Journal of Applied Mathematics and Physics, 2025, 13(4), 1555-1576
https://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Convergence Rate Analysis of Modified BiG-SAM for Solving Bi-Level Optimization Problems Based on S-FISTA

Nishi Xiaoyin*, Lin Yang#

Key Laboratory of Optimization Theory and Applications at China West Normal University of Sichuan Province, School of Mathematics and Information, China West Normal University, Nanchong, China
Email: 1124990720@qq.com, #1540260500@qq.com

How to cite this paper: Xiaoyin, N.S. and Yang, L. (2025) Convergence Rate Analysis of Modified BiG-SAM for Solving Bi-Level Optimization Problems Based on S-FISTA. *Journal of Applied Mathematics and Physics*, **13**, 1555-1576.
https://doi.org/10.4236/jamp.2025.134084

Received: March 29, 2025
Accepted: April 24, 2025
Published: April 27, 2025

Abstract

In this paper, we consider a more general bi-level optimization problem, where the inner objective function is consisted of three convex functions, involving a smooth and two non-smooth functions. The outer objective function is a classical strongly convex function which may not be smooth. Motivated by the smoothing approaches, we modify the classical bi-level gradient sequential averaging method to solve the bi-level optimization problem. Under some mild conditions, we obtain the convergence rate of the generated sequence, and then based on the analysis framework of S-FISTA, we show the global convergence rate of the proposed algorithm.

Keywords

Bi-Level Optimization, Convex Problems, First-Order Methods, Proximal Gradient Method, Sequential Averaging Method, Moreau Envelope

1. Introduction

In this paper, we mainly consider the bi-level optimization problems, which is derived from the Stackelberg game in game theorey. Bi-level optimization problems is a special kind of optimization problem that involves two levels, called outer level and inner level. This structure means that the goals and constraints of the outer level problem depend on the solution of the inner problem. Bi-level optimization problems have a wide range of applications in many fields, including economics, engineering design, transportation planning, machine learning and so on.

*First author.
#Corresponding author.

Recall that the classical bi-level optimization problem, where the outer and inner objective functions are convex. The outer objective function is a constrained minimization problem, it is

$$\min_{\mathbf{x}\in X^*} \omega(\mathbf{x}), \tag{OP}$$

where X^* is the set of minimizers of the inner objective function, which is a composite convex minimization problem, as follows,

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{\Phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}. \tag{P1}$$

In this case, ω is a strong convex and differentiable function, f is a convex and continuously differentiable function and g is an extended real-valued function on $\mathbb{R}^n$. Here, g maybe is a nonsmooth function. Problem (OP)-(P1) is called simple bi-level optimization in [1], which is opposed to the more general version of the problem, see in [2].

Note that, both inner problem (P1)and outer problem (OP) are classical convex optimization problems, which can be solved in different cases, by projected gradient, proximal gradient algorithm, forward-backward splitting algorithm, and so on. However, if we combine problem (OP) and (P1) together, it is difficult to handle.

For problem (OP)-(P1), we can solve it directly or indirectly. In general, we can transform the bi-level optimization problems into a simple optimization problem structure. Then, it can be solved indirectly and easily. The common method for solving the classical bi-level optimization problem is Tikhonov regularization [3], it is for some $\theta > 0$, solving the following regularized problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \{\Phi_\theta(\mathbf{x}) := \Phi(\mathbf{x}) + \theta\omega(\mathbf{x})\}. \tag{1.1}$$

Problem (OP) and (P1) can be traced back to the work of Managsarian and Meyer [4] in the process of developing efficient algorithms for large scale linear programs. They proposed a modification of the Tikhonov regularization technique [3], the underlying idea is called finite-perturbation property. It is that finding a parameter θ^* (Tikhonov perturbation parameter) such that for all $\theta \in [0, \theta^*]$,

$$\arg\min_{\mathbf{x}\in X^*} \omega(\mathbf{x}) = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \Phi_\theta(\mathbf{x}).$$

This property is proven by Managsarian initially, when the inner level problem is a linear program. Then, it was extended by Ferris *et al.* [5], where the inner objective problem is general convex optimization problem.

In [5], they considered the case that g is an indicator function of a closed convex set C, and under some restrictions, they demonstrated that the optimal solution of problem (1.1) is the optimal solution of problem (OP), when there exists a small enough $\theta^* > 0$. In practice, the value of θ^* is unknown, it means that solving problem (1.1) should depend on a sequence of regularizing parameters $\{\theta_n\}$, where $\theta_n \to 0$ as $n \to +\infty$. Solodov [6] showed that for $\sum_{n=1}^{\infty} \theta_n = \infty$,

there is no need to find the optimal solution of problem (1.1) with indicator θ_n. He proposed an explicit and more tractable proximal point method for the bi-level optimization problem (OP)-(P1), which is opposite to the algorithm proposed by Cabot [7], where the approximation scheme is only implicit thus making the method of Cabot [7] not easy to implement. Based on the proximal point algorithm, some researchers developed various proximal point algorithms to solve the problems under different types of framework, see [8] [9].

On the other hand, we can solve the bi-level problem (OP)-(P1) by a direct approach, called hybrid steepest descent method [10], where the sequence converges to the optimal solution according to $\sum_{n=1}^{\infty}\theta_n=\infty$ and $\theta_n\to 0\left(n\to\infty\right)$. Then, the hybrid steepest descent method was further extended by Neto *et al.* for solving a more general outer objective function.

Recently, Beck *et al.* [11] proposed a new direct first order method, which is called Minimal Norm Gradient, it can solve problem (OP). The author proved that in terms of the inner objective function, the algorithm has $\mathcal{O}\left(1/\sqrt{n}\right)$ convergence rate result. However, for some choice of outer objective function ω, the computation of this method is so expensive to get the optimal solution. Motivated by the minimal norm gradient method, Sabach [12] suggested a first order method, called BiG-SAM, to solve the bilevel optimization problem, which is based on existing viscosity approximation methods [13]. According to the convergence analysis of the BiG-SAM, they get $\mathcal{O}\left(1/n\right)$ global convergence rate of in the light of the inner level function. In addition, Yekini *et al.* combined inertial technique with Big-SAM and proposed an inertial BiG-SAM algorithm, more details see in [14].

In this paper, we consider a more general composite convex function as the inner objective function of the bi-level optimization problem. It is,

$$\min_{\mathbf{x}\in\mathbb{R}^n}\left\{H\left(\mathbf{x}\right):=f\left(\mathbf{x}\right)+h\left(\mathbf{x}\right)+g\left(\mathbf{x}\right)\right\}, \tag{P2}$$

where f is a continuously differentiable function with L_f-Lipschitz continuous gradient, h is a real-valued and convex and g is an extended valued function. It is rich enough to cover many interesting generic optimization models by appropriate choices of $\left(f,h,g\right)$. For more details about the assumption of the functions we will give in the following Section 2. Let $\mathbf{x}_{op}^{*}$ is the unique optimal solution of problem (OP).

This paper is organized as follows. In Section 2, we use smooth technique partially smooth inner level problem (P2), construct the inner objective function (Q) and give some useful lemmas for the convergence rate analysis. In Section 3, we introduce a new BiG-SAM algorithm for solving the bi-level optimization problem with outer level (OP). In Section 4, we investigate the convergence rate of BiG-SAM for non-smooth version of bi-level optimization problem.

2. Motivation and Construction

In this section, we will present the motivation and the process of our algorithm

design, as well as the useful lemma. Recall the bi-level optimization problem, where the outer level is problem (OP), the inner level is

$$\min_{\mathbf{x}\in\mathbb{R}^n}\left\{H(\mathbf{x}) := f(\mathbf{x}) + h(\mathbf{x}) + g(\mathbf{x})\right\}, \tag{P2}$$

where f, h and g are satisfy the following Assumption.

Assumption I:

i) $f:\mathbb{R}^n \to \mathbb{R}$ is a convex and continuous differentiable function, it has a Lipschitz continuous gradient with constant L_f, i.e.,

$$\left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right\| \le L_f \left\|\mathbf{x}-\mathbf{y}\right\|, \ \forall \mathbf{x},\mathbf{y}\in\mathbb{R}^n$$

ii) $h:\mathbb{R}^n \to \mathbb{R}$ is a (α,β)-smoothable function, $(\alpha,\beta>0)$. It is that for any $\mu>0$, h_μ denotes a $\frac{1}{\mu}$-smooth approximation of h with parameters (α,β).

iii) $g:\mathbb{R}^n \to (-\infty,\infty]$ is a proper, lower semicontinuous and convex function.

iv) H has bounded level sets. Specifically, for any $\delta>0$, there exists $R_\delta>0$ such that

$$\left\|\mathbf{x}\right\| \le R_\delta \text{ for any } \mathbf{x} \text{ satisfying } H(\mathbf{x}) \le \delta.$$

v) Let X_{P2}^* be the optimal set of problem (P2), and it is nonempty. Set H_{opt} as the optimal value of the problem (P2).

Definition 2.1. [15] A convex function $h:\mathbb{R}^n \to \mathbb{R}$ is called (α,β)-smoothable, $(\alpha,\beta>0)$ if for any $\mu>0$, there exists a convex differentiable function $h:\mathbb{R}^n \to \mathbb{R}$ such that the following holds:

a) $h_\mu(\mathbf{x}) \le h(\mathbf{x}) \le h_\mu(\mathbf{x}) + \beta\mu$ for all $\mathbf{x}\in\mathbb{R}^n$.

b) h_μ is $\frac{\alpha}{\mu}$-smooth.

The function h_μ is called a $\frac{1}{\mu}$-smooth approximation of h with parameters (α,β).

According to the definition of h, combined with the Definition 2.1, we can smooth h as a $1/\mu$-smooth function h_μ. Then, problem (P2) becomes into

$$\min_{\mathbf{x}\in\mathbb{R}^n}\left\{H_\mu(\mathbf{x}) := \underbrace{f(\mathbf{x}) + h_\mu(\mathbf{x})}_{F_\mu(\mathbf{x})} + g(\mathbf{x})\right\}.$$

For convenience, we write the above composite minimization problem as the following form,

$$\min_{\mathbf{x}\in\mathbb{R}^n}\left\{F_\mu(\mathbf{x}) + g(\mathbf{x})\right\}. \tag{Q}$$

Remark 2.1. Here, let X^* be the optimal solution set of problem (Q), which is non-empty and $X^* \subseteq X_{P2}^*$. When μ is small enough, the optimal solution set X^* is equal to $X_{P_2}^*$. This implies that when μ is small enough, the optimal solution of ω over X^* is equivalent to the optimal solution of ω over X_{P2}^*, i.e.,

$$\min_{\mathbf{x}\in X^*}\omega(\mathbf{x})=\min_{\mathbf{x}\in X_{P2}^*}\omega(\mathbf{x}).$$

Observe that problem (P2) is a non-smooth composite function, involving two non-smooth functions. A common methodology for solving non-smooth optimization problems is to replace the original problem by a sequence of approximating smooth problems, and then using direct and classical methods [16] to solve. The main idea is to transform the nondifferentiable problem into a smooth problem, there are many different smoothing approaches to various classes of non-smooth optimization problem, see in [17]-[19]. Motivated by the work of Beck *et al.* [15], we consider to partially smooth the inner objective function (P2) and transform it into a classical structure of convex optimization problem. The motivation of this approach is twofold. Firstly, according to the design and algorithmic analysis of the related schemes, it comes from the classical composite optimization problem formula, like (P1) where f is smooth and g is nonsmooth, can be solved by gradient-based algorithms, [20] [21]. Second, in many applications [22] [23], one of the non-smooth terms in (Q), plays a key role in describing a desirable property of the decision variable $\mathbf{x}$. If we smooth all non-smooth functions in (P2), it will destroy the property of $\mathbf{x}$.

Since h_μ is $\frac{\alpha}{\mu}$-smooth, it has $\frac{\alpha}{\mu}$-Lipschitz continuous gradient ∇h_u. Due to $F_\mu=f+h_\mu$ and f is also have L_f-Lipschitz continuous gradient, it implies the F_μ is a continuous differentiable convex function, the Lipschitz constant of the gradient is equal to $L_f+\frac{\alpha}{\mu}$. Thus, problem (Q) can be solved by the classical proximal gradient (PG) method or proximal forward-backward method, the iteration is as follow:

$$\mathbf{x}^{n+1}=\operatorname{prox}_{\lambda g}\left(\mathbf{x}^n-\lambda\nabla F_\mu\left(\mathbf{x}^n\right)\right),\ n\in\mathbb{N}, \tag{2.1}$$

where the stepsize is $\lambda=1\Big/\left(L_f+\frac{\alpha}{\mu}\right)$. Since g is a proper, lower semicontinuous and convex function, $\operatorname{prox}_{\lambda g}$ is called Moreau Proximal Mapping, which is defined as follow:

$$\operatorname{prox}_{\lambda g}(\mathbf{x}):=\arg\min_{\mathbf{u}\in\mathbb{R}^n}\left\{g(\mathbf{u})+\frac{1}{2\lambda}\|\mathbf{u}-\mathbf{x}\|^2\right\}. \tag{2.2}$$

In addition, the PG method (1) can be regarded as a fixed-point algorithm, it can be formulated as

$$T_\lambda(\mathbf{x}):=\operatorname{prox}_{\lambda g}\left(\mathbf{x}-\lambda\nabla F_\mu(\mathbf{x})\right), \tag{2.3}$$

it is called the **prox-grad** mapping (proximal-gradient mapping). Denote $\operatorname{Fix}(T_\lambda):=\left\{\mathbf{x}\in\mathbb{R}^n\mid T_\lambda(x)=x\right\}$, it is the fixed point set of T_λ. From [24] and [22], we have the following two crucial properties.

Lemma 2.1. [12]

i) The prox-grad mapping T_λ is nonexpansive for all $\lambda \in \left(0, 1/\left(L_f + \alpha/\mu\right)\right]$, that is,

$$T_\lambda(\mathbf{x}) - T_\lambda(\mathbf{y}) \le \|\mathbf{x} - \mathbf{y}\|, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \tag{2.4}$$

ii) Fixed points of the prox-grad mapping T_λ are optimal solutions of problem (Q) and the reverse is also true, i.e.,

$$\mathbf{x} \in X^* \ \Leftrightarrow \ \mathbf{x} = T_\lambda(\mathbf{x}) = \operatorname{prox}_{\lambda g}\left(\mathbf{x} - \lambda \nabla F_\mu(\mathbf{x})\right) \tag{2.5}$$

Therefore, we have that $\operatorname{Fix}(T_\lambda) = X^*$ for all $\lambda > 0$.

Now, we give a key proposition, which is a significant result in convergence rate analysis. Indeed, we consider the following quadratic approximation of $H_\mu(\mathbf{x}) := f(\mathbf{x}) + h_\mu(\mathbf{x}) + g(\mathbf{x})$ at $\mathbf{y}$, it is:

$$Q_\lambda(\mathbf{x}, \mathbf{y}) := f(\mathbf{y}) + h_\mu(\mathbf{y}) + \left\langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) + \nabla h_\mu(\mathbf{y}) \right\rangle + \frac{1}{2\lambda}\|\mathbf{x} - \mathbf{y}\|^2 + g(\mathbf{x}),$$

which admits a unique minimizer

$$p_\lambda(\mathbf{y}) := \arg\min\left\{Q_\lambda(\mathbf{x}, \mathbf{y}) : \mathbf{x} \in \mathbb{R}^n\right\}.$$

It implies that we have,

$$p_\lambda(\mathbf{y}) = \arg\min_{\mathbf{x} \in \mathbb{R}^n}\left\{g(\mathbf{x}) + \frac{1}{2\lambda}\mathbf{x} - \left(\mathbf{y} - \lambda \nabla f(\mathbf{y}) - \lambda \nabla h_\mu(\mathbf{y})\right)^2\right\}.$$

From the characterize of the optimality of $p_\lambda(\cdot)$, we have the following lemma.

Lemma 2.2. For any $\mathbf{y} \in \mathbb{R}^n$, one has $\mathbf{z} = p_\lambda(\mathbf{y})$ if and only if there exists $\gamma(\mathbf{y}) \in \partial g(\mathbf{z})$, the subdifferential of $g(\cdot)$, such that

$$\nabla f(\mathbf{y}) + \nabla h_\mu(\mathbf{y}) + \frac{1}{\lambda}(\mathbf{z} - \mathbf{y}) + \gamma(\mathbf{y}) = 0. \tag{2.6}$$

Then we have the following proposition.

Proposition 2.1. Suppose that **Assumption I** holds true. Let $\mathbf{y} \in \mathbb{R}^n$ and denote $p_\lambda(\mathbf{y}) = T_\lambda(\mathbf{y})$, such that

$$H_\mu(p_\lambda(\mathbf{y})) \le Q_\lambda(p_\lambda(\mathbf{y}), \mathbf{y}). \tag{2.7}$$

Then, for any $\lambda \le 1/\left(L_f + \alpha/\mu\right)$ and $x \in \mathbb{R}^n$, we have

$$H_\mu(\mathbf{x}) - H_\mu(p_\lambda(\mathbf{y})) \ge \frac{1}{2\lambda}\|p_\lambda(\mathbf{y}) - \mathbf{y}\|^2 + \frac{1}{\lambda}\langle \mathbf{y} - \mathbf{x}, p_\lambda(\mathbf{y}) - \mathbf{y}\rangle. \tag{2.8}$$

Proof. From (2.7), we have,

$$H_\mu(\mathbf{x}) - H_\mu(p_\lambda(\mathbf{y})) \ge H_\mu(\mathbf{x}) - Q_\lambda(p_\lambda(\mathbf{y}), \mathbf{y}). \tag{2.9}$$

Since f, h, and g are convex, it implies

$$f(\mathbf{x}) \ge f(\mathbf{y}) + \left\langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) \right\rangle,$$

$$h_\mu(\mathbf{x}) \ge h_\mu(\mathbf{y}) + \left\langle \mathbf{x} - \mathbf{y}, \nabla h_\mu(\mathbf{y}) \right\rangle,$$

$$g(\mathbf{x}) \ge g(p_\lambda(\mathbf{y})) + \left\langle \mathbf{x} - p_\lambda(\mathbf{y}), \gamma(\mathbf{y}) \right\rangle,$$

where the $\gamma(\mathbf{y})$ is defined from lemma 2.2. Now, Summing the above inequalities together, we have

$$H_\mu(\mathbf{x}) \geq f(\mathbf{y}) + h_\mu(\mathbf{y}) + \left\langle \mathbf{x}-\mathbf{y}, \nabla f(\mathbf{y}) + \nabla h_\mu(\mathbf{y}) \right\rangle + g\left(p_\lambda(\mathbf{y})\right) + \left\langle \mathbf{x} - p_\lambda(\mathbf{y}), \gamma(\mathbf{y}) \right\rangle. \tag{2.10}$$

On the other hand, from the definition of $Q_\lambda(\mathbf{x},\mathbf{y})$, let $\mathbf{x} := p_\lambda(\mathbf{y})$, we have

$$Q_\lambda\left(p_\lambda(\mathbf{y}),\mathbf{y}\right) = f(\mathbf{y}) + h_\mu(\mathbf{y}) + \left\langle p_\lambda(\mathbf{y}) - \mathbf{y}, \nabla f(\mathbf{y}) + \nabla h_\mu(\mathbf{y}) \right\rangle + \frac{1}{2\lambda}\left\| p_\lambda(\mathbf{y}) - \mathbf{y} \right\|^2 + g\left(p_\lambda(\mathbf{y})\right). \tag{2.11}$$

Now, combine (9) with (10) and (11), it follows that

$$\begin{aligned} H_\mu(\mathbf{x}) - H_\mu\left(p_\lambda(\mathbf{y})\right) &\geq H_\mu(\mathbf{x}) - Q_\lambda\left(p_\lambda(\mathbf{y}),\mathbf{y}\right) \\ &\geq \left\langle \mathbf{x} - p_\lambda(\mathbf{y}), \nabla f(\mathbf{y}) + \nabla h_\mu(\mathbf{y}) + \gamma(\mathbf{y}) \right\rangle - \frac{1}{2\lambda}\left\| p_\lambda(\mathbf{y}) - \mathbf{y} \right\|^2 \\ &= -\frac{1}{2\lambda}\left\| p_\lambda(\mathbf{y}) - \mathbf{y} \right\|^2 + \left\langle \mathbf{x} - p_\lambda(\mathbf{y}), -\frac{1}{\lambda}\left(p_\lambda(\mathbf{y}) - \mathbf{y}\right) \right\rangle \\ &= -\frac{1}{2\lambda}\left\| p_\lambda(\mathbf{y}) - \mathbf{y} \right\|^2 + \frac{1}{\lambda}\left\langle p_\lambda(\mathbf{y}) - \mathbf{y} + \mathbf{y} - \mathbf{x}, p_\lambda(\mathbf{y}) - \mathbf{y} \right\rangle \\ &= -\frac{1}{2\lambda}\left\| p_\lambda(\mathbf{y}) - \mathbf{y} \right\|^2 + \frac{1}{\lambda}\left\| p_\lambda(\mathbf{y}) - \mathbf{y} \right\|^2 + \frac{1}{\lambda}\left\langle \mathbf{y} - \mathbf{x}, p_\lambda(\mathbf{y}) - \mathbf{y} \right\rangle \\ &= \frac{1}{2\lambda}\left\| p_\lambda(\mathbf{y}) - \mathbf{y} \right\|^2 + \frac{1}{\lambda}\left\langle \mathbf{y} - \mathbf{x}, p_\lambda(\mathbf{y}) - \mathbf{y} \right\rangle, \end{aligned}$$

where the first equality is getting from we used (6). Thus, we complete the proof. □

Now, we turn to discuss the details of outer level problem (OP). Recall the formulation of (OP), it is a convex constraint optimization, where X^* is the optimal solution set of problem (Q). In general, we suppose that outer objective function (OP) satisfies the following assumptions.

Assumption II.

i) $\omega: \mathbb{R}^n \to \mathbb{R}$ is σ-strongly convex, $\sigma > 0$.

ii) ω is a continuously differentiable function such that $\nabla\omega$ is Lipschitz continuous with constant $L_\omega > 0$.

Due to ω is differential, we can use the gradient descent method solving the outer level problem (OP). Nevertheless, not all the outer function ω satisfies **Assumption II** (ii), that is, ω is nonsmooth. So, we assume ω satisfies the following property.

Assumption III: $\omega: \mathbb{R}^n \to \mathbb{R}$ is strong convex with parameter $\sigma > 0$ and ℓ_ω-Lipschitz continuous.

In this case, we can depend on the Moreau envelop of ω and solve the outer level problem, which is denoted by $M_{s\omega}$, the formula is as follow:

$$M_{s\omega}(\mathbf{x}) = \min_{\mathbf{u}\in\mathbb{R}^n}\left\{ \omega(\mathbf{u}) + \frac{1}{2s}\left\| \mathbf{u} - \mathbf{x} \right\|^2 \right\}. \tag{2.12}$$

It is well-known that $M_{s\omega}$ is continuously differentiable on $\mathbb{R}^n$ with an $1/s$-Lipschitz continuous gradient, which is given by

$$\nabla M_{s\omega}(\mathbf{x}) = \frac{1}{s}\left(\mathbf{x} - \operatorname{prox}_{s\omega}(\mathbf{x})\right). \tag{2.13}$$

Additionaly, the Moreau envelope has another useful property, that is:

Lemma 2.3. [12] Let $\omega : \mathbb{R}^n \to (-\infty, \infty]$ be a strongly convex function with strong convexity parameter σ and let $s > 0$. Then, the Moreau envelope $M_{s\omega}$ is strongly convex with parameter $\sigma/(1+s\sigma)$.

Definition 2.2. [12] A mapping $S : \mathbb{R}^n \to \mathbb{R}^n$ is said to be η-contraction if there exists $\eta < 1$ such that

$$\|S(\mathbf{x}) - S(\mathbf{y})\| \le \eta \|\mathbf{x} - \mathbf{y}\|, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

When ω satisfies **Assumption II**, the following result is crucial for our derivations.

Lemma 2.4. [12] Suppose that **Assumption II** holds and let I is a identity operator. Then, the mapping $S_s = I - s\nabla\omega$ is a contractive operator, for all $s \le 2/(L_\omega + \sigma)$, that is,

$$\left\|\mathbf{x} - s\nabla\omega(\mathbf{x}) - \left(\mathbf{y} - s\nabla\omega(\mathbf{y})\right)\right\| \le \sqrt{1 - \frac{2s\sigma L_\omega}{\sigma + L_\omega}}\|\mathbf{x} - \mathbf{y}\|, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{2.14}$$

3. BiG-SAM Algorithm for Smooth Bi-Level Optimization

In this section, we will introduce a new BiG-SAM algorithm for solving bi-level optimization. Firstly, we similarly construct a general framwork for the bi-level problem, consisting of inner problem (Q).

3.1. The General Framework

Motivated by Sabach *et al.* [12], our approach is also to use the Sequential Averaging Method (SAM), in which we can handle the fixed point problem, proposed in [13]. Right now, we will analyse how to use it for solving the bi-level optimization problems, which is made up of problem (OP) and (Q). The sequence $\{\mathbf{x}^n\}$ generated by SAM algorithm, converges to a solution of the fixed-point problem [13]. The iteration is

$$\mathbf{x}^n = \alpha_n S(\mathbf{x}^{n-1}) + (1-\alpha_n)T(\mathbf{x}^{n-1}),$$

where $\{\alpha_n\}_{n\in\mathbb{N}}$ is a carefully chosen sequence in $(0,1]$.

The above algorithm, designed in [13], is to find a fixed-point of a nonexpansive operator T, i.e. $x^* \in \operatorname{Fix}(T)$. This point also satisfies a variational inequality:

$$\left\langle \mathbf{x}^* - S(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \right\rangle \ge 0, \ \forall \mathbf{x} \in \operatorname{Fix}(T), \tag{3.1}$$

where S is a contraction mapping. Here, it means that x^* is the "better" fixed-point in $\operatorname{Fix}(T)$. Where $\{\alpha_n\}$ satisfies the following assumption.

Assumption IV. Let $\{\alpha_n\}_{n\in\mathbb{N}}$ be a sequence of real numbers in $(0,1]$ which satisfies $\lim_{n\to\infty}\alpha_n = 0$, $\sum_{n=1}^{\infty}\alpha_n = \infty$ and $\lim_{n\to\infty}\alpha_{n+1}/\alpha_n = 1$.

It should be noted that **Assumption IV** holds true for several choices of sequences $\{\alpha_n\}_{n\in\mathbb{N}}$ which include, for example, $\alpha_n = \alpha/n, n \in \mathbb{N}$ for any choice

of $\alpha \in (0,1]$.

The following lemma summarizes the key results on SAM, as established in ([13], Theorem 3.2), which serve as the foundation for this paper.

Lemma 3.1. [12] Assume that $S:\mathbb{R}^n \to \mathbb{R}^n$ is a η-contraction and that $T:\mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive mapping, for which $\mathrm{Fix}(T) \neq \varnothing$. Let $\{\mathbf{x}^n\}_{n\in\mathbb{N}}$ be the sequence generated by SAM. If **Assumption IV** holds true, then the following assertions are valid.

i) The sequence $\{\mathbf{x}^n\}_{n\in\mathbb{N}}$ is bounded, in particular, for any $\tilde{\mathbf{x}} \in \mathrm{Fix}(T)$ we have, for all $n \in \mathbb{N}$, that

$$\left\|\mathbf{x}^n - \tilde{\mathbf{x}}\right\| \le C_{\tilde{\mathbf{x}}} := \max\left\{\left\|\mathbf{x}^0 - \tilde{\mathbf{x}}\right\|, \frac{1}{1-\eta}\left\|(I-S)\tilde{\mathbf{x}}\right\|\right\}. \tag{3.2}$$

Moreover, for all $n \in \mathbb{N}$, we also have that

$$\left\|T(\mathbf{x}^n) - \tilde{\mathbf{x}}\right\| \le C_{\tilde{\mathbf{x}}} \text{ and } \left\|S(\mathbf{x}^n) - S(\tilde{\mathbf{x}})\right\| \le \eta C_{\tilde{\mathbf{x}}}.$$

ii) The sequence $\{\mathbf{x}^n\}_{n\in\mathbb{N}}$ converges to some $\mathbf{x}^* \in \mathrm{Fix}(T)$.

iii) The limit point $\mathbf{x}^*$ of $\{\mathbf{x}^n\}_{n\in\mathbb{N}}$, which the existence is ensured by (ii), satisfies the following variational inequality

$$\left\langle \mathbf{x}^* - S(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \right\rangle \ge 0, \ \forall \mathbf{x} \in \mathrm{Fix}(T). \tag{3.3}$$

3.2. SAM for Smooth Bi-Level Optimization Problem

From the Section 2, we know that the inner level optimization problem (P2) can be smoothed as problem (Q), it has a same structure of the inner level optimization problem in [12]. Inspired by the works of [12], we can match the outer problem (OP) and the inner problem (Q) with mapping S and T, respectively. Here, we know that,

i) The mapping T and its fixed-point set $\mathrm{Fix}(T)$ are related to problem (Q) with the composite function $H_\mu = F_\mu + g$ and the optimal solution set X^*.

ii) The mapping S is related to problem (OP) and the outer objective function ω.

Thus, we set T as the prox-grad mapping defined in (2.3), that is, for some $\lambda \in \left(0, 1/\left(L_f + \alpha/\mu\right)\right]$ we have

$$T(\mathbf{x}) := T_\lambda(\mathbf{x}) = \mathrm{prox}_{\lambda g}\left(\mathbf{x} - \lambda \nabla F_\mu(\mathbf{x})\right). \tag{3.4}$$

According to Lemma 2.1 and based on the **Assumption I**, it implies that T is nonexpansive and $\mathrm{Fix}(T) = X^*$. Then, from Lemma 2.4 and **Assumption II**, we can construct the η-contraction mapping S as follow:

$$S(\mathbf{x}) := \mathbf{x} - s\nabla\omega(\mathbf{x}), \tag{3.5}$$

where $s \in \left(0, 2/\left(\sigma + L_\omega\right)\right]$, and the contraction parameter is $\eta = \left(1 - 2sL_\omega\sigma/\left(L_\omega + \sigma\right)\right)^{1/2}$.

Similarly, we use the Sequential Averaging Method (SAM) to design a new BiG-SAM algorithm for solving the bi-level optimization problems(Q) and (OP). The iteration is as follow.

New Bi-level Gradient SAM (BiG-SAM)

Input: $\lambda \in \left(0,1/\left(L_f + \alpha/\mu\right)\right]$, $s \in \left(0,2/\left(L_\omega + \sigma\right)\right]$, and $\{\alpha_n\}_{n\in\mathbb{N}}$ satisfying Assumption **IV**.

Initialization: $\mathbf{x}^0 \in \mathbb{R}^n$.

General Step ($n = 1,2,\cdots$):

$$\mathbf{y}^n = \text{prox}_{\lambda g}\left(\mathbf{x}^{n-1} - \lambda \nabla F_\mu\left(\mathbf{x}^{n-1}\right)\right), \tag{3.6}$$

$$\mathbf{z}^n = \mathbf{x}^{n-1} - s\nabla\omega\left(\mathbf{x}^{n-1}\right), \tag{3.7}$$

$$\mathbf{x}^n = \alpha_n \mathbf{z}^n + \left(1-\alpha_n\right)\mathbf{y}^n. \tag{3.8}$$

Due to the new BiG-SAM algorithm is similar to the works in [12], we can get the similar convergence result.

Lemma 3.2. Let $\left\{\mathbf{x}^n\right\}_{n\in\mathbb{N}}$ be a sequence generated by the new BiG-SAM. Suppose that Assumptions **I**, **II** and **IV** hold true. Then, the sequence $\left\{\mathbf{x}^n\right\}_{n\in\mathbb{N}}$ converges to $\mathbf{x}^* \in X^*$ which satisfies

$$\left\langle \nabla\omega\left(\mathbf{x}^*\right), \mathbf{x} - \mathbf{x}^* \right\rangle \geq 0, \ \forall \mathbf{x} \in X^*, \tag{3.9}$$

and therefore, $\mathbf{x}^* = \mathbf{x}^*_{op}$ is the optimal solution of problem (OP).

The proof of lemma 3.2 is similar to the proof Proposition 5 in [12].

3.3. The Global Convergence Rate of BiG-SAM

In this section, we first prove a technical result on the convergence rate of the gap between successive SAM iterations for fixed-point problems, as described in Section 3.1. Then, we use this to derive our main result: a convergence rate for BiG-SAM in terms of the values of the inner objective function.

We first present a technical lemma which will assist us in the rate of convergence proof.

Lemma 3.3. [12] Let $M > 0$. Assume that $\{a_n\}_{n\in\mathbb{N}}$ is a sequence of nonnegative real numbers which satisfy $a_1 \leq M$ and

$$a_{n+1} \leq \left(1-\gamma b_{n+1}\right)a_n + \left(b_n - b_{n+1}\right)c_n, \ n \geq 1,$$

where $\gamma \in (0,1]$, $\{b_n\}_{n\in\mathbb{N}}$ is a sequence which is defined by $b_n = \min\left\{2/(\gamma n),1\right\}$, and $\{c_n\}_{n\in\mathbb{N}}$ is a sequence of real numbers such that $c_n \leq M < \infty$. Then, the sequence $\{a_n\}_{n\in\mathbb{N}}$ satisfies

$$a_n \leq \frac{MJ}{\gamma n}, \ n \geq 1,$$

where $J = \lfloor 2/\gamma \rfloor$.

For simplicity, we denote $\mathbf{y}^n = T\left(\mathbf{x}^{n-1}\right)$ and $\mathbf{z}^n = S\left(\mathbf{x}^{n-1}\right)$ for any $n \in \mathbb{N}$. The convergence analysis rate is divided into two parts, which ultimately lead to the main conclusions of Theorem 3.1 and Theorem 3.2. Lemma 3.4 provides useful inequalities, while Proposition 3.1 shows that by choosing an appropriate sequence $\{\alpha_n\}_{n\in\mathbb{N}}$, the distance between successive elements of $\left\{\mathbf{x}^n\right\}_{n\in\mathbb{N}}$ is

bounded by $\mathcal{O}(1/n)$, and the sequence converges to a fixed-point of T at the same rate.

Lemma 3.4. [12] Assume that $S:\mathbb{R}^n \to \mathbb{R}^n$ is a η-contraction and that $T:\mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive mapping, for which $\mathrm{Fix}(T) \neq \varnothing$. Let $\{\mathbf{x}^n\}_{n\in\mathbb{N}}, \{\mathbf{y}^n\}_{n\in\mathbb{N}}$ and $\{\mathbf{z}^n\}_{n\in\mathbb{N}}$ be sequences generated by SAM. Then, for any $n \geq 1$ and any $\tilde{\mathbf{x}} \in \mathrm{Fix}(T)$, defining $\tilde{\mathbf{z}} = S(\tilde{\mathbf{x}})$ the following inequalities hold true

$$\left\|\mathbf{y}^{n+1} - \mathbf{y}^n\right\| \leq \left\|\mathbf{x}^n - \mathbf{x}^{n-1}\right\|, \tag{3.10}$$

$$\left\|\mathbf{z}^{n+1} - \mathbf{z}^n\right\| \leq \eta\left\|\mathbf{x}^n - \mathbf{x}^{n-1}\right\|, \tag{3.11}$$

$$\left\|\mathbf{y}^n - \tilde{\mathbf{x}}\right\| \leq \left\|\mathbf{x}^{n-1} - \tilde{\mathbf{x}}\right\|, \tag{3.12}$$

$$\left\|\mathbf{z}^n - \tilde{\mathbf{z}}\right\| \leq \eta\left\|\mathbf{x}^{n-1} - \tilde{\mathbf{x}}\right\|. \tag{3.13}$$

Now, we prove the convergence rate of the sequence $\left\{\left\|\mathbf{x}^n - \mathbf{x}^{n-1}\right\|\right\}_{n\in\mathbb{N}}$, where $\{\mathbf{x}^n\}_{n\in\mathbb{N}}$ is generated by SAM and the averaging parameters $\alpha_n, n \in \mathbb{N}$, are chosen as follows.

$$\alpha_n = \min\left\{\frac{2\gamma}{n(1-\eta)}, 1\right\}, \ n \geq 1, \tag{3.14}$$

where $\gamma \in (0,1]$. For simplicity, we prove our results under the assumption that $\gamma = 1$. It is important to note that all the following results remain valid for any γ chosen from the interval $(0,1]$.

Proposition 3.1. Let $\{\mathbf{x}^n\}_{n\in\mathbb{N}}, \{\mathbf{y}^n\}_{n\in\mathbb{N}}$ and $\{\mathbf{z}^n\}_{n\in\mathbb{N}}$ be sequences generated by SAM where $\{\alpha_n\}_{n\in\mathbb{N}}$ is defined by (3.14). Then, for any $\tilde{\mathbf{x}} \in \mathrm{Fix}(T)$, the two sequences $\left\{\left\|\mathbf{x}^n - \mathbf{x}^{n-1}\right\|\right\}_{n\in\mathbb{N}}$ and $\left\{\left\|\mathbf{y}^n - \mathbf{x}^{n-1}\right\|\right\}_{n\in\mathbb{N}}$ converge to $\mathbf{0}$, and the rates of convergence are given by

$$\left\|\mathbf{x}^n - \mathbf{x}^{n-1}\right\| \leq \frac{2C_{\tilde{\mathbf{x}}}J}{(1-\eta)n}, \ n \geq 1, \tag{3.15}$$

and

$$\left\|\mathbf{y}^n - \mathbf{x}^{n-1}\right\| \leq \frac{2C_{\tilde{\mathbf{x}}}(J+2)}{(1-\eta)n}, \ n \geq 1, \tag{3.16}$$

where $C_{\tilde{\mathbf{x}}}$ is defined in (3.2), and $J = \lfloor 2/(1-\eta) \rfloor$.

Proof. From the definitions of $\mathbf{x}^n$ and $\mathbf{x}^{n+1}$, we directly obtain:

$$\begin{aligned}
\left\|\mathbf{x}^{n+1} - \mathbf{x}^n\right\| &= \left\|(1-\alpha_{n+1})\mathbf{y}^{n+1} + \alpha_{n+1}\mathbf{z}^{n+1} - \left((1-\alpha_n)\mathbf{y}^n + \alpha_n\mathbf{z}^n\right)\right\| \\
&= \left\|(1-\alpha_{n+1})\left(\mathbf{y}^{n+1} - \mathbf{y}^n\right) + \alpha_{n+1}\left(\mathbf{z}^{n+1} - \mathbf{z}^n\right) + (\alpha_n - \alpha_{n+1})\left(\mathbf{y}^n - \mathbf{z}^n\right)\right\| \\
&\leq (1-\alpha_{n+1})\left\|\mathbf{y}^{n+1} - \mathbf{y}^n\right\| + \alpha_{n+1}\left\|\mathbf{z}^{n+1} - \mathbf{z}^n\right\| + (\alpha_n - \alpha_{n+1})\left\|\mathbf{y}^n - \mathbf{z}^n\right\| \\
&\leq (1-\alpha_{n+1})\left\|\mathbf{x}^n - \mathbf{x}^{n-1}\right\| + \alpha_{n+1}\eta\left\|\mathbf{x}^n - \mathbf{x}^{n-1}\right\| + (\alpha_n - \alpha_{n+1})\left\|\mathbf{y}^n - \mathbf{z}^n\right\| \\
&= \left(1 - \alpha_{n+1}(1-\eta)\right)\left\|\mathbf{x}^n - \mathbf{x}^{n-1}\right\| + (\alpha_n - \alpha_{n+1})\left\|\mathbf{y}^n - \mathbf{z}^n\right\|,
\end{aligned} \tag{3.17}$$

where the second inequality follows from (3.10) and (3.11). Now, let $\tilde{\mathbf{x}} \in \mathrm{Fix}(T)$

and let $\tilde{\mathbf{z}} = S(\tilde{\mathbf{x}})$, then

$$\begin{aligned}\left\|\mathbf{y}^n - \mathbf{z}^n\right\| &= \left\|\mathbf{y}^n - \tilde{\mathbf{x}} + \tilde{\mathbf{x}} - \tilde{\mathbf{z}} + \tilde{\mathbf{z}} - \mathbf{z}^n\right\| \\ &\le \left\|\mathbf{y}^n - \tilde{\mathbf{x}}\right\| + \left\|\tilde{\mathbf{x}} - \tilde{\mathbf{z}}\right\| + \left\|\tilde{\mathbf{z}} - \mathbf{z}^n\right\| \\ &\le \left\|\mathbf{x}^{n-1} - \tilde{\mathbf{x}}\right\| + \left\|(I - S)\tilde{\mathbf{x}}\right\| + \eta\left\|\mathbf{x}^{n-1} - \tilde{\mathbf{x}}\right\| \\ &\le C_{\tilde{\mathbf{x}}} + (1-\eta)C_{\tilde{\mathbf{x}}} + \eta C_{\tilde{\mathbf{x}}} \\ &= 2C_{\tilde{\mathbf{x}}},\end{aligned} \tag{3.18}$$

where the second inequality follows from (3.12) and (3.13), as well as the definition of $\tilde{\mathbf{z}}$, and the last inequality follows from Lemma 3.1(i). Additionally, we have

$$\begin{aligned}\left\|\mathbf{x}^1 - \mathbf{x}^0\right\| &= \left\|\mathbf{x}^1 - \tilde{\mathbf{x}} + \tilde{\mathbf{x}} - \mathbf{x}^0\right\| \\ &\le \left\|\mathbf{x}^1 - \tilde{\mathbf{x}}\right\| + \left\|\mathbf{x}^0 - \tilde{\mathbf{x}}\right\| \\ &\le 2C_{\tilde{\mathbf{x}}},\end{aligned} \tag{3.19}$$

where the second inequality follows from Lemma 3.1(i). Let $a_n := \left\|\mathbf{x}^n - \mathbf{x}^{n-1}\right\|$, $b_n := \alpha_n, \gamma := 1-\eta$ and $c_n := \left\|\mathbf{y}^n - \mathbf{z}^n\right\|$. Then, it means that (3.17) is equal to

$$\begin{aligned}a_{n+1} = \left\|\mathbf{x}^{n+1} - \mathbf{x}^n\right\| &\le \left(1 - \alpha_{n+1}(1-\eta)\right)\left\|\mathbf{x}^n - \mathbf{x}^{n-1}\right\| + \left(\alpha_n - \alpha_{n+1}\right)\left\|\mathbf{y}^n - \mathbf{z}^n\right\| \\ &= \left(1 - b_{n+1}\gamma\right)a_n + \left(b_n - b_{n+1}\right)c_n.\end{aligned}$$

Note that, $c_n := \left\|\mathbf{y}^n - \mathbf{z}^n\right\|$ and combine it with (3.18), we know that $c_n \le 2C_{\tilde{\mathbf{x}}}$. Now, set $M := 2C_{\tilde{\mathbf{x}}}$. According to Lemma 3.3, we know that

$$a_n \le \frac{MJ}{\gamma n} = \frac{2C_{\tilde{\mathbf{x}}}J}{(1-\eta)n},$$

it means that we have (3.15). Then the convergence rate for $\left\{\left\|\mathbf{y}^n - \mathbf{x}^{n-1}\right\|\right\}_{n\in\mathbb{N}}$ is derived from the following arguments. Recall that $\mathbf{x}^n = \alpha\mathbf{z}^n + (1-\alpha)\mathbf{y}^n$, then

$$\begin{aligned}\left\|\mathbf{y}^n - \mathbf{x}^{n-1}\right\| &= \left\|\mathbf{y}^n - \mathbf{x}^n + \mathbf{x}^n - \mathbf{x}^{n-1}\right\| \\ &\le \left\|\mathbf{y}^n - \mathbf{x}^n\right\| + \left\|\mathbf{x}^n - \mathbf{x}^{n-1}\right\| \\ &= \alpha_n\left\|\mathbf{y}^n - \mathbf{z}^n\right\| + \left\|\mathbf{x}^n - \mathbf{x}^{n-1}\right\| \\ &\le \frac{2}{(1-\eta)n}2C_{\tilde{\mathbf{x}}} + \frac{2C_{\tilde{\mathbf{x}}}J}{(1-\eta)n} \\ &= \frac{2C_{\tilde{\mathbf{x}}}(J+2)}{(1-\eta)n},\end{aligned}$$

where the second inequality is due to the previous result as well as (3.18).

□

It is no hard to see from (3.16) that the sequence generated by BiG-SAM algorithm converges to an optimal solution of the inner problem (Q) with $\mathcal{O}(1/n)$. In the following, we will discuss the convergence an important result that is the convergence of $\left\{H_\mu\left(\mathbf{y}^n\right)\right\}_{n\in\mathbb{N}}$. Why not discuss the convergence of $\left\{H_\mu\left(\mathbf{x}^n\right)\right\}_{n\in\mathbb{N}}$ directly? Because of the domain of the function H_μ may not be feasible for $\left\{H_\mu\left(\mathbf{x}^n\right)\right\}$. However, due to $\left\|\mathbf{y}^n - \mathbf{x}^{n-1}\right\| \to 0$ as $n \to \infty$ and H_μ

is lower semicontinuous, we know that proving convergence of $\left\{H_\mu\left(\mathbf{y}^n\right)\right\}_{n\in\mathbb{N}}$ to the optimal value also means the convergence of $\left\{H_\mu\left(\mathbf{x}^n\right)\right\}_{n\in\mathbb{N}}$ to the same value. The global convergence rate result is as follow.

Theorem 3.1. Let $\left\{\mathbf{x}^n\right\}_{n\in\mathbb{N}}, \left\{\mathbf{y}^n\right\}_{n\in\mathbb{N}}$ and $\left\{\mathbf{z}^n\right\}_{n\in\mathbb{N}}$ be sequences generated by BiG-SAM. Let $\left\{\alpha_n\right\}_{n\in\mathbb{N}}$ be defined by (3.14). Then, for all $\lambda \le 1/\left(L_f+\alpha/\mu\right)$ and $n\in\mathbb{N}$, we have that

$$H_\mu\left(\mathbf{y}^n\right)-H_\mu\left(\mathbf{x}_{op}^*\right)\le\frac{2C_{\mathbf{x}_{op}^*}^2\left(J+2\right)}{\left(n+1\right)\left(1-\eta\right)\lambda},$$

where $C_{\mathbf{x}_{op}^*}=\max\left\{\left\|\mathbf{x}^0-\tilde{\mathbf{x}}\right\|,\frac{1}{1-\eta}\left\|\left(I-S\right)\tilde{\mathbf{x}}\right\|\right\}$, and $J=\left\lfloor 2/\left(1-\eta\right)\right\rfloor$.

Proof. From Proposition 2.1 we have, for any step-size $\lambda \le 1/\left(L_f+\alpha/\mu\right)$, that the following inequality that holds true,

$$H_\mu\left(\mathbf{y}^{n+1}\right)-H_\mu\left(\mathbf{x}_{op}^*\right)\le\frac{1}{\lambda}\left\langle\mathbf{x}^n-\mathbf{y}^{n+1},\mathbf{x}^n-\mathbf{x}_{op}^*\right\rangle-\frac{1}{2\lambda}\left\|\mathbf{x}^n-\mathbf{y}^{n+1}\right\|^2. \tag{3.20}$$

For $\mathbf{x}_{op}^*\in X^*=\mathrm{Fix}\left(T_\lambda\right)$, from Lemma 3.1(i) and Lemma 3.1, we obtain

$$\left\langle\mathbf{x}^n-\mathbf{y}^{n+1},\mathbf{x}^n-\mathbf{x}_{op}^*\right\rangle\le\left\|\mathbf{x}^n-\mathbf{y}^{n+1}\right\|\left\|\mathbf{x}^n-\mathbf{x}_{op}^*\right\|\le\frac{2C_{\mathbf{x}_{op}^*}^2\left(J+2\right)}{\left(1-\eta\right)\left(n+1\right)}. \tag{3.21}$$

Substituting (3.21) into (3.20), we get

$$H_\mu\left(\mathbf{y}^{n+1}\right)-H_\mu\left(\mathbf{x}_{op}^*\right)\le\frac{2C_{\mathbf{x}_{op}^*}^2\left(J+2\right)}{\left(n+1\right)\left(1-\eta\right)\lambda}. \tag{3.22}$$

We complete the proof.

□

Now, we will show the complexity result of BiG-SAM algorithm.

Theorem 3.2. Suppose that **Assumption I** holds. Let $\varepsilon\in\left(0,\bar{\varepsilon}\right)$ for some fixed $\bar{\varepsilon}>0$. Let $\left\{\mathbf{x}^n\right\}_{n\in\mathbb{N}}, \left\{\mathbf{y}^n\right\}_{n\in\mathbb{N}}$ and $\left\{\mathbf{z}^n\right\}_{n\in\mathbb{N}}$ be generated by BiG-SAM algorithm with smoothing parameter

$$\mu=\sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}}.$$

Then for any n satisfying

$$n\ge\frac{2\Gamma^2\left(J+2\right)\left(2\sqrt{\alpha\beta}+\sqrt{L_f\varepsilon}\right)^2}{\varepsilon^2\left(1-\eta\right)},$$

where $\Gamma=\max\left\{\left\|R_\delta-\mathbf{x}^0\right\|,\frac{1}{1-\eta}\left\|\left(I-S\right)R_\delta\right\|\right\}$, and $J=\left\lfloor 2/\left(1-\eta\right)\right\rfloor$, it holds that $H\left(\mathbf{y}^n\right)-H_{opt}\le\varepsilon$.

Proof. Using the $\frac{1}{\mu}$-smooth approximation property of h_μ with parameters $\left(\alpha,\beta\right)$, it follows that for any $\mathbf{y}\in\mathbb{R}^n$,

$$H_\mu(\mathbf{y}) \le H(\mathbf{y}) \le H_\mu(\mathbf{y}) + \beta\mu. \tag{3.23}$$

In particular, the following two inequalities hold:

$$H_{\text{opt}} \ge H_\mu\left(\mathbf{x}_{op}^*\right) \text{ and } H\left(\mathbf{y}^n\right) \le H_\mu\left(\mathbf{y}^n\right) + \beta\mu, n = 0,1,\cdots. \tag{3.24}$$

In which, combined with (3.22) and $\lambda = 1/\left(L_f + \alpha/\mu\right)$, it yields

$$\begin{aligned} H\left(\mathbf{y}^n\right) - H_{\text{opt}} &\le H_\mu\left(\mathbf{y}^n\right) + \beta\mu - H_\mu\left(\mathbf{x}_{op}^*\right) \\ &\le \frac{2C_{\mathbf{x}_{op}^*}^2 (J+2)}{(n+1)(1-\eta)\lambda} + \beta\mu \\ &= 2L_f \frac{C_{\mathbf{x}_{op}^*}^2 (J+2)}{(n+1)(1-\eta)} + \left(\frac{2\alpha C_{\mathbf{x}_{op}^*}^2 (J+2)}{(n+1)(1-\eta)}\right)\frac{1}{\mu} + \beta\mu \\ &\le 2L_f \frac{C_{\mathbf{x}_{op}^*}^2 (J+2)}{n(1-\eta)} + \left(\frac{2\alpha C_{\mathbf{x}_{op}^*}^2 (J+2)}{n(1-\eta)}\right)\frac{1}{\mu} + \beta\mu, \end{aligned}$$

where $C_{\mathbf{x}_{op}^*} := \max\left\{\left\|\mathbf{x}^0 - \tilde{\mathbf{x}}\right\|, \frac{1}{1-\eta}\left\|(I-S)\tilde{\mathbf{x}}\right\|\right\}$, and $J = \lfloor 2/(1-\eta) \rfloor$. Therefore, for a given $N > 0$, it holds that for any $n \ge N$,

$$H\left(\mathbf{y}^n\right) - H_{\text{opt}} \le 2L_f \frac{C_{\mathbf{x}_{op}^*}^2 (J+2)}{N(1-\eta)} + \left(\frac{2\alpha C_{\mathbf{x}_{op}^*}^2 (J+2)}{N(1-\eta)}\right)\frac{1}{\mu} + \beta\mu. \tag{3.25}$$

Minimizing the right-hand side w.r.t. μ, we obtain

$$\mu = \sqrt{\frac{2\alpha C_{\mathbf{x}_{op}^*}^2 (J+2)}{N\beta(1-\eta)}}. \tag{3.26}$$

Plugging (3.26) into (3.25), it implies that for any $n \ge N$,

$$H\left(\mathbf{y}^n\right) - H_{\text{opt}} \le 2L_f \frac{C_{\mathbf{x}_{op}^*}^2 (J+2)}{N(1-\eta)} + 2\sqrt{\frac{2\alpha\beta C_{\mathbf{x}_{op}^*}^2 (J+2)}{N(1-\eta)}}.$$

Thus, to make sure that $\mathbf{y}^n$ is an ε-optimal solution for any $n \ge N$, it is enough that N will satisfy

$$2L_f \frac{C_{\mathbf{x}_{op}^*}^2 (J+2)}{N(1-\eta)} + 2\sqrt{\frac{2\alpha\beta C_{\mathbf{x}_{op}^*}^2 (J+2)}{N(1-\eta)}} \le \varepsilon.$$

Setting $\tau^2 = \frac{2C_{\mathbf{x}_{op}^*}^2 (J+2)}{N(1-\eta)}$, the above inequality reduces to

$$L_f\tau^2 + 2\sqrt{\alpha\beta}\tau - \varepsilon \le 0,$$

which, by the fact that $\tau > 0$, is equivalent to

$$\sqrt{\frac{2C_{\mathbf{x}_{op}^*}^2 (J+2)}{N(1-\eta)}} = \tau \le \frac{-\sqrt{\alpha\beta} + \sqrt{\alpha\beta + L_f\varepsilon}}{L_f} = \frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta + L_f\varepsilon}}.$$

We conclude that N should satisfy

$$N \geq \frac{\left(2C_{\mathbf{x}_{op}^*}^2 (J+2)\right)\left(\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}\right)^2}{\varepsilon^2(1-\eta)}.$$

In particular, if we choose

$$N = N_1 \equiv \frac{\left(2C_{\mathbf{x}_{op}^*}^2 (J+2)\right)\left(\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}\right)^2}{\varepsilon^2(1-\eta)},$$

and μ according to (3.26), meaning that

$$\mu = \sqrt{\frac{2\alpha C_{\mathbf{x}_{op}^*}^2 (J+2)}{N_1\beta(1-\eta)}} = \sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}},$$

then for any $n \geq N_1$, it holds that $H\left(\mathbf{y}^n\right) - H_{\text{opt}} \leq \varepsilon$. By (3.23) and (3.24),

$$H\left(\mathbf{x}_{op}^*\right) - \beta\mu \leq H_\mu\left(\mathbf{x}_{op}^*\right) \leq H_{\text{opt}} \leq H\left(\mathbf{y}^0\right),$$

which along with the inequality

$$\mu = \sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}} \leq \sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta}+\sqrt{\alpha\beta}} \leq \frac{\bar{\varepsilon}}{2\beta},$$

implies that $H\left(\mathbf{x}_{op}^*\right) \leq H\left(\mathbf{y}^0\right) + \frac{\bar{\varepsilon}}{2}$, and hence, by Assumption **I** (iv), it follows that $\|\tilde{\mathbf{x}}\| \leq R_\delta$, where $\delta := H\left(\mathbf{y}^0\right) + \frac{\bar{\epsilon}}{2}$. Therefore,

$C_{\mathbf{x}_{op}^*} \leq \max\left\{\left\|R_\delta - \mathbf{x}^0\right\|, \frac{1}{1-\eta}\left\|(I-S)R_\delta\right\|\right\} = \Gamma$. Consequently,

$$\begin{aligned} N_1 &= \frac{\left(2C_{\mathbf{x}_{op}^*}^2 (J+2)\right)\left(\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}\right)^2}{\varepsilon^2(1-\eta)} \\ &\leq \frac{\left(2C_{\mathbf{x}_{op}^*}^2 (J+2)\right)\left(2\sqrt{\alpha\beta}+\sqrt{L_f\varepsilon}\right)^2}{\varepsilon^2(1-\eta)} \\ &\leq \frac{2\Gamma^2 (J+2)\left(2\sqrt{\alpha\beta}+\sqrt{L_f\varepsilon}\right)^2}{\varepsilon^2(1-\eta)} \\ &\equiv N_2. \end{aligned}$$

The second inequality follows from the fact that for any $\gamma, \delta \geq 0$, it holds that $\sqrt{\gamma+\delta} \leq \sqrt{\gamma}+\sqrt{\delta}$. Consequently, for any $n \geq N_2$, we have $H\left(\mathbf{y}^n\right) - H_{\text{opt}} \leq \varepsilon$, thus establishing the desired result.

□

4. BiG-SAM for Nonsmooth Bi-level Optimization Problems

In this section, we adopt the problem (OP) described in Section 2, where the

objective function ω does not necessarily satisfy Assumption **II**, which satisfies the **Assumption III**.

Note that, BiG-SAM cannot be directly applied to bi-level problems with **Assumption III**. However, we can handle this case indirectly. From the strong convexity of ω, we can smooth ω by the Moreau envelope $M_{s\omega}$. Recall the properties of Moreau envelope in Section 2, $M_{s\omega}$ is continuously differentiable, with a $1/s$-Lipschitz continuous gradient, $1/s>0$, and is strongly convex (see Lemma 2.3). Thus, $M_{s\omega}$ satisfies Assumption **II**, it maked BiG-SAM algorithm applicable. In this case, step (3.7) can be simplified as follow:

$$\begin{aligned}\mathbf{z}^n &= \mathbf{x}^{n-1} - s\nabla M_{s\omega}\left(\mathbf{x}^{n-1}\right)\\ &= \mathbf{x}^{n-1} - s\frac{1}{s}\left(\mathbf{x}^{n-1} - \operatorname{prox}_{s\omega}\left(\mathbf{x}^{n-1}\right)\right)\\ &= \operatorname{prox}_{s\omega}\left(\mathbf{x}^{n-1}\right),\end{aligned} \tag{4.1}$$

where the second equality follows from (2.13). This implies that computing $\mathbf{z}^n$ ($n\in\mathbb{N}$) requires evaluating the proximal mapping of ω.

Remark 1. Note that the proximal mapping of a strongly convex function is a contraction ([12], Lemma 6), making the theory in Section 3.1 applicable here. A direct consequence of Lemma 3.2 applies to the mappings:

$$S\left(\mathbf{x}\right)=\mathbf{x}-s\nabla M_{s\omega}\left(\mathbf{x}\right) \text{ and } T\left(\mathbf{x}\right)=\operatorname{prox}_{\lambda g}\left(\mathbf{x}-\lambda\nabla F_{\mu}\left(\mathbf{x}\right)\right),$$

where $s>0$ and $\lambda\in\left(0,1/\left(L_f+\alpha/\mu\right)\right]$.

Lemma 4.1. [12] Let $\left\{\mathbf{x}^n\right\}_{n\in\mathbb{N}}$ be a sequence generated by BiG-SAM. Under Assumptions **I**, **III** and **IV**, for $s>0$, the sequence $\left\{\mathbf{x}^n\right\}_{n\in\mathbb{N}}$ converges to $\mathbf{x}_s^*\in X^*$, where $\mathbf{x}_s^*$ satisfies:

$$\left\langle\nabla M_{s\omega}\left(\mathbf{x}_s^*\right),\mathbf{x}-\mathbf{x}_s^*\right\rangle\geq 0,\ \forall\mathbf{x}\in X^*. \tag{4.2}$$

Thus, $\mathbf{x}_s^*$ is the optimal solution of the problem (OP) with respect to the Moreau envelope $M_{s\omega}$, i.e.,

$$\mathbf{x}_s^*=\arg\min_{\mathbf{x}\in X^*} M_{s\omega}\left(\mathbf{x}\right),$$

where X^* is the set of optimal solutions of problem (Q).

Smoothing the ω seems to not affect the convergence rate, which is based on the inner function. From the works in [12], we know that the convergence rate depends on the contraction parameter η. We have the following result from ([12], Lemma 6),

$$\eta=\frac{1}{1+s\sigma}.$$

Let $\delta>0$ be the required uniform accuracy in terms of the outer objective function, that is,

$$\omega\left(\mathbf{x}^n\right)-M_{s\omega}\left(\mathbf{x}^n\right)\leq\delta,\ \forall n\in\mathbb{N} \tag{4.3}$$

where it should be noted that $\omega(\mathbf{x}^n) - M_{s\omega}(\mathbf{x}^n) \geq 0$ for all $n \in \mathbb{N}$. Now, we aim to determine the number of iterations N' required to achieve an ε-optimal solution for the inner problem, that is,

$$H(\mathbf{y}^{N'}) - H_{opt} \leq \varepsilon,$$

while keeping the uniform accuracy as given in (4.3). This means that N' depends on both ε and δ.

Proposition 4.1. Let $\varepsilon \in (0, \bar{\varepsilon})$ for some fixed $\bar{\varepsilon} > 0$. Let $\{\mathbf{x}^n\}_{n\in\mathbb{N}}$ and $\{\mathbf{y}^n\}_{n\in\mathbb{N}}$ be a sequence generated by BiG-SAM and suppose that Assumptions **I**, **III** and **IV** hold true. In addition, suppose that the smoothing parameter is chosen by

$$s = \frac{2\delta}{\ell_\omega^2}$$

and

$$\mu = \sqrt{\frac{\alpha}{\beta}} \frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta + L_f \varepsilon}}.$$

Then, (4.3) holds true and for

$$n \geq \frac{2\Gamma^2 \left(2\sqrt{\alpha\beta} + \sqrt{L_f \varepsilon}\right)^2}{\varepsilon^2} \left(2 + \frac{3\ell_\omega^2}{2\sigma\delta} + \frac{\ell_\omega^4}{4\sigma^2\delta^2}\right),$$

where $\Gamma = \max\left\{\left\|R_\delta - \mathbf{x}^0\right\|, \frac{1}{1-\eta}\left\|(I-S)R_\delta\right\|\right\}$, it holds that $H(\mathbf{y}^n) - H_{opt} \leq \varepsilon$.

Proof. Since ω is ℓ_ω-Lipschitz continuous (see **Assumption III**) it follows that the norms of the subgradients of ω are bounded from above by ℓ_ω. Thus, from ([15], Lemma 4.2) it follows, for all $\mathbf{x} \in \mathbb{R}^n$, that

$$\omega(\mathbf{x}) - \frac{s\ell_\omega^2}{2} \leq M_{s\omega}(\mathbf{x}) \leq \omega(\mathbf{x})$$

Therefore, for $s = 2\delta/\ell_\omega^2$, we obtain that

$$\omega(\mathbf{x}^n) - M_{s\omega}(\mathbf{x}^n) \leq \delta, \ \forall n \in \mathbb{N}.$$

Using the $\frac{1}{\mu}$-smooth approximation property of h_μ with parameters (α, β), it follows that for any $\mathbf{y} \in \mathbb{R}^n$,

$$H_\mu(\mathbf{y}) \leq H(\mathbf{y}) \leq H_\mu(\mathbf{y}) + \beta\mu. \tag{4.4}$$

In particular, the following two inequalities hold:

$$H_{\text{opt}} \geq H_\mu(\mathbf{x}_{op}^*) \text{ and } H(\mathbf{y}^n) \leq H_\mu(\mathbf{y}^n) + \beta\mu, n = 0, 1, \cdots, \tag{4.5}$$

which, combined with (3.22), yields

$$\begin{aligned} H\left(\mathbf{y}^n\right)-H_{\text{opt}} &\le H_\mu\left(\mathbf{y}^n\right)+\beta\mu-H_\mu\left(\mathbf{x}_{op}^*\right) \\ &\le \frac{2C_{\mathbf{x}_{op}^*}^2(J+2)}{(n+1)(1-\eta)\lambda}+\beta\mu \\ &= 2L_f\frac{C_{\mathbf{x}_{op}^*}^2(J+2)}{(n+1)(1-\eta)}+\left(\frac{2\alpha C_{\mathbf{x}_{op}^*}^2(J+2)}{(n+1)(1-\eta)}\right)\frac{1}{\mu}+\beta\mu \\ &\le 2L_f\frac{C_{\mathbf{x}_{op}^*}^2(J+2)}{n(1-\eta)}+\left(\frac{2\alpha C_{\mathbf{x}_{op}^*}^2(J+2)}{n(1-\eta)}\right)\frac{1}{\mu}+\beta\mu, \end{aligned}$$

where $C_{\mathbf{x}_{op}^*} := \max\left\{\left\|\mathbf{x}^0-\tilde{\mathbf{x}}\right\|, \frac{1}{1-\eta}\left\|(I-S)\tilde{\mathbf{x}}\right\|\right\}$, and $J=\lfloor 2/(1-\eta)\rfloor$. Therefore, for a given $N'>0$, it holds that for any $n\ge N'$,

$$H\left(\mathbf{y}^n\right)-H_{\text{opt}} \le 2L_f\frac{C_{\mathbf{x}_{op}^*}^2(J+2)}{N'(1-\eta)}+\left(\frac{2\alpha C_{\mathbf{x}_{op}^*}^2(J+2)}{N'(1-\eta)}\right)\frac{1}{\mu}+\beta\mu. \tag{4.6}$$

Minimizing the right-hand side w.r.t. μ, we obtain

$$\mu=\sqrt{\frac{2\alpha C_{\mathbf{x}_{op}^*}^2(J+2)}{N'\beta(1-\eta)}}. \tag{4.7}$$

Plugging the above expression into (4.6), we conclude that for any $n\ge N'$,

$$H\left(\mathbf{y}^n\right)-H_{\text{opt}} \le 2L_f\frac{C_{\mathbf{x}_{op}^*}^2(J+2)}{N'(1-\eta)}+2\sqrt{\frac{2\alpha\beta C_{\mathbf{x}_{op}^*}^2(J+2)}{N'(1-\eta)}}.$$

Thus, to guarantee that $\mathbf{y}^n$ is an ε-optimal solution for any $n\ge N'$, it is enough that N' will satisfy

$$2L_f\frac{C_{\mathbf{x}_{op}^*}^2(J+2)}{N'(1-\eta)}+2\sqrt{\frac{2\alpha\beta C_{\mathbf{x}_{op}^*}^2(J+2)}{N'(1-\eta)}}\le\varepsilon.$$

Denoting $\tau^2=\frac{2C_{\mathbf{x}_{op}^*}^2(J+2)}{N'(1-\eta)}$, the above inequality reduces to

$$L_f\tau^2+2\sqrt{\alpha\beta}\tau-\varepsilon\le 0,$$

which, by the fact that $\tau>0$, is equivalent to

$$\sqrt{\frac{2C_{\mathbf{x}_{op}^*}^2(J+2)}{N'(1-\eta)}}=\tau\le\frac{-\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}}{L_f}=\frac{\varepsilon}{\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}}.$$

We conclude that N' should satisfy

$$N' \geq \frac{\left(2C_{\mathbf{x}_{op}^*}^2 (J+2)\right)\left(\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}\right)^2}{\varepsilon^2(1-\eta)}.$$

In particular, if we choose

$$N' = N_3 \equiv \frac{\left(2C_{\mathbf{x}_{op}^*}^2 (J+2)\right)\left(\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}\right)^2}{\varepsilon^2(1-\eta)} \tag{4.8}$$

Now, substituting $J = \lfloor 2/(1-\eta) \rfloor$, $\eta = 1/(1+s\sigma)$, and $s = 2\delta/\ell_\omega^2$ into equation (4.8), we obtain

$$\begin{aligned}
N' = N_3 &\equiv \frac{\left(2C_{\mathbf{x}_{op}^*}^2 (J+2)\right)\left(\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}\right)^2}{\varepsilon^2(1-\eta)} \\
&= \frac{\left(2C_{\mathbf{x}_{op}^*}^2 \left(\frac{2}{1-\eta}+2\right)\right)\left(\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}\right)^2}{\varepsilon^2(1-\eta)} \\
&= \frac{4C_{\mathbf{x}_{op}^*}^2 \left(\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}\right)^2 (2-\eta)}{\varepsilon^2(1-\eta)^2} \\
&= \frac{4C_{\mathbf{x}_{op}^*}^2 \left(\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}\right)^2}{\varepsilon}\left(2+\frac{3}{s\sigma}+\frac{1}{(s\sigma)^2}\right) \\
&= \frac{4C_{\mathbf{x}_{op}^*}^2 \left(\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}\right)^2}{\varepsilon^2}\left(2+\frac{3\ell_\omega^2}{2\sigma\delta}+\frac{\ell_\omega^4}{4\sigma^2\delta^2}\right)
\end{aligned}$$

and μ according to (4.7), meaning that

$$\mu = \sqrt{\frac{2\alpha C_{\mathbf{x}_{op}^*}^2 (J+2)}{N_3\beta(1-\eta)}} = \sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}},$$

then for any $n \geq N_3$ it holds that $H(\mathbf{y}^n) - H_{\text{opt}} \leq \varepsilon$. By (4.4) and (4.5),

$$H(\mathbf{x}_{op}^*) - \beta\mu \leq H_\mu(\mathbf{x}_{op}^*) \leq H_{\text{opt}} \leq H(\mathbf{y}^0),$$

which along with the inequality

$$\mu = \sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}} \leq \sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta}+\sqrt{\alpha\beta}} \leq \frac{\overline{\varepsilon}}{2\beta},$$

implies that $H(\mathbf{x}_{op}^*) \leq H(\mathbf{y}^0) + \frac{\overline{\varepsilon}}{2}$, and hence, by Assumption **I** (iv), it follows that $\tilde{\mathbf{x}} \leq R_\delta$, where $\delta := H(\mathbf{y}^0) + \frac{\overline{\epsilon}}{2}$. Therefore,

$C_{\mathbf{x}_{op}^*} \le \max\left\{\left\|R_\delta - \mathbf{x}^0\right\|, \frac{1}{1-\eta}\left\|(I-S)R_\delta\right\|\right\} = \Gamma$. Consequently,

$$\begin{aligned} N_3 &= \frac{4C_{\mathbf{x}_{op}^*}^2\left(\sqrt{\alpha\beta}+\sqrt{\alpha\beta+L_f\varepsilon}\right)^2}{\varepsilon^2}\left(2+\frac{3\ell_\omega^2}{2\sigma\delta}+\frac{\ell_\omega^4}{4\sigma^2\delta^2}\right) \\ &\le \frac{4C_{\mathbf{x}_{op}^*}^2\left(2\sqrt{\alpha\beta}+\sqrt{L_f\varepsilon}\right)^2}{\varepsilon^2}\left(2+\frac{3\ell_\omega^2}{2\sigma\delta}+\frac{\ell_\omega^4}{4\sigma^2\delta^2}\right) \\ &\le \frac{2\Gamma^2\left(2\sqrt{\alpha\beta}+\sqrt{L_f\varepsilon}\right)^2}{\varepsilon^2}\left(2+\frac{3\ell_\omega^2}{2\sigma\delta}+\frac{\ell_\omega^4}{4\sigma^2\delta^2}\right) \\ &\equiv N_4. \end{aligned}$$

The second inequality follows from the fact that for any $\gamma, \delta \ge 0$, it holds that $\sqrt{\gamma+\delta} \le \sqrt{\gamma}+\sqrt{\delta}$. The desired result is achieved by selecting n as the upper bound derived above.

□

5. Conclusion

In this paper, we construct a novel bi-level gradient sequential averaging method (BiG-SAM) for solving a more composite convex bi-level optimization problem, where the inner level problem is to find the optimal solution of the sum of three functions, including two non-smooth function and one smoothable. We analyze the convergence rate of the BiG-SAM in two different cases, where the outer objective is smooth or non-smooth, the global convergence rate with respected to inner objective function is $\mathcal{O}(1/n)$. In the future, we could further explore the convergence rate and complexity analysis of the outer objective function. Additionally, we could design stochastic and parallel variants of BiG-SAM for high-dimensional data or distributed scenarios. This would help reduce computational complexity while ensuring convergence and scalability of both the outer and inner objectives in distributed environments.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Dempe, S., Dinh, N. and Dutta, J. (2010) Optimality Conditions for a Simple Convex Bilevel Programming Problem. In: Burachik, R. and Yao, J.C., Eds., *Variational Analysis and Generalized Differentiation in Optimization and Control*, Springer, 149-161. https://doi.org/10.1007/978-1-4419-0437-9_7

[2] Dempe, S. (2002) Foundations of Bilevel Programming. Springer Science & Business Media.

[3] Tikhonov, A.N. and Arsenin, V.I. (1977) Solutions of Ill-Posed Problems. Winston & Sons.

[4] Mangasarian, O.L. and Meyer, R.R. (1979) Nonlinear Perturbation of Linear Pro-

grams. *SIAM Journal on Control and Optimization*, **17**, 745-752. https://doi.org/10.1137/0317052

[5] Ferris, M.C. and Mangasarian, O.L. (1991) Finite Perturbation of Convex Programs. *Applied Mathematics & Optimization*, **23**, 263-273. https://doi.org/10.1007/bf01442401

[6] Solodov, M. (2007) An Explicit Descent Method for Bilevel Convex Optimization. *Journal of Convex Analysis*, **14**, 227-237.

[7] Cabot, A. (2005) Proximal Point Algorithm Controlled by a Slowly Vanishing Term: Applications to Hierarchical Minimization. *SIAM Journal on Optimization*, **15**, 555-572. https://doi.org/10.1137/s105262340343467x

[8] Boţ, R.I. and Nguyen, D. (2018) A Forward-Backward Penalty Scheme with Inertial Effects for Monotone Inclusions. Applications to Convex Bilevel Programming. *Optimization*, **68**, 1855-1880. https://doi.org/10.1080/02331934.2018.1556662

[9] Malitsky, Y. (2017). Chambolle-Pock and Tseng's Methods: Relationship and Extension to the Bilevel Optimization. https://optimization-online.org/2017/06/6103/

[10] Yamada, I., Yukawa, M. and Yamagishi, M. (2011) Minimizing the Moreau Envelope of Nonsmooth Convex Functions over the Fixed Point Set of Certain Quasi-Nonexpansive Mappings. In: Bauschke, H., Burachik, R., Combettes, P., Elser, V., Luke, D. and Wolkowicz, H., Eds., *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, Springer, 345-390. https://doi.org/10.1007/978-1-4419-9569-8_17

[11] Beck, A. and Sabach, S. (2013) A First Order Method for Finding Minimal Norm-Like Solutions of Convex Optimization Problems. *Mathematical Programming*, **147**, 25-46. https://doi.org/10.1007/s10107-013-0708-2

[12] Sabach, S. and Shtern, S. (2017) A First Order Method for Solving Convex Bilevel Optimization Problems. *SIAM Journal on Optimization*, **27**, 640-660. https://doi.org/10.1137/16m105592x

[13] Xu, H. (2004) Viscosity Approximation Methods for Nonexpansive Mappings. *Journal of Mathematical Analysis and Applications*, **298**, 279-291. https://doi.org/10.1016/j.jmaa.2004.04.059

[14] Shehu, Y., Vuong, P.T. and Zemkoho, A. (2019) An Inertial Extrapolation Method for Convex Simple Bilevel Optimization. *Optimization Methods and Software*, **36**, 1-19. https://doi.org/10.1080/10556788.2019.1619729

[15] Beck, A. and Teboulle, M. (2012) Smoothing and First Order Methods: A Unified Framework. *SIAM Journal on Optimization*, **22**, 557-580. https://doi.org/10.1137/100818327

[16] Shor, N.Z. (1985) Minimization Methods for Nondifferentiable Functions. Springer-Verlag.

[17] Ben-Tal, A. and Teboulle, M. (1989) A Smoothing Technique for Nondifferentiable Optimization Problems. In: Dolecki, S., Ed., *Optimization*, Springer, 1-11. https://doi.org/10.1007/bfb0083582

[18] Bertsekas, D.P. (1975) Nondifferentiable Optimization via Approximation. In: Balinski, M.L. and Wolfe, P., Eds., *Nondifferentiable Optimization*, Springer, 1-25. https://doi.org/10.1007/bfb0120696

[19] Bertsekas, D.P. (1982) Constrained Optimization and Lagrange Multiplier Methods. Academic Press.

[20] Beck, A. and Teboulle, M. (2009) A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, **2**, 183-202. https://doi.org/10.1137/080716542

[21] Nesterov, Y. (2012) Gradient Methods for Minimizing Composite Functions. *Mathematical Programming*, **140**, 125-161. https://doi.org/10.1007/s10107-012-0629-5

[22] Beck, A. and Teboulle, M. (2010) Gradient-Based Algorithms with Applications to Signal-Recovery Problems. *Journal of Convex Analysis*, **17**, 445-477.

[23] Combettes, P.L. and Pesquet, J. (2011) Proximal Splitting Methods in Signal Processing. In: Bauschke, H., Burachik, R., Combettes, P., Elser, V., Luke, D. and Wolkowicz, H., Ed., *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, Springer, 185-212. https://doi.org/10.1007/978-1-4419-9569-8_10

[24] Bauschke, H.H. and Combettes, P.L. (2019) Correction To: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. In: Bauschke, H.H. and Combettes, P.L., Eds., *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, C1-C4. https://doi.org/10.1007/978-3-319-48311-5_31