A Proximal ADMM for Multiblock Problems with Block Anti-Upper Triangular Constraints

Zhanwang Deng[1,2], Yuqiu Su[2], Wen Huang[2]

[1]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China

[2]School of Mathematical Sciences, Xiamen University, Xiamen, China

Problem Formulation

$$\min_{\mathbf{X}_1,\ldots,\mathbf{X}_n} \sum_{i=1}^{n} f_i(\mathbf{X}_i)$$
$$\text{s.t.}\quad \mathcal{A}_{j,1}\mathbf{X}_1 + \cdots + \mathcal{A}_{j,n-j+1}\mathbf{X}_{n-j+1} = \mathbf{b}_j, \quad j = 1,\ldots,m.$$
$$\begin{bmatrix} \mathcal{A}_{1,1} & \mathcal{A}_{1,2} & \cdots & \mathcal{A}_{1,n-1} & \mathcal{A}_{1,n} \\ \mathcal{A}_{2,1} & \mathcal{A}_{2,2} & \cdots & \mathcal{A}_{2,n-1} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathcal{A}_{m,1} & \mathcal{A}_{m,2} & \cdots & \mathcal{A}_{m,n-m+1} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{bmatrix}.$$

where $f_i : \mathbb{R}^{m_i\times p_i} \to (-\infty,\infty]$ is a proper closed function for $i = 1,\cdots,n$. $\mathcal{A}_{i,j} : \mathbb{R}^{m_j\times p_j} \to \mathbb{R}^{q_i}, i = 1,\cdots,m, j = 1,\cdots,n-i+1$ are linear operators, and $\mathbf{b}_i \in \mathbb{R}^{q_i}$, for $i = 1,\cdots,m$. It is assumed in this paper that $\mathrm{range}(\mathcal{A}_{i,j}) \subseteq \mathrm{range}(\mathcal{A}_{i,n-i+1})$ for $i = 1,\cdots,m$ and $j = 1,\cdots,n-i$. Problem (1) finds widespread applications across various domains. In what follows, we provide several instances where it is applicable.

Representative Structured Applications

Example 1: Consensus problem.

$$\min_{\mathbf{X}_0,\mathbf{X}_1,\cdots,\mathbf{X}_n} \sum_{i=1}^{n} f(\mathbf{X}_i) + h(\mathbf{X}_0), \qquad \text{s.t. } \mathbf{X}_0 = \mathbf{X}_i,\ i = 1,\cdots,n.$$

Example 2: Block-angular problem.

$$\min_{\mathbf{X}_i} \sum_{i=1}^{n+1} \left[\frac{1}{2}\langle \mathbf{X}_i, \mathcal{Q}_i(\mathbf{X}_i)\rangle + \langle \mathbf{C}_i, \mathbf{X}_i\rangle\right],$$
$$\text{s.t.} \begin{bmatrix} \mathcal{A}_{1,1} & \cdots & \mathcal{A}_{1,n} & \mathcal{A}_{1,n+1} \\ 0 & \cdots & 0 & \mathcal{A}_{2,n+1} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \mathcal{A}_{n,n} & \mathcal{A}_{n,n+1} \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_{n+1} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}.$$

Example 3: Low-rank and sparse image processing model.

$$\min_{\mathbf{X},\mathbf{W}} \sum_i \phi(\sigma_i(\sqrt{\mathbf{X}^{\mathrm{T}}\mathbf{X} + \varepsilon^2})) + \lambda p(\mathbf{W}), \qquad \text{s.t. } \mathbf{X} = \mathbf{W},\ \mathcal{P}_\Omega(\mathcal{A}(\mathbf{W})) = \mathcal{P}_\Omega(\mathbf{Y}).$$

Example 4: Robust regression problem.

$$\min_{\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_3} g(\mathbf{X}_1) + \alpha_1\left\{\alpha_2 P(\mathbf{X}_2) + \frac{1}{2}(1-\alpha_2)\|\mathbf{X}_3\|_2^2\right\}, \quad \text{s.t. } \mathcal{H}(\mathbf{X}_1) = \mathbf{X}_2,\ \mathcal{H}(\mathbf{X}_1) = \mathbf{X}_3.$$

Why the New Assumption Matters

Note that Problem (1) can be formulated as

$$\min_{\mathbf{X}_i} \quad f_1(\mathbf{X}_1) + \cdots + f_n(\mathbf{X}_n),$$
$$\text{s.t.} \quad \mathcal{A}_1(\mathbf{X}_1) + \cdots + \mathcal{A}_n(\mathbf{X}_n) = \mathbf{b},$$

by rewriting $\mathcal{A}_i = \left[\mathcal{A}_{1,i}^{\mathrm{T}} \ \ldots \ \mathcal{A}_{m,i}^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\mathbf{b} = \left[\mathbf{b}_1^{\mathrm{T}} \ \ldots \ \mathbf{b}_m^{\mathrm{T}}\right]^{\mathrm{T}}$. To the best of our knowledge, the commonly used methods, ADMM and its variants, require a "range assumption", i.e.,

$$\mathrm{range}(\mathcal{A}_i) \subseteq \mathrm{range}(\mathcal{A}_n), \qquad i = 1,\ldots,n-1,$$

for the global convergence analyses if $f_i, i = 1,\cdots,n$ is nonconvex. However, the range assumption fails for the consensus, block-angular, image-processing, and robust-regression examples. To address this limitation, we propose a weaker assumption, i.e.,

$$\boxed{\mathrm{range}(\mathcal{A}_{i,j}) \subseteq \mathrm{range}(\mathcal{A}_{i,n-i+1}), \qquad i = 1,\cdots,m,\ j = 1,\cdots,n-i.}$$

which holds for Problem (1). In the strongly convex setting, we discuss the necessity of the weaker assumption in the convergence analysis.

Nonconvex Related Works

Algorithm	Optimization problem	Assumptions
Proximal ADMM (this paper)	$\min_{x_i} \sum_{i=1}^n f_i(x_i)$, s.t. $\sum_{j=1}^{\min\{n-i+1,m\}} A_{i,j}x_i = 0,\ i = 1,\cdots,m$	Assumption 1.
Proximal linearized-ADMM [Yashtini (2022)]	$\min_{x,y} g(x,y) + f(x) + h(y)$, s.t. $Ax + By = 0$	g, h Lipschitz differentiable; $g+f+h$ lower bounded, coercive; B full column rank; $\mathrm{Im}(A) \subseteq \mathrm{Im}(B)$.
Proximal ADMM [Bolte et al. (2020)]	$\min_{x,y} g(x,y) + f(x) + h(y)$, s.t. $Ax + By = 0$	g, h Lipschitz differentiable; $g+f+h$ lower bounded, coercive; B full column rank; $\mathrm{Im}(A) \subseteq \mathrm{Im}(B)$.
Linearized ADMM [Liu et al. (2019)]	$\min_{x,y} g(x,y) + f(x) + h(y)$, s.t. $Ax + By = 0$	g, h Lipschitz differentiable; $g+f+h$ lower bounded, coercive; B full column rank; $\mathrm{Im}(A) \subseteq \mathrm{Im}(B)$.
ADMM [Wang et al. (2019)]	$\min_{x,y} g(x) + \sum_i f_i(x_i) + h(y)$, s.t. $Ax + By = 0$	g, h Lipschitz differentiable; f_0 lower semi-continuous, f_i prox-regular; $\mathrm{Im}(A) \subseteq \mathrm{Im}(B)$; subproblems have Lipschitz continuous solutions.

Strongly Convex Related Works

Algorithm	Optimization problem	Assumptions
Proximal ADMM (this paper)	$\min_{x_i} \sum_{i=1}^n f_i(x_i)$, s.t. $\sum_{j=1}^{\min\{n-i+1,m\}} A_{i,j}x_i = 0,\ i = 1,\cdots,m$	Assumptions 2 and 3, Tables 1 and 2.
ADMM [Lin et al. (2015)]	$\min_{\{x_i\}} \sum_{i=1}^n f_i(x_i)$, s.t. $\sum_{i=1}^n A_i x_i = 0$	Assumptions 2 and 3; strongly convex: $f_2,\ldots,f_N$; f_N is Lipschitz differentiable, A_N has full row rank, or f_N is strongly convex, $f_1,\ldots,f_N$ are Lipschitz differentiable, or $f_1,\ldots,f_N$ are Lipschitz differentiable, A_1 has full column rank.
Proximal ADMM [Deng & Yin (2016)]	$\min_{x,y} f(x) + g(y)$, s.t. $Ax + By = b$	Assumptions 2 and 3; Tables 1 and 2 for $m = 1$ and $n = 2$.
ADMM [Lin et al. (2018)]	$\min_{x_1,x_2,x_3} f_1(x_1) + f_2(x_2) + f_3(x_3)$, s.t. $A_1x_1 + A_2x_2 + A_3x_3 = b$	Assumptions 2 and 3; f_1 and f_2 are lower semi-continuous and coercive; f_3 is lower bounded and strongly convex; $f_3(x)$ is Lipschitz differentiable; A_1 and A_2 have full column rank; the condition number of f_3 is in $[1, 1.0798)$.
ADMM [Chen et al. (2025)]	$\min_{x_1,x_2,x_3} \theta_1(x_1) + \theta_2(x_2) + \theta_3(x_3)$, s.t. $A_1x_1 + A_2x_2 + A_3x_3 = b$	Assumptions 2 and 3. The sublevel set of the objective function is nonempty; θ_1 is weakly convex with modules μ_1, θ_2,θ_3 are strongly convex with modules μ_2 and μ_3, respectively; A_1 has full column rank; μ_1,μ_2, and μ_3 satisfy some given equalities.

Proximal ADMM

$$\mathcal{L}(\mathbf{X},\Lambda) = \sum_{i=1}^n f_i(\mathbf{X}_i) + \sum_{j=1}^m \left[\frac{\beta_j}{2}\left\|\sum_{\ell=1}^{n-j+1}\mathcal{A}_{j,\ell}\mathbf{X}_\ell - \mathbf{b}_j + \frac{\Lambda_j}{\beta_j}\right\|_{\mathrm{F}}^2 - \frac{1}{2\beta_j}\|\Lambda_j\|_{\mathrm{F}}^2\right].$$

$$\mathbf{X}_i^{(k+1)} \in \arg\min_{\mathbf{X}_i} \mathcal{L}(\mathbf{X}_1^{(k+1)},\ldots,\mathbf{X}_i,\mathbf{X}_{i+1}^{(k)},\ldots,\mathbf{X}_n^{(k)},\Lambda^{(k)}) + \frac{1}{2}\|\mathbf{X}_i - \mathbf{X}_i^{(k)}\|_{\mathcal{Q}_i^{(k)}}^2,$$

$$\Lambda_j^{(k+1)} = \Lambda_j^{(k)} + \beta_j\left(\sum_{\ell=1}^{n-j+1}\mathcal{A}_{j,\ell}\mathbf{X}_\ell^{(k+1)} - \mathbf{b}_j\right).$$

Let $\overline{\mathbf{X}}^{(k)} = [\mathbf{X}_1^{(k)},\cdots,\mathbf{X}_n^{(k)}]$ and $\overline{\mathbf{\Lambda}}^{(k)} = [\Lambda_1^{(k)},\cdots,\Lambda_m^{(k)}]$ denote the k-th iteration point for $\overline{\mathbf{X}}$ and $\overline{\mathbf{\Lambda}}$ respectively. In the primal update, $\mathbf{X}_i^{(k+1)}$ is a minimizer of the objective $\mathcal{L}(\overline{\mathbf{X}},\overline{\Lambda}) + \frac{1}{2}\left\|\mathbf{X}_i - \mathbf{X}_i^{(k)}\right\|_{\mathcal{Q}_i^{(k)}}^2$, where $\mathbf{X}_j = \mathbf{X}_j^{(k+1)}$ if $1 \le j \le i-1$ and $\mathbf{X}_j = \mathbf{X}_j^{(k)}$ if $i+1 \le j \le n$. The multiplier update is $\mathbf{\Lambda}_i^{(k+1)} = \mathbf{\Lambda}_i^{(k)} + \beta_i\left(\sum_{j=1}^{n-i+1}\mathcal{A}_{i,j}(\mathbf{X}_j^{(k+1)}) - \mathbf{b}_i\right)$.

Nonconvex Theory

Assumption 1. The lower semi-continuous functions $f_i, i = 1,\cdots,n-m$ are coercive and the functions $f_i, i = n-m+1,\cdots,n$ are L_{f_i} Lipschitz differentiable respectively and bounded from below; $\mathrm{range}(\mathcal{A}_{i,j}) \subseteq \mathrm{range}(\mathcal{A}_{i,n-i+1}), i = [m], j = [n-i]$; $\mathcal{A}_{i,n-i+1}$ is injective, $i = 1,\cdots,m$.

To guarantee the global convergence under the nonconvex setting, it is shown that $\sum_{i=1}^m \frac{1}{\beta_i}\|\Delta\Lambda_i^{(k+1)}\|_{\mathrm{F}}^2$ can be bounded by $\|\Delta\mathbf{X}_i^{(k)}\|$ and $\|\Delta\mathbf{X}_i^{(k+1)}\|, i = n-m+1,\cdots,n$. Specifically,

$$\sum_{i=1}^m \frac{1}{\beta_i}\|\Delta\Lambda_i^{(k+1)}\|_{\mathrm{F}}^2 \le \sum_{i=n-m+1}^n \theta_{0,i}\|\Delta\mathbf{X}_i^{(k)}\|_{\mathrm{F}}^2 + \sum_{i=n-m+1}^n \theta_{1,i}\|\Delta\mathbf{X}_i^{(k+1)}\|_{\mathrm{F}}^2.$$

Define the corrected merit function

$$\mathcal{R}^k = \mathcal{L}(\mathbf{X}^k,\Lambda^k) + \sum_{i=n-m+1}^n r\theta_{0,i}\|\Delta\mathbf{X}_i^k\|_{\mathrm{F}}^2, \qquad r < 1.$$

If Assumption 1 holds, it can be shown that the merit function is decreasing:

$$\mathcal{R}^{k+1} + c_X\|\Delta\mathbf{X}^{k+1}\|_{\mathrm{F}}^2 + c_\Lambda\|\Delta\Lambda^{k+1}\|_{\mathrm{F}}^2 \le \mathcal{R}^k.$$

Theorem [Global Convergence]. Suppose that Assumption 1 holds and $\mathcal{R}$ satisfies the KL property on the limit point set. Then

$$\sum_{k=0}^\infty \|\Delta\overline{\mathbf{X}}^{(k)}\|_{\mathrm{F}} + \|\Delta\overline{\mathbf{\Lambda}}^{(k)}\|_{\mathrm{F}} < \infty,$$

and consequently the whole sequence converges to a stationary point of (1).

Strongly Convex Theory

Theorem [Global convergence]. Suppose $\beta_{\max}$ satisfies

$$\beta_{\max} < \min_{i=2,\ldots,n-1}\left\{\frac{2\kappa_{f_i}}{a_i}\right\},$$

then in Scenario 1, cases 1 and 3, there exists $(\overline{\mathbf{X}}_*,\overline{\Lambda}_*) \in \Omega$ such that $(\mathcal{A}_1\mathbf{X}_1^{(k)},\mathbf{X}_2^{(k)},\cdots,\mathbf{X}_n^{(k)},\overline{\Lambda}^{(k)})$ converges to $(\mathcal{A}_1\mathbf{X}_{*,1},\mathbf{X}_{*,2},\cdots,\mathbf{X}_{*,n},\overline{\Lambda}_*)$. In Scenarios 1, 2, 3, and 4, it holds that $(\overline{\mathbf{X}}^{(k)},\overline{\Lambda}^{(k)})$ converges to some $(\overline{\mathbf{X}}_*,\overline{\Lambda}_*) \in \Omega$.

Theorem [Linear convergence]. Suppose the conditions in Table 1 hold. For four cases, if

$$\beta_{\max} < \min_{i=2,\ldots,n-1}\left\{\frac{\kappa_{f_i}}{a_i}\right\},$$

then

$$\tilde{r}^{(k)} \ge (1+\delta)\tilde{r}^{(k+1)}.$$

For four cases, define

$$\tilde{r}^{(k)} = \sum_{j=1}^m \beta_j \sum_{i=1}^{n-j+1}\left\|\sum_{l=i+1}^{n-j+1}\mathcal{A}_{j,l}(\mathbf{X}_l^{(k)} - \mathbf{X}_{*,l})\right\|_{\mathrm{F}}^2 + \sum_{i=1}^m \frac{1}{\beta_i}\|\Lambda_i^{(k)} - \Lambda_{*,i}\|_{\mathrm{F}}^2 + \frac{1}{2}\sum_{i=1}^n \|\mathbf{X}_i^{(k)} - \mathbf{X}_{*,i}\|_{\mathcal{Q}_i}^2.$$

The main difficulty is to bound $\sum_{i=1}^m \beta_i^{-1}\|\Lambda_i^{(k+1)} - \Lambda_{*,i}\|_{\mathrm{F}}^2$ under different scenarios.

R-linear consequence. By the definition of R-linear convergence, any part of a Q-linearly convergent sequence converges R-linearly.

Theorem. Under the same conditions in the linear-convergence theorem, $\mathbf{X}_n^{(k)}$, $\overline{\Lambda}^{(k)}$, and $\mathcal{A}_i\mathbf{X}_i^{(k)}, i = 1,\ldots,n-1$, converge R-linearly. Moreover, if $\mathcal{A}_i, i = 1,2,\ldots,n-1$ are further assumed to be injective, then $\mathbf{X}_i^{(k)}, i = 1,2,\ldots,n-1$ converge R-linearly, respectively.

Strongly Convex Scenarios

Scen.	Conditions
1	SC: $f_{\min\{2,n-m+1\}},\ldots,f_n$; sm.: $\nabla f_i,\ i = n-m+1,\ldots,n$; op.: row-wise range; assump.: $\mathcal{A}_1$ injective.
2	SC: $f_1,\ldots,f_n$; sm.: $\nabla f_i,\ i = n-m+1,\ldots,n$; op.: row-wise range; assump.: none.
3	SC: $f_2,\ldots,f_n$; sm.: $\nabla f_i,\ i = 1,\ldots,n$; op.: not needed; assump.: $\mathcal{A}_1$ injective.
4	SC: $f_1,\ldots,f_n$; sm.: $\nabla f_i,\ i = 1,\ldots,n$; op.: not needed; assump.: none.

Four Proximal-Term Cases

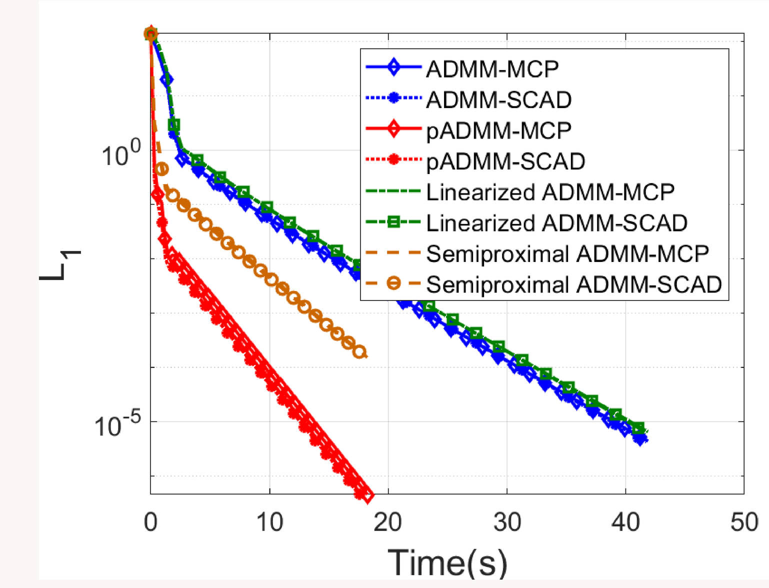
Case	$Q_i,\ i \le n-m$	$Q_i,\ i \ge n-m+1$	Q-linear sequence	R-linear sequence
1	$Q_i = 0$	$Q_i = 0$	$(\overline{H}^{(k)},\overline{\Lambda}^{(k)})$	$(X_n^{(k)},\overline{T}^{(k)},\overline{\Lambda}^{(k)})$
2	$Q_i = 0$	$Q_i \succ 0$	$((X_i^{(k)})_{i=n-m+1}^n,\overline{H}^{(k)},\overline{\Lambda}^{(k)})$	$((X_i^{(k)})_{i=n-m+1}^n,\overline{T}^{(k)},\overline{\Lambda}^{(k)})$
3	$Q_i \succ 0$	$Q_i = 0$	$((X_i^{(k)})_{i=1}^{n-m},\overline{H}^{(k)},\overline{T}^{(k)},\overline{\Lambda}^{(k)})$	$((X_i^{(k)})_{i=1}^{n-m},\overline{T}^{(k)},\overline{\Lambda}^{(k)})$
4	$Q_i \succ 0$	$Q_i \succ 0$	$((X_i^{(k)})_{i=1}^n,\overline{H}^{(k)},\overline{\Lambda}^{(k)})$	$((X_i^{(k)})_{i=1}^n,\overline{T}^{(k)},\overline{\Lambda}^{(k)})$

Notation. $\overline{\mathbf{H}}^{(k)}$ and $\overline{\mathbf{T}}^{(k)}$ are auxiliary variables derived from $\overline{\mathbf{X}}^{(k)}$. $\overline{\mathbf{H}}^{(k)} = (\mathbf{H}_2^{(k)},\cdots,\mathbf{H}_n^{(k)})$, where $\mathbf{H}_i^{(k)} = \sum_{j=i}^n \mathcal{A}_j\mathbf{X}_j^{(k)}$. $\overline{\mathbf{T}}^{(k)} = (\mathbf{T}_2^{(k)},\cdots,\mathbf{T}_n^{(k)})$, where $\mathbf{T}_i^{(k)} = \mathcal{A}_j\mathbf{X}_j^{(k)}$.

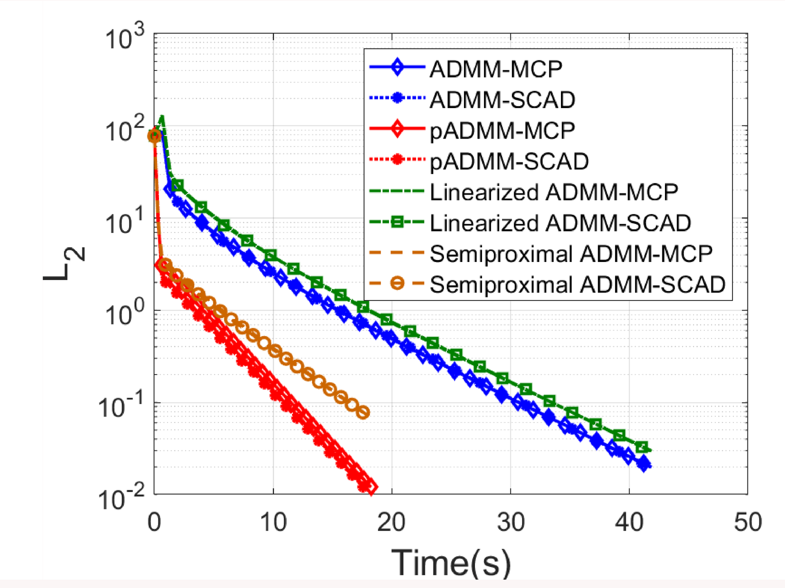
◇ Any part of a Q-linearly convergent sequence converges R-linearly.

Nonconvex Numerical Results

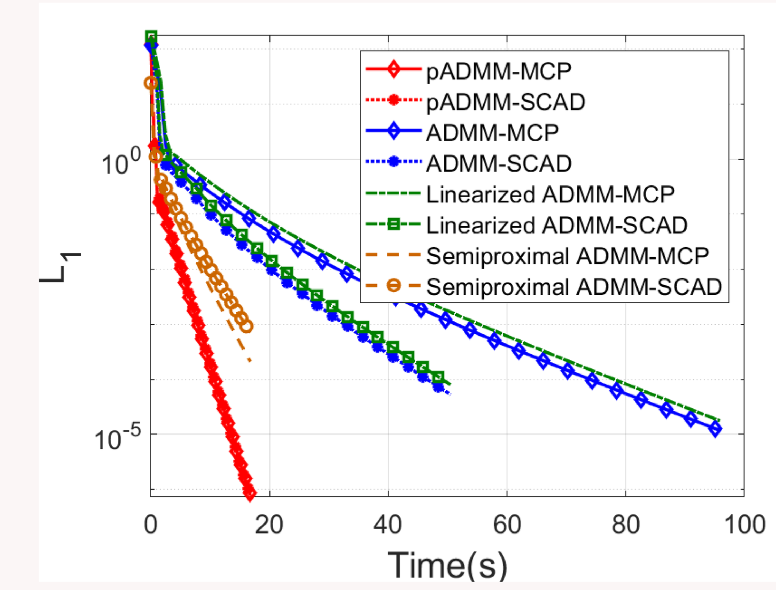
Datasets. Leukemia $(72, 7129)$ and MRI $(200, 4200)$. **Methods.** pADMM, classical ADMM, linearized ADMM, and semiproximal ADMM under MCP and SCAD regularizations. **Measures.** Primal residual L_1 and successive-iterate difference L_2 versus CPU time.



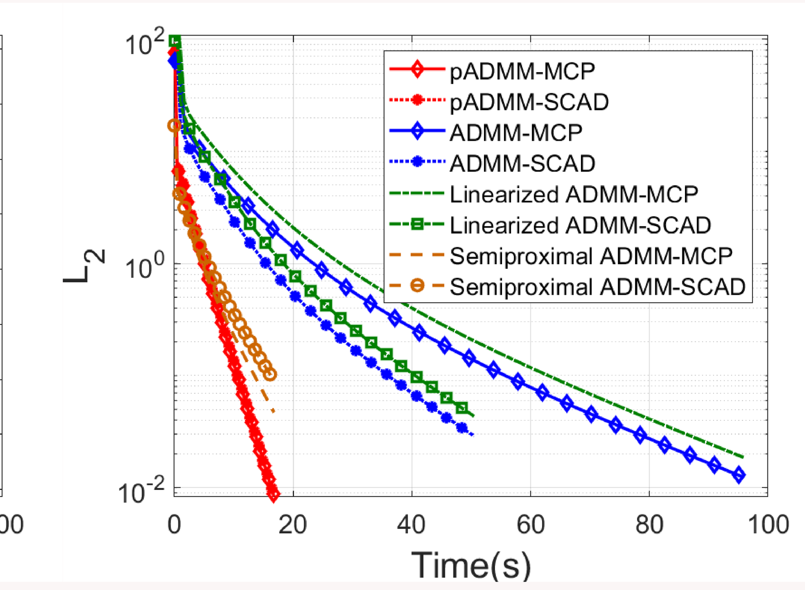
Leukemia: L_1 vs. CPU time



Leukemia: L_2 vs. CPU time



MRI: L_1 vs. CPU time

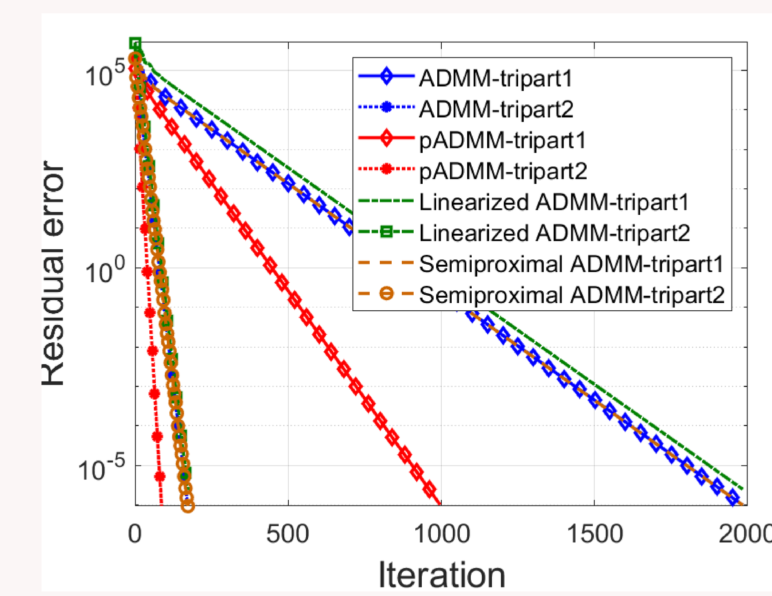


MRI: L_2 vs. CPU time

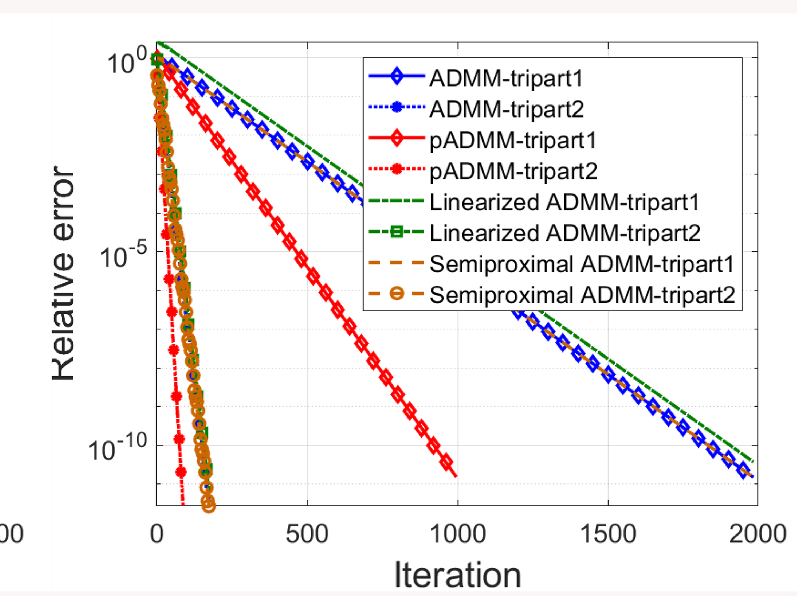
Figure 1 shows that pADMM decreases both L_1 and L_2 faster than the ADMM-type baselines in CPU time on both datasets, for both MCP and SCAD regularizations.

Strongly Convex Numerical Results

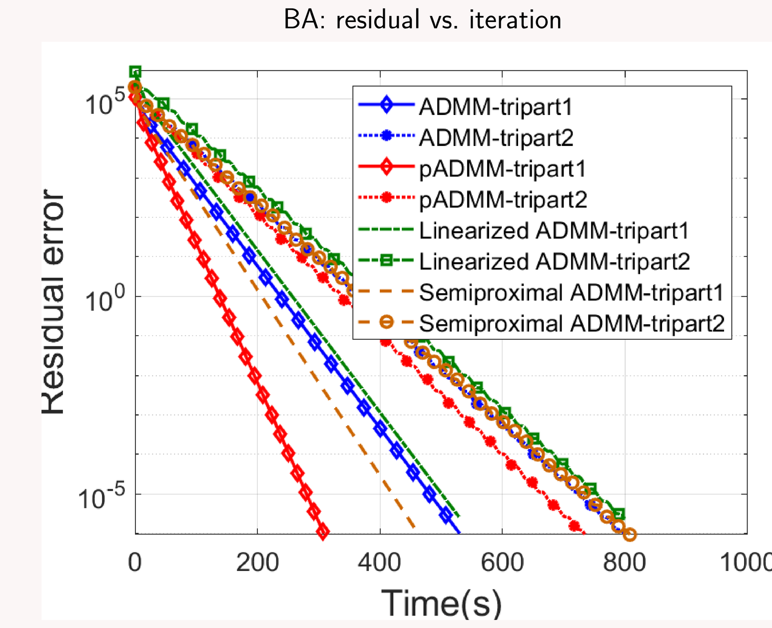
Instances. Synthetic BA1/BA2 and real NMCNF tripart1/tripart2. **Methods.** pADMM, classical ADMM, linearized ADMM, and semiproximal ADMM. **Measures.** Residual error and relative error versus iteration or CPU time; $Q_i = 0.1I$ on the NMCNF instances.



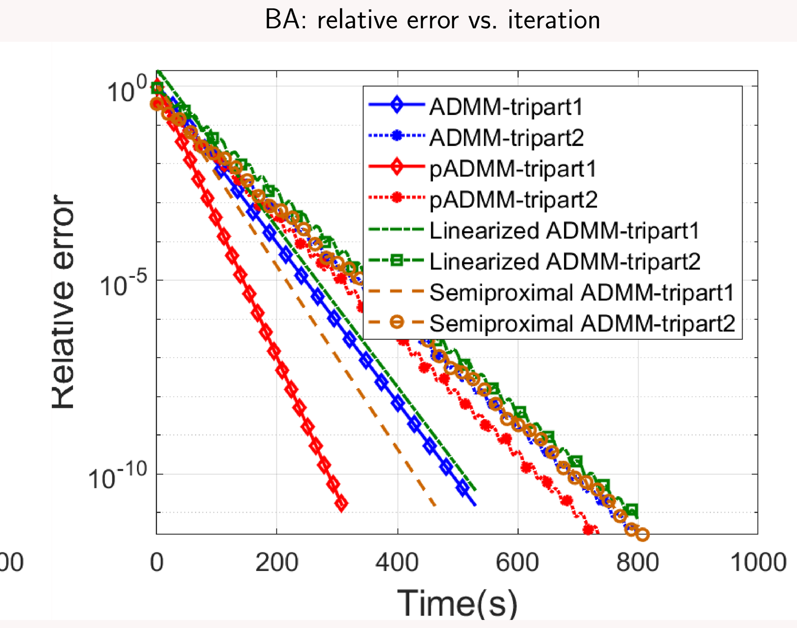
BA: residual vs. iteration



BA: relative error vs. iteration



NMCNF: residual vs. time



NMCNF: relative error vs. time

Figure 2 reports synthetic BA results in the first row and real NMCNF tripart results in the second row; pADMM reaches the target accuracy more efficiently than the compared ADMM-type baselines in both settings.