# A Subspace Acceleration Framework For Minimization Involving a Group Sparsity-Inducing Regularizer

Frank E. Curtis[1] Yutong Dai[1] Daniel P. Robinson[1]

[1]Industrial and Systems Engineering, Lehigh University

INFORMS Annual Meeting 2020

11/12/2020

# Outline

# Problem of Interest

## Sparse optimization problem

$$\min_{x\in\mathbb{R}^n} \quad f(x) + r(x) \quad \text{with} \quad r(x) = \sum_{i=1}^{K} \lambda_i \|x_{\mathcal{G}_i}\|_2 \qquad \big(\lambda_i > 0,\ \mathcal{G}_i \subset \{1,\ldots,n\}\big)$$

- $f$: loss function, assumed to be convex and differentiable:
  - logistic regression: $f(x) = \frac{1}{N}\sum_{i=1}^{N} \log(1 + e^{-y_i x^T d_i})$
- $r$: sparsity inducing regularizer is convex and nonsmooth:
  - group sparsity: $r(x) = \sum_{i=1}^{K} \lambda_i \|x_{\mathcal{G}_i}\|_p$ for $p \in [1, \infty)$
- problems arise in signal processing and machine learning applications
  - jointly select genes that regulate hormone levels
- sparsity in group structure imposes more optimization challenges

# Overview

## Two Pillars

1. space decomposition
   - Predict zero and non-zero groups of the solution $x^*$
2. subspace acceleration
   - Utilize second order information to improve convergence rate
   - Design a projected line search scheme to promote the sparsity of iterates

# Space decomposition

## Proximal Gradient (PG) Direction

1. choose PG parameter $\alpha_k > 0$
2. compute PG direction

$$s_k \leftarrow \arg\min_{x \in \mathbb{R}^n} \left\{ \frac{1}{2\alpha_k} \|x - \big(x_k - \alpha_k \nabla f(x_k)\big)\|_2^2 + r(x) \right\} - x_k$$

Properties:

- Repeated computation of $s_k$ recovers PG iterates
- The support of a PG iterate matches with that of $x^*$ after finite #iterations

**Use $s_k$ to do space decomposition.**

# Subspace acceleration

## Reduced Newton System

1. Pick a subset $I_k \subset \{1, 2, \cdots, n\}$ s.t. all groups of variables in $I_k$ are non-zero.
2. Set $g_k \leftarrow \nabla_{\mathcal{I}_k}(f + r)(x_k)$ and pick a positive-definite $H_k \in \mathbb{R}^{|\mathcal{I}_k| \times |\mathcal{I}_k|}$.
3. Obtain an inexact Newton direction $d_k$ by solving

$$H_k d_k \approx -g_k$$

with a CG method equipped with early termination rules.

Properties:

- The iterates $\{x_k\}$ (under assumptions) converges to $x^*$ at a superlinear/quadratic rate.

# Algorithmic Framework

**Algorithm** Fast Reduced-Space Algorithm for Group Sparsity (FaRSA-Group)

**for** $k = 0, 1, 2, \ldots$ **do**

Compute the PG direction $s_k$.

Divide the groups $\{\mathcal{G}_i\}$ into two sets: [How?]

$$\mathcal{I}_k^{\text{cg}} := \{\text{the groups that you think are nonzero at a solution}\}$$

$$\mathcal{I}_k^{\text{pg}} := \{\text{the groups that you think are zero at a solution}\}$$

Define measures of optimality:

$$\chi_k^{\text{cg}} := \left\| [s_k]_{\mathcal{I}_k^{\text{cg}}} \right\|_2 \quad \text{and} \quad \chi_k^{\text{pg}} := \left\| [s_k]_{\mathcal{I}_k^{\text{pg}}} \right\|_2$$

Terminate if $\max\{\chi_k^{\text{cg}}, \chi_k^{\text{pg}}\} \le \epsilon$.

**if** $\chi_k^{\text{pg}} \le \chi_k^{\text{cg}}$ **then**

Select $I_k \subseteq \mathcal{I}_k^{\text{cg}}$.

Apply CG method on reduced Newton system $H_k d \approx -g_k$ to obtain $d_k$.

Perform a reduced space projected line search using the direction $d_k$. [How?]

**else**

Select $I_k \subseteq \mathcal{I}_k^{\text{pg}}$.

Perform a reduced space backtracking Armijo linesearch along the direction $[s_k]_{I_k}$.

**end if**

Compute PG parameter $\alpha_{k+1}$.

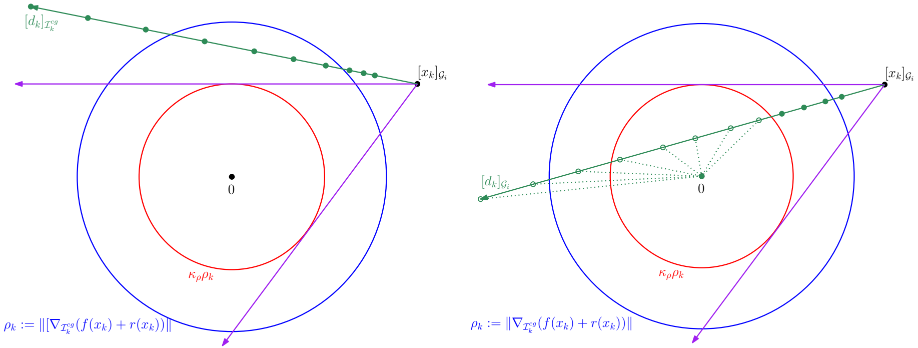
**end for**

# $\mathcal{I}_k^{\text{cg}}$ and $\mathcal{I}_k^{\text{pg}}$

- $\mathcal{I}_k^{\text{cg}}$ consists of all group of variables that are currently
  - non-zero
  - *sufficiently* far away from zero
    1. taking an unit-step along the $s_k$ remains non-zero
    2. distance to 0 proportional to the first order optimality measure

- $\mathcal{I}_k^{\text{pg}} = \{1, 2, \ldots, n\} \setminus \mathcal{I}_k^{\text{cg}}$

# Projected backtracking line search




(a) The reduced Newton-CG direction does not intersect the red sphere.

(b) The reduced Newton-CG direction does intersect the red sphere.

Figure: Projected backtrack lines search along the Newton-CG direction reduced-space .

# Global

Assumptions:

- $f$ and $r$ are convex, proper, and closed
- $f$ is a $C^1$ function with $\nabla f$ Lipschitz continuous
- $f + r$ is bounded below

**Theorem 1 (worst-case complexity)**

*For $\epsilon \in (0, \infty)$, the maximum number of iterations before $\max\{\chi_k^{\mathrm{cg}}, \chi_k^{\mathrm{pg}}\} \leq \epsilon$ is*

$$\mathcal{O}\left(\epsilon^{-(2+p)}\right)$$

**Remark:** For PG, the worst case complexity is $\mathcal{O}\left(\epsilon^{-2}\right)$.

# Local

Assumptions:

- $f$ is strongly convex, a $C^2$ function, and $\nabla^2 f$ is Lipschitz continuous
- non-degeneracy: $\|[\nabla f(x_*)]_{\mathcal{G}_i}\|_2 < \lambda_i$ for all $i$ such that $[x_*]_{\mathcal{G}_i} = 0$.

**Theorem 2 (support identification)**

*Let $\mathcal{S}_* := \{i : [x_*]_{\mathcal{G}_i} \neq 0\}$. For all sufficiently large $k$, it holds that*

$$[x_k]_{\mathcal{G}_i} \neq 0 \textit{ for all } i \in \mathcal{S}_* \quad \textit{and} \quad [x_k]_{\mathcal{G}_i} = 0 \textit{ for all } i \notin \mathcal{S}_*.$$

**Theorem 3 (local convergence rate)**

*The sequence $\{x_k\}$ converges to the unique minimizer $x^*$ at a superlinear / quadratic rate, depending on how accurately we solve the reduced Newton system.*

# Setup

- 25 binary classification datasets from LIBSVM
- 2 sparsity levels:
  - $\lambda_i = 0.1\lambda_{\min}\sqrt{|\mathcal{G}_i|}$
  - $\lambda_i = 0.01\lambda_{\min}\sqrt{|\mathcal{G}_i|}$

  where

  $$\lambda_{\min} = \min\{\lambda \geq 0\colon \text{ the solution with } \lambda_i = \lambda\sqrt{|\mathcal{G}_i|} \text{ is } x = 0\}$$
- 4 different settings for the number of groups:

  $$\text{number of groups} \in \{\lfloor 0.25n \rfloor, \lfloor 0.50n \rfloor, \lfloor 0.75n \rfloor, n\},$$
- Total of 200 problem instances are tested
- Compare our algorithm `FaRSA-Group`([1]) vs. `gglasso`([2])
- Max allowed time: 1000 seconds.

# Numerical results

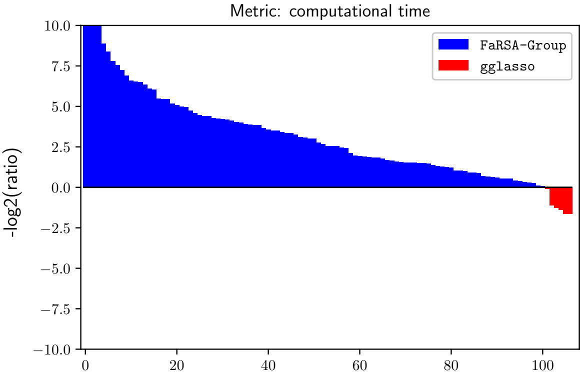



Figure: Performance profile of CPU time (seconds) on problem instances for which at least one algorithm takes at least 1 second.

- the height of the bar given by

$$-\log_2\left(\frac{\text{time required by }\texttt{FaRSA-Group}}{\text{time required by }\texttt{gglasso}}\right) \tag{1}$$

# Conclusions

- New framework for optimization problems with group-sparse regularization.
  - scalable: reduced-space subproblems
  - fast, efficient: reduced-space Newton-CG computation
- Global convergence with worst-case complexity result.
- Fast local convergence.
- State-of-the-art performance.

# References I


[1] F. E. Curtis, Y. Dai, and D. P. Robinson, *A subspace acceleration method for minimization involving a group sparsity-inducing regularizer*, 2020.

[2] Y. Yang and H. Zou, *A fast unified algorithm for solving group-lasso penalize learning problems*, Statistics and Computing, 25 (2015), pp. 1129–1141.

# Newton-CG direction

1. Define the model
$$m_k(d) := g_k^T d + \tfrac{1}{2} d^T H_k d$$
2. Compute the reference direction (an approximate minimizer of $m_k$) as
$$d_k^R \leftarrow -\beta_k g_k, \text{ where } \beta_k \leftarrow \|g_k\|_2^2/(g_k^T H_k g_k)$$
3. Choose $\mu_k \in (0,1]$ and then compute any $\bar{d}_k \approx \underset{d}{\operatorname{argmin}}\ m_k(d)$ that satisfies
$$\begin{aligned} g_k^T \bar{d}_k &\le g_k^T d_k^R \\ m_k(\bar{d}_k) &\le m_k(0) \quad \text{and} \\ \|H_k \bar{d}_k + g_k\|_2 &\le \mu_k \|g_k\|_2^q \end{aligned}$$

## $\mathcal{I}_k^{\text{cg}}$ and $\mathcal{I}_k^{\text{pg}}$

### How to choose?

1. Calculate a candidate set

$$\bar{\mathcal{I}}_k^{\text{cg}} := \{j \in \mathcal{G}_i : [x_k]_{\mathcal{G}_i} \neq 0, \ \ [x_k + s_k]_{\mathcal{G}_i} \neq 0, \text{ and } \|[x_k]_{\mathcal{G}_i}\|_2 \geq \kappa_1 \|\nabla_{\mathcal{G}_i}(f+r)(x_k)\|_2\} \tag{2}$$

for some $\kappa_1 \in (0, \infty)$.

2. Secondary screening

$$\mathcal{I}_k^{\text{small}} := \{j \in \mathcal{G}_i : \mathcal{G}_i \subseteq \bar{\mathcal{I}}_k^{\text{cg}} \text{ and } \|[x_k]_{\mathcal{G}_i}\|_2 < \kappa_2 \|\nabla_{\bar{\mathcal{I}}_k^{\text{cg}}}(f+r)(x_k)\|_2^p\} \tag{3}$$

for some $\{\kappa_2, p\} \subset (0, \infty)$

3. Finalize

$$\begin{aligned} \mathcal{I}_k^{\text{cg}} &:= \bar{\mathcal{I}}_k^{\text{cg}} \setminus \mathcal{I}_k^{\text{small}} \\ \mathcal{I}_k^{\text{pg}} &:= \{1, 2, \ldots, n\} \setminus \mathcal{I}_k^{\text{cg}} \end{aligned} \tag{4}$$