A Proximal Splitting Algorithm for Linearly Involved Generalized Minimax Concave Penalty Models

Isao Yamada
Tokyo Institute of Technology

Based on a joint work with Jiro Abe & Masao Yamagishi

2nd Workshop on Operator Splitting Methods in Data Analysis
@ Center for Computational Mathematics, Flatiron Institute, NY

March 20, 2019, New York
Thanks for invitation to this dreaming place!

One of my most favorite buildings on the globe!

It was 30 years ago when I came here for the 1st time.

I was so impressed with this beautiful shape that I took many photos from the very low angle lying on the street.
Thanks for invitation to this dreaming place!

I believe you also like this!

Why?
Thanks for invitation to this dreaming place!

I believe you also like this!

Why?

Because we are very strange persons who like optimization subject to severe constraints!
Very severe constraints found in ...
Many problems in sparsity-aware signal processing and data analytics have been translated into

$$\min_{\mathbf{x} \in \mathbb{R}^n} J_{\Psi \circ L}(\mathbf{x}) := \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 + \mu \Psi \circ L(\mathbf{x}), \mu > 0,$$

where \( \mathbf{y} \in \mathbb{R}^m, \mathbf{A} \in \mathbb{R}^{m \times n}, L \in \mathbb{R}^{l \times n} \) and \( \Psi : \mathbb{R}^l \rightarrow \mathbb{R}_+ \) is a certain approximation of \( \| \cdot \|_0 \).

Popular Examples of \( \Psi \):

$$\|(z_1, z_2, \ldots, z_l)\|_p := \left( \sum_{k=1}^l |z_k|^p \right)^{1/p}, 0 < p \leq 1.$$

\( \| \cdot \|_1 \) has been a standard choice due to \underline{convexity}.

..., LASSO [Tibishirani '96], TV [Rudin, Osher, Fatemi'92], ...
Convexity is certainly a key for global optimization

$$\min_{x \in \mathbb{R}^n} J_{\Psi \circ L}(x) := \frac{1}{2} \|y - Ax\|^2 + \mu \Psi \circ L(x), \quad \mu > 0,$$

But

Convex relaxation of $\| \cdot \|_0$ would not be an only realistic compromise for convexity of $J_{\Psi \circ L}$

For better sparsity, How about designing nonconvex $\Psi$ to ensure the overall convexity of $J_{\Psi \circ L}$?

\[ \ldots [\text{Blake, Zisserman'87}, [\text{Nikolova'98}, [\text{Nikolova'11]} \ldots
\]
\[ [\text{Möllenhoff, Strekalovskiy, Moeller, Cremers'15}] \ldots
\[ [\text{Bayram'16}, [\text{Lanza, Morigi, Sgallari'16}] \ldots [\text{Selesnick'17}] \ldots \]
Generalized Minimax Concave (GMC) Penalty [Selesnick '17]

\[(B \in \mathbb{R}^{m \times l}) \Psi_B : \mathbb{R}^l \rightarrow \mathbb{R}_+ : z \mapsto \|z\|_1 - \min_{v \in \mathbb{R}^l} \left[ \|v\|_1 + \frac{1}{2} \|B(z - v)\|^2 \right]\]

is a parametric penalty function satisfying

Nonconvex but includes \(\| \cdot \|_1\) as special case

\[\Psi_B(z) = \|z\|_1 - \frac{1}{2} \|Bz\|^2 \Leftrightarrow \|B^\top Bz\|_\infty \leq 1,\]

\[\Psi_B = \begin{cases} \| \cdot \|_1 & \text{if } B = O_{m \times l}, \\ \text{nonconvex} & \text{otherwise}. \end{cases}\]

Overall convexity for \(L = I_d\)

For \((A, B, \mu, y) \in \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n} \times \mathbb{R}_{++} \times \mathbb{R}^m,\)

\[A^\top A - \mu B^\top B \succeq O_n \Rightarrow \frac{1}{2} \|y - A(\cdot)\|^2 + \mu \Psi_B(\cdot) \in \Gamma_0(\mathbb{R}^n)\]
Generalized Minimax-Concave (GMC) Penalty


\[ (B \in \mathbb{R}^{m \times l}) \Psi_B(z) := \|z\|_1 - \min_{v \in \mathbb{R}^l} \left[ \|v\|_1 + \frac{1}{2} \|B(z - v)\|^2 \right] \]

Parametrized Multidimensional Extension

Minimax-Concave (MC) Penalty [Zhang '10]

\[ \psi : \mathbb{R} \rightarrow \mathbb{R} : z \mapsto \psi(z) := \begin{cases} |z| - \frac{1}{2} z^2, & \text{if } |z| \leq 1 \\ \frac{1}{2}, & \text{otherwise} \end{cases} \]

Generalized Minimax-Concave (GMC) Penalty


\[(B \in \mathbb{R}^{m \times l})\quad \Psi_B(z) := \|z\|_1 - \min_{v \in \mathbb{R}^l} \left[ \|v\|_1 + \frac{1}{2} \|B(z - v)\|^2 \right]\]

I like this beautiful idea of GMC penalty very much because it has a great potential for dealing with many nice nonconvex penalties under a single umbrella of the modern convex analysis.
We are interested in global optimization of

\[
\min_{x \in \mathbb{R}^n} J_{\Psi_B \circ L}(x) := \frac{1}{2} \|y - Ax\|^2 + \mu \Psi_B \circ L(x), \quad \mu > 0, \tag{1}
\]

where \(y \in \mathbb{R}^m\), \(A \in \mathbb{R}^{m \times n}\), \(L \in \mathbb{R}^{l \times n}\), \(B \in \mathbb{R}^{m \times l}\), and

\[
\Psi_B(\cdot) := \|\cdot\|_1 - \min_{\mathbf{v} \in \mathbb{R}^l} \left[ \|\mathbf{v}\|_1 + \frac{1}{2} \|B(\cdot - \mathbf{v})\|^2 \right],
\]

under as much general overall-convexity cond for \(B\) as possible!

Q1. What is general cond for \(B\) to ensure overall-convexity of (1)?

\[[\text{Selesnick '17}]\]

\[
A^\top A - \mu B^\top B \succeq O_n \Rightarrow \frac{1}{2} \|y - A(\cdot)\|^2 + \mu \Psi_B \circ \text{Id}(\cdot) \in \Gamma_0(\mathbb{R}^n)
\]

But

\[
\frac{1}{2} \|y - A(\cdot)\|^2 + \mu \Psi_B \circ L(\cdot) \in \Gamma_0(\mathbb{R}^n)
\]

does not seem to have been clarified yet.
We are interested in global optimization of

$$\min_{x \in \mathbb{R}^n} J_{\Psi_B \circ L}(x) := \frac{1}{2} \| y - Ax \|^2 + \mu \Psi_B \circ L(x), \quad \mu > 0,$$

(1)

where $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$, $L \in \mathbb{R}^{l \times n}$, $B \in \mathbb{R}^{m \times l}$, and

$$\Psi_B(\cdot) := \| \cdot \|_1 - \min_{v \in \mathbb{R}^l} \left[ \| v \|_1 + \frac{1}{2} \| B(\cdot - v) \|^2 \right],$$

under as much general overall-convexity condition for $B$ as possible!

Q2. Can we establish any iterative algorithm of guaranteed convergence to globally optimal solution of (1) under general overall-convexity condition?

[Selesnick '17]

A forward-backward splitting type algorithm is found but only for a special case: $L = \text{Id}$ and $B^T B = (\theta/\mu) A^T A$, $0 \leq \theta \leq 1$

Any algorithm applicable to fully general cases does not seem to have been reported yet.
We are interested in global optimization of

$$\min_{x \in \mathbb{R}^n} J_{\Psi_B \circ L}(x) := \frac{1}{2} \|y - Ax\|^2 + \mu \Psi_B \circ L(x), \quad \mu > 0, \quad (1)$$

where $\Psi_B(\cdot) := \| \cdot \|_1 - \min_{v \in \mathbb{R}^l} \left[ \| v \|_1 + \frac{1}{2} \| B(\cdot - v) \|^2 \right]$, 

Q1. What is general cond for $B$ to ensure overall-convexity of (1)?

Q2. Can we establish any iterative algorithm of guaranteed convergence to globally optimal solution of (1) under general overall-convexity condition?

Q3. Can we choose $B$ flexibly to ensure overall-convexity of (1)?

These questions would be keys for further evolutions as well as broader applications of the GMC penalties!

Straightforward limited applications of certain linearly involved GMC penalties have been reported very recently in [Du, Liu'18], [Zhong, Yi, Xiao, Zhang, Wu '18] without any theoretical consideration on these key questions.
We are interested in global optimization of

\[
\minimize_{x \in \mathbb{R}^n} J_{\Psi_B \circ L}(x) := \frac{1}{2} \|y - Ax\|^2 + \mu \Psi_B \circ L(x), \quad \mu > 0, \tag{1}
\]

where \( \Psi_B(\cdot) := \| \cdot \|_1 - \min_{v \in \mathbb{R}^l} \left[ \|v\|_1 + \frac{1}{2} \|B(\cdot - v)\|^2 \right] \).

Q1. What is general cond for \( B \) to ensure overall-convexity of (1)?

[AYY-ICASSP'19, Proposition 1]

\[
A^\top A - \mu L^\top B^\top BL \succeq O_n
\]

\[
\frac{1}{2} \|y - A(\cdot)\|^2 + \mu \Psi_B \circ L(\cdot) \in \Gamma_0(\mathbb{R}^n)
\]

We are interested in global optimization of

\[
\min_{x \in \mathbb{R}^n} J_{\Psi_B \circ L}(x) := \frac{1}{2} \|y - Ax\|^2 + \mu \Psi_B \circ L(x), \quad \mu > 0, \tag{1}
\]

where

\[
\Psi_B(\cdot) := \|\cdot\|_1 - \min_{v \in \mathbb{R}^l} \left[ \|v\|_1 + \frac{1}{2} \|B(\cdot - v)\|^2 \right],
\]

Q2. Can we establish any iterative algorithm of guaranteed convergence to globally optimal solution of (1) under general overall-convexity condition?

[AYY-ICASSP'19, Theorem 1]

Although \(\Psi_B\) is nonsmooth and nonconvex, we can express the set of all globally optimal solutions in terms of the fixed-point set of computable nonexpansive operator in a certain Hilbert space and therefore can solve (1).

Theorem 1 (ICASSP'19) Assume \( \frac{1}{2} \| y - A(\cdot) \|^2 + \mu \Psi_B \circ L(\cdot) \in \Gamma_0(\mathbb{R}^n) \) and define \( T_{LCP} : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^l : (x, v, w) \mapsto (\xi, \zeta, \eta) \) by

\[
\xi := \left[ \text{Id} - \frac{1}{\sigma} \left( A^\top A - \mu L^\top B^\top B L \right) \right] x - \frac{\mu}{\sigma} L^\top B^\top B v - \frac{\mu}{\sigma} L^\top w + \frac{1}{\sigma} A^\top y
\]

\[
\zeta := \text{Soft}_{\frac{\mu}{\tau}} \left[ \frac{2\mu}{\tau} B^\top B L \xi - \frac{\mu}{\tau} B^\top B L x + \left( \text{Id} - \frac{\mu}{\tau} B^\top B \right) v \right]
\]

\[
\eta := P_{[-1,1]^l} \left( 2L \xi - L x + w \right),
\]

where \( (\sigma, \tau, \kappa) \in \mathbb{R}_{++} \times \mathbb{R}_{++} \times (1, \infty) \) is chosen to satisfy

\[
\begin{bmatrix}
\sigma \text{Id} - \frac{\mu^2}{\tau} L^\top \left( B^\top B \right)^2 L - \mu L^\top L & \succ O_n \\
\sigma \text{Id} - \frac{\kappa}{2} A^\top A - \mu L^\top L & \succ O_n \\
\tau \geq \left( \frac{\kappa}{2} + \frac{2}{\kappa} \right) \mu \| B^\top B \|_2.
\end{bmatrix}
\]

\[
\arg\min J_{\Psi_B \circ L} = Q(\text{Fix}(T_{LCP})),
\]

where \( Q : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}^n : (x, v, w) \mapsto x \)

and \( \text{Fix}(T_{LCP}) := \{ (x, v, w) \mid T_{LCP}(x, v, w) = (x, v, w) \} \)
\[ P := \begin{bmatrix} \sigma \text{Id} & -\mu L^\top B^\top B & -\mu L^\top \\ -\mu B^\top B L & \tau \text{Id} & O_l \\ -\mu L & O_l & \mu \text{Id} \end{bmatrix} \succ O_{n+2l} \text{ and} \]

\[ T_{\text{LCP}} : \mathcal{H} (:= \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^l) \to \mathcal{H} \text{ is } \frac{\kappa}{2\kappa - 1} \text{-averaged nonexpansive in} \]

the Hilbert space \((\mathcal{H}, \langle \cdot, \cdot \rangle, \| \cdot \|_P), \text{ i.e.,} \)

\[(\forall z_1, z_2 \in \mathcal{H}) \]

\[ \| T_{\text{LCP}}(z_1) - T_{\text{LCP}}(z_2) \|_P \leq \| z_1 - z_2 \|_P - \frac{\kappa - 1}{\kappa} \| (\text{Id} - T_{\text{LCP}})(z_1) - (\text{Id} - T_{\text{LCP}})(z_2) \|_P \]

For any initial point \((x_0, v_0, w_0) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^l\),

the sequence \((x_n, v_n, w_n)_{n \in \mathbb{N}} \subset \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^l\) generated by

[Krasnosel'skii-Mann]

\[ (x_{n+1}, v_{n+1}, w_{n+1}) := T_{\text{LCP}}(x_n, v_n, w_n) \]

converges to a point \((x^*, v^*, w^*) \in \text{Fix}(T_{\text{LCP}})\) and

\[ \lim_{n \to \infty} x_n = x^* \in \text{argmin} \ J_{\Psi_B \circ L} \]
\[ P := \begin{bmatrix} \sigma \text{Id} & -\mu L^\top B^\top B & -\mu L^\top \\ -\mu B^\top B L & \tau \text{Id} & O_l \\ -\mu L & O_l & \mu \text{Id} \end{bmatrix} \succ O_{n+2l} \text{ and} \]

\[ T_{\text{LCP}} : \mathcal{H} := \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^l \rightarrow \mathcal{H} \text{ is } \frac{\kappa}{2\kappa-1} \text{-averaged nonexpansive in} \]

the Hilbert space \((\mathcal{H}, \langle \cdot, \cdot \rangle_P, \| \cdot \|_P)\), i.e.,

\[
(\forall z_1, z_2 \in \mathcal{H})
\]

\[
\|T_{\text{LCP}}(z_1) - T_{\text{LCP}}(z_2)\|_P \leq \|z_1 - z_2\|_P - \frac{\kappa - 1}{\kappa} \| (\text{Id} - T_{\text{LCP}})(z_1) - (\text{Id} - T_{\text{LCP}})(z_2)\|_P
\]

For any initial point \((x_0, v_0, w_0) \in \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^l\), the sequence \((x_n, v_n, w_n)_{n \in \mathbb{N}} \subset \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^l\) generated by

Further optimization is possible if you use HSDM!

[Krasnosolskii-Mann]

\[(x_{n+1}, v_{n+1}, w_{n+1}) := T_{\text{LCP}}(x_n, v_n, w_n)\]

converges to a point \((x^*, v^*, w^*) \in \text{Fix}(T_{\text{LCP}})\) and

\[
\lim_{n \to \infty} x_n = x^* \in \text{argmin} \ J_{\Psi_B \circ L}
\]
The proof sketch of [AYY-ICASSP'19,Theorem 1]

**STEP1**

Translate

$$\min_{x \in \mathbb{R}^n} J_{\Psi_B \circ L}(x) := \frac{1}{2} \|y - Ax\|^2 + \mu \Psi_B \circ L(x), \; \mu > 0,$$

**into a monotone inclusion** by using

Fact (Regular/Frèchet Subdifferential [e.g., Rockafellar-Wets’09])

Let $f \in \Gamma_0(\mathbb{R}^n)$. Assume $g : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable in a neighborhood of $\bar{x} \in \text{dom}(f)$ and $f + g \in \Gamma_0(\mathbb{R}^n)$.

$$\partial(f + g)(\bar{x}) = \hat{\partial}(f + g)(\bar{x}) = \hat{\partial}f(\bar{x}) + \nabla g(\bar{x}) = \partial f(\bar{x}) + \nabla g(\bar{x}),$$

where $\hat{\partial} \varphi(\bar{x}) := \left\{ u \in \mathbb{R}^n \mid \liminf_{x \to \bar{x}} \frac{\varphi(x) - \varphi(\bar{x}) - \langle u, x - \bar{x} \rangle}{\|x - \bar{x}\|} \geq 0 \right\}$. 

**STEP2** Follow [Condat’13/Vũ’13] + [Ogura-IY’02/Combettes-IY’15] to design an averaged operator whose fixed-point set can express the set of all solutions of the monotone inclusion.
How can we apply this optimization model?

\[
\min_{x \in \mathbb{R}^n} J_{\Psi_B \circ L}(x) := \frac{1}{2} \| y - Ax \|^2 + \mu \Psi_B \circ L(x), \quad \mu > 0,
\]

where \( \Psi_B(\cdot) := \| \cdot \|_1 - \min_{v \in \mathbb{R}^l} \left[ \| v \|_1 + \frac{1}{2} \| B(\cdot - v) \|^2 \right] \).

Q3. Can we choose \( B \) flexibly to ensure overall-convexity of (1)?

[AYY-ICASSP'19, Proposition 2]

Suppose \( L \in \mathbb{R}^{l \times n} \) satisfies \( \text{rank}(L) = l \).

Choose a nonsingular \( \tilde{L} \in \mathbb{R}^{n \times n} \), s.t., \[
\begin{bmatrix}
O_{l \times (n-l)} & I_l
\end{bmatrix}
\tilde{L} = L.
\]

\[
B_\theta := \sqrt{\frac{\theta}{\mu \Lambda^{1/2} U^T}}, \quad \theta \in [0, 1],
\]

ensures the convexity of \( J_{\Psi_{B_\theta \circ L}} \),

where \( \Lambda U U^T := \tilde{A}_2^T \tilde{A}_2 - \tilde{A}_2^T \tilde{A}_1 \left( \tilde{A}_1^T \tilde{A}_1 \right)^\dagger \tilde{A}_1^T \tilde{A}_2 \in \mathbb{R}^{l \times l} \)

is the EVD with \( A(\tilde{L})^{-1} = \begin{bmatrix} \tilde{A}_1 & \tilde{A}_2 \end{bmatrix} \).

Numerical Experiments (Convexity-Edge-Preserving smoother)

\[
\text{minimize} \quad J_{\Psi_{B_{\theta}}} \circ L(x) := \frac{1}{2} \| y - Ax \|^2 + \mu \Psi_{B_{\theta}} \circ L(x), \quad \mu > 0.
\]

We applied the proposed algorithm to

\[
y = Ax_{\star} + \varepsilon
\]

\[
\begin{aligned}
x_{\star} \in \mathbb{R}^{128} : & \text{ a piecewise constant signal} \\
A \in \mathbb{R}^{100 \times 128} : & \text{ drawn from } \mathcal{N}(0,1) \text{ and fixed} \\
\varepsilon \in \mathbb{R}^{100} : & \text{ AWG to achieve SNR (15dB)}
\end{aligned}
\]

To achieve convexity-edge-preserving smoother, we used

\[
L = D := \begin{bmatrix}
-1 & 1 \\
& \ddots & \ddots \\
& & -1 & 1
\end{bmatrix} \in \mathbb{R}^{(n-1) \times n}, \quad \tilde{L} = \tilde{D} := \begin{bmatrix}
1 & 0 & \cdots & 0 \\
0 & & & \ddots \\
\vdots & \ddots & & \\
0 & & & D
\end{bmatrix} \in \mathbb{R}^{n \times n}
\]

and \( B_{\theta} := \sqrt{\theta/\mu} \Lambda^{1/2} U^T \), \( \theta \in [0,1] \)

\( \theta := 0 \) (Standard TV)

\( \theta := 0.9 \) (Nonconvex)
minimize \( J_{\Psi_{B_{\theta}}} \circ D(x) := \frac{1}{2} \| y - Ax \|^2 + \mu \Psi_{B_{\theta}} \circ D(x) \), \( \mu > 0 \).

**Fig. 1** Dependency of recovering performance on \( \mu \).

MSE (mean squared error): Average of \( \| x_{20000} - x_* \|^2 \) over 100 independent realizations of \( \varepsilon \).
\[
\minimize_{x \in \mathbb{R}^n} J_{\Psi_{B_\theta} \circ D}(x) := \frac{1}{2} \| y - Ax \|^2 + \mu \Psi_{B_\theta} \circ D(x), \quad \mu > 0.
\]

**Fig. 1** Dependency of recovering performance on $\mu$.
MSE (mean squared error): Average of $\| x_{20000} - x_* \|^2$ over 100 independent realizations of $\varepsilon$. 
Fig. 2 $\|x_k - x_*\|^2$ vs the iteration number $k$
Fig. 3 Recovered signals by the proposed algorithm after 20000 iterations
Fig. 4 Noise suppression by Standard TV model and Proposed model: Entries in $y - Ax_\star$, $Ax_{TV} - Ax_\star$, and $Ax_{prop} - Ax_\star$
We have introduced a proximal splitting algorithm which has theoretical guarantee of convergence to a globally optimal solution of the linearly involved generalized minimax concave (GMC) penalty models.

The proposed algorithm can be applied to fully general cases of these beautiful nonconvex penalty models under overall-convexity condition.


Further extension of the proposed algorithm will be presented at ICCOPT'2019, Aug., 2019.
Conclusion and Open Questions

We have introduced a proximal splitting algorithm which has theoretical guarantee of convergence to a globally optimal solution of the linearly involved generalized minimax concave (GMC) penalty models.

The proposed algorithm can be applied to fully general cases of these beautiful nonconvex penalty models under overall-convexity condition.


Further extension of the proposed algorithm will be presented at ICCOPT'2019, Aug., 2019.

Is there any clever statistical strategy for tuning good parameters $B \in \mathbb{R}^{m \times l}$ and $\mu > 0$? How about extension of TREX to GMC penalty models?