External Division of Two Proximity Operators: An Application to Signal Recovery with Structured Sparsity

Kyohei Suzuki, Masahiro Yukawa

Department of Electronics and Electrical Engineering, Keio University, Japan

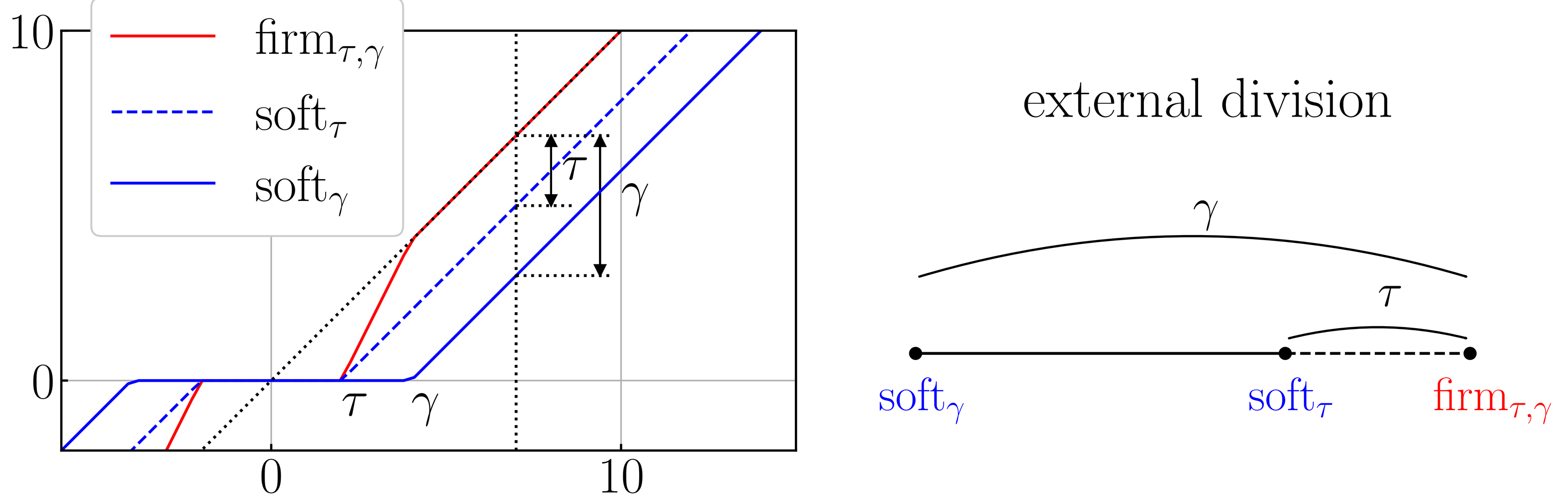
ICASSP 2024 in Seoul, Korea (14–19 Apr.)

Paper ID: 3640 (SPTM-P1.3)

Proposed Method

External Division Operator

For estimation of sparse signals, estimation accuracy: firm shrinkage [1] > soft shrinkage



External division operator

For any γ, τ ($\gamma > \tau > 0$),

$$\mathrm{firm}_{\tau,\gamma} = \frac{\gamma}{\gamma-\tau}\mathrm{soft}_\tau - \frac{\tau}{\gamma-\tau}\mathrm{soft}_\gamma \quad \text{(Proposition 1)}$$

$$= \frac{\gamma}{\gamma-\tau}\mathrm{Prox}_{\tau\|\cdot\|_1} - \frac{\tau}{\gamma-\tau}\mathrm{Prox}_{\gamma\|\cdot\|_1}$$

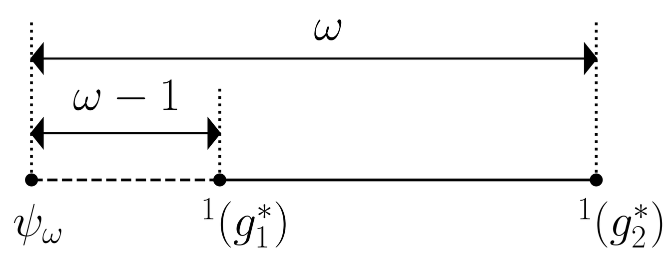
generalization ↓

$$T_\omega := \omega\mathrm{Prox}_{g_1} - (\omega-1)\mathrm{Prox}_{g_2}$$

($\omega > 1$, $g_1, g_2 : \mathbb{R}^n \to \mathbb{R}$: convex functions)

Proposition 2

Set $\psi_\omega := \omega({}^{1}(g_1^*)) - (\omega-1)({}^{1}(g_2^*))$.
Then, $T_\omega = \nabla\psi_\omega$.
If ψ_ω is convex, T_ω is ω-Lipschitz continuous ($\Leftrightarrow \omega^{-1}$-cocoercive).



- The Moreau envelope of $f : \mathbb{R}^n \to \mathbb{R}$ of index $\gamma > 0$ is defined as

$${}^{\gamma}f : \boldsymbol{x} \mapsto \min_{\boldsymbol{\xi}\in\mathbb{R}^n}\left(f(\boldsymbol{\xi}) + \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{\xi}\|_2^2\right)$$

- The Fenchel conjugate of convex function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$f^* : \boldsymbol{z} \mapsto \sup_{\boldsymbol{x}\in\mathbb{R}^n}\langle\boldsymbol{x},\boldsymbol{z}\rangle_2 - f(\boldsymbol{x})$$

Convergence Analysis

Suppress a given fidelity $f : \mathbb{R}^n \to \mathbb{R}$ while accomodating the prior information with the operator T_ω.

For a convex function ψ_ω, $T_\omega = \nabla\psi_\omega$ implies that [2]

$$T_\omega = \mathrm{Prox}_{\varphi_\omega} \qquad \left(\varphi_\omega := \psi_\omega^* - \frac{1}{2}\|\cdot\|_2^2 \text{ is } (1-\omega^{-1})\text{-weakly convex}\right)$$

Thanks to Proposition 2, if f is ρ-strongly convex, $\omega := (1-\mu\rho)^{-1} > 1$, and $\mu \in \left(0, \frac{2}{\sigma+\rho}\right)$, the sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ produced by

$$\boldsymbol{x}_{k+1} := T_\omega(\boldsymbol{x}_k - \mu\nabla f(\boldsymbol{x}_k)) \quad (\mu > 0: \text{ step size})$$

converges to a minimizer of the following problem [2]:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \mu f(\boldsymbol{x}) + \varphi_\omega(\boldsymbol{x}).$$

Can we guarantee convergence even for the underdetermined linear systems?

- $f(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$ for $\boldsymbol{y}\in\mathbb{R}^m$ and $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ with $m < n$
- f is strongly convex only on $\mathcal{M} := \mathrm{Null}^{\perp}(\boldsymbol{A})$

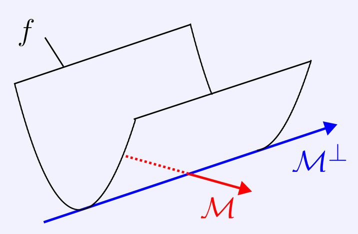
Proposition 3

Let $\mathcal{M} \subset \mathbb{R}^n$ be a subspace. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function such that
(i) $f - \frac{\rho}{2}\|P_{\mathcal{M}}\cdot\|_2^2$ is convex
(ii) $\nabla f(\boldsymbol{x}) \in \mathcal{M}$ for all $\boldsymbol{x}\in\mathbb{R}^n$
(iii) ∇f is σ-Lipschitz continuous for $\rho, \sigma > 0$ ($\sigma \geq \rho$).

Let $g_1, g_2 : \mathbb{R}^n \to \mathbb{R}$ be convex such that ψ_ω is convex. Let $\mu \in \left(0, \frac{2}{\sigma+\rho}\right)$ be the step size parameter, and set $\omega := (1-\mu\rho)^{-1} > 1$. Then, given an arbitrary $\boldsymbol{x}_0 \in \mathbb{R}^n$, the sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ produced by

$$\boldsymbol{x}_{k+1} := T_\omega(\boldsymbol{x}_k - \mu(\nabla f(\boldsymbol{x}_k) + \rho P_{\mathcal{M}^\perp}\boldsymbol{x}_k))$$

converges to a minimizer of the following cost function if exists:

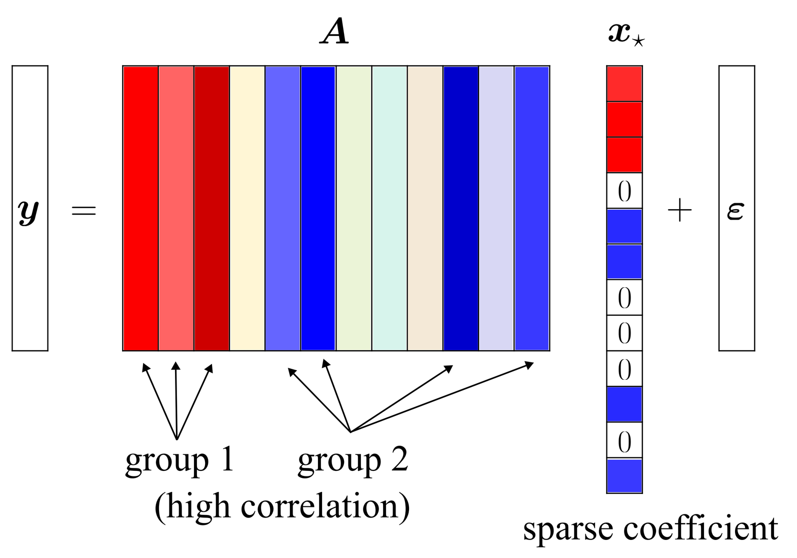


$$\underbrace{\mu f}_{\substack{\mu\rho\text{-strongly convex on } \mathcal{M},\\ \text{convex on } \mathcal{M}^\perp}} + \underbrace{\varphi_\omega + \frac{\mu\rho}{2}\|P_{\mathcal{M}^\perp}\cdot\|_2^2}_{\substack{\mu\rho\text{-weakly convex on } \mathcal{M},\\ \text{convex on } \mathcal{M}^\perp}}$$

The debiased effect is restricted on $\mathcal{M}$ while preserving the overall convexity.

Numerical Examples

Supervised Clustering



- Task: From given $\boldsymbol{y}$, $\boldsymbol{A}\in\mathbb{R}^{m\times n}$, estimate the sparse coefficients $\boldsymbol{x}_\star \in \mathbb{R}^n$
→ Group the important variables based on $\hat{\boldsymbol{x}}$
- Applications: gene expression analysis, brain imaging, protein-protein interaction networks analysis, etc

Lasso tends to select only one variable from a group of highly correlated inputs

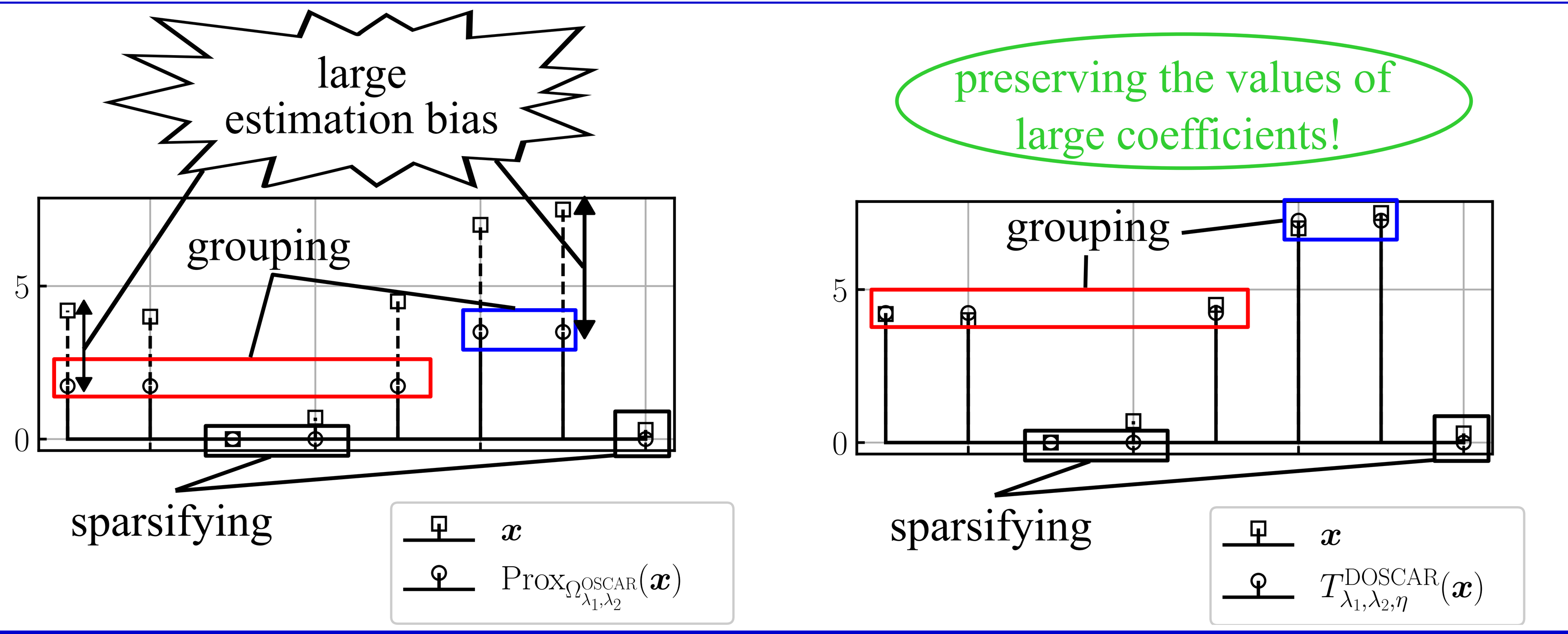
→ OSCAR (octagonal shrinkage and clustering algorithm for regression[3])

$$\Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}} : \boldsymbol{x} \mapsto \lambda_1\|\boldsymbol{x}\|_1 + \lambda_2\sum_{i<j}\max\{|x_i|, |x_j|\}$$

Large penalty for large coefficients

Specific example: Debiased OSCAR (DOSCAR)

$$T_{\lambda_1,\lambda_2,\omega,\eta}^{\mathrm{DOSCAR}} := \omega\,\mathrm{Prox}_{\Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}}} - (\omega-1)\,\mathrm{Prox}_{\eta\Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}}} \quad (\lambda_1,\lambda_2 > 0, \omega, \eta > 1)$$



Proposition 4

For any $\omega, \eta > 1$, $\psi_\omega := \omega({}^{1}((\Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}})^*)) - (\omega-1)({}^{1}((\eta\Omega_{\lambda_1,\lambda_2}^{\mathrm{OSCAR}})^*))$ is convex.

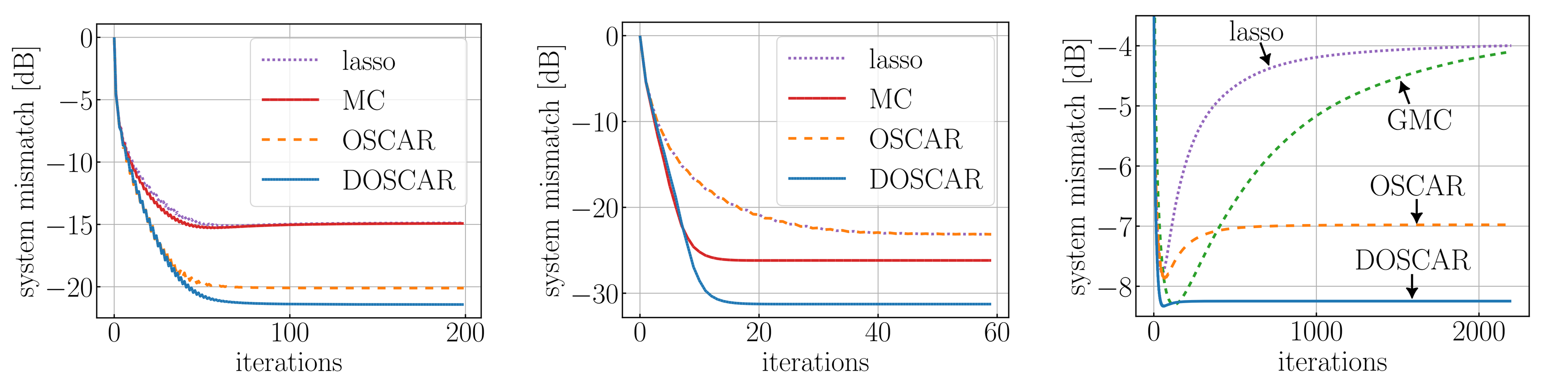
→ Convergence is guaranteed owing to Proposition 2.

Experiment

Measurement: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_\star + \boldsymbol{\varepsilon} \in \mathbb{R}^m$ ($\boldsymbol{\varepsilon} \sim$ i.i.d. $\mathcal{N}(0, \sigma_\varepsilon^2)$)

1. dataset A (overdetermined, high correlation among variables)
   - $\boldsymbol{A}\in\mathbb{R}^{m\times n}$: generated from Gaussian distribution with mean $\boldsymbol{0}$, covariance $\mathrm{cov}(\boldsymbol{a}_i, \boldsymbol{a}_j) = 0.7^{|i-j|}$ ($m = 100$, $n = 40$)
   - $\boldsymbol{x}_\star := [\underbrace{0\ldots0}_{10}, \underbrace{2\ldots2}_{10}, \underbrace{0\ldots0}_{10}, \underbrace{2\ldots2}_{10}]^\top \in \mathbb{R}^{40}$
2. dataset B (overdetermined, low correlation among variables)
   - Same as dataset A except that $\boldsymbol{A} \sim$ i.i.d. standard Gaussian distribution
3. dataset C (underdetermined)
   - Same as dataset A except that $m = 30$, $n = 60$ and $\boldsymbol{x}_\star := [\underbrace{0\ldots0}_{10}, \underbrace{2\ldots2}_{10}, \underbrace{0\ldots0}_{10}, \underbrace{2\ldots2}_{10}, \underbrace{0\ldots0}_{20}]^\top \in \mathbb{R}^{60}$

$\mathrm{SNR} := \frac{\|\boldsymbol{A}\boldsymbol{x}_\star\|_2^2}{\|\boldsymbol{\varepsilon}\|_2^2}$: 20 dB, system mismatch $:= \frac{\|\hat{\boldsymbol{x}}-\boldsymbol{x}_\star\|_2^2}{\|\boldsymbol{x}_\star\|_2^2}$ ($\hat{\boldsymbol{x}}$: estimate)



(a) dataset A (high correlation) (b) dataset B (low correlation) (c) dataset C (underdetermined case)

- The performance of OSCAR deteriorates when the correlation is low.
- Proposed method outperforms the other methods no matter if the explanatory variables have correlations.

Conclusion

1. We studied the properties of the external division operator and proposed a debiased estimator for signals with structured sparsity.
2. The convergence conditions for the algorithm based on the external division were provided.
3. Numerical examples demonstrated that the performance of the proposed operator exhibits a significant improvement over that of OSCAR.

References

[1] H.-Y. Gao and A. G. Bruce, "Waveshrink with firm shrinkage," *Statistica Sinica*, vol. 7, no. 4, pp. 855–874, 1997.
[2] M. Yukawa and I. Yamada, "Cocoercive Gradient Operator and Its Associated Weakly Convex Function: A Generalization of Moreau's Proximity Operator for Case of Unique Minimizer," *Proc. IEICE Signal Processing Symposium*, 2023.
[3] H.D. Bondell, and B.J. Reich, "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR." *Biometrics*, 2007.