Fast Projection onto the $\ell_{\infty,1}$-Mixed Norm Ball using Steffensen Root Search

Gustavo Chau † Brendt Wohlberg* Paul Rodriguez†

†Electrical Engineering Department, Pontificia Universidad Católica del Perú, Lima, Peru
*Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, USA

Abstract

- We present a **new algorithm for computing the projection onto the $\ell_{\infty,1}$ ball**.
- **Improvements**: Steffensen type root search technique, pruning strategy and initial guess of solution.
- **Simulations**: Average speedups of 4∼5 w.r.t. state of the art. Up to 14 times faster for very sparse solutions.
- **fMRI LASSO task**: Speedups of ~ 120.

Introduction

- Mixed norms are important in modeling group correlations [1]. Let $\mathbf{A} \in \mathbb{R}^{M \times N}$, where the rows represent the different groups. The $\ell_{\infty,1}$-norm is defined as $\|\mathbf{A}\|_{\infty,1} = \Sigma_{m=1}^{M} \|\mathbf{a}_m\|_\infty$
- The main contribution of this work is a computationally efficient algorithm for computing the projection onto the $\ell_{\infty,1}$ ball:

$$\mathrm{proj}_{\|\cdot\|_{\infty,1}}(\mathbf{B},\tau) := \underset{\mathbf{X}}{\mathrm{argmin}} \frac{1}{2}\|\mathbf{X}-\mathbf{B}\|_F^2 \text{ s.t. } \|\mathbf{X}\|_{\infty,1} \le \gamma \quad (1)$$

- Sra [2] proposed a general root search based algorithm for mixed-norm ball projection problems.
- We propose two significant improvements: (i) a feasible initial solution, and (ii) pruning.

References

[1] M. Kowalski, "Sparse regression using mixed norms," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 303–324, 2009.

[2] S. Sra, "Fast projections onto $\ell_{1,q}$-norm balls for grouped feature selection," *Machine learning and knowledge discovery in databases*, pp. 305–317, 2011.

[3] S. Amat, S. Busquier, Á. Magreñán, and L. Orcos, "An overview on Steffensen-type methods," in *Advances in Iterative Methods for Nonlinear Equations*. Springer, 2016, pp. 5–21.

[4] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," *science*, vol. 320, no. 5880, pp. 1191–1195, 2008.

Proposed method

- By duality, $\mathbf{X}^* + \mathbf{A}^* = \mathbf{B}$, ($X^*$ is solution to (1)) where:

$$[\mathbf{a}_1^*; \ldots; \mathbf{a}_M^*] = \mathbf{A}^* = \min_{\mathbf{A}} \frac{1}{2}\|\mathbf{A}-\mathbf{B}\|_F^2 + \lambda \cdot \|\mathbf{A}\|_{1,\infty} \quad (2)$$

If we had $\gamma^* = \|\mathbf{A}^*\|_{1,\infty}$ then the problem would be separable in each row, i.e., $\mathbf{a}_\mathbf{m}^* = \mathrm{proj}_{\|\cdot\|_1}(\mathbf{b}_m, \gamma^*)$.

- We define the search function

$$g(\gamma) = \sum \max(\mathbf{b_m} - \mathbf{a_m}(\gamma)) - \tau,$$

A^* is obtained with $g(\gamma^*) = 0$.

$$\mathbf{a}_m(\gamma) = \begin{cases} \mathbf{b}_m & \text{if } \|\mathbf{b}_m\|_1 < \gamma \\ \mathrm{shrink}(\mathbf{b}_m, \lambda(\gamma)) & \text{if } \|\mathbf{b}_m\|_1 \ge \gamma \; . \end{cases}$$

- **Pruning:** Only the $\mathbf{b_m}$ with $\|\mathbf{b_m}\|_1 \ge \gamma$ contribute to the sum in $g(\gamma)$.
- Problem reduces to finding γ^* though a root-finding procedure over g. We use Steffensen's root search [3]:

$$\gamma_{n+1} := \gamma_n + \gamma_n \frac{y_n - \gamma_n}{g(y_n) - g(\gamma_n)}, \quad y_n = \gamma_n + \delta_n |g(\gamma_n)| \; . \quad (3)$$

- **Initial Point:** Compute $\sigma_k = \|\mathrm{shrink}(b_k, \tau)\|_1$ for each row of B and take $\gamma_0 = \max_k(\sigma_k)$. It can be shown that $0 \le \gamma_0 \le \gamma^*$.

Results: simulations

- Synthetic $\mathbf{B} \sim \mathcal{U}([-0.5, 0.5])$ (100 realizations). Constraint $\tau = \alpha\|\mathbf{B}\|_{\infty,1}$, $\alpha \in \{10^{-4},\, 5 \times 10^{-4},\, 10^{-3}\}$

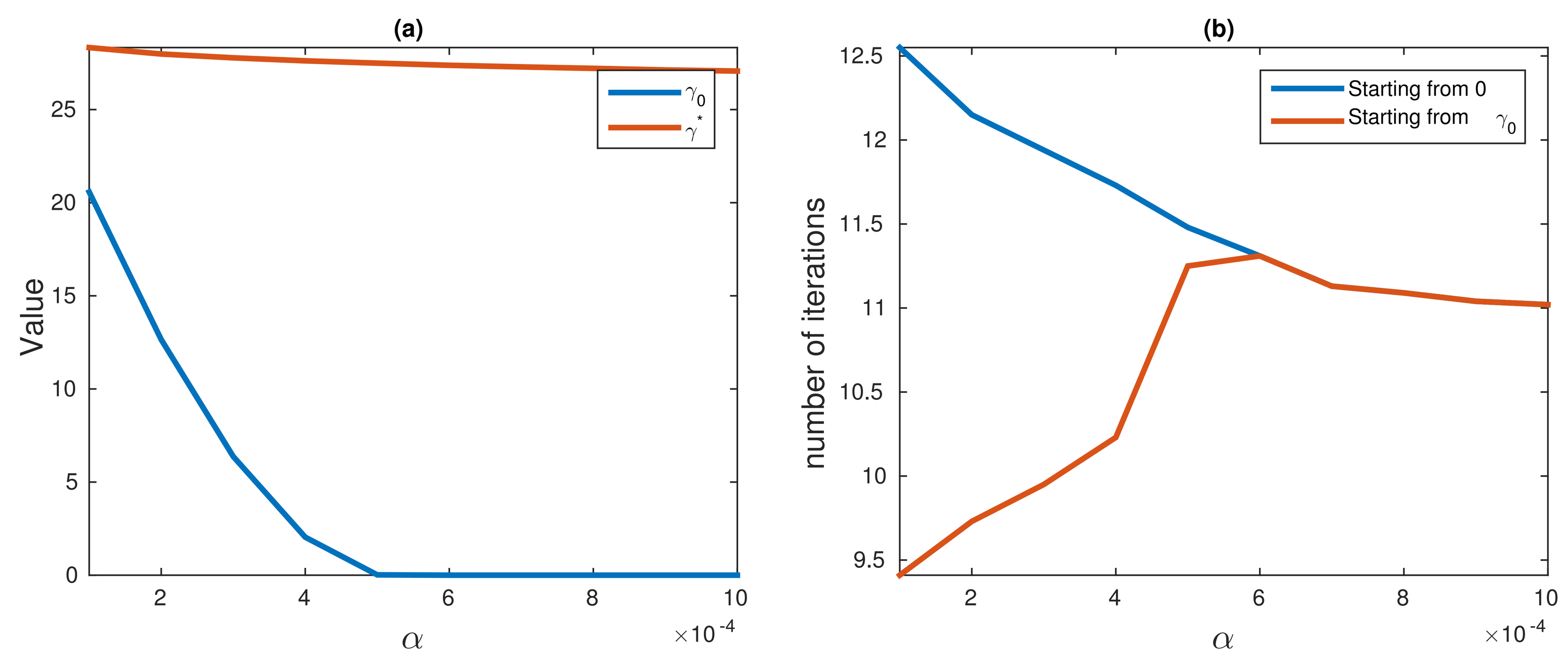

Figure 1: Impact of initial point

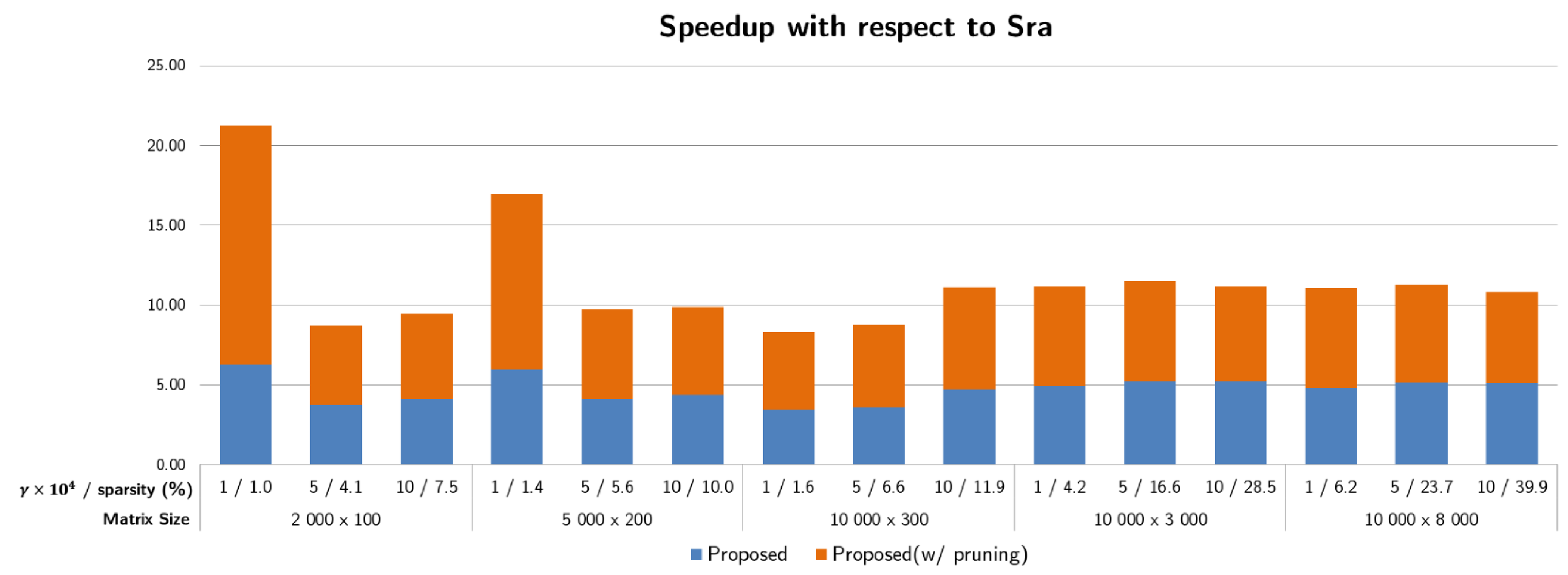

Figure 2: Speedup w.r.t. to Sra [2]

Results: fMRI task results

- Data from fMRI prediction of word response based on co-occurrence matrix [4].
- Solve $\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{Y} - \mathbf{XW}\|_2^2 \;\; s.t. \|\mathbf{W}\|_{\infty,1} \le \tau$ by projected gradient descent (PGD). In each step we use our proposed algorithm or Sra [2].

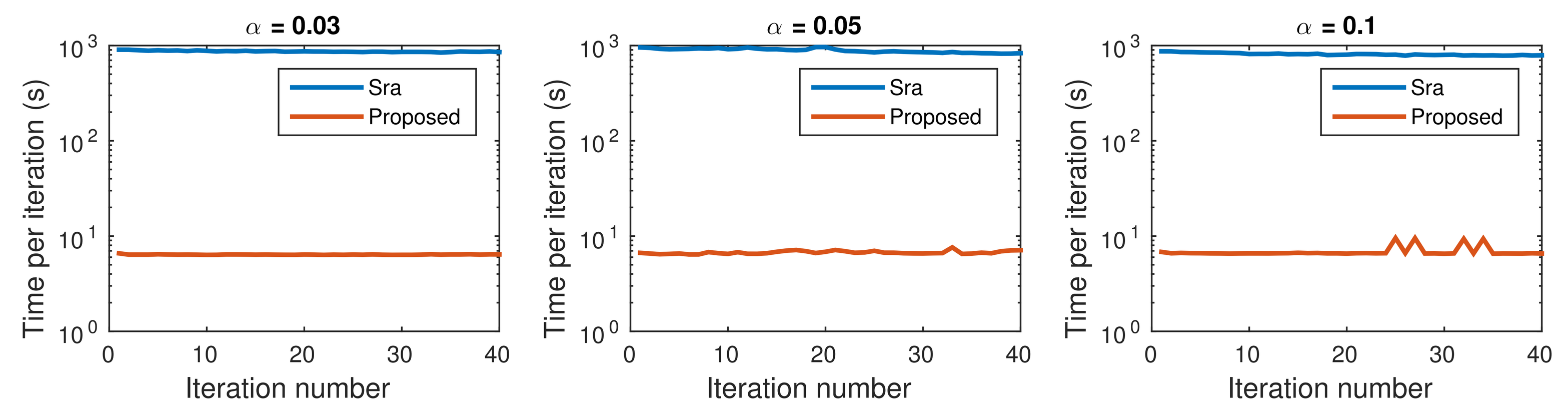

Figure 3: Time per PGD iteration for solving the $\ell_{\infty,1}$ projection problem. Speedups of ∼120 (10 hours → 3 minutes)

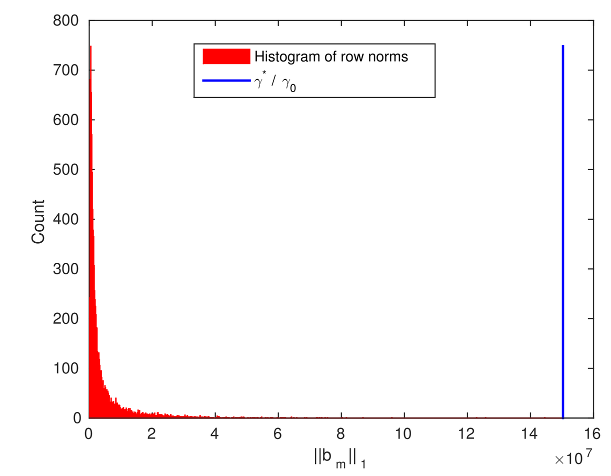

Figure 4: Distribution of $\|\mathbf{b_m}\|_1$ values (red) and optimal γ value (blue). This data distribution explains the higher speedups obtained in the fMRI dataset.

Conclusion

- New algorithm for projection onto the $\ell_{\infty,1}$-norm ball with speedups of around 5 – 6 times or more. Higher speedups with favorable sparsity conditions or data distributions.