FaRSA-Group: Efficient Structured Spares Problem Solver

Frank E. Curtis, Yutong Dai, Daniel P. Robinson
contact: yud319@lehigh.edu
Industrial and Systems Engineering, Lehigh University

General setup

$$\min_{x\in\mathbb{R}^n} \; f(x) + r(x)$$

- f: loss function; convex and differentiable:
- r: group sparsity inducing regularizer $r(x) = \sum_{i=1}^{K} \lambda_i \|x_{\mathcal{G}_i}\|_2$

Motivation

If we can somehow predict zero components of the solution x^*, we can use more efficient second-order method to solve a smooth problem in a low dimension space.

Contribution

- Partitioned variables in a way that incorporates the support prediction property of the PG method and tackles the challenge that the gradient of the function being optimized in the reduced space is not Lipschitz continuous.
- Designed a specialized projection procedure for the group ℓ_1-norm regularizer that allows us to prove convergence guarantees and obtain strong numerical performance.
- Proved a worst-case iteration complexity bound with a simple but principled way of adjusting the PG step size that allows for support identification in finite iterations.

Algorithm

Algorithm 1 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:

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

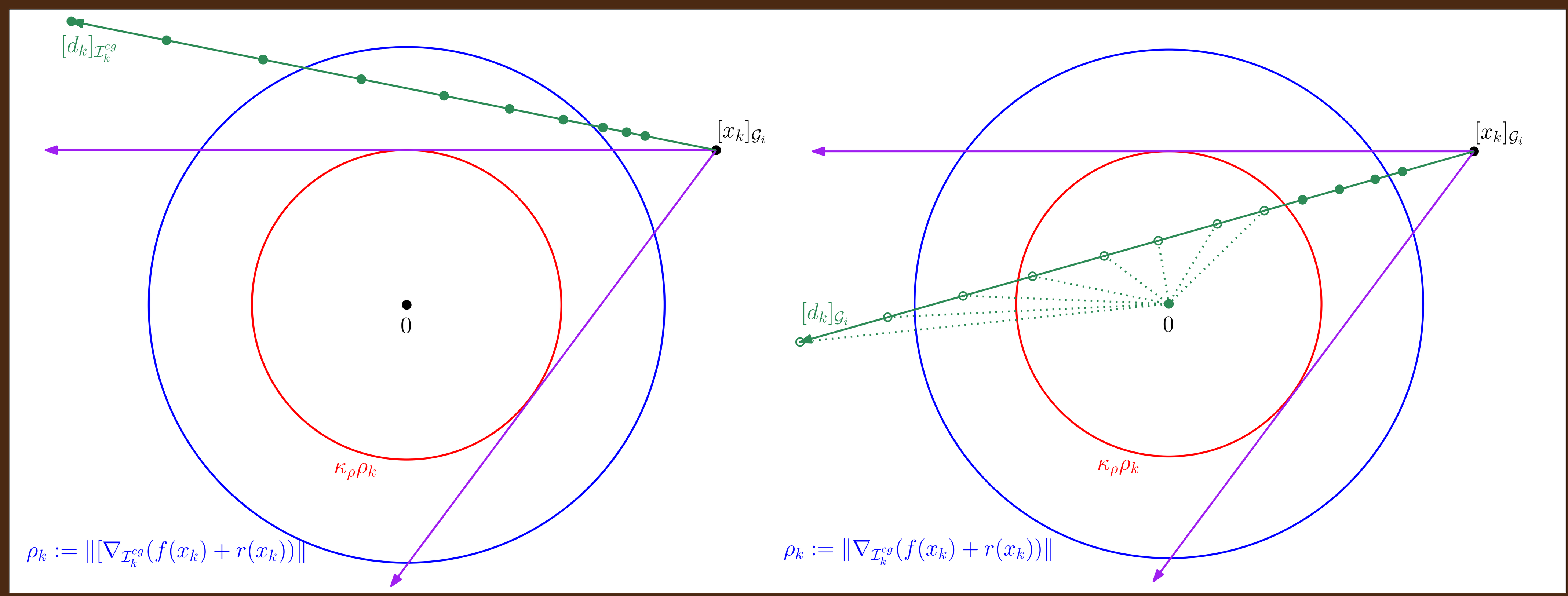
  Terminate if $\max\{\chi_k^{\text{cg}}, \chi_k^{\text{pg}}\} \le \epsilon$.
  if $\chi_k^{\text{pg}} \le \chi_k^{\text{cg}}$ **then**
    Select $I_k \subseteq \mathcal{I}_k^{\text{cg}}$.
    Apply CG method on reduced Newton system $H_k d \approx -g_k$ to obtain d_k.
    Perform a reduced space projected 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 α_{k+1}.
end for

Full-length paper!

1. Proximal gradient direction helps to identify the low-dimension manifold in finite many number of steps and can be used to do subspace decomposition.

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

2. Specialized backtracking line-search scheme with projection promotes the sparsity of iterates and speeds up the convergence.



Reference

1. A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202.
2. J. D. Lee, Y. Sun, and M. A. Saunders, Proximal newton-type methods for minimizing composite functions, SIAM Journal on Optimization, 24 (2014), pp. 1420–1443.
3. R. S. Dembo, S. C. Eisenstat, and T. Steihaug, Inexact newton methods, SIAM Journal on Numerical analysis, 19 (1982), pp. 400–408.
4. Y. Zhang, N. Zhang, D. Sun, and K.-C. Toh, An efficient Hessian based algorithm for solving large-scale sparse group Lasso problems, Math. Prog., 179 (2020), pp. 223–263.

Complexity results

Theorem 1 (worst-case complexity). *For $\epsilon \in (0, \infty)$, the maximum number of iterations before $\max\{\chi_k^{\text{cg}}, \chi_k^{\text{pg}}\} \le \epsilon$ is $\mathcal{O}\left(K\epsilon^{-(2+p)}\right)$.*

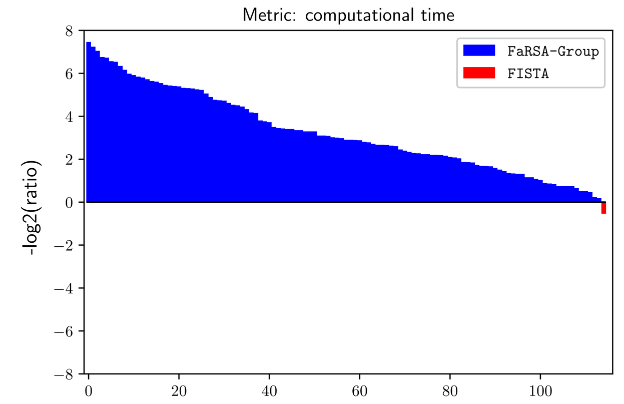
Theorem 2 (local convergence rate). *The sequence $\{x_k\}$ converges to the unique minimizer x^* at a superlinear / quadratic rate, depending on how accurately we solve the reduced Newton system.*

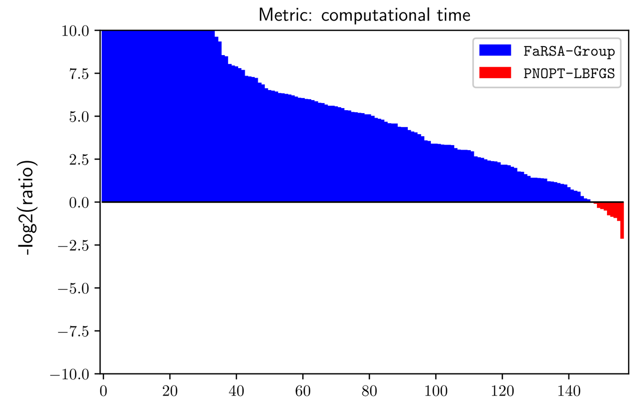
Theorem 3 (support identification). *Let $\mathcal{S}_* := \{i : [x_*]_{\mathcal{G}_i} \neq 0\}$. For all sufficiently large k, it holds that*

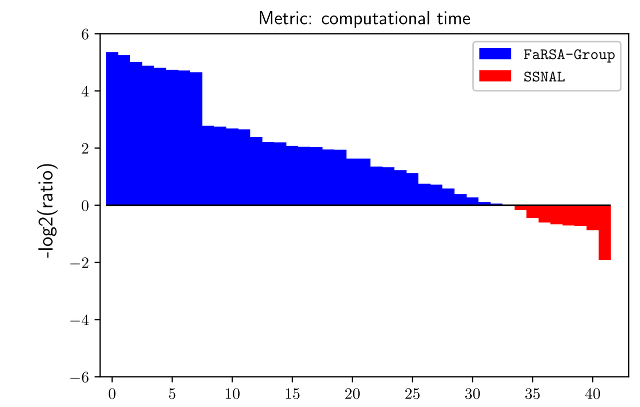
$$[x_k]_{\mathcal{G}_i} \neq 0 \text{ for all } i \in \mathcal{S}_*$$
$$[x_k]_{\mathcal{G}_i} = 0 \text{ for all } i \notin \mathcal{S}_*.$$

Experiments

- Total of 200/120 problem instances for logistic/linear regression problems;







Conclusion

- New framework for optimization problems with group-sparse regularization. Scalable, fast, efficient
- Global convergence with the worst-case complexity result
- Fast local convergence
- State-of-the-art performance