arXiv:2405.03233v1 [math.OC] 6 May 2024

Inertial Relaxed Proximal Linearized ADMM for Nonconvex Optimization under Minimal Continuity Assumption

Ganzhao Yuan [1]

Abstract

This paper proposes an Inertial Relaxed Proximal Linearized Alternating Direction Method of Multipliers (IRPL-ADMM) for solving general multi-block nonconvex composite optimization problems. Distinguishing itself from existing ADMM-style algorithms, our approach imposes a less stringent condition, specifically requiring continuity in only one block of the objective function. It incorporates an inertial strategy for primal variable updates, and a relaxed strategy for dual variable updates. The fundamental concept underlying our algorithm is based on novel *regular penalty update rules*, ensuring that the penalty increases but not excessively fast. We devise a novel potential function to facilitate our convergence analysis and extend our methods from deterministic optimization problems to finite-sum stochastic settings. We establish the iteration complexity for both scenarios for achieving an approximate stationary solution. Under the Kurdyka-Łojasiewicz (KŁ) inequality, we establish strong limit-point convergence results for the IRPL-ADMM algorithm. Finally, some experiments have been conducted on two machine learning tasks to show the effectiveness of our approaches.

1. Introduction

We consider the following multi-block nonconvex nonsmooth composite optimization problem:

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

where $\mathbf{b}\in\mathbb{R}^{m\times 1}$, $\mathbf{A}_i\in\mathbb{R}^{m\times \mathbf{d}_i}$, $\mathbf{x}_i\in\mathbb{R}^{\mathbf{d}_i\times 1}$, and $i\in[n]\triangleq\{1,2,\ldots,n\}$. We do not assume convexity of $f_i(\cdot)$ and $h_i(\cdot)$ for all $i\in[n]$. Furthermore, we require that the function $f_i(\cdot):\mathbb{R}^{\mathbf{d}_i\times 1}\mapsto(-\infty,\infty)$ is differentiable, while $h_i(\cdot):\mathbb{R}^{\mathbf{d}_i\times 1}\mapsto(-\infty,\infty]$ is potentially nonsmooth. However, its associated nonconvex operator $\min_{\mathbf{x}_i}\ \frac{\mu}{2}\|\mathbf{x}_i-\mathbf{x}_i'\|_2^2+h_i(\mathbf{x}_i)$ is well-defined and simple to compute for all $i\in[n]$, any $\mu>0$, and $\mathbf{x}_i'\in\mathbb{R}^{\mathbf{d}_i\times 1}$.

Problem (1) has a wide range of applications in machine learning. The function $f_i(\cdot)$ plays a crucial role in handling empirical loss, including neural network activation functions. Incorporating multiple nonsmooth regularization terms $h_i(\cdot)$ enables diverse prior information integration, including structured sparsity, low-rank, orthogonality, and non-negativity constraints, enhancing regularization model accuracy. These capabilities extend to various applications such as sparse PCA, overlapping group Lasso, graph-guided fused Lasso, and phase retrieval.

► **ADMM Literature**. The Alternating Direction Method of Multipliers (ADMM) is a versatile optimization tool suitable for solving composite constrained problems as in Problem (1), which pose challenges for other standard optimization methods, such as the accelerated proximal gradient method (Nesterov, 2013) and the augmented Lagrangian method (Zhu et al., 2023; Lin et al., 2022). The standard ADMM was initially introduced in (Gabay & Mercier, 1976), and its complexity analysis for the convex settings was first conducted in (He & Yuan, 2012; Monteiro & Svaiter, 2013). Since then, numerous papers have explored the iteration complexity of ADMM in diverse settings. These settings include acceleration through multi-step updates (Pock & Sabach, 2016; Li et al., 2016; Ouyang et al., 2015; Shen et al., 2017; Tran Dinh, 2018), asynchronous updates (Zhang & Kwok, 2014), Jacobi updates (Deng et al., 2017), non-Euclidean proximal updates (Gonçalves et al., 2017b), and extensions to handle more specific or general functions such as strongly convex functions (Nishihara et al., 2015; Lin et al., 2015b; Ouyang et al., 2015), nonlinear constrained functions (Lin et al., 2022), and multi-block composite functions (Lin et al., 2015a;a; Xu et al., 2017).

► **Nonconvex ADMM**. The convergence analysis of the nonconvex ADMM is challenging due to the absence of Fejér monotonicity in iterations. In the past decade, significant research has focused on exploring various nonconvex ADMM variants (Li & Pong, 2015; Hong et al., 2016; Yang et al., 2017). (Li & Pong, 2015) establishes the convergence

[1]Peng Cheng Laboratory, China. Correspondence to: Ganzhao Yuan <yuangzh@pcl.ac.cn>.

Table 1. Comparison of existing nonconvex ADMM approaches. `CVX`: convex. `NC`: nonconvex. `LCONT`: Lipschitz continuous. `WC`: weakly convex. `RWC`: restricted weakly convex. `Id`: $\mathbf{A}_n$ is identity. `Sj`: $\mathbf{A}_n$ is surjective with $\boldsymbol{\lambda}_{\min}(\mathbf{A}_n\mathbf{A}_n^\mathsf{T}) > 0$. `Bj`: $\mathbf{A}_n$ is bijective (both surjective and injective). `Ij`: $\mathbf{A}_n$ is injective. `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. **IN**: inertial strategy.

Reference	Optimization Problems and Main Assumptions			Iteration Complexity		Accelerated?
	Blocks	Functions $f_i(\cdot)$ and $h_i(\cdot)$ [a]	Matrices $\mathbf{A}_i$	Deterministic	Stochastic [b]	
(He & Yuan, 2012)	$n = 2$	`CVX`: $f_i, h_i, \forall i \in [2]$	feasible	$\mathcal{O}(1/\epsilon)$ [c]	unknown	$\sigma = 1$
(Li & Pong, 2015)	$n = 2$	`NC`: h_1, f_2; $f_1 = h_2 = 0$	`Sj`	$\mathcal{O}(1/\epsilon)$	unknown	$\sigma = 1$
(Yang et al., 2017) [d]	$n = 3$	`CVX`: h_1, f_3; `NC`: h_2; $f_1 = f_2 = h_3 = 0$	`Id`	$\mathcal{O}(1/\epsilon)$	unknown	$\sigma \in [1, 2)$
(Yashtini, 2022)	$n = 2$	`NC`: $f_i, h_i, \forall i \in [2]$; $h_2 = 0$	`Bj`	$\mathcal{O}(1/\epsilon)$	unknown	$\sigma \in (0, 1)$
(Yashtini, 2021)	$n \geq 2$	`WC`: $f_i, \forall i \in [n-1]$; $h_i = 0, \forall i \in [n]$	`Bj`, `Im`	$\mathcal{O}(1/\epsilon)$	unknown	$\sigma \in (0, 1)$
(Wang et al., 2019a)	$n \geq 2$	`RWC`: $h_i, \forall i \in [n-1]$, $h_n = 0$	`Ij`, `Im`	$\mathcal{O}(1/\epsilon)$	unknown	$\sigma = 1$
(Boţ et al., 2019)	$n = 2$	`NC`: $h_i, f_i, \forall i \in [n]$; $f_1 = h_2 = 0$	`Id`	$\mathcal{O}(1/\epsilon)$	unknown	$\sigma \in [1, 2)$
(Boţ & Nguyen, 2020)	$n = 2$	`NC`: $h_i, f_i, \forall i \in [n]$; $f_1 = h_2 = 0$	`Sj`	$\mathcal{O}(1/\epsilon)$	unknown	$\sigma \in (0, 1)$
(Huang et al., 2019)	$n \geq 2$	`CVX`: $h_i, \forall i \in [n]$; $h_n = 0$	`Bj` [e]	$\mathcal{O}(1/\epsilon)$	$\mathcal{O}(N + \sqrt{N}/\epsilon)$	$\sigma = 1$
This paper	$n \geq 2$	`NC`: $h_i, f_i, \forall i \in [n]$; `LCONT`: h_n, f_n	`Id`	$\mathcal{O}(1/\epsilon)$ [f]	$\mathcal{O}(N + \sqrt{N}/\epsilon)$ [f]	$\sigma \in [1, 2)$, **IN**
This paper	$n \geq 2$	`NC`: $h_i, f_i, \forall i \in [n]$; `LCONT`: h_n, f_n	`Sj`	$\mathcal{O}(1/\epsilon)$ [f]	$\mathcal{O}(N + \sqrt{N}/\epsilon)$ [f]	$\sigma \in (0, 1)$, **IN**

Note a: The notation $h_n = 0$ indicates that, for the n-th block, the non-smooth part is absent and the objective function is smooth.

Note b: N is the number of data points for the finite-sum structure (See Equation (102)).

Note c: The iteration complexity relies on the variational inequality of the convex problem.

Note d: 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 other two blocks to be injective.

Note e: Assumption 4 in (Wang et al., 2019a) claims that the matrix can exhibit either full row rank or full column rank. However, Equation (20) in their analysis relies on the matrix's surjectiveness, while Lemma 7 depends on its injectiveness.

Note f: The iteration complexity is contingent on the newly introduced approximate inertial nonconvex proximal point (see Section 4).

of a class of nonconvex problems when a specific potential function associated with the augmented Lagrangian satisfies the Kurdyka-Łojasiewicz (KŁ) inequality. (Yang et al., 2017) analyzes ADMM variants for solving low-rank and sparse optimization problems. (Hong et al., 2016) investigates ADMM variants for nonconvex consensus and sharing problems. Some researchers have examined ADMM variants under weaker conditions, such as restricted weak convexity (Wang et al., 2019a), restricted strong convexity (Barber & Sidky, 2020), and the Hoffman error bound (Zhang & Luo, 2020). However, existing methods all assume the smoothness of at least one block. In contrast, our approach imposes the fewest conditions on the objective function by employing an increasing penalty update strategy. The convergence of our proposed method is established through the use of KŁ inequalities (Attouch et al., 2010; Bolte et al., 2014; Li & Lin, 2015).

► **Accelerating Nonconvex ADMM**. There has been significant research interest in accelerating ADMM for nonconvex problems. Prior studies (Gonçalves et al., 2017a; Yang et al., 2017; Yashtini, 2022; 2021; Boţ & Nguyen, 2020) have analyzed ADMM, using under-relaxation stepsize $\sigma \in (0, 1)$ or over-relaxation stepsize $\sigma \in (1, 2)$ to update the dual variable, in contrast to previous fixed values of 1 or the golden ratio $(\sqrt{5} + 1)/2$. The work by (Hien et al., 2022) explores an inertial strategy to accelerate nonconvex ADMM. This method omits inertial updates for specific blocks to ensure convergence. Studies by (Huang et al., 2019; Bian et al., 2021; Liu et al., 2020) employ stochastic gradient descent to reduce the Incremental First-order Oracle (IFO) complexity when addressing composite problems with finite-sum structures. Inspired by these works, we apply an inertial strategy (Pock & Sabach, 2016; Le et al., 2020; Boţ et al., 2023; Phan & Gillis, 2023) for primal variable updates and employ a relaxed strategy for dual variable updates. Additionally, we extend our techniques to handle finite-sum stochastic settings and analyze the IFO complexity of our method.

We make a comparison of existing nonconvex ADMM approaches in Table 1.

► **Contributions.** Our main contributions are summarized as follows. ***(i)*** We propose IRPL-ADMM for solving the nonconvex optimization problem as in Problem (1). IRPL-ADMM imposes the least stringent condition, specifically requiring continuity in just one block of the objective function, while employing an increasing penalty update rule to ensure convergence. ***(ii)*** IRPL-ADMM exhibits both convergence and speed. In primal variable updates, it leverages inertial acceleration for fast convergence. In dual variable updates, it utilizes over-relaxation stepsize for faster convergence when the linear operator is an identity matrix, and under-relaxation stepsize for global convergence when the linear operator is surjective. ***(iii)*** We establish the convergence rate of IRPL-ADMM by introducing a novel concept of ϵ-INP point (*Inertial Nonconvex Proximal Point*). We prove that any ϵ-INP point is a critical point when $\epsilon = 0$, and show its convergence to an ϵ-INP point with a time complexity of $\mathcal{O}(1/\epsilon)$. Additionally, we establish strong limit-point convergence results for IRPL-ADMM under the

Kurdyka-Łojasiewicz (KŁ) inequality. ***(iv)*** We extend our method to stochastic settings and demonstrate its optimality in terms of IFO.

► **Assumptions.** Through this paper, we impose the following assumptions on Problem (1).

Assumption 1.1. Each function $f_i(\cdot)$ is L_i-smooth for all $i \in [n]$ such that $\|\nabla f_i(\mathbf{x}_i) - \nabla f_i(\hat{\mathbf{x}}_i)\| \leq L_i\|\mathbf{x}_i - \hat{\mathbf{x}}_i\|$ holds for all $\mathbf{x}_i \in \mathbb{R}^{\mathbf{d}_i \times 1}$ and $\hat{\mathbf{x}}_i \in \mathbb{R}^{\mathbf{d}_i \times 1}$. This implies that $|f_i(\mathbf{x}_i) - f_i(\hat{\mathbf{x}}_i) - \langle \nabla f_i(\hat{\mathbf{x}}_i), \mathbf{x}_i - \hat{\mathbf{x}}_i\rangle| \leq \frac{L_i}{2}\|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2$ (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. Either of these two conditions holds for matrix $\mathbf{A}_n$:
a) Condition $\boxed{\mathbb{I}}$: $\mathbf{A}_n$ is an identity matrix with $\mathbf{A}_n = \mathbf{I}_{\mathbf{d}_n}$.
b) Condition $\boxed{\mathbb{A}}$: $\mathbf{A}_n$ is surjective (i.e., $\lambda_{\min}(\mathbf{A}_n\mathbf{A}_n^\mathsf{T}) > 0$).

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 $\underline{\Theta} > -\infty$.

Remarks. ***(i)*** Assumption 1.1 is commonly used in the convergence analysis of nonconvex algorithms. ***(ii)*** Assumption 1.2 imposes a continuity assumption only for the last block, allowing other blocks of the function $h_i(\mathbf{x}_i)_{i=1}^{n-1}$ to be nonsmooth and non-Lipschitz, such as indicator functions of constraint sets. It ensures bounded (sub-)gradients for $f_n(\cdot)$ and $h_n(\cdot)$, a relatively mild requirement that has found use in stochastic optimization (Huang et al., 2019). ***(iii)*** Assumption 1.3 demands a condition on the linear matrix $\mathbf{A}_i$ for the last block ($i = n$), while leaving $\mathbf{A}_i$ unrestricted for $i \in [n-1]$. ***(iv)*** Assumption 1.4 ensures the well-defined nature of the penalty function associated with the problem, as has also been used in (Gonçalves et al., 2017a).

► **Notations.** We define $[n] \triangleq \{1, 2, \ldots, n\}$ and $\mathbf{x} \triangleq \mathbf{x}_{[n]} \triangleq \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. For any $j \geq i$, we denote $\mathbf{x}_{[i,j]} \triangleq \{\mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_j\}$. We define $\underline{\lambda}$ and $\overline{\lambda}$ as the smallest and largest eigenvalue of the matrix $\mathbf{A}_n\mathbf{A}_n^\mathsf{T} \in \mathbb{R}^{m\times m}$, respectively. We denote $\|\mathbf{A}_i\|$ as the spectral norm of the matrix $\mathbf{A}_i$. We denote $\mathbf{A}\mathbf{x} \triangleq \sum_{j=1}^n \mathbf{A}_j\mathbf{x}_j$. Further notations and technical preliminaries are provided in Appendix A.

2. The Proposed ADMM Algorithm

This section describes an Inertial Relaxed Proximal Linearized ADMM (IRPL-ADMM) algorithm for solving the nonconvex and nonsmooth optimization problem in Problem (1).

2.1. Regular Penalty Update Rule

We consider an increasing penalty update strategy, which plays a significant role in our algorithm. Natural choice for the penalty update rule is the ℓ_p family.

We introduce a novel concept of (β^0, ξ, p)-regular penalty update rule, as follows:

Definition 2.1. Given constants $p \in (0, 2]$, $\beta^0 > 0$, and $\xi > 0$. A penalty update rule $\{\beta^t\}_{t=0}^\infty$ is considered (β^0, ξ, p)-regular when the sequence $\{\beta^t\}_{t=0}^\infty$ is increasing, and the following condition holds for all $t \geq 0$ with some $\vartheta' > 0$:

$$\beta^0 + \vartheta'(t+1)^p \leq \beta^{t+1} \leq (1+\xi)\beta^t. \tag{2}$$

We provide four examples of (β^0, ξ, p)-regular penalty update rules that align with Definition 2.1.

Lemma 2.2. *(Proof in Appendix B.1,* Sublinear Rule*) Let $p \in (0, 1]$. The penalty update rule $\beta^t = \beta^0 + \vartheta t^p$, is (β^0, ξ, p)-regular if $\vartheta \leq \beta^0\xi$.*

Lemma 2.3. *(Proof in Appendix B.2,* Superlinear Rule*) Let $p \in (1, 2]$. The penalty update rule $\beta^t = \beta^0 + \vartheta t^p$, is (β^0, ξ, p)-regular if $\vartheta \leq \beta^0\xi^2/(1+\xi)$.*

Lemma 2.4. *(Proof in Appendix B.3,* Adaptive Sublinear Rule*). Let $p \in (0, 1]$. The penalty update rule $\beta^{t+1} = \beta^t + \min(\|\mathbf{A}\mathbf{x}^{t+1} - \mathbf{b}\| + \vartheta(t+1)^p - \vartheta t^p, \beta^t\xi)$, is (β^0, ξ, p)-regular if $\vartheta \leq \beta^0\xi$.*

Lemma 2.5. *(Proof in Appendix B.4,* Adaptive Superlinear Rule*) Let $p \in (1, 2]$. The penalty update rule $\beta^{t+1} = \beta^t + \min(\|\mathbf{A}\mathbf{x}^{t+1} - \mathbf{b}\| + \vartheta(t+1)^p - \vartheta t^p, \beta^t\xi)$, is (β^0, ξ, p)-regular if $\vartheta \leq \beta^0\xi^2/(1+\xi)$.*

Remarks ***(i)*** Increasing penalty updates are commonly used in subgradient methods (Davis & Drusvyatskiy, 2019; Li et al., 2021), smoothing gradient methods (Sun & Sun, 2023; Lei Yang, 2021; Böhm & Wright, 2021), and penalty decomposition methods (Lu & Zhang, 2013), but are less prevalent in ADMM. We examine this approach within ADMM but limit our discussion to specific conditions as in Inequality (2). ***(ii)*** Adaptive sublinear and superlinear rules can integrate the penalty error $\|\mathbf{A}\mathbf{x}^{t+1} - \mathbf{b}\|$ into the penalty update process. As $\|\mathbf{A}\mathbf{x}^{t+1} - \mathbf{b}\|$ diminishes, the increase in β^t also becomes less pronounced, giving rise to the term "adaptive" in these rules. ***(iii)*** The parameter ξ serves a dual purpose: it ensures theoretical convergence (see Theorem 4.5) and enhances practical performance. The increase in penalty might hinder the efficiency of ADMM. However, ξ offers control to prevent it from growing too rapidly. As demonstrated later in Algorithm 2 (Step 5), we set an upper bound for ξ, ensuring it remains at a sufficiently small positive constant to prevent excessive growth.

For $p \in (1, 2]$, the penalty update rule possesses a favorable property that streamlines our analysis. We have the following lemma.

Algorithm 1 IRPL-ADMM: The Proposed Inertial Relaxed Proximal Linearized ADMM for Solving Problem (1).

Initialize $\{\mathbf{x}^0, \mathbf{z}^0\}$. Let $\mathbf{x}^{-1} = \mathbf{x}^0$ and $\mathbf{y}^0 = \mathbf{x}^0$.
Use Algorithm 2 to choose suitable $\{\beta^0, \boldsymbol{\theta}, \boldsymbol{\alpha}, \xi, \sigma\}$.
for $t = 0$ to T **do**
  $\mathbf{x}_1^{t+1} \in \min_{\mathbf{x}_1} h_1(\mathbf{x}_1) + \frac{\boldsymbol{\theta}_1 \mathsf{L}_1^t}{2}\|\mathbf{x}_1 - \mathbf{y}_1^t\|_2^2 + \langle \mathbf{x}_1 - \mathbf{x}_1^t, \nabla_{\mathbf{x}_1} G(\mathbf{x}_{[1,n]}^t, \mathbf{z}^t; \beta^t)\rangle$
  $\mathbf{x}_2^{t+1} \in \min_{\mathbf{x}_2} h_2(\mathbf{x}_2) + \frac{\boldsymbol{\theta}_2 \mathsf{L}_2^t}{2}\|\mathbf{x}_2 - \mathbf{y}_2^t\|_2^2 + \langle \mathbf{x}_2 - \mathbf{x}_2^t, \nabla_{\mathbf{x}_2} G(\mathbf{x}_1^{t+1}, \mathbf{x}_{[2,n]}^t, \mathbf{z}^t; \beta^t)\rangle$
  ...
  $\mathbf{x}_n^{t+1} \in \min_{\mathbf{x}_n} h_n(\mathbf{x}_n) + \frac{\boldsymbol{\theta}_n \mathsf{L}_n^t}{2}\|\mathbf{x}_n - \mathbf{y}_n^t\|_2^2 + \langle \mathbf{x}_n - \mathbf{x}_n^t, \nabla_{\mathbf{x}_n} G(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_n^t, \mathbf{z}^t; \beta^t)\rangle$
  $\mathbf{y}_j^{t+1} = \mathbf{x}_j^{t+1} + \boldsymbol{\alpha}_j(\mathbf{x}_j^{t+1} - \mathbf{x}_j^t),\ \forall j \in [n]$
  $\mathbf{z}^{t+1} = \mathbf{z}^t + \sigma\beta^t([\sum_{j=1}^n \mathbf{A}_j \mathbf{x}_j^{t+1}] - \mathbf{b})$
  Use a (β^0, ξ, p)-regular penalty update rule to update β^{t+1} based on β^t.
end for

Lemma 2.6. *(Proof in Appendix B.5) Let $p \in (1, 2]$. We define $C_{\mathsf{b}} \triangleq \frac{1}{\beta^0} + \frac{p}{(p-1)\vartheta}$. The (β^0, ξ, p)-regular penalty update rule satisfies:*

$$\textstyle\sum_{t=0}^{\infty} \frac{1}{\beta^t} \le C_{\mathsf{b}}. \tag{3}$$

2.2. The Proposed IRPL-ADMM in Algorithm 1

This section provides the proposed **IRPL-ADMM** algorithm. We begin with providing the augmented Lagrangian function of Problem (1) as follows:

$$\mathcal{L}(\mathbf{x}, \mathbf{z}; \beta) \triangleq G(\mathbf{x}, \mathbf{z}; \beta) + \textstyle\sum_{i=1}^n h_i(\mathbf{x}_i), \tag{4}$$

where $G(\mathbf{x}, \mathbf{z}; \beta)$ represents the differential component of $\mathcal{L}(\mathbf{x}, \mathbf{z}; \beta)$ and is defined as:

$$\begin{aligned} G(\mathbf{x}, \mathbf{z}; \beta) \triangleq \textstyle\sum_{i=1}^n f_i(\mathbf{x}_i) + \langle [\sum_{i=1}^n \mathbf{A}_i \mathbf{x}_i] - \mathbf{b}, \mathbf{z}\rangle \\ + \tfrac{\beta}{2}\|[\textstyle\sum_{i=1}^n \mathbf{A}_i \mathbf{x}_i] - \mathbf{b}\|_2^2. \end{aligned} \tag{5}$$

Here, $\mathbf{z} \in \mathbb{R}^{m\times 1}$ and $\beta > 0$ are respectively the dual variable and penalty parameter. We employ an increasing penalty scheme throughout all iterations $t = \{0, 1, \ldots, \infty\}$. Notably, the function $G(\mathbf{x}^t, \mathbf{z}^t; \beta^t)$ is L_i^t-smooth *w.r.t.* $\mathbf{x}_i$ for all $i \in [m]$, where $\mathsf{L}_i^t = L_i + \beta^t\|\mathbf{A}_i\|_2^2$.

In each iteration, we use the proximal linearized method to cyclically update the variables $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. Specifically, we update each variable $\mathbf{x}_i$ by solving the following subproblem for all $i \in [n]$: $\mathbf{x}_i^{t+1} \approx \arg\min_{\mathbf{x}_i \in \mathbb{R}^{\mathbf{d}_i\times 1}} \mathcal{L}(\mathbf{x}_{[1,i-1]}^t, \mathbf{x}_i, \mathbf{x}_{[i+1,n]}^t, \mathbf{z}^t; \beta^t)$. To tackle the $\mathbf{x}_i$-subproblem, we employ an inertial proximal linearized minimization strategy (Pock & Sabach, 2016):

$$\begin{aligned} \mathbf{x}_i^{t+1} \in \arg\min_{\mathbf{x}_i} h_i(\mathbf{x}_i) + \tfrac{\boldsymbol{\theta}_i \mathsf{L}_i^t}{2}\|\mathbf{x}_i - \mathbf{y}_i^t\|_2^2 \\ + \langle \mathbf{x}_i - \bar{\mathbf{x}}_i, \nabla_{\mathbf{x}_i} G(\mathbf{x}_{[1,i+1]}^{t+1}, \mathbf{x}_i, \mathbf{x}_{[i+1,n]}^t, \mathbf{z}^t; \beta^t)\rangle, \end{aligned}$$

and $\mathbf{y}_j^t$ is updated via: $\mathbf{y}_j^{t+1} = \mathbf{x}_j^{t+1} + \boldsymbol{\alpha}_j(\mathbf{x}_j^{t+1} - \mathbf{x}_j^t),\ \forall j \in [n]$. Importantly, we introduce distinct inertial parameters $\boldsymbol{\alpha}_i$ and proximal parameters $\boldsymbol{\theta}_i$ with $i \in [n]$ for different blocks. Our algorithm updates the dual variable $\mathbf{z}^t$ using either under-relaxed stepsize ($\sigma \in (0, 1)$) or over-relaxed stepsize ($\sigma \in (1, 2)$). As the parameters $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \mathbf{y}, \mathbf{z})$ are updated sequentially, the penalty β^t is increased using a (β^0, ξ, p)-regular penalty update rule, with p set to the default value of 2.

We present **IRPL-ADMM** in Algorithm 1, a generalization of cyclic coordinate descent. It guarantees convergence when employing the parameter selection procedure outlined in Algorithm 2 for $\{\beta^0, \boldsymbol{\theta}, \boldsymbol{\alpha}, \xi, \sigma\}$.

Algorithm 2 A Procedure for Finding Suitable Parameters $\xi \in (0, \epsilon_1)$, $\boldsymbol{\alpha} \in (0, 1)^n$, $\boldsymbol{\theta} \in (1, \infty)^n$, $\sigma \in (0, 2)$, $\beta^0 \in (0, \infty)$ for Algorithm 1 (Deterministic Settings).

1: Choose suitable $(\epsilon_1, \epsilon_2, \epsilon_3)$. Default parameters:

$$\text{Cond. } \boxed{\mathbb{I}} : (\epsilon_1, \epsilon_2, \epsilon_3) = (0.01, 0.01, 0.001) \tag{6}$$

$$\text{Cond. } \boxed{\mathbb{A}} : (\epsilon_1, \epsilon_2, \epsilon_3) = (0.01, 1, 0.001) \tag{7}$$

2: For all $i \in [n]$, we define $\boldsymbol{\gamma}_i \triangleq \frac{1}{2}[\boldsymbol{\theta}_i - 1 - (2+\epsilon_1)\boldsymbol{\alpha}_i\boldsymbol{\theta}_i]$,

$$\boldsymbol{\gamma}_i' \triangleq \boldsymbol{\gamma}_i[1 - \epsilon_3]. \tag{8}$$

3: For the first $(n-1)$ blocks, find suitable parameters $\{\boldsymbol{\alpha}_i, \boldsymbol{\theta}_i\}_{i=1}^{n-1}$ such that $\boldsymbol{\gamma}_i' > 0$ for all $i \in [n-1]$.
4: For the last block, find suitable parameters $(\boldsymbol{\alpha}_n, \boldsymbol{\theta}_n, \sigma)$ such that (9) or (10) holds.
• Condition $\boxed{\mathbb{I}}$: Over-Relaxation Stepsize $\sigma \in [1, 2)$.

$$\begin{cases} \sigma \in [1, 2), \boldsymbol{\gamma}_n' > 0, \\ \underbrace{8\sigma_1\delta \cdot (1+\epsilon_3)}_{=4C_{\mathsf{u}}}[(\chi-1)^2 + \tau\chi] \le \boldsymbol{\gamma}_n'. \end{cases} \tag{9}$$

• Condition $\boxed{\mathbb{A}}$: Under-Relaxation Stepsize $\sigma \in (0, 1)$.

$$\begin{cases} \sigma \in (0, 1), \boldsymbol{\gamma}_n' > 0, \\ \underbrace{\overline{\lambda}/\underline{\lambda} \cdot 8\sigma\delta}_{=2\overline{\lambda}C_{\mathsf{u}}} \cdot (\chi^2 + \chi\tau) \le \boldsymbol{\gamma}_n'. \end{cases} \tag{10}$$

Here, $\{\delta, \chi, \boldsymbol{\alpha}_n'\}$ in (9) and (10) are defined as:

$$\delta \triangleq 1 + \epsilon_2, \chi \triangleq \boldsymbol{\theta}_n(1+\epsilon_3), \tau \triangleq \boldsymbol{\alpha}_n^2(1+\epsilon_1). \tag{11}$$

5: Choose β^0 and ξ satisfying Assumption 2.7 that: $\xi \le \min(\epsilon_1, \epsilon_2\sigma)$, $\beta^0 \ge L_i/(\epsilon_3\overline{\lambda})$ for all $i \in [n]$.

2.3. Choosing Suitable Parameters in Algorithm 2

This subsection discusses how to choose suitable parameters $\{\beta^0, \boldsymbol{\theta}, \boldsymbol{\alpha}, \xi, \sigma\}$ in Algorithm 2.

To simplify our discussions and derive more practical parameters, we make the following assumption concerning the parameters (β^0, ξ).

Assumption 2.7. Let $(\epsilon_1, \epsilon_2, \epsilon_3)$ be some small positive constants with $\max(\epsilon_1, \epsilon_2, \epsilon_3) \leq 1$. We assume:

$$\xi \leq \min(\epsilon_1, \sigma\epsilon_2), \text{ and } \beta^0 \geq L_i/(\epsilon_3\overline{\lambda}), \forall i \in [n]. \quad (12)$$

We use default parameters in Inequalities (6) and (7).

The following points are notable in Algorithm 2. ***(i)*** We impose a lower bound on β^0, which always holds as t becomes sufficiently large due to the increasing penalty update rule. ***(ii)*** When $\boldsymbol{\theta}_i \leq 1$ and $\boldsymbol{\alpha}_i > 0$ for some $i \in [n]$, we can never find a strictly positive $\boldsymbol{\gamma}_i > 0$ to guarantee convergence. ***(iii)*** For Cond. $\boxed{\mathbb{I}}$, consider the default parameter setting: $(\boldsymbol{\theta}_{[1:(n-1)]}, \boldsymbol{\theta}_n) = (1.05, 1.001)$, $(\boldsymbol{\alpha}_{[1:(n-1)]}, \boldsymbol{\alpha}_n) = (0.023, 0.0002)$, and $(\sigma, \xi) = (1.5, 0.01)$ [1] ***(iv)*** For Cond. $\boxed{\mathbb{A}}$, consider the default parameter setting: $(\boldsymbol{\theta}_{[1:(n-1)]}, \boldsymbol{\theta}_n) = (1.05, 1.5)$, $(\boldsymbol{\alpha}_{[1:(n-1)]}, \boldsymbol{\alpha}_n) = (0.023, 0.099)$, and $\sigma = \xi = 5/1000 \cdot \underline{\lambda}/\overline{\lambda}$ [2].

3. Global Convergence

This section provides global convergence for Algorithm 1.

Initially, we provide the following useful lemma.

Lemma 3.1. *(Proof in Appendix C.1,* Decrease for the Primal*) We define $\delta \triangleq 1 + \epsilon_2$. We have:*

$$\begin{aligned} &\mathcal{E}^{t+1} + \Theta_{\mathsf{o}}^{t+1} - \Theta_{\mathsf{o}}^{t} \\ \leq\ & \tfrac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 + \boldsymbol{\gamma}_n(\epsilon_3 - 1)\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2. \end{aligned} \quad (13)$$

where $\{\mathcal{E}^{t+1}, \Theta_{\mathsf{o}}^t, \mathbf{r}^t, \boldsymbol{\gamma}, \Delta_i^t, \mathsf{L}_i^t\}$ are respectively defined as:

$$\mathcal{E}^{t+1} \triangleq \tfrac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2 + \epsilon_3\textstyle\sum_{i=1}^n \boldsymbol{\gamma}_i\mathsf{L}_i^t\|\Delta_i^{t+1}\|_2^2, \quad (14)$$

$$\Theta_{\mathsf{o}}^t \triangleq \mathcal{L}(\mathbf{x}^t, \mathbf{z}^t; \beta^t) + \tfrac{1}{2}\textstyle\sum_{i=1}^n \boldsymbol{\theta}_i\boldsymbol{\alpha}_i\mathsf{L}_i^t\|\Delta_i^t\|_2^2 \quad (15)$$

$$\mathbf{r}^t \triangleq [\textstyle\sum_{i=1}^n \mathbf{A}_i\mathbf{x}_i^t] - \mathbf{b} \triangleq \mathbf{A}\mathbf{x}^t - \mathbf{b} \quad (16)$$

$$\boldsymbol{\gamma}_i \triangleq \tfrac{1}{2}[\boldsymbol{\theta}_i - 1 - (2+\epsilon_1)\boldsymbol{\alpha}_i\boldsymbol{\theta}_i],\ \forall i \in [n] \quad (17)$$

$$\Delta_i^t \triangleq \mathbf{x}_i^t - \mathbf{x}_i^{t-1}, \mathsf{L}_i^t = L_i + \beta^t\|\mathbf{A}_i\|_2^2. \quad (18)$$

Lemma 3.2. *(Proof in Appendix C.2,* First-Order Optimality Condition*) Assume $\sigma \in (0, 2)$. We let $i \in [n]$, $\mathbb{w}_i^{t+1} \in \partial h_i(\mathbf{x}_i^{t+1}) + \nabla f_i(\mathbf{x}_i^t)$, and $\mathbb{u}_i^{t+1} = \boldsymbol{\theta}_i\mathsf{L}_i^t(\mathbf{x}_i^{t+1} - \mathbf{x}_i^t - \boldsymbol{\alpha}_i(\mathbf{x}_i^t - \mathbf{x}_i^{t-1})) - \beta^t\mathbf{A}_i^\mathsf{T}[\sum_{j=i}^n \mathbf{A}_j(\mathbf{x}_j^{t+1} - \mathbf{x}_j^t)]$. For all $i \in [n]$, it holds that: $\mathbf{0} = \sigma\mathbf{A}_i^\mathsf{T}\mathbf{z}^t + \mathbf{A}_i^\mathsf{T}(\mathbf{z}^{t+1} - \mathbf{z}^t) + \sigma\mathbb{w}_i^{t+1} + \sigma\mathbb{u}_i^{t+1}$. We have the following two different identities:*

$$\text{Cond. } \boxed{\mathbb{I}}: \begin{cases} \mathbb{a}^{t+1} = (1-\sigma)\mathbb{a}^t + \sigma\mathbb{c}^t, \\ \mathbb{a}^{t+1} \triangleq \mathbf{A}_n^\mathsf{T}(\mathbf{z}^{t+1} - \mathbf{z}^t), \\ \mathbb{c}^t \triangleq \mathbb{u}_n^t - \mathbb{u}_n^{t+1} + \mathbb{w}_n^t - \mathbb{w}_n^{t+1}. \end{cases} \quad (19)$$

$$\text{Cond. } \boxed{\mathbb{A}}: \begin{cases} \mathbb{a}^{t+1} = (1-\sigma)\mathbb{a}^t + \sigma\mathbb{c}^t, \\ \mathbb{a}^{t+1} \triangleq \mathbf{A}_n^\mathsf{T}(\mathbf{z}^{t+1} - \mathbf{z}^t) + \sigma\mathbb{u}_n^{t+1}, \\ \mathbb{c}^t \triangleq \sigma\mathbb{u}_n^t + \mathbb{w}_n^t - \mathbb{w}_n^{t+1}. \end{cases} \quad (20)$$

Here, $\mathbb{u}_n^t \triangleq \mathbb{H}^{t-1}(\mathbf{x}_n^t - \mathbf{x}_n^{t-1}) - \boldsymbol{\theta}_n\boldsymbol{\alpha}_n\mathsf{L}_n^{t-1}(\mathbf{x}_n^{t-1} - \mathbf{x}_n^{t-2})$, where $\mathbb{H}^t \triangleq \boldsymbol{\theta}_n\mathsf{L}_n^t\mathbf{I} - \beta^t\mathbf{A}_n^\mathsf{T}\mathbf{A}_n$.

The following lemma bounds the terms $\|\mathbb{w}_n^{t+1} - \mathbb{w}_n^t\|_2^2$ and $\frac{1}{\beta^t}\|\mathbb{u}_n^{t+1}\|_2^2$.

Lemma 3.3. *(Proof in Appendix C.3) We define $\iota \triangleq 8C_h^2 + 8C_f^2$, $\chi \triangleq \boldsymbol{\theta}_n(1+\epsilon_3)$, $\tau = \boldsymbol{\alpha}_n^2(1+\epsilon_1)$, $\rho \triangleq 2\overline{\lambda}\chi\boldsymbol{\alpha}_n^2$, and $\Theta_{\mathsf{x}}^t \triangleq \rho\mathsf{L}_n^t\|\Delta_n^t\|_2^2$. We have:*

(a) $\|\mathbb{w}_n^{t+1} - \mathbb{w}_n^t\|_2^2 \leq \iota$.

(b) $\frac{1}{\beta^t}\|\mathbb{u}_n^{t+1}\|_2^2 \leq 2\overline{\lambda}\cdot\{(\chi - \underline{\lambda}/\overline{\lambda})^2 + \chi\tau\}\cdot\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 + \Theta_{\mathsf{x}}^t - \Theta_{\mathsf{x}}^{t+1}$.

We provide convergence analysis of Algorithm 1 under two conditions: Condition $\boxed{\mathbb{I}}$ using Formulation (19), and Condition $\boxed{\mathbb{A}}$ using Formulation (20).

We first the following parameters for different Conditions $\boxed{\mathbb{I}}$ and $\boxed{\mathbb{A}}$:

$$\boxed{\mathbb{I}}: \begin{cases} C_{\mathsf{a}} \triangleq \delta\sigma_2, C_{\mathsf{u}} \triangleq 2\delta\sigma_1(1+\epsilon_3), \\ C_{\mathsf{x}} \triangleq 2C_{\mathsf{u}}\rho, C_{\mathsf{w}} \triangleq \iota\delta\sigma_1(1 + \frac{1}{\epsilon_3}). \end{cases} \quad (21)$$

$$\boxed{\mathbb{A}}: \begin{cases} C_{\mathsf{a}} \triangleq 2\delta\sigma_2/\underline{\lambda}, C_{\mathsf{u}} \triangleq 4\delta\sigma/\underline{\lambda}, \\ C_{\mathsf{x}} \triangleq 4\rho\delta\sigma/\underline{\lambda}, C_{\mathsf{w}} \triangleq \iota 4\delta/(\sigma\underline{\lambda}). \end{cases} \quad (22)$$

Here, $\sigma \in (0, 2)$, and $\{\sigma_1, \sigma_2\}$ are defined as:

$$\sigma_1 \triangleq \tfrac{\sigma}{(1-|1-\sigma|)^2},\ \sigma_2 \triangleq \tfrac{|1-\sigma|}{\sigma(1-|1-\sigma|)}.$$

Using the parameters $\{C_{\mathsf{a}}, C_{\mathsf{u}}, C_{\mathsf{x}}\}$, we construct a sequence associated with the potential (or Lyapunov) function as follows:

$$\Theta^t = \Theta_{\mathsf{o}}^t + \tfrac{C_{\mathsf{a}}}{\beta^t}\|\mathbb{a}^t\|_2^2 + \tfrac{C_{\mathsf{u}}}{\beta^t}\|\mathbb{u}_n^t\|_2^2 + C_{\mathsf{x}}\mathsf{L}_n^t\|\Delta_n^t\|_2^2. \quad (23)$$

3.1. Analysis for Condition $\boxed{\mathbb{I}}$

We provide a convergence analysis of Algorithm 1 under Condition $\boxed{\mathbb{I}}$, where $\mathbf{A}_n$ is an identity matrix. We assume over-relaxation stepsize is used with $\sigma \in [1, 2)$.

The subsequent lemma utilizes Equation (19) to establish an upper bound for the term $\frac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2$.

[1] The auxiliary variable values are $(\gamma_n', \chi, \tau, \delta, C_{\mathsf{u}}) = (2.521 \cdot 10^{-4}, 1.002, 6.119 \cdot 10^{-8}, 1.01, 12.132)$.

[2] The auxiliary variable values are $(\gamma_n', \chi, \tau, \delta, C_{\mathsf{u}}\underline{\lambda}) = (0.203, 1.502, 0.01, 2, 0.0014)$.

Lemma 3.4. *(Proof in Appendix C.6,* Bounding Dual Using Primal*) We define* $\delta \triangleq 1+\epsilon_2$, $\chi \triangleq \boldsymbol{\theta}_n(1+\epsilon_3)$, *and* $\tau \triangleq \boldsymbol{\alpha}_n^2(1+\epsilon_1)$. *We have:*

$$\begin{aligned}&\tfrac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1}-\mathbf{z}^t\|_2^2\\ &\le \Theta_{\mathsf{z}}^t-\Theta_{\mathsf{z}}^{t+1}+\tfrac{C_{\mathsf{w}}}{\beta^t}+\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2\cdot 4C_{\mathsf{u}}[(\chi-1)^2+\chi\tau],\end{aligned}$$

where $\Theta_{\mathsf{z}}^t \triangleq \frac{C_{\mathsf{a}}}{\beta^t}\|\mathbb{a}^t\|_2^2+\frac{C_{\mathsf{u}}}{\beta^t}\|\mathbb{u}_n^t\|_2^2+C_{\mathsf{x}}\mathsf{L}_n^t\|\Delta_n^t\|_2^2$, *and* $\{C_{\mathsf{a}},C_{\mathsf{u}},C_{\mathsf{x}},C_{\mathsf{w}}\}$ *are defined in Equation (21).*

Theorem 3.5. *(Proof in Appendix C.7,* Decrease on a Potential Function*) For all* $t\ge 0$, *we have:*

$$\mathcal{E}^{t+1}\le \Theta^t-\Theta^{t+1}+\tfrac{C_{\mathsf{w}}}{\beta^t}.$$

3.2. Analysis for Condition $\boxed{\mathbb{A}}$

We provide the convergence analysis under Condition $\boxed{\mathbb{A}}$, where $\mathbf{A}$ is a full-row rank matrix with $\underline{\lambda}>0$. We assume under-relaxation stepsize is used with $\sigma\in(0,1)$.

The following lemma utilizes Equation (20) to establish an upper bound for the term $\frac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1}-\mathbf{z}^t\|_2^2$.

Lemma 3.6. *(Proof in Appendix C.8,* Bounding Dual Using Primal*) We define* $\delta \triangleq 1+\epsilon_2$, $\chi \triangleq \boldsymbol{\theta}_n(1+\epsilon_3)$, *and* $\tau \triangleq \boldsymbol{\alpha}_n^2(1+\epsilon_1)$. *We have:*

$$\begin{aligned}&\tfrac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1}-\mathbf{z}^t\|_2^2\\ &\le \Theta_{\mathsf{z}}^t-\Theta_{\mathsf{z}}^{t+1}+\tfrac{C_{\mathsf{w}}}{\beta}+\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2\cdot 2\overline{\lambda}C_{\mathsf{u}}(\chi^2+\chi\tau),\end{aligned}$$

where $\Theta_{\mathsf{z}}^t \triangleq \frac{C_{\mathsf{a}}}{\beta^t}\|\mathbb{a}^t\|_2^2+\frac{C_{\mathsf{u}}}{\beta^t}\|\mathbb{u}_n^t\|_2^2+C_{\mathsf{x}}\mathsf{L}_n^t\|\Delta_n^t\|_2^2$, *and* $\{C_{\mathsf{a}},C_{\mathsf{u}},C_{\mathsf{x}},C_{\mathsf{w}}\}$ *are defined in Equation (22).*

Theorem 3.7. *(Proof in Appendix C.9,* Decrease on a Potential Function*). For all* $t\ge 0$, *we have:*

$$\mathcal{E}^{t+1}\le \Theta^t-\Theta^{t+1}+\tfrac{C_{\mathsf{w}}}{\beta^t}.$$

3.3. Continuing Analysis for Conditions $\boxed{\mathbb{I}}$ and $\boxed{\mathbb{A}}$

We first obtain the following lemma.

Lemma 3.8. *(Proof in Appendix C.4) We have* $\Theta^t\ge\underline{\Theta}$, *where* $\underline{\Theta}$ *is defined in Assumption 1.4.*

Finally, we have the following corollary from Theorems 3.5 and 3.7.

Corollary 3.9. *(Proof in Appendix C.5,* a Square-Summable Property*) Assume* $p\in(1,2]$. *We have:* $\sum_{t=0}^{\infty}\mathcal{E}^{t+1}\le C_{\mathsf{p}}\triangleq \Theta^0-\underline{\Theta}+C_{\mathsf{w}}C_{\mathsf{b}}$, *where* C_{w} *is defined in (21) or (22), and* C_{b} *is defined in (3).*

Remarks. ***(i)*** In view of Corollary 3.9 and the definition of $\mathcal{E}^{t+1}$ in Equation (14), as $t\to\infty$, we observe that: $\beta^t\|\mathbf{A}\mathbf{x}^{t+1}-\mathbf{b}\|_2^2\to 0$ and $\beta^t\|\mathbf{x}_i^{t+1}-\mathbf{x}_i^t\|_2^2\to 0$ for all $i\in[n]$. This observation implies the convergence of Algorithm 1. ***(ii)*** Consider the penalty error term $\frac{\beta^t}{2}\|\mathbf{A}\mathbf{x}^{t+1}-\mathbf{b}\|_2^2$ in the argument Lagrangian function. Despite increasing β^t, the term $\frac{1}{2}\|\mathbf{A}\mathbf{x}^{t+1}-\mathbf{b}\|_2^2$ decreases faster than $1/\beta^t$, leading to the gradual disappearance of the term $\frac{\beta^t}{2}\|\mathbf{A}\mathbf{x}^{t+1}-\mathbf{b}\|_2^2$.

4. Iteration Complexity

This section provides iteration complexity of Algorithm 1.

► **Surrogate Stationarity Measure**. Due to the use of the increasing penalty and inertial update strategy, the standard critical point may not be an appropriate stationarity measure. Inspired by recent research (Davis & Drusvyatskiy, 2019; Li et al., 2021), we use a surrogate stationarity measure that can monitor the progress of Algorithm 1.

Given parameters $\{\beta,\boldsymbol{\theta},\boldsymbol{\alpha},\sigma\}$, Steps 4-9 in Algorithm 1 form a fixed-point procedure, from which we introduce the following definition of ϵ-Inertial Nonconvex Proximal Point.

Definition 4.1. (ϵ-Inertial Nonconvex Proximal Point, or ϵ-INP Point for short) Given $\beta\in(0,\infty)$, $\boldsymbol{\theta}>1$, $\boldsymbol{\alpha}\ge\mathbf{0}$, $\sigma\in(0,2)$, and a point $(\ddot{\mathbf{x}},\ddot{\mathbf{y}},\ddot{\mathbf{z}})$. For all $i\in[n]$, we define $\mathsf{L}_i \triangleq L_i+\beta\|\mathbf{A}_i\|_2^2$, and let $\ddot{\mathbf{x}}_i^+\in\arg\min_{\mathbf{x}_i} h_i(\mathbf{x}_i)+\frac{1}{2}\boldsymbol{\theta}_i\mathsf{L}_i^t\|\mathbf{x}_i-\ddot{\mathbf{y}}_i\|_2^2+\langle\mathbf{x}_i-\ddot{\mathbf{x}}_i,\nabla_{\mathbf{x}_i}G(\ddot{\mathbf{x}}^+_{[1,i-1]},\ddot{\mathbf{x}}_{[i,n]},\ddot{\mathbf{z}};\beta)\rangle$, $\ddot{\mathbf{y}}_i^+=\ddot{\mathbf{x}}_i^+-\boldsymbol{\alpha}_i(\ddot{\mathbf{x}}_i^+-\ddot{\mathbf{x}}_i)$. The point $(\ddot{\mathbf{x}},\ddot{\mathbf{y}},\ddot{\mathbf{z}})$ is an ϵ-INP point if it holds that: $\beta\|\mathbf{A}\ddot{\mathbf{x}}^+-\mathbf{b}\|_2^2+\beta\sum_{i=1}^n[\|\ddot{\mathbf{x}}_i^+-\ddot{\mathbf{x}}_i\|_2^2+\|\ddot{\mathbf{y}}_i^+-\ddot{\mathbf{y}}_i\|_2^2]\le\epsilon$.

To illustrate the connection with the existing definition of optimality conditions, we define approximated critical, and directional points.

Definition 4.2. (ϵ-Critical Point) A solution $(\check{\mathbf{x}},\check{\mathbf{z}})$ is an ϵ-critical point if it holds that for all $i\in[n]$: $\mathrm{dist}^2(\mathbf{0},\nabla f_i(\check{\mathbf{x}}_i)+\partial h_i(\check{\mathbf{x}}_i)+[\mathbf{A}_i]^{\mathsf{T}}\check{\mathbf{z}})\le\epsilon$, and $\|\mathbf{A}_i\check{\mathbf{x}}-\mathbf{b}\|_2^2\le\epsilon$, where $\mathrm{dist}^2(\Omega,\Omega')\triangleq\inf_{\mathbf{w}\in\Omega,\mathbf{w}'\in\Omega'}\|\mathbf{w}-\mathbf{w}'\|_2^2$ denotes the squared distance between two sets.

Definition 4.3. (ϵ-Directional Point) A solution $(\grave{\mathbf{x}},\grave{\mathbf{z}})$ is an ϵ-directional point if, for some $\beta>0$, for all $i\in[n]$ with $\mathbf{x}_i\in\mathrm{dom}(\mathcal{L}_i)$, we have: $\|\mathbf{A}_i\grave{\mathbf{x}}-\mathbf{b}\|_2^2\le\epsilon$, and $\mathcal{L}_i'(\mathbf{x}_i-\grave{\mathbf{x}}_i;\grave{\mathbf{x}},\grave{\mathbf{z}})\ge-\epsilon$. Here, $\mathcal{J}_i'(\Delta;\grave{\mathbf{x}},\grave{\mathbf{z}})\triangleq\lim_{t\to0}\frac{1}{t}[\mathcal{L}(\grave{\mathbf{x}}_{[1,i-1]},\grave{\mathbf{x}}_i+\Delta,\grave{\mathbf{x}}_{[i+1,n]},\grave{\mathbf{z}};\beta)-\mathcal{L}(\grave{\mathbf{x}},\grave{\mathbf{z}};\beta)]$, and $\mathrm{dom}(\mathcal{L}_i)\triangleq\{\mathbf{x}_i\,|\,\mathcal{L}(\grave{\mathbf{x}}_{[1,i-1]},\mathbf{x}_i,\grave{\mathbf{x}}_{[i+1,n]},\grave{\mathbf{z}};\beta)|<\infty\}$.

Remarks. Note that we apply the standard definition of direction point (Pang et al., 2017; Rockafellar & Wets., 2009) to each block of the augmented Lagrangian function.

The following theorem establishes their hierarchy at $\epsilon=0$.

Theorem 4.4. *(Proof in Appendix D.1, Optimality Hierarchy) The following results hold if* $\epsilon=0$: ***(i)*** Any INP-point is a critical-point and a directional-point, while the reverse is not necessarily true. ***(ii)*** Any optimal point is an INP-point.

Remarks. ***(i)*** Our method identifies stationary points that

are stronger than critical points and directional points. This contrasts with commonly used approaches, such as multi-stage convex relaxation (Zhang, 2010) and DC programming methods, which only find critical points of Problem (1). Such results are based on the assumption that the subproblem for the nonconvex operator can be solved globally. ***(ii)*** When $\epsilon \neq 0$, establishing relations among different optimality conditions becomes challenging.

We now demonstrate that Algorithm 1 converges to INP-points of Problem (1) at a rate of $\mathcal{O}(1/\epsilon)$. We have the following theorem.

Theorem 4.5. *(Proof in Appendix C.10) Let the sequence $\{\mathbf{x}^t, \mathbf{y}^t, \mathbf{z}^t\}_{t=0}^T$ be generated by Algorithm 1. There exists an index $\bar{t}$ with $0 \leq \bar{t} \leq T$ such that $\beta^{\bar{t}}\|\mathbf{r}^{\bar{t}+1}\|_2^2 + \beta^{\bar{t}}\sum_{i=1}^n[\|\mathbf{x}_i^{\bar{t}+1}-\mathbf{x}_i^{\bar{t}}\|_2^2+\|\mathbf{y}_i^{\bar{t}+1}-\mathbf{y}_i^{\bar{t}}\|_2^2] \leq \frac{C_{\mathsf{p}}\max(1/c_1,1/c_2)}{T}$, where $c_0 \triangleq \epsilon_3 \min_{i=1}^n \boldsymbol{\gamma}_i\|\mathbf{A}_i\|$, $c_1 \triangleq \frac{c_0}{17}$, $c_2 \triangleq \frac{\xi}{2}$, and C_{p} is defined in Corollary (3.9). It implies that Algorithm 1 finds an ϵ-INP point of Problem (1) in at most T iterations, where $T \leq \lceil \frac{C_{\mathsf{p}}\max(1/c_1,1/c_2)}{\epsilon} \rceil = \mathcal{O}(\epsilon^{-1})$.*

Remarks. The sequence $\beta^t(\|\mathbf{r}^{t+1}\|_2^2 + \sum_{i=1}^n[\|\mathbf{x}_i^{t+1} - \mathbf{x}_i^t\|_2^2 + \|\mathbf{y}_i^{t+1} - \mathbf{y}_i^t\|_2^2])$ converging to 0 for any $\beta^t \in (0, \infty)$ implies that Algorithm 1 converges to an INP point of Problem (1), which is also a critical point.

5. Strong Limit-Point Convergence

This section provides strong limit-point convergence of IRPL-ADMM. Our analyses are based on a non-convex analysis tool called Kurdyka-Łojasiewicz (KŁ) inequality (Attouch et al., 2010; Bolte et al., 2014; Li et al., 2023).

We denote $\mathbb{X} \triangleq \{\mathbf{x}, \mathbf{z}, \mathbf{x}', \mathbf{x}''\}$. For different Conditions $\boxed{\mathbb{I}}$ and $\boxed{\mathbb{A}}$, we define the Lyapunov function as:

$$\Theta(\mathbb{X}; \beta, \beta') \triangleq \begin{cases} \frac{C_{\mathsf{a}}}{\beta}\|\sigma\beta\mathbf{A}_n^{\mathsf{T}}\mathbf{r}\|_2^2 + \upsilon, & \boxed{\mathbb{I}}; \\ \frac{C_{\mathsf{a}}}{\beta}\|\sigma\beta\mathbf{A}_n^{\mathsf{T}}\mathbf{r} + \sigma\mathfrak{u}\|_2^2 + \upsilon, & \boxed{\mathbb{A}}. \end{cases}$$

Here, $\upsilon = \mathcal{L}(\mathbf{x}, \mathbf{z}; \beta) + \frac{C_{\mathfrak{u}}}{\beta}\|\mathfrak{u}\|_2^2 + \frac{1}{2}\sum_{i=1}^n \boldsymbol{\eta}_i\|\mathbf{x}_i - \mathbf{x}_i'\|_2^2 + C_{\mathsf{x}}\mathsf{L}_n\|\mathbf{x}_n - \mathbf{x}_n'\|_2^2$, and $\mathbf{r} = \mathbf{A}\mathbf{x} - \mathbf{b}$. Furthermore, $\mathfrak{u} = \mathbb{H}(\mathbf{x}_n - \mathbf{x}_n') - \boldsymbol{\eta}_n(\mathbf{x}_n' - \mathbf{x}_n'')$, $\mathbb{H} \triangleq \boldsymbol{\theta}_n\mathsf{L}_n'\mathbf{I} - \beta'\mathbf{A}_n^{\mathsf{T}}\mathbf{A}_n$, $\boldsymbol{\eta}_i \triangleq \boldsymbol{\theta}_i\boldsymbol{\alpha}_i\mathsf{L}_i'$, $\mathsf{L}_i \triangleq L_i + \beta\|\mathbf{A}_i\|_2^2$, $\mathsf{L}_i' \triangleq L_i + \beta'\|\mathbf{A}_i\|_2^2$, $\forall i$. Clearly, we have $\Theta^t = \Theta(\mathbf{x}^t, \mathbf{z}^t, \mathbf{x}^{t-1}, \mathbf{x}^{t-2}; \beta^t, \beta^{t-1})$.

We define $\mathbb{X}^t \triangleq \{\mathbf{x}^t, \mathbf{z}^t, \mathbf{x}^{t-1}, \mathbf{x}^{t-2}\}$, and let $F(\mathbb{X}^t) \triangleq \Theta(\mathbb{X}^t; \beta^t, \beta^{t+1})$. We denote $\mathbb{X}^*$ as a limiting point of $\{\mathbb{X}^t\}_{t=0}^\infty$. We let $F(\mathbb{X}^*) \triangleq \Theta(\mathbb{X}^*; \beta, \beta')$, where β and β' are the associated penalties for $\mathbb{X}^*$.

We make the following additional assumption.

Assumption 5.1. (Kurdyka-Łojasiewicz Inequality). Consider a semi-algebraic function $F(\mathbb{X}^t)$ *w.r.t.* $\mathbb{X}^t$ with $\mathbb{X}^t \in \mathrm{dom}(F)$. There exist $\tilde{\sigma} \in [0, 1)$, $\tilde{\eta} \in (0, +\infty]$, a neighborhood Υ of $\mathbb{X}^*$, and a continuous and concave desingularization function $\varphi(t) \triangleq ct^{1-\tilde{\sigma}}$ with $c > 0$ and $t \in [0, \tilde{\eta})$ such that, for all $\mathbb{X}^t \in \Upsilon$ satisfying $F(\mathbb{X}^t) \in (F(\mathbb{X}^*), F(\mathbb{X}^*) + \tilde{\eta})$, it holds that: $\mathrm{dist}(\mathbf{0}, \partial F(\mathbb{X}^t)) \cdot \varphi'(F(\mathbb{X}^t) - F(\mathbb{X}^*)) \geq 1$.

Semi-algebraic functions, including real polynomial functions, finite combinations, products, and indicator functions of semi-algebraic sets, commonly exhibit the KŁ property and find extensive use in various applications.

We have the following useful lemma.

Lemma 5.2. *(Proof in Section E.1,* Subgradient Bounds for Conditions $\boxed{\mathbb{I}}$ and $\boxed{\mathbb{A}}$*) There exists a constant $K > 0$ such that: $|||\partial\Theta(\mathbf{x}^t, \mathbf{z}^t, \mathbf{x}^{t-1}, \mathbf{x}^{t-2}; \beta^t, \beta^{t-1})||| \leq \beta^t K\{\sum_{i=1}^n[\|\mathbf{x}_i^{t-1} - \mathbf{x}_i^{t-2}\| + \|\mathbf{x}_i^t - \mathbf{x}_i^{t-1}\|] + \|\mathbf{r}^t\|\}$.*

Finally, we have the following convergence results.

Theorem 5.3. *(Proof in Section E.2,* A Finite Length Property*) We define $e^t \triangleq \|\mathbf{r}^t\| + \sum_{i=1}^n \|\Delta_i^t\|$. Then the sequence $\{\mathbb{X}^t\}_{t=0}^\infty$ has the following finite length property:*

$$[\textstyle\sum_{t=0}^\infty e^t] \leq C_{\mathsf{e}} < +\infty,$$

where $C_{\mathsf{e}} \triangleq \frac{K\cdot\varphi(F(\mathbb{X}^0)-F(\mathbb{X}^*))}{2(n+1)V} + \frac{C_{\mathsf{w}}C_{\mathsf{b}}}{2(n+1)V\beta^0} + \frac{3}{2}e^0 + \frac{1}{2}e^{-1}$. Furthermore, $V \triangleq \min(\frac{\xi}{2}, \epsilon_3 \min_{i=1}^n \boldsymbol{\gamma}_i\|\mathbf{A}_i\|_2^2)$, K is defined in Lemma 5.2, and $\varphi(\cdot)$ is the desingularization function defined in Assumption 5.1.

Remarks. ***(i)*** Lemma 5.2 significantly differs from prior work that used a constant penalty due to the crucial role played by the increasing penalty. ***(ii)*** The finite-length property in Theorem 5.3 are much stronger convergence results compared to those in Theorem 4.5. ***(iii)*** While the work of (Li et al., 2023) establishes convergence rate for the iterates with diminishing step sizes by further exploring the KŁ exponent $\tilde{\sigma}$ in Assumption 5.1, we plan to extend their analysis to establish stronger convergence results for our problem, which we leave for future research.

6. Experiments

In this section, we compare the proposed algorithm IRPL-ADMM with existing non-convex optimization algorithms on two applications, namely the sparse PCA problem and noise sparse recovery problem.

All methods are implemented in MATLAB on an Intel 2.6 GHz CPU with 64 GB RAM. Appendix Section H.3 describes how to generate the data used in the experiments. We provide our code in the supplemental material.

We compare IRPL-ADMM with ***(i)*** the Subgradient method (SubGrad), ***(ii)*** the Penalty Decomposition Method (PDM), and ***(iii)*** the standard ADMM that does not use the inertial strategy with $\boldsymbol{\alpha} = \mathbf{0}$. All algorithms start with the common initial solution $\mathbf{x}^0$ drawn from a standard normal distribution. We use the theoretical default parameters for $\{\boldsymbol{\theta}, \boldsymbol{\alpha}, \xi, \sigma\}$ discussed in Section 2.3, and set $\beta^0 = 10$. We use the

adaptive (β^0, ξ, p) regular penalty update rule with $p = 2$ to update β^t for all methods.

6.1. Sparse PCA

Sparse Principal Component Analysis (Sparse PCA) extends traditional PCA by emphasizing a subset of informative variables with sparse loadings, reducing model complexity and enhancing interpretability. It is formulated as follows:

$$\min_{\mathbf{V}\in\mathbb{R}^{d'\times r'}} \frac{1}{2m'}\|\mathbf{D}-\mathbf{D}\mathbf{V}\mathbf{V}^\top\|_{\mathsf{F}}^2 + \rho\tilde{h}(\mathbf{V}),\ s.t.\ \mathbf{V}\in\mathcal{M},$$

where $\mathcal{M} \triangleq \{\mathbf{V}\,|\,\mathbf{V}^\top\mathbf{V}=\mathbf{I}\}$, $\mathbf{D}\in\mathbb{R}^{m'\times d'}$ is the data matrix, and $\tilde{h}(\mathbf{V}) = \hat{h}(\mathrm{vec}(\mathbf{V}))$ is the sparse-inducing function such as DC ℓ_1-largest-k function (Gotoh et al., 2018): $\hat{h}(\mathbf{y}) = \|\mathbf{y}\|_1 - \|\mathbf{y}\|_{[k]}$.

Introducing extra parameter $\mathbf{Y}$, this problem can be formulated as: $\min_{\mathbf{V},\mathbf{Y}} \frac{1}{2m'}\|\mathbf{D}-\mathbf{D}\mathbf{V}\mathbf{V}^\top\|_{\mathsf{F}}^2 + \rho\tilde{h}(\mathbf{V}),\ s.t.\ \mathbf{V}=\mathbf{Y},\ \mathbf{Y}\in\mathcal{M}$. It coincides with Problem (1) with $f_2(\mathbf{x}_1) = \frac{1}{2m'}\|\mathbf{D}-\mathbf{D}\mathbf{V}\mathbf{V}^\top\|_{\mathsf{F}}^2$, $h_2(\mathbf{x}_2) = \rho\tilde{h}(\mathbf{V})$, $f_1(\mathbf{x}_1) = 0$, $h_1(\mathbf{x}_1) = \mathcal{I}_{\mathcal{M}}(\mathbf{Y})$, and $\mathbf{A}_1 = \mathbf{A}_2 = \mathbf{I}$ with Condition $\boxed{\mathbb{I}}$.

The experimental results in Figure 1 show the following: ***(i)*** Both ADMM and IRPL-ADMM exhibit convergence when using the regular penalty update rule. ***(ii)*** IRPL-ADMM demonstrates faster convergence compared to other methods.

6.2. Noisy Sparse Recovery

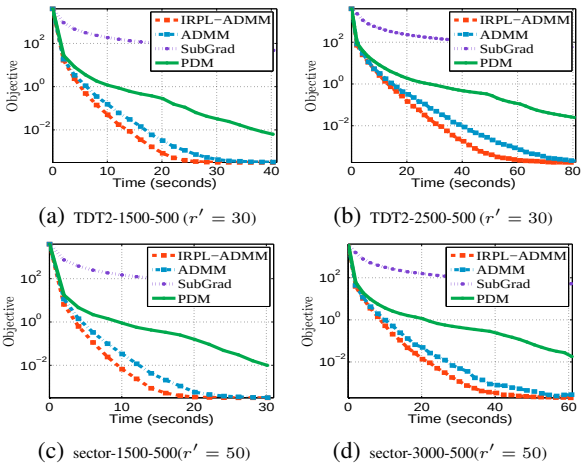
Noisy sparse recovery, a signal processing technique, acquires and reconstructs signals effectively by solving an underdetermined linear system. Given a design matrix $\mathbf{D}\in\mathbb{R}^{m'\times d'}$ and an observation vector $\mathbf{y}\in\mathbb{R}^{m'\times 1}$, it is formulated as follows (Lei Yang, 2021):

$$\min_{\mathbf{v}\in\mathbb{R}^{d'\times 1}} \|\mathbf{x}\|_q^q,\ s.t.\ \|\mathbf{D}\mathbf{v}-\mathbf{y}\|\le\tau, \tag{24}$$

with $q\in(0,1)$. Here, we set $q=\frac{1}{2}$. Introducing additional parameter $\mathbf{y}$, we have: $\min_{\mathbf{x},\mathbf{y}} \|\mathbf{x}\|_p^p,\ s.t.\ \|\mathbf{y}\|_2\le\tau,\ \mathbf{A}\mathbf{x}-\mathbf{y}=\mathbf{b}$. This formulation coincides with Problem (1) with $f_2(\mathbf{x}_2)=0$, $h_2(\mathbf{x}_2)=\|\mathbf{x}_2\|_p^p$, $f_1(\mathbf{x}_1)=0$, $h_1(\mathbf{x}_1)=\mathcal{I}_\Omega(\mathbf{x}_1)$, $\Omega=\{\mathbf{x}\,|\,\|\mathbf{x}\|\le\tau\}$, and $\mathbf{A}_1=\mathbf{I}$, $\mathbf{A}_2=\mathbf{D}$ with Condition $\boxed{\mathbb{A}}$.

As Problem (24) lacks Lipschitz continuity, we use an alternative Lipschitz function denoted as $F(\mathbf{v}) \triangleq \|\mathbf{v}\|_q^q + 10^3\times\max(0,\|\mathbf{G}\mathbf{v}-\mathbf{u}\|-\tau)$, to assess the quality of the solution.

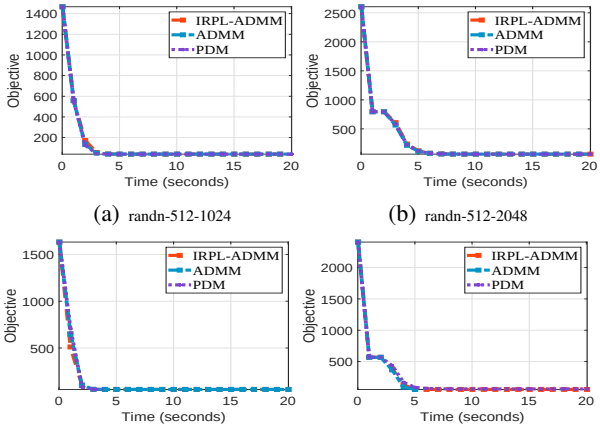
The experimental results in Figure 2 reveals the following findings: (ii) All three methods exhibit global convergence. (ii) The three methods yield similar results, with IRPL-ADMM and ADMM not demonstrating faster convergence than PDM. We attribute this to the possibility of the stepsize σ strictly satisfying Inequality (10) being too small and conservative.



(a) TDT2-1500-500 ($r'=30$) (b) TDT2-2500-500 ($r'=30$)

(c) sector-1500-500($r'=50$) (d) sector-3000-500($r'=50$)

Figure 1. The convergence curve of the compared methods for solving the Sparse PCA problem with $\rho = 10$.



(a) randn-512-1024 (b) randn-512-2048

(c) dct-512-1024 (d) dct-512-2048

Figure 2. The convergence curve of the compared methods for the noise sparse recovery problem.

7. Conclusions

In this paper, we introduce Inertial Proximal Linearized ADMM (IRPL-ADMM) for solving general multi-block nonconvex composite optimization problems. IRPL-ADMM operates under a relatively relaxed condition, requiring continuity in just one block of the objective function. It incorporates inertial strategies for primal variable updates and relaxed strategies for dual variable updates. We use a novel regular penalty update rule to control its growth rate and introduce a Lyapunov function for convergence analysis. We also derive the iteration complexity of IRPL-ADMM and extend it from deterministic to stochastic optimization. Finally, we conduct experiments to demonstrate the effectiveness of our approaches.

References

Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-lojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

Barber, R. F. and Sidky, E. Y. Convergence for nonconvex admm, with applications to ct imaging. *arXiv preprint arXiv:2006.07278*, 2020.

Bertsekas, D. *Convex optimization algorithms*. Athena Scientific, 2015.

Bian, F., Liang, J., and Zhang, X. A stochastic alternating direction method of multipliers for non-smooth and non-convex optimization. *Inverse Problems*, 37(7):075009, 2021.

Boţ, R. I. and Nguyen, D.-K. The proximal alternating direction method of multipliers in the nonconvex setting: convergence analysis and rates. *Mathematics of Operations Research*, 45(2):682–712, 2020.

Boţ, R. I., Csetnek, E. R., and Nguyen, D.-K. A proximal minimization algorithm for structured nonconvex and nonsmooth problems. *SIAM Journal on Optimization*, 29 (2):1300–1328, 2019. doi: 10.1137/18M1190689.

Böhm, A. and Wright, S. J. Variable smoothing for weakly convex composite functions. *Journal of Optimization Theory and Applications*, 188(3):628–649, 2021.

Bolte, J., Sabach, S., and Teboulle, M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

Boţ, R. I., Dao, M. N., and Li, G. Inertial proximal block coordinate method for a class of nonsmooth sum-of-ratios optimization problems. *SIAM Journal on Optimization*, 33(2):361–393, 2023.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Deng, W., Lai, M.-J., Peng, Z., and Yin, W. Parallel multi-block admm with o (1/k) convergence. *Journal of Scientific Computing*, 71:712–736, 2017.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018a.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018b.

Gabay, D. and Mercier, B. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & mathematics with applications*, 2(1):17–40, 1976.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155 (1-2):267–305, 2016.

Gonçalves, M. L., Melo, J. G., and Monteiro, R. D. Convergence rate bounds for a proximal admm with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems. *arXiv preprint arXiv:1702.01850*, 2017a.

Gonçalves, M. L., Melo, J. G., and Monteiro, R. D. Improved pointwise iteration-complexity of a regularized admm and of a regularized non-euclidean hpe framework. *SIAM Journal on Optimization*, 27(1):379–407, 2017b.

Gotoh, J., Takeda, A., and Tono, K. Dc formulations and algorithms for sparse optimization problems. *Mathematical Programming*, 169(1):141–176, 2018.

He, B. and Yuan, X. On the $\mathcal{O}(1/n)$ convergence rate of the douglas-rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

Hien, L. T. K., Phan, D. N., and Gillis, N. Inertial alternating direction method of multipliers for non-convex non-smooth optimization. *Computational Optimization and Applications*, 83(1):247–285, 2022.

Hong, M., Luo, Z.-Q., and Razaviyayn, M. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1):337–364, 2016.

Huang, F., Chen, S., and Huang, H. Faster stochastic alternating direction method of multipliers for nonconvex optimization. In *International Conference on Machine Learning (ICML)*, volume 97, pp. 2839–2848, 2019.

J Reddi, S., Sra, S., Poczos, B., and Smola, A. J. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. *Advances in neural information processing systems*, 29, 2016.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Lai, R. and Osher, S. A splitting method for orthogonality constrained problems. *Journal of Scientific Computing*, 58(2):431–449, 2014.

Le, H., Gillis, N., and Patrinos, P. Inertial block proximal methods for non-convex non-smooth optimization. In *International Conference on Machine Learning*, pp. 5671–5681. PMLR, 2020.

Lei Yang, Xiaojun Chen, S. X. Sparse solutions of a class of constrained optimization problems. *Mathematics of Operations Research*, 2021.

Li, G. and Pong, T. K. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4):2434–2460, 2015.

Li, H. and Lin, Z. Accelerated proximal gradient methods for nonconvex programming. *Advances in neural information processing systems*, 28, 2015.

Li, M., Sun, D., and Toh, K.-C. A majorized admm with indefinite proximal terms for linearly constrained convex composite optimization. *SIAM Journal on Optimization*, 26(2):922–950, 2016.

Li, Q., Zhou, Y., Liang, Y., and Varshney, P. K. Convergence analysis of proximal gradient with momentum for nonconvex optimization. In *International Conference on Machine Learning*, pp. 2111–2119. PMLR, 2017.

Li, X., Chen, S., Deng, Z., Qu, Q., Zhu, Z., and Man-Cho So, A. Weakly convex optimization over stiefel manifold using riemannian subgradient-type methods. *SIAM Journal on Optimization*, 31(3):1605–1634, 2021.

Li, X., Milzarek, A., and Qiu, J. Convergence of random reshuffling under the kurdyka–łojasiewicz inequality. *SIAM Journal on Optimization*, 33(2):1092–1120, 2023.

Lin, Q., Ma, R., and Xu, Y. Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. *Computational optimization and applications*, 82(1):175–224, 2022.

Lin, T., Ma, S., and Zhang, S. On the global linear convergence of the admm with multiblock variables. *SIAM Journal on Optimization*, 25(3):1478–1497, 2015a.

Lin, T.-Y., Ma, S.-Q., and Zhang, S.-Z. On the sublinear convergence rate of multi-block admm. *Journal of the Operations Research Society of China*, 3:251–274, 2015b.

Liu, Y., Shang, F., Liu, H., Kong, L., Jiao, L., and Lin, Z. Accelerated variance reduction stochastic admm for large-scale machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4242–4255, 2020.

Lu, Z. and Zhang, Y. Sparse approximation via penalty decomposition methods. *SIAM Journal on Optimization*, 23(4):2448–2478, 2013.

Monteiro, R. D. and Svaiter, B. F. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization*, 23(1):475–507, 2013.

Mordukhovich, B. S. Variational analysis and generalized differentiation i: Basic theory. *Berlin Springer*, 330, 2006.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Nesterov, Y. E. *Introductory lectures on convex optimization: a basic course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, 2003.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takác. Sarah: A novel method for machine learning problems using stochastic recursive gradient.

Nishihara, R., Lessard, L., Recht, B., Packard, A., and Jordan, M. A general analysis of the convergence of admm. In *International Conference on Machine Learning*, pp. 343–352. PMLR, 2015.

Ouyang, Y., Chen, Y., Lan, G., and Pasiliao Jr, E. An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences*, 8(1): 644–681, 2015.

Pang, J., Razaviyayn, M., and Alvarado, A. Computing b-stationary points of nonsmooth DC programs. *Mathematics of Operations Research*, 42(1):95–118, 2017.

Phan, D. N. and Gillis, N. An inertial block majorization minimization framework for nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 24: 1–41, 2023.

Pock, T. and Sabach, S. Inertial proximal alternating linearized minimization (ipalm) for nonconvex and nonsmooth problems. *SIAM journal on imaging sciences*, 9 (4):1756–1787, 2016.

Rockafellar, R. T. and Wets., R. J.-B. Variational analysis. *Springer Science & Business Media*, 317, 2009.

Shen, L., Liu, W., Yuan, G., and Ma, S. Gsos: Gauss-seidel operator splitting algorithm for multi-term nonsmooth convex composite optimization. In *International Conference on Machine Learning*, pp. 3125–3134. PMLR, 2017.

Sun, K. and Sun, X. A. Algorithms for difference-of-convex programs based on difference-of-moreau-envelopes smoothing. *INFORMS Journal on Optimization*, 5(4):321–339, 2023.

Suzuki, T. Stochastic dual coordinate ascent with alternating direction method of multipliers. In *International Conference on Machine Learning*, pp. 736–744. PMLR, 2014.

Tran Dinh, Q. Non-ergodic alternating proximal augmented lagrangian algorithms with optimal rates. *Advances in Neural Information Processing Systems*, 31, 2018.

Wang, Y., Yin, W., and Zeng, J. Global convergence of admm in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 78(1):29–63, 2019a.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. Spiderboost and momentum: Faster variance reduction algorithms. *Advances in Neural Information Processing Systems*, 32, 2019b.

Xu, Y., Liu, M., Lin, Q., and Yang, T. Admm without a fixed penalty parameter: Faster convergence with new adaptive penalization. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Xu, Z., Chang, X., Xu, F., and Zhang, H. $l_{1/2}$ regularization: A thresholding representation theory and a fast solver. *IEEE Transactions on Neural Networks and Learning Systems*, 23(7):1013–1027, 2012.

Yang, L., Pong, T. K., and Chen, X. Alternating direction method of multipliers for a class of nonconvex and nonsmooth problems with applications to background/foreground extraction. *SIAM Journal on Imaging Sciences*, 10(1):74–110, 2017.

Yashtini, M. Multi-block nonconvex nonsmooth proximal admm: Convergence and rates under kurdyka–*l*ojasiewicz property. *Journal of Optimization Theory and Applications*, 190(3):966–998, 2021.

Yashtini, M. Convergence and rate analysis of a proximal linearized ADMM for nonconvex nonsmooth optimization. *Journal of Global Optimization*, 84(4):913–939, 2022.

Zhang, J. and Luo, Z.-Q. A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. *SIAM Journal on Optimization*, 30(3): 2272–2302, 2020.

Zhang, R. and Kwok, J. Asynchronous distributed admm for consensus optimization. In *International conference on machine learning*, pp. 1701–1709. PMLR, 2014.

Zhang, T. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107, 2010.

Zhou, D., Xu, P., and Gu, Q. Stochastic nested variance reduction for nonconvex optimization. *The Journal of Machine Learning Research*, 21(1):4130–4192, 2020.

Zhu, D., Zhao, L., and Zhang, S. A first-order primal-dual method for nonconvex constrained optimization based on the augmented lagrangian. *Mathematics of Operations Research*, 2023.

Appendix

The organization of the appendix is as follows:

Appendix A covers notations, technical preliminaries, and relevant lemmas.

Appendix B contains proofs related to Section 2.

Appendix C contains proofs related to Section 3.

Appendix D contains proofs related to Section 4.

Appendix E contains proofs related to Section 5.

Appendix F offers our extension to Stochastic IRPL-ADMM.

Appendix G provides proofs for Stochastic IRPL-ADMM.

Appendix H includes additional experiments for the proposed algorithms.

A. Notations, Technical Preliminaries, and Relevant Lemmas

A.1. Notations

We use the following notations in this paper.

- $[n]$: $\{1, 2, ..., n\}$.
- $\mathbf{x}$: $\mathbf{x} \triangleq \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} = \mathbf{x}_{[n]}$.
- $\mathbf{x}_{[i,j]}$: $\mathbf{x}_{[i,j]} \triangleq \{\mathbf{x}_i, \mathbf{x}_{i+1}, \mathbf{x}_{i+2}, \ldots, \mathbf{x}_j\}$, where $j \geq i$.
- Δ_i^t: $\Delta_i^t \triangleq \mathbf{x}_i^t - \mathbf{x}_i^{t-1}$.
- L_i^t: $\mathsf{L}_i^t = L_i + \beta^t \|\mathbf{A}_i\|_2^2$. Note that the function $G(\mathbf{x}^t, \mathbf{z}^t; \beta^t)$ is L_i^t-smooth.
- σ_1: $\sigma_1 \triangleq \frac{\sigma}{(1-|1-\sigma|)^2} \in \mathbb{R}$, where $\sigma \in (0, 2)$. Refer to Lemma A.5.
- σ_2: $\sigma_2 \triangleq \frac{|1-\sigma|}{\sigma(1-|1-\sigma|)} \in \mathbb{R}$, where $\sigma \in (0, 2)$. Refer to Lemma A.5.
- $\|\mathbf{x}\|$: Euclidean norm: $\|\mathbf{x}\| = \|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.
- $\langle \mathbf{a}, \mathbf{b} \rangle$: Euclidean inner product, i.e., $\langle \mathbf{a}, \mathbf{b} \rangle = \sum_i \mathbf{a}_i \mathbf{b}_i$.
- $\mathbf{x}_i$: the i-th block of the vector $\mathbf{x} \in \mathbb{R}^{(\mathbf{d}_1 + \mathbf{d}_2 + \ldots + \mathbf{d}_n) \times 1}$ with $\mathbf{x}_i \in \mathbb{R}^{\mathbf{d}_i \times 1}$.
- $\underline{\lambda}$: the smallest eigenvalue of the matrix $\mathbf{A}_n \mathbf{A}_n^\mathsf{T}$.
- $\overline{\lambda}$: the largest eigenvalue of the matrix $\mathbf{A}_n \mathbf{A}_n^\mathsf{T}$.
- $\|\mathbf{A}\|$: the spectral norm of the matrix $\mathbf{A}$: the largest singular value of $\mathbf{A}$.
- $\mathbf{A}^\mathsf{T}$: the transpose of the matrix $\mathbf{A}$.
- $\mathbf{I}_r$: $\mathbf{I}_r \in \mathbb{R}^{r \times r}$, Identity matrix; the subscript is omitted sometimes.
- $\partial F(\mathbf{x})$: classical (limiting) Euclidean subdifferential of $F(\mathbf{x})$ at $\mathbf{x}$.
- $(|||\partial\Theta(\mathbf{x}, \mathbf{z}, \mathbf{x}', \mathbf{x}''; \beta, \beta'|||)^2$: $\|\partial\Theta_{\mathbf{x}}(\mathbf{x}, \mathbf{z}, \mathbf{x}', \mathbf{x}''; \beta, \beta'\|_2^2 + \|\partial\Theta_{\mathbf{z}}(\mathbf{x}, \mathbf{z}, \mathbf{x}', \mathbf{x}''; \beta, \beta'\|_2^2 + \|\partial\Theta_{\mathbf{x}'}(\mathbf{x}, \mathbf{z}, \mathbf{x}', \mathbf{x}''; \beta, \beta'\|_2^2 + \|\partial\Theta_{\mathbf{x}''}(\mathbf{x}, \mathbf{z}, \mathbf{x}', \mathbf{x}''; \beta, \beta'\|_2^2$
- $\mathcal{I}_\Xi(\mathbf{x})$: the indicator function of a set Ξ with $\mathcal{I}_\Xi(\mathbf{x}) = 0$ if $\mathbf{x} \in \Xi$ and otherwise $+\infty$.
- $\mathrm{vec}(\mathbf{V})$: $\mathrm{vec}(\mathbf{V}) \in \mathbb{R}^{d' \times r'}$, the vector formed by stacking the column vectors of $\mathbf{V}$.
- $\mathrm{mat}(\mathbf{x})$: $\mathrm{mat}(\mathbf{x}) \in \mathbb{R}^{d' \times r'}$, Convert $\mathbf{x} \in \mathbb{R}^{(d' \cdot r') \times 1}$ into a matrix with $\mathrm{mat}(\mathrm{vec}(\mathbf{V})) = \mathbf{V}$.
- $\mathcal{M}$: Orthogonality constraint set: $\mathcal{M} = \{\mathbf{V} \,|\, \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}\}$.
- $\|\mathbf{y}\|_{[k]}$: the ℓ_1 norm of the k largest (in magnitude) elements of the vector $\mathbf{y}$.
- $\mathrm{dist}^2(\Omega, \Omega')$: the squared distance between two sets with $\mathrm{dist}^2(\Omega, \Omega') \triangleq \inf_{\mathbf{w} \in \Omega, \mathbf{w}' \in \Omega'} \|\mathbf{w} - \mathbf{w}'\|_2^2$.
- b: denotes the mini-batch size of stochastic gradient for **IRPL-ADMM-SPIDER**.

A.2. Technical Preliminaries

We present some tools in non-smooth analysis including Fréchet subdifferential, limiting (Fréchet) subdifferential, and directional derivative (Mordukhovich, 2006; Rockafellar & Wets., 2009; Bertsekas, 2015). For any extended real-valued (not necessarily convex) function $F : \mathbb{R}^n \to (-\infty, +\infty]$, its domain is defined by $\mathrm{dom}(F) \triangleq \{\mathbf{x} \in \mathbb{R}^n : |F(\mathbf{x})| < +\infty\}$. The Fréchet subdifferential of F at $\mathbf{x} \in \mathrm{dom}(F)$, denoted as $\hat{\partial}F(\mathbf{x})$, is defined as $\hat{\partial}F(\mathbf{x}) \triangleq \{\mathbf{v} \in \mathbb{R}^n : \lim_{\mathbf{z}\to\mathbf{x}} \inf_{\mathbf{z}\neq\mathbf{x}} \frac{F(\mathbf{z})-F(\mathbf{x})-\langle \mathbf{v}, \mathbf{z}-\mathbf{x}\rangle}{\|\mathbf{z}-\mathbf{x}\|} \geq 0\}$. The limiting subdifferential of $F(\mathbf{x})$ at $\mathbf{x} \in \mathrm{dom}(F)$ is defined as: $\partial F(\mathbf{x}) \triangleq \{\mathbf{v} \in \mathbb{R}^n : \exists \mathbf{x}^k \to \mathbf{x}, F(\mathbf{x}^k) \to F(\mathbf{x}), \mathbf{v}^k \in \hat{\partial}F(\mathbf{x}^k) \to \mathbf{v}, \forall k\}$. Note that $\hat{\partial}F(\mathbf{x}) \subseteq \partial F(\mathbf{x})$. If $F(\cdot)$ is differentiable at $\mathbf{x}$, then $\hat{\partial}F(\mathbf{x}) = \partial F(\mathbf{x}) = \{\nabla F(\mathbf{x})\}$ with $\nabla F(\mathbf{x})$ being the gradient of $F(\cdot)$ at $\mathbf{x}$. When $F(\cdot)$ is convex, $\hat{\partial}F(\mathbf{x})$ and $\partial F(\mathbf{x})$ reduce to the classical subdifferential for convex functions, i.e., $\hat{\partial}F(\mathbf{x}) = \partial F(\mathbf{x}) = \{\mathbf{v} \in \mathbf{R}^n : F(\mathbf{z}) - F(\mathbf{x}) - \langle \mathbf{v}, \mathbf{z}-\mathbf{x}\rangle \geq 0, \forall \mathbf{z} \in \mathbb{R}^n\}$. The directional derivative of $F(\cdot)$ at $\mathbf{x}$ in the direction $\mathbf{v}$ is defined (if it exists) by $F'(\mathbf{x}; \mathbf{v}) \triangleq \lim_{t\to 0^+} \frac{1}{t}(F(\mathbf{x}+t\mathbf{v}) - F(\mathbf{x}))$.

A.3. Relevant Lemmas

We introduce several useful lemmas that will be utilized in this paper.

Lemma A.1. (Pythagoras Relation) For any vectors $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{c} \in \mathbb{R}^n$, we have:

$$\frac{1}{2}\|\mathbf{a}-\mathbf{b}\|_2^2 - \frac{1}{2}\|\mathbf{c}-\mathbf{b}\|_2^2 = \frac{1}{2}\|\mathbf{a}-\mathbf{c}\|_2^2 + \langle \mathbf{b}-\mathbf{c}, \mathbf{c}-\mathbf{a}\rangle. \tag{25}$$

$$\frac{1}{2}\|\mathbf{b}\|_2^2 - \frac{1}{2}\|\mathbf{c}-\mathbf{b}\|_2^2 = \frac{1}{2}\|\mathbf{c}\|_2^2 + \langle \mathbf{b}-\mathbf{c}, \mathbf{c}\rangle. \tag{26}$$

Lemma A.2. *For any constants $\xi > 0$ and $a \geq 0$, it holds that: $-\frac{1}{\xi} - a\xi \leq -2\sqrt{a}$.*

Proof. We have: $4a \leq 4a + (\frac{1}{\xi} - a\xi)^2 = 4a + \frac{1}{\xi^2} - 2a + \xi^2 a^2 = (\frac{1}{\xi} + a\xi)^2$. Taking the square root of both sides, we conclude this lemma. □

Lemma A.3. *Assume $p \in (0, 1]$. Let $a \geq 0$ and $b \geq 0$. It holds that: $(a+b)^p \leq a^p + b^p$.*

Proof. We define $h(t) = (1+t)^p - t^p$. We have: $\nabla h(t) = p(1-t)^{p-1} - pt^{p-1} < 0$ for all $p \in (0, 1)$ and $t \in (0, \infty)$. Therefore, $h(t)$ is decreasing on $t \in (0, \infty)$. Letting $t = a/b$, we have: $(1 + \frac{a}{b})^p - (\frac{a}{b})^p \leq h(0) = 1$, leading to: $(a+b)^p \leq a^p + b^p$. □

Lemma A.4. *Assume $p \in (0, 2]$. For all $t \geq 0$, it holds that $(t+1)^p - t^p \leq 1 + 2t^{p/2}$.*

Proof. We have: $((t+1)^2)^{p/2} = (t^2 + 1 + 2t)^{p/2} \overset{①}{\leq} (t^2)^{p/2} + (1)^{p/2} + (2t)^{p/2} = t^p + 1 + (2t)^{p/2} \overset{②}{\leq} t^p + 1 + 2(t)^{p/2}$, where step ① uses Lemma A.3 since $p/2 \in (0, 1]$; step ② uses $p/2 \leq 1$. □

Lemma A.5. *For any vectors $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^n$, and any constant $\theta > 0$, we have: $\langle \mathbf{a}, \mathbf{b}\rangle \leq \frac{\theta}{2}\|\mathbf{a}\|_2^2 + \frac{1}{2\theta}\|\mathbf{b}\|_2^2$, and $\|\mathbf{a}+\mathbf{b}\|_2^2 \leq (1+\theta)\|\mathbf{a}\|_2^2 + (1+\frac{1}{\theta})\|\mathbf{b}\|_2^2$.*

Proof. **(a)** For any $\theta > 0$, we have: $0 \leq \frac{1}{2}\|\sqrt{\theta}\mathbf{a} - \frac{1}{\sqrt{\theta}}\mathbf{b}\|_2^2 = \frac{\theta}{2}\|\mathbf{a}\|_2^2 - \langle \mathbf{a}, \mathbf{b}\rangle + \frac{1}{2\theta}\|\mathbf{b}\|_2^2$. **(b)** For any $\theta > 0$, we have: $\|\mathbf{a}+\mathbf{b}\|_2^2 = \|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2 + 2\langle \mathbf{a}, \mathbf{b}\rangle \leq (1+\theta)\|\mathbf{a}\|_2^2 + (1+\frac{1}{\theta})\|\mathbf{b}\|_2^2$, where the inequality uses **Part (a)** of this lemma.

□

Lemma A.6. *Assume $\sigma \in (0, 2)$. Let $\mathbf{c} = \sigma\mathbf{a} + (1-\sigma)\mathbf{b}$, where $\mathbf{c} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^n$, and $\mathbf{a} \in \mathbb{R}^n$. We have:*

$$\frac{1}{\sigma}\|\mathbf{c}\|_2^2 \leq \sigma_1\|\mathbf{a}\|_2^2 + \sigma_2(\|\mathbf{b}\|_2^2 - \|\mathbf{c}\|_2^2),$$

where $\sigma_1 \triangleq \frac{\sigma}{(1-|1-\sigma|)^2}$, and $\sigma_2 \triangleq \frac{|1-\sigma|}{\sigma(1-|1-\sigma|)}$.

Proof. **(a)** When $\sigma = 1$, we have $\sigma_1 = 1$, $\sigma_2 = 0$, and $\mathbf{c} = \mathbf{a}$. The conclusion of this lemma clearly holds.

(b) We now focus on the case when $\sigma \neq 1$. Noticing $|1-\sigma| \neq 0$ and $1-|1-\sigma| \neq 0$, we rewrite $\mathbf{c} = (1-\sigma)\mathbf{b} + \sigma\mathbf{a}$ into the following equivalent equality

$$\mathbf{c} = (1-|1-\sigma|) \cdot \tfrac{\sigma\mathbf{a}}{1-|1-\sigma|} + |1-\sigma| \cdot \tfrac{(1-\sigma)\mathbf{b}}{|1-\sigma|}.$$

Using the fact that the function $\|\cdot\|_2^2$ is convex and $|1-\sigma| \in (0,1)$, we derive the following results:

$$\begin{aligned} \|\mathbf{c}\|_2^2 &\leq (1-|1-\sigma|) \cdot \|\tfrac{\sigma\mathbf{a}}{1-|1-\sigma|}\|_2^2 + |1-\sigma| \cdot \|\tfrac{(1-\sigma)\mathbf{b}}{|1-\sigma|}\|_2^2 \\ &\leq \tfrac{\sigma^2}{1-|1-\sigma|} \cdot \|\mathbf{a}\|_2^2 + |1-\sigma| \cdot \|\mathbf{b}\|_2^2. \end{aligned}$$

Subtracting $(|1-\sigma| \cdot \|\mathbf{c}\|_2^2)$ from both sides of the above inequality, we have:

$$(1-|1-\sigma|)\|\mathbf{c}\|_2^2 \leq \tfrac{\sigma^2}{1-|1-\sigma|} \cdot \|\mathbf{a}\|_2^2 + |1-\sigma|(\|\mathbf{b}\|_2^2 - \|\mathbf{c}\|_2^2).$$

Dividing both sides by $\sigma(1-|1-\sigma|)$, we have:

$$\tfrac{1}{\sigma}\|\mathbf{c}\|_2^2 \leq \tfrac{\sigma}{(1-|1-\sigma|)^2}\|\mathbf{a}\|_2^2 + \tfrac{|1-\sigma|}{\sigma(1-|1-\sigma|)}(\|\mathbf{b}\|_2^2 - \|\mathbf{c}\|_2^2).$$

Using the definition of σ_1 and σ_2, we finish the proof of this lemma.

□

Lemma A.7. *For any positive constants c_1 and c_2, and nonnegative sequences $\{\varphi_1^t, \varphi_2^t\}_{t=1}^T$, we have:*

$$\min_{t=1}^T(c_1\varphi_1^t, c_2\varphi_2^t) \geq \tfrac{1}{c_0} \cdot \min_{t=1}^T(\varphi_1^t, \varphi_2^t),$$

where $c_0 \triangleq \max(\frac{1}{c_1}, \frac{1}{c_2})$, and $\min_{t=1}^T(\varphi_1^t, \varphi_2^t) \triangleq \min_{t=1}^T[\min(\varphi_1^t, \varphi_2^t)]$.

Proof. We have the following inequalities:

$$\begin{aligned} \min_{t=1}^T\{\varphi_1^t, \varphi_2^t\} &\leq \min_{t=1}^T\{\max(\tfrac{c_1}{c_1}\varphi_1^t, \tfrac{c_1}{c_2}\varphi_1^t), \max(\tfrac{c_2}{c_1}\varphi_2^t, \tfrac{c_2}{c_2}\varphi_2^t)\} \\ &= \min_{t=1}^T\{c_0c_1\varphi_1^t, c_0c_2\varphi_2^t\} \overset{①}{=} c_0 \min_{t=1}^T(c_1\varphi_1^t, c_2\varphi_2^t), \end{aligned}$$

where step ① uses the nonnegativity of both the constants $\{c_0, c_1, c_2\}$ and the sequences $\{\varphi_1^t, \varphi_2^t\}_{t=1}^T$.

□

Lemma A.8. *For any three nonnegative sequences $\{e^t, w^t, p^t\}_{t=0}^\infty$ satisfying $e^{t+1} \leq \sqrt{w^t(e^t - e^{t-1}) + p^t}$, we have:*

$$\textstyle\sum_{t=0}^T e^t \leq \frac{3}{2}e^0 + \frac{1}{2}e^{-1} + \frac{1}{2}\sum_{t=0}^T(w^t + p^t). \tag{27}$$

Proof. For any $\alpha_1 > 0$ and $\alpha_2 > 0$, we derive the following inequalities:

$$\begin{aligned} e^{t+1} &\leq \sqrt{w^t(e^t - e^{t-1}) + p^t} \\ &\overset{①}{\leq} \sqrt{\alpha_1(e^t)^2 + \tfrac{1}{4\alpha_1}(w^t)^2 + \alpha_2(e^{t-1})^2 + \tfrac{1}{4\alpha_2}(w^t)^2 + p^t} \\ &\overset{②}{\leq} \sqrt{\alpha_1}e^t + \tfrac{w^t}{2\sqrt{\alpha_1}} + \sqrt{\alpha_2}e^{t-1} + \tfrac{w^t}{2\sqrt{\alpha_2}} + p^t, \end{aligned}$$

where step ① uses the fact that $ab \leq \xi a^2 + \frac{1}{4\xi}b^2$ for all $\xi > 0$; step ② uses the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. We further obtain:

$$e^{t+1} + \sqrt{\alpha_2}e^t \leq (\sqrt{\alpha_1} + \sqrt{\alpha_2})e^t + \sqrt{\alpha_2}e^{t-1} + (\tfrac{1}{2\sqrt{\alpha_1}} + \tfrac{1}{2\sqrt{\alpha_2}}) \cdot w^t + p^t.$$

Telescoping the inequality above over t from 0 to T, we obtain:

$$\begin{aligned} &\{e^{T+1} + \sqrt{\alpha_2}e^T\} - \{(\sqrt{\alpha_1} + \sqrt{\alpha_2})e^0 + \sqrt{\alpha_2}e^{-1}\} + \textstyle\sum_{t=1}^T\{(\sqrt{\alpha_1} + \sqrt{\alpha_2} - 1)e^t + \sqrt{\alpha_2}e^{t-1}\} \\ \leq\ &\textstyle\sum_{t=0}^T\{(\frac{1}{2\sqrt{\alpha_1}} + \frac{1}{2\sqrt{\alpha_2}}) \cdot w^t + p^t\}. \end{aligned}$$

Letting $\sqrt{\alpha_1}+\sqrt{\alpha_2}-1=1$, $\alpha_1=1$, and $\alpha_2=1$, we have:

$$\textstyle\sum_{t=1}^{T} e^t + e^{t-1} \leq -e^{T+1} - e^T + 2e^0 + e^{-1} + \sum_{t=0}^{T}(w^t + p^t)$$

Adding $e^0 + e^T$ to both sides yields:

$$\begin{aligned} 2\textstyle\sum_{t=0}^{T} e^t &\leq e^0 + e^T - e^{T+1} - e^T + 2e^0 + e^{-1} + \sum_{t=0}^{T}(w^t + p^t) \\ &\leq 3e^0 + e^{-1} + \textstyle\sum_{t=0}^{T}(w^t + p^t). \end{aligned}$$

We further obtain:

$$\textstyle\sum_{t=0}^{T} e^t \leq \frac{3}{2}e^0 + \frac{1}{2}e^{-1} + \frac{1}{2}\sum_{t=0}^{T}(w^t + p^t).$$

□

B. Proofs for Section 2

B.1. Proof of Lemma 2.2

Proof. Let $p \in (0, 1]$. Consider the update rule $\beta^t = \beta^0 + \vartheta t^p$ for all $t \geq 0$, where $\vartheta \leq \beta^0 \xi$.

(a) The lower bound for β^{t+1} in Inequality (2), is clearly satisfied with $\vartheta' = \vartheta$.

(b) We now focus on establishing the upper bound for β^{t+1}. We have:

$$\beta^{t+1} - \beta^t - \xi\beta^t \overset{①}{=} \vartheta((t+1)^p - t^p) - \xi\beta^0 \overset{②}{\leq} \vartheta - \beta^0\xi \overset{③}{\leq} 0,$$

where step ① uses the update rule $\beta^t = \beta^0 + \vartheta t^p$; step ② uses the fact that the function $h(t) \triangleq (t+1)^p - t^p$ is monotonically decreasing *w.r.t.* t that: $h(t) \leq h(0) = 1$; step ③ uses $\vartheta \leq \beta^0\xi$.

Therefore, this update rule aligns with Definition 2.1.

□

B.2. Proof of Lemma 2.3

Proof. We let $p \in (1, 2]$. Consider the update rule $\beta^{t+1} = \beta^0 + \vartheta(t+1)^p$ for all $t \geq 0$, where $\vartheta \leq \frac{\beta^0\xi^2}{1+\xi}$.

(a) The lower bound for β^{t+1} in Inequality (2), is clearly satisfied with $\vartheta' = \vartheta$.

(b) We now focus on establishing the upper bound for β^{t+1}. We derive:

$$\begin{aligned} \tfrac{1}{\vartheta}(\beta^{t+1} - \beta^t - \xi\beta^t) &\overset{①}{=} (t+1)^p - t^p - (\tfrac{\xi}{\vartheta}\cdot\beta^0 + \tfrac{\xi}{\vartheta}\cdot\vartheta t^p) \\ &\overset{②}{\leq} (t+1)^p - t^p - 1 - \tfrac{1}{\xi} - \xi t^p \\ &\overset{③}{\leq} (t+1)^p - t^p - 1 - 2t^{p/2} \\ &\leq 0, \end{aligned}$$

where step ① uses the update rule $\beta^{t+1} = \beta^0 + \vartheta(t+1)^2$; step ② uses $\frac{\beta^0\xi}{\vartheta} \geq \frac{1+\xi}{\xi}$, $\vartheta > 0$, and $\xi > 0$; step ③ uses the fact $-\frac{1}{\xi} - a\xi \leq -2\sqrt{a}$ for all $\xi > 0$ and $a \geq 0$, which is due to Lemma A.2; step ④ uses Lemma A.4 since $p \in (1, 2] \in (0, 2]$.

Therefore, this update rule aligns with Definition 2.1.

□

B.3. Proof of Lemma 2.4

Proof. Let $p \in (0, 1)$. We define $\tilde{e}^t \triangleq \|\sum_{i=1}^{n} \mathbf{A}_i\mathbf{x}_i^{t+1} + \mathbf{b}\|$. We let $h(t) \triangleq (t+1)^p - t^p$.

Consider the update rule $\beta^{t+1} = \beta^t + \min(\tilde{e}^t + \vartheta h(t), \beta^t \xi)$ for all $t \geq 0$, where $\vartheta \leq \beta^0 \xi$.

(a) The upper bound of β^{t+1} in Inequality (2), is evidently met, given that $\beta^{t+1} = \beta^t + \min(\beta^t \xi, h(t) + \tilde{e}^t) \leq \beta^t + \beta^t \xi$.

(b) We now focus on the lower bound for β^{t+1}. Using the penalty parameter update rule, we have:

$$\begin{aligned}
\beta^{t+1} &\geq \beta^t + \min\{\xi\beta^t, \vartheta h(t) + \tilde{e}^t\} \\
&\overset{①}{\geq} \beta^t + \min\{\xi\beta^t, \vartheta h(t)\} \\
&\overset{②}{\geq} \beta^t + \min\{\xi\beta^0 h(t), \vartheta h(t)\} \\
&\overset{③}{\geq} \beta^t + \min\{\vartheta h(t), \vartheta h(t)\},
\end{aligned}$$

where step ① uses the fact that $\min(a, b + \tilde{e}^t) \geq \min(a, b)$ for all $a \in \mathbb{R}$ and $b \in \mathbb{R}$ since $\tilde{e}^t \geq 0$; step ② uses the fact that the sequence $\{\beta^t\}_{t=0}^{\infty}$ is monotonically increasing, and the function $h(t)$ is monotonically decreasing *w.r.t.* t that $1 = h(0) \geq h(t)$. Therefore, for all $t \geq 0$, we have:

$$\beta^{t+1} \geq \beta^t + \vartheta(t+1)^p - \vartheta t^p$$

Telescoping this inequality over t from 0 to T, we have:

$$\beta^{T+1} - \beta^0 \geq \vartheta(T+1)^p - \vartheta 0^p = \vartheta(T+1)^p$$

Hence, we establish the lower bound for β^{t+1} that: $\beta^{t+1} - \beta^0 \geq \vartheta(t+1)^p$.

Therefore, this update rule aligns with Definition 2.1.

□

B.4. Proof of Lemma 2.5

Proof. Let $p \in (1, 2]$. We define $\tilde{e}^t \triangleq \| \sum_{i=1}^n \mathbf{A}_i \mathbf{x}_i^{t+1} + \mathbf{b} \|$. We let $h(t) \triangleq (t+1)^p - t^p$.

Consider the update rule $\beta^{t+1} = \beta^t + \min(\tilde{e}^t + \vartheta h(t), \beta^t \xi)$ for all $t \geq 0$, where $\vartheta \leq \beta^0 \xi^2/(1+\xi)$.

Initially, we have the following results:

$$(t+1)^p - t^p - \xi t^p - \tfrac{1+\xi}{\xi} \overset{①}{\leq} 1 + 2t^{p/2} - \xi t^p - 1 - \tfrac{1}{\xi} = 2t^{p/2} - \xi t^p - \tfrac{1}{\xi} \overset{②}{\leq} 2t^{p/2} - 2\sqrt{t^p} = 0,$$

where step ① uses Lemma A.4 given $p \in (1, 2] \in (0, 2]$; step ② uses the inequality $-\frac{1}{\xi} - a\xi \leq -2\sqrt{a}$ for all $a \geq 0$ and $\xi > 0$, which is due to Lemma A.2. Therefore, for all $t \geq 0$, we have:

$$h(t) - \xi t^p \leq \tfrac{1+\xi}{\xi}. \tag{28}$$

(a) The upper bound of β^{t+1} in Inequality (2), is evidently met, given that $\beta^{t+1} = \beta^t + \min(\beta^t \xi, \vartheta h(t) + \tilde{e}^t) \leq \beta^t + \beta^t \xi$.

(b) We now focus on establishing the lower bound for β^{t+1}. Using the penalty parameter update rule, we have:

$$\begin{aligned}
\beta^{t+1} &\geq \beta^t + \min\{\xi\beta^t, \vartheta h(t) + \tilde{e}^t\} \\
&\overset{①}{\geq} \beta^t + \min\{\xi\beta^t, \vartheta h(t)\}, \\
&\overset{②}{\geq} \beta^t + \vartheta h(t),
\end{aligned} \tag{29}$$

where step ① uses the fact that $\min(a, b + \tilde{e}^t) \geq \min(a, b)$ for all $a \in \mathbb{R}$ and $b \in \mathbb{R}$ since $\tilde{e}^t \geq 0$; and step ② uses the following inequality:

$$\vartheta h(t) - \beta^t \xi \leq 0. \tag{30}$$

In what follows, we prove that Inequality (30) always holds if $\vartheta \leq \frac{\beta^0 \xi^2}{1+\xi}$. This can be achieved by iteratively lower bounding the parameter β^{t+1}.

Case (i). Consider $t = 0$. We derive: $\vartheta h(t) - \beta^t\xi \overset{①}{=} \vartheta \cdot h(0) - \beta^0\xi \overset{②}{\leq} \beta^0\xi \cdot \frac{\xi}{1+\xi} \cdot \frac{1+\xi}{\xi} - \beta^0\xi \overset{②}{=} 0$, where step ① uses $t = 0$; step ② uses uses $\vartheta \leq \frac{\beta^0\xi^2}{1+\xi}$, and the fact that $h(0) - \xi 0^p \leq \frac{1+\xi}{\xi}$ which is implied by Inequality (28). Hence, Inequality (30) holds for $t = 0$. Additionally, we obtain from Inequality (29):

$$\beta^1 \geq \beta^0 + \vartheta h(0) = \beta^0 + \vartheta 1^p. \tag{31}$$

Case (ii). Consider $t = 1$. We derive: $\vartheta h(t) - \beta^t\xi \overset{①}{=} \vartheta h(1) - \beta^1\xi \overset{②}{\leq} \vartheta \cdot \{h(1) - 1^p\xi\} - \beta^0\xi \overset{③}{\leq} \frac{\beta^0\xi^2}{1+\xi} \cdot \frac{1+\xi}{\xi} - \beta^0\xi = 0$, where step ① uses $t = 1$; step ② uses Inequality (31); step ③ uses $\vartheta \leq \frac{\beta^0\xi^2}{1+\xi}$ and Inequality (28). Hence, Inequality (30) holds for $t = 1$. Additionally, we obtain from Inequality (29):

$$\beta^2 \geq \beta^1 + \vartheta h(1) \geq \beta^0 + \vartheta h(0) + \vartheta h(1) = \beta^0 + \vartheta 2^p. \tag{32}$$

Case (iii). Consider $t = 2$, we derive: $\vartheta h(t) - \beta^t\xi \overset{①}{=} \vartheta h(2) - \beta^2\xi \leq \vartheta \cdot \{h(2) - \xi t^p\} - \beta^0\xi \overset{③}{\leq} \frac{\beta^0\xi^2}{1+\xi} \cdot \frac{1+\xi}{\xi} - \beta^0\xi = 0$, where step ① uses $t = 2$; step ② uses Inequality (32); and step ③ uses $\vartheta \leq \frac{\beta^0\xi^2}{1+\xi}$ and Inequality (28). Hence, Inequality (30) holds for $t = 2$. Additionally, we obtain from Inequality (29):

$$\beta^3 \geq \beta^2 + \vartheta h(2) \geq \beta^0 + \vartheta 2^p + \vartheta h(2) = \beta^0 + \vartheta 3^p.$$

Using similar strategies, we can recursively conclude that Inequality (30) and Inequality (29) hold for $t \geq 0$.

Telescoping Inequality (29) over t from 0 to T, we have:

$$\beta^{T+1} - \beta^0 \geq \vartheta(T+1)^p - \vartheta 0^p = \vartheta(T+1)^p.$$

Hence, we establish the lower bound for β^{t+1} that: $\beta^{t+1} - \beta^0 \geq \vartheta(t+1)^p$.

Therefore, this update rule aligns with Definition 2.1.

□

B.5. Proof of Lemma 2.6

Proof. We use a direct approach involving integral comparison. By comparing the series $\sum_{t=1}^{\infty} \frac{1}{t^p}$ with the integral $\int_1^{\infty} \frac{1}{t^p} dt$, where the function $f(t) = \frac{1}{t^p}$ is decreasing and positive for $t \geq 1$ and $p > 1$, we have the following inequality: $\sum_{t=2}^{\infty} \frac{1}{t^p} \leq \int_1^{\infty} \frac{1}{t^p} dt$, leading to the following results:

$$\sum_{t=1}^{\infty} \frac{1}{t^p} = 1 + \sum_{t=2}^{\infty} \frac{1}{t^p} \overset{①}{\leq} 1 + \frac{1}{p-1} = \frac{p}{p-1}, \tag{33}$$

where step ① uses the fact that $\int_1^{\infty} \frac{1}{t^p} dt = \frac{1}{p-1}$. Notably, Inequality (33) is a well-established result for convergent p-series with $p > 1$. Finally, we obtain:

$$\sum_{t=0}^{\infty} \frac{1}{\beta^t} \overset{①}{=} \frac{1}{\beta^0} + \sum_{t=1}^{\infty} \frac{1}{\beta^0 + \vartheta t^2} \leq \frac{1}{\beta^0} + \sum_{t=1}^{\infty} \frac{1}{\vartheta t^p} \overset{②}{\leq} \frac{1}{\beta^0} + \frac{p}{(p-1)\vartheta},$$

where step ① uses $\beta^t \geq \beta^0 + \vartheta t^2$; step ② uses Inequality (33) for all $t \geq 1$.

□

C. Proofs for Section 3

C.1. Proof of Lemma 3.1

Proof. **(a)** We now establish the decrease in the objective function value for the subproblem of the i-th block with $i \in [n]$. First, noticing the function $G(\mathbf{x}_{[1,i-1]}^{t+1}, \mathbf{x}_i, \mathbf{x}_{[i+1,n]}^t, \mathbf{z}^t; \beta^t)$ is L_i^t-smooth *w.r.t.* $\mathbf{x}_i$ for the t-th iteration, we have:

$$\begin{aligned} & G(\mathbf{x}_{[1,i-1]}^{t+1}, \mathbf{x}_i^{t+1}, \mathbf{x}_{[i+1,n]}^t, \mathbf{z}^t; \beta^t) \\ \leq\ & G(\mathbf{x}_{[1,i-1]}^{t+1}, \mathbf{x}_i^t, \mathbf{x}_{[i+1,n]}^t, \mathbf{z}^t; \beta^t) + \langle \mathbf{x}_i^{t+1} - \mathbf{x}_i^t, \nabla_{\mathbf{x}_i} G(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_i^t, \mathbf{x}_{[i+1,n]}^t, \mathbf{z}^t; \beta^t) \rangle + \tfrac{\mathsf{L}_i^t}{2} \|\mathbf{x}_i^{t+1} - \mathbf{x}_i^t\|_2^2. \end{aligned} \tag{34}$$

Second, we notice that $\mathbf{x}_i^{t+1}$ is the minimizer of the following optimization problem:

$$\mathbf{x}_i^{t+1} \in \arg\min_{\mathbf{x}_i}\ h_i(\mathbf{x}_i) + \langle \mathbf{x}_i - \mathbf{y}_i^t, \nabla_{\mathbf{x}_i} G(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_i^t, \mathbf{x}_{[i+1,n]}^t, \mathbf{z}^t; \beta^t)\rangle + \tfrac{\boldsymbol{\theta}_i \mathsf{L}_i^t}{2}\|\mathbf{x}_i - \mathbf{y}_i^t\|_2^2. \tag{35}$$

Using the optimality of $\mathbf{x}_i^{t+1}$ in (35), we have:

$$\begin{aligned} & h_i(\mathbf{x}_i^{t+1}) - h_i(\mathbf{x}_i^t) + \langle \mathbf{x}_i^{t+1} - \mathbf{x}_i^t, \nabla_{\mathbf{x}_i} G(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_i^t, \mathbf{x}_{[i+1,n]}^t, \mathbf{z}^t; \beta^t)\rangle \\ \leq\ & \tfrac{\boldsymbol{\theta}_i \mathsf{L}_i^t}{2}\|\mathbf{x}_i^t - \mathbf{y}_i^t\|_2^2 - \tfrac{\boldsymbol{\theta}_i \mathsf{L}_i^t}{2}\|\mathbf{x}_i^{t+1} - \mathbf{y}_i^t\|_2^2 \\ \overset{①}{=}\ & -\tfrac{\boldsymbol{\theta}_i \mathsf{L}_i^t}{2}\|\mathbf{x}_i^{t+1} - \mathbf{x}_i^t\|_2^2 - \boldsymbol{\theta}_i \mathsf{L}_i^t \langle \mathbf{x}_i^{t+1} - \mathbf{x}_i^t, \mathbf{x}_i^t - \mathbf{y}_i^t\rangle, \end{aligned} \tag{36}$$

where step ① uses the Pythagoras relation as presented in Lemma A.1. Combining equations (34) and (36), we derive the following expressions:

$$\begin{aligned} & h_i(\mathbf{x}_i^{t+1}) + G(\mathbf{x}_{[1,i-1]}^{t+1}, \mathbf{x}_i^{t+1}, \mathbf{x}_{[i+1,n]}^t, \mathbf{z}^t; \beta^t) - h_i(\mathbf{x}_i^t) - G(\mathbf{x}_{[1,i-1]}^{t+1}, \mathbf{x}_i^t, \mathbf{x}_{[i+1,n]}^t, \mathbf{z}^t; \beta^t) \\ \leq\ & -\tfrac{(\boldsymbol{\theta}_i - 1)L_i^t}{2}\|\mathbf{x}_i^{t+1} - \mathbf{x}_i^t\|_2^2 - \boldsymbol{\theta}_i \mathsf{L}_i^t \langle \mathbf{x}_i^{t+1} - \mathbf{x}_i^t, \mathbf{x}_i^t - \mathbf{y}_i^t\rangle \\ \overset{①}{=}\ & -\tfrac{(\boldsymbol{\theta}_i - 1)\mathsf{L}_i^t}{2}\|\mathbf{x}_i^{t+1} - \mathbf{x}_i^t\|_2^2 + \boldsymbol{\theta}_i \mathsf{L}_i^t \|\mathbf{x}_i^{t+1} - \mathbf{x}_i^t\| \|\boldsymbol{\alpha}_i(\mathbf{x}_i^t - \mathbf{x}_i^{t-1})\| \\ \overset{②}{\leq}\ & -\tfrac{(\boldsymbol{\theta}_i - 1)\mathsf{L}_i^t}{2}\|\mathbf{x}_i^{t+1} - \mathbf{x}_i^t\|_2^2 + \boldsymbol{\theta}_i \boldsymbol{\alpha}_i \mathsf{L}_i^t \left(\tfrac{1}{2}\|\mathbf{x}_i^{t+1} - \mathbf{x}_i^t\|_2^2 + \tfrac{1}{2}\|\mathbf{x}_i^t - \mathbf{x}_i^{t-1}\|_2^2\right) \\ \overset{③}{=}\ & -(\boldsymbol{\theta}_i - 1 - \boldsymbol{\theta}_i\boldsymbol{\alpha}_i)\cdot \tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^{t+1}\|_2^2 + \boldsymbol{\theta}_i\boldsymbol{\alpha}_i \cdot (\tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^t\|_2^2 - \tfrac{\mathsf{L}_i^{t+1}}{2}\|\Delta_i^{t+1}\|_2^2) + \boldsymbol{\theta}_i\boldsymbol{\alpha}_i \cdot \tfrac{\mathsf{L}_i^{t+1}}{2}\|\Delta_i^{t+1}\|_2^2 \\ \overset{④}{\leq}\ & -(\boldsymbol{\theta}_i - 1 - \boldsymbol{\theta}_i\boldsymbol{\alpha}_i)\cdot \tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^{t+1}\|_2^2 + \boldsymbol{\theta}_i\boldsymbol{\alpha}_i \cdot (\tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^t\|_2^2 - \tfrac{\mathsf{L}_i^{t+1}}{2}\|\Delta_i^{t+1}\|_2^2) + \boldsymbol{\theta}_i\boldsymbol{\alpha}_i \cdot \tfrac{(1+\xi)\mathsf{L}_i^t}{2}\|\Delta_i^{t+1}\|_2^2 \\ =\ & -(\boldsymbol{\theta}_i - 1 - \boldsymbol{\theta}_i\boldsymbol{\alpha}_i(2+\xi))\cdot \tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^{t+1}\|_2^2 + \boldsymbol{\theta}_i\boldsymbol{\alpha}_i \cdot (\tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^t\|_2^2 - \tfrac{\mathsf{L}_i^{t+1}}{2}\|\Delta_i^{t+1}\|_2^2) \end{aligned} \tag{37}$$

where step ① uses the Cauchy-Schwarz inequality and the update rule for $\mathbf{y}_i^{t+1}$ that $\mathbf{y}_i^{t+1} = \mathbf{x}_i^{t+1} + \boldsymbol{\alpha}_i(\mathbf{x}_i^{t+1} - \mathbf{x}_i^t)$ for all $i \in [n]$; step ② uses the inequality that $ab \leq \frac{a^2}{2} + \frac{b^2}{2}, \forall a, b$; step ⑤ uses the definition $\Delta_i^t \triangleq \mathbf{x}_i^t - \mathbf{x}_i^{t-1}$; step ⑥ uses $\mathsf{L}_i^{t+1} \leq (1+\xi)\mathsf{L}_i^t$, which is implied by $\beta^{t+1} \leq \beta^t(1+\xi)$ and $\mathsf{L}_i^t \triangleq L_i + \beta^t\|\mathbf{A}_i\|_2^2$.

(b) We now establish the decrease for $\mathcal{L}(\mathbf{x}^{t+1}, \mathbf{z}^t; \beta^t) - \mathcal{L}(\mathbf{x}^t, \mathbf{z}^t; \beta^t)$. In view of Inequality (37), we define

$$\Lambda_i^t \triangleq -(\boldsymbol{\theta}_i - 1 - \boldsymbol{\theta}_i\boldsymbol{\alpha}_i(2+\xi))\cdot \tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^{t+1}\|_2^2 + \boldsymbol{\theta}_i\boldsymbol{\alpha}_i \cdot (\tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^t\|_2^2 - \tfrac{\mathsf{L}_i^{t+1}}{2}\|\Delta_i^{t+1}\|_2^2) - h_i(\mathbf{x}_i^{t+1}) + h_i(\mathbf{x}_i^t). \tag{38}$$

We have the following inequalities:

$$\begin{aligned} & i = 1, \quad G(\mathbf{x}_{[1,1-1]}^{t+1}, \mathbf{x}_1^{t+1}, \mathbf{x}_{[1+1,n]}^t, \mathbf{z}^t; \beta^t) - G(\mathbf{x}_{[1,1-1]}^{t+1}, \mathbf{x}_1^t, \mathbf{x}_{[1+1,n]}^t, \mathbf{z}^t; \beta^t) \leq \Lambda_1^t \\ & i = 2, \quad G(\mathbf{x}_{[1,2-1]}^{t+1}, \mathbf{x}_2^{t+1}, \mathbf{x}_{[2+1,n]}^t, \mathbf{z}^t; \beta^t) - G(\mathbf{x}_{[1,2-1]}^{t+1}, \mathbf{x}_2^t, \mathbf{x}_{[2+1,n]}^t, \mathbf{z}^t; \beta^t) \leq \Lambda_2^t \\ & \quad \cdots \\ & i = n, \quad G(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_n^{t+1}, \mathbf{x}_{[n+1,n]}^t, \mathbf{z}^t; \beta^t) - G(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_n^t, \mathbf{x}_{[n+1,n]}^t, \mathbf{z}^t; \beta^t) \leq \Lambda_n^t. \end{aligned}$$

Summing up all these inequalities together, we have:

$$G(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_n^{t+1}, \mathbf{x}_{[n+1,n]}^t, \mathbf{z}^t; \beta^t) - G(\mathbf{x}_{[1,1-1]}^{t+1}, \mathbf{x}_1^t, \mathbf{x}_{[1+1,n]}^t, \mathbf{z}^t; \beta^t) \leq \textstyle\sum_{i=1}^n \Lambda_i^t.$$

We further obtain the following inequality:

$$G(\mathbf{x}^{t+1}, \mathbf{z}^t; \beta^t) - G(\mathbf{x}^t, \mathbf{z}^t; \beta^t) \leq \textstyle\sum_{i=1}^n \Lambda_i^t. \tag{39}$$

Uses the definition of $\mathcal{L}(\mathbf{x}, \mathbf{z}; \beta) \triangleq G(\mathbf{x}, \mathbf{z}; \beta) + \sum_{i=1}^n h_i(\mathbf{x}_i)$, we derive the following results:

$$\begin{aligned} & \mathcal{L}(\mathbf{x}^{t+1}, \mathbf{z}^t; \beta^t) - \mathcal{L}(\mathbf{x}^t, \mathbf{z}^t; \beta^t) \\ =\ & \{G(\mathbf{x}^{t+1}, \mathbf{z}^t; \beta^t) + \textstyle\sum_{i=1}^n h_i(\mathbf{x}_i^{t+1})\} - \{G(\mathbf{x}^t, \mathbf{z}^t; \beta^t) + \textstyle\sum_{i=1}^n h_i(\mathbf{x}_i^t)\} \\ \overset{①}{\leq}\ & \textstyle\sum_{i=1}^n \{\Lambda_i^t + h_i(\mathbf{x}_i^{t+1}) - h_i(\mathbf{x}_i^t)\} \\ \overset{②}{=}\ & \textstyle\sum_{i=1}^n \{-(\boldsymbol{\theta}_i - 1 - \boldsymbol{\theta}_i\boldsymbol{\alpha}_i(2+\xi))\cdot \tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^{t+1}\|_2^2 + \tfrac{1}{2}\boldsymbol{\theta}_i\boldsymbol{\alpha}_i(\mathsf{L}_i^t\|\Delta_i^t\|_2^2 - \mathsf{L}_i^{t+1}\|\Delta_i^{t+1}\|_2^2)\} \end{aligned} \tag{40}$$

where step ① uses Inequality (39); and step ② uses the definition of Λ_i^t in Equation (38).

(c) For notation convenience, we define:

$$\begin{aligned} \mathbf{r}^{t+1} &\triangleq [\textstyle\sum_{i=1}^n \mathbf{A}_i\mathbf{x}_i^{t+1}] - \mathbf{b} \\ \Theta_{\circ}^t &\triangleq \mathcal{L}(\mathbf{x}^t, \mathbf{z}^t; \beta^t) + \tfrac{1}{2}\textstyle\sum_{i=1}^n \boldsymbol{\theta}_i\boldsymbol{\alpha}_i\mathsf{L}_i^t\|\Delta_i^t\|_2^2 \end{aligned} \tag{41}$$

We have the following inequalities:

$$\begin{aligned} & \tfrac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2 + \mathcal{L}(\mathbf{x}^{t+1}, \mathbf{z}^{t+1}; \beta^{t+1}) - \mathcal{L}(\mathbf{x}^{t+1}, \mathbf{z}^t; \beta^t) \\ \overset{①}{=} & \langle \mathbf{z}^{t+1} - \mathbf{z}^t, \mathbf{r}^{t+1}\rangle + (\tfrac{\xi\beta^t}{2} + \tfrac{\beta^{t+1}-\beta^t}{2})\|\mathbf{r}^{t+1}\|_2^2 \\ \overset{②}{=} & \langle \mathbf{z}^{t+1} - \mathbf{z}^t, (\mathbf{z}^{t+1} - \mathbf{z}^t)\cdot\tfrac{1}{\sigma\beta^t}\rangle + (\tfrac{\xi\beta^t}{2} + \tfrac{\beta^{t+1}-\beta^t}{2})\|\tfrac{1}{\sigma\beta^t}(\mathbf{z}^{t+1} - \mathbf{z}^t)\|_2^2 \\ = & \|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 \cdot \tfrac{1}{\sigma\beta^t}\cdot\{1 + \tfrac{\xi+\beta^{t+1}/\beta^t - 1}{2\sigma}\} \\ \overset{③}{\leq} & \|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 \cdot \tfrac{1}{\sigma\beta^t}\cdot\{1 + \tfrac{\xi+\xi}{2\sigma}\} \\ \overset{④}{\leq} & \|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 \cdot \tfrac{1}{\sigma\beta^t}\cdot\{1 + \epsilon_2\}, \end{aligned} \tag{42}$$

where step ① uses the definition of $\mathcal{L}(\mathbf{x}^{t+1}, \mathbf{z}^{t+1}; \beta^{t+1})$ and $\mathcal{L}(\mathbf{x}^{t+1}, \mathbf{z}^t; \beta^t)$; step ② uses the update rule of $\mathbf{z}^{t+1}$ that: $\frac{1}{\sigma\beta^t}(\mathbf{z}^{t+1} - \mathbf{z}^t) = (\sum_{i=1}^n \mathbf{A}_i\mathbf{x}_i^{t+1}) - \mathbf{b} = \mathbf{r}^{t+1}$; step ③ uses $\beta^t \geq \beta^0 \geq 2$ and $\frac{\beta^{t+1}}{\beta^t} - 1 \leq \xi$, and step ⑥ uses $\frac{\xi}{\sigma} \leq \epsilon_2$ as shown in Assumption 2.7.

In view of Inequalities (40), (41), and (42), we have the following inequalities:

$$\begin{aligned} & \tfrac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2 + \Theta_{\circ}^{t+1} - \Theta_{\circ}^t \\ \leq & \tfrac{1+\epsilon_2}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 - \textstyle\sum_{i=1}^n(\boldsymbol{\theta}_i - 1 - \boldsymbol{\theta}_i\boldsymbol{\alpha}_i(2+\xi))\cdot\frac{\mathsf{L}_i^t}{2}\|\Delta_i^{t+1}\|_2^2 \\ \overset{①}{\leq} & \tfrac{1+\epsilon_2}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 - \textstyle\sum_{i=1}^n \underbrace{\frac{1}{2}(\boldsymbol{\theta}_i - 1 - \boldsymbol{\theta}_i\boldsymbol{\alpha}_i(2+\epsilon_1))}_{\triangleq\gamma_i}\cdot\mathsf{L}_i^t\|\Delta_i^{t+1}\|_2^2, \end{aligned} \tag{43}$$

where step ① uses $\xi \leq \epsilon_1$ as shown in Assumption 2.7. We further obtain:

$$\begin{aligned} & \tfrac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2 + \epsilon_3[\textstyle\sum_{i=1}^n \boldsymbol{\gamma}_i\mathsf{L}_i^t\|\Delta_i^{t+1}\|_2^2] + \Theta_{\circ}^{t+1} - \Theta_{\circ}^t \\ \overset{①}{\leq} & \tfrac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2 + [\textstyle\sum_{i=1}^{n-1} \boldsymbol{\gamma}_i\mathsf{L}_i^t\|\Delta_i^{t+1}\|_2^2] + \epsilon_3\boldsymbol{\gamma}_n\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 + \Theta_{\circ}^{t+1} - \Theta_{\circ}^t \\ \overset{②}{\leq} & \tfrac{1+\epsilon_2}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 - \textstyle\sum_{i=1}^n \boldsymbol{\gamma}_i\mathsf{L}_i^t\|\Delta_i^{t+1}\|_2^2 + [\sum_{i=1}^{n-1} \boldsymbol{\gamma}_i\mathsf{L}_i^t\|\Delta_i^{t+1}\|_2^2] + \epsilon_3\boldsymbol{\gamma}_n\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 \\ = & \tfrac{1+\epsilon_2}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 + \boldsymbol{\gamma}_n(\epsilon_3 - 1)\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2, \end{aligned}$$

where step ① uses $\epsilon_3 \leq 1$ as shown in Assumption 2.7; and step ② uses Inequality (43).

□

C.2. Proof of Lemma 3.2

Proof. For any $i \in [n]$, we define $\mathbb{u}_i^{t+1} \triangleq \boldsymbol{\theta}_i\mathsf{L}_i^t[\mathbf{x}_i^{t+1} - \mathbf{x}_i^t - \boldsymbol{\alpha}_i(\mathbf{x}_i^t - \mathbf{x}_i^{t-1})] - \beta^t\mathbf{A}_i^\mathsf{T}[\sum_{j=i}^n \mathbf{A}_j(\mathbf{x}_j^{t+1} - \mathbf{x}_j^t)]$, and let $\mathbb{w}_i^{t+1} \in \partial h_i(\mathbf{x}_i^{t+1}) + \nabla f_i(\mathbf{x}_i^t)$.

We notice that $\mathbf{x}_i^{t+1}$ is the minimizer of the following problem:

$$\mathbf{x}_i^{t+1} \in \arg\min_{\mathbf{x}_i} \tfrac{\theta\mathsf{L}_i^t}{2}\|\mathbf{x}_i - \mathbf{y}_i^t\|_2^2 + h_i(\mathbf{x}_i) + \langle \mathbf{x}_i - \mathbf{y}_i^t, \nabla_{\mathbf{x}_i}G(\mathbf{x}_{[1,i-1]}^{t+1}, \mathbf{x}_{[i,n]}^t, \mathbf{z}^t; \beta^t)\rangle.$$

Using the necessary first-order optimality condition of the solution $\mathbf{x}_i^{t+1}$, we have:

$$\begin{aligned} \nabla_{\mathbf{x}_i}G(\mathbf{x}_{[1,i-1]}^{t+1}, \mathbf{x}_{[i,n]}^t, \mathbf{z}^t; \beta^t) &\in -\partial h_i(\mathbf{x}_i^{t+1}) - \theta\mathsf{L}_i^t(\mathbf{x}_i^{t+1} - \mathbf{y}_i^t) \\ &\overset{①}{=} -\partial h_i(\mathbf{x}_i^{t+1}) - \theta\mathsf{L}_i^t(\mathbf{x}_i^{t+1} - \mathbf{x}_i^t - \boldsymbol{\alpha}_i(\mathbf{x}_i^t - \mathbf{x}_i^{t-1})), \end{aligned} \tag{44}$$

where step ① uses the update rule of $\mathbf{y}_i^{t+1} = \mathbf{x}_i^{t+1} + \boldsymbol{\alpha}_i(\mathbf{x}_i^{t+1} - \mathbf{x}_i^t)$ for all $i \in [n]$.

Using the definition of the function $G(\mathbf{x}, \mathbf{z}; \beta) \triangleq \langle [\sum_{j=1}^n \mathbf{A}_j\mathbf{x}_j] - \mathbf{b}, \mathbf{z}\rangle + \frac{\beta}{2}\|[\sum_{j=1}^n \mathbf{A}_j\mathbf{x}_j] - \mathbf{b}\|_2^2 + \sum_{j=1}^n f_j(\mathbf{x}_j)$, we have:

$$
\begin{aligned}
& \nabla_{\mathbf{x}_i} G(\mathbf{x}_{[1,i-1]}^{t+1}, \mathbf{x}_{[i,n]}^t, \mathbf{z}^t; \beta^t) \\
= \;& \nabla f_i(\mathbf{x}_i^t) + \mathbf{A}_i^\mathsf{T}\mathbf{z}^t + \beta^t \mathbf{A}_i^\mathsf{T}\{[\textstyle\sum_{j=1}^{i-1} \mathbf{A}_j\mathbf{x}_j^{t+1}] + [\sum_{j=i}^n \mathbf{A}_j\mathbf{x}_j^t] - \mathbf{b}\} \\
= \;& \nabla f_i(\mathbf{x}_i^t) + \mathbf{A}_i^\mathsf{T}\mathbf{z}^t + \beta^t \mathbf{A}_i^\mathsf{T}\{[\textstyle\sum_{j=1}^{i-1} \mathbf{A}_j\mathbf{x}_j^{t+1}] + [\sum_{j=i}^n \mathbf{A}_j\mathbf{x}_j^{t+1}] - [\sum_{j=i}^n \mathbf{A}_j\mathbf{x}_j^{t+1}] + [\sum_{j=i}^n \mathbf{A}_j\mathbf{x}_j^t] - \mathbf{b}\} \\
= \;& \nabla f_i(\mathbf{x}_i^t) + \mathbf{A}_i^\mathsf{T}\mathbf{z}^t + \beta^t \mathbf{A}_i^\mathsf{T}\{\textstyle\sum_{j=1}^n \mathbf{A}_j\mathbf{x}_j^{t+1} - \mathbf{b} + [\sum_{j=i}^n \mathbf{A}_j(\mathbf{x}_j^t - \mathbf{x}_j^{t+1})]\} \\
\overset{①}{=} \;& \nabla f_i(\mathbf{x}_i^t) + \mathbf{A}_i^\mathsf{T}\mathbf{z}^t + \tfrac{1}{\sigma}\mathbf{A}_i^\mathsf{T}(\mathbf{z}^{t+1} - \mathbf{z}^t) + \beta^t \mathbf{A}_i^\mathsf{T}\{\textstyle\sum_{j=i}^n \mathbf{A}_j(\mathbf{x}_j^t - \mathbf{x}_j^{t+1})\},
\end{aligned} \tag{45}
$$

where step ① uses the update rule of $\mathbf{z}^{t+1}$ that $\mathbf{z}^{t+1} - \mathbf{z}^t = \sigma\beta^t(\sum_{i=1}^n \mathbf{A}_i\mathbf{x}_i^{t+1} - \mathbf{b})$. Combining the Equalities (44) and (45), we obtain the following result:

$$
\begin{aligned}
\mathbf{0} \;\in\; & \partial h_i(\mathbf{x}_i^{t+1}) + \boldsymbol{\theta}_i \mathsf{L}_i^t[\mathbf{x}_i^{t+1} - \mathbf{x}_i^t - \boldsymbol{\alpha}_i(\mathbf{x}_i^t - \mathbf{x}_i^{t-1})] + \nabla f_i(\mathbf{x}_i^t) \\
& + \mathbf{A}_i^\mathsf{T}\mathbf{z}^t + \beta^t \mathbf{A}_i^\mathsf{T}[\textstyle\sum_{j=i}^n \mathbf{A}_j(\mathbf{x}_j^t - \mathbf{x}_j^{t+1})] + \frac{1}{\sigma}\mathbf{A}_i^\mathsf{T}(\mathbf{z}^{t+1} - \mathbf{z}^t)
\end{aligned}
$$

Using the definition of $\mathbb{w}_i^{t+1}$ and $\mathbb{u}_i^{t+1}$ for all $i \in [n]$, we have: $\mathbf{0} = \mathbb{w}_i^{t+1} + \mathbb{u}_i^{t+1} + \mathbf{A}_i^\mathsf{T}\mathbf{z}^t + \frac{1}{\sigma}\mathbf{A}_i^\mathsf{T}(\mathbf{z}^{t+1} - \mathbf{z}^t)$. Multiplying both sides by $\sigma \in (0, 2)$, we have:

$$
\mathbf{0} = \sigma\mathbb{w}_i^{t+1} + \sigma\mathbf{A}_i^\mathsf{T}\mathbf{z}^t + \mathbf{A}_i^\mathsf{T}(\mathbf{z}^{t+1} - \mathbf{z}^t) + \sigma\mathbb{u}_i^{t+1}. \tag{46}
$$

Given that t can take on any integer value, we derive the following:

$$
\mathbf{0} = \sigma\mathbb{w}_i^t + \sigma\mathbf{A}_i^\mathsf{T}\mathbf{z}^{t-1} + \mathbf{A}_i^\mathsf{T}(\mathbf{z}^t - \mathbf{z}^{t-1}) + \sigma\mathbb{u}_i^t. \tag{47}
$$

Combining Equality (46) and Equality (47), we have:

$$
\mathbf{A}_i^\mathsf{T}(\mathbf{z}^{t+1} - \mathbf{z}^t) = (1-\sigma)\mathbf{A}_i^\mathsf{T}(\mathbf{z}^t - \mathbf{z}^{t-1}) - \sigma(\mathbb{w}_i^{t+1} - \mathbb{w}_i^t) - \sigma(\mathbb{u}_i^{t+1} - \mathbb{u}_i^t) \tag{48}
$$

In view of (48), we let $i = n$ and arrive at the following three distinct identities:

$$
\text{Condition } \boxed{\mathbb{I}}: \quad \underbrace{\mathbf{A}_n^\mathsf{T}(\mathbf{z}^{t+1} - \mathbf{z}^t)}_{\triangleq \mathbb{a}^{t+1}} = (1-\sigma)\underbrace{(\mathbf{A}_n^\mathsf{T}(\mathbf{z}^t - \mathbf{z}^{t-1}))}_{\triangleq \mathbb{a}^t} + \sigma\underbrace{(\mathbb{u}_n^t - \mathbb{u}_n^{t+1} + \mathbb{w}_n^t - \mathbb{w}_n^{t+1})}_{\mathbb{c}^t}. \tag{49}
$$

$$
\begin{aligned}
\text{Condition } \boxed{\mathbb{A}}: \quad & \underbrace{\mathbf{A}_n^\mathsf{T}(\mathbf{z}^{t+1} - \mathbf{z}^t) + \sigma\mathbb{u}_n^{t+1}}_{\triangleq \mathbb{a}^{t+1}} = (1-\sigma)\underbrace{(\mathbf{A}_n^\mathsf{T}(\mathbf{z}^t - \mathbf{z}^{t-1}) + \sigma\mathbb{u}_n^t)}_{\triangleq \mathbb{a}^t} + \sigma(\underbrace{\sigma\mathbb{u}_n^t + \mathbb{w}_n^t - \mathbb{w}_n^{t+1}}_{\triangleq \mathbb{c}^t}). \\
& \mathbf{A}_n^\mathsf{T}(\mathbf{z}^{t+1} - \mathbf{z}^t) + \sigma\mathbb{w}_n^{t+1} = (1-\sigma)(\mathbf{A}_n^\mathsf{T}(\mathbf{z}^t - \mathbf{z}^{t-1}) + \sigma\mathbb{w}_n^t) + \sigma(\sigma\mathbb{w}_n^t + \mathbb{u}_n^t - \mathbb{u}_n^{t+1}).
\end{aligned} \tag{50}
$$

Notably, our attention is specifically directed towards Formulations (49) and (50).

□

C.3. Proof of Lemma 3.3

Proof. We define $\chi \triangleq \boldsymbol{\theta}_n(1+\epsilon_3)$, $\tau = \boldsymbol{\alpha}_n^2(1+\epsilon_1)$, $\rho \triangleq 2\overline{\lambda}\chi\boldsymbol{\alpha}_n^2$, and $\Theta_{\mathsf{x}}^t \triangleq \rho\mathsf{L}_n^t\|\Delta_n^t\|_2^2 = 2\overline{\lambda}\boldsymbol{\theta}_n(1+\epsilon_3)\boldsymbol{\alpha}_n^2\mathsf{L}_n^t\|\Delta_n^t\|_2^2$.

We define $\mathbb{H}^t \triangleq \boldsymbol{\theta}_n\mathsf{L}_n^t\mathbf{I} - \beta^t\mathbf{A}_n^\mathsf{T}\mathbf{A}_n$, and $\mathbb{u}_i^{t+1} \triangleq \boldsymbol{\theta}_i\mathsf{L}_i^t(\mathbf{x}_i^{t+1} - \mathbf{x}_i^t - \boldsymbol{\alpha}_i(\mathbf{x}_i^t - \mathbf{x}_i^{t-1})) + \beta^t\mathbf{A}_i^\mathsf{T}[\sum_{j=i}^n \mathbf{A}_j(\mathbf{x}_j^t - \mathbf{x}_j^{t+1})]$.

(a) we bound the term $\|\mathbb{w}_n^{t+1} - \mathbb{w}_n^t\|_2^2$.

$$
\begin{aligned}
\|\mathbb{w}_n^{t+1} - \mathbb{w}_n^t\|_2^2 &= \|\partial h_n(\mathbf{x}_n^{t+1}) + \nabla f_n(\mathbf{x}_n^t) - \partial h_n(\mathbf{x}_n^t) - \nabla f_n(\mathbf{x}_n^{t-1})\|_2^2 \\
&\leq 4\|\partial h_n(\mathbf{x}_n^{t+1})\|_2^2 + 4\|\nabla f_n(\mathbf{x}_n^t)\|_2^2 + 4\|\partial h_n(\mathbf{x}_n^t)\|_2^2 + 4\|\nabla f_n(\mathbf{x}_n^{t-1})\|_2^2 \\
&\overset{②}{=} 4C_h^2 + 4C_f^2 + 4C_h^2 + 4C_f^2 \\
&\leq 8C_h^2 + 8C_f^2 \triangleq \iota,
\end{aligned}
$$

where step ① uses Assumption 1.2.

(b) We now bound the term $\frac{1}{\beta^t}\|\mathbb{u}_n^{t+1}\|_2^2$. First, using the definition of $\mathbb{u}_i^{t+1}$ and $\mathbb{H}^t$, we have:

$$\begin{aligned}\mathbb{u}_n^{t+1} &= \boldsymbol{\theta}_n \mathsf{L}_n^t(\mathbf{x}_n^{t+1}-\mathbf{x}_n^t-\boldsymbol{\alpha}_n(\mathbf{x}_n^t-\mathbf{x}_n^{t-1}))+\beta^t\mathbf{A}_n^\mathsf{T}[\textstyle\sum_{j=n}^n\mathbf{A}_j(\mathbf{x}_j^t-\mathbf{x}_j^{t+1})]\\ &= (\boldsymbol{\theta}_n\mathsf{L}_n^t\mathbf{I}-\beta^t\mathbf{A}_n^\mathsf{T}\mathbf{A}_n)(\mathbf{x}_n^{t+1}-\mathbf{x}_n^t)-\boldsymbol{\theta}_n\boldsymbol{\alpha}_n\mathsf{L}_n^t(\mathbf{x}_n^t-\mathbf{x}_n^{t-1})\\ &= \mathbb{H}^t\Delta_n^{t+1}-\boldsymbol{\theta}_n\boldsymbol{\alpha}_n\mathsf{L}_n^t\Delta_n^t.\end{aligned}$$

Second, we bound the term L_i^t using the following inequalities:

$$\mathsf{L}_i^t = L_i+\beta^t\overline{\lambda}\overset{①}{\leq}\beta^t[\tfrac{L_i}{\beta^0}+\overline{\lambda}]\overset{②}{\leq}\beta^t[\tfrac{L_i}{L_i/(\epsilon_3\overline{\lambda})}+\overline{\lambda}]=\beta^t\overline{\lambda}(1+\epsilon_3), \tag{51}$$

where step ① uses $\beta^t\geq\beta^0$; step ② uses $\beta^0\geq L_i/(\epsilon_3\overline{\lambda})$ as shown in Assumption 2.7.

Third, we bound the term $\|\mathbb{H}^t\|$. We assume that $\mathbf{A}_n^\mathsf{T}\mathbf{A}_n\in\mathbb{R}^{\mathbf{d}_i\times\mathbf{d}_i}$ has the singular value decomposition that $\mathbf{A}_n^\mathsf{T}\mathbf{A}_n=\tilde{\mathbf{U}}^\mathsf{T}\mathrm{diag}(\boldsymbol{\lambda})\tilde{\mathbf{U}}$ with $\tilde{\mathbf{U}}\in\mathbb{R}^{\mathbf{d}_i\times\mathbf{d}_i}$, $\boldsymbol{\lambda}\in\mathbb{R}^{\mathbf{d}_i\times 1}$, and $\tilde{\mathbf{U}}^\mathsf{T}\tilde{\mathbf{U}}=\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\mathsf{T}=\mathbf{I}_{\mathbf{d}_i}$, where $\mathrm{diag}(\boldsymbol{\lambda})$ denotes a diagonal matrix with $\boldsymbol{\lambda}$ as the main diagonal entries. Using the definition of $\mathbb{H}^t$, we have:

$$\begin{aligned}\|\mathbb{H}^t\| &= \|\boldsymbol{\theta}_n\mathsf{L}_n^t\mathbf{I}-\beta^t\mathbf{A}_n^\mathsf{T}\mathbf{A}_n\|\\ &\overset{①}{=} \|\boldsymbol{\theta}_n\mathsf{L}_n^t-\beta^t\boldsymbol{\lambda}\|_\infty\\ &\overset{②}{=} \boldsymbol{\theta}_n\mathsf{L}_n^t-\min(\beta^t\boldsymbol{\lambda})\\ &\overset{③}{\leq} \boldsymbol{\theta}_n\beta^t(1+\epsilon_3)\overline{\lambda}-\beta^t\underline{\lambda}\\ &= \overline{\lambda}\beta^t(\underbrace{\boldsymbol{\theta}_n(1+\epsilon_3)}_{\triangleq\chi}-\underline{\lambda}/\overline{\lambda}),\end{aligned} \tag{52}$$

where step ① uses $\|\boldsymbol{\theta}_n\mathsf{L}_n^t\mathbf{I}-\beta^t\mathbf{A}_n^\mathsf{T}\mathbf{A}_n\|=\|\tilde{\mathbf{U}}^\mathsf{T}\mathrm{diag}(\boldsymbol{\theta}_n\mathsf{L}_n^t-\beta^t\boldsymbol{\lambda})\tilde{\mathbf{U}}\|$; step ② uses the fact that $\|\rho-\mathbf{x}\|_\infty=\max(\rho-\mathbf{x})=\rho-\min(\mathbf{x})$ whenever $\rho\geq\max(\mathbf{x})$ for all ρ and $\mathbf{x}$; step ③ uses Inequality (51).

Fourth, we bound the term $\frac{2}{\beta^t}\|\boldsymbol{\theta}_n\boldsymbol{\alpha}_n\mathsf{L}_n^t\Delta_n^t\|_2^2$. We have:

$$\begin{aligned}\tfrac{2}{\beta^t}\|\boldsymbol{\theta}_n\boldsymbol{\alpha}_n\mathsf{L}_n^t\Delta_n^t\|_2^2 &= \tfrac{\mathsf{L}_n^t}{\beta^t}\cdot 2\boldsymbol{\theta}_n^2\boldsymbol{\alpha}_n^2\mathsf{L}_n^t\|\Delta_n^t\|_2^2\\ &\overset{①}{\leq} \overline{\lambda}(1+\epsilon_3)\cdot 2\boldsymbol{\theta}_n\boldsymbol{\alpha}_n^2\mathsf{L}_n^t\|\Delta_n^t\|_2^2\triangleq\Theta_\mathsf{x}^t\\ &\overset{②}{=} \Theta_\mathsf{x}^t-\Theta_\mathsf{x}^{t+1}+\overline{\lambda}(1+\epsilon_3)\cdot 2\boldsymbol{\theta}_n\boldsymbol{\alpha}_n^2\mathsf{L}_n^{t+1}\|\Delta_n^{t+1}\|_2^2\\ &\overset{③}{\leq} \Theta_\mathsf{x}^t-\Theta_\mathsf{x}^{t+1}+2\overline{\lambda}\underbrace{(1+\epsilon_3)\boldsymbol{\theta}_n}_{\triangleq\chi}\cdot\underbrace{\boldsymbol{\alpha}_n^2(1+\epsilon_1)}_{\triangleq\tau}\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2,\end{aligned} \tag{53}$$

where step ① uses Inequality (51); step ② uses the definition of Θ_x^t; step ③ uses the fact that $\mathsf{L}_n^{t+1}\leq(1+\xi)\mathsf{L}_n^t\leq(1+\epsilon_1)\mathsf{L}_n^t$.

Finally, we bound the term $\frac{1}{\beta^t}\|\mathbb{u}_n^{t+1}\|_2^2$ using the following inequalities.

$$\begin{aligned}\tfrac{1}{\beta^t}\|\mathbb{u}_n^{t+1}\|_2^2 &\overset{①}{=} \tfrac{1}{\beta^t}\|\mathbb{H}^t\Delta_n^{t+1}-\boldsymbol{\theta}_n\boldsymbol{\alpha}_n\mathsf{L}_n^t\Delta_n^t\|_2^2\\ &\overset{②}{\leq} \tfrac{2}{\beta^t}\|\boldsymbol{\theta}_n\boldsymbol{\alpha}_n\mathsf{L}_n^t\Delta_n^t\|_2^2+\tfrac{2}{\beta^t}\|\mathbb{H}^t\Delta_n^{t+1}\|_2^2\\ &\overset{③}{\leq} \Theta_\mathsf{x}^t-\Theta_\mathsf{x}^{t+1}+\{2\overline{\lambda}\chi\tau+2\beta^t\cdot[\overline{\lambda}(\chi-\underline{\lambda}/\overline{\lambda})]^2\cdot\tfrac{1}{\mathsf{L}_n^t}\}\cdot\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2\\ &\overset{④}{\leq} \Theta_\mathsf{x}^t-\Theta_\mathsf{x}^{t+1}+\{2\overline{\lambda}\chi\tau+2\overline{\lambda}(\chi-\underline{\lambda}/\overline{\lambda})^2\}\cdot\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2,\end{aligned} \tag{54}$$

where step ① uses the definition of $\mathbb{u}^{t+1}$ as shown in (51); step ② uses the inequality $\|\mathbf{a}+\mathbf{b}\|_2^2\leq 2\|\mathbf{a}\|_2^2+2\|\mathbf{b}\|_2^2$; step ③ uses Inequality (53); step ④ uses $\beta^t\overline{\lambda}\leq L_n+\beta^t\|\mathbf{A}_n\|_2^2\triangleq\mathsf{L}_n^t$.

□

C.4. Proof of Lemma 3.8

Proof. Let $\underline{\Theta}$ be defined in Assumption 1.4.

Initially, using the definition of Θ^t as presented in Equation (114), we have the following inequalities:

$$\begin{aligned}\Theta^t &= \mathcal{L}(\mathbf{x}^t,\mathbf{z}^t;\beta^t)+\tfrac{1}{2}\textstyle\sum_{i=1}^n \boldsymbol{\theta}_i\boldsymbol{\alpha}_i\mathsf{L}_i^t\|\Delta_i^t\|_2^2+\frac{C_{\mathsf{a}}}{\beta^t}\|\mathbb{a}^t\|_2^2+\frac{C_{\mathsf{u}}}{\beta^t}\|\mathbb{u}_n^t\|_2^2+C_{\mathsf{x}}\beta^t\|\Delta_n^t\|_2^2\\ &\geq \mathcal{L}(\mathbf{x}^t,\mathbf{z}^t;\beta^t)\geq\mathcal{L}(\mathbf{x}^t,\mathbf{z}^t;0).\end{aligned}\tag{55}$$

We now conclude the proof of this lemma through contradiction. Suppose that there exists $t_0\geq 1$ such that $\Theta^{t_0}<\underline{\Theta}$. We derive the following inequalities:

$$\begin{aligned}\textstyle\sum_{t=1}^T(\Theta^t-\underline{\Theta}) &= [\textstyle\sum_{t=1}^{t_0-1}(\Theta^t-\underline{\Theta})]+[\textstyle\sum_{t=t_0}^{T}(\Theta^t-\underline{\Theta})]\\ &\leq [\textstyle\sum_{t=1}^{t_0-1}(\Theta^t-\underline{\Theta})]+(T+1-t_0)\cdot\max_{t=t_0}^T(\Theta^t-\underline{\Theta})\\ &\overset{①}{\leq} [\textstyle\sum_{t=0}^{t_0-1}(\Theta^t-\underline{\Theta})]+(T+1-t_0)\cdot(\Theta^{t_0}-\underline{\Theta}),\end{aligned}\tag{56}$$

where step ① uses $\Theta^t\leq\Theta^{t_0}$ for all $t\geq t_0$. We closely examine Inequality (56). As t_0 is finite, the sum $\sum_{t=1}^{t_0-1}(\Theta^t-\underline{\Theta})$ is upper bounded. Considering the negativity of the term $(\Theta^{t_0}-\underline{\Theta})$, we deduce from Inequality (56):

$$\lim_{T\to\infty}\textstyle\sum_{t=1}^T(\Theta^t-\underline{\Theta})=-\infty.\tag{57}$$

Meanwhile, for all $t\geq 0$, the following inequalities hold:

$$\begin{aligned}\Theta^t-\underline{\Theta} &\overset{①}{\geq} \mathcal{L}(\mathbf{x}^t,\mathbf{z}^t;0)-\underline{\Theta}\\ &\overset{②}{=} \textstyle\sum_{i=1}^n[f_i(\mathbf{x}_i^t)+h_i(\mathbf{x}_i^t)]+\langle[\sum_{i=1}^n\mathbf{A}_i\mathbf{x}_i^t]-\mathbf{b},\mathbf{z}^t\rangle+\frac{\beta^t}{2}\|[\sum_{i=1}^n\mathbf{A}_i\mathbf{x}_i^t]-\mathbf{b}\|_2^2-\underline{\Theta}\\ &\overset{③}{\geq} \langle[\textstyle\sum_{i=1}^n\mathbf{A}_i\mathbf{x}_i^t]-\mathbf{b},\mathbf{z}^t\rangle\\ &\overset{④}{=} \tfrac{1}{\sigma\beta^{t-1}}\langle\mathbf{z}^t-\mathbf{z}^{t-1},\mathbf{z}^t\rangle\\ &\overset{⑤}{=} \tfrac{1}{2\sigma}\{\tfrac{1}{\beta^{t-1}}\|\mathbf{z}^t\|_2^2-\tfrac{1}{\beta^{t-1}}\|\mathbf{z}^{t-1}\|_2^2+\tfrac{1}{\beta^{t-1}}\|\mathbf{z}^t-\mathbf{z}^{t-1}\|_2^2\}\\ &\overset{⑥}{\geq} \tfrac{1}{2\sigma}\{\tfrac{1}{\beta^{t-1}}\|\mathbf{z}^t\|_2^2-\tfrac{1}{\beta^{t-2}}\|\mathbf{z}^{t-1}\|_2^2+0\},\end{aligned}\tag{58}$$

where step ① uses Inequality (55); step ② uses the definition of $\mathcal{L}(\mathbf{x}^t,\mathbf{z}^t;\beta)$ with $\beta=0$; step ③ uses Assumption 1.4; step ④ uses $\mathbf{z}^{t+1}=\mathbf{z}^t+\sigma\beta^t[(\sum_{i=1}^n\mathbf{A}_i\mathbf{x}_i)-\mathbf{b}]$; step ⑤ uses the Pythagoras relation presented in (26); step ⑥ uses $\frac{1}{\beta^t}\leq\frac{1}{\beta^{t-1}}$.

Telescoping Inequality (58) over t from 1 to T, we have:

$$\textstyle\sum_{t=1}^T(\Theta^t-\underline{\Theta})\geq\frac{1}{2\sigma}\cdot\{\frac{1}{\beta^T}\|\mathbf{z}^T\|_2^2-\frac{1}{\beta^0}\|\mathbf{z}^0\|_2^2\}\geq-\frac{1}{2\sigma\beta^0}\|\mathbf{z}^0\|_2^2,$$

which contradicts with (57). Therefore, we conclude that $\Theta^t\geq\underline{\Theta}$ for all $t\geq 0$.

□

C.5. Proof of Corollary 3.9

For both conditions $\boxed{\mathbb{I}}$ and $\boxed{\mathbb{A}}$, we have from Theorem (3.5) and Theorem (3.7):

$$\mathcal{E}^{t+1}+\Theta^{t+1}\leq\Theta^t+\tfrac{C_{\mathsf{w}}}{\beta^t}$$

Telescoping this inequality over t from 0 to T, we have:

$$\mathcal{E}^{T+1}\leq\Theta^0-\Theta^{T+1}+\textstyle\sum_{t=0}^T\frac{C_{\mathsf{w}}}{\beta^t}\overset{①}{\leq}\Theta^0-\underline{\Theta}+C_{\mathsf{w}}C_{\mathsf{b}},$$

where step ① uses Lemma 3.8 that $\Theta^t\geq\underline{\Theta}$ for all t; step ② uses Inequality (3) that $\sum_{t=0}^\infty\frac{1}{\beta^t}\leq C_{\mathsf{b}}<+\infty$.

C.6. Proof of Lemma 3.4

Proof. For any $\sigma \in [1,2)$, we define $\sigma_1 \triangleq \frac{\sigma}{(1-|1-\sigma|)^2}$, and $\sigma_2 \triangleq \frac{|1-\sigma|}{\sigma(1-|1-\sigma|)}$.

We define $\mathbb{a}^{t+1} \triangleq \mathbf{A}_n^{\mathsf{T}}(\mathbf{z}^{t+1}-\mathbf{z}^t)$, and $\mathbb{c}^t \triangleq \mathbb{u}_n^t - \mathbb{u}_n^{t+1} + \mathbb{w}_n^t - \mathbb{w}_n^{t+1}$, where $\mathbb{w}_n^{t+1} \in \partial h_n(\mathbf{x}_n^{t+1}) + \nabla f_n(\mathbf{x}_n^t)$.

We define $C_{\mathsf{a}} = \delta\sigma_2$, $C_{\mathsf{u}} = 2\delta\sigma_1(1+\epsilon_3)$, $C_{\mathsf{x}} = 2C_{\mathsf{u}}\rho$, and $C_{\mathsf{w}} = \iota\delta\sigma_1(1+1/\epsilon_3)$, where $\iota \triangleq 8C_f^2 + 8C_h^2$.

We define $\Theta_{\mathsf{a}}^t \triangleq \frac{C_{\mathsf{a}}}{\beta^t}\|\mathbb{a}^t\|_2^2$, and $\Theta_{\mathsf{u}}^t \triangleq \frac{C_{\mathsf{u}}}{\beta^t}\|\mathbb{u}^t\|_2^2$.

We derive the following inequalities:

$$
\begin{aligned}
& \tfrac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1}-\mathbf{z}^t\|_2^2 \\
\overset{①}{=}\ & \tfrac{\delta}{\sigma\beta^t}\|\mathbb{a}^{t+1}\|_2^2 \\
\overset{②}{\leq}\ & \sigma_2\delta\big(\tfrac{1}{\beta^t}\|\mathbb{a}^t\|_2^2 - \tfrac{1}{\beta^t}\|\mathbb{a}^{t+1}\|_2^2\big) + \tfrac{\delta\sigma_1}{\beta^t}\|\mathbb{c}^t\|_2^2 \\
\overset{③}{\leq}\ & \underbrace{\sigma_2\delta}_{\triangleq C_{\mathsf{a}}}\big(\tfrac{1}{\beta^t}\|\mathbb{a}^t\|_2^2 - \tfrac{1}{\beta^{t+1}}\|\mathbb{a}^{t+1}\|_2^2\big) + \tfrac{\delta\sigma_1}{\beta^t}\|\mathbb{c}^t\|_2^2 \\
\overset{④}{=}\ & \Theta_{\mathsf{a}}^t - \Theta_{\mathsf{a}}^{t+1} + \tfrac{\delta\sigma_1}{\beta^t}\|(\mathbb{u}_n^t - \mathbb{u}_n^{t+1}) + (\mathbb{w}_n^t - \mathbb{w}_n^{t+1})\|_2^2 \\
\overset{⑤}{\leq}\ & \Theta_{\mathsf{a}}^t - \Theta_{\mathsf{a}}^{t+1} + \delta\sigma_1\{\tfrac{1+1/\epsilon_3}{\beta^t}\|\mathbb{w}_n^t - \mathbb{w}_n^{t+1}\|_2^2\} + \delta\sigma_1\{\tfrac{1+\epsilon_3}{\beta^t}\|\mathbb{u}_n^{t+1} - \mathbb{u}_n^t\|_2^2\} \\
\overset{⑥}{\leq}\ & \Theta_{\mathsf{a}}^t - \Theta_{\mathsf{a}}^{t+1} + \iota\delta\sigma_1(1+1/\epsilon_3)\tfrac{1}{\beta^t} + \delta\sigma_1(1+\epsilon_3)\{\tfrac{1}{\beta^t}\|\mathbb{u}_n^{t+1} - \mathbb{u}_n^t\|_2^2\} \\
\overset{⑦}{\leq}\ & \Theta_{\mathsf{a}}^t - \Theta_{\mathsf{a}}^{t+1} + \tfrac{C_{\mathsf{w}}}{\beta^t} + \underbrace{2\delta\sigma_1(1+\epsilon_3)}_{\triangleq C_{\mathsf{u}}}\{\tfrac{1}{\beta^t}\|\mathbb{u}_n^{t+1}\|_2^2 + \tfrac{1}{\beta^t}\|\mathbb{u}_n^t\|_2^2\} \\
\overset{⑧}{=}\ & \Theta_{\mathsf{a}}^t - \Theta_{\mathsf{a}}^{t+1} + \tfrac{C_{\mathsf{w}}}{\beta^t} + \Theta_{\mathsf{u}}^t - \Theta_{\mathsf{u}}^{t+1} + C_{\mathsf{u}}\{(\tfrac{1}{\beta^t} + \tfrac{1}{\beta^{t+1}})\|\mathbb{u}_n^{t+1}\|_2^2\} \\
\overset{⑨}{=}\ & \Theta_{\mathsf{a}}^t - \Theta_{\mathsf{a}}^{t+1} + \tfrac{C_{\mathsf{w}}}{\beta^t} + \Theta_{\mathsf{u}}^t - \Theta_{\mathsf{u}}^{t+1} + 2C_{\mathsf{u}}\{\tfrac{1}{\beta^t}\|\mathbb{u}_n^{t+1}\|_2^2\},
\end{aligned} \tag{59}
$$

where step ① uses the fact that $\mathbf{A}_n$ is an identity matrix; step ② uses Lemma A.6 with $\mathbf{c} = \mathbb{a}^{t+1}$, $\mathbf{b} = \mathbb{a}^t$ and $\mathbf{a} = \mathbb{c}^t$ that:

$$
\tfrac{1}{\sigma\beta^t}\|\mathbb{a}^{t+1}\|_2^2 \leq \tfrac{\sigma_2}{\beta^t}(\|\mathbb{a}^t\|_2^2 - \|\mathbb{a}^{t+1}\|_2^2) + \tfrac{\sigma_1}{\beta^t}\|\mathbb{c}^t\|_2^2;
$$

step ③ uses $-\frac{1}{\beta^t} \leq -\frac{1}{\beta^{t+1}}$; step ④ uses the definitions of Θ_{a}^t and $\mathbb{c}^t$; step ⑤ uses **Part (b)** of Lemma A.5 with $\mathbf{a} = \mathbb{u}^t - \mathbb{u}^{t+1}$, $\mathbf{b} = \mathbb{w}^t - \mathbb{w}^{t+1}$, and $\theta = \epsilon_3$; step ⑥ uses $\|\mathbb{w}_n^{t+1} - \mathbb{w}_n^t\|_2^2 \leq \iota$ as shown in Lemma C.3; step ⑦ uses the definition of C_{w}, and the inequality $\|\mathbf{a}+\mathbf{b}\|_2^2 \leq 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2$ for all $\mathbf{a}$ and $\mathbf{b}$; step ⑧ uses the definition of Θ_{u}^t; step ⑨ uses $\frac{1}{\beta^{t+1}} \leq \frac{1}{\beta^t}$.

Using **Part (b)** of Lemma 3.3, we have:

$$
2C_{\mathsf{u}}\tfrac{1}{\beta^t}\|\mathbb{u}^{t+1}\|_2^2 \leq 2C_{\mathsf{u}}\cdot 2\overline{\lambda}\cdot\{(\chi - \underline{\lambda}/\overline{\lambda})^2 + \chi\tau\}\cdot \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 + \underbrace{2C_{\mathsf{u}}\rho}_{\triangleq C_{\mathsf{x}}}\cdot\mathsf{L}_n^t\|\Delta_n^t\|_2^2 - 2C_{\mathsf{u}}\rho\mathsf{L}_n^{t+1}\|\Delta_n^{t+1}\|_2^2. \tag{60}
$$

We define $\Theta_{\mathsf{z}}^t \triangleq \frac{C_{\mathsf{a}}}{\beta^t}\|\mathbb{a}^t\|_2^2 + \frac{C_{\mathsf{u}}}{\beta^t}\|\mathbb{u}_n^t\|_2^2 + C_{\mathsf{x}}\mathsf{L}_n^t\|\Delta_n^t\|_2^2$. Combining Inequalities (59) and (60), we have:

$$
\begin{aligned}
& \tfrac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1}-\mathbf{z}^t\|_2^2 + \Theta_{\mathsf{z}}^{t+1} - \Theta_{\mathsf{z}}^t \\
\leq\ & \{\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2\}\cdot 2C_{\mathsf{u}}\cdot 2\overline{\lambda}\cdot\{(\chi - \underline{\lambda}/\overline{\lambda})^2 + \chi\tau\} \\
\overset{①}{\leq}\ & \{\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2\}\cdot 4C_{\mathsf{u}}\cdot\{(\chi-1)^2 + \chi\tau\},
\end{aligned}
$$

where step ① uses $\underline{\lambda} = \overline{\lambda} = 1$ and $\xi \leq \epsilon_1$ as shown in Assumption 2.7.

□

C.7. Proof of Theorem 3.5

Proof. We define $\boldsymbol{\gamma}_i' \triangleq \boldsymbol{\gamma}_i[1-\epsilon_3]$, and $\boldsymbol{\gamma}_i \triangleq \frac{1}{2}[\boldsymbol{\theta}_i - 1 - (2+\epsilon_1)\boldsymbol{\alpha}_i\boldsymbol{\theta}_i]$ for all $i \in [n]$.

We let $\Theta^t \triangleq \Theta_{\mathsf{z}}^t + \Theta_{\mathsf{o}}^t$, $\mathbf{r}^t \triangleq [\sum_{i=1}^n \mathbf{A}_i\mathbf{x}_i^t] - \mathbf{b}$, and $\mathcal{E}^{t+1} \triangleq \frac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2 + \epsilon_3\sum_{i=1}^n \boldsymbol{\gamma}_i \mathsf{L}_i^t\|\Delta_i^{t+1}\|_2^2$.

First, based on Lemma 3.1, we have:

$$\mathcal{E}^{t+1} + \Theta_{\mathsf{o}}^{t+1} - \Theta_{\mathsf{o}}^t - \tfrac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1}-\mathbf{z}^t\|_2^2 \leq (\epsilon_3-1)\boldsymbol{\gamma}_n\cdot\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 = -\boldsymbol{\gamma}_n'\cdot\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 \tag{61}$$

Second, using Lemma 3.4, we obtain:

$$\tfrac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1}-\mathbf{z}^t\|_2^2 + \Theta_{\mathsf{z}}^{t+1} - \Theta_z^t \leq \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2\cdot 4C_{\mathsf{u}}\cdot\{(\chi-1)^2+\chi\tau\}. \tag{62}$$

Adding Inequalities (61) and (62) together, we have:

$$\begin{aligned}\mathcal{E}^{t+1} + \Theta^{t+1} - \Theta^t &\leq \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2\cdot\{4C_{\mathsf{u}}\cdot\{(\chi-1)^2+\chi\tau\} - \boldsymbol{\gamma}_n'\}\\ &\overset{①}{\leq} 0,\end{aligned}$$

where step ① uses Inequality (9).

□

C.8. Proof of Lemma 3.6

Proof. For any $\sigma \in (0,1)$, we define $\sigma_1 \triangleq \frac{\sigma}{(1-|1-\sigma|)^2}$, and $\sigma_2 \triangleq \frac{|1-\sigma|}{\sigma(1-|1-\sigma|)}$.

We define $\mathbb{a}^{t+1} \triangleq \mathbf{A}_n^\mathsf{T}(\mathbf{z}^{t+1}-\mathbf{z}^t) + \sigma\mathbb{u}_n^t$, and $\mathbb{c}^t \triangleq \sigma\mathbb{u}_n^t + \mathbb{w}_n^t - \mathbb{w}_n^{t+1}$, where $\mathbb{w}_n^{t+1} \in \partial h_n(\mathbf{x}_n^{t+1}) + \nabla f_n(\mathbf{x}_n^t)$.

We define $C_{\mathsf{a}} \triangleq 2\delta\sigma_2/\underline{\lambda}$, $C_{\mathsf{u}} \triangleq 4\delta\sigma/\underline{\lambda}$, $C_{\mathsf{x}} \triangleq 4\rho\delta\sigma/\underline{\lambda}$, and $C_{\mathsf{w}} \triangleq \iota 4\delta/(\sigma\underline{\lambda})$, where $\iota \triangleq 8C_f^2 + 8C_h^2$.

We define $\Theta_{\mathsf{a}}^t \triangleq \frac{C_{\mathsf{a}}}{\beta^t}\|\mathbb{a}^t\|_2^2$, and $\Theta_{\mathsf{u}}^t \triangleq \frac{C_{\mathsf{u}}}{\beta^t}\|\mathbb{u}^t\|_2^2$.

We derive the following inequalities:

$$\begin{aligned}
&\tfrac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1}-\mathbf{z}^t\|_2^2\\
\overset{①}{\leq}\ &\tfrac{\delta}{\sigma\beta^t\cdot\underline{\lambda}}\|\mathbf{A}_n^\mathsf{T}(\mathbf{z}^{t+1}-\mathbf{z}^t)\|_2^2\\
\overset{②}{=}\ &\tfrac{\delta}{\underline{\lambda}}\cdot\tfrac{1}{\sigma\beta^t}\|\mathbb{a}^{t+1}-\sigma\mathbb{u}_n^{t+1}\|_2^2\\
\overset{③}{\leq}\ &\tfrac{2\delta}{\underline{\lambda}}\cdot\{\tfrac{1}{\sigma\beta^t}\|\mathbb{a}^{t+1}\|_2^2 + \tfrac{\sigma}{\beta^t}\|\mathbb{u}_n^{t+1}\|_2^2\}\\
\overset{④}{\leq}\ &\tfrac{2\delta}{\underline{\lambda}}\cdot\{\tfrac{\sigma_2}{\beta^t}\|\mathbb{a}^t\|_2^2 - \tfrac{\sigma_2}{\beta^t}\|\mathbb{a}^{t+1}\|_2^2 + \tfrac{\sigma_1}{\beta^t}\|\mathbb{c}^t\|_2^2\} + \tfrac{2\delta\sigma}{\beta^t\underline{\lambda}}\|\mathbb{u}_n^{t+1}\|_2^2\\
\overset{⑤}{\leq}\ &\underbrace{\tfrac{2\delta}{\underline{\lambda}}\tfrac{\sigma_2}{\beta^t}\|\mathbb{a}^t\|_2^2}_{\triangleq\Theta_{\mathsf{a}}^t} - \tfrac{2\delta}{\underline{\lambda}}\tfrac{\sigma_2}{\beta^{t+1}}\|\mathbb{a}^{t+1}\|_2^2 + \tfrac{2\delta}{\underline{\lambda}}\tfrac{1}{\sigma\beta^t}\|\mathbb{c}^t\|_2^2 + \tfrac{2\delta\sigma}{\beta^t\underline{\lambda}}\|\mathbb{u}_n^{t+1}\|_2^2\\
\overset{⑥}{\leq}\ &\Theta_{\mathsf{a}}^t - \Theta_{\mathsf{a}}^{t+1} + \tfrac{2\delta}{\sigma\underline{\lambda}}\cdot\{\tfrac{1}{\beta^t}\|\sigma\mathbb{u}^t + (\mathbb{w}_n^t - \mathbb{w}_n^{t+1})\|_2^2\} + \tfrac{2\delta\sigma}{\beta^t\underline{\lambda}}\|\mathbb{u}_n^{t+1}\|_2^2\\
\overset{⑦}{\leq}\ &\Theta_{\mathsf{a}}^t - \Theta_{\mathsf{a}}^{t+1} + \underbrace{\tfrac{2\delta}{\sigma\underline{\lambda}}\cdot\tfrac{2}{\beta^t}\|\sigma\mathbb{u}_n^t\|_2^2}_{\triangleq\Theta_{\mathsf{u}}^t} + \tfrac{2\delta}{\sigma\underline{\lambda}}\cdot\tfrac{2}{\beta^t}\|\mathbb{w}_n^t - \mathbb{w}_n^{t+1}\|_2^2 + \tfrac{2\delta\sigma}{\beta^t\underline{\lambda}}\|\mathbb{u}_n^{t+1}\|_2^2\\
\overset{⑧}{=}\ &\Theta_{\mathsf{a}}^t - \Theta_{\mathsf{a}}^{t+1} + \Theta_{\mathsf{u}}^t - \Theta_{\mathsf{u}}^{t+1} + \tfrac{4\delta}{\sigma\underline{\lambda}\beta^t}\cdot\iota + \tfrac{2\delta\sigma}{\underline{\lambda}}\cdot(\tfrac{1}{\beta^t} + \tfrac{1}{\beta^{t+1}})\|\mathbb{u}_n^{t+1}\|_2^2\\
\overset{⑨}{\leq}\ &\Theta_{\mathsf{a}}^t - \Theta_{\mathsf{a}}^{t+1} + \Theta_{\mathsf{u}}^t - \Theta_{\mathsf{u}}^{t+1} + \underbrace{\tfrac{4\delta\iota}{\sigma\underline{\lambda}}}_{\triangleq C_{\mathsf{w}}}\cdot\tfrac{1}{\beta^t} + \underbrace{\tfrac{4\delta\sigma}{\underline{\lambda}}}_{\triangleq C_{\mathsf{u}}}\cdot\tfrac{1}{\beta^t}\|\mathbb{u}_n^{t+1}\|_2^2,
\end{aligned}\tag{63}$$

where step ① uses the fact that $\underline{\lambda}\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}_n^\mathsf{T}\mathbf{x}\|_2^2$ for all $\mathbf{x}$; step ② uses the definition of $\mathbb{a}^{t+1}$; step ③ uses the inequality $\|\mathbf{a}+\mathbf{b}\|_2^2 \leq 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2$ for all $\mathbf{a}$ and $\mathbf{b}$; step ④ uses Lemma A.6 with $\mathbf{c} = \mathbb{a}^{t+1}$, $\mathbf{b} = \mathbb{a}^t$, and $\mathbf{a} = \mathbb{c}^t$ that

$$\tfrac{1}{\sigma\beta^t}\|\mathbb{a}^{t+1}\|_2^2 \quad \leq \quad \tfrac{\sigma_1}{\beta^t}\|\mathbb{c}^t\|_2^2 + \tfrac{\sigma_2}{\beta^t}(\|\mathbb{a}^t\|_2^2 - \|\mathbb{a}^{t+1}\|_2^2);$$

step ⑤ uses $-\frac{1}{\beta^t} \leq -\frac{1}{\beta^{t+1}}$ and $\sigma_1 = \frac{1}{\sigma}$ when $\sigma \in (0,1)$; step ⑥ uses the definition of Θ_{a}^t and $\mathbb{c}$; step ⑦ uses the inequality $\|\mathbf{a}+\mathbf{b}\|_2^2 \leq 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2$ for all $\mathbf{a}$ and $\mathbf{b}$; step ⑧ uses $\|\mathbb{w}_n^t - \mathbb{w}_n^{t+1}\|_2^2 \leq \iota$ which is presented in Lemma 3.3; step ⑨ uses $\frac{1}{\beta^{t+1}} \leq \frac{1}{\beta^t}$, and the definition of $\{C_{\mathsf{w}}, C_{\mathsf{u}}\}$.

Using **Part (b)** of Lemma 3.3, we have:

$$C_{\mathsf{u}}\tfrac{1}{\beta^t}\|\mathbb{u}_n^{t+1}\|_2^2 \leq C_{\mathsf{u}} \cdot \{(\chi - \underline{\lambda}/\overline{\lambda})^2 + \chi\tau\} \cdot 2\overline{\lambda}\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 + C_{\mathsf{u}}\Theta_{\mathsf{x}}^t - C_{\mathsf{u}}\Theta_{\mathsf{x}}^{t+1}. \tag{64}$$

We define $\Theta_{\mathsf{z}}^t \triangleq \Theta_{\mathsf{z}}^t + \Theta_{\mathsf{u}}^t + C_{\mathsf{u}}\Theta_{\mathsf{x}}^t$. Combining Inequalities (63) and (64), we have:

$$\begin{aligned} & \tfrac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 + \Theta_z^{t+1} - \Theta_z^t - \tfrac{C_{\mathsf{w}}}{\beta^t} \\ \leq \; & \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 \cdot C_{\mathsf{u}} \cdot 2\overline{\lambda} \cdot \{(\chi - \underline{\lambda}/\overline{\lambda})^2 + \chi\tau\} \\ \overset{①}{\leq} \; & \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 \cdot C_{\mathsf{u}} \cdot 2\overline{\lambda} \cdot \{\chi^2 + \chi\tau\}, \end{aligned}$$

where step ① uses $-\underline{\lambda}/\overline{\lambda} \leq 0$.

□

C.9. Proof of Theorem 3.7

Proof. We define $\boldsymbol{\gamma}_i' \triangleq \boldsymbol{\gamma}_i[1-\epsilon_3]$, and $\boldsymbol{\gamma}_i \triangleq \frac{1}{2}[\boldsymbol{\theta}_i - 1 - (2+\epsilon_1)\boldsymbol{\alpha}_i\boldsymbol{\theta}_i]$ for all $i \in [n]$.

We let $\Theta^t \triangleq \Theta_{\mathsf{z}}^t + \Theta_{\mathsf{o}}^t$, $\mathbf{r}^t \triangleq [\sum_{i=1}^n \mathbf{A}_i\mathbf{x}_i^t] - \mathbf{b}$, and $\mathcal{E}^{t+1} \triangleq \frac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2 + \epsilon_3\sum_{i=1}^n \boldsymbol{\gamma}_i\mathsf{L}_i^t\|\Delta_i^{t+1}\|_2^2$.

First, based on Lemma 3.1, we have:

$$\mathcal{E}^{t+1} + \Theta_{\mathsf{o}}^{t+1} - \Theta_{\mathsf{o}}^t - \tfrac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 \leq (\epsilon_3 - 1)\boldsymbol{\gamma}_n \cdot \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 = -\boldsymbol{\gamma}_n' \cdot \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 \tag{65}$$

Second, using **Part (b)** of Lemma 3.6, we obtain:

$$\tfrac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 + \Theta_{\mathsf{z}}^{t+1} - \Theta_{\mathsf{z}}^t \leq \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 \cdot \tfrac{\overline{\lambda}}{\underline{\lambda}} \cdot 8\delta\sigma \cdot (\chi^2 + \chi\tau). \tag{66}$$

Adding Inequalities (65) and (66) together, we have:

$$\begin{aligned} & \mathcal{E}^{t+1} + \Theta^{t+1} - \Theta^t \\ \leq \; & \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 \cdot \{\tfrac{\overline{\lambda}}{\underline{\lambda}} \cdot 8\delta\sigma \cdot (\chi^2 + \chi\tau) - \boldsymbol{\gamma}_n'\} \\ \overset{①}{\leq} \; & 0, \end{aligned}$$

where step ① step ② use Inequality (10).

□

C.10. Proof of Theorem 4.5

To finish the proof of this theorem, we first prove the following lemma.

Lemma C.1. *We define* $\mathcal{X}^{t+1} \triangleq \beta^t\sum_{i=1}^n \|\mathbf{x}_i^{t+1} - \mathbf{x}_i^t\|_2^2$, *and* $\mathcal{Y}^{t+1} \triangleq \beta^t\sum_{i=1}^n \|\mathbf{y}_i^{t+1} - \mathbf{y}_i^t\|_2^2$. *We have:* $\sum_{t=0}^T \mathcal{X}^{t+1} \geq \frac{1}{16}\sum_{t=1}^T \mathcal{Y}^{t+1}$.

Proof. Initially, we have the following results:

$$\begin{aligned}\beta^t\|\mathbf{y}_i^{t+1}-\mathbf{y}_i^t\|_2^2 &\overset{①}{=} \beta^t\|[\mathbf{x}_i^{t+1}+\boldsymbol{\alpha}_i(\mathbf{x}_i^{t+1}-\mathbf{x}_i^t)]-[\mathbf{x}_i^t+\boldsymbol{\alpha}_i(\mathbf{x}_i^t-\mathbf{x}_i^{t-1})]\|_2^2\\ &= \beta^t\|(1+\boldsymbol{\alpha}_i)(\mathbf{x}_i^{t+1}-\mathbf{x}_i^t)-\boldsymbol{\alpha}_i(\mathbf{x}_i^t-\mathbf{x}_i^{t-1})\|_2^2\\ &\overset{②}{\leq} 2(1+\boldsymbol{\alpha}_i)^2\beta^t\|\mathbf{x}_i^{t+1}-\mathbf{x}_i^t\|_2^2+2\boldsymbol{\alpha}_i^2\beta^{t-1}(1+\xi)\|\mathbf{x}_i^t-\mathbf{x}_i^{t-1}\|_2^2\\ &\overset{③}{\leq} 8\cdot\left(\beta^t\|\mathbf{x}_i^{t+1}-\mathbf{x}_i^t\|_2^2+\beta^{t-1}\|\mathbf{x}_i^t-\mathbf{x}_i^{t-1}\|_2^2\right),\end{aligned} \tag{67}$$

where step ① uses the update rule for $\mathbf{y}_i^{t+1}$ that $\mathbf{y}_i^{t+1}=\mathbf{x}_i^{t+1}+\boldsymbol{\alpha}_i(\mathbf{x}_i^{t+1}-\mathbf{x}_i^t)$, $\forall i\in[n]$; step ② uses the inequality $\|\mathbf{a}-\mathbf{b}\|_2^2\leq 2\|\mathbf{a}\|_2^2+2\|\mathbf{b}\|_2^2$, and the fact that $\beta^t\leq\beta^{t-1}(1+\xi)$; step ③ uses $\boldsymbol{\alpha}_i<1$ and $\xi<1$.

Telescoping Inequality (67) over i from 1 to n, we have:

$$8\cdot\{\mathcal{X}^{t+1}+\mathcal{X}^t\}\geq\mathcal{Y}^{t+1}. \tag{68}$$

We derive the following inequalities:

$$\begin{aligned}\textstyle\sum_{t=0}^T\mathcal{X}^{t+1} &= \tfrac{1}{2}\cdot\{\textstyle\sum_{t=0}^T\mathcal{X}^{t+1}+\sum_{t=1}^{T+1}\mathcal{X}^t\}\\ &\geq \tfrac{1}{2}\cdot\{\textstyle\sum_{t=1}^T\mathcal{X}^{t+1}+\sum_{t=1}^{T}\mathcal{X}^t\}\\ &\overset{①}{\geq} \tfrac{1}{2}\cdot\tfrac{1}{8}\cdot\textstyle\sum_{t=1}^T\mathcal{Y}^{t+1}\\ &= \tfrac{1}{16}\textstyle\sum_{t=1}^T\mathcal{Y}^{t+1},\end{aligned}$$

where step ① uses Inequality (68).

□

Now, we proceed to prove the theorem.

Proof. We define $\mathcal{X}^{t+1}\triangleq\beta^t\sum_{i=1}^n\|\mathbf{x}_i^{t+1}-\mathbf{x}_i^t\|_2^2$, $\mathcal{Y}^{t+1}\triangleq\beta^t\sum_{i=1}^n\|\mathbf{y}_i^{t+1}-\mathbf{y}_i^t\|_2^2$.

We define $\varphi_1^t=\mathcal{X}^{t+1}+\mathcal{Y}^{t+1}$, and $\varphi_2^t=\beta^t\|\mathbf{r}^{t+1}\|_2^2$.

We define $c_0\triangleq\epsilon_3\min_{i=1}^n\boldsymbol{\gamma}_i\|\mathbf{A}_i\|$, $c_1\triangleq\frac{c_0}{17}$, and $c_2\triangleq\frac{\xi}{2}$.

We derive the following inequalities:

$$\begin{aligned}C_{\mathsf{p}}\triangleq\Theta^0-\underline{\Theta}+C_{\mathsf{b}}C_{\mathsf{w}} &\overset{①}{\geq} \textstyle\sum_{t=0}^T\mathcal{E}^{t+1}\\ &\overset{②}{=} \textstyle\sum_{t=0}^T\{\frac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2+\epsilon_3\sum_{i=1}^n\boldsymbol{\gamma}_i\mathsf{L}_i^t\|\Delta_i^{t+1}\|_2^2\}\\ &\overset{③}{\geq} \textstyle\sum_{t=0}^T\{\frac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2+c_0\sum_{i=1}^n\beta^t\|\Delta_i^{t+1}\|_2^2\}\\ &= \textstyle\sum_{t=1}^T\{\frac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2+\frac{c_0}{17}\mathcal{X}^{t+1}+\frac{16c_0}{17}\mathcal{X}^{t+1}\}\\ &\overset{④}{\geq} \textstyle\sum_{t=1}^T\{\frac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2+\frac{c_0}{17}\mathcal{X}^{t+1}+\frac{c_0}{17}\mathcal{Y}^{t+1}\}\\ &\overset{⑤}{=} \textstyle\sum_{t=1}^T\{c_1\phi_1^t+c_2\phi_2^t\}\\ &= \textstyle\min_{t=1}^T\{c_1\phi_1^t+c_2\phi_2^t\}\cdot T\\ &\overset{⑥}{\geq} \textstyle\min_{t=1}^T\{\phi_1^t+\phi_2^t\}\cdot\frac{1}{1/c_1+1/c_2}\cdot T,\end{aligned}$$

where step ① uses Theorem 3.5 and Theorem 3.7; step ② uses the definition of $\mathcal{E}^{t+1}$ as shown in Equation (14); step ③ uses the definition of c_0; step ④ uses Lemma C.1; step ⑤ uses the definitions of $\{c_1,c_2,\phi_1^t,\phi_2^t\}$; step ⑥ uses Lemma A.7.

As a result, there exists an index $\bar{t}$ with $1\leq\bar{t}\leq T$ such that

$$\beta^{\bar{t}}\|\mathbf{r}^{\bar{t}+1}\|_2^2+\beta^{\bar{t}}\textstyle\sum_{i=1}^n[\|\mathbf{x}_i^{\bar{t}+1}-\mathbf{x}_i^{\bar{t}}\|_2^2+\|\mathbf{y}_i^{\bar{t}+1}-\mathbf{y}_i^{\bar{t}}\|_2^2]\leq\frac{C_{\mathsf{p}}\cdot\max(1/c_1,1/c_2)}{T}.$$

We conclude that Algorithm 1 finds an ϵ-INP point of Problem 1 in at most T iterations, where $T \leq \lceil \frac{C_{\mathrm{p}} \cdot \max(1/c_1, 1/c_2)}{\epsilon} \rceil = \mathcal{O}(\epsilon^{-1})$.

□

D. Proofs for Section 4

D.1. Proof of Theorem 4.4

Initially, we prove the following important lemma, which establishes the quadratic growth condition for any INP-point.

Lemma D.1. *Let* $\mathsf{L}_i^\star \triangleq L_i + \beta\|\mathbf{A}_i\|_2^2$ *for all* $i \in [n]$. *If* $(\ddot{\mathbf{x}}, \ddot{\mathbf{y}}, \ddot{\mathbf{z}})$ *is an* ϵ-INP point with $\epsilon = 0$, it holds that for all $\Delta_i \in \mathbb{R}^{\mathbf{d}_i \times 1}$: $\mathcal{L}(\ddot{\mathbf{x}}, \ddot{\mathbf{z}}; \beta) - \mathcal{L}(\ddot{\mathbf{x}}_1 + \Delta_1, \ldots, \ddot{\mathbf{x}}_n + \Delta_n, \ddot{\mathbf{z}}; \beta) \leq \sum_{i=1}^n \frac{L_i + \boldsymbol{\theta}_i \mathsf{L}_i^\star}{2} \|\Delta_i\|_2^2$.

Proof. We let $i \in [n]$. For any $\mathbf{x}$, we define $\mathbf{A}\mathbf{x} \triangleq \sum_{i=1}^n \mathbf{A}_i \mathbf{x}_i$. We let $\Delta_i \in \mathbb{R}^{\mathbf{d}_i \times 1}$.

Given $f_i(\mathbf{x}_i)$ is L_i-smooth, we have the following inequality (cf. Lemma 1.2.3 in (Nesterov, 2003)):

$$\forall \mathbf{x}_i, \dot{\mathbf{x}}_i, \ | - f_i(\mathbf{x}_i) + f_i(\dot{\mathbf{x}}_i) + \langle \nabla f_i(\dot{\mathbf{x}}_i), \mathbf{x}_i - \dot{\mathbf{x}}_i \rangle | \leq \tfrac{L_i}{2} \|\mathbf{x}_i - \dot{\mathbf{x}}_i\|_2^2.$$

Applying the inequality above with $\mathbf{x}_i = \ddot{\mathbf{x}}_i + \Delta_i$ and $\dot{\mathbf{x}}_i = \ddot{\mathbf{x}}$, we have:

$$\begin{aligned} f_i(\ddot{\mathbf{x}}_i) - f_i(\ddot{\mathbf{x}}_i + \Delta_i) &\leq \langle (\ddot{\mathbf{x}}_i + \Delta_i) - \ddot{\mathbf{x}}_i, -\nabla f_i(\ddot{\mathbf{x}}_i) \rangle + \tfrac{L_i}{2} \|(\ddot{\mathbf{x}}_i + \Delta_i) - \ddot{\mathbf{x}}_i\|_2^2 \\ &= -\langle \Delta_i, \nabla f_i(\ddot{\mathbf{x}}_i) \rangle + \tfrac{L_i}{2} \|\Delta_i\|_2^2. \end{aligned} \tag{69}$$

Using the optimality of the $\mathbf{x}$-subproblem, we have: $h_i(\ddot{\mathbf{x}}_i) + \langle \ddot{\mathbf{x}}_i, \nabla f_i(\ddot{\mathbf{x}}_i) + \mathbf{A}^\mathsf{T} \ddot{\mathbf{z}} \rangle + \frac{\boldsymbol{\theta}_i \mathsf{L}_i^\star}{2} \|\ddot{\mathbf{x}}_i - \ddot{\mathbf{x}}_i\|_2^2 \leq h_i(\mathbf{x}_i) + \langle \mathbf{x}_i, \nabla f_i(\ddot{\mathbf{x}}_i) + \mathbf{A}^\mathsf{T} \ddot{\mathbf{z}} \rangle + \frac{\boldsymbol{\theta}_i \mathsf{L}_i^\star}{2} \|\mathbf{x}_i - \ddot{\mathbf{x}}_i\|_2^2, \forall \mathbf{x}_i$. Letting $\mathbf{x}_i = \ddot{\mathbf{x}}_i + \Delta_i$, we have:

$$\begin{aligned} & h_i(\ddot{\mathbf{x}}_i) - h_i(\ddot{\mathbf{x}}_i + \Delta_i) \\ \leq\ & \langle (\ddot{\mathbf{x}}_i + \Delta_i) - \ddot{\mathbf{x}}_i, \nabla f_i(\ddot{\mathbf{x}}_i) + \mathbf{A}^\mathsf{T} \ddot{\mathbf{z}} \rangle + \tfrac{\boldsymbol{\theta}_i \mathsf{L}_i^\star}{2} \|(\ddot{\mathbf{x}}_i + \Delta_i) - \ddot{\mathbf{x}}_i\|_2^2 \\ =\ & \langle \Delta_i, \nabla f_i(\ddot{\mathbf{x}}_i) + \mathbf{A}^\mathsf{T} \ddot{\mathbf{z}} \rangle + \tfrac{\boldsymbol{\theta}_i \mathsf{L}_i^\star}{2} \|\Delta_i\|_2^2. \end{aligned} \tag{70}$$

Adding (69) and (70) together, we have:

$$f_i(\ddot{\mathbf{x}}_i) + h_i(\ddot{\mathbf{x}}_i) - f_i(\ddot{\mathbf{x}}_i + \Delta_i) - h_i(\ddot{\mathbf{x}}_i + \Delta_i) \leq \tfrac{L_i + \boldsymbol{\theta}_i \mathsf{L}_i^\star}{2} \|\Delta_i\|_2^2 + \langle \mathbf{A}\Delta_i, \ddot{\mathbf{z}} \rangle. \tag{71}$$

Telescoping Inequality (71) over i from 1 to n, we have:

$$\begin{aligned} & \textstyle\sum_{i=1}^n [f_i(\ddot{\mathbf{x}}_i) + h_i(\ddot{\mathbf{x}}_i)] - \sum_{i=1}^n [f_i(\ddot{\mathbf{x}}_i + \Delta_i) + h_i(\ddot{\mathbf{x}}_i + \Delta_i)] - \sum_{i=1}^n \langle \mathbf{A}_i \Delta_i, \ddot{\mathbf{z}} \rangle \\ \leq\ & \textstyle\sum_{i=1}^n \frac{L_i + \boldsymbol{\theta}_i \mathsf{L}_i^\star}{2} \|\Delta_i\|_2^2. \end{aligned} \tag{72}$$

We derive the following inequalities:

$$\begin{aligned} & \mathcal{L}(\ddot{\mathbf{x}}, \ddot{\mathbf{z}}; \beta) - \mathcal{L}(\ddot{\mathbf{x}}_1 + \Delta_1, \ldots, \ddot{\mathbf{x}}_n + \Delta_n, \ddot{\mathbf{z}}; \beta) \\ \overset{①}{=}\ & \textstyle\sum_{i=1}^n [h_i(\ddot{\mathbf{x}}_i) + f_i(\ddot{\mathbf{x}}_i)] + \langle \mathbf{A}\ddot{\mathbf{x}} - \mathbf{b}, \mathbf{z} \rangle + \frac{\beta}{2} \|\mathbf{A}\ddot{\mathbf{x}} - \mathbf{b}\|_2^2 \\ & \textstyle -(\sum_{i=1}^n [h_i(\ddot{\mathbf{x}}_i + \Delta_i) + f_i(\ddot{\mathbf{x}}_i + \Delta_i)] + \langle \mathbf{A}\mathbf{x} + [\sum_{j=1}^n \mathbf{A}_i \Delta_i] - \mathbf{b}, \mathbf{z} \rangle + \frac{\beta}{2} \|\mathbf{A}\ddot{\mathbf{x}} + [\sum_{j=1}^n \mathbf{A}_i \Delta_i] - \mathbf{b}\|_2^2) \\ \overset{②}{=}\ & \textstyle\sum_{i=1}^n [h_i(\ddot{\mathbf{x}}_i) + f_i(\ddot{\mathbf{x}}_i)] - (\sum_{i=1}^n [h_i(\ddot{\mathbf{x}}_i + \Delta_i) + f_i(\ddot{\mathbf{x}}_i + \Delta_i)] + \langle [\sum_{j=1}^n \mathbf{A}_i \Delta_i], \mathbf{z} \rangle + \frac{\beta}{2} \|[\sum_{j=1}^n \mathbf{A}_i \Delta_i]\|_2^2) \\ \overset{③}{\leq}\ & \textstyle\sum_{i=1}^n \frac{L_i + \boldsymbol{\theta}_i \mathsf{L}_i^\star}{2} \|\Delta_i\|_2^2 - \frac{\beta}{2} \|[\sum_{j=1}^n \mathbf{A}_i \Delta_i]\|_2^2 \\ \leq\ & \textstyle\sum_{i=1}^n \frac{L_i + \boldsymbol{\theta}_i \mathsf{L}_i^\star}{2} \|\Delta_i\|_2^2, \end{aligned}$$

where step ① uses the definition of $\mathcal{L}(\mathbf{x}, \mathbf{z}; \beta) \triangleq \sum_{i=1}^n [h_i(\mathbf{x}_i) + f_i(\mathbf{x}_i)] + \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{z} \rangle + \frac{\beta}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$; step ② uses $\mathbf{A}\mathbf{x} = \mathbf{b}$; step ③ uses Inequality (72).

□

Now, we proceed to prove the theorem.

For any $\mathbf{x}$, we denote $\mathbf{A}\mathbf{x} \triangleq \sum_{j=1}^{n} \mathbf{A}_j\mathbf{x}_j$. We let $\mathsf{L}_i^\star \triangleq L_i + \beta\|\mathbf{A}_i\|_2^2$ for all $i \in [n]$. We only consider $\epsilon = 0$.

Firstly, using the definition of INP-point, we have for all $i \in [n]$ that:

$$\ddot{\mathbf{x}}_i^+ = \ddot{\mathbf{x}}_i,\ \ddot{\mathbf{y}}_i^+ = \ddot{\mathbf{y}}_i, \text{ and } \mathbf{A}\ddot{\mathbf{x}}^+ = \mathbf{b}. \tag{73}$$

Secondly, we have the following equalities:

$$\ddot{\mathbf{x}}_i^+ - \ddot{\mathbf{y}}_i \overset{①}{=} \ddot{\mathbf{x}}_i^+ - \ddot{\mathbf{y}}_i^+ \overset{②}{=} -\boldsymbol{\alpha}_i(\ddot{\mathbf{x}}_i^+ - \ddot{\mathbf{x}}_i) \overset{③}{=} \mathbf{0}, \tag{74}$$

where step ① uses the second equality in (73); step ② uses the update rule $\ddot{\mathbf{y}}_i^+ = \ddot{\mathbf{x}}_i^+ + \boldsymbol{\alpha}_i(\ddot{\mathbf{x}}_i^+ - \ddot{\mathbf{x}}_i)$ for all $i \in [n]$; step ③ uses the first equality in (73).

Thirdly, we notice that $\mathbf{x}_i^+$ is the gloabl optimal solution of the following nonconvex problem:

$$\ddot{\mathbf{x}}_i^+ \in \arg\min_{\mathbf{x}_i}\ h_i(\mathbf{x}_i) + \tfrac{1}{2}\boldsymbol{\theta}_i\mathsf{L}_i^t\|\mathbf{x}_i - \ddot{\mathbf{y}}_i\|_2^2 + \langle \mathbf{x}_i - \ddot{\mathbf{x}}_i, \nabla_{\mathbf{x}_i} G(\ddot{\mathbf{x}}_{[1,i-1]}^+, \ddot{\mathbf{x}}_{[i,n]}, \ddot{\mathbf{z}}; \beta)\rangle, \tag{75}$$

where $\nabla_{\mathbf{x}_i} G(\ddot{\mathbf{x}}_{[1,i-1]}^+, \ddot{\mathbf{x}}_{[i,n]}, \ddot{\mathbf{z}}; \beta)$ represents the gradient of the function $G(\ddot{\mathbf{x}}_{[1,i-1]}^+, \ddot{\mathbf{x}}_i, \ddot{\mathbf{x}}_{[i+1,n]}, \ddot{\mathbf{z}})$ *w.r.t.* $\mathbf{x}_i$ at the point $\ddot{\mathbf{x}}_i$ which can be computed as:

$$\nabla_{\mathbf{x}_i} G(\ddot{\mathbf{x}}_{[1,i-1]}^+, \ddot{\mathbf{x}}_{[i,n]}, \ddot{\mathbf{z}}; \beta) = \nabla_{\mathbf{x}_i} f_i(\ddot{\mathbf{x}}_i) + \mathbf{A}_i^\mathsf{T}\ddot{\mathbf{z}} + \beta\mathbf{A}_i^\mathsf{T}([\textstyle\sum_{j=1}^{i-1}\mathbf{A}_j\ddot{\mathbf{x}}_j^+] + [\sum_{j=i}^{n}\mathbf{A}_j\ddot{\mathbf{x}}_j] - \mathbf{b}). \tag{76}$$

Fourth, using the following first-order optimality condition for $\ddot{\mathbf{x}}_i^+$, we have:

$$\mathbf{0} \in \partial h_i(\ddot{\mathbf{x}}_i^+) + \boldsymbol{\theta}_i\mathsf{L}_i^t(\ddot{\mathbf{x}}_i^+ - \ddot{\mathbf{y}}_i) + \nabla_{\mathbf{x}_i} G(\ddot{\mathbf{x}}_{[1,i-1]}^+, \ddot{\mathbf{x}}_{[i,n]}, \ddot{\mathbf{z}}; \beta). \tag{77}$$

(a) We first show that any INP-point must be a C-point, while the reverse is not necessarily true. We derive the following results:

$$\begin{aligned}
& \|\ddot{\mathbf{h}} + \nabla_{\mathbf{x}_i} f_i(\ddot{\mathbf{x}}_i) + \mathbf{A}_i^\mathsf{T}\ddot{\mathbf{z}}\|,\ \forall \ddot{\mathbf{h}} \in \partial h_i(\ddot{\mathbf{x}}_i) \\
\overset{①}{\leq}\ & \|\beta\mathbf{A}_i^\mathsf{T}(\mathbf{A}\ddot{\mathbf{x}}^+ - \mathbf{b})\| + \|\ddot{\mathbf{h}} + \nabla_{\mathbf{x}_i} f_i(\ddot{\mathbf{x}}_i) + \mathbf{A}_i^\mathsf{T}\ddot{\mathbf{z}} + \beta\mathbf{A}_i^\mathsf{T}(\mathbf{A}\ddot{\mathbf{x}}^+ - \mathbf{b})\| \\
=\ & \|\beta\mathbf{A}_i^\mathsf{T}(\mathbf{A}\ddot{\mathbf{x}}^+ - \mathbf{b})\| + \|\ddot{\mathbf{h}} + \nabla_{\mathbf{x}_i} G(\ddot{\mathbf{x}}_{[1,i-1]}^+, \ddot{\mathbf{x}}_{[i,n]}, \ddot{\mathbf{z}}; \beta) + \beta\{\mathbf{A}_i^\mathsf{T}\textstyle\sum_{j=i}^{n}(\mathbf{A}_j\ddot{\mathbf{x}}_j^+ - \mathbf{A}_j\ddot{\mathbf{x}}_j)\}\| \\
\overset{②}{=}\ & \|\beta\mathbf{A}_i^\mathsf{T}(\mathbf{A}\ddot{\mathbf{x}}^+ - \mathbf{b})\| + \|\ddot{\mathbf{h}} + \nabla_{\mathbf{x}_i} G(\ddot{\mathbf{x}}_{[1,i-1]}^+, \ddot{\mathbf{x}}_{[i,n]}, \ddot{\mathbf{z}}; \beta)\| + \beta\|\mathbf{A}_i^\mathsf{T}\textstyle\sum_{j=i}^{n}(\mathbf{A}_j\ddot{\mathbf{x}}_j^+ - \mathbf{A}_j\ddot{\mathbf{x}}_j)\| \\
\overset{③}{=}\ & \|\beta\mathbf{A}_i^\mathsf{T}(\mathbf{A}\ddot{\mathbf{x}}^+ - \mathbf{b})\| + \|\boldsymbol{\theta}_i\mathsf{L}_i^t(\ddot{\mathbf{x}}_i^+ - \ddot{\mathbf{y}}_i)\| + \beta\|\{\mathbf{A}_i^\mathsf{T}\textstyle\sum_{j=i}^{n}(\mathbf{A}_j\ddot{\mathbf{x}}_j^+ - \mathbf{A}_j\ddot{\mathbf{x}}_j)\}\| \\
\overset{④}{=}\ & 0,
\end{aligned} \tag{78}$$

where step ① and step ② use the triangle inequality; step ③ uses the optimality condition in (77); step ④ uses the third equality in (73), Inequality (74), the first equality in (73), and the fact that both L_i^t and $\boldsymbol{\theta}_i$ are finite.

In view of Equation (78), we conclude that any INP-point must be a C-point. Note that the reverse is not true since the condition in (77) is necessary but not sufficient for the global optimal solution in (75).

(b) We now show that any INP-point must be a D-point, while the reverse is not necessarily true. Since $(\ddot{\mathbf{x}}, \ddot{\mathbf{y}}, \ddot{\mathbf{z}})$ is an INP-point, we have from Lemma D.1:

$$\mathcal{L}(\ddot{\mathbf{x}}, \ddot{\mathbf{z}}; \beta) - \mathcal{L}(\ddot{\mathbf{x}}_1 + \Delta_1, \ldots, \ddot{\mathbf{x}}_n + \Delta_n, \ddot{\mathbf{z}}; \beta) \leq \textstyle\sum_{i=1}^{n} \frac{L_i + \boldsymbol{\theta}_i\mathsf{L}_i^\star}{2}\|\Delta_i\|_2^2. \tag{79}$$

For any $\mathbf{x}_i \in \mathrm{dom}(\mathcal{L}_i)$, we let $\Delta_i = t(\mathbf{x}_i - \dot{\mathbf{x}}_i)$. We obtain:

$$\begin{aligned}
& \lim_{t\to 0}[\mathcal{L}(\dot{\mathbf{x}}_1 + t(\mathbf{x}_1 - \dot{\mathbf{x}}_1), \ldots, \dot{\mathbf{x}}_n + t(\mathbf{x}_n - \dot{\mathbf{x}}_n), \dot{\mathbf{z}}; \beta) - \mathcal{L}(\dot{\mathbf{x}}, \dot{\mathbf{z}}; \beta)] \cdot \tfrac{1}{t} \\
\overset{①}{\geq}\ & \lim_{t\to 0} -\textstyle\sum_{i=1}^{n} \frac{L_i + \boldsymbol{\theta}_i\mathsf{L}_i^\star}{2}\|\Delta_i\|_2^2 \cdot \frac{1}{t} \\
=\ & \lim_{t\to 0} -\textstyle\sum_{i=1}^{n} \frac{L_i + \boldsymbol{\theta}_i\mathsf{L}_i^\star}{2}\|\mathbf{x}_i - \dot{\mathbf{x}}_i\|_2^2 \cdot \frac{t^2}{t} = 0,
\end{aligned}$$

where step ① uses (79). Combining with the fact that $\mathbf{A}\ddot{\mathbf{x}} = \mathbf{b}$, we conclude that any INP-point must be a D-point. The reverse is not true since a D-point is necessary but not sufficient for the global optimal solution in (75).

(c) We now show that any optimal point must be an INP-point. Since $(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}})$ is the optimal solution, for all $\mathbf{x}$ and $\mathbf{z}$ with $\mathbf{A}\bar{\mathbf{x}} = \mathbf{b}$, we have:

$$\begin{aligned} & \textstyle\sum_{i=1}^n h_i(\bar{\mathbf{x}}_i) \\ \leq\ & \textstyle\sum_{i=1}^n h_i(\mathbf{x}_i) + \sum_{i=1}^n [f_i(\mathbf{x}_i) - f_i(\bar{\mathbf{x}}_i)],\ \forall \mathbf{x} \\ \overset{①}{\leq}\ & \textstyle\sum_{i=1}^n h_i(\mathbf{x}_i) + \sum_{i=1}^n [\langle \mathbf{x}_i - \bar{\mathbf{x}}_i, \nabla f_i(\bar{\mathbf{x}}_i)\rangle + \frac{\boldsymbol{\theta}_i \mathsf{L}_i^\star}{2}\|\mathbf{x}_i - \bar{\mathbf{x}}_i\|_2^2] \\ \overset{②}{\leq}\ & \textstyle\sum_{i=1}^n h_i(\mathbf{x}_i) + \sum_{i=1}^n [\langle \mathbf{x}_i - \bar{\mathbf{x}}_i, \nabla f_i(\bar{\mathbf{x}}_i) + \beta \mathbf{A}_i^\mathsf{T}(\mathbf{A}\bar{\mathbf{x}} - \mathbf{b})\rangle + \frac{\boldsymbol{\theta}_i \mathsf{L}_i^\star}{2}\|\mathbf{x}_i - \bar{\mathbf{x}}_i\|_2^2], \end{aligned} \tag{80}$$

where step ① uses the Lipschitz continuity of $f_i(\mathbf{x}_i)$ for all $i \in [n]$; step ② uses the fact that $\mathbf{A}\bar{\mathbf{x}} = \mathbf{b}$ for any optimal solution $(\bar{\mathbf{x}}, \bar{\mathbf{z}})$. From Problem (80), we have for all $i \in [n]$:

$$\bar{\mathbf{x}}_i \in \arg\min_{\mathbf{x}_i} h_i(\mathbf{x}_i) + \langle \mathbf{x}_i, \nabla f_i(\bar{\mathbf{x}}_i) + \beta \mathbf{A}_i^\mathsf{T}(\mathbf{A}\bar{\mathbf{x}} - \mathbf{b})\rangle + \tfrac{\boldsymbol{\theta}_i \mathsf{L}_i^\star}{2}\|\mathbf{x}_i - \bar{\mathbf{x}}_i\|_2^2. \tag{81}$$

Problem (81) essentially coincides with Problem (75) with $\bar{\mathbf{x}} = \ddot{\mathbf{x}}_i = \ddot{\mathbf{x}}_i^+ = \ddot{\mathbf{y}}_i = \ddot{\mathbf{y}}_i^+$. Therefore, any optimal point $(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{z}})$ must be an INP-point.

E. Proofs for Section 5

E.1. Proof of Lemma 5.2

We begin by presenting the following six useful lemmas: Lemma E.1, Lemma E.2, Lemma E.3, Lemma E.4, Lemma E.5, and Lemma E.6.

Lemma E.1. *For* Condition $\boxed{\mathbb{I}}$, we define the Lyapunov function as: $\Theta(\mathbf{x}, \mathbf{z}, \mathbf{x}', \mathbf{x}''; \beta, \beta') \triangleq \frac{C_\mathsf{a}}{\beta}\|\sigma\beta\mathbf{A}_n^\mathsf{T}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_2^2 + \mathcal{L}(\mathbf{x}, \mathbf{z}; \beta) + \frac{C_\mathsf{u}}{\beta}\|\mathbb{H}(\mathbf{x}_n - \mathbf{x}_n') - \boldsymbol{\eta}_n(\mathbf{x}_n' - \mathbf{x}_n'')\|_2^2 + \{\frac{1}{2}\sum_{i=1}^n \boldsymbol{\eta}_i\|\mathbf{x}_i - \mathbf{x}_i'\|_2^2\} + C_\mathsf{x}\mathsf{L}_n\|\mathbf{x}_n - \mathbf{x}_n'\|_2^2$, where $\mathcal{L}(\mathbf{x}, \mathbf{z}; \beta) \triangleq \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{z}\rangle + \frac{\beta}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \sum_{i=1}^n [f_i(\mathbf{x}_i) + h_i(\mathbf{x}_i)]$, $\mathsf{L}_i \triangleq L_i + \beta\|\mathbf{A}_i\|_2^2$, $\mathsf{L}_i' \triangleq L_i + \beta'\|\mathbf{A}_i\|_2^2$, $\mathbb{H} \triangleq \boldsymbol{\theta}_n\mathsf{L}_n'\mathbf{I} - \beta'\mathbf{A}_n^\mathsf{T}\mathbf{A}_n$, $\boldsymbol{\eta}_i \triangleq \boldsymbol{\theta}_i\boldsymbol{\alpha}_i\mathsf{L}_i'$, $\forall i \in [n]$. We have:

$$(\mathbf{d}_{\mathbf{x}_1}, \ldots, \mathbf{d}_{\mathbf{x}_n}, \mathbf{d}_\mathbf{z}, \mathbf{d}_{\mathbf{x}_1'}, \ldots, \mathbf{d}_{\mathbf{x}_n'}, \mathbf{d}_{\mathbf{x}_1''}, \ldots, \mathbf{d}_{\mathbf{x}_n''}) \in \partial\Theta(\mathbf{x}^t, \mathbf{z}^t, \mathbf{x}^{t-1}, \mathbf{x}^{t-2}; \beta^t, \beta^{t-1}), \tag{82}$$

where

$$\mathbf{d}_{\mathbf{x}_i} \triangleq \begin{cases} \mathfrak{q}_i^t, & i \neq n; \\ \mathfrak{q}_i^t + \mathbb{v}_n^t, & i = n. \end{cases}, \quad \mathbf{d}_\mathbf{z} \triangleq \mathbf{A}\mathbf{x}^t - \mathbf{b},$$

$$\mathbf{d}_{\mathbf{x}_i'} \triangleq \begin{cases} -\boldsymbol{\eta}_i^{t-1}\Delta_i^t, & i \neq n; \\ -\boldsymbol{\eta}_i^{t-1}\Delta_i^t + \mathbb{y}_n^t, & i = n. \end{cases}, \quad \text{and} \quad \mathbf{d}_{\mathbf{x}_i''} \triangleq \begin{cases} \mathbf{0}, & i \neq n; \\ \mathbf{z}_n^t, & i = n. \end{cases}.$$

Here, $\mathfrak{q}_i^t \triangleq \nabla f_i(\mathbf{x}_i^t) + \nabla h_i(\mathbf{x}_i^t) + \mathbf{A}_i^\mathsf{T}\mathbf{z}^t + \beta^t\mathbf{A}_i^\mathsf{T}\mathbf{r}^t + 2C_\mathsf{a}\beta^t\sigma^2\mathbf{A}_i^\mathsf{T}\mathbf{A}_n\mathbf{A}_n^\mathsf{T}\mathbf{r}^t + \boldsymbol{\eta}_i^{t-1}\Delta_i^t$,
$\mathbb{v}_n^t \triangleq \frac{2C_\mathsf{u}}{\beta^t}[\mathbb{H}^{t-1}]^\mathsf{T}\mathbb{u}_n^t + 2C_\mathsf{x}\mathsf{L}_n^t\Delta_n^t$, $\mathbb{y}_n^t \triangleq -\frac{2C_\mathsf{u}}{\beta^t}\cdot[\mathbb{H}^{t-1} + \boldsymbol{\eta}_n^{t-1}\mathbf{I}]^\mathsf{T}\mathbb{u}_n^t - 2C_\mathsf{x}\mathsf{L}_n^t\Delta_n^t$,
$\mathbf{z}_n^t \triangleq \frac{2C_\mathsf{u}}{\beta^t}\cdot\boldsymbol{\eta}_n^{t-1}\mathbb{u}_n^t$, and $\mathbb{u}_n^t \triangleq \mathbb{H}^{t-1}(\mathbf{x}_n^t - \mathbf{x}_n^{t-1}) - \boldsymbol{\eta}_n^{t-1}(\mathbf{x}_n^{t-1} - \mathbf{x}_n^{t-2}) \triangleq \mathbb{H}^{t-1}\Delta_n^t - \boldsymbol{\eta}_n^{t-1}\Delta_n^{t-1}$.
Furthermore, $\mathsf{L}_i^t \triangleq L_i + \beta^t\|\mathbf{A}_i\|_2^2$, $\mathbb{H}^{t-1} \triangleq \boldsymbol{\theta}_n\mathsf{L}_n^{t-1}\mathbf{I} - \beta^{t-1}\mathbf{A}_n^\mathsf{T}\mathbf{A}_n$, and $\boldsymbol{\eta}_i^{t-1} \triangleq \boldsymbol{\theta}_i\boldsymbol{\alpha}_i\mathsf{L}_i^{t-1}$ for all $i \in [n]$.

Proof. These results are based on very basic deductions. □

Lemma E.2. *For* Condition $\boxed{\mathbb{A}}$, we define the Lyapunov function as: $\Theta(\mathbf{x}, \mathbf{z}, \mathbf{x}', \mathbf{x}''; \beta) \triangleq \frac{C_\mathsf{a}}{\beta}\|\sigma\beta\mathbf{A}_n^\mathsf{T}(\mathbf{A}\mathbf{x}-\mathbf{b}) + \sigma(\mathbb{H}(\mathbf{x}_n - \mathbf{x}_n') - \boldsymbol{\eta}_n(\mathbf{x}_n - \mathbf{x}_n'))\|_2^2 + \mathcal{L}(\mathbf{x}, \mathbf{z}; \beta) + \frac{C_\mathsf{u}}{\beta}\|\mathbb{H}(\mathbf{x}_n - \mathbf{x}_n') - \boldsymbol{\eta}_n(\mathbf{x}_n' - \mathbf{x}_n'')\|_2^2 + \{\frac{1}{2}\sum_{i=1}^n \boldsymbol{\eta}_i\|\mathbf{x}_i - \mathbf{x}_i'\|_2^2\} + C_\mathsf{x}\mathsf{L}_n\|\mathbf{x}_n - \mathbf{x}_n'\|_2^2$, where $\mathcal{L}(\mathbf{x}, \mathbf{z}; \beta) \triangleq \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{z}\rangle + \frac{\beta}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \sum_{i=1}^n [f_i(\mathbf{x}_i) + h_i(\mathbf{x}_i)]$, $\mathsf{L}_i \triangleq L_i + \beta\|\mathbf{A}_i\|_2^2$, $\mathsf{L}_i' \triangleq L_i + \beta'\|\mathbf{A}_i\|_2^2$, $\mathbb{H} \triangleq \boldsymbol{\theta}_n\mathsf{L}_n'\mathbf{I} - \beta'\mathbf{A}_n^\mathsf{T}\mathbf{A}_n$, $\boldsymbol{\eta}_i \triangleq \boldsymbol{\theta}_i'\boldsymbol{\alpha}_i\mathsf{L}_i$, $\forall i \in [n]$. We have:

$$(\mathbf{d}_{\mathbf{x}_1}, \ldots, \mathbf{d}_{\mathbf{x}_n}, \mathbf{d}_\mathbf{z}, \mathbf{d}_{\mathbf{x}_1'}, \ldots, \mathbf{d}_{\mathbf{x}_n'}, \mathbf{d}_{\mathbf{x}_1''}, \ldots, \mathbf{d}_{\mathbf{x}_n''}) \in \partial\Theta(\mathbf{x}^t, \mathbf{z}^t, \mathbf{x}^{t-1}, \mathbf{x}^{t-2}; \beta^t, \beta^{t-1}), \tag{83}$$

where

$$\mathbf{d}_{\mathbf{x}_i} \triangleq \begin{cases} \mathfrak{q}_i^t, & i \neq n; \\ \mathfrak{q}_i^t + \mathbb{v}_n^t, & i = n. \end{cases}, \quad \mathbf{d}_{\mathbf{z}} \triangleq \mathbf{A}\mathbf{x}^t - \mathbf{b},$$
$$\mathbf{d}_{\mathbf{x}'_i} \triangleq \begin{cases} -\boldsymbol{\eta}_i^{t-1}\Delta_i^t, & i \neq n; \\ -\boldsymbol{\eta}_i^{t-1}\Delta_i^t + \mathbb{y}_n^t, & i = n. \end{cases}, \quad \text{and} \quad \mathbf{d}_{\mathbf{x}''_i} \triangleq \begin{cases} \mathbf{0}, & i \neq n; \\ \mathbb{z}_n^t, & i = n. \end{cases}.$$

Here, $\mathfrak{q}_i^t \triangleq 2C_{\mathsf{a}}\mathbf{A}_i^{\mathsf{T}}\mathbf{A}_n\sigma^2\{\beta^t\mathbf{A}_n^{\mathsf{T}}\mathbf{r}^t + \mathbb{u}_n^t\} + \nabla f_i(\mathbf{x}_i^t) + \nabla h_i(\mathbf{x}_i^t) + \mathbf{A}_i^{\mathsf{T}}\mathbf{z}^t + \beta^t\mathbf{A}_i^{\mathsf{T}}\mathbf{r}^t + \boldsymbol{\eta}_i^{t-1}\Delta_i^t$,
$\mathbb{v}_n^t \triangleq \frac{2C_{\mathsf{a}}\sigma^2}{\beta^t}[\mathbb{H}^{t-1}]^{\mathsf{T}}\{\beta^t\mathbf{A}_n^{\mathsf{T}}\mathbf{r}^t + \mathbb{u}_n^t\} + \frac{2C_{\mathsf{u}}}{\beta^t}[\mathbb{H}^{t-1}]^{\mathsf{T}}\mathbb{u}_n^t + 2C_{\mathsf{x}}\mathsf{L}_n^t\Delta_n^t$,
$\mathbb{y}_n^t \triangleq -\frac{2C_{\mathsf{a}}}{\beta^t}(\sigma\mathbb{H}^{t-1} + \sigma\boldsymbol{\eta}_n^{t-1}\mathbf{I})^{\mathsf{T}}(\sigma\beta^t\mathbf{A}_n^{\mathsf{T}}\mathbf{r}^t + \sigma\mathbb{u}_n^t) - \frac{2C_{\mathsf{u}}}{\beta^t}[\mathbb{H}^{t-1} + \boldsymbol{\eta}_n^{t-1}\mathbf{I}]^{\mathsf{T}}\mathbb{u}_n^t - 2C_{\mathsf{x}}\mathsf{L}_n^t\Delta_n^t$,
$\mathbb{z}_n^t \triangleq \frac{2C_{\mathsf{a}}}{\beta^t}\sigma\boldsymbol{\eta}_n^{t-1}(\sigma\beta^t\mathbf{A}_n^{\mathsf{T}}\mathbf{r}^t + \sigma\mathbb{u}_n^t) + \frac{2C_{\mathsf{u}}}{\beta^t}\boldsymbol{\eta}_n^{t-1}\mathbb{u}_n^t$.
$\mathbb{u}_n^t \triangleq \mathbb{H}^{t-1}(\mathbf{x}_n^t - \mathbf{x}_n^{t-1}) - \boldsymbol{\eta}_n^{t-1}(\mathbf{x}_n^{t-1} - \mathbf{x}_n^{t-2}) \triangleq \mathbb{H}^{t-1}\nabla_n^t - \boldsymbol{\eta}_n^{t-1}\nabla_n^{t-1}$.
Furthermore, $\mathsf{L}_i^t \triangleq L_i + \beta^t\|\mathbf{A}_i\|_2^2$, $\mathbb{H}^{t-1} \triangleq \boldsymbol{\theta}_n\mathsf{L}_n^{t-1}\mathbf{I} - \beta^{t-1}\mathbf{A}_n^{\mathsf{T}}\mathbf{A}_n$, and $\boldsymbol{\eta}_i^{t-1} \triangleq \boldsymbol{\theta}_i\boldsymbol{\alpha}_i\mathsf{L}_i^{t-1}$ for all $i \in [n]$.

Proof. These results are based on very basic deductions. □

Lemma E.3. *We denote $\lambda_\star \triangleq \max_{i=1}^n \|\mathbf{A}_i\|_2^2$ and $\theta_\star \triangleq \max(\boldsymbol{\theta})$. For all $t \geq 0$, it holds that:*

$$\tfrac{1}{\beta^t}L_i \leq \lambda_\star, \quad \tfrac{1}{\beta^t}\mathsf{L}_i^t \leq 2\lambda_\star, \quad \tfrac{1}{\beta^t}\|\mathbb{H}^t\| \leq 2\theta_\star\lambda_\star, \quad \tfrac{1}{\beta^t}\boldsymbol{\eta}_i^t \leq 2\theta_\star\lambda_\star.$$

Proof. For all $i \in [n]$, we have:

$$\tfrac{1}{\beta^t}\cdot L_i \overset{①}{\leq} \overline{\lambda}\epsilon_3 \overset{④}{\leq} \lambda_\star.$$
$$\tfrac{1}{\beta^t}\cdot \mathsf{L}_i^t \overset{②}{\leq} \overline{\lambda}(1+\epsilon_3) \overset{⑤}{\leq} 2\lambda_\star.$$
$$\tfrac{1}{\beta^t}\cdot \|\mathbb{H}^t\| \overset{③}{\leq} (1+\epsilon_3)\boldsymbol{\theta}_n\overline{\lambda} \overset{⑥}{\leq} 2\theta_\star\lambda_\star$$
$$\tfrac{1}{\beta^t}\cdot \boldsymbol{\eta}_i^t = \tfrac{\mathsf{L}_i^t}{\beta^t}\cdot\boldsymbol{\theta}_i\boldsymbol{\alpha}_i \leq (1+\epsilon_3)\overline{\lambda}\cdot\boldsymbol{\theta}_i\boldsymbol{\alpha}_i \leq 2\theta_\star\lambda_\star$$

Here, step ① uses $L_i \leq \beta^0\epsilon_3\overline{\lambda} \leq \beta^t\epsilon_3\overline{\lambda}$; step ② uses Inequality (51); step ③ uses Inequality (52); step ④, step ⑤, and step ⑥ use $\epsilon_3 \leq 1$, $\boldsymbol{\theta}_n \leq \theta_\star$, and $\overline{\lambda} \leq \lambda_\star$.

□

Lemma E.4. *For any $i \in [n]$, we define $\mathbb{u}_i^t \triangleq \boldsymbol{\theta}_i\mathsf{L}_i^{t-1}(\mathbf{x}_i^t - \mathbf{x}_i^{t-1} - \boldsymbol{\alpha}_i(\mathbf{x}_i^{t-1} - \mathbf{x}_i^{t-2})) - \beta^{t-1}\mathbf{A}_i^{\mathsf{T}}[\sum_{j=i}^n \mathbf{A}_j(\mathbf{x}_j^t - \mathbf{x}_j^{t-1})]$, and we let $\mathbb{o}_i^t \in \nabla f_i(\mathbf{x}_i^t) + \partial h_i(\mathbf{x}_i^t) + \mathbf{A}_i^{\mathsf{T}}\mathbf{z}^t + \beta^t\mathbf{A}_i^{\mathsf{T}}\mathbf{r}^t$. We have:*

$$\begin{aligned} \tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\mathbb{u}_i^t\| &\leq c_1\cdot\textstyle\sum_{i=1}^n\|\Delta_i^{t-1}\| + c_1\cdot\textstyle\sum_{i=1}^n\|\Delta_i^t\| \\ \tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\mathbb{o}_i^t\| &\leq c_1\cdot\textstyle\sum_{i=1}^n\|\Delta_i^{t-1}\| + c_1\cdot\textstyle\sum_{i=1}^n\|\Delta_i^t\| + c_2\cdot\|\mathbf{r}^t\| \\ \tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\boldsymbol{\eta}_i^{t-1}\Delta_i^t\| &\leq c_3\cdot\textstyle\sum_{i=1}^n\|\Delta_i^{t-1}\|. \end{aligned} \tag{84}$$

Here, $c_1 \triangleq 2\theta_\star\lambda_\star + n\lambda_\star + \lambda_\star$, $c_2 \triangleq 2\sqrt{\lambda_\star}n$, and $c_3 \triangleq 2\theta_\star\lambda_\star$, where $\lambda_\star \triangleq \max_{i=1}^n\|\mathbf{A}_i\|_2^2$ and $\theta_\star \triangleq \max(\boldsymbol{\theta})$.

Proof. We denote $\Gamma_1 \triangleq \sum_{i=1}^n\|\Delta_i^{t-1}\|$, and $\Gamma_2 \triangleq \sum_{i=1}^n\|\Delta_i^t\|$.

Using the optimality of $\mathbf{x}_i^{t+1}$ as presented in Lemma 3.2, we have:

$$\begin{aligned} & -\partial h_i(\mathbf{x}_i^{t+1}) - \nabla f_i(\mathbf{x}_i^t) - \mathbf{A}_i^{\mathsf{T}}\mathbf{z}^t - \tfrac{1}{\sigma}\mathbf{A}_i^{\mathsf{T}}(\mathbf{z}^{t+1} - \mathbf{z}^t) \\ \ni\ & \boldsymbol{\theta}_i\mathsf{L}_i^t(\mathbf{x}_i^{t+1} - \mathbf{x}_i^t - \boldsymbol{\alpha}_i(\mathbf{x}_i^t - \mathbf{x}_i^{t-1})) - \beta^t\mathbf{A}_i^{\mathsf{T}}[\textstyle\sum_{j=i}^n\mathbf{A}_j(\mathbf{x}_j^{t+1} - \mathbf{x}_j^t)] = \mathbb{u}_i^{t+1}. \end{aligned} \tag{85}$$

(a) We bound the term $\frac{1}{\beta^t}\sum_{i=1}^n \|\mathbb{u}_i^t\|$:

$$\begin{aligned}
&\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\mathbb{u}_i^t\| \\
\overset{①}{=}\ &\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\boldsymbol{\theta}_i \mathsf{L}_i^{t-1}(\mathbf{x}_i^t - \mathbf{x}_i^{t-1} - \boldsymbol{\alpha}_i(\mathbf{x}_i^{t-1} - \mathbf{x}_i^{t-2})) - \beta^{t-1}\mathbf{A}_i^\mathsf{T}[\textstyle\sum_{j=i}^n \mathbf{A}_j(\mathbf{x}_j^t - \mathbf{x}_j^{t-1})]\| \\
\overset{②}{\leq}\ &\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \{\|\boldsymbol{\theta}_i \mathsf{L}_i^{t-1}(\mathbf{x}_i^t - \mathbf{x}_i^{t-1})\| + \|\boldsymbol{\theta}_i \mathsf{L}_i^{t-1}\boldsymbol{\alpha}_i(\mathbf{x}_i^{t-1} - \mathbf{x}_i^{t-2})\| + \|\beta^{t-1}\mathbf{A}_i^\mathsf{T}[\textstyle\sum_{j=i}^n \mathbf{A}_j(\mathbf{x}_j^t - \mathbf{x}_j^{t-1})]\|\} \\
\overset{③}{\leq}\ &\tfrac{\theta_\star \mathsf{L}_i^{t-1}}{\beta^t}\Gamma_1 + \max(\boldsymbol{\alpha})\tfrac{\theta_\star \mathsf{L}_i^{t-1}}{\beta^t}\Gamma_2 + \tfrac{\beta^{t-1}}{\beta^t} n\lambda_\star\Gamma_1 \\
\overset{④}{\leq}\ &\tfrac{\theta_\star \cdot 2\lambda_\star \beta^{t-1}}{\beta^t}\Gamma_1 + \max(\boldsymbol{\alpha})\tfrac{\theta_\star 2\lambda_\star\beta^{t-1}}{\beta^t}\Gamma_2 + \tfrac{\beta^{t-1}}{\beta^t} n\lambda_\star\Gamma_1 \\
\overset{⑤}{\leq}\ &2\theta_\star\lambda_\star\Gamma_1 + 2\theta_\star\lambda_\star\Gamma_2 + n\lambda_\star\Gamma_1 \\
<\ &c_1(\Gamma_1 + \Gamma_2),
\end{aligned} \tag{86}$$

where step ① uses the definition of $\mathbb{u}_i^{t+1}$; step ② uses the norm inequality; step ③ uses the definition of Γ_1 and Γ_2; step ④ uses Lemma E.3; step ⑤ uses $\beta^{t-1} \leq \beta^t$.

(b) We bound the term $\frac{1}{\beta^t}\sum_{i=1}^n \|\mathbb{o}_i^t\|$:

$$\begin{aligned}
&\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\mathbb{O}_i^t\| \\
\overset{①}{=}\ &\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\nabla f_i(\mathbf{x}_i^t) + \partial h_i(\mathbf{x}_i^t) + \mathbf{A}_i^\mathsf{T}\mathbf{z}^t + \beta^t\mathbf{A}_i^\mathsf{T}\mathbf{r}^t\| \\
\overset{②}{=}\ &\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\nabla f_i(\mathbf{x}_i^t) + (1-\tfrac{1}{\sigma})\mathbf{A}_i^\mathsf{T}(\mathbf{z}^t - \mathbf{z}^{t-1}) + \beta^t\mathbf{A}_i^\mathsf{T}\mathbf{r}^t - \mathbb{u}_i^t - \nabla f_i(\mathbf{x}_i^{t-1})\| \\
\overset{③}{=}\ &\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\nabla f_i(\mathbf{x}_i^t) + (\sigma-1)\beta^{t-1}\mathbf{A}_i^\mathsf{T}\mathbf{r}^t + \beta^t\mathbf{A}_i^\mathsf{T}\mathbf{r}^t - \mathbb{u}_i^t - \nabla f_i(\mathbf{x}_i^{t-1})\| \\
\overset{④}{\leq}\ &\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\mathbb{u}_i^t\| + \tfrac{1}{\beta^t}\sum_{i=1}^n \|\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\mathbf{x}_i^{t-1})\| + \tfrac{\beta^{t-1}}{\beta^t}|\sigma-1|\cdot n\sqrt{\lambda_\star}\|\mathbf{r}^{t+1}\| + n\sqrt{\lambda_\star}\|\mathbf{r}^{t+1}\| \\
\overset{⑤}{\leq}\ &\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\mathbb{u}_i^t\| + \tfrac{1}{\beta^t}\sum_{i=1}^n L_i\|\mathbf{x}_i^t - \mathbf{x}_i^{t-1}\| + 2n\sqrt{\lambda_\star}\|\mathbf{r}^{t+1}\| \\
\overset{⑥}{\leq}\ &\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\mathbb{u}_i^t\| + \lambda_\star\Gamma_1 + 2n\sqrt{\lambda_\star}\|\mathbf{r}^{t+1}\| \\
\overset{⑦}{\leq}\ &2\theta_\star\lambda_\star\Gamma_1 + 2\theta_\star\lambda_\star\Gamma_2 + n\lambda_\star\Gamma_1 + \lambda_\star\Gamma_1 + 2n\sqrt{\lambda_\star}\|\mathbf{r}^{t+1}\| \\
\overset{⑧}{\leq}\ &c_1\Gamma_1 + c_1\Gamma_2 + c_2\|\mathbf{r}^{t+1}\|,
\end{aligned}$$

where step ① uses the definition of $\mathbb{o}_i^{t+1}$; step ② uses

$$\partial h_i(\mathbf{x}_i^t) \ni -\mathbb{u}_i^t - \nabla f_i(\mathbf{x}_i^{t-1}) - \mathbf{A}_i^\mathsf{T}\mathbf{z}^{t-1} - \tfrac{1}{\sigma}\mathbf{A}_i^\mathsf{T}(\mathbf{z}^t - \mathbf{z}^{t-1}), \tag{87}$$

which is due to Equation (85); step ③ uses $\mathbf{z}^{t+1} - \mathbf{z}^t = \sigma\beta^t\mathbf{r}^{t+1}$; step ④ uses the norm inequality; step ⑤ uses the fact that $\|\nabla f_i(\mathbf{x}_i^{t+1}) - \nabla f_i(\mathbf{x}_i^t)\| \leq L_i\|\mathbf{x}_i^{t+1} - \mathbf{x}_i^t\|$, and the condition $\sigma \in (0,2)$; step ⑥ uses Lemma E.3; step ⑦ uses Inequality (86); step ⑧ uses $c_1 \triangleq 2\theta_\star\lambda_\star + n\lambda_\star + \lambda_\star$.

(c) We now bound the term $\frac{1}{\beta^t}\sum_{i=1}^n \|\boldsymbol{\eta}_i^{t-1}\Delta_i^t\|$ (as per Lemma E.1):

$$\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\boldsymbol{\eta}_i^{t-1}\Delta_i^t\| \overset{①}{\leq} 2\tfrac{\beta^{t-1}}{\beta^t}\theta_\star\lambda_\star\sum_{i=1}^n\|\Delta_i^t\| \overset{②}{\leq} 2\theta_\star\lambda_\star\sum_{i=1}^n\|\Delta_i^t\| = 2\theta_\star\lambda_\star\Gamma_1 \overset{③}{=} c_3\Gamma_1, \tag{88}$$

where step ① uses Lemma E.3; step ② uses $\beta^{t-1} \leq \beta^t$; step ③ uses the definition of c_3.

□

Lemma E.5. *For* Condition Ⅱ, we define $\{\mathbf{d}_{\mathbf{x}_i}, \mathbf{d}_\mathbf{z}, \mathbf{d}_{\mathbf{x}_i'}, \mathbf{d}_{\mathbf{x}_i''}\}$ as in Lemma E.1. It holds that:

$$\begin{aligned}
\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\mathbf{d}_{\mathbf{x}_i}\| &\leq p_1\sum_{i=1}^n\|\Delta_i^{t-1}\| + s_1\sum_{i=1}^n\|\Delta_i^t\| + u_1\|\mathbf{r}^t\|, \\
\tfrac{1}{\beta^t}\|\mathbf{d}_\mathbf{z}\| &\leq p_2\textstyle\sum_{i=1}^n\|\Delta_i^{t-1}\| + s_2\sum_{i=1}^n\|\Delta_i^t\| + u_2\|\mathbf{r}^t\|, \\
\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\mathbf{d}_{\mathbf{x}_i'}\| &\leq p_3\sum_{i=1}^n\|\Delta_i^{t-1}\| + s_3\sum_{i=1}^n\|\Delta_i^t\| + u_3\|\mathbf{r}^t\|, \\
\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\mathbf{d}_{\mathbf{x}_i''}\| &\leq p_4\sum_{i=1}^n\|\Delta_i^{t-1}\| + s_4\sum_{i=1}^n\|\Delta_i^t\| + u_4\|\mathbf{r}^t\|.
\end{aligned}$$

Here, we define: $p_1 \triangleq c_1 + c_5$, $s_1 \triangleq c_1 + c_3 + c_5 + c_6$, $u_1 \triangleq c_2 + c_4$;
$p_2 \triangleq 0$, $s_2 \triangleq 0$, $u_2 \triangleq 1/\beta^0$;
$p_3 \triangleq 2c_5 + c_7 + c_3$, $s_3 \triangleq 2c_5$, $u_3 \triangleq 0$;
$p_4 \triangleq c_5$, $s_4 \triangleq c_5$, $u_4 \triangleq 0$;
Also, $c_5 \triangleq 4c_1\theta_\star\lambda_\star C_{\mathsf{u}}$, $c_6 \triangleq 4C_{\mathsf{x}}\lambda_\star$, and $c_7 \triangleq 4C_{\mathsf{u}}\lambda_\star$.
Additionally, $\lambda_\star \triangleq \max_{i=1}^n \|\mathbf{A}_i\|_2^2$, $\theta_\star \triangleq \max(\boldsymbol{\theta})$, and $c_4 \triangleq 8nC_{\mathsf{a}}{\lambda_\star}^{3/2}$.
Lastly, note that $\{c_1, c_2, c_3\}$ are defined in Lemma E.4, and $\{C_{\mathsf{u}}, C_{\mathsf{x}}, C_{\mathsf{a}}\}$ are defined in Equation (21).

Proof. For any $i \in [n]$, we define $\mathbb{u}_i^t \triangleq \boldsymbol{\theta}_i \mathsf{L}_i^{t-1}(\mathbf{x}_i^t - \mathbf{x}_i^{t-1} - \boldsymbol{\alpha}_i(\mathbf{x}_i^{t-1} - \mathbf{x}_i^{t-2})) - \beta^{t-1}\mathbf{A}_i^\mathsf{T}[\sum_{j=i}^n \mathbf{A}_j(\mathbf{x}_j^t - \mathbf{x}_j^{t-1})]$, and we let $\mathbb{o}_i^t \in \nabla f_i(\mathbf{x}_i^t) + \partial h_i(\mathbf{x}_i^t) + \mathbf{A}_i^\mathsf{T}\mathbf{z}^t + \beta^t\mathbf{A}_i^\mathsf{T}\mathbf{r}^t$.

We define $\Gamma_1 = \sum_{i=1}^n \|\Delta_i^{t-1}\|$ and $\Gamma_2 = \sum_{i=1}^n \|\Delta_i^t\|$.

We bound the terms $\frac{1}{\beta^t}\sum_{i=1}^n \|\mathbb{q}_i^t\|$ (as per Lemma E.1):

$$\begin{aligned}
\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\mathbb{q}_i^t\| &= \tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \{\|\mathbb{o}_i^t + 2C_{\mathsf{a}}\beta^t\sigma^2\mathbf{A}_i^\mathsf{T}\mathbf{A}_n\mathbf{A}_n^\mathsf{T}\mathbf{r}^t + \boldsymbol{\eta}_i^{t-1}\nabla_i^t\|\} \\
&\overset{①}{\leq} \tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\mathbb{o}_i^t\| + \{8C_{\mathsf{a}}\textstyle\sum_{i=1}^n \|\mathbf{A}_i^\mathsf{T}\mathbf{A}_n\mathbf{A}_n^\mathsf{T}\mathbf{r}^{t+1}\|\} + \tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\boldsymbol{\eta}_i^{t-1}\nabla_i^t\| \\
&\overset{②}{\leq} c_1(\Gamma_1 + \Gamma_2) + c_2\|\mathbf{r}^{t+1}\| + c_4\|\mathbf{r}^{t+1}\| + c_3\Gamma_2, \qquad (89)
\end{aligned}$$

where step ① uses the norm inequality, and the condition $\sigma \leq 2$; step ② uses Lemma E.4, and the definition of $\{c_3, c_4\}$.

We now bound the term $\frac{1}{\beta^t}\|\mathbb{v}_n^t\|$ (as per Lemma E.1):

$$\begin{aligned}
\tfrac{1}{\beta^t}\|\mathbb{v}_n^t\| &= \tfrac{1}{\beta^t}\|\tfrac{2C_{\mathsf{u}}}{\beta^t}[\mathbb{H}^{t-1}]^\mathsf{T}\mathbb{u}_n^t + 2C_{\mathsf{x}}\mathsf{L}_n^t\Delta_n^t\| \\
&\overset{①}{\leq} c_1\{\Gamma_1 + \Gamma_2\}\cdot 2C_{\mathsf{u}}\cdot\tfrac{\|\mathbb{H}^{t-1}\|}{\beta^t} + 2C_{\mathsf{x}}\cdot\tfrac{\mathsf{L}_n^t}{\beta^t}\cdot\textstyle\sum_{i=1}^n\|\Delta_i^t\| \\
&\overset{②}{\leq} c_1\{\Gamma_1 + \Gamma_2\}\cdot 2C_{\mathsf{u}}\cdot 2\theta_\star\lambda_\star\tfrac{\beta^{t-1}}{\beta^t} + 2C_{\mathsf{x}}\cdot 2\lambda_\star\cdot\Gamma_2 \\
&\overset{③}{\leq} (4c_1\theta_\star\lambda_\star C_{\mathsf{u}})\cdot\Gamma_1 + (4c_1\theta_\star\lambda_\star C_{\mathsf{u}} + 4C_{\mathsf{x}}\lambda_\star)\Gamma_2 \\
&\overset{④}{\leq} c_5\Gamma_1 + (c_5 + c_6)\Gamma_2, \qquad (90)
\end{aligned}$$

where step ① uses the norm inequality and Lemma E.4; step ② uses Lemma E.3; step ③ uses $\beta^{t-1} \leq \beta^t$; step ④ uses the definition of $\{c_5, c_6\}$.

We now bound the term $\frac{1}{\beta^t}\|\mathbb{y}_n^t\|$ (as per Lemma E.1):

$$\begin{aligned}
\tfrac{1}{\beta^t}\|\mathbb{y}_n^t\| &= \tfrac{1}{\beta^t}\|\tfrac{2C_{\mathsf{u}}}{\beta^t}[\mathbb{H}^{t-1} + \boldsymbol{\eta}_n^{t-1}\mathbf{I}]^\mathsf{T}\mathbb{u}_n^t + 2C_{\mathsf{x}}\mathsf{L}_n^t\Delta_n^t\| \\
&\overset{①}{\leq} c_1\{\Gamma_1 + \Gamma_2\}\cdot 2C_{\mathsf{u}}\cdot\tfrac{\|\mathbb{H}^{t-1}\| + \boldsymbol{\eta}_n^{t-1}}{\beta^t} + 2C_{\mathsf{u}}\cdot\tfrac{\mathsf{L}_n^t}{\beta^2}\cdot\textstyle\sum_{i=1}^n\|\Delta_i^t\| \\
&\overset{②}{\leq} c_1\{\Gamma_1 + \Gamma_2\}\cdot 8C_{\mathsf{u}}\theta_\star\lambda_\star\cdot\tfrac{\beta^{t-1}}{\beta^t} + 4C_{\mathsf{u}}\lambda_\star\cdot\Gamma_1 \\
&\overset{③}{\leq} (8C_{\mathsf{u}}\theta_\star\lambda_\star c_1 + 4C_{\mathsf{u}}\lambda_\star)\Gamma_1 + (8C_{\mathsf{u}}\theta_\star\lambda_\star c_1)\Gamma_2, \\
&\overset{④}{=} 2c_5\Gamma_1 + c_7\Gamma_1 + 2c_5\Gamma_2, \qquad (91)
\end{aligned}$$

where step ① uses the norm inequality and Lemma E.4; step ② uses Lemma E.3; step ③ uses $\beta^{t-1} \leq \beta^t$; step ④ uses the definition of $\{c_5, c_7\}$.

We now bound the term $\frac{1}{\beta^t}\|\mathbf{z}_n^t\|$ (as per Lemma E.1):

$$\begin{aligned}\tfrac{1}{\beta^t}\|\mathbf{z}_n^t\| &\triangleq \tfrac{1}{\beta^t}\|\tfrac{2C_{\mathsf{u}}}{\beta^t}\cdot\boldsymbol{\eta}_n^{t-1}\mathbb{u}_n^t\| \\ &\overset{①}{\leq} 2C_{\mathsf{u}}\cdot\tfrac{\boldsymbol{\eta}_n^{t-1}}{\beta^t}\cdot\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n\|\mathbb{u}_i^t\| \\ &\overset{②}{\leq} 2C_{\mathsf{u}}\cdot 2\theta_\star\lambda_\star\cdot c_1(\Gamma_1+\Gamma_1) \\ &\overset{③}{\leq} c_5(\Gamma_1+\Gamma_2),\end{aligned} \tag{92}$$

where step ① uses the norm inequality; step ② uses Lemma E.4, Lemma E.3, and $\beta^{t-1}\leq\beta^t$; step ③ uses the definition of c_5.

(a) We bound the term $\frac{1}{\beta^t}\sum_{i=1}^n\|\mathbf{d}_{\mathbf{x}_i}\|$:

$$\begin{aligned}\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n\|\mathbf{d}_{\mathbf{x}_i}\| &\leq \tfrac{1}{\beta^t}\|\mathbb{v}_n^t\|+\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n\|\mathbb{q}_i^t\| \\ &\overset{①}{\leq} \{c_5\Gamma_1+(c_5+c_6)\Gamma_2\}+\{c_1(\Gamma_1+\Gamma_2)+c_2\|\mathbf{r}^t\|+c_4\|\mathbf{r}^t\|+c_3\Gamma_2\} \\ &\overset{②}{\leq} p_1\Gamma_1+s_1\Gamma_2+u_1\|\mathbf{r}^t\|,\end{aligned}$$

where step ① uses Inequality (90); step ② uses the definitions: $p_1\triangleq c_1+c_5$, $s_1\triangleq c_1+c_3+c_5+c_6$, and $u_1\triangleq c_2+c_4$.

(b) We bound the term $\frac{1}{\beta^t}\|\mathbf{d}_{\mathbf{z}}\|$:

$$\tfrac{1}{\beta^t}\|\mathbf{d}_{\mathbf{z}}\|=\tfrac{1}{\beta^t}\|\mathbf{A}\mathbf{x}^t-\mathbf{b}\|\overset{①}{\leq}\tfrac{1}{\beta^0}\cdot\|\mathbf{r}^t\|\overset{②}{=}u_2\|\mathbf{r}^t\|,$$

where step ① uses $\beta^0\leq\beta^t$; step ② uses the definitions $u_2\triangleq 1/\beta^0$.

(c) We bound the term $\frac{1}{\beta^t}\sum_{i=1}^n\|\mathbf{d}_{\mathbf{x}_i'}\|$:

$$\begin{aligned}\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n\|\mathbf{d}_{\mathbf{x}_i'}\| &\leq \tfrac{1}{\beta^t}\|\mathbb{y}_n^t\|+\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n\|\boldsymbol{\eta}_i^{t-1}\Delta_i^t\| \\ &\overset{①}{\leq} \{2c_5\Gamma_1+c_7\Gamma_1+2c_5\Gamma_2\}+c_3\Gamma_1, \\ &\overset{②}{=} p_3\Gamma_1+s_3\Gamma_2,\end{aligned}$$

where step ① uses Inequalities (91) and (88); step ② uses the definitions: $p_3\triangleq 2c_5+c_7+c_3$, and $s_3\triangleq 2c_5$.

(d) We bound the term $\frac{1}{\beta^t}\sum_{i=1}^n\|\mathbf{d}_{\mathbf{x}_i''}\|$:

$$\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n\|\mathbf{d}_{\mathbf{x}_i''}\|\leq\tfrac{1}{\beta^t}\|\mathbf{z}_n^t\|\overset{①}{\leq}c_5\Gamma_1+c_5\Gamma_2\overset{②}{=}p_4\Gamma_1+s_4\Gamma_2,$$

where step ① uses Inequality (92); step ② uses the definitions: $p_4\triangleq c_5$, and $s_4\triangleq c_5$.

□

Lemma E.6. *For* Condition $\boxed{\mathbb{A}}$, we define $\{\mathbf{d}_{\mathbf{x}_i},\mathbf{d}_{\mathbf{z}},\mathbf{d}_{\mathbf{x}_i'},\mathbf{d}_{\mathbf{x}_i''}\}$ as in Lemma E.2. It holds that:

$$\begin{aligned}\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n\|\mathbf{d}_{\mathbf{x}_i}\| &\leq p_1\textstyle\sum_{i=1}^n\|\Delta_i^{t-1}\|+s_1\sum_{i=1}^n\|\Delta_i^t\|+u_1\|\mathbf{r}^t\|, \\ \tfrac{1}{\beta^t}\|\mathbf{d}_{\mathbf{z}}\| &\leq p_2\textstyle\sum_{i=1}^n\|\Delta_i^{t-1}\|+s_2\sum_{i=1}^n\|\Delta_i^t\|+u_2\|\mathbf{r}^t\|, \\ \tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n\|\mathbf{d}_{\mathbf{x}_i'}\| &\leq p_3\textstyle\sum_{i=1}^n\|\Delta_i^{t-1}\|+s_3\sum_{i=1}^n\|\Delta_i^t\|+u_3\|\mathbf{r}^t\|, \\ \tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n\|\mathbf{d}_{\mathbf{x}_i''}\| &\leq p_4\textstyle\sum_{i=1}^n\|\Delta_i^{t-1}\|+s_4\sum_{i=1}^n\|\Delta_i^t\|+u_4\|\mathbf{r}^t\|.\end{aligned}$$

Here, we define: $p_1\triangleq c_1+c_3+c_5+c_6$, $s_1\triangleq c_1+c_5+c_6+c_7$, $u_1\triangleq c_4+c_8$;
$p_2\triangleq 0$, $s_2\triangleq 0$, $u_2\triangleq 1/\beta^0$;
$p_3\triangleq 2c_6+c_3$, $s_3\triangleq 2c_6+c_7$, $u_3\triangleq 2c_8$;

$p_4 \triangleq c_6 + c_7$, $s_4 \triangleq c_6 + c_7$, $u_4 \triangleq c_8$.
Also, $c_5 \triangleq 8C_{\mathsf{a}}\lambda_\star c_1$, $c_6 \triangleq (16C_{\mathsf{a}} + 4C_{\mathsf{u}})\theta_\star\lambda_\star c_1$, $c_7 \triangleq 4C_{\mathsf{x}}\theta_\star\lambda_\star$, and $c_8 \triangleq 16C_{\mathsf{a}}\theta_\star\lambda_\star^{3/2}$.
Additionally, $\lambda_\star \triangleq \max_{i=1}^n \|\mathbf{A}_i\|_2^2$, $\theta_\star \triangleq \max(\boldsymbol{\theta})$, and $c_4 \triangleq 8nC_{\mathsf{a}}{\lambda_\star}^{3/2}$.
Lastly, note that $\{c_1, c_2, c_3\}$ are defined in Lemma E.4, and $\{C_{\mathsf{u}}, C_{\mathsf{x}}, C_{\mathsf{a}}\}$ are defined in Equation (22).

Proof. For any $i \in [n]$, we define $\mathbb{u}_i^t \triangleq \boldsymbol{\theta}_i \mathsf{L}_i^{t-1}(\mathbf{x}_i^t - \mathbf{x}_i^{t-1} - \boldsymbol{\alpha}_i(\mathbf{x}_i^{t-1} - \mathbf{x}_i^{t-2})) - \beta^{t-1}\mathbf{A}_i^\mathsf{T}[\sum_{j=i}^n \mathbf{A}_j(\mathbf{x}_j^t - \mathbf{x}_j^{t-1})]$, and we let $\mathbb{o}_i^t \in \nabla f_i(\mathbf{x}_i^t) + \partial h_i(\mathbf{x}_i^t) + \mathbf{A}_i^\mathsf{T}\mathbf{z}^t + \beta^t\mathbf{A}_i^\mathsf{T}\mathbf{r}^t$.

We define $\Gamma_1 = \sum_{i=1}^n \|\Delta_i^{t-1}\|$ and $\Gamma_2 = \sum_{i=1}^n \|\Delta_i^t\|$.

We bound the terms $\frac{1}{\beta^t}\sum_{i=1}^n \|\mathbb{q}_i^t\|$ (as per Lemma E.2):

$$\begin{aligned}
\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\mathbb{q}_i^t\| &= \tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\mathbb{o}_i^t + 2C_{\mathsf{a}}\mathbf{A}_i^\mathsf{T}\mathbf{A}_n\sigma^2\{\beta^t\mathbf{A}_n^\mathsf{T}\mathbf{r}^t + \mathbb{u}_n^t\} + \boldsymbol{\eta}_i^{t-1}\Delta_i^t\| \\
&\overset{①}{\leq} \tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \|\mathbb{o}_i^t\| + 8nC_{\mathsf{a}}\lambda_\star^{3/2}\|\mathbf{r}^t\| + 8C_{\mathsf{a}}\lambda_\star \cdot \tfrac{1}{\beta}\textstyle\sum_{i=1}^n \|\mathbb{u}_i^t\| + \tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n \boldsymbol{\eta}_i^{t-1}\|\Delta_i^t\| \\
&\overset{②}{\leq} \{c_1(\Gamma_1 + \Gamma_2) + c_2\|\mathbf{r}^t\|\} + \{c_4\|\mathbf{r}^t\|\} + \{8C_{\mathsf{a}}\lambda_\star \cdot c_1(\Gamma_1 + \Gamma_2)\} + c_4\Gamma_1 \\
&= (c_1 + 8C_{\mathsf{a}}\lambda_\star c_1 + c_3)\Gamma_1 + (c_1 + 8C_{\mathsf{a}}\lambda_\star c_1)\Gamma_2 + c_4\|\mathbf{r}^t\|, \\
&\overset{③}{=} (c_1 + c_5 + c_3)\Gamma_1 + (c_1 + c_5)\Gamma_2 + c_4\|\mathbf{r}^t\|, \qquad (93)
\end{aligned}$$

where step ① uses the norm inequality and the fact that $\sigma \leq 2$; step ② uses Lemma E.4; step ③ uses the definition of c_5.

We now bound the term $\frac{1}{\beta^t}\|\mathbb{v}_n^t\|$ (as per Lemma E.2):

$$\begin{aligned}
\tfrac{1}{\beta^t}\|\mathbb{v}_n^t\| &= \tfrac{1}{\beta^t}\|\tfrac{2C_{\mathsf{a}}\sigma^2 + 2C_{\mathsf{u}}}{\beta^t}[\mathbb{H}^{t-1}]^\mathsf{T}\mathbb{u}_n^t + 2C_{\mathsf{a}}\sigma^2[\mathbb{H}^{t-1}]^\mathsf{T}\mathbf{A}_n^\mathsf{T}\mathbf{r}^t + 2C_{\mathsf{x}}\mathsf{L}_n^t\Delta_n^t\| \\
&\overset{①}{\leq} (8C_{\mathsf{a}} + 2C_{\mathsf{u}}) \cdot \tfrac{\|\mathbb{H}^{t-1}\|}{\beta^t} \cdot \tfrac{1}{\beta^t}\|\mathbb{u}_n^t\| + 8C_{\mathsf{a}} \cdot \tfrac{\|\mathbb{H}^{t-1}\|}{\beta^t} \cdot \sqrt{\lambda_\star} \cdot \|\mathbf{r}^t\| + 2C_{\mathsf{x}} \cdot \tfrac{\mathsf{L}_n^t}{\beta^t} \cdot \Gamma_2 \\
&\overset{②}{\leq} (8C_{\mathsf{a}} + 2C_{\mathsf{u}}) \cdot 2\theta_\star\lambda_\star \cdot c_1(\Gamma_1 + \Gamma_2) + 8C_{\mathsf{a}} \cdot 2\theta_\star\lambda_\star \cdot \sqrt{\lambda_\star} \cdot \|\mathbf{r}^t\| + 2C_{\mathsf{x}} \cdot 2\theta_\star\lambda_\star \cdot \Gamma_2 \\
&= (16C_{\mathsf{a}} + 4C_{\mathsf{u}})\theta_\star\lambda_\star c_1 \cdot \Gamma_1 + \{(16C_{\mathsf{a}} + 4C_{\mathsf{u}})\theta_\star\lambda_\star c_1 + 4C_{\mathsf{x}}\theta_\star\lambda_\star\} \cdot \Gamma_2 + 16C_{\mathsf{a}}\theta_\star\lambda_\star^{3/2}\|\mathbf{r}^t\|, \\
&\overset{③}{=} c_6\Gamma_1 + (c_6 + c_7)\Gamma_2 + c_8\|\mathbf{r}^t\|, \qquad (94)
\end{aligned}$$

where step ① uses the norm inequality, and the fact that $\sigma \leq 2$; step ② uses Lemma E.4 and Lemma E.3; step ③ uses the definition of $\{c_6, c_7, c_8\}$.

We now bound the term $\frac{1}{\beta^t}\|\mathbb{y}_n^t\|$ (as per Lemma E.2):

$$\begin{aligned}
\tfrac{1}{\beta^t}\|\mathbb{y}_n^t\| &= \tfrac{1}{\beta^t}\|2C_{\mathsf{a}}\sigma^2(\mathbb{H}^{t-1} + \boldsymbol{\eta}_n\mathbf{I})^\mathsf{T} \cdot \mathbf{A}_n^\mathsf{T}\mathbf{r}^t + \tfrac{2C_{\mathsf{a}}\sigma^2 + 2C_{\mathsf{u}}}{\beta^t}[\mathbb{H}^{t-1} + \boldsymbol{\eta}_n^{t-1}\mathbf{I}]^\mathsf{T}\mathbb{u}_n^t + 2C_{\mathsf{x}}\mathsf{L}_n^t\Delta_n^t\| \\
&\overset{①}{\leq} 8C_{\mathsf{a}} \cdot \tfrac{1}{\beta^t}(\|\mathbb{H}^{t-1}\| + \boldsymbol{\eta}_n^{t-1}) \cdot \sqrt{\lambda_\star} \cdot \|\mathbf{r}^t\| + (8C_{\mathsf{a}} + 2C_{\mathsf{u}}) \cdot \tfrac{1}{\beta^t}(\|\mathbb{H}^{t-1}\| + \boldsymbol{\eta}_n^{t-1}) \cdot \tfrac{1}{\beta}\|\mathbb{u}_n^t\| + 2C_{\mathsf{x}} \cdot \tfrac{\mathsf{L}_n^t}{\beta^t} \cdot \Gamma_2 \\
&\overset{②}{\leq} 8C_{\mathsf{a}} \cdot 4\lambda_\star\theta_\star \cdot \sqrt{\lambda_\star} \cdot \|\mathbf{r}^t\| + (8C_{\mathsf{a}} + 2C_{\mathsf{u}}) \cdot 4\lambda_\star\theta_\star \cdot (c_1\Gamma_1 + c_1\Gamma_2) + 2C_{\mathsf{x}} \cdot 2\lambda_\star \cdot \Gamma_2 \\
&= 32C_{\mathsf{a}}\lambda_\star^{3/2}\theta_\star\|\mathbf{r}^t\| + (32C_{\mathsf{a}} + 8C_{\mathsf{u}}) \cdot \lambda_\star\theta_\star c_1 \cdot \Gamma_1 + \{(32C_{\mathsf{a}} + 8C_{\mathsf{u}})\lambda_\star\theta_\star c_1 + 4C_{\mathsf{x}}\lambda_\star\} \cdot \Gamma_2 \\
&= 2c_8\|\mathbf{r}^t\| + 2c_6\Gamma_1 + (2c_6 + c_7)\Gamma_2, \qquad (95)
\end{aligned}$$

where step ① uses the norm inequality, and the fact that $\sigma \in (0, 2)$; step ② uses Lemma E.3 and Lemma E.4; step ③ uses the definition of $\{c_6, c_7, c_8\}$.

We now bound the term $\frac{1}{\beta^t}\|\mathbf{z}_n^t\|$ (as per Lemma E.2):

$$\begin{aligned}
\tfrac{1}{\beta^t}\|\mathbf{z}_n^t\| &= \tfrac{1}{\beta^t}\|\tfrac{2C_{\mathsf{a}}}{\beta^t}\sigma\boldsymbol{\eta}_n^{t-1}(\sigma\beta^t\mathbf{A}_n^{\mathsf{T}}\mathbf{r}^t+\sigma\mathbb{u}_n^t)+\tfrac{2C_{\mathsf{u}}}{\beta^t}\boldsymbol{\eta}_n^{t-1}\mathbb{u}_n^t\| \\
&= \tfrac{1}{\beta^t}\|2C_{\mathsf{a}}\sigma^2\boldsymbol{\eta}_n^{t-1}\mathbf{A}_n^{\mathsf{T}}\mathbf{r}^t+\tfrac{2C_{\mathsf{a}}\sigma^2+2C_{\mathsf{u}}}{\beta^t}\boldsymbol{\eta}_n^{t-1}\mathbb{u}_n^t\| \\
&\overset{①}{\leq} 8C_{\mathsf{a}}\cdot\tfrac{\boldsymbol{\eta}_n^{t-1}}{\beta}\cdot\|\mathbf{A}_n\|\|\mathbf{r}^t\|+(8C_{\mathsf{a}}+2C_{\mathsf{u}})\cdot\tfrac{\boldsymbol{\eta}_n^{t-1}}{\beta^t}\cdot\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n\|\mathbb{u}_i^t\|, \\
&\overset{②}{\leq} 8C_{\mathsf{a}}\cdot 2\theta_\star\lambda_\star\cdot\sqrt{\lambda_\star}\|\mathbf{r}^t\|+(8C_{\mathsf{a}}+2C_{\mathsf{u}})\cdot 2\lambda_\star\theta_\star\cdot(c_1\Gamma_1+c_1\Gamma_2) \\
&\overset{③}{=} c_8\|\mathbf{r}^t\|+(c_6+c_7)(\Gamma_1+\Gamma_2), \qquad (96)
\end{aligned}$$

where step ① uses the norm inequality and the condition that $\sigma\in(0,2)$; step ② uses Lemma E.3 and Lemma E.4; step ③ uses the definition of $\{c_6,c_7,c_8\}$.

(a) We bound the term $\frac{1}{\beta^t}\sum_{i=1}^n\|\mathbf{d}_{\mathbf{x}_i}\|$:

$$\begin{aligned}
&\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n\|\mathbf{d}_{\mathbf{x}_i}\| \\
\overset{①}{\leq}\ &\tfrac{1}{\beta^t}\|\mathbb{y}_n^t\|+\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n\|\mathbb{q}_i^t\| \\
\overset{②}{\leq}\ &\{c_6\Gamma_1+(c_6+c_7)\Gamma_2+c_8\|\mathbf{r}^t\|\}+\{(c_1+c_5+c_3)\Gamma_1+(c_1+c_5)\Gamma_2+c_4\|\mathbf{r}^t\|\}, \\
\overset{③}{=}\ &p_1\Gamma_1+s_1\Gamma_2+u_1\|\mathbf{r}^t\|,
\end{aligned}$$

where step ① uses the norm inequality; step ② uses Inequalities (93) and (94); step ③ uses the definition $p_1\triangleq c_1+c_3+c_5+c_6$, $s_1\triangleq c_1+c_5+c_6+c_7$, and $u_1\triangleq c_4+c_8$.

(b) We bound the term $\frac{1}{\beta^t}\|\mathbf{d}_{\mathbf{z}}\|$:

$$\tfrac{1}{\beta^t}\|\mathbf{d}_{\mathbf{z}}\|=\tfrac{1}{\beta^t}\|\mathbf{A}\mathbf{x}^t-\mathbf{b}\|\overset{①}{\leq}\tfrac{1}{\beta^0}\cdot\|\mathbf{r}^t\|\overset{②}{=}u_2\|\mathbf{r}^t\|,$$

where step ① uses $\beta^0\leq\beta^t$; step ② uses the definitions $u_2\triangleq 1/\beta^0$.

(c) We bound the term $\frac{1}{\beta^t}\sum_{i=1}^n\|\mathbf{d}_{\mathbf{x}_i'}\|$:

$$\begin{aligned}
\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n\|\mathbf{d}_{\mathbf{x}_i'}\| &\overset{①}{\leq} \tfrac{1}{\beta^t}\|\mathbb{y}_n^t\|+\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n\|\boldsymbol{\eta}_i^{t-1}\Delta_i^t\| \\
&\overset{②}{\leq} 2c_8\|\mathbf{r}^t\|+2c_6\Gamma_1+(2c_6+c_7)\Gamma_2+c_3\Gamma_1 \\
&\overset{③}{=} p_3\Gamma_1+s_3\Gamma_2+u_3\|\mathbf{r}^t\|,
\end{aligned}$$

where step ① uses the norm inequality; step ② uses Inequality (95) and Lemma E.4; step ③ uses the definition $p_3\triangleq 2c_6+c_3$, $s_3\triangleq 2c_6+c_7$, and $u_3\triangleq 2c_8$.

(d) We bound the term $\frac{1}{\beta^t}\sum_{i=1}^n\|\mathbf{d}_{\mathbf{x}_i''}\|$:

$$\tfrac{1}{\beta^t}\textstyle\sum_{i=1}^n\|\mathbf{d}_{\mathbf{x}_i''}\| = \tfrac{1}{\beta^t}\|\mathbf{z}_n^t\|\leq c_8\|\mathbf{r}^t\|+(c_6+c_7)(\Gamma_1+\Gamma_2)=p_4\Gamma_1+s_4\Gamma_2+u_4\|\mathbf{r}^t\|,$$

where step ① uses Inequality (96); step ③ uses the definition $p_4\triangleq c_6+c_7$, $s_4\triangleq c_6+c_7$, and $u_4\triangleq c_8$.

□

Now, we proceed to prove the main result of this lemma.

Lemma E.7. *(*Subgradient Bounds for Conditions $\boxed{\mathbb{I}}$ and $\boxed{\mathbb{A}}$*) We let $K\triangleq\max\{[\sum_{i=1}^4 p_i],[\sum_{i=1}^4 s_i],[\sum_{i=1}^4 u_i]\}$, where $\{p_i,s_i,u_i\}_{i=1}^4$ are defined in Lemma E.5 or Lemma E.6. We have:*

$$\tfrac{1}{\beta^t}|||\partial\Theta(\mathbf{x}^t,\mathbf{z}^t,\mathbf{x}^{t-1},\mathbf{x}^{t-2};\beta^t,\beta^{t-1})|||\leq K\{\textstyle\sum_{i=1}^n\|\mathbf{x}_i^{t-1}-\mathbf{x}_i^{t-2}\|+\sum_{i=1}^n\|\mathbf{x}_i^t-\mathbf{x}_i^{t-1}\|+\|\mathbf{r}^t\|\}.$$

Here, we define $\mathbb{X}\triangleq\{\mathbf{x},\mathbf{z},\mathbf{x}',\mathbf{x}''\}$, and $[|||\partial\Theta(\mathbb{X};\beta)|||]^2\triangleq\|\partial_{\mathbf{z}}\Theta(\mathbb{X};\beta)\|_2^2+\sum_{i=1}^n[\|\partial_{\mathbf{x}_i}\Theta(\mathbb{X};\beta)\|_2^2+\|\partial_{\mathbf{x}_i'}\Theta(\mathbb{X};\beta)\|_2^2+\|\partial_{\mathbf{x}_i''}\Theta(\mathbb{X};\beta)\|_2^2]$.

Proof. We define $\Gamma_1 \triangleq \sum_{i=1}^n \|\mathbf{x}_i^{t-1} - \mathbf{x}_i^{t-2}\|$ and $\Gamma_2 \triangleq \sum_{i=1}^n \|\mathbf{x}_i^{t} - \mathbf{x}_i^{t-1}\|$. We have:

$$\begin{aligned}
& |||\partial\Theta(\mathbf{x}^t, \mathbf{z}^t, \mathbf{x}^{t-1}, \mathbf{x}^{t-2}; \beta^t, \beta^{t-1})||| \\
= \; & \{\textstyle\sum_{i=1}^n [\|\mathbf{d}_{\mathbf{x}_i}\|_2^2 + \|\mathbf{d}_{\mathbf{x}_i'}\|_2^2 + \|\mathbf{d}_{\mathbf{x}_i''}\|_2^2] + \|\mathbf{d}_{\mathbf{z}}\|_2^2\}^{1/2} \\
\overset{①}{\leq} \; & \{\textstyle\sum_{i=1}^n \|\mathbf{d}_{\mathbf{x}_i}\|\} + \{\sum_{i=1}^n \|\mathbf{d}_{\mathbf{x}_i'}\|\} + \{\sum_{i=1}^n \|\mathbf{d}_{\mathbf{x}_i''}\|\} + \|\mathbf{d}_{\mathbf{z}}\| \\
\overset{②}{\leq} \; & [\textstyle\sum_{j=1}^4 p_j]\Gamma_1 + [\sum_{j=1}^4 s_j]\Gamma_2 + [\sum_{j=1}^4 u_j]\|\mathbf{r}^{t+1}\| \\
\overset{③}{\leq} \; & K\Gamma_1 + K\Gamma_2 + K\|\mathbf{r}^{t+1}\|,
\end{aligned}$$

where step ① uses $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a \geq 0$ and $b \geq 0$; step ② uses Lemmas E.5 and E.6; step ③ uses the definition of K.

□

E.2. Proof of Theorem 5.3

Proof. We define $\mathbb{X}^t \triangleq \{\mathbf{x}^t, \mathbf{z}^t, \mathbf{x}^{t-1}, \mathbf{x}^{t-2}\}$, $F(\mathbb{X}^t) \triangleq \Theta(\mathbb{X}^t; \beta^t, \beta^{t+1})$. We denote $\mathbb{X}^\star$ as a limiting point of $\{\mathbb{X}^t\}_{t=0}^\infty$. We let $F(\mathbb{X}^\star) \triangleq \Theta(\mathbb{X}^\star; \beta, \beta')$, where β and β' are the associated penalties for $\mathbb{X}^*$.

For simplicity, we denote $F^t \triangleq F(\mathbb{X}^t)$ and $F^\star \triangleq F(\mathbb{X}^\star)$.

Firstly, using Assumption 5.1, we have:

$$\tfrac{1}{\varphi'(F(\mathbb{X}^t) - F(\mathbb{X}^\star))} \leq \operatorname{dist}(\mathbf{0}, \partial F(\mathbb{X}^t)), \tag{97}$$

Secondly, given the desingularization function $\varphi(\cdot)$ is concave, for any $a \in \mathbb{R}$ and $b \in \mathbb{R}$, we have:

$$\varphi(b) + (a-b)\varphi'(a) \leq \varphi(a).$$

Applying the inequality above with $a = F^t - F^\star$ and $b = F^{t+1} - F^\star$, we have:

$$(F^t - F^{t+1}) \cdot \varphi'(F^t - F^\star) \leq \varphi(F^t - F^\star) - \varphi(F^{t+1} - F^\star) \triangleq \Delta_F^t. \tag{98}$$

We derive the following inequalities:

$$\begin{aligned}
& V\beta^t \cdot \{\|\mathbf{r}^{t+1}\|_2^2 + \textstyle\sum_{i=1}^n \|\Delta_i^{t+1}\|_2^2\} \\
\overset{①}{\leq} \; & \tfrac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2 + \epsilon_3 \textstyle\sum_{i=1}^n \boldsymbol{\gamma}_i \mathsf{L}_i^t \|\Delta_i^{t+1}\|_2^2 = \mathcal{E}^{t+1} \\
\overset{②}{\leq} \; & \Theta^t - \Theta^{t+1} + \tfrac{C_{\mathsf{w}}}{\beta^t} = F^t - F^{t+1} + \tfrac{C_{\mathsf{w}}}{\beta^t} \\
\overset{③}{\leq} \; & \tfrac{\Delta_F^t}{\varphi'(\Theta^t - \Theta^*))} + \tfrac{C_{\mathsf{w}}}{\beta^t} \\
\overset{④}{\leq} \; & \Delta_F^t \cdot \operatorname{dist}(\mathbf{0}, \partial\Theta(\mathbb{X}^t; \beta^t, \beta^{t-1})) + \tfrac{C_{\mathsf{w}}}{\beta^t} \\
\overset{⑤}{\leq} \; & \Delta_F^t K\{\beta^t\|\mathbf{r}^t\| + \beta^t \textstyle\sum_{i=1}^n \|\Delta_i^t\| + \beta^t \sum_{i=1}^n \|\Delta_i^{t-1}\|\} + \tfrac{C_{\mathsf{w}}}{\beta^t},
\end{aligned}$$

where step ① uses the definition of $V \triangleq \min(\frac{\xi}{2}, \epsilon_3 \min_{i=1}^n \boldsymbol{\gamma}_i \|\mathbf{A}_i\|_2^2)$; step ② uses $\mathcal{E}^{t+1} \leq \Theta^t - \Theta^{t+1} + \frac{C_{\mathsf{w}}}{\beta^t}$ which is due to Theorems 3.5 and 3.7; step ③ uses Inequality (98); step ④ uses Inequality (97); step ⑤ uses Lemma E.7. Dividing both sides by $V\beta^t$, we have:

$$\begin{aligned}
& \|\mathbf{r}^{t+1}\|_2^2 + \textstyle\sum_{i=1}^n \|\Delta_i^{t+1}\|_2^2 \\
\leq \; & \tfrac{K}{V} \cdot \Delta_F^t \{\|\mathbf{r}^t\| + \textstyle\sum_{i=1}^n \|\Delta_i^t\|\} + \sum_{i=1}^n \|\Delta_i^{t-1}\|\} + \tfrac{C_{\mathsf{w}}}{V} \cdot \tfrac{1}{(\beta^t)^2} \\
\overset{①}{\leq} \; & \tfrac{K}{V} \cdot \Delta_F^t \{\underbrace{\|\mathbf{r}^t\| + \textstyle\sum_{i=1}^n \|\Delta_i^t\|}_{\triangleq e^t} + \underbrace{\|\mathbf{r}^{t-1}\| + \textstyle\sum_{i=1}^n \|\Delta_i^{t-1}\|}_{\triangleq e^{t-1}}\} + \tfrac{C_{\mathsf{w}}}{\beta^0 V} \cdot \tfrac{1}{\beta^t},
\end{aligned} \tag{99}$$

where step ① uses $\|\mathbf{r}^{t-1}\| \geq 0$ and $\beta^t \geq \beta^0$.

Additionally, we notice that:

$$e^{t+1} \triangleq \|\mathbf{r}^{t+1}\| + \sum_{i=1}^n \|\Delta_i^{t+1}\| \overset{①}{\leq} \frac{1}{\sqrt{n+1}} \cdot \sqrt{\|\mathbf{r}^{t+1}\|_2^2 + \sum_{i=1}^n \|\Delta_i^{t+1}\|_2^2}, \tag{100}$$

where step ① uses the norm inequality that $\|\mathbf{a}\|_1 \leq \sqrt{n}\|\mathbf{a}\|_2$ for any $\mathbf{a} \in \mathbb{R}^n$.

Combining Inequalities (99) and (100) together, we have:

$$e^{t+1} \leq \frac{1}{\sqrt{n+1}} \sqrt{\Delta_F^t \frac{K}{V}(e^t + e^{t-1}) + \frac{C_{\mathsf{w}}}{V\beta^0} \cdot \frac{1}{\beta^t}}. \tag{101}$$

Applying Lemma A.8 with $w^t \triangleq \frac{K}{(n+1)V} \cdot \Delta_F^t$ and $p^t \triangleq \frac{C_{\mathsf{w}}}{(n+1)\beta^0 V} \cdot \frac{1}{\beta^t}$, we have from Inequality (101):

$$\begin{aligned}
\sum_{t=0}^T e^t &\leq \tfrac{3}{2}e^0 + \tfrac{1}{2}e^{-1} + \tfrac{1}{2}\sum_{t=0}^T (w^t + p^t) \\
&\leq \tfrac{3}{2}e^0 + \tfrac{1}{2}e^{-1} + \tfrac{K}{2(n+1)V}\sum_{t=0}^T \{\varphi(F^t - F^\star) - \varphi(F^{t+1} - F^\star)\} + \tfrac{C_{\mathsf{w}}}{2(n+1)V\beta^0} \cdot \sum_{t=0}^T \tfrac{1}{\beta^t} \\
&\overset{①}{\leq} \tfrac{3}{2}e^0 + \tfrac{1}{2}e^{-1} + \tfrac{K}{2(n+1)V} \cdot \{\varphi(F^0 - F^\star) - \varphi(F^{T+1} - F^\star)\} + \tfrac{C_{\mathsf{w}}}{2(n+1)V\beta^0} \cdot C_{\mathsf{b}} \\
&\overset{②}{\leq} \tfrac{3}{2}e^0 + \tfrac{1}{2}e^{-1} + \tfrac{K}{2(n+1)V} \cdot \varphi(F^0 - F^\star) + \tfrac{C_{\mathsf{w}}C_{\mathsf{b}}}{2(n+1)V\beta^0} \triangleq C_{\mathsf{e}},
\end{aligned}$$

where step ① uses Inequality (3); step ② uses the fact the desingularization function $\varphi(\cdot)$ is positive. □

F. Extension: Stochastic IRPL-ADMM

This section extends the proposed IRPL-ADMM algorithm to stochastic settings.

Stochastic optimization methods are potent tools for addressing large-scale problems in machine learning. Stochastic Gradient Descent (SGD) efficiently tackles finite-sum optimization problems by computing gradients for individual samples in each iteration. Various accelerated versions of SGD have been successfully introduced to reduce variance in convex composite minimization (Defazio et al., 2014), non-convex smooth minimization (Johnson & Zhang, 2013; Nguyen et al.; Fang et al., 2018a; Zhou et al., 2020), and nonconvex composite minimization (Johnson & Zhang, 2013; Ghadimi et al., 2016; J Reddi et al., 2016; Li et al., 2017). Furthermore, stochastic gradient descent has been integrated into the ADMM framework to address a broader range of convex and nonconvex composite optimization problems (Suzuki, 2014; Zhang & Kwok, 2014; Huang et al., 2019).

Given that the Stochastic Path Integrated Differential Estimator (SPIDER) estimator has been demonstrated to possess nearly optimal computational complexity bounds, we follow the approach outlined in (Huang et al., 2019) by integrating the methods presented in (Fang et al., 2018b; Wang et al., 2019b) into our ADMM algorithm, resulting in IRPL-ADMM-SPIDER.

Building upon Problem (1), we introduce the following additional assumption.

Assumption F.1. The function $f_n(\mathbf{x}_n)$ takes the following form which is of finite-sum structure:

$$f_n(\mathbf{x}_n) = \frac{1}{N}\sum_{j=1}^N f_{n,j}(\mathbf{x}_n). \tag{102}$$

Additionally, each $f_{n,j}(\cdot)$ is L_n-smooth, meaning that $\|\nabla f_{n,j}(\mathbf{x}_n) - \nabla f_{n,j}(\tilde{\mathbf{x}}_n)\| \leq L_n\|\mathbf{x}_n - \tilde{\mathbf{x}}_n\|$ for all $j \in [N]$. This property extends to $f_n(\mathbf{x}_n)$, which is also L_n-smooth. Furthermore, $f_n(\mathbf{x}_n)$ is Lipschitz continuous with a constant C_f such that $\|\nabla f_n(\mathbf{x}_n)\| \leq C_f$.

Remarks. We assume that the smooth function $f_n(\cdot)$ has a finite-sum structure, but our algorithm and convergence analysis are also applicable to cases where another block of smooth function exhibits this structure.

F.1. The Proposed IRPL-ADMM-SPIDER Algorithm

We denote r_t be the unique integer such that $(r_t - 1)q \leq t \leq r_t q - 1$.

Algorithm 3 IRPL-ADMM-SPIDER: The Proposed Inertial Relaxed Proximal Linearized ADMM based on SPIDER.

Initialize $\{\mathbf{x}^0, \mathbf{z}^0\}$. Let $\mathbf{x}^{-1} = \mathbf{x}^0$ and $\mathbf{y}^0 = \mathbf{x}^0$.
Use Algorithm 2 to choose suitable $\{\beta^0, \boldsymbol{\theta}, \boldsymbol{\alpha}, \xi, \sigma\}$.
for $t = 0$ to T **do**
  Compute $\mathbf{v}^t$ using Formula (103).
  $\mathbf{x}_1^{t+1} \in \min_{\mathbf{x}_1} h_1(\mathbf{x}_1) + \frac{\boldsymbol{\theta}_1 \mathsf{L}_1^t}{2}\|\mathbf{x}_1 - \mathbf{y}_1^t\|_2^2 + \langle \mathbf{x}_1 - \mathbf{x}_1^t, \nabla_{\mathbf{x}_1}\tilde{G}(\mathbf{x}_{[1,n]}^t, \mathbf{z}^t; \beta^t)\rangle$
  $\mathbf{x}_2^{t+1} \in \min_{\mathbf{x}_2} h_2(\mathbf{x}_2) + \frac{\boldsymbol{\theta}_2 \mathsf{L}_2^t}{2}\|\mathbf{x}_2 - \mathbf{y}_2^t\|_2^2 + \langle \mathbf{x}_2 - \mathbf{x}_2^t, \nabla_{\mathbf{x}_2}\tilde{G}(\mathbf{x}_1^{t+1}, \mathbf{x}_{[2,n]}^t, \mathbf{z}^t; \beta^t)\rangle$
  ...
  $\mathbf{x}_n^{t+1} \in \min_{\mathbf{x}_n} h_n(\mathbf{x}_n) + \frac{\boldsymbol{\theta}_n \mathsf{L}_n^t}{2}\|\mathbf{x}_n - \mathbf{y}_n^t\|_2^2 + \langle \mathbf{x}_n - \mathbf{x}_n^t, \nabla_{\mathbf{x}_n}\tilde{G}(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_n^t, \mathbf{z}^t; \beta^t)\rangle$
  $\mathbf{y}_j^{t+1} = \mathbf{x}_j^{t+1} + \boldsymbol{\alpha}_j(\mathbf{x}_j^{t+1} - \mathbf{x}_j^t),\ \forall j \in [n]$
  $\mathbf{z}^{t+1} = \mathbf{z}^t + \sigma\beta^t((\sum_{i=1}^n \mathbf{A}_i\mathbf{x}_i) - \mathbf{b})$
  Use a (β^0, ξ, p)-regular penalty update rule to update β^{t+1} based on the current value β^t.
end for

First, we introduce the SPIDER estimator as follows (Fang et al., 2018a):

$$\mathbf{v}^t = \begin{cases} \nabla f_n(\mathbf{x}_n^t), & \mathrm{mod}(t, q) = 0; \\ \mathbf{v}^{t-1} + \nabla f_n(\mathbf{x}_n^t; \mathcal{I}^t) - \nabla f_n(\mathbf{x}_n^{t-1}; \mathcal{I}^{t-1}), & \text{else.} \end{cases} \tag{103}$$

Here, $\nabla f_n(\mathbf{x}_n^t; \mathcal{I}^t)$ denotes the average gradient of the examples $\mathcal{I}^t$, and $\mathcal{I}^t$ is a mini-batch which is picked uniformly and randomly (with replacement) from $\{1, 2, ..., N\}$ with $|\mathcal{I}^t| = b$ for all t.

Second, we employ the following stochastic gradient for all $j \in [n]$:

$$\nabla_{\mathbf{x}}\tilde{G}(\mathbf{x}_{[1,j-1]}^{t+1}, \mathbf{x}_{[j,n]}^t, \mathbf{z}^t; \beta^t) \triangleq \begin{cases} \mathbf{A}_j^\top \mathbf{z}^t + \beta^t \mathbf{A}_j^\top\{[\sum_{i=1}^{j-1} \mathbf{A}_i\mathbf{x}_i^{t+1}] + [\sum_{i=j}^n \mathbf{A}_i\mathbf{x}_i^t] - \mathbf{b}\} + \nabla f_j(\mathbf{x}_j^t), & j \neq n. \\ \mathbf{A}_j^\top \mathbf{z}^t + \beta^t \mathbf{A}_j^\top\{[\sum_{i=1}^{n-1} \mathbf{A}_i\mathbf{x}_i^{t+1}] + [\sum_{i=n}^n \mathbf{A}_i\mathbf{x}_i^t] - \mathbf{b}\} + \mathbf{v}^t, & j = n; \end{cases} \tag{104}$$

Notably, when $j \neq n$, we use the gradient, which has the same form as for the deterministic settings. However, when $j = n$, we replace the true gradient $\nabla f_j(\mathbf{x}_j^t)$ with $\mathbf{v}^t$, the unbiased estimate of $\nabla f_j(\mathbf{x}_j^{t+1})$ using SPIDER.

Thirdly, we provide the proposed IRPL-ADMM-SPIDER in Algorithm F.4. We conduct a comprehensive analysis of the convergence properties of the IRPL-ADMM-SPIDER algorithm and establish its optimality in terms of the Incremental First-order Oracle (IFO). The expectation of a random variable is denoted as $\mathbb{E}[\cdot]$.

F.2. Some Pre-convergence Results

Initially, we introduce a useful lemma from (Fang et al., 2018a).

Lemma F.2. *(Fang et al., 2018a) The SPIDER estimator generates stochastic gradient $\mathbf{v}^t$ satisfies for all $(r_t - 1)q + 1 \leq t \leq r_t q - 1$ that:*

$$\mathbb{E}[\|\mathbf{v}^t - \nabla f_n(\mathbf{x}_n^t)\|_2^2] - \mathbb{E}[\|\mathbf{v}^{t-1} - \nabla f_n(\mathbf{x}_n^{t-1})\|_2^2] \leq \tfrac{(L_n)^2}{b}\mathbb{E}[\|\mathbf{x}_n^t - \mathbf{x}_n^{t-1}\|_2^2]. \tag{111}$$

We have the following useful lemma concerning the decrease in the primal.

Lemma F.3. *(Proof in Appendix G.1,* Decrease for the Primal under Stochastic Settings*) We define $\boldsymbol{\gamma}_i \triangleq \frac{1}{2}[\boldsymbol{\theta}_i - 1 - (2 + \epsilon_1)\boldsymbol{\alpha}_i\boldsymbol{\theta}_i]$ for all $i \in [n]$. We have:*

$$\begin{aligned} & \mathcal{E}^{t+1} + \Theta_\circ^{t+1} - \Theta_\circ^t \\ \leq\ & \tfrac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 - \{(1 - \epsilon_3)\boldsymbol{\gamma}_n - \tfrac{\epsilon_3}{2}\}\mathsf{L}_n^t\|\mathbf{x}_n^{t+1} - \mathbf{x}_n^t\|^2 + \tfrac{1}{2\epsilon_3\mathsf{L}_n^t}\|\nabla f_n(\mathbf{x}_n^t) - \mathbf{v}^t\|_2^2, \end{aligned} \tag{112}$$

where $\delta \triangleq 1 + \epsilon_2$, and $\{\Theta_\circ^t, \mathcal{E}^{t+1}\}$ are respectively defined in Equations (14) and (15).

We have the following lemma that provides an upper bound for the critical term $\frac{1}{2\epsilon_3\mathsf{L}_n^t}\|\nabla f_n(\mathbf{x}_n^t) - \mathbf{v}^t\|_2^2$ in Lemma F.3.

Algorithm 4 A Procedure for Finding Suitable Parameters $\xi \in (0,\epsilon_1)$, $\boldsymbol{\alpha} \in (0,1)^n$, $\boldsymbol{\theta} \in (1,\infty)^n$, $\sigma \in (0,2)$, $\beta^0 \in (0,\infty)$ for Algorithm 3 (Stochastic Settings).

1: Choose suitable $(\epsilon_1,\epsilon_2,\epsilon_3)$. Default parameters:

$$\text{Condition } \boxed{\mathbb{I}} : (\epsilon_1,\epsilon_2,\epsilon_3) = (0.01, 0.01, 0.001) \tag{105}$$

$$\text{Condition } \boxed{\mathbb{A}} : (\epsilon_1,\epsilon_2,\epsilon_3) = (0.01, 1, 0.001) \tag{106}$$

2: We define $\gamma_i \triangleq \frac{1}{2}[\boldsymbol{\theta}_i - 1 - (2+\epsilon_1)\boldsymbol{\alpha}_i\boldsymbol{\theta}_i]$ and :

$$\boldsymbol{\gamma}_i' \triangleq \begin{cases} \gamma_i[1-\epsilon_3], & i \in [n-1]; \\ \gamma_i[1-\epsilon_3] - \frac{\epsilon_3}{2} - \frac{\epsilon_3 q}{2b}, & i = n. \end{cases} \tag{107}$$

3: For the first $(n-1)$ blocks, find suitable parameters $\{\boldsymbol{\alpha}_i, \boldsymbol{\theta}_i\}_{i=1}^{n-1}$ such that $\boldsymbol{\gamma}_i' > 0$ for all $i \in [n-1]$.

4: For the last block, find suitable parameters $(\boldsymbol{\alpha}_n, \boldsymbol{\theta}_n, \sigma)$ such that (9) or (10) holds.

• Condition $\boxed{\mathbb{I}}$: Over-Relaxation Stepsize $\sigma \in [1,2)$.

$$\sigma \in [1,2), \boldsymbol{\gamma}_n' > 0, \underbrace{8\sigma_1\delta \cdot (1+\epsilon_3)}_{=4C_{\mathsf{u}}}[(\chi-1)^2 + \tau\chi] \leq \boldsymbol{\gamma}_n'. \tag{108}$$

• Condition $\boxed{\mathbb{A}}$: Under-Relaxation Stepsize $\sigma \in (0,1)$.

$$\sigma \in (0,1),\ \boldsymbol{\gamma}_n' > 0, \underbrace{\overline{\lambda}/\underline{\lambda} \cdot 8\sigma\delta}_{=2\overline{\lambda}C_{\mathsf{u}}} \cdot (\chi^2 + \chi\tau) \leq \boldsymbol{\gamma}_n'. \tag{109}$$

Here, $\{\delta, \chi, \boldsymbol{\alpha}_n'\}$ in (9) and (10) are defined as:

$$\delta \triangleq 1 + \epsilon_2, \chi \triangleq \boldsymbol{\theta}_n(1+\epsilon_3), \tau \triangleq \boldsymbol{\alpha}_n^2(1+\epsilon_1). \tag{110}$$

5: Choose β^0 and ξ satisfying Assumption 2.7 that: $\xi \leq \min(\epsilon_1, \epsilon_2\sigma)$, $\beta^0 \geq L_i/(\epsilon_3\overline{\lambda})$ for all $i \in [n]$.

Lemma F.4. *(Proof in Section G.2) We let $\mathsf{L}_i^t \triangleq L_i + \beta^t\|\mathbf{A}_i\|_2^2$ for all $i \in [n]$. The SPIDER estimator generates stochastic gradient $\mathbf{v}^t$ satisfies for all $(r_t-1)q \leq t \leq r_t q - 1$ that:*

$$\mathbb{E}[\tfrac{1}{2\epsilon_3 L_n^t}\|\nabla f_n(\mathbf{x}_n^t) - \mathbf{v}^t\|_2^2] \leq \tfrac{\epsilon_3}{2b}\textstyle\sum_{i=(r_t-1)q}^{t-1}\mathbb{E}[L_n^t\|\Delta_n^{i+1}\|_2^2]. \tag{113}$$

We present the following lemma on the first-order optimality condition.

Lemma F.5. (First-Order Optimality Condition under Stochastic Settings) Assume $\sigma \in (0,2)$. We let $\mathbb{w}_n^{t+1} \in \partial h_n(\mathbf{x}_n^{t+1}) + \mathbf{v}^{t+1}$, and $\mathbb{u}_n^{t+1} = \boldsymbol{\theta}_n\mathsf{L}_n^t(\mathbf{x}_n^{t+1} - \mathbf{x}_n^t - \boldsymbol{\alpha}_n(\mathbf{x}_n^t - \mathbf{x}_n^{t-1})) + \beta^t\mathbf{A}_n^{\mathsf{T}}[\mathbf{A}_n(\mathbf{x}_n^t - \mathbf{x}_n^{t+1})]$, where $\mathbf{v}^{t+1}$ is a unbiased estimation of $\nabla f_n(\mathbf{x}_n^{t+1})$. We have: $\mathbf{0} = \sigma\mathbb{w}_n^{t+1} + \sigma\mathbf{A}_n^{\mathsf{T}}\mathbf{z}^t + \mathbf{A}_n^{\mathsf{T}}(\mathbf{z}^{t+1} - \mathbf{z}^t) + \sigma\mathbb{u}_n^{t+1}$. Furthermore, we obtain two different identities as shown in Equations (19) and (20).

Remarks. Lemmas F.5 and 3.2 share a close resemblance, differing primarily in the choice of gradients. In Lemma 3.2, we utilize a deterministic gradient for $\mathbb{w}_n^{t+1} \in \partial h_n(\mathbf{x}_n^{t+1}) + \nabla f_n(\mathbf{x}_n^t)$, while in Lemma F.5, we opt for a stochastic gradient for $\mathbb{w}_i^{t+1} \in \partial h_n(\mathbf{x}_n^{t+1}) + \mathbf{v}^t$. We omit the proof for brevity.

The following lemma bounds the terms $\|\mathbb{w}_n^{t+1} - \mathbb{w}_n^t\|_2^2$ and $\frac{1}{\beta^t}\|\mathbb{u}_n^{t+1}\|_2^2$.

Lemma F.6. *(Proof in Appendix G.3) We define $\iota \triangleq 8C_h^2 + 16qC_f^2$, $\chi \triangleq \boldsymbol{\theta}_n(1+\epsilon_3)$, $\tau = \boldsymbol{\alpha}_n^2(1+\epsilon_1)$, $\rho \triangleq 2\overline{\lambda}\chi\boldsymbol{\alpha}_n^2$, and $\Theta_{\mathsf{x}}^t \triangleq \rho\mathsf{L}_n^t\|\Delta_n^t\|_2^2$. We have:*

(a) $\|\mathbb{w}_n^{t+1} - \mathbb{w}_n^t\|_2^2 \leq \iota$.

(b) $\frac{1}{\beta^t}\|\mathbb{u}_n^{t+1}\|_2^2 \leq 2\overline{\lambda} \cdot \{(\chi - \underline{\lambda}/\overline{\lambda})^2 + \chi\tau\} \cdot \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 + \Theta_{\mathsf{x}}^t - \Theta_{\mathsf{x}}^{t+1}$.

Utilizing the parameters $\{C_{\mathsf{a}}, C_{\mathsf{u}}, C_{\mathsf{x}}\}$, we construct a sequence that is associated with the potential (or Lyapunov) function

as follows:

$$\Theta^t = \Theta_{\mathsf{o}}^t + \tfrac{C_{\mathsf{a}}}{\beta^t}\|\mathbb{a}^t\|_2^2 + \tfrac{C_{\mathsf{u}}}{\beta^t}\|\mathbb{u}_n^t\|_2^2 + C_{\mathsf{x}}\mathsf{L}_n^t\|\Delta_n^t\|_2^2. \tag{114}$$

Here, the parameters $\{C_{\mathsf{a}}, C_{\mathsf{u}}, C_{\mathsf{x}}\}$ remain consistent with those defined in (21) for Condition $\boxed{\mathbb{I}}$ and in (22) for Condition $\boxed{\mathbb{A}}$. The sole difference is the utilization of an alternate ι, as shown in **Part (a)** of Lemma F.6.

F.3. Analysis for Condition $\boxed{\mathbb{I}}$ (Stochastic Settings)

We offer a convergence analysis under Condition $\boxed{\mathbb{I}}$, where $\mathbf{A}_n$ is an identity matrix. We assume that $\sigma \in [1, 2)$.

The following lemma provides an upper bound for the term $\frac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2$ using Equation (19).

Lemma F.7. *(*Bounding Dual Using Primal for Stochastic Settings) We define $\delta \triangleq 1 + \epsilon_2$, $\chi \triangleq \boldsymbol{\theta}_n(1 + \epsilon_3)$, and $\tau \triangleq \boldsymbol{\alpha}_n^2(1 + \epsilon_1)$. We have:

$$\begin{aligned}&\tfrac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2\\ \leq\ & \Theta_{\mathsf{z}}^t - \Theta_{\mathsf{z}}^{t+1} + \tfrac{C_{\mathsf{w}}}{\beta^t} + \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 \cdot 4C_{\mathsf{u}}\{(\chi - 1)^2 + \chi\tau\},\end{aligned}$$

where $\Theta_{\mathsf{z}}^t \triangleq \frac{C_{\mathsf{a}}}{\beta^t}\|\mathbb{a}^t\|_2^2 + \frac{C_{\mathsf{u}}}{\beta^t}\|\mathbb{u}_n^t\|_2^2 + C_{\mathsf{x}}\mathsf{L}_n^t\|\Delta_n^t\|_2^2$, and $\{C_{\mathsf{a}}, C_{\mathsf{u}}, C_{\mathsf{x}}, C_{\mathsf{w}}\}$ are defined in Equation (21).

Remarks. Lemma F.7 closely parallels Lemma 3.4, with the sole difference being the use of a distinct coefficient C_{w} since it depends on the coefficient ι as shown in **Part (a)** of Lemma F.6. The proof is omitted for brevity.

We have the following theorem.

Theorem F.8. *(Proof in Appendix G.4,* Decrease on a Potential Function and a Square-Summable Property under Stochastic Settings) Let $p \in (1, 2]$. For all t, we obtain:

$$\mathbb{E}[\textstyle\sum_{i=0}^{\infty}\mathcal{E}^{i+1}] \leq \Theta^0 - \underline{\Theta} + C_{\mathsf{w}}C_{\mathsf{b}} \triangleq C_{\mathsf{p}}. \tag{115}$$

F.4. Analysis for Condition $\boxed{\mathbb{A}}$ (Stochastic Settings)

We provide the convergence analysis under Condition $\boxed{\mathbb{A}}$, where $\mathbf{A}_n$ is a full-row rank matrix with $\underline{\lambda} > 0$. We assume $\sigma \in (0, 1)$.

The following lemma establishes an upper bound for the term $\frac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2$ using Equation (20).

Lemma F.9. *(*Bounding Dual Using Primal for Stochastic Settings) We define $\delta \triangleq 1 + \epsilon_2$, $\chi \triangleq \boldsymbol{\theta}_n(1 + \epsilon_3)$, and $\tau \triangleq \boldsymbol{\alpha}_n^2(1 + \epsilon_1)$. We have:

$$\begin{aligned}&\tfrac{\delta}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2\\ \leq\ & \Theta_{\mathsf{z}}^t - \Theta_{\mathsf{z}}^{t+1} + \tfrac{C_{\mathsf{w}}}{\beta} + \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 \cdot 2\overline{\lambda}C_{\mathsf{u}} \cdot (\chi^2 + \chi\tau),\end{aligned}$$

where $\Theta_{\mathsf{z}}^t \triangleq \frac{C_{\mathsf{a}}}{\beta^t}\|\mathbb{a}^t\|_2^2 + \frac{C_{\mathsf{u}}}{\beta^t}\|\mathbb{u}_n^t\|_2^2 + C_{\mathsf{x}}\mathsf{L}_n^t\|\Delta_n^t\|_2^2$, and $\{C_{\mathsf{a}}, C_{\mathsf{u}}, C_{\mathsf{x}}, C_{\mathsf{w}}\}$ are defined in Equation (22).

Remarks. Lemma F.9 closely parallels Lemma 3.6, with the sole distinction being the utilization of a distinct coefficient C_{w} since it depends on the coefficient ι as shown in **Part (a)** of Lemma F.6. The proof is omitted for brevity.

We obtain the following theorem.

Theorem F.10. *(Proof in Appendix G.5,* Decrease on a Potential Function and a Square-Summable Property under Stochastic Settings) Let $p \in (1, 2]$. For all t, we obtain:

$$\mathbb{E}[\textstyle\sum_{i=0}^{\infty}\mathcal{E}^{i+1}] \leq \Theta^0 - \underline{\Theta} + C_{\mathsf{w}}C_{\mathsf{b}} \triangleq C_{\mathsf{p}}. \tag{116}$$

F.5. Continuing Analysis for Conditions $\boxed{\mathbb{I}}$ and $\boxed{\mathbb{A}}$ (Stochastic Settings)

Finally, we have the following theorem.

Theorem F.11. *(Proof in Appendix G.6) Let the sequence $\{\mathbf{x}^t, \mathbf{y}^t, \mathbf{z}^t\}_{t=0}^T$ be generated by Algorithm 3. Let $p \in (1, 2]$. We have:*

***(b)** There exists an index $\bar{t}$ with $0 \leq \bar{t} \leq T$ such that $\mathbb{E}[\beta^{\bar{t}}\|\mathbf{r}^{\bar{t}+1}\|_2^2 + \beta^{\bar{t}}\sum_{i=1}^n[\|\mathbf{x}_i^{\bar{t}+1} - \mathbf{x}_i^{\bar{t}}\|_2^2 + \|\mathbf{y}_i^{\bar{t}+1} - \mathbf{y}_i^{\bar{t}}\|_2^2]] \leq \frac{C_{\mathsf{p}}\max(1/c_1, 1/c_2)}{T}$, where $c_0 \triangleq \epsilon_3 \min_{i=1}^n \boldsymbol{\gamma}_i\|\mathbf{A}_i\|$, $c_1 \triangleq \frac{c_0}{17}$, $c_2 \triangleq \frac{\xi}{2}$, and C_{p} is defined in Inequality (115) or Inequality (116). It implies that Algorithm 3 finds an ϵ-INP point of Problem (1) in at most T iterations in the sense of expectation, where $T \leq \lceil \frac{\max(1/c_1, 1/c_2, 1/c_3)C_{\mathsf{p}}}{\epsilon} \rceil = \mathcal{O}(\epsilon^{-1})$.*

(c) Let N denote the number of data points for the finite-sum structure as in Equation (102). Let $q = \sqrt{N}$ and $b = \sqrt{N}$ for Algorithm 3. The overall stochastic first-order oracl complexity is $\mathcal{O}(\sqrt{N}\epsilon^{-1} + N)$.

G. Proofs for Stochastic IRPL-ADMM

G.1. Proof of Lemma F.3

Proof. The proof of this lemma shares a similar structure with that of Lemma 3.1. To keep it concise, we will primarily focus on highlighting the main difference.

For notation convenience, we define $\Gamma^t \triangleq \langle \mathbf{x}_n^{t+1} - \mathbf{x}_n^t, \nabla_{\mathbf{x}_n} G(\mathbf{x}_{[1,s-1]}^{t+1}, \mathbf{x}_{[n,n]}^t, \mathbf{z}^t; \beta^t) - \nabla_{\mathbf{x}_n}\tilde{G}(\mathbf{x}_{[1,s-1]}^{t+1}, \mathbf{x}_{[n,n]}^t, \mathbf{z}^t; \beta^t)\rangle$.

Initially, we bound the term Γ^t using the following inequalities:

$$\begin{aligned}
\Gamma_n^t &= \langle \mathbf{x}_n^{t+1} - \mathbf{x}_n^t, \nabla_{\mathbf{x}_n} G(\mathbf{x}_{[1,s-1]}^{t+1}, \mathbf{x}_{[n,n]}^t, \mathbf{z}^t; \beta^t) - \nabla_{\mathbf{x}_n}\tilde{G}(\mathbf{x}_{[1,s-1]}^{t+1}, \mathbf{x}_{[n,n]}^t, \mathbf{z}^t; \beta^t)\rangle \\
&= \langle \mathbf{x}_n^{t+1} - \mathbf{x}_n^t, \nabla f_n(\mathbf{x}_n^t) - \mathbf{v}^t\rangle \\
&\overset{①}{\leq} \tfrac{\epsilon_3 L_n^t}{2}\|\mathbf{x}_n^{t+1} - \mathbf{x}_n^t\|_2^2 + \tfrac{1}{2\epsilon_3 L_n^t}\|\nabla f_n(\mathbf{x}_n^t) - \mathbf{v}^t\|_2^2,
\end{aligned} \tag{117}$$

where step ① uses Lemma A.5.

(a) We now establish the decrease in the objective function value for the subproblem of the n-th block. Noticing the function $G(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_n^t, \mathbf{z}^t; \beta^t)$ is L_n^t-smooth *w.r.t.* $\mathbf{x}_n$ for the t-th iteration, we have

$$\begin{aligned}
& G(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_n^{t+1}, \mathbf{z}^t; \beta^t) \\
\leq\ & G(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_n^t, \mathbf{z}^t; \beta^t) + \langle \mathbf{x}_n^{t+1} - \mathbf{x}_n^t, \nabla_{\mathbf{x}_n} G(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_n^t, \mathbf{z}^t; \beta^t)\rangle + \tfrac{\mathsf{L}_n^t}{2}\|\mathbf{x}_n^{t+1} - \mathbf{x}_n^t\|_2^2.
\end{aligned} \tag{118}$$

Second, we notice that $\mathbf{x}_n^{t+1}$ is the minimizer of the following optimization problem:

$$\mathbf{x}_n^{t+1} \in \arg\min_{\mathbf{x}_n}\ h_n(\mathbf{x}_n) + \langle \mathbf{x}_n - \mathbf{x}_n^t, \nabla_{\mathbf{x}_n}\tilde{G}(\mathbf{x}_{[1,s-1]}^{t+1}, \mathbf{x}_{[n,n]}^t, \mathbf{z}^t; \beta^t)\rangle + \tfrac{\boldsymbol{\theta}_n \mathsf{L}_n^t}{2}\|\mathbf{x}_n - \mathbf{y}_n^t\|_2^2. \tag{119}$$

Here, $\nabla_{\mathbf{x}_n}\tilde{G}(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_{[n]}^t, \mathbf{z}^t; \beta^t)$ can be viewed as a unbiased estimation of the true value of $\nabla_{\mathbf{x}_n} G(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_{[n]}^t, \mathbf{z}^t; \beta^t) = \nabla f_n(\mathbf{x}_n^t) + \mathbf{A}_n^\mathsf{T}\mathbf{z}^t + \beta^t \mathbf{A}_n^\mathsf{T}(\mathbf{A}_n\mathbf{x}_n^t + [\sum_{i=1}^{n-1}\mathbf{A}_i\mathbf{x}_i^{t+1}] - \mathbf{b})$.

Using the optimality of $\mathbf{x}_n^{t+1}$ as in (119), we have

$$\begin{aligned}
& h_n(\mathbf{x}_n^{t+1}) - h_n(\mathbf{x}_n^t) + \langle \mathbf{x}_n^{t+1} - \mathbf{x}_n^t, \nabla_{\mathbf{x}_n}\tilde{G}(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_{[n]}^t, \mathbf{z}^t; \beta^t)\rangle \\
\leq\ & \tfrac{\boldsymbol{\theta}_n L_n^t}{2}\|\mathbf{x}_n^t - \mathbf{y}_n^t\|_2^2 - \tfrac{\boldsymbol{\theta}_n \mathsf{L}_n^t}{2}\|\mathbf{x}_n^{t+1} - \mathbf{y}_n^t\|_2^2 \\
=\ & -\tfrac{\boldsymbol{\theta}_n \mathsf{L}_n^t}{2}\|\mathbf{x}_n^{t+1} - \mathbf{x}_n^t\|_2^2 - \boldsymbol{\theta}_n \mathsf{L}_n^t\langle \mathbf{x}_n^t - \mathbf{y}_n^t, \mathbf{x}_n^{t+1} - \mathbf{x}_n^t\rangle.
\end{aligned} \tag{120}$$

Adding (118) and (120) together, we obtain the decrease in the objective function value for the subproblem of the n-th block:

$$\begin{aligned}
& h_n(\mathbf{x}_n^{t+1}) + G(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_n^{t+1}, \mathbf{z}^t; \beta^t) - h_n(\mathbf{x}_n^t) - G(\mathbf{x}_{[1,n-1]}^{t+1}, \mathbf{x}_n^t, \mathbf{z}^t; \beta^t) \\
\leq\ & -\tfrac{(\boldsymbol{\theta}_n - 1)\mathsf{L}_n^t}{2}\|\mathbf{x}_n^{t+1} - \mathbf{x}_n^t\|_2^2 - \boldsymbol{\theta}_n \mathsf{L}_n^t\langle \mathbf{x}_n^t - \mathbf{y}_n^t, \mathbf{x}_n^{t+1} - \mathbf{x}_n^t\rangle + \Gamma_n^t \\
\overset{①}{=}\ & -(\boldsymbol{\theta}_n - 1 - \boldsymbol{\theta}_n\boldsymbol{\alpha}_n(2 + \xi)) \cdot \tfrac{L_n^t}{2}\|\Delta_n^{t+1}\|_2^2 + \boldsymbol{\theta}_n\boldsymbol{\alpha}_n \cdot (\tfrac{\mathsf{L}_n^t}{2}\|\Delta_n^t\|_2^2 - \tfrac{\mathsf{L}_n^{t+1}}{2}\|\Delta_n^{t+1}\|_2^2) + \Gamma^t,
\end{aligned} \tag{121}$$

where step ① uses the same strategy as in deriving Inequality (37).

(b) We now establish the decrease in the objective function value for the subproblem of the i-th block with $i \neq s$. Using the same strategy as in deriving Inequality (37), we have the following inequality for all $i \neq n$:

$$\begin{aligned} & h_i(\mathbf{x}_i^{t+1}) + G(\mathbf{x}_{[1,i-1]}^{t+1}, \mathbf{x}_i^{t+1}, \mathbf{x}_{[i+1,n]}^{t}, \mathbf{z}^t; \beta^t) - h_i(\mathbf{x}_i^t) - G(\mathbf{x}_{[1,i-1]}^{t+1}, \mathbf{x}_i^{t}, \mathbf{x}_{[i+1,n]}^{t}, \mathbf{z}^t; \beta^t) \\ \leq & -(\boldsymbol{\theta}_i - 1 - \boldsymbol{\theta}_i\boldsymbol{\alpha}_i(2+\xi)) \cdot \tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^{t+1}\|_2^2 + \boldsymbol{\theta}_i\boldsymbol{\alpha}_i \cdot (\tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^t\|_2^2 - \tfrac{\mathsf{L}_i^{t+1}}{2}\|\Delta_i^{t+1}\|_2^2). \end{aligned} \tag{122}$$

(c) We now establish the decrease for $\mathcal{L}(\mathbf{x}^{t+1}, \mathbf{z}^t; \beta^t) - \mathcal{L}(\mathbf{x}^t, \mathbf{z}^t; \beta^t)$. In view of Inequality (121) and (122), for all $i \in [n]$, we define

$$\Lambda_i^t \triangleq -(\boldsymbol{\theta}_i - 1 - \boldsymbol{\theta}_i\boldsymbol{\alpha}_i(2+\xi)) \cdot \tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^{t+1}\|_2^2 + \boldsymbol{\theta}_i\boldsymbol{\alpha}_i \cdot (\tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^t\|_2^2 - \tfrac{\mathsf{L}_i^{t+1}}{2}\|\Delta_i^{t+1}\|_2^2) - h_i(\mathbf{x}_i^{t+1}) + h_i(\mathbf{x}_i^t). \tag{123}$$

Using the same strategy as in deriving Inequality (39), we have the following inequality:

$$G(\mathbf{x}^{t+1}, \mathbf{z}^t; \beta^t) - G(\mathbf{x}^t, \mathbf{z}^t; \beta^t) \leq \Gamma^t + \textstyle\sum_{i=1}^n \Lambda_i^t.$$

Uses the definition of $\mathcal{L}(\mathbf{x}, \mathbf{z}; \beta) \triangleq G(\mathbf{x}, \mathbf{z}; \beta) + \sum_{i=1}^n h_i(\mathbf{x}_i)$, we derive the following results:

$$\begin{aligned} & \mathcal{L}(\mathbf{x}^{t+1}, \mathbf{z}^t; \beta^t) - \mathcal{L}(\mathbf{x}^t, \mathbf{z}^t; \beta^t) \\ = & \{G(\mathbf{x}^{t+1}, \mathbf{z}^t; \beta^t) + \textstyle\sum_{i=1}^n h_i(\mathbf{x}_i^{t+1})\} - \{G(\mathbf{x}^t, \mathbf{z}^t; \beta^t) + \textstyle\sum_{i=1}^n h_i(\mathbf{x}_i^t)\} \\ \overset{①}{\leq} & \Gamma^t + \textstyle\sum_{i=1}^n \{\Lambda_i^t + h_i(\mathbf{x}_i^{t+1}) - h_i(\mathbf{x}_i^t)\} \\ \overset{②}{=} & \Gamma^t + \textstyle\sum_{i=1}^n \{-(\boldsymbol{\theta}_i - 1 - \boldsymbol{\theta}_i\boldsymbol{\alpha}_i(2+\xi)) \cdot \tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^{t+1}\|_2^2 + \tfrac{1}{2}\boldsymbol{\theta}_i\boldsymbol{\alpha}_i(\mathsf{L}_i^t\|\Delta_i^t\|_2^2 - \mathsf{L}_i^{t+1}\|\Delta_i^{t+1}\|_2^2)\} \end{aligned} \tag{124}$$

where step ① uses Inequality (124); and step ② uses the definition of Λ_i^t in Equation (123).

Using the same strategy as in deriving Inequality (42), we have:

$$\tfrac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2 + \mathcal{L}(\mathbf{x}^{t+1}, \mathbf{z}^{t+1}; \beta^{t+1}) - \mathcal{L}(\mathbf{x}^{t+1}, \mathbf{z}^t; \beta^t) \leq \|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 \cdot \tfrac{1}{\sigma\beta^t} \cdot \{1 + \epsilon_2\}. \tag{125}$$

We define $\mathbf{r}^{t+1} \triangleq [\sum_{i=1}^n \mathbf{A}_i\mathbf{x}_i^{t+1}] - \mathbf{b}$ and $\Theta_\circ^t \triangleq \mathcal{L}(\mathbf{x}^t, \mathbf{z}^t; \beta^t) + \frac{1}{2}\sum_{i=1}^n \boldsymbol{\theta}_i\boldsymbol{\alpha}_i\mathsf{L}_i^t\|\Delta_i^t\|_2^2$. Combining Inequalities (124) and (125), we have the following inequalities:

$$\begin{aligned} & \tfrac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2 + \Theta_\circ^{t+1} - \Theta_\circ^t \\ \leq & \tfrac{1+\epsilon_2}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 - \{\textstyle\sum_{i=1}^n (\boldsymbol{\theta}_i - 1 - \boldsymbol{\theta}_i\boldsymbol{\alpha}_i(2+\xi)) \cdot \tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^{t+1}\|_2^2\} + \Gamma^t \\ \overset{①}{\leq} & \tfrac{1+\epsilon_2}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 - \{\textstyle\sum_{i=1}^n (\boldsymbol{\theta}_i - 1 - \boldsymbol{\theta}_i\boldsymbol{\alpha}_i(2+\epsilon_1)) \cdot \tfrac{\mathsf{L}_i^t}{2}\|\Delta_i^{t+1}\|_2^2\} + \Gamma^t \\ \overset{②}{=} & \tfrac{1+\epsilon_2}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 - \{\textstyle\sum_{i=1}^n \boldsymbol{\gamma}_i\mathsf{L}_i^t\|\Delta_i^{t+1}\|_2^2\} + \Gamma^t, \end{aligned} \tag{126}$$

where step ① uses $\xi \leq \epsilon_1$ as shown in Assumption 2.7, and step ② uses the definition of $\boldsymbol{\gamma}_i \triangleq \frac{1}{2} \cdot (\boldsymbol{\theta}_i - 1 - \boldsymbol{\theta}_i\boldsymbol{\alpha}_i(2+\epsilon_1))$ for all $i \in [n]$.

Finally, we obtain:

$$\begin{aligned} & \tfrac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2 + \epsilon_3[\textstyle\sum_{i=1}^n \boldsymbol{\gamma}_i\mathsf{L}_i^t\|\Delta_i^{t+1}\|_2^2] + \Theta_\circ^{t+1} - \Theta_\circ^t \\ \overset{①}{\leq} & \tfrac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2 + [\textstyle\sum_{i=1}^{n-1} \boldsymbol{\gamma}_i\mathsf{L}_i^t\|\Delta_i^{t+1}\|_2^2] + \epsilon_3\boldsymbol{\gamma}_n\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 + \Theta_\circ^{t+1} - \Theta_\circ^t \\ \overset{②}{\leq} & \tfrac{1+\epsilon_2}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 + (\epsilon_3 - 1)\boldsymbol{\gamma}_n\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 + \Theta_\circ^{t+1} - \Theta_\circ^t + \tfrac{\epsilon_3 L_n^t}{2}\|\mathbf{x}_n^{t+1} - \mathbf{x}_n^t\|_2^2 + \tfrac{1}{2\epsilon_3 L_n^t}\|\nabla f_n(\mathbf{x}_n^t) - \mathbf{v}^t\|_2^2 \\ \overset{③}{=} & \tfrac{1+\epsilon_2}{\sigma\beta^t}\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2^2 - \{(1-\epsilon_3)\boldsymbol{\gamma}_n - \tfrac{\epsilon_3}{2}\}\mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 + \Theta_\circ^{t+1} - \Theta_\circ^t + \tfrac{1}{2\epsilon_3\mathsf{L}_n^t}\|\nabla f_n(\mathbf{x}_n^t) - \mathbf{v}^t\|_2^2, \end{aligned}$$

where step ① uses $\epsilon_3 \leq 1$ as shown in Assumption 2.7; and step ② uses Inequality (126).

□

G.2. Proof of Lemma F.4

Proof. Initially, we derive the following inequalities:

$$\frac{L_n}{\mathsf{L}_n^t} \overset{①}{=} \frac{L_n}{L_n+\beta^t\|\mathbf{A}_n\|_2^2} \overset{②}{\leq} \frac{L_n}{\beta^0\|\mathbf{A}_n\|_2^2} \overset{③}{\leq} \frac{L_n}{\|\mathbf{A}_n\|_2^2\cdot(L_n/\epsilon_3\|\mathbf{A}_n\|_2^2)} = \epsilon_3, \tag{127}$$

where step ① uses the definition of $\mathsf{L}_i^t \triangleq L_i+\beta^t\|\mathbf{A}_i\|_2^2$; step ② uses $\beta^t \geq \beta^0$; step ③ uses the choice of $\beta^0 \geq L_n/(\epsilon_3\|\mathbf{A}_n\|_2^2)$.

Telescoping Inequality (111) over t from $(r_t-1)q+1$ to t, where $t \leq r_tq-1$, we obtain

$$\begin{aligned}
& \tfrac{1}{2\epsilon_3\mathsf{L}_n^t}\cdot\mathbb{E}[\|\nabla f_n(\mathbf{x}_n^t)-\mathbf{v}^t\|_2^2] \\
\leq\ & \tfrac{1}{2\epsilon_3\mathsf{L}_n^t}\cdot\left(\mathbb{E}[\|\mathbf{v}^{(r_t-1)q}-\nabla f_n(\mathbf{x}_n^{(r_t-1)q})\|_2^2]+\tfrac{(L_n)^2}{b}\textstyle\sum_{i=(r_t-1)q}^{t-1}\mathbb{E}[\|\mathbf{x}_n^{i+1}-\mathbf{x}_n^i\|_2^2]\right) \\
\overset{①}{=}\ & 0+\tfrac{1}{2b\epsilon_3}\cdot\tfrac{(L_n)^2}{(\mathsf{L}_n^t)^2}\cdot\mathsf{L}_n^t\textstyle\sum_{i=(r_t-1)q}^{t-1}\mathbb{E}[\|\mathbf{x}_n^{i+1}-\mathbf{x}_n^i\|_2^2] \\
\overset{④}{=}\ & \tfrac{\epsilon_3}{2b}\textstyle\sum_{i=(r_t-1)q}^{t-1}\mathbb{E}[\mathsf{L}_n^t\|\mathbf{x}_n^{i+1}-\mathbf{x}_n^i\|_2^2],
\end{aligned} \tag{128}$$

where step ① uses $\mathbf{v}^j=\nabla f_n(\mathbf{x}_n^j)$ for all $j=(r_t-1)q$; step ② uses Inequality (127).

We notice that (128) holds true for $t=(r_t-1)q$, which can be checked by plugging $t=(r_t-1)q$ into the inequality. Therefore, (128) holds true for all $(r_t-1)q \leq t \leq r_tq-1$.

□

G.3. Proof of Lemma F.6

Proof. We denote $r_t \geq 1$ be the unique integer such that $(r_t-1)q \leq t \leq \mathbf{r}_tq-1$.

Using the update rule of SPIDER: $\mathbf{v}^t=\begin{cases}\nabla f_n(\mathbf{x}_n^t), & \text{mod}(t,q)=0;\\ \mathbf{v}^{t-1}+\nabla f_n(\mathbf{x}_n^t;\mathcal{I}^t)-\nabla f_n(\mathbf{x}_n^{t-1};\mathcal{I}^{t-1}), & \text{else.}\end{cases}$, for any t with $(r_t-1)q \leq t \leq r_tq-1$, we have:

$$\mathbb{v}^t=\nabla f_n(\mathbf{x}_n^{r_t})+\textstyle\sum_{i=1+(r_t-1)q}^{t}\{\nabla f_n(\mathbf{x}_n^i;\mathcal{I}^i)-\nabla f_n(\mathbf{x}_n^{i-1};\mathcal{I}^{i-1})\}. \tag{129}$$

We bound the term $\|\mathbf{v}^t\|_2^2$. We derive the following inequalities for any t:

$$\begin{aligned}
\|\mathbb{v}^t\|_2^2 &\overset{①}{=} \|\nabla f_n(\mathbf{x}_n^{r_t})+\textstyle\sum_{i=(r_t-1)q+1}^{t}(\nabla f_n(\mathbf{x}_n^t;\mathcal{I}^t)-\nabla f_n(\mathbf{x}_n^{t-1};\mathcal{I}^{t-1}))\|_2^2 \\
&\overset{②}{\leq} \{1+2\times(q-1)\}C_f^2, \\
&\leq 2qC_f^2,
\end{aligned} \tag{130}$$

where step ① uses Equality (129); step ② uses the fact that the term $\sum_{i=(r_t-1)q+1}^{t}(\mathbf{r}^i-\mathbf{r}^{i-1})$ involves at most $2(q-1)$ evaluations of the gradient for the mini-batch of data, and the inequality that $\sum_{i=1}^n\|\mathbf{a}_i\|_2^2 \leq n\max_{i=1}^n\|\mathbf{a}_i\|_2^2$ for any vectors $\{\mathbf{a}_1,\ldots,\mathbf{a}_n\}$.

We have:

$$\begin{aligned}
\|\mathbb{w}_n^{t+1}-\mathbb{w}_n^t\|_2^2 &\overset{①}{=} \|\partial h_n(\mathbf{x}_n^{t+1})+\mathbb{v}^{t+1}-\partial h_n(\mathbf{x}_n^t)-\mathbf{v}^t\|_2^2 \\
&\leq 4\|\partial h_n(\mathbf{x}_n^{t+1})\|_2^2+4\|\partial h_n(\mathbf{x}_n^t)\|_2^2+4\|\mathbb{v}^{t+1}\|_2^2+4\|\mathbb{v}^t\|_2^2 \\
&\overset{②}{\leq} 4C_h^2+4C_h^2+8qC_f^2+8qC_f^2 \\
&\leq 8C_h^2+16qC_f^2,
\end{aligned}$$

where step ① uses $\mathbb{w}_n^{t+1} \in \partial h_n(\mathbf{x}_n^{t+1})+\mathbf{v}^{t+1}$; step ② uses Inequality (130).

The second part of this lemma is identical to **Part (b)** in Lemma 3.3; hence, we omit its proof for brevity.

□

G.4. Proof of Theorem F.8

Proof. We define $\boldsymbol{\gamma}_n' \triangleq (1-\epsilon_3)\boldsymbol{\gamma}_n - \frac{\epsilon_3}{2}$, and $W \triangleq 4C_{\mathsf{u}}\{(\chi-1)^2+\chi\tau\} - \{(1-\epsilon_3)\boldsymbol{\gamma}_n - \frac{\epsilon_3}{2}\}$

We define $\mathcal{E}^{t+1} \triangleq \frac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2 + \epsilon_3[\sum_{i=1}^n \boldsymbol{\gamma}_i \mathsf{L}_i^t \|\Delta_i^{t+1}\|_2^2]$, $\mathcal{E}_\star^{t+1} \triangleq \mathcal{E}^{t+1} - \frac{C_{\mathsf{w}}}{\beta^t}$, and $\Theta^t \triangleq \Theta_{\mathsf{o}}^t + \Theta_{\mathsf{z}}^t$.

Using Lemma F.3 and Lemma F.7, we have:

$$
\begin{aligned}
& \mathcal{E}_\star^{t+1} + \Theta^{t+1} - \Theta^t \\
\leq\ & -\{(1-\epsilon_3)\boldsymbol{\gamma}_n - \tfrac{\epsilon_3}{2}\}\mathsf{L}_n^t\|\mathbf{x}_n^{t+1}-\mathbf{x}_n^t\|^2 + \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 \cdot 4C_{\mathsf{u}}\{(\chi-1)^2+\chi\tau\} + \tfrac{1}{2\epsilon_3\mathsf{L}_n^t}\|\nabla f_n(\mathbf{x}_n^t) - \mathbf{v}^t\|_2^2 \\
\overset{①}{\leq}\ & -\{(1-\epsilon_3)\boldsymbol{\gamma}_n - \tfrac{\epsilon_3}{2}\}\mathsf{L}_n^t\|\mathbf{x}_n^{t+1}-\mathbf{x}_n^t\|^2 + \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 \cdot 4C_{\mathsf{u}}\{(\chi-1)^2+\chi\tau\} + \tfrac{\epsilon_3}{2b}\textstyle\sum_{i=(r_t-1)q}^{t-1}\mathbb{E}[L_n^t\|\Delta_n^{i+1}\|_2^2] \\
\overset{②}{\leq}\ & W\mathsf{L}_n^t\|\mathbf{x}_n^{t+1}-\mathbf{x}_n^t\|^2 + \tfrac{\epsilon_3}{2b}\textstyle\sum_{i=(r_t-1)q}^{t-1}\mathbb{E}[L_n^t\|\Delta_n^{i+1}\|_2^2], \qquad (131)
\end{aligned}
$$

where step ① uses the upper bound of $\frac{1}{2\epsilon_3\mathsf{L}_n^t}\|\nabla f_n(\mathbf{x}_n^t) - \mathbf{v}^t\|_2^2$ as presented in Lemma F.4; step ② uses the definition of W.

Telescoping the inequality in (131) over t from $(r_t-1)q$ to t where $t \leq r_t q - 1$, we have:

$$
\begin{aligned}
& \mathbb{E}[\textstyle\sum_{j=(r_t-1)q}^t \mathcal{E}_\star^{j+1} + \sum_{j=(r_t-1)q}^t(\Theta^{j+1}-\Theta^j)] \\
=\ & \tfrac{\epsilon_3}{2b}\textstyle\sum_{j=(r_t-1)q}^t\sum_{i=[r_j-1]q}^{j-1}\mathbb{E}[\mathsf{L}_n^j\|\mathbf{x}_n^{i+1}-\mathbf{x}_n^i\|_2^2] + \sum_{j=(r_t-1)q}^t W\mathsf{L}_n^j\|\Delta_n^{j+1}\|_2^2 \\
\overset{①}{=}\ & \tfrac{\epsilon_3}{2b}\textstyle\sum_{j=(r_t-1)q}^t\sum_{i=[r_t-1]q}^{j-1}\mathbb{E}[\mathsf{L}_n^j\|\mathbf{x}_n^{i+1}-\mathbf{x}_n^i\|_2^2] + \sum_{j=(r_t-1)q}^t W\mathsf{L}_n^j\|\Delta_n^{j+1}\|_2^2 \\
\overset{②}{\leq}\ & \tfrac{\epsilon_3}{2b}\textstyle\sum_{j=(r_t-1)q}^t\sum_{i=[r_t-1]q}^{t}\mathbb{E}[\mathsf{L}_n^j\|\mathbf{x}_n^{i+1}-\mathbf{x}_n^i\|_2^2] + \sum_{j=(r_t-1)q}^t W\mathsf{L}_n^j\|\Delta_n^{j+1}\|_2^2 \\
=\ & \tfrac{\epsilon_3}{2b}\textstyle\sum_{j=(r_t-1)q}^t(t-j+1)\mathbb{E}[\mathsf{L}_n^j\|\mathbf{x}_n^{j+1}-\mathbf{x}_n^j\|_2^2] + \sum_{j=(r_t-1)q}^t W\mathsf{L}_n^j\|\Delta_n^{j+1}\|_2^2 \\
\overset{③}{\leq}\ & \tfrac{\epsilon_3}{2b}\textstyle\sum_{j=(r_t-1)q}^t q\mathsf{L}_n^j\|\Delta_n^{j+1}\|_2^2 + \sum_{j=(r_t-1)q}^t W\mathsf{L}_n^j\|\Delta_n^{j+1}\|_2^2 \\
\overset{④}{\leq}\ & 0,
\end{aligned}
$$

where step ① uses $r_j = r_t$ for all $(r_t-1)q \leq j \leq r_t q - 1$; step ② uses the extension the summation of the second term from $j-1$ to t since $j-1 < j \leq t$; step ③ uses $t-j+1 \leq q$; step ④ uses the inequality that:

$$
\tfrac{\epsilon_3 q}{2b} \leq W \triangleq 4C_{\mathsf{u}}\{(\chi-1)^2+\chi\tau\} - \{(1-\epsilon_3)\boldsymbol{\gamma}_n - \tfrac{\epsilon_3}{2}\}, \qquad (132)
$$

which holds due to Inequality (108).

Using the fact that $\sum_{j=(r_t-1)q}^t(\Theta^{j+1}-\Theta^j) = \Theta^{t+1} - \Theta^{(r_t-1)q}$, we have:

$$
\mathbb{E}[\Theta^{t+1} - \Theta^{(r_t-1)q}] \leq \mathbb{E}[-\textstyle\sum_{j=(r_t-1)q}^t \mathcal{E}_\star^{j+1}]. \qquad (133)
$$

Telescoping the inequality above over t from 0 to T, we have:

$$
\begin{aligned}
\mathbb{E}[\Theta^{T+1} - \Theta^0] &= \mathbb{E}[(\Theta^q - \Theta^0) + (\Theta^{2q} - \Theta^q) + \ldots + (\Theta^t - \Theta^{(r_t-1)q})] \\
&\leq \mathbb{E}[\Big(-\textstyle\sum_{i=0}^{q-1} - \mathcal{E}_\star^{i+1}\Big) + \Big(-\sum_{i=q}^{2q-1}\mathcal{E}_\star^{i+1}\Big) + \ldots \Big(-\sum_{i=(r_t-1)q}^{t-1}\mathcal{E}_\star^{i+1}\Big)] \\
&= \mathbb{E}[-\textstyle\sum_{i=0}^t \mathcal{E}_\star^{i+1}] \\
&\overset{①}{=} \mathbb{E}[\textstyle\sum_{i=0}^t(\frac{C_{\mathsf{w}}}{\beta^t} - \mathcal{E}^{t+1})] \\
&\overset{②}{\leq} \mathbb{E}[-\textstyle\sum_{i=0}^t \mathcal{E}^{i+1} + C_{\mathsf{w}}C_{\mathsf{b}}],
\end{aligned}
$$

where step ① uses the definition of $\mathcal{E}_\star^{t+1} \triangleq \mathcal{E}^{t+1} - \frac{C_{\mathsf{w}}}{\beta^t}$; step ② uses Inequality (3). Using the fact that $\Theta^{T+1} \geq \underline{\Theta}$ as shown in Lemma 3.8, we conclude this lemma. □

G.5. Proof of Theorem F.10

Proof. We define $\boldsymbol{\gamma}_n' \triangleq (1-\epsilon_3)\boldsymbol{\gamma}_n - \frac{\epsilon_3}{2}$, and $W \triangleq 2\overline{\lambda}C_{\mathsf{u}} \cdot (\chi^2+\chi\tau) - \{(1-\epsilon_3)\boldsymbol{\gamma}_n - \frac{\epsilon_3}{2}\}$

We define $\mathcal{E}^{t+1} \triangleq \frac{\xi\beta^t}{2}\|\mathbf{r}^{t+1}\|_2^2 + \epsilon_3[\sum_{i=1}^n \boldsymbol{\gamma}_i \mathsf{L}_i^t \|\Delta_i^{t+1}\|_2^2]$, $\mathcal{E}_\star^{t+1} \triangleq \mathcal{E}^{t+1} - \frac{C_{\mathsf{w}}}{\beta^t}$, and $\Theta^t \triangleq \Theta_{\mathsf{o}}^t + \Theta_{\mathsf{z}}^t$.

Leveraging Lemma F.3 and Lemma F.9, we have:

$$
\begin{aligned}
& \mathcal{E}_\star^{t+1} + \Theta^{t+1} - \Theta^t \\
\leq\ & -\{(1-\epsilon_3)\boldsymbol{\gamma}_n - \tfrac{\epsilon_3}{2}\}\mathsf{L}_n^t\|\mathbf{x}_n^{t+1}-\mathbf{x}_n^t\|^2 + \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 \cdot 2\overline{\lambda}C_{\mathsf{u}} \cdot (\chi^2+\chi\tau) + \tfrac{1}{2\epsilon_3\mathsf{L}_n^t}\|\nabla f_n(\mathbf{x}_n^t) - \mathbf{v}^t\|_2^2 \\
\overset{①}{\leq}\ & -\{(1-\epsilon_3)\boldsymbol{\gamma}_n - \tfrac{\epsilon_3}{2}\}\mathsf{L}_n^t\|\mathbf{x}_n^{t+1}-\mathbf{x}_n^t\|^2 + \mathsf{L}_n^t\|\Delta_n^{t+1}\|_2^2 \cdot 2\overline{\lambda}C_{\mathsf{u}} \cdot (\chi^2+\chi\tau) + \tfrac{\epsilon_3}{2b}\textstyle\sum_{i=(r_t-1)q}^{t-1}\mathbb{E}[L_n^t\|\Delta_n^{i+1}\|_2^2] \\
\overset{②}{\leq}\ & W\mathsf{L}_n^t\|\mathbf{x}_n^{t+1}-\mathbf{x}_n^t\|^2 + \tfrac{\epsilon_3}{2b}\textstyle\sum_{i=(r_t-1)q}^{t-1}\mathbb{E}[L_n^t\|\Delta_n^{i+1}\|_2^2], \qquad (134)
\end{aligned}
$$

where step ① uses the upper bound of $\frac{1}{2\epsilon_3\mathsf{L}_n^t}\|\nabla f_n(\mathbf{x}_n^t) - \mathbf{v}^t\|_2^2$ as presented in Lemma F.4; step ② uses the definition of W.

Telescoping the inequality in (134) over t from $(r_t-1)q$ to t where $t \leq r_t q - 1$, we have:

$$
\begin{aligned}
& \mathbb{E}[\textstyle\sum_{j=(r_t-1)q}^t \mathcal{E}_\star^{j+1} + \sum_{j=(r_t-1)q}^t(\Theta^{j+1}-\Theta^j)] \\
=\ & \tfrac{\epsilon_3}{2b}\textstyle\sum_{j=(r_t-1)q}^t\sum_{i=[r_j-1]q}^{j-1}\mathbb{E}[\mathsf{L}_n^j\|\mathbf{x}_n^{i+1}-\mathbf{x}_n^i\|_2^2] + \sum_{j=(r_t-1)q}^t W\mathsf{L}_n^j\|\Delta_n^{j+1}\|_2^2 \\
\overset{①}{=}\ & \tfrac{\epsilon_3}{2b}\textstyle\sum_{j=(r_t-1)q}^t\sum_{i=[r_t-1]q}^{j-1}\mathbb{E}[\mathsf{L}_n^j\|\mathbf{x}_n^{i+1}-\mathbf{x}_n^i\|_2^2] + \sum_{j=(r_t-1)q}^t W\mathsf{L}_n^j\|\Delta_n^{j+1}\|_2^2 \\
\overset{②}{\leq}\ & \tfrac{\epsilon_3}{2b}\textstyle\sum_{j=(r_t-1)q}^t\sum_{i=[r_t-1]q}^{t}\mathbb{E}[\mathsf{L}_n^j\|\mathbf{x}_n^{i+1}-\mathbf{x}_n^i\|_2^2] + \sum_{j=(r_t-1)q}^t W\mathsf{L}_n^j\|\Delta_n^{j+1}\|_2^2 \\
=\ & \tfrac{\epsilon_3}{2b}\textstyle\sum_{j=(r_t-1)q}^t(t-j+1)\mathbb{E}[\mathsf{L}_n^j\|\mathbf{x}_n^{j+1}-\mathbf{x}_n^j\|_2^2] + \sum_{j=(r_t-1)q}^t W\mathsf{L}_n^j\|\Delta_n^{j+1}\|_2^2 \\
\overset{③}{\leq}\ & \tfrac{\epsilon_3}{2b}\textstyle\sum_{j=(r_t-1)q}^t q\mathsf{L}_n^j\|\Delta_n^{j+1}\|_2^2 + \sum_{j=(r_t-1)q}^t W\mathsf{L}_n^j\|\Delta_n^{j+1}\|_2^2 \\
\overset{④}{\leq}\ & 0,
\end{aligned}
$$

where step ① uses $r_j = r_t$ for all $(r_t-1)q \leq j \leq r_t q - 1$; step ② uses the extension the summation of the second term from $j-1$ to t since $j-1 < j \leq t$; step ③ uses $t-j+1 \leq q$; step ④ uses the inequality that:

$$
\tfrac{\epsilon_3 q}{2b} \leq W \triangleq 2\overline{\lambda}C_{\mathsf{u}} \cdot (\chi^2+\chi\tau) - \{(1-\epsilon_3)\boldsymbol{\gamma}_n - \tfrac{\epsilon_3}{2}\}, \qquad (135)
$$

which holds due to Inequality (109).

Using the fact that $\sum_{j=(r_t-1)q}^t(\Theta^{j+1}-\Theta^j) = \Theta^{t+1} - \Theta^{(r_t-1)q}$, we have:

$$
\mathbb{E}[\Theta^{t+1} - \Theta^{(r_t-1)q}] \leq \mathbb{E}[-\textstyle\sum_{j=(r_t-1)q}^t \mathcal{E}_\star^{j+1}]. \qquad (136)
$$

Telescoping the inequality above over t from 0 to T, we have:

$$
\begin{aligned}
\mathbb{E}[\Theta^{T+1} - \Theta^0] &= \mathbb{E}[(\Theta^q - \Theta^0) + (\Theta^{2q} - \Theta^q) + \ldots + (\Theta^t - \Theta^{(r_t-1)q})] \\
&\leq \mathbb{E}[\Big(-\textstyle\sum_{i=0}^{q-1} - \mathcal{E}_\star^{i+1}\Big) + \Big(-\sum_{i=q}^{2q-1}\mathcal{E}_\star^{i+1}\Big) + \ldots \Big(-\sum_{i=(r_t-1)q}^{t-1}\mathcal{E}_\star^{i+1}\Big)] \\
&= \mathbb{E}[-\textstyle\sum_{i=0}^t \mathcal{E}_\star^{i+1}] \\
&\overset{①}{=} \mathbb{E}[\textstyle\sum_{i=0}^t(\frac{C_{\mathsf{w}}}{\beta^t} - \mathcal{E}^{t+1})] \\
&\overset{②}{\leq} \mathbb{E}[-\textstyle\sum_{i=0}^t \mathcal{E}^{i+1} + C_{\mathsf{w}}C_{\mathsf{b}}],
\end{aligned}
$$

where step ① uses the definition of $\mathcal{E}_\star^{t+1} \triangleq \mathcal{E}^{t+1} - \frac{C_{\mathsf{w}}}{\beta^t}$; step ② uses Inequality (3). Using the fact that $\Theta^{T+1} \geq \underline{\Theta}$ as shown in Lemma 3.8, we conclude this lemma. □

G.6. Proof of Theorem F.11

Proof. The proof of **Part (a)** of this theorem is similar to that of Theorem 3.7. We omit the proof for brevity.

(b) We have shown that it takes at most $T = \mathcal{O}(\epsilon^{-1})$ iterations for Algorithm 3 to find an ϵ-INP. Therefore, the total stochastic first-order oracle complexity is given by $\lceil \frac{T}{q} \rceil N + T \cdot b$ in the sense of expectation, where b is the size of the mini-batch. We further derive:

$$\lceil \tfrac{T}{q} \rceil N + T \cdot b \leq \tfrac{T+q}{q} N + Tb = T\tfrac{N}{q} + N + Tb \overset{①}{\leq} T\tfrac{N}{\sqrt{N}} + N + T\sqrt{N} \overset{②}{=} \mathcal{O}(\sqrt{N}\epsilon^{-1} + N),$$

where step ① uses the choice that $q = b = \sqrt{N}$; step ② uses the fact that $T = \mathcal{O}(\epsilon^{-1})$.

□

H. Additional Experiments and Details

H.1. Additional Experiments for Deterministic Settings

We provide additional experiment results for the Sparse PCA and Noisy Sparse Recovery problems under deterministic settings. As depicted in Figure **??** and Figure **??**, our proposed method, IRPL-ADMM, exhibits convergence for both tasks, generally outperforming other methods in terms of speed for the Sparse PCA problem. These results reinforce our earlier findings.

H.2. Experiments for Stochastic Settings

We compare the proposed algorithm, **IRPL-ADMM-SPIDER**, with IRPL-ADMM and standard ADMM. Figure **??** demonstrates that **IRPL-ADMM-SPIDER** significantly outperforms IRPL-ADMM and ADMM, in line with our theoretical analysis assigning a complexity of $N + \sqrt{N}/\epsilon$ to our method, compared to the N/ϵ complexity of IRPL-ADMM and ADMM.

H.3. Datasets

We incorporate six datasets in our experiments, which include both randomly generated data and publicly available real-world data. These datasets serve as our data matrices $\mathbf{D} \in \mathbb{R}^{m' \times d'}$. The dataset names are as follows: 'CnnCaltech-m'-d'', 'TDT2-m'-d'', 'sector-m'-d'', 'mnist-m'-d'', 'randn-m'-d'', and 'dct-m'-d''. Here, $\mathrm{randn}(m, n)$ represents a function that generates a standard Gaussian random matrix with dimensions $m \times n$, while $\mathrm{dct}(m, n)$ refers to a function that produces a random matrix sampled from the discrete cosine transform. The matrix $\mathbf{D} \in \mathbb{R}^{m' \times d'}$ is constructed by randomly selecting m' examples and d' dimensions from the original real-world dataset (http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html,https://www.csie.ntu.edu.tw/~cjlin/libsvm/). We normalize each column of $\mathbf{D}$ to have a unit norm and center the data by subtracting the mean (represented as $\mathbf{D} \Leftarrow \mathbf{D} - \mathbf{1}\mathbf{1}^\mathsf{T}\mathbf{D}$).

For the noisy sparse recovery problem, we create the original signal $\check{\mathbf{v}}$ with an s-sparse structure as follows: We randomly select a support set $\mathtt{S}$ with a cardinality of $0.1 \times n$ and set $\check{\mathbf{v}}_{\mathtt{S}} = \mathrm{randn}(|\mathtt{S}|, 1)$, while $\check{\mathbf{v}}_{\{1,\ldots,n\}\setminus \mathtt{S}}$ is set to $\mathbf{0}$. Additionally, the observation vector is generated as $\mathbf{y} = \mathbf{D}\check{\mathbf{v}} + 0.1\|\mathbf{D}\check{\mathbf{v}}\| \times \mathrm{randn}(d', 1)$. Finally, we set $\tau = \|\mathbf{y} - \mathbf{D}\check{\mathbf{v}}\|$.

H.4. Nonconvex Proximal Operators

In this subsection, we demonstrate how to compute the nonconvex proximal operator for various functions $h(\mathbf{x})$ involved in this paper, given $\mathbf{x}' \in \mathbb{R}^{d \times 1}$ and $\mu > 0$. The proximal operator is defined as follows:

$$\min_{\mathbf{x} \in \mathbb{R}^{d \times 1}} \; h(\mathbf{x}) + \tfrac{\mu}{2}\|\mathbf{x} - \mathbf{x}'\|_2^2. \tag{137}$$

H.4.1. ℓ_q Norm Function with $q = 1/2$

When $h(\mathbf{x}) = \|\mathbf{x}\|_q^q$ with $q = 1/2$, Problem (137) reduces to the following optimization problem:

$$\bar{\mathbf{x}} \in \arg\min_{\mathbf{x}} \tfrac{\mu}{2}\|\mathbf{x} - \mathbf{x}'\|_2^2 + \lambda\|\mathbf{x}\|_{1/2}^{1/2}.$$

We utilize a computationally efficient closed-form solver to calculate the ℓ_q norm proximal operator (Xu et al., 2012).

H.4.2. ORTHOGONALITY CONSTRAINT

When $h(\mathbf{x}) = \mathcal{I}_{\mathcal{M}}(\mathrm{mat}(\mathbf{x}))$, Problem (137) reduces to the following optimization problem:

$$\bar{\mathbf{x}} \in \arg\min_{\mathbf{x}} \tfrac{\mu}{2}\|\mathbf{x}-\mathbf{x}'\|_2^2,\ s.t.\ \mathrm{mat}(\mathbf{x}) \in \mathcal{M} \triangleq \{\mathbf{V} \mid \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}\}.$$

This is the nearest orthogonality matrix problem, and the optimal solution can be computed as $\bar{\mathbf{x}} = \mathrm{vec}(\hat{\mathbf{U}}\hat{\mathbf{V}}^\mathsf{T})$, where $\mathrm{mat}(\mathbf{x}') = \hat{\mathbf{U}}\mathrm{Diag}(\mathbf{s})\hat{\mathbf{U}}^\mathsf{T}$ is the singular value decomposition of the matrix $\mathrm{mat}(\mathbf{x}')$. Refer to (Lai & Osher, 2014).

H.4.3. DC ℓ_1-LARGEST-k FUNCTION

When the function $h(\mathbf{x})$ is defined as: $h(\mathbf{x}) = \lambda(\|\mathbf{x}\|_1 - \sum_{i=1}^{k}|\mathbf{x}_{[i]}|)$ with $\lambda \geq 0$ being a constant, Problem (137) reduces to the following optimization problem:

$$\bar{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^{d\times 1}} \tfrac{\mu}{2}\|\mathbf{x}-\mathbf{x}'\|_2^2 + \lambda(\|\mathbf{x}\|_1 - \textstyle\sum_{i=1}^{k}|\mathbf{x}_{[i]}|).$$

Here, $\mathbf{x}_{[i]}$ is the i-th largest component of $\mathbf{x} \in \mathbb{R}^d$ in magnitude. Furthermore, we define $\mathbf{x}_{\{j\}}$ is the j-th *smallest* component of $\mathbf{x} \in \mathbb{R}^d$ in magnitude. We can rewrite this problem using the fact that $\|\mathbf{x}\|_1 - \sum_{i=1}^{k}|\mathbf{x}_{[i]}| = \sum_{j=1}^{d-k}|\mathbf{x}_{\{j\}}|$, resulting in the equivalent problem:

$$\bar{\mathbf{x}} \in \arg\min_{\mathbf{x}} \tfrac{\mu}{2}\|\mathbf{x}-\mathbf{b}\|_2^2 + \lambda\textstyle\sum_{j=1}^{d-k}|\mathbf{x}_{\{j\}}|. \tag{138}$$

Notably, for any optimal solution $\bar{\mathbf{x}}$, the following relation holds for all i and j:

$$(|\mathbf{b}_i| > |\mathbf{b}_j|) \Rightarrow (|\bar{\mathbf{x}}_i| > |\bar{\mathbf{x}}_j|).$$

We denote $\mathtt{I}$ as the index of the largest k elements of $\mathbf{x}'$ in magnitude, and $\mathtt{J} = \{1,\ldots,d\} \setminus \mathtt{I}$ as the index of the smallest $(d-k)$ elements of $\mathbf{x}'$ in magnitude. We have $|\mathtt{I}| = k$ and $|\mathtt{J}| = d-k$. The optimal solution to Problem (138) can be decomposed into two dependent sub-problems:

$$\begin{aligned}
\bar{\mathbf{x}}_{\mathtt{I}} &= \arg\min_{\mathbf{x}_I} \tfrac{\mu}{2}\|\mathbf{x}_{\mathtt{I}} - \mathbf{x}'_{\mathtt{I}}\|_2^2 + 0 \\
\bar{\mathbf{x}}_{\mathtt{J}} &= \arg\min_{\mathbf{x}_J} \tfrac{\mu}{2}\|\mathbf{x}_{\mathtt{J}} - \mathbf{x}'_{\mathtt{J}}\|_2^2 + \lambda\|\mathbf{x}_{\mathtt{J}}\|_1.
\end{aligned}$$