Smoothing Proximal Gradient Methods for Nonsmooth Sparsity Constrained Optimization: Optimality Conditions and Global Convergence

Ganzhao Yuan [1]

Abstract

Nonsmooth sparsity constrained optimization encompasses a broad spectrum of applications in machine learning. This problem is generally nonconvex and NP-hard. Existing solutions to this problem exhibit several notable limitations, including their inability to address general nonsmooth problems, tendency to yield weaker optimality conditions, and lack of comprehensive convergence analysis. This paper considers Smoothing Proximal Gradient Methods (SPGM) as solutions to nonsmooth sparsity constrained optimization problems. Two specific variants of SPGM are explored: one based on Iterative Hard Thresholding (SPGM-IHT) and the other on Block Coordinate Decomposition (SPGM-BCD). It is shown that the SPGM-BCD algorithm finds stronger stationary points compared to previous methods. Additionally, novel theories for analyzing the convergence rates to approximate global optimal solutions of both the SPGM-IHT and SPGM-BCD algorithms are developed. Our theoretical bounds, capitalizing on the intrinsic sparsity of the optimization problem, are on par with the best-known error bounds available to date. Finally, numerical experiments reveal that SPGM-IHT performs comparably to current IHT-style methods, while SPGM-BCD consistently surpasses them.

1. Introduction

This paper mainly focuses on the following nonsmooth sparsity constrained optimization problem ('$\triangleq$' means define):

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) \triangleq f(\mathbf{x}) + h(\mathbf{A}\mathbf{x}-\mathbf{b}),\ \mathrm{s.t.}\ \|\mathbf{x}\|_0 \leq s. \quad (1)$$

Here, $\mathbf{A}\in\mathbb{R}^{m\times n}$, $\mathbf{b}\in\mathbb{R}^m$, $s\in[n]$ is a positive integer, $f(\mathbf{x}):\mathbb{R}^n\mapsto\mathbb{R}$ is a smooth convex function, and $h(\mathbf{y}):\mathbb{R}^m\mapsto\mathbb{R}$ is a convex but not necessarily smooth function. For any vector $\mathbf{c}\in\mathbb{R}^m$ and any positive constant $\mu\in\mathbb{R}$, we assume that the following proximal operator of $h(\cdot)$ can be computed efficiently:

$$\mathbb{P}_\mu(\mathbf{c}) \triangleq \arg\min_{\mathbf{y}} h(\mathbf{y}) + \tfrac{1}{2\mu}\|\mathbf{c}-\mathbf{y}\|_2^2. \quad (2)$$

Problem (1) captures a diverse range of applications in machine learning. To illustrate, nonsmooth functions including $h(\mathbf{x})\triangleq\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_1$, $h(\mathbf{x})\triangleq\|\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{x}-\mathbf{b})\|_\infty$, and $h(\mathbf{x})\triangleq\|\max(0,\mathbf{A}\mathbf{x}-\mathbf{b})\|_1$ have been used in robust regression, Digzig selector computation, and support vector machines, respectively (Yuan et al., 2020b). Furthermore, Problem (1) covers a multitude of significant applications, such as sparse logistic regression (Bahmani et al., 2013), sparse censored regression (Bian & Chen, 2020), impulse noise removal (Yuan & Ghanem, 2019), sparse isotonic regression (Chen & Banerjee, 2018), and sparse quantile regression (Bian & Chen, 2020), as specific instances.

Solving Problem (1) presents a challenge primarily due to the combinatorial nature of the cardinality constraint. A conventional approach involves replacing the non-convex ℓ_0 norm with its convex relaxations, such as the ℓ_1 norm (Candes & Tao, 2005) and top-k norm relaxation. However, studies have revealed that non-convex approximation techniques, such as the Schatten ℓ_p norm (Xu et al., 2012; Zeng et al., 2016) and reweighted ℓ_1 norm (Candes et al., 2008), often yield superior accuracy compared to their convex counterparts (Zhang, 2010; Yuan & Ghanem, 2019). Furthermore, alternative strategies like multi-stage convex relaxation techniques have been introduced (Zhang, 2010; Bi et al., 2014), aiming to refine solutions obtained through convex methods. Recent efforts have primarily focused on directly minimizing the non-convex formulation in Problem (1). Greedy pursuit methods (Bahmani et al., 2013; Tropp & Gilbert, 2007) selectively choose a variable coordinate to update, leading to optimality guarantees in certain scenarios. Iterative Hard Thresholding (IHT) methods (Bahmani et al., 2013; Nguyen et al., 2017) maintain sparsity by iteratively zeroing out small magnitude elements in a gradient descent fashion. Convergence rates and parameter estimation errors for IHT-style methods have been rigorously established under restricted smoothness and strong convexity conditions (Yuan et al., 2017; Jain et al., 2014). The works of (Beck

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

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Table 1: Comparison among existing nonsmooth sparsity constrained optimization methods. t denotes the iteration counter, L_F is the Lipschitz constant for $F(\mathbf{x})$, and $\bar{\mathbf{x}}$ is the *global optimal solution* satisfying $\|\bar{\mathbf{x}}\|_0 \leq s$. The notation $\tilde{\mathcal{O}}(\cdot)$ hides polylogarithmic factors, while $\mathcal{O}(\cdot)$ hides constants.

	General Nonsmooth	Optimality Conditions	Convergence Rate to Approximate Optimal Solutions
PALM (Bolte et al., 2014)	✗	Critical Point	Unknown
GHTP (Yuan et al., 2017)	✗	Lipschitz Stat. Point	$\|\mathbf{x}^t - \bar{\mathbf{x}}\| \leq \mathcal{O}(\vartheta^t) + c_0\sqrt{s}\|\nabla F(\bar{\mathbf{x}})\|_\infty^a$
PDM (Lu & Zhang, 2013)	✓	Lipschitz Stat. Point	Unknown
DIHT (Yuan et al., 2020b)	✗	Lipschitz Stat. Point	Unknown[b]
PSGD (Liu et al., 2019)	✓✗[c]	Lipschitz Stat. Point	$[\min_{i=1}^t F(\mathbf{x}^i)] - F(\bar{\mathbf{x}}) \leq \mathcal{O}(\frac{1}{t}) + L_F\|\bar{\mathbf{x}}\|$
SPGM-IHT [ours]	✓	Lipschitz Stat. Point	$[\min_{i=1}^t F(\mathbf{x}^i)] - F(\bar{\mathbf{x}}) \leq \tilde{\mathcal{O}}(\frac{1}{t}) + L_F\|\bar{\mathbf{x}}\|^d$
SPGM-BCD [ours]	✓	Block-k Stat. Point	$\mathbb{E}_{\xi^t}[\min_{i=1}^t F(\mathbf{x}^i)] - F(\bar{\mathbf{x}}) \leq \tilde{\mathcal{O}}(\frac{1}{t}) + 3L_F\|\bar{\mathbf{x}}\|^e$

Note a: **GHTP** addresses only the special case of Problem (1) with $h(\cdot) = 0$. Here, $\vartheta \in (0, 1)$, and c_0 is a constant related to the stepsize.

Note b: **DIHT** establishes the convergence rate only for the primal-dual gap rate $\mathcal{O}(\frac{1}{\sqrt{t}})$, without addressing the primal convergence rate.

Note c: **PSGD** is less versatile, unable to solve nonsmooth problems when the function $h(\mathbf{x})$ lacks Lipschitz continuity.

Note d: The irreducible estimation error term $L_F\|\bar{\mathbf{x}}\|$ precisely aligns with the **PSGD** bound. Refer to Theorems 4.6, and 4.7.

Note e: The irreducible estimation error term $3L_F\|\bar{\mathbf{x}}\|$ is three times that of the **PSGD** bound. Refer to Theorems 4.13,and 4.14.

& Eldar, 2013; Beck & Vaisbourd, 2016; Beck & Hallak, 2016; 2019) have introduced a novel optimality condition based on coordinate-wise optimality for sparsity constrained optimization. This condition is proven to be stronger than the IHT-based optimality condition. Additionally, a new block coordinate optimality condition (Yuan et al., 2020a; 2019) is introduced for general sparse optimization, which is more powerful than the coordinate-wise optimality condition, encompassing it as a special case.

Another challenge in solving Problem (1) arises from the nonsmooth nature of the objective function. The Alternating Direction Method of Multipliers (ADMM) (He & Yuan, 2012) is a versatile tool and effectively handles nonsmooth and nonseparable problems, such as Problem (1), which pose challenges for other standard optimization methods like Proximal Alternating Linearized Minimization (PALM) (Bolte et al., 2014). ADMM introduces dual variables to address linear constraints, iteratively optimizing primal variables while keeping other primal and dual variables static, and employs a gradient ascent strategy to update the dual variables. However, it has been noted in (Lu & Zhang, 2013) that ADMM often yields unsatisfactory solution quality. This observation has motivated the exploration of Penalty Decomposition Methods (PDM) for solving generally nonlinear sparsity constrained optimization problems (Lu & Zhang, 2013). Additionally, Projective Subgradient Descent (PSGD) methods have been proposed for solving nonsmooth one-bit compressed sensing problems (Liu et al., 2019), operating by iteratively projecting the intermediate solution onto the nonconvex sparsity constraint after each sub-gradient descent update. Furthermore, Dual Iterative Hard Thresholding (DIHT) (Yuan et al., 2020b) applies projective subgradient methods to the dual of sparsity constraint optimization problems, offering proven guarantees on primal-dual gap convergence and sparsity recovery. Their duality theory establishes sufficient and necessary conditions for solving the original non-convex problem equivalently or approximately through a concave dual approach.

We summarize three main limitations of existing methods for solving Problem (1). ***(i)*** Inability to handle general nonsmooth problems. Proximal Alternating Linearized Minimization (PALM) methods are limited to solving nonconvex problems that allow for efficient closed-form proximal subproblems (Bolte et al., 2014). Block decomposition (Yuan et al., 2020a) and dual Iterative Hard Thresholding (IHT) (Yuan et al., 2020b) methods are restricted to smooth sparsity constrained problems. In contrast, PSGD methods are only applicable to objectives that are Lipschitz continuous. These methods struggle with general nonsmooth, nonseparable problems, which are better addressed by penalty decomposition (Lu & Zhang, 2013) or smoothing proximal gradient methods (Bian & Chen, 2020; Chen, 2012). ***(ii)*** Tendency to yield weaker optimality conditions. PALM methods primarily focus on identifying critical points of Problem (1). Additionally, by predominantly relying on IHT, current methods often result in suboptimal optimality guarantees of Lipschitz stationary points, leading to subpar practical accuracy (Beck & Eldar, 2013; Yuan et al., 2020a; Yuan, 2023a;b). ***(iii)*** Lack of comprehensive convergence analysis. The work in (Bolte et al., 2014) outlines the convergence rates to critical points by employing the Kurdyka-Łojasiewicz inequality. Although IHT-style methods have been incorporated into PDM (Lu & Zhang, 2013), it is important to note that comprehensive convergence analysis remains absent. Additionally, the duality theory presented in (Yuan et al., 2020b) is constrained by its assumption of smooth objective functions, as demonstrated in Theorems 15 and 17 of the same reference.

To address the aforementioned three limitations, we establish and achieve the following three goals (details in Table 1). (***i***) *Toward general nonsmooth optimization algorithms*. We consider Smoothing Proximal Gradient Methods (SPGM) for nonsmooth sparsity constrained optimization, featuring two SPGM variants: SPGM based on Iterative Hard Thresholding (**SPGM-IHT**) and SPGM based on Block Coordinate Decomposition (**SPGM-BCD**). These methods, rooted in smoothing techniques, tackle a wide range of nonsmooth problems. (***ii***) *Toward a stronger optimality condition*. The proposed **SPGM-BCD** targets a stronger block-k stationary point of Problem (1). It achieves this by employing a simple and efficient local combinatorial search strategy (Yuan et al., 2020a; 2019) for the small-sized subproblem. Specifically, **SPGM-BCD** first systematically enumerates the full binary tree and then solves 2^k small linear equations to identify all potential candidates; finally, it selects the one with the lowest objective value as the optimal solution. (***iii***) *Toward stronger global convergence*. Our research aims to establish the convergence rate to approximate global optimal solutions. We prove that the degree of approximation is influenced by the Lipschitz constant of the objective function L_F, and the ℓ_2 norm of the global optimal solution $\|\bar{\mathbf{x}}\|_2$. Our theoretical bounds leverage the inherent sparsity of the optimization problem, matching the best-known error bounds (Liu et al., 2019) currently available.

Contributions. The contributions of this paper are threefold. (***i***) We explore Smoothing Proximal Gradient Methods (SPGM) for solving Problem (1), including **SPGM-IHT** and **SPGM-BCD** (see Section 2). We offer smooth and optimality analyses for the smoothing reformulation problem, demonstrating that **SPGM-BCD** achieves stronger stationary points compared to existing solutions (see Section 3). (***ii***) We develop novel theories to analyze the convergence rate of both **SPGM-IHT** and **SPGM-BCD** (see Section 4). (***iii***) We have conducted experiments on two nonsmooth sparsity constrained optimization tasks to show the superiority of our methods (see Section 5).

Notations. All vectors are column vectors, with superscript $^{\mathsf{T}}$ indicating transpose. For a vector $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}_i$ represents its i-th component for any $i \in [n] \triangleq \{1, 2, ..., n\}$. The Euclidean inner product between vectors $\mathbf{x}$ and $\mathbf{x}'$ is expressed as $\langle \mathbf{x}, \mathbf{x}' \rangle$ or $\mathbf{x}^{\mathsf{T}}(\mathbf{x}')$. The identity matrix in $\mathbb{R}^{n\times n}$ is denoted by $\mathbf{I}_n$. $\|\mathbf{A}\|$ represents the spectral norm of $\mathbf{A}$. The notations $\mathbf{C} \succeq \mathbf{0}$ and $\mathbf{C} \succ \mathbf{0}$ indicate positive semidefiniteness and definiteness of $\mathbf{C}$, respectively. For any $\mathbf{C}$ with $\mathbf{C} \succeq \mathbf{0}$, we define $\|\mathbf{x}\|_{\mathbf{C}} \triangleq \sqrt{\mathbf{x}^{\mathsf{T}}\mathbf{C}\mathbf{x}}$ as a generalized vector norm, and denote $\boldsymbol{\lambda}_{\max}(\mathbf{C})$ and $\boldsymbol{\lambda}_{\min}(\mathbf{C})$ as respectively the largest and smallest eigenvalue of $\mathbf{C}$. If β is a constant, β^t refers to its t-th power, while if β is an optimization variable, β^t signifies the value in the t-th iteration. The subdifferential of the function $h : \mathbb{R}^n \mapsto (-\infty, +\infty]$ at $\mathbf{x}$, defined as $\partial h(\mathbf{x}) \triangleq \{\mathbf{g} : h(\mathbf{y}) \geq h(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle\}$, includes all subgradients of $h(\mathbf{x})$. The squared distance between sets Ξ and Ξ' is defined as $\operatorname{dist}^2(\Xi, \Xi') \triangleq \inf_{\mathbf{v}\in\Xi, \mathbf{v}'\in\Xi'} \|\mathbf{v} - \mathbf{v}'\|_2^2$.

For a set $\mathtt{B} \in \mathbb{N}^k$ containing k unique integers selected from $\{1, 2, ..., n\}$, we define $\mathtt{B}^c \triangleq \{1, 2, ..., n\} \setminus \mathtt{B}$, and denote $\mathbf{C}_{\mathtt{BB}}$ as the sub-matrix of $\mathbf{C}$ indexed by $\mathtt{B}$. C_n^k counts the combinations to select k items from n without repetition. $\Omega_n^k \triangleq \{\mathtt{B}_{(1)}, \mathtt{B}_{(2)}, \ldots, \mathtt{B}_{(C_n^k)}\}$ represents the set of all index vector combinations for this selection, with each $\mathtt{B}_{(i)} \in \mathbb{N}^k$.

2. Smoothing Proximal Gradient Methods

This section explores Smoothing Proximal Gradient Methods (SPGM) for Problem (1), detailing two versions: **SPGM-IHT**, using Iterative Hard Thresholding (Blumensath & Davies, 2008; 2009), and **SPGM-BCD**, employing Block Coordinate Decomposition (Yuan et al., 2020a; 2019).

Initially, we impose the following assumptions on Problem (1).

Assumption 2.1. The functions $f(\cdot)$ and $h(\cdot)$ are Lipschitz continuous with some constants L_f and L_h, satisfying $\|\nabla f(\mathbf{x})\| \leq L_f$ for all $\|\mathbf{x}\|_0 \leq s$ and $\|\partial h(\mathbf{y})\| \leq L_h$ for all $\mathbf{y} \in \mathbb{R}^m$. Consequently, $F(\mathbf{x})$ is Lipschitz continuous with constant $L_F \triangleq L_f + \|\mathbf{A}\| L_h$.

Assumption 2.2. The function $f(\cdot)$ is restricted V_s-strongly convex and restricted M_s-smooth, such that for all $\|\mathbf{x}\|_0 \leq s$ and $\|\mathbf{x}'\|_0 \leq s$, we have:

$$\tfrac{V_s}{2}\|\mathbf{x} - \mathbf{x}'\|_2^2 \leq \mathcal{Q}(\mathbf{x}, \mathbf{x}') \leq \tfrac{M_s}{2}\|\mathbf{x} - \mathbf{x}'\|_2^2,$$

where $\mathcal{Q}(\mathbf{x}, \mathbf{x}') \triangleq f(\mathbf{x}') - f(\mathbf{x}) - \langle \mathbf{x}' - \mathbf{x}, \nabla f(\mathbf{x}) \rangle$. Additionally, for all $\|\mathbf{x}\|_0 \leq s$ and $\|\mathbf{x}'\|_0 \leq s$, a symmetric matrix $\tilde{\mathbf{M}} \in \mathbb{R}^{n\times n}$ exists, fulfilling $\mathbf{0} \prec V_s\mathbf{I}_n \preceq \tilde{\mathbf{M}} \preceq M_s\mathbf{I}_n$ and

$$\mathcal{Q}(\mathbf{x}, \mathbf{x}') \leq \tfrac{1}{2}\|\mathbf{x} - \mathbf{x}'\|_{\tilde{\mathbf{M}}}^2. \tag{3}$$

Assumption 2.3. A constant $A_s > 0$ exists, ensuring $\|\mathbf{A}(\mathbf{x} - \mathbf{x}')\| \leq A_s\|\mathbf{x} - \mathbf{x}'\|$ for all $\mathbf{x} \in \mathbb{R}^n, \mathbf{x}' \in \mathbb{R}^n$ with $\|\mathbf{x}\|_0 \leq s$, $\|\mathbf{x}'\|_0 \leq s$.

Remarks. (***i***) Assumptions 2.1, 2.2, and 2.3 are broadly applicable, meeting conditions of various applications like robust regression and support vector machines (see (Yuan et al., 2017)). (***ii***) Assumption 2.3 is less stringent than $\|\mathbf{A}(\mathbf{x} - \mathbf{x}')\| \leq \|\mathbf{A}\|\|\mathbf{x} - \mathbf{x}'\|$. (***iii***) Common choices for nonsmooth $h(\mathbf{y})$ include $\{\|\mathbf{y}\|_1, \|\max(0, \mathbf{y})\|_1, \|\mathbf{y}\|_\infty\}$, with their corresponding L_h values being $\{\sqrt{m}, \sqrt{m}, 1\}$, respectively. (***iv***) When $f(\mathbf{x})$ takes the form of a quadratic function with $f(\mathbf{x}) \triangleq \frac{1}{2}\mathbf{x}^{\mathsf{T}}\hat{\mathbf{Q}}\mathbf{x} + \mathbf{x}^{\mathsf{T}}\hat{\mathbf{p}}$ for some $\hat{\mathbf{Q}} \in \mathbb{R}^{n\times n}$ and $\hat{\mathbf{p}} \in \mathbb{R}^n$, Inequality (3) holds with $\mathcal{Q}(\mathbf{x}, \mathbf{x}') = \frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|_{\tilde{\mathbf{M}}}^2$, where $\tilde{\mathbf{M}} = \hat{\mathbf{Q}}$.

Subsequently, we introduce a new variable $\mathbf{y} \in \mathbb{R}^m$ and reframe Problem (1) as:

$$\min_{\mathbf{x},\mathbf{y}} f(\mathbf{x}) + h(\mathbf{y}),\ \mathrm{s.t.}\ \mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{y},\ \|\mathbf{x}\|_0 \leq s.$$

In SPGM, a smoothing parameter $\mu \to 0$ is incorporated to penalize the squared error in the linear constraints, leading to the following minimization problem:

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{y}} \mathcal{J}(\mathbf{x},\mathbf{y};\mu) \triangleq \mathcal{R}(\mathbf{x},\mathbf{y};\mu) + h(\mathbf{y}) + \delta(\mathbf{x}), \\ &\text{where}\, \mathcal{R}(\mathbf{x},\mathbf{y};\mu) \triangleq f(\mathbf{x}) + \tfrac{1}{2\mu}\|\mathbf{A}\mathbf{x} - \mathbf{b} - \mathbf{y}\|_2^2, \end{aligned} \tag{4}$$

and $\delta(\mathbf{x}) \triangleq \left\{ \begin{smallmatrix} 0, & \|\mathbf{x}\|_0 \leq s \\ \infty, & \text{else} \end{smallmatrix} \right.$. In each iteration, we employ proximal alternating linearized minimization strategies (Bolte et al., 2014) to alternatively minimize *w.r.t.* $\mathbf{x}$ and $\mathbf{y}$. Notably, SPGM is closely related to block coordinate descent methods (Tseng & Yun, 2009; Xu & Yin, 2013), and penalty decomposition methods (Lu & Zhang, 2013) in the literature.

We summarize the SPGM algorithm in Algorithm 1.

Algorithm 1 Smoothing Proximal Gradient Methods with IHT or BCD Strategies.

```
Input: initial feasible solution $\mathbf{x}^1$; working set size $k \in \{2, 3, \ldots, n\}$; positive proximal point parameters $\{\theta, \theta_1, \theta_2\}$; positive smoothing parameter $\{\mu^1\}$;
for $t = 1$ to $T$ do
  (S1) Solve the $\mathbf{x}$-subproblem using IHT or BCD strategy.
  ► Option I (IHT): Solve the following problem globally (Blumensath & Davies, 2008):
      $\mathbf{x}^{t+1} \in \arg\min_{\|\mathbf{x}\|_0 \leq s} \dot{\mathcal{M}}(\mathbf{x}, \mathbf{x}^t, \mathbf{y}^t; \mu^t) \triangleq R^t$
      $+ \frac{H^t}{2}\|\mathbf{x} - \mathbf{x}^t\|_2^2 + \langle \mathbf{x} - \mathbf{x}^t, \mathbf{r}^t \rangle,$    (5)
    where $H^t \triangleq A_s^2/\mu^t + M_s + \theta \in \mathbb{R}$, and $R^t \triangleq \mathcal{R}(\mathbf{x}^t, \mathbf{y}^t; \mu^t)$.
  ► Option II (BCD): Use a random or/and a greedy method to find a working set $\mathtt{B}^t$ of size $k$ for the $t$-th iteration. Let $\mathtt{B} = \mathtt{B}^t$ and $\mathtt{B}^c \triangleq \{1, ..., n\} \setminus \mathtt{B}$. Solve the following problem globally (Yuan et al., 2020a):
      $\mathbf{x}^{t+1} \in \arg\min_{\|\mathbf{x}\|_0 \leq s, \mathbf{x}_{\mathtt{B}^c} = \mathbf{x}^t_{\mathtt{B}^c}} \ddot{\mathcal{M}}(\mathbf{x}, \mathbf{x}^t, \mathbf{y}^t; \mu^t) \triangleq R^t$
      $+ \frac{1}{2}\|\mathbf{x} - \mathbf{x}^t\|_{\mathbf{H}^t}^2 + \langle \mathbf{x} - \mathbf{x}^t, \mathbf{r}^t \rangle,$    (6)
    where $\mathbf{H}^t \triangleq (\mathbf{A}^\mathsf{T}\mathbf{A} + \theta_1 \mathbf{I}_n)/\mu^t + \tilde{\mathbf{M}} + \theta_2 \mathbf{I}_n \in \mathbb{R}^{n \times n}$, and $R^t \triangleq \mathcal{R}(\mathbf{x}^t, \mathbf{y}^t; \mu^t)$.
  (S2) Solve the $\mathbf{y}$-subproblem $\mathbf{y}^{t+1} = \mathbb{P}_{\mu^t}(\mathbf{A}\mathbf{x}^{t+1} - \mathbf{b})$.
  (S3) Choose a new parameter $\mu^{t+1}$ with $\mu^{t+1} \leq \mu^t$.
end for
```

► $\mathbf{x}$-subproblem. Keeping parameters $\mathbf{y}^t$ and μ^t constant at their current values, we minimize $\mathcal{J}(\mathbf{x}, \mathbf{y}^t; \mu^t)$ *w.r.t.* $\mathbf{x}$, resulting in the subsequent optimization problem:

$$\min_{\mathbf{x}} \mathcal{R}(\mathbf{x}, \mathbf{y}^t; \mu^t),\ \mathrm{s.t.}\ \|\mathbf{x}\|_0 \leq s.$$

The function $\mathcal{R}(\mathbf{x}, \mathbf{y}^t; \mu^t)$ is differentiable in $\mathbf{x}$, with its gradient at $\mathbf{x}^t$ given by:

$$\nabla_{\mathbf{x}} \mathcal{R}(\mathbf{x}^t, \mathbf{y}^t; \mu^t) = \nabla f(\mathbf{x}^t) + \tfrac{1}{\mu^t}\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{x}^t - \mathbf{b} - \mathbf{y}^t) \triangleq \mathbf{r}^t.$$

To solve the $\mathbf{x}$-subproblem, we consider state-of-the-art sparse optimization methods, including the IHT strategy (Yuan et al., 2017; 2020b; Jain et al., 2014; Lu, 2014) and the BCD strategy (Yuan et al., 2020a; 2019).

$\boxed{\text{IHT strategy.}}$ We notice the following inequality consistently holds for all $\|\mathbf{x}\|_0 \leq s$:

$$\mathcal{R}(\mathbf{x}, \mathbf{y}^t; \mu^t) \leq \dot{\mathcal{M}}(\mathbf{x}, \mathbf{x}^t, \mathbf{y}^t; \mu^t), \tag{7}$$

where $\dot{\mathcal{M}}(\mathbf{x}, \mathbf{x}^t, \mathbf{y}^t; \mu^t)$ is defined in Problem (5), and $\theta > 0$ is a constant. The IHT strategy aims to minimize the majorization function $\dot{\mathcal{M}}(\mathbf{x}, \mathbf{x}^t, \mathbf{y}^t; \mu^t)$, while adhering to the sparsity constraint. This approach simultaneously reduces the objective function and identifies the active variables, as indicated by the update in Problem (5). We note that Problem (5) is equivalent to the following problem:

$$\mathbf{x}^{t+1} \in \arg\min_{\|\mathbf{x}\|_0 \leq s} \tfrac{1}{2}\|\mathbf{x} - \mathbf{x}^t_+\|_2^2 = \Pi_s(\mathbf{x}^t_+), \tag{8}$$

where $\mathbf{x}^t_+ \triangleq \mathbf{x}^t - \mathbf{r}^t/H^t$, $H^t \triangleq A_s^2/\mu^t + M_s + \theta$, and $\Pi_s(\mathbf{x})$ is an operator that sets all but the largest (in magnitude) s elements of $\mathbf{x}$ to zero.

$\boxed{\text{BCD strategy.}}$ We observe that the following condition always holds for all $\|\mathbf{x}\|_0 \leq s$:

$$\mathcal{R}(\mathbf{x}, \mathbf{y}^t; \mu^t) \leq \ddot{\mathcal{M}}(\mathbf{x}, \mathbf{x}^t, \mathbf{y}^t; \mu^t), \tag{9}$$

where $\ddot{\mathcal{M}}(\mathbf{x}, \mathbf{x}^t, \mathbf{y}^t; \mu^t)$ is defined in Problem (6), and $\{\theta_1, \theta_2\}$ are given positive constants. The **BCD** strategy minimizes the majorization function $\ddot{\mathcal{M}}(\mathbf{x}, \mathbf{x}^t, \mathbf{y}^t; \mu^t)$ using a block coordinate descent approach. It employs either a random method or a greedy method to select a subset of coordinates of size k as the working set $\mathtt{B}$, where $k \geq 2$. It then conducts a global combinatorial search over this working set, based on the quadratic majorization function, as indicated by the update in Problem (6). Problem (6) can be equivalently rewritten as: $\mathbf{x}^{t+1}_{\mathtt{B}} \in \arg\min_{\mathbf{z}_{\mathtt{B}}} \ddot{\mathcal{M}}(\mathrm{U}_{\mathtt{B}}\mathbf{z}_{\mathtt{B}} + \mathrm{U}_{\mathtt{B}^c}\mathbf{x}^t_{\mathtt{B}^c}, \mathbf{x}^t, \mathbf{y}^t; \mu^t) + \delta(\mathrm{U}_{\mathtt{B}}\mathbf{z}_{\mathtt{B}} + \mathrm{U}_{\mathtt{B}^c}\mathbf{x}^t_{\mathtt{B}^c})$, where $\mathtt{B}^c \triangleq \{1, ..., n\} \setminus \mathtt{B}$, $\mathrm{U}_{\mathtt{B}} \in \mathbb{R}^{n \times k}$, $\mathrm{U}_{\mathtt{B}^c} \in \mathbb{R}^{n \times (n-k)}$, and

$$[\mathrm{U}_{\mathtt{B}}]_{ji} = \left\{ \begin{array}{ll} 1, & \mathtt{B}_i = j; \\ 0, & \text{else.} \end{array} \right.,\ [\mathrm{U}_{\mathtt{B}^c}]_{ji} = \left\{ \begin{array}{ll} 1, & \mathtt{B}^c_i = j; \\ 0, & \text{else.} \end{array} \right..$$

We have: $\mathbf{x} = (\mathbf{U}_{\mathrm{B}^c}\mathbf{U}_{\mathrm{B}^c}^{\mathsf{T}} + \mathbf{U}_{\mathrm{B}}\mathbf{U}_{\mathrm{B}}^{\mathsf{T}})\mathbf{x} = \mathbf{U}_{\mathrm{B}}\mathbf{x}_{\mathrm{B}} + \mathbf{U}_{\mathrm{B}^c}\mathbf{x}_{\mathrm{B}^c}$, and $\mathbf{x}_{\mathrm{B}} = \mathbf{U}_{\mathrm{B}}^{\mathsf{T}}\mathbf{x}$. Thus, Problem (6) reduces to the following problem:

$$\begin{aligned} \mathbf{x}_{\mathrm{B}}^{t+1} \in \arg\min\nolimits_{\mathbf{z}_{\mathrm{B}}\in\mathbb{R}^k}\ & \tfrac{1}{2}(\mathbf{z}_{\mathrm{B}} - \mathbf{x}_{\mathrm{B}}^t)^{\mathsf{T}}[\mathbf{H}_{\mathrm{BB}}^t](\mathbf{z}_{\mathrm{B}} - \mathbf{x}_{\mathrm{B}}^t), \\ & + \langle \mathbf{z}_{\mathrm{B}} - \mathbf{x}_{\mathrm{B}}^t,\, \mathbf{r}_{\mathrm{B}}^t \rangle,\ \mathrm{s.t.}\ \|\mathbf{z}_{\mathrm{B}}\|_0 + \|\mathbf{x}_{\mathrm{B}^c}^t\|_0 \leq s, \end{aligned} \tag{10}$$

where $\mathbf{H}^t \triangleq (\mathbf{A}^{\mathsf{T}}\mathbf{A} + \theta_1\mathbf{I}_n)/\mu^t + \tilde{\mathbf{M}} + \theta_2\mathbf{I}_n$. Problem (10) involves k unknown decision variables, and can be tackled by solving a set of 2^k linear equations. The BCD strategy combines the efficacy of combinatorial search methods with the efficiency of coordinate descent methods, allowing it to efficiently identify stronger stationary points than the IHT strategy when minimizing smooth functions under sparsity constraints (Yuan et al., 2020a; 2019).

► $\mathbf{y}$-subproblem. With the parameters $\mathbf{x}^{t+1}$ and μ^t fixed at their current estimates, we encounter an optimization problem *w.r.t.* $\mathbf{y}$ as: $\mathbf{y}^{t+1} = \arg\min_{\mathbf{y}} \mathcal{J}(\mathbf{x}^{t+1}, \mathbf{y}; \mu^t) = \mathbb{P}_{\mu^t}(\mathbf{A}\mathbf{x}^{t+1} - \mathbf{b})$, which is equivalent to the computation of the proximal operator as described in Equation (2).

3. Smooth and Optimality Analyses

This section provides smooth and optimality analyses for the smoothing function as in Problem (4).

3.1. Smooth Analysis

We conduct a smooth analysis for Problem (4). By eliminating $\mathbf{y}$, Problem (4) simplifies to:

$$\begin{aligned} \min_{\mathbf{x}}\ & \mathcal{G}(\mathbf{x}; \mu) \triangleq f(\mathbf{x}) + h(\mathbb{P}_\mu(\mathbf{A}\mathbf{x} - \mathbf{b})) \\ & + \tfrac{1}{2\mu}\|\mathbf{A}\mathbf{x} - \mathbf{b} - \mathbb{P}_\mu(\mathbf{A}\mathbf{x} - \mathbf{b})\|_2^2,\ \mathrm{s.t.}\ \|\mathbf{x}\|_0 \leq s. \end{aligned} \tag{11}$$

The function $\mathcal{G}(\mathbf{x}; \mu)$ is differentiable *w.r.t.* $\mathbf{x}$ and its gradient is given by:

$$\nabla_{\mathbf{x}}\mathcal{G}(\mathbf{x}; \mu) = \nabla f(\mathbf{x}) + \tfrac{1}{\mu}\mathbf{A}^{\mathsf{T}}(\mathbf{A}\mathbf{x} - \mathbf{b} - \mathbb{P}_\mu(\mathbf{A}\mathbf{x} - \mathbf{b})).$$

We have the following useful lemmas [1].

Lemma 3.1. *(Proof in Appendix B.1) Fix* $\mathbf{x}$ *with* $\|\mathbf{x}\|_0 \leq s$. *The function* $\psi(\mu) \triangleq \mathcal{G}(\mathbf{x}; \mu)$ *is decreasing and* $(\frac{1}{2}L_h^2)$*-Lipschitz continuous for all* $\mu > 0$. *In other words, for all* $0 < \mu_1 < \mu_2$, *we have:* $0 \leq \frac{\psi(\mu_1) - \psi(\mu_2)}{\mu_2 - \mu_1} \leq \frac{1}{2}L_h^2$.

Lemma 3.2. *(Proof in Appendix B.2) Fix* $\mu > 0$. *For all* $\mathbf{x}$ *with* $\|\mathbf{x}\|_0 \leq s$, *we have:*

(a) *It holds that* $F(\mathbf{x}) - \frac{\mu}{2}L_h^2 \leq \mathcal{J}(\mathbf{x}, \mathbb{P}_\mu(\mathbf{A}\mathbf{x} - \mathbf{b}); \mu) = \mathcal{G}(\mathbf{x}; \mu) \leq F(\mathbf{x})$.

(b) *It holds that* $\|\partial F(\mathbf{x})\| \leq L_F$, $\|\nabla\mathcal{G}(\mathbf{x}; \mu)\| \leq L_F$, *where* $L_F \triangleq L_f + L_h\|\mathbf{A}\|$.

(c) $\mathcal{G}(\mathbf{x}; \mu)$ *is restricted* V_s*-strongly convex and restricted* $(M_s + A_s\|\mathbf{A}\|/\mu)$*-smooth.*

Remarks. ***(i)*** Lemmas 3.1 and 3.2 can be derived using Assumptions 2.1, 2.2, and 2.3, along with the optimality of the proximal operator $\mathbb{P}_\mu(\mathbf{c})$ for any $\mathbf{c}$. ***(ii)*** The inequalities in Lemma 3.1 and **Part (a)** of Lemma 3.2 are closely linked to smooth approximation functions as discussed in (Chen, 2012) and the Moreau-Yosida approximation (Bauschke et al., 2011) in the literature. These properties play a crucial role in the development of smoothing methods for nonsmooth optimization. ***(iii)*** Given that $\mathcal{G}(\mathbf{x}^t; \mu^{t-1})$ serves as a smooth approximation function for $F(\mathbf{x}^t)$, we can assess the convergence rate of $F(\mathbf{x}^t)$ by estimating the convergence rate of $\mathcal{G}(\mathbf{x}^t; \mu^{t-1})$.

3.2. Optimality Analysis

We provide an optimality analysis for SPGM.

As μ tends to 0, Problem (4) becomes equivalent to the original optimization problem in Problem (1). This equivalence can be represented as: $[\min_{\mathbf{x}} F(\mathbf{x}) + \delta(\mathbf{x})] \equiv [\min_{\mathbf{x}} \lim_{\mu\to 0} \mathcal{G}(\mathbf{x}; \mu)] \equiv [\min_{\mathbf{x},\mathbf{y}} \lim_{\mu\to 0} \mathcal{J}(\mathbf{x}, \mathbf{y}; \mu)]$. Thus, we can perform an optimality analysis using the smooth function $\mathcal{J}(\mathbf{x}, \mathbf{y}; \mu)$ with a sufficiently small μ.

We introduce the following fundamental definitions.

Definition 3.3. (Critical Point, or Basic Stationary Point (Beck & Eldar, 2013)) A solution $\check{\mathbf{x}}$ is a critical point if the following condition is met: $F(\check{\mathbf{x}}) = \min_{\mathbf{x}} F(\mathbf{x}),\ \mathrm{s.t.}\ [\mathbf{x}]_{\mathrm{J}^c} = \mathbf{0}$. Here, $\mathrm{J}^c \triangleq \{1, ..., n\} \setminus \mathrm{J}$, where J represents the known support set of the solution $\check{\mathbf{x}}$ with $|\mathrm{J}| \leq s$.

Remarks. When the support set is restricted, the NP-hard problem in Problem (1) reduces to a convex problem. The basic stationary point implies that the solution attains global optimality for the reduced convex problem when the support set is fixed.

Definition 3.4. (Lipschitz Stationary Point) Fix $\mu > 0$ as a sufficiently small constant. A solution $(\dot{\mathbf{x}}, \dot{\mathbf{y}})$ is a Lipschitz stationary point if the following condition holds:

$$\begin{aligned} \dot{\mathbf{y}} &= \arg\min_{\mathbf{y}} \mathcal{J}(\dot{\mathbf{x}}, \mathbf{y}; \mu), \\ \dot{\mathbf{x}} &\in \arg\min_{\mathbf{x}} \dot{\mathcal{M}}(\mathbf{x}, \dot{\mathbf{x}}, \dot{\mathbf{y}}; \mu) + \delta(\mathbf{x}), \end{aligned}$$

where $\dot{\mathcal{M}}(\mathbf{x}, \mathbf{x}^t, \mathbf{y}^t; \mu^t)$ is defined in Problem (5).

Remarks. The Lipschitz stationary point states that if we minimize the smoothing function $\mathcal{J}(\dot{\mathbf{x}}, \mathbf{y}; \mu)$ over $\mathbf{y}$ and the majorization function $\dot{\mathcal{M}}(\mathbf{x}, \dot{\mathbf{x}}, \dot{\mathbf{y}}; \mu)$ over $\mathbf{x}$, the quality of the solution $(\dot{\mathbf{x}}, \dot{\mathbf{y}})$ cannot be further improved.

Definition 3.5. (Block-k Stationary Point) Fix $\mu > 0$ as a sufficiently small constant. We denote $\mathrm{B}^c \triangleq \{1, ..., n\} \backslash \mathrm{B}$. A

[1] All proofs can be found in the Appendix.

solution $(\ddot{\mathbf{x}}, \ddot{\mathbf{y}})$ is a block-k stationary point if the following condition is met:

$$\ddot{\mathbf{y}} = \arg\min_{\mathbf{y}} \mathcal{J}(\ddot{\mathbf{x}}, \mathbf{y}; \mu),$$

$$\ddot{\mathbf{x}}_{\mathtt{B}} \in \arg\min_{\mathbf{z}_{\mathtt{B}}, \|\mathbf{z}_{\mathtt{B}}\|_0 + \|\ddot{\mathbf{x}}_{\mathtt{B}^c}\|_0 \leq s} \ddot{\mathcal{M}}(\mathbf{U}_{\mathtt{B}}\mathbf{z}_{\mathtt{B}} + \mathbf{U}_{\mathtt{B}^c}\ddot{\mathbf{x}}_{\mathtt{B}^c}, \ddot{\mathbf{x}}, \ddot{\mathbf{y}}; \mu)$$

for all $\mathtt{B} \in \Omega_n^k$. Here, $\Omega_n^k \triangleq \{\mathtt{B}_{(i)}\}_{i=1}^{C_n^k}$ denotes all the combinations of the index vector choosing k items from n without repetition, and $\ddot{\mathcal{M}}(\mathbf{x}, \mathbf{x}^t, \mathbf{y}^t; \mu^t)$ is defined in Problem (6).

Remarks. (***i***) Block-k stationary point captures more intrinsic structures of the nonconvex problem than Lipschitz stationary points, and it holds that $\ddot{\mathcal{M}}(\mathbf{x}, \mathbf{x}^t, \mathbf{y}^t; \mu^t) \leq \dot{\mathcal{M}}(\mathbf{x}, \mathbf{x}^t, \mathbf{y}^t; \mu^t)$ for all $\mathbf{x}$. (***ii***) Deterministically finding a block-k stationary point requires evaluating C_n^k subproblems, which can be time-consuming. However, using a random strategy to select the working set $\mathtt{B}$ from the C_n^k combinations allows for an expected block-k stationary point.

The following proposition states the relation between different types of the stationary point above.

Proposition 3.6. ***Optimality Hierarchy (Yuan et al., 2020a).*** *We denote the sets $\{\check{\mathbf{x}}\}$ (basic stationary points), $\{\dot{\mathbf{x}}\}$ (Lipschitz stationary points), $\{\ddot{\mathbf{x}}_{[k]}\}$ (block-k stationary points), and $\{\bar{\mathbf{x}}\}$ (global optimal points). The following relation holds for all $2 \leq k \leq n-1$:*

$$\{\bar{\mathbf{x}}\} \equiv \{\ddot{\mathbf{x}}_{[n]}\} \subseteq \{\ddot{\mathbf{x}}_{[k+1]}\} \subseteq \{\ddot{\mathbf{x}}_{[k]}\} \subseteq \{\dot{\mathbf{x}}\} \subseteq \{\check{\mathbf{x}}\}.$$

We establish the optimality hierarchy among the optimality conditions by directly applying the results of Proposition 1 in (Yuan et al., 2020a), which addresses the minimization of smooth functions under sparsity constraints.

4. Convergence Analysis

In this section, we develop novel theories to analyze the convergence rate of **SPGM-IHT** and **SPGM-BCD**.

In our analysis, we consider two strategies for updating μ^t for all $t = 1, 2, \ldots, \infty$.

- $\mu^t = \bar{\mu}$, where $\bar{\mu} > 0$ is a sufficiently small constant.
- $\mu^t = \frac{\eta}{t+t_0}$, where $\eta > 0$ and $t_0 \geq 1$ are constants.

We notice the following relation between $\nabla_{\mathbf{x}}\mathcal{G}(\mathbf{x}^t; \mu^{t-1})$ and $\nabla_{\mathbf{x}}\mathcal{R}(\mathbf{x}^t, \mathbf{y}^t; \mu^t)$:

$$\begin{aligned} \mathbf{g}^t &\triangleq \nabla\mathcal{G}(\mathbf{x}^t; \mu^{t-1}) = \nabla_{\mathbf{x}}\mathcal{R}(\mathbf{x}^t, \mathbf{y}^{t+1}; \mu^{t-1}) \\ &= \nabla f(\mathbf{x}^t) + \tfrac{1}{\mu^{t-1}}\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{x}^t - \mathbf{b} - \mathbf{y}^t) \\ &= \underbrace{\nabla_{\mathbf{x}}\mathcal{R}(\mathbf{x}^t, \mathbf{y}^t; \mu^t)}_{\triangleq \mathbf{r}^t} + \underbrace{(\tfrac{1}{\mu^{t-1}} - \tfrac{1}{\mu^t})\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{x}^t - \mathbf{b} - \mathbf{y}^t)}_{\triangleq \boldsymbol{\varepsilon}^t}. \end{aligned}$$

We derive the following lemmas for both **SPGM-IHT** and **SPGM-BCD**, which are independent of the choice of strategies for solving the $\mathbf{x}$-subproblem, and hold deterministically.

Lemma 4.1. *(Proof in Appendix B.3) For all $t = 1, 2, ..., \infty$, we have:*

(a) $\|\mathbf{A}\mathbf{x}^t - \mathbf{y}^t - \mathbf{b}\| \leq L_h\mu^{t-1}$.

(b) $\|\mathbf{y}^{t+1} - \mathbf{y}^t\| \leq \|\mathbf{A}\|\|\mathbf{x}^{t+1} - \mathbf{x}^t\| + 2L_h\mu^{t-1}$.

(c) $\|\mathbf{r}^t\| \leq \begin{cases} L_F \triangleq L_f + L_h\|\mathbf{A}\|, & \mu^t = \bar{\mu}; \\ L_F' \triangleq L_f + \frac{t_0+1}{t_0}L_h\|\mathbf{A}\|, & \mu^t = \eta/(t+t_0) \end{cases}.$

(d) $\|\boldsymbol{\varepsilon}^t\| \leq (\frac{\mu^{t-1}}{\mu^t} - 1)\|\mathbf{A}\|L_h$.

(e) $\mathcal{J}(\mathbf{x}^t, \mathbf{y}^t; \mu^t) - \mathcal{J}(\mathbf{x}^t, \mathbf{y}^t; \mu^{t-1}) \leq \Psi^t$, *where* $\Psi^t \triangleq \frac{L_h^2}{2}(\frac{(\mu^{t-1})^2}{\mu^t} - \mu^{t-1})$.

(f) $[\sum_{t=1}^\infty \Psi^t] \leq \begin{cases} 0, & \mu^t = \bar{\mu}; \\ \eta L_h^2, & \mu^t = \eta/(t+t_0) \end{cases}.$

Remarks. (***i***) Given our choices of μ^t and the fact that $\|\mathbf{A}\mathbf{x}^t - \mathbf{y}^t - \mathbf{b}\| \leq L_h\mu^{t-1}$, we have: $\|\mathbf{A}\mathbf{x}^t - \mathbf{y}^t - \mathbf{b}\| \to 0$. (***ii***) When $\|\mathbf{x}^{t+1} - \mathbf{x}^t\| \to 0$, it holds that $\|\mathbf{y}^{t+1} - \mathbf{y}^t\| \to 0$. (***iii***) We notice that $\|\mathbf{r}^t\|$ is upper bounded, and $L_F' \to L_F$ as $t_0 \to \infty$. (***iv***) We observe $\mu^{t-1} \to 0$ and $\frac{\mu^{t-1}}{\mu^t} \to 1$ as $t \to \infty$, resulting in $\|\boldsymbol{\varepsilon}^t\| \to 0$. (***v***) We focus on the term $\frac{1}{2\mu}\|\mathbf{A}\mathbf{x} - \mathbf{b} - \mathbf{y}\|_2^2$ in Problem (4). We have $\frac{1}{2\mu^t}\|\mathbf{A}\mathbf{x}^{t+1} - \mathbf{b} - \mathbf{y}^{t+1}\|_2^2 \leq \frac{1}{2\mu^t}(L_h\mu^t)^2 = \frac{L_h^2\mu^t}{2} \to 0$.

The following lemma is useful in our subsequent analysis.

Lemma 4.2. *(Proof in Appendix B.4) Let $\bar{\mathbf{x}}$ be any global optimal solution of Problem (1). We define:*

$$\Upsilon^t \triangleq F(\bar{\mathbf{x}}) - F(\mathbf{x}^t) + \tfrac{\mu^{t-1}L_h^2}{2} + \tfrac{2L_FL_h}{V_s}\|\mathbf{A}\|(\tfrac{\mu^{t-1}}{\mu^t} - 1).$$

We have the following results:

(a) $\langle\mathbf{r}^t, \bar{\mathbf{x}} - \mathbf{x}^t\rangle \leq \Upsilon^t - \frac{V_s}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2$.

(b) *If $\mu^t = \bar{\mu}$, we have: $\Upsilon^t \leq \frac{1}{2}\bar{\mu}L_h^2 - [\min_{i=1}^t F(\mathbf{x}^i)] + F(\bar{\mathbf{x}})$.*

(c) *If $\mu^t = \frac{\eta}{t+t_0}$, we have: $\sum_{i=1}^t \Upsilon^i \leq C_\Upsilon(\ln(t) + 1) - t[\min_{i=1}^t F(\mathbf{x}^i)] + tF(\bar{\mathbf{x}})$, where $C_\Upsilon \triangleq \frac{2L_FL_h}{V_s}\|\mathbf{A}\| + \frac{\eta L_h^2}{2}$.*

Remarks. Noticing that $F(\bar{\mathbf{x}}) - F(\mathbf{x}^t) \leq 0$ and $\frac{\mu^{t-1}}{\mu^t} - 1 \to 0$, we have $\Upsilon^t \to 0$ as $t \to \infty$. This results in: $\langle\mathbf{r}^t, \mathbf{x}^t - \bar{\mathbf{x}}\rangle \leq -\frac{V_s}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2$ in the limit.

4.1. Convergence Rate for SPGM-IHT

In this subsection, we assume that IHT strategy is used for solving the $\mathbf{x}$-subproblem.

We denote any limit point of **SPGM-IHT** as $(\dot{\mathbf{x}}, \dot{\mathbf{y}})$ and present the following useful definition.

Definition 4.3. (Approximate Lipschitz Stationary Point) Given any constant $\epsilon > 0$. Fix $\mu > 0$ to be a sufficiently small constant. A solution $(\dot{\mathbf{x}}, \dot{\mathbf{y}})$ is a ϵ-approximate Lipschitz stationary point if: $\mathrm{dist}^2(\dot{\mathbf{y}}, \arg\min_{\mathbf{y}} \mathcal{J}(\dot{\mathbf{x}}, \mathbf{y}; \mu)) + \mathrm{dist}^2(\dot{\mathbf{x}}, \arg\min_{\mathbf{x}} \delta(\mathbf{x}) + \dot{\mathcal{M}}(\mathbf{x}, \dot{\mathbf{x}}, \dot{\mathbf{y}}; \mu)) \leq \epsilon$, where $\dot{\mathcal{M}}(\cdot, \cdot, \cdot; \cdot)$ is defined in Problem (5).

The following theorem establishes the convergence of **SPGM-IHT**.

Theorem 4.4. *(Proof in Appendix C.1)* ***Convergence to Lipschitz Stationary Solutions****. We let $\mathcal{J}^{t+1} \triangleq \mathcal{J}(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}; \mu^t)$. We define Ψ^t as in Lemma 4.1. We have:*

(a) $\frac{\theta}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2 + \frac{1}{2\mu^0}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|_2^2 \leq \Psi^t + \mathcal{J}^t - \mathcal{J}^{t+1}$.

(b) $\sum_{t=1}^{T}[\frac{\theta}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2 + \frac{1}{2\mu^0}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|_2^2] \leq \mathcal{J}^1 - \mathcal{J}^{T+1} + \eta L_h^2 \triangleq C < \infty$.

(c) *Algorithm 1 finds an ϵ-approximate Lipschitz stationary point of Problem (1) in at most T iterations, where $T \leq \lceil \frac{2C}{\epsilon \min(\theta, (\mu^0)^{-1})} \rceil = \mathcal{O}(\epsilon^{-1})$.*

Remarks. The introduction of parameter $\theta > 0$ is important since it guarantees sufficient decrease condition and global convergence of Algorithm 1.

In what follows, we present enhanced convergence results for **SPGM-IHT**, leading to the attainment of the approximate global optimal solution $\bar{\mathbf{x}}$. We use the following quantities to measure the distance between $\mathbf{x}^t$ and $\bar{\mathbf{x}}$:

$$\Delta_{\mathbf{x}}^t \triangleq \|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2, \text{ and } \Delta_F^t \triangleq [\min_{i=1}^{t} F(\mathbf{x}^i)] - F(\bar{\mathbf{x}}).$$

We first have the following useful lemma.

Lemma 4.5. *(Proof in Appendix C.2) We define $H^t \triangleq A_s^2/\mu^t + M_s + \theta$. We have:*

(a) $\|\mathbf{x}^{t+1} - \mathbf{x}^t\| \leq \frac{2}{H^t}\|\mathbf{r}^t\|$.

(b) $\langle \mathbf{x}_+^t, \mathbf{x}_+^t \rangle \geq \langle \mathbf{x}_+^t, \mathbf{x}^{t+1} \rangle$*, where* $\mathbf{x}_+^t \triangleq \mathbf{x}^t - \mathbf{r}^t/H^t$.

(c) $\frac{1}{2}H^t\Delta_{\mathbf{x}}^{t+1} \leq \frac{1}{2}(H^t - V_s)\Delta_{\mathbf{x}}^t + \frac{3\|\mathbf{r}^t\|_2^2}{H^t} + \Upsilon^t + \|\mathbf{r}^t\|\|\bar{\mathbf{x}}\|$.

Remarks. As $\mu^t \to 0$, we have $H^t \to +\infty$, leading to $\|\mathbf{x}^{t+1} - \mathbf{x}^t\| \to 0$.

The following theorems establish the convergence of **SPGM-IHT** to the approximate global optimal solution $\bar{\mathbf{x}}$.

Theorem 4.6. *(Proof in Appendix C.3)* ***Convergence to the Approximate Global Optimal Solutions for Constant Stepsizes****. Assume constant stepsizes are used with $\mu^t = \bar{\mu}$. We define $\gamma \triangleq 1 - V_s/H$, and $H \triangleq A_s^2/\bar{\mu} + M_s + \theta$.*

(a) *We have the following recursive inequality:*

$$\tfrac{1}{2}\Delta_{\mathbf{x}}^{t+1} - \tfrac{\gamma}{2}\Delta_{\mathbf{x}}^t \leq \tfrac{L_F}{H}\|\bar{\mathbf{x}}\| + \tfrac{3L_F^2}{(H)^2} + (\tfrac{1}{2}\bar{\mu}L_h^2 - \Delta_F^t)\tfrac{1}{H}.$$

(b) *The following inequalities hold:*

$$\Delta_F^t \leq K_1\gamma^t + D_1\bar{\mu} + L_F\|\bar{\mathbf{x}}\|, \tag{12}$$
$$\Delta_{\mathbf{x}}^{t+1} \leq \left(K_1\gamma^t + D_1\bar{\mu} + L_F\|\bar{\mathbf{x}}\|\right)\tfrac{2}{V_s}, \tag{13}$$

where $K_1 \triangleq \frac{V_s}{2}\Delta_{\mathbf{x}}^1$, and $D_1 \triangleq \frac{3L_F^2}{A_s^2} + \frac{1}{2}L_h^2$.

Theorem 4.7. *(Proof in Appendix C.4)* ***Convergence to Approximate Global Optimal Solutions for Diminishing Stepsizes****. Assume diminishing stepsizes are used with $\mu^t = \frac{\eta}{t+t_0}$, where $\eta = A_s^2/V_s$. We let $L_F' \triangleq L_f + \frac{t_0+1}{t_0}L_h\|\mathbf{A}\|$, and $H^t \triangleq A_s^2/\mu^t + M_s + \theta$. We define Υ^t as in Lemma 4.2.*

(a) *We have the following recursive inequality:*

$$\begin{aligned}&\tfrac{1}{2}(H^{t+1} - V_s)\Delta_{\mathbf{x}}^{t+1}\\ &\leq \tfrac{1}{2}(H^t - V_s)\Delta_{\mathbf{x}}^t + 3(L_F')^2/(V_s \cdot t) + \Upsilon^t + L_F'\|\bar{\mathbf{x}}\|.\end{aligned}$$

(b) *The following inequalities hold:*

$$\Delta_F^t \leq \tfrac{K_2 + D_2(\ln(t)+1)}{t} + L_F'\|\bar{\mathbf{x}}\|, \tag{14}$$
$$\Delta_{\mathbf{x}}^{t+1} \leq (\tfrac{K_2 + D_2(\ln(t)+1)}{t+1} + L_F'\|\bar{\mathbf{x}}\|)\tfrac{2}{V_s}, \tag{15}$$

where $K_2 \triangleq \frac{H^1}{2}\Delta_{\mathbf{x}}^1$, $D_2 \triangleq \frac{3(L_F')^2}{V_s} + C_\Upsilon$, and C_Υ is defined in Lemma 4.2.

Remarks. (***i***) As $t_0 \to +\infty$, we have $L_F' = L_f + \frac{t_0+1}{t_0}L_h\|\mathbf{A}\| \to L_F$. (***ii***) The irreducible error terms for **SPGM-IHT** $\{L_F\|\bar{\mathbf{x}}\|, 2L_F\|\bar{\mathbf{x}}\|/V_s, L_F'\|\bar{\mathbf{x}}\|, 2L_F\|\bar{\mathbf{x}}\|/V_s$, as presented respectively in Inequalities (12), (13), (14), and (15), are influenced by the Lipschitz constant L_F, the strong convexity parameter V_s, and the ℓ_2 norm of the optimal solution $\|\bar{\mathbf{x}}\|$. This highlights the challenges inherent in solving such NP-hard problems. **SPGM-IHT** shows a higher likelihood of converging to the global optimum when $\|\bar{\mathbf{x}}\|$ and L_F are small while V_s is large. (***iii***) Given $\|\bar{\mathbf{x}}\|_0 \leq s$, we obtain: $\|\bar{\mathbf{x}}\| \leq \sqrt{s}\|\bar{\mathbf{x}}\|_\infty$, rendering the irreducible error terms small. Hence, our theoretical bounds can exploit the inherent sparsity structure of the problem. (***iv***) Our theoretical results do not rely on the Restricted Isometry Property (RIP) condition (Candes & Tao, 2005). Instead, they primarily hinge on the Lipschitz continuity of $F(\mathbf{x})$ (see Assumption 2.1), along with the restricted strong convexity and smoothness of the objective function (see Assumption 2.2). (***v***) We do not claim to achieve the exact global optimal solution of Problem (1), as that would imply solving an NP-hard problem outright. Instead, we aim to identify an approximate solution that closely approaches the global optimum. (***vi***) We compare our results in Inequalities (14) and (15) with the following estimates generated by the **PSGD** method (Liu et al., 2019):

$$[\min_{i=1}^{t} F(\mathbf{x}^i)] - F(\bar{\mathbf{x}}) \leq \mathcal{O}(1/t) + L_F\|\bar{\mathbf{x}}\|, \tag{16}$$
$$\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2 \leq \mathcal{O}(1/t) + L_F\|\bar{\mathbf{x}}\| \cdot \tfrac{2}{V_s}. \tag{17}$$

Inequalities (16) and (17) have been established in **Corollary III.4** and **Corollary III.8** of (Liu et al., 2019), respectively. We conclude that the irreducible estimation error terms $L'_F\|\bar{\mathbf{x}}\|$ and $L'_F\|\bar{\mathbf{x}}\| \cdot \frac{2}{V_s}$, as specified in Inequalities (14) and (15), match the best-known error bounds presented in (Liu et al., 2019) for this nonconvex NP-hard problem. ***(vii)*** We also draw comparisons with the following estimate generated by the **GHTP** method (Yuan et al., 2017):

$$\|\mathbf{x}^t - \bar{\mathbf{x}}\| \leq \mathcal{O}(\vartheta^t) + c_0\sqrt{s}\|\nabla F(\bar{\mathbf{x}})\|_\infty, \tag{18}$$

where $\vartheta \in (0,1)$, and c_0 is a constant related to the stepsize. Inequality (18) has been established in Theorem 2 of (Yuan et al., 2017). Noting that $\|\bar{\mathbf{x}}\| \leq \sqrt{s}\|\bar{\mathbf{x}}\|_\infty$, we derive from Inequalities (14) and (15):

$$[\min_{i=1}^{t} F(\mathbf{x}^i)] - F(\bar{\mathbf{x}}) \leq \tilde{\mathcal{O}}(1/t) + L'_F\sqrt{s}\|\bar{\mathbf{x}}\|_\infty, \tag{19}$$
$$\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2 \leq \tilde{\mathcal{O}}(1/t) + L'_F\sqrt{s}\|\bar{\mathbf{x}}\|_\infty \cdot \tfrac{2}{V_s}. \tag{20}$$

We conclude that our results in Inequalities (19) and (20) are in some sense analogous to those in (Yuan et al., 2017).

4.2. Convergence Rate for SPGM-BCD

In this subsection, we assume that BCD strategy is used for solving the $\mathbf{x}$-subproblem.

We assume that the working set B is selected randomly and uniformly from $\Omega_n^k \triangleq \{\mathrm{B}_1, \mathrm{B}_2, ..., \mathrm{B}_{C_n^k}\}$. **SPGM-BCD** generates a random output $\mathbf{x}^t$ with $t = 1, 2, ...$, based on the observed realization of the random variable $\xi^{t-1} \triangleq \{\mathrm{B}^1, \mathrm{B}^2, ..., \mathrm{B}^{t-1}\}$. The expectation of a random variable is denoted by $\mathbb{E}_{\xi^t}[\cdot]$. The following lemma is useful in this context.

Lemma 4.8. *(Proof in Appendix C.5) For any $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{z} \in \mathbb{R}^n$, we have $\frac{1}{C_n^k}\sum_{\mathrm{B}\in\Omega_n^k} \mathbf{x}^\top(\mathrm{U}_\mathrm{B}\mathrm{U}_\mathrm{B}^\top)\mathbf{z} = Z_n^k\langle\mathbf{x},\mathbf{z}\rangle$, and $\mathbb{E}_\mathrm{B}\|\mathbf{x}_\mathrm{B}\|_2^2 = Z_n^k\|\mathbf{x}\|_2^2$, where $Z_n^k \triangleq \frac{k}{n}$.*

We denote any limit point of **SPGM-BCD** as $(\ddot{\mathbf{x}}, \ddot{\mathbf{y}})$ and offer the following useful definition.

Definition 4.9. (Approximate block-k Stationary Point) Given any constant $\epsilon > 0$. Fix $\mu > 0$ to be a sufficiently small constant. A solution $(\ddot{\mathbf{x}}, \ddot{\mathbf{y}})$ is an ϵ-approximate block-k stationary point if: $\mathrm{dist}^2(\ddot{\mathbf{y}}, \arg\min_\mathbf{y} \mathcal{J}(\ddot{\mathbf{x}}, \mathbf{y}; \mu)) + \frac{1}{C_n^k}\sum_{\mathrm{B}\in\Omega_n^k} \mathrm{dist}^2(\ddot{\mathbf{x}}_\mathrm{B}, \arg\min_{\mathbf{x}_\mathrm{B}} \delta(\mathrm{U}_\mathrm{B}\mathbf{x}_\mathrm{B} + \mathrm{U}_{\mathrm{B}^c}\ddot{\mathbf{x}}_{\mathrm{B}^c}) + \ddot{\mathcal{M}}(\mathrm{U}_\mathrm{B}\mathbf{x}_\mathrm{B} + \mathrm{U}_{\mathrm{B}^c}\ddot{\mathbf{x}}_{\mathrm{B}^c}, \ddot{\mathbf{x}}, \ddot{\mathbf{y}}; \mu)) \leq \epsilon$, where $\ddot{\mathcal{M}}(\cdot,\cdot,\cdot;\cdot)$ is defined in Problem (6).

The following theorem establishes the convergence of **SPGM-BCD**.

Theorem 4.10. *(Proof in Appendix C.6)* ***Convergence to Block-k Stationary Solutions****. We let $\theta \triangleq \frac{\theta_1}{\mu^1} + \theta_2$, and $\mathcal{J}^{t+1} \triangleq \mathbb{E}_{\xi^t}[\mathcal{J}(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}; \mu^t)]$. We define Ψ as in Lemma 4.1. We have:*

(a) $\mathbb{E}_{\xi^t}[\frac{\theta}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2 + \frac{1}{2\mu^1}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|_2^2] \leq \Psi^t + \mathcal{J}^t - \mathcal{J}^{t+1}$.

(b) $\mathbb{E}_{\xi^T}[\sum_{t=1}^T \frac{\theta}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2 + \frac{1}{2\mu^1}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|_2^2] \leq \mathcal{J}^1 - \mathcal{J}^{T+1} + \eta L_h^2 \triangleq C < \infty$.

(c) *Algorithm 1 finds an ϵ-approximate block-k stationary point of Problem (1) in at most T iterations in the sense of expectation, where $T \leq \lceil\frac{2C}{\epsilon\min(\theta,(\mu^0)^{-1})}\rceil = \mathcal{O}(\epsilon^{-1})$.*

Remarks. Theorem 4.10 resembles Theorem 4.4, with the key distinction being that **SPGM-IHT** deterministically converges to a Lipschitz stationary point, whereas **SPGM-BCD** converges to a block-k stationary point in expectation.

In what follows, we present enhanced convergence results for **SPGM-BCD**, leading to the attainment of the approximate global optimal solution $\bar{\mathbf{x}}$. For notation convenience, we define

$$\Delta_\mathbf{x}^t \triangleq \mathbb{E}_{\xi^t}[\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2],\ \Delta_F^t \triangleq \mathbb{E}_{\xi^t}[(\min_{i=1}^{t} F(\mathbf{x}^i)) - F(\bar{\mathbf{x}})],$$
$$\overline{\mathrm{V}} \triangleq \max_{\mathrm{B}\in\Omega_n^k} \boldsymbol{\lambda}_{\max}(\tilde{\mathbf{M}}_{\mathrm{BB}}),\ \underline{\mathrm{V}} \triangleq \min_{\mathrm{B}\in\Omega_n^k} \boldsymbol{\lambda}_{\min}(\tilde{\mathbf{M}}_{\mathrm{BB}}),$$
$$\overline{\mathrm{A}} \triangleq \max_{\mathrm{B}\in\Omega_n^k} \boldsymbol{\lambda}_{\max}(\mathbf{A}^\top\mathbf{A})_{\mathrm{BB}}),\ \underline{\mathrm{A}} \triangleq \min_{\mathrm{B}\in\Omega_n^k} \boldsymbol{\lambda}_{\min}((\mathbf{A}^\top\mathbf{A})_{\mathrm{BB}}),$$
$$\overline{\mathrm{H}}^t \triangleq \frac{\overline{\mathrm{A}}+\theta_1}{\mu^t} + \overline{\mathrm{V}} + \theta_2,\ \underline{\mathrm{H}}^t \triangleq \frac{\underline{\mathrm{A}}+\theta_1}{\mu^t} + \underline{\mathrm{V}} + \theta_2, \kappa^t \triangleq \frac{\overline{\mathrm{H}}^t}{\underline{\mathrm{H}}^t},$$

where $\underline{\mathrm{V}} \geq V_s$, and $\underline{\mathrm{A}}$ could be zero.

We first have the following two useful lemmas.

Lemma 4.11. *(Proof in Appendix C.8) Given any constant $\tilde{\epsilon} > 0$. If θ_1 and θ_2 are sufficiently large such that $\theta_1 \geq \mathcal{T}_1(\tilde{\epsilon}) \triangleq \frac{\overline{\mathrm{A}}-\underline{\mathrm{A}}(1+\tilde{\epsilon})}{\tilde{\epsilon}}$ and $\theta_2 \geq \mathcal{T}_2(\tilde{\epsilon}) \triangleq \frac{\overline{\mathrm{V}}-\underline{\mathrm{V}}(1+\tilde{\epsilon})}{\tilde{\epsilon}}$, we have:*

$$\kappa^t \triangleq \frac{\overline{\mathrm{H}}^t}{\underline{\mathrm{H}}^t} \leq 1 + \tilde{\epsilon}.$$

Lemma 4.12. *(Proof in Appendix C.7) We define Z_n^k as in Lemma 4.8. We define:*

$$\begin{aligned}\mathbf{H}^t &\triangleq (\mathbf{A}^\top\mathbf{A} + \theta_1\mathbf{I}_n)/\mu^t + \tilde{\mathbf{M}} + \theta_2\mathbf{I}_n,\\ \mathbf{H}_*^t &\triangleq \mathrm{U}_{\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\top\mathbf{H}^t\mathrm{U}_{\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\top.\end{aligned}$$

For all $t \geq 1$, we have the following results:

(a) $\mathbb{E}_{\xi^t}[\|\mathbf{x}^{t+1} - \mathbf{x}^t\|] \leq \frac{2}{\underline{\mathrm{H}}^t}\mathbb{E}_{\xi^t}[\|\mathbf{r}^t\|]$.

(b) $\mathbb{E}_{\xi^t}[\langle[[\mathbf{H}^t]_{\mathrm{BB}}](\mathbf{x}_\mathrm{B}^{t+1} - \mathbf{x}_\mathrm{B}^t), \mathbf{x}_\mathrm{B}^{t+1}\rangle] = -Z_n^k\langle\mathbf{r}^t, \mathbf{x}^{t+1}\rangle$.

(c) $\mathbb{E}_{\xi^t}[\frac{1}{2}\|\mathbf{x}^{t+1} - \bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2 - \frac{1}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2] \leq Z_n^k \cdot \{\Upsilon^t + 2/\underline{\mathrm{H}}^t\|\mathbf{r}^t\|_2^2 + (1 + 2\kappa^t)\|\mathbf{r}^t\|\|\bar{\mathbf{x}}\| - \frac{V_s}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2\}$.

Remarks. When $\mu^t \to 0$, we have $\underline{\mathrm{H}}^t \to +\infty$, leading to $\mathbb{E}_{\xi^t}[\|\mathbf{x}^{t+1} - \mathbf{x}^t\|] \to 0$.

The following theorems establish the convergence of **SPGM-BCD** to the approximate global optimal solution $\bar{\mathbf{x}}$.

Theorem 4.13. *(Proof in Appendix C.9)* ***Convergence to Approximate Global Optimal Solutions for Constant Stepsizes****. Assume constant stepsizes are used with $\mu^t = \bar{\mu}$. Given any constant $\tilde{\epsilon} > 0$. Assume that $\theta_1 \geq \mathcal{T}_1(\tilde{\epsilon})$ and $\theta_2 \geq \mathcal{T}_2(\tilde{\epsilon})$, where $\mathcal{T}_1(\cdot)$ and $\mathcal{T}_2(\cdot)$ are define in Lemma 4.11. We let $\overline{\mathrm{H}} \triangleq \frac{\overline{\mathrm{A}}+\theta_1}{\bar{\mu}} + \overline{\mathrm{V}} + \theta_2$, $\underline{\mathrm{H}} \triangleq \frac{\underline{\mathrm{A}}+\theta_1}{\bar{\mu}} + \underline{\mathrm{V}} + \theta_2$, and $\gamma \triangleq 1 - \frac{V_s}{\overline{\mathrm{H}}} \in (0,1)$. We define $\mathbf{H}_*^t$ as in Lemma 4.12, and Υ^t as in Lemma 4.2.*

(a) *We have the following recursive inequality:*

$$\mathbb{E}_{\xi^{t+1}}[\tfrac{1}{2}\|\mathbf{x}^{t+1} - \bar{\mathbf{x}}\|_{\mathbf{H}_*^{t+1}}^2] - \gamma \mathbb{E}_{\xi^t}[\tfrac{1}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2]$$
$$\leq Z_n^k[\Upsilon^t + \tfrac{2(L_F)^2}{\underline{\mathrm{H}}} + (3+2\tilde{\epsilon})L_F\|\bar{\mathbf{x}}\|].$$

(b) *The following inequalities hold:*

$$\Delta_F^t \leq \; K_3\gamma^t + D_3\bar{\mu} + (3+2\tilde{\epsilon})L_F\|\bar{\mathbf{x}}\|, \tag{21}$$
$$\Delta_{\mathbf{x}}^{t+1} \leq \; (K_3\gamma^t + D_3\bar{\mu} + (3+2\tilde{\epsilon})L_F\|\bar{\mathbf{x}}\|)\tfrac{2}{V_s}, \tag{22}$$

where $K_3 \triangleq \frac{V_s(1+\tilde{\epsilon})}{2}\Delta_{\mathbf{x}}^1$, and $D_3 \triangleq \frac{2(L_F)^2}{\theta_1+\overline{\mathrm{A}}} + \frac{L_h^2}{2}$.

Theorem 4.14. *(Proof in Appendix C.10)* ***Convergence to Approximate Global Optimal Solutions for Diminishing Stepsizes****. Assume diminishing stepsizes are used with $\mu^t = \frac{\eta}{t+t_0}$, where $\eta = \frac{\overline{\mathrm{A}}+\theta_1}{V_s}$. Given any constant $\tilde{\epsilon} > 0$. Assume that $\theta_1 \geq \mathcal{T}_1(\tilde{\epsilon})$ and $\theta_2 \geq \mathcal{T}_2(\tilde{\epsilon})$, where $\mathcal{T}_1(\cdot)$ and $\mathcal{T}_2(\cdot)$ are define in Lemma 4.11. We let $\overline{\mathrm{H}}^t \triangleq \frac{\overline{\mathrm{A}}+\theta_1}{\mu^t} + \overline{\mathrm{V}} + \theta_2$, $\underline{\mathrm{H}}^t \triangleq \frac{\underline{\mathrm{A}}+\theta_1}{\mu^t} + \underline{\mathrm{V}} + \theta_2$, and $L_F' \triangleq L_f + \frac{t_0+1}{t_0}L_h\|\mathbf{A}\|$. We define $\mathbf{H}_*^t$ as in Lemma 4.12, and Υ^t as in Lemma 4.2.*

(a) *We have the following recursive inequality:*

$$\Phi^{t+1} - \Phi^t \leq Z_n^k\Upsilon^t + \tfrac{2Z_n^k(L_F')^2}{\underline{\mathrm{H}}^t} + (3+2\tilde{\epsilon})Z_n^k L_F'\|\bar{\mathbf{x}}\|,$$

where $\Phi^t \triangleq \mathbb{E}_{\xi^t}[\frac{1}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2] - Z_n^k\frac{V_s}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2$.

(b) *The following inequalities hold:*

$$\Delta_F^t \leq \tfrac{K_4+D_4(1+\ln(t))}{t} + (3+2\tilde{\epsilon})L_F'\|\bar{\mathbf{x}}\|, \tag{23}$$
$$\Delta_{\mathbf{x}}^{t+1} \leq (\tfrac{K_4+D_4(\ln(t)+1)}{t+1} + (3+2\tilde{\epsilon})L_F'\|\bar{\mathbf{x}}\|)\tfrac{2}{V_s}, \tag{24}$$

where $K_4 \triangleq \frac{\overline{\mathrm{H}}^1}{2}\Delta_{\mathbf{x}}^1$, $D_4 \triangleq \frac{2(L_F')^2}{V_s} + C_\Upsilon$, and C_Υ is defined in Lemma 4.2.

Remarks. (***i***) The convergence rates in Theorems 4.13 and 4.6 are similar, as are those in Theorems 4.14 and 4.7. However, analyzing **SPGM-BCD** is more intricate than **SPGM-IHT** due to the utilization of a general Hessian matrix $\mathbf{H}^t$ and a stochastic mechanism of **SPGM-BCD**, in contrast to the utilization of a scaled identity matrix $H^t\mathbf{I}_n$ and a deterministic mechanism of **SPGM-IHT**. Consequently, their theoretical derivations differ significantly. (***ii***) As $\tilde{\epsilon} \to 0$ and $t_0 \to +\infty$, the irreducible estimation error terms for Δ_F^t in Inequalities (21) and (23) simplify to $3L_F\|\bar{\mathbf{x}}\|$, which is three times the bound of **PSGD** in (Liu et al., 2019). Our bounds leverage the inherent sparsity of the problem.

5. Experiments

This section evaluates the effectiveness of **SPGM-IHT** and **SPGM-BCD**, comparing them with five state-of-the-art nonsmooth sparsity constrained optimization algorithms: (***i***) Projective Subgradient Descent (PSGD) (Liu et al., 2019). (***ii***) Alternating Direction Method of Mutipliers based on IHT (ADMM-IHT) (He & Yuan, 2012). (***iii***) Dual Iterative Hard Thresholding(DIHT) (Yuan et al., 2020b). (***iv***) Convex ℓ_1 Approximation Method (CVX-ℓ_1) (Candes & Tao, 2005). (***v***) Nonconvex ℓ_p Approximation Method (NCVX-ℓ_p) (Xu et al., 2012).

Our experiments reveal that **SPGM-IHT** is on par with existing IHT-style methods, and **SPGM-BCD** consistently delivers the best performance. This outcome is expected as **SPGM-IHT** is an IHT-style method itself, while **SPGM-BCD** excels in identifying stronger stationary points compared to other approaches.

Due to space constraints, detailed experiment results are provided in Appendix Section D. Our MATLAB code is available on the authors research webpage.

6. Conclusions

This paper explores Smoothing Proximal Gradient Methods (SPGM) for solving nonsmooth sparsity constrained optimization problems. We discuss two specific variants of SPGM: one based on Iterative Hard Thresholding (**SPGM-IHT**) and the other on Block Coordinate Decomposition (**SPGM-BCD**). We provide both smooth and optimality analyses for the smoothing functions, demonstrating that **SPGM-BCD** discovers stronger stationary points of the nonsmooth nonconvex problem. We offer theoretical insights into the convergence rates of the **SPGM-IHT** and **SPGM-BCD** algorithms. Our bounds depend on the Lipschitz constant of the objective function, the strong convexity parameter of its smooth component, and the ℓ_2 norm of the global optimal point. Leveraging the inherent sparsity of the optimization problem, our bounds align with the most competitive error estimates in the field. Finally, numerical experiments demonstrate that **SPGM-IHT** performs on par with existing IHT-style methods, while **SPGM-BCD** consistently delivers state-of-the-art numerical performance.

Acknowledgments

This work was supported by NSFC (12271278, 61772570), Guangdong Natural Science Funds for Distinguished Young Scholar (2018B030306025), and the Major Key Project of PCL (PCL2021A13).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bahmani, S., Raj, B., and Boufounos, P. T. Greedy sparsity-constrained optimization. *The Journal of Machine Learning Research*, 14(1):807–841, 2013.

Bauschke, H. H., Combettes, P. L., et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

Beck, A. and Eldar, Y. C. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

Beck, A. and Hallak, N. On the minimization over sparse symmetric sets: projections, optimality conditions, and algorithms. *Mathematics of Operations Research*, 41(1): 196–223, 2016.

Beck, A. and Hallak, N. Optimization problems involving group sparsity terms. *Mathematical Programming*, 178: 39–67, 2019.

Beck, A. and Vaisbourd, Y. The sparse principal component analysis problem: Optimality conditions and algorithms. *Journal of Optimization Theory and Applications*, 170(1): 119–143, 2016. ISSN 1573-2878.

Bi, S., Liu, X., and Pan, S. Exact penalty decomposition method for zero-norm minimization based on mpec formulation. *SIAM Journal on Scientific Computing*, 36(4): A1451–A1477, 2014.

Bian, W. and Chen, X. A smoothing proximal gradient algorithm for nonsmooth convex regression with cardinality penalty. *SIAM Journal on Numerical Analysis*, 58(1): 858–883, 2020.

Blumensath, T. and Davies, M. E. Gradient pursuits. *IEEE Transactions on Signal Processing*, 56(6):2370–2382, 2008.

Blumensath, T. and Davies, M. E. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265 – 274, 2009.

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

Candes, E. J. and Tao, T. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Candes, E. J., Wakin, M. B., and Boyd, S. P. Enhancing s-parsity by reweighted l1 minimization. *Journal of Fourier analysis and applications*, 14:877–905, 2008.

Chen, S. and Banerjee, A. Sparse linear isotonic models. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 84:1270–1279, 2018.

Chen, X. Smoothing methods for nonsmooth, nonconvex minimization. *Mathematical programming*, 134(1):71–99, 2012.

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.

Jain, P., Tewari, A., and Kar, P. On iterative hard thresholding methods for high-dimensional m-estimation. *Neural Information Processing Systems (NeurIPS)*, pp. 685–693, 2014.

Liu, D., Li, S., and Shen, Y. One-bit compressive sensing with projected subgradient method under sparsity constraints. *IEEE Transactions on Information Theory*, 65 (10):6650–6663, 2019.

Lu, Z. Iterative hard thresholding methods for ℓ_0 regularized convex cone programming. *Mathematical Programming*, 147(1-2):125–154, 2014.

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

Nguyen, N., Needell, D., and Woolf, T. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Transactions on Information Theory*, 63 (11):6869–6895, 2017.

Tropp, J. A. and Gilbert, A. C. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

Tseng, P. and Yun, S. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423, 2009.

Xu, Y. and Yin, W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013.

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.

Yuan, G. Coordinate descent methods for fractional minimization. In *International Conference on Machine Learning (ICML)*, pp. 40488–40518. PMLR, 2023a.

Yuan, G. Coordinate descent methods for dc minimization: Optimality conditions and global convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pp. 11034–11042, 2023b.

Yuan, G. and Ghanem, B. $\ell_0 tv$: A sparse optimization method for impulse noise image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (2):352–364, 2019.

Yuan, G., Shen, L., and Zheng, W. A decomposition algorithm for the sparse generalized eigenvalue problem. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6113–6122, 2019.

Yuan, G., Shen, L., and Zheng, W.-S. A block decomposition algorithm for sparse optimization. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2020a.

Yuan, X., Li, P., and Zhang, T. Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18:166:1–166:43, 2017.

Yuan, X.-T., Liu, B., Wang, L., Liu, Q., and Metaxas, D. N. Dual iterative hard thresholding. *Journal of Machine Learning Research*, 21(152):1–50, 2020b.

Zeng, J., Lin, S., and Xu, Z. Sparse regularization: Convergence of iterative jumping thresholding algorithm. *IEEE Transactions on Signal Processing*, 64(19):5106–5118, 2016.

Zhang, T. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11(35):1081–1107, 2010.

Appendix

The appendix is organized as follows.

Appendix A contains some useful lemmas.

Appendix B includes the proofs for Section 3.

Appendix C presents the proofs for Section 4.

Appendix D provides the experimental results.

A. Some Useful Lemmas

We present some useful lemmas that will be used subsequently.

Lemma A.1. (**Pythagoras Relation**) For any symmetric matrix $\mathbf{H} \in \mathbb{R}^{n\times n}$ with $\mathbf{H}^{\mathsf{T}} = \mathbf{H}$ and any vectors $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{c} \in \mathbb{R}^n$, we have:

$$\frac{1}{2}\|\mathbf{a}-\mathbf{b}\|_{\mathbf{H}}^2 - \frac{1}{2}\|\mathbf{c}-\mathbf{b}\|_{\mathbf{H}}^2 = \frac{1}{2}\|\mathbf{a}-\mathbf{c}\|_{\mathbf{H}}^2 - \langle \mathbf{a}-\mathbf{c}, \mathbf{H}(\mathbf{b}-\mathbf{c})\rangle.$$

Lemma A.2. *Assume $\gamma \in (0,1)$. Denote γ^t as the t-th power of γ. Let $\{\Phi^t\}_{t=1}^{\infty}$ and $\{\Lambda^t\}_{t=1}^{\infty}$ be any two non-negative sequences. If the following inequality is satisfied for all t: $\Phi^{t+1} \leq \gamma\Phi^t + \Lambda^t$. Then, it implies the following inequality: $\Phi^{t+1} \leq \gamma^t\Phi^1 + \frac{\max_{i=1}^t(\Lambda^i)}{1-\gamma}$.*

Proof. Using basic induction, we have the following results:

$$\begin{aligned}
&t=1,\ \Phi^2 \leq \gamma\Phi^1 + \Lambda^1\\
&t=2,\ \Phi^3 \leq \gamma\Phi^2 + \Lambda^2 \leq \gamma(\gamma\Phi^1 + \Lambda^2) + \Lambda^1 = \gamma^2\Phi^1 + (\Lambda^2 + \gamma\Lambda^1)\\
&t=3,\ \Phi^4 \leq \gamma\Phi^3 + \Lambda^3 \leq \gamma(\gamma^2\Phi^1 + (\Lambda^2 + \gamma\Lambda^1)) + \Lambda^3 = \gamma^3\Phi^1 + (\Lambda^3 + \gamma\Lambda^2 + \gamma^2\Lambda^1)\\
&...
\end{aligned}$$

Therefore, we obtain:

$$\Phi^{T+1} \leq \gamma^T\Phi^1 + \sum_{i=1}^{T}\Lambda^i\gamma^{T-i} \overset{①}{\leq} \gamma^T\Phi^1 + (\max_{i=1}^{T}\Lambda^i)\cdot(\sum_{i=1}^{T}\gamma^{T-i}) \overset{②}{\leq} \gamma^T\Phi^1 + \frac{(\max_{i=1}^T\Lambda^i)}{1-\gamma},$$

where step ① uses the Cauchy-Schwarz Inequality; step ② uses the fact that

$$\sum_{i=1}^{t}\gamma^{t-i} = 1 + \gamma^1 + \gamma^2 + ... + \gamma^{t-1} = \frac{1-\gamma^t}{1-\gamma} < \frac{1}{1-\gamma}.$$

□

B. Proofs for Section 3

B.1. Proof of Lemma 3.1

Proof. Without loss of generality, we assume $\mu_1 < \mu_2$. For all $\mathbf{x}$ with $\|\mathbf{x}\|_0 \leq s$, we define:

$$\psi(\mu) \triangleq \mathbf{G}(\mathbf{x};\mu) = f(\mathbf{x}) + h(\mathbb{P}_\mu(\mathbf{c})) + \tfrac{1}{2\mu}\|\mathbf{c} - \mathbb{P}_\mu(\mathbf{c})\|_2^2 \text{ with } \mathbf{c} \triangleq \mathbf{A}\mathbf{x} - \mathbf{b}. \quad (25)$$

Using the definition of $\mathbb{P}_\mu(\mathbf{c})$ as shown in Equation (2), we have for any given μ_1 and μ_2:

$$\mathbb{P}_{\mu_1}(\mathbf{c}) = \arg\min_{\mathbf{y}}\ h(\mathbf{y}) + \tfrac{1}{2\mu_1}\|\mathbf{y}-\mathbf{c}\|_2^2, \text{ and } \mathbb{P}_{\mu_2}(\mathbf{c}) = \arg\min_{\mathbf{y}}\ h(\mathbf{y}) + \tfrac{1}{2\mu_2}\|\mathbf{y}-\mathbf{c}\|_2^2.$$

By the optimality of $\mathbb{P}_{\mu_1}(\mathbf{c})$ and $\mathbb{P}_{\mu_2}(\mathbf{c})$, we obtain:

$$\mathbf{c} - \mathbb{P}_{\mu_1}(\mathbf{c}) \in \mu_1 \partial h(\mathbb{P}_{\mu_1}(\mathbf{c})), \;\; \text{and} \;\; \mathbf{c} - \mathbb{P}_{\mu_2}(\mathbf{c}) \in \mu_2 \partial h(\mathbb{P}_{\mu_2}(\mathbf{c})). \tag{26}$$

(a) We now prove that $\psi(\mu)$ is a decreasing function. For any $\mathbf{p}_1 \in \partial h(\mathbb{P}_{\mu_1}(\mathbf{c}))$ and $\mathbf{p}_2 \in \partial h(\mathbb{P}_{\mu_2}(\mathbf{c}))$, we have:

$$\begin{aligned}
\psi(\mu_2) - \psi(\mu_1) &\overset{①}{=} h(\mathbb{P}_{\mu_2}(\mathbf{c})) + \tfrac{1}{2\mu_2}\|\mathbf{c} - \mathbb{P}_{\mu_2}(\mathbf{c})\|_2^2 - h(\mathbb{P}_{\mu_1}(\mathbf{c})) - \tfrac{1}{2\mu_1}\|\mathbf{c} - \mathbb{P}_{\mu_1}(\mathbf{c})\|_2^2 \\
&\overset{②}{\leq} \langle \mathbb{P}_{\mu_2}(\mathbf{c}) - \mathbb{P}_{\mu_1}(\mathbf{c}), \mathbf{p}_2 \rangle + \tfrac{1}{2\mu_2}\|\mathbf{c} - \mathbb{P}_{\mu_2}(\mathbf{c})\|_2^2 - \tfrac{1}{2\mu_1}\|\mathbf{c} - \mathbb{P}_{\mu_1}(\mathbf{c})\|_2^2 \\
&\overset{③}{=} \langle \mu_1 \mathbf{p}_1 - \mu_2 \mathbf{p}_2, \mathbf{p}_2 \rangle + \tfrac{\mu_2}{2}\|\mathbf{p}_2\|_2^2 - \tfrac{\mu_1}{2}\|\mathbf{p}_1\|_2^2 \\
&= \langle \mu_1 \mathbf{p}_1, \mathbf{p}_2 \rangle - \tfrac{\mu_2}{2}\|\mathbf{p}_2\|_2^2 - \tfrac{\mu_1}{2}\|\mathbf{p}_1\|_2^2 \\
&\overset{④}{\leq} \langle \mu_1 \mathbf{p}_1, \mathbf{p}_2 \rangle - \tfrac{\mu_1}{2}\|\mathbf{p}_2\|_2^2 - \tfrac{\mu_1}{2}\|\mathbf{p}_1\|_2^2 \\
&= -\tfrac{\mu_1}{2}\|\mathbf{p}_1 - \mathbf{p}_2\|_2^2 \leq 0,
\end{aligned}$$

where step ① uses the definition of $\psi(\mu)$ in Equation (25); step ② uses the convexity of $h(\cdot)$; step ③ uses the optimality of $\mathbb{P}_{\mu_1}(\mathbf{c})$ and $\mathbb{P}_{\mu_2}(\mathbf{c})$ in Equation (26); step ④ uses $0 < \mu_1 < \mu_2$.

(b) We now prove that $\psi(\mu)$ is $(\frac{1}{2}L_h^2)$-Lipschitz. For any $\mathbf{p}_1 \in \partial h(\mathbb{P}_{\mu_1}(\mathbf{c}))$ and $\mathbf{p}_2 \in \partial h(\mathbb{P}_{\mu_2}(\mathbf{c}))$, we have:

$$\begin{aligned}
\psi(\mu_1) - \psi(\mu_2) &\overset{①}{=} h(\mathbb{P}_{\mu_1}(\mathbf{c})) + \tfrac{1}{2\mu_1}\|\mathbf{c} - \mathbb{P}_{\mu_1}(\mathbf{c})\|_2^2 - h(\mathbb{P}_{\mu_2}(\mathbf{c})) - \tfrac{1}{2\mu_2}\|\mathbf{c} - \mathbb{P}_{\mu_2}(\mathbf{c})\|_2^2 \\
&\overset{②}{\leq} \langle \mathbb{P}_{\mu_1}(\mathbf{c}) - \mathbb{P}_{\mu_2}(\mathbf{c}), \mathbf{p}_1 \rangle + \tfrac{1}{2\mu_1}\|\mathbf{c} - \mathbb{P}_{\mu_1}(\mathbf{c})\|_2^2 - \tfrac{1}{2\mu_2}\|\mathbf{c} - \mathbb{P}_{\mu_2}(\mathbf{c})\|_2^2 \\
&\overset{③}{=} \langle [\mathbf{c} - \mathbb{P}_{\mu_2}(\mathbf{c})] - [\mathbf{c} - \mathbb{P}_{\mu_1}(\mathbf{c})], \mathbf{p}_1 \rangle + \tfrac{1}{2\mu_1}\|\mathbf{c} - \mathbb{P}_{\mu_1}(\mathbf{c})\|_2^2 - \tfrac{1}{2\mu_2}\|\mathbf{c} - \mathbb{P}_{\mu_2}(\mathbf{c})\|_2^2 \\
&\overset{④}{=} \langle \mu_2 \mathbf{p}_2 - \mu_1 \mathbf{p}_1, \mathbf{p}_1 \rangle + \tfrac{\mu_1}{2}\|\mathbf{p}_1\|_2^2 - \tfrac{\mu_2}{2}\|\mathbf{p}_2\|_2^2 \\
&= -\tfrac{\mu_2}{2}\|\mathbf{p}_2\|_2^2 + \mu_2 \langle \mathbf{p}_1, \mathbf{p}_2 \rangle - \tfrac{\mu_1}{2}\|\mathbf{p}_1\|_2^2 \\
&\overset{⑤}{\leq} \tfrac{\mu_2}{2}\|\mathbf{p}_1\|_2^2 - \tfrac{\mu_1}{2}\|\mathbf{p}_1\|_2^2 \\
&\overset{⑥}{\leq} \tfrac{\mu_2 - \mu_1}{2} \cdot L_h^2,
\end{aligned}$$

where step ① uses the definition of $\psi(\mu)$ in Equation (25); step ② uses the convexity of $h(\cdot)$; step ③ uses the fact that $\mathbb{P}_{\mu_1}(\mathbf{c}) - \mathbb{P}_{\mu_2}(\mathbf{c}) = [\mathbf{c} - \mathbb{P}_{\mu_2}(\mathbf{c})] - [\mathbf{c} - \mathbb{P}_{\mu_1}(\mathbf{c})]$; step ④ uses the optimality of $\mathbb{P}_{\mu_1}(\mathbf{c})$ and $\mathbb{P}_{\mu_2}(\mathbf{c})$ in Equation (26); step ⑤ uses the inequality that $-\frac{\mu}{2}\|\mathbf{p}_2\|_2^2 + \mu\langle \mathbf{p}_1, \mathbf{p}_2 \rangle \leq \frac{\mu}{2}\|\mathbf{p}_1\|_2^2$ for all $\mu > 0$ and for all $\mathbf{p}_1, \mathbf{p}_2 \in \mathbb{R}^m$; step ⑥ uses $\|\mathbf{p}_1\| \leq L_h$. Dividing both sides by $(\mu_2 - \mu_1)$, we conclude that $\psi(\mu)$ is $(\frac{1}{2}L_h^2)$-Lipschitz.

□

B.2. Proof of Lemma 3.2

Proof. We fix $\mu > 0$ to be a constant. For any given $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^n$ with $\|\mathbf{x}\|_0 \leq s$ and $\|\mathbf{x}'\|_0 \leq s$, we define

$$\mathbf{c} \triangleq \mathbf{A}\mathbf{x} - \mathbf{b}, \text{ and } \mathbf{c}' \triangleq \mathbf{A}\mathbf{x}' - \mathbf{b}. \tag{27}$$

Using the definition of $\mathbb{P}_\mu(\cdot)$ as shown in Equation (2), we have:

$$\mathbb{P}_\mu(\mathbf{c}) = \arg\min_{\mathbf{y}} h(\mathbf{y}) + \tfrac{1}{2\mu}\|\mathbf{y} - \mathbf{c}\|_2^2, \text{ and } \mathbb{P}_\mu(\mathbf{c}') = \arg\min_{\mathbf{y}} h(\mathbf{y}) + \tfrac{1}{2\mu}\|\mathbf{y} - \mathbf{c}'\|_2^2.$$

By the optimality condition of $\mathbb{P}_\mu(\mathbf{c})$ and $\mathbb{P}_\mu(\mathbf{c}')$, we have:

$$\mathbf{c} - \mathbb{P}_\mu(\mathbf{c}) \in \mu \partial h(\mathbb{P}_\mu(\mathbf{c})), \text{ and } \mathbf{c}' - \mathbb{P}_\mu(\mathbf{c}') \in \mu \partial h(\mathbb{P}_\mu(\mathbf{c})). \tag{28}$$

The function $\mathcal{G}(\mathbf{x}; \mu)$ defined in Equation (11) is differentiable and its gradient at $\mathbf{x}$ and $\mathbf{x}'$ can be respectively computed as:

$$\nabla_{\mathbf{x}} \mathcal{G}(\mathbf{x}; \mu) = \nabla f(\mathbf{x}) + \tfrac{1}{\mu}\mathbf{A}^\mathsf{T}(\mathbf{c} - \mathbb{P}_\mu(\mathbf{c})) \;\; \text{and} \;\; \nabla_{\mathbf{x}} \mathcal{G}(\mathbf{x}'; \mu) = \nabla f(\mathbf{x}') + \tfrac{1}{\mu}\mathbf{A}^\mathsf{T}(\mathbf{c}' - \mathbb{P}_\mu(\mathbf{c}')). \tag{29}$$

(a) We notice that $F(\mathbf{x}) = \lim_{\bar{\mu}\to 0} \mathcal{G}(\mathbf{x};\bar{\mu})$ and $\mathcal{G}(\mathbf{x};\mu)$ is a decreasing function *w.r.t.* μ. The inequality $F(\mathbf{x}) \geq \mathcal{G}(\mathbf{x};\mu)$ clearly holds. We now prove that $F(\mathbf{x}) - \frac{\mu}{2}L_h^2 \leq \mathcal{G}(\mathbf{x};\mu)$. For any $\mathbf{x}$ with $\|\mathbf{x}\|_0 \leq s$ and $\mathbf{p} \in \partial h(\mathbb{P}_\mu(\mathbf{c}))$, we obtain:

$$\begin{aligned}
F(\mathbf{x}) - \mathcal{G}(\mathbf{x};\mu) &\overset{\text{①}}{=} [f(\mathbf{x}) + h(\mathbf{A}\mathbf{x}-\mathbf{b})] - [f(\mathbf{x}) + h(\mathbb{P}_\mu(\mathbf{A}\mathbf{x}-\mathbf{b})) + \tfrac{1}{2\mu}\|\mathbf{A}\mathbf{x}-\mathbf{b}-\mathbb{P}_\mu(\mathbf{A}\mathbf{x}-\mathbf{b})\|_2^2] \\
&\overset{\text{②}}{=} h(\mathbf{c}) - h(\mathbb{P}_\mu(\mathbf{c})) - \tfrac{1}{2\mu}\|\mathbf{c}-\mathbb{P}_\mu(\mathbf{c})\|_2^2 \\
&\overset{\text{③}}{\leq} \langle \mathbf{c}-\mathbb{P}_\mu(\mathbf{c}), \partial h(\mathbb{P}_\mu(\mathbf{c}))\rangle - \tfrac{1}{2\mu}\|\mu\partial h(\mathbb{P}_\mu(\mathbf{c}))\|_2^2 \\
&\leq \langle \mathbf{c}-\mathbb{P}_\mu(\mathbf{c}), \mathbf{p}\rangle - \tfrac{\mu}{2}\|\mathbf{p}\|_2^2 \\
&\overset{\text{④}}{\leq} \tfrac{1}{2\mu}\|\mathbf{c}-\mathbb{P}_\mu(\mathbf{c})\|_2^2 \\
&\overset{\text{⑤}}{=} \tfrac{1}{2\mu}\|\mu\mathbf{p}\|_2^2 \\
&\overset{\text{⑥}}{\leq} \tfrac{\mu}{2}L_h^2,
\end{aligned}$$

where step ① uses the definition of $F(\mathbf{x}) \triangleq f(\mathbf{x}) + h(\mathbf{A}\mathbf{x}-\mathbf{b})$ in Problem (1) and the definition of $\mathcal{G}(\mathbf{x};\mu)$ in Equation (11); step ② uses $\mathbf{A}\mathbf{x}-\mathbf{b} = \mathbf{c}$; step ③ uses the convexity of $h(\cdot)$ and the optimality of $\mathbb{P}_\mu(\mathbf{c})$ as shown in Equation (28); step ④ uses the inequality $-\frac{\mu}{2}\|\mathbf{p}\|_2^2 + \langle \mathbf{v},\mathbf{p}\rangle \leq \frac{1}{2\mu}\|\mathbf{v}\|_2^2$ for all $\mathbf{v}$ and $\mu > 0$; step ⑤ uses Equation (28); step ⑥ uses $\|\mathbf{p}\|_2 \leq L_h$.

(b) We now prove that $F(\mathbf{x})$ is $(L_f + L_h\|\mathbf{A}\|)$-Lipschitz. We have:

$$\begin{aligned}
\|\partial F(\mathbf{x})\| &\overset{\text{①}}{=} \|\nabla f(\mathbf{x}) + \mathbf{A}^\mathsf{T}\partial h(\mathbf{A}\mathbf{x}-\mathbf{b})\| \\
&\overset{\text{②}}{\leq} \|\nabla f(\mathbf{x})\| + \|\mathbf{A}\|\|\partial h(\mathbf{A}\mathbf{x}-\mathbf{b})\| \\
&\overset{\text{③}}{\leq} L_f + L_h\|\mathbf{A}\|,
\end{aligned}$$

where step ① uses $\partial F(\mathbf{x}) = \nabla f(\mathbf{x}) + \mathbf{A}^\mathsf{T}\partial h(\mathbf{A}\mathbf{x}-\mathbf{b})$; step ② uses the norm inequality; step ③ uses the fact that $h(\cdot)$ is L_h-Lipschitz and $f(\cdot)$ is L_f-Lipschitz.

We now prove that $\mathcal{G}(\mathbf{x},\mu)$ is $(L_f + L_h\|\mathbf{A}\|)$-Lipschitz. We obtain:

$$\begin{aligned}
\|\nabla_\mathbf{x}\mathcal{G}(\mathbf{x};\mu)\| &= \|\nabla f(\mathbf{x}) + \tfrac{1}{\mu}\mathbf{A}^\mathsf{T}(\mathbf{c}-\mathbb{P}_\mu(\mathbf{c}))\| \\
&\leq \|\nabla f(\mathbf{x})\| + \|\mathbf{A}\| \cdot \|\tfrac{1}{\mu}(\mathbf{c}-\mathbb{P}_\mu(\mathbf{c}))\| \\
&\overset{\text{①}}{=} \|\nabla f(\mathbf{x})\| + \|\mathbf{A}\| \cdot \|\partial h(\mathbb{P}_\mu(\mathbf{c}))\| \\
&\leq L_f + L_h\|\mathbf{A}\|,
\end{aligned}$$

where step ① uses the optimality condition of $\mathbb{P}_\mu(\mathbf{c})$ as shown in Equation (28) that $\mathbf{c} - \mathbb{P}_\mu(\mathbf{c}) \in \mu\partial h(\mathbb{P}_\mu(\mathbf{c}))$.

(c) Noticing $f(\mathbf{x})$ is restricted V_s-strongly convex, we directly conclude that $\mathcal{G}(\mathbf{x},\mu)$ is also restricted V_s-strongly convex. We now prove that the function $\mathcal{G}(\mathbf{x},\mu)$ is restricted $(M_s + \frac{A_s\|\mathbf{A}\|}{\mu})$-smooth. For any $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}' \in \mathbb{R}^n$, $\mathbf{p}_1 \in \partial h(\mathbb{P}_\mu(\mathbf{c}))$, and $\mathbf{p}_2 \in \partial h(\mathbb{P}_\mu(\mathbf{c}'))$, we derive:

$$\begin{aligned}
&\|[\mathbf{A}\mathbf{x}' - \mathbf{A}\mathbf{x}] + [\mathbb{P}_\mu(\mathbf{c}) - \mathbb{P}_\mu(\mathbf{c}')]\|_2^2 \\
\overset{\text{①}}{=}\ & \|\mathbf{A}\mathbf{x}' - \mathbf{A}\mathbf{x}\|_2^2 + \|\mathbb{P}_\mu(\mathbf{c}) - \mathbb{P}_\mu(\mathbf{c}')\|_2^2 + 2\langle \mathbb{P}_\mu(\mathbf{c}) - \mathbb{P}_\mu(\mathbf{c}'), (\mathbf{A}\mathbf{x}'-\mathbf{b}) - (\mathbf{A}\mathbf{x}-\mathbf{b})\rangle \\
\overset{\text{②}}{=}\ & A_s^2\|\mathbf{x}'-\mathbf{x}\|_2^2 + \|\mathbb{P}_\mu(\mathbf{c}) - \mathbb{P}_\mu(\mathbf{c}')\|_2^2 + 2\langle \mathbb{P}_\mu(\mathbf{c}) - \mathbb{P}_\mu(\mathbf{c}'), [\mu\mathbf{p}_1 + \mathbb{P}_\mu(\mathbf{c}')] - [\mu\mathbf{p}_1 + \mathbb{P}_\mu(\mathbf{c})]\rangle \\
=\ & A_s^2\|\mathbf{x}'-\mathbf{x}\|_2^2 - \|\mathbb{P}_\mu(\mathbf{c}) - \mathbb{P}_\mu(\mathbf{c}')\|_2^2 + 2\langle \mathbb{P}_\mu(\mathbf{c}) - \mathbb{P}_\mu(\mathbf{c}'), \mu\mathbf{p}_2 - \mu\mathbf{p}_1\rangle \\
\overset{\text{③}}{\leq}\ & A_s^2\|\mathbf{x}'-\mathbf{x}\|_2^2 + 0 + 0,
\end{aligned} \tag{30}$$

where step ① uses the Pythagoras relation; step ② uses Assumption 2.3 and the optimality conditions in (28); step ③ uses the convexity of $h(\cdot)$ that $\langle \mathbf{y}' - \mathbf{y}, \partial h(\mathbf{y}') - \partial h(\mathbf{y})\rangle \geq 0$ for all $\mathbf{y}$ and $\mathbf{y}'$.

Finally, we have the following inequalities:

$$\begin{aligned}
& \|\nabla_{\mathbf{x}}\mathcal{G}(\mathbf{x}';\mu) - \nabla_{\mathbf{x}}\mathcal{G}(\mathbf{x};\mu)\| \\
\overset{\text{①}}{=} \;& \|[\nabla f(\mathbf{x}') + \tfrac{1}{\mu}\mathbf{A}^{\mathsf{T}}(\mathbf{c}' - \mathbb{P}_{\mu}(\mathbf{c}'))] - [\nabla f(\mathbf{x}) + \tfrac{1}{\mu}\mathbf{A}^{\mathsf{T}}(\mathbf{c} - \mathbb{P}_{\mu}(\mathbf{c}))]\| \\
\overset{\text{②}}{\leq} \;& \|\nabla f(\mathbf{x}') - \nabla f(\mathbf{x})\| + \|\tfrac{1}{\mu}\mathbf{A}^{\mathsf{T}}(\mathbf{A}\mathbf{x}' - \mathbf{b} - \mathbb{P}_{\mu}(\mathbf{c}')) - \tfrac{1}{\mu}\mathbf{A}^{\mathsf{T}}(\mathbf{A}\mathbf{x} - \mathbf{b} - \mathbb{P}_{\mu}(\mathbf{c}))]\| \\
\overset{\text{③}}{\leq} \;& M_s\|\mathbf{x} - \mathbf{x}'\| + \tfrac{1}{\mu}\cdot\|\mathbf{A}\|\cdot\|[\mathbf{A}\mathbf{x}' - \mathbf{A}\mathbf{x}] + [\mathbb{P}_{\mu}(\mathbf{c}) - \mathbb{P}_{\mu}(\mathbf{c}')]\| \\
\overset{\text{④}}{\leq} \;& M_s\|\mathbf{x} - \mathbf{x}'\| + \tfrac{1}{\mu}\cdot\|\mathbf{A}\|\cdot A_s\cdot\|\mathbf{x} - \mathbf{x}'\|,
\end{aligned}$$

where step ① uses the definition of $\nabla_{\mathbf{x}}\mathcal{G}(\mathbf{x};\mu)$ in Equation (29); step ② uses the norm inequality; step ③ uses the fact that $f(\mathbf{x})$ is restricted M_s-smooth as shown in Assumption 2.2 and norm inequality; step ④ uses Inequality (30).

□

B.3. Proof of Lemma 4.1

Proof. **(a)** We now bound $\|\mathbf{y}^{t+1} + \mathbf{b} - \mathbf{A}\mathbf{x}^{t+1}\|$ using these inequalities:

$$\|\mathbf{y}^{t+1} + \mathbf{b} - \mathbf{A}\mathbf{x}^{t+1}\| \overset{\text{①}}{=} \mu^t\|\partial h(\mathbf{y}^{t+1})\| \overset{\text{②}}{\leq} L_h\mu^t, \tag{31}$$

where step ① uses the optimality condition of $\mathbf{y}^{t+1}$ with $\mathbf{y}^{t+1} = \arg\min_{\mathbf{y}}\, h(\mathbf{y}) + \frac{1}{2\mu^t}\|\mathbf{A}\mathbf{x}^{t+1} - \mathbf{b} - \mathbf{y}\|_2^2$, which yields:

$$\mathbf{A}\mathbf{x}^{t+1} - \mathbf{b} - \mathbf{y}^{t+1} \in \mu^t\partial h(\mathbf{y}^{t+1}); \tag{32}$$

step ② uses Assumption 2.1.

(b) We now bound $\|\mathbf{y}^{t+1} - \mathbf{y}^t\|$ using these inequalities:

$$\begin{aligned}
\|\mathbf{y}^{t+1} - \mathbf{y}^t\| \;&\overset{\text{①}}{=}\; \|\mathbf{A}(\mathbf{x}^{t+1} - \mathbf{x}^t) + \mu^{t-1}\partial h(\mathbf{y}^t) - \mu^t\partial h(\mathbf{y}^{t+1})\| \\
&\overset{\text{②}}{\leq}\; \|\mathbf{A}\|\|\mathbf{x}^{t+1} - \mathbf{x}^t\| + \|\mu^t\partial h(\mathbf{y}^{t+1})\| + \|\mu^{t-1}\partial h(\mathbf{y}^t)\| \\
&\overset{\text{③}}{\leq}\; \|\mathbf{A}\|\|\mathbf{x}^{t+1} - \mathbf{x}^t\| + 2\mu^{t-1}L_h,
\end{aligned}$$

where step ① uses (32); step ② uses the triangle inequality and norm inequality; step ③ uses $\|\partial h(\mathbf{y})\| \leq L_h$ and $\mu^t \leq \mu^{t-1}$.

(c) We now bound $\|\nabla_{\mathbf{x}}\mathcal{R}(\mathbf{x}^t, \mathbf{y}^t; \mu^t)\|$ using these inequalities:

$$\begin{aligned}
\|\nabla_{\mathbf{x}}\mathcal{R}(\mathbf{x}^t, \mathbf{y}^t; \mu^t)\| \;&\overset{\text{①}}{=}\; \|\nabla f(\mathbf{x}^t) + \tfrac{1}{\mu^t}\mathbf{A}^{\mathsf{T}}(\mathbf{A}\mathbf{x}^t - \mathbf{y}^t - \mathbf{b})\| \\
&\leq\; \|\nabla f(\mathbf{x}^t)\| + \tfrac{1}{\mu^t}\|\mathbf{A}\|\|\mathbf{A}\mathbf{x}^t - \mathbf{y}^t - \mathbf{b}\| \\
&\overset{\text{②}}{\leq}\; L_f + \tfrac{\mu^{t-1}}{\mu^t}L_h\|\mathbf{A}\|,
\end{aligned}$$

where step ① uses the definition of $\nabla_{\mathbf{x}}\mathcal{R}(\mathbf{x}^t, \mathbf{y}^t; \mu^t)$ as in Equation (4); step ② uses **Part (a)** of this lemma.

If we choose $\mu^t = \bar{\mu}$, we have: $\frac{\mu^{t-1}}{\mu^t} = 1$, and $\|\nabla_{\mathbf{x}}\mathcal{R}(\mathbf{x}^t, \mathbf{y}^t; \mu^t)\| \leq L_f + L_h\|\mathbf{A}\|$.
If we choose $\mu^t = \frac{\eta}{t+t_0}$, we have $\frac{\mu^{t-1}}{\mu^t} = \frac{t+t_0}{t-1+t_0} \leq \max_{i=1}^{\infty}\frac{i+t_0}{i-1+t_0} \leq \frac{t_0+1}{t_0}$, and $\|\nabla_{\mathbf{x}}\mathcal{R}(\mathbf{x}^t, \mathbf{y}^t; \mu^t)\| \leq L_f + \frac{t_0+1}{t_0}L_h\|\mathbf{A}\|$.

(d) We now bound the term $\|\varepsilon^t\|$ using these inequalities:

$$\begin{aligned}
\|\varepsilon^t\| \;&=\; \|(\tfrac{1}{\mu^{t-1}} - \tfrac{1}{\mu^t})\cdot\mathbf{A}^{\mathsf{T}}(\mathbf{A}\mathbf{x}^t - \mathbf{b} - \mathbf{y}^t)\| \\
&\overset{\text{①}}{\leq}\; (\tfrac{1}{\mu^t} - \tfrac{1}{\mu^{t-1}})\|\mathbf{A}\|\|\mathbf{A}\mathbf{x}^t - \mathbf{b} - \mathbf{y}^t\| \\
&\overset{\text{②}}{=}\; (\tfrac{1}{\mu^t} - \tfrac{1}{\mu^{t-1}})\|\mathbf{A}\|\cdot L_h\mu^{t-1},
\end{aligned}$$

where step ① uses the norm inequality; step ② uses Inequality (31).

(e) We now bound $\mathcal{J}(\mathbf{x}^t, \mathbf{y}^t; \mu^t) - \mathcal{J}(\mathbf{x}^t, \mathbf{y}^t; \mu^{t-1})$ using these inequalities:

$$\begin{aligned}
\mathcal{J}(\mathbf{x}^t, \mathbf{y}^t; \mu^t) - \mathcal{J}(\mathbf{x}^t, \mathbf{y}^t; \mu^{t-1}) &\overset{①}{=} \tfrac{1}{2}\|\mathbf{A}\mathbf{x}^t - \mathbf{b} - \mathbf{y}^t\|_2^2 \cdot (\tfrac{1}{\mu^t} - \tfrac{1}{\mu^{t-1}}) \\
&\overset{②}{\leq} \tfrac{1}{2}(\mu^{t-1}L_h)^2(\tfrac{1}{\mu^t} - \tfrac{1}{\mu^{t-1}}) \\
&= \tfrac{1}{2}L_h^2(\tfrac{(\mu^{t-1})^2}{\mu^t} - \mu^{t-1}) \triangleq \Psi^t,
\end{aligned}$$

where step ① uses the definition of $\mathcal{J}(\mathbf{x}^t, \mathbf{y}^t; \mu^t) \triangleq f(\mathbf{x}^t) + \frac{1}{2\mu^t}\|\mathbf{A}\mathbf{x}^t - \mathbf{b} - \mathbf{y}^t\|_2^2$; step ② uses **Part (a)** of Lemma 4.1.

(f) We now prove that $[\sum_{t=1}^{\infty} \Psi^t] \leq \left\{ \begin{smallmatrix} 0, & \mu^t = \bar{\mu}; \\ \eta L_h^2, & \mu^t = \eta(t+t_0)^{-1} \end{smallmatrix} \right.$, where $\Psi^t \triangleq \frac{1}{2}L_h^2(\frac{(\mu^{t-1})^2}{\mu^t} - \mu^{t-1})$. We discuss two cases for μ^t.

Case 1). When $\mu^t = \bar{\mu}$, we have:

$$[\sum_{t=1}^{\infty} \Psi^t] \triangleq \tfrac{1}{2}L_h^2 \sum_{t=0}^{\infty}(\tfrac{(\mu^{t-1})^2}{\mu^t} - \mu^{t-1}) = \tfrac{1}{2}L_h^2 \sum_{t=0}^{\infty}(\bar{\mu} - \bar{\mu}) = 0.$$

Case 2). When $\mu^t = \frac{\eta}{t+t_0}$, we have:

$$\begin{aligned}
[\sum_{t=1}^{\infty} \Psi^t] &\overset{①}{=} (\tfrac{1}{2}L_h^2) \cdot \sum_{t=1}^{\infty}(\tfrac{(\mu^{t-1})^2}{\mu^t} - \mu^{t-1}) \overset{②}{=} (\tfrac{1}{2}L_h^2) \cdot (\sum_{t=1}^{\infty} \frac{\eta}{(t+t_0-1)^2}) \\
&\overset{③}{\leq} (\tfrac{1}{2}L_h^2) \cdot (\sum_{t=1}^{\infty} \frac{\eta}{t^2}) \overset{④}{<} (\tfrac{1}{2}L_h^2) \cdot 2\eta,
\end{aligned}$$

where step ① uses the definition of $\Psi^t \triangleq \frac{L_h^2}{2}(\frac{(\mu^{t-1})^2}{\mu^t} - \mu^{t-1})$; step ② uses $\mu^t = \frac{\eta}{t+t_0}$; step ③ uses $t_0 \geq 1$; step ④ uses $\sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6} < 2$.

□

B.4. Proof of Lemma 4.2

Proof. **(a)** We first now bound the term $\|\mathbf{x}^t - \bar{\mathbf{x}}\|$ using these inequalities:

$$\tfrac{V_s}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2 \overset{①}{\leq} F(\bar{\mathbf{x}}) - F(\mathbf{x}^t) - \langle \bar{\mathbf{x}} - \mathbf{x}^t, \partial F(\mathbf{x}^t)\rangle \overset{②}{\leq} 0 + L_F\|\bar{\mathbf{x}} - \mathbf{x}^t\|,$$

where step ① uses the restricted strong convexity of $F(\cdot)$; step ② uses $F(\bar{\mathbf{x}}) \leq F(\mathbf{x}^t)$ and $\|\partial F(\mathbf{x})\| \leq L_F$. Dividing both sides by $(\frac{V_s}{2}\|\bar{\mathbf{x}} - \mathbf{x}^t\|)$, we have:

$$\|\mathbf{x}^t - \bar{\mathbf{x}}\| \leq \tfrac{2L_F}{V_s}. \tag{33}$$

We now now bound the term $\frac{V_s}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2$ using these inequalities:

$$\begin{aligned}
&\tfrac{V_s}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2 \\
\overset{①}{\leq}\ & \langle \nabla_{\mathbf{x}}\mathcal{G}(\mathbf{x}^t; \mu^{t-1}),\, \mathbf{x}^t - \bar{\mathbf{x}}\rangle + \mathcal{G}(\bar{\mathbf{x}}; \mu^{t-1}) - \mathcal{G}(\mathbf{x}^t; \mu^{t-1}) \\
\overset{②}{\leq}\ & \langle \mathbf{r}^t + \boldsymbol{\varepsilon}^t,\, \mathbf{x}^t - \bar{\mathbf{x}}\rangle + [F(\bar{\mathbf{x}}) - F(\mathbf{x}^t) + \tfrac{\mu^{t-1}}{2}L_h^2] \\
\overset{③}{\leq}\ & \langle \mathbf{r}^t, \mathbf{x}^t - \bar{\mathbf{x}}\rangle + \|\boldsymbol{\varepsilon}^t\| \cdot \|\bar{\mathbf{x}} - \mathbf{x}^t\| + [F(\bar{\mathbf{x}}) - F(\mathbf{x}^t) + \tfrac{\mu^{t-1}}{2}L_h^2] \\
\overset{④}{\leq}\ & \langle \mathbf{r}^t, \mathbf{x}^t - \bar{\mathbf{x}}\rangle + \underbrace{(\tfrac{1}{\mu^t} - \tfrac{1}{\mu^{t-1}})L_h\mu^{t-1}\|\mathbf{A}\| \cdot \tfrac{2L_F}{V_s} + [F(\bar{\mathbf{x}}) - F(\mathbf{x}^t) + \tfrac{\mu^{t-1}}{2}L_h^2]}_{\triangleq \Upsilon^t}
\end{aligned}$$

where step ① uses the restricted strong convexity of $\mathcal{G}(\mathbf{x};\mu^{t-1})$; step ② uses the the relation between $\nabla_{\mathbf{x}}\mathcal{G}(\mathbf{x}^t;\mu^{t-1})$ and $\nabla_{\mathbf{x}}\mathcal{R}(\mathbf{x}^t,\mathbf{y}^t;\mu^t)$ and **Part (a)** in Lemma (3.2) that

$$\mathcal{G}(\bar{\mathbf{x}};\mu^{t-1}) \leq F(\bar{\mathbf{x}}), \quad \mathcal{G}(\mathbf{x}^t;\mu^{t-1}) \geq F(\mathbf{x}^t) - \frac{\mu^{t-1}}{2}L_h^2;$$

step ③ uses the norm inequality; step ④ uses Inequality (33) and the inequality in **Part (d)** of Lemma 4.1 that $\|\varepsilon^t\| \leq (\frac{1}{\mu^t} - \frac{1}{\mu^{t-1}})L_h\mu^{t-1}\|\mathbf{A}\|$.

(b) When $\mu^t = \bar{\mu}$, we have the following results:

$$\begin{aligned}
\Upsilon^t &\triangleq F(\bar{\mathbf{x}}) - F(\mathbf{x}^t) + \tfrac{1}{2}\mu^{t-1}L_h^2 + \tfrac{2L_FL_h}{V_s}\|\mathbf{A}\|(\tfrac{\mu^{t-1}}{\mu^t} - 1) \\
&\overset{\textcircled{1}}{=} F(\bar{\mathbf{x}}) - F(\mathbf{x}^t) + \tfrac{1}{2}\bar{\mu}L_h^2. \\
&\overset{\textcircled{2}}{\leq} F(\bar{\mathbf{x}}) - [\min_{i=1}^{t} F(\mathbf{x}^i)] + \tfrac{1}{2}\bar{\mu}L_h^2,
\end{aligned}$$

where step ① uses $\frac{\mu^{t+1}}{\mu^t} = 1$; step ② uses $[\min_{i=1}^t F(\mathbf{x}^i)] \leq F(\mathbf{x}^t)$.

(c) When $\mu^t = \frac{\eta}{t+t_0}$, we have the following results:

$$\begin{aligned}
\sum_{t=1}^{T}\Upsilon^t &\overset{\textcircled{1}}{=} \sum_{t=1}^{T}\Big(F(\bar{\mathbf{x}}) - F(\mathbf{x}^t) + \tfrac{1}{2}\mu^{t-1}L_h^2 + \tfrac{2L_FL_h}{V_s}\|\mathbf{A}\|(\tfrac{\mu^{t-1}}{\mu^t} - 1)\Big) \\
&\overset{\textcircled{2}}{\leq} TF(\bar{\mathbf{x}}) - T[\min_{t=1}^{T} F(\mathbf{x}^t)] + \sum_{t=1}^{T}\left(\frac{L_h^2}{2}\frac{\eta}{t+t_0-1} + \tfrac{2L_FL_h}{V_s}\|\mathbf{A}\|\frac{1}{t+t_0-1}\right) \\
&\overset{\textcircled{3}}{\leq} TF(\bar{\mathbf{x}}) - T[\min_{t=1}^{T} F(\mathbf{x}^t)] + [1+\ln(T)]\cdot\left(\frac{\eta L_h^2}{2} + \tfrac{2L_FL_h}{V_s}\|\mathbf{A}\|\right),
\end{aligned}$$

where step ① uses the definition of Υ^t as shown in Lemma 4.2; step ② uses $\max_{t=1}^T[-F(\mathbf{x}^t)] = -[\min_{t=1}^T F(\mathbf{x}^t)]$, the inequality that $\frac{\mu^{t-1}}{\mu^t} - 1 = \frac{t+t_0}{t+t_0-1} - 1 = \frac{1}{t+t_0-1}$; step ③ uses $t_0 \geq 1$, and the fact that

$$\sum_{t=1}^{T}\frac{1}{t+t_0-1} \leq \sum_{t=1}^{T}\frac{1}{t} \leq 1 + \ln(T).$$

Using the definition of $C_\Upsilon \triangleq \frac{\eta L_h^2}{2} + \frac{2L_FL_h}{V_s}\|\mathbf{A}\|$, we finish the proof of this lemma.

□

C. Proofs for Section 4

C.1. Proof of Theorem 4.4

Proof. We denote $\mathbf{r}^t \triangleq \nabla_{\mathbf{x}}\mathcal{R}(\mathbf{x}^t,\mathbf{y}^t;\mu^t)$ and $H^t = A_s^2/\mu^t + M_s + \theta$.

(a) We focus on the $\mathbf{x}$-subproblem. We have from Problem (5) that

$$\langle \mathbf{r}^t,\, \mathbf{x}^{t+1} - \mathbf{x}^t\rangle + \tfrac{H^t}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2 \leq \langle \mathbf{r}^t,\, \mathbf{x}^t - \mathbf{x}^t\rangle + \tfrac{H^t}{2}\|\mathbf{x}^t - \mathbf{x}^t\|_2^2 = 0.$$

Since $\mathcal{R}(\mathbf{x}^t,\mathbf{y}^t;\mu^t)$ is restricted $(A_s^2/\mu^t + M_s)$-smooth *w.r.t.* $\mathbf{x}$, we have:

$$\mathcal{R}(\mathbf{x}^{t+1},\mathbf{y}^t;\mu^t) \leq \mathcal{R}(\mathbf{x}^t,\mathbf{y}^t;\mu^t) + \langle \mathbf{r}^t, \mathbf{x}^{t+1} - \mathbf{x}^t\rangle + \tfrac{A_s^2/\mu^t + M_s}{2}\|\mathbf{x}^t - \mathbf{x}^{t+1}\|_2^2.$$

We observe that the following equality holds:

$$\mathcal{R}(\mathbf{x}^{t+1},\mathbf{y}^t;\mu^t) - \mathcal{R}(\mathbf{x}^t,\mathbf{y}^t;\mu^t) = \mathcal{J}(\mathbf{x}^{t+1},\mathbf{y}^t;\mu^t) - \mathcal{J}(\mathbf{x}^t,\mathbf{y}^t;\mu^t)$$

Summing these three inequalities, we have:

$$\mathcal{J}(\mathbf{x}^{t+1},\mathbf{y}^t;\mu^t) - \mathcal{J}(\mathbf{x}^t,\mathbf{y}^t;\mu^t) \leq -\tfrac{\theta}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_2^2. \tag{34}$$

We now focus on the $\mathbf{y}$-subproblem. We derive the following inequalities for all $\mathbf{y} \in \mathbb{R}^m$:

$$\begin{aligned} & \mathcal{J}(\mathbf{x}^{t+1},\mathbf{y}^{t+1};\mu^t) - \mathcal{J}(\mathbf{x}^{t+1},\mathbf{y};\mu^t) \\ \overset{\textcircled{1}}{\leq} \; & -\frac{1}{2\mu^t}\|\mathbf{y}^{t+1}-\mathbf{y}\|_2^2 - \langle \mathbf{y}-\mathbf{y}^{t+1}, \partial_{\mathbf{y}}\mathcal{J}(\mathbf{x}^{t+1},\mathbf{y}^{t+1};\mu^t)\rangle \\ \overset{\textcircled{2}}{=} \; & -\frac{1}{2\mu^t}\|\mathbf{y}^{t+1}-\mathbf{y}\|_2^2 \\ \overset{\textcircled{3}}{\leq} \; & -\frac{1}{2\mu^1}\|\mathbf{y}^{t+1}-\mathbf{y}\|_2^2, \end{aligned}$$

where step ① uses the fact that $\mathcal{J}(\mathbf{x}^{t+1},\mathbf{y};\mu^t)$ is $\frac{1}{\mu}$-strongly convex *w.r.t.* $\mathbf{y}$; step ② uses the optimality of $\mathbf{y}^{t+1}$ that $\mathbf{0} \in \partial_{\mathbf{y}}\mathcal{J}(\mathbf{x}^{t+1},\mathbf{y}^{t+1};\mu^t)$; step ③ uses the fact that the sequence $\{\mu^t\}_{t=1}^{\infty}$ is non-increasing. Letting $\mathbf{y}=\mathbf{y}^t$, we obtain:

$$\mathcal{J}(\mathbf{x}^{t+1},\mathbf{y}^{t+1};\mu^t) - \mathcal{J}(\mathbf{x}^{t+1},\mathbf{y}^t;\mu^t) \leq -\frac{1}{2\mu^1}\|\mathbf{y}^{t+1}-\mathbf{y}^t\|_2^2 \tag{35}$$

Using the continuity of $\mathcal{J}(\mathbf{x}^t,\mathbf{y}^t;\mu)$ *w.r.t.* μ as shown in **Part (e)** of Lemma 4.1, we obtain:

$$0 \leq \mathcal{J}(\mathbf{x}^t,\mathbf{y}^t;\mu^t) - \mathcal{J}(\mathbf{x}^t,\mathbf{y}^t;\mu^{t-1}) \leq \tfrac{L_h^2}{2}(\tfrac{(\mu^{t-1})^2}{\mu^t} - \mu^{t-1}) \triangleq \Psi^t. \tag{36}$$

Summing Inequalities (34), (35), and (36) together, we have:

$$\begin{aligned} & \tfrac{1}{2\mu^1}\|\mathbf{y}^{t+1}-\mathbf{y}^t\|_2^2 + \tfrac{\theta}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_2^2 \\ \leq \; & \mathcal{J}(\mathbf{x}^t,\mathbf{y}^t;\mu^{t-1}) - \mathcal{J}(\mathbf{x}^{t+1},\mathbf{y}^{t+1};\mu^t) + \tfrac{1}{2}L_h^2(\tfrac{(\mu^{t-1})^2}{\mu^t} - \mu^{t-1}) \\ \overset{\textcircled{1}}{=} \; & \mathcal{J}^t - \mathcal{J}^{t+1} + \Psi^t, \end{aligned} \tag{37}$$

where step ① uses the definitions of Ψ^t and $\mathcal{J}^{t+1} \triangleq \mathcal{J}(\mathbf{x}^{t+1},\mathbf{y}^{t+1};\mu^t)$.

(b) Summing Inequality (37) over t from 1 to T, we obtain:

$$\begin{aligned} & \textstyle\sum_{t=1}^T \tfrac{1}{2\mu^1}\|\mathbf{y}^{t+1}-\mathbf{y}^t\|_2^2 + \sum_{t=1}^T \tfrac{\theta}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_2^2 \\ \leq \; & \mathcal{J}^1 - \mathcal{J}^{T+1} + [\textstyle\sum_{t=1}^T \Psi^t] \overset{\textcircled{1}}{\leq} \mathcal{J}^1 - \mathcal{J}^{T+1} + \eta L_h^2 \triangleq C < +\infty, \end{aligned}$$

where step ① uses $\sum_{t=1}^T \Psi^t < \sum_{t=1}^{\infty} \Psi^t \leq \eta L_h^2$ which is shown in **Part (f)** of Lemma 4.1.

(c) As a result, there exists an index $\bar{t}$ with $1 \leq \bar{t} \leq T$ such that $\frac{1}{2\mu^1}\|\mathbf{y}^{\bar{t}+1}-\mathbf{y}^{\bar{t}}\|_2^2 + \frac{\theta}{2}\|\mathbf{x}^{\bar{t}+1}-\mathbf{x}^{\bar{t}}\|_2^2 \leq \frac{C}{T}$, leading to:

$$\|\mathbf{y}^{\bar{t}+1}-\mathbf{y}^{\bar{t}}\|_2^2 + \|\mathbf{x}^{\bar{t}+1}-\mathbf{x}^{\bar{t}}\|_2^2 \leq \frac{2C}{T\cdot\min(\theta,(\mu^1)^{-1})}. \tag{38}$$

Letting $\Gamma_x(\mathbf{x},\mathbf{y};\mu) \triangleq \operatorname{dist}^2(\mathbf{x}, \arg\min_{\mathbf{x}'} \mathcal{M}(\mathbf{x}',\mathbf{x},\mathbf{y};\mu))$ and $\Gamma_y(\mathbf{x},\mathbf{y};\mu) \triangleq \operatorname{dist}^2(\mathbf{y}, \arg\min_{\mathbf{y}'} \mathcal{J}(\mathbf{x},\mathbf{y}';\mu))$, we have:

$$\|\mathbf{x}^{\bar{t}+1}-\mathbf{x}^{\bar{t}}\|_2^2 + \|\mathbf{y}^{\bar{t}+1}-\mathbf{y}^{\bar{t}}\|_2^2 \geq \Gamma_x(\mathbf{x}^{\bar{t}},\mathbf{y}^{\bar{t}};\mu) + \Gamma_y(\mathbf{x}^{\bar{t}},\mathbf{y}^{\bar{t}};\mu) \tag{39}$$

for all $\bar{t} \geq 1$ and some sufficiently small $\mu = \mu^{\bar{t}} > 0$. Combining Inequalities (38) and (39), we have:

$$\Gamma_x(\mathbf{x}^{\bar{t}},\mathbf{y}^{\bar{t}};\mu^{\bar{t}}) + \Gamma_y(\mathbf{x}^{\bar{t}},\mathbf{y}^{\bar{t}};\mu^{\bar{t}}) \leq \frac{2C}{T\cdot\min(\theta,(\mu^1)^{-1})}$$

Therefore, we conclude that Algorithm 1 finds an ϵ-approximate Lipschitz stationary point of Problem (1) in at most T iterations, where $T \leq \lceil \frac{2C}{\epsilon\min(\theta,(\mu^1)^{-1})} \rceil$.

□

C.2. Proof of Lemma 4.5

Proof. We define $H^t \triangleq A_s^2/\mu^t + M_s + \theta$, $\mathbf{x}_+^t \triangleq \mathbf{x}^t - \mathbf{r}^t/H^t \in \mathbb{R}^n$, $\mathrm{J} \triangleq \{i \,|\, \mathbf{x}_i^{t+1} \neq 0\}$, and $\mathrm{J}^c \triangleq \{i \,|\, \mathbf{x}_i^{t+1} = 0\}$.

(a) Due to the optimality of $\mathbf{x}^{t+1}$ in Problem (5) that $\mathbf{x}^{t+1} = \arg\min_{\|\mathbf{x}\|_0 \leq s} \frac{1}{2}\|\mathbf{x} - \mathbf{x}_+^t\|_2^2$, we have $\|\mathbf{x}^{t+1} - \mathbf{x}_+^t\| \leq \|\mathbf{x} - \mathbf{x}_+^t\|$ for all $\|\mathbf{x}\|_0 \leq s$. Given that $\|\mathbf{x}^t\|_0 \leq s$, we let $\mathbf{x} = \mathbf{x}^t$, resulting in:

$$\|\mathbf{x}^{t+1} - \mathbf{x}_+^t\| \leq \|\mathbf{x}^t - \mathbf{x}_+^t\|. \tag{40}$$

We derive the following inequalities:

$$\|\mathbf{x}^{t+1} - \mathbf{x}^t\| \overset{①}{\leq} \|\mathbf{x}^{t+1} - \mathbf{x}_+^t\| + \|\mathbf{x}_+^t - \mathbf{x}^t\| \overset{②}{\leq} \tfrac{1}{H^t}\|\mathbf{r}^t\| + \tfrac{1}{H^t}\|\mathbf{r}^t\| = \tfrac{2}{H^t}\|\mathbf{r}^t\|,$$

where step ① uses the triangle inequality; step ② uses Inequality (40).

(b) We have the following inequalities:

$$\begin{aligned}
\langle \mathbf{x}_+^t, \mathbf{x}_+^t \rangle &= \langle [\mathbf{x}_+^t]_{\mathrm{J}}, [\mathbf{x}_+^t]_{\mathrm{J}} \rangle + \|[\mathbf{x}_+^t]_{\mathrm{J}^c}\|_2^2 \\
&\geq \langle [\mathbf{x}_+^t]_{\mathrm{J}}, [\mathbf{x}_+^t]_{\mathrm{J}} \rangle + 0 \\
&\overset{①}{=} \langle [\mathbf{x}_+^t]_{\mathrm{J}}, [\mathbf{x}^{t+1}]_{\mathrm{J}} \rangle \\
&\overset{②}{=} \langle [\mathbf{x}_+^t]_{\mathrm{J}}, [\mathbf{x}^{t+1}]_{\mathrm{J}} \rangle + \langle [\mathbf{x}_+^t]_{\mathrm{J}^c}, [\mathbf{x}^{t+1}]_{\mathrm{J}^c} \rangle \\
&\overset{③}{=} \langle \mathbf{x}_+^t, \mathbf{x}^{t+1} \rangle,
\end{aligned} \tag{41}$$

where step ① uses the fact $[\mathbf{x}_+^t]_{\mathrm{J}} = [\mathbf{x}^{t+1}]_{\mathrm{J}}$; step ② uses $[\mathbf{x}^{t+1}]_{\mathrm{J}^c} = \mathbf{0}$; step ③ uses $\mathrm{J} \cup \mathrm{J}^c = \{1, 2, ..., n\}$.

(c) We derive the following inequalities:

$$\begin{aligned}
&\tfrac{H^t}{2}\|\mathbf{x}^{t+1} - \bar{\mathbf{x}}\|_2^2 - \tfrac{H^t}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2 - \tfrac{H^t}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2 \\
\overset{①}{=}\; & H^t\langle \mathbf{x}^t - \mathbf{x}^{t+1}, \bar{\mathbf{x}} - \mathbf{x}^t \rangle \\
\overset{②}{=}\; & H^t\langle \mathbf{x}_+^t - \mathbf{x}^{t+1}, \bar{\mathbf{x}} - \mathbf{x}^t \rangle + H^t\langle \mathbf{r}^t/H^t, \bar{\mathbf{x}} - \mathbf{x}^t \rangle \\
\overset{③}{\leq}\; & H^t\langle \mathbf{x}_+^t - \mathbf{x}^{t+1}, \bar{\mathbf{x}} - \mathbf{x}^t \rangle + \Upsilon^t - \tfrac{V_s}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2
\end{aligned} \tag{42}$$

where step ① uses the Pythagoras relation that $\|\mathbf{a} - \mathbf{b}\|_2^2 - \|\mathbf{c} - \mathbf{b}\|_2^2 = \|\mathbf{a} - \mathbf{c}\|_2^2 - 2\langle \mathbf{a} - \mathbf{c}, \mathbf{b} - \mathbf{c} \rangle$ for all $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$; step ② uses $\mathbf{x}^t = \mathbf{x}_+^t + \mathbf{r}^t/H^t$; step ③ uses **Part (b)** of this lemma.

We now bound the first term of the right-hand side in Inequality (42) using the following inequalities:

$$\begin{aligned}
H^t\langle \mathbf{x}_+^t - \mathbf{x}^{t+1}, \bar{\mathbf{x}} - \mathbf{x}^t \rangle &\overset{①}{=} H^t\langle \mathbf{x}_+^t - \mathbf{x}^{t+1}, \bar{\mathbf{x}} - \mathbf{x}^t - \mathbf{x}_+^t + \mathbf{x}_+^t \rangle \\
&\overset{②}{\leq} H^t\langle \mathbf{x}_+^t - \mathbf{x}^{t+1}, \bar{\mathbf{x}} - \mathbf{x}^t + \mathbf{x}_+^t \rangle \\
&= H^t\langle \mathbf{x}_+^t - \mathbf{x}^{t+1}, \bar{\mathbf{x}} \rangle + H^t\langle \mathbf{x}_+^t - \mathbf{x}^{t+1}, \mathbf{x}_+^t - \mathbf{x}^t \rangle \\
&\overset{③}{\leq} H^t(\|\mathbf{x}_+^t - \mathbf{x}^{t+1}\|\|\bar{\mathbf{x}}\| + \|\mathbf{x}_+^t - \mathbf{x}^{t+1}\|\|\mathbf{x}_+^t - \mathbf{x}^t\|) \\
&\overset{④}{\leq} H^t\|\mathbf{x}_+^t - \mathbf{x}^t\|(\|\bar{\mathbf{x}}\| + \|\mathbf{x}_+^t - \mathbf{x}^t\|) \\
&\overset{⑤}{=} \|\mathbf{r}^t\|(\|\bar{\mathbf{x}}\| + \tfrac{\|\mathbf{r}^t\|}{H^t}),
\end{aligned} \tag{43}$$

where step ① uses the fact that $\bar{\mathbf{x}} - \mathbf{x}^t = (\bar{\mathbf{x}} - \mathbf{x}_+^t) + (\mathbf{x}_+^t - \mathbf{x}^t)$; step ② uses Inequality (41); step ③ uses the Cauchy-Schwarz Inequality, step ④ uses Inequality (40); step ⑤ uses $\|\mathbf{x}_+^t - \mathbf{x}^t\| = \frac{1}{H^t}\|\mathbf{r}^t\|$.

Finally, we derive from Inequalities (42) and (43):

$$\begin{aligned}
&\tfrac{H^t}{2}\|\mathbf{x}^{t+1} - \bar{\mathbf{x}}\|_2^2 - (\tfrac{H^t}{2} - \tfrac{V_s}{2})\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2 \\
\leq\; & \tfrac{H^t}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2 + \tfrac{1}{H^t}\|\mathbf{r}^t\|_2^2 + \Upsilon^t + \|\mathbf{r}^t\|\|\bar{\mathbf{x}}\| \\
\overset{①}{\leq}\; & \tfrac{2+1}{H^t}\|\mathbf{r}^t\|_2^2 + \Upsilon^t + \|\mathbf{r}^t\|\|\bar{\mathbf{x}}\|,
\end{aligned}$$

where step ① uses $\|\mathbf{x}^{t+1} - \mathbf{x}^t\| \leq \frac{2}{H^t}\|\mathbf{r}^t\|$ as shown in **Part (a)** of this lemma.

□

C.3. Proof of Theorem 4.6

Proof. Assume constant stepsizes are used with $\mu^t = \bar{\mu}$ for all $t \geq 1$.

We define $H \triangleq A_s^2/\bar{\mu} + M_s + \theta$, $\gamma \triangleq 1 - \frac{V_s}{H}$, $\Delta_{\mathbf{x}}^t \triangleq \|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2$, and $\Delta_F^t \triangleq [\min_{i=1}^t F(\mathbf{x}^i)] - F(\bar{\mathbf{x}})$.

First, using **Part (c)** in Lemma 4.1, we have: $\|\mathbf{r}^t\| \leq L_F$.

Second, using **Part (b)** Lemma 4.2, we have the upper bound of Υ^t that $\forall t,\ \Upsilon^t \leq \frac{\bar{\mu}}{2}L_h^2 - \Delta_F^t$.

Third, it holds that $H^t = H$ for all $t \geq 1$.

(a) Using the inequality in **Part (c)** in Lemma 4.5, we have the following recursive formulation:

$$\begin{aligned}
\Delta_{\mathbf{x}}^{t+1} &\leq (1 - \tfrac{V_s}{H})\Delta_{\mathbf{x}}^t + \tfrac{6}{H^2}\|\mathbf{r}^t\|_2^2 + \tfrac{2}{H}\Upsilon^t + \tfrac{2}{H}\|\mathbf{r}^t\| \cdot \|\bar{\mathbf{x}}\| \\
&\overset{①}{\leq} \gamma\Delta_{\mathbf{x}}^t + \tfrac{6L_F^2}{H^2} + \tfrac{\bar{\mu}L_h^2 - 2\Delta_F^t}{H} + \tfrac{2L_F\|\bar{\mathbf{x}}\|}{H},
\end{aligned}$$

where step ① uses the definition of $\gamma \triangleq 1 - \frac{V_s}{H}$, $\|\mathbf{r}^t\| \leq L_F$, and $\Upsilon^t \leq \frac{1}{2}\bar{\mu}L_h^2 - \Delta_F^t$.

(b) Let $T \geq 1$ be any integer. Applying Lemma A.2 with $\Phi^t = \Delta_{\mathbf{x}}^t$ and $\Lambda^t = \frac{6L_F^2}{H^2} + \frac{\bar{\mu}L_h^2 - 2\Delta_F^t}{H} + \frac{2L_F\|\bar{\mathbf{x}}\|}{H}$, we have:

$$\begin{aligned}
\Delta_{\mathbf{x}}^{T+1} &\leq \gamma^T\Delta_{\mathbf{x}}^1 + \tfrac{1}{1-\gamma} \cdot \max_{t=1}^{T}\left(\tfrac{6L_F^2}{H^2} + \tfrac{\bar{\mu}L_h^2 - 2\Delta_F^t}{H} + \tfrac{2L_F\|\bar{\mathbf{x}}\|}{H}\right) \\
&\overset{①}{=} \gamma^T\Delta_{\mathbf{x}}^1 + \tfrac{6L_F^2}{V_sH} + \tfrac{\bar{\mu}L_h^2}{V_s} - \tfrac{2\Delta_F^T}{V_s} + \tfrac{2L_F\|\bar{\mathbf{x}}\|}{V_s} \\
&\overset{②}{\leq} \gamma^T\Delta_{\mathbf{x}}^1 + \tfrac{6L_F^2}{V_sA_s^2}\bar{\mu} + \tfrac{L_h^2}{V_s}\bar{\mu} - \tfrac{2\Delta_F^T}{V_s} + \tfrac{2L_F\|\bar{\mathbf{x}}\|}{V_s} \\
&\overset{③}{=} \tfrac{2}{V_s}\left(K_1\gamma^T + D_1\bar{\mu} - \Delta_F^T + L_F\|\bar{\mathbf{x}}\|\right),
\end{aligned} \tag{44}$$

where step ① uses $\max_{t=1}^T[-\Delta_F^t] = -\Delta_F^T$ since $\Delta_F^1 \geq \Delta_F^2 \geq ... \geq \Delta_F^T \geq 0$ and $\gamma \triangleq 1 - \frac{V_s}{H}$; step ② uses $H \triangleq A_s^2/\bar{\mu} + M_s + \theta \geq A_s^2/\bar{\mu}$; step ③ uses the definitions of $K_1 \triangleq \frac{1}{2}V_s\Delta_{\mathbf{x}}^1$ and $D_1 \triangleq \frac{3L_F^2}{A_s^2} + \frac{1}{2}L_h^2$.

We now focus on Inequality (44). Using the fact that $\Delta_{\mathbf{x}}^{T+1} \geq 0$, we obtain: $\Delta_F^T \leq K_1\gamma^t + D_1\bar{\mu} + L_F\|\bar{\mathbf{x}}\|$.

Using the fact that $\Delta_F^T \geq 0$, we obtain: $\Delta_{\mathbf{x}}^{T+1} \leq \left(K_1\gamma^T + D_1\bar{\mu} + L_F\|\bar{\mathbf{x}}\|\right)\frac{2}{V_s}$.

□

C.4. Proof of Theorem 4.7

Proof. Assume diminishing stepsizes are used with $\mu^t = \frac{\eta}{t+t_0}$ for all $t \geq 1$, where $\eta = \frac{A_s^2}{V_s}$.

We define $H^t \triangleq A_s^2/\mu^t + M_s + \theta$, $\Delta_{\mathbf{x}}^t \triangleq \|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2$, and $\Delta_F^t \triangleq [\min_{i=1}^t F(\mathbf{x}^i)] - F(\bar{\mathbf{x}})$.

First, using **Part (c)** in Lemma 4.1, we have: $\|\mathbf{r}^t\| \leq L'_F$.

Second, using Lemma 4.2, we have: $\sum_{t=1}^T \Upsilon^t \leq C_\Upsilon\left(1 + \ln(T)\right) - T\Delta_F^T$ for any $T \geq 1$.

Third, using the definition of H^t and the choice of $\eta = \frac{A_s^2}{V_s}$, we have:

$$H^t = V_s\eta/\mu^t + M_s + \theta = V_s(t + t_0) + M_s + \theta, \tag{45}$$

(a) We have the following inequalities:

$$\begin{aligned} \tfrac{1}{2}(H^{t+1}-V_s)\Delta_{\mathbf{x}}^{t+1} &\overset{\text{①}}{\leq} \tfrac{1}{2}H^t\Delta_{\mathbf{x}}^{t+1} \\ &\overset{\text{②}}{\leq} \tfrac{1}{2}(H^t-V_s)\Delta_{\mathbf{x}}^{t} + \tfrac{3\|\mathbf{r}^t\|_2^2}{H^t} + \Upsilon^t + \|\mathbf{r}^t\|\|\bar{\mathbf{x}}\| \\ &\overset{\text{③}}{\leq} \tfrac{1}{2}(H^t-V_s)\Delta_{\mathbf{x}}^{t} + \tfrac{3(L_F')^2}{V_s\cdot t} + \Upsilon^t + L_F'\|\bar{\mathbf{x}}\|, \end{aligned} \tag{46}$$

where step ① uses $H^{t+1}=H^t+V_s$, which can be implied by Equality (45); step ② uses **Part (b)** in Lemma 4.5; step ③ uses $\|\mathbf{r}^t\|\leq L_F'$ and $H^t \geq A_s^2/\mu^t = A_s^2(t+t_0)/\eta \geq A_s^2 t/\eta = V_s t$.

(b) Let $T\geq 1$ be any integer. Summing Inequality (46) over t from 1 to T, we obtain:

$$\begin{aligned} 0 &\leq -\tfrac{1}{2}(H^{T+1}-V_s)\Delta_{\mathbf{x}}^{T+1} + \tfrac{1}{2}(H^1-V_s)\Delta_{\mathbf{x}}^{1} + \tfrac{3(L_F')^2}{V_s}\textstyle\sum_{t=1}^T \tfrac{1}{t} + \sum_{t=1}^T \Upsilon^t + TL_F'\|\bar{\mathbf{x}}\| \\ &\overset{\text{①}}{\leq} -\tfrac{V_s}{2}(T+1)\Delta_{\mathbf{x}}^{T+1} + \tfrac{H^1}{2}\Delta_{\mathbf{x}}^1 + [\tfrac{3(L_F')^2}{V_s} + C_\Upsilon](\ln(T)+1) - T\Delta_F^T + TL_F'\|\bar{\mathbf{x}}\| \\ &\overset{\text{②}}{\leq} -\tfrac{V_s}{2}(T+1)\Delta_{\mathbf{x}}^{T+1} + K_2 + D_2(\ln(T)+1) - T\Delta_F^T + TL_F'\|\bar{\mathbf{x}}\|, \end{aligned} \tag{47}$$

where step ① uses $H^{T+1}-V_s = V_s(T+1+t_0)+M_s+\theta-V_s \geq V_s(T+1)+M_s+\theta \geq V_s(T+1)$, $-\frac{1}{2}V_s\Delta_{\mathbf{x}}^1\leq 0$, the fact that $\sum_{t=1}^T \frac{1}{t}\leq \ln(T)+1$, and the upper bound $\sum_{t=1}^T \Upsilon^t \leq C_\Upsilon\,(1+\ln(T)) - T\Delta_F^T$; step ② uses the definition of $K_2 \triangleq \frac{H^1}{2}\Delta_{\mathbf{x}}^1$, and the definition of $D_2 \triangleq \frac{3(L_F')^2}{V_s}+C_\Upsilon$.

We now focus on Inequality (47). Using the fact that $\frac{V_s}{2}(T+1)\Delta_{\mathbf{x}}^{T+1}\geq 0$, we obtain: $T\Delta_F^T \leq K_2 + TL_F'\|\bar{\mathbf{x}}\| + D_2(\ln(T)+1)$, leading to $\Delta_F^T \leq \frac{K_2}{T} + L_F'\|\bar{\mathbf{x}}\| + D_2\frac{\ln(T)+1}{T}$.

Using the fact that $T\Delta_F^T\geq 0$, we obtain: $\frac{V_s}{2}(T+1)\Delta_{\mathbf{x}}^{T+1} \leq K_2 + TL_F'\|\bar{\mathbf{x}}\| + D_2(\ln(T)+1)$, leading to $\Delta_{\mathbf{x}}^{T+1} \leq (\frac{K_2}{T+1} + D_2\frac{\ln(T)+1}{T+1} + L_F'\|\bar{\mathbf{x}}\|)\frac{2}{V_s}$.

□

C.5. Proofs for Lemma 4.8

Proof. We denote $\Omega_n^k \triangleq \{\mathrm{B}_{(i)}\}_{i=1}^{C_n^k}$ as all the possible combinations of the index vectors choosing k items from n with $\mathrm{B}_i\in\mathbb{N}^k$, $\forall i$. For any vector $\mathbf{x}\in\mathbb{R}^n$, we have:

$$\textstyle\sum_{\mathrm{B}\in\Omega_n^k} \mathbf{x}^\mathsf{T}(\mathrm{U}_\mathrm{B}\mathrm{U}_\mathrm{B}^\mathsf{T})\mathbf{z} \overset{\text{①}}{=} \sum_{\mathrm{B}\in\Omega_n^k}\langle \mathbf{x}_\mathrm{B},\mathbf{z}_\mathrm{B}\rangle \overset{\text{②}}{=} C_n^k\frac{k}{n}\langle\mathbf{x},\mathbf{z}\rangle,$$

where step ① uses $\mathrm{U}_\mathrm{B}^\mathsf{T}\mathbf{x}=\mathbf{x}_\mathrm{B}$ and $\mathrm{U}_\mathrm{B}^\mathsf{T}\mathbf{z}=\mathbf{z}_\mathrm{B}$; step ② uses the basic induction that every entry $(\mathbf{x}_i\cdot\mathbf{z}_i)$ is present within the term $(\sum_{\mathrm{B}\in\Omega_n^k}\langle\mathbf{x}_\mathrm{B},\mathbf{z}_\mathrm{B}\rangle)$ for a total of $(C_n^k\cdot\frac{k}{n})$ times for all $i\in[n]$.

Given B is chosen from Ω_n^k randomly and uniformly, we have: $\mathbb{E}_\mathrm{B}[\|\mathbf{x}_\mathrm{B}\|_2^2] = \frac{1}{C_n^k}\sum_{i=1}^{C_n^k}\|\mathbf{x}_{\mathrm{B}_i}\|_2^2 = \frac{k}{n}\|\mathbf{x}\|_2^2$.

□

C.6. Proof of Theorem 4.10

Proof. We denote $\mathbf{r}^t \triangleq \nabla_\mathbf{x}\mathcal{R}(\mathbf{x}^t,\mathbf{y}^t;\mu^t)$, $\mathbf{H}^t = (\mathbf{A}^\mathsf{T}\mathbf{A}+\theta_1\mathbf{I}_n)/\mu^t + \tilde{\mathbf{M}} + \theta_2\mathbf{I}_n$, and $\theta = \frac{\theta_1}{\mu^1}+\theta_2$.

(a) We focus on the $\mathbf{x}$-subproblem. We have from Problem (6) that

$$\mathbb{E}_{\xi^t}[\langle\mathbf{r}^t,\mathbf{x}^{t+1}-\mathbf{x}^t\rangle + \tfrac{1}{2}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|_{\mathbf{H}^t}^2] \leq \mathbb{E}_{\xi^t}[\langle\mathbf{r}^t,\mathbf{x}^{t}-\mathbf{x}^t\rangle + \tfrac{1}{2}\|\mathbf{x}^{t}-\mathbf{x}^t\|_{\mathbf{H}^t}^2] = 0.$$

Using Assumption 2.2 and the inherent structure of the function $\mathcal{R}(\mathbf{x}^t,\mathbf{y}^t;\mu^t)$, we have:

$$\mathcal{R}(\mathbf{x}^{t+1},\mathbf{y}^t;\mu^t) \leq \mathcal{R}(\mathbf{x}^t,\mathbf{y}^t;\mu^t) + \langle\mathbf{r}^t,\mathbf{x}^{t+1}-\mathbf{x}^t\rangle + \tfrac{1}{2}\|\mathbf{x}^t-\mathbf{x}^{t+1}\|_{[\mathbf{A}^\mathsf{T}\mathbf{A}/\mu^t+\tilde{\mathbf{M}}]}^2.$$

We observe that the following equality holds:

$$\mathcal{R}(\mathbf{x}^{t+1}, \mathbf{y}^t; \mu^t) - \mathcal{R}(\mathbf{x}^t, \mathbf{y}^t; \mu^t) = \mathcal{J}(\mathbf{x}^{t+1}, \mathbf{y}^t; \mu^t) - \mathcal{J}(\mathbf{x}^t, \mathbf{y}^t; \mu^t).$$

Summing these three inequalities, we obtain:

$$\begin{aligned} \mathbb{E}_{\xi^t}[\mathcal{J}(\mathbf{x}^{t+1}, \mathbf{y}^t; \mu^t) - \mathcal{J}(\mathbf{x}^t, \mathbf{y}^t; \mu^t)] &\leq \mathbb{E}_{\xi^t}[-(\tfrac{\theta_1}{\mu^t} + \theta_2) \cdot \frac{1}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2] \\ &\overset{①}{\leq} \mathbb{E}_{\xi^t}[-\theta \cdot \frac{1}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2], \end{aligned} \tag{48}$$

where step ① uses $\frac{\theta_1}{\mu^t} + \theta_2 \leq \frac{\theta_1}{\mu^1} + \theta_2 \triangleq \theta$ as the sequence $\{\mu^t\}_{t=1}^{\infty}$ is non-increasing.

We now focus on the $\mathbf{y}$-subproblem. Similar to the proof for Theorem 4.4, we have:

$$\mathcal{J}(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}; \mu^t) - \mathcal{J}(\mathbf{x}^{t+1}, \mathbf{y}^t; \mu^t) \leq -\frac{1}{2\mu^1}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|_2^2. \tag{49}$$

Using the continuity of $\mathcal{J}(\mathbf{x}^t, \mathbf{y}^t; \mu)$ *w.r.t.* μ as detailed in **Part (e)** of Lemma 4.1, we obtain:

$$0 \leq \mathbb{E}_{\xi^t}[\mathcal{J}(\mathbf{x}^t, \mathbf{y}^t; \mu^t) - \mathcal{J}(\mathbf{x}^t, \mathbf{y}^t; \mu^{t-1})] \leq \tfrac{L_h^2}{2}(\tfrac{(\mu^{t-1})^2}{\mu^t} - \mu^{t-1}) \triangleq \Psi^t. \tag{50}$$

Summing Inequalities (48), (49), and (50) together, we have:

$$\begin{aligned} & \mathbb{E}_{\xi^T}[\frac{1}{2\mu^1}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|_2^2 + \tfrac{\theta}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2] \\ \leq\ & \mathbb{E}_{\xi^{t-1}}[\mathcal{J}(\mathbf{x}^t, \mathbf{y}^t; \mu^{t-1})] - \mathbb{E}_{\xi^t}[\mathcal{J}(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}; \mu^t)] + \tfrac{1}{2}L_h^2(\tfrac{(\mu^{t-1})^2}{\mu^t} - \mu^{t-1}) \\ \overset{①}{=}\ & \mathcal{J}^t - \mathcal{J}^{t+1} + \Psi^t, \end{aligned} \tag{51}$$

where step ① uses the definitions of Ψ^t and $\mathcal{J}^{t+1} \triangleq \mathbb{E}_{\xi^t}[\mathcal{J}(\mathbf{x}^{t+1}, \mathbf{y}^{t+1}; \mu^t)]$.

(b) Summing Inequality (51) over t from 1 to T, we have:

$$\begin{aligned} & \textstyle\sum_{t=1}^T \frac{1}{2\mu^1}\|\mathbf{y}^{t+1} - \mathbf{y}^t\|_2^2 + \sum_{t=1}^T \frac{\theta}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2 \\ \leq\ & \mathcal{J}^1 - \mathcal{J}^{T+1} + [\textstyle\sum_{t=1}^T \Psi^t] \overset{①}{\leq} \mathcal{J}^1 - \mathcal{J}^{T+1} + \eta L_h^2 = C < +\infty, \end{aligned}$$

where step ① uses $\sum_{t=1}^T \Psi^t < \sum_{t=1}^{\infty} \Psi^t \leq \eta L_h^2$, as demonstrated in **Part (f)** of Lemma 4.1.

(c) As a result, there exists an index $\bar{t}$ with $1 \leq \bar{t} \leq T$ such that $\frac{1}{2\mu^1}\|\mathbf{y}^{\bar{t}+1} - \mathbf{y}^{\bar{t}}\|_2^2 + \frac{\theta}{2}\|\mathbf{x}^{\bar{t}+1} - \mathbf{x}^{\bar{t}}\|_2^2 \leq \frac{C}{T}$, leading to:

$$\|\mathbf{y}^{\bar{t}+1} - \mathbf{y}^{\bar{t}}\|_2^2 + \|\mathbf{x}^{\bar{t}+1} - \mathbf{x}^{\bar{t}}\|_2^2 \leq \frac{2C}{T \cdot \min(\theta, (\mu^1)^{-1})}. \tag{52}$$

We define $\Gamma_x(\mathbf{x}, \mathbf{y}; \mu) \triangleq \frac{1}{C_n^k}\sum_{\mathrm{B} \in \Omega_n^k} \mathrm{dist}^2(\mathbf{x}_{\mathrm{B}}, \arg\min_{\mathbf{z}_{\mathrm{B}}} \delta(\mathbf{U}_{\mathrm{B}}\mathbf{z}_{\mathrm{B}} + \mathbf{U}_{\mathrm{B}^c}\mathbf{x}_{\mathrm{B}^c}) + \ddot{\mathcal{M}}(\mathbf{U}_{\mathrm{B}}\mathbf{z}_{\mathrm{B}} + \mathbf{U}_{\mathrm{B}^c}\mathbf{x}_{\mathrm{B}^c}, \mathbf{x}, \mathbf{y}; \mu)$ and $\Gamma_y(\mathbf{x}, \mathbf{y}; \mu) \triangleq \mathrm{dist}^2(\mathbf{y}, \arg\min_{\mathbf{y}'} \mathcal{J}(\mathbf{x}, \mathbf{y}'; \mu))$. It is important to note that $\mathbf{x}^{\bar{t}+1}$ and $\mathbf{x}^{\bar{t}}$ differ in at most k coordinates. We have:

$$\|\mathbf{x}^{\bar{t}+1} - \mathbf{x}^{\bar{t}}\|_2^2 + \|\mathbf{y}^{\bar{t}+1} - \mathbf{y}^{\bar{t}}\|_2^2 \geq \Gamma_x(\mathbf{x}^{\bar{t}}, \mathbf{y}^{\bar{t}}; \mu) + \Gamma_y(\mathbf{x}^{\bar{t}}, \mathbf{y}^{\bar{t}}; \mu) \tag{53}$$

for all $\bar{t} \geq 1$ and some sufficiently small $\mu = \mu^{\bar{t}} > 0$. Combining Inequalities (52) and (53), we have:

$$\Gamma_x(\mathbf{x}^{\bar{t}}, \mathbf{y}^{\bar{t}}; \mu^{\bar{t}}) + \Gamma_y(\mathbf{x}^{\bar{t}}, \mathbf{y}^{\bar{t}}; \mu^{\bar{t}}) \leq \frac{2C}{T \cdot \min(\theta, (\mu^1)^{-1})}.$$

Therefore, we conclude that Algorithm 1 finds an ϵ-approximate block-k stationary point of Problem (1) in at most T iterations in the sense of expectation, where $T \leq \lceil \frac{2C}{\epsilon \min(\theta, (\mu^1)^{-1})} \rceil = \mathcal{O}(\epsilon^{-1})$.

□

C.7. Proof of Lemma 4.12

Proof. For notation simplicity, we denote $\mathtt{B} = \mathtt{B}^t$, where t can be inferred from the context.

We define $\mathbf{H}^t \triangleq (\mathbf{A}^\mathsf{T}\mathbf{A} + \theta_1 \mathbf{I}_n)/\mu^t + \tilde{\mathbf{M}} + \theta_2 \mathbf{I}_n \in \mathbb{R}^{n\times n}$, and $\mathbf{H}^t_* \triangleq \mathrm{U}_{\mathtt{B}^t}\mathrm{U}_{\mathtt{B}^t}^\mathsf{T}\mathbf{H}^t\mathrm{U}_{\mathtt{B}^t}\mathrm{U}_{\mathtt{B}^t}^\mathsf{T} \in \mathbb{R}^{n\times n}$.

We define $\overline{\mathrm{H}}^t \triangleq \overline{\mathrm{V}} + \theta + \frac{\theta_1 + \overline{\mathrm{A}}}{\mu^t}$, and $\underline{\mathrm{H}}^t \triangleq \underline{\mathrm{V}} + \theta + \frac{\theta_1 + \underline{\mathrm{A}}}{\mu^t}$.

(a) Problem (6) in Algorithm 1 is equivalent to solving the following optimization problem:

$$\mathbf{x}_{\mathtt{B}}^{t+1} \in \arg\min_{\mathbf{z}_{\mathtt{B}}\in\mathbb{R}^k} \mathcal{W}(\mathbf{z}_{\mathtt{B}}) \ \mathrm{s.\,t.}\ \|\mathbf{z}_{\mathtt{B}}\|_0 + \|\mathbf{x}^t_{\mathtt{B}^c}\|_0 \leq s, \tag{54}$$

where $\mathcal{W}(\mathbf{z}_{\mathtt{B}}) \triangleq \langle \mathbf{z}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}}, \mathbf{r}^t_{\mathtt{B}}\rangle + \frac{1}{2}(\mathbf{z}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}})^\mathsf{T}[\mathbf{H}^t]_{\mathtt{BB}}(\mathbf{z}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}})$. By the optimality of $\mathbf{x}^{t+1}_{\mathtt{B}}$, we have: $\mathcal{W}(\mathbf{x}^{t+1}_{\mathtt{B}}) \leq \mathcal{W}(\mathbf{x}^t_{\mathtt{B}}) = 0$, leading to:

$$\langle \mathbf{x}^{t+1}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}}, \mathbf{r}^t_{\mathtt{B}}\rangle + \frac{1}{2}(\mathbf{x}^{t+1}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}})^\mathsf{T}[\mathbf{H}^t]_{\mathtt{BB}}(\mathbf{x}^{t+1}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}}) \leq 0. \tag{55}$$

We derive the following inequalities:

$$\frac{1}{2}\underline{\mathrm{H}}^t\|\mathbf{x}^{t+1}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}}\|_2^2 \overset{\text{①}}{\leq} \frac{1}{2}(\mathbf{x}^{t+1}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}})^\mathsf{T}[\mathbf{H}^t]_{\mathtt{BB}}(\mathbf{x}^{t+1}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}}) \overset{\text{②}}{\leq} -\langle \mathbf{x}^{t+1}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}}, \mathbf{r}^t_{\mathtt{B}}\rangle \overset{\text{③}}{=} \|\mathbf{x}^{t+1}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}}\|\|\mathbf{r}^t_{\mathtt{B}}\|,$$

where step ① uses $\underline{\mathrm{H}}^t\mathbf{I}_k \preceq [\mathbf{H}^t]_{\mathtt{BB}} \preceq \overline{\mathrm{H}}^t\mathbf{I}_k$; step ② uses Inequalities (55); step ③ uses the Cauchy-Schwarz Inequality. Dividing both sides by $\|\mathbf{x}^{t+1}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}}\|$, we have: $\mathbb{E}_{\xi^t}[\|\mathbf{x}^{t+1}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}}\|] \leq \mathbb{E}_{\xi^t}[\frac{2}{\underline{\mathrm{H}}^t}\|\mathbf{r}^t_{\mathtt{B}}\|]$. Using the result in Lemma 4.8, we have:

$$Z^k_n\mathbb{E}_{\xi^t}[\|\mathbf{x}^{t+1} - \mathbf{x}^t\|] \leq Z^k_n\mathbb{E}_{\xi^t}[\tfrac{2}{\underline{\mathrm{H}}^t}\|\mathbf{r}^t\|]. \tag{56}$$

(b) For notation convenience, we define:

$$\mathtt{B}_1 \triangleq \{i \,|\, \mathbf{x}^{t+1}_i \neq 0,\ i \in \mathtt{B}\},\ \text{and}\ \mathtt{B}_2 \triangleq \{i \,|\, \mathbf{x}^{t+1}_i = 0,\ i \in \mathtt{B}\}.$$

The solution $\mathbf{x}^{t+1}_{\mathtt{B}} \in \mathbb{R}^k$ is a local minimizer for Problem (54) if and only if $[\nabla\mathcal{W}(\mathbf{x}^{t+1}_{\mathtt{B}})]_{\mathtt{B}_1} = 0$. We have:

$$\begin{aligned}
\mathbf{0} \ &= \ \mathbf{r}^t_{\mathtt{B}_1} + [[\mathbf{H}^t_{\mathtt{BB}}]_{\mathtt{B}_1\mathtt{B}}](\mathbf{x}^{t+1}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}}) \\
&\overset{\text{①}}{=} \ \mathbf{r}^t_{\mathtt{B}_1} + [[\mathbf{H}^t_{\mathtt{BB}}]_{\mathtt{B}_1\mathtt{B}_1}](\mathbf{x}^{t+1}_{\mathtt{B}_1} - \mathbf{x}^t_{\mathtt{B}_1}) + [[\mathbf{H}^t_{\mathtt{BB}}]_{\mathtt{B}_1\mathtt{B}_2}](\mathbf{x}^{t+1}_{\mathtt{B}_2} - \mathbf{x}^t_{\mathtt{B}_2}),
\end{aligned} \tag{57}$$

where step ① uses $\mathtt{B} = \mathtt{B}_1 \cup \mathtt{B}_2$. We derive the following equalities:

$$\begin{aligned}
& \mathbb{E}_{\xi^t}[\langle [[\mathbf{H}^t]_{\mathtt{BB}}](\mathbf{x}^{t+1}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}}), \mathbf{x}^{t+1}_{\mathtt{B}}\rangle] \\
= \ & \mathbb{E}_{\xi^t}[\begin{bmatrix} \mathbf{x}^{t+1}_{\mathtt{B}_1} - \mathbf{x}^t_{\mathtt{B}_1} \\ \mathbf{x}^{t+1}_{\mathtt{B}_2} - \mathbf{x}^t_{\mathtt{B}_2} \end{bmatrix}^\mathsf{T} \begin{bmatrix} [\mathbf{H}^t_{\mathtt{BB}}]_{\mathtt{B}_1\mathtt{B}_1} & [\mathbf{H}^t_{\mathtt{BB}}]_{\mathtt{B}_1\mathtt{B}_2} \\ [\mathbf{H}^t_{\mathtt{BB}}]_{\mathtt{B}_2\mathtt{B}_1} & [\mathbf{H}^t_{\mathtt{BB}}]_{\mathtt{B}_2\mathtt{B}_2} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{t+1}_{\mathtt{B}_1} \\ \mathbf{x}^{t+1}_{\mathtt{B}_2} \end{bmatrix}] \\
\overset{\text{①}}{=} \ & \mathbb{E}_{\xi^t}[\,\langle [[\mathbf{H}^t_{\mathtt{BB}}]_{\mathtt{B}_1\mathtt{B}_1}](\mathbf{x}^{t+1}_{\mathtt{B}_1} - \mathbf{x}^t_{\mathtt{B}_1}), \mathbf{x}^{t+1}_{\mathtt{B}_1}\rangle + \langle [[\mathbf{H}^t_{\mathtt{BB}}]_{\mathtt{B}_1\mathtt{B}_2}](\mathbf{x}^{t+1}_{\mathtt{B}_2} - \mathbf{x}^t_{\mathtt{B}_2}), \mathbf{x}^{t+1}_{\mathtt{B}_1}\rangle] + 0 + 0 \\
\overset{\text{②}}{=} \ & \mathbb{E}_{\xi^t}[\,\langle -\mathbf{r}^t_{\mathtt{B}_1} - [[\mathbf{H}^t_{\mathtt{BB}}]_{\mathtt{B}_1\mathtt{B}_2}](\mathbf{x}^{t+1}_{\mathtt{B}_2} - \mathbf{x}^t_{\mathtt{B}_2}), \mathbf{x}^{t+1}_{\mathtt{B}_1}\rangle + \langle [[\mathbf{H}^t_{\mathtt{BB}}]_{\mathtt{B}_1\mathtt{B}_2}](\mathbf{x}^{t+1}_{\mathtt{B}_2} - \mathbf{x}^t_{\mathtt{B}_2}), \mathbf{x}^{t+1}_{\mathtt{B}_1}\rangle] \\
= \ & \mathbb{E}_{\xi^t}[\,\langle -\mathbf{r}^t_{\mathtt{B}_1}, \mathbf{x}^{t+1}_{\mathtt{B}_1}\rangle] \\
\overset{\text{③}}{=} \ & \mathbb{E}_{\xi^t}[-\langle \mathbf{r}^t_{\mathtt{B}_1}, \mathbf{x}^{t+1}_{\mathtt{B}_1}\rangle - \langle \mathbf{r}^t_{\mathtt{B}_2}, \mathbf{x}^{t+1}_{\mathtt{B}_2}\rangle] = \mathbb{E}[-\langle \mathbf{r}^t_{\mathtt{B}}, \mathbf{x}^{t+1}_{\mathtt{B}}\rangle] \\
\overset{\text{④}}{=} \ & -Z^k_n\langle \mathbf{r}^t, \mathbf{x}^{t+1}\rangle,
\end{aligned} \tag{58}$$

where step ① uses the fact that $[\mathbf{x}^{t+1}]_{\mathtt{B}_2} = \mathbf{0}$; step ② uses the optimality condition as in Equality (57); step ③ uses $\mathtt{B} = [\mathtt{B}_1; \mathtt{B}_2]$ and the fact that $[\mathbf{x}^{t+1}]_{\mathtt{B}_2} = \mathbf{0}$; step ④ uses Lemma 4.8 with $Z^k_n = \frac{k}{n}$.

(c) We derive the following equalities:

$$\begin{aligned}
& \mathbb{E}_{\xi^t}[\frac{1}{2}\|\mathbf{x}^{t+1} - \bar{\mathbf{x}}\|^2_{\mathbf{H}^t_*}] - \mathbb{E}_{\xi^t}[\frac{1}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|^2_{\mathbf{H}^t_*}] + \mathbb{E}_{\xi^t}[\frac{1}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2_{\mathbf{H}^t_*}] \\
\overset{\text{①}}{=} \ & \mathbb{E}_{\xi^t}[\langle \mathbf{H}^t_*(\mathbf{x}^{t+1} - \mathbf{x}^t), \mathbf{x}^{t+1} - \bar{\mathbf{x}}\rangle] \\
\overset{\text{②}}{=} \ & \mathbb{E}_{\xi^t}[\langle [[\mathbf{H}^t]_{\mathtt{BB}}](\mathbf{x}^{t+1}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}}), \mathbf{x}^{t+1}_{\mathtt{B}} - \bar{\mathbf{x}}_{\mathtt{B}}\rangle] \\
= \ & \underbrace{\mathbb{E}_{\xi^t}[\langle [[\mathbf{H}^t]_{\mathtt{BB}}](\mathbf{x}^t_{\mathtt{B}} - \mathbf{x}^{t+1}_{\mathtt{B}}), \bar{\mathbf{x}}_{\mathtt{B}}\rangle]}_{\Gamma_1} + \underbrace{\mathbb{E}_{\xi^t}[\langle [[\mathbf{H}^t]_{\mathtt{BB}}](\mathbf{x}^{t+1}_{\mathtt{B}} - \mathbf{x}^t_{\mathtt{B}}), \mathbf{x}^{t+1}_{\mathtt{B}}\rangle]}_{\Gamma_2},
\end{aligned} \tag{59}$$

where step ① uses the Pythagoras relation; step ② uses $[\mathbf{H}_*^t]_{\mathrm{BB}} = [\mathbf{H}^t]_{\mathrm{BB}}$ and $\mathbf{x}_{\mathrm{B}^c}^{t+1} - \mathbf{x}_{\mathrm{B}^c}^t = \mathbf{0}$.

We first bound the term Γ_1 in Equality (59) using the following inequalities:

$$\begin{aligned}
\Gamma_1 &= \mathbb{E}_{\xi^t}[\langle [[\mathbf{H}^t]_{\mathrm{BB}}](\mathbf{x}_{\mathrm{B}}^t - \mathbf{x}_{\mathrm{B}}^{t+1}), \bar{\mathbf{x}}_{\mathrm{B}} \rangle] \\
&\overset{\text{①}}{\leq} \mathbb{E}_{\xi^t}[\| [[\mathbf{H}^t]_{\mathrm{BB}}](\mathbf{x}_{\mathrm{B}}^t - \mathbf{x}_{\mathrm{B}}^{t+1}) \| \cdot \|\bar{\mathbf{x}}_{\mathrm{B}}\|] \\
&\overset{\text{②}}{\leq} \mathbb{E}_{\xi^t}[\overline{\mathrm{H}}^t \|\mathbf{x}_{\mathrm{B}}^t - \mathbf{x}_{\mathrm{B}}^{t+1}\| \cdot \|\bar{\mathbf{x}}_{\mathrm{B}}\|] \\
&\overset{\text{③}}{=} Z_n^k \overline{\mathrm{H}}^t \|\mathbf{x}^t - \mathbf{x}^{t+1}\| \cdot \|\bar{\mathbf{x}}\| \\
&\overset{\text{④}}{\leq} 2 Z_n^k \|\mathbf{r}^t\| \|\bar{\mathbf{x}}\| \overline{\mathrm{H}}^t / \underline{\mathrm{H}}^t = 2 Z_n^k \|\mathbf{r}^t\| \|\bar{\mathbf{x}}\| \kappa^t, \qquad (60)
\end{aligned}$$

where step ① uses the Cauchy-Schwarz Inequality; step ② uses $\underline{\mathrm{H}}^t \mathbf{I}_k \preceq [\mathbf{H}^t]_{\mathrm{BB}} \preceq \overline{\mathrm{H}}^t \mathbf{I}_k$; step ③ uses Lemma 4.8 with $Z_n^k = \frac{k}{n}$; step ④ uses Inequality (56).

We now bound the term Γ_2 in Equality (59) using the following inequalities:

$$\begin{aligned}
\Gamma_2 &\overset{\text{①}}{\leq} -Z_n^k \langle \mathbf{r}^t, \mathbf{x}^{t+1} \rangle \\
&= Z_n^k \langle \mathbf{r}^t, \bar{\mathbf{x}} - \mathbf{x}^t \rangle + Z_n^k \langle \mathbf{r}^t, \mathbf{x}^t - \mathbf{x}^{t+1} \rangle + Z_n^k \langle \mathbf{r}^t, -\bar{\mathbf{x}} \rangle \\
&\overset{\text{②}}{\leq} Z_n^k (\Upsilon^t - \tfrac{V_s}{2} \|\bar{\mathbf{x}} - \mathbf{x}^t\|_2^2) + Z_n^k \|\mathbf{r}^t\| \|\mathbf{x}^t - \mathbf{x}^{t+1}\| + Z_n^k \|\mathbf{r}^t\| \|\bar{\mathbf{x}}\| \\
&\overset{\text{③}}{\leq} Z_n^k (\Upsilon^t - \tfrac{V_s}{2} \|\bar{\mathbf{x}} - \mathbf{x}^t\|_2^2) + \tfrac{2 Z_n^k}{\underline{\mathrm{H}}^t} \|\mathbf{r}^t\|_2^2 + Z_n^k \|\mathbf{r}^t\| \|\bar{\mathbf{x}}\|, \qquad (61)
\end{aligned}$$

where step ① uses Equality (58); step ② uses Lemma 4.2 that $\langle \mathbf{r}^t, \bar{\mathbf{x}} - \mathbf{x}^t \rangle \leq \Upsilon^t - \frac{V_s}{2} \|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2$, and the Cauchy-Schwarz Inequality; step ③ uses Inequality (56).

In view of Inequalities (59), (60), and (61), we have:

$$\begin{aligned}
& \mathbb{E}_{\xi^t}[\tfrac{1}{2} \|\mathbf{x}^{t+1} - \bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2 - \tfrac{1}{2} \|\mathbf{x}^t - \bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2] \\
\leq\ & -\mathbb{E}_{\xi^t}[\tfrac{1}{2} \|\mathbf{x}^{t+1} - \mathbf{x}^t\|_{\mathbf{H}_*^t}^2] + \Gamma_1 + \Gamma_2 \\
\overset{\text{①}}{\leq}\ & 0 + Z_n^k (\Upsilon^t - \tfrac{V_s}{2} \|\bar{\mathbf{x}} - \mathbf{x}^t\|_2^2) + \tfrac{2 Z_n^k}{\underline{\mathrm{H}}^t} \|\mathbf{r}^t\|_2^2 + (1 + 2\kappa^t) Z_n^k \|\mathbf{r}^t\| \|\bar{\mathbf{x}}\|,
\end{aligned}$$

where step ① uses $-\mathbb{E}_{\xi^t}[\frac{1}{2} \|\mathbf{x}^{t+1} - \mathbf{x}^t\|_{\mathbf{H}_*^t}^2] \leq 0$.

□

C.8. Proof of Lemma 4.11

Proof. We initially establish the subsequent inequality:

$$\frac{a+b}{c+d} \leq \max(\frac{a}{c}, \frac{b}{d}), \forall a \geq 0, b \geq 0, c > 0, d > 0. \qquad (62)$$

We consider two cases: (***i***) $\frac{a}{c} \geq \frac{b}{d}$. we have: $b \leq \frac{ad}{c}$, leading to $\frac{a+b}{c+d} \leq \frac{a + \frac{ad}{c}}{c+d} = \frac{a}{c} \cdot \frac{c+d}{c+d} = \frac{a}{c}$. (***ii***) $\frac{a}{c} < \frac{b}{d}$. We have: $a \leq \frac{bc}{d}$, resulting in $\frac{a+b}{c+d} \leq \frac{\frac{bc}{d} + b}{c+d} = \frac{b}{d} \cdot \frac{c+d}{c+d} = \frac{b}{d}$. Therefore, Inequality (62) holds.

Using the definition of $\overline{\mathrm{H}}^t$ and $\underline{\mathrm{H}}^t$, we have:

$$\frac{\overline{\mathrm{H}}^t}{\underline{\mathrm{H}}^t} = \frac{\frac{\overline{\mathrm{A}} + \theta_1}{\mu^t} + \overline{\mathrm{V}} + \theta_2}{\frac{\underline{\mathrm{A}} + \theta_1}{\mu^t} + \underline{\mathrm{V}} + \theta_2} \overset{\text{①}}{\leq} \max(\frac{\overline{\mathrm{A}} + \theta_1}{\underline{\mathrm{A}} + \theta_1}, \frac{\overline{\mathrm{V}} + \theta_2}{\underline{\mathrm{V}} + \theta_2}) \overset{\text{②}}{\leq} 1 + \tilde{\epsilon},$$

where step ① uses Inequality (62); step ② uses the fact that $\frac{\overline{\mathrm{A}} + \theta_1}{\underline{\mathrm{A}} + \theta_1} \leq 1 + \tilde{\epsilon}$ if $\theta_1 \geq \frac{\overline{\mathrm{A}} - \underline{\mathrm{A}}(1 + \tilde{\epsilon})}{\tilde{\epsilon}}$, and $\frac{\overline{\mathrm{V}} + \theta_2}{\underline{\mathrm{V}} + \theta_2} \leq 1 + \tilde{\epsilon}$ if $\theta_2 \geq \frac{\overline{\mathrm{V}} - \underline{\mathrm{V}}(1 + \tilde{\epsilon})}{\tilde{\epsilon}}$.

□

C.9. Proof of Theorem 4.13

Proof. We consider constant stepsizes with $\mu^t = \bar{\mu}$ for all $t \geq 1$.

We define $\mathbf{H} \triangleq (\mathbf{A}^\mathsf{T}\mathbf{A} + \theta_1 \mathbf{I}_n)/\bar{\mu} + \tilde{\mathbf{M}} + \theta_2 \mathbf{I}_n \in \mathbb{R}^{n\times n}$, and $\mathbf{H}_*^t \triangleq \mathrm{U}_{\mathtt{B}^t}\mathrm{U}_{\mathtt{B}^t}^\mathsf{T}\mathbf{H}\mathrm{U}_{\mathtt{B}^t}\mathrm{U}_{\mathtt{B}^t}^\mathsf{T} \in \mathbb{R}^{n\times n}$.

We let $\underline{\mathrm{H}} \triangleq \frac{\underline{\mathrm{A}}+\theta_1}{\bar{\mu}} + \underline{\mathrm{V}} + \theta_2$, and $\overline{\mathrm{H}} \triangleq \frac{\overline{\mathrm{A}}+\theta_1}{\bar{\mu}} + \overline{\mathrm{V}} + \theta_2$, $\Delta_{\mathbf{x}}^t \triangleq \mathbb{E}_{\xi^t}[\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2]$, $\Delta_F^t \triangleq \mathbb{E}_{\xi^t}[(\min_{i=1}^t F(\mathbf{x}^i)) - F(\bar{\mathbf{x}})]$.

First, using **Part (c)** in Lemma 4.1, we have: $\|\mathbf{r}^t\| \leq L_F$.

Second, using **Part (b)** of Lemma 4.2, we have the upper bound of Υ^t that $\forall t,\ \Upsilon^t \leq \frac{1}{2}\bar{\mu}L_h^2 - \Delta_F^t$.

Third, with $\mathtt{B}^t$ and $\mathtt{B}^{t+1}$ randomly and uniformly chosen, for any $\mathbf{z} \in \mathbb{R}^n$, the following holds:

$$\begin{aligned} \mathbb{E}_{\mathtt{B}^t}[\|\mathbf{z}\|_{\mathbf{H}_*^t}^2] &= \mathbb{E}_{\mathtt{B}^t}[\mathbf{z}^\mathsf{T}[\mathbf{H}_*^t]\mathbf{z}] = \mathbb{E}_{\mathtt{B}^t}[\mathbf{z}^\mathsf{T}\mathrm{U}_{\mathtt{B}^t}\mathrm{U}_{\mathtt{B}^t}^\mathsf{T}\mathbf{H}\mathrm{U}_{\mathtt{B}^t}\mathrm{U}_{\mathtt{B}^t}^\mathsf{T}]\mathbf{z} \\ &= \mathbb{E}_{\mathtt{B}^{t+1}}[\mathbf{z}^\mathsf{T}[\mathrm{U}_{\mathtt{B}^{t+1}}\mathrm{U}_{\mathtt{B}^{t+1}}^\mathsf{T}\mathbf{H}\mathrm{U}_{\mathtt{B}^{t+1}}\mathrm{U}_{\mathtt{B}^{t+1}}^\mathsf{T}]\mathbf{z}] \\ &= \mathbb{E}_{\mathtt{B}^{t+1}}[\|\mathbf{z}\|_{\mathbf{H}_*^{t+1}}^2]. \end{aligned} \tag{63}$$

(a) Building upon our prior discussions, we derive the following inequalities:

$$\begin{aligned} & \mathbb{E}_{\xi^{t+1}}[\tfrac{1}{2}\|\mathbf{x}^{t+1} - \bar{\mathbf{x}}\|_{\mathbf{H}_*^{t+1}}^2] \\ \overset{①}{=} \;& \mathbb{E}_{\xi^t}[\tfrac{1}{2}\|\mathbf{x}^{t+1} - \bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2] \\ \overset{②}{\leq} \;& \mathbb{E}_{\xi^t}[\tfrac{1}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2] - Z_n^k \tfrac{V_s}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2 + Z_n^k\{\tfrac{2}{\underline{\mathrm{H}}^t}\|\mathbf{r}^t\|_2^2 + \Upsilon^t + (1+2\kappa^t)\|\mathbf{r}^t\|\|\bar{\mathbf{x}}\|\} \\ \overset{③}{\leq} \;& \mathbb{E}_{\xi^t}[(1 - \tfrac{V_s}{\overline{\mathrm{H}}}) \cdot \tfrac{1}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2] + Z_n^k\{\tfrac{2}{\underline{\mathrm{H}}^t}(L_F)^2 + \tfrac{\bar{\mu}L_h^2}{2} - \Delta_F^t + (3+2\tilde{\epsilon})L_F\|\bar{\mathbf{x}}\|\}, \end{aligned} \tag{64}$$

where step ① uses Equality (63) with $\mathbf{z} = \mathbf{x}^{t+1} - \bar{\mathbf{x}}$, leading to $\mathbb{E}_{\xi^t}[\|\mathbf{x}^{t+1} - \bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2] = \mathbb{E}_{\xi^{t+1}}[\|\mathbf{x}^{t+1} - \bar{\mathbf{x}}\|_{\mathbf{H}_*^{t+1}}^2]$; step ② uses the inequality in **Part (b)** of Lemma (4.12); step ③ uses $Z_n^k \frac{V_s}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_2^2 \geq \mathbb{E}_{\xi^t}[\frac{V_s}{2\overline{\mathrm{H}}}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2]$.

(b) Based on (64), we apply Lemma A.2 with the following definitions:

$$\gamma \triangleq 1 - \tfrac{V_s}{\overline{\mathrm{H}}},\ \Phi^t \triangleq \mathbb{E}_{\xi^t}[\tfrac{1}{2}\|\mathbf{x}^t - \bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2],\ \Lambda^t \triangleq Z_n^k\{\tfrac{2}{\underline{\mathrm{H}}^t}(L_F)^2 + \tfrac{\bar{\mu}}{2}L_h^2 - \Delta_F^t + (3+2\tilde{\epsilon})L_F\|\bar{\mathbf{x}}\|\}. \tag{65}$$

This results in the subsequent inequality for any integer $T \geq 1$:

$$\begin{aligned} & \mathbb{E}_{\xi^{T+1}}[\tfrac{1}{2}\|\mathbf{x}^{T+1} - \bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2] \\ \leq \;& \gamma^T \mathbb{E}_{\xi^1}[\frac{1}{2}\|\mathbf{x}^1 - \bar{\mathbf{x}}\|_{\mathbf{H}_*^1}^2] + \tfrac{Z_n^k}{1-\gamma}\max_{t=1}^T\{\tfrac{2}{\underline{\mathrm{H}}}(L_F)^2 + \tfrac{\bar{\mu}}{2}L_h^2 - \Delta_F^t + (3+2\tilde{\epsilon})L_F\|\bar{\mathbf{x}}\|\} \end{aligned} \tag{66}$$

We further derive the following inequalities:

$$\begin{aligned} & \mathbb{E}_{\xi^{T+1}}[\tfrac{1}{2}\|\mathbf{x}^{T+1} - \bar{\mathbf{x}}\|_2^2] \\ \overset{①}{\leq} \;& \tfrac{1}{\underline{\mathrm{H}}Z_n^k}\mathbb{E}[\tfrac{1}{2}\|\mathbf{x}^{T+1} - \bar{\mathbf{x}}\|_{\mathbf{H}_*^{T+1}}^2] \\ \overset{②}{\leq} \;& \tfrac{\gamma^T}{\underline{\mathrm{H}}Z_n^k} \cdot \mathbb{E}_{\xi^1}[\tfrac{1}{2}\|\mathbf{x}^1 - \bar{\mathbf{x}}\|_{\mathbf{H}_*^1}^2] + \tfrac{1}{\underline{\mathrm{H}}(1-\gamma)}\max_{t=1}^T\{\tfrac{2(L_F)^2}{\underline{\mathrm{H}}^t} + \tfrac{1}{2}\bar{\mu}L_h^2 - \Delta_F^t + (3+2\tilde{\epsilon})L_F\|\bar{\mathbf{x}}\|\} \\ \overset{③}{\leq} \;& \tfrac{\gamma^T\overline{\mathrm{H}}}{\underline{\mathrm{H}}} \cdot \mathbb{E}_{\xi^1}[\tfrac{1}{2}\|\mathbf{x}^1 - \bar{\mathbf{x}}\|_2^2] + \tfrac{1}{\underline{\mathrm{H}}(1-\gamma)} \cdot \max_{t=1}^T\{\tfrac{2(L_F)^2}{\underline{\mathrm{H}}^t} + \tfrac{1}{2}\bar{\mu}L_h^2 - \Delta_F^t + (3+2\tilde{\epsilon})L_F\|\bar{\mathbf{x}}\|\} \\ \overset{④}{\leq} \;& \gamma^T \cdot (1+\tilde{\epsilon}) \cdot \mathbb{E}_{\xi^1}[\tfrac{1}{2}\|\mathbf{x}^1 - \bar{\mathbf{x}}\|_2^2] + \tfrac{1}{V_s}\{2(L_F)^2 \cdot \tfrac{\bar{\mu}}{\overline{\mathrm{A}}+\theta_1} + \tfrac{1}{2}\bar{\mu}L_h^2 - \Delta_F^T + (3+2\tilde{\epsilon})L_F\|\bar{\mathbf{x}}\|\} \\ \overset{⑤}{=} \;& \tfrac{1}{V_s}\{K_3\gamma^T + D_3\bar{\mu} - \Delta_F^T + (3+2\tilde{\epsilon})L_F\|\bar{\mathbf{x}}\|\}, \end{aligned} \tag{67}$$

where step ① uses $\underline{\mathrm{H}}Z_n^k\mathbb{E}_{\xi^{T+1}}[\|\mathbf{x}^{T+1} - \bar{\mathbf{x}}\|_2^2] \leq \mathbb{E}_{\xi^{T+1}}[\|\mathbf{x}^{T+1} - \bar{\mathbf{x}}\|_{\mathbf{H}_*^{T+1}}^2]$; step ② uses Inequality (66); step ③ uses $\mathbb{E}_{\xi^1}[\|\mathbf{x}^1 - \bar{\mathbf{x}}\|_{\mathbf{H}_*^1}^2] \leq Z_n^k\overline{\mathrm{H}}\mathbb{E}_{\xi^1}[\|\mathbf{x}^1 - \bar{\mathbf{x}}\|_2^2]$; step ④ uses $\frac{\overline{\mathrm{H}}}{\underline{\mathrm{H}}} = \kappa \leq 1 + \tilde{\epsilon}$, $\gamma \triangleq 1 - \frac{V_s}{\overline{\mathrm{H}}}$, and $\overline{\mathrm{H}} \triangleq \frac{\underline{\mathrm{A}}+\theta_1}{\bar{\mu}} + \underline{\mathrm{V}} + \theta_2 \geq \frac{\overline{\mathrm{A}}+\theta_1}{\bar{\mu}}$; step ⑤ uses $K_3 \triangleq \mathbb{E}_{\xi^1}[(1+\tilde{\epsilon}) \cdot \frac{V_s}{2}\|\mathbf{x}^1 - \bar{\mathbf{x}}\|_2^2] = \frac{V_s(1+\tilde{\epsilon})}{2}\Delta_{\mathbf{x}}^1$ and $D_3 \triangleq \frac{2(L_F)^2}{\theta_1 + \overline{\mathrm{A}}} + \frac{L_h^2}{2}$.

We now focus on Inequality (67). Using the fact that $\mathbb{E}_{\xi^{T+1}}[\frac{1}{2}\|\mathbf{x}^{T+1}-\bar{\mathbf{x}}\|_2^2] \geq 0$, we obtain: $\Delta_F^T \leq K_3\gamma^T + D_3\bar{\mu} + (3+2\tilde{\epsilon})L_F\|\bar{\mathbf{x}}\|$.

Using the fact that $\Delta_F^T \geq 0$, we have: $\Delta_{\mathbf{x}}^{T+1} \leq \frac{2}{V_s}\left(K_3\gamma^T + D_3\bar{\mu} + (3+2\tilde{\epsilon})L_F\|\bar{\mathbf{x}}\|\right)$.

□

C.10. Proof of Theorem 4.14

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

Lemma C.1. *Assume $\mu^t = \frac{\eta}{t+t_0}$ with $\eta = \frac{\overline{\mathrm{A}}+\theta_1}{V_s}$. We have:*

$$\mathbb{E}_{\xi^{t+1}}[\tfrac{1}{2}\|\mathbf{x}^{t+1}-\bar{\mathbf{x}}\|_{\mathbf{H}_*^{t+1}}^2] \leq \mathbb{E}_{\xi^t}[\tfrac{1}{2}\|\mathbf{x}^{t+1}-\bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2] + \tfrac{V_s Z_n^k}{2}\|\mathbf{x}^{t+1}-\bar{\mathbf{x}}\|_2^2.$$

Proof. We denote $\mathbf{H}_*^t \triangleq \mathrm{U}_{\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\mathsf{T}\mathbf{H}^t\mathrm{U}_{\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\mathsf{T} \in \mathbb{R}^{n\times n}$, where $\mathbf{H}^t \triangleq (\mathbf{A}^\mathsf{T}\mathbf{A}+\theta_1\mathbf{I}_n)/\mu^t + \tilde{\mathbf{M}} + \theta_2\mathbf{I}_n \in \mathbb{R}^{n\times n}$.

We have the following inequalities for all $\mathbf{z} \triangleq \mathbf{x}^{t+1}-\bar{\mathbf{x}} \in \mathbb{R}^n$:

$$\begin{aligned}
& \mathbb{E}_{\xi^{t+1}}[\|\mathbf{z}\|_{\mathbf{H}_*^{t+1}}^2] - \mathbb{E}_{\xi^t}[\|\mathbf{z}\|_{\mathbf{H}_*^t}^2] \\
\overset{\textcircled{1}}{=} \; & \mathbb{E}_{\xi^{t+1}}[\mathbf{z}^\mathsf{T}(\mathrm{U}_{\mathrm{B}^{t+1}}\mathrm{U}_{\mathrm{B}^{t+1}}^\mathsf{T}\mathbf{H}^{t+1}\mathrm{U}_{\mathrm{B}^{t+1}}\mathrm{U}_{\mathrm{B}^{t+1}}^\mathsf{T})\mathbf{z}] - \mathbb{E}_{\xi^t}[\mathbf{z}^\mathsf{T}(\mathrm{U}_{\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\mathsf{T}\mathbf{H}^t\mathrm{U}_{\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\mathsf{T})\mathbf{z}] \\
\overset{\textcircled{2}}{=} \; & \mathbb{E}_{\xi^t}[\mathbf{z}^\mathsf{T}(\mathrm{U}_{\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\mathsf{T}\mathbf{H}^{t+1}\mathrm{U}_{\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\mathsf{T})\mathbf{z}] - \mathbb{E}_{\xi^t}[\mathbf{z}^\mathsf{T}(\mathrm{U}_{\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\mathsf{T}\mathbf{H}^t\mathrm{U}_{\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\mathsf{T})\mathbf{z}] \\
= \; & \mathbb{E}_{\xi^t}[\mathbf{z}^\mathsf{T}(\mathrm{U}_{\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\mathsf{T}[\mathbf{H}^{t+1}-\mathbf{H}^t]\mathrm{U}_{\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\mathsf{T})\mathbf{z}] \\
\overset{\textcircled{3}}{=} \; & (\tfrac{1}{\mu^{t+1}} - \tfrac{1}{\mu^t})\mathbb{E}_{\xi^t}[\mathbf{z}^\mathsf{T}(\mathrm{U}_{\mathrm{B}^t}[\mathbf{A}^\mathsf{T}\mathbf{A}+\theta_1\mathbf{I}_n]_{\mathrm{B}^t\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\mathsf{T})\mathbf{z}] \\
\overset{\textcircled{4}}{\leq} \; & \tfrac{1}{\eta}\mathbb{E}_{\xi^t}[\mathbf{z}^\mathsf{T}\mathrm{U}_{\mathrm{B}^t}\left(\overline{\mathrm{A}}\mathbf{I}_k+\theta_1\mathbf{I}_k\right)\mathrm{U}_{\mathrm{B}^t}^\mathsf{T}\mathbf{z}] = \tfrac{\overline{\mathrm{A}}+\theta_1}{\eta}\mathbb{E}_{\mathrm{B}^t}[\mathbf{z}^\mathsf{T}\mathrm{U}_{\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\mathsf{T}\mathbf{z}] \\
\overset{\textcircled{5}}{=} \; & Z_n^k\tfrac{\overline{\mathrm{A}}+\theta_1}{\eta}\|\mathbf{z}\|_2^2 \\
\overset{\textcircled{6}}{=} \; & Z_n^k V_s\|\mathbf{z}\|_2^2,
\end{aligned}$$

where step ① uses the definition of $\mathbf{H}_*^t$; step ② uses the fact that both B^t and B^{t+1} are choosen randomly and uniformly; step ③ uses the choice $\mu^t = \frac{\eta}{t+t_0}$ that $\frac{1}{\mu^{t+1}} - \frac{1}{\mu^t} = \frac{1}{\eta}\cdot((t+t_0+1)-(t+t_0)) = \frac{1}{\eta}$; step ④ uses $[\mathbf{A}^\mathsf{T}\mathbf{A}]_{\mathrm{B}^t\mathrm{B}^t} \preceq \overline{\mathrm{A}}\mathbf{I}_k$; step ⑤ uses Lemma 4.8; step ⑥ uses the choice $\eta = \frac{\overline{\mathrm{V}}+\theta_1}{V_s}$.

□

We now prove the proof of this theorem.

Proof. We consider diminishing stepsizes with $\mu^t = \frac{\eta}{t+t_0}$ for all $t \geq 1$, where $\eta = \frac{\overline{\mathrm{A}}+\theta_1}{V_s}$.

We define $\mathbf{H}^t \triangleq (\mathbf{A}^\mathsf{T}\mathbf{A}+\theta_1\mathbf{I}_n)/\mu^t + \tilde{\mathbf{M}} + \theta_2\mathbf{I}_n \in \mathbb{R}^{n\times n}$, $\mathbf{H}_*^t \triangleq \mathrm{U}_{\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\mathsf{T}\mathbf{H}^t\mathrm{U}_{\mathrm{B}^t}\mathrm{U}_{\mathrm{B}^t}^\mathsf{T} \in \mathbb{R}^{n\times n}$.

We let $\underline{\mathrm{H}}^t \triangleq \frac{\underline{\mathrm{A}}+\theta_1}{\mu^t} + \underline{\mathrm{V}} + \theta_2$, and $\overline{\mathrm{H}}^t \triangleq \frac{\overline{\mathrm{A}}+\theta_1}{\mu^t} + \overline{\mathrm{V}} + \theta_2$.

We let $\Delta_{\mathbf{x}}^t \triangleq \mathbb{E}_{\xi^t}[\|\mathbf{x}^t-\bar{\mathbf{x}}\|_2^2]$, $\Delta_F^t \triangleq \mathbb{E}_{\xi^t}[(\min_{i=1}^t F(\mathbf{x}^i)) - F(\bar{\mathbf{x}})]$, and $\Phi^t \triangleq \mathbb{E}_{\xi^t}[\frac{1}{2}\|\mathbf{x}^t-\bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2] - Z_n^k V_s\frac{1}{2}\|\mathbf{x}^t-\bar{\mathbf{x}}\|_2^2$.

First, using **Part (c)** in Lemma 4.1, we have: $\|\mathbf{r}^t\| \leq L_F'$.

Second, using Lemma 4.2, we have: $\sum_{t=1}^T \Upsilon^t \leq C_\Upsilon\left(1+\ln(T)\right) - T\Delta_F^T$ for any $T \geq 1$.

Third, using the definition of $\underline{\mathrm{H}}^t$, we have:

$$\underline{\mathrm{H}}^t \triangleq \frac{\underline{\mathrm{A}}+\theta_1}{\mu^t} + \underline{\mathrm{V}} + \theta_2 \geq \frac{\underline{\mathrm{A}}+\theta_1}{\mu^t} + V_s = \frac{(\underline{\mathrm{A}}+\theta_1)(t+t_0)}{\eta} + V_s = V_s(t+t_0) + V_s \geq V_s(t+1). \tag{68}$$

Fourth, we establish the upper bound for $(-\Phi^{T+1}+\Phi^1)$ using the following inequalities:

$$\begin{aligned}
& -\Phi^{T+1}+\Phi^1 \\
= \;& -\{\mathbb{E}_{\xi^{T+1}}[\tfrac{1}{2}\|\mathbf{x}^{T+1}-\bar{\mathbf{x}}\|_{\mathbf{H}_*^{T+1}}^2] - Z_n^k\tfrac{V_s}{2}\|\mathbf{x}^{T+1}-\bar{\mathbf{x}}\|_2^2\} \\
& +\{\mathbb{E}_{\xi^1}[\tfrac{1}{2}\|\mathbf{x}^1-\bar{\mathbf{x}}\|_{\mathbf{H}_*^1}^2] - Z_n^k\tfrac{V_s}{2}\|\mathbf{x}^1-\bar{\mathbf{x}}\|_2^2\} \\
\overset{\text{①}}{\leq} \;& -\underline{\mathrm{H}}^{T+1}\mathbb{E}_{\xi^{T+1}}[\tfrac{1}{2}\|\mathbf{x}^{T+1}-\bar{\mathbf{x}}\|_2^2] + Z_n^k\tfrac{V_s}{2}\|\mathbf{x}^{T+1}-\bar{\mathbf{x}}\|_2^2 + \overline{\mathrm{H}}^1\mathbb{E}_{\xi^1}[\tfrac{1}{2}\|\mathbf{x}^1-\bar{\mathbf{x}}\|_2^2] \\
\overset{\text{②}}{\leq} \;& -Z_n^k[V_s(T+2)-V_s]\cdot\tfrac{1}{2}\|\mathbf{x}^{T+1}-\bar{\mathbf{x}}\|_2^2 + Z_n^k\overline{\mathrm{H}}^1\tfrac{1}{2}\Delta_{\mathbf{x}}^1 \\
\overset{\text{③}}{\leq} \;& Z_n^k\{-(T+1)\tfrac{V_s}{2}\Delta_{\mathbf{x}}^{T+1}+K_4\},
\end{aligned} \tag{69}$$

where step ① uses $\underline{\mathrm{H}}^t\mathbb{E}_{\xi^t}[\frac{1}{2}\|\mathbf{x}^t-\bar{\mathbf{x}}\|_2^2] \leq \mathbb{E}_{\xi^t}[\frac{1}{2}\|\mathbf{x}^t-\bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2] \leq \overline{\mathrm{H}}^t\mathbb{E}_{\xi^t}[\frac{1}{2}\|\mathbf{x}^t-\bar{\mathbf{x}}\|_2^2]$ for all $t \geq 1$; step ② uses $\mathbb{E}_{\xi^{T+1}}[\frac{1}{2}\|\mathbf{x}^{T+1}-\bar{\mathbf{x}}\|_2^2] = Z_n^k\frac{1}{2}\|\mathbf{x}^{T+1}-\bar{\mathbf{x}}\|_2^2$ and $\underline{\mathrm{H}}^{T+1} \geq (T+2)\frac{V_s}{2}\|\mathbf{x}^{T+1}-\bar{\mathbf{x}}\|_2^2$; step ③ uses the definition of $K_4 \triangleq \frac{1}{2}\overline{\mathrm{H}}^1\Delta_{\mathbf{x}}^1$.

(a) Using the inequality in **Part (c)** of Lemma 4.12, we have:

$$\begin{aligned}
& \mathbb{E}_{\xi^t}[\tfrac{1}{2}\|\mathbf{x}^{t+1}-\bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2] - \mathbb{E}_{\xi^t}[\tfrac{1}{2}\|\mathbf{x}^t-\bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2] + Z_n^k\tfrac{V_s}{2}\|\mathbf{x}^t-\bar{\mathbf{x}}\|_2^2 \\
\leq \;& Z_n^k\Upsilon^t + (2/\underline{\mathrm{H}}^t)Z_n^k\|\mathbf{r}^t\|_2^2 + (1+2\kappa^t)Z_n^k\|\mathbf{r}^t\|\|\bar{\mathbf{x}}\| \\
\overset{\text{①}}{\leq} \;& Z_n^k\Upsilon^t + \tfrac{2Z_n^k(L_F')^2}{V_s t} + (3+2\tilde{\epsilon})Z_n^kL_F'\|\bar{\mathbf{x}}\|,
\end{aligned} \tag{70}$$

where step ① uses $\underline{\mathrm{H}}^t \geq V_s(t+1) > V_s t$ as shown in Inequality (68), $\|\mathbf{r}^t\| \leq L_F'$, and $\kappa^t \leq 1+\tilde{\epsilon}$.

Using the results in Lemma C.1, we have:

$$\mathbb{E}_{\xi^{t+1}}[\tfrac{1}{2}\|\mathbf{x}^{t+1}-\bar{\mathbf{x}}\|_{\mathbf{H}_*^{t+1}}^2] - \mathbb{E}_{\xi^t}[\tfrac{1}{2}\|\mathbf{x}^{t+1}-\bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2] \leq Z_n^k\tfrac{V_s}{2}\|\mathbf{x}^{t+1}-\bar{\mathbf{x}}\|_2^2. \tag{71}$$

We define $\Phi^t \triangleq \mathbb{E}_{\xi^t}[\frac{1}{2}\|\mathbf{x}^t-\bar{\mathbf{x}}\|_{\mathbf{H}_*^t}^2] - Z_n^k\frac{V_s}{2}\|\mathbf{x}^t-\bar{\mathbf{x}}\|_2^2$. Adding Inequalities (70) and (71) together, we have:

$$\Phi^{t+1}-\Phi^t \leq Z_n^k(\tfrac{2(L_F')^2}{V_s}\cdot\tfrac{1}{t} + \Upsilon^t + (3+2\tilde{\epsilon})L_F'\|\bar{\mathbf{x}}\|). \tag{72}$$

(b) Let $T \geq 1$ be any integer. Summing Inequality (72) over t from 1 to T, we have:

$$\begin{aligned}
0 \;\leq\;& -\Phi^{T+1}+\Phi^1+Z_n^k\{(\textstyle\sum_{t=1}^T\frac{1}{t})\cdot\frac{2(L_F')^2}{V_s}+\sum_{t=1}^T\Upsilon^t+T(3+2\tilde{\epsilon})L_F'\|\bar{\mathbf{x}}\|\} \\
\overset{\text{①}}{\leq}\;& -\Phi^{T+1}+\Phi^1+Z_n^k\{(\ln(T)+1)(\tfrac{2(L_F')^2}{V_s}+C_\Upsilon)-T\Delta_F^T+T(3+2\tilde{\epsilon})L_F'\|\bar{\mathbf{x}}\|\} \\
\overset{\text{②}}{=}\;& -\Phi^{T+1}+\Phi^1+Z_n^k\{(\ln(T)+1)D_4-T\Delta_F^T+T(3+2\tilde{\epsilon})L_F'\|\bar{\mathbf{x}}\|\} \\
\overset{\text{③}}{\leq}\;& Z_n^k\{-(T+1)\tfrac{V_s}{2}\Delta_{\mathbf{x}}^{T+1}+K_4+D_4(1+\ln(T))-T\Delta_F^T+T(3+2\tilde{\epsilon})L_F'\|\bar{\mathbf{x}}\|)\},
\end{aligned} \tag{73}$$

where step ① uses $\sum_{t+1}^T\frac{1}{t} \leq \ln(T)+1$ and the upper bound for $\sum_{t=1}^T\Upsilon^t \leq C_\Upsilon\,(1+\ln(T))-T\Delta_F^T$; step ② uses the definition of $D_4 \triangleq \frac{2(L_F')^2}{V_s}+C_\Upsilon$; step ③ uses Inequality (69).

We now focus on Inequality (73). Using the fact that $\Delta_{\mathbf{x}}^{T+1} \geq 0$, we obtain: $\Delta_F^T \leq \frac{K_4}{T}+\frac{D_4\cdot(1+\ln(T))}{T}+(3+2\tilde{\epsilon})L_F'\|\bar{\mathbf{x}}\|$.

Using the fact that $\Delta_F^T \geq 0$, we have: $\Delta_{\mathbf{x}}^{T+1} \leq \left(\frac{K_4}{T+1}+\frac{D_4\cdot(1+\ln(T))}{T+1}+(3+2\tilde{\epsilon})L_F'\|\bar{\mathbf{x}}\|\right)\frac{2}{V_s}$.

□

D. Experiments

This section demonstrates the effectiveness and efficiency of Algorithm 1 on two nonsmooth sparsity constrained optimization tasks, namely the sparsity constrained ℓ_1 regression and sparsity constrained ℓ_∞ regression. Given an arbitrary design matrix

$\mathbf{A} \in \mathbb{R}^{m\times n}$ and an observation vector $\mathbf{b} \in \mathbb{R}^m$, we aim to solve the following optimization problems:

$$\min_{\mathbf{x}} \frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1, \text{ s. t. } \|\mathbf{x}\|_0 \leq s,$$

$$\text{and} \qquad \min_{\mathbf{x}} \frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_\infty, \text{ s. t. } \|\mathbf{x}\|_0 \leq s,$$

where s and λ are given parameters.

► **Datasets**. Following (Yuan et al., 2020a), we examine four types of datasets for the design matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$. **(i)** 'random-m-n': The matrix of size $m \times n$ is generated by sampling from a standard Gaussian distribution. **(ii)** 'e2006-m-n': We select m examples and n dimensions from the original real-world dataset 'e2006', available for download at: `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets`. This dataset contains 16087 examples and 150360 dimensions. **(iii)** 'random-m-n-C': We create a matrix $\mathcal{V}(\mathbf{A}) \in \mathbb{R}^{m\times n}$ to verify the robustness of the algorithms. Here, $\mathcal{V}(\mathbf{A})$ is a noisy version of $\mathbf{A} \in \mathbb{R}^{m\times n}$, with 2% of entries in $\mathbf{A}$ corrupted by scaling the original values by 100 times (Yuan et al., 2020a). **(iv)** 'e2006-m-n-C': We employ the same method to generate corrupted real-world data as used in the 'random-m-n-C' dataset. We generate the observation vector $\mathbf{b}$ in $\mathbb{R}^m$ as follows: a sparse signal $\bar{\mathbf{x}}$ in $\mathbb{R}^n$ is created by randomly selecting a support set of size 100, with values sampled from a standard Gaussian distribution. The observation vector $\mathbf{b}$ is then computed as $\mathbf{b} = \mathbf{A}\bar{\mathbf{x}} + 10 \times \text{randn}(m, 1)$.

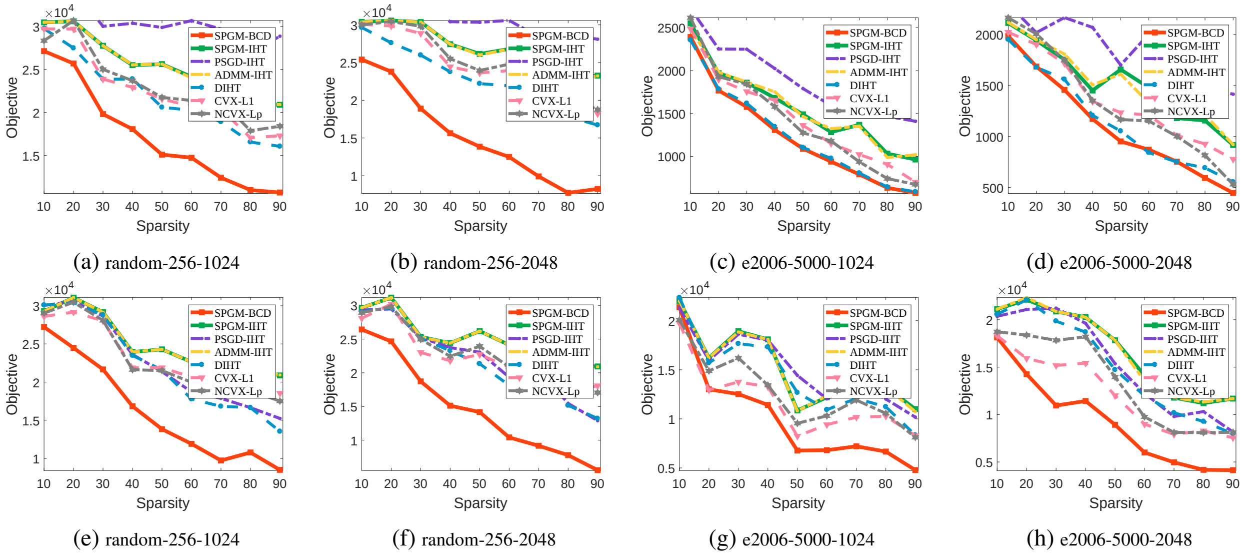
► **Compared Methods**. We compare **SPGM-IHT** and **SPGM-BCD** with 5 state-of-the-art nonsmooth sparsity constrained optimization algorithms: (***i***) Projective Subgradient Descent (PSGD) (Liu et al., 2019), (***ii***) Alternating Direction Method of Mutipliers based on IHT (ADMM-IHT) (He & Yuan, 2012), (***iii***) Dual Iterative Hard Thresholding(DIHT)(Yuan et al., 2020b), (***iv***) Convex ℓ_1 Approximation Method (CVX-ℓ_1) (Candes & Tao, 2005), and (***v***) Nonconvex ℓ_p Approximation Method (NCVX-ℓ_p) (Xu et al., 2012). For CVX-ℓ_1 and NCVX-ℓ_p, we use standard linearized ADMM to solve nonsmooth ℓ_1 norm and $\ell_{1/2}$ norm regularized problems $\min_{\mathbf{x}} F(\mathbf{x}) + \sigma\|\mathbf{x}\|_p^p$ with $p \in \{1, \frac{1}{2}\}$, sweeping the regularization parameter σ over a range or values ($\sigma = \{2^{-9}, 2^{-7}, ..., 2^9\}$). We run these two algorithms for 10 parameters, selecting the solution that leads to the smallest objective after hard thresholding projection and re-optimization over the support set. We employ an efficient closed-form solver to compute the ℓ_p norm proximal operator (Xu et al., 2012).

► **Experimental Settings**. We update the smoothing parameter μ every $K = 10$ iterations by halving it: $\mu \Leftarrow \mu \times \frac{1}{2}$. For **SPGM-BCD**, the random strategy ensures a strong optimality guarantee by maintaining the block-k stationary condition. However, the greedy strategy often yields faster convergence in practice. Therefore, we combine both methods, selecting 8 coordinates using the random strategy and 2 coordinates using the greedy strategy (Yuan et al., 2020a). We keep a record of the relative changes of the objective function values by $d_t = |F(\mathbf{x}^t) - F(\mathbf{x}^{t+1})|/(1 + |F(\mathbf{x}^t)|)$. We let SPGM run up to T iterations and stop it at iteration $t < T$ if $\text{mean}([d_{t-\min(t,v)+1}, d_{t-min(t,v)+2}, ..., d_t]) \leq \epsilon$. We use the default value $(\theta, \epsilon, v, T) = (10^{-3}, 10^{-5}, 100, 1000)$ for SPGM. All code was implemented in Matlab on an Intel 3.20GHz CPU with 8 GB RAM. We assess the quality of the solution by comparing the objective values across different methods. Recognizing that the optimal solution is expected to be sparse, we initialize the solutions for all methods as $10^{-3} \times \text{randn}(n, 1)$ and project them to feasible solutions. We vary $s = \{5, 10, 20, ..., 80, 90\}$ for different datasets and present the average results based on 5 random initial points.

► **Computational Effectiveness**. We demonstrate the computational effectiveness of **SPGM-IHT** and **SPGM-BCD** by comparing them to a set of methods {PSGD-IHT, ADMM-IHT, DIHT, CVX-ℓ_1, NCVX-ℓ_p}. Several observations can be made from Figure 1 and Figure 2. (***i***) DIHT achieves comparable results with **SPGM-BCD** on random-256-1024 and random-256-2048 in the ℓ_1 regression. (***ii***) CVX-ℓ_1 and NCVX-ℓ_p exhibit similar performance, generally outperforming others methods except **SPGM-BCD**. They achieve this by solving the relaxation problem ten times and fine-tuning the hyperparameter σ to obtain k-sparsity solutions. (***iii***) PSGD-IHT generally yields worse results in our experiments. (***iv***) **SPGM-IHT** performs similarly to ADMM-IHT. (***v***) **SPGM-BCD** significantly outperforms most methods due to its ability to find stronger stationary points, which aligns with our theoretical results.

► **Computational Efficiency**. We present runtime comparisons for all the methods on various datasets for solving the sparsity constrained ℓ_1 regression problem. Table 2 displays the average CPU times from three runs. (***i***) The convex and nonconvex relaxation methods are slightly slower than IHT-style methods because they need to run ten times to find the best regularization parameter. (***ii***) The computational efficiency of **SPGM-IHT** is comparable to that of other IHT-style methods since it is itself another IHT-style method. (***iii***) **SPGM-DEC** is slower than the other methods and typically takes about 20 seconds to converge in all instances while achieving better accuracy. (***iv***) Overall, the efficiency of both **SPGM-DEC** and

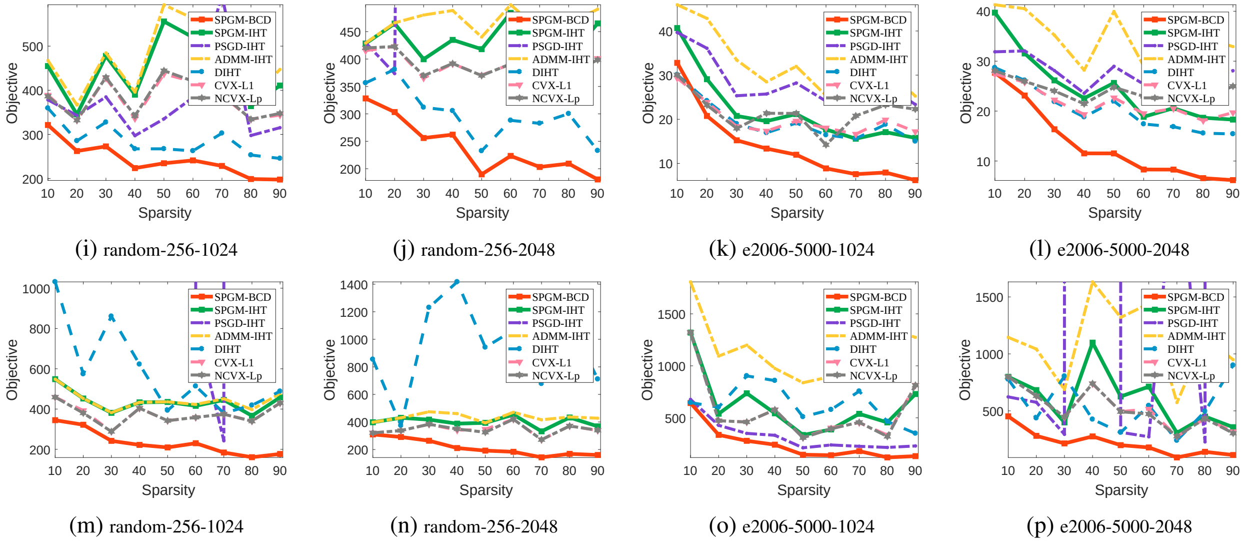
SPGM-IHT is on par with existing methods. This is expected since they are block coordinate descent algorithms.



(a) random-256-1024 (b) random-256-2048 (c) e2006-5000-1024 (d) e2006-5000-2048

(e) random-256-1024 (f) random-256-2048 (g) e2006-5000-1024 (h) e2006-5000-2048

Figure 1: Experimental results on sparsity constrained ℓ_1 regression problems on different datasets with varying the sparsity of the solution.



(i) random-256-1024 (j) random-256-2048 (k) e2006-5000-1024 (l) e2006-5000-2048

(m) random-256-1024 (n) random-256-2048 (o) e2006-5000-1024 (p) e2006-5000-2048

Figure 2: Experimental results on sparsity constrained ℓ_∞ regression problems on different datasets with varying the sparsity of the solution.

	PSGD-IHT	ADMM-IHT	DIHT	CVX-ℓ_1	NCVX-ℓ_p	**SPGM-IHT**	**SPGM-BCD**
random-256-1024	1 ± 1	2 ± 3	1 ± 2	4 ± 1	2 ± 1	2 ± 1	14 ± 3
random-256-2048	1 ± 1	2 ± 1	3 ± 2	3 ± 1	2 ± 1	2 ± 1	15 ± 3
e2006-5000-1024	4 ± 1	2 ± 1	2 ± 1	5 ± 1	4 ± 1	2 ± 1	21 ± 5
e2006-5000-2048	5 ± 1	3 ± 2	3 ± 3	5 ± 2	4 ± 1	2 ± 1	22 ± 5

Table 2: Comparisons of average times (in seconds) of all the methods on different datasets.