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
Sham Kakade
May 5, 2016
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 $F$
  - Want to find minimum

- Often, hard to find minimum for all coordinates, but easy for one coordinate

- Coordinate descent:
  - How do we pick a coordinate?
  - When does this converge to optimum?
Soft Thresholding

\[ \min_{\beta_j} \text{obj.} \left( \beta_1, \ldots, \beta_j, \ldots, \beta_p \right) \]

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

\[ c_j \propto \text{cov}(x_j, r_j) \]

\[ a_j = \mathbb{E}[x_j^2] \]

If \( X'X = I \)

\[ \beta_j = \frac{c_j}{a_j} \]

From Kevin Murphy textbook
Stochastic Coordinate Descent for LASSO (aka Shooting Algorithm)

• Repeat until convergence
  – Pick a coordinate $j$ at random
  • Set:
    $\hat{\beta}_j = \begin{cases} 
    (c_j + \lambda)/a_j & c_j < -\lambda \\
    0 & c_j \in [-\lambda, \lambda] \\
    (c_j - \lambda)/a_j & c_j > \lambda 
    \end{cases}$

• Where:
  $a_j = 2 \sum_{i=1}^{N} (x^i_j)^2$
  $c_j = 2 \sum_{i=1}^{N} x^i_j (y^i - \beta'_j x^i_{-j})$
Analysis of SCD [Shalev-Shwartz, Tewari ’09/’11]

• Analysis works for LASSO, L1 regularized logistic regression, and other objectives!

• For (coordinate-wise) strongly convex functions:
  • Theorem:
    – Starting from
    – After T iterations
    – Where $E[\cdot]$ is wrt random coordinate choices of SCD

• Natural question: How does SCD & SGD convergence rates differ?
**Shooting: Sequential SCD**

Lasso: \( \min F(\ ) \) where \( F(\ ) = \| Xy \|_2^2 + \| \|_1 \)

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

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

\[
F(\ ) \text{ contour}
\]

©Sham Kakade 2016
Shotgun: Parallel SCD [Bradley et al ‘11]

Lasso: \[ \min F(\cdot) \text{ where } F(\cdot) = \| X y \|_2^2 + \| \cdot \|_1 \]

**Shotgun (Parallel SCD)**

While not converged,
- On each of \( P \) processors,
  - Choose random coordinate \( j \),
  - Update \( \beta_j \) (same as for Shooting)
Is SCD inherently sequentially?

**Lasso:** \( \min F(\cdot) \) where \( F(\cdot) = \| X y \|_2^2 + \| \cdot \|_1 \)

**Coordinate update:**

\[
\begin{align*}
  j &\leftarrow j + j & (\text{closed-form minimization})
\end{align*}
\]

**Collective update:**

\[
\begin{pmatrix}
  0 \\
  0 \\
  0 \\
  i
\end{pmatrix} = \\
\begin{pmatrix}
  0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0 \\
  0 & 0 & 0 & 0
\end{pmatrix}
\]

©Sham Kakade 2016
Convergence Analysis

Lasso: \( \min F(\cdot) \) where \( F(\cdot) = \| X y \|_2^2 + \| \cdot \|_1 \)

Theorem: Shotgun Convergence
Assume \( P < \frac{\rho}{1 + 1} \)
where \( \rho = \) spectral radius of \( X^T X \)

\[
E F(T) = F(\star) \\
\frac{\rho \left( \frac{1}{2} \| \star \|_2^2 + F(0) \right)}{TP}
\]

Nice case: Uncorrelated features
\( = \_ \) \( P_{\text{max}} = \_ \)

Bad case: Correlated features
\( = \_ \) \( P_{\text{max}} = \_ \) (at worst)
Stepping Back...

• Stochastic coordinate ascent
  – Optimization:
  – Parallel SCD:
  – Issue:
  – Solution:

• Natural counterpart:
  – Optimization:
  – Parallel
  – Issue:
  – Solution:
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)
Case Study 3: fMRI Prediction

“Scalable” LASSO Solvers:
Parallel SCD (Shotgun)
Parallel SGD
Averaging Solutions
ADMM

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Sham Kakade
May 5th, 2016
Stepping Back...

• Stochastic coordinate ascent
  – Optimization:
  – Parallel SCD:
  – Issue:
  – Solution:

• Natural counterpart:
  – Optimization:
  – Parallel
  – Issue:
  – Solution:
Parallel SGD with No Locks
[e.g., Hogwild!, Niu et al. ‘11]

- Each processor in parallel:
  - Pick data point \( i \) at random
  - For \( j = 1 \ldots p \):

- Assume atomicity of:
Addressing Interference in Parallel SGD

• Key issues:
  – Old gradients
  – Processors overwrite each other’s work

• Nonetheless:
  – Can achieve convergence and some parallel speedups
  – Proof uses weak interactions, but through sparsity of data points
Problem with Parallel SCD and SGD

• Both Parallel SCD & SGD assume access to current estimate of weight vector

• Works well on shared memory machines

• Very difficult to implement efficiently in distributed memory

• Open problem: Good parallel SGD and SCD for distributed setting...
  – Let’s look at a trivial approach
Simplest Distributed Optimization Algorithm Ever Made

• Given $N$ data points & $P$ machines
• Stochastic optimization problem:
• Distribute data:
  • Solve problems independently
  • Merge solutions
• Why should this work at all????
For Convex Functions...

- Convexity:

- Thus:
Hopefully...

- Convexity only guarantees:

- But, estimates from independent data!

Figure from John Duchi

©Sham Kakade 2016
Analysis of Distribute-then-Average

[Zhang et al. ‘12]

- Under some conditions, including strong convexity, lots of smoothness, etc.
- If all data were in one machine, converge at rate:

- With $P$ machines, converge at a rate:
Tradeoffs, tradeoffs, tradeoffs,…

- Distribute-then-Average:
  - “Minimum possible” communication
  - Bias term can be a killer with finite data
    - Issue definitely observed in practice
  - Significant issues for L1 problems:

- Parallel SCD or SGD
  - Can have much better convergence in practice for multicore setting
  - Preserves sparsity (especially SCD)
  - But, hard to implement in distributed setting
Alternating Directions Method of Multipliers

• A tool for solving convex problems with separable objectives:

• LASSO example:

• Know how to minimize $f(\beta)$ or $g(\beta)$ separately
ADMM Insight

• Try this instead:

• Solve using method of multipliers
• Define the augmented Lagrangian:
  
  – Issue: L2 penalty destroys separability of Lagrangian
  – Solution: Replace minimization over (x, z) by alternating minimization
ADMM Algorithm

• Augmented Lagrangian:

\[ L_\rho(x, z, y) = f(x) + g(z) + y^T(x - z) + \frac{\rho}{2}||x - z||^2_2 \]

• Alternate between:

1. \( x \leftarrow \)

2. \( z \leftarrow \)

1. \( y \leftarrow \)
ADMM for LASSO

\[ L_\rho(x, z, y) = f(x) + g(z) + y^T(x - z) + \frac{\rho}{2}||x - z||^2_2 \]

• Objective:

• Augmented Lagrangian:

\[ L_\rho(\beta, z, a) = \]

• Alternate between:

1. \( \beta \leftarrow \)

2. \( z \leftarrow \)

1. \( a \leftarrow \)
ADMM Wrap-Up

• When does ADMM converge?
  – Under very mild conditions
  – Basically, $f$ and $g$ must be convex

• ADMM is useful in cases where
  – $f(x) + g(x)$ is challenging to solve due to coupling
  – We can minimize
    • $f(x) + (x-a)^2$
    • $g(x) + (x-a)^2$

• Reference
What you need to know

• 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
  – General idea
  – Application to LASSO