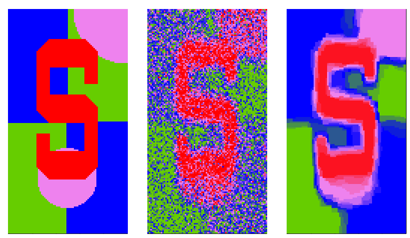
$$D_i = (0, 0, \ldots \underset{s}{-1}, \ldots, \underset{t}{1}, \ldots 0),$$

$\lambda \sum_{(s,t) \in E} |\beta_t - \beta_s|$

where adjacency is defined according to a graph $\mathcal{G}$. For a 2d grid, this is the 2d fused lasso.

noisy image

true



encourages constant regions

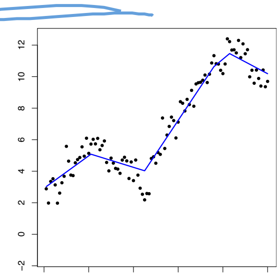
recovered image

 15

Generalized LASSO

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^n}{\text{argmin}} \ \frac{1}{2}\|y - \beta\|_2^2 + \lambda\|D\beta\|_1$$

Let $D = \begin{bmatrix} -1 & 2 & -1 & 0 & \ldots \\ 0 & -1 & 2 & -1 & \ldots \\ 0 & 0 & -1 & 2 & \ldots \\ \vdots & & & & \end{bmatrix}$. This is linear trend filtering.



β_1 β_2 β_3

$\beta_3 - \beta_2 = \beta_2 - \beta_1$

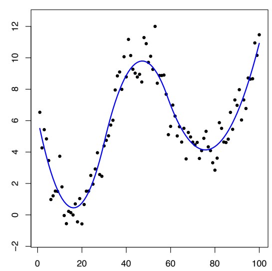
$\Rightarrow 2\beta_2 - \beta_1 - \beta_3 = 0$

 16

Generalized LASSO

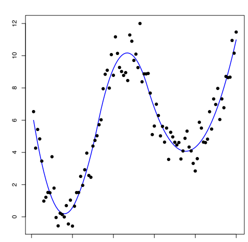
$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2}\|y - \beta\|_2^2 + \lambda \|D\beta\|_1$$

Let $D = \begin{bmatrix} -1 & 3 & -3 & 1 & \dots \\ 0 & -1 & 3 & -3 & \dots \\ 0 & 0 & -1 & 3 & \dots \\ \vdots & & & & \end{bmatrix}$. Get quadratic trend filtering.

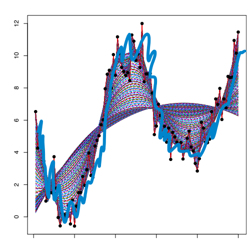


Generalized LASSO

- Tracing out the fits as a function of the regularization parameter



$\hat{\beta}_\lambda$ for $\lambda = 25$



$\hat{\beta}_\lambda$ for $\lambda \in [0, \infty]$

Acknowledgements

- Some material relating to the fused/generalized LASSO slides was provided by Ryan Tibshirani

Case Study 3: fMRI Prediction

LASSO Solvers—Part 1:
LARS

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
April 28th, 2015

LASSO Algorithms

- So far: Standard convex optimizer
- Now: Least angle regression (LAR)
  - Efron et al. 2004
  - Computes entire path of solutions
  - State-of-the-art until 2008
- Next up:
  - Pathwise coordinate descent ("shooting") – new
  - Parallel (approx.) methods

LARS ← shrinkage

now

later

©Emily Fox 2015 21

LARS – Efron et al. 2004

- LAR is an efficient stepwise variable selection algorithm
  - "useful and less greedy version of traditional forward selection methods"
- Can be modified to compute regularization path of LASSO
  - → LARS (Least angle regression and *shrinkage*)
- Increasing upper bound *B*, coefficients gradually "turn on"
  - Few critical values of *B* where support changes
  - Non-zero coefficients increase or decrease linearly between critical points
  - Can solve for critical values analytically

key to providing full regularization path/coeff.

- Complexity:

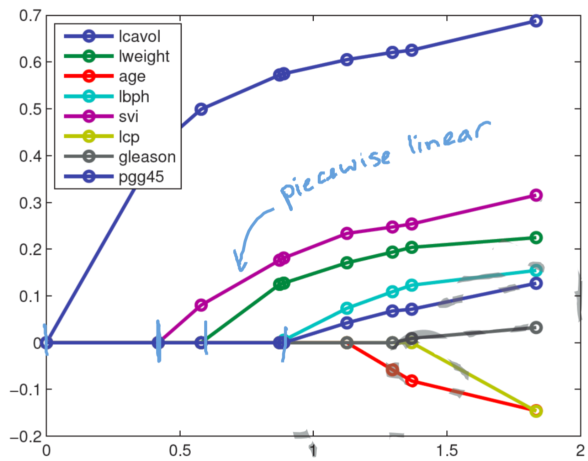
$O(\min(Np^2, pN^3))$ = cost of a single LS soln

# of obs

# of covariates

©Emily Fox 2015 22

LASSO Coefficient Path



From Kevin Murphy textbook

LARS – Algorithm Overview

- Start with all coefficient estimates $\hat{\beta}_1 = \hat{\beta}_2 = \ldots = \hat{\beta}_p = 0$
- Let $\mathcal{A}$ be the "active set" of covariates most correlated with the "current" residual ← based on covariates already in the model
- Initially, $\mathcal{A} = \{x_{j_1}\}$ for some covariate x_{j_1}
- Take the largest possible step in the direction of x_{j_1} until another covariate x_{j_2} enters $\mathcal{A}$
- Continue in the direction equiangular between x_{j_1} and x_{j_2} until a third covariate x_{j_3} enters $\mathcal{A}$
- Continue in the direction equiangular between $x_{j_1}, x_{j_2}, x_{j_3}$ until a fourth covariate x_{j_4} enters $\mathcal{A}$
- This procedure continues until all covariates are added

Comments

- LARS increases $\mathcal{A}$, but LASSO allows it to decrease
- Only involves a single index at a time
- If *p* > *N*, LASSO returns at most *N* variables
- If group of variables are highly correlated, LASSO tends to choose one to include rather arbitrarily
  - Straightforward to observe from LARS algorithm….Sensitive to noise.

beware of interpreting the variables included

More Comments

- In general, can't solve analytically for GLM (e.g., logistic reg.)
  - Gradually decrease λ and use efficiency of computing $\hat{\beta}(\lambda_k)$ from $\hat{\beta}(\lambda_{k-1})$ = warm-start strategy
  - See Friedman et al. 2010 for coordinate ascent + warm-start strategy
- If *N* > *p*, but variables are correlated, ridge regression tends to have better predictive performance than LASSO (Zou & Hastie 2005)
  - Elastic net is hybrid between LASSO and ridge regression

$$\|y - X\beta\|_2^2 + \lambda_1 \sum |\beta_j| + \lambda_2 \|\beta\|_2^2$$

(there are still some issues... see KM book)

Case Study 3: fMRI Prediction

LASSO Solvers – Part 2:

SCD for LASSO (Shooting)
Parallel SCD (Shotgun)
Parallel SGD
Averaging Solutions

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox

April 28th, 2015

Scaling Up LASSO Solvers

- Another way to solve LASSO problem:
  - Stochastic Coordinate Descent (SCD)
  - Minimizing a coordinate in LASSO
- A simple SCD for LASSO (Shooting)
  - Your HW, a more efficient implementation! ☺
  - Analysis of SCD
- Parallel SCD (Shotgun)
- Other parallel learning approaches for linear models
  - Parallel stochastic gradient descent (SGD)
  - Parallel independent solutions then averaging
- ADMM

Coordinate Descent

- Given a function F(β) $\leftarrow F(\beta_1, \dots, \beta_p)$
  - Want to find minimum $\beta^* \in \arg\min_{\beta} F(\beta)$
- Often, hard to find minimum for all coordinates, but easy for one coordinate

  1D optimization problem
- Coordinate descent:

  while not converged

  pick coordinate j

  $\beta_j \leftarrow \arg\min_b F(\beta_1, \beta_2, \dots, \beta_{j-1}, b, \beta_{j+1}, \dots, \beta_p)$ varying jth coord.
- How do we pick a coordinate?

  Round robin, random, smartly, ...
- When does this converge to optimum?

  e.g. strongly convex (separability)

 29

Soft Threshholding

Fixing all other coord., soln for jth coord: strength of penalty

$c_j \propto corr(\underline{x}_j, r_{-j})$ — all examples of feature j covariate; residual from model w/o using jth covariate

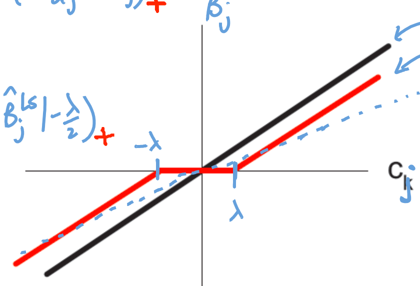
$$\hat{\beta}_j = \begin{cases} (c_j + \lambda)/a_j & c_j < -\lambda \\ 0 & c_j \in [-\lambda, \lambda] \\ (c_j - \lambda)/a_j & c_j > \lambda \end{cases} \leftarrow \arg\min_b F(\dots, \beta_{j-1}, b, \beta_{j+1})$$

$$= \text{sign}\left(\frac{c_j}{a_j}\right)\left(\frac{|c_j|}{a_j} - \frac{\lambda}{a_j}\right)_+ = \text{sign}(c_j)\frac{(|c_j| - \lambda)_+}{a_j} = \hat{\beta}_j$$

If $X^T X = I$

$$\hat{\beta}_j^{lasso} = \text{sign}(\hat{\beta}_j^{LS})\left(|\hat{\beta}_j^{LS}| - \frac{\lambda}{2}\right)_+$$

$$\hat{\beta}_j^{ridge} = \frac{\hat{\beta}^{LS}}{1+\lambda}$$



From Kevin Murphy textbook

In LASSO, all coeff. are shrunk relative to $\hat{\beta}^{LS}$

 30

Stochastic Coordinate Descent for LASSO (aka Shooting Algorithm)

$\underset{b}{\operatorname{argmin}} \; F(\beta_1, \ldots, \beta_{j-1}, b, \beta_{j+1}, \ldots, \beta_p)$

- Repeat until convergence
  - Pick a coordinate *j* at random
    - Set:

$$\hat{\beta}_j = \begin{cases} (c_j + \lambda)/a_j & c_j < -\lambda \\ 0 & c_j \in [-\lambda, \lambda] \\ (c_j - \lambda)/a_j & c_j > \lambda \end{cases} = \operatorname{sign}(c_j)\frac{(|c_j| - \lambda)_+}{a_j}$$

- Where: cache

$$a_j = 2\sum_{i=1}^{N}(x_j^i)^2 \qquad c_j = 2\sum_{i=1}^{N} x_j^i (y^i - \beta'_{-j} x^i_{-j})$$

omitting j^{th} cov $[\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p]$

cost per iteration $O(N)$

Can be done more efficiently. Proof: Your HW!

Analysis of SCD [Shalev-Shwartz, Tewari '09/'11]

$e_j := \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \leftarrow j^{th}$ dim

- Analysis works for LASSO, L1 regularized logistic regression, and other objectives!
- For (coordinate-wise) strongly convex functions:

$$F(\beta + \Delta\beta) \leq F(\beta) + \partial\beta_j (\nabla F(\beta))_j + \frac{\gamma(\partial\beta_j)^2}{2}$$

$$\Delta\beta = \partial\beta_j \cdot e_j$$

Lasso: $\gamma = 1$

Logistic reg $\gamma = \frac{1}{4}$

- Theorem:
  - Starting from $\beta^{(0)}$
  - After T iterations

$$E[F(\beta^{(T)})] - F(\beta^*) \leq \frac{p(\gamma \|\beta^*\|_2^2 + 2F(\beta^{(0)}))}{T+1}$$

dim; how hard; where we started from

gets linearly better with iterations

  - Where E[ ] is wrt random coordinate choices of SCD

- Natural question: How does SCD & SGD convergence rates differ?

see paper: SCD → faster w/ large p ← no params to tune

SGD → faster w/ large N ← need η

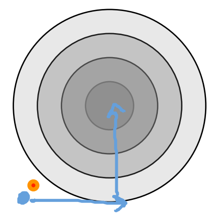
Shooting: Sequential SCD

Lasso: $\min_{\beta} F(\beta)$ where $F(\beta) = \| X\beta - \mathbf{y} \|_2^2 + \lambda \|\beta\|_1$

Stochastic Coordinate Descent (SCD)
(e.g., Shalev-Shwartz & Tewari, 2009)

While not converged,
- Choose random coordinate j,
- Update β_j (closed-form minimization)

$F(\beta)$ contour



How do we measure?
- annoying
- over a time window? has anything changed?

- do a round robin iter to msr convergence
measure

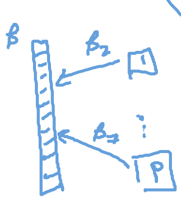
Shotgun: Parallel SCD [Bradley et al '11]

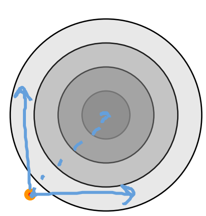
Lasso: $\min_{\beta} F(\beta)$ where $F(\beta) = \| X\beta - \mathbf{y} \|_2^2 + \lambda \|\beta\|_1$

Shotgun (Parallel SCD)

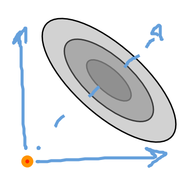
While not converged,
- On each of P processors,
  - Choose random coordinate j,
  - Update β_j (same as for Shooting)

independently





yes!
features are uncorrelated



no :(
Feature are highly correlated

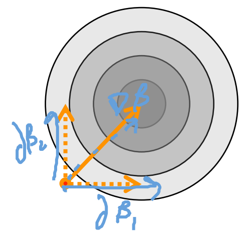
Is SCD inherently sequential?

Lasso: $\min_{\beta} F(\beta)$ where $F(\beta) = \| X\beta - \mathbf{y} \|_2^2 + \lambda \| \beta \|_1$

Coordinate update:

$$\beta_j \leftarrow \beta_j + \underline{\underline{\delta\beta_j}}$$

(closed-form minimization)



Collective update:

$$\Delta\beta = \begin{pmatrix} \delta\beta_i \\ 0 \\ 0 \\ \delta\beta_j \\ 0 \end{pmatrix}$$

there are interferences in these update if features are correlated.

Can we quantify this?

 35

Is SCD inherently sequential?

Lasso: $\min_{\beta} F(\beta)$ where $F(\beta) = \| X\beta - \mathbf{y} \|_2^2 + \lambda \| \beta \|_1$

Theorem: If X is normalized s.t. diag($X^T X$)=1,

$$F(\beta + \Delta\beta) - F(\beta)$$

$$\leq \underbrace{-\sum_{i_j \in \mathcal{P}} \left(\delta\beta_{i_j}\right)^2}_{\text{"positive" progress}} + \underbrace{\sum_{\substack{i_j, i_k \in \mathcal{P}, \\ j \neq k}} \left(X^{\mathrm{T}} X\right)_{i_j, i_k} \delta\beta_{i_j} \delta\beta_{i_k}}_{\text{could be pos. or neg.}}$$

"interference" or "bias" from parallelism

 36

Is SCD inherently sequential?

Theorem: If X is normalized s.t. diag(X^TX)=1,

$$F(\beta+\Delta\beta)-F(\beta)$$

$$\le -\sum_{i_j\in\mathcal{P}}\left(\delta\beta_{i_j}\right)^2 + \sum_{\substack{i_j,i_k\in\mathcal{P},\\ j\neq k}} \overbrace{\left(X^{\mathrm{T}}X\right)_{i_j,i_k}}^{\text{key term}} \delta\beta_{i_j}\delta\beta_{i_k}$$

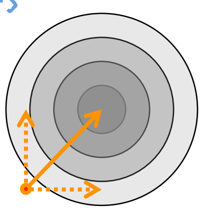
msrs magnitude of interference

$(X^TX)_{jk}=0$

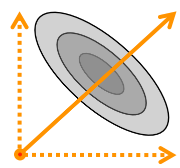
corr bt. x_j & x_k

$(X^TX)_{jk}\neq 0$

"interference"



Nice case:
Uncorrelated
features



Bad case:
Correlated
features

Shotgun: Convergence Analysis

Lasso: $\min_\beta F(\beta)$ where $F(\beta) = \|X\beta - \mathbf{y}\|_2^2 + \lambda\|\beta\|_1$

Assume # parallel updates $P < p/\rho + 1$

dim

spectral radius (X^TX)

$$\underbrace{E[F(\beta^{(T)})]}_{\text{where we are}} - \underbrace{F(\beta^*)}_{\text{opt}} \le \frac{\overset{\text{dim}}{p}\left(\|\beta^*\|_2^2 + 2F(\beta^{(0)})\right)}{T\cdot P}$$

# of iters

# of processors

Generalizes bounds for Shooting (Shalev-Shwartz & Tewari, 2009)

Convergence Analysis

Lasso: $\min_{\beta} F(\beta)$ where $F(\beta) = \| X\beta - \mathbf{y} \|_2^2 + \lambda \| \beta \|_1$

Theorem: Shotgun Convergence

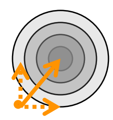
Assume $P < p/\rho + 1$

where ρ = spectral radius of $\mathbf{X}^T\mathbf{X}$

$$E\left[F(\beta^{(T)})\right] - F(\beta^*) \leq \frac{p\left(\frac{1}{2}\| \beta^* \|_2^2 + F(\beta^{(0)})\right)}{TP}$$

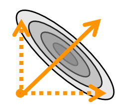
linear speed up, up to P processors

Nice case: Uncorrelated features



$\rho = \underline{1} \Rightarrow P_{max} = \underline{p}$

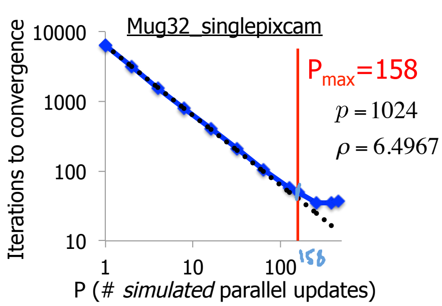
Bad case: Correlated features

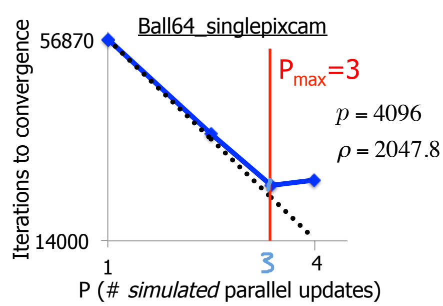


$\rho = \underline{p} \Rightarrow P_{max} = \underline{1}$ (at worst)

Empirical Evaluation

2 classical compressed sensing datasets





What you need to know

- Sparsistency
- Fused LASSO
- LASSO Solvers
  - LARS
  - A simple SCD for LASSO (Shooting)
    - Your HW, a more efficient implementation! ☺
    - Analysis of SCD
  - Parallel SCD (Shotgun)