ICLR 2025

ADMM for Nonconvex Optimization under Minimal Continuity Assumption

Ganzhao Yuan
Peng Cheng Laboratory, China

1. Multi-block Nonconvex Nonsmooth Composite Optimization Problem

$$\min_{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n} \sum_{i=1}^{n} [f_i(\mathbf{x}_i) + h_i(\mathbf{x}_i)],\ s.t.\ [\sum_{i=1}^{n} \mathbf{A}_i \mathbf{x}_i] = \mathbf{b}$$

2. Assumptions

Assumption 1.1. *Each function $f_i(\cdot)$ is L_i-smooth for all $i \in [n]$ such that $\|\nabla f_i(\mathbf{x}_i) - \nabla f_i(\grave{\mathbf{x}}_i)\| \leq L_i\|\mathbf{x}_i - \grave{\mathbf{x}}_i\|$ holds for all $\mathbf{x}_i,\ \grave{\mathbf{x}}_i \in \mathbb{R}^{\mathbf{d}_i \times 1}$. This implies that $|f_i(\mathbf{x}_i) - f_i(\grave{\mathbf{x}}_i) - \langle \nabla f_i(\grave{\mathbf{x}}_i), \mathbf{x}_i - \grave{\mathbf{x}}_i\rangle| \leq \frac{L_i}{2}\|\mathbf{x}_i - \grave{\mathbf{x}}_i\|_2^2$ (cf. Lemma 1.2.3 in (Nesterov, 2003)).*

Assumption 1.2. *The functions $f_n(\cdot)$ and $h_n(\cdot)$ are Lipschitz continuous with some constants C_f and C_h, satisfying $\|\nabla f_n(\mathbf{x}_n)\| \leq C_f$ and $\|\partial h_n(\mathbf{x}_n)\| \leq C_h$ for all $\mathbf{x}_n$.*

Assumption 1.3. *We define $\overline{\lambda} \triangleq \lambda_{max}(\mathbf{A}_n\mathbf{A}_n^{\mathsf{T}})$, $\underline{\lambda} \triangleq \lambda_{min}(\mathbf{A}_n\mathbf{A}_n^{\mathsf{T}})$, $\underline{\lambda}' = \lambda_{min}(\mathbf{A}_n^{\mathsf{T}}\mathbf{A}_n)$. Either of these two conditions holds for matrix $\mathbf{A}_n$:*
a) *Condition $\mathbb{BI}$: $\mathbf{A}_n$ is bijective (i.e., $\underline{\lambda} = \underline{\lambda}' > 0$), and it holds that $\kappa \triangleq \overline{\lambda}/\underline{\lambda} < 2$.*
b) *Condition $\mathbb{SU}$: $\mathbf{A}_n$ is surjective (i.e., $\underline{\lambda} > 0$, and $\underline{\lambda}'$ could be zero).*

Assumption 1.4. *Given any constant $\bar{\beta} \geq 0$, we let $\underline{\Theta}' \triangleq \inf_{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n} \sum_{i=1}^{n}[f_i(\mathbf{x}_i) + h_i(\mathbf{x}_i)] + \frac{\bar{\beta}}{2}\|[\sum_{i=1}^{n} \mathbf{A}_i\mathbf{x}_i] - \mathbf{b}\|_2^2$. We assert that $\underline{\Theta}' > -\infty$.*

Assumption 1.5. *For all $i \in [n]$, the proximal operator $\mathrm{Prox}_i(\mathbf{x}_i;\mu) \triangleq \min_{\mathbf{x}_i'} \frac{\mu}{2}\|\mathbf{x}_i' - \mathbf{x}_i\|_2^2 + h_i(\mathbf{x}_i')$ can be computed efficiently and exactly for any given $\mathbf{x}_i \in \mathbb{R}^{\mathbf{d}_i \times 1}$ and $\mu > 0$.*

Assumption 1.6. *If $\sum_{i=1}^{n}[f_i(\mathbf{x}_i) + h_i(\mathbf{x}_i)] < +\infty$, it follows that $\|\mathbf{x}_i\| < +\infty$ for all $i \in [n]$.*

Assumption 1.7. *For any $i \in [n]$, if the vector $\mathbf{x}_i \in \mathbb{R}^{\mathbf{d}_i \times 1}$ is bounded, then the set $\mathrm{Prox}_i(\mathbf{x}_i;\mu)$ is also bounded for all $\mu \in (0,\infty)$.*

3. Comparisons With Existing Works

Table 1: Comparison of existing ADMM approaches for solving the nonconvex problem in Problem (1). CVX: convex. NC: nonconvex. LCONT: Lipschitz continuous. WC: weakly convex. RWC: restricted weakly convex. $\mathbb{F}$: the constraint set is non-empty. $\mathbb{I}$: $\mathbf{A}_n$ is identity. $\mathbb{SU}$: $\mathbf{A}_n$ is surjective with $\lambda_{\min}(\mathbf{A}_n\mathbf{A}_n^{\mathsf{T}}) > 0$. $\mathbb{IN}$: $\mathbf{A}_n$ is injective with $\lambda_{\min}(\mathbf{A}_n^{\mathsf{T}}\mathbf{A}_n) > 0$. $\mathbb{BI}$: $\mathbf{A}_n$ is bijective (both surjective and injective). $\mathbb{IM}$: $\mathrm{Im}([\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_{n-1}]) \subseteq \mathrm{Im}(\mathbf{A}_n)$ with Im being the image of the matrix.

Reference	Optimization Problems and Main Assumptions			Complexity	Parameter σ
	Blocks	Functions $f_i(\cdot)$ and $h_i(\cdot)$[a]	Matrices $\mathbf{A}_i$		
(He & Yuan, 2012)	$n=2$	CVX: $f_i, h_i, \forall i \in [2]$	$\mathbb{F}$	$\mathcal{O}(\epsilon^{-2})$[b]	$\sigma = 1$
(Li & Pong, 2015)	$n=2$	NC: h_1, f_2; $f_1 = 0$; $h_2 = 0$	$\mathbb{SU}$	$\mathcal{O}(\epsilon^{-2})$	$\sigma = 1$
(Yang et al., 2017)[c]	$n=3$	CVX: h_1, f_3; NC: h_2; $f_1 = f_2 = 0$; $h_3 = 0$	$\mathbb{I}$	$\mathcal{O}(\epsilon^{-2})$	$\sigma \in [1,2)$
(Yashtini, 2022)	$n=2$	NC: $f_{[1,2]}, h_{[1,2]}$; $h_2 = 0$	$\mathbb{BI}$	$\mathcal{O}(\epsilon^{-2})$	$\sigma \in (0,1)$
(Yashtini, 2021)	$n \geq 2$	WC: $f_{[1,n-1]}$; $h_{[1,n-1]} = 0$; $h_n = 0$	$\mathbb{BI}$, $\mathbb{IM}$	$\mathcal{O}(\epsilon^{-2})$	$\sigma \in (0,1)$
(Wang et al., 2019b)	$n \geq 2$	RWC: $h_{[1,n-1]}$; $h_n = 0$	$\mathbb{IN}$, $\mathbb{IM}$	$\mathcal{O}(\epsilon^{-2})$	$\sigma = 1$
(Boţ & Nguyen, 2020)	$n=2$	NC: $h_{[1,n]}, f_{[1,n]}$; $f_1 = 0$; $h_2 = 0$	$\mathbb{I}$	$\mathcal{O}(\epsilon^{-2})$	$\sigma \in [1,2)$
(Boţ et al., 2019)	$n=2$	NC: $h_{[1,n]}, f_{[1,n]}$; $f_1 = 0$; $h_2 = 0$	$\mathbb{SU}$	$\mathcal{O}(\epsilon^{-2})$	$\sigma \in (0,1)$
(Huang et al., 2019)	$n \geq 2$	CVX: $h_{[1,n]}$; $h_n = 0$	$\mathbb{BI}$	$\mathcal{O}(\epsilon^{-2})$	$\sigma = 1$
(Li et al., 2022)[d]	$n=2$	NC: f_1, h_1; CVX: h_2; $f_2 = 0$; LCONT: $h_2 \neq 0$	$\mathbb{I}$	$\mathcal{O}(\epsilon^{-4})$	$\sigma = 1$
This work	$n \geq 2$	NC: $h_{[1,n-1]}, f_{[1,n]}$; CVX: h_n; LCONT: $f_n, h_n \neq 0$	$\mathbb{BI}$	$\mathcal{O}(\epsilon^{-3})$	$\sigma \in [1,2)$
This work	$n \geq 2$	NC: $h_{[1,n-1]}, f_{[1,n]}$; CVX: h_n; LCONT: $f_n, h_n \neq 0$	$\mathbb{SU}$	$\mathcal{O}(\epsilon^{-3})$	$\sigma \in (0,1)$

Note a: $h_n = 0$ denotes that the n-th block has no non-smooth part, making the objective function smooth.
Note b: The iteration complexity relies on the variational inequality of the convex problem.
Note c: We adapt their application model into our optimization framework in Equation (1) with $(L, S, Z) = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$, as their model additionally requires the linear operator for the first two blocks to be injective.
Note d: This paper studies manifold optimization with a fixed large penalty and small stepsize.

4. The Proposed Algorithm

- **Smoothing the nonsmooth function of the last block:**

$$\min_{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n} h_n(\mathbf{x}_i;\mu) + [\textstyle\sum_{i=1}^{n-1} h_i(\mathbf{x}_i)] + [\sum_{i=1}^{n} f_i(\mathbf{x}_i)],$$
$$s.t.\ [\textstyle\sum_{i=1}^{n} \mathbf{A}_i\mathbf{x}_i] = \mathbf{b}$$

- **The resulting augmented Lagrangian function:**

$$\mathcal{L}(\mathbf{x},\mathbf{z};\beta,\mu) \triangleq h_n(\mathbf{x}_n;\mu) + \{\textstyle\sum_{i=1}^{n-1} h_i(\mathbf{x}_i)\} + G(\mathbf{x},\mathbf{z};\beta)$$

$$G(\mathbf{x},\mathbf{z};\beta) \triangleq \textstyle\sum_{i=1}^{n} f_i(\mathbf{x}_i) + \langle[\sum_{i=1}^{n} \mathbf{A}_i\mathbf{x}_i] - \mathbf{b}, \mathbf{z}\rangle + \frac{\beta}{2}\|[\sum_{i=1}^{n} \mathbf{A}_i\mathbf{x}_i] - \mathbf{b}\|_2^2$$

- **Our Key Strategy: Increasing Penalization and Decreasing Smoothing**

$$\beta^t = \mathcal{O}(\sqrt[3]{t}),\quad \mu^t = \mathcal{O}(1/\sqrt[3]{t})$$

- **The Proposed IPDS-ADMM:**

Algorithm 1: IPDS-ADMM: The Proposed Proximal Linearized ADMM for Problem (1).

Choose suitable parameters $\{p,\xi,\delta\}$ and $\{\sigma,\theta_1,\theta_2\}$ using Formula (5) or Formula (6).
Initialize $\{\mathbf{x}^0,\mathbf{z}^0\}$. Choose $\beta^0 \geq L_n/(\delta\overline{\lambda})$.
for t from 0 to T **do**
  S1) IPDS Strategy: Set $\beta^t = \beta^0(1+\xi t^p)$, $\mu^t = 1/(\overline{\lambda}\delta\beta^t)$.
  We define $\ddot{\mathbf{g}}_i^t \triangleq \nabla_{\mathbf{x}_i} G(\mathbf{x}_{[1,i-1]}^{t+1}, \mathbf{x}_i^t, \mathbf{x}_{[i+1,n]}^t, \mathbf{z}^t;\beta^t)$.
  S2) $\mathbf{x}_1^{t+1} \in \arg\min_{\mathbf{x}_1} h_1(\mathbf{x}_1) + \langle\mathbf{x}_1 - \mathbf{x}_1^t, \ddot{\mathbf{g}}_1^t\rangle + \frac{\theta_1\mathsf{L}_1^t}{2}\|\mathbf{x}_1 - \mathbf{x}_1^t\|_2^2$
  S3) $\mathbf{x}_2^{t+1} \in \arg\min_{\mathbf{x}_2} h_2(\mathbf{x}_2) + \langle\mathbf{x}_2 - \mathbf{x}_2^t, \ddot{\mathbf{g}}_2^t\rangle + \frac{\theta_1\mathsf{L}_2^t}{2}\|\mathbf{x}_2 - \mathbf{x}_2^t\|_2^2$
  ...
  S4) $\mathbf{x}_{n-1}^{t+1} \in \arg\min_{\mathbf{x}_{n-1}} h_{n-1}(\mathbf{x}_{n-1}) + \langle\mathbf{x}_{n-1} - \mathbf{x}_{n-1}^t, \ddot{\mathbf{g}}_{n-1}^t\rangle + \frac{\theta_1\mathsf{L}_{n-1}^t}{2}\|\mathbf{x}_{n-1} - \mathbf{x}_{n-1}^t\|_2^2$
  S5) $\mathbf{x}_n^{t+1} \in \arg\min_{\mathbf{x}_n} h_n(\mathbf{x}_n;\mu) + \langle\mathbf{x}_n - \mathbf{x}_n^t, \ddot{\mathbf{g}}_n^t\rangle + \frac{\theta_2\mathsf{L}_n^t}{2}\|\mathbf{x}_n - \mathbf{x}_n^t\|_2^2$. It can be solved using Lemma 3.6 as $\mathbf{x}_n^{t+1} = \frac{1}{1+\mu\rho}(\breve{\mathbf{x}}_n^{t+1} + \mu\rho\mathbf{c})$, where $\breve{\mathbf{x}}_n^{t+1} = \mathrm{Prox}_n(\mathbf{c};\mu+1/\rho)$, $\mu = \mu^t$, $\rho \triangleq \theta_2\mathsf{L}_n^t$, and $\mathbf{c} \triangleq \mathbf{x}_n^t - \ddot{\mathbf{g}}_n^t/\rho$.
  S6) $\mathbf{z}^{t+1} = \mathbf{z}^t + \sigma\beta^t([\sum_{j=1}^{n} \mathbf{A}_j\mathbf{x}_j^{t+1}] - \mathbf{b})$
end

- **On Choosing the Parameters:**

$$\mathbb{BI}: p = \tfrac{1}{3},\ \xi \in (0,\infty),\ \delta \in (0, \tfrac{1}{3}(\tfrac{2}{\kappa} - 1)), \sigma \in [1,2), \theta_1 = 1.01, \theta_2 = \tfrac{1/\kappa - \delta}{1+\delta} + \tfrac{1}{2\varrho(1+\delta)^2}. \quad (5)$$

$$\mathbb{SU}: p = \tfrac{1}{3},\ \xi = \delta = \sigma = \tfrac{0.01}{\kappa},\ \theta_1 = 1.01, \theta_2 = 1.5. \quad (6)$$

Here, $\varrho \triangleq 6\omega\sigma_1\kappa$, $\sigma_1 \triangleq \frac{\sigma}{(1-|1-\sigma|)^2}$, and $\omega \triangleq 1 + \frac{\xi}{2\sigma} + \sigma\xi$. Notably, θ_2 in (5) depends on (ξ,δ,σ).

5. Iteration Complexity

- **Controlling Dual using Primal:**

$$\tfrac{\omega}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 \leq \Theta_+^t - \Theta_+^{t+1} + \chi\mathsf{L}_n^t\|\mathbf{x}_n^{t+1} - \mathbf{x}_n^t\|_2^2 + \mathbb{U}^t$$

$$\Theta_+^t \triangleq a\mathbb{A}^t + b\mathbb{B}^t \qquad \mathbb{U}^t \triangleq C_h^2\tfrac{b}{\beta^t}\cdot(\tfrac{\mu^{t-1}}{\mu^t} - 1)^2$$

- **The Associated Lyapunov Function:**

$$\mathbb{BI}:\ \Theta^t = \Theta_L^t + \underbrace{\tfrac{\omega\sigma_2}{\lambda}}_{\triangleq a}\cdot\underbrace{\tfrac{1}{\beta^t}\|\mathbb{a}^t\|_2^2}_{\triangleq\mathbb{A}^t} + \underbrace{\tfrac{3\omega\sigma_1}{\lambda}}_{\triangleq b}\cdot\underbrace{\tfrac{1}{\beta^t}(L_n\|\mathbf{x}_n^t - \mathbf{x}_n^{t-1}\| + \|\mathbb{u}_n^t\|)^2}_{\triangleq\mathbb{B}^t}.$$

$$\mathbb{SU}:\ \Theta^t = \Theta_L^t + \underbrace{\tfrac{2\omega\sigma_2}{\lambda}}_{\triangleq a}\cdot\underbrace{\tfrac{1}{\beta^t}\|\mathbb{a}^t\|_2^2}_{\triangleq\mathbb{A}^t} + \underbrace{\tfrac{6\omega\sigma_1}{\lambda}}_{\triangleq b}\cdot\underbrace{\tfrac{1}{\beta^t}(L_n\|\mathbf{x}_n^t - \mathbf{x}_n^{t-1}\| + \sigma\|\mathbb{u}_n^t\|)^2}_{\triangleq\mathbb{B}^t}.$$

$$\Theta_L^t \triangleq \mathcal{L}(\mathbf{x}^t,\mathbf{z}^t;\beta^t,\mu^t) + \tfrac{1}{2}C_h\mu^t,\ \mathsf{L}_i^t = L_i + \beta^t\|\mathbf{A}_i\|_2^2$$

$$\mathcal{E}^{t+1} \triangleq [\varepsilon_1\textstyle\sum_{i=1}^{n-1}\mathsf{L}_i^t\|\mathbf{x}_i^{t+1} - \mathbf{x}_i^t\|_2^2] + \varepsilon_2\mathsf{L}_n^t\|\mathbf{x}_n^{t+1} - \mathbf{x}_n^t\|_2^2 + \frac{\varepsilon_3}{\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2$$

- **A Summerable Property:**

$$\mathcal{E}^{t+1} \leq \Theta^t - \Theta^{t+1} + \mathbb{U}^t$$

$$\textstyle\sum_{t=1}^{\infty}\mathbb{U}^t \leq \overline{\mathrm{U}} \qquad \sum_{t=1}^{\infty}\mathcal{E}^{t+1} \leq \overline{\mathcal{E}}$$

- **Approximate Critical Point**

$$\mathrm{Crit}(\breve{\mathbf{x}},\breve{\mathbf{z}}) \leq \epsilon.$$

$$\mathrm{Crit}(\breve{\mathbf{x}},\breve{\mathbf{z}}) \triangleq \|\mathbf{A}\breve{\mathbf{x}} - \mathbf{b}\| + \textstyle\sum_{i=1}^{n}\mathrm{dist}(\mathbf{0}, \nabla f_i(\breve{\mathbf{x}}_i) + \partial h_i(\breve{\mathbf{x}}_i) + \mathbf{A}_i^{\mathsf{T}}\breve{\mathbf{z}}).$$

- **The Complexity Result**

We define $\mathbf{q}^t \triangleq \{\mathbf{x}_1^t, \mathbf{x}_2^t, \ldots, \mathbf{x}_{n-1}^t, \breve{\mathbf{x}}_n^t\}$

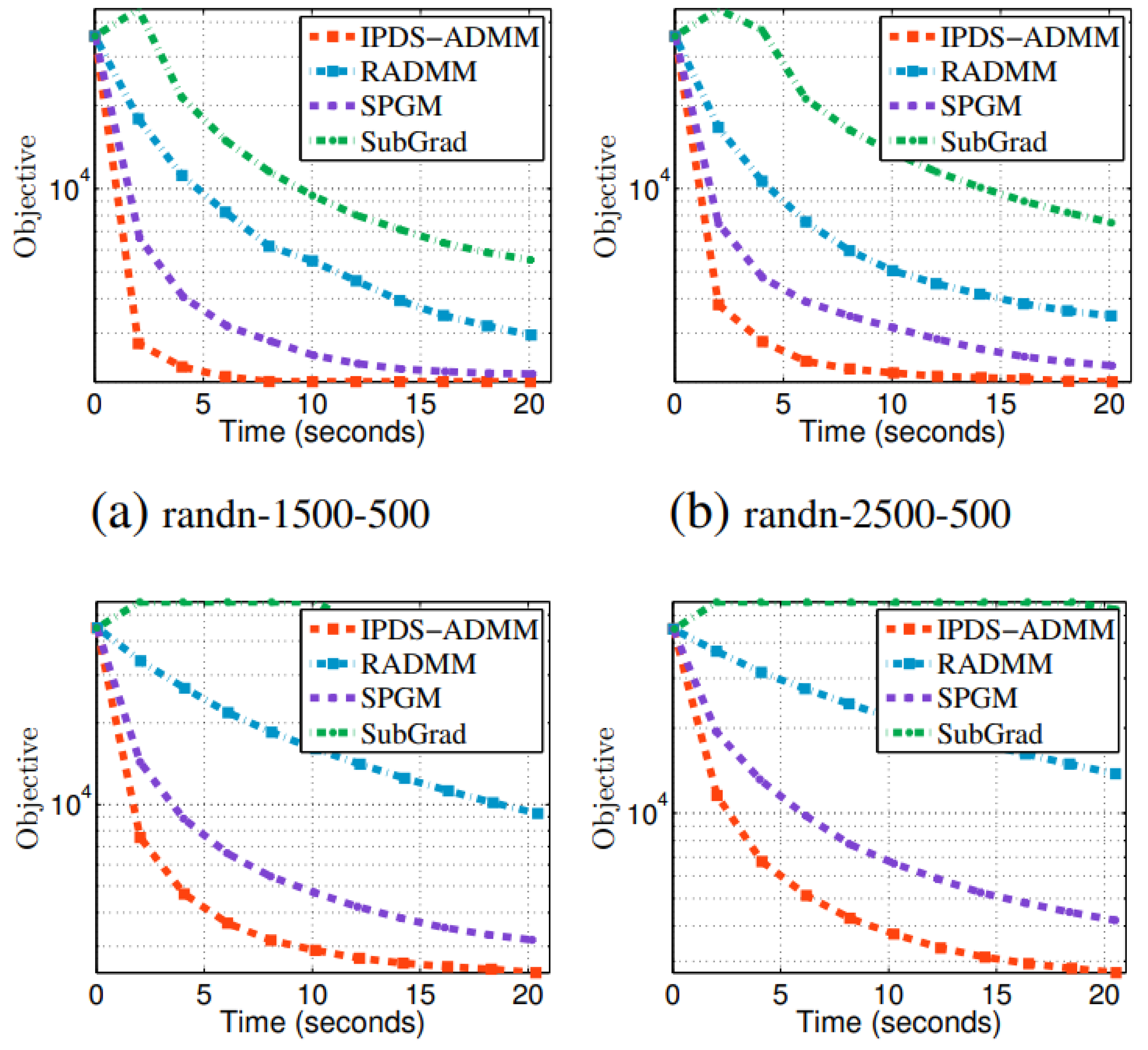
$$\tfrac{1}{T}\textstyle\sum_{t=1}^{T}\mathrm{Crit}(\mathbf{q}^{t+1},\mathbf{z}^{t+1}) \leq \mathcal{O}(T^{(p-1)/2}) + \mathcal{O}(T^{-p})$$

with the choice $p = 1/3$, we have $\frac{1}{T}\sum_{t=1}^{T}\mathrm{Crit}(\mathbf{q}^{t+1},\mathbf{z}^{t+1}) \leq \mathcal{O}(T^{-1/3})$.

6. Experiment Results

$$\min_{\mathbf{V}\in\mathbb{R}^{d\times\dot{r}}} \tfrac{1}{2\dot{m}}\|\mathbf{D} - \mathbf{D}\mathbf{V}\mathbf{V}^{\mathsf{T}}\|_{\mathsf{F}}^2 + \dot{\rho}\|\mathbf{V}\|_1$$

$$s.t.\ \mathbf{V} \in \mathcal{M} \triangleq \{\mathbf{V}\,|\,\mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}\}$$



(a) randn-1500-500
(b) randn-2500-500
(c) mnist-1500-780
(d) mnist-2500-780

Conclusions:

(a) Sub-Grad tends to be less efficient in comparison to other methods.

(b) SPGM, utilizing a variable smoothing strategy, generally demonstrates slower performance than the multiplier-based variable splitting method. This observation corroborates the widely accepted notion that primal-dual methods are typically more robust and quicker than primal-only methods.

(c) The proposed IPDS-ADMM consistently achieves the lowest objective function values among all methods examined.